Protein İkincil Yapı Tahmini İçin Makine Öğrenmesi Yöntemlerinin Karşılaştırılması

Comparison of Machine Learning Classifiers for Protein Secondary Structure Prediction

Zafer AYDIN
Bilgisayar Mühendisliği
Abdullah Gül Üniversitesi
Kayseri, Türkiye
zafer.aydin@agu.edu.tr

Oğuz KAYNAR, Yasin GÖRMEZ ve Yunus Emre IŞIK
Yönetim Bilişim Sistemleri
Cumhuriyet Üniversitesi
Sivas, Türkiye
okaynar@cumhuriyet.edu.tr,
yasingormez@cumhuriyet.edu.tr, yeisik@cumhuriyet.edu.tr

***Özetçe*—Proteinlerin üç boyutlu yapılarının tahmin edilmesi teorik kimya ve biyoenformatik için önemli problemlerden biridir. Protein yapı tahmininin en önemli aşamalarından biri ise ikincil yapı tahminidir. Protein veritabanlarındaki verilerin hızlı artışı ve yakın zamanda geliştirilen farklı öznitelik çıkarma yöntemleri neticesinde ikincil yapı tahmini için kullanılan veri setleri boyut ve örnek sayısı bakımından büyümektedir. Bu nedenle hızlı çalışan ve belirli bir doğruluk oranını sahip tahmin algoritmaların kullanılması önem kazanmaktadır. Bu çalışmada iki aşamalı hibrit bir sınıflandırıcının ikinci aşaması için çeşitli sınıflama algoritmaları, EVAset veri seti kullanılarak hem orijinal boyutlu uzayda hem de bilgi kazancı metriği ile boyutu düşürülen uzayda optimize edilmiştir. Elde edilen sonuçlar doğrultusunda en başarılı tahmin yöntemi destek vektör makinası olurken model eğitme süresi bakımından en hızlı yöntem aşırı öğrenme makinası olarak elde edilmiştir.**

Anahtar Kelimeler — İkincil Yapı Tahmini; Protein Yapı Tahmini; Öznitelik Seçimi; Makine Öğrenmesi

***Abstract*— Three-dimensional structure prediction is one of the important problems in bioinformatics and theoretical chemistry. One of the most important steps in the three-dimensional structure prediction is the estimation of secondary structure. Due to rapidly growing databases and recent feature extraction methods datasets used for predicting secondary structure can potentially contain a large number of samples and dimensions. For this reason, it is important to use algorithms that are fast and accurate. In this study, various classification algorithms have been optimized for the second phase of a two-stage classifier on EVAset benchmark both in the original input space and in the space reduced using the information gain metric. The most accurate classifier is obtained as the support vector machine while the extreme learning machine is significantly faster in model training.**

Keywords — Secondary Structure Prediction; Protein Structure Prediction; Feature Selection; Machine Learning

I. GİRİŞ

Son zamanlarda büyük çaplı DNA dizilim çalışmalarından çok fazla sayıda protein veri dizisi üretilmiştir. Ancak bu proteinlerin birçoğunun yapısı deneysel olarak çözülmemiştir. X-ışını kristalografisi ve Nükleer Manyetik Rezonans (NMR) gibi proteinin yapısını deneysel olarak ortaya çıkarmaya yarayan yöntemler zaman alabilmekte ve masraflı olabilmektedirler. Proteinin yapısı ile işlevi arasında güçlü bir bağlantı olduğu da düşünüldüğünde, protein yapı tahmini (PYP) son zamanların en popüler konularından biri haline gelmiştir. Ayrıca ilaç tasarımı probleminde ilaç moleküllerinin bağlanacağı proteinlerin yapılarının tespit edilmesi için deneysel yöntemlerin yetersiz kaldığı durumlarda PYP kullanılmaktadır. PYP denince akla üç boyutlu (3D) yapı tahmini gelmektedir. Üç boyutlu yapının direkt olarak tahmin edilmesi zor bir problem olduğundan ilk olarak ikincil yapı ve çözücü erişilirlik gibi hedef proteinin çeşitli yapısal özellikleri tahmin edilir. Protein ikincil yapı tahmini (PİYP), üç boyutlu yapıyı tahmin eden yöntemlerde yaygın olarak kullanılmaktadır. PİYP probleminde proteini oluşturan her bir amino aside karşılık gelen ikincil yapı sınıfının tahmin edilmesi amaçlanır.

