Protein İkincil Yapı Tahmini için NR ve UniClust Veri Tabanlarının Karşılaştırılması

Comparison of NR and UniClust Databases for Protein Secondary Structure Prediction

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

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

***Özetçe*— Proteinlerin üç boyutlu yapılarının tahmin edilmesi teorik kimya ve biyoenformatik için önemli problemlerden biridir. Üç boyutlu yapı tahminin en önemli aşamalarından biri ise ikincil yapı tahminidir. İkincil yapı tahmininde başarı oranının artırılması kullanılan sınıflama algoritması kadar, hesaplanan özniteliklere de bağlı olmaktadır. Öznitelik çıkarmak için sıkça kullanılan çoklu hizalama yöntemlerinde ise hesaplanan değerler, hizalama için kullanılan veri tabanına göre farklılık göstermektedir. Bu nedenle öznitelik matrisleri oluşturulurken uygun veri tabanın seçilmesi önem kazanmaktadır. Bu çalışmada CB513 veri seti kullanılarak iki farklı hizalama yöntemi ve üç farklı veri tabanı yardımı ile 5 farklı veri seti oluşturulmuş ve bu veri setleri iki aşamalı hibrit bir sınıflandırıcı kullanılarak karşılaştırılmıştır. Elde edilen sonuçlar doğrultusunda en iyi başarı oranı HHBlits hizalama yönteminin ilk aşamasında hesaplanacak PSSM değerleri için UniClust ve yapısal profil matrisleri için yine HHBlits'in ilk aşamasında NR veri tabanı kullanıldığında elde edilmiştir.**

Anahtar Kelimeler — İkincil Yapı Tahmini; Protein Yapı Tahmini; Çoklu Hizalama, Protein Veri Tabanı

***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. Improving the accuracy rate in protein secondary structure prediction depends on computed attributes as well as the classification algorithms. In multiple alignment methods, which are often used to extract an attribute, the calculated values differ according to the database used for the alignment. For this reason, it is important to use a suitable database against which the target proteins are aligned to compute profile feature vectors. In this study, 5 different datasets are generated for the CB513 benchmark with the aid of two different alignment methods and three different databases. The profile features are fed as input to a two-stage hybrid classifier. According to the experimental results, the highest accuracy rate is obtained when UniClust database is used at the first stage of HHBlits alignment to calculate PSSM values and NR database is used at the first stage of HHBlits alignment to calculate structural profile matrices.**

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

Keywords — Secondary Structure Prediction; Protein Structure Prediction; Multi Alignment; Protein Database

I. GİRİŞ

Canlı organizmalar için hayati öneme sahip olan proteinlerin, üç boyutlu (3D) yapısı ile işlevi arasında yakın bir ilişki bulunmaktadır. Bu nedenle proteinin 3D yapısı ne kadar iyi bilinirse, işlevi hakkında da o kadar iyi bilgi sahibi olunabilmektedir. X-ışını kristalografisi ve Nükleer Manyetik Rezonans (NMR) gibi yöntemler kullanılarak proteinin üç boyutlu yapısını deneysel olarak çözümlemek mümkündür ancak bu yöntemler masraflı olabilmekte ve zaman alabilmektedir. Bu nedenle protein yapı tahmini (PYT), biyoenformatik alanında önemli konulardan biri haline gelmiştir. Ayrıca ilaç tasarımı problemlerinde ilaç moleküllerinin bağlanacağı proteinlerin yapılarının tespit edilmesi için deneysel yöntemlerin yetersiz kaldığı durumlarda PYT kullanılmaktadır. 3D yapının tahmini yapılmadan önce ikincil yapı, torsion açı, çözücü erişilirlik gibi hedef proteinin çeşitli yapısal özellikleri tahmin edilir. Bu yapısal özelliklerden biri olan protein ikincil yapı tahmini (PİYT) ise PYT'nin en önemli aşamalarından biridir. PİYT probleminde proteinin her bir amino asidine karşılık gelen ikincil yapı etiketini bulmak amaçlanmaktadır.

