

Proteinlerin İkincil Yapılarının Tahmininde Heuristik Yöntemler

^{*1}Burcu Çarklı Yavuz, ¹Nilüfer Yurtay, ²Hikmet Benan Dinçtürk Botofte ve ¹Melike Şeyda Arslan
^{*1}Bilgisayar ve Bilişim Bilimleri Fakültesi, Bilgisayar Mühendisliği Bölümü, Sakarya Üniversitesi, Türkiye
²Fen- Edebiyat Fakültesi, Biyoloji Bölümü, Sakarya Üniversitesi, Türkiye

Özet

Proteinler canlı organizmalar için hayati önem taşırlar. Proteinlerin üç boyutlu yapılarının bilinmesi proteinin işlevini anlamak açısından önemlidir. Günümüzde proteinlerin birincil yapıları, yani amino asit dizilimleri genetik kodun evrensel olması nedeniyle gen klonlama çalışmaları veya N-terminal protein dizi analizi yöntemleriyle kolaylıkla bulunabilir. Ancak ikincil, üçüncül ve dördüncül yapıların belirlenmesi için bunu söylemek mümkün değildir. X-ışını kırınımı gibi kristalografik yöntemlerle her proteinde başarı sağlanamadığı ve bu çalışmalar uzun süre gerektirdiğinden birbirleriyle ilişkili olan bu yapılar üzerinde birincil yapı verileri kullanarak heuristik yöntemlerle tahmin çalışmaları yaygındır.

Bu çalışmada makine öğrenmesi, destek vektör makineleri, yapay sinir ağları, karar ağaçları, genetik algoritmalar ve benzeri heuristik yöntemlerin protein ikincil yapı tahmininde kullanımı ve başarı oranları ayrıntılı olarak incelenmiştir. Bakış açısını genişleterek protein ikincil yapı tahmininde kullanılabilecek daha farklı yöntemlerin belirlenebilmesi ve tahmin başarısının maksimum düzeye çıkartabilmesi için yapılacak bir projenin başlangıcı olması nedeniyle çalışmamız önem kazanmaktadır.

Anahtar Kelimeler: Protein İkincil Yapı Tahmini, Yapay Sinir Ağları, Karar Ağaçları, Genetik Algoritmalar, Destek Vektör Makinaları

1. Giriş

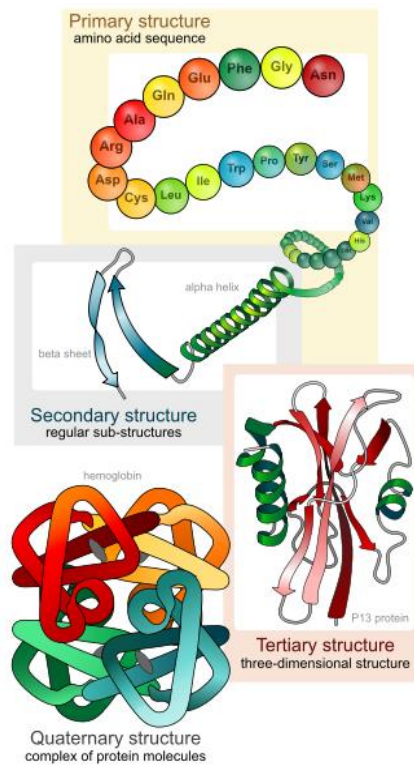
Herhangi bir hücreye mikroskop ile bakıldığında, ya da bu hücrenin elektriksel ya da biyokimyasal davranışları incelendiğinde asıl olarak incelenmekte ve gözlemlenmekte olan yapı proteindir. Hücrelerin kuru kütlelerinin neredeyse tamamını oluşturan proteinler, hücrelerin yapı taşları olmakla kalmayıp, hücre içerisindeki faaliyetlerin birçoğunu yerine getiren elemanlardır. Bu nedenle vücuttaki kaslar, üreme sistemleri v.b. tüm bileşenlerin faaliyetlerini kavrayabilmek, nasıl çalıştıklarını çözebilmek için proteinlerin derinlemesine anlaşılabilmesi gerekmektedir [1].