Makine öğrenmesi yöntemlerinin birçok alanda kullanılmaya başlanması ile birlikte PİYP için de çeşitli çalışmalar yapılmıştır. Bu çalışmaların çoğunu denetimli öğrenme yaklaşımları oluşturmaktadır. Salamov ve Solovyev, skorlama matrisi üzerine yapay sinir ağları (YSA) ve en yakın k komşu (k-nn) uygulayarak yaptıkları çalışmada %72.2 Q3 başarı oranı elde etmişlerdir [1]. Jones pozisyona özgü puanlama matrisini (Position-Specific Scoring Matrix - PSSM), PSI-BLAST algoritması ile hesaplayarak elde ettiği veri seti üzerine YSA uygulamış ve %76.5 ile %78.3 arasında Q3 başarı oranı elde etmiştir [2]. Jian-wei ve diğerleri PİYP için bir YSA modeli önermişler ve bu modeli klasik geri yayılımlı öğrenme algoritması ile karşılaştırmışlardır. Önermiş oldukları bu model klasik algoritmaya göre %9 daha iyi sonuç elde etmiştir [3]. Mirabello ve Pollastri çift yönlü yinelemeli sinir ağı (bidirectional recurrent neural networks) kullanarak yaptıkları iki uygulamaya Porter 4.0 ve Paleale 4.0 isimlerini vermişlerdir. Porter 4.0 %82.2, Paleale 4.0 ise %80 Q3 başarı oranı elde etmiştir [4]. Aydın ve diğerleri CB513 veri seti üzerine dinamik Bayes ağları ve destek vektör makineleri (Support Vector Machines - SVM) uyguladıkları çalışmada %80.3 başarı oranı elde etmişlerdir [5]. Huang ve Chen PSSM değerlerini, net yük, doğrulama parametreleri, yan zincir kütlesi ve hidrofobik olmak

978-1-5386-1501-0/18/$31.00 ©2018 IEEE

üzere dört fizikokimyasal özellikle birleştirerek oluşturdukları veri seti üzerine destek vektör makineleri uygulayarak %79.52 Q3 başarı oranı elde etmişlerdir [6].

Protein yapı tahminin başarısı yalnızca sınıflama algoritmalarının geliştirilmesi ile sınırlandırılmamaktadır. PİYP için geliştirilen ilk yöntemler, her bir amino asidin sarmal ya da yaprak oluşturma eğilimlerine dayanıyordu. Buna ek olarak ikincil yapısal etiketlerinin oluşum enerjisini tahmin eden kuralları da kullanan yöntemler 3-halli ikincil yapı tahmininde %60 başarı oranları elde etmekteydi. Daha sonra çoklu hizalama yöntemlerinden faydalanılarak geliştirilen yeni öznitelik vektörleri sayesinde bu başarı oranı %80-82'lere ulaşmıştır [5], [7]. Bu hizalamalara ek olarak yapısal profillerin protein bilgilerini özetlemek için öznitelik olarak kullanıldığı durumlarda başarı oranı %84-85'e ulaşmıştır [8], [9]. Bu çalışmalar sonucunda bilgilendirici özniteliklerin çıkarılmasının, başarı oranını artırdığı gözlemlenmiştir. Bu öznitelikleri çıkarmakta genellikle PSI-BLAST algoritması kullanılsa da, HHBLITS algoritması kullanılarak çıkarılan özniteliklerin de eklenmesi ile sınıflama başarısında artış sağlandığı görülmüştür [5], [10]. PİYP için öznitelik çıkarma işleminde örnek sayısı amino asit sayısına eşittir ve bu sayı veri tabanlarındaki büyümeye paralel olarak artmaktadır. Ayrıca bir amino asit için öznitelik vektörü oluşturulurken bir boyutlu dizilime göre o amino asitten önce ya da sonra gelen amino asitlerin öznitelik parametreleri de kayan pencere yaklaşımı ile birleştirilmektedir [5], [11]. Dolayısıyla farklı öznitelik çıkarımı yöntemlerinin birlikte kullanılması, veri tabanlarındaki veri miktarının hızlı artması ve öznitelik vektörü oluşturmak için kullanılan pencerenin genişliği gibi faktörler neticesinde model eğitmek ve tahmin başarısını sınamak için oluşturulan veri seti yüksek boyutlara sahip olabilmektedir. Bunun neticesinde kullanılan sınıflama algoritmasının model eğitme süresi önem kazanmaktadır. Diğer taraftan ilaç tasarımı ve protein fonksiyon tahmini gibi problemlerde etkin sonuçların alınabilmesi için yapı tahmin başarısı belirli bir seviyenin üzerinde olmalıdır.