Son zamanların güncel konularından olan makine öğrenmesi yöntemleri, PİYT problemi için de sıkça kullanılmıştır. Bu yöntemler arasında yapay sinir ağları (YSA) [1], [2], dinamik bayes ağları (DBA) [3], [4], gizli Markov modelleri (GMM) [5], destek vektör makineleri (DVM) [6], [7], en yakın k komşu (k-nn) [8], [9], derin öğrenme yaklaşımları [10]–[12] gibi çeşitli makine öğrenmesi yöntemleri bulunmaktadır.

Yapı tahminin başarısı sınıflama algoritmasına bağlı olduğu kadar, eğitim için elde edilen özniteliklere de bağlıdır. Çoklu hizalama yöntemleri kullanılarak elde edilen öznitelik vektörlerinin sayesinde tahmin başarı oranı %80-82'lere ulaşmıştır [3], [13]. Bu vektörlere yapısal profil matrisleri de

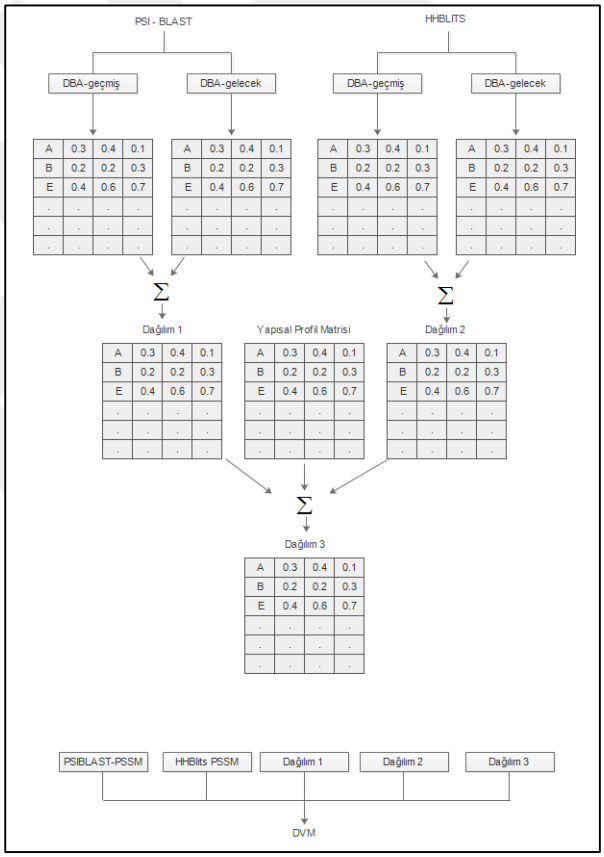
öznitelik olarak eklendiğinde başarı oranı %84-85'lere yükselmiştir [14], [15]. Çoklu hizalama için PSI-BLAST algoritması sıkça kullanılsa da, HHBlits algoritması kullanılarak çıkarılan özniteliklerin de eklenmesinin tahmin başarısını artırdığı gözlemlenmiştir [3], [16].