Yunanca kökenli bir sözcük olan protein asıl olarak “ilk element” anlamına gelmektedir [2]. Yukarıda anlatıldığı gibi proteinler hücre içindeki tüm biyolojik faaliyetlerde etkin bir çalışma göstermekte olan bilinen en karmaşık ve en işlevli makro moleküllerdir. Proteinlerde her biri farklı kimyasal özellikte yaklaşık 20 çeşit amino asit bulunmaktadır. Protein molekülleri her biri

*Sorumlu yazar: Adres: Sakarya Üniversitesi Bilgisayar ve Bilişim Bilimleri Fakültesi Bilgisayar Müh. Bölümü 54187, Sakarya TÜRKİYE. E-posta: bcarkli@sakarya.edu.tr, Telefon: +902642955009 Fax: +902642955031

komşusuna kovalan peptid bağı ile bağlı olan uzun bir aminoasit zincirinden oluşur. Proteinin yapısı belirlenirken Hidrojen bağları, iyonik bağlar, van der Waals bağları gibi bağlar da önemli rol oynarlar. Bu bağların varlığı veya yokluğu, proteinlerin katlanma şekil ve formasyonları proteinlerin yapılarını belirler. Biyologlar tarafından bir protein dizisi dört düzeyde tanımlanmaktadır [1]:

- 1) Proteinlerin amino asit dizisi; birincil yapı(primer)
- 2) α sarmal ve β levhaları oluşturan uzun polipeptid zincirleri; ikincil yapı(sekonder)
- 3) Polipeptid zincirinin tüm üç boyutlu organizasyonları; üçüncül yapı(tersiye)
- 4) Birden fazla polipeptid zincirinden oluşmuş protein yapı; dördüncül yapı(kuarterner)



Şekil 1. Proteinin birincil, ikincil, üçüncül ve dördüncül yapısı [3]

Protein dizilerinde genellikle 2 farklı katlanma örüntüsü mevcuttur: α sarmal ve β levhaları. α sarmal ilk olarak α keratin proteininde ve β levhaları ise ilk olarak fibroin proteininde bulunmuş olan katlanma örüntüleridir. α sarmal ve β levhalarının farklı birleşimlerinin her biri protein katlama biçimi olarak bilinirler. Günümüze kadar 1000'den farklı katlanma biçimi keşfedilmiştir [1]. İşte bu farklı katlanma ve amino asit organizasyonları protein yapılarını belirleyen etmenler olarak öne çıkmaktadır.

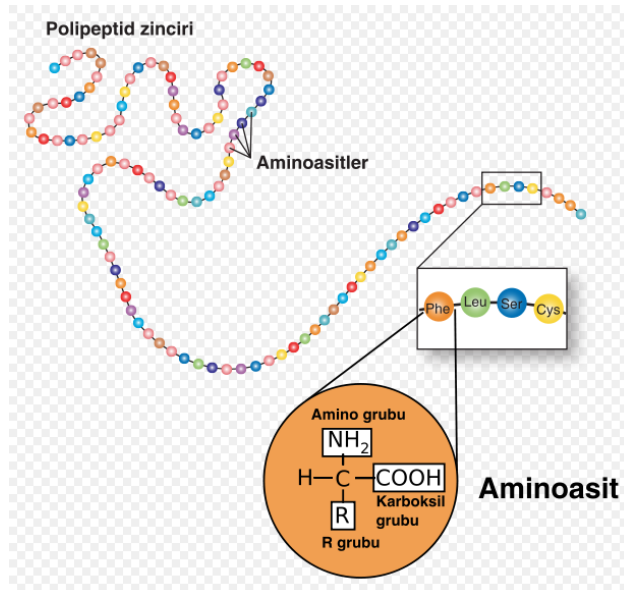
Günümüzde birincilden dördüncül yapıya kadar olan farklı protein yapılarını tahmin edebilmek, dolayısıyla genetik bozukluklara, hastalıklara müdahale edebilmek ve tedaviler geliştirebilmek

için bilgisayar yazılımları geliştirilerek amino asit dizilerinin bilinen protein katlanma biçimleri ile ilişkilendirilmesine olanak sağlanmıştır. Biyokimya ve bilgisayar bilimleri biyoinformatik konusu altında birleşerek bu konuda çalışır hale gelmiştir. Protein yapılarının tahmininde özellikle doğa esintili hesaplama yöntemleri (heuristik yöntemler) sıklıkla kullanılmıştır.