Bu çalışmada protein ikincil yapı tahmini için iki aşamalı hibrit bir sınıflandırıcı (DSPRED) kullanılmıştır [11]. DSPRED yönteminin ikinci aşaması için aşırı öğrenme makineleri (Extreme Learning Machine - ELM), k-NN, rastgele ağaçlar (Random Forest - RF), YSA ve SVM yöntemleri kullanılmıştır. Yöntemler EVAset standart veri kümesi kullanılarak gerek olası bütün öznitelikleri içeren orijinal girdi uzayında gerekse öznitelik seçimi ile boyutu düşürülen uzayda eğitilmiştir. Her iki deneyde de sınıflandırma algoritmaları başarı oranları ve hız bakımından karşılaştırılmıştır. Literatürde protein ikincil yapı tahmini için incelenen bu yöntemleri karşılaştıran başka bir çalışma bulunmamaktadır.

II. YÖNTEMLER

A. Öznitelik Çıkarma

Bu çalışmada ilk olarak veri setindeki amino asitlerin etiket bilgilerini çıkarmak için ikincil yapı birimleri (sarmal - H, beta iplik - E ve döngü - L) proteinin PDB [12] veri tabanındaki üç boyutlu yapısından başlayarak DSSP programı ile çıkarılmıştır[13]. Daha sonra her bir amino asit için öznitelikler PSI-BLAST [14] ve HHBLITS [15] olmak üzere iki farklı hizalama yöntemi ile DSPRED yönteminin dinamik Bayes ağlarını (DBN) kullanan ilk tahmin aşaması ile çıkarılmıştır [11]. Bu işlem sonucunda her bir protein için $20 \times N$ boyutunda iki adet PSSM matrisi ile $3 \times N$ boyutunda üç adet ikincil yapı tahmini profil matrisi oluşturulmuştur. N hedef proteinin amino asit dizilimi uzunluğunu, "3" değeri sınıf sayısını, 20 değeri ise doğada bulunan amino asit çeşidi sayısını temsil etmektedir. Son olarak amino asitlerin çevresindeki amino asitlerle kimyasal reaksiyonlara girmesinden kaynaklanan etkileri ölçebilmek için 11 birim uzunluğunda simetrik bir pencere kullanılmış ve her bir amino asit için toplamda 539 adet öznitelik çıkarılmıştır.

B. Öznitelik Seçimi

Çalışmada kazanım oranı (information gain ratio) metriği [16] süzücü öznitelik seçim yöntemi olarak kullanılmıştır. PİYP için bu metriği kullanan öznitelik seçme yöntemlerin, diğer öznitelik seçme yöntemlerine göre daha yüksek başarı/hız oranı elde ettiği Görmez tarafından yapılan tez çalışmasında tespit edilmiştir [11]. Bu nedenle bu metriği tek başına kullanan öznitelik seçim yöntemi tercih edilmiştir. Bu amaç doğrultusunda ilk olarak, her bir öznitelik için bir entropi değeri hesaplanır. Daha sonra bu değer, ilgili öznitelik için hesaplanan bölünme bilgisi değerine bölünerek bir skor değeri hesaplanır. Daha sonra belirlenen eşik değerinin altında skor elde eden, ya da skor sıralamasına göre belirli bir sıranın altında kalan öznitelikler elenerek yeni veri setleri oluşturulur. Bu çalışmada veri setinden kaynaklı hız farklarının önüne geçmek için, her bir veri setindeki en yüksek skora sahip 200 öznitelik seçilerek boyutu düşmüş veri setleri elde edilmiştir.

C. Sınıflandırma

PİYP, tahmin edilecek sınıf sayısına bağlı olarak 8-halli ya da 3-halli olmak üzere iki farklı şekilde yapılabilmektedir. 3-halli olarak yapılan öngörü işlemi, 8-halli olarak oluşturulan veri setinde benzer etiketlerin bir araya getirilerek 3'e indirgenmesi sonucunda oluşturulan veri seti kullanılarak yapılmaktadır. Bu çalışmada, 3-halli olarak oluşturulan veri seti üzerinde DSPRED yönteminin ikinci aşaması için ELM [17], k-nn [18], RF [19], YSA [20] ve SVM [21] yöntemleri kullanılmıştır.