Çoklu hizalama işleminde hedef protein, veri tabanındaki yapısı bilinen proteinler ile hizalanır. Daha sonra belirli bir eşik değerinin üstüne kalan proteinler, hedef proteinin özniteliklerini belirlemek üzere seçilir. Bu nedenle kullanılan hizalama yönteminin yanı sıra, kullanılan protein veri tabanı da, tahmin başarısını artırmada önemli bir unsur olmaktadır. Bu çalışmada PSI-BLAST [17] yöntemi kullanılarak elde edilen özniteliklere, HHBlits [18] yönteminin ilk aşamasında NR [19] veri tabanına hizalayan yöntem kullanılarak elde edilen öznitelikler eklenmiş ve PİYT yapılmıştır. Daha sonra HHBlits [18] yönteminin ilk aşamasında NR [19] veri tabanı yerine yakın zamanda geliştirilen UniClust [20] veri tabanı kullanılarak öznitelik vektörleri yeniden elde edilmiş ve PİYT tekrar yapılarak elde edilen sonuçlar kıyaslanmıştır. İki aşaması olan HHBlits [18] yöntemi için ikinci aşamada her zaman PDB [21] veri tabanı kullanılmıştır. Bu ikinci aşamadan elde edilen taslak proteinler sadece yapısal profil matrislerinin hesaplanması için kullanılmış, PSSM profil vektörlerinin hesaplanması için kullanılmamıştır. Öznitelikler çıkarıldıktan sonra, sınıflama algoritmalarından doğacak üstünlüğün önüne geçmek için DSPRED [3] yöntemi kullanılmıştır. Bu yöntemde ilk aşamada Dinamik Bayes ağları ikinci aşamada ise destek vektör makinesi kullanılmaktadır.

II. YÖNTEMLER

A. DSPRED

DSPRED yöntemi iki aşamalı hibrit bir sınıflandırıcıdır [3]. İlk aşamada PSI-BLAST [17] ve HHBlits [18] hizalama yöntemleri ile elde edilen pozisyona özgü puanlama matrisleri (Position-Specific Scoring Matrix – PSSM) Dinamik Bayes ağlarına girdi özniteliği (input feature) olarak gönderilir (Şekil 1). Bu matrislere dizi profil matrisi de denmektedir. Yapısı tahmin edilecek proteindeki amino asit sayısı N ise bu matrislerin boyutu $20 \times N$ olmaktadır. Bu matristeki sütunlar proteinlerde bulunan amino asitleri satırlar ise her amino asite hizalanabilecek 20 amino aside ait skorları göstermektedir. Bir proteinde ortalama 100-200 amino asit bulunduğundan ve her proteindeki amino asit sayısı farklı olabildiğinden matristeki değerlerin hepsinin bir seferde sınıflandırma modellerinde kullanılması model eğitme süresinin artmasına neden olmaktadır. Ayrıca öznitelik sayısı yüksek olduğunda modelin aşırı uyum davranışına (overfitting) yakalanma riski yüksektir ve bu durum tahmin başarısını olumsuz etkilemektedir. Bu nedenle yapı sınıfı tahmin edilecek her amino asidin etrafında lokal bir pencere alınır ve bu pencerede bulunan değerler öznitelik olarak kullanılır. Burada penceredeki değerler profil matrisinin belirli sütun aralığını almaya karşılık gelir ve kayan pencere yaklaşımıyla bütün amino asitler için öznitelik vektörleri oluşturulur. Dinamik Bayes ağlarında kullanılan öznitelik vektörlerinde tek taraflı ve asimetrik bir pencere kullanılmaktadır. DBA-geçmiş modelinde bir amino asite ait matristen elde edilen 20 özniteliğe ek olarak o amino asitten önce gelen L_A amino asidin öznitelik değerleri uç uca eklenerek birleştirilir. DBA-gelecek modelinde ise bir amino asite ait matristen elde edilen 20 özniteliğe ek olarak o amino asitten sonra gelen L_A amino asidin öznitelik değerleri ucuca eklenerek birleştirilir. Burada amino asitlerin öncelik ve sonralık sırası proteinin N-terminal ucundan C-terminal ucuna doğru gidildiği zamanki sıraya göredir. Bu çalışmada L_A parametresi 5 alınmıştır. PSI-BLAST PSSM öznitelikleri bu sayede hem DBA-geçmiş hem de DBA-gelecek modellerinde kullanılır. Aynı şekilde HHBlits hizalamalarından elde edilen HHMAKE öznitelikleri de iki DBA modelinde kullanılmaktadır (Şekil 1). Bu sayede toplam dört DBA modeli kullanılmaktadır. PSIBLAST PSSM özniteliklerinin kullanıldığı DBA-geçmiş ve DBA-gelecek modelleri her amino asidin ikincil yapı sınıfını bir olasılık dağılımı olarak tahmin eder. Bu tahminlerin ortalaması alındığında Dağılım 1 elde edilir. Benzer şekilde HHMAKE PSSM öznitelikleri için elde edilen tahminlerin ortalaması alındığında Dağılım 2 elde edilir. Dağılım 1, Dağılım 2 ve yapısal profil matrislerinin ortalaması alındığında ise Dağılım 3 elde edilir. Yapısal profil matrisleri (YPM) ise HHBlits [18] yönteminin ikinci aşaması sonucunda elde edilmektedir.