Proteinlerin birincil yapıları, yani amino asit dizimleri genetik kodun evrensel olması nedeniyle gen klonlama çalışmaları veya N-terminal protein dizi analizi yöntemleriyle kolaylıkla bulunabilir. Ancak ikincil, üçüncül ve dördüncül yapıların belirlenmesi için bunu söylemek mümkün değildir. X-ışını kırınımı gibi kristalografik yöntemlerle her proteinde başarı sağlanamadığı ve bu çalışmalar uzun süre gerektirdiğinden birbirleriyle ilişkili olan bu yapılar üzerinde birincil yapı verileri kullanarak heuristik yöntemlerle tahmin çalışmaları yaygındır.

Bu çalışmada makine öğrenmesi, destek vektör makineleri, yapay sinir ağları, karar ağaçları, genetik algoritmalar ve benzeri heuristik yöntemlerin protein ikincil yapı tahmininde kullanımı ve başarı oranları ayrıntılı olarak incelenmiştir. Bakış açısını genişleterek protein ikincil yapı tahmininde kullanılabilecek daha farklı yöntemlerin belirlenebilmesi ve tahmin başarısının maksimum düzeye çıkartabilmesi için yapılacak bir projenin başlangıcı olması nedeniyle çalışmamız önem kazanmaktadır.

2. Aminoasitler ve Protein Oluşumu



Şekil 2. Aminoasit yapısı ve polipeptid zinciri [4]

Bir protein molekülü, her biri komşusuna kovalan peptid bağı ile bağlı uzun bir aminoasit zincirinden oluşur. Yani, proteinler = polipeptidler denilebilir. Her aminoasit yapısında Karbon atomuna bağlı amino grubu, karboksil grubu ve hidrojen atomu sabittir. Yan zincir değişerek

farklı aminoasitlerin türemesini sağlar. Proteinlerin oluşumunda 20 farklı amino asit rol alır. Bu amino asitlerin farklı sıra ve uzunlukta bir araya gelmeleriyle farklı proteinler meydana gelir. Proteinler 20 harfli bir alfabe ile oluşturulan kelimeler şeklinde düşünülebilir. [5]

3. Protein İkincil Yapı Tahmini

Protein ikincil yapı tahmini yıllardır üzerinde çalışılmakta olan bir konudur. 90 lı yıllarda protein ikincil yapılarının %60 civarı doğrulukla tahmini çok başarılı kabul edilirken günümüzde %90 lara varan doğrulukta tahmin oranları başarılı mertebesinde sayılmaktadır.

Bu bölümde son yıllarda protein ikincil yapı tahmini üzerine yapılan çalışmalar irdelenmiş ve bu çalışmaların bazıları Tablo 1’de sunulmuştur. Bu çalışmalarda kullanılan yöntemler incelendiğinde dokuzunda destek vektör makinesi, üçünde karar ağacı, beşinde yapay sinir ağı, ikisinde ise genetik algoritma yöntemlerinin ağırlıklı olarak kullanıldığı görülmektedir.

Tablo 1. Protein İkincil Yapı Tahmin Çalışmalarına Genel Bakış

Referans	Anahtar Kelimeler
[6]	Protein ikincil yapı tahmini, local protein yapısı, destek vektör makinesi
[7]	Protein ikincil yapı tahmini (PSSP), Yapay sinir ağı (ANN), α -helix, β -sheet, coil, Pozisyona özgü puanlama matrisi (PSSM)
[8]	Protein ikincil yapı tahmini, Makine öğrenmesi yaklaşımı, Destek vektör regresyonu, Çok amaçlı genetik algoritma
[9]	Protein ikincil yapı tahmini, Makine öğrenmesi yaklaşımı, Destek vector makineleri, Paralel hiyerarşik grid arama, Ağırlıklı çekirdek fonksiyonu
[10]	Fonksiyon tahmini, Protein etkileşimleri, Choquet-Integral
[11]	Protein ikincil yapı tahmini, Fizikokimyasal özellikler, Mixed-modal SVM, Bileşik piramit model
[12]	Protein ikincil yapı tahmini, Bilgi keşfi, Bileşik piramit model
[13]	Protein ikincil yapı tahmini, Fizikokimyasal özellikler, Multi-modal BP, Bileşik piramit model
[14]	Derinlik öncelikli arama, Protein yapı tahmini, Genetik algoritma, HP lattice, HCP modeli
[15]	Yapay bağışıklık sistemi, Protein lokalizasyonu, k en yakın komşu sınıflayıcısı, E. coli veri tabanı
[16]	İkincil yapı, tekli dizilim, Yapay sinir ağı, Saklı markov modeli, Grafiksel modeller, Tahmin, Melezleme
[17]	Karar ağacı, Protein yapısı, Kural çıkarma, Destek vektör makinesi (SVM)
[18]	Çoklu sınıflayıcı, Denetimli sınıflandırma, Makina öğrenmesi, Yığılmış genelleme, Bayes ağı, Protein ikincil yapı tahmini, Pazzani-EDA
[19]	Protein ikincil yapı tahmini, SVM, Kodlama düzeni, CB513

