KIRIKKALE ÜNİVERSİTESİ

FEN BİLİMLERİ ENSTİTÜSÜ

BİLGİSAYAR MÜHENDİSLİĞİ ANABİLİM DALI

DOKTORA TEZİ

KÜMELENMİŞ PROTEİN DİZİLERİ KULLANARAK YAPISAL ÖZELLİK TAHMİNİ YAPAN YÖNTEMLERE ÖZELLİK VEKTÖRÜ TASARLAMAK

Sema ATASEVER

ARALIK 2019

Bilgisayar Mühendisliği Anabilim Dalında Sema ATASEVER tarafından hazırlanan KÜMELENMİŞ PROTEİN DİZİLERİ KULLANARAK YAPISAL ÖZELLİK TAHMİNİ YAPAN YÖNTEMLERE ÖZELLİK VEKTÖRÜ TASARLAMAK adlı Doktora Tezinin Anabilim Dalı standartlarına uygun olduğunu onaylarım.

Doç. Dr. Atilla ERGÜZEN
Anabilim Dalı Başkanı

Bu tezi okuduğumu ve tezin **Doktora Tezi** olarak bütün gereklilikleri yerine getirdiğini onaylarım.

Dr. Öğr. Üyesi Zafer AYDIN — Ortak Danışman

Prof. Dr. Hasan ERBAY — Danışman

Jüri Üyeleri

Başkan : Doç. Dr. Celal ÖZTÜRK _____

Üye (Danışman) : Prof. Dr. Hasan ERBAY _____

Üye : Dr. Öğr. Üyesi Bülent Gürsel EMİROĞLU _____

Üye : Dr. Öğr. Üyesi Ebubekir KAYA _____

Üye : Dr. Öğr. Üyesi Hakan KÖR _____

13.12.2019

Bu tez ile Kırıkkale Üniversitesi Fen Bilimleri Enstitüsü Yönetim Kurulu Doktora derecesini onaylamıştır.

Prof. Dr. Recep ÇALIN
Fen Bilimleri Enstitüsü Müdürü

ÖZET

KÜMELENMİŞ PROTEİN DİZİLERİ KULLANARAK YAPISAL ÖZELLİK TAHMİNİ YAPAN YÖNTEMLERE ÖZELLİK VEKTÖRÜ TASARLAMAK

ATASEVER, Sema
Kırıkkale Üniversitesi
Fen Bilimleri Enstitüsü
Bilgisayar Mühendisliği Anabilim Dalı, Doktora tezi
Danışman: Prof. Dr. Hasan ERBAY
Ortak Danışman: Dr. Öğr. Üyesi Zafer AYDIN
Aralık 2019, 56 sayfa

Protein yapıları ve fonksiyonları için her yıl büyük miktarlarda veri üretilmektedir. Elde edilen bu bilgilerin oluşturduğu protein veri tabanları modern biyolojinin önemli bir parçasıdır. Boyutları sürekli olarak artan bu büyük boyutlu veri tabanları ile Destek Vektör Makinesi (SVM) eğitimi karesel optimizasyon nedeniyle uzun zaman almaktadır. Bu problem durumu ile başa çıkabilmek için bu tez çalışmasında, tahmin başarısını azaltmadan mümkün olduğunca eğitim veri kümesini azaltarak eğitim sürecini kısaltmaya yarayacak yöntemler denenmiştir. Çalışmamızda, eğitilerek optimize edilen Dinamik Bayes Ağı (DBN) ve SVM kullanan iki aşamalı hibrit bir sınıflandırıcının (DSPRED), protein ikincil yapı tahmini için gelişmiş tahmin doğruluğu sağladığı gösterilmiştir. SVM eğitiminde kullanılacak olan veri kümesindeki örnek sayısını azaltmak için 7 kat çapraz doğrulama uygulanmış CB513 veri kümesi üzerinde iki farklı yöntem denenmiştir. Tabakalı örnekleme seçim stratejisinin kullanıldığı ilk yöntemde, eğitim veri kümesinden değişen oranlarda rastgele ve eşsiz veri örnekleri seçilmiştir. Sonuç olarak veri örneklerinin %50'si atılsa bile doğruluk oranını önemli ölçüde azaltmadan, model eğitim süresinde ortalama %73,38'lik bir iyileşme söz konusu olmuştur. İkinci yöntem, eğitim süresinin iyileştirilmesi amacıyla, veri örneklerini hiyerarşik bir kümeleme algoritması ile sınıflandırarak eğitim veri kümesindeki örnekleri küme merkezine en yakın komşularıyla değiştirmektedir. Öznitelik vektörlerini kümelemek için, validasyon

setindeki tahmin doğruluğunu hesaplayarak, küme sayısı ve en yakın komşu sayısı gibi hiper parametrelerin optimize edildiği hiyerarşik kümeleme yöntemi uygulanmıştır. Sonuç olarak, ikinci yöntemde tahmin doğruluğunu azaltmadan eğitim veri kümesinin %26 oranında azaltılabileceği sonucu elde edilmiştir. Kullanılan hiyerarşik kümeleme teknikleri arasında ward yönteminin en iyi kümeleme sonucunu sağladığı gözlenmiştir.

Anahtar Kelimeler: Protein İkincil Yapı Tahmini, Destek Vektör Makinesi, Bayes Ağı, Tabakalı Örnekleme, Hiyerarşik Kümeleme.

ABSTRACT

DESIGNING FEATURE VECTOR FOR METHODS WHICH PREDICT PROTEIN STRUCTURE BY USING CLUSTERED PROTEIN SEQUENCES

ATASEVER, Sema

Kırıkkale University

Graduate School of Natural and Applied Sciences

Department of Computer Engineering, Ph. D. Thesis

Supervisor: Prof. Dr. Hasan ERBAY

Co-Supervisor: Asst. Prof. Dr. Zafer AYDIN

December 2019, 56 pages

Large amounts of data regarding protein structures and functions are being produced each year, and the protein databases gathered through these data form an important part of modern biology. Support vector machine training with these large-sized databases, which are constantly increasing in size, takes a long time due to quadratic optimization. In order to cope with this problem, the methods which would be helpful to shorten the training time were used by reducing the educational dataset as much as possible without reducing the accuracy of the prediction. In our study, it was revealed that a two-stage hybrid classifier using a trained and optimized Dynamic Bayesian Network (DBN) and a Support Vector Machine (SVM) provided improved prediction accuracy for protein secondary structure prediction. In order to reduce the number of samples in the dataset to be used in support vector machine training, two different methods were tested on CB513 dataset with 7-fold cross validation. In the first method stratified sampling strategy was used, and unique samples were selected randomly and in varying ratios from the training dataset. As a result, in the case of discarding 50% of data samples, there was approximately 73.38% improvement in model training time without a significant reduction in accuracy. The second method classifies the data samples through a hierarchical clustering algorithm in order to improve the training time and replaces the samples in the training dataset with the neighbors closest to the cluster center. For clustering feature vectors, the hierarchical clustering method, which

requires the optimization of hyper parameters like number of clusters and number of nearest neighbors by calculating the accuracy of prediction in the validation set, was employed. With regard to the second method the results indicated that the training dataset could be decreased by 26% without reducing the accuracy of prediction. Among the hierarchical clustering techniques used, it was observed that the ward method provided the best clustering result.

Key Words: Protein Secondary Structure Prediction, Support Vector Machine, Bayesian Network, Stratified Sampling, Hierarchical Clustering.

TEŞEKKÜR

Tezimin hazırlanması esnasında zamanını, emeğini ve desteğini esirgemeyen danışmanlarım Sayın Prof. Dr. Hasan ERBAY'a ve Dr. Öğr. Üyesi Zafer AYDIN'a teşekkür ederim.

Tezimin hazırlanması sürecindeki öneri ve katkılarından dolayı Doç. Dr. Celal ÖZTÜRK'e, tezimi okuyup önerilerde bulunan Dr. Öğr. Üyesi Ebubekir KAYA ve Arş. Gör. Dr. Nuh AZGINOĞLU'na, tez özetinin İngilizcesini düzenleyen Dr. Yelda Sarıkaya ERDEM'e ve Mühendislik Mimarlık Fakültesi'ndeki yöneticilerime ve çalışma arkadaşlarıma teşekkürü borç bilirim.