Şekil 1. *DSPRED yöntemi adımları*

DSPRED yönteminin ikinci aşamasında PSI-BLAST PSSM öznitelikleri, HHMAKE PSSM öznitelikleri ve ilk aşamada elde edilen Dağılım 1, Dağılım 2 ve Dağılım 3 değerleri uç uca eklenerek öznitelik vektörü oluşturulur. Amino asitler arasındaki lokal etkileşimleri modelleyebilmek ve öznitelik sayısını makul bir seviyede tutabilmek için her amino asidin etrafında iki taraflı ve simetrik bir pencere alınır ve elde edilen öznitelik vektörü destek vektör makinesi (DVM) ile

sınıflandırılır (Şekil 1). Bu çalışmada simetrik pencere uzunluğunu 11 olarak seçilmiştir. Pencere uzunluğu değerlerini daha önceki çalışmalarımızda optimize edilmişti [3]. Seçilen pencere uzunluğu için elde edilen öznitelik vektörünün boyutu 539 olmuştur. İkincil yapı tahmini probleminde üç adet sınıf etiketi bulunduğundan bu çalışmada DVM için bire karşı bir (BKB) tekniği kullanılmıştır. BKB'de her bir sınıf ikilisi için ayrı ayrı DVM modelleri eğitilir ve bu modellerden gelen tahminler çeşitli istatiksel yöntemler ile birleştirilerek çoklu sınıflama yapan model oluşturulur.

B. Dizi Hizalama, Veritabanları, Öznitelik Çıkarımı ve Veri Setleri

Çalışmada 84119 amino asit içeren CB513 [22] veri seti kullanılmıştır. Bu veri setindeki proteinlere ait her bir amino asidin ikincil yapı sınıfı etiketleri (sarmal – H, beta iplik – E ve döngü – L) PDB veri tabanındaki üç boyutlu yapısından başlanarak DSSP [23] programı ile çıkarılmıştır.

Sınıflandırma modellerinde kullanılacak özniteliklerinin hesaplanması için yapısı tahmin edilecek proteinler veri tabanlarındaki amino asit dizileri ile ikili olarak hizalanır. Daha sonra bu ikili hizalamalar birleştirilerek çoklu hizalama (multiple alignment) elde edilir ve bu çoklu hizalamadaki istatistiksel skorlar hesaplanarak PSSM matrisleri veya saklı Markov modelleri oluşturulur. Bu çalışmada öznitelik vektörlerinin hesaplanması için PSI-BLAST ve HHBlits hizalama yöntemleri kullanılmıştır. PSI-BLAST yöntemi ile PSI-BLAST PSSM öznitelik vektörleri hesaplanmıştır. PSIBLAST yöntemi ile hizalama yaparken NR veri tabanı [19] kullanılmıştır. NR veri tabanında milyonlarca protein dizisi bulunmaktadır. HHBlits yöntemi ise saklı Markov modellerine dayandığından PSI-BLAST yönteminden daha doğru hizalama yapabilmektedir. Bu çalışmada HHBlits yöntemi HHMAKE PSSM öznitelikleri ile yapısal profil matrislerinin hesaplanması için kullanılmıştır. HHMAKE PSSM özniteliklerinin hesaplanması için HHBlits yönteminin ilk aşaması koşturulmuş ve CB513 veri setindeki proteinler NR veri tabanının indirgenmiş versiyonu olan NR20 veya yakın zamanda geliştirilen UniClust veri tabanlarından biri ile hizalanmıştır. Yapısal profil matrislerinin elde edilmesi içinse HHBlits yönteminin iki aşaması da çalıştırılmış ve ilk aşamada CB513 proteinleri NR20 veya yakın zamanda geliştirilen ve proteinlerin fonksiyonel özelliklerine göre kümelenmesine dayanan UniClust [20] veri tabanlarından biri ile hizalanmıştır. HHBlits yönteminin ikinci aşamasında ise ilk aşamada elde edilen saklı Markov modeli PDB [21] proteinlerinin saklı Markov modellerine hizalanmıştır. Daha sonra eşleşen proteinlerinin etiket bilgilerinin frekanslarının normalize edilmesi ile yapısal profil matrisleri elde edilmiştir.