Yin Fai ve arkadaşlarının 2012’de yapmış oldukları araştırmanın [6] amacı RS126 dataseti üzerinde makine öğrenmesi algoritmaları kullanılarak protein ikincil yapı tahmini yapmaktır. Bu çalışmada uygulanan makine öğrenmesi algoritması destek Vektör Makinesidir (Support Vector Machine – SVM). Bu çalışmada SVM kullanılmasının nedeni ikili karar için çok güçlü bir algoritma olarak bilinmesidir. RS126 makine öğrenmesi algoritmalarıyla eğitim ve test yapmak için önemli bir datasettir, çünkü içerdiği 126 proteinin, 80 rezidüden daha uzun olanlarında %25 den daha fazla benzerlik vardır. Benzerlik oranı küçük bir seviyede olmasına rağmen gerçek dünya ortamına oldukça yakın bir protein dizimi mevcuttur ve protein ikincil yapı tahmini için oldukça uygun bir veri setidir.

DSSP (Dictionary of Secondary Structure of Proteins) programı ikincil yapıyı standardize etmek için Kabsch ve Sander tarafından tasarlanmıştır. DSSP, protein veri bankasındaki tüm proteinler için standart bir ikincil yapı sunan veri tabanıdır, hidrojen bağlarının modeline bağlı olarak protein ikincil yapısında 8 türe sahiptir; H,B,E,G,I,T,S,L ve bu 8 tür genelde 3 ana sınıfta toplanır; H (heliks yada sarmal), E (sheet yada tabaka), C (coil yada iplik).

10 kat çapraz doğrulama kullanılarak veri seti 10 parçaya bölünmüştür. Çapraz doğrulama 10 kez tekrarlanmıştır, elde edilen tüm sonuçlar tahmin için kullanılmıştır. RBF kernel uygulanmıştır. Sonuçta bu hibrit algoritmanın konvansiyonel metotlara göre daha iyi bir tahmin yüzdesi (%70) sunduğu ortaya çıkmıştır.

Alirezade ve arkadaşlarının 2012 tarihli çalışmalarında [7] yazarlar protein ikincil yapı tahmininde bulunabilmek için iki sıralı ileri beslemeli yapay sinir ağlarını eğiterek çalışmalarda bulunmuşlardır. Birinci YSA sonuçları göstermektedir ki amino asitlerin ikincil yapılarının yakın sınırı için tahmin için bir yapıdan diğerine geçiş yapmak genellikle başarısız olur, bu yüzden ilk ağ çıkışlarının gözden geçirilmesi gerekmektedir. Ayrıca tahmin sonuçlarını geliştirmek için başka bir ağ kullanmak da tahmin doğruluğuna katkı verir. Bu çalışmada kullanılan veri kümesi Rost ve Sander tarafından önerilmiş RS126 veri setidir. RS126 verisinde yedi kat çapraz doğrulama kullanılmıştır. Bu metotta veri setleri 7 eşit parçaya A,B,...,G ayrılmıştır ve belirlenen eğitim dışında kalan ve test kümesi sırayla kullanılmıştır. İlk ağ çıkışları tahminleri %73,28, kademeli olarak başka bir ağ tarafından çıkışlar düzeltildikten sonra kullanıldığında ise % 75.22 oranında doğruluk göstermiştir. Yani, ikinci bir ağ kullanımıyla tahmin oranlarında yaklaşık % 2 düzelme sağlanabilmiştir.

Zangoeei ve Jalili çalışmalarında [8] Gerçek değerleri tam sayıya yuvarlamak için haritalama noktaları bulmak maksadıyla baskın olmayan sıralamalı genetik algoritma yöntemi (NSGAIL) uygulamışlardır. Ayrıca NSGAIL, SVR kernel parametrelerini bulup ayarlayarak performansı arttırmak ve daha iyi sonuçlar elde etmek için de kullanılmıştır. RS126 ve CB513 datasetleri üzerinde %85.79 ve %84.94 başarılı protein ikincil yapı tahmini elde edilmiştir.