Yol arkadaşım, eşim Umut'a, en değerlim biricik oğlum Ömer'e, annem, babam ve kardeşlerime her zaman ve her koşulda yanımda oldukları için teşekkür ederim.

Bu çalışma, Türkiye Bilimsel ve Teknolojik Araştırma Kurumu (TÜBİTAK) 3501 Kariyer Geliştirme Programı 113E550 sözleşme numaralı proje ile desteklenmiştir. Bu destek için TÜBİTAK'a teşekkür ederim.

Bu araştırmada yer alan kısmi nümerik hesaplamalar TÜBİTAK ULAKBİM, Yüksek Başarım ve Grid Hesaplama Merkezi'nde (TRUBA kaynaklarında) gerçekleştirilmiştir.

İÇİNDEKİLER DİZİNİ

Sayfa

ŞEKİLLER DİZİNİ

ÇİZELGELER DİZİNİ

KISALTMALAR DİZİNİ

BLAST	Basic Local Alignment Search Tool Temel Yerel Hizalama Arama Aracı
DBN	Dynamic Bayesian Network – Dinamik Bayes Ağı
DNA	Deoxyribonucleic Acid - Deoksiribo Nükleik Asit
DSSP	Dictionary of Secondary Structure of Proteins Protein İkincil Yapı Sözlüğü
E-VALUE	Expectation Value – Beklenen Değer
FSSP	Families Of Structurally Similar Proteins – Yapısal Olarak Benzer Protein Aileleri
HMM	Hidden Markov Model – Saklı Markov Model
MCA	Multiple Sequence Alignment – Çoklu Dizi Hizalama
MCL	Markov Clustering – Markov Kümeleme
MLP	Multi-Layer Perceptron – Çok Katmanlı Algılayıcı
NCBI	The National Center for Biotechnology Information - Ulusal Biyoteknoloji Bilgi Merkezi
NMR	Nuclear Magnetic Resonance - Nükleer Manyetik Rezonans
NR	Non-redundant Protein Sequence Database Yinelenmeyen Protein Dizi Veri Tabanı
PCA	Principal Components Analysis - Temel Bileşenler Analizi
PDB	Protein Data Bank - Protein Veri Bankası
PSI BLAST	Position-Specific Iterative Basic Local Alignment Search Tool Pozisyona Özel İteratif Temel Yerel Hizalama Arama Aracı
PSSM	Position-Specific Scoring Matrix Pozisyona Özel Profil Matrisi
PSSP	Protein Secondary Structure Prediction – Protein İkincil Yapı Tahmini
RBF	Radial Basis Function - Radyal Temelli Çekirdek Fonksiyonu

RMSD	Root Mean Square Deviation Sapmaların Ortalama Kare Kökü
RNA	Ribonucleic Acid - Ribo Nükleik Asit
SCOP	Structural Classification of Proteins Proteinlerin Yapısal Sınıflandırması
SCPS	Spectral Clustering of Protein Sequences Protein Dizilerinin Spektral Kümelenmesi
SVM	Support Vector Machine - Destek Vektör Makinesi

1. GİRİŞ

Çok sayıda amino asit biriminden oluşan, hücrenin her kısmında bulunabilen en yaygın biyolojik makro moleküllerden olan proteinler bütün canlı sistemlerde bulunurlar ve biyolojik katalizör, taşıyıcı molekül, mekanik destek ve bağışıklık koruması, sinir uyarılarının iletilmesi gibi geniş, işlevsel biyolojik fonksiyonlara sahiptirler [1,2]. Hücredeki en önemli yapısal, enzimatik, taşıma ve düzenleme işlevlerini yerine getirirler [3].

Her biri bir kovalent peptit bağı vasıtasıyla komşusuna bağlanan amino asitler uzun bir zincir yapısı oluştururlar. Bu zincir yapısı, fonksiyonel bir protein olarak katlanabilmektedir. Amino asitlerin birbirleriyle etkileşimlerinin bir sonucu olarak, her bir protein belirli bir üç boyutlu şekle veya konformasyona katlanır [4]. Elde edilen bu benzersiz üç boyutlu konformasyon, zincirdeki amino asitlerin diziliş sırasına, yani birincil yapılarına göre belirlenir [5]. Proteinin birincil yapısını kullanarak üçüncül yapısının tahmin edilmesi problemi moleküler biyolojideki en büyük zorluklardan biri olmuştur. Bu hedefe (katlanmış şekil veya konformasyon) ulaşmak için ikincil yapı tahmini gibi protein yapısının farklı seviyelerini ve yönlerini hedef alan ara adımlar kullanılmıştır [6].

Proteinlerin amino asit dizileri en yaygın kullanılan biyolojik bilgi türüdür [3]. Öyle ki, ilk protein yapıları X-ışını kristalografisi ile çözüldüğünden bu yana, proteinlerin ikincil yapısını (α-sarmal, β-iplik gibi) amino asit dizilerinden tahmin etmek için bu süreçte pek çok çalışma gerçekleştirilmiştir [7].

Çizelge 1.1.'de listelenen 20 çeşit amino asit vardır. Amino asitler serbest halde ya da peptit ve proteinlerde doğrusal zincirler halinde bulunurlar [8].

Çizelge 1.1. Amino asit çeşitleri [8]

	Sembol		
Amino Asit	**3-Harfli Kod**	**1-Harfli Kod**	**İnsanlar için gerekli mi?**
Asidik amino asitler			
Aspartik asit	ASP	D	Hayır
Glutamik asit	GLU	E	Hayır
Nötr amino asitler			
Alanin	ALA	A	Hayır
Asparajin	ASN	N	Hayır
Sistein	CYS	C	Hayır
Glutamin	GLN	Q	Hayır
Glisin	GLY	G	Hayır
Izolösin	ILE	I	Evet
Lösin	LEU	L	Evet
Metiyonin	MET	M	Evet
Fenilalanin	PHE	F	Evet
Serin	SER	S	Hayır
Treonin	THR	T	Evet
Triptofan	TRP	W	Evet
Tirozin	TYR	Y	Hayır
Valin	VAL	V	Evet

Çizelge 1.1. (devam) Amino asit çeşitleri [8]

Basit amino asitler			
Arjinin	ARG	R	Evet
Histidin	HIS	H	Evet
Lizin	LYS	K	Evet
İmino asit			
Prolin	PRO	P	Hayır

Canlı hücrelerde gerçekleşen dinamik süreçlerin altında yatan olağanüstü fonksiyonların gerçekleşmesi için pek çok proteinin görevlerini etkili bir şekilde yerine getirmeleri gerekmektedir. Bir protein ancak doğru bir üç boyutlu yapıya sahip olduğunda etkili bir şekilde çalışabilir [9].

Pek çok proteinin amino asit dizisi kendi aralarında bir dereceye kadar benzerliğe sahiptir [10]. Birbirine yakın amino asit dizilimine sahip proteinlerin yapıları dolayısıyla yerine yetirdikleri fonksiyonlar da benzer olacağından [11], fonksiyonu bilinen proteinler ile fonksiyonu bilinmeyen proteinler arasındaki yapısal benzerliklere göre tahmin yürütülür [12].