HHMAKE PSSM özniteliklerinin çıkartılması aşamasında iki farklı veri tabanı kullanılabildiğinden (NR20 veya UniClust) ve yapısal profil matrislerinin oluşturulmasında HHBlits'in ilk aşamasında yine iki farklı veri tabanı kullanılabildiğinden (NR20 veya UniClust) toplam dört farklı kombinasyon bulunmaktadır. Bu nedenle CB513 veri setinin dört farklı versiyonu elde edilmiştir. Bu versiyonlar Tablo I'de özetlenmiştir. Bu tabloda sütunlar öznitelik vektörlerini, satırlar ise farklı DSPRED modelini eğitmek için kullanılacak CB513 veri setinin farklı versiyonlarını göstermektedir. Örneğin uniclust_pssm_nr_ypm modelinde HHMAKE PSSM özniteliklerini elde etmek için UniClust veri tabanı, yapısal profil matrislerini elde etmek içinse NR veritabanı kullanılmıştır. Tablo I'de özetlenen dört veri setinde HHBlits yönteminin iterasyon sayıları ilk aşamada 2, ikinci aşamada 1 olacak şekilde seçilmiştir. Bu veri setlerine ek olarak en iyi başarı oranı elde edilen veri seti olan uniclust_pssm_nr_ypm, HHBlits iterasyon sayıları iki aşamada da 3 olacak şekilde ayarlanarak yeniden oluşturulmuş ve beşinci veri seti uniclust_pssm_nr_ypm_3_3 olarak adlandırılmıştır.

TABLO I. DSPRED YÖNTEMİNİN ÖZNİTELİK VEKTÖRLERİNİ HESAPLAMAK İÇİN KULLANILAN VERİTABANI KOMBİNASYONLARI

Veriseti \ Öznitelik	***HHMAKE PSSM***	***YPM***
nr_pssm_nr_ypm	NR	NR
nr_pssm_uniclust_ypm	NR	UniClust
uniclust_pssm_nr_ypm	UniClust	NR
uniclust_pssm_uniclust_ypm	UniClust	UniClust

III. UYGULAMA

CB513 veri setinde 7-katlı çapraz doğrulama (cross-validation) yapılmıştır. Bunun için 7 farklı eğitim ve test seti oluşturulmuştur. İlk olarak PSI-BLAST [17] algoritması kullanılarak PSI-BLAST PSSM öznitelikleri çıkarılmıştır. Daha sonra HHBlits [18] yöntemi kullanılarak HHMAKE PSSM ve yapısal profil matrisleri hesaplanmış ve öznitelik vektörleri ve veri setleri bu matrislerden yukarıda açıklandığı gibi oluşturulmuştur. Veri setleri oluşturulduktan sonra DBA modelleri çapraz doğrulama deneyinin eğitim kümelerinde eğitilmiştir. DSPRED'in ilk aşamasında hesaplanması gereken Dağılım 1, 2, ve 3'ü eğitim kümeleri üzerinde tahmin edebilmek ve bu tahminleri destek vektör makinasında girdi olarak kullanabilmek için her eğitim kümesi üzerinde 2 katlı-çapraz doğrulama deneyi yapılmıştır. Bu sayede aşırı uyum davranışına takılmadan DVM modeli eğitilebilmiştir. Bu çalışmada DBA modelleri için GMTK programı [24] ve DVM modeli içinse libSVM [25] programı kullanılmıştır. DVM modelinde daha önce Aydın ve diğerleri tarafından optimize edilen gama parametresi 0.00781, C parametresi ise 1 olarak kullanılmıştır [3].