Zangoei ve Jalili çalışmalarında [9] yeni bir paralel çok sınıf (PMC) yöntemi, paralel hiyerarşik grid arama (PHGS), çapraz doğrulama (CV) tekniği ve ağırlıklı çekirdek füzyon (WKF) yöntemini birleştirerek hibrit bir yöntem önermişlerdir. Sunulan paralel hiyerarşik grid arama yöntemi (PHGS), doğrulukta önemli bir etkisi olan SVM nin çekirdek fonksiyonunu düzenlemek için kullanılmıştır. Ayrıca yöntemi geliştirmek için, Pozisyon skorlama matrisi (PSSM) profilleri giriş bilgisi olarak kullanılmıştır. Bu çalışmanın amaçları proteinin ikincil yapı sınıflarını doğru belirlemek için kernel fonksiyonu parametrelerini ayarlamak ve farklı çekirdek fonksiyonlarının sonuçlarını birleştirmektir. Birleştirme yönteminin doğru seçilmesi üstün bir performans için önemli bir konudur, dolayısıyla yazarlar lineer olmayan analiz sistemi tabanlı dinamik ağırlık tahsisi yöntemini önermişlerdir. Bu yöntem sonucu elde edilen çıktılar RS126 ve CB513 datasetleri için sırasıyla %84.65 ve %83.94 tür.

Hou ve Chi protein ikincil yapı tahmini için İteratif Fonksiyonel Toplama Algoritması (IFAA) olarak adlandırdıkları yöntemi kullanmışlardır [10]. Deneyleri için yazarlar S. Cerevisiae protein etkileşimi veri kümesini kullanmışlardır. Bu veri kümesi 232.239 protein etkileşimini içermektedir. Yazarlar değerlendirmelerinde sadece fiziksel etkileşimleri kullanmıştır. Yöntemin etkinliğini değerlendirmek için, iki tahmin yöntemi ile IFAA yöntemi karşılaştırılmıştır. IFAA konvansiyonel yöntemlere göre yaklaşık %40 daha başarılı tahminde bulunmuştur.

Yang ve arkadaşları hibrit yöntemlerin yüksek doğruluk yüzdeleri tahminler yapabiliyor olmalarına karşın kararlı olmadıklarını ve farklı veri setlerinde her zaman beklenen tahmin başarısını elde edemediklerini iddia etmişlerdir [11,12]. Yazarlar Compound Pyramid Modeli (CPM) ile en başarılı tahminlerin yapılabildiğini ve bu modeli kullanarak çalışmalar yaptıklarını belirtmişlerdir. RS126, CB513 ve CASP8 veri setleri çalışmalar için seçilmiş; RS126 için %83.99, CB513 için %85.58, CASP8 için %72.13 tahmin başarısı oranlarına ulaşılmıştır.

Qu ve arkadaşları 2011 tarihli çalışmalarında [13] benzer şekilde iyi bir veri madenciliği ve makine öğrenmesi yöntemi olan CPM'i (Compound Pyramid Model) protein ikincil yapı tahmininde MMBP (mixed-modal BP), MMS (mixed-modal SVM), SAC (structural association classifier) ve PAC(Property association classifier) ile entegreli olarak kullanmış, sonuç olarak %86.13 mertebesinde doğruluk oranlarına ulaşmışlardır.

Hoque ve arkadaşlarının çalışmasında [14] iyi bir çaprazlama yöntemi olan Genetik Algoritma (GA) kullanılmıştır. Derin Öncelikli Arama Yöntemi (DFS) tüm olasılıkları derinlemesine araştırdığı ve çaprazlamalardaki tüm problemleri eşleşmeleri ortaya kolayca çıkarabildiği GA altında tercih edilen yöntem olmuştur. 2 ve 3 boyutlu simülasyonlarda DFS test edilmiş ve hem doğruluk oranı hem de tahmin yüzdesi yüksekliği bakımından diğer GA yöntemlerine göre öne çıkmıştır.

Türkoğlu ve Kaymaz'ın çalışmalarında [15] daha iyi protein ikincil yapı tanıma doğruluğu için k-en yakın komşu (k-NN), ikili karar ağaçları ve Bayes sınıflandırıcı gibi çeşitli sınıflandırma algoritmaları kullanılmıştır. K-NN bu yöntemlere kıyasla %86'lık en iyi tahmin doğruluğunu elde etmiştir. Öte yandan K-NN'den çıkan verilerin olasılıksal yapay sinir ağlarında yeniden değerlendirilmesi ile tahmin doğruluk yüzdesinin %90lar seviyelerine ulaştığı görülmüştür. Bu da sonuç olarak protein ikincil yapı tahmininde hibrit yöntemlerin başarı seviyesini önemli ölçüde arttırdığının başka bir kanıtıdır.

