T.C.

ONDOKUZ MAYIS ÜNİVERSİTESİ

SAĞLIK BİLİMLERİ ENSTİTÜSÜ

TIBBİ BİYOLOJİ

ANABİLİM DALI

DİZİSİ BİLİNEN PROTEİNLERİN 3D BENZETİM MODELLERİNİN KURULMASI VE MODELİN ETKİLİ SNP'LERİN SAPTANMASINDA KULLANILMASI

DOKTORA TEZİ

Muhammed Kamil TURAN

Samsun

Şubat-2011

T.C.

ONDOKUZ MAYIS ÜNİVERSİTESİ

SAĞLIK BİLİMLERİ ENSTİTÜSÜ

TIBBİ BİYOLOJİ

ANABİLİM DALI

DİZİSİ BİLİNEN PROTEİNLERİN 3D BENZETİM MODELLERİNİN KURULMASI VE MODELİN ETKİLİ SNP'LERİN SAPTANMASINDA KULLANILMASI

DOKTORA TEZİ

Muhammed Kamil TURAN

Danışman: Prof. Dr. Hasan BAĞCI

Samsun

Şubat-2011

T.C.
ONDOKUZ MAYIS ÜNİVERSİTESİ
SAĞLIK BİLİMLERİ ENSTİTÜSÜ

Bu çalışma jürimiz tarafından Tıbbi Biyoloji Programında doktora tezi olarak kabul edilmiştir.

Başkan : Prof. Dr. Kazım ÖZDAMAR (Osmangazi Üniversitesi Tıp Fakültesi)

Üye : Prof. Dr. Prof. Dr. Hasan BAĞCI (Ondokuz Mayıs Üniversitesi Tıp Fakültesi)

Üye : Doç. Dr. Nurten KARA (Ondokuz Mayıs Üniversitesi Tıp Fakültesi)

Üye : Yrd. Doç. Dr. Sezgin GÜNEŞ (Ondokuz Mayıs Üniversitesi Tıp Fakültesi)

Üye : Yrd. Doç. Dr. Sedat DOĞAN (Ondokuz Mayıs Üniversitesi Mühendislik Fakültesi)

Bu tez, Enstitü Yönetim Kurul'unca belirlenen yukarıdaki jüri üyeleri tarafından uygun görülmüştür.

Prof.Dr.Süleyman KAPLAN
Enstitü Müdürü

İÇİNDEKİLER

TEŞEKKÜR

Ondokuz Mayıs Üniversitesi (OMÜ), Tıp Fakültesi Tıbbi Biyoloji Anabilim Dalı'nda doktora eğitimim ve tez çalışmam sırasında bana her türlü desteği veren değerli danışmanım Tıbbi Biyoloji Anabilim Dalı Başkanı Öğretim üyesi Prof. Dr. Hasan BAĞCI'ya öncelikle teşekkür ederim.

Doktora eğitimim ve tez çalışmalarımın tüm aşamalarımda bana her konuda büyük destek veren ve büyük emeği olan OMU Harita Mühendisliği AD öğretim üyesi Yrd. Doc. Dr. Sedat DOĞAN 'a ve Rukiye DOĞAN teşekkür ederim.

Doktora eğitimim sırasında bilgilerini esirgemeyen OMU Tıp Fakültesi Tıbbi Biyoloji AD öğretim üyeleri Prof. Dr. Gülsen ÖKTEN'e, Prof. Dr. Mehmet ELBİSTAN'a ,Doç. Dr. Nurten KARA'ya ve Yrd. Doç. Dr. Sezgin GÜNEŞ 'e teşekkür ederim.

Tez çalışmamda her zaman gece ve gündüz beni destekleyen, arkadaşım, dostum ve büyük dert ortağım Dr. Emre TAŞKIN 'a ve Zeynep YEĞİN 'e ile Yavuz AÇIKGÖZ 'e teşekkür ederim.

Değerli çalışma arkadaşlarım Dr. Kenan KOŞAR, Dr. Dilara GÖK, Dr. Hayri Fatih METİNYURT, Metehan ÇAMOĞLU ile Sağlık Grup Başkanımız olan Uzm.Dr. Ali ÇOŞKUN 'a manevi destekleri için teşekkür ederim.

Değerli annem Sevim TURAN 'a ve babam Ahmet TURAN ile eşim Sema TURAN ve 'baba ders' diyerek büyüyen oğlum Ahmet Mert TURAN (BabülSan) 'a, kuzenim Ali Alper AK ile Teyzem Hatice AK ve değerli eşi Prof. Dr. İsmail Ak 'a teşekkür ederim.

1. GİRİŞ

Proteinler organizmanın temel yapı taşlarıdır. Organizmada pek çok hayati görevde rol alırlar. Reaksiyonları başlatmak, bitirmek, hızlarını düzenlemek, kasların haraket etmesini sağlamak, nöronal iletiyi sağlamak, oksijen taşımak v.b. Proteinlerin yapılarını tahmin etmek ve üç boyutlu bir şekilde onları görsel hale çevirmek biyoinformatiğin büyük problemlerinden bir tanesidir. Proteinlerin üç boyutlu benzetim modellerini kurmak, onların fonksiyonlarını tahmin etmekte, mutasyon ya da 'single nucleotide polymorphism' (SNP; tek nükleotid polimorfizmi) etkisiyle ne tip değişiklikler olduğunu anlamakta bize yardımcı olur. Proteinlerin üç boyutlu benzetim modellerini oluşturmak çok büyük bir sorundur. Büyük yatırımlar, geniş çalışma grupları ve multidisipliner yaklaşımlar gerektirir. Biyoinformatikçilerin çözümü en güç sorularından birisidir.

Proteinlerin üç boyutlu yapısını tahmin etmek için deneysel ve hesaplamalı biyoloji yaklaşımları kullanılır. Deneysel yaklaşımlar iki tanedir. Birincisi, X-ışını Kırınım Spektrometresi (X-Ray Diffraction Spectrometre), ikincisi ise Çekirdek Manyetik Titreşim Spektroskopi (Nucleer Magnetic Resonance Spectroscopy, NMR) 'dır. Her iki yöntem de pahalı olmaları, çalışılmasının zor olması, ileri derecede hesap yükü getirmesi ve uzun dizilere uygulama problemlerine sahip olduklarından insan genom projesi gibi büyük projelerde çok etkili olamamışlardır (Arjuvan ve ark., 2001).

Hesaplamalı biyoloji yaklaşımları ise iki büyük model ile çözüme ulaşmaya çalışır. Bunlardan ilki fiziksel gerçekliği simüle edip kuantum fiziği ve kuantum matematiği notasyonlarını kullanmak; ikincisi ise yapay sinir ağı, gentik algortimalar, istatistiki yaklaşım modelleri gibi benzetimler ile ikincil yapıyı tahmin etmektir. Görüldüğü üzere her iki yol da, ciddi bilgisayar desteği gerektirmektedir.

Yapay sinir ağları kullanılarak yapılan tahminlere bahsedilen hesaplama yükünün yanına bir de yapay sinir ağı modellerinin ihtiyaç duyduğu verileri toplama sorunu eklenir. Biyolojik bilginin toplanması, organize edilmesi, her an güncel ve kullanıma hazır bir şekilde tutulması biyoinformatikçilerin ikinci büyük sorunudur. Çünkü genetik bilgi uzayı artık kontrol edilemez derecede büyümüştür. 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). Saatler içinde yapısı değişen, eklemeler ve güncellemeler yapılan veri tabanları her an değişen sunum formatları, referans numaraları, notasyonlar, doğal dil kullanımı gibi pek çok aşılması gereken sorun bulunmaktadır (Cochrane ve ark., 2009).

Bu tezin amacı;

1) Proteinlerin üç boyutlu benzetim modellerini kurmak,
2) Onların fonksiyonlarını tahmin etmek,
3) Mutasyon ya da SNP etkisiyle ne tip değişiklikler olduğunu anlamak,
4) Yukarıda üç maddede belirtilen amaçları gerçekleştirmek için kullanılacak, dağılım kurgulamalarını, varsayımları ve kısıtları açıklamak, benzetimde kullanılan matematiksel modelleri açıklamak, veri türetimi ve analizinde kullanılan algoritması özgün olarak belirlenmiş ve açıklamaları Türkçe, çıktıları Türkçe olan bilgisayar programları geliştirmek ve veri analizlerinde uygulamak,
5) Bundan sonra bu alanda çalışacak araştırıcılara özgün programlar sunmaktır.

2. GENEL BİLGİLER

2.1 Amino asitler

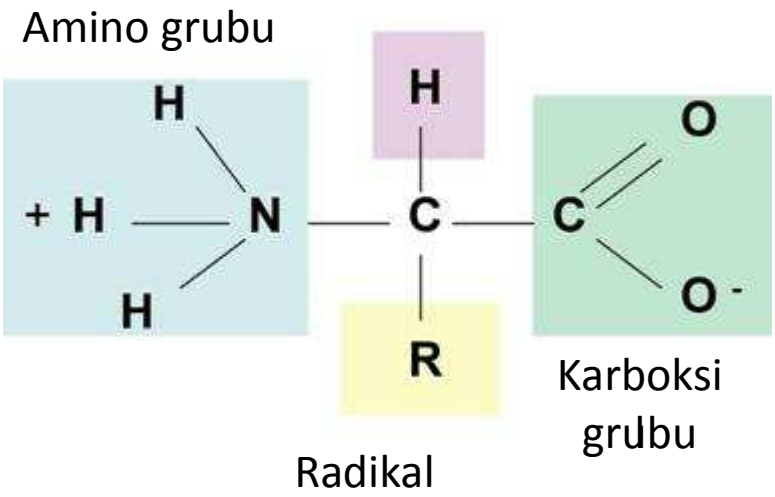
Proteinler, amino asitlerin zincirler halinde birleşmesinden oluşmuş organik polimer moleküllerdir. Proteinler yaşamın kaynağını oluştururlar. Organizmanın yaşamı için gerekli tüm kimyasal süreçlerin hemen her basamağında proteinler düzenleyici, başlatıcı ya da sonlandırıcı görevini üstlenirler. Yağlar ve karbonhidratlar gibi özel olarak depo edilmezler. Açlık durumunda enerji kaynağı olabilmeleri için öncelikle organizmanın hemen tüm karbonhidrat ve yağ depolarının bitmiş olması gerekir. Kimyasal sindirimleri midede başlar.

Proteinler, amino asitlerden oluşmuşlardır. Proteinler, amino asitlerin birbiri ardına bağlar yaparak sıralanması neticesinde oluşan polimer yapılardır. Proteinlerin yapısına giren 20 çeşit amino asit tanımlanmıştır. Bu yirmi çeşit amino asit, protein çeşitliliğinden dolaylı olarak hayatın çeşitliliğinden de sorumludur.

2.2. Amino asitler ve özellikleri

Amino asitler, proteinlerin yapıtaşlarıdır. Amino asit, yapısında bir karboksil grubu ve bir amino grubu bulundurduğundan amino asit olarak adlandırılmıştır. İlk bulunan amino asit 19 yüzyılın erken dönemlerinde Louis-Nicolas Vauquelin ve Pierre Jean Robiquet tarafından *Asparagus officinalis* bitkisinden izole edilmiş olan asparajindir. Daha sonra sırası ile sistein, glisin ve lösin izole edilmiştir.

Genel olarak bakıldığında her amino asit merkezcil bir karbon atomu etrafında sıralanmış bir amino grubu, bir karboksil grubu, bir radikal grup ile bir protondan (hidrojen atomu) oluşmuştur. Amino asitlerde çeşitliliğin nedeni yapıdaki radikal gruptur. Merkezcil olan karbon atomu α-karbonu olarak adlandırılır. Bu genel yapı Şekil 1 ‘de gösterilmiştir.



Şekil 1: Genel olarak bir amino asit

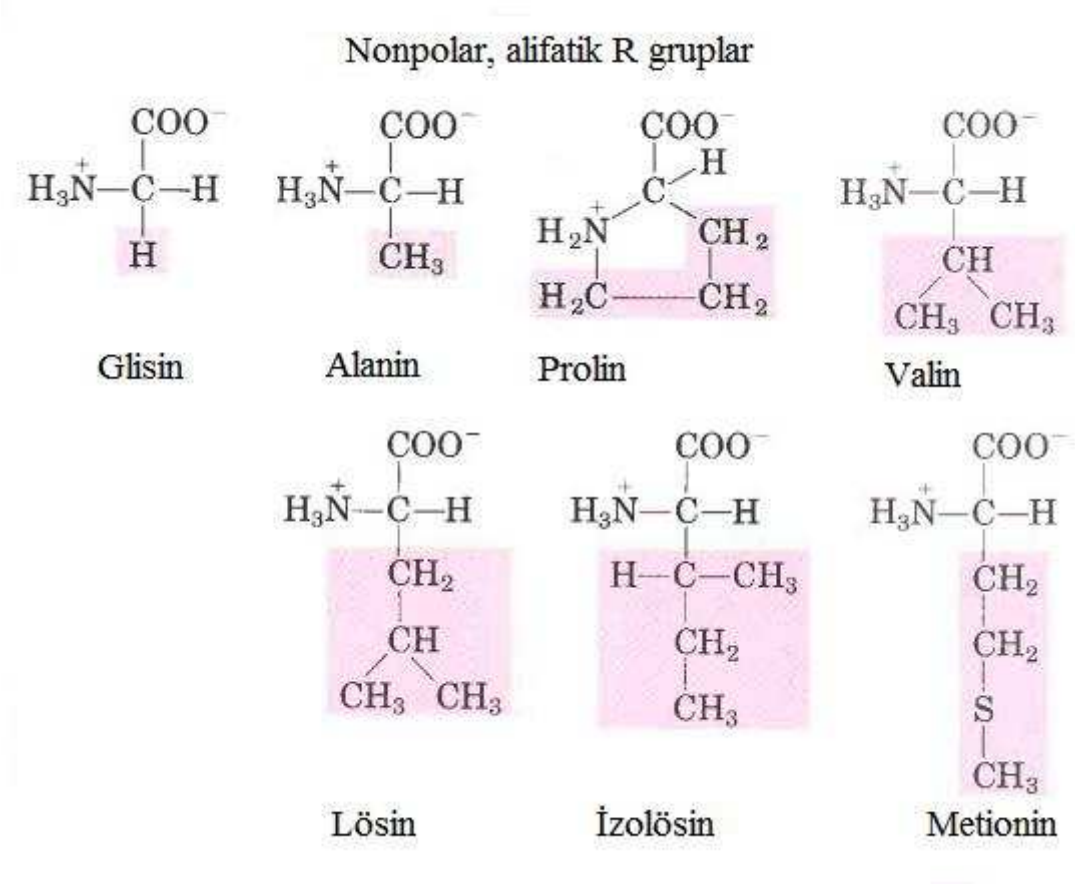
Burada merkezcil karbon atomuna α-karbon atomu da denir. Radikal grup ‘R’ ile gösterilmiştir. Amino asitlerin bu merkezcil karbon atomuna göre optik izomerleri bulunur, bu özellik steroizomeri olarak adlandırılır. Steroizomeriye göre oluşan moleküllere ise enantiomer adı verilmektedir. İnsan vücudundaki proteinlerin yapı taşı olan tüm amino asitler steroizomeri bakımından L enantiomeridir (L enantiomerinin izomeri ise D enantiomer olarak adlandırılır). Tüm amino asit yapısı bu karbon atomuna göre isimlendirilir. Zira merkezcil karbon atomu eğer α-karbonu olarak adlandırılırsa radikal grup karbon atomları β, Ω, γ, δ, ε ve sırasıyla, eğer 1. karbon atomu olarak kabul edilirse R grubu karbon atomları 2,3,4,… sırası ile isimlendirilirler (Davies ¸ 2007).

Amino asitler sahip oldukları R gruplarına göre isimlendirilmiş ve gruplandırılmıştır. İsimlendirme de üç harf metodu (bu metotta amino asitler üçlü harf grupları şeklinde ifade edilirler Met, Lys, Leu vb. gibi) ya da tek harf metodu (bu metotta amino asitler bir harf şeklinde ifade edilirler M, K, L vb. gibi) kullanılmıştır. Bu tez boyunca tek harf metodu gereğince isimlendirme yapılacaktır. Benzer şekilde amino asitlerin fizikokimyasal özellikleri R gruplarından gelir. R gruplarına göre protein yapısına

giren 20 çeşit amino asit bulunur. 20 amino asitin 3-harf, 1-harf kodlamaları, molekül ağırlıkları (dalton cinsinden), hidrofobisite indeksleri ve sahip oldukları R gruplarına göre fizikokimyasal özelliklerine göre dahil oldukları gruplar Tablo 1 ‘de verilmiştir. Amino asitlerin R gruplarına göre kimyasal gösterimleri Şekil 2a, 2b, 2c, 2d, 2e ‘de gösterilmiştir (Dixon ve ark., 1984; Nelson ve Cox, 2008).

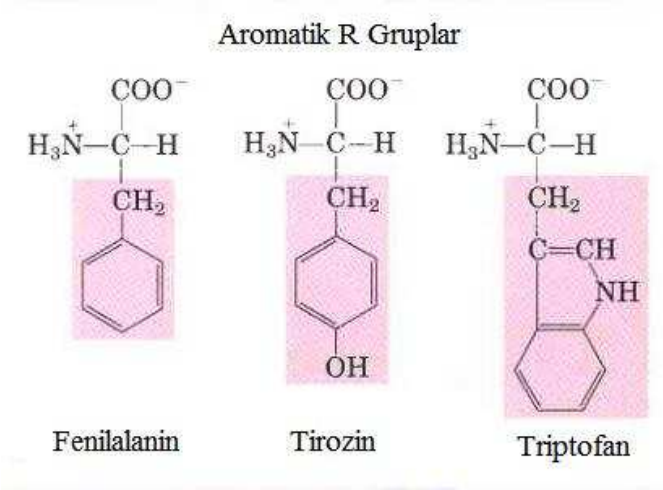
Tablo 1: Amino asitlerin değişik özelliklerinin gösterimleri

Amino asit	**3-Harf**	**1-Harf**	**MolekülAğırlığı**	**Hidrofobisite**	**R Grup Özelliği**
Glisin	Gly	G	75,0666	-0,4	Nonpolar, alifatik
Alanin	Ala	A	89,0932	1,8	Nonpolar, alifatik
Prolin	Pro	P	115,1305	-1,6	Nonpolar, alifatik
Valin	Val	V	117,1463	4,2	Nonpolar, alifatik
Lösin	Leu	L	131,1729	3,8	Nonpolar, alifatik
İzolösin	Ile	I	131,1729	4,5	Nonpolar, alifatik
Metionin	Met	M	149,2113	1,9	Nonpolar, alifatik
Fenilalanin	Phe	F	165,1891	2,8	Aromatik grup
Tirozin	Tyr	Y	181,1885	-1,3	Aromatik grup
Triptofan	Trp	W	204,2252	-0,9	Aromatik grup
Serin	Ser	S	105,0926	-0,8	Polar, yüksüz
Treonin	Thr	T	119,1192	-0,7	Polar, yüksüz
Sistein	Cys	C	121,1582	2,5	Polar, yüksüz
Asparajin	Asn	N	132,1179	-3,5	Polar, yüksüz
Glutamin	Gln	Q	146,1445	-3,5	Polar, yüksüz
Lizin	Lys	K	146,1876	-3,9	Pozitif yüklü
Histidin	His	H	155,1546	-3,2	Pozitif yüklü
Arjinin	Arg	R	174,201	-4,5	Pozitif yüklü
Asparttat	Asp	D	133,1027	-3,5	Negatif yüklü
Glutamat	Glu	E	147,1293	-3,5	Negatif yüklü



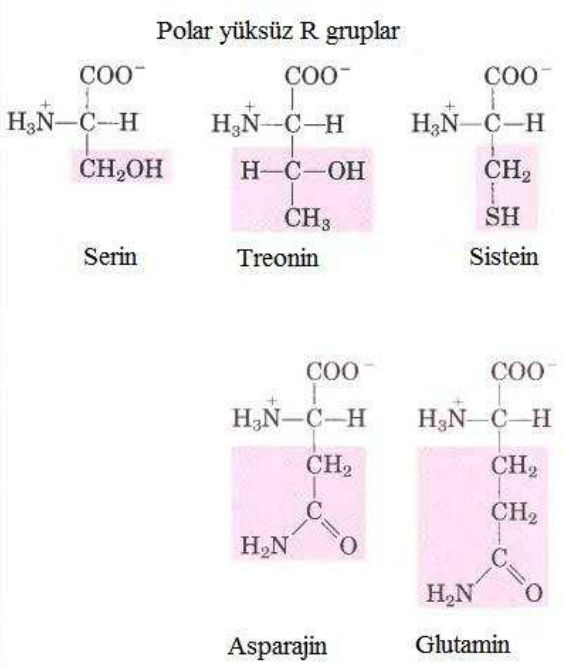
Şekil 2a: Nonpolar alifatik R gruba sahip amino asitler

(http://web2.tmu.edu.tw/m110093011/DNA2protein.htm'den uyarlanarak alınmıştır) .



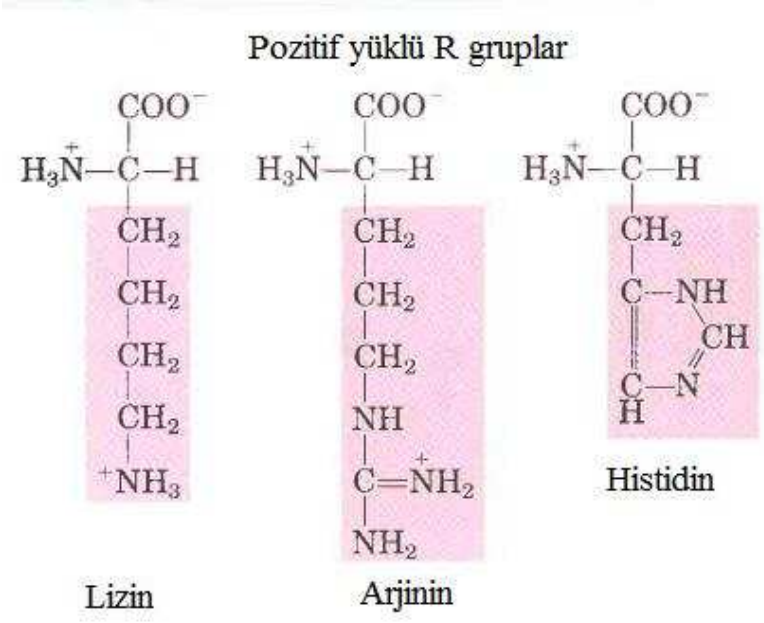
Şekil 2b: Aromatik R gruba sahip amino asitler

(http://web2.tmu.edu.tw/m110093011/DNA2protein.htm'den uyarlanarak alınmıştır) .



Şekil 2c: Polar yüksüz R gruba sahip amino asitler

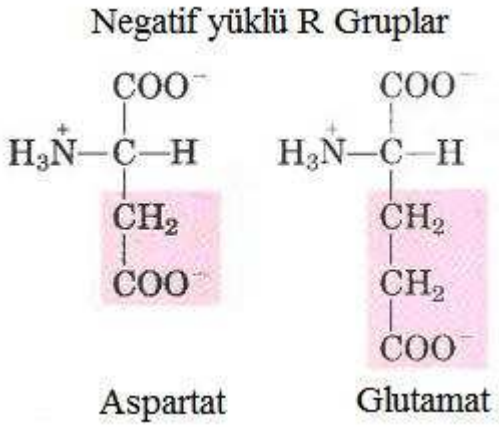
(http://web2.tmu.edu.tw/m110093011/DNA2protein.htm'den uyarlanarak alınmıştır) .



Şekil 2d: Pozitif yüklü R gruba sahip amino asitler

(http://web2.tmu.edu.tw/m110093011/DNA2protein.htm‘den uyarlanarak alınmıştır) .

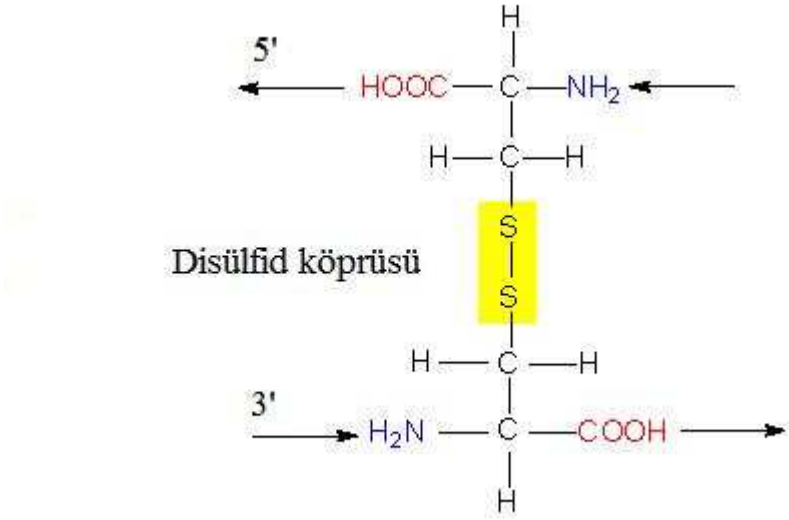
.



Şekil 2e: Negatif yüklü R gruba sahip amino asitler

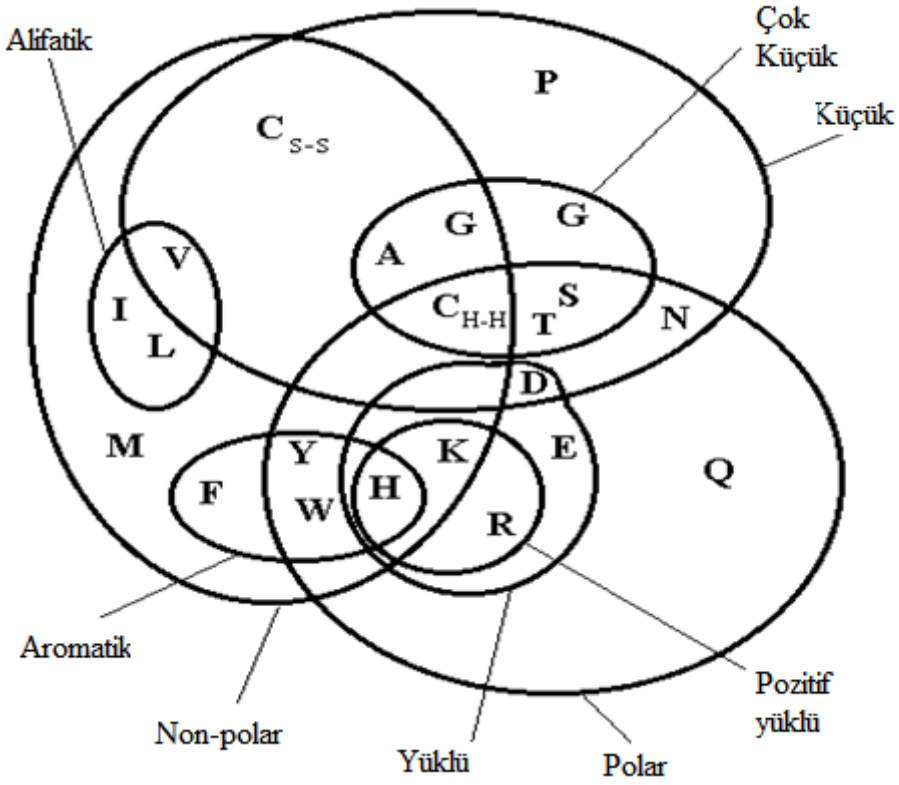
(http://web2.tmu.edu.tw/m110093011/DNA2protein.htm‘den uyarlanarak alınmıştır) .

Nonpolar alifatik R grupları olan amino asitler, alanin, valin, lösin ve izolösin proteinlerde genel olarak birlikte bulunma eğilimindedirler. Bu eğilim neticesinde hidrofobik etkileşimler doğar ve bu etkileşim proteinlerin yapısını stabil tutan kuvvetlerden biridir. Aromatik R grubu olan fenilalanin, tirozin ve triptofan ise R gruplarında aromatik bir halka taşırlar. Polar, yüksüz R grubu olan amino asitler ki R gruplarında aktif H atomu bulundururlar bu nedenle suda oldukça iyi çözünürler. Bu özellik hidrofili olarak adlandırılır. Hidrofilik olan amino asitler su ile etkileşime girer. Pozitif yüklü R grubuna sahip olan amino asitler ise nötral çözücülerde proton kazanarak pozitif yüklü hale geçerler; negatif yüklü amino asitler çözücülerde proton kaybederek negatif yüklü hale geçerler. R grubunda sülfür bulunduran amino asitler ise etkileşimde bulunarak disülfid köprüleri oluştururlar. R grubunda sülfür bulunduran amino asitler sistein ve metionin amino asitleridir. Disülfid köprüleri de proteinlerin stabilitesi açısından önemli bağlardandır. Disülfid köprüsü Şekil 3 ‘de gösterilmiştir (Nelson ve Cox, 2008).



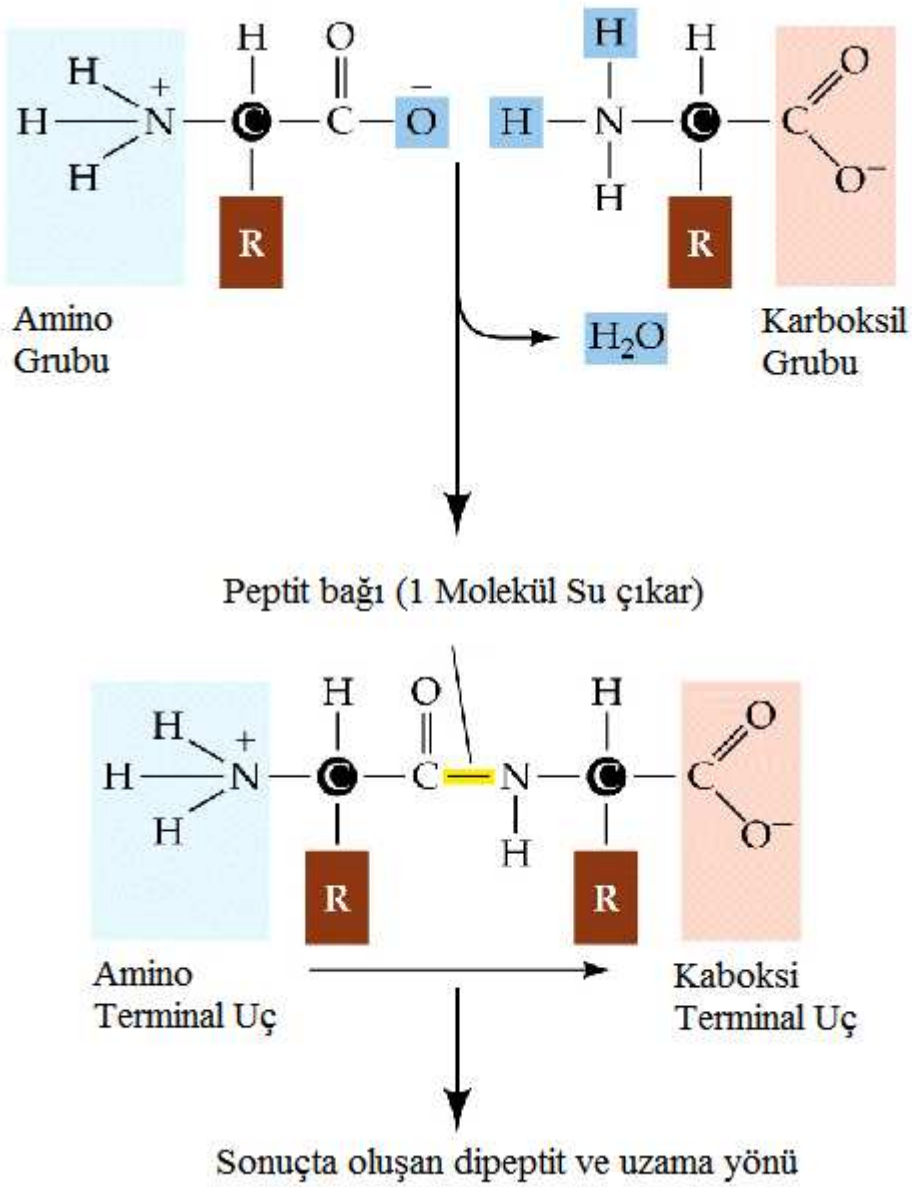
Şekil 3: Disulfid bağı.

Amino asitlerin R gruplarına göre fizikokimyasal özellikleri için genel kabul görmüş venn şeması kullanılmıştır. Bu venn şeması Şekil 4 ‘de gösterilmiştir (Taylor, 1986).



Şekil 4: R gruplarına göre amino asitlerin dağılımları ve dağılımın venn şeması ile ifadesi.

İki amino asit birbirine peptit bağı denen kovalan bir bağ sayesinde bağlanır. Bu şekilde bir peptit bağı ile bağlanmış iki amino asit bir dipeptit yapısı oluşturur. Benzer şekilde üç amino asitten oluşan yapıya tripeptit, yirmi ile elli arasında amino asitten oluşan ve amino asitleri birbirlerine peptit bağları ile bağlı haldeki yapıya oligopeptit, pek çok amino asidin peptit bağları ile bağlanıp oluşturdukları yapıya ise polipeptit adı verilir. Proteinler peptit bağı ile birbirine bağlanmış durumdaki polipeptit zincirlerinden oluşur. Şekil 5 ‘de iki amino asitten oluşan bir dipeptit örneği ve peptit bağının oluşumu gösterilmiştir.



Şekil 5: İki amino asit arasında oluşan peptit bağı ve dipeptit oluşumu (http://www.bothbrainsandbeauty.com/page/3 ‘den uyarlanarak alınmıştır).

Şekil 5 ‘de görüldüğü gibi birinci amino asitin negatif yüklü karboksil grubu ile, ikinci amino asitin pozitif yüklü amino grubu arasındaki kondensasyon reaksiyonu neticesinde bir molekül su çıkarak peptit bağı, karboksil grubu ile amino grubu arasında oluşur. Oluşan bu yapı dipeptit adını alır. Bu şekilde amino asitlerin birbiri ardına eklenmesi ile uzayıp giden proteinler meydana gelir.

2.3 Proteinler

Protein sözcüğü, Yunanca birincil öneme sahip anlamını taşıyan *prota* sözcüğünden köken almıştır. Bu isim, proteinleri 1838'de ilk tanımlayan Jöns Jakob Berzelius tarafından verilmiştir. Yapısı çözülen ilk proteinler arasında insülin ve miyoglobin bulunur ki, insülin için Sir Frederick Sanger 1958'de, miyoglobin için de Max Perutz ve Sir John Cowdery Kendrew 1962'de Nobel Kimya Ödülü kazanmıştır (Kendrew ve ark., 1958).

Proteinler canlıların en önemli yapısal ve fonksiyonel makromolekülleridir. Hücre içinde aynı anda gerçekleşen binlerce kimyasal reaksiyonu katalizleyen enzimler, metabolik olayları düzenleyen hormonlar, vücut savunmasında görev alan antikorlar vb. pek çok hayati fonksiyon proteinler tarafından yürütülür (Nelson ve Cox, 2008).

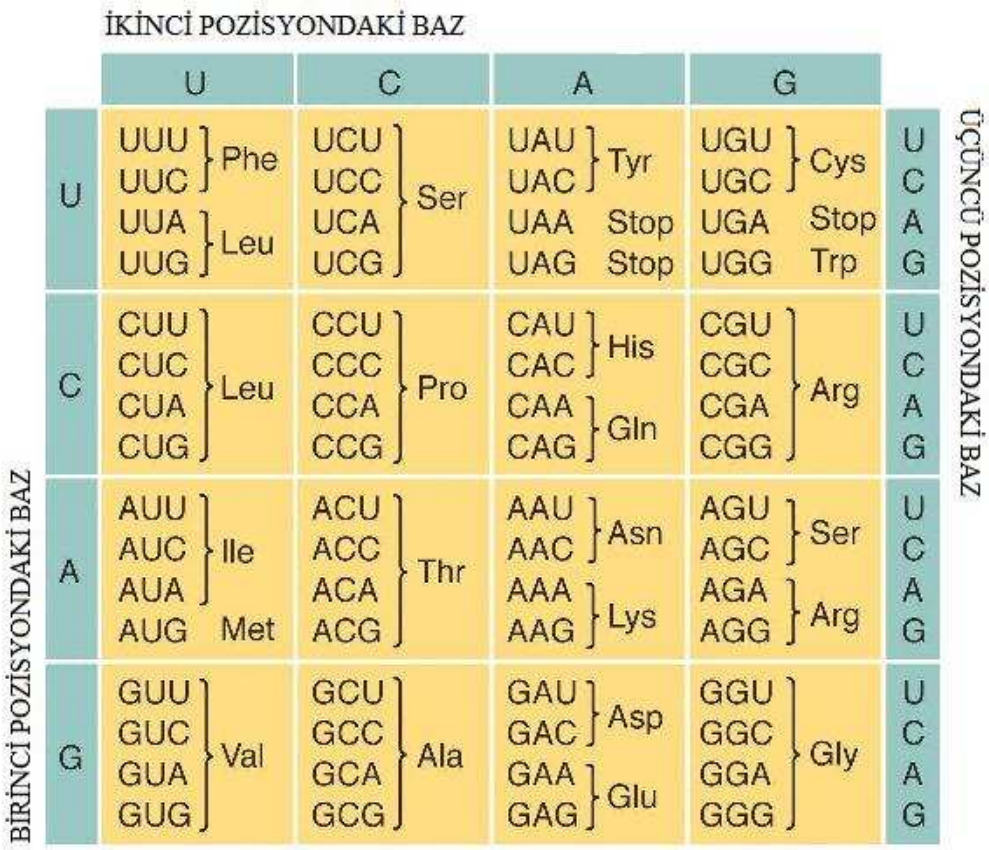
2.3.1 Proteinlerin sentezlenmesi

Protein sentezi ribozomlarda gerçekleşir. Basitçe hücre içindeki amino asitlerin ribozom üzerine taşıyıcı ribonükleik asit (tRNA) molekülleri ile getirilerek, aralarında peptit bağı kurulması ve bu olayın bir zincir şeklinde uzayıp gitmesi şeklinde tanımlanabilir. DNA üzerinde protein kodlayan gen dizisi bir mesajcı RNA (mRNA) ‘ya aktarılır. Bu işlemin adı transkripsiyon olarak adlandırılır. Daha sonra gen dizisini taşıyan mRNA hücre çekirdeğinden hücre sitoplazmasına geçer. Hücre içinde 20 çeşit amino asit bulunur. Amino asitler tRNA tarafından yakalanır. Amino asitlerin taşınması esnasında amino asitlere özgün olan tRNA’lar kullanılır. Amino asitlerin özgün tRNA’sı tarafından yakalanması işlemine yükleme denir. Yükleme işlemi hücre içinde aminoaçil tRNA sentetaz adı verilen bir enzim tarafından gerçekleştirilir. Hücre içinde 20 çeşit amino asit bulunduğu düşünülürse en azından 20 çeşitte tRNA ‘nın var olması gerekir. Fakat bir amino asitin kodlanması 3 adet deoksiribo nükleik asit (DNA) bazının[1] yan yana gelmesi

[1] DNA bazların Adenin (A), Timin (T), Guanin (G), Sitozin (C) ‘dir.

ile mümkün olduğundan (yan yana gelen üç DNA bazı ile oluşan bu özgün kodlama şekline kodon adı verilir, her bir tRNA üzerinde ise kendi amino asitinin kodonuna özel olmak üzere yine 3 RNA bazından[2] oluşan bir antikodon yapısı bulunur) teorik olarak 61 çeşit tRNA ’nın bulunması gerektiği düşünülür. Yapılan araştırmalar hücre içinde 32 çeşit tRNA bulunduğunu göstermiştir. Bu farklılığın temel nedeni kodon yapısında 3. bazın esnek olmasıdır. Bu durum biyolojide wobble hipotezi olarak adlandırılır (Crick , 1966; Lodish ve ark., 2004; Klug ve Cummings, 2003). Şekil 6 ‘da amino asitlere özgün kodonlar gösterilmiştir.

[2] RNA bazları Adenin (A), Urasil (U), Guanin (G), Sitozin (C) ‘dir



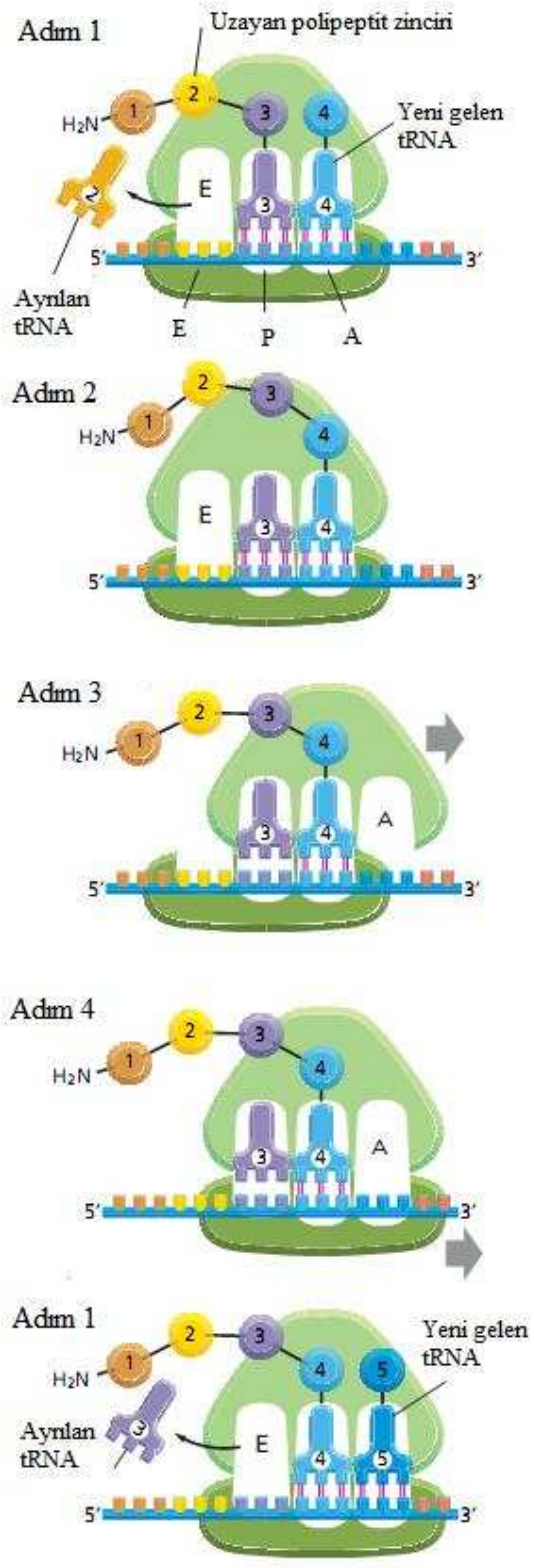
Şekil 6: Standart genetik kod tablosu, amino asit kodonları ve üçüncü bazın esnekliği (http://www.mun.ca/biology/scarr/MGA2_03-20.html 'den uyarlanarak alınmıştır).

Yüklenmiş tRNA ribozom üzerindeki P bölgesine[3] oturur, sonra ikinci tRNA taşıdığı amino asit ile A bölgesine[4] gelir. Bu sayede iki amino asit yan yana gelmiş olur. Bu esnada iki amino asit arasında kondensasyon tepkimesi meydana gelir. Bu tepkime dehidratasyon tepkimesi olarak da adlandırılır. İki amino asitin yana yana gelip dehidratasyon tepkimesi ile birleşmesi ile peptit bağı oluşur. Peptit bağı n amino asitinin karboksil ucu ile n+1 amino asitinin amino ucu arasında oluşan bir bağdır (n+2 amino asiti

[3] Ribozom üzerindeki Peptidil bölgesi.
[4] Ribozom üzerindeki Aminoaçil bölgesi

ribozoma getirildiğinde n amino asiti ribozomu E[5] bölgesinden terk eder). Bu esnada n+1 amino asitine peptit bağı ile sıkıca bağlanmıştır. Dehidratasyon tepkimesi olduğundan zincire ne zaman yeni bir amino asit eklenecek olsa bir molekül su çıkışı olur. mRNA kalıbı üzerinden tRNA 'ların uygun amino asiti getirip zincire yerleştirmesi şeklinde uzayıp giden sentez durdurucu kodonlardan biri gelene kadar devam eder. Durdurucu kodonlar 3 tane olup UAA, UGA, UAG 'dir (durdurucu kodonlar Şekil 6'da da görüldüğü üzere herhangi bir amino asiti kodlamazlar ve wooble hipotezi bu kodonlar içinde geçerlidir UAA - UAG gibi). Şekil 7 'de protein sentezinin basamakları gösterilmiştir.

[5] Ribozom üzerindeki tünel bölgesi ya da exit bölgesi. P bölgesindeki amino asit A bölgesindeki amino asitle peptit bağ oluşturduktan sonra ribozomu E bölgesinden terk eder.



Şekil 7: Protein sentezi aşamaları ve peptit zincirinin uzaması, amino asitlerin ribozom bölgelerindeki hareketleri.

2.3.2 Proteinlerin yapısı

Proteinlerin doğal durum katlanmalarına ulaşıp özgün görevlerini yapma sürecine etki eden olayları ortaya çıkarmak biyoinformatik biliminin en temel fakat en zor sorularından birisidir. Bu sorunun cevabını ararken yapısal olarak proteinler birincil yapı, ikincil yapı, üçüncül yapı ve dördüncül yapı olarak seviyelere ayrılmıştır (Nelson ve Cox, 2008).

2.3.2.1. Birincil yapı

Proteinlerin herhangi bir katlanma yapmadan sergiledikleri yapıdır (teorik olarak lineer sıralanma). Proteinlerin aralarında peptit bağı olduğu halde amino asitlerin yan yana lineer bir şekilde dizilmesi ile oluşan yapısı onların birincil yapılarını tanımlar. Birincil yapı gösterimi amino asitlerin yan yana yazılması ile sağlanır. Bu yazım esnasında amino asitler amino ucundan, karboksil ucuna doğru yazılırlar.

NTER|Met-Ala-Ser-Leu-Gly-His-Ile-Leu-Val-Phe-Cys-Val-Gly-Leu-Leu-Thr-Met-Ala-Lys-Ala-Glu-Ser-Pro-Lys-Glu-His-Asp-Pro-Phe-Thr-Tyr-Asp-Tyr-Gln|CTER şeklinde 3-harf kodlamasına uygun olarak proteinlerin birincil yapısı gösterilebilir. Benzer şekilde proteinlerin birincil yapıları 1-harf yöntemine uygun olarak NTER|MASLGHILVFCVGLLTMAKAESPKEHDPFTYIAGILFI|CTER şeklinde de gösterilebilir. Fakat biyoinformatik biliminde yaşanan gelişmeler, ve biyolojik bilgi artışı 3-harf kullanımını nerede ise imkansız hale getirmiştir. Dizi başındaki NTER ibaresi peptit zincirinin amino terminal ucu (n-terminal) ve dizi sonundaki CTER ibaresi peptit zincirinin karboksi terminal ucu (c-terminal) için kullanılmış kısaltmalardır. Bu teorik protein için uzunluk 38 amino asittir. Genel olarak 38 aa şeklinde gösterilir.

Birincil yapıdaki amino asitlerin farklı fizikokimyasal özellikler sergileyen R grupları nedeni ile direkt olarak oluşan proteinin yapısına ve özellikleri ile proteinin daha

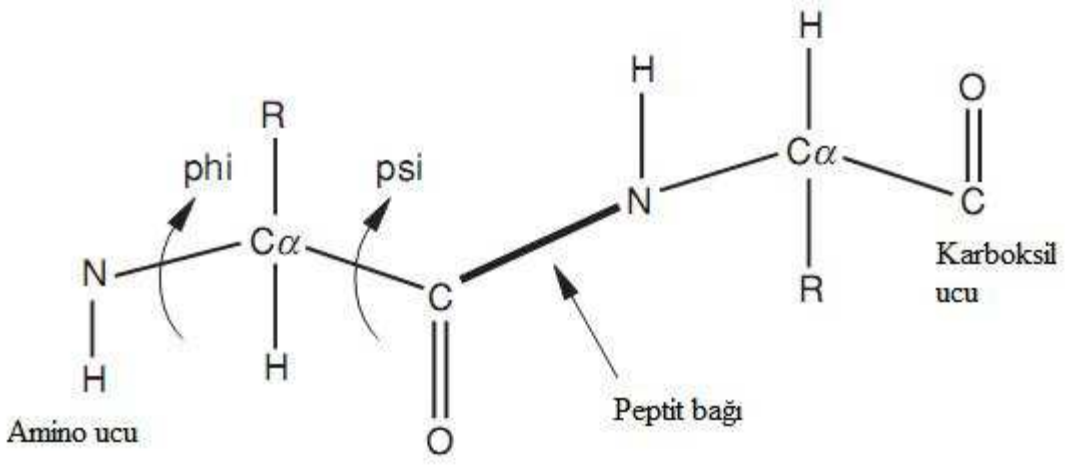
yüksek seviyeden yapısına direkt olarak etki eder. Buradan hareketle proteinlerin doğal durum katlanmalarına, dolayısı ile, fonksiyonuna direkt olarak onun birincil yapısı etki eder hipotezi kabul edilir. Bu hipotez dogması olarak bilinir. Anfinsen dogması termodinamik kanunlarının moleküler biyolojideki yansımasıdır (Anfinsen, 1973).

2.3.2.2 İkincil yapı

Birincil yapının kendi üstüne katlanması ile oluşan ikinci seviye protein yapısına proteinlerin ikincil yapısı adı verilir. İkincil yapı sürekli tekrar eden farklı üç boyutlu primerlerden oluşur. İlk bulunan primerler α-heliks ve β-tabaka yapılarıdır (Pauling ve Corey; 1951).

İkincil yapı ayrıntılarına girmeden önce ikincil yapının kaynağı olan peptit bağına ve bu bağın özelliklerine bakmak gerekir. Peptit bağı, n amino asitinin karboksil ucu ile n+1 amino asitinin amino ucu arasındaki dehidrasyon tepkimesi sonucunda bir molekül suyun ayrılması ile iki amino asitin birbirine sıkıca bağlanmasını sağlayan kovalent yapıya verilen addır. Peptit bağı, kendisini oluşturan amino asitlerin R gruplarının farklı olmasından dolayı özgün bir geometriye sahiptir. Peptit bağları ile uzayıp giden polipeptitin tamamına bakıldığında n-terminal ucundan hemen sonra alfa karbonu ve alfa karbonuna bağlı karboksil karbonu bunu izleyen peptit bağı ve sonrasında ikinci amino asitin amino grubuna bağlı azot ve tekrar alfa karbonu şeklinde devam eder (CORN yapısı). İşte R grupları yapıya dahil edilmeksizin karşımıza çıkan bu geometriye omurga (backbone) adı verilir. Amino asit bilindiği gibi sabit üç boyutlu yapıya sahiptir. Çünkü omurgaya eklenen bir R grubu ile ancak farklı bir amino asit yazılabilir. O halde proteinlerde ki inanılmaz çeşitliliğin en önemli nedeni peptit bağ açıları ve sahip oldukları

farklı R gruplarıdır. Peptit bağı ve peptit bağ açıları Şekil 8‘de gösterilmiştir (Branden ve Tooze, 1991).



Şekil 8: Peptit bağı ve peptit bağ açıları.

Görüldüğü üzere iki amino asit peptit bağı ile bağlanıp bir dipeptit formasyonu oluşturmuştur. Alfa karbon atomu üzerinde iki dönüklük açısı Phi ve Psi olarak verilmiştir. Phi dönüklüğü alfa karbon atomunun amino ucundaki N ile olan; Psi dönüklüğü ise alfa karbon atomunun karboksil ucundaki karbon atomu ile yaptığı dönüklük açıları tanımlamaktadır. Peptit yapısının bu geometrisi tüm ikincil yapı elementlerinin temsili için yeterlidir. İkincil yapının temel elementleri alfa heliks, beta tabaka ve dönüşlerdir (Sander ve ark., 1983).

2.3.2.3 Heliks yapısı

Alfa heliks yapısı düzenli dönüşlerden meydana gelir. Dönüş üzerinde bulunan amino asit sayısı ve dönüşü sağlayan hidrojen bağlarının lokalizasyonuna göre de isimlendirilirler. Alfa heliks yapısı proteinlerin yapısında en çok görülen ikincil yapı elementidir. Alfa heliks için Phi ve Psi açıları -60 ile -50 arasındadır. Alfa heliks oluşturabilmek için bir tam dönüş (360 derece) içinde 3.6 amino asit bulunmalıdır.

Hidrojen bağlarının yerleşimi ise *i* amino asitinin amino grubu ile *i+3* amino asitinin karboksil grubu arasında arasındadır. Alanin, glutamik asit, lösin ve metionin heliks yapıda bulunmayı tercih ederken prolin, serin, tirozin ve glisin alfa heliks yapısında çok nadir olarak bulunurlar (Branden ve Tooze, 1991; Pace ve Scholtz, 1998).

Alfa heliks yapısının uzunluğu en az 3 amino asit iken 40 amino asite kadar çıkan heliksler de tanımlanmıştır. Ortalama bir heliks boyu 10 amino asittir (Sander ve ark,1993; Arjuvan ve ark., 2001).

Alfa heliks yapısının bir özelliği de parsiyel olarak amino ucunda pozitif, karboksil ucunda negatif yüklü olmasıdır. Bu parsiyel yük heliks yapısına etki eder. Bu etkinin adı dipol momenti olarak bilinir. Dipol momenti heliks eksenine paralel olarak uzanan bir elektro manyetik alan oluşumuna neden olur. Bu kuvvet alanı makroskopik dipol momenti olarak tanımlanır, bu moment mikro dipol momentler toplamıdır. Makroskopik dipol momenti direkt olarak protein yapı ve fonksiyonuna etki eden kuvvetlerden birisi olup heliks yapısının stabilizasyonuna da etki eder (Sengupta ve ark., 2005).

Hidrojen bağlarının lokalizasyonuna göre heliks yapısı 3π-heliks, 3-10π-heliks, 5π-heliks olarak ayrılmıştır. Heliks yapıları ikincil yapı sözlüğün (Dictionary of Secondary Structure, DSSP) ‘de ; 3-dönüş (3-turn), 4-dönüş (4-turn) ve 5-dönüş (5-turn) olarak adlandırılmıştır. Tüm ikincil yapı elementleri ve DSSP kısaltmaları Tablo 2 ‘de gösterilmiştir (Sander ve ark., 1983).

Tablo 2: DSSP ‘de yer alan ikincil yapı elementleri

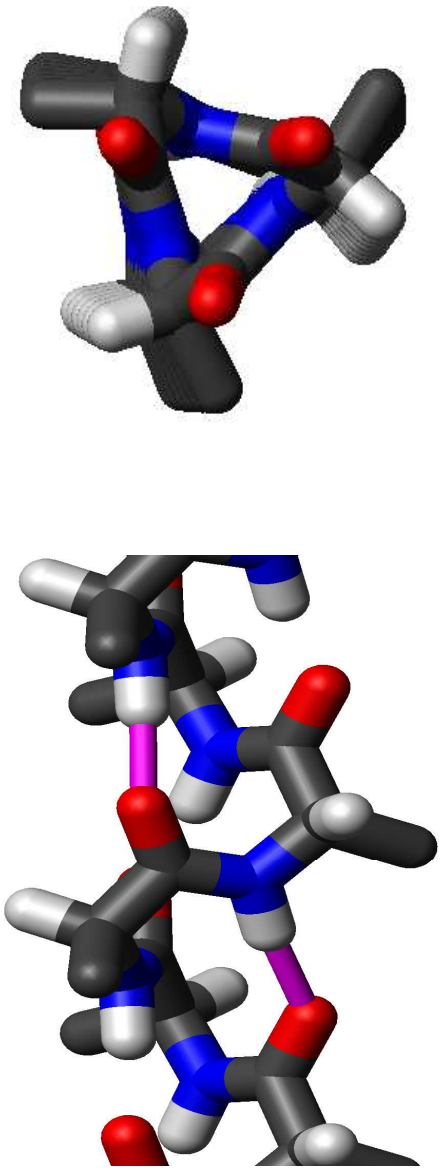
Kısaltma	**Açıklama**	**Genel Grup**
H	Alfa heliks yapısı	Heliks
B	Beta tabaka	Beta Tabaka
E	Genişlemiş beta tabaka ya da beta merdiven yapısı	Heliks
G	3-heliks ya da 3/10 heliks yapısı	Heliks
I	5 heliks ya da pi heliks	Heliks
T	Dönüş yapısı	Dönüş
S	Kuşak yapısı	Zincir

Tablo 2 ’de ikincil yapı elementlerinin hidrojen bağlanmalarına göre yapılmış sınıflandırılması, kullanılan kısaltmalar ve bu kısaltmaların kısa açıklamaları gösterilmiştir. Tablo 2’de görüldüğü gibi genel gruplar üç tanedir. Heliks (H), Beta tabaka (S), dönüş (T), zincir ya da kuşak (C).

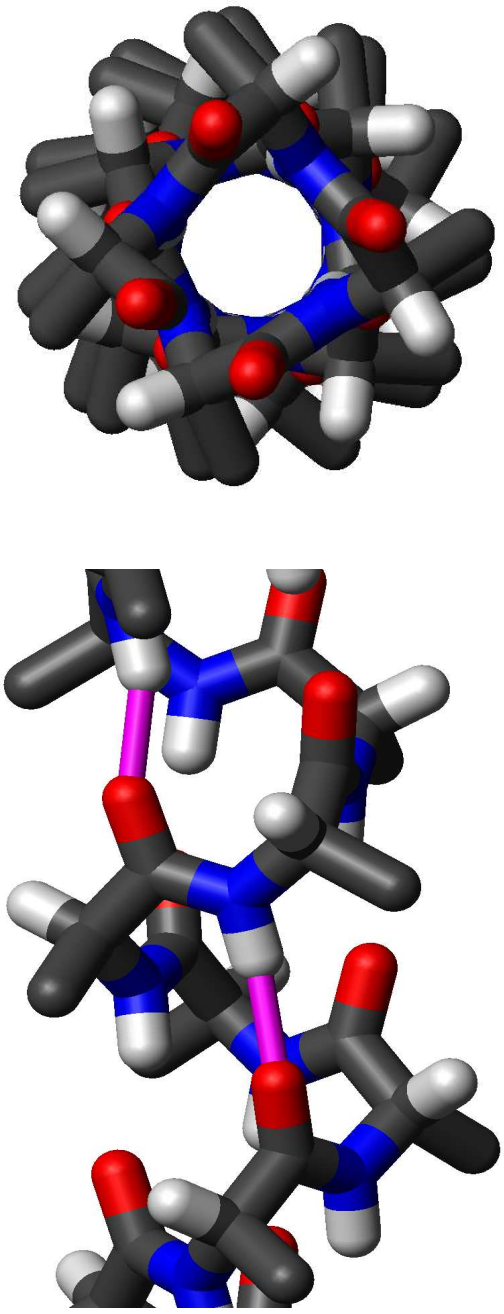
3-dönüş yapısında ; *i* amino asitinin karboksil ucundaki hidrojen atomu ile *i+3* amino asitinin amino ucundaki hidrojen atomu arasında hidrojen bağı bulunur. Bu yapı *Hbond(i, i+3)*[6] olarak yazılır. 4-dönüş yapısında; *i* amino asitinin karboksil ucundaki hidrojen atomu ile *i+4* amino asitinin amino ucundaki hidrojen atomu arasında hidrojen bağı bulunur. Bu yapı *Hbond(i,i+4)* olarak gösterilir. 5-dönüş yapısında; *i* amino asitinin karboksil ucundaki hidrojen atomu ile *i+5* amino asitinin amino ucundaki hidrojen atomu arasında hidrojen bağı bulunur. Bu yapı *Hbond(i,i+5)* olarak gösterilir. Bahsi geçen heliks yapılarının genel hali n={1,2,3,…} iken *n-turn (i) = HBond(i,i+n)* olarak sembolize edilmiştir (Sander ve ark., 1983).