Tablo II sınıflandırma deneyleri ile elde edilen deney sonuçlarını göstermektedir. Sonuçlar incelediğinde en iyi başarı oranının uniclust_pssm_nr_pm veri seti ile elde edildiği görülmektedir. Bu verisetindeki başarı DSPRED yönteminde standart olarak kullanılan nr_pssm_nr_ypm veri setindeki başarıya kıyasla %0.28 daha yüksektir. Bu farkın istatistiksel olarak anlamlı olup olmadığını ölçmek için Z-test yöntemi kullanılmıştır [26]. Elde edilen iyileşme Tek taraflı Z-test'e göre %93 güven seviyesinde istatistiksel olarak anlamlıdır. Bu iyileşmenin sebebi olarak farklı ve heterojen bilgi kaynakları kullanılarak eğitilen modellerin birbirini tamamlayıcı nitelikte olması gösterilebilir. İterasyon sayısının başarı oranına etkisini ölçmek için oluşturulan uniclust_pssm_nr_pm_3_3 veri seti ile elde edilen sonuçlara bakıldığında ise hedef proteini NR veri tabanına hizalayarak PSSM ve yapısal matrisleri elde eden veri setine göre daha iyi sonuçlar almasına rağmen iterasyon sayısı

ilk aşamada 2 ikinci aşamada 1 olan veri setine göre daha düşük sonuçlar almıştır. Bunun sebebi iterasyon sayısı arttıkça hizalama sonuçlarında yanlış eşleşmelerin artması olabilir.

TABLO II. CB513 VERİ SETİ İÇİN DESTEK VEKTÖR MAKİNELERİ İLE ELDE EDİLEN 7 KAT ÇAPRAZ DOĞRULAMA DENEY SONUÇLARI

Veri seti	*Acc (%)*	*SOV(%)*	*MCC 'H'*	*MCC 'E'*	*MCC 'L'*
nr_pssm_nr _ypm	82.78	78.33	0.80	0.73	0.67
nr_pssm _uniclust_ypm	81.66	77.27	0.80	0.70	0.65
uniclust_pssm _nr_ypm	**83.06**	**78.66**	**0.81**	**0.74**	**0.67**
uniclust_pssm _uniclust_ypm	81.95	77.45	0.80	0.71	0.65
uniclust_pssm _nr_ypm_3_3	82.84	78.32	0.80	0.74	0.67

IV. SONUÇLAR

Bu çalışmada DSPRED yöntemi ile protein ikincil yapı tahmini yapılmıştır. CB513 setindeki proteinlerin özniteliklerini hesaplamak için dizi hizalama aşamasında farklı veri tabanları kullanılarak ve HHBlits yönteminin iterasyon sayısı da değiştirilerek 5 veri seti oluşturulmuş ve bu setler üzerinde eğitilen modellerin tahmin başarısı karşılaştırılmıştır. Sonuçlar incelendiğinde HHBlits yönteminin ilk aşamasında hedef proteini UniClust veri tabanına hizalayarak hesaplanan PSSM ve hedef proteini NR veri tabanına hizalayarak hesaplanan yapısal profil matrisleri kullanıldığında elde edilen başarı oranının en yüksek olduğu gözlemlenmiştir. Bu nedenle PİYT için DSPRED yöntemi kullanılarak oluşturulacak bir veri setinde HHBlits yönteminin ilk aşamasında hedef proteini NR veri tabanı yerine UniClust veri tabanına hizalayarak hesaplanan PSSM değerlerinin kullanılmasının ve iterasyon sayısının ilk aşamada 2, ikinci aşamada ise 1 olarak ayarlanmasının uygun olacağı anlaşılmaktadır.