Bidargaddi ve arkadaşlarının çalışmasında [16] primer diziden ikincil yapı tahmini için iki farklı perspektif, generatif modeli perspektif (Bayesian ağları) ve sınıflandırma perspektifini (yapay sinir ağları) birleştirilip melez ve modüler mimarinin faydaları araştırılmıştır. Yazarlar evrimsel bilgileri kullanmadan aminoasitlerin fiziksel-kimyasal ve yapısal özellikleri ile birlikte ikincil yapı için, Markov Modeli ve sinir ağlarına dayalı iki tabakalı bir model oluşturmaktadır. Önerilen model, sinir ağları ile protein ikincil yapı tahmini (BSPSS-NN1) için Bayesian sınıflandırılması ve generatif modellerin bir birleşimidir. Bu yeni hibrit mimarisi ile mevcut tek sıra tahmin yöntemleri karşılaştırıldığında PDB_SELECT verisetinde % 69'un üzerinde bir doğruluk değerine ulaşılır. Doğruluktaki %1 gelişim sinir ağlarının ek girişleri gibi fiziksel kimyasal özellikleri ile başarılmıştır. Ayrıca CB513 veri setinde ise %72.9 seviyeleri görülür.

He ve arkadaşlarının çalışmasında [17] genel anlamda iyi anlatılıp ifade edilmediği için çok fazla anlaşılmayan Destek Vektör Makinesi karar ağaçları ile birlikte kullanılarak protein ikincil yapı tahmini yapılmaya çalışılmıştır. SVM_DT adı verilen yeni bir algoritma ile SVM ile karar ağaçları birleştirilmiştir. Bu algoritma önce SVM ile eğitmektedir. SVM'nin çıktılarından dikkatle seçilen sonuçlar ise karar ağaçlarının eğitimi için bir set olarak kurgulanmaktadır. RS126 veri seti üzerinde yapılan çalışmalar SVM_DT ile yapılan eğitimlerin SVM'den daha doğru tahminler yapmasına olanak sağladığını göstermiştir. SVM_DT'nin genelleştirilebilme özelliği hem C4.5 karar ağaçlarından hem de SVM'den daha iyidir. Bu nedenle SVM_DT sadece protein tahmininde değil biyolojik deneylerde de kullanılabilecek bir algoritmadır denebilir.

Robles ve diğerleri çalışmalarında [18] internet üzerinden ulaşılabilen en iyi sınıflandırıcıların tahminlerini bir araya getiren birkaç yeni sınıflandırıcı sunmaktadırlar. Makale sonuçları birkaç sınıflandırıcının sonuçlarını birleştirerek yapılan tahminlerin daha verimli olduğunu göstermiştir. Bu çoklu sınıflandırıcılar Bayesian ağlarını kullanmaktadırlar (naive Bayes, Interval Estimation Naive Bayes (IENB) and the idea of Pazzani of joining attributes in naive Bayes). 9 ayrı veri seti ile birlikte değerlendirilmişlerdir. Bunlar HS1771, CB513, RS126, EVA1,...EVA6 dır. Tahmin başarı oranları mevcut en iyi protein ikincil yapı tahmin sistemlerinin ortalama %1.21 daha üzerinde çıkmıştır. En iyi sonuçları Pazzani-EDA isimli sınıflandırıcılarıyla almışlardır. Java tabanlı bir sistemde, internetten aldıkları verilerle, daha önce tahminleri yapılmış olan yapıların tekrar tahminini yaparak sistemlerini doğrulamışlardır. Yazarlar bu çalışmada klasik yöntemlerin

aksine 7 farklı seviye-1 sınıflandırıcı kullanmışlardır. 7 farklı seviye-1 sınıflandırıcı 0. Seviyede kullanıp bunların sonuçlarını seviye-1'de bütün olarak değerlendirmiş ve daha iyi bir sonuç elde etmeyi başarmışlardır. Sistemlerini test etmek için 9 farklı veri seti kullanmışlardır. Internal Estimation Naive Bayes (IENB) yaklaşımı en iyi sonuçları vermiştir.