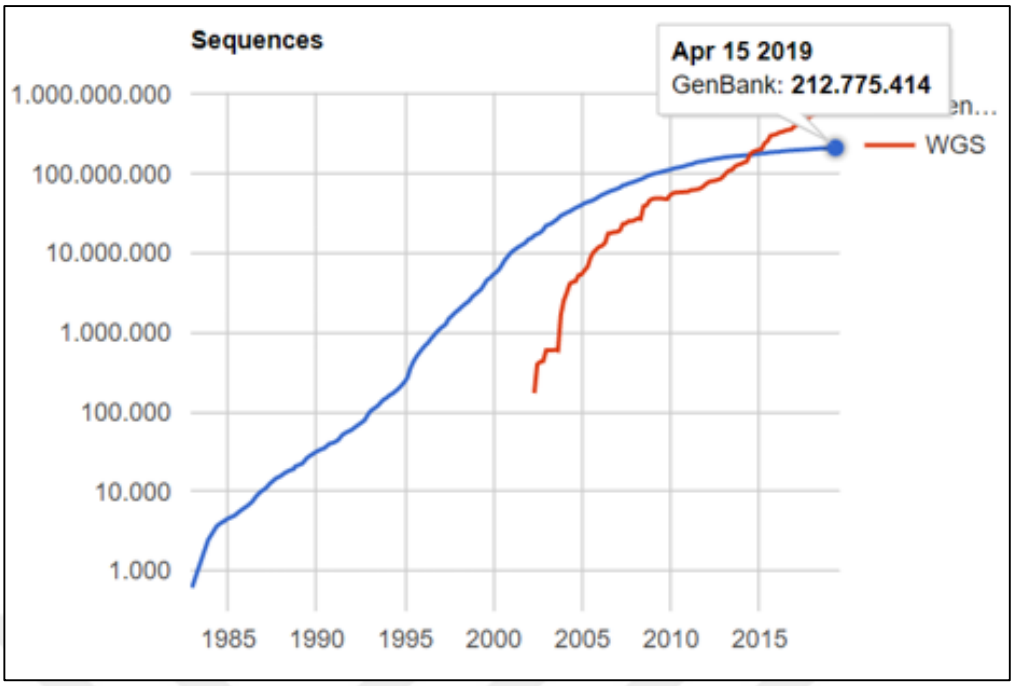
Bir proteinin biyolojik olarak aktif olabilmesi, spesifik bir üç boyutlu şekle sahip olacak şekilde katlanmasını gerektirmektedir [13]. Genetik ve çevresel faktörlerin etkisi ile, Alzheimer, Parkinson, Huntington gibi nörodejeneratif rahatsızlıklar ve Tip II diyabet gibi çeşitli sistemik bozukluklar da dahil olmak üzere proteinlerin yanlış katlanması ile ilişkili pek çok rahatsızlık mevcuttur [14]. Proteinlerin birincil yapı bilgisinden yararlanarak ne şekilde katlanacağını yüksek bir doğrulukla tahmin edebilirsek, amino asit dizisindeki mutasyonlar sonucu ya da normal katlanma işleminin çevresel faktörler tarafından bozulması gibi nedenlerle proteinlerin hatalı katlanmasına bağlı olduğuna inanılan Alzheimer, Parkinson, Huntington gibi çok sayıda hastalığın [15] hangi koşullar altında başladığı, nasıl engellenebileceği gibi genel olarak uygulanabilecek çözümleri bulmak kolaylaşacaktır [16]. Moleküler düzeyde meydana gelen protein katlanma problemi biyofizik ve biyokimyanın en

zorlayıcı konularından birisi olagelmiştir [16]. Proteinler sentezlenmeleri tamamlandıktan sonra işlevlerini yerine getirebilmek için kendilerine has üç boyutlu yapılarına katlanırlar. Proteinlerin bu üç boyutlu yapılarının belirlenmesinde kullanılan en yaygın iki deneysel yöntem: X-ışını kristalografisi ve Nükleer Manyetik Rezonans (NMR) spektroskopisidir [17]. Kullanılan bir diğer deneysel yöntem ise elektron mikroskobu yöntemidir. Bu yöntemden elde edilen sonuçlar ile X-ışını kristalografisiden elde edilen sonuçlar birleştirilerek, proteinin fonksiyonunu anlama çabasında beraber kullanılabilirler. Bu sayede, farklı tekniklerden gelen verilerin birleştirilmesi ile daha anlamlı sonuçlar elde edilebilir [18]. Protein kristalizasyonu uygulaması zor bir işlemdir [19] ayrıca bütün proteinler deneysel olarak moleküler yapılarının belirlenmesine elverişli değildir [20]. Dolayısıyla, x-ışını kristalografisi ve NMR spektroskopisi gibi adı geçen bu deneysel yöntemlerin uygulanması zaman almaktadır ve maliyetlidirler [21]. Bu sorun için olası bir çözüm, deneysel yöntemlerden çok daha hızlı ve ucuz olacak olan hesaplamalı yöntemler geliştirmektir. Bu olası çözüm özellikle ilaç sektörü için önem arz etmektedir. Çünkü yeni ilaçların tasarım sürecinde protein yapısının hızlı bir şekilde tahmin edilmesi, tek bir ilacın geliştirilmesinde gerekli olan zahmetli ve pahalı deneylerden tasarruf edilmesi beklenmektedir [22].

Biyoinformatikte araştırmacıların kullanımına sunulan, kapsam ve amaç bakımından özelleştirilmiş çeşitli veri tabanları bulunmaktadır. Bu veri tabanlarının bazı dezavantajları mevcuttur. Bunlar, fazla bilgi içermeleri, birden fazla veri tabanına yayılmış verinin varlığı, eksik bilgi, veri tabanı adlandırma kurallarının açıkça tanımlanmamış olmasıdır. Bu ve bunun gibi nedenler bu veri tabanlarından anlamlı bilgi çıkarımını zorlaştırmaktadır [23]. İnsan Genom Projesi gibi çalışmalardan elde edilen ve çeşitli biyolojik sistemlerdeki bilgileri temsil eden biyolojik veriler petabyte (PB) hatta exabyte (EB) boyutlarına ulaşmıştır [24]. Bu ve bunun gibi dünyanın dört bir yanındaki araştırma gruplarının çalışmalarından elde edilen biyolojik veriler, içerdikleri bilgi türüne göre birincil, ikincil ya da kompozit olmak üzere farklı sınıflara ayrılan çok sayıda biyoinformatik veri tabanlarında arşivlenmektedirler. Veri tabanlarında arşivlenen bu veriler organize edilip, erişime açılmıştır [25]. Böylece erişime açılan bu veriler, araştırmacılara geniş bir biyolojik bilgi hazinesi sunarak, onları bu verilerden çeşitli çıkarımlar yapacak yeni araçlar geliştirmeye zorlamıştır [3].

Birincil veri tabanları deneysel olarak türetilmiş yapısal verileri içermektedir. Swiss-Prot, UniProt, PIR, GenBank, EMBL, DDBJ ve Protein Veri Bankası (PDB) birincil veritabanlarındandır [25]. Bu veritabanlarından UniProt, Swiss-Prot, TrEMBL ve PIR-PSD diğer birçok veritabanından veri toplayan kapsamlı bir havuzdur [26].

Ulusal Biyoteknoloji Bilgi Merkezi (NCBI) web sitesi üzerinden halka açık erişimli GenBank® Nisan 2019 tarihi itibariyle 212 milyonu aşkın protein dizisi (Şekil 1.1.) içermektedir [27]. Dizi sekans analizinde merkezi bir rol oynayan ve genom araştırmacıları tarafından rutin olarak kullanılan [28] GenBank, 1982'deki 3. sürümden bu yana, yaklaşık her 18 ayda bir iki katına çıkarak üssel bir oranda büyümektedir [27]. Bu rutin kullanımın sonucu olarak, her nükleotit sekansının erişime açılmadan önce araştırmacılar tarafından GenBank'a girilmesi ve yapılan yeni araştırma sonuçlarının varlığı gibi nedenlerle bu veri tabanı hızla büyümeye devam etmektedir [28]. Yine biyoinformatiğin sık kullanılan veri bankalarından bir diğeri olan PDB, proteinler, peptitler ve nükleik asitler gibi biyolojik makro molekülerin deneysel olarak belirlenmiş üç boyutlu (3D) yapılarına ait koordinatların halka açık olarak saklandığı dünyadaki tek arşivdir [29]. Ağustos 2019 yılı itibariyle PDB [30] veri bankası, yılda yaklaşık 8,700 artışla yaklaşık 143,000 yapı ihtiva etmektedir [31]. Bu veri bankasına her yıl binlerce yeni yapı sunulduğundan büyümeye devam etmektedir ve sürekli güncellenmektedir [32]. Bahsedilen veri tabanlarında yer alan, yapısı bilinen binlerce proteinin varlığına rağmen hala deneysel olarak yapısı çözülememiş pek çok protein vardır [21]. Şöyle ki, GenBank biyolojik veri bankasında milyonlarca dizi sekansı yer alırken, yapısı bilinen makro moleküllerin arşivlendiği PDB veri bankasında sadece binlerce yapıya ait bilgi yer almaktadır. Bu durum deneysel yöntemlerin yetersiz kaldığı ve maliyetli olduğu, yapısı bilinmeyen proteinlerin yapı tahmininde çeşitli hesaplama ve makine öğrenmesi yöntemlerinin kullanımının kaçınılmaz olduğu sonucunu doğurmaktadır [33]. Protein ikincil yapı tahmini gibi bir boyutlu yerel yapı özelliklerinin tahmini için geliştirilen bu hesaplama ve makine öğrenmesi yöntemlerinden elde edilen doğruluk oranları günümüze kadar sürekli bir iyileşme göstermiştir.