III. UYGULAMA

Çalışmada 584595 amino asit içeren EVAset standart veri kümesi [22] üzerinde protein ikincil yapı tahmini yapılmıştır. Bu veri seti çapraz doğrulama (cross-validation) [23] kullanılarak 10 adet farklı eğitim ve test veri setleri oluşturulmuştur. Deneyler 2.6 Ghz 32 çekirdekli işlemcisi, 128 GB RAM'i bulunan bir iş istasyonunda yapılmıştır. Deneyler sırasında her bir veri seti için sadece tek çekirdek kullanılmıştır. ELM deneyleri için Matlab [24] platformu, SVM deneyleri için libSVM [25] programı, YSA, k-NN, RF ve kazanım oranı deneyleri için ise Weka [26] programı kullanılmıştır.

Çalışmada ilk olarak öznitelik çıkarma işlemi bölüm II A'da anlatıldığı gibi yapılmıştır. Daha sonra çapraz doğrulama için oluşturulan her bir eğitim kümesindeki proteinlerin %18'i eğitim amaçlı %5'i de test (validasyon) amaçlı kullanılmak üzere rastgele ve örtüşmeyecek şekilde seçilmiş ve 10 adet optimizasyon amaçlı eğitim kümesi ile 10 adet optimizasyon amaçlı validasyon kümesi elde edilmiştir. Optimizasyon için kullanılacak veri kümelerinin daha az sayıda veri örneği barındıracak sekilde oluşturulmasının sebebi optimizasyon aşamasında çok miktarda model eğitme ve test etme işleminin yapılmasıdır. Bu sayede optimizasyon aşamasında tahmin

başarısından fazla ödün vermeden model eğitme süreleri önemli ölçüde kısaltılabilmektedir. Elde edilen bu veri setleri kullanılarak SVM hariç tüm sınıflama algoritmaları için optimizasyon işlemi yapılmış, daha sonra kazanım oranı metriği yardımı ile her bir veri seti için en etkili 200 öznitelik bulunarak elde edilen yeni veri setleri için optimizasyon deneyleri tekrar edilmiştir. SVM yönteminin C ve gamma parametreleri daha önceki bir çalışmada optimize edilmişti [5]. Bu deneyler esnasında ELM için gizli katmandaki nöron sayısı (ns) 25 ile 800 arasında 25'er artacak şekilde, k-nn için en yakın komşu değeri (k) 1 ile 51 arasında 2'şer artacak şekilde, RF için ağaç sayısı (a) 10 ile 350 arasında 10'ar artacak şekilde YSA için ise iterasyon sayısı (i) 50 ile 200 arasında 50'şer artacak ve gizli katman nöron sayısı (ns) 25 ile 500 arasında 25'er artacak şekilde ızgara araması (grid-search) yaklaşımı ile optimize edilmiştir. Tablo 1'de her bir yöntem için en iyi optimizasyon-test başarı oranı elde eden parametre değerleri, çalışma süreleri ve başarı oranları 539 ve 200 öznitelik içeren veri setleri için ayrı ayrı gösterilmiştir. Çalışma süresi, çapraz doğrulama sonucunda elde edilen 10 veri setinin toplam çalışma zamanının saat cinsinden değeri, başarı oranı ise 10 veri setinin ortalama başarı oranı değerlerini göstermektedir. Parametre değerleri ise 10 veri setinin her biri için sırası ile gösterilmiştir.