Bu heliks yapılarının üç boyutlu benzetimleri üstten ve yandan görünüm olmak üzere Şekil 9a, Şekil 9b, Şekil 9c ‘de sırası ile 3-dönüş, 4- dönüş ve 5- dönüş gösterilmiştir.

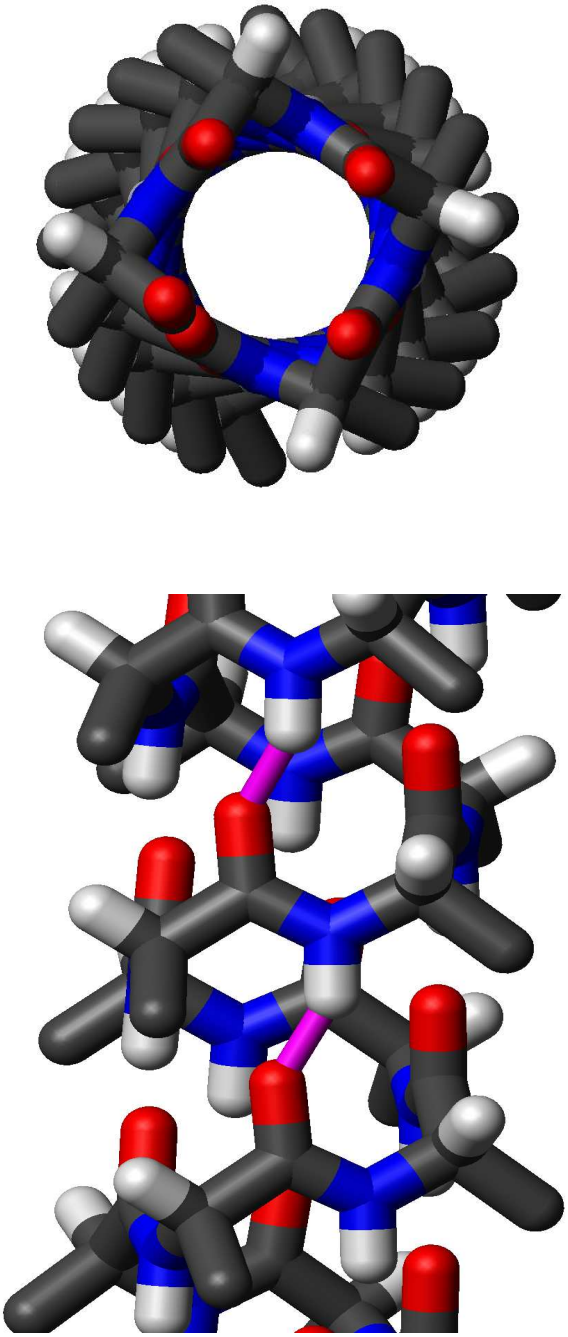
[6] Bu yazım şekli orijinal DSSP sözlüğü yazım şekli olduğundan değiştirilmemiştir.



Şekil 9a: 3/10 heliks ya da 3pi heliks ya da 3-turn heliks yapısı. Burada i rezüdüsü ile i+3 rezüdüsü arasında hidrojen bağı oluşmuştur. Sağ el sarmalı helikal yapısıdır. Her amino asit bir öncekine göre 120 derece kadar dönüktür. Teorik açılar -49,-26 arasındadır. 0.2nm aksa sahiptir. Hidrojen bağları pembe renk ile gösterilmiştir (www.wikipedia.org ‘den uyarlanmıştır).



Şekil 9b: 4/10 heliks ya da alfa heliks ya da 4-turn heliks. Burada i rezüdüsü ile i+4 rezüdüsü arasında bir hidrojen bağı vardır. Sağ el sarmalı helikal yapıya sahiptir. Her amino asit bir öncekine göre yaklaşık 100 derece kadar dönüktür. Aksı 1.5 A ‘dur. Proteinlerin yapısında bulunan helikal formların majör grubunu oluşturur. Hidrojen bağları pembe ile gösterilmiştir (www.wikipedia.org ‘den uyarlanmıştır).



Şekil 9c: 5/10 heliks, 5pi heliks ya da 5-turn heliks. Burada i rezüdüsü ile i+5 rezüdüsü arasından hidrojen bağı kurulmuştur. Her amino asit bir öncekine göre yaklaşık 87 derece dönüktür. Helikal yapısı sağ el sarmalıdır. Aksı yaklaşık olarak 1.15 A ‘dur. Hidrojen bağları pembe ile gösterilmiştir (www.wikipedia.org ‘den uyarlanmıştır).

2.3.2.4 Beta Tabaka

Beta tabaka proteinlerin temel ikincil yapı elementlerinden heliks yapısından sonra gelen ikinci en sık rastlanan yapı formudur. Heliks yapısında olduğu gibi amino asitlerin lineer dizilimi olan birincil yapının bir seviye katlanması ile oluşur. Fakat helikal yapıdan farkı karboksil grubu ile amino grubu arasında oluşan hidrojen bağının lokalizasyonudur.

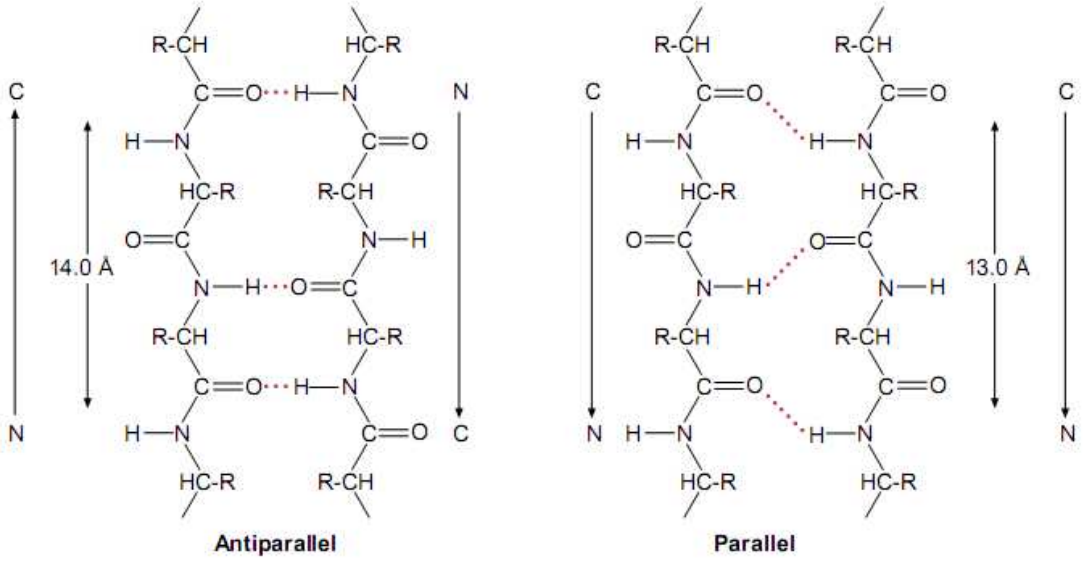
Beta tabakanın oluşabilmesi için iki ya da daha fazla sayıda fakat en az iki lineer zincirin yan yana gelmesi gereklidir. Daha sonra yan yana gelen bu iki zincir arasında hidrojen bağları kurulur. Beta tabaka iki çeşittir. Bunlardan ilki paralel beta tabaka olarak, ikincisi ise anti paralel beta tabaka olarak adlandırılır. Paralel beta tabakanın oluşması için polipeptit zincirlerinin aynı yönde uzanması gereklidir. Anti paralel olan beta tabakanın oluşması için polipeptit zincirlerinin birbirinin zıttı istikamette uzanması gereklidir. Anti paralel olan beta tabaka paralel olan beta tabakaya nazaran daha kararlı bir yapıdır. Bu kararlılıktan dolayı proteinlerin yapısında daha çok bulunur. Karışık (mix type) tipte beta tabaka formasyonuna nadiren rastlanır (Davies, 2008; Sander ve ark., 1983).

Hidrojen bağlanmaları ise :

$$Paralel\ beta\ tabaka(i,j) = \begin{cases} HBond(i-1,j)\ ve\ HBond(j,i+1) \\ HBond(j-1,i)\ ve\ HBond(i,j+1) \end{cases}$$

$$Antiparalel\ beta\ tabaka(i,j) = \begin{cases} HBond(i,j)\ ve\ HBond(j,i) \\ HBond(i-1,j+1)\ ve\ HBond(j-1,i+1) \end{cases}$$

Şeklinde ifade edilmiştir. Şekil 10 ‘ de beta tabaka ve tipleri ile bu tabakalardaki hidrojen bağları gösterilmiştir.

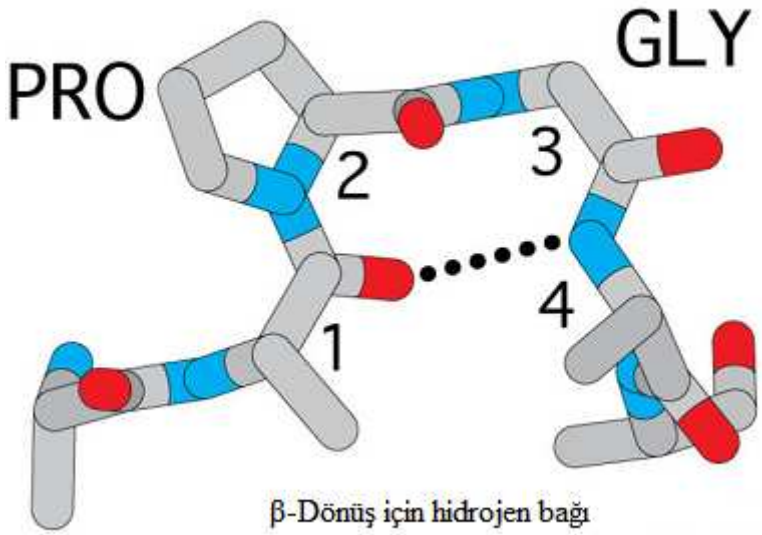


Şekil 10: Antiparalel ve paralel beta tabakalar ve bu tabakaların sahip oldukları hidrojen bağlanmaları.

2.3.2.5. Dönüş yapısı

İkincil yapının, heliks ve beta tabakadan sonra gelen elementidir. Hidrojen bağlanmalarının lokalizasyonuna göre beta dönüşleri sınıflandırılmıştır. Bu sınıflamaya göre 5 tip dönüş yapısı bulunur. Bu dönüşler sırası ile β-dönüş, α- dönüş, γ- dönüş, π-dönüş ve δ- dönüş olarak adlandırılmıştır.

β- dönüş yapısında i rezüdüsü ile i ± 3 rezüdüsü ya da arasında hidrojen bağı bulunur. α- dönüş yapısında i rezüdüsü ile i±4 rezüdüsü arasında, γ- dönüş i rezüdüsü ile i±2 rezüdüsü arasında, π- dönüş i rezüdüsü ile i±5 rezüdüsü arasında ve δ- dönüş ise i rezüdüsü ile i±1 rezüdüsü arasında hidrojen bağları bulunur. Bu noktalarda protein dönerek kendi üzerine katlanır (Hutchinson ve Thornton, 1994). Şekil 11 ’de beta dönüş yapısı gösterilmiştir.



Şekil 11: Prolin ve glisin amino asitleri arasında i rezüdüsü olan 1 ile i+3 rezüdüsü arasında kurulan hidrojen bağı ile oluşmuş beta dönüş yapısı.

2.3.3 Üçüncül yapı

Proteinlerin ikincil yapısının bir derece katlanması ile oluşan formasyonuna, proteinlerin üçüncül yapısı denir. Bu yapıda ikincil yapıdaki heliks, beta tabaka ve dönüş formasyonlarının birbirlerine göre dönüklükleri ile protein tek bir molekül olarak gösterilir.

Üçüncül yapının stabilitesi genelde lokal olmayan kuvvetler sayesinde olur. Bu kuvvetlerden bazıları hidrofobik çekirdek oluşumu, tuz köprüleri, hidrojen bağları ve disülfid köprüleri ile sentez sonrası değişimler (post translasyonel modifikasyon, PTM) ısı, asidite, çözücü etkisi olarak sayılabilir.

2.3.4 Dördüncül yapı

Proteinlerin dördüncül yapıları, birden çok proteinin ya da polipeptit yapısının birlikte ve tek bir molekül olarak üç boyutlu yapı kazanması sonucunda oluşur. Bu yapıya

güzel bir örnek hemoglabin molekülüdür. Hemoglobin molekülünde alfa ve beta alt ünite polipeptit dizileri ile yapıdaki demir içeren hem grupları dördüncül yapıyı oluştururlar.

2.4 Genetik bilgi uzayı ve veri tabanları

Genetik bilginin her gün değişen ve gelişen doğası bugün önemli bir soruna neden olmuştur. Aradığınız bilgi internette nerededir? 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. Var olan 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. Diğer yandan 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 daha zorlaştırmaktadır. İhtiyaç duyulan bilginin elde var olan verilerden çıkartılıp karşılaştırılmalarının, hesaplamaların uygulanabilmesi, yeni kazanımlar için depolanabilir kılınması ilişkisel veri modellerine olan ihtiyacı artırmıştır. Verilerin üzerinde her an değişiklik yapabilmek, yenilerini eklemek, yeni ilişkiler tanımlamak, hipotezleri 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 gerekli hale gelmiştir.

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 aşağıda sıralanmıştır (Baxevanis, 2006).

1) Gen Bankası (GenBank)
2) Avrupa Moleküler Biyoloji Laboratuvarı (The European Molecular Biology Laboratory, *EMBL*)
3) Tek Nokta Değişimleri Veri Tabanı (Database of Single Nucleotide Polymorphisim, dbSNP)
4) Ulusal Biyoteknoloji Merkezi (National Center For Biotechnology Information, NCBI)
5) Ulusal Tıp Kütüphanesi (National Library of Medicine, NLM)
6) Ulusal Sağlık Enstitüsü (National Institutes of Health, NIH)

GenBank kısaca; kapsamlı, herkese açık kullanıma sahip, nükleotid veri bankasıdır. GenBank ayrıca bibliyografik biyolojik notları da kullanıcısına sunar. Bunların yanında bu veri tabanından Genom Araştırma Dizisi (Genome Survey Sequence, GSS), İfade Edilen Dizi Etiketleri (Expressed Sequence Tags, EST) ve Tam Genom Dizileme (Whole Genome Sequencing, WGS) gibi pek çok bilgiye de referansları ile ulaşmak mümkündür (Benson ve ark., 2008).

Sadece bu üç veri tabanı (GenBank, EMBL, dbSNP) 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, 03.02.2009).

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, 24.01.2010).

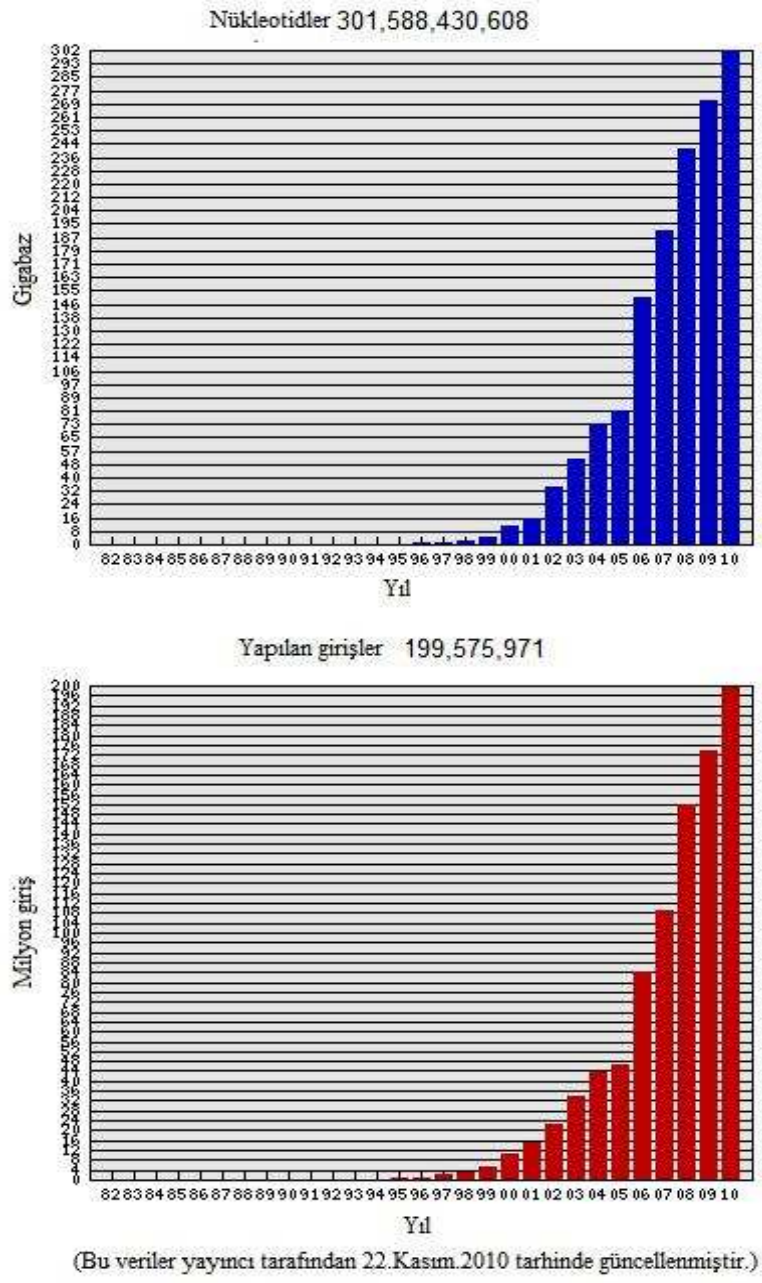
GenBank altında yine oldukça sık ziyaret edilen diğer bir veri tabanı olan dbSNP ise 130 farklı organizma hakkında polimorfizm verilerini bize sunmaktadır. dbSNP'de var olan 130 organizmadan birisi olan insan hakkında 24 Ocak 2010 itibarı ile toplam 12.878.918 giriş bulunmaktadır (http://www.ncbi.nlm.nih.gov/SNP/snp_summary.cgi, 24.01.2010).

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, 01.11.2009). Medline ise 1966 'dan bu yana 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, 24.01.2010).

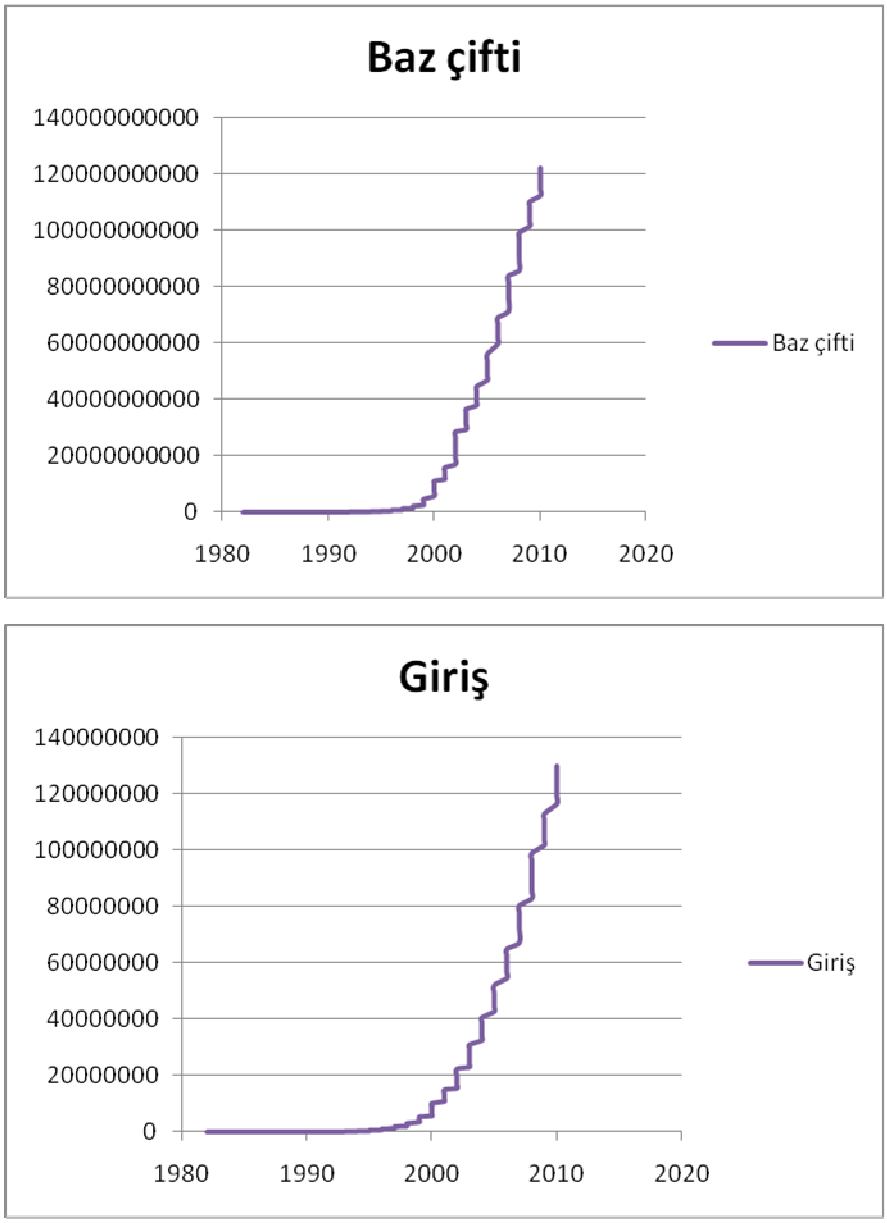
Görüldüğü gibi bu rakamlar her saniye değişmektedir. Nükleotid dizileme makinelerinin artması, teknolojilerinin her geçen gün değişmesi ve iyileşmesi, biyoinformatik yazılımlar ve bunların güncel teknoloji ile buluşması, yürütülegelen çalışmalar ve sonuçları gibi pek çok etken nedeniyle genetik bilgi 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. Genetik veri uzayının büyüyen ve gelişen dünyası Şekil 12a, Şekil 12b üzerinde gösterilmiştir.



Şekil 12a: EMBL için nükleotid ve veri girişlerinin yıllara göre nükleotid ve adet bazında karşılığı. http://www.ebi.ac.uk/embl/Services/DBStats/ adresinden uyarlanarak alınmıştır. Veriler 05.12.2011 tarihine aittir, lakin yayıncı tarafından 22.11.2010 tarihinde güncellenmiştir.



Şekil 12b: GenBank veri tabanındaki büyüme. Bu grafikler oluşturulurken 1980 yılından 2010 yılı sonuna kadar olan veri tabanına kayıtlı toplam nükleotid sayısı ve veri tabanına kayıtlı toplam giriş sayısı değerleri kullanılmıştır. Bu veriler GenBank 181 'inci sürümünün sürüm notlarında, 2.2.8 Growth of GenBank bölümünden elde edilen ham verilerdir. Bu veriler GenBank 181 sürümü için ftp://ftp.ncbi.nih.gov/genbank/gbrel.txt adresinden alınmış grafik şekline dönüştürülüp uyarlanmıştır.

Birkaç on yıl önce bilim adamları tarafından biyolojik bilgi birikimini tek merkezden yönetmek ve araştırmacılara uzun dönemli veri olarak sunmak amacıyla çalışmalara başlanılmıştır. Erken dönemlerinde bu çalışmada kelime işlemciler ve hesap tabloları kullanılarak bu işlem yapılmaya çalışılmıştır. Bu çabalar sınırlı da olsa verilerin depolanmasına ve araştırmacılar arasında değiş tokuşuna müsaade edilmiştir. Fakat bilginin inanılmaz artışı karşısında bu yöntemler, verimli ve etkili depolama, paylaşma ve yararlanmak amacıyla sonra kullanma gibi ihtiyaçlara cevap veremez olmuştur. Bu nedenle web tabanlı daha karmaşık organizasyona sahip yönetim sistemlerinin geliştirilmesi gerekmiştir. Web tabanlı olarak verileri sunmak günümüz koşullarında 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 içerir. 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 şeklinde kullanıcısına 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, dbSNP, 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.

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 veri tabanı 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 formatlarda (text, imaj, PDF dosyası vb.) saklanmış özelliklerini yönetmeyi sağlayan ilk örnektir (Hong ve Wong, 2005).

SnpHunter ise, seçtiğimiz bir gendeki tüm SNP noktalarını bilgisayarımıza indirip, filtre edip üzerinde çalışabileceğimiz 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).

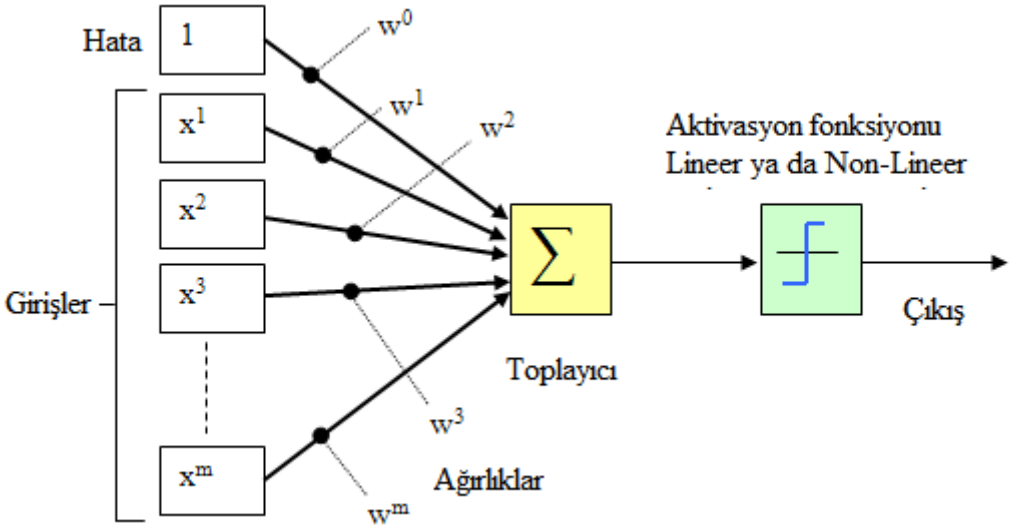
GeneKeyDB ise Atlas gibi farkı veri tabanlarından alınan ham veriyi veri madenciliği kuralları dahilinde işleyerek kullanıcısına sunan bir diğer yazılımdır (Kirov, 2005).

2.5 Yapay sinir ağları

Yapay sinir ağları bir optimizasyon modelidir. Doğası bilinmeyen ya da denklemler vasıtası ile açıkça matematiksel modeli ortaya konamayan problemlerin optimum çözümlerine ulaşmak için kullanılırlar. Yapay sinir ağları çözüme ulaşırken, eldeki veriler ile bu verilere karşılık gelen fiziksel gerçeklikten elde edilen verileri karşılaştırır ve oluşan hataların en aza indirir. Bu yönüyle yapay sinir ağları bir minimizasyon algoritması olarak karşımıza çıkar. Yapay sinir ağları, adını insan beyninin nöronal organizasyon ve hesaplama yeteneğine benzemesi yönü ile almıştır. Bu yapıda nöronların yerlerini nodlar, sinapsların yerlerini bağlantılar ve hafızanın yerini de ağırlıklar kullanılmıştır (Haykin, 1999). Yapay nöron (nod) fikri ilk defa 1958 yılında Cornell Aeronautical Laboratory ‘de Frank Rosenblatt tarafından ortaya atılmış ve bu modele perceptron adı verilmiştir. Her ne kadar perceptron ilk tanımlanan yapay nöron olarak anlatılsa da, perceptron aslında 1957 ile 1962 yılları arasında yapay sinir ağı ve öğrenebilen sistem çalışmalarının genel adı olarak karşımıza çıkar (Rosenbalt, 1958).

2.5.1. Perceptron

Teknik anlam olarak perceptron, giriş değerlerinin ağırlıklı toplamlarının bir fonksiyon üzerine projeksiyonunu sağlayan birime verilen addır. Bu bağlamda bir perceptron girişler, ağırlıklar, toplayıcı ve perceptron aktivasyon fonksiyonundan oluşur. En basit şekli ile bir perceptron Şekil 13 ‘de gösterilmiştir.



Şekil 13: Giriş, ağırlıklar, toplayıcı ve aktivasyon fonksiyonu kısımlarından oluşan genel bir perceptron.

Şekil 13 ‘de x ile perceptrona olan giriş değerleri, w ile perceptronun sahip olduğu ağırlık değerleri sembolize edilmiştir. Şekil 13 ‘deki hata biti (bias bit) tüm girişlerin 0 olması durumunda girişin bir bilgi değeri taşımamasını ve ağın konvergent (örtüşür) olması aşamasında olumsuz etkilenmemesi için konulmuştur. Bias sistematik bir hata olarak kabul edilir. Girişlere uygulanır. Görüldüğü gibi toplayıcı sisteme üzerlerinde w ağırlık değerleri olan bağlantılar ile bağlanmış durumdadır. Bu w ağırlıkları hafıza değeri olarak adlandırılabilir. Toplayıcı sistem her koldaki ağırlık değeri ile o kola gelen giriş değerini çarparak o kolun ağırlıklı değerini elde ederek bunların yığışmasını sağlar.

Toplam giriş aktivasyon fonksiyonundan geçirilerek perceptron biriminin çıkışı elde edilir. Aktivasyon fonksiyonları çok çeşitli seçilebilir. Lineer bir fonksiyon, nonlineer bir fonksiyon, bir parçalı fonksiyon burada görev alabilir. Genel bir kural olarak ağı oluşturacak perceptron ya da nodların hepsinde aynı aktivasyon fonksiyonu kullanılır. Farklı aktivasyon fonksiyonlarının farklı nodlarda kullanımı da teorik olarak mümkündür. Çok sık olarak kullanılan aktivasyon fonksiyonları aşağıda sıralanmıştır (Fausett, 1993).

1) Kimlik fonksiyonu

2) *K* eşikleme değerine sahip olan ikili adım fonksiyonları

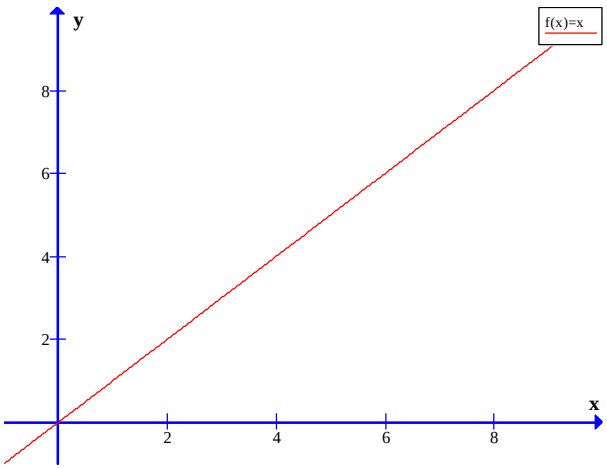
3) Sigmoid fonksiyon

4) Bipolar sigmoid fonksiyon

5)Hiperbolik tanjant fonksiyonu

Kimlik fonksiyonu oldukça basit lineer bir fonksiyondur. En genel hali Denklem 1 de gösterilmiştir. Bu fonksiyon genel olarak tek katmanlı ve tek perceptronlu ağlarda tercih edilir. Grafiksel şekli Şekil 14 ‘de gösterildiği gibidir.

$$f(x) = x \tag{1}$$



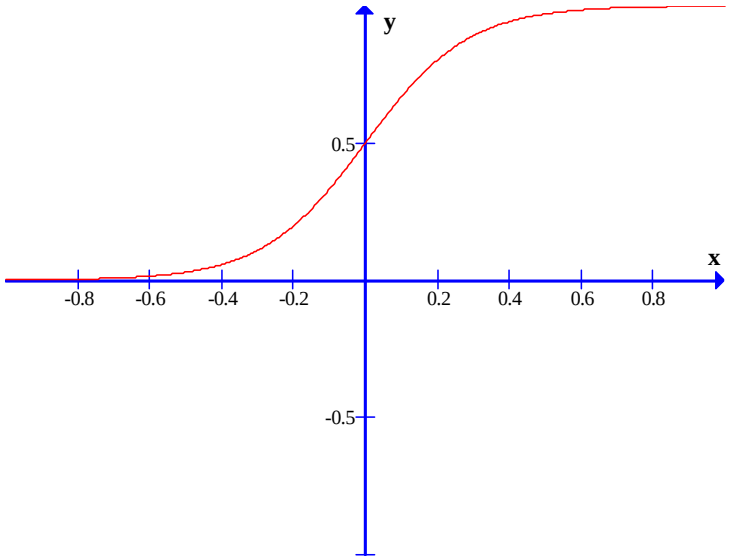
Şekil 14: Kimlik fonksiyonu grafiği

K eşik değerine sahip ikili adım fonksiyonu ise Denklem 2 ile gösterilmiştir. Bu denklem ile çıkışın; eğer giriş değeri belli bir *k* eşik değerinin altında ise 0 değeri, o eşik değerin üstünde ise 1 değerini alması sağlanmaktadır.

$$f(x) = \begin{cases} 1, x < k \\ ve \\ 0, x \geq k \end{cases} \quad (2)$$

Sigmoid fonksiyon sıklıkla kullanılan bir aktivasyon fonksiyonudur. Nonlineer bir yapıya sahiptir. Sigmoid fonksiyonun tanım aralığı (0,1) olup, bu aralıktaki her noktada türevlenebilen sürekli bir fonksiyon olması ile özel bir anlam ifade eder. Sigmoid fonksiyon Denklem 3 ile gösterilmiştir. Grafiksel şekli ise Şekil 15 'deki gibidir.

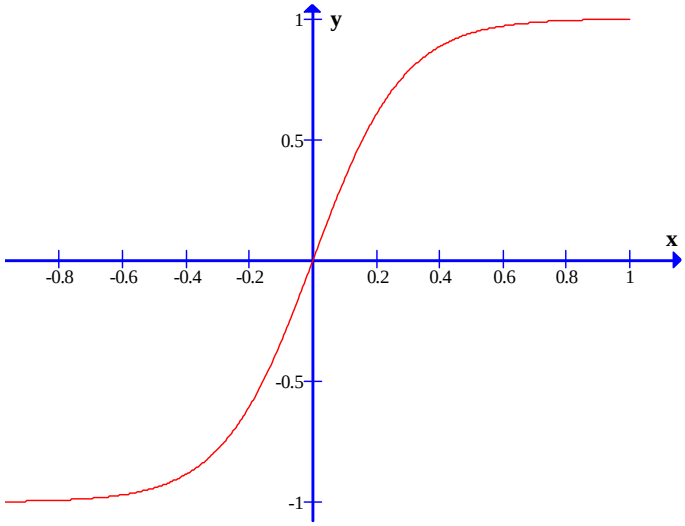
$$f(x) = \frac{1}{1+e^{-x}} \quad (3)$$



Şekil 15: Sigmoid fonksiyon grafiği

Bipolar sigmoid fonksiyon Denklem 4 ile gösterildiği gibi ifade edilir ve grafiksel şekli Şekil 16 'deki gibidir.

$$f(x) = \frac{1-e^{-x}}{1+e^{-x}} \quad (4)$$



Şekil 16: Bipolar sigmoid fonksiyon

Hiperbolik tanjant fonksiyonu ise Denklem 5 ile ifade edilir.

$$f(x) = \frac{e^{x} - e^{-x}}{e^{x} + e^{-x}} \tag{5}$$

2.5.2. Matematiksel model ve öğrenme kuralları

Toplayıcı sistem her koldaki ağırlık değeri ile o kola gelen giriş değerini çarparak o kolun ağırlıklı değerini elde ederek bunların yığışmasını sağlar. Bu toplama işlemi Denklem 3 ‘de gösterilmiştir.

$$T = x^{1}w^{1} + x^{2}w^{2} + \cdots + x^{m}w^{m} \tag{6}$$

$$T = \sum_{l=1}^{m} x^l w^l$$

T, toplam değeri ifade eder. Bu bir perceptrona olan net girişi ifade eder. X ile girişler ve w ile bu girişlerin sahip oldukları ağırlıklar sembolize edilmiştir. *T* değeri aktivasyon fonksiyonundan geçirilerek bir perceptron biriminin çıkışı elde edilir.

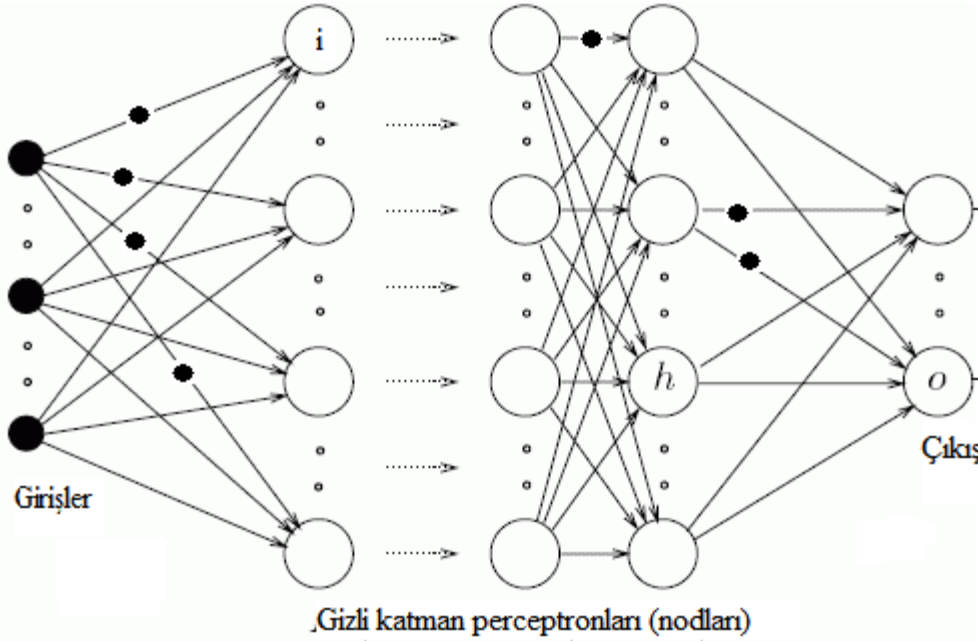
Denklem 3 ‘de verilen toplam giriş üzerinde Denklem 3 ile verilen sigmoid aktivasyon fonksiyonu perceptron ‘un ya da nodun çıkışını belirler. Denklem 7 ‘de bir nodun çıkışı gösterilmiştir.

$$O_x = \frac{1}{1+e^{T_x}} \qquad (7)$$

Burada $T_{x,}$ x noduna giren toplam giriş verisini; O_x ise x nodu için toplam giriş miktarına karşılık gelen çıkışı ifade eder. Buna aktivasyon çıkışı ya da nod çıkışı adı verilir.

Perceptronların katmanlar şeklinde organize olması ile modern anlamda yapay sinir ağları oluşmuştur. Yapay sinir ağlarında katmanlar üç çeşittir. Bunlardan ilki giriş katmanıdır. Giriş katmanı sadece bir tanedir. Giriş katmanından sonra orta katman yer alır. Orta katmana aynı zamanda gizli katman adı da verilir. Gizli katman en az bir tane olmak üzere teorik olarak bir sınırı yoktur. Çıkış katmanı gizli katmandan hemen sonra gelen katmandır ve bir tanedir. Bu anlatılan genel yapıya ağ topolojisi adı verilir. Ağ topolojisinin parametreleri giriş katmanındaki nod sayısı, gizli katman sayısı ve bu gizli katman ya da katmanların her birinde bulunan nod sayısı ile çıkış katmanının nod sayısıdır. Perceptronların bu şekilde katmanlar oluşturmuş ve belli bir topoloji altında

şekillendirilmiş hali çok katlı perceptron ya da güncel olarak çok katlı yapay sinir ağı adı verilir (Rossi ve Guez, 2005) . Şekil 17 ‘de perceptronlardan oluşmuş çok katlı yapay sinir ağının topolojisi gösterilmiştir.



Şekil 17: Çok katmanlı Perceptron ağı topolojisi

Perceptron denen nodların (işlem birimleri) birbiri ardına ve bir sonraki katmandaki nodlar ile birebir örten bir şekilde (nodlar arasında aynı katman içinde bir bağlantı olmadan) bağlantıda olduğu çok katlı yapay sinir ağı topolojisi. Şekil 17 ‘de siyah küçük nokta olarak gösterilen elemanlar nod ağırlıklarıdır. *i* ile verilen nodlar giriş birimleri, *h* ile gösterilen nodlar gizli katman birimleri, *o* ile gösterilen nodlar çıkış katmanı işlem birimleridir.

Tüm yapay sinir ağları temel olarak iki alt birime ayrılır. Bu birimler sırası ile danışmanlı ve danışmansız yapay sinir ağlarıdır. Danışmanlı olanlar ağa giren her bir giriş için verilen topolojiye uygun bir çıkış değeri üretir ve bunu fiziksel gerçeklikten gelen gerçek çıkış değerleri ile karşılaştırır. Danışmansız ağlar ise yine ağa gelen her giriş için

bir çıkış üretilir yalnız bu çıkış bir değer ile karşılaştırılmadan öğrenme kuralları doğrultusunda bir nod kazanan ya da doğru öğrenen nod olarak ağa etki eder. Danışmanlı öğrenme kuralı kullanan ağlara örnek olarak perceptron, geri yayılım ağları; danışmansız öğrenme kuralı kullanan ağlara örnek olarak Hopfield ağları, Kohonen ağları (öz örgütlemeli ağlar), Boltzman ağları örnek olarak verilebilir (Elmas, 2003).

Çok katmanlı geri yayılım ağı bir giriş değerine karşı bir çıkış üretmektedir. Bu çıkış hatalıdır ve istenilen çıkış değerinden de farklılık gösterir. Bu hatanın kaynağı nedir? Ağa etki eden her giriş bir çıkış değeri oluşturur. Bu çıkış değeri ise ağa etki eden girişlerin aktivayon fonksiyonundan geçirilmesi ile ortaya çıkar. Ve bu süreç giriş katmanından çıkış katmanına doğru devam eder. Bu şekilde gerçekleşen ileri yayılım sonucunda elde edilen ağ çıkışları ile fiziksel gerçeklikten gelen gerçek çıkış değerleri arasında fark vardır. Buna ağ hatası denir. Şu halde 3, 6, 9 denklemleri düşünüldüğünde bu hata ya girişlerden ya da ağ ağırlıklarından kaynaklanacaktır. Giriş değerlerinden kaynaklandığı düşünülürse ki giriş değerleri eğer ölçüm sonuçları olarak alınırsa hata olması doğaldır. Bu hatanın ağ topolojisi içinde düzeltilmesine geri yayılım ağlarında imkan yoktur. O halde hata ağırlıklardan kaynaklanmaktadır. Zira Denklem 9 ‘da çıkış aktivasyonunda bu hatanın giriş katmanından çıkış katmanına kadar taşındığı görülmektedir. İşte geri yayılım algoritmasının temel mantığı budur. Giriş nodlarından aktivasyon fonksiyonu ile giriş verilerinin ağırlıklardan kaynaklanan hatalarının ağa bu sefer ters yönde (çıkıştan girişe, buna geri yayılım denir) dağıtılarak taşınması algoritmanın çatısını oluşturur. Öğrenme algoritması da bu temel üzerine kurulmuştur. Denklem 10 ile öğrenme algoritması (hatanın ağa geri yayılması) gösterilmiştir.

$$\Delta = -\eta \frac{dE}{dw} \qquad (8)$$

Burada *E* ile yapay sinir çıkış katmanındaki nodların hataları, *w* ile ağın ağırlıkları ve η ile sabit kabul edilen öğrenme oranı ifade edilmiştir.

Görüldüğü üzere ağın ağılıklarının hata karşısındaki değişiminin minimum olması Denklem 10 ‘da söylenmektedir.

Ağa olan girişler *x* ile gösterilsin. O halde giriş katmanındaki nod sayısı *i* olarak alınırsa bu girişleri $x_o, x_1, x_2, \cdots, x_i$ şeklinde ifade edilir. Bir nodun aktivasyon çıkışı *y* olarak alınır ise giriş katmanında ki nodlar için çıkış değerleri $y_o, y_1, y_2, \cdots, y_i$ olarak tanımlanır. İlk gizli katmanda ki nod sayısı *j* ile verilir ise ve ağırlıkların da *w* olması durumunda; giriş katmanı birinci nodu ile ilk gizli katman arasında var olan ağırlıklar $w_{1,0}, w_{2,0}, w_{3,0}, \cdots, w_{j,0}$ olarak verilir. Benzer şekilde giriş katmanı *i* nodu ile birinci gizli katman *j* nodu arasında $w_{j,0}, w_{j,1}, w_{j,2}, \cdots, w_{j,i}$ ağırlıkları olacaktır. Çıkış katmanı nod sayısı *k* ile verildiğinde bu katmana olan girişler $x_o, x_1, x_2, \cdots, x_k$ ve bu katmandan olan çıkışlar $y_o, y_1, y_2, \cdots, y_k$ olacaktır. Her çıkış değerine karşılık gelen istenilen gerçeklik değerleri *d* ile ifade edilir ise çıkış katmanı istenilen değerleri $d_o, d_1, d_2, \cdots, d_k$ olarak yazılabilir. Çıkış katmanındaki bir nodun yaptığı hata miktarı *e* olarak alınırsa, bu hata Denklem 9 ile gösterildiği gibi yazılabilir.

$$e_k = d_k - y_k \tag{9}$$

Burada d ile istenilen çıkış değeri, y ile ağın çıkışı ve e ile o nodun hatası sembolize edilmiştir. Bu katmanın karesel hata fonksiyonu bu durumda Denklem 10 ile gösterildiği gibi yazılabilir.

$$E_k = \frac{1}{2} e_k^2 \tag{10}$$

Burada E ile karesel hata, e ile ise Denklem 9 ‘de verilen hata değeri sembolize edilmiştir. İlk gizli katman girişleri ve çıkış katmanı girişleri Denklem 6‘den yola çıkarak ve notasyona uygun bir şekilde Denklem 11 ‘de gösterildiği gibi yazılabilir.

$$x_j = \sum_i y_i w_{j,i} \tag{11}$$

$$x_k = \sum_j y_j w_{k,j}$$

Burada x ile bahsedilen noda olan toplam giriş, w ile ağırlıklar ve y ile bir önceki katmanın çıkış değerleri sembolize edilmiştir. Bu durumda *i* giriş katmanındaki, *j* gizli katmanındaki ve *k* çıkış katmanındaki nodların aktivasyon çıkışları olan *y* değerleri Denklem 12 ‘de gösterildiği gibi yazılabilir.

$$y_i = f(x_i)$$

$$y_j = f(x_j)$$

$$y_k = f(x_k) \tag{12}$$

Denklem 8 ‘de gösterilen değişim miktarı hatanın kaynağı olduğunda iteratif bir minimizasyon ile hesaplanması gerekecektir. Bu değer aynı zamanda ağırlık değişimini (hata karşısında) ifade etmektedir. O halde Denklem 8 ‘dan yola çıkarak Denklem 13 yazılabilir.

$$\Delta w_{k,j} = -\eta \frac{dE_k}{dw_{k,j}} \tag{13}$$

Burada η ile öğrenme oranı sembolize edilmiştir. Denklem 13 ‘de gösterilen kural ile hata son gizli katman ile çıkış katmanındaki ağırlıklara dağıtılır. Denklem 13 ‘deki işleme devam edilir ve Denklem 14 aşağıda verildiği şekilde yazılabilir.

$$\Delta w_{k,j} = -\eta \frac{dE_k}{dw_{k,j}} = -\eta \left[\frac{d}{dw_{k,j}}\left(\frac{1}{2}e_k^2\right)\right] = -\eta \left[\frac{d}{de_k}\left(\frac{1}{2}e_k^2\right)\frac{de_k}{dw_{k,j}}\right]$$

$$= -\eta \frac{1}{2} 2e_k \frac{de_k}{dw_{k,j}} = -\eta e_k \left[\frac{d}{dw_{k,j}}(d_k - y_k)\right]$$

$$\Delta w_{k,j} = -\eta e_k \left[\frac{d}{dw_{k,j}}(d_k - y_k)\right] = -\eta e_k \left[\frac{d}{dy_k}(d_k - y_k)\frac{dy_k}{dw_{k,j}}\right]$$

$$= -\eta e_k (-1)\frac{dy_k}{dw_{k,j}} = \eta e_k \frac{dy_k}{dw_{k,j}} = \eta e_k \left[\frac{d}{dw_{k,j}}\left(F(x_k)\right)\right]$$

$$= \eta e_k \left[\frac{d}{dx_k}\left(F(x_k)\right)\frac{dx_k}{dw_{k,j}}\right] = \eta e_k F'_{(x_k)}\frac{dx_k}{dw_{k,j}}$$

(14)

Aktivasyon fonksiyonu olarak eğer sigmoid fonksiyon Denklem 6 ‘ da gösterildiği şekli ile alınır ise, biz tezimizde aktivasyon fonksiyonu olarak tüm nodlarımızda aynı olmak üzere sigmoid fonksiyonu kullandık, bu fonksiyonun türevi olan $F'_{(x_k)}$ Denklem 15 de gösterildiği gibi yazılabilir.

$$F'(x) = \frac{1}{1 + e^{-x}}$$

$$= \frac{d}{dx}\left(F(x)\right)$$

$$= \frac{1(1 + e^{-x}) - (1 + e^{-x})'}{(1 + e^{-x})^2}$$

$$= \frac{-(-e^{-x})}{(1+e^{-x})^2}$$

$$= \frac{e^{-x}}{(1+e^{-x})^2}$$

$$F'(x) = \frac{1}{1+e^{-x}}\left(1 - \frac{1}{1+e^{-x}}\right)$$

$$= F(x)(1 - F(x))$$

(15)

Denklem 14 ‘e kalındığı yerden devam edilir ise ;

$$\Delta w_{k,j} = \eta e_k \left[\frac{d}{dx_k}\left(F(x_k)\right)\frac{dx_k}{dw_{k,j}}\right] = \eta e_k F'_{(x_k)}\frac{dx_k}{dw_{k,j}}$$

$$= \eta e_k F'_{(x_k)}\left[\frac{d}{dw_{k,j}}\left(\sum_j y_j w_{k,j}\right)\right]$$

$$= \eta e_k F'_{(x_k)}\left[\frac{d}{dw_{k,j}}\left(y_1 w_{k,1} + y_2 w_{k,2} + \cdots + y_j w_{k,j}\right)\right]$$

$$= \eta e_k F'_{(x_k)} y_j$$

O halde çıkış katmanı ile son gizli katman arasındaki ağırlıkların her iterasyonda bir öncekinden daha az hata yapacak şekilde yeniden ayarlanması için kullanılan $\Delta w_{k,j}$ için Denklem 16 kullanılabilir.

$$\Delta w_{k,j} = \eta e_k F'_{(x_k)} y_j$$

$\alpha_k = e_k F'_{(x_k)}$ buradan dönüşümü yapılır ise

$$\Delta w_{k,j} = \eta \alpha_k y_j \quad (16)$$

Ağırlıkların hesaplanan değer üzerinde yeniden ayarlanması Denklem 17 ‘den yola çıkarak Denklem 17 ‘de gösterildiği gibi yazılabilir.

$$\Delta_{w_{k,j}}^{yeni} = \Delta_{w_{k,j}}^{eski} + \Delta w_{k,j} \quad (17)$$

Bu tezde kullanılan modelin veri yapısı tasarlanırken her katmandaki nod sayısının kullanıcının isteği doğrultusunda ayarlanabilecek, gizli katman sayısının en az bir olmak üzere birden çok da kullanılabilecek şekilde oldukça geliştirilebilir şekilde yazılmıştır. Ayrıca sınıf yapısına uygun tasarlandığından birden çok ağı da tek bir ağmışçasına yönetebilme özgürlüğünü bize kazandırmıştır. Bu bağlamda matematiksel alt yapının daha net anlaşılması için tek bir gizli katman varmışçasına devam edilecektir.

Gizli katman ile giriş katmanı arasında var olan $w_{j,i}$ ağırlıkları için Denklem 10 ’dan yola çıkarak Denklem 18 yazılabilir.

$$\Delta w_{j,i} = -\eta \frac{dE_j}{dw_{j,i}} \tag{18}$$

Doğal olarak gizli katmanın kendine ait olan çıkışları y_j değerleri için fiziksel gerçeklikten gelme d_j istenilen değerleri olmadığından hata terimi yazmak gerekli olacaktır. Gizli katmandaki *j* nodunun hatasını yazmak için öncelikle bu nod için çıkış katmanındaki nodlar ve bunların hataları olan e_k değerleri düşünülmelidir. *j* nodu kendinden sonra gelen çıkış katmanındaki *k* noddan etkilendiğine göre ve bunlarında tahmini hatası E_j için denklem , Denklem 19 düzenlenebilir.

$$E_j = \sum_k \frac{1}{2} e_k^2 \tag{19}$$

Denklem 18 ve Denklem 19 ‘den faydalanarak işlemlere devam edilir ise Denklem 21 aşağıdaki gibi yazılabilir.

$$\Delta w_{j,i} = -\eta \frac{dE_j}{dw_{j,i}} = -\eta \frac{d}{dw_{j,i}} \left[\sum_k \frac{1}{2} e_k^2 \right]$$

$$= -\eta \frac{d}{de_k} \left(\sum_k \frac{1}{2} e_k^2 \right) \frac{de_k}{dw_{j,i}}$$

$$= -\eta \frac{d}{de_k} \left(\frac{1}{2} e_0^2 + \frac{1}{2} e_1^2 + \cdots + \frac{1}{2} e_k^2 \right) \frac{de_k}{dw_{j,i}}$$

$$= -\eta \frac{d}{de_k} (e_0 + e_1 + \cdots + e_k) \frac{de_k}{dw_{j,i}}$$

$$= -\eta \sum_k e_k \left[\frac{d}{dw_{j,i}} (d_k - y_k) \right]$$

$$= -\eta \sum_k e_k \left[\frac{d}{dy_k} (d_k - y_k) \frac{dy_k}{dw_{j,i}} \right]$$

$$= \eta \sum_k e_k \left[\frac{d}{dw_{j,i}} \left(F(x_k) \right) \right]$$

$$= \eta \sum_k e_k \left[\frac{d}{dx_x} \left(F(x_k) \right) \frac{dx_k}{dw_{j,i}} \right]$$

$$= \eta \sum_k e_k \cdot F'_{(x_k)} \left[\frac{d}{dw_{j,i}} \left(\sum_j y_j w_{k,j} \right) \right]$$

$$= \eta \sum_k e_k \cdot F'_{(x_k)} \left[\frac{d}{dy_i} \left(\sum_j y_j w_{k,j} \right) \frac{dy_i}{dw_{j,i}} \right]$$

$$\Delta w_{j,i} = \eta \sum_k e_k \,.\, F'_{(x_k)} \left[\frac{d}{dy_j} \left(y_0 w_{k,0} + y_1 w_{k,1} + \cdots + y_j w_{k,j} \right) \frac{dy_j}{dw_{j,i}} \right]$$

$$= \eta \sum_k e_k \,.\, F'_{(x_k)} w_{k,j} \left[\frac{d}{dw_{j,i}} \left(F(x_j) \right) \right]$$

$$= \eta \sum_k e_k \,.\, F'_{(x_k)} w_{k,j} \left[\frac{d}{dx_j} \left(F(x_j) \right) \frac{dx_j}{dw_{j,i}} \right]$$

$$= \eta \sum_k e_k \,.\, F'_{(x_k)} w_{k,j} F'_{(x_j)} \frac{d}{dw_{j,i}} \left(\sum_i y_i w_{j,i} \right)$$

$$= \eta \sum_k e_k \,.\, F'_{(x_k)} w_{k,j} F'_{(x_j)} \frac{d}{dw_{j,i}} \left(y_0 w_{j,0} + y_1 w_{j,1} \right.$$

$$\left. + \cdots y_i w_{j,i} \right) = \eta \sum_k e_k \,.\, F'_{(x_k)} w_{k,j} F'_{(x_j)} y_i \qquad (20)$$

Denklem 16 ‘deki gibi bir dönüşüm yapılır ise Denklem 21 elde edilir.

$$\alpha_k = e_k F'_{(x_k)}$$

$$\Delta w_{j,i} = \eta \sum_k \alpha_k \, w_{k,j} F'_{(x_j)} y_i$$

$$\alpha_j = \sum_k \alpha_k w_{k,j} F'_{(x_j)} \quad (21)$$

Giriş katmanı ile gizli katman arasındaki $w_{j,i}$ ağırlıklarının hata karşındaki değişimleri olan $\Delta w_{j,i}$ Denklem 22 'de gösterildiği gibi yazılabilir.

$$\Delta w_{j,i} = \eta \alpha_j y_i \quad (22)$$

Denklem 18 'den yola çıkarak giriş katmanı ile gizli katman arasındaki ağırlıkların yeniden düzenlenmesi için Denklem 23 yazılabilir.

$$\Delta_{w_{j,i}}^{yeni} = \Delta_{w_{j,i}}^{eski} + \Delta w_{j,i} \quad (23)$$

Buradan haraket ile birden çok gizli katman bulunması halinde ve yine gizli katmanlar alt indisi *j* olduğu halde tüm gizli katmanlar *b* tane şeklinde bir kabullenme ile Denklem 24 yazılabilir.

$$\Delta w_{j,j-1} = \eta \alpha_{j-1} y_{j-2}$$

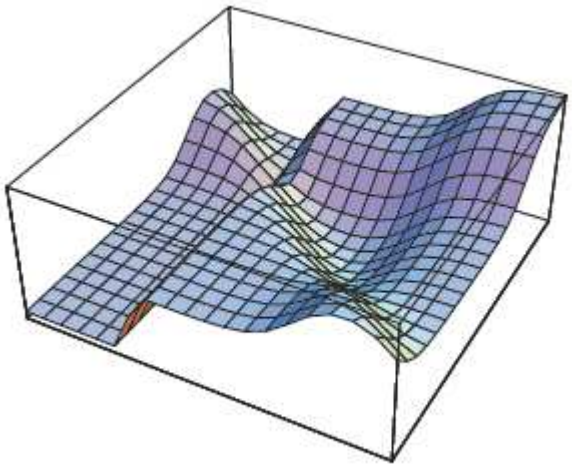
$$\Delta w_{j-1,j-2} = \eta \alpha_{j-2} y_{j-3}$$

$$\vdots$$

$$\Delta w_{j-b,i} = \eta \alpha_{j-b} y_i \quad (24)$$

3.6.3 Lokal minimum ve ezberleme

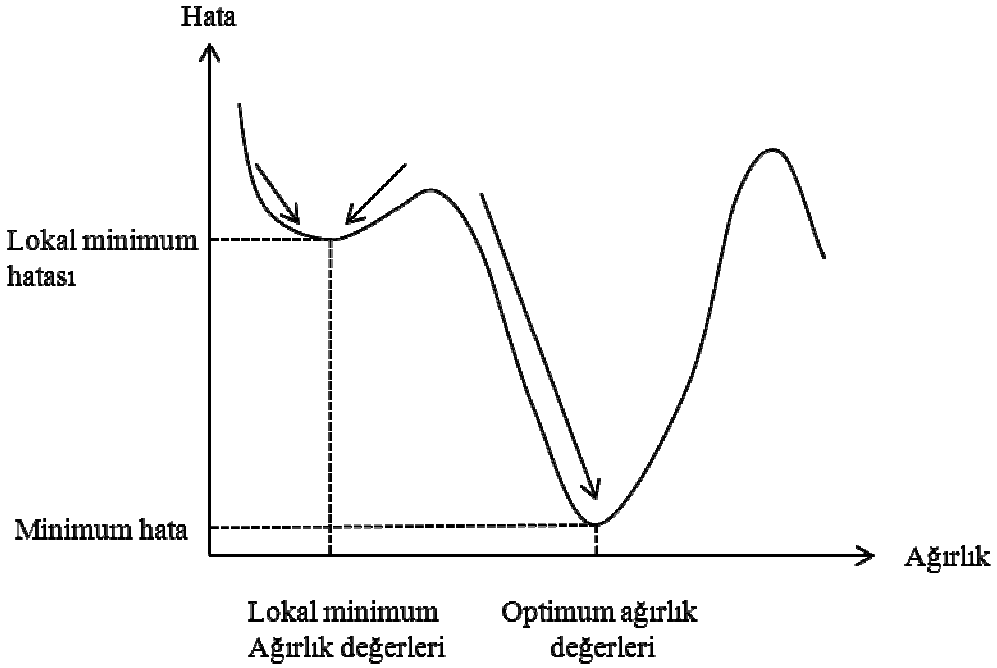
Bilindiği gibi yapay sinir ağları hatanın ağırlık yüzeyi üzerinde minimum hale getirilmesini amaçlar ve bunun için türevleri kullanır (geri yayılım ağları v.b). Ağırlık yüzeyinin son şekli ya da bağımlı olduğu değişkenlere göre kestirilebilir olması kanaatimce mümkün gibi görünmemektedir. Ağırlık yüzeyi problemin doğasına, giriş değerlerine, ağırlıkların başlangıçta atanan rastlantısal ilk değerlerine, iterasyon sayısına, kullanılan aktivasyon fonksiyonuna göre değişmektedir (Denklem 18 ve Denklem 13 uyarınca). Teorik bir yüzey Şekil 18 ‘ de gösterilmiştir.