Şekil 1.1. GenBank'ın yıllara göre büyüme grafiği [27]

Tıpkı gerçek hayatta uyguladığımız gibi, protein yapı tahmini probleminin çözümünde de, büyük ve karmaşık sorunların üstesinden daha kolay gelebilmek için böl ve yönet stratejileri (divide and conquer strategies) kullanılır [33]. Bu stratejide asıl problem daha küçük, çözülebilir alt problem parçalarına ayrılarak, çözüm arayışına gidilir, böylece asıl problemin çözümüne giden yolda önemli bir adım atılmış olur. Yan zincir konumlandırma (side chain positioning) [34], proteinlerin bir boyutlu yapısal özelliklerinin (örneğin protein ikincil yapısı) sekansa dayalı tahmini [35], bir proteinin polipeptit omurgasının yerel konformasyonu [33] gibi çalışma konuları, genel protein yapısının tahminine katkı sağlayan, en çok bilinen önemli alt problemlerdendir. Bu amaçla bu tez çalışmasında, bahsedilen alt problemlerden biri olan, bir boyutlu yapı tahmin elemanlarından ikincil yapı tahmini konusu çalışılarak, daha ayrıntılı üç boyutlu genel protein yapısının tahminine katkı sağlanması amaçlanmaktadır.

1.1. Literatür

Bu bölümde, ikincil yapı tahmini için SVM'leri kullanan literatür ve örneklem ile SVM'nin eğitim süresini kısaltmaya yönelik yöntemler öneren çalışmalar hakkında kısa bir derleme sunulmaktadır.

Lin ve arkadaşları [36], ikincil yapı tahmini performansını artırmak için bir çoklu SVM topluluğu (multi-SVM ensemble) önermişlerdir. Metotları iki katman içermektedir: ilk katman beş sınıflandırıcı topluluğu ve ikinci katman üç SVM tarafından oluşturulmuştur. Çoklu SVM topluluğu, bootstrap örneklemesi yoluyla eğitim veri kümesini yeniden örneklemek için bagging yöntemini kullanarak, RS126 veri kümesinde yedi kat çapraz doğrulama yapıp ikincil yapı tahmininde gelişmiş performans elde etmişlerdir.

Hua ve arkadaşları [37], SVM'ye dayanan yeni bir protein ikincil yapı tahmini yöntemi önererek, CB513 veri kümesinde yedi kat çapraz doğrulama ile % 73,5'lik üç durumlu tahmin doğruluğu (Q3) elde etmişlerdir.

Literatürde protein ikincil yapı tahmini için SVM kullanan birçok yayın olmasına rağmen, bu yayınlarda SVM'nin eğitim süresini iyileştirmek için eğitim veri kümesinin azaltılması yöntemi denenmemiştir. Bu nedenle diğer problemlerde SVM'nin model eğitim süresini iyileştiren yöntemlerden bahsedilecektir.

Jun [38], eğitim kümesinden örneklerin bir alt kümesini seçmek için tabakalı örnekleme (stratified sampling) yöntemini kullanmıştır. Çalışmasında yazar, her sınıftan örneklerin % 10'unu seçmiştir ki bu da eğitim kümesinin boyutunu 10 kat azaltmıştır. Ardından, azaltılmış veri kümesini kullanarak SVM'yi eğitmiştir. Yöntem, UCI Machine Learning Repository sitesinde yer alan dört veri kümesine uygulanmıştır. Adult ve iris veri kümelerinin %10'luk tabakalı örneklemeyle eğitilmiş modellerinin tahmin doğruluğu korunsa da, harf görüntü tanıma ve protein lokasyon site veri kümeleri için tüm örneklerin kullanılmasına kıyasla doğruluk oranları önemli ölçüde azalmıştır.

Bir başka çalışmada, Hens ve Tiwari [39] F-score'la özniteliklerin sayısını azaltarak, kredi puanlama probleminin hesaplama süresini azaltmak için tabakalı örnekleme stratejisini kullanmışlardır. Sonuçta kredi puanlama modeli için önerdikleri yeni yöntemin doğruluk oranının diğer yöntemlere göre daha rekabetçi ve daha az hesaplama süresine sahip olduğunu göstermişlerdir.

Örnekleme stratejilerine ek olarak, eğitim veri kümesinin örnek sayısını azaltmak için kümelemeyi kullanan yöntemler de vardır.

Awad ve arkadaşları [40], özellikle büyük veri kümeleri için SVM'nin eğitim süresini iyileştirmek için hiyerarşik bir kümeleme yaklaşımı kullanmışlardır. Model eğitimi için verimli çalıştıkları gösterilen TCT-SVM, TCTD-SVM ve OTC-SVM adlı üç teknik önermişlerdir. Bunların arasında TCT-SVM, doğruluk açısından diğerlerinden daha iyi performans göstermiştir ancak daha yüksek bir model eğitim süresine sahiptir.

Yu ve arkadaşları [41], yüksek sınıflandırma doğruluğuna sahip, büyük veri kümeleri için ölçeklenebilir bir kümeleme yöntemini bütünleştiren Kümeleme Tabanlı SVM (CB-SVM) adı verilen yeni bir yöntem önermişlerdir. Yazarlar, yapay ve gerçek veri setleri kullanarak CB-SVM algoritmasının performansını test etmişlerdir. Kümeleme bazlı eğitim örnekleriyle kullanılan CB-SVM'nin , aynı sayıda rastgele veri kümesi ile eğitilen standart SVM'den daha iyi performans gösterdiği sonucunu elde etmişlerdir.

Protein ikincil yapı tahmini (PSSP) çalışmaları, herhangi bir protein yapısı daha çözülmeden evvel, 1951 yılında Pauling ve Corey tarafından deneysel olarak başlamıştır [33]. O yıllardan günümüze kadar PSSP problemi çeşitli yöntemler ve algoritmalar kullanılarak üzerinde çalışan ve biyoinformatiğin önemli problemlerinden biri haline gelen bir konu olmuştur [33,42]. Bu problemin çözümünde kullanılan veri kümelerinden birisi de CB513'tür. Bu veri kümesi kullanılarak elde edilen protein yapı tahmini çalışmalarına ait sonuçlardan bazıları şu şekildedir :

Rashid ve arkadaşları [43] yaptıkları çalışmada CB513 veri kümesinden sezgisel tabanlı bir yaklaşım kullanarak 55 proteinlik bir veri kümesi seçmişlerdir. Bu veri kümesini Fully Complex-valued Relaxation Network (FCRN) sınıflandırıcısı ile eğitmişlerdir. Modelin performansını çapraz doğrulama ile değerlendirip, G Switch

proteinlerinin bir veri kümesi üzerinde test edip yaklaşık olarak %81 doğruluk oranı elde etmişlerdir. Özetle çalışmada bahsedilen modeli, literatürdeki bazı tekniklerle karşılaştırdıklarında daha iyi sonuçlar elde ettiklerini ifade etmişlerdir.