TABLO I. EVASET VERİ KÜMESİNDE YAPILACAK 10-KATLI ÇAPRAZ DOĞRULAMA DENEYİNİN HER EĞİTİM KÜMESİ ÜZERİNDE YAPILMIŞ HİPER-PARAMETRE OPTİMİZASYONU SONUÇLARI

Yöntem	*Öznitelik sayısı*	*Başarı oranı*	*Çalışma süresi*	*En iyi parametre değerleri*
ELM	539	0.8163	0.7 sa.	ns = {800, 750, 750, 625, 775, 725, 750, 750, 550, 750}
k-nn	539	0.8049	461 sa.	k = {45, 31, 25, 33, 37, 23, 31, 45, 51, 25}
RF	539	0.8209	28.2 sa.	a = {160, 350, 300, 230, 190, 330, 250, 350, 300, 300}
YSA	539	0.8049	5120 sa.	i = {50, 50, 50, 50, 50, 50, 50, 50, 50, 100} ns = {25, 25, 25, 25, 25, 25, 25, 500, 425, 450}
ELM	200	0.8255	0.49 sa.	ns = {750, 650, 700, 750, 800, 800, 750, 775, 800, 750}
k-nn	200	0.8137	128 sa.	k = {43, 25, 21, 23, 27, 21, 21, 33, 21, 31}
RF	200	0.8267	24.6 sa.	a = {270, 350, 310, 240, 280, 230, 260, 320, 340, 130}
YSA	200	0.8196	2234 sa.	i = {50, 50, 50, 50, 50, 50, 50, 50, 100, 50} ns = {25, 25, 25, 25, 25, 25, 25, 25, 25, 25}

Optimizasyon sonuçları incelediğinde, 539 öznitelik içeren veri setleri içinde en yüksek başarı oranı RF algoritması ile elde edilirken, en düşük başarı oranı YSA ve k-nn algoritmaları ile elde edilmiştir. Çalışma süresi bakımından incelendiğinde, ELM yönteminin en hızlı, YSA yönteminin ise en yavaş çalışan yöntemler oldukları gözlemlenmiştir. 200 öznitelik içeren veri setleri içinde ise en yüksek başarı oranı RF algoritması ile elde edilirken, en düşük başarı oranı k-nn algoritması ile elde edilmiştir. Çalışma süresi bakımından incelendiğinde, ELM yönteminin en hızlı, YSA yönteminin ise en yavaş çalışan yöntemler oldukları gözlemlenmiştir. Her iki boyuttaki veri setlerini kullanan modeller için de ELM başarı oranı bakımından ikinci en yüksek başarı oranı elde etmesinin yanı sıra, diğer algoritmalardan en az 40 kat daha hızlı optimizasyon işlemini bitirmiştir. Bu işlemlerden sonra, belirlenen en iyi parametreler kullanılarak EVAset'in 539 özniteliğe sahip orijinal eğitim ve test verileri ile 10-katlı çağraz doğrulama deneyi yapılmıştır. Bu deneyde ELM, k-nn, RF, YSA ve SVM modelleri eğitilmiştir. SVM modelinde daha önce Aydin ve diğerleri tarafından optimize edilen gama parametresi 0.00781, C parametresi ise 1 olarak kullanılmıştır [5]. Tablo 2'de her bir yöntem için 10-katlı çapraz doğrulama için elde edilen ortalama başarı oranı, 10 adet başarı oranı arasındaki standart sapma değeri, ortalama MCC skoru, ortalama F skoru ve toplam çalışma zamanı gösterilmiştir.

TABLO II. 539 ÖZNİTELİK İÇEREN EVASET İÇİN 10 KATLI ÇAPRAZ DOĞRULAMA İLE ELDE EDİLEN DENEY SONUÇLARI

Yöntem	*Başarı oranı*	*Standart sapma*	*MCC*	*F skor*	*Çalışma süresi*
ELM	0.8173	0.639	0.716	0.8177	1.88 sa.
k-nn	0.8152	0.530	0.710	0.8155	291.9 sa.
RF	0.8280	0.602	0.730	0.8284	12.09 sa.
YSA	0.8207	1.058	0.718	0.8208	119.5 sa.
SVM	0.8380	0.595	0.744	0.8383	777.6 sa.

539 öznitelik içeren veri setleri için sınıflama sonuçları incelediğinde en iyi başarı oranı veren algoritma SVM olurken, optimizasyon işleminde de olduğu gibi ELM en hızlı çalışan algoritma olmuştur. Burada çalışma süresi model eğitme ve tahmin hesaplama zamanını içermektedir.