Çalışma sonunda üzerinde çalışılması gereken konuları ise şöyle sınıflandırmışlardır:

- İkinci seviye stratejileri için yeni sınıflandırma yöntemlerinin değerlendirilmesi
- Meta-sınıflandırıcıların açık erişimli bir web server olarak yayınlanması
- Protein ikincil yapı tahmini için var olan mevcut serverlara erişim sağlayacak bir portal oluşturulması. Bu sayede kullanıcılar tek bir noktadan bir çok veriye daha kolay ulaşabilirler.

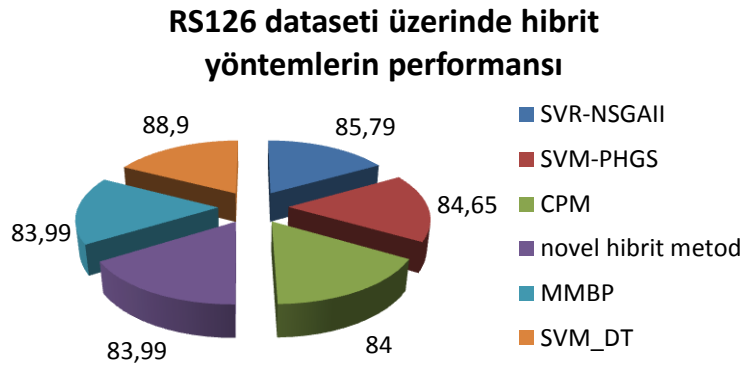
Wang ve arkadaşlarının çalışmasında [19] Destek Vektör Makinesi yöntemi ve RS 126, CB513 veri setleri kullanılmıştır. Bu çalışmada eski çalışmalardan farklı olarak, fiziksel-kimyasal özelliklere ve proteinlerin yapısal özelliklerine bağlı olan yeni bir kodlama sistemi gündeme getirilmiştir. Çalışmada kullanılan program DSSP'dir. DSSP'de sekiz ikincil yapı sınıfı mevcuttur. 5 farklı metodla bu 8 sınıfı 3'e indirmek mümkündür. Bu çalışmada 2. Metod kullanılmıştır. Yani H(α -helix), G(310-helix)'i H'ye, E(β -strand), B(isolated β -bridge)'yi E'ye geri kalanları da I(π -helix), T(turn), S(bend) ve -(rest)) C'ye indirgeme metodu. Tatmin edici sonuçlar alabilmek için kullanılan veri setleri 7 katlamalı çapraz doğrulama teste tabi tutulmuştur. Bu yöntemde her veri seti benzer sayı ve yapıda 7 katlamaya bölünmüştür. Her seferinde bir set teste tabi tutulurken diğerleri eğitim seti olmuştur. Bu iterasyon 7 kez tekrarlanmıştır. Sonuç olarak binary sınıflandırıcıların tahminde çok verimli olduğu ancak tahmin veriminin artırılması için çok daha fazla aminoasit verisine ihtiyaç duyulduğu, bu veri sayısı arttıkça doğruluk yüzdesinin arttığı görülmüştür. Ancak aminoasit verisi arttıkça uzay karmaşıklığı da artma gösterdiğinden aşırı fazla veri kullanılmasının da SVM'de sıkıntı yaratabileceği söylenmiştir.

Dowe ve arkadaşlarının 90lı yıllarda hem elle hem de yazılımlarla gerçekleştirdikleri protein ikincil yapı tahminlerinde [20] Stember, GOR, Golem, Chou, Fassman, Lim ve Fust gibi daha eski yöntemler kullanılmış en çok %60lar civarında doğruluk yüzdelere ulaşılabilmiştir.

4. Sonuçlar

Yapılan yakın tarihli çalışmalar dikkatle incelendiğinde farklı yöntemlerin birlikte kullanılmasıyla başarı oranlarını arttığını görülmektedir. Hibrit yöntemler adı verilebilecek bu yöntemler temelde bir yöntemle yapılan tahminin diğer yöntemle tekrar değerlendirilmesi sonucu doğruluk oranını arttırmayı amaçlamaktadır. Gerek yapay sinir ağları gerekse vektör destek makinaları gibi heuristik yöntemler kullanmak doğruluk yüzdeleri yükseltirken, bu yöntemlere

eklenecek yapay bağışıklık uygulamaları %90lar seviyelerine gelmiş olan protein ikincil yapı tahmin yüzdelerini daha da yukarı mertebelere çekebilecektir.