Wang ve arkadaşları [44], mevcut yöntemlerden farklı olarak, amino asitlerin fiziksel-kimyasal özelliklerini ve yapı özelliklerini hesaba katan SVM'ye dayalı yeni bir yöntem önermişlerdir. Önerdikleri bu yöntemi popüler veri kümelerinden biri olan CB513'te test ettiklerinde, Q3 doğruluğunu %78.4 olarak bulmuşlardır.

Aydın ve arkadaşları [45], üzerinde yedi kat çapraz doğrulama gerçekleştirdikleri, 513 protein zinciri ve 84,119 amino asit içeren, iyi bilinen ve zor bir kıyaslama dizi veri kümesi olan CB513'ü kullanarak her bir kalıntı için doğruluk (per-residue accuracy) değerini %80.3 olarak bulmuşlardır.

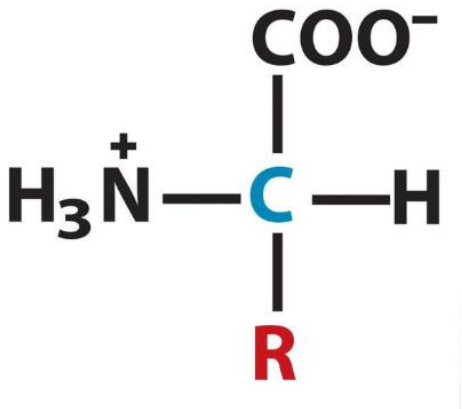
Bu tez çalışmasında örnek indirgeme yöntemleri, DSPRED [44] isimli iki aşamalı hibrit bir sınıflandırıcı ile kullanılarak protein ikincil yapı tahmini için gelişmiş tahmin doğruluğu elde edilmiştir. DSPRED [44] yönteminin ikinci aşaması olan SVM sınıflandırıcısının eğitim süresinin iyileştirilerek, protein ikincil yapı tahmini için gelişmiş tahmin doğruluğu elde edilmesi amaçlanmıştır. Çalışmaya ait elde edilen ilk sonuçlar 3rd World Conference on Big Data [46] isimli uluslararası konferansta tanıtılarak, bildiri özeti şeklinde sunulmuştur. Elde edilen nihai sonuçlar SCIE kapsamında taranan Applied Sciences [47] isimli dergide yayınlanarak literatüre kazandırılmıştır.

1.2. PROTEİN YAPISI

Proteinler tüm hücrelerde bulunan, enzimatik, yapısal ve karmaşık rolleri olan en yaygın biyolojik makromoleküllerdendir [1,48]. Her biri kendine özgü amino asit dizisine sahip binlerce farklı tipte protein bilinmektedir ve bu proteinler hemen hemen her sürece aracılık ederler. Proteinler, her biri bir kovalent peptit bağıyla komşusuna bağlanarak uzun bir amino asit zincirinden meydana gelirler [4]. Bazı amino asitler proteinlerde diğer amino asitlerden daha fazla miktarda bulunur. Örneğin sistein, triptofan ve metiyonin protein yapısında nadir bulunan amino asitlerden iken, lösin, serin, lisin ve glutamik asit ise protein yapısında en bol bulunan amino asitlerdendir [9].

Şekil 1.2.'de gösterildiği üzere tüm amino asitler bir amino grubuna, bir karboksil grubuna ve bir hidrojen atomuna bağlanmış, alfa karbon olarak adlandırılan merkezi bir karbon atomundan oluşan bir yapıya sahiptir [22].

Bir amino asit yapısındaki α-karboksil grubunun başka bir amino asit yapısındaki α-amino grubuna peptit bağı ile bağlanmasıyla polimer yapı oluşur. İki amino asittin oluşturduğu bir dipeptit oluşumunda ise bir su molekülünün kaybı olur [2]. Peptit bağları ile birbirine bağlanan amino asitler polipeptit zincirleri oluştururlar. Polipeptitler ise üç boyutlu uzayda katlanarak serbest enerjilerinin en aza indirildiği yapıyı oluştururlar [4].

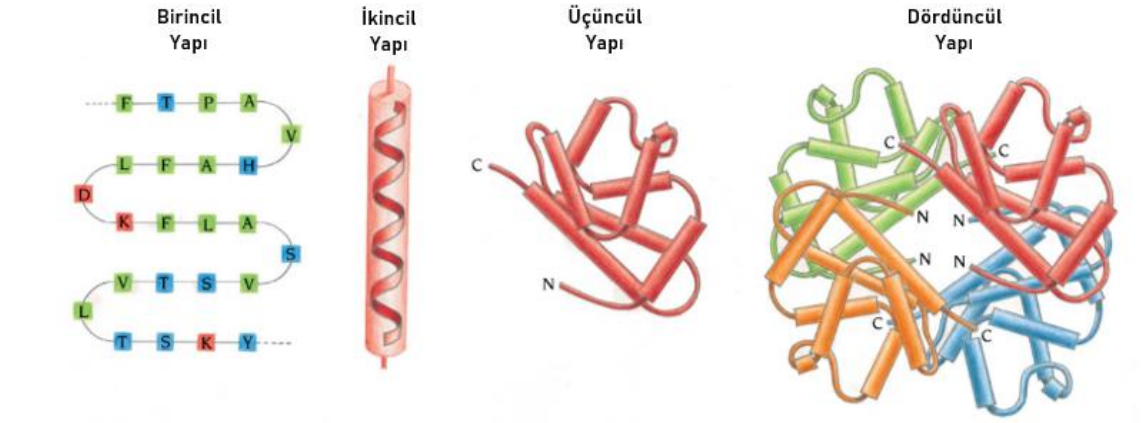


Şekil 1.2. Amino asitin genel yapısı [1]

1.3. Birincil Yapı

Protein yapısı, birincil, ikincil, üçüncül ve dördüncül (kuaterner) olarak adlandırılan, hiyerarşik bir şekilde organize edilmiş dört yapı düzeyinden oluşan karmaşık yapılar içermektedir [22] (Şekil 1.3.).

"Temel yapı" olarak adlandırılan bu hiyerarşinin ilk seviyesi, protein zincirini oluşturan sıralı amino asit dizisidir yani birincil yapıdır. Amino asitler kovalent peptit bağları ile birbirine bağlanarak polipeptit zincirleri oluştururlar. Proteinin amino asit dizilimi, proteinin üç boyutlu yapısını ve dolayısıyla fonksiyonunu belirlediği için önemlidir [22].



Şekil 1.3. Protein yapı düzeyleri [49]

1.4. İkincil Yapı

Birincil yapının üzerindeki protein hiyerarşisinin ikinci seviyesi, düzenli yerel alt yapılardır. Ana zincir karboksil ve amino grupları arasındaki hidrojen bağları ile stabilize edilmiş, α-sarmal, β-strand [2] gibi tekrarlayan ikincil yapı elemanlarının varlığı ile karakterize edilir [3]. Amino asit dizisi ve mevcut hücresel koşullar polipeptitin nasıl katlanacağını ve proteinin yapısını nasıl kazanacağını belirleyen faktörlerdendir [26].

Proteinin üçüncül yapısını, bir araya getirilen ikincil yapı elemanları olarak düşündüğümüzde, ikincil yapının öngörülmesinin, 3D yapı tahmini için önemli bir aşama olduğu sonucu ortaya çıkmaktadır [7].

1.5. Üçüncül Yapı

Genel zincir, hiyerarşinin üçüncü seviyesini oluşturan kompakt, üç boyutlu üçüncül bir yapıya katlanma eğilimindedir. Üçüncül yapı, proteinin en stabil şeklidir, çünkü zinciri oluşturan farklı amino asitler arasındaki çeşitli çekim kuvvetlerini optimize eder. Ayrıca, üçüncül yapı aynı zamanda proteinin biyolojik olarak aktif şeklidir ve bozulması, proteini kısmen veya tamamen inaktif hale getirir. Bu nedenle, üçüncül yapıya sıklıkla proteinin "doğal yapısı" da denilmektedir [22].

Üçüncül yapıları bilinen proteinler, nükleik asitler ve diğer biyolojik moleküllerin üç boyutlu yapıları PDB veri bankası arşivinde depolanmaktadır [26].