Şekil 18: Teorik Hata ve ağırlık.

Başlangıç hata değerinden adım azaltarak hatayı ağırlık değişimi üzerinden minimuma indirmek isteyen geri yayılım ağının teorik hata yüzeyinin herhangi bir *t* iterasyon anında alınan projeksiyonu Şekil 19 verildiği gibi olsun. Burada başlangıç hatasından azalırken algoritmanın geçmesi gereken bir lokal minimum değeri lokal minimum ağırlık değerleri noktasında göze çarpmaktadır. Burada lokal arama yapan

algoritmanın bu hata düzeyinden kurtulup optimum hatanın oluştuğu optimum ağırlık değerleri noktasına gelememe ihtimali her zaman vardır. İşte bu durum geri yayılım ağlarında lokal minimuma düşme olarak tanımlanmıştır. Lokal minimuma düşmeyi önleyecek ve algoritmanın performansına etki edecek bazı yöntemler momentum terimi ilavesi, öğrenme oranının yeniden tayin edilmesi, öğrenme oranının adaptif bir şekilde çevrim içi hesaplanması bunlardan önemli görülen bazılarıdır (Baldi ve ark., 1989; Freeman, 1991).



Şekil 19: Geri yayılım ağlarında lokal minimuma düşme durumu

Ağ lokal minimum noktalarında ya da optimum ağırlık değerlerinde hata oranı çok çok küçük bir değer olana kadar, aynı set ile, yüksek iterasyon değerlerinde çalıştırılmaya devam edilir ise bu durumda ezberleme durumu ortaya çıkar. Bu tip bir ağ eğitim setinde %100 ‘e varan başarı elde ederken test setinde bu başarının çok çok altında kalır (Freeman, 1991).

2.6 Tek nükleotid polimorfizmi

İki insan genomu arasındaki benzerlik oranı %99.9 ‘dur (Cooper ve ark., 1985). 3.2 milyar baz çifti içinde bu oran 3.2 milyon farklılık anlamına gelmektedir. Bu farklılıkların çoğu SNP kökenli farklılıklardır. SNP ‘lerin pek çoğunun herhangi bir biyolojik etkisi olmasa da SNP’ler canlılık açısından çeşitliliğin temelini oluşturur. Bugün için 6.944.000 adet SNP ‘i tanımlanmıştır (http://www.ncbi.nlm.nih.gov/SNP/snp_summary.cgi, 07.02.2011). SNP ‘ler toplumda %1 ‘den daha sık olarak görülen DNA ‘daki tek baz değişimlerini simgeler. Farklı değişim tipleri nedeni ile farklı SNP tipleri tanımlanmıştır (Vignal ve ark., 2002). Bu değişim tipleri aşağıda gösterilmiştir.

1) Pürin-Pürin değişimi (A – G),
2) Pirimidin-Pirimidin değişimi (C – T),
3) Pürin-Pirimidin (A – C, A – T),
4) Pirimidin-Pürin değişimleri (G – C, G – T).

Genetik sonuçları açısından bakıldığında sinonim SNP ‘ler ve sinonim olmayan SNP ‘ler olarak ayrılırlar. Sinonim olmayan SNP ‘ler DNA ‘ da meydana gelen değişikliğin bir öncekine göre protein seviyesinde (kodlanan amino asit değişir) değişime neden olma durumunu anlatır. Sinonim SNP ‘lerde ise bu değişim olmaz.

2.7 Dizi hizalama

Dizi hizalaması, DNA, RNA veya protein dizilerini düzenleyerek benzer bölgelerin tespit edilmesidir. Bu bölgelerin benzer olması, diziler arasında işlevsel, yapısal veya evrimsel bir ilişki olduğu anlamına gelir (Mount, 2004). Hizalanmış nükleotit veya amino asit rezidü dizileri tipik olarak bir matriksin satırları olarak gösterilir. Kimyasal rezidüleri temsil eden harflerin arasına boşluklar konarak ardışık sütunlarda yer alan aynı veya

benzer harflerin bir hizada olması (alt alta gelmesi) sağlanır. Hizalama Şekil 20 ‘de gösterilmiştir.

```
AAB24882        TYHMCQFHCRYVNNHSGEKLYECNERSKAFSCPSHLQCHKRRQIGEKTHEHNQCGKAFPT 60
AAB24881        --------------------YECNQCGKAFAQHSSLKCHYRTHIGEKPYECNQCGKAFSK 40
                                     ****: .***:  * *:** * :****.:* *******..

AAB24882        PSHLQYHERTHTGEKPYECHQCGQAFKKCSLLQRHKRTHTGEKPYE-CNQCGKAFAQ- 116
AAB24881        HSHLQCHKRTHTGEKPYECNQCGKAFSQHGLLQRHKRTHTGEKPYMNVINMVKPLHNS 98
                 **** *:***********:***:**.: .***************    :  *.: :
```

Şekil 20: Clustal W ile elde edilmiş iki insan çinko parmak proteininin hizalama sonucu (Clustal W ile elde edilmiştir).

İki çeşit hizalama algoritması bulunmaktadır. Bunlardan ilki yerel dizi hizalama algoritmasıdır (local sequence alignment), diziler segmentler şeklinde yüksek hizalama skorlarını türetecek biçimde hizalanmaya çalışılır. Bu şekilde yerel hizalamaya ayrıca Smith Waterman algoritması da denir (Smith ve Waterman, 1981). İkincisi ise global dizi hizalama algoritmasıdır (global sequence alignment), diziler burada bir bütün şeklinde yüksek hizalama skorlarını türetecek biçimde hizalanmaya çalışılır. Bu şekilde global hizalamaya ayrıca Needleman Wunsch algoritması da denir (Needleman ve Wunsch, 1970). Şekil 21 ‘de benzer iki dizinin yerel ve global hizalama sonuçları gösterilmiştir.

```
GLOBAL  FTFTALILLAVAV
        F--TAL-LLA-AV

YEREL   FTFTALILL-AVAV
        --FTAL-LLAAV--
```

Şekil 21: Global ve yerel hizalama ve sonuçları

2.8 Proteinlerin ikincil yapılarının tahminine yönelik çalışmalar

Proteinlerin üçüncül yapılarının aslında birincil yapının bir derece katlanması ve ikincil yapıyı oluşturması ve ikincil yapının da bir derece katlanması neticesinde oluştuğunu bir önceki bölümde değinmiştik. Bu nedenle ikincil yapının etkili bir şekilde tahmin edilebilmesi oldukça önem kazanmış ve bu soru biyoinformatikçilerin aklına geldiği günden beri ikincil yapı tahmini konusunda pek çok çalışma yapılmıştır. İkincil yapının tahmini bu nedenle biyoinformatiğin çözümü en zor olan konularından birisi halini alıp “Holly Grail” (kutsal kase) takma ismiyle anılmasına neden olmuştur.

Gelişmiş moleküler biyoloji laboratuvarlarında DNA ‘nın hızlı bir şekilde dizilenmesi ve protein dizilerinin de benzer şekilde ortaya konması neticesinde genom projesinden sonra önümüzde ikinci bir kapı açmıştır. Aynı zamanda proteinlerin yapısını belirlemek için gereken süre de kısalmıştır. Bugün için kullanılan iki teknikten ilki X-ışını kırınımı (X-ray) olup pahalı, uzun zaman isteyen bir işlemdir. İkincisi, yüksek çözünürlüklü nükleer manyetik rezonans (NMR) tekniği olup oldukça pahalı, uzun zaman alan, oldukça karmaşık hesaplama kabiliyeti gerektiren, kısa polipeptit yapılarında etkili (uzunluk arttıkça güvenilirliği düşen) ve yüksek miktarlarda saf protein örneği isteyen bir yöntemdir. Her iki yöntemin de icrasında belli başlı iki problem para ve zamandır. Bu nedenle hesaplamalı biyolojik yöntemlere olan ilgi giderek artmıştır (Stephen ve ark., 1990).

İkincil yapıyı tahmin etmek için pek çok yöntem kullanılmıştır. Amino asitlerin ikincil yapı elementlerini tercih edip etmemeleri üzerine kurulu olan istatistiki yöntemler, yapay sinir ağları, ve moleküler dinamik ve minimum enerji simülasyonları, genetik algoritmalar bunlardan dikkati çeken bazılarıdır.

2.8.1 Tahmin yöntemleri

İstatistiki yöntemler üzerinde en çok durulan Chou-Fasman metodudur. Bu yöntem amino asitlerin ikincil yapı elementlerinde bulunmalarının göreceli olasılıkları ile bu olasılıklar üzerinden yazılmış ampirik kuralların işletilmesi ile ikincil yapıyı tahmin etmeye çalışır.

Proteinleri oluşturan amino asitlerin pozisyon seçimleri üzerine yapılan bağıl frekans analizleri neticesinde Chou-Fasman yapısal parametreleri elde edilmiştir. Pek çok yazar tarafından veri kümeleri büyütülerek parametreler tekrar hesaplanmış olup bu özelliği ile istatistiki yöntemler içinde üzerinde en çok durulanı olmuştur.

Orijinal parametreler, yapısı X-ray ile aydınlatılmış 15 protein ve bu proteinlerden elde edilen toplan 2473 adet amino asit üzerinde yapılmıştır. Bu çalışmada ikincil yapı elementleri olarak alfa heliks, beta tabakalanma ve halka yapısal formasyonu olmak üzere üç sınıf oluşturulmuştur. Sınıflara göre amino asitlerin frekansları bulunduktan sonra bir "k" yapısal sınıfındaki bir "a" rezüdüsü için frekans denklemi olarak Denklem 25 ve benzer şekilde ortalama frekansı da Denklem 26 'de gösterildiği gibi formüle edilerek hesaplamalar yapılmış; bu yöntem ile Chou-Fasman yapısal parametreleri bulunmuştur. Yapısal parametreler Tablo 3'de gösterilmiştir.

$$F_k^a = \frac{n_{a,k}}{\sum_k n_k} \tag{25}$$

$$\langle F_a \rangle = \frac{n_a}{\sum_a n_a} \tag{26}$$

Burada $n_{a,k}$, a 'nın k yapısal sınıfında bulunma frekansını ve n_k ise k yapısal sınıfının toplam sayısını gösterir. n_a, a 'nın toplam sayısını ifade eder.

Tablo 3: Chou-Fasman yapısal parametreleri (*)

Amino asitler	P(a)	P(b)	P(t)	f(i)	f(i+1)	f(i+2)	f(i+3)
Alanin	142	83	66	0,060	0.076	0.035	0.058
Arjinin	98	93	95	0,070	0.106	0.099	0.085
Asparajin	67	89	156	0,161	0.083	0.191	0.091
Aspartik asit	101	54	146	0,147	0.110	0.179	0.081
Sistein	70	119	119	0,149	0.050	0.117	0.128
Glutamik asit	151	37	74	0,056	0.060	0.077	0.064
Glutamin	111	110	98	0,074	0.098	0.037	0.098
Glisin	57	75	156	0,102	0.085	0.190	0.152
Histidin	100	87	95	0,140	0.047	0.093	0.054
Izolösin	108	160	47	0,043	0.034	0.013	0.056
Lösin	121	130	59	0,061	0.025	0.036	0.070
Lizin	114	74	101	0,055	0.115	0.072	0.095
Methionin	145	105	60	0,068	0.082	0.014	0.055
Fenilalanin	113	138	60	0,059	0.041	0.065	0.065
Prolin	57	55	152	0,102	0.301	0.034	0.068
Serin	77	75	143	0,120	0.139	0.125	0.106
Threonin	83	119	96	0,086	0.108	0.065	0.079
Triptofan	108	137	96	0,077	0.013	0.064	0.167
Tirozin	69	147	114	0,082	0.065	0.114	0.125
Valin	106	170	50	0,062	0.048	0.028	0.053

*(http://prowl.rockefeller.edu Amino Acid Information aainfo/contents.htm bölümden ve http://expasy.org/tools/protscale adresinden uyarlanarak alınmıştır).

Bir proteinin ikincil yapısını bu yöntemle tahmin edebilmek için aşılması gereken büyük bir engel bulunur. Bu engel yapının büyümesidir. Bir yapıya karar verildikten sonra yapıya eklenen her amino asit için tekrar ve tekrar işletilmek üzere ampirik kurallar yazılmıştır. Bu sayede ikincil yapı tahmin edilmiştir. Bu kurallar dizisi Chou - Fasman Ampirik Kuralları olarak bilinir. Bu kurallar özet olarak aşağıda sunulmuştur.

Chou-Fasman ampirik kuralları:

Kural 1: Polipeptit zinciri boyunca 4 rezüdünün heliks yapısal parametresi eğer 1 ‘den büyük ise bu durumda burada helikal bir çekirdeklenme vardır.

Kural 1.1: Helikal segment her iki yöne doğru tetrapeptit için heliks yapısal parametresi değeri 1 ‘in altına düşene dek yayılımına devam eder. Bu değer 1 ‘in altına düştüğü anda helikal segment o noktada sonlanır.

Kural 1.2: Prolin özel olarak helikal segmentin c-terminal uç noktasında iç helikal segment içinde bulunmayı pek tercih etmediğinden heliks kırıcı olarak bilinir.

Kural 1.3: Pro, Asp, Glu amino asitleri n-terminal uç sahasını tercih etme eğiliminde iken ; His, Lys, Arg amino asitleri c-terminal uç sahasında bulunmayı tercih ederler.

Süperpoze (üstüste çakışan) olan özel segmentler için diğer bir ampirik kural ise;

Bir polipeptit zinciri boyunca eğer heliks yapısal parametresi 1.03 değerinden büyük iken heliks yapısal parametresi beta tabaka yapısal parametresinden büyük ise bu bölge heliks yapısındadır.

Kural 2: 3 rezüdünün beta tabaka yapısal parametresi 1 ‘den büyük ise bu noktada beta tabaka çekirdeklenmesi vardır.

Kural 2.1: Beta tabaka çekirdeklenmesi her iki yöne doğru yayılma eğilimdedir. Bu yayılım esnasında beta tabaka yapısal parametresi ortalaması 1 ‘in altına düştüğü anda beta tabakalanma son bulur.

Kural 2.2: Glu çok nadir olarak beta tabaka formasyonunda bulunma eğilimindedir. Pro benzer bir tercih yapar fakat yine de beta tabaka formasyonunda görülebilir.

Kural 2.3: Yüklü rezüdüler çok nadir olarak beta tabaka içinde bulunma eğilimindedirler. Fakat beta tabakanın n-terminal ve c-terminal uç noktalarında nadir olarak bulunurlar.

Süperpoze olan özel segmentler için bir diğer ampirik kural:

Bir polipeptit zinciri içinde 5 ya da daha çok amino asit içeren bir segment için beta tabaka yapısal parametresi 1.05 değerinden büyük iken beta tabaka yapısal parametresi eğer heliks tabaka parametresinden büyük ise bu segment için beta tabakalanma gösterir tahmini yapılmalıdır.

Bu şekilde ampirik kurallar ve bağıl olasılıklara bağlı olarak hesaplanmış yapısal parametreler ile yapılan ikincil yapı tahmininin başarı oranı %50 olarak hesaplanmıştır (Chou ve Fasman, 1974a, 1974b; Chou ve Fasman 1978a, 1978b).

İstatistiki yöntemlerde üzerinde çok çalışılmış bir diğer yöntem ise GOR algoritmasıdır. GOR adını araştırmacıları olan Garnier – Osguthorpe - Robson ‘dan almıştır. İlk defa 1978 yılında önerilmiştir. Erken dönem ikincil yapı tahmin algoritmaları arasında kendisini günümüze taşıyabilmiş tek örnek olması nedeniyle önemlidir. GOR-I ile başlayan sürümler günümüze GOR-IV olarak taşınmıştır.

Algoritmaların temel prensibi informasyonel teoriyi kullanarak yapısı aydınlatılmış 26 protein (GOR-I) üzerinden elde edilen frekanslardan yola çıkarak bir amino asidin bir yapısal elementte bulunma eğilimini hesaplamaya dayanır. Bir amino asidin herhangi bir k yapısal formunda bulunma eğilimi Denklem 27 ‘de gösterilmiştir (Garnier ve ark. 1978).

$$I(k,i) = ln\frac{P_{k,i}}{P_k} \quad (27)$$

GOR algoritması tek bir amino asitin k durumunda bulunma eğilimine göre çalışmaz. Belli boyutlarda ve opsiyonel olan pencerelerin dizi üzerinde kaydırılması ile sonuca gider. Bu pencerenin merkezi konumunda i rezüdüsü ise tahminin yapıldığı rezüdüdür. GOR algoritmalarının ortalama başarı oranları %60 olarak tahmin edilmiştir. 267 adet protein ve bu proteinlerdeki ikincil yapı elementleri pencere boyutu 17 seçilmek suretiyle amino asidin herhangi bir k yapısal formunda bulunma eğilimi hesaplanmış; bu veriler üzerinde yapılan tahminlerde GOR-IV 'ün başarısı %64.4 olarak hesaplanmıştır (Garnier ve ark., 1996).

Chou-Fasman ve GOR yöntemleri istatistiki yöntemlerin temel halleridir. Daha sonra veri kümeleri artırılıp yeniden çalışılarak farklı yapısal parametreler hesaplanmış, farklı bağıl eğilimler hesaplanmış olmasına rağmen uzun menzildeki etkileşimler ve tahmin edilecek dizinin uzun olması gibi durumlarda bu yöntemler başarısız olmuştur.

Yapay sinir ağları, doğası bilinmeyen ve fakat elde problemin ve sonuçlarının olduğu, neden ve sonuç ilişkisinin deterministik olarak ortaya konulamadığı durumlarda; çözümü, hataların her seferinde bir önceki noktadan daha az olmasının sağlanması kuralı ile bulmaya çalışan algoritmalardır. İkincil yapının birincil yapıdan nasıl oluyor da katlanıyor tümcesi deterministik bir şekilde açıklanmaz fakat elde birincil yapıları ve ikincil yapıları aydınlatılmış pek çok örnek bulunur. Yapay sinir ağları (YSA) bu örnekleri alarak uygun bir şekilde her iterasyonda (tekrarda) hatayı minimum yapacak olan YSA ağırlıklarını hesaplamak suretiyle çözüme ulaşmaya çalışır. Problemin değişken doğası

düşünülür ve ardından YSA ‘ların değişken durumlarda eldeki veri kümesi üzerinde tahmin yapabilme güçleri de hesaba katılırsa YSA problemin çözümü için uygun bir yöntem olarak karşımıza çıkar. Yapay sinir ağları yapıları itibari ile biyoinformatiğin zor sorularına her basamakta en iyi çözüm gibi görülmektedir (Mutasem ve ark., 2009).

Bu yöntemlerden iyi bilinen bir tanesi ikincil yapı tahmini için kullanılan PHD algoritmasıdır. PHD sistemi çoklu dizi hizalama sonuçları ile yapay sinir ağlarının bir kombinasyonunu kullanarak ikincil yapı tahmini yapmaktadır. PHD ‘ye örnek bir protein verildiği anda PHD bunun tüm homologlarını bulur. Bu sayede ağ için sadece amino asit dizisinin taşıdığından daha fazla bilgi sağlanmış olur. Burada iki katmanlı YSA kullanılmıştır. Yapısı bilinen proteinlerden 130 protein seçilmiştir. Bu proteinler veri kümesi addedilip diziden yapıya tahmin sonra yapıdan yapıya tahmin için eğitilen ağları kullanmıştır. Bu ağların çıktısı bir jüri sisteminden geçirilerek nihai tahmin elde edilmiştir. PHD ‘nin globüler yapıdaki proteinler için başarısı %70.8 olarak hesaplanmıştır (Rost ve Sander, 1993). Benzer şekilde transmembran proteinlerindeki heliks yapı tahmin edilmeye çalışılmış ve burada yaklaşık %95 başarı elde edilmiştir (Rost ve ark., 1994).

Bir çeşit yapay sinir algoritması olan geri yayılım algoritması ile globüler proteinlerin ikincil yapıları aydınlatılmaya çalışılmıştır. İkincil yapılar heliks, beta tabaka ve sarmal şeklinde düşünüldükten sonra yapay sinir ağı ile işlenmiş ve sonuçta %64.3 ortalama başarı elde edilmiştir (Qian ve Sejnowski, 1988). Bu çalışmada toplam 106 protein kullanılmıştır. Eğitim seti için toplam 18105 rezüdü seçilmiştir. Ağ topolojisi çok katlı geri yayılım ağına uygun olacak şekilde 13 amino asitlik pencerenin ana dizi üzerinde kaydırılmasından elde edilen giriş katmanı ile gizli katmanda toplam 40 nod çıkış katmanında ise toplam 3 nod olacak şekilde ayarlanmıştır.

Başka bir çalışmada 62 protein kullanılmıştır. İlk 48 protein eğitim setini oluşturmuş (eğitim setinde toplam 8315 rezüdü bulunur), son 12 protein ise test seti olarak kabul edilmiştir (test setinde toplam 2441 rezüdü bulunur). Pencere boyutu olarak 17 amino asit kullanılmış ve medyan rezüdünün sınıfı tahmin edilmeye çalışılmıştır. Başarı oranı %63 olarak hesaplanmıştır (Holley ve Karplus, 1988).

Java Object Oriented Neural Network Engine (JOONE), kullanılarak yapılan bir çalışmada heliks yapılanması için tahmin başarısı %71, beta tabaka yapılanması için tahmin başarısı %65 olarak hesaplanmıştır (Mottalib ve ark., 2010).

Standart grupların dışında geliştirilmiş gruplarda tahmin yapmaya dayalı bir başka çalışmada sınıflar alfa heliks, beta tabaka, paralel beta tabaka, anti paralel beta tabaka, beta köprü, 3/10 heliks, pi heliks ve dönüşler olarak bildirilmiştir. Bu çalışmada yapay sinir ağları ve geri yayılım algoritması kullanılmış olup ortalama hatalar 0.08 ile 0.02 arasında olup öz tutarlılık testi ile total ortalama hata 0.022, bağımsız veri kümesi ile yapılan testlerde total ortalama hata 0.025 olarak verilmiştir (Cai ve ark., 2002; Cai ve ark., 2003).

Geri yayılım algoritması kullanan bir diğer çalışmada, üç yapısal sınıf olan H, S, C tahmin edilmeye çalışılmıştır. 11 proteinin kullanıldığı bu çalışmada başarı oranı %79 olarak verilmiştir (Yadav ve ark., 2010).

Bu öncü çalışmalar doğrultusunda pek çok çalışma yapılmış bunlardan bazıları ve bu çalışmaların kısa tanımları ile başarı oranları Tablo 4’de gösterilmiştir (Burkhard ve ark., 1999; Burkhard R., 2001; Yılmaz, 2003).

Tablo 4: İkincil yapı tahmini için kullanılan yöntemler ve bunların başarısı

Metod Adı	**Başarı oranı**	**Kısa açıklama**
PROF	77.0	Çoklu basamaklı sınıflandırıcı, lineer ve ikinicil derece dağılımları kullanır.
SSpro	76.3	Yapay sinir ağı tabanlı ikincil yapı tahmini yapar. İki yönlü yapay sinir ağı sınıflandırıcısı kullanır.
PHD	71.9	Yapay sinir ağı tabanlı ikincil yapı tahmini yapar. Diziden yapıya ve yapıdan yapıya seviyesinde kombine yapay sinir ağı kullanılmıştır.
PHDsec	72.2	Hidrojen bağlarının formasyonları ve çoklu dizi hizalama bilgilerini yapay sinir ağı sınıflandırıcıları ile işler.
GOR-IV	64.5	17 amino asitlik pencerelerde mümkün olan tüm ikili frekansların hesaplanması şeklinde informasyonel teoriyi kullanır.
SOPM	70.0	Çeşitli tahmin programları ortak olarak kullanılmıştır.
SSPRED	70.0	İstatistiki metodlar ve çoklu dizi hizalama yöntemlerini kullanarak ikincil yapı tahmini yapar.
PSIPRED	76.5	Yapay sinir ağı tabanlı ikincil yapı tahmin algoritmasıdır.
JPred2	75.2	Çeşitli tahmin programları ortak olarak kullanılmıştır.
PHDpsi	75.1	Dizi hizalama tabanlı tahmin yapar
NSSP	71.0	Çoklu dizi hizalama ve en yakın komşuluk yöntemlerinin bir kombinasyonudur.
GOR V	73.5	GOR-IV algoritmasındaki 17 amino asitlik sabit pencere boyutu farklı boyutlara da imkan tanıyacak şekilde değiştirilmiştir. Bu sayede GOR-IV ‘ün gelişmiş bir sürümü halini almıştır. Tahmin yöntemi informasyonel teoridir.
DSC	70.0	Rezüdülerin konformasyonel pozisyonları seçip seçmemesi üzerine kurulu bir yöntemdir.
NNPREDICT	65	Yapay sinir ağı tabanlı ikincil yapı tahmin algoritmasıdır.

Genetik algoritmalar kullanılarak ikincil yapı tahmini yapılmaya çalışılmıştır. Genetik algoritmalar devasa çözüm kümesi bulunan örneklerde çözüm kümesinin evrimsel yöntemler kullanılarak, belirlenmiş kurallara göre en iyileme yapılarak olası çözümlerin azaltılmasını amaçlar. Öncelikle çözüm evrenini temsil ettiğine inanılan bir çözüm kümesi

kromozomlara yüklenir. Kromozomlarda mutasyonlar ve delesyonlar ile duplikasyonlar ve karşılıklı parça değiş tokuşu yapılarak eldeki kümeden yeni çözüm kümesi elde edilir. Çözüm kümesi enerji fonksiyonundan geçirilir. Belirlenen eşik değeri yakalamayan bireyler popülasyondan silinir. Eldeki yeni oluşturulmuş elit çözüm kümesi üzerinde rekombinasyonlar ile yeni nesiller oluşturulur. Oluşan yeni nesiller tekrar genetik operatörlere maruz bırakılır. Bu sayede iterasyonel olarak baştaki çözüm kümesi daha yüksek skor ya da daha düşük enerji seviyesine sahip olacak ve bir öncekinin de bir alt kümesi olarak kalacak şekilde daraltılır (Özkaçar, 1998).

Bu yöntemde genelde amino asitler hidrofobik ve hidrofilik olarak sınıflandırılır. Sınırları belirlenmiş bir kafes içinde rastlantısal olarak dağıtılır. Bu rastlantısal dağılım skoru hesaplanır. Skor hesaplanırken genelde yazılı kurallar kullanılır. Bu kurallardan en iyi bilineni hidrofobik amino asitlerin protein yapının merkezinde ve çözücüden uzakta, hidrofilik amino asitlerin de proteinin yüzeyinde ve çözücüye yakın olarak konumlanmasıdır. Bu yazılı kurallar genelde hidrofobik amino asitlerin birbirine komşu olması halinde skorun belli bir oranda artması, hidrofilik olanların yan yana olması durumunda skorun belirlenen bir oranda artması ve her iki durum için aksi bir formasyon varsa skorun belli oranlarda azalması ve sonuçta bu iki skorun toplanması ile tek bir skor elde edilmesi şeklindedir. Yüksek skorlu bireyler bir sonraki jenerasyon içine alınır. Diğerleri popülasyondan silinir. Genetik algoritmaların belki de en büyük avantajı tek bir sonuç değil de olası en iyi sonuçları toplu şekilde elde etmeye olanak sağlamasıdır (Unger ve Moult, 1993; Lesh ve ark., 2002; Huang, 2004; Custodio ve ark., 2004; Unger, 2004).

İkincil yapının oluşması için amino asitlerin kararlı ikincil yapı elementlerini oluşturması gereklidir. Bu kararlılık molekülü oluşturacak olan amino asitlerin uzaysal koordinatları ve bu koordinatlarda sahip olduğu enerji düzeyine bağlıdır. Bu süreç yani

proteinin doğal şekline kavuşması protein katlanması olarak adlandırılır. Bu olayı bilgisayar ortamında canladırmak için kullanılan yöntemlere ise *ab initio* (başlangıçtan başlayarak moleküler dinamik kurallarını uygulamak) yöntemler adı verilir. Bu metodun amacı protein katlanmasındaki biyolojik işlemleri taklit ederek kararlı bir üç boyutlu yapı önerebilmektir. Sentez tek tek amino asitlerin gelmesi ile başlar; farklı kaynaklardan kaynaklanan ve potansiyel enerji fonksiyonunu yazmak için kullandığımız kuvvetlerin oluşturduğu kuvvet alanında, enerjinin minimum olduğu konformasyonel pozisyon aranır .Enerjinin minimum hale getirilmesi için Newton kuralları ve Monte Carlo simülasyonları kullanılır. Moleküler dinamik kuralları üzerinden ikincil yapı tahmininde ikinci yol ise ampirik enerji fonksiyonu yazmaktır. Bu yöntemde proteine bir başlangıç konformasyonu atanır ve bu yapı ampirik enerji fonksiyonu ile her seferinde ampirik enerji minimum olacak şekilde değiştirilmeye çalışılır. Bu tip yöntemlerin hemen hepsinde iki büyük problem dikkati çekmiştir. Bunlardan ilki enerji fonksiyonunun seçimi ve bu seçim esnasında rol alacak kuvvetlerin belirlenmesidir (hidrojen bağları, kovalan bağlanmalar, disülfid bağları, hidrofobik kuvvetler, hidrofilik kuvvetler vb.). ikinci büyük problem ise inanılmaz derecede çok ve farklı uzaysal konformasyonun bulunmasıdır.

Proteinlerin doğal durum katlanmaları bilindiği üzere fiziksel gerçekliğe uygun bir şekilde olmaktadır. Protein yapısına katılan her amino asit için birbirleri ile yaptığı bağların çeşitliliği bir tarafa bırakılır ve bu olası konformasyonların sadece 3 tane olduğu varsayılırsa; karşımıza şu tablo çıkacaktır. Örnek olarak 101 amino asit ve her amino asitin bağ yapma çeşitliliği de 3 olmak üzere bu polipeptitte toplam 3^{100} farklı olası konformasyon bulunacaktır (Levinthal, 1969). 3^{100} farklı konformasyondan 10^{13} tanesi 1 saniyede tasarlanmış olsa dahi; tüm konfügürasyonu hesaplamak $3x10^{20}$ saniye alacaktır ($5x10^{47}$ saniiye). Buda yaklaşık olarak 10^{27} ($5x10^{47}/10^{13}$ yıl) yıla karşılık gelecektir. Oysa

hücre içinde ilerleyen protein sentezinde, proteinin minimum enerjili doğal durum konformasyonuna kavuşması milisaniyeler almaktadır. İşte bu bir paradokstur. Bu paradoks “Levinthal Paradooksu” olarak bilinir. Ab initio yöntemlerin önündeki en büyük problem budur (Zwanzig ve ark., 1991).

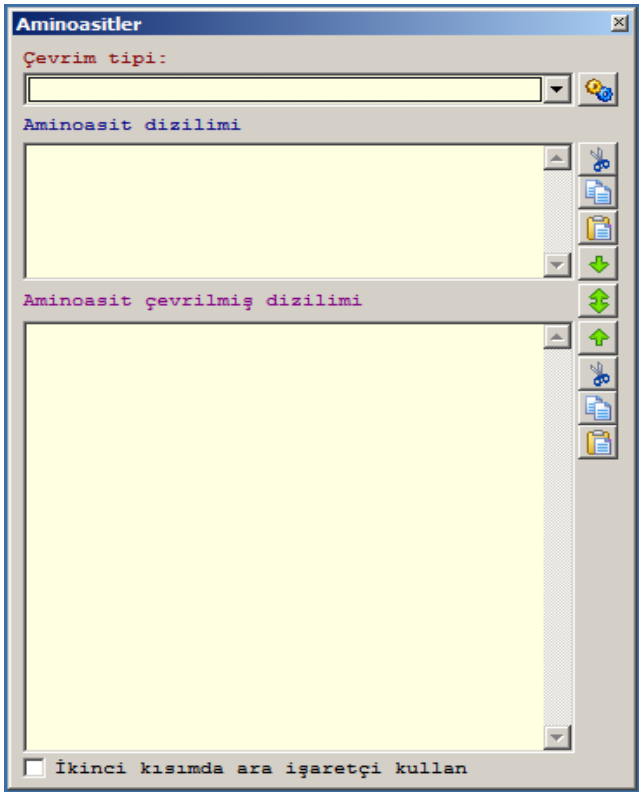
Ab initio yöntemler ikincil yapının tahmin edilmesinden çok proteinlerin doğal durumlarının fiziksel gerçekliğe uygun bir şekilde oluşturulmasında ya da protein katlanmasının benzetiminde kullanılmıştır.

3. GEREÇ VE YÖNTEM

Tez boyunca kullanılan tüm yazılımlar Türkçe bir şekilde kodlanmıştır. Bu yazılımlarım çıktılarının ise yine Türkçe olmasına özen gösterilmiştir. Tüm yazılımlar C++ dilinde sınıf mantığına uygun bir şekilde kodlanmıştır. Kodların derlenmesinde ise C++ derleyicisi olan Borland C++ 6.0 sürümü kullanılmıştır.

3.1. AminoAsit

Amino asitlerin 1-harf ve 3-harf kodlamaları ile bunların birbirine dönüştürülmelerine tez boyunca sık sık ihtiyaç duyulmuştur. Bu dönüşümlerin yanı sıra bir dizideki amino asitlerin frekanslarına, hidrofob çekirdek ve hidrofil dış kısım grafiklerine, dizinin teorik molekül ağırlığına, dizinin izoelektrik hat değerine de sık sık ihtiyaç duyulmuştur. Bu ihtiyacı karşılamak üzere AminoAsit adı verilen bir yazılım geliştirilmiştir. Bu yazılım Türkçe çıktısı olacak şekilde kodlanmıştır. AminoAsit yazılımı 1-harf kodlamadan, 3-harf kodlamaya, 3-harf kodlamadan 1-harf kodlamaya çevirmeleri yapabilmektedir. Benzer şekilde verilen dizinin hidrofobisite grafiklerini, izoelektrik hat değerini ve dizinin teorik moleküler ağırlığını hesaplayabilmektedir. AminoAsit yazılımının genel ekran görüntüsü Şekil 22 ‘de gösterilmiştir.

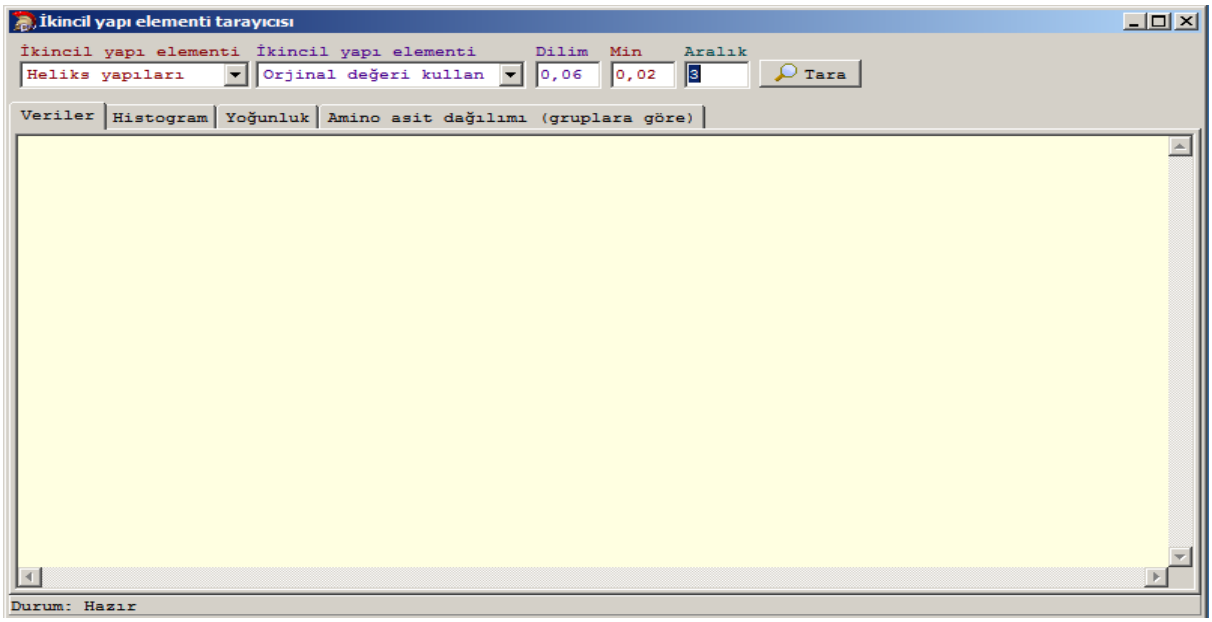


Şekil 22: AminoAsit yazılımının genel ekran görüntüsü.

3.2. İkincil yapı element tarayıcı

Tez boyunca hangi amino asitin hangi ikincil yapı elementinde ve hangi oranda bulunduğunu bilgisine ihtiyaç duyulmuştur. Bu nedenle İkincil Yapı Element Tarayıcı isimli bir yazılım geliştirilmiştir. Bu yazılım belirli bir veri tabanında yer alan proteinlerdeki ikincil yapı elementlerini tarayarak hangi amino asitin hangi ikincil yapı elementinde ve hangi oranda bulunduğunu hesaplayarak hem sayısal (taranan amino asit sayısı, taranan element sayısı, maksimum frekanslı amino asit, minimum frekanslı amino asit, en uzun ve en kısa element tipleri ve bu tipleri barındıran protein ya da proteinler gibi) hem de grafiksel (histogram, olasılık yoğunluk, dağılım histogramları) olarak gösterebilmektedir. İkincil Yapı Element Tarayıcısı isimli yazılım kendi sonuçlarını belirli filtrelerden geçirerek (örneğin maksimum ve minimum filtre ile eşik değerleme

fonksiyonu) verilerini gruplanmış veriler olarak da ifade edebilmektedir. İkincil Yapı Element Tarayıcısının genel ekran görüntüsü Şekil 23 ‘de gösterilmiştir.



Şekil 23:İkincil Yapı Element Tarayıcısı genel ekran görüntüsü.

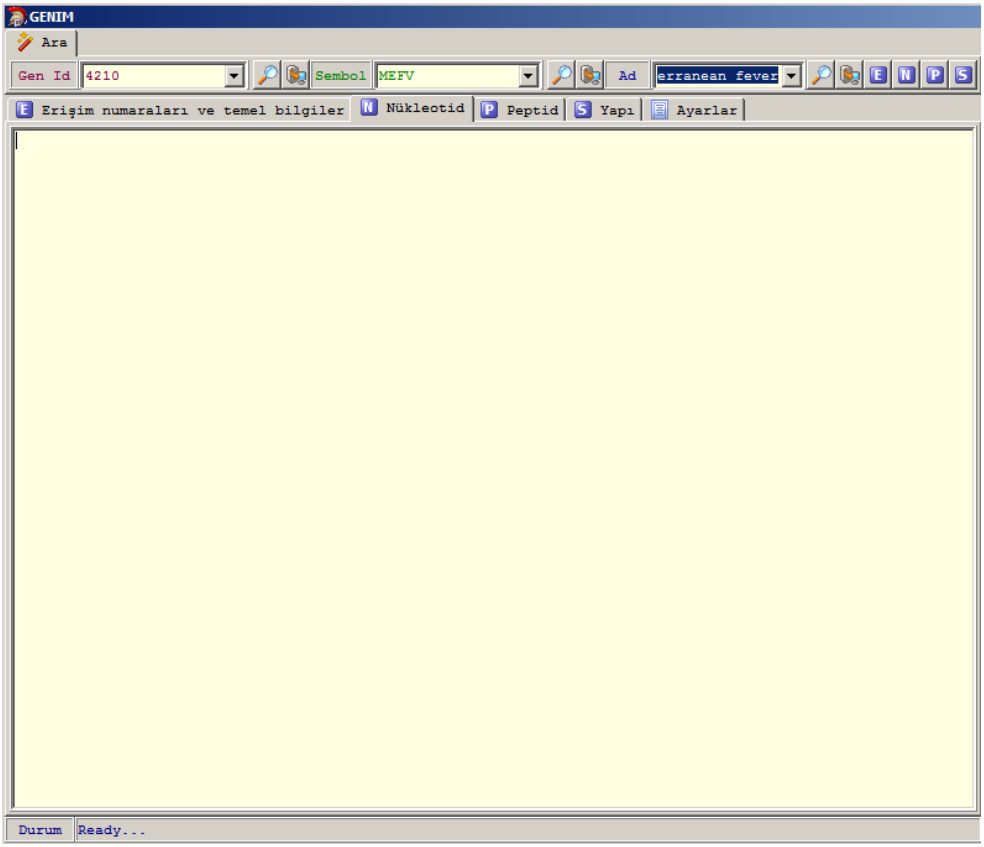
3.3. Genim

Tez boyunca farklı farklı proteinlerin farklı yanları ile çalışmak gerektiğinden her seferinde eldeki proteinin referans numaralarının farklı bir veri tabanlarındaki karşılıklarını araştırıp o veri tabanlarından bilgi almak uzun zaman ve çaba gerektirdiği için Genim isimli bir yazılım geliştirilmiştir. Genim, herhangi bir proteine ait herhangi bir referans numarasının bilinmesi durumunda bu referans numarasını diğer bazı popüler veri tabanlarının referans numaralarına çevirebilmektedir. Bu sayede Genim 17 ayrı bilgiyi gösterebilmektedir. Bu bilgiler aşağıda listelenmiştir.

1) İnsan Gen Terminoloji Komitesi (Human Gene Nomenclature Committee, HGNC)
2) Çevrimiçi Mendilian Kalıtım (Online Mendelian Inheritence In Man, OMIM)

3) Avrupa Biyoinformatik Enstitüsü (European Bioinformatic Institute, ENSEMBL / EBI)
4) Ulusal Biyoteknoloji ve Enformasyon Merkezi (National Center Of Biotechnology Information, NCBI)
5) Ulusal Tıp Kütüphanesi (National Library of Medicine, PubMed)
6) İnsan Proteini Referans Veri Tabanı (Human Protein Reference Database, HPRD)
7) Kyoto Gen ve Genom Ansiklopedisi (Kyoto Encyclopedia Genes and Genomes, KEGG)
8) Bilgi Tabanlı Protein Veri Tabanı (Protein Knowledgebase, UniProtKB)
9) SNP Veri Tabanı (SNP Database, dbSNP)
10) Tekil Dizi İşaretleri Veri Tabanı (Unique Sequence Tagged Sitesi, dbUST)
11) Ortak Kodlanan Segmentler Veri Tabanı (Consesnsus CDS Database, dbCDS)
12) Referans Dizi Veri Tabanı (Reference Sequence Database, RefSeq)
13) Gen durumu (NCBI)
14) Genin kısa açıklaması (NCBI)
15) Soy ağacı (HGNC, NCBI ve KEGG)
16) Organizma (UniProtId, EBI ve ENSEMBL ile NCBI)
17) Eski kullanılan semboller ve isimleri (HGNC ve NCBI)

Genim, genel ekran görüntüsü Şekil 24 ‘de gösterilmiştir.



Şekil 24: Genim genel ekran görüntüsü

Genim bahsedilen bu 17 sonucun içinde, sonuç maddesindeki veri tabanı taranırken elde edilen önemli bazı sonuçları da listelemektedir. UniProtKB taranırken aynı zamanda protein protein etkileşimi de listelenmektedir. Bu bilgi Protein-Protein Etkileşim Veri tabanı (interPro) taranarak elde edilmektedir. Bunların yanı sıra Genim hızlı bir şekilde verilen erişim numarası üzerinden gen nükleotid dizisine, protein amino asit dizisine, ikincil yapı elementlerine de tek seferde erişim sağlamaktadır.

3.4. Solucan

Tez boyunca genler ve proteinler hakkında pek çok bilgi gerekli olmuştur. Bunlardan bazıları şunlardır:

1) Gen adları, sembolleri kısa açıklamalar,

2) Kromozomal yerleşimleri
3) Gen lokus tipleri
4) Gen aileleri
5) Erişim numaraları
6) Nükleotid dizileri
7) Amino asit dizilimleri
8) İkincil yapı element dizileri, ikincil yapı elementlerinin hangi proteinde nerede başladığı ve nerede bittiği
9) İkincil yapı elementlerin amino asit dizilimleri

Bu bilgiler farklı farklı sunucularda bulunur. Farklı sunum formatlarında araştırmacılara ulaştırılır. Web tabanlı olduklarından şekilsel sunumları devamlı değişim ve gelişim göstermektedir. Bu sayılan etkilerden ötürü gerekli olan her seferde çevrim içi şekilde bu bilgilere ulaşmak hem zor hem de uzun zaman alacağından bu işlemleri otomatik olarak yapan bir yazılım olan Solucan geliştirilmiştir. Solucan istediğimiz verileri, istediğimiz ham veri kaynaklarından, istediğimiz kurallar doğrultusunda bilgiye dönüştürerek elde etmeyi ve istediğimiz bir biçimde depolamayı saplayan biyolojik veri madencisidir.

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, vb. 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 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 (web varlıkların, SQL varlıklarının, kural dizilerinin tanımlanması).

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, birden çok veri tabanında birden çok sistemi de desteklemektedir. Alt yapı malzemesi olarak kullanılacak veri tabanı yönetim sistemi uygulamasının mutlaka yapısal sorgulama dilini (SQL) tam desteklemesi gerekmektedir.

Tez boyunca görsel veri tabanlı tasarım yazılımı olarak serbest, açık kaynak kodlu lisansa sahip olması nedeniyle DbDesigner 4 kullanılmıştır (http://fabforce.net/dbdesigner4/, 22.10.2010).

İlişkisel veri tabanı yönetim sistemi olarak serbest, açık kaynak kodlu olması ve SQL'i tam desteklemesi nedeniyle MySQL 5.1 kullanılmıştır (http://www.mysql, 22.10.2010).

DbDesigner 4 ve MySQL 5.1'in Türkçe değişken isimleri tanımlamasına izin vermemesi nedeniyle karşılaştırmalarda kolaylık sağlamak için, uygulama ve programlama da İngilizce değişken isimleri kullanılmıştır.

Öncelikle genlere istediğimiz zaman ulaşmamıza izin verecek tekil bir numara tanımlanmıştır. Tanımlanan bu numara HgncID 'dir. Human Gene Nomenclature Committee (HGNC) her gene akılda kalıcı ve sadece o gene özgü bir isim ve sadece o gene ait olacak tekil bir kimlik numarası vermek için kurulmuştur (http://www.genenames.org/, 22.10.2010). HgncId, bu komite tarafından atanan numaranın aynısıdır.

Genlerin ait oldukları lokus tipleri ve genlerin ait oldukları gen aileleri ile gen sembolü ve gen ismi diğer önemli bilgilerdendir. Ayrıca genin kısa olarak açıklaması, ne işe yaradığı bilgisi, kromozomal lokalizasyonu bilgisini de (kromozomal lokalizasyon parçalanarak veri tabanına alınmıştır; yerleşim; kromozom, kol, bant ve alt bant seviyesinde parçalanmıştır) Solucan tarafından toplanması istenen bilgilerdendir.

Tez boyunca genlerin erişim numaralarına sık sık ihtiyaç duyulmuştur. Bu numaralar bize genler hakkında farklı veri tabanlarından farklı verileri almamızı sağlamıştır. Örnek olarak eldeki gen sembolü ya da eldeki gen adı ya da sahip olduğunuz HgncId ile genin mRNA dizisine ulaşmak biyoinformatik açısından bakıldığında pek kolay olmayacaktır. Bu nedenle hızlı ve doğru bir şekilde bilgiye ulaşmak için bazı erişim numaralarına önem verilmiş, Solucan tarafından toplanması istenen bilgiler arasına eklenmiştir. Bu erişim numaraları:

1) Entrez Gene Id (http://www.ncbi.nlm.nih.gov/gene),
2) UniProtId (http://www.ncbi.nlm.nih.gov/gene),
3) OmimId (http://www.ncbi.nlm.nih.gov/omim),

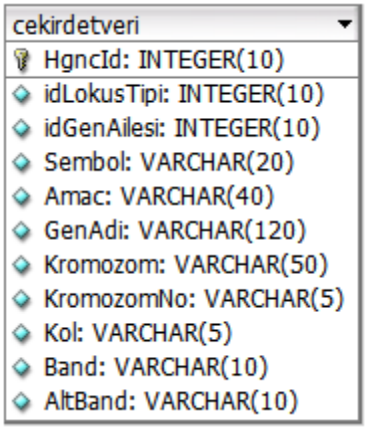
4) EmsemblId (http://www.ensembl.org),

5) RefSeqId (http://www.ncbi.nlm.nih.gov),

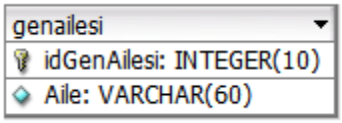
6) UcscId (http://genome.cse.ucsc.edu).

Gen lokus tipi, gen ailesi, genin sembolü, kısa tanımı, cinsi, kromozomal yerleşimi ve HgncId numarasını ihtiva eden tablo “Çekirdek Veri” olarak adlandırılmıştır. Tablo 5a çekirdek veri parametrelerini içerir. Gen ailesi isimleri “Gen Ailesi” isimli tabloda Tablo 5b, lokus tipleri “Lokus Tipi” isimli tabloda Tablo 5c, amino asit kodlamalarını ise “AABin” isimli tabloda Tablo 5d, erişim numaralarını “Erisim” isimli tabloda Tablo 5e ‘de toplanmıştır ve bunları birbirleri ile HgncId, idLokusTipi, idGenAilesi numaraları ile ilişkilendirilmiştir.

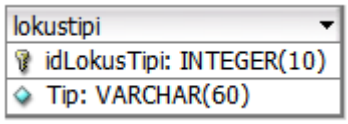
Tablo 5a: Çekirdek veri tablosu ve yer alan parametreler.



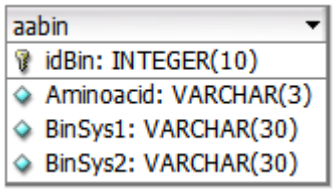
Tablo 5b: Gen ailesi isimleri tablosu ve yer alan parametreler.



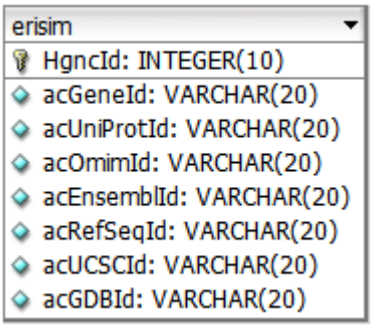
Tablo 5c: Gen ailesi isimleri tablosu ve yer alan parametreler.



Tablo 5d: Amino asitlerin kodlarını içeren tablo ve yer alan parametreler.



Tablo 5e: Genlerin belirlenen erişim numaralarını içeren tablo ve yer alan parametreler.



Tüm bu veriler bu tez boyunca Solucan veri tabanı olarak adlandırılacaktır.

3.4.1. 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.

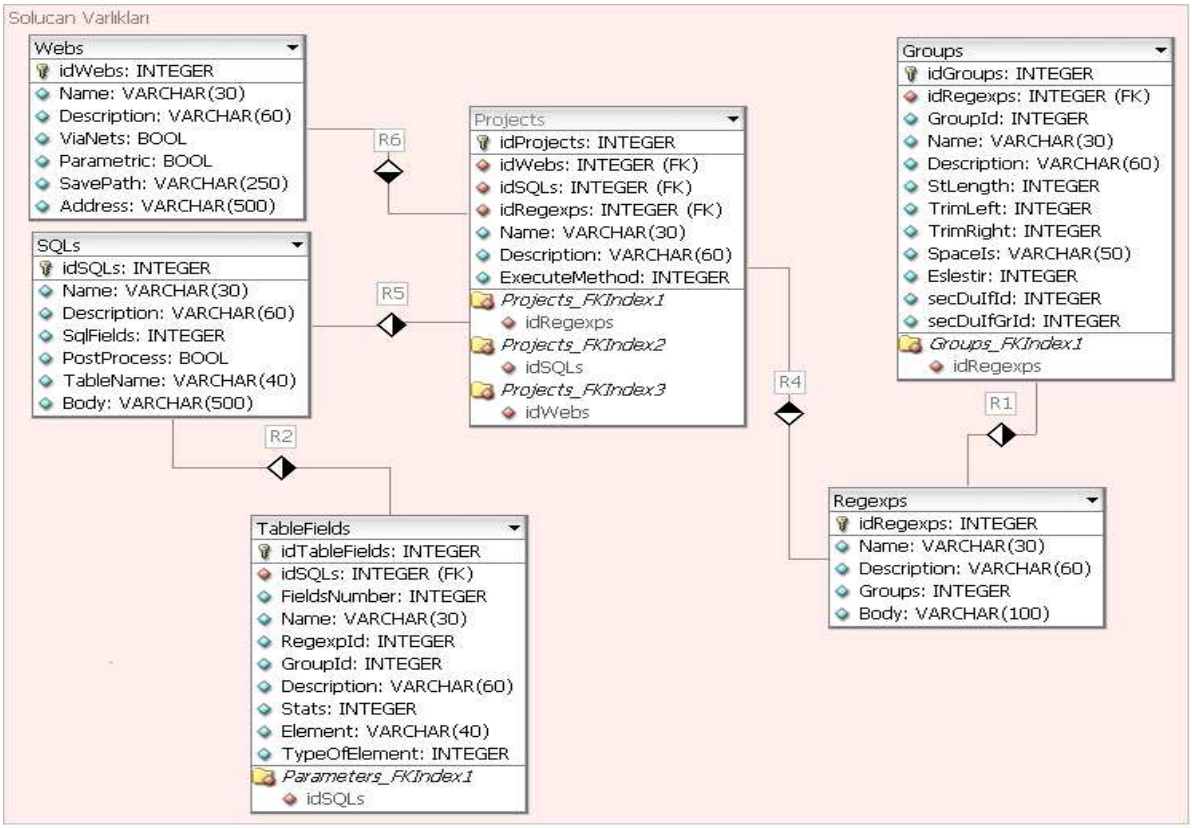
Bu üç ana varlık sırası ile şunlardır:

1) Web varlığı,

2) Düzenli ifade varlığı,

3) SQL varlığı.

Ayrıca düzenli ifade varlığının grup bağlantı varlıkları ile SQL varlığının alan bağlantı varlıkları, üç ana varlığı ilişkisel olarak birbirine bağlanmıştır. SOLUCAN varlıkları için kullanılan veri tabanının adı SOLUCANInfo olarak verilmiştir. Varlıkları tanımlayan tablolar SOLUCANInfo veritabanı içerisinde DbDesigner 4 kullanılarak ilişkilendirilip görsel hale getirilmiştir. Şekil 25 ‘de SOLUCANInfo veri tabanı, bağlantıları ve veri yapıları görülmektedir.

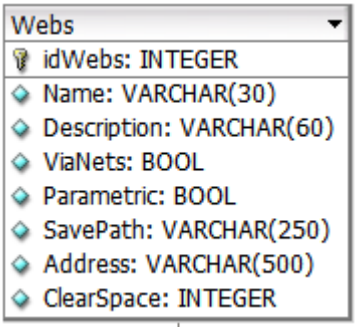


Şekil 25: SOLUCANInfo veritabanı için ilişkilendirilmiş görsel veri modeli**.**

3.4.2. Web varlığı

Web varlığı ham verinin alınacağı yeri belirler. Ham veri Web varlığı tarafından bir Web sunucusundan, dosya transfer protokolü (File transfer protocol, 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. Tablo 6 'da web varlık tablosu ve web varlığı tanımlayabilmek için gerekli parametreler gösterilmiştir.

Tablo 6: Web varlığı ve parametreleri



Yukarıdan aşağıya doğru parametre tanımları ise:

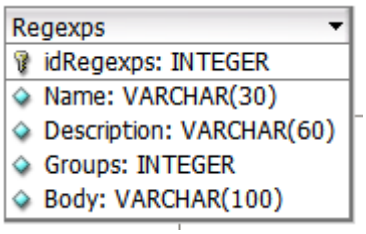
1) Web varlığı kimlik no (idWebs)
2) Varlığın adı (Name)
3) Varlık hakkında kullanıcının kısa hatırlatma ya da tanımlama notu (Description)
4) Varlığın bilgiyi internet ortamından alıp almayacağının belirlenmesi (ViaNets)
5) Varlığın parametrik bir şekilde derlenip derlenmeyeceğini belirler. Parametrik derleme, kullanıcı tarafından oluşturulan parametreleri bir web adresi köküne istenilen formatta yazılması ile oluşturulan listedeki tüm adreslerden veri indirmek için kullanılır (Parametric)

6) İndirilen verinin daha sonra işlenme ihtimaline karşın hangi lokal adrese kaydedileceğini ve kayıt ismini belirler (SavePath). Bu parametre boş olarak geçilirse kayıt işlemi yapılmaz.
7) Ham verinin hangi Web adresinden alınacağını belirler (Address).
8) Ham verinin olduğu gibi mi yoksa boşluklarının temizlenmiş halde mi işleme sunulacağını belirler (ClearSpace).

3.4.3. 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 istatistikî işlemler de yapabilir. Aslında ana düzenli ifade grubuna dahil olmayan bu şekilde ki gruplar Solucan içerisinde sanal gruplar olarak adlandırılmıştır. Düzenli ifade varlığı tablosu ve yer alan parametreler Tablo 11a ‘da sunulmuştur.