KAYNAKLAR

[1] G. Pollastri ve A. McLysaght, "Porter: a new, accurate server for protein secondary structure prediction", Bioinformatics, c. 21, sayı 8, ss. 1719–1720, Nis. 2005.

[2] D. T. Jones, "Protein secondary structure prediction based on position-specific scoring matrices11Edited by G. Von Heijne", J. Mol. Biol., c. 292, sayı 2, ss. 195–202, Eyl. 1999.

[3] 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.

[4] X.-Q. Yao, H. Zhu, ve Z.-S. She, "A dynamic Bayesian network approach to protein secondary structure prediction", BMC Bioinformatics, c. 9, s. 49, Oca. 2008.

[5] J. Martin, J.-F. Gibrat, ve F. Rodolphe, "Analysis of an optimal hidden Markov model for secondary structure prediction", BMC Struct. Biol., c. 6, s. 25, Ara. 2006.

[6] B. Yang, Q. Wu, Z. Ying, ve H. Sui, "Predicting protein secondary structure using a mixed-modal SVM method in a compound pyramid model", Knowl.-Based Syst., c. 24, sayı 2, ss. 304–313, Mar. 2011.

[7] M. H. Zangooei ve S. Jalili, "Protein secondary structure prediction using DWKF based on SVR-NSGAII", Neurocomputing, c. 94, ss. 87–101, Eki. 2012.

[8] W. Yang, K. Wang, ve W. Zuo, "A fast and efficient nearest neighbor method for protein secondary structure prediction", içinde 2011 3rd International Conference on Advanced Computer Control, 2011, ss. 224–227.

[9] 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.

[10] A. R. Johansen, C. K. Sønderby, S. K. Sønderby, ve O. Winther, "Deep Recurrent Conditional Random Field Network for Protein Secondary Prediction", içinde Proceedings of the 8th ACM International Conference on Bioinformatics, Computational Biology,and Health Informatics, New York, NY, USA, 2017, ss. 73–78.

[11] X. Pan, P. Rijnbeek, J. Yan, ve H.-B. Shen, "Prediction of RNA-protein sequence and structure binding preferences using deep convolutional and recurrent neural networks", bioRxiv, s. 146175, Haz. 2017.

[12] L. Zheng, H. Li, N. Wu, ve L. Ao, "Protein Secondary Structure Prediction Based on Deep Learning", DEStech Trans. Eng. Technol. Res., c. 0, sayı ismii, 2017.

[13] 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.

[14] 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, s. 201, Haz. 2007.

[15] 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.

[16] 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.

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

[18] 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.

[19] "RefSeq: NCBI Reference Sequence Database", https://www.ncbi.nlm.nih.gov/refseq/, 2018.

[20] M. Mirdita, L. von den Driesch, C. Galiez, M. J. Martin, J. Söding, ve M. Steinegger, "Uniclust databases of clustered and deeply annotated protein sequences and alignments", Nucleic Acids Res., c. 45, sayı D1, ss. D170–D176, Oca. 2017.

[21] "RCSB Protein Data Bank - RCSB PDB", https://www.rcsb.org/pdb/home/home.do, 2018.

[22] J. A. Cuff ve G. J. Barton, "Evaluation and improvement of multiple sequence methods for protein secondary structure prediction", Proteins Struct. Funct. Bioinforma., c. 34, sayı 4, ss. 508–519, Mar. 1999.

[23] 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.

[24] J. Bilmes, ve G. Zweig, "The Graphical Models Toolkit: An Open Source Software System for Speech and Time-Series Processing", Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing, 2002.

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

[26] Z-test for comparing two proportions: https://onlinecourses.science.psu.edu/stat414/node/268