1.6. Dördüncül Yapı

Dördüncü seviye farklı polipeptit zincirlerinin yapısal organizasyonunu içerir [22]. Her polipeptit zincirine alt birim adı verilir. Genellikle değişken sayılarda olmak üzere birden fazla alt birim bulunabilir. Dördüncül yapı, alt birimlerin uzamsal düzenini ve etkileşimlerinin doğasını ifade eder [2].

2. BİYOİNFORMATİKTE KULLANILAN YÖNTEMLER

2.1. Dizi Hizalama

Protein dizi hizalaması moleküler biyologlar için önemli biyolojik bilgilerin analizlerinde kullanılan mevcut en güçlü hesaplama aracıdır. Her zaman başarılı olunamasa da bir dizinin bilinmeyen bir yapı ve işleve sahip olması durumunda, hem yapı hem de işlevde iyi karakterize edilmiş bir başka diziye olan hizalaması, birinci dizinin yapı ve işlevini öğrenmek için kullanılabilmektedir [50]. Yaklaşım, dizi çiftlerinin genel benzerlik skorunu maksimuma çıkaran, mümkün olan en uygun çözümü sunan dinamik programlama yöntemleri kullanılarak hizalanmasına ve biyolojik dizi uyumlarının çıkarılmasına dayanır [51]. Yaygın olarak kullanılan optimum global hizalama yöntemi Needleman-Wunsch algoritması [52] ile yerel dizi hizalamaları üretmek için kullanılan Smith-Waterman algoritması gibi dinamik programlama tekniğine dayanan yöntemler iki dizi arasındaki benzer bölgeleri belirlemek için kullanılmaktadır.

2.2. Çoklu Dizi Hizalama

DNA, RNA ve protein dizilerinin çoklu dizi hizalaması, birden fazla biyolojik dizinin dizi hizalamasıdır ve yaygın olarak kullanılan bir hesaplama prosedürüdür. Doğru bir şekilde elde edilen çoklu dizi hizalama sonuçları, farklı diziler arasındaki benzerliği gösteren faydalı bilgiler sağlayarak, biyolojik modelleme yöntemlerinin başarısına katkı sağlar [53]. Mevcut yöntemlerin hiçbiri biyolojik olarak mükemmel çoklu dizi hizalama sonuçlarını verememektedir. Biyolojik veri tabanlarındaki büyüme ile birlikte çoklu dizi hizalamalarının biyolojik araştırmalarda kullanımı gerekli bir araç haline gelmesine paralel olarak, doğru çoklu dizi hizalama sonuçları elde edebilmeyi sağlayan biyolojik araştırma alanlarında artış gözlenmektedir [54].

Çoklu dizi hizalama için kullanılan çeşitli temel yaklaşımlar mevcuttur. İlk yaklaşım iyi bir hizalama sonucu veren fakat maliyetli bir hesaplamaya neden olan kaba kuvvet metodunu (brute-force method) kullanır. Hesaplama süresi n dizi için O (Uzunluk^n)

zaman gerektirecektir [55,56], bu yüzden bu yaklaşımla çoklu dizi hizalamayı hesaplamak az sayıda dizi için kullanıldığında süre ve hafıza anlamında verimli olmaktadır. Bu yüzden bu yaklaşım yerine, ilerleyici hizalama algoritmaları (progressive alignment algorithms) gibi hesaplama yöntemleri kullanılabilir. Bu yaklaşımda, ilk işlem olası bütün ikili hizalamaların oluşturulması işlemidir. n sayıda dizi için, n x (n-1)/2 dizi çifti elde edilir. Elde edilen çift yönlü hizalamalar, aritmetik ortalama gibi mesafeye dayalı bir algoritma kullanarak filogenetik bir ağacı tahmin etmek için kullanılır ki bu kısım bu yaklaşımın en fazla zaman alan kısmını oluşturmaktadır. Tahmin edilen bu ağaç yapısına göre, en benzer diziler ikili bir algoritma kullanarak birbirine hizalanır. Yaygın kullanılan çoklu hizalama programlarından olan ClustalW ve Needleman-Wunsch algoritmasında bu yaklaşım kullanılmaktadır [54]. Umut vaat eden diğer bir yaklaşım ise protein ikincil yapı tahmini ve çoklu dizi hizalaması gibi problemlerde başarıyla kullanılan Saklı Markov Model (HMM) olasılık yaklaşımıdır [57].

2.3. HHblits

HHblits [58] , HMM kullanarak UniProt gibi büyük veri tabanlarında homolog dizileri iteratif olarak arayabilen yani uzak benzerlik tespiti yapabilen açık kaynaklı kodlu bir araçtır. HMM'ler, gen dizimi, çoklu dizi hizalaması ya da protein ikincil yapı tahmini gibi moleküler biyolojideki birçok farklı problemde kullanılmaktadır [26]. Bu yöntemle elde edilen hizalama profillerinin doğruluğu ve hassasiyeti protein fonksiyon ve yapı tahmininin doğruluğunu da kritik olarak etkilemektedir [58].

HHblits önce tek bir dizi ya da çoklu dizi sıralamasını bir HMM'ye dönüştürür ve iteratif olarak uniprot20 ya da NR20 gibi HMM veri tabanlarında aramalar yapar, bir önceki arama sonuçlarına anlamlı derecede benzeyen dizileri ekleyerek bir sonraki iterasyon için HMM sorgusunu günceller. PSI-BLAST ile kıyaslandığında, HHblits hızlı, hassas ve daha doğru hizalamalar üretir. HMM veri tabanını kullanarak iteratif aramalar yapan HHblits, açık kaynak kodlu HHsuite'in bir parçasıdır [58,59].

2.4. Kümeleme Analizi, Yöntemleri ve Algoritmaları

Kümeleme, etiketlenmemiş veriler arasındaki ilişkinin açıkça bilinmediği ilginç kalıpları bulmaya yarayan denetimsiz bir öğrenme yöntemidir [60]. Makine öğrenmesi, veri madenciliği, örüntü tanıma, görüntü analizi, biyoinformatik gibi bir çok alanda kullanılan güçlü bir öğrenme aracıdır [61,62].

Aynı küme içerisinde birbirine benzer özelliklere sahip iken, diğer kümelerdeki nesnelere çok benzemeyen özelliklere sahip öğelerin oluşturduğu bir gruptur küme. Kümeleme analizinde amaç, özelliklerine göre nesneleri gruplandırarak, gruplar arasındaki benzerlik ya da farkların belirlenmesine yönelik veri vektörünün alt kümelere bölünmesidir. Bir kümenin alt kümelere bölünmesi için kullanılan kriter genellikle kümeler içindeki elemanların nasıl benzer olduklarını ya da kümeler arasındaki elemanların nasıl farklı olduklarını ölçen benzerlik ve uzaklık kavramlarından yararlanılarak belirlenir [63]. Kümeleme analizi, farklı bilim dallarında da kullanılmakla birlikte, benzer ifade kalıplarına sahip gen gruplarını belirlemek gibi çeşitli çalışma konularıyla biyoinformatik alanında önemli bir uygulama alanına sahiptir [64].

Kümeleme algoritmalarının çoğu girdi olarak küme sayısını gerektirir ve veri kümesindeki tüm nesneler özelliklerine göre genellikle kümelerden birine atanır [65].

Kümeleme yöntemlerinin biyoinformatikteki kullanımı önem arz etmektedir. Biyolojik veri tabanlarının büyüklüğü ve karmaşıklığı içerilen veri kümesini anlama ve yorumlama zorluklarını beraberinde getirmektedir. Bu nedenle kümelenme tekniklerinin kullanılarak altta yatan verilerdeki ilginç kalıpları belirlemek bu zorlukların ele alınmasına yönelik ilk adımdır [66].

Çizelge 2.1.'de biyoinformatikte kullanılan farklı kümeleme algoritmalarının listesi yer almaktadır [67].