Şekil 3. Bazı hibrit yöntemlerin RS126 veri seti üzerindeki performansları

Protein ikincil yapı tahmini üzerine yapılan çalışmaların tamamına yakınında RS126 ve CB513 veri setlerinin kullanıldığı görülmüştür. Bu veri setlerinin kullanımı performansların karşılaştırılmasını kolaylaştırmakla birlikte farklı proteinler üzerinde bu algoritmaların nasıl sonuçlar vereceği muammadır. İleriki çalışmalarda, RS126 ve CB513 verisetleri üzerinde başarılı sonuçlar elde edilmiş olan destek vektör makinesi, karar ağaçları, yapay sinir ağları yöntemlerinin yeni proteinler üzerinde nasıl tahmin başarısı yakalayacağı üzerine araştırmalar yapılmalıdır.

Kaynaklar

- [1] Bruce Alberts v.d., Hücrenin Moleküler Biyolojisi, Garland Science, Taylor and Francis, 2008.
- [2] Nedret KILIÇ, Lehninger Biyokimyanın İlkeleri, Palme Yayıncılık, 2005.
- [3] http://tr.wikipedia.org/wiki/Dosya:Main_protein_structure_levels_en.svg, Erişim tarihi 16.04.2013.
- [4] <http://tr.wikipedia.org/w/index.php?title=Dosya:Aminoasit.svg&filetimestamp=20121005153945>, Erişim tarihi 16.04.2013.
- [5] Turan Aksel, SuPred:Yapay sinir ağları ve saklı markov model kullanarak protein ikincil yapı tahmin yöntemi, 2006 IEEE.
- [6] Chin Yin Fai v.d., Protein Secondary Structure Prediction Using Optimal Local Protein Structure and Support Vector Machine, International Journal of Bio-science and Bio-Technology, 4(2), pp.35-44, 2012.
- [7] Maryam Alirezaee v.d., Predicting the secondary structure of proteins by cascading neural networks, International Journal of Artificial Intelligence & Applications, 3(6), pp.39-47, 2012.
- [8] Mohammad Hossein Zangooei , Saeed Jalili, Protein secondary structure prediction using DWKF based on SVR-NSGAI, Neurocomputing, 94, pp.87-101, 2012.

- [9] Mohammad Hossein Zangooei, Saeed Jalili, PSSP with dynamic weighted kernel fusion based on SVM-PHGS, *Knowledge Based Systems*, 27, pp.424-442, 2012.
- [10] Jingyu Hou, Xiaoxiao Chi, Predicting protein functions from PPI networks using functional aggregation, *Mathematical, Biosciences*, 240, pp.63-69, 2012.
- [11] Bingru Yang v.d., Predicting protein secondary structure using a mixed-modal SVM method in a compound pyramid model, *Knowledge-Based Systems*, 24, pp.304-313, 2011.
- [12] Bingru Yang v.d., Predicting protein second structure using a novel hybrid method, *Expert Systems with Applications*, 38, pp. 11657-11664, 2011.
- [13] Wu Qu v.d., Improving protein secondary structure prediction using a multi-modal BP method, *Computers in Biology and Medicine*, 41, 946-959, 2011.
- [14] Md Tamjidul Hoque v.d., DFS-generated pathways in GA crossover for protein structure prediction, *Neurocomputing*, 73, pp.2308-2316, 2010.
- [15] Ibrahim Turkoglu , E. Didem Kaymaz, A hybrid method based on artificial immune system and k-NN algorithm for better prediction of protein cellular localization sites, *Applied Soft Computing*, 9, pp.497-502, 2009.
- [16] Niranjana P. Bidargaddi v.d., Combining segmental semi-Markov models with neural networks for protein secondary structure prediction, *Neurocomputing*, 72, 3943-3950, 2009.
- [17] Jieyue He v.d., Rule Generation for Protein Secondary Structure Prediction With Support Vector Machines and Decision Tree, *IEEE Transactions on Nanobioscience*, 5(1), pp.46-53, 2006.
- [18] Victor Robles v.d., Bayesian network multi-classifiers for protein secondary structure prediction, *Artificial Intelligence in Medicine*, 31, 117-136, 2004.
- [19] Long-Hui Wang v.d., Predicting Protein Secondary Structure by a Support Vector Machine Based on a New Coding Scheme, *Genome Informatics*, 15(2), 181-190, 2004.
- [20] David L. Dowe v.d., A Decision Graph Explanation of Protein Secondary Structure Prediction, 26th Hawaii Int. Conf. Sys. Sci., Vol 1, pp. 669-678, 1993.