Tablo 11a: Düzenli ifade varlığı ve yer alan parametreler



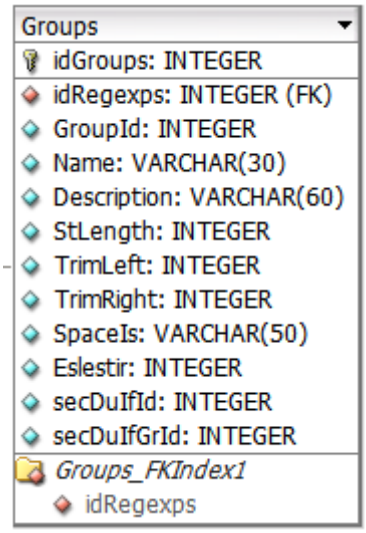
Yukarıdan aşağıya doğru parametre tanımları ise:

1) Varlık kimlik no (idRegexps).
2) Varlık ismi (Name).
3) Varlık için kısa tanımlama ya da hatırlatıcı not (Description).

4) Varlığın ham veriyi bilgi grupları şeklinde organize edebilmesi için tanımlanan grup sayısını belirler (Groups)
5) Varlık gövde metni. Düzenli ifade metni bu alana yazılır (Body).

Görüldüğü üzere düzenli ifade varlığı ham veriyi gruplara böler. Her grup varlığı, grup şeklinde organize edilmiş veriye gerekli olabilecek metin işleme işlemlerinin uygulanıp uygulanmayacağını belirleyen bir dizi parametre daha içerir. Bu parametreler Tablo 11b ‘de gösterilmiştir.

Tablo 11b: Düzenli ifade grubu ve ek parametreleri



Yukarıdan aşağıya doğru parametre tanımları ise:

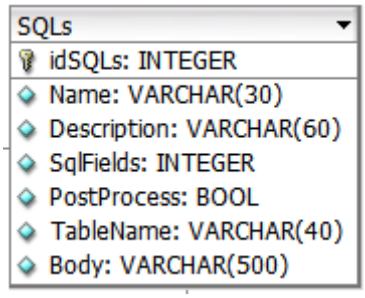
1) Grup kimlik no (idGroups)
2) Grubun ait olduğu düzenli ifade varlığının kimlik numarası (idRegexps)
3) Hangi grubun tanımlanacağını belirleyen grup numarası (GroupId)
4) Grup adı (Name)
5) Grubun kısa tanımı ya da hatırlatıcı not (Description)

6) Grubun sahip olduğu bilginin maksimum uzunluğu. Eğer bir şekilde buraya girilen değerden daha yüksek veri içeren bir veri alınır ise, fazla kısım göz ardı edilir (stLength)
7) Grubun içerdiği verinin soldan itibaren kaç karakterlik kısmının kesileceğini belirler. -1 olması durumunda kesme işlemi yapılmaz (TrimLeft)
8) Grubun içerdiği verinin sağdan itibaren kaç karakterlik kısmının kesileceğini belirler.-1 olması durumunda kesme işlemi yapılmaz (TrimRight)
9) Grubun içerdiği verinin veri uzunluğunun 0 olması durumunda (boş karakter) bu veri grubuna verilecek ortak değeri belirler (SpaceIs).
10) Ham veriden elde edilen grubun içerdiği veri ham veri olarak atanıp atanmayacağını belirler. Eğer ham veri olarak atanır ise başka bir düzenli ifade ile eşleştirme yapılır. Bu sayede kendi üzerine dönüşlü bir veri yapısı tanımlanmış olur. Kromozomal lokalizasyonun parçalanmasında bu tip bir kendi üzerine dönüşlü yapı kullanılmıştır (Eslestir).
11) SecDuIdId ve secDuIfGrId SOLUCAN tarafından kendi üzerine dönüşlü veri modelini kurabilmek için otomatik olarak doldurulan parametrelerdir.

3.4.4. SQL varlığı

Ham veri, bilgi gruplarına, Solucan sanal gruplarına parçalandıktan sonra, hedef ilişkisel veri tabanına SQL varlığı kullanılarak kaydedilir. Her SQL varlığı bilgi grupları ile veri tabanı alanları arasında ilişkileri içerir. Her SQL varlığı alan adlarını içerir. Bu alan adları düzenli ifade varlığının grupları ile kayıt edilecek veri tabanı arasındaki bağlantıyı sağlar. SQL varlığını tanımlamak için gerekli olan parametreler Tablo 12a ‘da sunulmuştur.

Tablo 12a:SQL varlığı ve yer alan parametreler

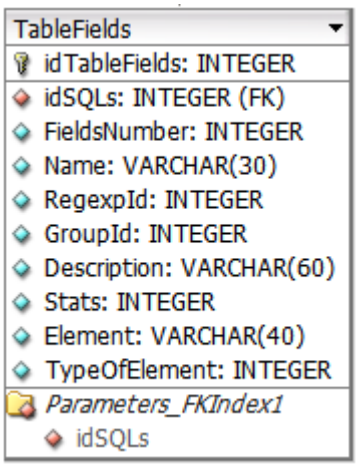


Yukarıdan aşağıya doğru parametre tanımları ise:

1) SQL varlığı kimlik no (idSQLs)
2) Varlık adı (Name)
3) Varlık için kısa tanımlama ya da hatırlatma notu (Description)
4) SQL varlığının sahip olduğu alan sayısı (SqlFields)
5) SQL varlığı ile düzenli ifade varlıkları arasında bağlantı kurulduktan ve gerekli olan veriler gerekli olan yerlere yazıldıktan sonra; yazılan veriler üzerinde herhangi bir kayıt sonrası işlem yapılıp yapılmayacağını belirler. Bu değerin 1 olması halinde o tablodaki mükerrer kayıtlar silinir, endeksler yeniden ayarlanır, boşluklar doldurulur ve ilişkiler kontrol edilir (PostProcess)
6) Düzenli ifade varlığı verilerinin hangi tabloya yazılacağını belirler. Bu tablo aynı zamanda kayıt sonrası düzenlemelerin yapılacağı tablodur (TableName).
7) SQL metni bu kısma yazılır (Body)

Her SQL varlığı düzenli ifade varlığı gibi ek bir takım parametrelere ihtiyaç duyar. Bu parametreler alan parametreleri olarak adlandırılmıştır. Alan parametreleri tablosu Tablo 12b ‘de gösterilmiştir.

Tablo 12b: SQL varlıkları için alan parametreleri



Yukarıdan aşağıya doğru parametre tanımları ise:

1) Alan parametreleri için kimlik numarası (idTableFields).
2) Alan parametrelerinin ait olduğu SQL varlığı için kimlik numarası (idSQLs)
3) Hangi alan parametresinin tanımlandığı bu kısma yazılır (FieldsNumber)
4) Alan parametresi adı (Name)
5) Alan parametresi ile düzenli ifade varlığı arasındaki bağlantıyı tanımlar. Hangi düzenli ifade varlığından gelen bilginin tanımlanan alana yazılacağına dair ilişki bu kısımla belirlenir (RegexpId)
6) Bilginin düzenli ifadenin hangi grubundan alınacağı bu kısımda belirlenir. Bu kısma girilen parametre ilişkili düzenli ifade varlığının gruplarından biri olmalıdır (GroupId).
7) Alan parametresi hakkında tanımlayıcı bilgi ya da hatırlatıcı not bu kısma yazılır (Description)
8) Bu alanda düzenli ifade grubundan gelen verilerin olduğu gibi mi yazılacağı yoksa belli bir takım sayma, frekans hesaplama, ortalama alma gibi istatistiki işlemler uygulandıktan sonra mı yazılacağını belirler. Bu kısma 0 girilmesi

durumunda bilgi aynen yazılır, 1 değeri girilmesi durumda sayma işlemine tabii tutulur (Stats)

9) Sayma işlemi uygulanacak ise bu hangi alanda yapılacaksa bu elementin adı bu kısma verilmelidir (Element)
10) Dokuzuncu maddede bahsedilen elementin tipi buraya yazılır (String, Integer, Double, Float vb)

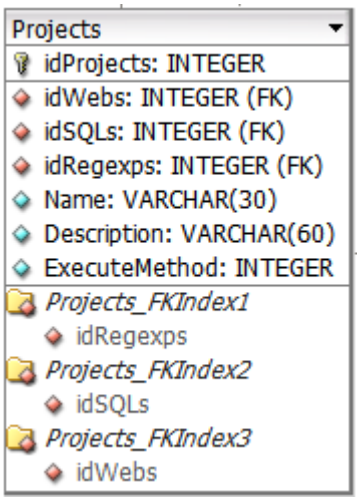
Dokuzuncu ve onuncu parametreler yeni SOLUCAN sürümünde SOLUCAN tarafından otomatik olarak belirlendiğinden varsayılan değerleri 0 olarak atanır.

3.4.5. Solucan projeleri

SOLUCAN projesi, yukarıda vurgulanan üç varlığın bileşimi ve tanımlanan projenin SOLUCAN tarafından ne şekilde derleneceğini içeren kullanıcı direktifidir. SOLUCAN, bir projeyi, ya tek kaynaktan tek bir ham verinin alınıp içindeki tüm bilgi gruplarının çıkartılıp kaydedilmesi (düzenli derleme) şeklinde; ya da birden çok kaynaktan birden çok kere indirilen farklı ham verilerin içindeki tüm bilgi gruplarının çıkartılıp kaydedilmesi şeklinde (parametrik derleme) derler. 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 edilmiştir.

Bir SOLUCAN projesi tanımlamak için gerekli olan parametreler Tablo 13‘de sunulmuştur.

Tablo 13: SOLUCAN projesi parametreleri



Yukarıdan aşağıya doğru parametre tanımları ise:

1) SOLUCAN projesi için kimlik numarası (idProjects)
2) SOLUCAN projesi için web varlığı kimlik numarası (idWebs)
3) SOLUCAN projesi için SQL varlığı kimlik numarası (idSQLs)
4) SOLUCAN projesi için düzenli ifade kimlik numarası (idRegexps)
5) SOLUCAN projesi adı (Name)
6) SOLUCAN projesi hakkında kısa açıklama ya da hatırlatıcı not (Description)
7) SOLUCAN projesinin ne şekilde derleneceği bilgi bu kısımda belirlenir. 0 değeri alması durumunda düzenli derleme, 1 değerini alması durumunda parametrik derleme yapılır (ExecuteMethod)

Bu şekilde tanımlanan bir ya da birden çok proje SOLUCAN tarafından belirlenen metot üzerinden derlenir.

Solucan pek çok projeyi derleyecek şekilde yazılmıştır. Bu sayede elle yapılması oldukça zor ve zaman alıcı olan veri tabanlarında bilgi arama, bulunan bilgiyi ham veri

şeklinde alma ve kurallar doğrultusunda bilgi grupları şeklinde organize edip istenilen şekilde depolama sorununa çözüm oluşturmuştur.

3.5. Düzenli ifadeler

Bir string verisi içinde başka bir string verisinin hızlı bir şekilde bulunması, karşılaştırılması amacına hizmet eden bir araçtır. Örnek olarak “ab+” paterni “ab” sabit olmak üzere string dizi içinde ki “ab”, “abb”, “abbb” v.b. diziler ile uyuşacaktır. Bu bağlamda düzenli ifadeler veri madenciliği konusunda oldukça kullanışlı bir hale gelmişlerdir.

Deel.h kütüphanesi tezimizde düzenli ifade derleyicisi olarak kullanılacaktır (http://www.regexplab.com, 23.10.2010). Deel.h kütüphanesi açık kaynak kodlu C++ uyumlu olduğundan tercih edilmiştir.

3.5.1. Ortak karakterler

Harfler, numaralar, noktalama işaretleri ile özel olmayan tanımlamalar ortak karakterler adını alırlar. Örneğin “c” paterni “abcd” metni içinde aransın. Bu durumda (Başlangıç, Bitiş) formatında (3,4) verisi dönecektir. Aranılan ifade metinde 3. karaketerde başlar ve 4. karakterde sonlanır anlamındadır.

3.5.2. Kaçış karakterleri

Basılmayan karakterler kaçış karakterleri olarak anılır. Bunlar:

a) \t : Tab

b) \n: Yeni satır

c) \\: Kendisi ile eşleştirme

d) \r: Taşıyıcı döndür (return)

Kaçış karakterlerinden “\” bizim için oldukça önemlidir. Bu karakter “.”, “,”, “:” gibi deelx.h kütüphanesinde özel anlamı bulunan karakterlerin kendileri ile eşleşmesine olanak sağlarlar.

3.5.3. Çok karakter ile uyum sağlayan ifadeler

Bazı ifadeler kuralları gereği bir ya da birden çok karakter ile direkt eşleşebilirler. \d ifadesi tek karakter şeklinde herhangi bir sayı ile eşleşir “0123456789”. \d\d ifadesi iki basamaklı herhangi bir sayı ile eşleşebilir.

Örneğin ifade “a\d” şeklinde ise metinde “a” ile başlayan ve sonrasında herhangi bir rakam olan ifadelerin tümü ile eşleşir. “a0”, “a1”, …, “a9” gibi. Benzer şekilde harfler ve diğer karakterler ile eşleşen ifadeler:

a) \d : 0 ile 9 arasındaki rakamlar ile eşleşir
b) \w : herhangi bir karakter ile eşleşir. Bu karakterin büyük ya da küçük harf olması gerekmez. Metin içindeki rakamlar ile eşleşmezler. Noktalama işaretleri ile eşleşmezler.
c) \s: metin içindeki boşluk karakterleri ile eşleşirler. \t tab karakteri ile eşleşmezler.
d) . : Bu karakter hemen her şey ile eşleşir. Herhangi bir karakter anlamındadır fakat \n, \r gibi satır sonu satır başı karakterleri ile eşleşmezler.

3.5.4. Kullanıcı tanımlı ifadeler

“[“ ile başlayan ve “]” ile biten tanımlamalar; kullanıcıların kendi isteklerine göre eşleşme kümesi tanımlamasına izin verirler. Bu şekilde birden çok karakterle eşleşen geniş eşleşme potansiyeli olan ya da sadece birkaç karakter ile eşleşen özel kümeler tanımlanabilir. [f-k]: “f” harfinden başlayarak “k” harfine kadar olan harfler ile eşleşen bir

ifade yazılabilir. [^abc], “abc” stringinin olmadığı her string ile eşleşir. “abarbbbbc56” metinindeki “r” ve “56” nın bizim için önemli olduğunu varsayalım; bu durumda [^abc] ifadesi kullanılabilir. Dönüt “r” ve “5” ve “6” olacaktır.

3.5.5. Miktar belirleyiciler

Eşleşmelerden bahsedildikten sonra eşleşmelerin miktarlarının belirlenmesi de önemli bir özelliktir. Örneğin “kkkk” ifadesi “dghegdhjgewhjdewkkkgfdgefdgefdhe” metini içinde aranmakta olsun. Önceki bilgiler eşliğinde \w ifadesi kullanılabilir ya da [^dghrjwh] gibi bir ifade kullanılabilir. Fakat bunun rakamsal değerinin de önemli olup “kkkk” ifadesinin “kkkkkk” ya da “kk” ifadesi ile aynı olmadığı durumlarda, sayısal betimleyicilere başvurulur.{n}, sabit bir sayıda tekrarlar için kullanılır. [k]{4} ifadesi metin içinde sadece 4 adet “k” karakterinin yan yana bulunduğu alan ile eşleşecektir. {m,n} ifadesi minimum m adet maksimum n adet olmak kaydı ile eşleşecektir. Örneğin bizim ikin “k” karaketerinin metin içinde “k”, “kk”, “kkk” olması önemli ise ve özellikle bu diziler bulunmak isteniyorsa bu durumda ifade [k]{1,3} şeklinde düzeltilebilir. Bunun herhangi bir harf olması istenirse \w{1,3} ile değiştirilmelidir.

{m, } n sayısının null olması durumunda en az m tane olması ve üst sınırının ise önemsiz olması anlamına gelir. Bu ifade ile metin içinde “kk”, “kkk”, “kkkk”, “kkkkk” gibi artan sayılar önemli ise ve “k” göz ardı edilecekse [k]{2, } ifadesi kullanılır.

“+” işareti sayısal betimleme için kullanıldığında sadece 1 anlamındadır. [k]+ sadece “k” karakterinin ve sadece bir kez bulunduğu tüm noktalar ile eşleşir.

“*” ifadesi sayısal betimleme olarak bir üst sınır tanımlamaz, alt sınırı da yoktur. Esnek bir ifadedir. Aranılan karakter metinde olmasa da olabilir. Bu geneli bozamaz. Örneğin kromozomlar ve lokalizasyonları düşünüldüğünde ifade farklılıkları göze

çarpacaktır. Her kromozom için kromozom, kol, bant, alt bant gibi bir tanımlamaya her zaman metin içinde rastlanmayabilir. Örneğin 19q, 3q21.3-q25.2, 4p16.3, 1q32, 9pter, vb gibi. Böyle durumlarda "*" karakteri uygun sahalarda kullanılabilir. SOLUCAN kromozomları parçalar iken

HGNC:(\d{1, })\t([\dXYxy]{1,2})([pq]*)([\dtercen]*)\.*([\dtercen]*)

benzeri düzenli ifade kullanımıştır. Burada [pq]* "p" ya da "q" ile eşlecek ve fakat bunlar birden çok kere geçebilir veya geçmeyebilir. Hiç olmaması yanlış olmaz anlamını katmaktadır.

|, ifadesi veya anlamındadır. Opsiyonlu dizileri metin içinde bulmaya yarar. "a12", "a24" ifadeleri bir metin içinde aranıyor ise bu durumda "a" kesin olacağında bu özel katara dahil edilir [a], yanında 12 ya da 24 olabilir bu da katara eklenir [a12|24] bu ifadeler ile eşleşecektir (Regexp Laboratory, http://www.regexlab.com/en/deelx/, 08.01.2011). SOLUCAN, ile çalışırken pek çok düzenli ifade kullanılmıştır. Bu ifadeler liste şeklinde Tablo 14'de gösterilmiştir.

Tablo 14: SOLUCAN'da kullanılan düzenli ifadelerden bazıları.

İfade	**Metin**
1	FT\s*(STRAND|HELIX|TURN)\s*(\d{1, })\s*(\d{1, })
2	HGNC:(\d{1, })\t([\dXYxy]{1,2})([pq]*)([\dtercen]*)\.*([\dtercen]*)
3	>(NC_.*)<
4	(>>.*\n)(.*)(<)
5	HGNC:(\d{1, })\t(.*)
6	HGNC:(\d{1, })\t(.*)\t(.*)\t(.*)\t(.*)\t(.*)
7	FT {1,4}(\w{1, }) {1, }(\d*) {1, }(\d*)
8	>.*(NG_\d{1, }\.*.\d*).*<
9	(\d{1, })\t(\d{1, }|X|Y)(p*q*[ter]*[cen]*)(\d*)\.*(\d*)
10	(\d{1, })\.\.(\d{1, })<
11	SQ.* (\d{1, }) MW
12	SQ.* (\d{1, }) AA
13	FT\s*(STRAND|HELIX|TURN|DISULFID)\s*(\d{1, })\s*(\d{1, })

Deelx.h hakkında detaylı bilgiye, http://www.regexlab.com ve http://www.codeproject.com adreslerinden ulaşılabilir.

3.6. MySQL

MySQL, SQL için yazılmış açık kaynak kodlu, serbest dolaşımdan elde edilebilen derleyicidir. Açık kaynak kodlu olması, lokal medya aygıtlarına kolay yüklenmesi, sık tercih edilir olması, hızlı şekilde sonuç vermesi, serbest kullanıcı lisansına sahip olması, güvenlik açısından etkin algoritmalarının olması nedeni ile tercih edilmiştir. Alternatif olarak Microsoft SQL Server (Microsoft Corp., http://www.microsoft.com/sqlserver/, 05.01.2011), Postgresql (PostgreSQL Database Management System, http://www.postgresql.org/, 05.01.2011), Oracle (Oracle Corp., http://www.oracle.com/, 05.01.2011), Access (Microsoft Corp, http://office.microsoft.com/tr-tr/ , 05.01.2011) dikkat çeken derleyicilerdendir.

3.6.1. Komut kullanımı

SQL komutları basit ve anlaşılırdır. Her SQL komut satırı “;” ile biter. SELECT * FROM tablo komutu; seçer, neyi seçer “*” ile her şeyi , nereden seçer “FROM” ile tablo adını taşıyan özel tablomuzdan. Her şeyi seçer tanımı tablo altındaki tüm sütunları seçer anlamında kullanılmıştır. Belirli özellikleri taşıyan kayıtların özellikle seçilmesi isteniyor ise bu durumda :

```
SELECT * FROM tablo

WHERE Yer = ‘Safranbolu’
```

Komutları ile tablo adındaki tablodan, tablonun bir sütunu olan Yer verisinin sadece ‘Safranbolu’ olduğu özeli kayıtları seçer. Bu şekilde basit bir sorgu yapılmış olur.

Örnek olarak SOLUCAN veri tabanında 19. kromozomda olan genler görüntülenmek istenmektedir. Bu durumda kullanılacak komut dizisi:

SELECT * FROM CekirdetVeri

WHERE KromozomNo = 19

Benzer şekilde yukarıdaki sorguya ek olarak sadece “p” kolundaki kayıtlar listelenmek istenseydi o zaman komut dizisi:

SELECT * FROM CekirdetVeri

WHERE KromozomNo = 19 AND Kol = ‘p’

Şeklinde düzenlenmeliydi. Daha karmaşık sorgu örneği olarak, bant değerinin 13 olduğu ve alt bant değerinin 5’den az olduğu kayıtlar için komut dizisi:

SELECT * FROM CekirdetVeri

WHERE KromozomNo = 19 AND Kol = ‘p’ AND Bant = 13 ABD AltBant < 5

Eğer tablo içinden sadece bazı sütun verilerine ihtiyaç duyulur ise. Örneğin sadece sembol ve gen isimlerini almak istersek bu durumda komut dizisi:

SELECT Sembol, GenAdi FROM CekirdetVeri

WHERE KromozomNo = 19 AND Kol = ‘p’ AND Bant = 13 ABD AltBant < 5

Şeklinde olacaktır.

Herhangi bir tabloya belirli bazı verileri eklemek istediğimizde INSERT INTO, komutu kullanılır.

```
INSERT INTO tablo (sütun1, sütun2, sütun3)

VALUES (1, ‘Kamil’, ‘Safranbolu’);
```

Komut dizisi tablo isimli tablomuza 1, Kamil , Safranbolu satırını kaydeder. Integer tipli veriler her zaman SQL kurallarına göre “ ’ ” kullanılmadan direkt yazılır ama string tipli veriler tırnak işaretleri arasında verilmelidir.

3.6.2. Operatörler

SQL dilinde operatör terimi “=”, “<”, “>”, “<=”, “>=”, “!=” için kullanılır. Bunların anlamları:

a) “=” : Eşittir. Değerin eşit olup olmadığını sorgular
b) “<” : Küçüktür. Değerin küçük olup olmadığını sorgular
c) “>”: Büyüktür. Değerin büyük olup olmadığını sorgular
d) “<=”: Küçük eşit. Değerin küçük eşit olup olmadığını sorgular
e) “>=”: Büyük eşit. Değerin büyük eşit olup olmadığını sorgular
f) “!=” : Eşit değildir. Değerin Eşitdeğildir olup olmadığını sorgular

Bu operatörler WHERE cümlesi ile kullanılırlar.

3.6.3. Mantıksal operatörler

Bunlar SQL dilinde “AND”, “OR”, “LIKE”, “NOT LIKE”, “IS NULL”, “IS NOT NULL”, “IN”, “BETWEEN” operatörleridir. Bunların anlamları:

a) AND : Ve anlamındadır. Bir önceki işlem ile ve mantıksal işleminin yapılacağını belirtir.

b) OR : Veya anlamındadır. Bir önceki işlem işlem ile veya mantıksal işleminin yapılacağını belirtir.

c) LIKE : Patern sorgulaması yapar. Düzenli ifade yazılabilir.

d) NOT LIKE : Patern sorgulamasını yapar değilini döndürür.

e) IS NULL : Eğer boşsa anlamındadır.

f) IS NOT NULL : Eğer boşsa sorgusunun değilini döndürür.

g) IN : Ardıl OR operatörüdür.

h) BETWEEN : Arasında anlamıyla ardıl AND operatörüdür.

3.6.4. Sıralama

SQL sorgu sonuçları belirli şekillerde sıralama yapılarak döndürülebilir. Bunun için en çok tercih edilen komut ORDER BY ‘dır. İsmi içinde “Binding site” bulunan proteinlerin KromozomNo ‘ya göre sıralanarak sorgu sonucu almak istiyorsak bu durumda komut dizisi:

```
SELECT * FROM Cekirdek Veri

WHERE GenAdi LIKE "%Binding Site%"

ORDER BY KromozomNo
```

Şeklinde olmalıdır.

Eğer azalan bir sırada görülmek istenirse bu durumda DESC, artan bir sırada görülmek istenirse ASC kullanılır.

```
SELECT * FROM Cekirdek Veri

WHERE GenAdi LIKE "%Binding Site%"

ORDER BY KromozomNo DESC
```

Azalan sırada görüntüler, ya da

```
SELECT * FROM Cekirdek Veri

WHERE GenAdi LIKE "%Binding Site%"

ORDER BY KromozomNo ASC
```

Artan sırada görüntüler.

3.6.5. Özel fonksiyonlar

SQL dilinde bazı özel komutlar bulunur. Bunlardan bazıları “DISTINC”, “GROUP BY”, “MAX”, “MIN”, “AVG”, “STD”, “SUM”, “COUNT” şeklindedir. Bunların anlamları:

a) DISTINC : Aynı olan sütunları göstermez. Sadece ilk kayıt listelenir.
b) GROUP BY: Kayıtlar listenirken herhangi bir sütuna göre gruplanmak istenirse bu konut kullanılır.
c) MAX : Sütun verisi içinde en yüksek değer ne ise onu döndürür
d) MIN: Sütun verisi içinde en küçük değer ne ise onu döndürür.
e) AVG: Sütün verilerinin aritmetik ortalaması döndürür.
f) STD : Sütun verilerinin standart sapma değerini döndürür.
g) SUM : Sütun verilerinin toplamını döndürür.
h) COUNT : Sütun verilerini sayar ve sonucunu döndürür.

MySQL kurulum ve kullanım ile komut seti ve daha detaylı açıklamalar için http://www.mysql.com/ adresine başvurunuz. Ayrıca yine açık kaynak kodlu bir proje olan ve MySQL kullanım ara yüzü olarak tanımlanabilecek HeidiSQL http://www.heidisql.com/ adresinden yararlanılabilir.

3.7. BOT

Yapay sinir ağı için C++ dilinde yazılmış olan ve ticari şekle dönüştürülmüş yapay sinir ağı motorları bulunmaktadır. Bunlardan bazıları NeuronDotNet, ve MATLAB BP ‘dir. Fakat bu uygulamaların veri yapıları ve SOLUCAN veri tabanından gelen veri yapıları arasındaki farklılıklar, uyarlama sorunları ve kullandığımız derleyicinin Borland C++ Builder 6.0 (Borland 2009 yılında C++ Builder ‘ı Embarcadero Inc. ‘ye devretmiştir, bu nedenle bu derleyici Embarcadero lisansı altında güncel olarak anılmaktadır) olması nedeni ile yeniden tam uyumlu bir geri yayılım algoritması ile çalışan çok katlı yapay sinir ağı motoru yazmamız gerekmiştir. Bu motor C++ sınıfları şekline dönüştürülmüştür. Yapay sinir ağı motorunun adı mktANN olarak verilmiştir. mktANN ‘nin özet veri yapısı Şekil 26 ‘da gösterilmiştir.

```
typedef struct
            {
                  int    NodeId;
                  double Input;
                  double Output;
                  double Error;
                  double LocaGradient;
                  double Desired;
                  double * Weight;
                  double * dWeight;
            }__node;

typedef struct
            {
                  int LayerId;
                  int Connection;
                  __node * Nodes;
            }__layer;

typedef struct
            {
                  int AnnId;
                  __layer * Layers;
            }__ann;
```

Şekil 26: Yapay sinir ağımızda kullandığımız veri yapısı.

Sınıfsal özellikler ise Şekil 27 ‘ de gösterilmiştir.

```
__property int    LayerCount
__property int    NodesCount[int a]
__property double LearningRate
__property double MomentumRate
__property double StopCriteria
__property int    WeightStatus
__property int    WeightFactor
__property double Error
__property double ErrorForGraph
__property double MeanErrorRange
__property double Inputs[int a]
__property double Desireds[int a]
__property double LearningStatus
__property int    Iteration
__property int    BatchMode
__property int    CorrectOutput
__property int    IncorrectOutput
__property double AnswerError[int a]
__property double Weights[int l][int n][int w]
__property AnsiString ClassGropus[int a]
__property AnsiString Predictions[int a]
```

Şekil 27: Yapay sinir ağı sınıfında kullandığımız özellik listesi.

Yukarıdan aşağı doğru özellik listesi açıklamaları ile birlikte şu şekildedir:

1) Yapay sinir ağında giriş katmanı ve çıkış katmanı da dahil olmak üzere kaç tane katman vardır (LayerCount) ?
2) Katman indisleri 0 ‘dan başlamak ve LayerCount-1 ‘e dek devam etmek üzere her katmandaki nod sayısı NodesCount[int a] ile verilir.
3) Öğrenme oranı sabit ya da adaptif kullanılabilir. Öğrenme oranı eğer sabit ise LearningRate değeri ile verilir. Aksi halde adaptif şekliyle kullanılır.
4) Momentum katsayısı ile birlikte geliştirilmiş geri yayılım algortiması kullanılacak ise MomentumRate ile oran verilir. Aksi halde 0 olarak kabul edilir.
5) Yapay sinir ağının nerede öğrenme olayı sonlandıracağı StopCriteria ile belirlenir.

6) Ağırlıkların başlangıç değerleri WeightStatus ile verilmiştir. 0 olması durumunda tüm ağırlıklar 0 (ağırlıkları 0 olarak atanması sadece deneysel amaçlı olarak ağdaki bağlantıların doğruluğunu test etmek için kullanılır), 1 olması durumunda tüm ağırlıklar 1 ve 2 olması haline ağırlıkları verilen kriterler doğrultusunda rastlantısal olarak atanacaktır.
7) Ağırlıkların (-1,+1) aralığında ve rastlantısal atanması halinde ağırlık çarpanı WeightFactor ile tanımlanmalıdır. 1000 olması halinde ağırlık çarpanı 1/1000 şeklinde modele yansıtılır.
8) Ağın yaptığı toplam hata miktarı ve bu hatanın grafik şekilde verilmesi Error ve ErrorForGraph ile belirlenir.
9) Girişler, istenilen çıkışlar ve ağırlıklar sırası ile Input [int a] , Desired[int a], Weight[int l][int n][intw] özellikleri ile belirlenir.

Diğer parametreler yapay sinir motoru çalışırken online veri toplama amacına hizmet için bulunur. BOT, sayesinde çok katlı yapay sinir ağları parametrik biçimde oluşturulmuş ve derlenmiştir. Birden çok yapay sinir ağının tek bir ağmışçasına çalışmasına müsaade edecek biçimde geliştirilmiştir.

3.8. Fare

Solucan veri tabanı genel bilgi ihtiyaçlarını karşılamak ve özel bir takım verilere ulaşmak için kullanılan bilgileri depolamaktadır. Bu bilgilerden yola çıkarak daha özel bilgileri toparlamasına ihtiyaç duyulmuştur. İkincil yapı elementleri, bunların protein içindeki yerleşimleri, disülfit bağlanmaları, protein içindeki varyasyonlar ile mutasyonlar ve SNP noktaları, molekül ağırlıkları, NMR ve X-RAY dosyaları ile atomsal koordinatlar bu özel bilgilerden bazılarıdır. Fare, Solucan tabanına sahip bir yazılım olup bahsedilen özel bilgilerin otomatik olarak toparlanıp derlenmesi için geliştirilmiştir. Ayrıca Fare,

yapay sinir ağları ile veri tabanları arasındaki ilişkiyi de sağlamaktadır. Fare dört program parçasından oluşmuştur. Bu program parçaları aşağıda sıralanmıştır:

1) Protein listeleri: Solucan veri tabanında bulunup hakkında Fare tarafından özel bilgileri edinilip düzenlenmiş olan proteinlerin listesidir,
2) Proteinlere ait ikincil yapı elementleri ve bunların dizileri,
3) Yapay sinir ağı tasarım modülü,
4) Yapay sinir ağı eğitim ve sonuçları

Yapay sinir ağlarının farklı modeller ile eğitilmesi ve eğitilen modelin test edilip ikincil yapıların sonuçlarının gösterilmesi ve proteine ait üçüncül yapının tahmini içinde Fare adlı yazılım kullanılmıştır.

3.9. Hizalayıcı

Fare yazılımının ürettiği ikincil yapının yapısı bilinen ve üç boyutlu NMR çalışmaları yapılmış proteinler ile karşılaştırılması yerel ve global hizalama yapabilen özel bir yazılıma ihtiyaç duyulmuştur. Bu nedenle Hizalayıcı isimli yazılım geliştirilmiştir.

3.10. Amino asitlerin sayısallaştırılması

Amino asitlerin yapay sinir ağına girdi olabilmeleri için sayısallaştırılmaları gerekmiştir. Sayısallaştırma modelinde genel olarak kabul gören istenilen durumların (doğru, true) 1 ile, istenmeyen durumların (yanlış, false) ise 0 ile gösterildiği ikili sayısallaştırma kullanılmıştır. Tüm sayısallaştırma yöntemlerinde iki temel kural bulunmaktadır. Bunlardan ilki amino asitler hakkında çok fazla bilgiyi taşıyor olması (özellik çıkartımı), ikincisi ise her amino asitin yalnızca bir kodla ifade ediliyor olmasıdır (tek anlamlılık).

İki model tasarlanmıştır. Birinci model kurulurken amino asitlerin fiziko kimyasal yapılarına göre sınıflandırılmasında kullanılan ve Şekil-4 ‘de de gösterilen venn

şemasından ilham alınmıştır. Amino asitler sırası ile yazıldıktan sonra dahil oldukları gruplara karşılık gelen sahalar “1” ile, dahil olmadığı diğer durumlar ise “0” ile ifade edilmek üzere sayısallaştırma tamamlanmıştır. Amino asitlerin belirlenen bu modele göre sayısallaştırması sonucunda elde edilen sayısal amino asit modeli Tablo 15 ‘de gösterilmiştir.

Tablo 15: Amino asitlerin gruplara göre sınıflandırmasından yola çıkarak elde edilen sayısallaştırma tablosu

Grup	**P**	**N**	**D**	**S**	**C**	**T**	**G**	**A**	**V**	**I**	**L**	**M**	**F**	**W**	**Y**	**H**	**K**	**R**	**E**	**Q**
Küçük	1	1	1	1	1	1	1	1	1	0	0	0	0	0	0	0	0	0	0	0
Hidrofobik	0	0	0	0	1	1	1	1	1	1	1	1	1	1	1	1	1	0	0	0
Polar	0	1	1	1	1	1	0	0	0	0	0	0	0	1	1	1	1	1	1	1
Çok Küçük	0	0	0	1	1	0	1	1	0	0	0	0	0	0	0	0	0	0	0	0
Alifatik	0	0	0	0	0	0	0	0	1	1	1	0	0	0	0	0	0	0	0	0
Aromatik	0	0	0	0	0	0	0	0	0	0	0	0	1	1	1	1	0	0	0	0
Pozitif	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	1	1	0	0
Negatif	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0
Ek Bit 1	0	0	0	0	0	0	0	1	0	0	1	0	0	0	1	0	0	0	0	0

Burada kullanılan Ek Bit 1 isimli son bit; en küçük alt grupta iki amino asitin birlikte ifade edilmesi durumu için, iki amino asitin sayısal olarak birbirinden ayırt edilmesi amacıyla konulmuştur. Tüm amino asitlerin sayısallaştırılmış hali ise Tablo 16a ‘da gösterilmiştir.

Tablo 16a: Amino asitlerin sayısallaştırılmış halleri ile standart olmayan amino asitler (Solucan adlı programın ekran çıktısından alınıştır).

idBin	Aminoacid	BinSys1
1	P	100000000
2	N	101000000
3	D	101000010
4	S	101100000
5	C	111100000
6	T	111000000
7	G	110100000
8	A	110100001
9	V	110010000
10	I	010010000
11	L	010010001
12	M	010000000
13	F	010001000
14	W	011001000
15	Y	011001001
16	H	011001100
17	K	011000100
18	R	001000100
19	E	001000010
20	Q	001000000
21	O	100000000
22	U	111100000
23	B	101000000
24	Z	001000000
25	X	100000000

Tablo 16a incelendiğinde, bazı amino asitlerin standart kodlamaya uymadıkları görülebilir. Bu kodlar ve bu kodların açıklamaları ise Tablo 16b ‘da yapılmıştır. Expasy/Swiss-Prot güncel sürümünde yeni eklenen amino asit kodları sırası ile O, U, B , Z, X olarak verilmiştir. Üç harf kodlamaları Pyl, Sec, Asc, Glx, Xaa şeklindedir. O, pirolizin amino asitidir. U, selenosistein amino asitidir. B aspartik asit ya da asparajin amino asitleridir. Z, glutamik asit ya da glutamin amino asitleridir. X, ise herhangi bir amino asit yerine kullanılan tek harf kodlamasıdır (Expasy/Swiss-Prot, http://expasy.org/sprot/userman.html, 01.01.2011.)

Tablo 16b: Standart olmayan amino asitler ve kodlaması (http://expasy.org/sprot/userman.html ‘dan uyarlanarak alınmıştır).

Tek Harf Kodu	**Üç harf kodu**	**Açıklama**
O	Pyl	Pirolizin
U	Sec	Selenosistein
B	Asx	Asparajin ya da aspartik asit
Z	Glx	Glutamin ya da glutamik asit
X	Xaa	Herhangi bir amino asit için kullanılır

Benzer bir şekilde amino asitler benzer gruplara dahil olanlar alt alta yazılmak şartı ile elde edilen Tablo 1 ‘den yola çıkılarak 20 bitlik bir ikilik sayısallaştırma yapılmıştır. Bu sayede giriş örüntülerinin hem daha basit olması hem de amino asitlerin fizikokimyasal özelliklerini ifade etmesi sağlanmıştır. Tablo 17 ‘de bu ikinci model için amino asitler ve amino asitlerin sayısallaştırılmış halleri gösterilmiştir.

Tablo 17: Amino asitlerin sayısallaştırılmış halleri ile standart olmayan amino asitler (Tablo-1 ‘dan yola çıkarak hazırlanmıştır. Solucan adlı programın ekran çıktısından alınmıştır).

idBin	Aminoacid	BinSys2
1	P	10000000000000000000
2	N	01000000000000000000
3	D	00100000000000000000
4	S	00010000000000000000
5	C	00001000000000000000
6	T	00000100000000000000
7	G	00000010000000000000
8	A	00000001000000000000
9	V	00000000100000000000
10	I	00000000010000000000
11	L	00000000001000000000
12	M	00000000000100000000
13	F	00000000000010000000
14	W	00000000000001000000
15	Y	00000000000000100000
16	H	00000000000000010000
17	K	00000000000000001000
18	R	00000000000000000100
19	E	00000000000000000010
20	Q	00000000000000000001
21	O	10000000000000000000
22	U	00001000000000000000
23	B	01000000000000000000
24	Z	00000000000000000001
25	X	10000000000000000000

3.11. Yapay sinir ağı topolojileri

Yapay sinir ağlarının giriş değerleri amino asitlerin sayısallaştırma modeline uygun olarak ayarlanmıştır. Yapay sinir ağlarının çıkış değerleri istenilen değerler her zaman “1” ile istenmeyen değerler ise her zaman “0” ile temsil edilmiştir. Tüm modellerin ortak tahmin sınıfları için istenilen çıkışlar şunlardır:

1) Heliks yapısındaki amino asitler için istenilen çıkış matrisi $d = [01]$ şeklindedir.

1) Beta tabaka yapısındaki amino asitler için istenilen çıkış matrisi $d = [10]$ şeklindedir.

Burada *d* matrisi yapay sinir ağının istenilen çıkışı simgeler.

Kanat boyutu ile belirlenen pencere kayması modeline göre elde edilmiş pencere dizisi ağlara, amino asitler sayısallaştırıldıktan sonra, giriş olarak sunulmuştur. Bu katmanda toplam 2 adet ağ bulunmaktadır. Bu ağlar sırası ile Ann1, Ann2 olarak adlandırılmıştır..

Ann1, bu ağ ikincil yapıyı iki seviyede tahmin etmek için dizayn edilmiştir. İki adet çıkışı bulunur. Eğer gelen amino asit heliks yapısal sınıfından ise 1, beta tabaka yapısal sınıfından ise 0 değerini almaktadır. Tablo 18 ‘de Ann1 ağı için çıkış ve çıkışa ait kodlamalar gösterilmiştir.

Tablo 18: Ann1 ağı için çıkış sınıfları

Yapısal sınıf	**Çıkış 1**	**Çıkış 2**
Heliks	1	0
Beta tabaka	0	1

Ann2, bu ağ ikincil yapıyı iki seviyede tahmin etmek için dizayn edilmiştir. Girişlerini Ann1 ağının çıkışlarından alır. İki adet çıkışı bulunur. Eğer gelen amino asit heliks yapısal sınıfından ise 1, beta tabaka ya da zincir yapısal sınıfından ise 0 değerini almaktadır. Tablo 19 ‘da Ann2 ağı için çıkış ve çıkışa ait kodlamalar gösterilmiştir.

Tablo 19: Ann2 ağı için çıkış sınıfları

Yapısal sınıf	**Çıkıış 1**	**Çıkış 2**
Heliks	1	0
Beta tabaka	0	1

4. BULGULAR

İkincil yapı element tarayıcısı kullanılarak amino asitlerin hangi ikincil yapı elementlerinde hangi oranlarda bulunduğu hesaplanmıştır. Bu sayede amino asitlerin belirgin bir şekilde ikincil yapı elementlerinde bulunma oranlarının farklı olduğu sonucuna ulaşılmıştır. Sonuçlar aşağıda sunulmuştur.

4.1 İkincil yapı element tarayıcısı ile elde edilen bulgular

İkincil yapı tarayıcısı ile Solucan ve Fare kişisel veri tabanlarında yer alan ikincil yapı elementleri ve bunların amino asit dizileri analiz edilmiştir. Bu sayede hangi amino asitin hangi ikincil yapı elementinde daha fazla bulunmaya eğilimli olduğu ortaya konulmuştur.

4.1.1. Heliks yapısı ve amino asitleri

Bu tez için hazırlanmış kişisel veri tabanları (Solucan ve Fare veri tabanları) kullanılarak heliks yapı ve yapıda bulunan amino asitler incelenmiştir. Bu incelemede 27391 adet heliks yapısı taranmıştır. Toplam rezüdü (rezüdü, protein zincirinde yer alan amino asitler, ya da, DNA ve RNA zincirlerinde yer alan nükleotidler için kullanılır) miktarı 219993 adettir. En küçük heliks yapısının uzunluğu 2 rezüdü ve en büyük heliks yapısının uzunluğu 107 rezüdüdür. 107 rezüdülük dev bir helikse sahip olan protein kodlayan bir gen olan NDEL1 simgeli 17p13.1 lokalizasyonunda bulunan nud E nuclear distribution gene E homolog like 1 isimli gendir. Ortalama heliks boyu 8.0313 rezüdü olarak hesaplanmıştır. Tablo 20a, Tablo 20b, Tablo 20c ‘de amino asitlerin ayrıntılı frekans bilgileri gösterilmiştir.

Tablo 20a: Heliks yapısındaki amino asitler ve bulunma oranları

Amino asit	**Heliks yapıdaki sayısı**	**Bulunma oranı**
G	4916	0,0223461655598133
A	18191	0,0826889946498298
P	2261	0,01027759974181
V	12395	0,0563427018132395
L	24597	0,111808102985095
I	11820	0,0537289822858
M	5619	0,0255417217820567
F	8306	0,037755746773761
Y	6217	0,0282599900905938
W	2766	0,0125731273267786
S	9667	0,0439423072552308
T	7598	0,0345374625556268
C	3571	0,0162323346651939
N	5874	0,0267008495724864
Q	10993	0,0499697717654653
K	13585	0,0617519648352448
H	4563	0,0207415690499243
R	11893	0,0540608110258054
D	9184	0,0417467828521817
E	18586	0,0844845063252013

Tablo 2a ‘da veriler yüzde cinsinden ifade edilmiştir. İkincil Yapı Element Tarayıcısı ile elde edilen sonuçlara göre heliks yapısında en çok bulunan amino asitler (%6 ‘lık dilimden daha fazla orana sahip olanlar) L (%11), E (%8), A(%8), K(%6) ‘dır. Nadir bulunanlar (%2 ‘den az orana sahip olanlar) ise P(%1), W(%1), C(%1) ‘dir.

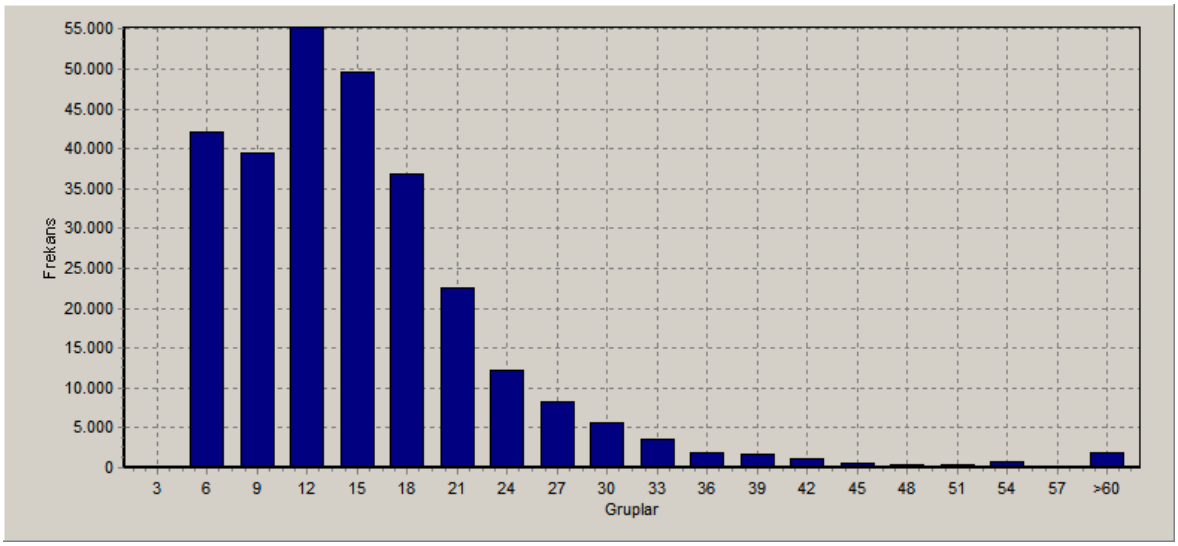
Tablo 20b: Heliks yapıda %6’dan çok bulunan amino asitler.

Amino asit	**Heliks yapıdaki sayısı**	**Bulunma oranı**
A	18191	0,0826889946498298
L	24597	0,111808102985095
K	13585	0,0617519648352448
E	18586	0,0844845063252013

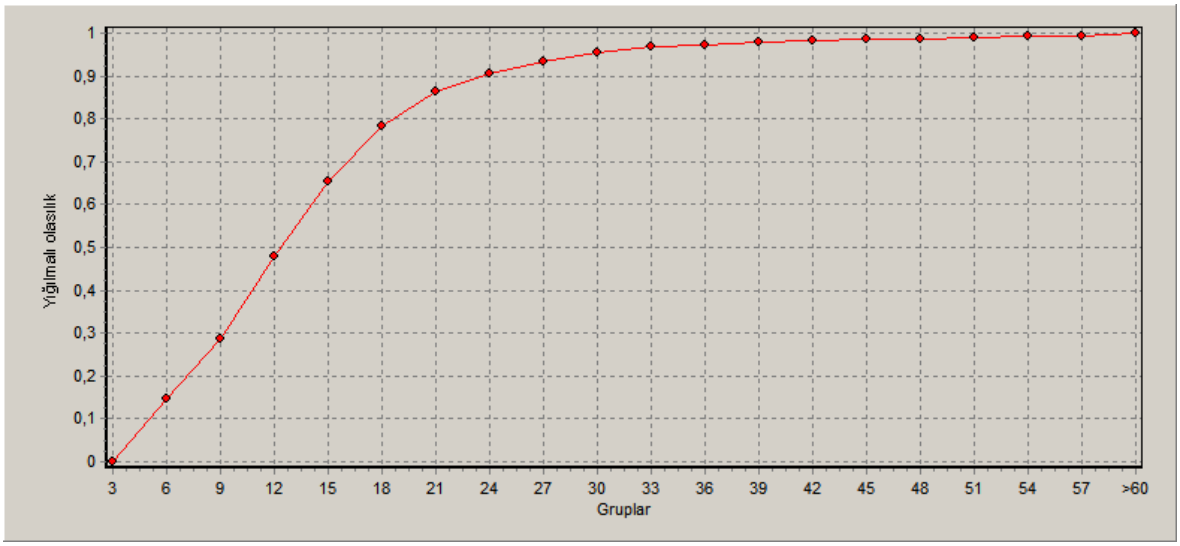
Tablo 20c: Heliks yapıda %2 ‘den az bulunan amino asitler.

Amino asit	**Heliks yapıdaki sayısı**	**Bulunma oranı**
P	2261	0,01027759974181
W	2766	0,0125731273267786
C	3571	0,0162323346651939

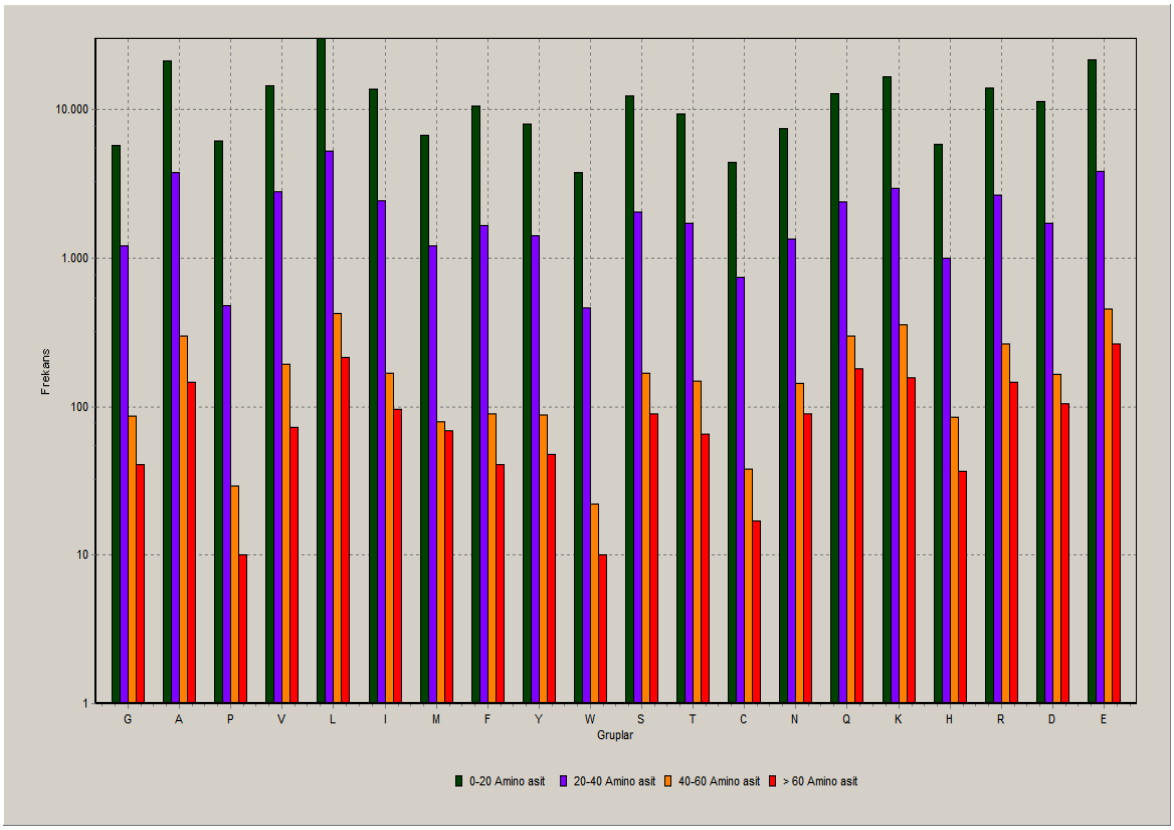
Heliks yapılarının uzunlukları ile ilgili gruplanmış histogram grafiği Şekil 28a ‘da, olasılık yoğunluk grafiği Şekil 28b ve gruplanmış heliks boylarında amino asitlerin bulunma frekansları ile ilgili histogram grafikleri Şekil 28c ‘ de gösterilmiştir.



Şekil 28a: Uzunluk olarak 3 artırmalı olarak oluşturulmuş gruplarda heliks yapının uzunluğu ile ilgili olarak hazırlanmış histogram (İkincil Yapı Element Tarayıcı kullanılarak elde edilmiştir).



Şekil 28b: Uzunluk olarak 3 artırmalı olarak oluşturulmuş gruplarda heliks yapının uzunluğu ile ilgili olarak hazırlanmış olasılık yoğunluk grafiği (İkincil Yapı Element Tarayıcı kullanılarak elde edilmiştir).



Şekil 28c: Amino asitlerin uzunluklarına göre gruplanmış heliks yapılarında bulunmalarına ilişkin histogram (İkincil Yapı Element Tarayıcı kullanılarak elde edilmiştir).

4.1.2 Beta tabaka

Toplam beta tabaka sayısı 32447 adettir. Bu beta tabakalarda toplam rezüdü sayısı 189646 adettir. En küçük beta tabaka formasyonu 3 rezüdüden, en büyük beta tabaka formasyonu ise 117 rezüdüden oluşmaktadır. Rezüdü uzunluklarının aritmetik ortalaması 5.8485 ve standart sapmaları ise 2.9515 olarak bulunmuştur. Amino asitlerin bu yapıdaki bulunma frekansları için Tablo 21a, hangi amino asitlerin hangi yapıyı daha sık tercih ettiklerini görmek için Tablo 21b ve hangi amino asitlerin hangi yapıda nadir olduğu bilgisi için Tablo 21c hazırlanmıştır. Ayrıca Şekil 29a, Şekil 29b ve Şekil 29c 'de ise beta tabaka boyları ve beta tabaka boylarının olasılık yoğunluk grafikleri ile beta tabakalarda amino asitlerin bulunma sıklıklarını gruplanmış veriler üzerinde histogram şeklinde hazırlanmak suretiyle gösterilmiştir.

Tablo 21a: Beta tabaka yapısında amino asitlerin bulunma frekansları (İkincil Yapı Element Tarayıcısı ile hazırlanmıştır).

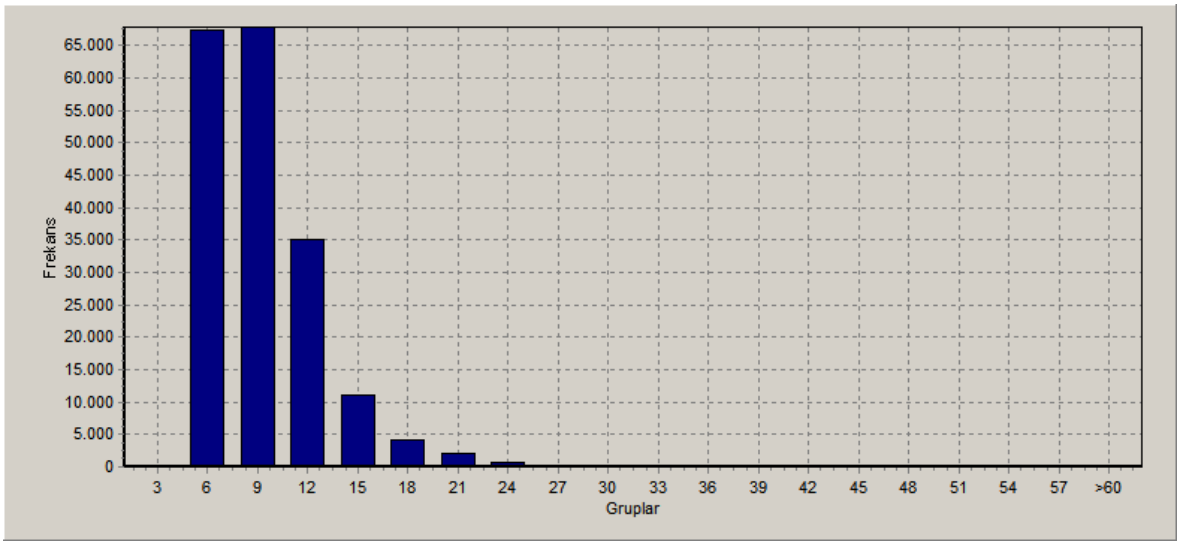
Amino asit	**Beta tabakadaki frekansı**	**Bulunma oranı**
G	10067	0,0530833926546969
A	10002	0,052740646998339
P	4805	0,0253368135199979
V	21899	0,115473648132036
L	19425	0,102428221150044
I	14940	0,0787787708613462
M	3978	0,0209760341691054
F	10917	0,0575654512378391
Y	8702	0,0458857338711804
W	3380	0,0178227741306125
S	10703	0,0564370270769069
T	11557	0,0609401777004403
C	5624	0,0296554087901078
N	5793	0,0305465474966385
Q	6417	0,0338369057976746
K	10061	0,05305175459411
H	4827	0,0254528197421498
R	9194	0,048480054839305
D	7305	0,0385193387645337
E	10046	0,0529726594426428

Tablo 21b: Beta tabaka yapısında %6'dan daha sık bulunana amino asitler (İkincil Yapı Element Tarayıcısı ile hazırlanmıştır).

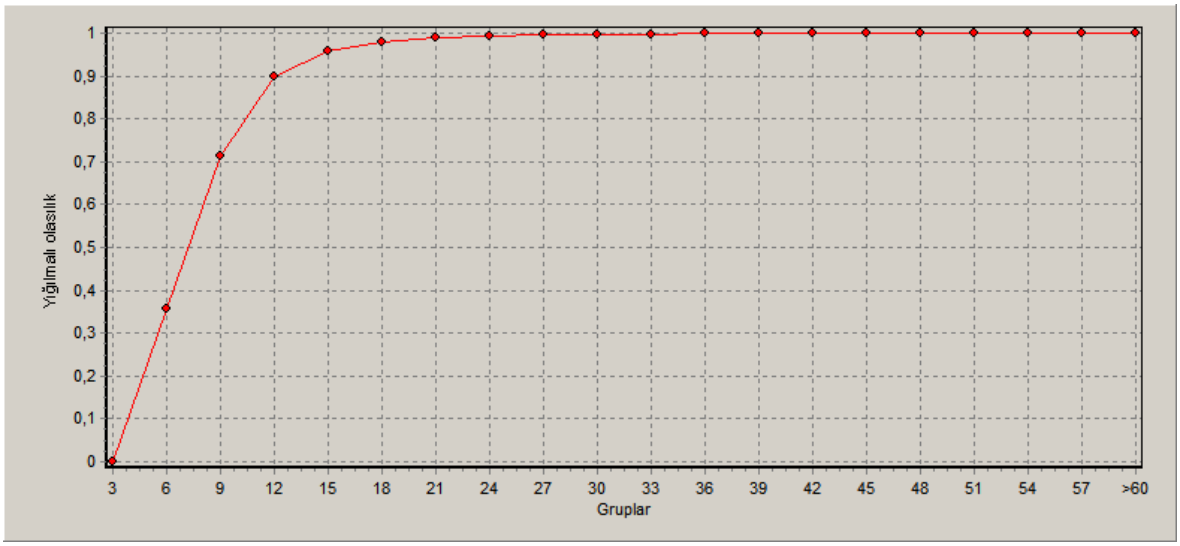
Amino asit	Beta tabakadaki frekansı	Bulunma oranı
V	21899	0,115473648132036
L	19425	0,102428221150044
I	14940	0,0787787708613462
T	11557	0,0609401777004403

Tablo 21c: Beta tabaka yapısında %2.5'den daha az sıklıkta bulunan amino asitler (İkincil Yapı Element Tarayıcısı ile hazırlanmıştır).