Bu sınıflama işleminden sonra bilgi kazancı ile belirlenen en etkili 200 öznitelik kullanılarak yeni eğitim ve test veri setleri oluşturulmuş ve modeller tekrar eğitilmiştir. Bilgi kazancı ile en etkili öznitelikleri belirleyerek 10 veri setini yeniden oluşturma işlemi toplamda 4.4 saat sürmüştür. Yeni veri setleri ile eğitimler sırasında ELM, k-nn, RF ve YSA modellerinde 200 öznitelik için optimize edilen parametre değerleri, SVM modelinde ise Aydin ve diğerleri tarafından optimize edilen parametreler kullanılmıştır [5]. Tablo 3'de 200 özniteliği olan EVAset üzerinde yapılan 10-katlı çapraz doğulama deneyinin ortalama başarı oranı, 10 adet başarı oranının standart sapma değeri, ortalama MCC skoru, ortalama F skoru ve toplam çalışma zamanı gösterilmektedir.

TABLO III. 200 ÖZNİTELİK İÇEREN VERİ SETLERİ İÇİN 10 KAT ÇAPRAZ DOĞRULAMA İLE ELDE EDİLEN DENEY SONUÇLARI

Yöntem	*Başarı oranı*	*Standart sapma*	*MCC*	*F skor*	*Çalışma süresi*
ELM	0.8274	0.646	0.728	0.8277	0.882 sa.
k-nn	0.8210	0.582	0.719	0.8213	91.18 sa.
RF	0.8318	0.637	0.735	0.8322	9.59 sa.
YSA	0.8308	0.656	0.734	0.8309	6.14 sa.
SVM	0.8345	0.643	0.742	0.8348	279.50 sa.

Sonuçlar incelendiğinde 200 öznitelik içeren veri setlerini kullanan modellerde de en yüksek başarı oranı SVM ile edilirken, en hızlı çalışan yöntem ELM olmuştur.

IV. SONUÇLAR

Bu çalışmada protein ikincil yapı tahmini için ELM, k-nn, RF ve YSA olmak üzere 4 farklı sınıflama algoritması optimize edilmiş ve daha önce optimize edilen SVM yöntemi ile karşılaştırılmıştır. Yöntemler hem orijinal boyuttaki veri setleri hem de kazanım oranı metriği yardımı ile boyut düşürülerek elde edilen veri setleri üzerinde eğitilmiştir. Sonuç olarak en yüksek başarı oranı veren yöntemin SVM olduğu, en hızlı çalışan yöntemin ise ELM olduğu tespit edilmiştir. ELM yönteminin diğer yöntemlere göre çok daha hızlı çalışıyor olması ve diğer yöntemlere yakın başarı oranları alıyor olması nedeni ile PİYP için oluşturulacak bir modelde doğruluk oranının daha az önemli olduğu uygulamalarda sınıflama algoritması olarak kullanılmasının uygun olacağı anlaşılmaktadır. Bunun yanı sıra 200 öznitelik içeren veri setleri kullanılarak oluşturulan modeller, 539 öznitelik içeren veri setleri kullanılarak oluşturulan tüm modellere göre daha hızlı model eğitme imkanı sağlamış ve SVM hariç diğer tüm sınıflama algoritmalarında 539 öznitelik içeren modellere göre daha yüksek başarı oranı elde etmişler. Bu nedenle PİYP için öznitelik seçiminin yapılmasının da sınıflama algoritmasının etkinliğini olumlu yönde etkilediği söylenebilir. İlerleyen zamanlarda aynı yöntemler çözücü erişilirlik ve torsion açı tahmini için de yapılarak bu yöntemlerin 3 boyutlu protein yapı tahminine olan etkilerinin analiz edilmesi planlanmaktadır.

KAYNAKLAR

[1] A. A. Salamov ve V. V. Solovyev, "Prediction of Protein Secondary Structure by Combining Nearest-neighbor Algorithms and Multiple Sequence Alignments", J. Mol. Biol., c. 247, sayı 1, ss. 11–15, Mar. 1995.

[2] D. T. Jones, "Protein secondary structure prediction based on position-specific scoring matrices", J. Mol. Biol., c. 292, sayı 2, ss. 195–202, Eyl. 1999.

[3] L. Jian-wei, C. Guang-hui, L. Hai-en, L. Yuan, ve L. Xiong-lin, "Prediction of protein secondary structure using multilayer feed-forward neural networks", içinde 2013 25th Chinese Control and Decision Conference (CCDC), 2013, ss. 1346–1351.

[4] C. Mirabello ve G. Pollastri, "Porter, PaleAle 4.0: high-accuracy prediction of protein secondary structure and relative solvent accessibility", Bioinformatics, c. 29, sayı 16, ss. 2056–2058, Ağu. 2013.