Çizelge 2.1. Biyoinformatikte kullanılan farklı kümeleme algoritmalarının listesi [67]

Algoritma ismi	**Algoritma tipi**
Mothur	Hiyerarşik
UCLUST	Greedy heuristic (Açgözlü sezgisel)
UPARSE	Greedy heuristic (Açgözlü sezgisel)
CD-HIT	Greedy heuristic (Açgözlü sezgisel)
ESPRIT	Hiyerarşik
ESPRIT-Tree	Hiyerarşik fakat ikili karşılaştırmalar kapsamlı değildir.
CROP	Bayes yaklaşımı
TSC	Adım 1: hiyerarşik 2: greedy heuristic (Açgözlü sezgisel)
M-pic	Modülerlik tabanlı
MSClust	Açgözlü sezgisel (Greedy heuristic)
SWARM	Kümelenmiş

Literatürde, kümeleme çalışmasının amacına ve kullanılacak veri tipine göre farklılıklar gösteren pek çok algoritma olmakla beraber, kümeleme algoritmaları aşağıdaki iki ana kategoriye ayrılabilir [68]:

2.4.1. Sıralı Algoritmalar

Bu algoritmalar basit, hızlı ve anlaşılması kolay yöntemlerdir. K-ortalama algoritması kümeleme problemlerinin çözümünde kullanılan en basit denetimsiz öğrenme algoritmalarından biri [69] olduğu için bu tarz bir kümelemeye örnek olarak verilebilir [70].

2.4.2. Hiyerarşik Kümeleme

Hiyerarşik kümeleme, tek bir küme yerine kümeleme hiyerarşisi oluşturduğu için yani veriler tek bir adımda belirli sayıda kümeye ayrılmadığı için, sıralı kümelemeden farklıdır [68]. Hiyerarşik kümeleme stratejileri aşağıdan yukarıya ya da yukarıdan aşağıya olmak üzere yığınsal (Agglomerative) ve bölücü (Divisive) diye adlandırılan genellikle iki kategoriye ayrılır. Yığınsal yaklaşımda her bir örnek başlangıçta ayrı bir kümeye atanmaktadır ve kümelerin aşağıdan yukarıya hiyerarşisini oluşturmak için her seferinde iki kümeyi birleştirmeye devam eder. Bölücü hiyerarşik kümelemede ise veri kümesi başlangıçta tek bir kümeye atanır ve sürekli olarak iki gruba bölünerek kümelerin yukarıdan aşağıya hiyerarşisini oluşturur [71].

3. MATERYAL VE YÖNTEM

3.1. Bir Boyutlu Protein Yapı Tahmini

Protein yapı tahmininde genel işleyiş, öncelikle bir boyutlu (1D) yapıları tahmin etmek, daha sonra bu bilgiyi kullanarak protein 3D yapısını öngörmek şeklindedir. Yani alt problemin çözümünden başlayıp asıl problemin çözümüne ulaşmak şeklinde özetlenebilir. Bu bağlamda proteinlerin bir boyutlu yapısal özelliklerinin amino asit dizisine dayalı tahmini, protein yapı tahmini genel probleminin bir alt problemi olagelmiştir.

Kabsch ve Sander tarafından tanımlanan Protein İkincil Yapı Sözlüğü (DSSP) [72], protein ikincil yapılarının tek harfli kodlar ile tanımlandığı, yaygın olarak kullanılan ikincil yapı atama yöntemlerinden bir tanesidir. STRIDE, DEFINE, PSEA ve P-Curve diğer atama yöntemlerindendir [73]. DSSP'ye göre sekiz durumlu ikincil yapı etiketi tanımlanmıştır. Bunlar şu şekilde isimlendirilmektedir : (H,B,E,G,I,T,S,' '). H (sarmal), E (beta iplik) ve L (döngü) olmak üzere bu sekiz sınıf sadeleştirilip, gruplandırılarak Çizelge 3.1.'deki gibi 3 sınıfa indirgenir [45].

Çizelge 3.1. 8 sınıflı DSSP etiketlerinin 3 sınıflı gösterimi [45]

8 Sınıflı DSSP Etiketleri	3 Sınıflı DSSP Etiketleri	İkincil Yapı Etiketi
H	H	Sarmal (Helix)
G		
I		
E	E	Beta İplik (Beta Strand)
B		
S	L	Döngü (Loop)
T		
' '		

Protein yapı tahmini için mevcut olan pek çok hesaplama yöntemi, iki genel yaklaşım halinde gruplandırılabilir. Birincisi, bilinen üç boyutlu yapı proteinlerinden gelen bilgileri kullanarak hedef proteini buna göre modellemeye çalışan karşılaştırma modelleme veya şablon tabanlı yöntemlerdir. İkinci kategori, yapının sıfırdan tahmin edildiği ab initio (de novo) yöntemidir [22]. Ab-initio yöntemleri karşılaştırmalı modelleme teknikleriyle birlikte kullanıldığında daha yararlı hale gelmektedir [74].

3.1.1. Problem Tanımı

Bir amino asit dizisinden başlayarak, ikincil yapı tahmin probleminde hedef, proteinin her amino asitine 3 harfli bir alfabeden (H: sarmal, E: beta iplik, L: döngü) bir yapısal sınıf etiketi atamaktır.

Şekil 3.1.'de örnek bir ikincil yapı etiketlemesi gösterilmektedir. Birinci satır proteine ait amino asit dizisini, ikinci satır ise her bir amino asite karşılık gelen ikincil yapı etiketini temsil etmektedir.

```
MSNTTWGLQRDITPRLGARLVQEGNQLHYLA
LLLLEEEEELLLHHHHHHHLLLLLEEEEELL
```

Şekil 3.1. Her bir amino asite karşılık gelen 3-durumlu örnek ikincil yapı etiketleri

3.2. Veri Kümesi

Bu tez çalışmasında, ikincil yapı tahmin yöntemi olarak kullandığımız DSPRED yönteminin performansını test etmek için, 513 protein ve 84.119 amino asit içeren Cuff ve Barton [75] tarafından oluşturulan halka açık erişimli CB513 veri kümesi kullanılmıştır.

3.3. Öznitelik Çıkarımı

DSPRED tahmin yöntemimizin giriş öznitelikleri, PSI-BLAST [76], HHMAKE PSSM'leri ve yapısal profil matrisleri tarafından türetilmiş olan pozisyona özel profil matrisleri (PSSM) [77] formundaki dizi profillerini içermektedirler. PSSM öznitelikleri ya tek başına ya da diğer protein özellikleriyle birlikte makine öğrenmesi algoritmalarında girdi olarak kullanılmaktadırlar [43].

3.3.1. PSI-BLAST PSSM Öznitelikleri

PSSM profil matrislerini hesaplamak için, CB513 veri kümesindeki her bir hedef protein, PSI-BLAST [76] programı vasıtası ile NCBI'nın NR [78] veri tabanındaki proteinler ile hizalanarak .psiblast uzantılı, belirli bir puan eşiğinin üzerinde tespit edilen dizilerin çoklu dizi hizalamalarından elde edilen profil matrisleri oluşturulmuştur. Bir sonraki adımda, hedefe benzeyen proteinler çoklu hizalama algoritması ile hizalanarak, amino asitlerin oluşum sıklığı Aydın ve arkadaşlarının [45] çalışmasında olduğu gibi sigmoidal dönüşüm uygulanarak [0,1] aralığına çekilerek ve normalize edilerek Nx20 boyutlu PSSM profil matrisleri elde edilmiştir [76]. Elde edilen matristeki N, hedef proteinin amino asit sayısına, 20 rakamı ise amino asit çeşitine karşılık gelmektedir. Her bir PSI-BLAST satırı, 20 amino asitten birini hedefin belirli bir amino asitinde gözlemleme eğilimini içermektedir.

Şekil 3.2. ve Şekil 3.3. PSI-BLAST programının out_ascii_pssm parametresi kullanılarak veri kümesindeki her bir hedef protein için elde edilen metin tabanlı örnek .alignment uzantılı dosyayı ve .psiblast uzantılı Nx20 boyutlu PSSM matrisini göstermektedir.