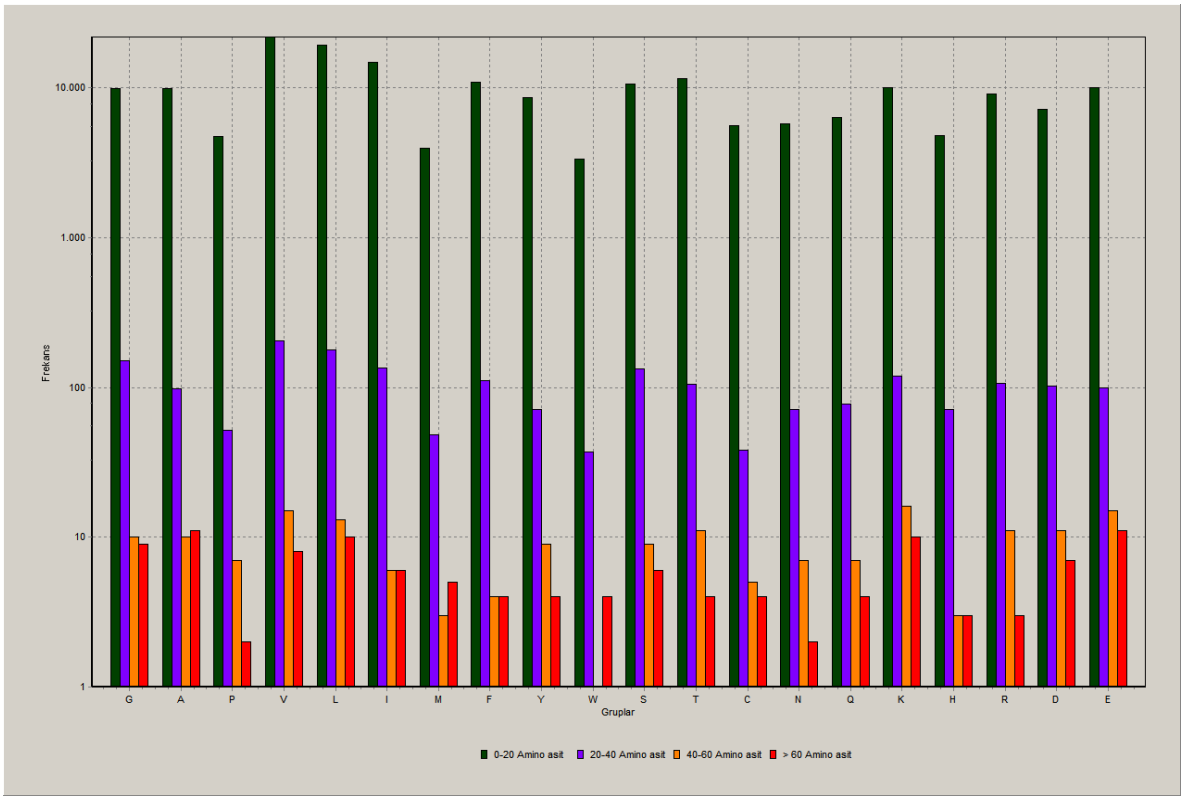
Amino asit	Beta tabakadaki frekansı	Bulunma oranı
M	3978	0,0209760341691054
W	3380	0,0178227741306125



Şekil 29a: Beta tabakaların gruplanmış verilerde sahip oldukları rezüdü sayıları (İkincil Yapı Element Tarayıcısı ile hazırlanmıştır).



Şekil 29b: Beta tabakaların gruplanmış verilerde sahip oldukları rezüdü sayılarının olasılık yoğunluk fonksiyonları (İkincil Yapı Element Tarayıcısı ile hazırlanmıştır).



Şekil 29c: Amino asitlerin uzunluklarına göre gruplanmış beta tabaka yapılarında bulunmalarına ilişkin histogram (İkincil Yapı Element Tarayıcı kullanılarak elde edilmiştir).

4.1.3. Dönüş yapısı

Toplam 7710 adet dönüş yapısı incelenmiştir. Bu dönüş yapılarındaki toplam rezüdü miktarı 27211 ‘dir. Minimum dönüş yapısı 1 amino asit içerirken maksimum dönüş formasyonunda toplam 50 amino asit bulunmuştur. Amino asitlerin bu yapıdaki bulunma frekansları için Tablo 22a, hangi amino asitlerin hangi yapıyı daha sık tercih ettiklerini görmek için Tablo 22b ve hangi amino asitlerin hangi yapıda nadir oluğu bilgisi için Tablo 22c hazırlanmıştır. Dönüş yapısındaki ortalama amino asit miktarı 3.5302 iken standart sapma değeri 1.191 olarak bulunmuştur. Ayrıca Şekil 30a, Şekil 30b ve Şekil 30c ‘de ise dönüş yapısı boyları ve dönüş yapısı boylarının olasılık yoğunluk grafikleri ile dönüş yapısında amino asitlerin bulunma sıklıkları gruplanmış veriler üzerinde histogram şeklinde hazırlanmak suretiyle gösterilmiştir.

Tablo 22a: Dönüş yapılarında amino asitlerin frekansları ve bulunma oranları (İkincil Yapı Element Tarayıcı kullanılarak elde edilmiştir).

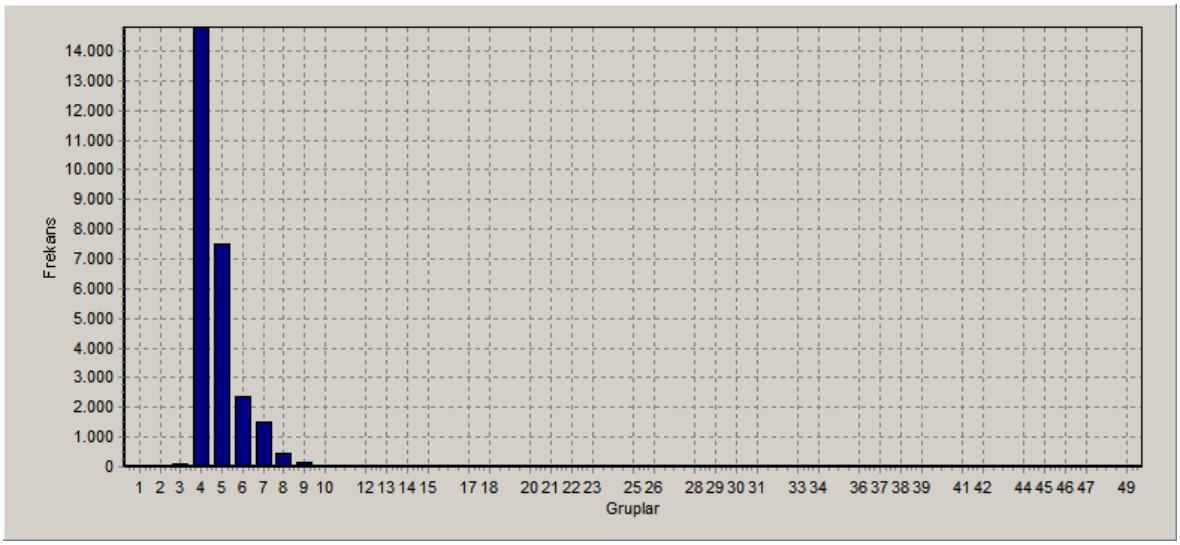
Amino asit	**Dönüş yapısındaki sayısı**	**Bulunma oranı**
G	2308	0,084818639520782
A	1643	0,0603799933850281
P	1664	0,0611517401051046
V	1033	0,0379625886589982
L	1924	0,0707066994965271
I	801	0,0294366248943442
M	498	0,0183014222189556
F	1046	0,0384403366285693
Y	875	0,0321561133365183
W	293	0,010767704237257
S	1859	0,0683179596486715
T	1534	0,0563742604093933
C	720	0,0264598875454779
N	1526	0,0560802616588879
Q	1156	0,0424828194480173
K	1972	0,072470691999559
H	762	0,0280033809856308
R	1535	0,0564110102532064
D	1853	0,0680974605857925
E	2208	0,0811436551394657

Tablo 22b: Dönüş yapısında %6 ‘dan daha sık bulunan amino asitler (İkincil Yapı Element Tarayıcı kullanılarak elde edilmiştir).

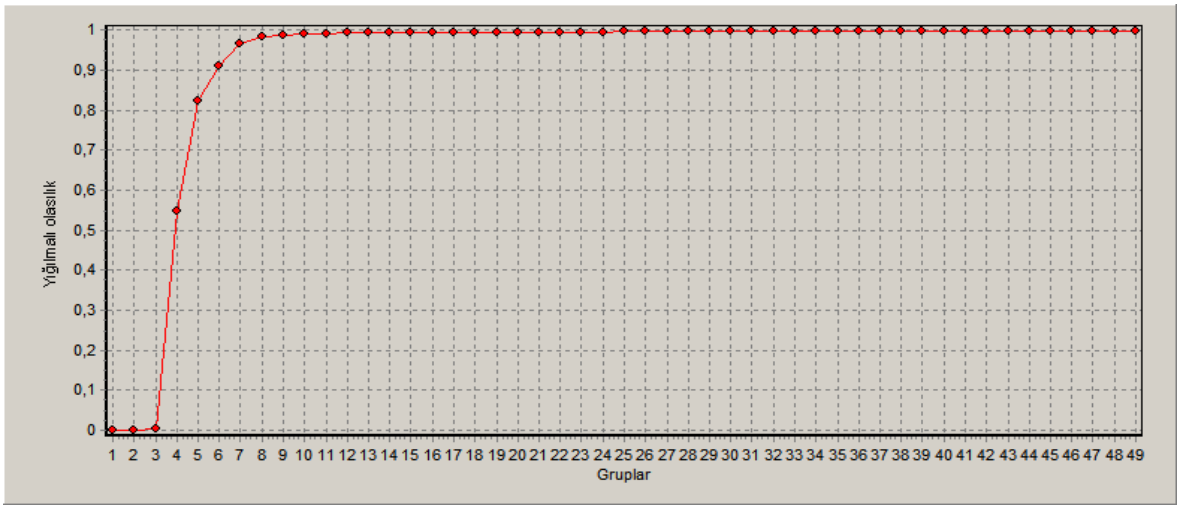
Amino asit	**Dönüş yapısındaki sayısı**	**Bulunma oranı**
G	2308	0,084818639520782
A	1643	0,0603799933850281
P	1664	0,0611517401051046
L	1924	0,0707066994965271
S	1859	0,0683179596486715
K	1972	0,072470691999559
D	1853	0,0680974605857925
E	2208	0,0811436551394657

Tablo 22c: Dönüş yapısında %2 ‘dan daha az bulunan amino asitler (İkincil Yapı Element Tarayıcı kullanılarak elde edilmiştir).

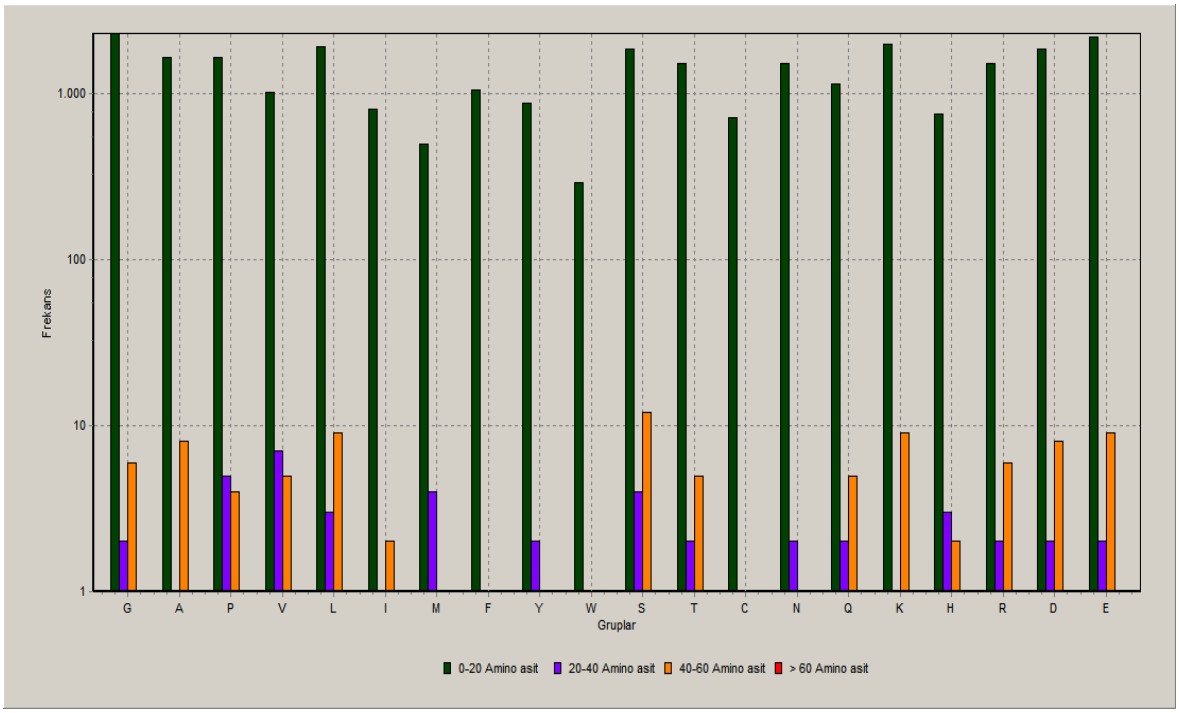
Amino asit	**Dönüş yapısındaki sayısı**	**Bulunma oranı**
M	498	0,0183014222189556
W	293	0,010767704237257



Şekil 30a: Dönüş yapılarının uzunluklara göre gruplanmış verilerinde bulunan toplam rezüdü miktarını gösteren histogram (İkincil Yapı Element Tarayıcı kullanılarak elde edilmiştir).



Şekil 30b: Dönüş yapılarının uzunluklara göre gruplanmış verilerinde bu gruplara göre olasılık yoğunluk grafiği (İkincil Yapı Element Tarayıcı kullanılarak elde edilmiştir).



Şekil 30c: Amino asitlerin uzunluklarına göre gruplanmış dönüş yapılarında bulunmalarına ilişkin histogram (İkincil Yapı Element Tarayıcı kullanılarak elde edilmiştir).

Amino asitlerin tümünün her üç yapı formunda bulunma oranları toplu bir şekilde Tablo 23 ‘da gösterilmiştir.

Tablo 23: Amino asitlerin İkincil Yapı Element Tarayıcısı kullanılarak elde edilmiş her üç yapıda ayrı ayrı bulunma oranları.

Amino asit	**Heliks yapı**	**Beta tabaka**	**Dönüş yapısı**
G	0,0223461655598133	0,0530833926546969	0,084818639520782
A	0,0826889946498298	0,052740646998339	0,0603799933850281
P	0,01027759974181	0,0253368135199979	0,0611517401051046
V	0,0563427018132395	0,115473648132036	0,0379625886589982
L	0,111808102985095	0,102428221150044	0,0707066994965271
I	0,0537289822858	0,0787787708613462	0,0294366248943442
M	0,0255417217820567	0,0209760341691054	0,0183014222189556
F	0,037755746773761	0,0575654512378391	0,0384403366285693
Y	0,0282599900905938	0,0458857338711804	0,0321561133365183
W	0,0125731273267786	0,0178227741306125	0,010767704237257
S	0,0439423072552308	0,0564370270769069	0,0683179596486715
T	0,0345374625556268	0,0609401777004403	0,0563742604093933
C	0,0162323346651939	0,0296554087901078	0,0264598875454779
N	0,0267008495724864	0,0305465474966385	0,0560802616588879
Q	0,0499697717654653	0,0338369057976746	0,0424828194480173
K	0,0617519648352448	0,05305175459411	0,072470691999559
H	0,0207415690499243	0,0254528197421498	0,0280033809856308
R	0,0540608110258054	0,048480054839305	0,0564110102532064
D	0,0417467828521817	0,0385193387645337	0,0680974605857925
E	0,0844845063252013	0,0529726594426428	0,0811436551394657

Bu oranlar üzerinde yapılan testlerde amino asitlerin ikincil yapılarda bulunma oranları arasında önemli fark olduğu Two Samples KS test kullanılarak saptanmıştır ($P<0.001$).

4.2. Amino Asit

Amino Asit isimli yazılım ile bazı teorik proteinler tasarlanmıştır. Bu teorik proteinlerin 1-harf kodlama sisteminden 3-harf kodlama sistemine ve bunun tersi olan dönüşümleri yapılmıştır. Örnek olarak 98 amino asit uzunluğundaki

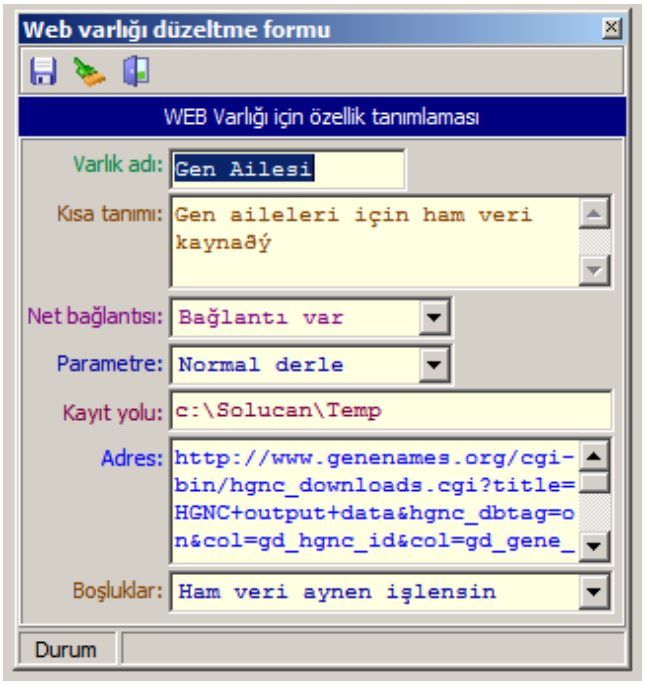
‘IAKPVSLLEKAAPQWCQGKLQAHLVAQTNLLRNQAEEELIKAQKVFEEMNVDLQ EELPSLWNSRVGFYVNTFQSIAGLEENFHKEMSKLNQNLNDVLV’ dizisi verilmiştir.

Bu dizilimin 3-harf kodlama sisteminde olan karşılığı Ile-Ala-Lys-Pro-Val-Ser-Leu-Leu-Glu-Lys-Ala-Ala-Pro-Gln-Trp-Cys-Gln-Gly-Lys-Leu-Gln-Ala-His-Leu-Val-Ala-Gln-Thr-Asn-Leu-Leu-Arg-Asn-Gln-Ala-Glu-Glu-Glu-Leu-Ile-Lys-Ala-Gln-Lys-Val-Phe-Glu-Glu-Met-Asn-Val-Asp-Leu-Gln-Glu-Glu-Leu-Pro-Ser-Leu-Trp-Asn-Ser-Arg-Val-Gly-Phe-Tyr-Val-Asn-Thr-Phe-Gln-Ser-Ile-Ala-Gly-Leu-Glu-Glu-Asn-Phe-His-Lys-Glu-Met-Ser-Lys-Leu-Asn-Gln-Asn-Leu-Asn-Asp-Val-Leu-Val şeklindedir. Bu dizilimin molekül ağırlığı 11207,6537 kD ‘dır. İzoelektrik noktası olan pI = 10,5299999999998 ‘dır.

4.3. Solucan

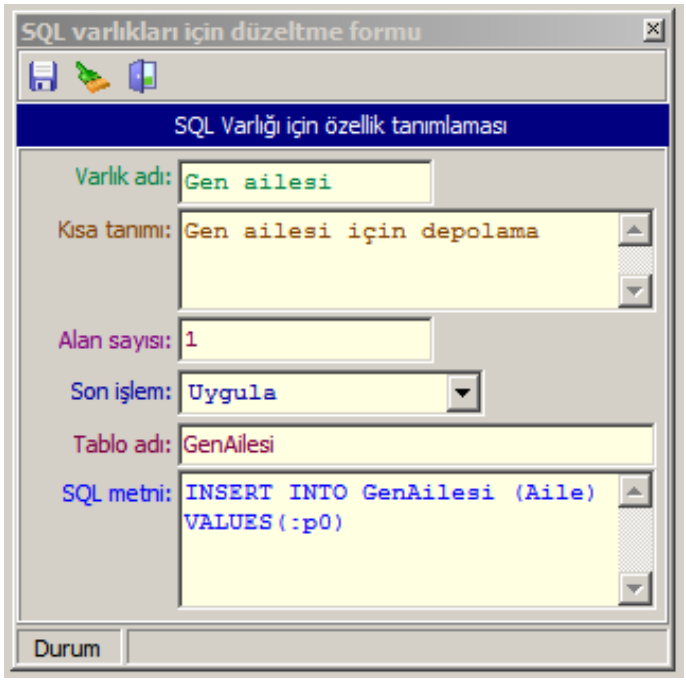
Tablo 9a, Tablo 9b, Tablo 9c, Tablo 9d ve Tablo 9e ‘de verilen bilgilerin yerli yerinde elde edilmesi için varlıkların yazılması gereklidir.

Gen ailesi isimlerinin istenilen adresten alınıp veri tabanının istediğimiz kısmına yazılabilmesi için öncelikle Web varlığı tanımlanmalıdır. Web varlığı tanımlamasına bir örnek Şekil 31 ‘da gösterilmiştir.

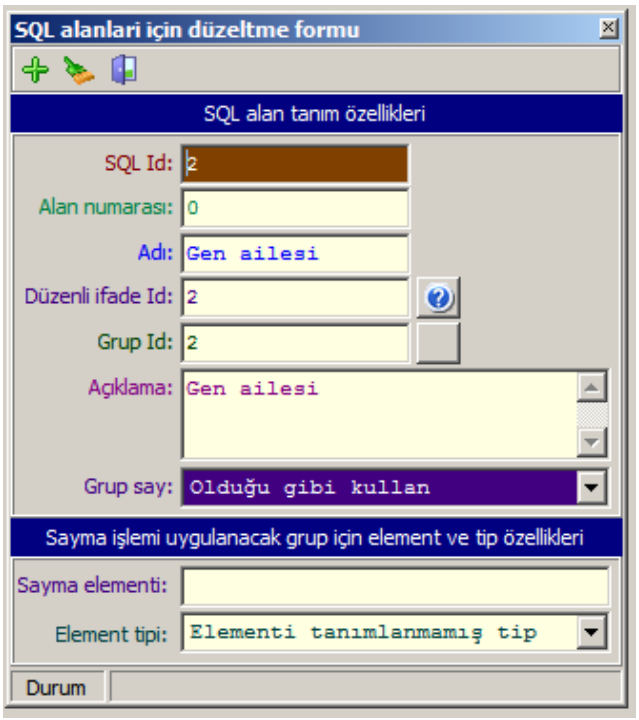


Şekil 31: Gen aileleri için ham veri kaynağı olan web varlığının parametreleri (SOLUCAN adlı programın ekran çıktısından alınmıştır).

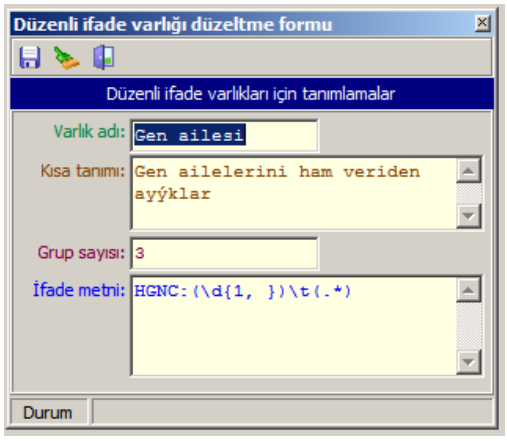
Web varlığı ile ham veri kaynağı Solucan ’a tanıtılmış olur. Solucan ‘ın istenilen işi yapabilmesi için düzenli ifade varlığının ve gruplarının hemen ardından SQL varlığı ve varlık alan parametrelerinin tamamlanmış olması gerekir. Şekil 32a, Şekil 32b ‘de sırası ile düzenli ifade varlığı ve Şekil 32c ve Şekil 32d ‘de grup parametreleri verilmiştir.



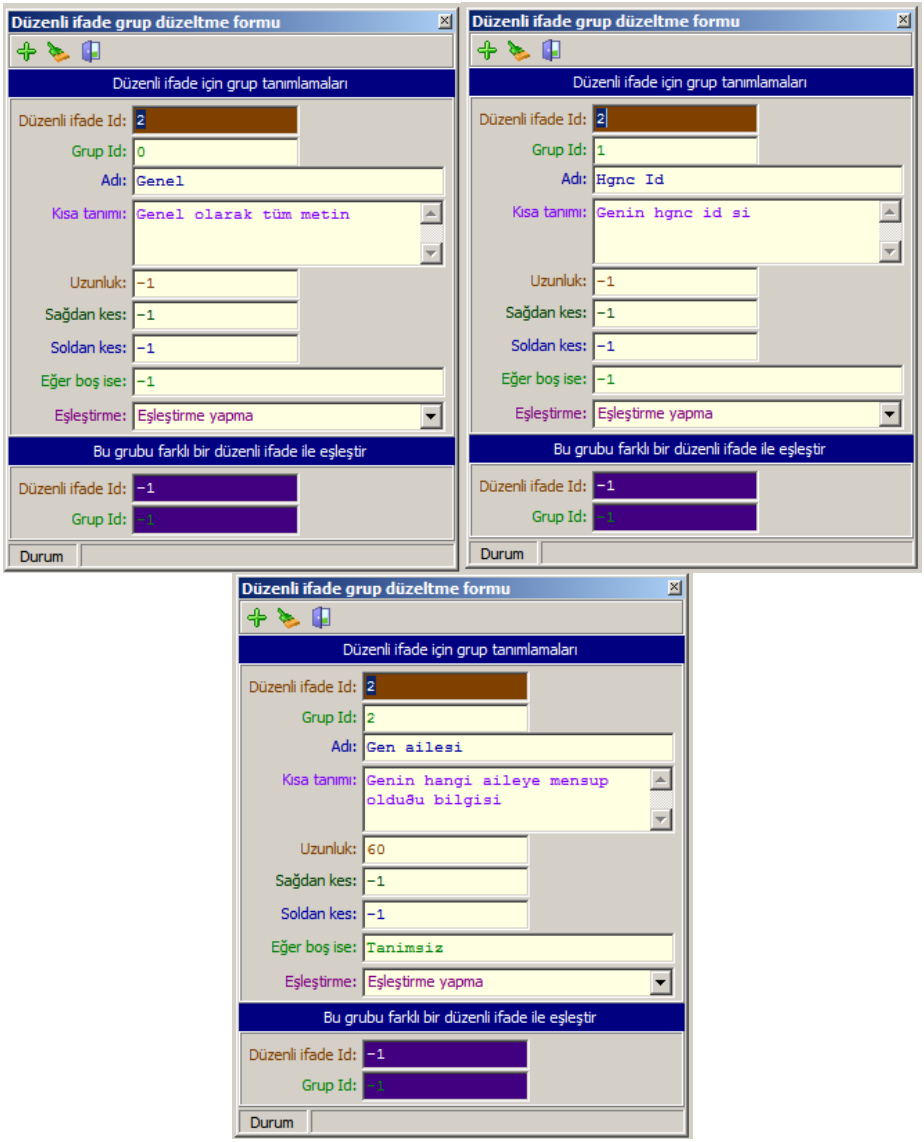
Şekil 32a: Lokus tiplerini ham veriden ayıklamak için gerekli olan düzenli ifade varlığı ve parametre değerleri (SOLUCAN adlı programın ekran çıktısından alınmıştır).



Şekil 32b: SQL varlığının alan parametreleri (SOLUCAN adlı programın ekran çıktısından alınmıştır).



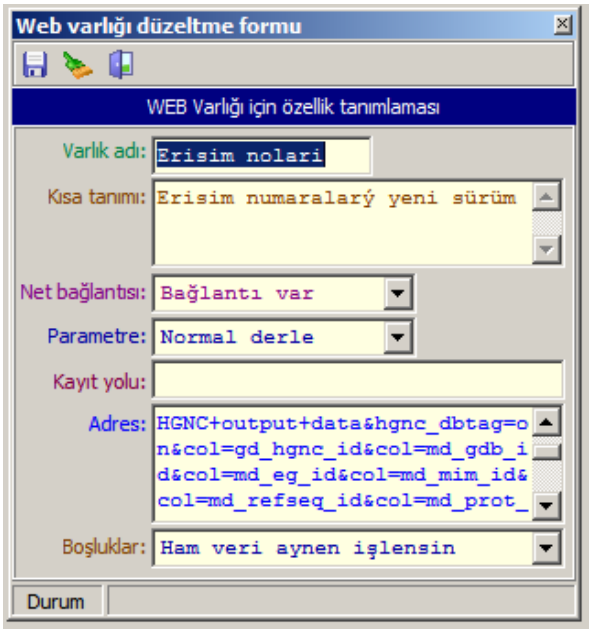
Şekil 32c: Gen aileleri için düzenli ifade varlığı ve parametreleri (SOLUCAN adlı programın ekran çıktısından alınmıştır).



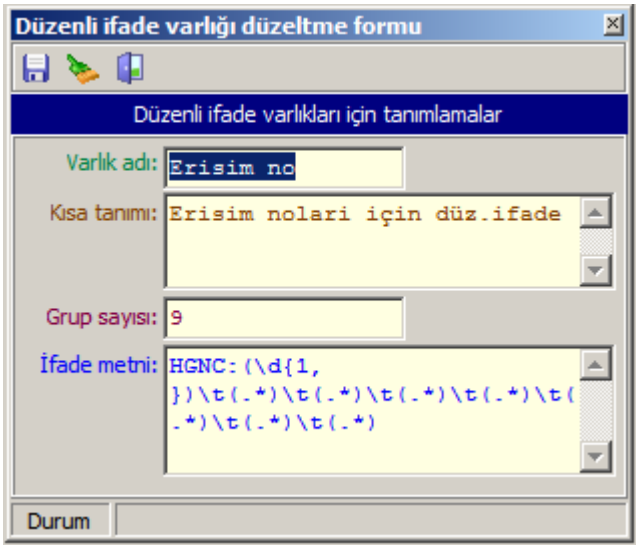
Şekil 32d: Düzenli ifade grup parametreleri. Grup 0, Grup 1 ve grup iki soldan sağa yazılmıştır (SOLUCAN adlı programın ekran çıktısından alınmıştır).

Yukarıdaki Şekil 32a, Şekil 32b, Şekil 32c ve Şekil 32d ‘deki parametrelere benzer fakat parametre adları ile parametre tanımlamaları lokus tipine göre ayarlanarak Solucan ‘ın lokus tipi bilgisini toplaması sağlanmıştır.

Oldukça önemli olan genlere ve proteinlere ait erişim numaralarının toplanması için Solucan ‘a Şekil 33a, Şekil 33b parametre setleri girilmiştir.



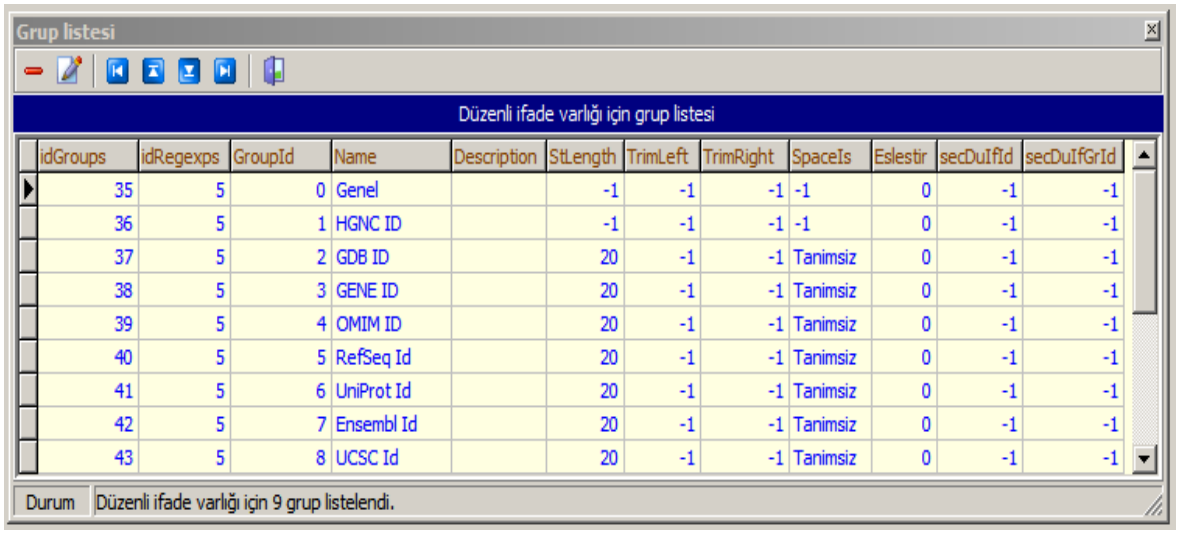
Şekil 33a: Erişim numaraları için ham veri kaynağı rolü ile web varlığı, ve varlık parametreleri (SOLUCAN adlı programın ekran çıktısından alınmıştır).



Şekil 3b: Erişim numaraları için düzenli ifade varlığı (SOLUCAN adlı programın ekran çıktısından alınmıştır).

Görüldüğü gibi burada ham veri kaynağından `HGNC:(\d{1,})\t(.*)\t(.*)\t(.*)\t(.*)\t(.*)\t(.*)` şeklinde düzenlenen bir düzenli ifade ile 9 ayrı grup elde edilmiştir. Bu dokuz grup ve düzenli ifade grup parametreleri Tablo 24 ‘da sunulmuştur.

Tablo 24: Düzenli ifadenin grup parametreleri. (SOLUCAN adlı programın ekran çıktısından alınmıştır).



Grup listesi

Düzenli ifade varlığı için grup listesi

idGroups	idRegexps	GroupId	Name	Description	StLength	TrimLeft	TrimRight	SpaceIs	Eslestir	secDuIfId	secDuIfGrId
35	5	0	Genel		-1	-1	-1	-1	0	-1	-1
36	5	1	HGNC ID		-1	-1	-1	-1	0	-1	-1
37	5	2	GDB ID		20	-1	-1	Tanimsiz	0	-1	-1
38	5	3	GENE ID		20	-1	-1	Tanimsiz	0	-1	-1
39	5	4	OMIM ID		20	-1	-1	Tanimsiz	0	-1	-1
40	5	5	RefSeq Id		20	-1	-1	Tanimsiz	0	-1	-1
41	5	6	UniProt Id		20	-1	-1	Tanimsiz	0	-1	-1
42	5	7	Ensembl Id		20	-1	-1	Tanimsiz	0	-1	-1
43	5	8	UCSC Id		20	-1	-1	Tanimsiz	0	-1	-1

Durum | Düzenli ifade varlığı için 9 grup listelendi.

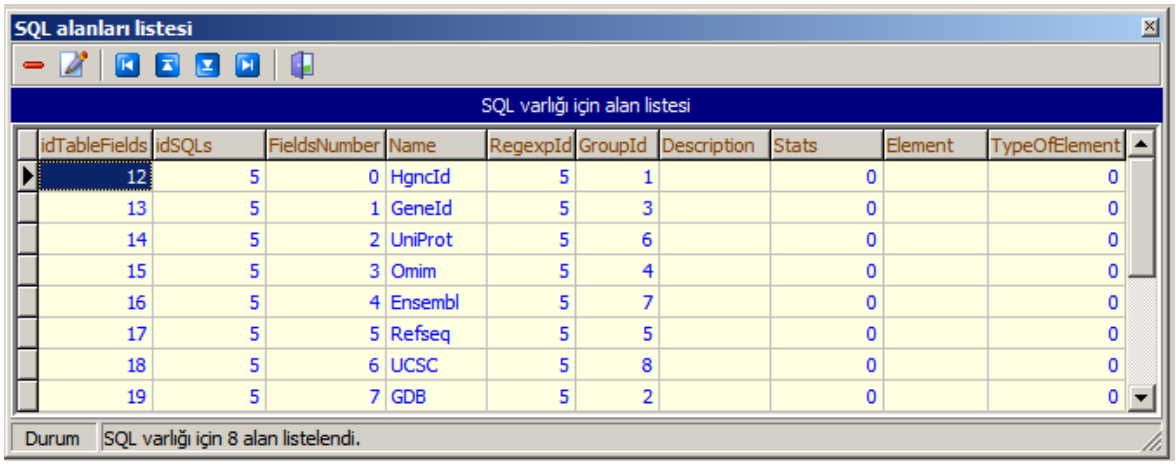
Burada ilk Grup 0 genel bir grup olup parçalanacak metni ifade eder. Grup 1 ‘de HgncId, Grup 2 ‘de GDBId, Grup 3 ‘de GeneId, Grup 4 ‘de OmimId, Grup 5 ‘de RefSeqId, Grup 6 ‘da UniProtId, Grup 7’de EnsemblId, Grup 8 ‘de UCSCId bilgileri bulunmaktadır. Uzunluk bilgileri stLength sütününda, kelime işleme parametreleri TrimLeft, TrimRight, SpaceIs sütunlarında verilmiştir. Geri kalan son 3 sütün SOLUCAN tarafından otomatik olarak kendi üzerinde dönüşlü bir grup oluşturmayacak şekilde (bu derlemede gerekli olmadığından) düzenlenmiştir.

Tablo 24 ’de verilen gruplardaki veriyi gerekli veri tablosuna yazdırmak için kullanılan SQL metni:

```
INSERT INTO erisim (HgncId, acGeneId,
acUniProtId, acOmimId, acEnsemblId, acRefSeqId,
acUCSCId, acGDBId)

VALUES( :p0,:p1,:p2,:p3,:p4,:p5,:p6,:p7)
```

yukarıda gösterildiği gibidir. Burada görüldüğü üzere toplam 8 adet grup 8 adet SQL alanı ile ilişkilendirilmiştir. Bu ilişkilere ait SQL varlığı alan parametreleri Tablo 25 ’de gösterilmiştir.

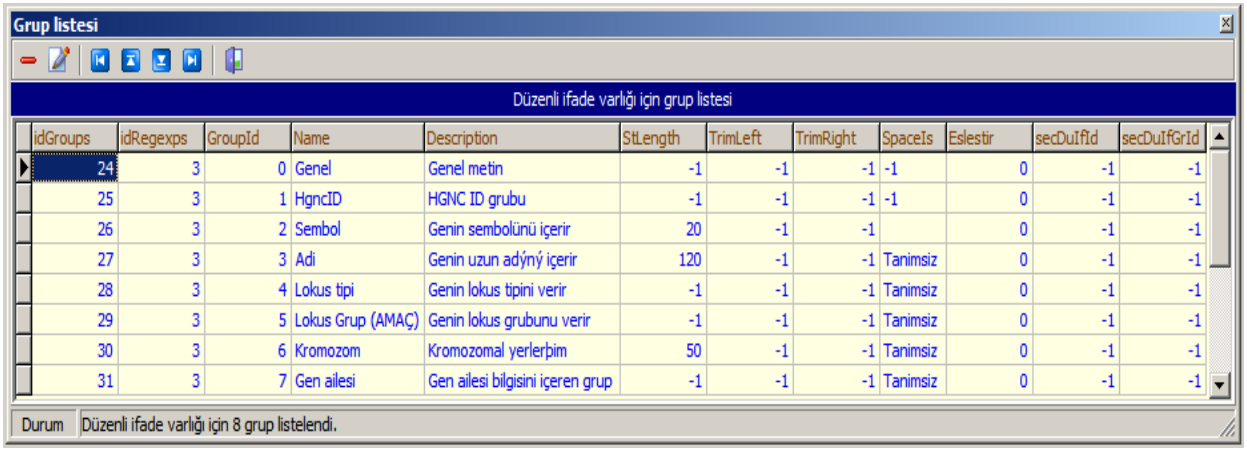
Tablo 25: SQL varlığının 8 alanı ile düzenli ifade varlığının 9 alanı arasındaki ilişki (SOLUCAN adlı programın ekran çıktısından alınmıştır).



idTableFields	idSQLs	FieldsNumber	Name	RegexpId	GroupId	Description	Stats	Element	TypeOfElement
12	5	0	HgncId	5	1		0		0
13	5	1	GeneId	5	3		0		0
14	5	2	UniProt	5	6		0		0
15	5	3	Omim	5	4		0		0
16	5	4	Ensembl	5	7		0		0
17	5	5	Refseq	5	5		0		0
18	5	6	UCSC	5	8		0		0
19	5	7	GDB	5	2		0		0

Çekirdek veriyi ham veriden bilgi şeklinde çıkartmak için `HGNC:(\d{1,})\t(.*)\t(.*)\t(.*)\t(.*)\t(.*)\t(.*)` şeklinde bir düzenli ifade kullanılmıştır. Metinden de anlaşılacağı üzere bu düzenli ifade de toplam 8 grup bulunmaktadır. Düzenli ifade grupları ve parametre değerleri Tablo 26 ‘de gösterilmiştir.

Tablo 26: Düzenli ifadenin 8 grubu (SOLUCAN adlı programın ekran çıktısından alınmıştır).



Grup listesi

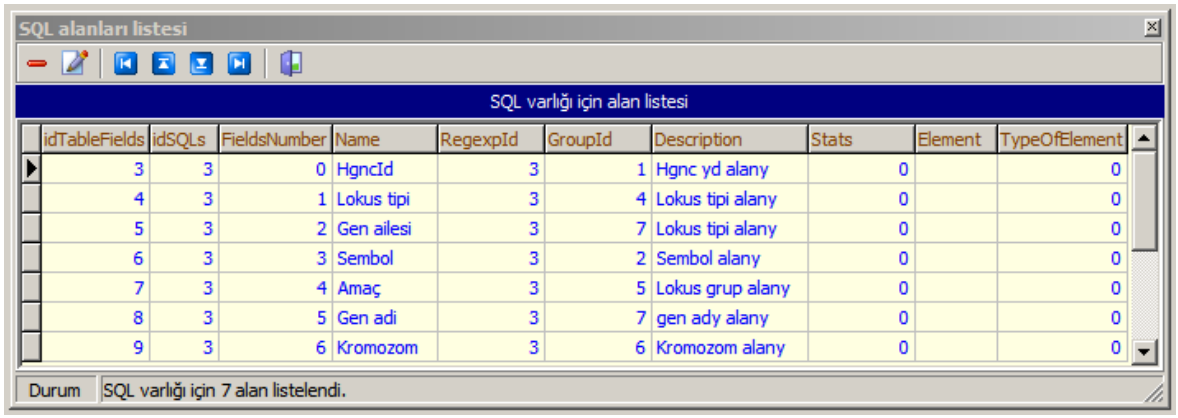
Düzenli ifade varlığı için grup listesi

idGroups	idRegexps	GroupId	Name	Description	StLength	TrimLeft	TrimRight	SpaceIs	Eslestir	secDuIfId	secDuIfGrId
24	3	0	Genel	Genel metin	-1	-1	-1	-1	0	-1	-1
25	3	1	HgncID	HGNC ID grubu	-1	-1	-1	-1	0	-1	-1
26	3	2	Sembol	Genin sembolünü içerir	20	-1	-1		0	-1	-1
27	3	3	Adi	Genin uzun adýný içerir	120	-1	-1	Tanimsiz	0	-1	-1
28	3	4	Lokus tipi	Genin lokus tipini verir	-1	-1	-1	Tanimsiz	0	-1	-1
29	3	5	Lokus Grup (AMAÇ)	Genin lokus grubunu verir	-1	-1	-1	Tanimsiz	0	-1	-1
30	3	6	Kromozom	Kromozomal yerleþim	50	-1	-1	Tanimsiz	0	-1	-1
31	3	7	Gen ailesi	Gen ailesi bilgisini içeren grup	-1	-1	-1	Tanimsiz	0	-1	-1

Durum | Düzenli ifade varlığı için 8 grup listelendi.

Burada toplam 8 düzenli ifade grubu, 7 adet SQL alanı ile ilişkilendirilmiştir. Bu ilişkilere ait SQL varlığı alan parametreleri Tablo 27'de gösterilmiştir.

Tablo 27: SQL alan parametreleri ile düzenli ifadeler grup parametreleri arasında kurulmuş ilişki (SOLUCAN adlı programın ekran çıktısından alınmıştır).



SQL alanları listesi

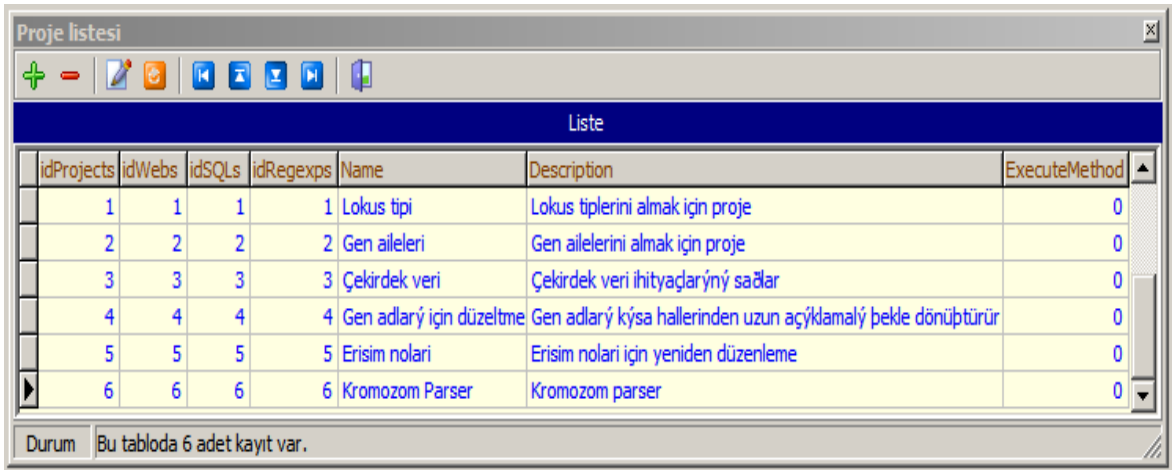
SQL varlığı için alan listesi

idTableFields	idSQLs	FieldsNumber	Name	RegexpId	GroupId	Description	Stats	Element	TypeOfElement
3	3	0	HgncId	3	1	Hgnc yd alany	0		0
4	3	1	Lokus tipi	3	4	Lokus tipi alany	0		0
5	3	2	Gen ailesi	3	7	Lokus tipi alany	0		0
6	3	3	Sembol	3	2	Sembol alany	0		0
7	3	4	Amaç	3	5	Lokus grup alany	0		0
8	3	5	Gen adi	3	7	gen ady alany	0		0
9	3	6	Kromozom	3	6	Kromozom alany	0		0

Durum | SQL varlığı için 7 alan listelendi.

Tablo 28 'de çekirdek veri projeleri gösterilmiştir.

Tablo 28: Çekirdek veri (SOLUCAN adlı programın ekran çıktısından alınmıştır).


Proje listesi

Liste

idProjects	idWebs	idSQLs	idRegexps	Name	Description	ExecuteMethod
1	1	1	1	Lokus tipi	Lokus tiplerini almak için proje	0
2	2	2	2	Gen aileleri	Gen ailelerini almak için proje	0
3	3	3	3	Çekirdek veri	Çekirdek veri ihityaçlarýný sağlar	0
4	4	4	4	Gen adlarý için düzeltme	Gen adlarý kýsa hallerinden uzun açýklamalý þekle dönüþtürür	0
5	5	5	5	Erisim nolari	Erisim nolari için yeniden düzenleme	0
6	6	6	6	Kromozom Parser	Kromozom parser	0

Durum Bu tabloda 6 adet kayıt var.

Çekirdek veri projesi için web varlıkları, SQL varlıkları ve düzenli ifade varlıkları düzenlendikten sonra 6 adet SOLUCAN projesi haline dönüştürülmüştür. idProjects ile proje kimlik numarası, idWebs ile Web varlığı kimlik numarası, idSQLs ile SQL varlığı kimlik numarası, idRFegexps ile düzenli ifade kimlik numarası, Name başlığı altında proje adı, Description başlığı altında projenin ne işe yaradığı ve hakkında kısa açıklama, ExecuteMethod ile parametrik mi, düzenli mi derleme yapılacağı bilgileri verilmiştir. Bu parametreler bir bütün halinde ele alındığında SOLUCAN Projesini oluşturur.

Lokus tipleri veri tablosunda toplam 32 lokus tipi verisine ulaşılmıştır. Bu tipler ve listesi Tablo 29 ‘de gösterilmiştir.

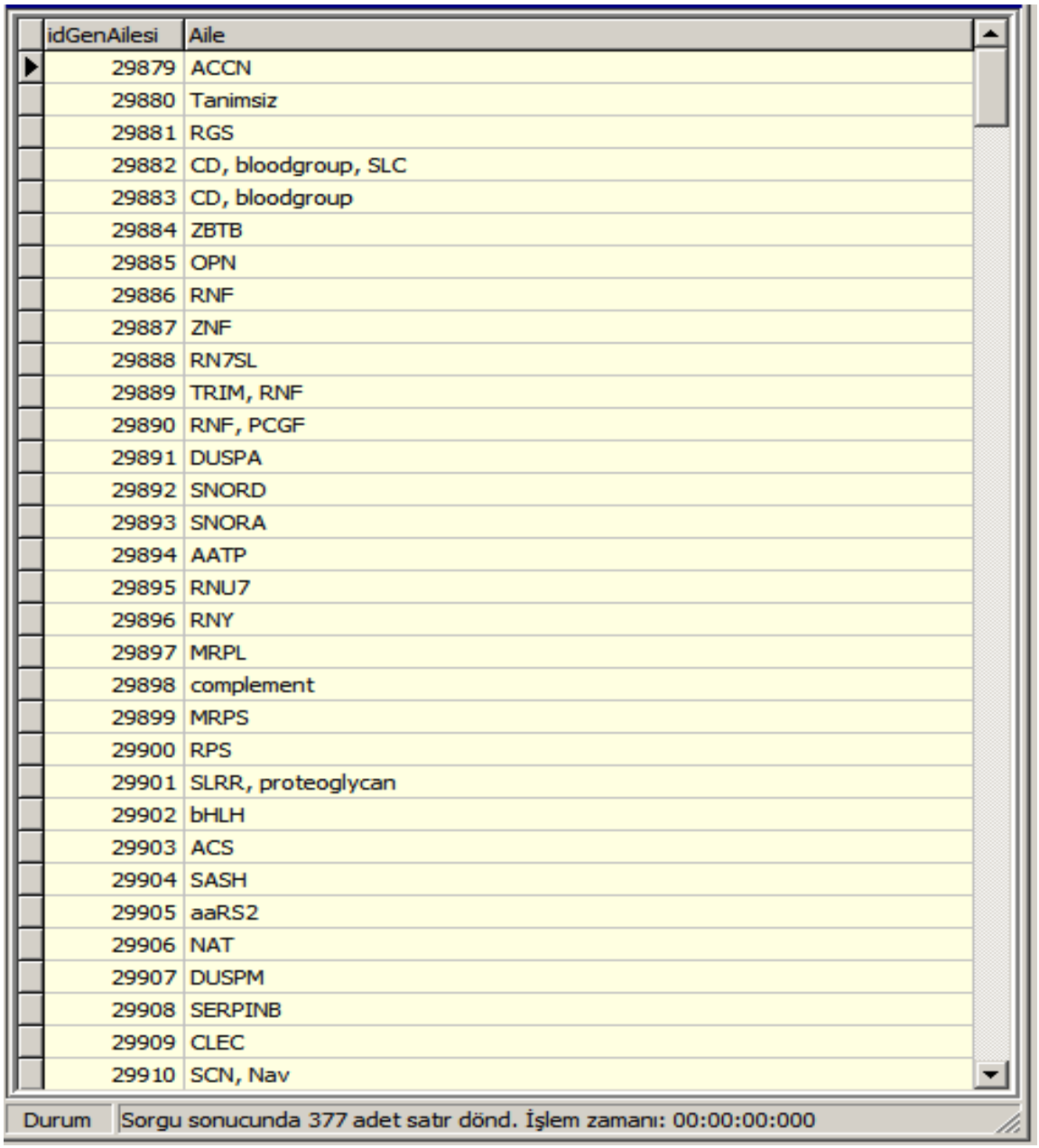
Tablo 29: Lokus tipleri ve lokus tipi kimlik numaraları. (SOLUCAN adlı programın ekran çıktısından alınmıştır, örnek veriyi içerir).

idLokusTipi	Tip
29879	gene with protein product
29880	pseudogene
29881	phenotype only
29882	RNA, RNase
29883	RNA, cluster
29884	RNA, small nuclear
29885	RNA, small cytoplasmic
29886	RNA, pseudogene
29887	RNA, small nucleolar
29888	gene with no protein product
29889	RNA, Y
29890	unknown
29891	endogenous retrovirus
29892	region
29893	RNA, antisense
29894	RNA, telomerase
29895	T cell receptor gene
29896	T cell receptor pseudogene
29897	RNA, transfer
29898	complex locus constituent
29899	RNA, vault
29900	virus integration site
29901	RNA, micro
29902	protocadherin
29903	immunoglobulin pseudogene
29904	transposable element
29905	fragile site
29906	readthrough
29907	RNA, Cajal body-specific
29908	RNA, ribosomal
29909	immunoglobulin gene
29910	Tanimsiz

Durum | Sorgu sonucunda 32 adet satır dönd. İşlem zamanı: 00:00:00:000

Gen aileleri ile ilgili verileri derlemesi sonucunda toplam 377 adet veriye ulaşmıştır. Örnek veriler Tablo 30 ‘da sunulmuştur.

Tablo 30: Gen aileleri (SOLUCAN adlı programın ekran çıktısından alınmıştır, örnek veriyi içerir).



idGenAilesi	Aile
29879	ACCN
29880	Tanimsiz
29881	RGS
29882	CD, bloodgroup, SLC
29883	CD, bloodgroup
29884	ZBTB
29885	OPN
29886	RNF
29887	ZNF
29888	RN7SL
29889	TRIM, RNF
29890	RNF, PCGF
29891	DUSPA
29892	SNORD
29893	SNORA
29894	AATP
29895	RNU7
29896	RNY
29897	MRPL
29898	complement
29899	MRPS
29900	RPS
29901	SLRR, proteoglycan
29902	bHLH
29903	ACS
29904	SASH
29905	aaRS2
29906	NAT
29907	DUSPM
29908	SERPINB
29909	CLEC
29910	SCN, Nav

Durum | Sorgu sonucunda 377 adet satır dönd. İşlem zamanı: 00:00:00:000

Çekirdek veri projesinin SOLUCAN tarafından derlenmesi sonucunda toplam 29891 adet veriye ulaşılmıştır. Örnek veriler Tablo 31 ‘de sunulmuştur.

Tablo 31: Çekirdek veri (SOLUCAN adlı programın ekran çıktısından alınmıştır, örnek veriyi

içerir).

HgncId	idLokusTipi	idGenAilesi	Sembol	GenAdi	Amac	Kromozom	KromozomNo	Kol	Band	AltBand
57	29879	30042	ABCC6	ATP-binding cassette, sub-family C (CFTR/MRP)	protein-coding gene	16p13.11	16	p	13	11
59	29879	30042	ABCC8	ATP-binding cassette, sub-family C (CFTR/MRP)	protein-coding gene	11p15.1	11	p	15	1
60	29879	30042	ABCC9	ATP-binding cassette, sub-family C (CFTR/MRP)	protein-coding gene	12p12.1	12	p	12	1
61	29879	30042	ABCD1	ATP-binding cassette, sub-family D (ALD), memb	protein-coding gene	Xq28	X	q	28	-
62	29880	30042	ABCD1P1	ATP-binding cassette, sub-family D (ALD), memb	pseudogene	2p11	2	p	11	-
63	29880	30042	ABCD1P2	ATP-binding cassette, sub-family D (ALD), memb	pseudogene	10p11.1	10	p	11	1
64	29880	30042	ABCD1P3	ATP-binding cassette, sub-family D (ALD), memb	pseudogene	16p11	16	p	11	-
65	29880	30042	ABCD1P4	ATP-binding cassette, sub-family D (ALD), memb	pseudogene	22q11	22	q	11	-
66	29879	30042	ABCD2	ATP-binding cassette, sub-family D (ALD), memb	protein-coding gene	12q11-q12	12	q	11	-
67	29879	30042	ABCD3	ATP-binding cassette, sub-family D (ALD), memb	protein-coding gene	1p22-p21	1	p	22	-
68	29879	30042	ABCD4	ATP-binding cassette, sub-family D (ALD), memb	protein-coding gene	14q24	14	q	24	-
69	29879	30042	ABCE1	ATP-binding cassette, sub-family E (OABP), men	protein-coding gene	4q31	4	q	31	-
70	29879	30042	ABCF1	ATP-binding cassette, sub-family F (GCN20), me	protein-coding gene	6p21.33	6	p	21	33
71	29879	30042	ABCF2	ATP-binding cassette, sub-family F (GCN20), me	protein-coding gene	7q35-q36	7	q	35	-
72	29879	30042	ABCF3	ATP-binding cassette, sub-family F (GCN20), me	protein-coding gene	3q25.1-q25.2	3	q	25	1
73	29879	30042	ABCG1	ATP-binding cassette, sub-family G (WHITE), me	protein-coding gene	21q22.3	21	q	22	3
74	29879	30203	ABCG2	ATP-binding cassette, sub-family G (WHITE), me	protein-coding gene	4q22-q23	4	q	22	-
76	29879	29880	ABL1	c-abl oncogene 1, non-receptor tyrosine kinase	protein-coding gene	9q34.1	9	q	34	1
77	29879	29880	ABL2	v-abl Abelson murine leukemia viral oncogene ho	protein-coding gene	1q25.2	1	q	25	2
78	29879	29880	ABLIM1	actin binding LIM protein 1	protein-coding gene	10q25	10	q	25	-
79	29879	29926	ABO	ABO blood group (transferase A, alpha 1-3-N-ac	protein-coding gene	9q34.2	9	q	34	2
80	29879	29880	ABP1	amiloride binding protein 1 (amine oxidase (copp	protein-coding gene	7q34-qter	7	q	34	-
81	29879	29880	ABR	active BCR-related gene	protein-coding gene	17p13	17	p	13	-
82	29879	29880	ACAA1	acetyl-CoA acyltransferase 1	protein-coding gene	3p23-p22	3	p	23	-
83	29879	29880	ACAA2	acetyl-CoA acyltransferase 2	protein-coding gene	18q21	18	q	21	-
84	29879	29880	ACACA	acetyl-CoA carboxylase alpha	protein-coding gene	17q21	17	q	21	-
85	29879	29880	ACACB	acetyl-CoA carboxylase beta	protein-coding gene	12q24.1	12	q	24	1

Durum | Sorgu sonucunda 29891 adet satır dönd. İşlem zamanı: 00:00:00:407

Bunun sonucunda genlerin Hgnc numaraları, lokus tipleri, hangi gen ailesinden oldukları, gen sembolleri, gen adları, amaç başlığı altında genin protein kodlayan bir gen mi, bir pseudogen mi, fenotip tanımlayıp tanımlamadığı, RNA kodlamayan bir gen olup olmadığı hakkında bilgi, kromozomal lokalizasyonu ve bu lokalizasyonun kromozom numarası, kromozom kolu, başlangıç bandı ve başlangıç alt bandı verilerine ulaşılmıştır. SOLUCAN ‘ın bu projeyi derlemesi yaklaşık olarak 47 dakika almıştır.