[5] Z. Aydin, A. Singh, J. Bilmes, ve W. S. Noble, "Learning sparse models for a dynamic Bayesian network classifier of protein secondary structure", BMC Bioinformatics, c. 12, s. 154, May. 2011.

[6] Y. F. Huang ve S. Y. Chen, "Protein secondary structure prediction based on physicochemical features and PSSM by SVM", içinde 2013 IEEE Symposium on Computational Intelligence in Bioinformatics and Computational Biology (CIBCB), 2013, ss. 9–15.

[7] J. Cheng, A. N. Tegge, ve P. Baldi, "Machine Learning Methods for Protein Structure Prediction", IEEE Rev. Biomed. Eng., c. 1, ss. 41–49, 2008.

[8] D. Li, T. Li, P. Cong, W. Xiong, ve J. Sun, "A novel structural position-specific scoring matrix for the prediction of protein secondary structures", Bioinformatics, c. 28, sayı 1, ss. 32–39, Oca. 2012.

[9] G. Pollastri, A. J. Martin, C. Mooney, ve A. Vullo, "Accurate prediction of protein secondary structure and solvent accessibility by consensus combiners of sequence and structure information", BMC Bioinformatics, c. 8, sayı 1, s. 201, Ara. 2007.

[10] Z. Aydin, D. Baker, ve W. S. Noble, "Constructing structural profiles for protein torsion angle prediction", sunulan 6th International Conference on Bioinformatics Models, Methods and Algorithms, BIOINFORMATICS 2015, 2015.

[11] Y. Görmez, "Dimensionality reduction for protein secondary structure prediction", Abdullah Gül Üniversitesi, 2017.

[12] "RCSB Protein Data Bank - RCSB PDB", 2017, https://www.rcsb.org/pdb/home/home.do.

[13] W. Kabsch ve C. Sander, "Dictionary of protein secondary structure: Pattern recognition of hydrogen-bonded and geometrical features", Biopolymers, c. 22, sayı 12, ss. 2577–2637, Ara. 1983.

[14] "PSI-BLAST", 2017, https://blast.ncbi.nlm.nih.gov/Blast.cgi?CMD=Web&PAGE=Proteins.

[15] M. Remmert, A. Biegert, A. Hauser, ve J. Söding, "HHblits: lightning-fast iterative protein sequence searching by HMM-HMM alignment", Nat. Methods, c. 9, sayı 2, s. 173, Şub. 2012.

[16] K. O. Jeppson, "Modeling the influence of the transistor gain ratio and the input-to-output coupling capacitance on the CMOS inverter delay", IEEE J. Solid-State Circuits, c. 29, sayı 6, ss. 646–654, Haz. 1994.

[17] G.-B. Huang, Q.-Y. Zhu, ve C.-K. Siew, "Extreme learning machine: a new learning scheme of feedforward neural networks", içinde 2004 IEEE International Joint Conference on Neural Networks (IEEE Cat. No.04CH37541), 2004, c. 2, ss. 985–990 c.2.

[18] D. T. Larose, "k-Nearest Neighbor Algorithm", içinde Discovering Knowledge in Data, John Wiley & Sons, Inc., 2004, ss. 90–106.

[19] M. Pal, "Random forest classifier for remote sensing classification", Int. J. Remote Sens., c. 26, sayı 1, ss. 217–222, Oca. 2005.

[20] J. E. Dayhoff ve J. M. DeLeo, "Artificial neural networks", Cancer, c. 91, sayı S8, ss. 1615–1635, Nis. 2001.

[21] V. Vapnik, The Nature of Statistical Learning Theory. Springer Science & Business Media, 2013.

[22] I. Y. Y. Koh vd., "EVA: evaluation of protein structure prediction servers", Nucleic Acids Res., c. 31, sayı 13, ss. 3311–3315, Tem. 2003.

[23] "Cross-validation", 2017, https://en.wikipedia.org/wiki/Cross-validation_(statistics).

[24] "Basic ELM Algorithms", 2017, http://www.ntu.edu.sg/home/egbhuang/elm_codes.html.

[25] "LIBSVM -- A Library for Support Vector Machines", 2017, https://www.csie.ntu.edu.tw/~cjlin/libsvm/.

[26] "WEKA", 2017, https://weka.wikispaces.com/.