Önem verdiğimiz erişim numaralarına ulaşmak için erişim numaraları projesi SOLUCAN tarafından derlenmiş ve bu derleme sonrasında 29891 adet genin 8 farklı veri

tabanına ait erişim numaraları SOLUCAN veri tabanına katılmıştır. Tablo 32 ‘de elde edilen verilerden örnekler gösterilmiştir.

Tablo 32: Erişim numaralarına ait örnek veriler (SOLUCAN adlı programın ekran çıktısından alınmıştır, örnek veriyi içerir).

HgncId	acGeneId	acUniProtId	acOmimId	acEnsemblId	acRefSeqId	acUCSCId	acGDBId
5	1	P04217	138670	ENSG00000121410	NM_130786	uc002qsd.3	GDB:119638
7	2	P01023	103950	ENSG00000175899	NM_000014	uc001qvk.1	GDB:119639
8	3	Tanimsiz	Tanimsiz	Tanimsiz	NG_001067	Tanimsiz	GDB:128103
15	11	Tanimsiz	Tanimsiz	Tanimsiz	NG_004857	Tanimsiz	GDB:132838
16	12	Q6NSC9	107280	ENSG00000196136	NM_001085	uc001ydp.2	GDB:118955
17	13	P22760	600338	ENSG00000114771	NM_001086	uc003eze.2	GDB:392587
18	14	Q13685	603488	ENSG00000127837	NM_001087	uc002vhk.2	GDB:4573993
19	15	Q16613	600950	ENSG00000129673	NM_001088	uc002jro.2	GDB:700076
20	16	P49588	601065	ENSG00000090861	NM_001605	uc002eyn.1	GDB:595485
21	9625	Q6ZMQ8	605276	ENSG00000181409	NM_001080395	uc010dia.2	GDB:9957818
22	17	Tanimsiz	102699	Tanimsiz	Tanimsiz	Tanimsiz	GDB:125369
23	18	P80404	137150	ENSG00000183044	NM_000663	uc002czc.3	GDB:581658
29	19	Q9UN09	600046	ENSG00000165029	NM_005502	uc004bcl.2	GDB:305294
30	10349	Q8WWZ4	612508	ENSG00000154263	NM_080282	uc010dfa.1	GDB:9956420
31	79963	Q4W5N1	Tanimsiz	ENSG00000182903	NR_002451	uc010ibd.1	GDB:9956416
32	20	Q9BZC7	600047	ENSG00000107331	NM_001606	uc011mel.1	GDB:305295
33	21	Q99758	601615	ENSG00000167972	NM_001089	uc002cpy.1	GDB:3770735
34	24	P78363	601691	ENSG00000198691	NM_000350	uc001dqh.2	GDB:370748
35	23461	Q8WWZ7	612503	ENSG00000154265	NM_018672	uc002jig.2	GDB:9964240
36	23460	Q8N139	612504	ENSG00000154262	NM_080284	uc002jhw.1	GDB:9964242
37	10347	Q8IZY2	605414	ENSG00000064687	NM_019112	uc002lqw.3	GDB:9956413
38	10351	O94911	612505	ENSG00000141338	NM_007168	uc002jhp.2	GDB:9956431
39	10350	Q8IUA7	612507	ENSG00000154258	NM_080283	uc002jhu.2	GDB:9956426
40	5243	P08183	171050	ENSG00000085563	NM_000927	uc003uiz.1	GDB:120712
41	23456	Q9NRK6	605454	ENSG00000135776	NM_012089	uc001htp.3	GDB:9964250
42	8647	O95342	603201	ENSG00000073734	NM_003742	uc002ueo.1	GDB:9864786
43	6890	Q03518	170260	ENSG00000168394	NM_000593	uc003ocg.2	GDB:132668

Durum | Sorgu sonucunda 29891 adet satır dönd. İşlem zamanı: 00:00:00:359

Toplam 29891 adet genin 8 farklı veri tabanına olan erişim numaraları gösterilmiştir. SOLUCAN ‘ın bu projeyi derlemesi yaklaşık olarak 19 dakika almıştır.

SOLUCAN ‘ın tüm projeleri (yaklaşık derleme süresi 101 dakika) derleme zamanları ve derlemesi sonrasında ortaya çıkan veri adetleri ile ilgili bilgiler Tablo 33 ‘da sunulmuştur.

Tablo 33: SOLUCAN projelerinin derleme zamanları

Proje	**Derleme süresi**	**Yakalanan bilgi**
Gen aileleri	17 dakika	377
Lokus tipleri	18 dakika	32
Çekirdek veri	47 dakika	29891
Erişim numaraları	19 dakika	29891

Zaman değerleri yaklaşık olarak ve dakika cinsinden, derleme sonucunda elde edilmiş işleme sonrası ulaşılan veri sayıları adet cinsinden verilmiştir.

Solucan ve diğer benzer görev yapan yazılımlar karşılaştırılmıştır. Sonuçlar Tablo-34 ‘de gösterilmiştir (Wang ve ark., 2005).

Tablo 34: Solucan ve benzeri yazılımların karşılaştırılması.

Yazılım	**Tip**	**Gereklilikler**	**SNP arama**	**Grafik Gösterim**	**SNP seçme**	**Primer dizaynı**
dbSNP	Web sunucu	Sadece sunulan veriler	Var	Var	Yok	Yok
SNPper	Web sunucu	Sadece sunulan veriler	Var	Yok	Yok	Var
SnpHUNTER	Web istemci	Sadece SNP ve ilişki veriler	Var	Var	Var	Yok
SNPicker	Web istemci	Sadece sunulan veriler	Yok	Yok	Yok	Yok
SNPbox	Web istemci	Sadece sunulan veriler	Yok	Yok	Yok	Yok
viewGene	Web istemci	Sadece sunulan veriler	Yok	Var	Yok	Yok
Solucan	Web sunucu istemci	İstenilen veriler	Var	Var	Var	Yok

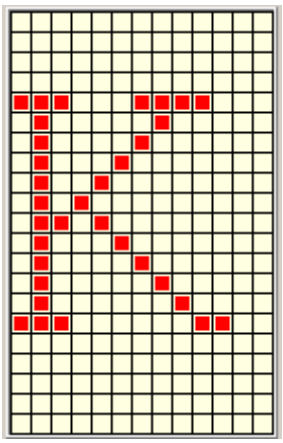
Solucan programı C++ dilinde yazılmıştır. Gerekli olan veri tabanı yazılımı MySQL olup görsel tasarım aracı olarak DbDesigner kullanılmıştır. Geliştirildiği

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/).

4.4. BOT

BOT, mktANN isimli sınıfın doğruluk ve performansını doğası iyi bilinen, iyi çalışılmış bir örnek olan bozuk örüntüleri tanıma konusunda test etmek için yazılmıştır. Kullanılan programlama dili C++ ‘dır. BOT ‘un genel amacı A,B,C,D, E,F,G,H,I,J,K,L,M,N,O,P,R ‘den oluşan bir alfabede harf örüntülerinin %5, %7, %10 deformasyon yapıldıktan sonra onları doğru bir şekilde tanımak amacıyla mktANN ‘nin eğitimidir.

Girişler 13x21 ‘lik matrisler şeklinde verilmiştir. 273 elemandan oluşan bu matris içinde bir harf Şekil 34 ‘ de gösterildiği şekliyle oluşturulmuştur.



Şekil 34: 13x21 matris içinde K harfi. Harfin dolu olan bitleri kırmızı ile gösterilip 1 ile ifade olunmuştur. Harfin boş olan yerleri ise krem (renk kodu clInfoBK) olarak gösterilip 0 ile ifade olunmuştur.

Her patern için bir çıkış değeri yine ikilik sistemde oluşturulmuştur. Harfler ve bu harflere karşılık gelen çıkışlar Tablo 35 gösterilmiştir.

Tablo 35: Harfler ve harflere karşılık gelen ikilik sistemde istenilen çıkış değerleri

Harf	İstenilen çıkış değeri
A	10000000000000000
B	01000000000000000
C	00100000000000000
D	00010000000000000
E	00001000000000000
F	00000100000000000
G	00000010000000000
H	00000001000000000
I	00000000100000000
J	00000000010000000
K	00000000001000000
L	00000000000100000
M	00000000000010000
N	00000000000001000
O	00000000000000100
P	00000000000000010
R	00000000000000001

Deforme edilmiş ‘K’ paternine ilişkin ikilik karşılıklar ve bu harfe ilişkin istenilen çıkış değerleri Tablo 36 ‘de gösterilmiştir.

Tablo 36: Sadece ‘K’ harfine ilişkin bozulmuş örüntü için ikilik karşılık ve bu harfin istenilen değerlerinin ikilik karşılıkları.

Harf	Deformasyon	İkilik karşılık	İstenilen Çıkış
K	%5	00000000000000000100000000 00000000000000000000000001 10101111000000000100000010 00010010000100011000000100 10000000001010000000000001100 00000000011001000001001000 10000000100000100000010100 01000001100000011100000000 00000000000000000000000000 00000000000100000100010000 000	00000000001000000
K	%7	01100000001100000000000000 00000100010000000000000011 10001111000010000001000010 00010000001100000000000100 10000000001010000000001101 00000000010001000000011000 10000000100000100000100000 01000011100000010000000000 10000000000000000000001000000 00000000000001000001001000000 000	00000000001000000
K	%10	00000000000000000010000000 00001010000001100000000111 10001111000100000100000010 00010000011100010000100100 10000000001000000010001101 00000100010001100000001001 0 11000000101000110010010 0100 01000011100000011000100000 0 00000010100000100010100000 0 10000000000000000000000000 000	00000000001000000

Bu şekilde A,B,C,D,E,F,G,H,I,J,K,L,M,N,O,P,R harflerinden oluşan set tanımlanmış, her harften 50 adet %5, 50 adet %7 ve 50 adet %10 deforme edilmiş halleri otomatik olarak elde edilmiş ve sonuçta 2550 paternden oluşan veri tabanı hazırlanmıştır.

Bu paternlerde koşturulmak üzere BOT ‘a 3 adet ayrı ağ topoloji farklı öğrenme oranları ve momentum terimleri eşliğinde oluşturtulmuştur. Ağ modelleri Tablo 37 ‘de gösterilmiştir.

Tablo 37: Katman sayıları, ağ parametreleri ve gizli katmandaki nod sayıları

Ağ Adı	**Katman sayısı**	**Öğrenme oranı**	**Momentum terimi**	**Gizli katman sayısı**	**Nod sayısı**
Deneme1	3	0,2	0	1	34
Deneme2	3	0,2	0	1	64
Deneme3	3	0,2	0	1	100
Deneme4	3	0,3	0	1	34
Deneme5	3	0,3	0	1	64
Deneme6	3	0,3	0	1	100
Deneme7	3	0,4	0	1	34
Deneme8	3	0,4	0	1	64
Deneme9	3	0,4	0	1	100
Deneme10	3	0,2	0,5	1	34
Deneme11	3	0,2	0,5	1	64
Deneme12	3	0,2	0,5	1	100
Deneme13	3	0,3	0,5	1	34
Deneme14	3	0,3	0,5	1	64
Deneme15	3	0,3	0,5	1	100
Deneme16	3	0,4	0,5	1	34
Deneme17	3	0,4	0,5	1	64
Deneme18	3	0,4	0,5	1	100
Deneme19	3	0,2	0	1	140
Deneme20	3	0,3	0	1	140
Deneme21	3	0,4	0	1	140
Deneme22	3	0,3	0,5	1	140
Deneme23	3	0,3	0,5	1	140
Deneme24	3	0,4	0,5	1	140
Deneme25	4	0,2	0	2	17,17
Deneme26	4	0,2	0	2	31,32
Deneme27	4	0,2	0	2	50,50
Deneme28	4	0,2	0	2	70,70

Bütün Deneme ağlarında giriş katmanı nod sayısı 273, çıkış katmanı nod sayısı 17 (ki bu değer sınıf sayısıdır) olarak alınmıştır. Ağırlıklar (-1, +1) aralığında faktör 1000

olmak üzere rastlantısal olarak seçilmiş ve atanmıştır. Bu sayede bağımsız parametreler öğrenme oranı ve gizli katman sayısı ve katmanlardaki nod sayısı ile momentum terimi olarak kalmıştır.

Öğrenme oranı, 0 ile 1 arasında keyfi olarak seçilen bir sayıdır. Her seferinde bu sayı kadar ağırlıkların azaltma ya da artırma yapılacağını garanti eden bir çarpandır. Küçük öğrenme oranı değerleri için eğitim işleminin iterasyon sayısı artarken bu değerin büyümesi halinde iterasyon sayısı değerleri azalmaktadır. Buna karşın büyük öğrenme oranı değerlerinde iterasyon sayısı oldukça azalmakta, öğrenme performansı düşmektedir. Bu şekilde paternleri birbirinden ayırma yetisini kaybeden ağ, çok düşük öğrenme oranlarında ise ezberlemeye meyillidir. İterasyon sayısının aşırı yüksek değerleri de aynı şeye yol açmaktadır. Her problemde deneme yanılma metodu ile hesaplanması gerekir. Genel olarak 0.2 alınmaktadır.

Tablo 38 ‘de öğrenme oranı ile iterasyon arasındaki bağlantı ve korelasyon değerleri verilmiştir.

Tablo 38: Farklı gizli katman nod sayılarına karşılık öğrenme oranlarının her grupta farklı seçilip denenmesi ile elde edilen iterasyon rakamları ile öğrenme oranı arasındaki korelasyon.

Gizli Katman	Öğrenme Oranı	İterasyon	Korelasyon
34	0,2	37251	-0,9821634
	0,3	27316	
	0,4	22329	
100	0,2	26193	-0,9998095
	0,3	20629	
	0,4	15429	
64	0,2	33188	-0,995206
	0,3	27316	
	0,4	19035	
140	0,2	28788	-0,995859
	0,3	20245	
	0,4	8493	

Gizli katman nod sayısı teorik olarak 1 ‘den sonsuza kadar artırılabilir. Fakat problemin doğasına göre, giriş katmanındaki nodların sayısına ve çıkış katmanındaki nodların sayısına göre ayarlanması gerekmektedir. BOT üzerinde yaptığımız denemelerin sonuçları Tablo 39a, Tablo 39b ‘ de gösterilmiştir.

Tablo 39a: Gizli katman nod sayısı ile ortalama iterasyon sayışları arasındaki ilişki

Gizli katman nod sayısı	**Ortalama iterasyon**	**Korelasyon**
34	28965,33333	-0,9723005
64	26513	
100	20750,33333	
140	19175,33333	

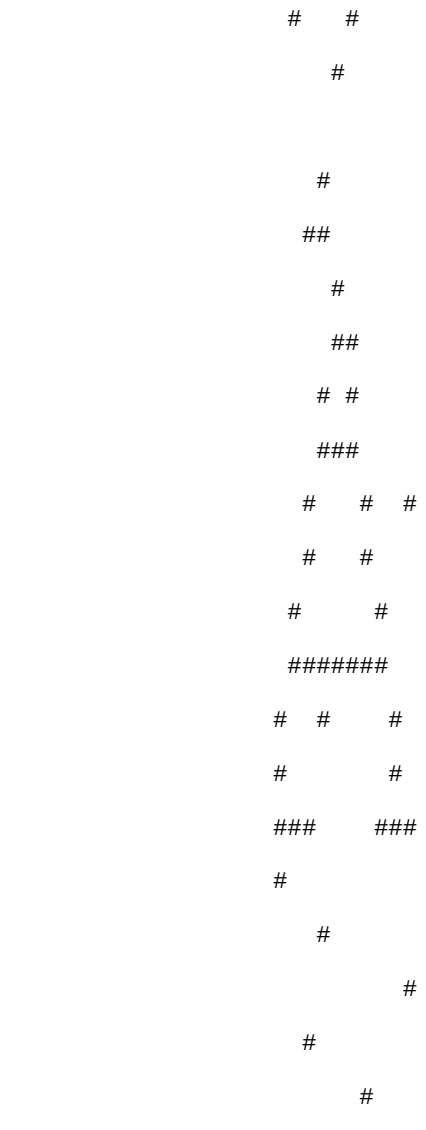
Tablo 39b: Gizli katman nod sayısı ile ortalama ağ performansı arasındaki ilişki.

Gizli katman nod sayısı	**Ortalama başarı oranı**	**Korelasyon**
34	96,67	-0,7083332
64	96,05	
100	94,37	
140	95,35	

Tablo 39a ve Tablo 39b ‘de gösterilen korelasyonlar nedeni ile gizli katman nod sayısı seçilirken problemin doğasına uygun olarak deneme yanılma yönteminin kullanılması oldukça önem arz etmektedir. Zira nod sayısı arttıkça iterasyon sayısı kuvvetli ve anlamlı bir şekilde azalırken, başarı oranı da kuvvetli ve anlamlı bir şekilde azalmaktadır.

BOT, rastlantısal olarak seçilen 50 örüntü üzerinde test edilmiştir. 50 örüntüden 27 tanesinin test sonuçlarının çıktısı aşağıda verilmiştir.

\\----------| AĞA SUNULAN PATERN |



\\----------| AĞIN VERDİĞİ CEVAP |

Tahmin grupları = ABCDEFGHIJKLMNOPR

|||||||||||||||||

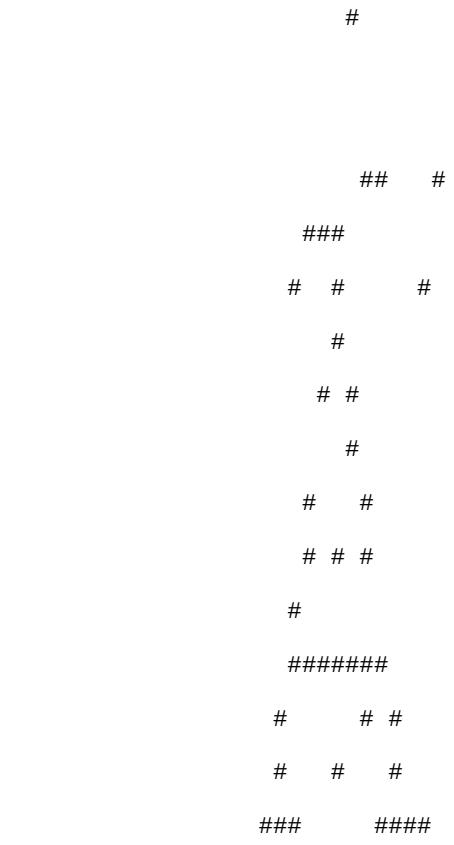
Ağın verdiği cevap = 10000000000000000 Yanlış cevap ihtimali = 1,53249547044125

1. en iyi tahmin = 00000000000100000 Yanlış cevap ihtimali = 32,7781980246774

2. en iyi tahmin = 00000000001000000 Yanlış cevap ihtimali = 33,4237385455886

3. en iyi tahmin = 00000000010000000 Yanlış cevap ihtimali = 33,4611037362244

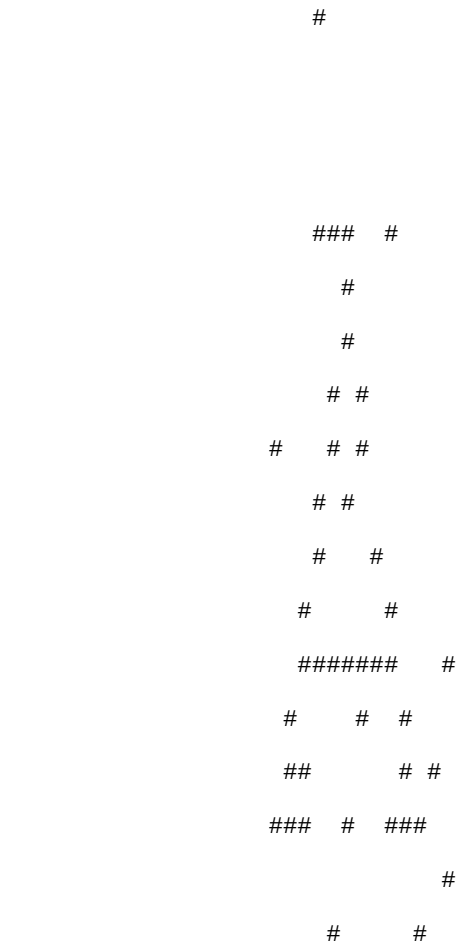
\\----------| AĞA SUNULAN PATERN |



\\----------| AĞIN VERDİĞİ CEVAP |

```
   Tahmin grupları = ABCDEFGHIJKLMNOPR
                     ||||||||||||||||
Ağın verdiği cevap = 1000000000000000 Yanlış cevap ihtimali = 1,4030509453393
  1. en iyi tahmin = 0000000000100000 Yanlış cevap ihtimali = 32,9416838935086
  2. en iyi tahmin = 0000000001000000 Yanlış cevap ihtimali = 33,3839721208989
  3. en iyi tahmin = 0000000010000000 Yanlış cevap ihtimali = 33,4419485101347
```

\\----------| AĞA SUNULAN PATERN |



\\----------| AĞIN VERDİĞİ CEVAP |

Tahmin grupları = ABCDEFGHIJKLMNOPR

|||||||||||||||||

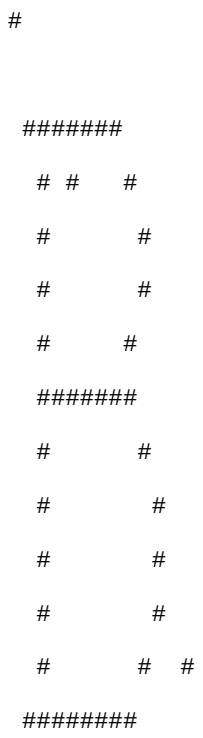
Ağın verdiği cevap = 1000000000000000 Yanlış cevap ihtimali = 1,79687049120175

1. en iyi tahmin = 00000000000100000 Yanlış cevap ihtimali = 32,507312152569

2. en iyi tahmin = 00000000001000000 Yanlış cevap ihtimali = 33,3392759206815

3. en iyi tahmin = 00000000010000000 Yanlış cevap ihtimali = 33,3719104708597

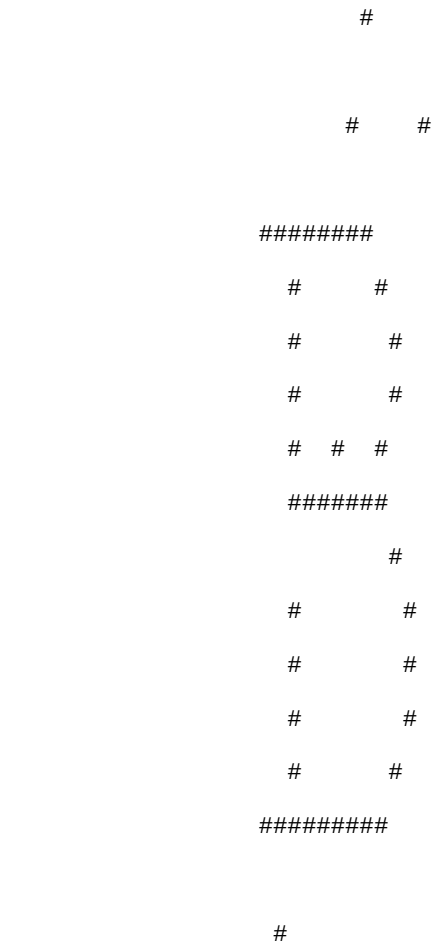
\\----------| AĞA SUNULAN PATERN |



\\----------| AĞIN VERDİĞİ CEVAP |

```
   Tahmin grupları = ABCDEFGHIJKLMNOPR
                     ||||||||||||||||
Ağın verdiği cevap = 01000000000000000 Yanlış cevap ihtimali = 2,02723684249957
  1. en iyi tahmin = 00000000000000010 Yanlış cevap ihtimali = 32,8300356066177
  2. en iyi tahmin = 00000000000100000 Yanlış cevap ihtimali = 33,3638109480676
  3. en iyi tahmin = 00010000000000000 Yanlış cevap ihtimali = 33,5275939640103
```

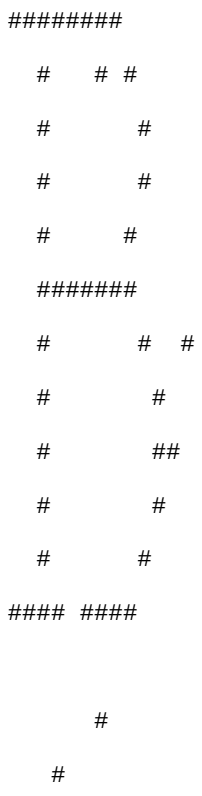
\\----------| AĞA SUNULAN PATERN |



\\----------| AĞIN VERDİĞİ CEVAP |

```
   Tahmin grupları = ABCDEFGHIJKLMNOPR
                     ||||||||||||||||
Ağın verdiği cevap = 01000000000000000 Yanlış cevap ihtimali = 2,06236851918379
  1. en iyi tahmin = 00000000000000010 Yanlış cevap ihtimali = 32,8369251538532
  2. en iyi tahmin = 00000000000100000 Yanlış cevap ihtimali = 33,4077810736376
  3. en iyi tahmin = 00010000000000000 Yanlış cevap ihtimali = 33,548849297859
```

\\----------| AĞA SUNULAN PATERN |



\\----------| AĞIN VERDİĞİ CEVAP |

Tahmin grupları = ABCDEFGHIJKLMNOPR

|||||||||||||||||

Ağın verdiği cevap = 01000000000000000 Yanlış cevap ihtimali = 2,08696824546816

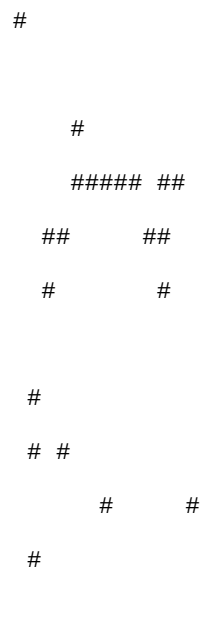
1. en iyi tahmin = 00000000000000010 Yanlış cevap ihtimali = 32,7831117608056

2. en iyi tahmin = 00000000000100000 Yanlış cevap ihtimali = 33,4358802747123

3. en iyi tahmin = 00010000000000000 Yanlış cevap ihtimali = 33,5870244314641

```
\\-----------------------------------------------------| AĞA SUNULAN PATERN |
                          #    #
                          #


                         #   ##### # #
                           ##  #  ###
                           #       ##
                          #
                          #
                          # #
                         ##         #
                          #       #
                          #
                           #       #
                           ##    ##
                             #####


                              #   #
                         # #   #

\\-----------------------------------------------------| AĞIN VERDİĞİ CEVAP |
   Tahmin grupları = ABCDEFGHIJKLMNOPR
                     ||||||||||||||||
Ağın verdiği cevap = 0010000000000000 Yanlış cevap ihtimali = 2,04901926548618
  1. en iyi tahmin = 0000000000001000 Yanlış cevap ihtimali = 32,3824983181918
  2. en iyi tahmin = 0000010000000000 Yanlış cevap ihtimali = 32,8933821417629
  3. en iyi tahmin = 0000100000000000 Yanlış cevap ihtimali = 32,9499200854601
```

\\---| AĞA SUNULAN PATERN |



\\---| AĞIN VERDİĞİ CEVAP |

Tahmin grupları = ABCDEFGHIJKLMNOPR

|||||||||||||||||

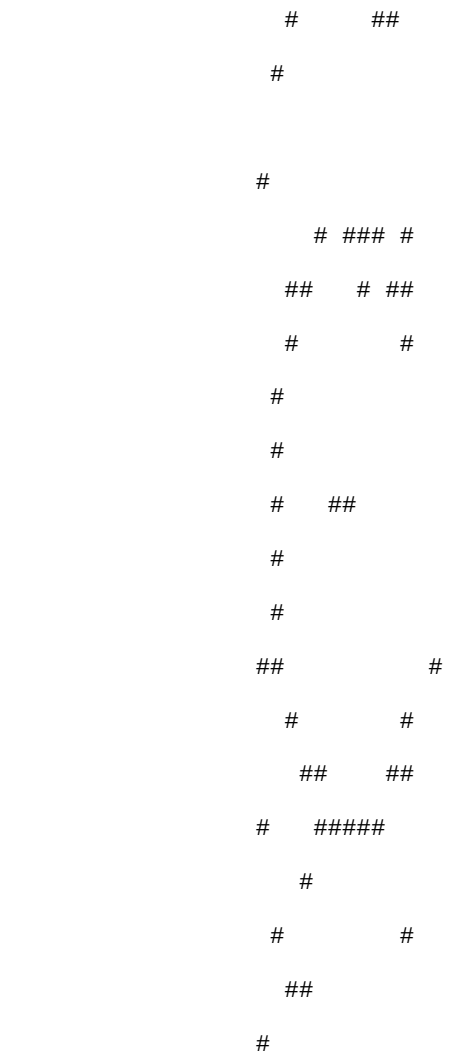
Ağın verdiği cevap = 00100000000000000 Yanlış cevap ihtimali = 1,99839249153058

1. en iyi tahmin = 00000000000001000 Yanlış cevap ihtimali = 32,4402185195182

2. en iyi tahmin = 00000100000000000 Yanlış cevap ihtimali = 32,9377476443813

3. en iyi tahmin = 00000010000000000 Yanlış cevap ihtimali = 32,9624420787254

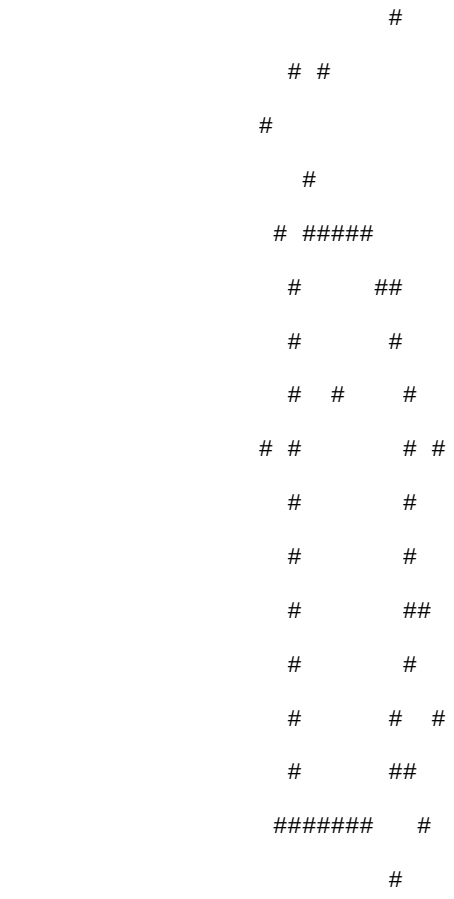
\\----------| AĞA SUNULAN PATERN |



\\----------| AĞIN VERDİĞİ CEVAP |

```
   Tahmin grupları = ABCDEFGHIJKLMNOPR
                     ||||||||||||||||
Ağın verdiği cevap = 00100000000000000 Yanlış cevap ihtimali = 1,91030493045817
  1. en iyi tahmin = 00000000000001000 Yanlış cevap ihtimali = 32,6001879103045
  2. en iyi tahmin = 00000100000000000 Yanlış cevap ihtimali = 32,8455836153465
  3. en iyi tahmin = 00000010000000000 Yanlış cevap ihtimali = 33,2439489306962
```

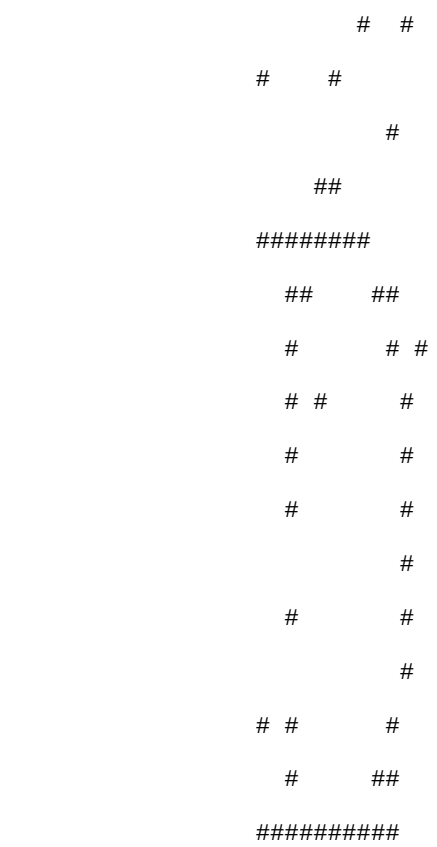
\\----------| AĞA SUNULAN PATERN |



\\----------| AĞIN VERDİĞİ CEVAP |

```
   Tahmin grupları = ABCDEFGHIJKLMNOPR
                     ||||||||||||||||
Ağın verdiği cevap = 00010000000000000 Yanlış cevap ihtimali = 15,1231444074635
  1. en iyi tahmin = 00000000000000100 Yanlış cevap ihtimali = 23,0322062887012
  2. en iyi tahmin = 00000000100000000 Yanlış cevap ihtimali = 34,51957585999
  3. en iyi tahmin = 00000000000100000 Yanlış cevap ihtimali = 35,0457897225252
```

\\----------| AĞA SUNULAN PATERN |



\\----------| AĞIN VERDİĞİ CEVAP |

Tahmin grupları = ABCDEFGHIJKLMNOPR

|||||||||||||||||

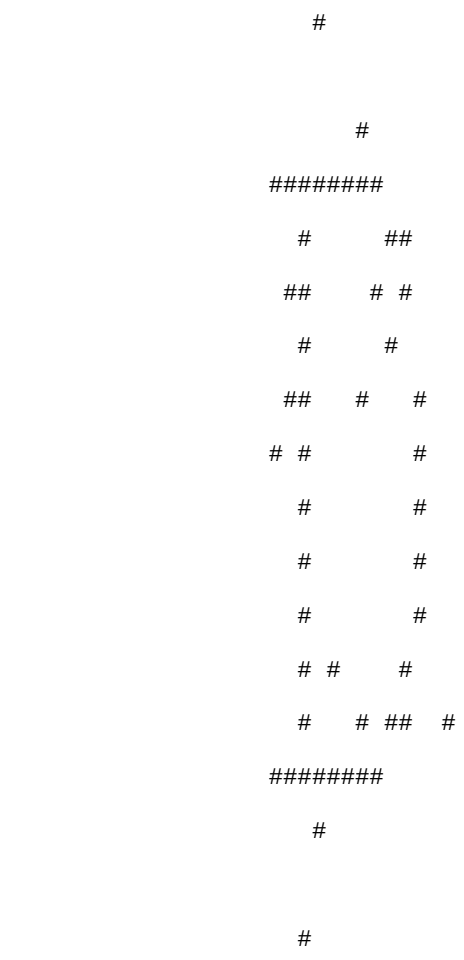
Ağın verdiği cevap = 00010000000000000 Yanlış cevap ihtimali = 2,35172804819125

1. en iyi tahmin = 00000000000100000 Yanlış cevap ihtimali = 32,1916113082968

2. en iyi tahmin = 00000000000000100 Yanlış cevap ihtimali = 32,4531055139915

3. en iyi tahmin = 10000000000000000 Yanlış cevap ihtimali = 32,9064166397417

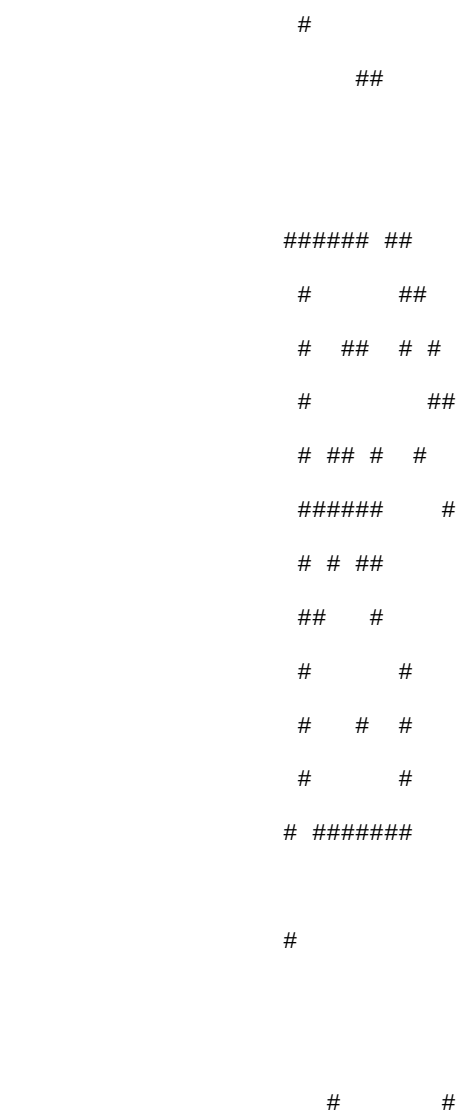
\\----------| AĞA SUNULAN PATERN |



\\----------| AĞIN VERDİĞİ CEVAP |

```
   Tahmin grupları = ABCDEFGHIJKLMNOPR
                     ||||||||||||||||
Ağın verdiği cevap = 00010000000000000 Yanlış cevap ihtimali = 2,42941917738434
  1. en iyi tahmin = 00000000000100000 Yanlış cevap ihtimali = 32,1464941130057
  2. en iyi tahmin = 00000000000000100 Yanlış cevap ihtimali = 32,3499291214692
  3. en iyi tahmin = 10000000000000000 Yanlış cevap ihtimali = 32,8374518495394
```

\\----------| AĞA SUNULAN PATERN |



\\----------| AĞIN VERDİĞİ CEVAP |

Tahmin grupları = ABCDEFGHIJKLMNOPR

|||||||||||||||||

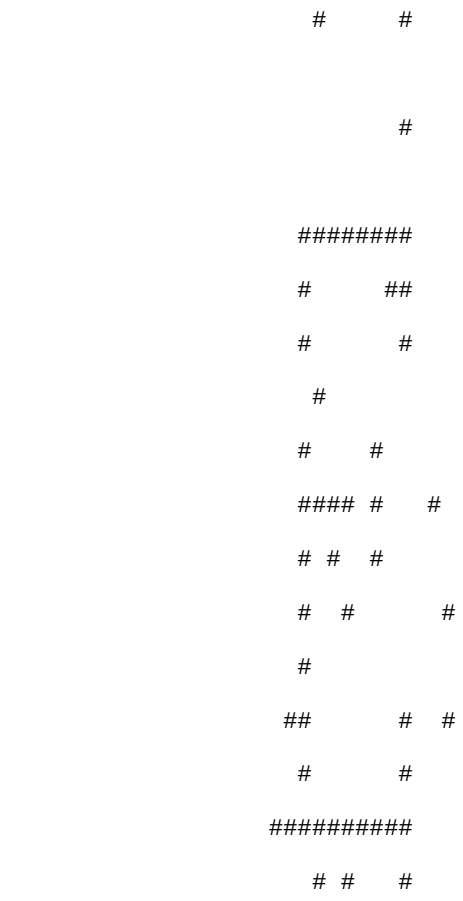
Ağın verdiği cevap = 00001000000000000 Yanlış cevap ihtimali = 3,45931627402516

1. en iyi tahmin = 00000000010000000 Yanlış cevap ihtimali = 31,3600574333715

2. en iyi tahmin = 00000000000100000 Yanlış cevap ihtimali = 31,5002824057353

3. en iyi tahmin = 00000000000000010 Yanlış cevap ihtimali = 31,8814318293001

\\----------| AĞA SUNULAN PATERN |

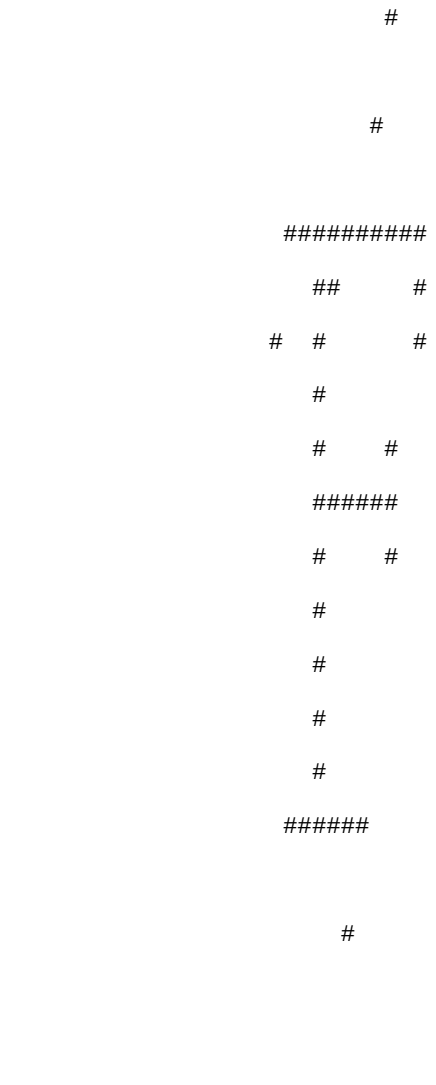


\\----------| AĞIN VERDİĞİ CEVAP |

```
   Tahmin grupları = ABCDEFGHIJKLMNOPR
                     ||||||||||||||||
Ağın verdiği cevap = 00001000000000000 Yanlış cevap ihtimali = 5,46422058947701
  1. en iyi tahmin = 00000000000100000 Yanlış cevap ihtimali = 29,6873706416896
  2. en iyi tahmin = 00000000010000000 Yanlış cevap ihtimali = 29,7215100949921
  3. en iyi tahmin = 00000000000000010 Yanlış cevap ihtimali = 31,2745787964732
```

```
\\-----------------------------------------------------| AĞA SUNULAN PATERN |
                        # # #


                                  #

                             # #

                        ## #######

                          #    # # #

                          #      #

                          #

                          # #  # #

                          ######   #

                         ##  # #

                          #       #

                          ##        #

                          #      #

                          ##     # #

                         ##########

                               #

                        #


                          #

                              #
\\-----------------------------------------------------| AĞIN VERDİĞİ CEVAP |
   Tahmin grupları = ABCDEFGHIJKLMNOPR
                     ||||||||||||||||
Ağın verdiği cevap = 00001000000000000 Yanlış cevap ihtimali = 3,52024323540261
  1. en iyi tahmin = 00000000000000010 Yanlış cevap ihtimali = 31,1645936109654
  2. en iyi tahmin = 00000000010000000 Yanlış cevap ihtimali = 31,7748098614008
  3. en iyi tahmin = 00000000000100000 Yanlış cevap ihtimali = 31,7912474413173
```

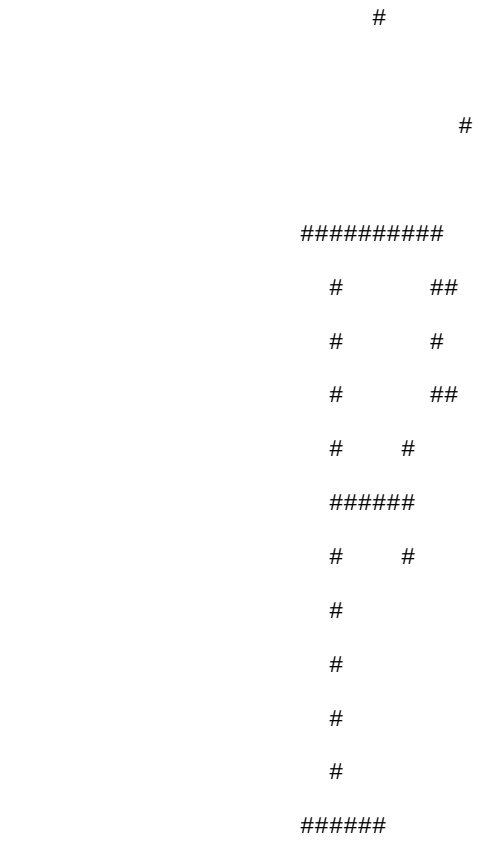
\\----------| AĞA SUNULAN PATERN |



\\----------| AĞIN VERDİĞİ CEVAP |

```
   Tahmin grupları = ABCDEFGHIJKLMNOPR
                     ||||||||||||||||
Ağın verdiği cevap = 00000100000000000 Yanlış cevap ihtimali = 2,55880506966119
  1. en iyi tahmin = 00100000000000000 Yanlış cevap ihtimali = 32,2096258559773
  2. en iyi tahmin = 00000000000000010 Yanlış cevap ihtimali = 32,2144996471537
  3. en iyi tahmin = 00000000001000000 Yanlış cevap ihtimali = 32,785779934056
```

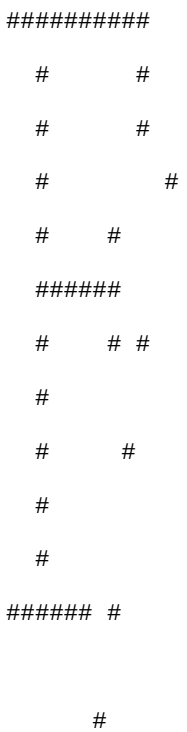
\\----------| AĞA SUNULAN PATERN |



\\----------| AĞIN VERDİĞİ CEVAP |

```
   Tahmin grupları = ABCDEFGHIJKLMNOPR
                     ||||||||||||||||
Ağın verdiği cevap = 00000100000000000 Yanlış cevap ihtimali = 2,59425762721333
  1. en iyi tahmin = 00000000000000010 Yanlış cevap ihtimali = 32,1712456567092
  2. en iyi tahmin = 00100000000000000 Yanlış cevap ihtimali = 32,2012364975635
  3. en iyi tahmin = 00000000001000000 Yanlış cevap ihtimali = 32,7815168611122
```

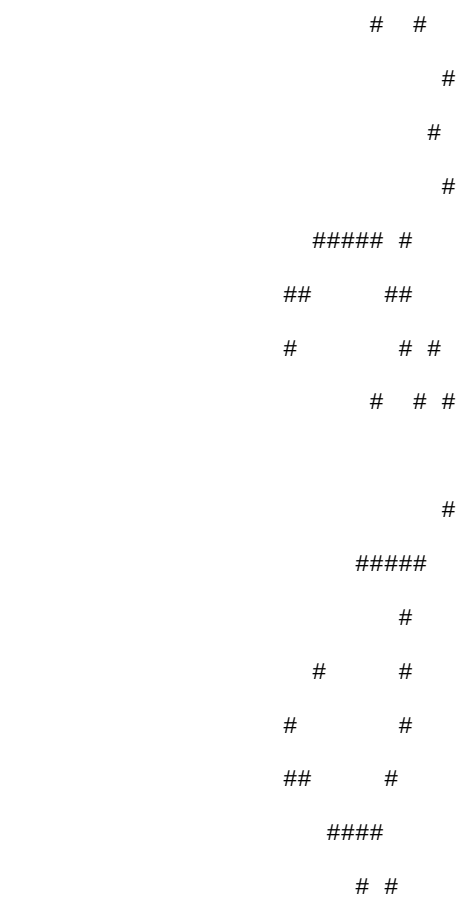
\\----------| AĞA SUNULAN PATERN |



\\----------| AĞIN VERDİĞİ CEVAP |

```
   Tahmin grupları = ABCDEFGHIJKLMNOPR
                     ||||||||||||||||
Ağın verdiği cevap = 00000100000000000 Yanlış cevap ihtimali = 2,63911545539834
  1. en iyi tahmin = 00100000000000000 Yanlış cevap ihtimali = 32,0718764167069
  2. en iyi tahmin = 00000000000000010 Yanlış cevap ihtimali = 32,1535357291772
  3. en iyi tahmin = 00000000001000000 Yanlış cevap ihtimali = 32,6131932440927
```

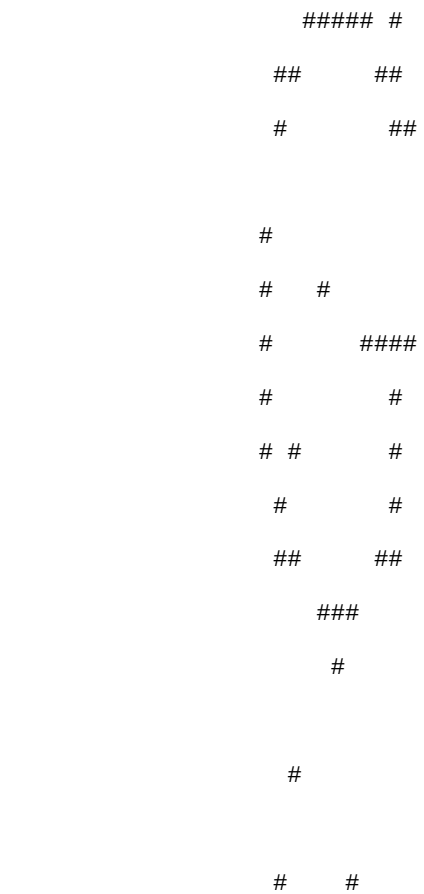
\\----------| AĞA SUNULAN PATERN |



\\----------| AĞIN VERDİĞİ CEVAP |

```
   Tahmin grupları = ABCDEFGHIJKLMNOPR
                     ||||||||||||||||
Ağın verdiği cevap = 00000010000000000 Yanlış cevap ihtimali = 1,9577650954212
  1. en iyi tahmin = 00000000010000000 Yanlış cevap ihtimali = 32,6410301728865
  2. en iyi tahmin = 00100000000000000 Yanlış cevap ihtimali = 32,9386545585969
  3. en iyi tahmin = 00000000000000100 Yanlış cevap ihtimali = 33,2939343127095
```

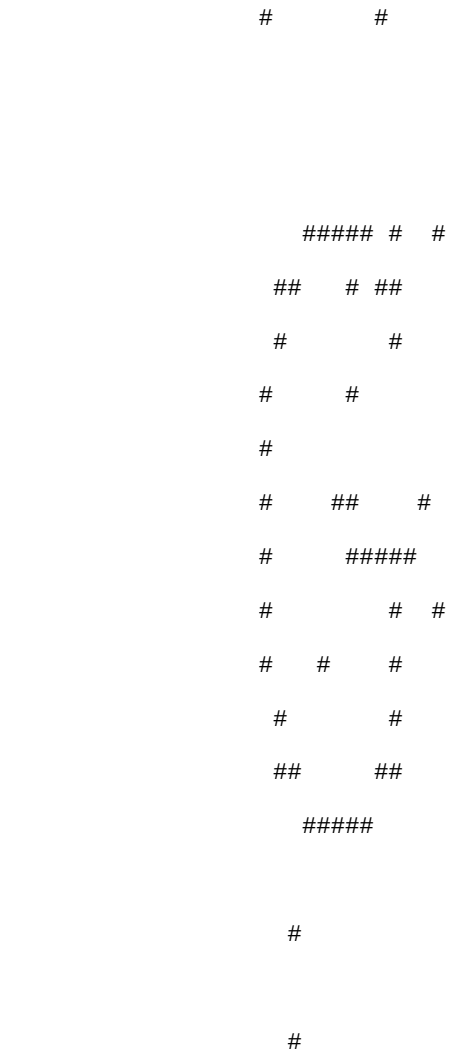
\\----------| AĞA SUNULAN PATERN |



\\----------| AĞIN VERDİĞİ CEVAP |

```
   Tahmin grupları = ABCDEFGHIJKLMNOPR
                     ||||||||||||||||
Ağın verdiği cevap = 00000010000000000 Yanlış cevap ihtimali = 1,55163464974366
  1. en iyi tahmin = 00000000010000000 Yanlış cevap ihtimali = 33,1090119758136
  2. en iyi tahmin = 00100000000000000 Yanlış cevap ihtimali = 33,1655597206705
  3. en iyi tahmin = 00000000000000100 Yanlış cevap ihtimali = 33,1658808931927
```

\\----------| AĞA SUNULAN PATERN |



\\----------| AĞIN VERDİĞİ CEVAP |

Tahmin grupları = ABCDEFGHIJKLMNOPR

||||||||||||||||

Ağın verdiği cevap = 00000010000000000 Yanlış cevap ihtimali = 1,585055652364

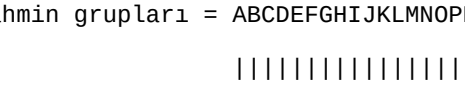
1. en iyi tahmin = 00000000000000100 Yanlış cevap ihtimali = 32,8515990540154

2. en iyi tahmin = 00100000000000000 Yanlış cevap ihtimali = 33,3671389589345

3. en iyi tahmin = 00000000010000000 Yanlış cevap ihtimali = 33,454351318487

\\----------| AĞA SUNULAN PATERN |

\\----------| AĞIN VERDİĞİ CEVAP |



Tahmin grupları = ABCDEFGHIJKLMNOPR

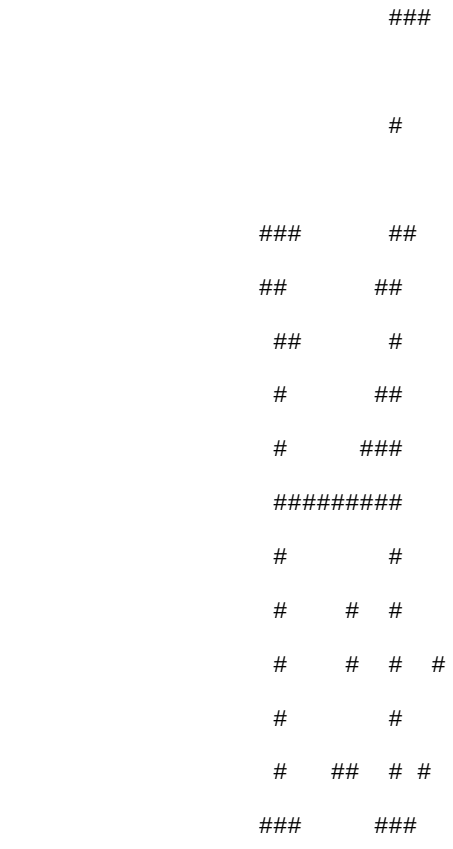
Ağın verdiği cevap = 00000001000000000 Yanlış cevap ihtimali = 11,5221557961279

1. en iyi tahmin = 00000000000010000 Yanlış cevap ihtimali = 25,5576530899151

2. en iyi tahmin = 00000000000000001 Yanlış cevap ihtimali = 27,4222779604904

3. en iyi tahmin = 00000000001000000 Yanlış cevap ihtimali = 27,5063711495366

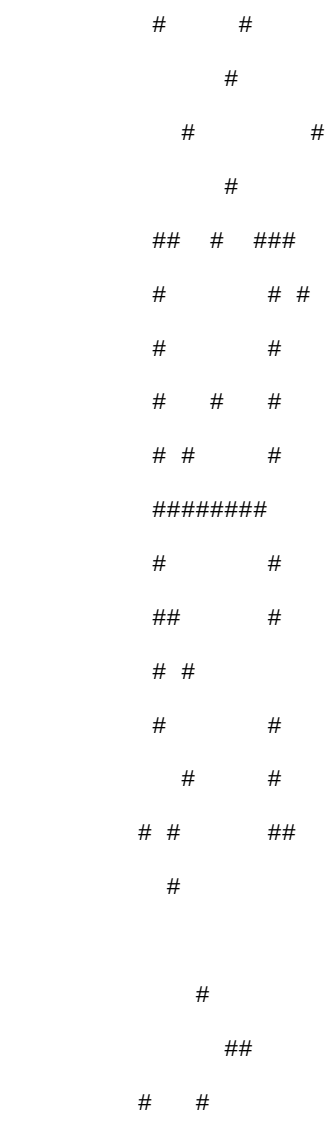
\\---| AĞA SUNULAN PATERN |



\\---| AĞIN VERDİĞİ CEVAP |

```
   Tahmin grupları = ABCDEFGHIJKLMNOPR
                     ||||||||||||||||
Ağın verdiği cevap = 0000000100000000 Yanlış cevap ihtimali = 3,77739903053096
  1. en iyi tahmin = 0000000000010000 Yanlış cevap ihtimali = 30,7354090168876
  2. en iyi tahmin = 0000000000000001 Yanlış cevap ihtimali = 31,6074995351002
  3. en iyi tahmin = 0000000000000010 Yanlış cevap ihtimali = 32,1891211840455
```

\\---| AĞA SUNULAN PATERN |



\\---| AĞIN VERDİĞİ CEVAP |

Tahmin grupları = ABCDEFGHIJKLMNOPR

||||||||||||||||

Ağın verdiği cevap = 00000001000000000 Yanlış cevap ihtimali = 4,03148930230376

1. en iyi tahmin = 00000000000010000 Yanlış cevap ihtimali = 30,5612968086582

2. en iyi tahmin = 00000000000000001 Yanlış cevap ihtimali = 31,3669917582437

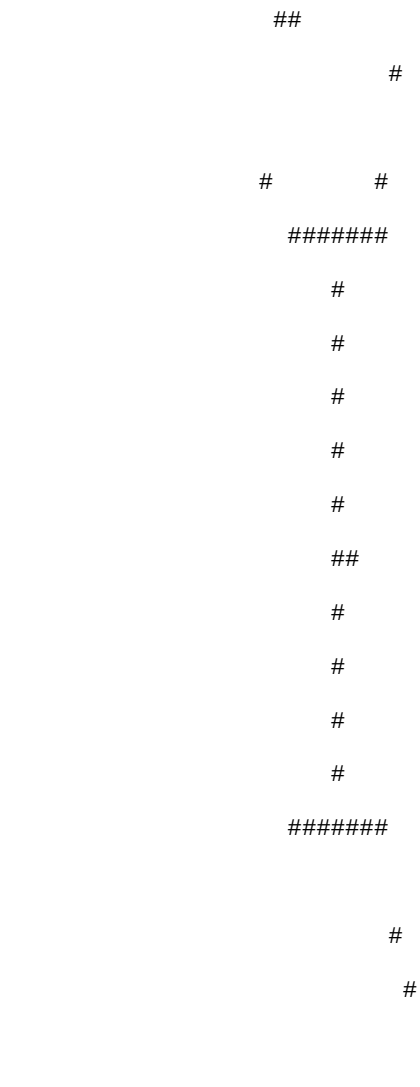
3. en iyi tahmin = 00000000001000000 Yanlış cevap ihtimali = 31,9432373064934

\\----------| AĞA SUNULAN PATERN |

\\----------| AĞIN VERDİĞİ CEVAP |

```
   Tahmin grupları = ABCDEFGHIJKLMNOPR
                     ||||||||||||||||
Ağın verdiği cevap = 0000000010000000 Yanlış cevap ihtimali = 4,16500261644246
  1. en iyi tahmin = 0000000000000100 Yanlış cevap ihtimali = 30,4131426860569
  2. en iyi tahmin = 0000010000000000 Yanlış cevap ihtimali = 31,6959819452498
  3. en iyi tahmin = 0000000010000000 Yanlış cevap ihtimali = 31,7109545768434
```

\\----------| AĞA SUNULAN PATERN |



\\----------| AĞIN VERDİĞİ CEVAP |

Tahmin grupları = ABCDEFGHIJKLMNOPR

|||||||||||||||||

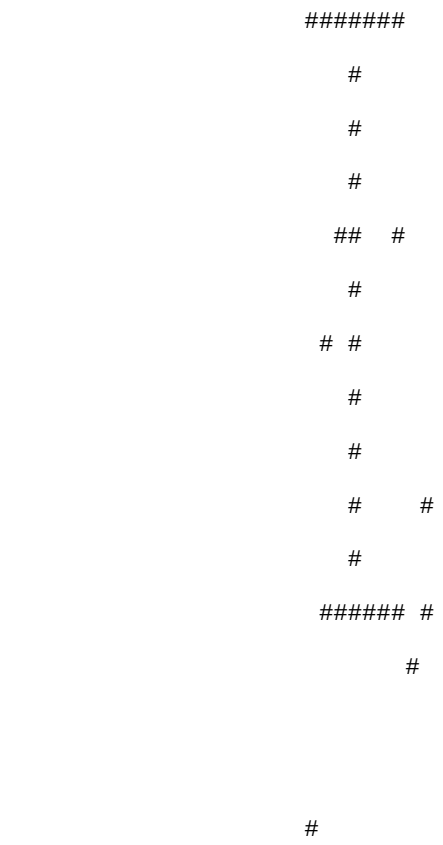
Ağın verdiği cevap = 00000000100000000 Yanlış cevap ihtimali = 2,85302881681579

1. en iyi tahmin = 00000000000000100 Yanlış cevap ihtimali = 31,8347768335112

2. en iyi tahmin = 00100000000000000 Yanlış cevap ihtimali = 32,2154952926805

3. en iyi tahmin = 00000000010000000 Yanlış cevap ihtimali = 32,4122104988364

\\----------| AĞA SUNULAN PATERN |



\\----------| AĞIN VERDİĞİ CEVAP |

Tahmin grupları = ABCDEFGHIJKLMNOPR

|||||||||||||||||

Ağın verdiği cevap = 00000000100000000 Yanlış cevap ihtimali = 2,35845899949856

1. en iyi tahmin = 00000000000000100 Yanlış cevap ihtimali = 32,4633473585344

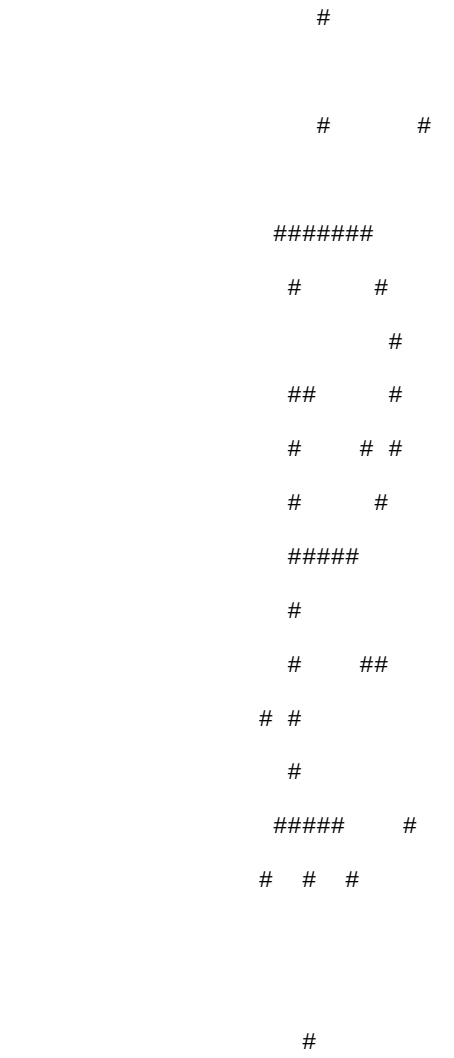
2. en iyi tahmin = 00100000000000000 Yanlış cevap ihtimali = 32,5472548143304

3. en iyi tahmin = 00000000010000000 Yanlış cevap ihtimali = 32,9606906423535

```
\\-------------------------------------------------------| TOPLU TEST SONUCU |

          Toplam test maddesi = 51
     Ağın verdiği doğru cevap = 50
    Ağın verdiği yanlış cevap = 1
         Yanlış / Doğru oranı = 1/50
Düzeltilmiş doğru cevap puanı = 51
            Ağın başarı oranı = 99.8  (Yaklaşık)
     Düzeltilmiş başarı oranı = 95,5345161793218
```

Benzer şekilde BOT benzer örüntülerin tanınması konusunda da denenmiştir. 300 adet daha önce ne eğitim setinde nede test setinde olmayan P ve R harfi tanımlanmıştır. Bu örüntüler %5, %7, %10 oranlarında bozunuma uğratılmıştır. BOT bu örüntülerin tümü üzerinden test edilmiştir. 300 adet test örüntüsünün 20 tanesinin sonuçları ile, testin genel sonuçları aşağıda gösterilmiştir.

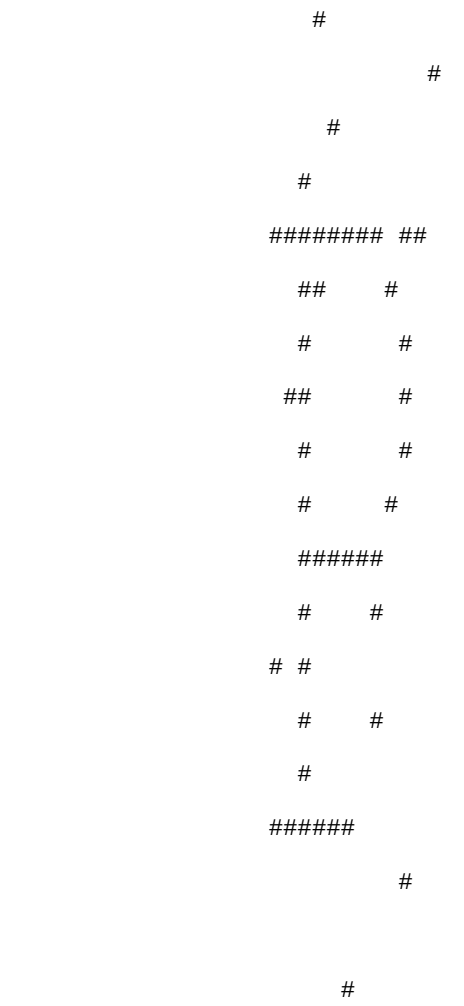
\\----------| AĞA SUNULAN PATERN |



\\----------| AĞIN VERDİĞİ CEVAP |

```
   Tahmin grupları = ABCDEFGHIJKLMNOPR
                     ||||||||||||||||
Ağın verdiği cevap = 0000000000000010 Yanlış cevap ihtimali = 2,74325923877133
  1. en iyi tahmin = 0000000000000001 Yanlış cevap ihtimali = 31,6454896976705
  2. en iyi tahmin = 0000000100000000 Yanlış cevap ihtimali = 32,5706392333045
  3. en iyi tahmin = 0100000000000000 Yanlış cevap ihtimali = 32,7366425183224
```

\\----------| AĞA SUNULAN PATERN |

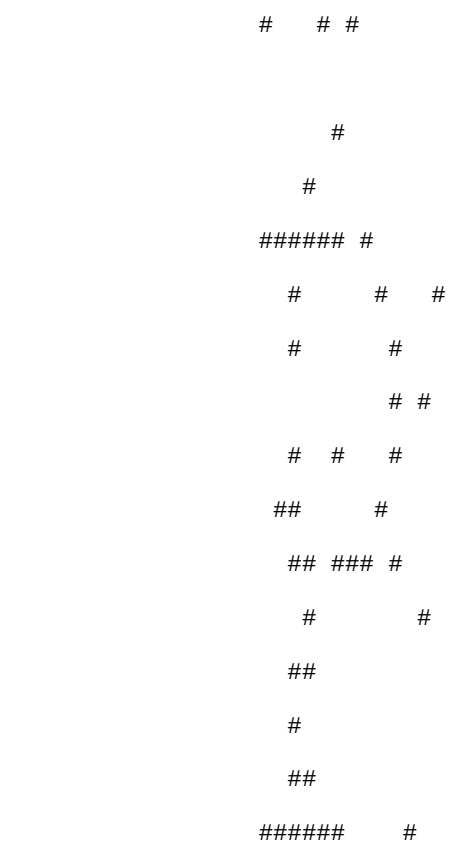


\\----------| AĞIN VERDİĞİ CEVAP |

```
   Tahmin grupları = ABCDEFGHIJKLMNOPR
                     ||||||||||||||||
Ağın verdiği cevap = 0000000000000010 Yanlış cevap ihtimali = 2,21884201100991
  1. en iyi tahmin = 0000000000000001 Yanlış cevap ihtimali = 32,1887064426149
  2. en iyi tahmin = 0000000100000000 Yanlış cevap ihtimali = 32,9104818594393
  3. en iyi tahmin = 0001000000000000 Yanlış cevap ihtimali = 32,9275001523719
```

```
\\-----------------------------------------------------| AĞA SUNULAN PATERN |
                        ##      #




                        ##### ##
                          #   #   # #
                          # #    #
                          #      #
                          ##     #
                          ##    # ##
                          ######  #
                          #
                          #      #
                          #
                          #
                        # ####
                                 #




                        #     ##
\\-----------------------------------------------------| AĞIN VERDİĞİ CEVAP |
   Tahmin grupları = ABCDEFGHIJKLMNOPR
                     ||||||||||||||||
Ağın verdiği cevap = 00000000000000010 Yanlış cevap ihtimali = 2,38673458670145
  1. en iyi tahmin = 00000000000000001 Yanlış cevap ihtimali = 32,220181059221
  2. en iyi tahmin = 01000000000000000 Yanlış cevap ihtimali = 32,5761420869512
  3. en iyi tahmin = 00010000000000000 Yanlış cevap ihtimali = 32,5799139604047
```

\\----------| AĞA SUNULAN PATERN |



\\----------| AĞIN VERDİĞİ CEVAP |

Tahmin grupları = ABCDEFGHIJKLMNOPR

||||||||||||||||

Ağın verdiği cevap = 00000000000000010 Yanlış cevap ihtimali = 3,8996259357425

1. en iyi tahmin = 00000000000000001 Yanlış cevap ihtimali = 30,6782039179351

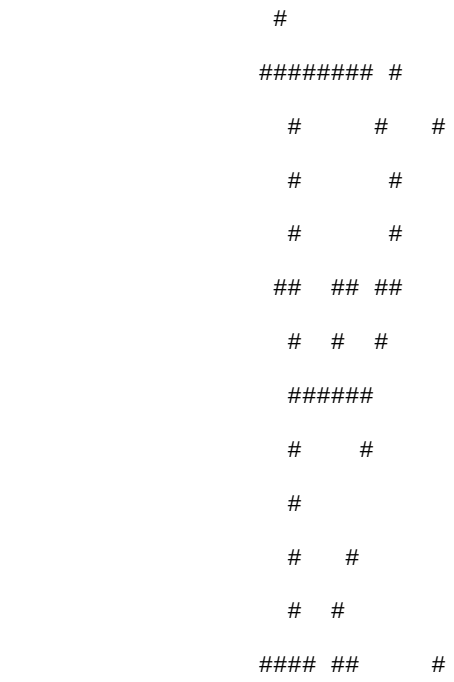
2. en iyi tahmin = 01000000000000000 Yanlış cevap ihtimali = 31,7564451446584

3. en iyi tahmin = 00000000000100000 Yanlış cevap ihtimali = 31,8106347417516

```
\\-----------------------------------------------------| AĞA SUNULAN PATERN |
                                  #
                                ##
                                  #   #


                                ######
                                #   ###
                                #      #
                                #      #
                                #      #
                                #     #
                                ######
                                #
                                #
                                #       #
                                #
                               ######     #



                                  #
                                    #
                                  #
\\-----------------------------------------------------| AĞIN VERDİĞİ CEVAP |
   Tahmin grupları = ABCDEFGHIJKLMNOPR
                     ||||||||||||||||
Ağın verdiği cevap = 0000000000000010 Yanlış cevap ihtimali = 1,85915583469264
  1. en iyi tahmin = 0000000000000001 Yanlış cevap ihtimali = 32,6969521483342
  2. en iyi tahmin = 0000000100000000 Yanlış cevap ihtimali = 32,8534801984948
  3. en iyi tahmin = 0100000000000000 Yanlış cevap ihtimali = 33,0706907501391
```

```
\\-----------------------------------------------------| AĞA SUNULAN PATERN |
                                #
                            ### #####
                             ##     #
                              #      #
                                      #
                              #     ##
                              #     #
                              ######
                              # #     #
                              #

                              #
                            #######     #
                              # #      #
                             ##
                                    #

                               #
\\-----------------------------------------------------| AĞIN VERDİĞİ CEVAP |
   Tahmin grupları = ABCDEFGHIJKLMNOPR
                     ||||||||||||||||
Ağın verdiği cevap = 00000000000000010 Yanlış cevap ihtimali = 2,91119655405042
  1. en iyi tahmin = 00000000100000000 Yanlış cevap ihtimali = 32,0766896031257
  2. en iyi tahmin = 01000000000000000 Yanlış cevap ihtimali = 32,0782481588346
  3. en iyi tahmin = 00000000000000001 Yanlış cevap ihtimali = 32,0787894130205
```

\\----------| AĞA SUNULAN PATERN |



\\----------| AĞIN VERDİĞİ CEVAP |

Tahmin grupları = ABCDEFGHIJKLMNOPR

|||||||||||||||||

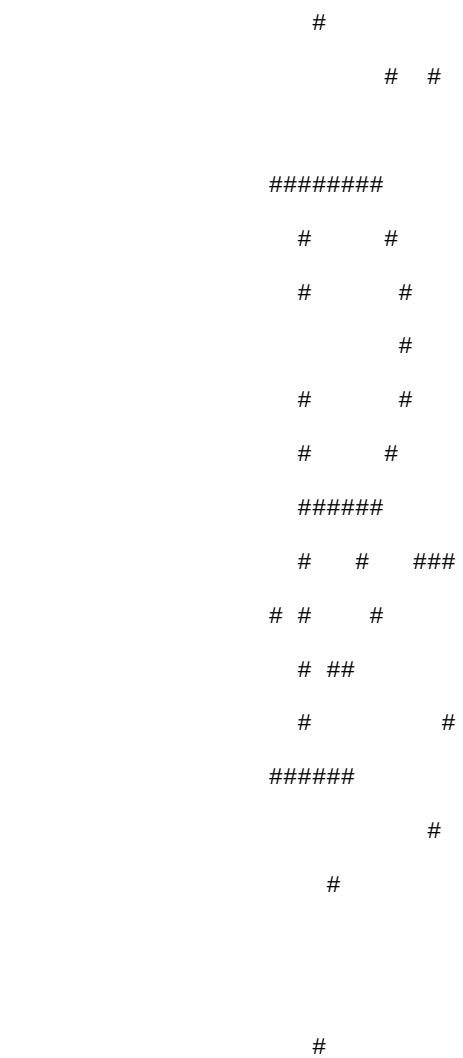
Ağın verdiği cevap = 00000000000000010 Yanlış cevap ihtimali = 2,33553631587935

1. en iyi tahmin = 00000000100000000 Yanlış cevap ihtimali = 32,423762553199

2. en iyi tahmin = 00000000000000001 Yanlış cevap ihtimali = 32,4281161046941

3. en iyi tahmin = 01000000000000000 Yanlış cevap ihtimali = 32,5664813947257

\\----------| AĞA SUNULAN PATERN |



\\----------| AĞIN VERDİĞİ CEVAP |

Tahmin grupları = ABCDEFGHIJKLMNOPR

|||||||||||||||||

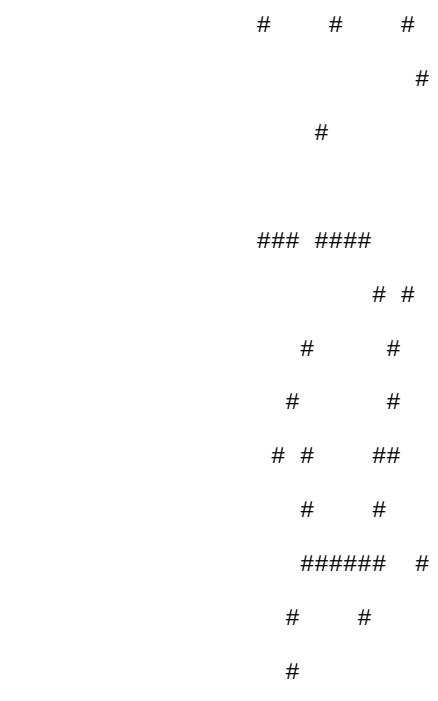
Ağın verdiği cevap = 0000000000000010 Yanlış cevap ihtimali = 3,75287468588492

1. en iyi tahmin = 0000000000000001 Yanlış cevap ihtimali = 30,6757780050778

2. en iyi tahmin = 0001000000000000 Yanlış cevap ihtimali = 32,4070097777878

3. en iyi tahmin = 0000000100000000 Yanlış cevap ihtimali = 32,6274937325752

\\---| AĞA SUNULAN PATERN |



\\---| AĞIN VERDİĞİ CEVAP |

Tahmin grupları = ABCDEFGHIJKLMNOPR

|||||||||||||||||

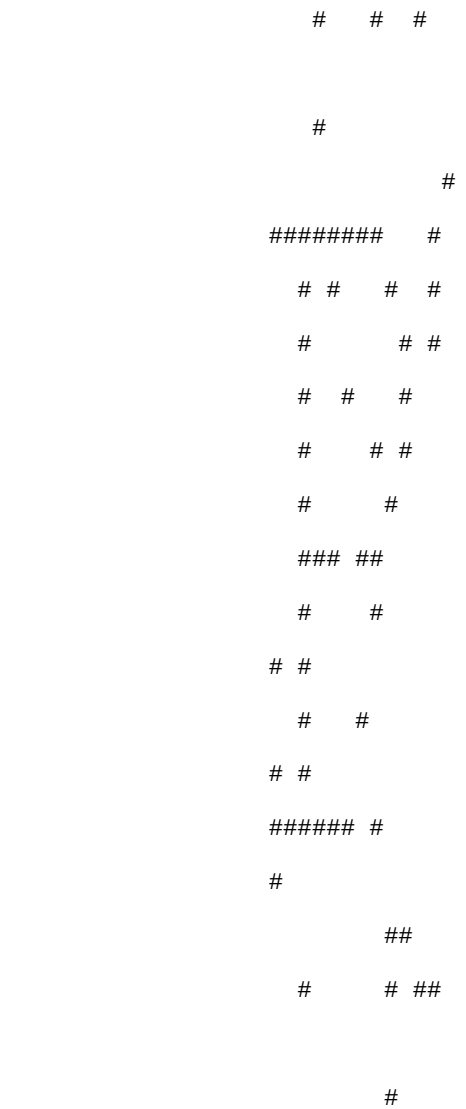
Ağın verdiği cevap = 00000000000000010 Yanlış cevap ihtimali = 3,41686850207277

1. en iyi tahmin = 00000100000000000 Yanlış cevap ihtimali = 31,3516579505398

2. en iyi tahmin = 00000000100000000 Yanlış cevap ihtimali = 31,8952736506951

3. en iyi tahmin = 00000000000000001 Yanlış cevap ihtimali = 31,9388242653855

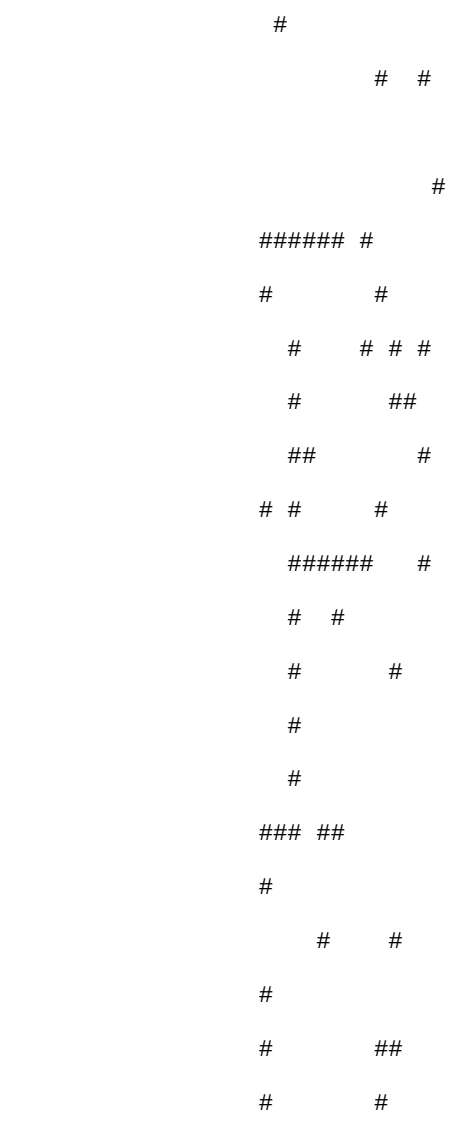
\\---| AĞA SUNULAN PATERN |



\\---| AĞIN VERDİĞİ CEVAP |

```
   Tahmin grupları = ABCDEFGHIJKLMNOPR
                     ||||||||||||||||
Ağın verdiği cevap = 0000000000000010 Yanlış cevap ihtimali = 1,90828267057442
  1. en iyi tahmin = 0000000000000001 Yanlış cevap ihtimali = 32,7836076994992
  2. en iyi tahmin = 0000000100000000 Yanlış cevap ihtimali = 32,8044041427968
  3. en iyi tahmin = 0000000000100000 Yanlış cevap ihtimali = 32,9535914111392
```

\\----------| AĞA SUNULAN PATERN |



\\----------| AĞIN VERDİĞİ CEVAP |

```
   Tahmin grupları = ABCDEFGHIJKLMNOPR
                     ||||||||||||||||
Ağın verdiği cevap = 0000000000000010 Yanlış cevap ihtimali = 2,05073620989326
  1. en iyi tahmin = 0000000000000001 Yanlış cevap ihtimali = 32,5053309117537
  2. en iyi tahmin = 0000000100000000 Yanlış cevap ihtimali = 32,7608192502205
  3. en iyi tahmin = 0000000000100000 Yanlış cevap ihtimali = 32,8426875573363
```

\\---| AĞA SUNULAN PATERN |

\\---| AĞIN VERDİĞİ CEVAP |

Tahmin grupları = ABCDEFGHIJKLMNOPR

|||||||||||||||||

Ağın verdiği cevap = 00000000000000010 Yanlış cevap ihtimali = 2,77530013313446

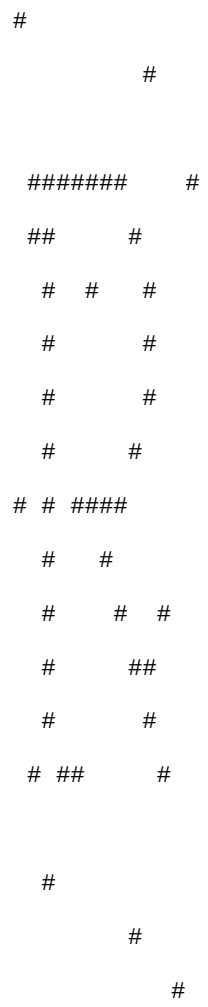
1. en iyi tahmin = 00000000000000001 Yanlış cevap ihtimali = 31,7840194570057

2. en iyi tahmin = 00000000010000000 Yanlış cevap ihtimali = 32,4113710013321

3. en iyi tahmin = 00000000100000000 Yanlış cevap ihtimali = 32,4206953386243

```
\\-----------------------------------------------------| AĞA SUNULAN PATERN |
                            ##



                         #######
                           #     #
                           #  #   ##
                           #      # #
                          ##  ##  #
                           ### # #
                           ### ##
                           #  ###   #
                           # #
                           #   #
                           #
                         ######


                               ##
                                #
                             #
                                ##
\\-----------------------------------------------------| AĞIN VERDİĞİ CEVAP |
   Tahmin grupları = ABCDEFGHIJKLMNOPR
                     ||||||||||||||||
Ağın verdiği cevap = 00000000000000010 Yanlış cevap ihtimali = 3,41568026750555
  1. en iyi tahmin = 00000000000000001 Yanlış cevap ihtimali = 31,2449454245601
  2. en iyi tahmin = 01000000000000000 Yanlış cevap ihtimali = 31,7695459901335
  3. en iyi tahmin = 00000000000100000 Yanlış cevap ihtimali = 32,0418171871407
```

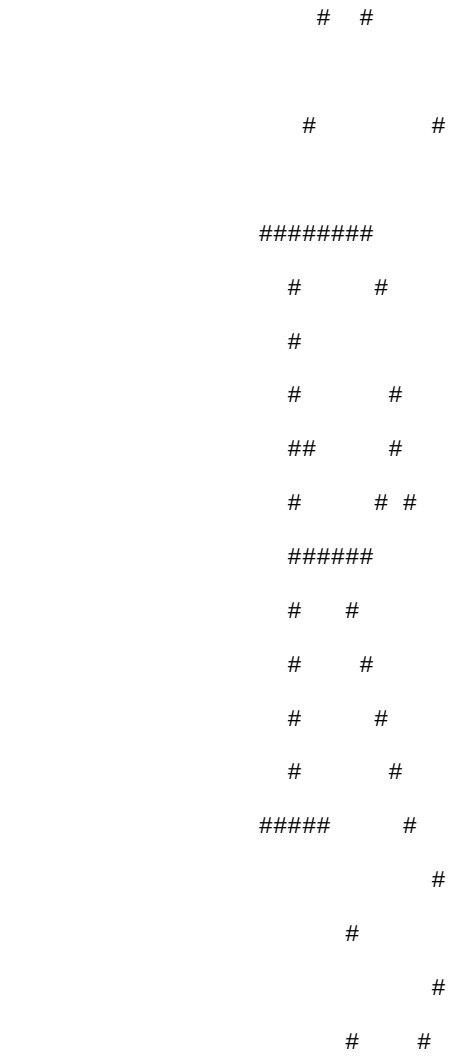
\\----------| AĞA SUNULAN PATERN |



\\----------| AĞIN VERDİĞİ CEVAP |

```
   Tahmin grupları = ABCDEFGHIJKLMNOPR
                     ||||||||||||||||
Ağın verdiği cevap = 0000000000000001 Yanlış cevap ihtimali = 10,2871113212896
  1. en iyi tahmin = 0000000000000010 Yanlış cevap ihtimali = 26,189041334697
  2. en iyi tahmin = 0001000000000000 Yanlış cevap ihtimali = 27,4548769105424
  3. en iyi tahmin = 0000000000100000 Yanlış cevap ihtimali = 28,0919990897431
```

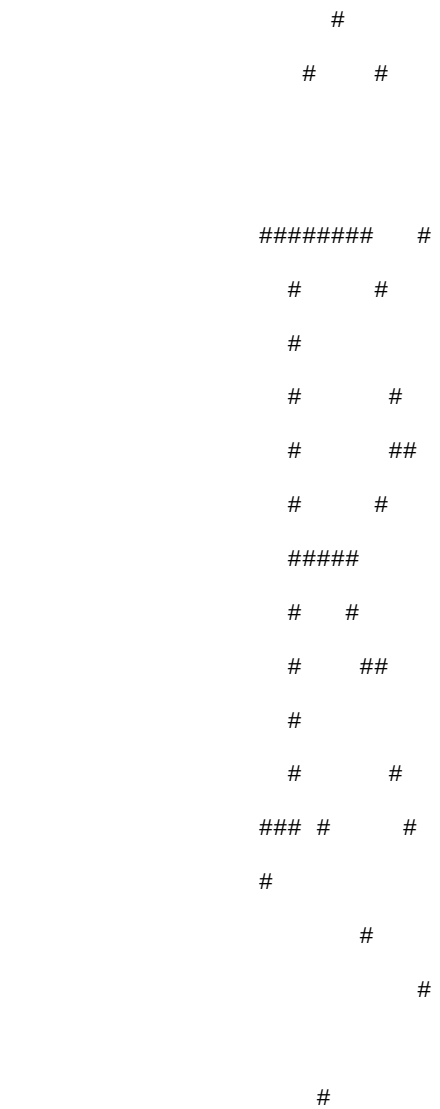
\\---| AĞA SUNULAN PATERN |



\\---| AĞIN VERDİĞİ CEVAP |

```
   Tahmin grupları = ABCDEFGHIJKLMNOPR
                     ||||||||||||||||
Ağın verdiği cevap = 0000000000000001 Yanlış cevap ihtimali = 12,5684821475281
  1. en iyi tahmin = 0000000000000010 Yanlış cevap ihtimali = 22,7174935474352
  2. en iyi tahmin = 0000000000100000 Yanlış cevap ihtimali = 32,1737964191388
  3. en iyi tahmin = 0001000000000000 Yanlış cevap ihtimali = 32,2946430812113
```

\\----------| AĞA SUNULAN PATERN |



\\----------| AĞIN VERDİĞİ CEVAP |

Tahmin grupları = ABCDEFGHIJKLMNOPR

||||||||||||||||

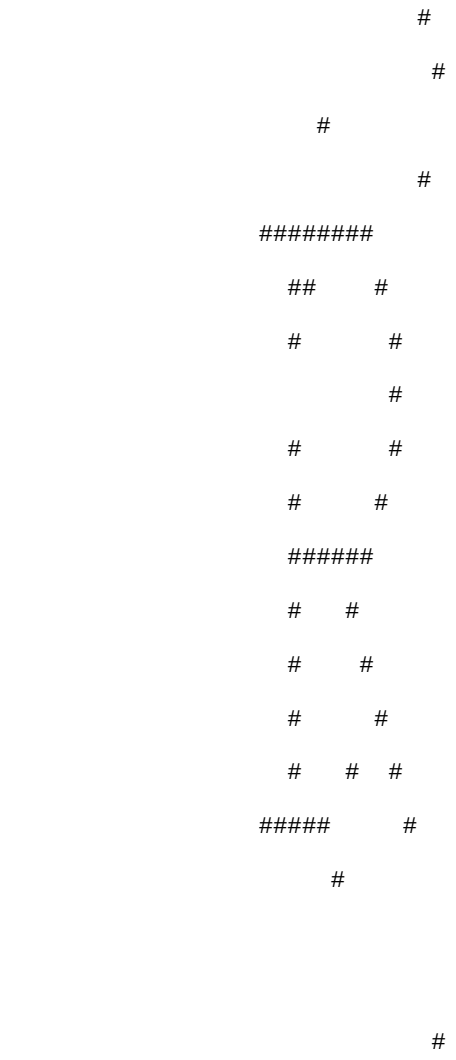
Ağın verdiği cevap = 0000000000000001 Yanlış cevap ihtimali = 14,7325398829829

1. en iyi tahmin = 0000000000000010 Yanlış cevap ihtimali = 19,8009756900232

2. en iyi tahmin = 0000000000100000 Yanlış cevap ihtimali = 30,0495735157513

3. en iyi tahmin = 0001000000000000 Yanlış cevap ihtimali = 30,4586791710998

\\----------| AĞA SUNULAN PATERN |



\\----------| AĞIN VERDİĞİ CEVAP |

Tahmin grupları = ABCDEFGHIJKLMNOPR

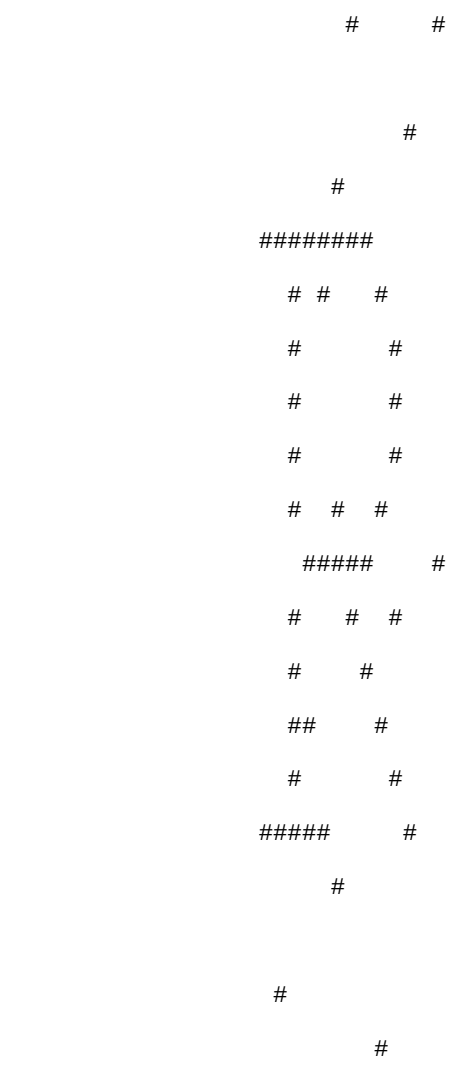
Ağın verdiği cevap = 00000000000000001 Yanlış cevap ihtimali = 9,51889517816077

1. en iyi tahmin = 00000000000000010 Yanlış cevap ihtimali = 25,2195118002971

2. en iyi tahmin = 00000000000100000 Yanlış cevap ihtimali = 31,9329973623107

3. en iyi tahmin = 10000000000000000 Yanlış cevap ihtimali = 31,9983364985283

\\----------| AĞA SUNULAN PATERN |



\\----------| AĞIN VERDİĞİ CEVAP |

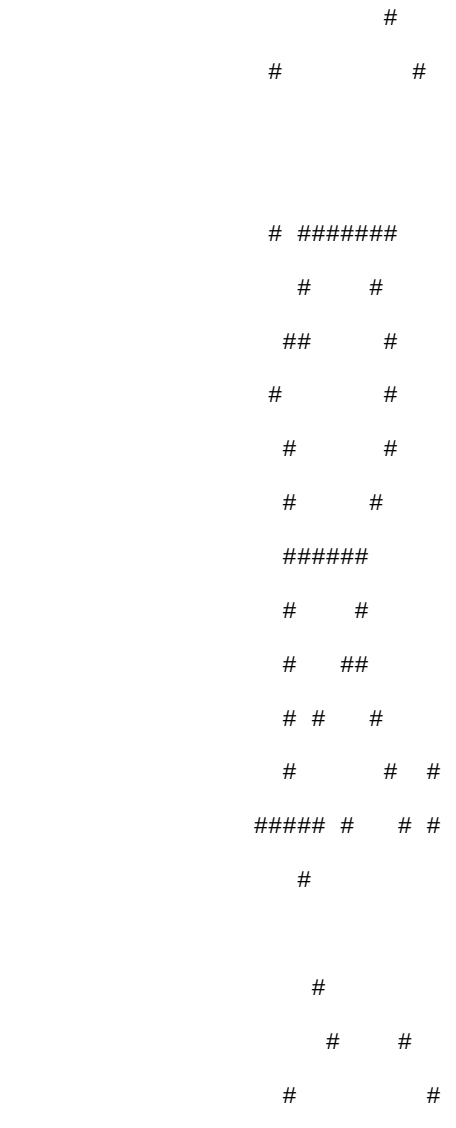
```
   Tahmin grupları = ABCDEFGHIJKLMNOPR
                     ||||||||||||||||
Ağın verdiği cevap = 0000000000000001 Yanlış cevap ihtimali = 7,14229932529347
  1. en iyi tahmin = 0000000000000010 Yanlış cevap ihtimali = 27,1826784784222
  2. en iyi tahmin = 0000000000100000 Yanlış cevap ihtimali = 31,4200072225809
  3. en iyi tahmin = 1000000000000000 Yanlış cevap ihtimali = 31,4219331220959
```

\\----------| AĞA SUNULAN PATERN |

\\----------| AĞIN VERDİĞİ CEVAP |

```
   Tahmin grupları = ABCDEFGHIJKLMNOPR
                     ||||||||||||||||
Ağın verdiği cevap = 0000000000000001 Yanlış cevap ihtimali = 16,3457641642733
  1. en iyi tahmin = 0000000000000010 Yanlış cevap ihtimali = 18,0713954999402
  2. en iyi tahmin = 0000000000100000 Yanlış cevap ihtimali = 28,4045264889505
  3. en iyi tahmin = 1000000000000000 Yanlış cevap ihtimali = 28,9057284770179
```

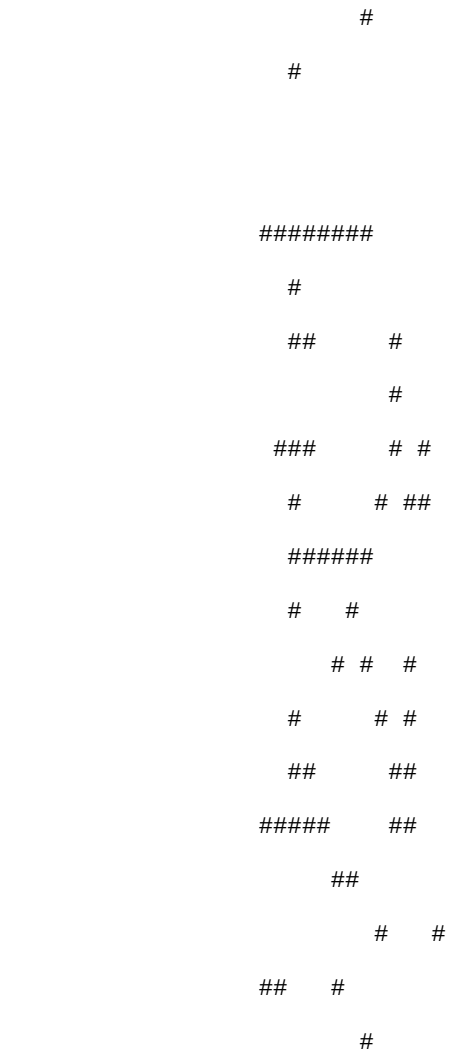
\\----------| AĞA SUNULAN PATERN |



\\----------| AĞIN VERDİĞİ CEVAP |

```
   Tahmin grupları = ABCDEFGHIJKLMNOPR
                     ||||||||||||||||
Ağın verdiği cevap = 0000000000000010 Yanlış cevap ihtimali = 16,9804021289822
  1. en iyi tahmin = 0000000000000001 Yanlış cevap ihtimali = 19,6160627637488
  2. en iyi tahmin = 0000000001000000 Yanlış cevap ihtimali = 26,3322679703962
  3. en iyi tahmin = 1000000000000000 Yanlış cevap ihtimali = 26,3494590024978
               NOT = *** AĞ CEVABI YANLIŞ ***
```

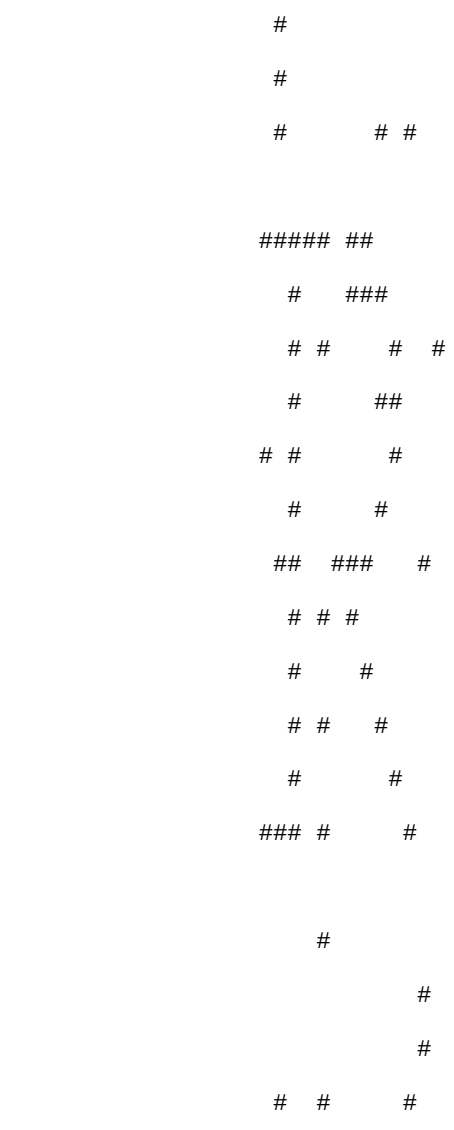
\\---| AĞA SUNULAN PATERN |



\\---| AĞIN VERDİĞİ CEVAP |

```
   Tahmin grupları = ABCDEFGHIJKLMNOPR
                     ||||||||||||||||
Ağın verdiği cevap = 0000000000000001 Yanlış cevap ihtimali = 10,0647218253259
  1. en iyi tahmin = 0000000000000010 Yanlış cevap ihtimali = 25,1076935518423
  2. en iyi tahmin = 0000000000100000 Yanlış cevap ihtimali = 28,4222127601241
  3. en iyi tahmin = 0000000000010000 Yanlış cevap ihtimali = 28,7386968204373
```

\\----------| AĞA SUNULAN PATERN |



\\----------| AĞIN VERDİĞİ CEVAP |

Tahmin grupları = ABCDEFGHIJKLMNOPR

||||||||||||||||

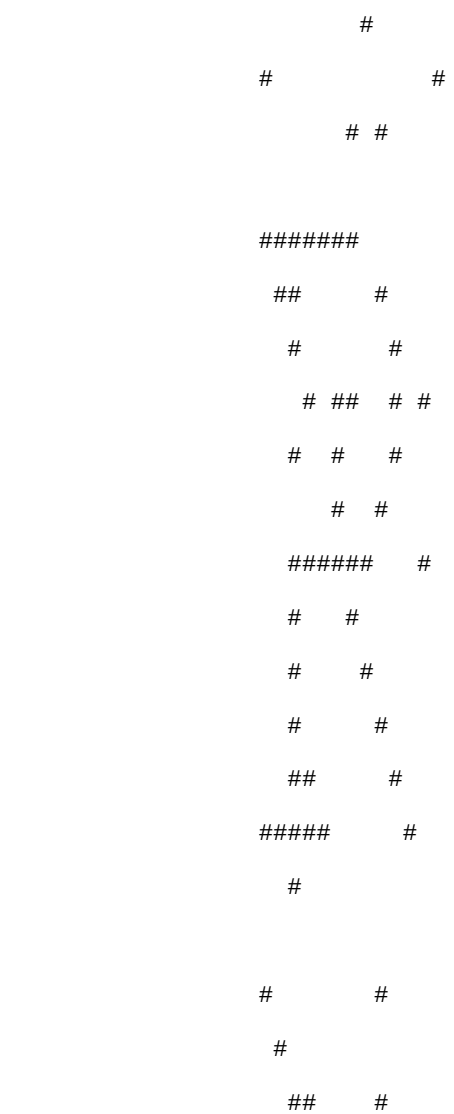
Ağın verdiği cevap = 0000000000000001 Yanlış cevap ihtimali = 6,9602762376494

1. en iyi tahmin = 0000000000000010 Yanlış cevap ihtimali = 27,4586932908244

2. en iyi tahmin = 0000000000100000 Yanlış cevap ihtimali = 30,600552779337

3. en iyi tahmin = 0001000000000000 Yanlış cevap ihtimali = 30,7085290882995

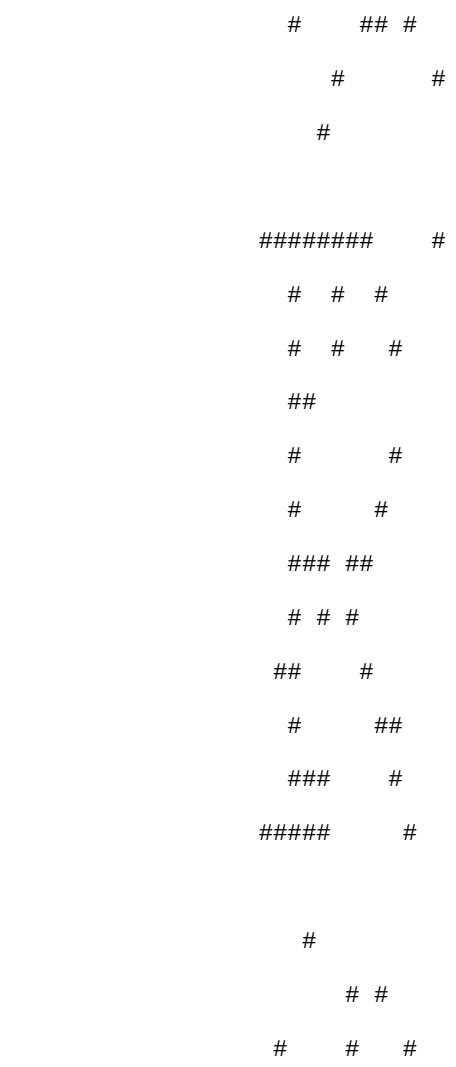
\\----------| AĞA SUNULAN PATERN |



\\----------| AĞIN VERDİĞİ CEVAP |

```
   Tahmin grupları = ABCDEFGHIJKLMNOPR
                     ||||||||||||||||
Ağın verdiği cevap = 0000000000000001 Yanlış cevap ihtimali = 8,06184033929896
  1. en iyi tahmin = 0000000000000010 Yanlış cevap ihtimali = 26,4987332991307
  2. en iyi tahmin = 0000000000100000 Yanlış cevap ihtimali = 29,9678939146532
  3. en iyi tahmin = 1000000000000000 Yanlış cevap ihtimali = 29,9829607543665
```

\\----------| AĞA SUNULAN PATERN |



\\----------| AĞIN VERDİĞİ CEVAP |

```
   Tahmin grupları = ABCDEFGHIJKLMNOPR
                     ||||||||||||||||
Ağın verdiği cevap = 0000000000000001 Yanlış cevap ihtimali = 6,22759099598094
  1. en iyi tahmin = 0000000000000010 Yanlış cevap ihtimali = 28,4503326232855
  2. en iyi tahmin = 0000000000100000 Yanlış cevap ihtimali = 30,1248047107071
  3. en iyi tahmin = 1000000000000000 Yanlış cevap ihtimali = 30,6563543740886
```

```
\\-------------------------------------------------------| TOPLU TEST SONUCU |

          Toplam test maddesi = 300
     Ağın verdiği doğru cevap = 287
    Ağın verdiği yanlış cevap = 13
         Yanlış / Doğru oranı = 0,0452961672473868
Düzeltilmiş doğru cevap puanı = 283,75
            Ağın başarı oranı = 95,6666666666667
     Düzeltilmiş başarı oranı = 94,5833333333333
```

Bu ve benzeri testler farklı örüntüleri kapsayacak şekilde tekrar tekrar yapılmıştır. Bu örüntüler, örüntü sayıları, ağın düzeltilmiş başarı oranları Tablo 40 ‘da gösterilmiştir.

Tablo 40: Ağın ayrım gücü ve başarı oranları.

Örüntü ikilisi	**Örüntü sayısı**	**Başarı oranı**
P ve R	300	94,583333
O ve P	300	98,234566
I ve J	300	94,6733332
J ve Y	300	95,1113333
L ve J	300	95,999299
K ve M	300	97,563456

4.4. Fare

Solucan veri tabanı genel bilgi ihtiyaçlarını karşılamak ve özel bir takım verilere ulaşmak için kullanılan bilgileri depolamaktadır. Bu bilgilerden yola çıkarak daha özel bilgileri toparlamasına ihtiyaç duyulmuştur. İkincil yapı elementleri, bunların protein içindeki yerleşimleri, disülfit bağlanmaları, protein içindeki varyasyonlar ile mutasyonlar ve SNP noktaları, molekül ağırlıkları, NMR ve X-RAY dosyaları ile atomsal koordinatlar bu özel bilgilerden bazılarıdır. Fare, Solucan tabanına sahip bir yazılım olup bahsedilen özel bilgilerin otomatik olarak toparlanıp derlenmesi için geliştirilmiştir. Ayrıca Fare,

yapay sinir ağları ile veri tabanları arasındaki ilişkiyi de sağlamaktadır. Bu şekliyle tez için entegre bir çözüm oluşturmuştur. Protein tablosu Tablo 41 ‘de gösterilmiştir.

Tablo 41: Fare programının proteinlere ait veri tabanı (Fare programının ekran çıktısından alınmıştır).

HgncId	Sembol	ErisimNo	Tipi	Uzunluk	MolekulAgirligi	Helix	BetaTabaka	Donus	Zincir	Disulfid
5986	IL18	Q14116	1	195	22326	3	82	4	106	0
6051	IMPA2	O14732	1	290	31321	94	59	4	133	0
6118	IRF3	Q14653	1	429	47219	97	106	3	223	1
6177	ITPK1	Q13572	1	416	45621	117	78	7	214	0
6207	JUP	P14923	1	747	81745	357	3	10	377	0
6357	KLK1	P06870	1	264	28890	28	90	3	143	5
6368	KLK7	P49862	1	255	27525	27	86	3	139	6
6407	KRAS	P01116	1	191	21656	46	48	5	92	0
6480	LALBA	P00709	1	144	16225	51	9	10	74	4
6508	LANCL1	O43813	1	401	45283	182	3	13	203	0
6568	LGALS7	P47929	1	138	15075	3	68	4	63	0
6598	LIG1	P18858	1	921	101736	247	142	28	504	0
6701	LRPAP1	P30533	1	359	41466	187	11	3	158	0
6709	LTA	P01374	1	207	22297	7	83	3	114	0
6737	LYPLA1	O75608	1	232	24670	79	52	4	97	0
6838	MAP1LC3A	Q9H492	1	123	14272	38	25	3	57	0
6842	MAP2K2	P36507	1	402	44424	122	48	9	223	0
6871	MAPK1	P28482	1	362	41390	135	35	18	174	0
6872	MAPK10	P53779	1	466	52585	132	63	16	255	0
6873	MAPK11	Q15759	1	366	41357	129	38	10	189	0
6874	MAPK12	P53778	1	369	41940	129	44	15	181	0
6875	MAPK13	O15264	1	367	42090	127	63	20	157	0
6876	MAPK14	Q16539	1	362	41293	130	46	10	176	0
6881	MAPK8	P45983	1	429	48296	130	61	16	222	0
6984	ME2	P23368	1	586	65444	262	55	24	245	0
7030	METTL1	Q9UBP6	1	278	31471	69	50	12	147	0
7155	MMP1	P03956	1	471	54007	82	134	19	236	1
7166	MMP2	P08253	1	662	73882	99	204	36	323	7
7373	MSN	P26038	1	579	67820	205	93	16	265	0
7413	MTAP	Q13126	1	285	31236	86	74	6	119	0
7423	MTCP1	P56278	1	109	12600	4	53	3	49	0
7450	MTMR2	Q13614	1	645	73381	193	106	45	301	0
7631	NAGA	P17050	1	413	46565	108	69	18	218	4
7645	NAT1	P18440	1	292	33899	84	82	4	122	0
7659	NCBP2	P52298	1	158	18001	43	33	10	72	0
7660	NCF1	Q9BXI8	1	392	44683	83	94	5	210	0
7861	NNMT	P40261	1	266	29574	90	53	3	120	0
7866	NOG	Q13253	1	234	25774	47	66	3	118	4
7945	NPR3	P17342	1	543	59808	158	102	28	255	3
7968	NR1I2	Q9UJ26	1	436	49762	167	23	4	242	0
8048	NUDT1	P36639	1	199	22520	23	67	3	106	0

Burada proteinlerin HgncId’leri, sembolleri, erişim numaraları, tipleri, teorik moleküler ağırlıkları, üç yapısal sınıf ile disülfid bağlarının sayısı, zincir başlığı altında ise üç yapısal sınıfa dahil olmayan rezüdü sayıları verilmiştir

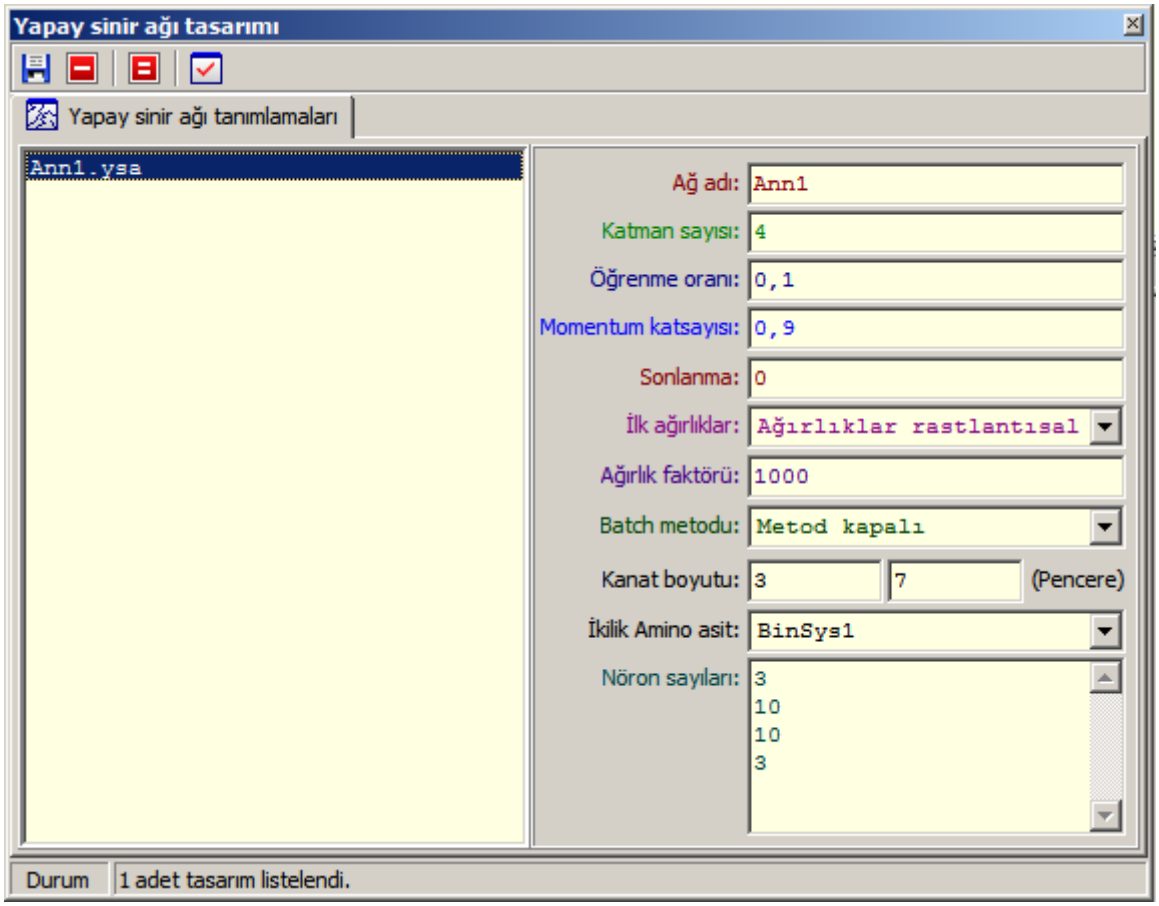
Tablo 42’de ayrıntılı yapı bilgilerinin listelendiği yapı tablosu gösterilmiştir.

Tablo 42: Proteinlerin yapısal ayrıntılarının gösterildiği Yapı tablosu (Fare programının ekran çıktısından alınmıştır).

idYapi	HgncId	Baslangic	Bitis	Tip	Dizi
82	51	644	653	STRAND	ITVRNATFTW
83	51	660	668	STRAND	TLNGITFSI
84	51	673	677	STRAND	LVAVV
85	51	684	691	HELIX	KSSLLSAL
86	51	695	704	STRAND	MDKVEGHVAI
87	51	708	711	STRAND	VAYV
88	51	719	721	STRAND	NDS
89	51	722	727	HELIX	LRENIL
90	51	736	743	HELIX	YYRSVIQA
91	51	747	750	HELIX	LPDL
92	51	756	758	HELIX	GDR
93	51	759	762	STRAND	TEIG
94	51	770	784	HELIX	GGQKQRVSLARAVYS
95	51	787	793	STRAND	DIYLFDD
96	51	794	797	TURN	PLSA
97	51	800	809	HELIX	AHVGKHIFEN
98	51	816	819	TURN	MLKN
99	51	820	825	STRAND	KTRILV
100	51	832	834	HELIX	LPQ
101	51	835	842	STRAND	VDVIIVMS
102	51	845	850	STRAND	KISEMG
103	51	852	858	HELIX	YQELLAR
104	51	861	868	HELIX	AFAEFLRT

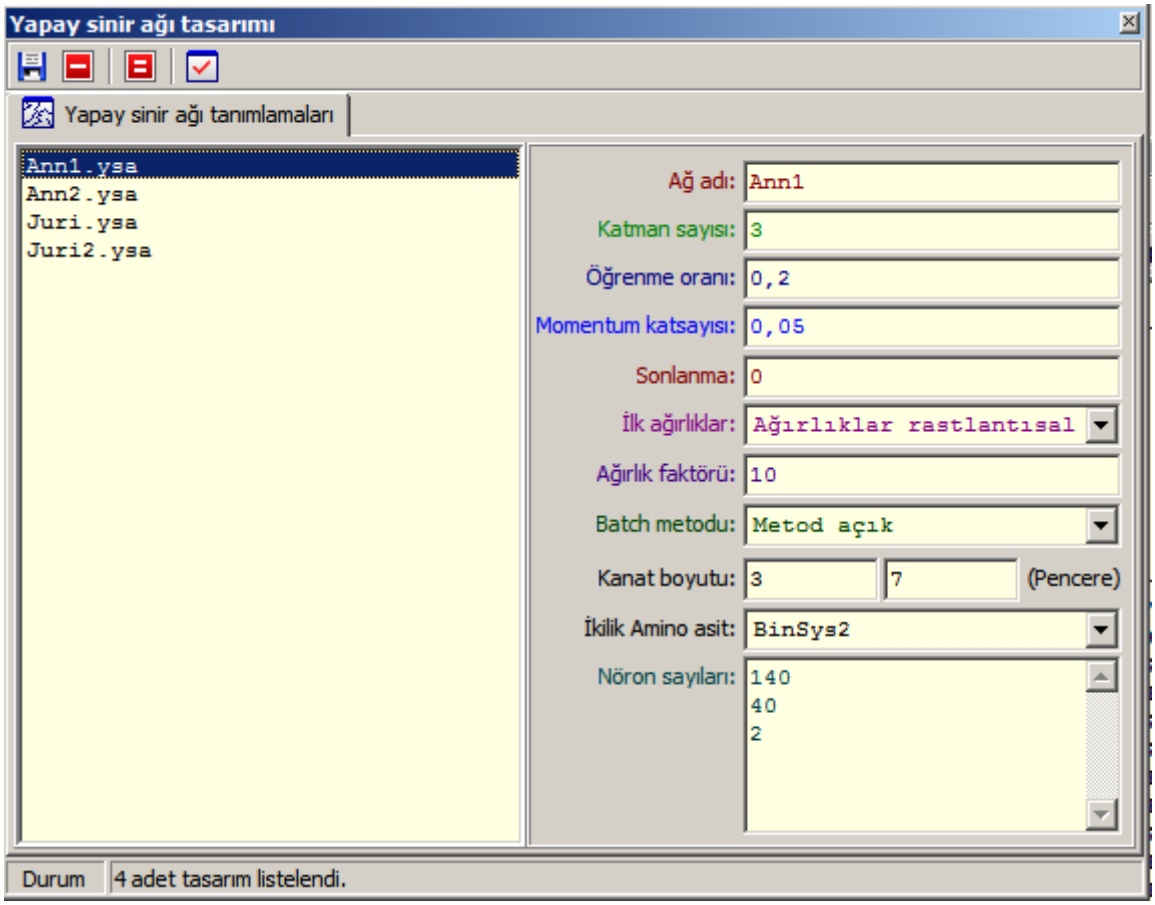
Burada Protein tablosu ile bağlantı kurmak için HgncId, başlangıç ve bitiş noktaları, bu aralıkta hangi yapısal formun bulunduğu ve bu yapısal formun amino asit dizisi gösterilmiştir. Yukarıdaki çıktı ABCC1 isimli gene aittir. Fare, bu sayısal verilerin yanında görsel olarak ta protein ve ikincil yapı elementleri hakkında bilgiler verir.

Fare, yapısında BOT ‘un omurgasını oluşturan yapay sinir ağı motorunu da içermektedir. Gerekli parametreler girildikten sonra pek çok yapay sinir ağını tasarlayıp yönetebilir. Şekil 35 ‘de yapay sinir ağı tasarım ekranı ve bu vesile ile tasarım parametreleri gösterilmiştir.



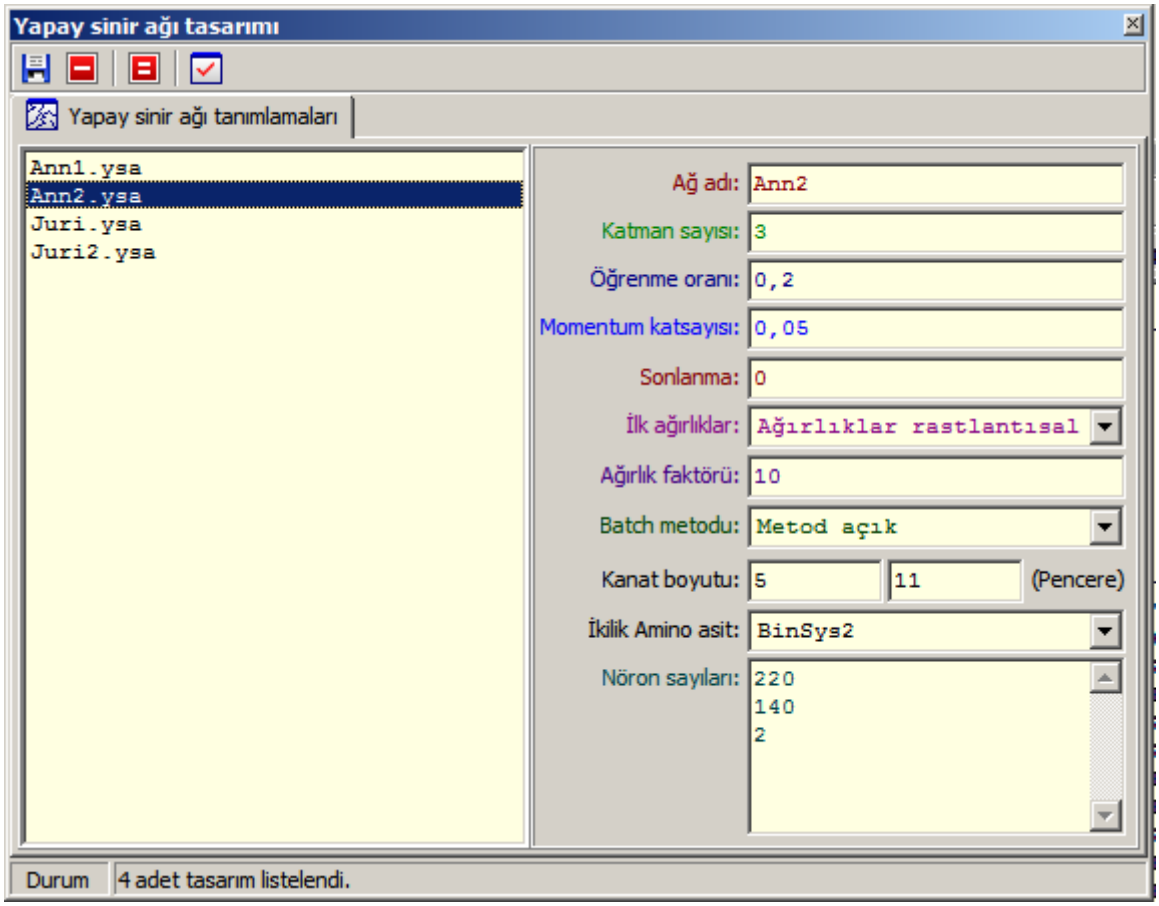
Şekil 35: Yapay sinir ağı tasarım ekranı (Fare programının ekran çıktısından alınmıştır).

Fare için iki ayrı ağ topolojisi tasarlanmıştır. Bu ağlar Ann1 ve Ann2 ağları sonuçları ise Tahmin1 ve Tahmin2 olarak isimlendirilmiştir. Bu ağların topolojileri Şekil 36a ve Şekil 36b ‘de gösterilmiştir.



Şekil 36a: Ann1 ağına ait topolojik parametreler

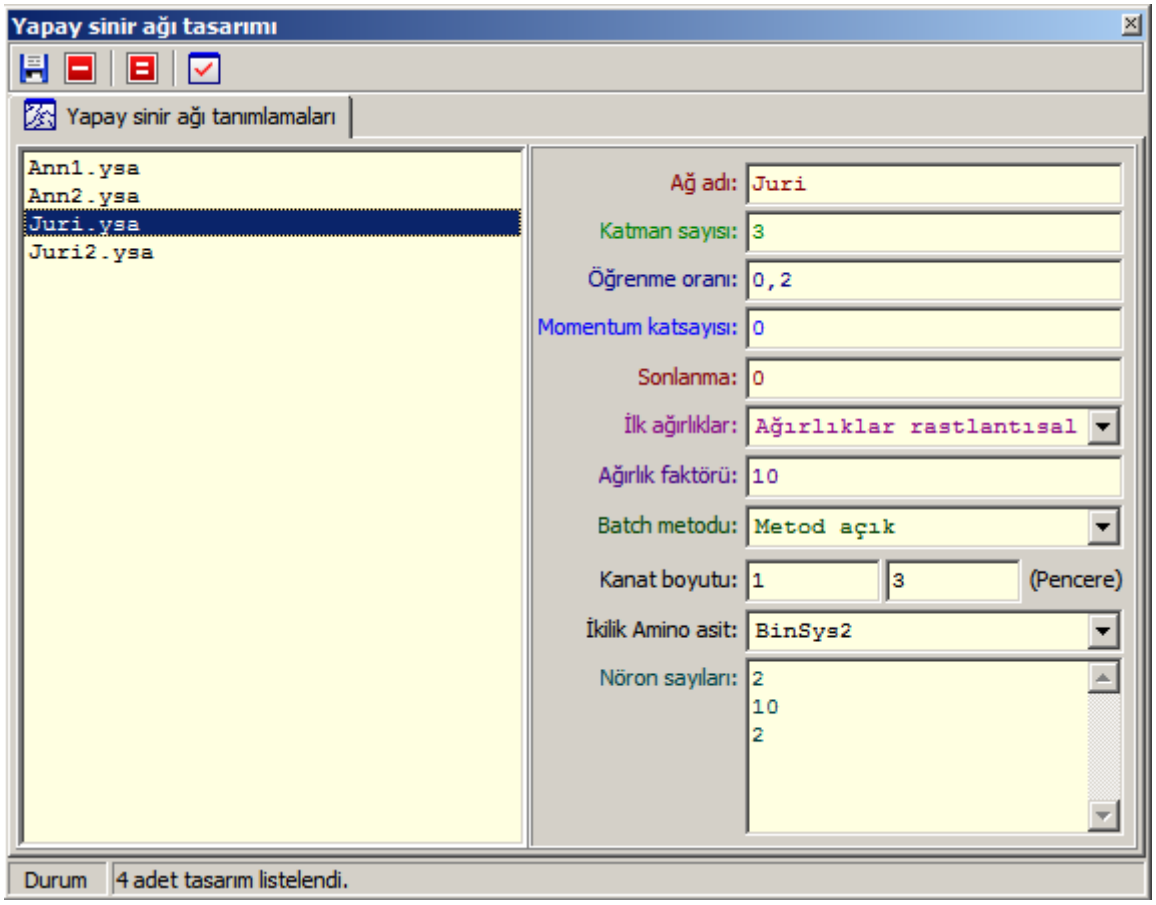
Görüldüğü üzere Ann1 ağının 140 girişi, 40 noddan oluşan bir gizli katmanı ve 2 çıkışı bulunmaktadır.



Şekil 36b: Ann2 ağına ait topolojik parametreler.

Görüldüğü üzere Ann2 ağının 220 girişi, 140 noddan oluşan bir gizli katmanı ve 2 çıkışı bulunmaktadır.

Ann1 ve Ann2 ağları diziden ikincil yapıya olan tahmini gerçekleştirmektedir. Ann1 ve Ann2 ağının sonuçlarını giriş kabul eden Juri1 ve Juri2 ağı ise yapıdan yapıya bir tahmin gerçekleştirmektedir. Juri1 ve Juri2 ağlarının topolojileri aynı olup Şekil 36c ‘de gösterilmiştir.



Şekil 36c: Juri1 ve Juri2 ağlarının topolojik parametreleri

Örnek bir protein olarak MEFV geni proteini pyrin seçilmiştir. Pyrin 783 amino asitten oluşmuştur. Erişim numarası O15553 ‘dır. Hgnc kimlik numarası 6998 ve NCBI Gene kimlik numarası 4210 ‘dur. Bu protein Fare ‘ye girdi olarak verilmiş ve ikincil yapı tahminlerini içeren sonuç çıktısı aşağıda gösterilmiştir.

RAPOR

Erişim no: O15553

Hgnc Id: 6998

Sembol: Pyrin

Uzunluk: 783

Tipi: 1

Adı: Mediterranean fever protein

TAHMİN SONUÇLARI (ORJİNAL İKİNCİL YAPI / TAHMİN 1)

```
                    1         2         3         4
           1234567890123456789012345678901234567890123456789
Tahmin 1 :    HHHHHHHHHHHHHHHHSSHHHHHHHHHHHHHHHHHHHHSHHHHHHHHHH
Tahmin 1 : HHHHHHHHHHHHHHHHHSSHHHHHHHHHHHHSHHHHHHHHHSHSSSHHHHSSS
Tahmin 1 : HHHHHHHHHHHHHHHHSSSHHSSSSSSSSSSSSHHHHHHHHHHHHHHHHHHHH
Tahmin 1 : HHHHHHSSSSSHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHSSSSSHHHHH
Tahmin 1 : SSSSSSSSSSSSSSSSSHHSSSSSHHHSSSSSSSSSHHHHHHHHHSSSSSHH
Tahmin 1 : HHHHHHSSSHHSHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHSHHHHHH
Tahmin 1 : HHHHHHHHSSSSSHHHHHHHHSHSHHHHHHHHHHHHHHHHHHHHHHHHHH
Tahmin 1 : HHHHHHHHHHHSSSSSHHHHHHHHHHSHSHHHHHHHHHHHHHHHHHHHSH
Tahmin 1 : SSSSSSSSSSSSSSSSSSHSHHHHHHHHSHSHSSSSHSSSSSSSSSSSSS
```

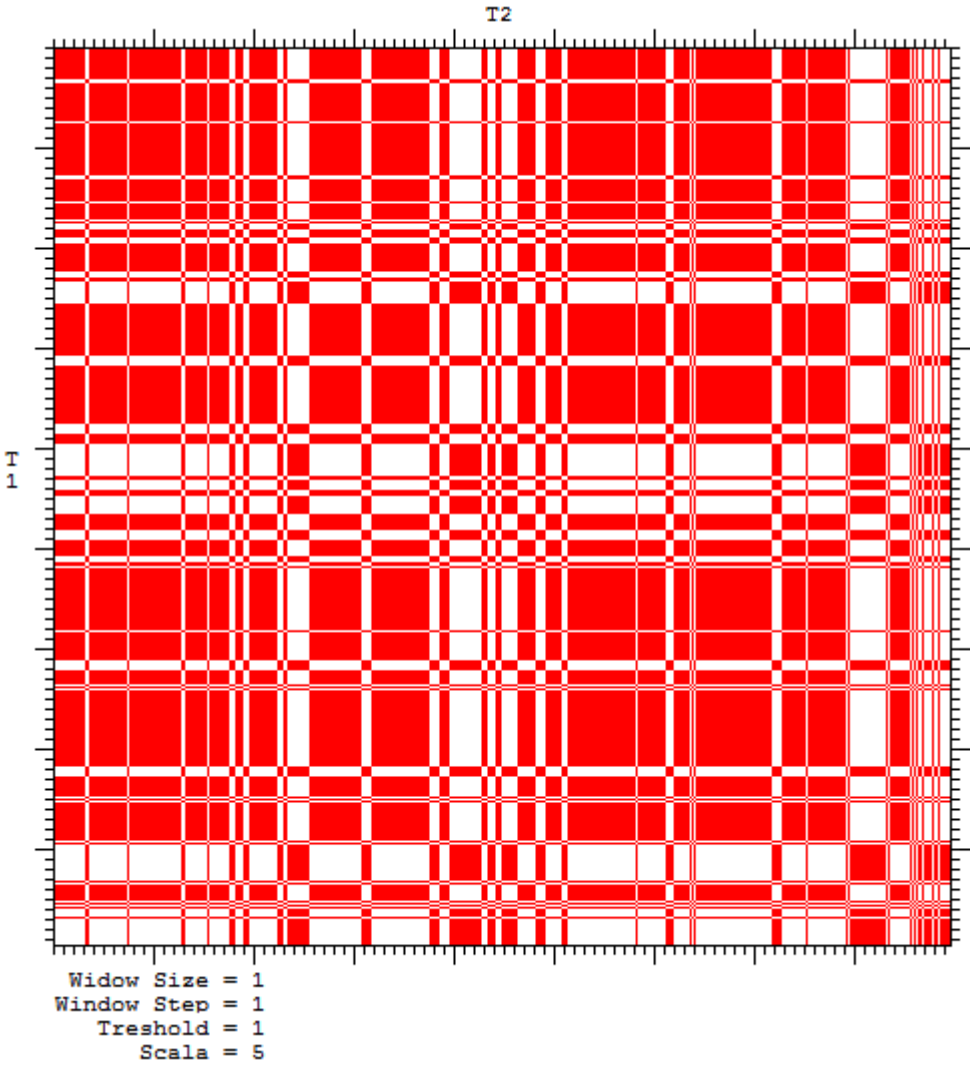
TAHMİN SONUÇLARI (ORJİNAL İKİNCİL YAPI / TAHMİN 2)

```
                    1         2         3         4
           1234567890123456789012345678901234567890123456789
Tahmin 2 :    HHHHHHHHHHHHHHHHSSHHHHHHHHHHHHHHHHHHHHSHHHHHHHHHH
Tahmin 2 : HHHHHHHHHHHHHHHHHSSHHHHHHHHHHHHSHHHHHHHHHSHSSSHHHHSSS
Tahmin 2 : HHHHHHHHHHHHHHHHSSSHHSSSSSSSSSSSSHHHHHHHHHHHHHHHHHHHH
Tahmin 2 : HHHHHHSSSSSHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHSSSSSHHHHH
Tahmin 2 : SSSSSSSSSSSSSSSSSHHSSSSSHHHSSSSSSSSSHHHHHHHHHSSSSSHH
Tahmin 2 : HHHHHHSSSHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHSHHHHHH
Tahmin 2 : HHHHHHHHHSSSSHHHHHHHHSHSHHHHHHHHHHHHHHHHHHHHHHHHHH
```

```
Tahmin 2 : HHHHHHHHHHHSSSSSHHHHHHHHHHHHSHHHHHHHHHHHHHHHHHHSH
Tahmin 2 : SSSSSSSSSSSSSSSSSSHSHHHHHHHHHHSHSHSSHSSSSHSSHSSSSS

HİZALAMA SONUÇLARI (TAHMİN1 / TAHMİN 2)
------------------------------------------------------------
O15553_T1:   HHHHHHHHHHHHHHHHSSHHHHHHHHHHHHHHHHHHHHSHHHHHHHHHH
 Operator:||||||||||||||||||||||||||||||||||||||||||||||||||||
O15553_T2:   HHHHHHHHHHHHHHHHSSHHHHHHHHHHHHHHHHHHHHSHHHHHHHHHH

O15553_T1:HHHHHHHSSSSSSHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHSSHH
 Operator:||||||||||||||||||||||||||||||||||||||||||||||||||||
O15553_T2:HHHHHHHSSSSSSHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHSSHH

O15553_T1:HHHHHHHHHSHHHHHHHHS-HSSSHHHHSSSSSSSSSSSSHHHHS-HHHH
 Operator:||||||||||||||||||  ||||||||||||||||||||||||  ||||
O15553_T2:HHHHHHHHHSHHHHHHHH-HHSSSHHHHSSSSSSSSSSSSHHHH-HHHHH

O15553_T1:HHHHSS-HHHHHHHHHHHHHHHHHHHSSSHHSSSSSSSSSSSHHHHHHH
 Operator:|||| | ||||||||||||||||||||||||||||||||||||||||||||
O15553_T2:HHHH-SHHHHHHHHHHHHHHHHHHHSSSHHSSSSSSSSSSSHHHHHHH

O15553_T1:HHHHHHHHHHHHHHHHHHHSSSSHHHHHHHSS-HS-SSSHSSSHHHHHH
 Operator:|||||||||||||||||||||||||||||| | |  ||||||||||||
O15553_T2:HHHHHHHHHHHHHHHHHHHSSSSHHHHHHH-SHH-HSSSHSSSHHHHHH

O15553_T1:HSSSSSHHHHHHHHHHHHHHHHHHHHHHHHHHHHHSSSSSHHHHHHHSS
 Operator:||||||||||||||||||||||||||||||||||||||||||||||||||||
O15553_T2:HSSSSSHHHHHHHHHHHHHHHHHHHHHHHHHHHHHSSSSSHHHHHHHSS

O15553_T1:SSSSSSSSSSSSHHHHHHHHHHHHHSSSSSSSSSSSSSSSSHHS-SSSSH
 Operator:|||||||||||||||||||||||||||||||||||||||||||  |||||
O15553_T2:SSSSSSSSSSSSHHHHHHHHHHHHHSSSSSSSSSSSSSSSSHH-HSSSSH

O15553_T1:HHSSSSSSSSS-HHHHHHHHSSSSSHHHHHHHHHHHHHHHHHHHHHH
```

```
 Operator:|| |||||||| ||||||||||||||||||||||||||||||||||||
O15553_T2:HH-SSSSSSSSHHHHHHHHHHSSSSSHHHHHHHHHHHHHHHHHHHHHHHHH

O15553_T1:HHSSSHHHHHHHHHSSSHHS-HHHHHHHHHHHHHHHHHHHHHHHHHHHHH
 Operator:||||||||||||||||||  ||||||||||||||||||||||||||||||
O15553_T2:HHSSSHHHHHHHHHSSSHH-HHHHHHHHHHHHHHHHHHHHHHHHHHHHH

O15553_T1:HSHHHHHHHHHHHHHHHHHHHHHHHHHHSSHHHHHHHHHHHHHHHHSSSS
 Operator:|||||||||||||||||||||||||||||||||||||||||||||||| |||
O15553_T2:HSHHHHHHHHHHHHHHHHHHHHHHHHHHSSHHHHHHHHHHHHHHHH-SSS

O15553_T1:S-HHHHHHHSHSHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHSS
 Operator:| ||||||||||||||||||||||||||||||||||||||||||||||||
O15553_T2:SHHHHHHHHSHSHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHSS

O15553_T1:SSSHHHHS-SHHHHHHHHHHHHHHHHHHHHSSSSSHHHHHHHHHHS-HS
 Operator:|||||||  ||||||||||||||||||||||||||||||||||  ||
O15553_T2:SSSHHHH-HSHHHHHHHHHHHHHHHHHHHHSSSSSHHHHHHHHHH-HHS

O15553_T1:HHHHHHHHHHHHHHHHHHHSHSHS-HHHHHHSSHHHHHHHHSHHHS-HH
 Operator:|||||||||||||||||||||||  ||||||||||||||||||  ||
O15553_T2:HHHHHHHHHHHHHHHHHHHSHSH-HHHHHHHSSHHHHHHHHSHHH-HHH

O15553_T1:SSS-SSSSSSSSSSSSSSSSSSHSHHHHHHHHS-HSHSS-SSHSSSSS-S
 Operator: || |||||||||||||||||||||||||||||  ||| | ||| ||||
O15553_T2:-SSHSSSSSSSSSSSSSSSSSSHSHHHHHHHH-HHSH-SHSSH-SSSSH-

O15553_T1:SS-SSSSSSSSHHHHHHHHSSSSSSHHHHHHHHSSSSSSSSSSSSSSSS
 Operator:|| ||||||||||||||||||||||||||||||||||||||||||||||||
O15553_T2:SSHSSSSSSSSSHHHHHHHHSSSSSSHHHHHHHHSSSSSSSSSSSSSSSS

Hizalama Skoru : 1556:1496 Hizalama oranı : 0,961439588688946
```

İki tahmin sonucu arasındaki benzerliğin dot matrix gösterimi ise Şekil 37 ‘de gösterilmiştir.



Şekil 37: T1 ve T2 arasındaki dot matrix grafiği (Hizalayıcı tarafından oluşturulmuştur).

Ann1 ve Ann2 ağlarına ilişkin tahminler sırası ile Ann1 için T1 ve Ann2 için ise T2 olarak gösterilmiştir. Her ikisinin birbirine benzerlik oranı yaklaşık olarak %96,14 ‘dür. Her iki çıktıyı kullanarak Fare pyrin proteini için üç boyutlu en iyi üç tane model önermiştir. Modeller önerilir iken yapısı NMR ile aydınlatılmış 1507 adet protein ve bunların ikincil yapıları ile tahmin edilen ikincil yapı hizalanmış ve en yüksek skora sahip

olan üç tanesi seçilmiştir. Önerilen modeller ve bu modellerin erişim numaralarını içeren çıktı aşağıda gösterilmiştir.

```
İKİNCİL YAPI HİZALAMA SKORLARI (TÜM GENOM)
--------------------------------------------------------------------------------
  Hizalanacak dizi toplamı : 1505
   Hazırlanan dizi toplamı : 1507
 En çok benzeyen 3D yapı 1 : 1AWE
 En çok benzeyen 3D yapı 2 : 2GFU
 En çok benzeyen 3D yapı 3 : 2E9G
```

Bunlara ilişkin görsel çıktılar ise EBI tarafından açık kaynak kodlu ve serbest dolaşım lisanslı olarak sunulan görüntüleyici olan AstexViewer kullanılmıştır. Sırası ile ağın en iyi birinci tahmini Şekil 38a, ikinci en iyi tahmin Şekil 38b, üçüncü en iyi tahmin ise Şekil 38c ‘de gösterilmiştir.

Şekil 38a: Ağın birinci en iyi tahmini.

Şekil 38b:Ağın ikinci en iyi tahmini.

Şekil 38c: Ağın üçüncü en iyi tahmini.

Pyrin proteininin NMR verileri eşliğinde ortaya atılmış gerçek üç boyutlu görüntüsü ise Şekil 38d ‘de gösterilmiştir.

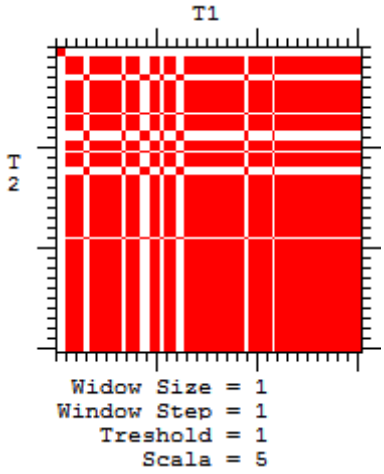
Şekil 38d: Pyrin proteinine ait gerçek üç boyutlu görüntü.

Diğer bir örnek protein olarak APOA1 gen ürünü apolipoprotein A1. Bu protein 269 amino asitten oluşmuştur. Erişim numarası P02647 ‘dir. Hgnc kimlik numarası 600 ‘dür. Bu protein Fare ‘ye girdi olarak verilmiş ve ikincil yapı tahminlerini içeren sonuç çıktısı aşağıda gösterilmiştir.

```
RAPOR
--------------------------------------------------------------------------------
Erişim no: P02647
  Hgnc Id: 600
   Sembol: ApoA1
  Uzunluk: 269
     Tipi: 1
      Adı: apolipoprotein A-I

TAHMİN SONUÇLARI (ORJİNAL İKİNCİL YAPI / TAHMİN 1)
--------------------------------------------------
                    2         3         4         5
           123456789012345678901234567890123456789
Tahmin 1 :    HHHHHHHHHSSSHHHHHHHHHHHHHHHHHSSHHHHHHHSSSSSHHHHH
Tahmin 1 : SSHHHHHHSSSSHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHSSHHHHHH
Tahmin 1 : HHHHHHSHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHH

TAHMİN SONUÇLARI (ORJİNAL İKİNCİL YAPI / TAHMİN 2)
--------------------------------------------------
                    2         3         4         5
           123456789012345678901234567890123456789
Tahmin 2 :    HHHHHHHHHSSSHHHHHHHHHHHHHHHHHSHHHHHHHSSSSSHHHHH
Tahmin 2 : SHHHHHHSSSSHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHSHHHHHH
Tahmin 2 : HHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHH
```

```
HİZALAMA SONUÇLARI (TAHMİN1 / TAHMİN 2)
--------------------------------------------------------------------------------
P02647_T1:   HHHHHHHHHSSSHHHHHHHHHHHHHHHHSS-HHHHHHHSSSSSHHHH
 Operator:|||||||||||||||||||||||||||||||| | ||||||||||||||||
P02647_T2:   HHHHHHHHHSSSHHHHHHHHHHHHHHHHH-SHHHHHHHHSSSSSHHHH

P02647_T1:HSSSHHSSHHHHHHHHHHHHHHHHHHSSSSSSS-HHHHHHSSSSHHHHHH
 Operator:||||||||||||||||||||||||||| |||||| ||||||||||||||||
P02647_T2:HSSSHHSSHHHHHHHHHHHHHHHHHH-SSSSSSHHHHHHHSSSSHHHHHH

P02647_T1:HHHHHHHHHHHHHHHHHHHHHHHHS-SHHHHHHHHHHHHHHHHHHHHHH
 Operator:|||||||||||||||||||||||||  |||||||||||||||||||||||
P02647_T2:HHHHHHHHHHHHHHHHHHHHHHHHH-HSHHHHHHHHHHHHHHHHHHHHHH

P02647_T1:HHHHHHHHHHHHHHHHHHHS-HHHHHHHHHHHHHHHHHHHHHHHHHHHH
 Operator:||||||||||||||||||||  ||||||||||||||||||||||||||||
P02647_T2:HHHHHHHHHHHHHHHHHHHH-HHHHHHHHHHHHHHHHHHHHHHHHHHHH

P02647_T1:HHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHS-HSH
 Operator:|||||||||||||||||||||||||||||||||||||||||||||||  |||
P02647_T2:HHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHH-HHSH
Hizalama Skoru : 528:513 Hizalama oranı : 0,971590909090909
```

İki tahmin sonucu arasındaki benzerliğin dot matrix gösterimi ise Şekil 39 ‘da gösterilmiştir.



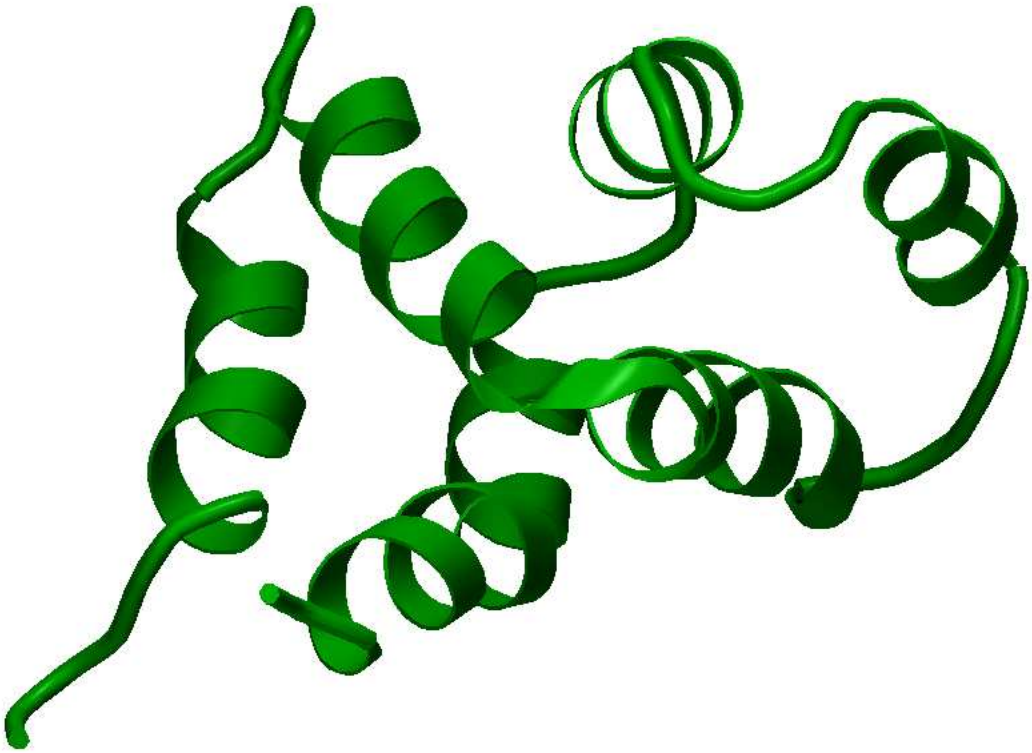
Şekil 39: T1 ve T2 arasındaki dot matrix grafiği (Hizalayıcı tarafından oluşturulmuştur).

Ann1 ve Ann2 ağlarına ilişkin tahminler sırası ile Ann1 için T1 ve Ann2 için ise T2 olarak gösterilmiştir. Her ikisinin birbirine benzerlik oranı yaklaşık olarak %93,64 ‘dür. Her iki çıktıyı kullanarak Fare ApoA1 için üç boyutlu en iyi üç tane model önermiştir. Modeller önerilir iken yapısı NMR ile aydınlatılmış 1507 adet protein ve bunların ikincil yapıları ile tahmin edilen ikincil yapı hizalanmış ve en yüksek skora sahip olan üç tanesi seçilmiştir. Önerilen modeller ve bu modellerin erişim numaralarını içeren çıktı aşağıda gösterilmiştir.

```
İKİNCİL YAPI HİZALAMA SKORLARI (TÜM GENOM)
--------------------------------------------------------------------------------
 Hizalanacak dizi toplamı : 1505
 Hazırlanan dizi toplamı : 1507
En çok benzeyen 3D yapı 0 : 1C15
En çok benzeyen 3D yapı 1 : 2CT4
En çok benzeyen 3D yapı 2 : 2COM
```

Bunlara ilişkin görsel çıktılar ise EBI tarafından açık kaynak kodlu ve serbest dolaşım lisanslı olarak sunulan görüntüleyici olan AstexViewer kullanılmıştır. Sırası ile

ağın en iyi birinci tahmini Şekil 40a, ikinci en iyi tahmin Şekil 40b, üçüncü en iyi tahmin ise Şekil 40c ‘de gösterilmiştir.

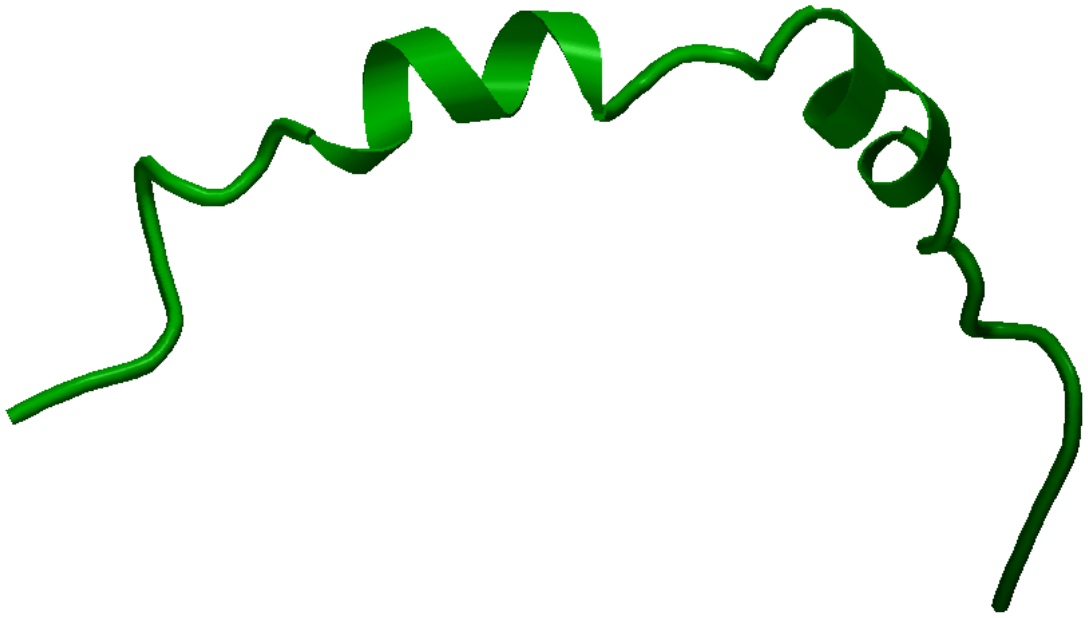


Şekil 40a: Ağın birinci en iyi tahmini.

Şekil 40b: Ağın ikinci en iyi tahmini.

Şekil 40c: Ağın üçüncü en iyi tahmini.

ApoA1 proteininin NMR verileri eşliğinde ortaya atılmış gerçek üç boyutlu görüntüsü ise Şekil 40d ‘de gösterilmiştir.



Şekil 40d: ApoA1 için gerçek üç boyutlu görüntü.

5. TARTIŞMA

Bu tez boyunca pek çok sorunla karşılaşılmıştır. Bu sorunlar biyoinformatik alanının büyük sorunları olarak görülmektedir. Verilerin toplanması bu verilerin değişen sunum formatlarından en az etkilenir şekilde tutulması ve istenildiğinde tekrar düzenlenmesi ya da önemli görülen ilişkilendirmelerin kurulabilmesi bu sorunlardan bazılarıdır. İkincil yapının tahmin edilmesi ise biyoinformatiğin başına gelen en büyük sorun olan kutsal kase unvanını almıştır. Solucan tez boyunca etkili bir şekilde kullanılmıştır. Sadece istediğimiz verilerin nerede olduğunu, nasıl bir düzenli ifade metni oluşturduğunu, hangi bilgi gruplarını almak istediğimizi ve bunları kişisel veri tabanımızda nereye yazmak istediğimizi bildirmek yeterli olmuştur. Bu nedenle bilgisayar başında veri toplama için harcadığımız zamanı minimuma indirmiştir. Günümüzde çok devasa boyutlara ulaşan genetik bilgi uzayını artık bir organizma olarak düşünmek gerekmektedir. Bu şekilde bir organizma hakkında yeniden bir genom projesi başlatmak ve belki de her şeyi baştan, tekrar bu sefer tek anlamlı ve doğal dilinde etkilerinden bir derece azade tasarlamak ütopikte olsa mümkün olup, sonuçları açısından karmaşıklığa bir derece son verilebileceği düşünülmektedir. Solucan genetik bilgi uzayının yanında veri yoğunluğu düşünüldüğünde çok ama çok küçük kalmaktadır (her ne kadar tüm genom projesinin protein katmanın tüm bilgilerine ve temel gen, erişim numarası bilgisine sahip olsa da). Bu dahi bilgiye ulaşmakta geçmek zorunda oluğumuz aşamaları en az indirmiştir. Zaman içinde devamlı değişerek tez boyunca ihtiyaç olan verilere göre özelleşmiştir. Bu sayede bu tez için özel bir veri tabanı oluşmuştur. Her kullanıcının Solucan benzeri kişisel veri tabanları kullandığı düşünülürse, elimizde pek çok özel veri tabanı olacağı görülebilir. Bunlar arasındaki organizasyonu ise Solucan veri modelinin dağıtık yapısı sağlayacağından neyin nerede ve hangi bilgisayarda hangi tablonun hangi alanında olduğu

rahat bir şekilde takip edilebilir olacaktır. Bu bağlamda denemelere devam etmek hatta bir ulusal ağ alt yapısı geliştirmek son derece faydalı olabilir.

Solucan ile bu çalışmada toplam 2 veri tabanı oluşturulmuştur. Bunlardan ilki Solucan veri tabanı olup toplam 5 tablo ve 40000 ‘in üstünde veri satırı içermektedir, ikincisi ise Fare veri tabanı olup toplam 7 tablo ve 90000 ‘in üstünde veri satırı içermektedir.

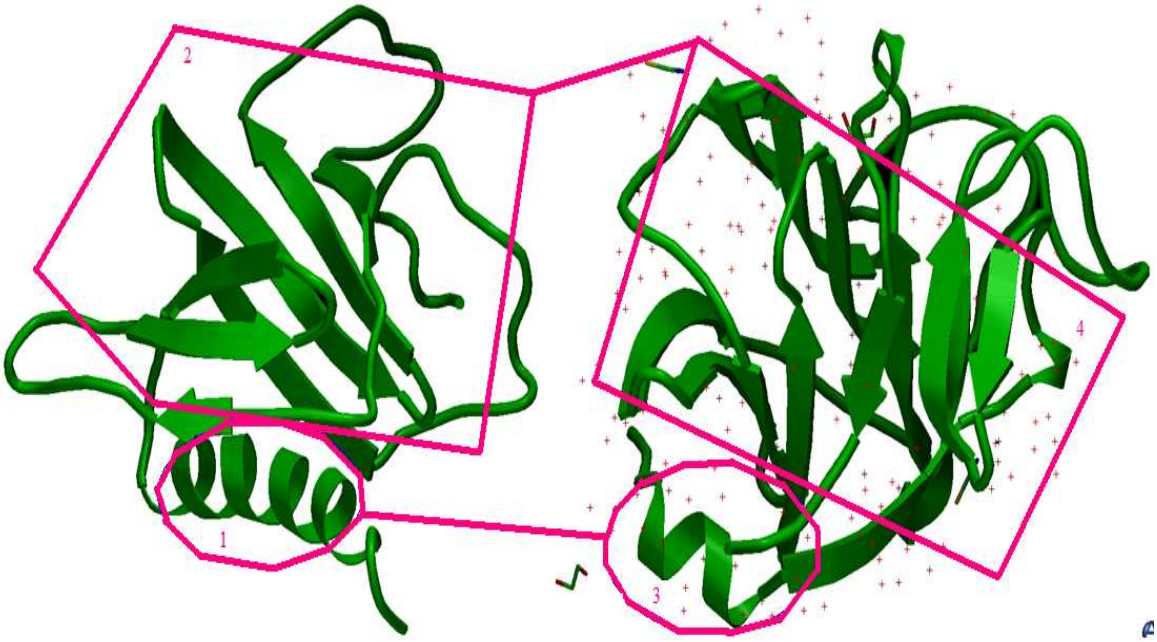
dbSNP ile sadece SNP verilerine ve bunlar ile ilişkili verilere ulaşılmaktadır. Web sunucusu olarak görev yapan bu veri tabanına kullanıcılar tarafından etki etmek mümkün değildir. SQL sorguları ile özelleştirilebilir (http://www.ncbi.nlm.nih.gov/books/NBK21088/, 01.01.2011). Solucan sadece belirli bir grup veriyi toplamaz, özelleştirilebilir bir yapısı bulunmaktadır. Bu şekli ile dbSNP’den ayrılmıştır. SNPper, dbSNP benzeri fakat detay bakımında fakir olan bir başka sunucudur. Sadece SNP ‘ler ile ilgili bilgileri toplamaya müsaade etmektedir. Web tabanlı bir program olup, çıktıları herhangi bir veri tabanına aktarılmaya uygun değildir (Riva ve Kohane, 2002). Solucan ise sonuçlarının veri tabanlarına aktarılmaya müsait olması, devamlı güncel tutulabilmesi, istenilen sonuçları içermesi bakımından bu uygulamadan ayrılmaktadır. SNPHunter, benzer bir şekilde lokus numarası üzerinden SNP ‘leri bulup, listeler, sonuçları sabit bir veri tabanına aktarılmış durumda olduğundan sorgulanmaya da müsaittir (Wang ve ark., 2005). Solucan bu yönü ile SNPHunter ‘a benzemektedir. Sonuçların araştırmacı tarafından dizayn edilen veri tabanlarına yazması nedeni ile SQL sorgu sonuçları daha etkili olmaktadır. Aynı zamanda dinamik yapısı ile SNPHunter ‘dan ayrılmaktadır.

dbSNP, SNPPer, SNPHunter ‘ın veri ulaşım zamanları Solucan ‘a nazaran oldukça kısadır. Solucan ‘ın çekirdek veri projesini derlemesi yaklaşık olarak 2 saat almıştır. Fare

‘nin proteinler hakkındaki özel verileri bir bütün şekilde toparlaması ise yaklaşık olarak 4.12 saat almıştır.

Fare ikincil büyük sorun olan kutsal kase için geliştirilmiş Solucan tabanına sahip bir yapay sinir ağı motoru olarak tahminleri yapmak ve modelleri geliştirmek için kullanılmıştır. Model sadece ikincil yapı tahmini üzerinden çalışmayıp evrimsel süreçten gelen verileri de hesaba katan dizi hizalama yöntemlerini de kullanmaktadır. Bu nedenle literatürdeki diğer pür yapay sinir ağı tabanlı tahmin programlarından ayrılmaktadır. Bu bağlamda Fare benzerleri arasında en çok PHD ‘ye benzemektedir. PHD önce hizalama yapmakta ve sonra ikincil yapının tahminini gerçekleştirmektedir (Rost ve Sander, 1993). Fare ise öncelikle ikincil yapıyı tahmin etmekte, sonra bu tahmini veri tabanlarında ki ikincil yapısı aydınlatılmış protein dizileri ile hizalamakta, en yüksek skorlu olanları PDB dosyası şeklinde düzenleyerek AstexViewer yardımı ile tahmini üç boyutlu yapıyı göstermektedir.

Örneklerden pyrinin tahmini yapısı ile gerçek yapısı arasındaki benzerlikler oldukça dikkat çekicidir. Bu benzerlikler Şekil 41 ‘da gösterilmiştir.



Şekil 41: Pyrinin tahmini yapısı ile NMR görüntüsü arasındaki benzerlikler.

Fare ile elde edilen Şekil 41 ‘da Alan 1 ile Alan 3 arasında heliks segment olması nedeni ile benzerlik bulunmaktadır. Alan 2 ile Alan 4 arasında ise her iki alanında beta barel olarak tanımlanan (dönel bir şekilde bir çap üstüne sıralanmış antiparalel beta tabaka yapıları) süper ikincil yapı elementi oluşturması açısından benzerlik bulunmaktadır. Şekil 41 ‘de sol tarafta görülen yapı tahmin edilen yapıyı, sağdaki ise orijinal yapıyı temsil eden resimlerdir.

Gende meydana gelen bir SNP üçüncül yapıya kadar intikal edemeyebilir. Bunun olası nedenleri arasında DNA ‘da meydana gelen değişimin mRNA ömrünü azaltması, mRNA ‘nın kendi üzerine katlanmasına neden olması, etki sahasında ki ikincil yapı elementinin yapısını bozacak kadar etkili bir değişim meydana getirmemesi olabilir. Fare ile MEFV genindeki SNP ‘ler denenmiş olup etkili bir üç boyut değişimine neden olmadığı da görülmüştür. Bu bağlamda tüm genlerin ve bun genlerdeki tüm SNP noktalarının tek tek ve kombinatör (haplotipler halinde) bir şekilde Fare ile tekrar modellenmesinin faydası olabilir.

6. SONUÇ

İkincil yapı tahmini yaptığımız bu çalışmamızda elde ettiğimiz sonuçlar doğrultusunda, yapıda meydana getirilen değişiklikler ve bu değişikliklerin etkileri gösterilmiştir. SOLUCAN ve FARE yazılımlarının yardımı ile gerekli veriler toplanmış, geliştirdiğimiz araçlar ile yorumlanmıştır. Solucan ‘ın istenilen verileri istenilen şekilde toplaması ve kullanılmak üzere kendi tasarladığımız veri tabanlarına yazması ve bunları güncel tutması sağlanmıştır. Fare ‘nin ikincil yapıyı tahmin etmesi, hizalama sonuçlarına göre üçüncül yapıyı tahmin ederek görsel bir şekilde sonuçları ifade etmesi sağlanmıştır.

Bu tez konusunu oluşturan 3D yapı tahmin yaklaşımı, bildiğimiz kadarıyla, ülkemizde ilk kez 2005 yılında bir Yüksek Lisans tez çalışması olarak denenmiştir (Kurt, 2005). Bu bağlamda ikinci örneği oluşturan bu tez çalışmasında ulaşılan sonuç (Şekil 41’de görülen pyrinin tahmini yapısı ile NMR görüntüsü arasındaki benzerlikler), yeni ve daha güçlü yazılımların yazılması ile sanal ortamda yürütülecek olan yapı-fonksiyon çalışmalarının laboratuar çalışmalarının önüne geçebileceğine olan inancımızı güçlendirmektedir.

7. KAYNAKLAR

Anfinsen CB. (1973). Principles that govern the folding of protein chains. Science **181** (96): 223–230

Arjunan SNV, Deris S, Rosli MD. (2001). Literature survey of protein secondary structure prediction. Jurnal Teknologi, **34**, 63–72.

Access. Microsoft Corporation, http://office.microsoft.com/tr-tr/ , 05.01.2011

Benson DA, Karsch-Mizrachi I, Lipman DJ, Ostell J, and Wheeler DL. (2008). Genbank. Nucleic Acids Research, **36**(Database issue), D25-30. Epub 2007 Dec 11.

Branden C, Tooze J. (1991). Introduction to Protein structure. Garland, New York.

Baxevanis AD. (2009). The importance of biological databases in biological discovery. Current Protocols in Bioinformatics, Chapter 1: Unit 1.1, **27**, 1.1.1-1.1.6. John Wiley & Sons, Inc.

Burkhard, R. (1999). Twilight Zone Of Protein Sequence Alignments. Protein Engineering. **12**(2): 85-94.

Burkhard, R. (2001). Review: Protein Secondary Structure Prediction Continues to Rise. Journal of Structural Biology. **134**(2-3):204-18.

Cai YD, Liu XJ, Xu XB, Chou KC. (2002). Artificial neural network method for predicting protein secondary structure content. Journal of Computational Chemistry, 26(4), 347-350.

Cai YD, Liu XJ, Chou KC. (2003). Prediction of protein secondary structure content by artificial neural network. Journal of Computational Chemistry, 24(6), 727-731.

Chairman: H. B. F. DIXON (UK); Secretary: A. CORNISH-BOWDEN (UK);Members: C. LIEBECQ (Belgium — as Chairman of TUB Committee of Editors of Biochemical Journals); K. L. LOENING (USA); G. P. MOSS (UK); J. REEDIJK (Netherlands); S. F. VELICK (USA); J. F. G. VLIEGENTHART (Netherlands).Additional contributors to the formulation of these recommendations: Nomenclature Committee of IUB (NC—IUB) (those additional to JCBN):H. BIELKA (GDR); N. SHARON (Israel); E. C. WEBB (Australia).P. KARLSON (FRG, Past Chairman of JCBN); B. KEIL (France, a former Member of NC-TUB); W. E. COHN (USA); J. T. EDSALL (USA); J. S. MORLEY (UK);G. T. YOUNG (UK); and Members of IUPAC Commission on Nomenclature of Organic Chemistry.IUPAC, IUB IUPAC-IUB Joint Commission on Biochemical Nomenclature (JCBN). (1984). Nomenclature and symbolism for amino acids and peptides. Recommendations 1983. Biochemical Journal, **219**, 345–373.

Chou PY, Fasman GD (1974a). "Prediction of protein conformation". Biochemistry **13** (2): 222–245.

Chou PY, Fasman GD (1974b). "Conformational parameters for amino acids in helical, beta-sheet, and random coil regions calculated from proteins". Biochemistry **13** (2): 211–222.

Chou PY, Fasman GD (1978a). "Empirical predictions of protein conformation". Annu Review Of Biochemistry **47**: 251–276.

Chou PY, Fasman GD (1978b). "Prediction of the secondary structure of proteins from their amino acid sequence". Advances in Enzymology and Related Areas of Molecular Biology **47**: 45–148.

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 PTT, Vaughan R, Zalunin V, and Birney E. (2009). Petabyte-scale innovations at the European nucleotide archive. Nucleic Acids Research, **37**(Database issue), D19–25.
Cooper, D.N., Smith, B.A., Cooke, H.J., Niemann, S., Schmidtke, J. (1985). An estimate of unique DNA sequence heterozygosity in the human genome. Human Genetics. **69**:201-205.

Cochrane GR, and Galperin MY. (2010). The 2010 Nucleic Acids Research Database Issue and online Database Collection: a community of data resources. Nucleic Acids Research, **38**(Database issue), D1-4. Epub 2009 Dec 3.

Crick F. (1966). Codon–anticodon pairing: the wobble hypothesis. Journal of Molecular Biology **19**(2), 548–555.

Custódio FL, Barbosa HJC, and Dardenne LE. (2004). Investigation of the three-dimensional lattice HP protein folding model using a genetic algorithm. Genetics and Molecular Biology, **27**(4), 611-615.

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**(16), 2883–2885.

DBDesigner 4. (2010). http://fabforce.net/dbdesigner4, 19.12.2010

DbSNP. (2010). http://www.ncbi.nlm.nih.gov/SNP/snp_summary.cgi, 22.12.2010

DbSNP. Hand Book. (2011). http://www.ncbi.nlm.nih.gov/books/NBK21088. 01.01.2011

Deelx Regular Expresşion Engine V1.2. (2010). http://www.regexlab.com/en/deelx. 24.12.2010

Ellis LB, Attwood TK. (2001). Molecular biology databases: today and tomorrow. Drug Discovery Today, **6**(10), 09-513.

Elmas Ç. (2003). Yapay sinir ağları. Seçkin Yayıncılık. 1. Baskı. Ankara.

EMBL. (2010). http://www.ebi.ac.uk/embl/Services/DBStats/, 29.12.2010

Expasy/Swiss-Prot, http://expasy.org/sprot/userman.html , 01.01.2011.

Fausett LV. (1993). Fundamentals of Neural networks. Prentice Hall, NJ, USA.

Garnier J, Gibrat J-F, Robson B. (1996). Secondary structure prediction method version IV. Methods in Enzymology, **266**, 540-553.

Garnier J, Osguthorpe DJ, Robson B. (1978). Analysis of the accuracy and implications of simple methods for predicting the secondary structure of globular proteins. Journal of Molecular Biology, **120**(1), 97-120.

GenBank. (2010). http://www.ncbi.nlm.nih.gov/Genbank/genbankstats.html

Haykin, S.(1999). Neural Networks: A Comprehensive Foundation, 2nd ed. Englewood Cliffs, NJ: Prentice-Hall.

Holley HL, Karplus M. (1988). Protein secondary structure prediction with neural network. Proceedings of the National Academy of Sciences of the USA, **86**, 152-156.

Hong P, and Wong WH. (2005). Genenotes–a novel information management software for biologists. BioMed Central Bioinformatics, **6**, 20.

Hoop TP, and Woods KR. (1981). Prediction of protein antigenic determinants from amino acid sequences. Proceedings of the National Academy of Sciences of the USA, **78,** 3824-3828.

Huang YY. (2004). Protein folding prediction with genetic algorithms. Master of Science. Department of computer science and engineering National Sun Yat-sen University

Hutchinson EG, Thornton JM. (1999). A revised set of potentials for ,turn formation in proteins. Protein Science. **3**:2207-2216

J. S. Davies, Amino acids, peptides, and proteins, Royal Society of Chemistry, 2006

Kabsch W, Sander C . (1983). Dictionary of protein secondary structure: pattern recognition of hydrogen-bonded and geometrical features. Biopolymers 22(12), 2577–637.

Kendrew JC, Bodo G, Dintzis HM, Parrish RG, Wyckoff H, Phillips DC. (1958). A three-dimensional model of the myoglobin molecule obtained by x-ray analysis. Nature, 181(4610), 662-666.

Kirov SA, Peng X,Baker E,Schmoyer D,Zhang B, Snoddy J. (2005). GeneKeyDB: A lightweight, gene-centric, relational database to support data mining environments. BioMed Central Bioinformatic. **6**:72

Klug WS, Cummings MR. (2003). Genetik Kavramlar. Altıncı Baskı, Gözden geçirilmiş 2. Türkçe Baskı. Palme Yayıncılık. Ankara.

Kyte J, and Doolittle R. (1982). A simple method for displaying the hydropathic character of a Protein. Journal of Molecular Biology, **157**, 105-132.

Lesh N, Mitzenmacher M, Whitesides S. (2003). A complete and effective move set for simplified protein folding. Proceedings of the Sventh Annual International Conference on Computational Biology, April 10-13, 2003, Berlin, Germany. ACM, pp188-195.

Levinthal C. (1969).How to Fold Graciously Mossbauer Spectroscopy in Biological Systems: Proceedings of a meeting held at Allerton House, Monticello, Illinois*: 22–24.*

Lodish H, Berk A, Matsudaira P, Kaiser CA, Krieger M, Scott MP, Zipursky SL, Darnell J. (2003). Molecular Biology of the Cell. WH Freeman, New York, NY. 5th edition, USA.

MEDLINE. (2010). http://www.nlm.nih.gov/bsd/history/tsld024.htm, 06.12.2010

Microsoft SQL Server. (2011). http://www.microsoft.com/sqlserver/, 05.01.2011

MySQL 5.1. (2010). http://www.mysql, 11.12.2010

Mottalib MA, Md. Safiur , A.B.M. Zunaid Haque, S.M. Al Mamun and Hawlader Abdullah Al-Mamun. (2010). Protein Secondary structure Prediction using Feed-Forward Neural network. Journal of computer and Information Technology, (JCIT), Manuscript code: 100713

Mutasem KSA, Khairuddin BO, Shahrul AN. (2009). Back Propagation Algorithm: The Best AlgorithmAmong the Multi-layer Perceptron Algorithm. International Journal of Computer Science and Network Security. **9**(4)

Needleman, S.B.; Wunsch, C.D. (1970).A general method applicable to the search for similarities in the amino acid sequence of two proteins. Journal of Molecular Biology **48** (3): 443–53.

Nelson D, and Cox M. (2003). Lehninger Principles of Biochemistry, 5th Edition,W.H. Freeman & Company, USA.

Oracle Database Management, Oracle Corporation., http://www.oracle.com/, 05.01.2011

Özkaçar N. (1998). Genetik algoritmalar. İ.Ü. İşletme Fakültesi Dergisi, **27**(1), 69-82.

Pace CN, Scholtz JM. (1998). A helix propensity scale based on experimental studies of peptides and proteins. Biophysics Journal, **75**(1), 422–427.

Pauling L, Corey RB. (1951). The pleated sheet, a new layer configuration of polypeptide chains. Proceedings of the National Academy of Sciences, U S A, **37**(5), 251-256.

Peng A, Baker X, Schmoyer E, Zhang D, and Snoddy BJ. (2005). Genekeydb: a lightweight, gene-centric, relational database to support data mining environments. BioMed Central Bioinformatics, **6**, 72.

Philippi S, and Kohler J. (2006). Addressing the problems with life-science databases for traditional uses and systems biology. Nature Reviews of Genetics, **7**(6), 482–488.

PubMed. (2010). http://www.ncbi.nlm.nih.gov/About/tools/restable_stat_pubmeddata.html

Qian N, Sejnowski TJ. (1998). Predicting the secondary structure of globular proteins using neural network models. Journal of Molecular Biology, **202**(4), 865-84.
Richardson JS. (1981). The Anatomy and Taxonomy of Protein Structure. Advances in Protein Chemistry, **34**, 167-339.

Riva A., Kohane S.I. (2002). SNPper: retrieval and analysis of human SNPs. Bioinformatics, Vol. **18** no. 12 2002, 1681–1685.

Rosenblatt, Frank (1957), The Perceptron: A perceiving and recognizing automaton. Report **85-460-1**, Cornell Aeronautical Laboratory.

Rost B, Sander C. (1993). Prediction of protein secondary structure at better than 70% accuracy. Journal of Molecular Biology, **232**(2), 584-599.

Rost B, Casadio R, Fariselli P, Sander C. (1995). Transmembrane helices predicted at %95 accuracy. Protein Science, **4**, 521-533.

Sengupta D, Behera NR, Smith JC, Ullmann GN. (2005). The α Helix Dipole: Screened Out. Structure, **13**(6), 849-855.

Shah SP, Huang Y, Xu T, Yuen MM, Ling J, and Ouellette BF. (2005). Atlas - a data warehouse for integrative bioinformatics. BioMed Central Bioinformatics, **6**, 34.

Smith, Temple F.; and Waterman, Michael S. (1981). Identification of Common Molecular Subsequences. *Journal of Molecular Biology* **147**: 195–197.

Swiss Institute of Bioinformatics Expert Protein Analysis System. (2011). http://expasy.org.

Taylor WR. (1986). The classification of amino acid conservation. Journal of Theoretical Biology, **119**(2), 205-218.

Unger R. (2004). The Genetic Algorithm Approach to Protein Structure Prediction. Structure & Bonding, **110**, 2697-2699.

Unger R, Moult J. (1993). Genetic algorithm for protein folding simulation. Journal of Molecular Biology, **231**, 75-81.

Vignal A., Milan D., SanCristobal M, Eggen A. (2002). A review on SNP and other types of molecular markers and their use in animal genetics.Genetics Selection Evolution May-Jun;**34**(3):275-305.

Volkan K. (2005). “Protein Structure Prediction using Decision Lists”, Yüksek Lisans Tezi. Mühendislik ve Bilgisayar Bilimleri Bölümü. Koç Üniversitesi, İstanbul

Walsh CT, Tsodikova SG, Gatto GJ. (2005). Protein Posttranslational Modifications: The Chemistry of Proteome Diversifications. Angewandte Chemie International Edition, **44**, 7342 – 7372.

Wang L., Liu S, Niu T, Xu X. (2005). SNPHunter: a bioinformatic software for single nucleotide polymorphism data acquisition and management. BioMed Central Bioinformatics, **6**, 60.

Wikipedia The free encyclopedia. (2010). http://en.wikipedia.org/wiki/Amino_acid

Yadav BS, Pokhariyal M, Ratta B, Rai G, Saxena M. (2010). Predicting Secondary Structure of Oxidoreductase Protein Family Using Bayesian Regularization Feed-forward Backpropagation ANN Technique. Journal of Proteomics and Bioinformatics, **3**, 179-182.

Yılmaz Ö. (2003). Two stage mathematical programming algorithm for predicting secondary structure of protein. Yüksek Lisans Tezi. (Danışmanlar: Metin Türkay, Selçuk Savaş). Endüstri Mühendisliği ve Operasyonlar Yönetimi'nde Lisansüstü Programlar. Koç Üniversitesi, İstanbul.

Zwanzig R, Szabo A, Bagchi B. (1992). Levinthal's paradox. Proceedings of National Academy of Sciences, USA, 89, 20-22.

ÖZGEÇMİŞ

Adı ve soyadı: Muhammed Kamil TURAN

Doğum yeri: Safranbolu / Karabük

Doğum tarihi: 1976

Medeni durumu: Evli

Yabancı dili: İngilizce

Elektronik posta: mkamilturan@gmail.com

Eğitim Durumu:

Zati Ağar ilköğretim okulu 1982,

Safranbolu Orta Okulu 1987,

Karabük Demir Çelik Lisesi 1990,

OMU Tıp Fakültesi 2000.

Görev Yeri: Ünye Özel Çakırtepe Hastanesi Acil Servis

Görevi: Doktor

Doktora: 2003 OMU tıp fakültesi Tıbbi Biyoloji AD.

Tez konusu: Dizisi bilinen proteinlerin 3D benzetim modellerinin kurulması ve modelin etkili SNP'lerin saptanmasında kullanılması