Kullanılan psi blast komut satırı:

psiblast –query $fasta_filename –out $alignment_filename -out_ascii_pssm $pssm_filename -line_length 50000 -num_iterations 3 -evalue 10 -inclusion_ethresh 0.001 –db $blast_db_dir $blast_db_root -num_threads 16

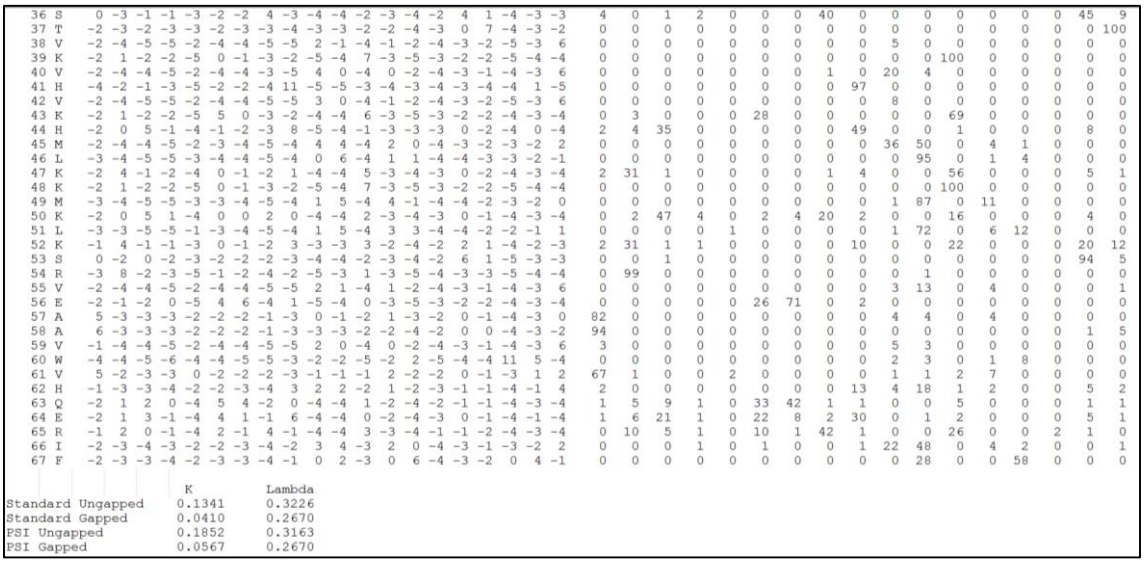
```
  Number of letters in database: 543,065,305
  Number of sequences in database:  1,427,391

Lambda      K       H
   0.318    0.187    0.640

Lambda      K       H
   0.267   0.0567    0.140


Matrix: BLOSUM62
Gap Penalties: Existence: 11, Extension: 1
Number of Hits to DB: 21,041,427,200
Number of Sequences: 93586724
Number of extensions: 1096490863
Number of successful extensions: 3526924
Number of sequences better than 10.0: 500
Number of HSP's better than 10.0 without gapping: 471475
Number of HSP's successfully gapped in prelim test: 8776
Number of HSP's that attempted gapping in prelim test: 3045863
Number of HSP's gapped (non-prelim): 484945
length of query: 67
length of database: 34,056,882,717
effective HSP length: 39
effective length of query: 28
effective length of database: 34,701,967,777
effective search space: 971655097756
effective search space used: 971655097756
T: 11
A: 40
X1: 16 ( 7.3 bits)
X2: 38 (14.6 bits)
X3: 64 (24.7 bits)
S1: 41 (21.2 bits)
S2: 85 (36.9 bits)
```

Şekil 3.2. PSI-BLAST hizalaması sonucu elde edilen .alignment uzantılı örnek bir dosyaya ait ekran görüntüsü



		K	Lambda
Standard Ungapped		0.1341	0.3226
Standard Gapped		0.0410	0.2670
PSI Ungapped		0.1852	0.3163
PSI Gapped		0.0567	0.2670

Şekil 3.3. PSI-BLAST hizalaması sonucu elde edilen .psiblast uzantılı Nx20 boyutlu PSSM matrisi

3.3.2. HHMAKE PSSM Öznitelikleri

HMM'den türetilen profiller, PSSM öznitelikleri ile birlikte protein yapı tahmini öngörmede girdi öznitelikleri olarak kullanılmaktadır [79]. Bu tez çalışmasında tahmin yöntemimizde girdi özniteliği olarak kullanılacak HMM profil matrislerini üretmek için, CB513 veri kümesindeki hedef proteinler, HHblits programının ilk basamağı kullanılarak NR20 veri tabanı ile hizalanıp, her bir hedef protein için HMM profil matrisleri hesaplanmıştır. Bir sonraki adımda, hedefin HMM profili, HHblits yönteminin ikinci basamağı kullanılarak PDB70 [80] veri tabanındaki (%70'e kadar dizi benzerliği içeren PDB veri kümesi) HMM profillerine hizalanarak .hhr uzantılı dosyalar elde edilmiştir (Şekil 3.4.). Elde edilen PSSM değerleri sigmoidal bir dönüşümle [0,1] aralığında normalize edilerek, tahmin yöntemimizin ikinci öznitelik girdisi elde edilmiştir.

No	Hit	Prob	E-value	P-value	Score	SS	Cols	Query HMM	Template HMM
1	1A0CA	100.0	4E-132	6E-136	1026.5	54.5	437	1-437	1-437 (437)
2	3XINA	100.0	6.7E-61	9.9E-65	488.3	42.7	346	41-411	8-361 (393)
3	9XIMA	100.0	4.8E-60	7.1E-64	481.9	43.1	346	41-411	8-361 (393)
4	3XIMA	100.0	5E-59	7.4E-63	474.4	43.0	346	41-411	8-361 (393)
5	2XINA	100.0	2.7E-58	4E-62	468.8	43.2	344	41-410	8-360 (393)
6	1XINA	100.0	2.6E-58	3.9E-62	468.9	43.0	346	41-411	8-361 (393)
7	4HHLA	100.0	1.5E-58	2.2E-62	470.2	41.1	322	40-384	8-332 (388)
8	1XLMA	100.0	3.8E-58	5.7E-62	468.0	43.8	348	36-410	4-361 (394)
9	1XIMA	100.0	3.1E-58	4.6E-62	468.2	43.1	346	41-411	7-360 (392)
10	2XIMA	100.0	9.9E-58	1.5E-61	464.5	42.8	346	41-411	8-361 (393)
11	5XINA	100.0	2.4E-57	3.5E-61	461.8	42.8	346	41-411	8-361 (393)
12	1XYLA	100.0	1.5E-57	2.2E-61	462.5	40.4	329	36-389	4-335 (386)
13	1BXCA	100.0	3.2E-57	4.7E-61	460.2	40.0	321	41-384	8-331 (387)
14	1BXBA	100.0	4.2E-57	6.2E-61	459.2	40.2	329	36-389	4-335 (387)
15	1MUWA	100.0	4.5E-57	6.7E-61	458.8	40.2	325	41-389	8-335 (386)
16	1XLAA	100.0	7.2E-57	1.1E-60	458.2	41.2	348	36-410	4-361 (394)
17	1GW9A	100.0	6.2E-57	9.2E-61	458.0	40.5	323	39-384	7-332 (388)
18	1OADA	100.0	9.1E-57	1.3E-60	456.9	40.8	323	39-384	7-332 (388)
19	4XIAA	100.0	8.2E-57	1.2E-60	457.6	40.4	344	41-410	7-360 (393)
20	2GLKA	100.0	1.8E-56	2.7E-60	454.7	40.6	323	39-384	7-332 (388)
21	6XIAA	100.0	2.8E-56	4.1E-60	453.0	40.6	321	41-384	8-331 (387)
22	1QT1A	100.0	2.4E-56	3.5E-60	453.8	39.9	321	41-383	8-330 (387)
23	1CLKA	100.0	6.3E-56	9.3E-60	450.6	40.1	325	41-389	8-335 (387)
24	2GYIA	100.0	2.8E-55	4.1E-59	445.6	40.3	325	41-389	8-335 (386)
25	4HHMA	100.0	2.1E-54	3.2E-58	439.3	40.2	326	40-389	8-336 (388)
26	3ITVA	100.0	4.3E-42	6.2E-46	353.1	35.8	302	35-382	45-365 (438)
27	3M0YA	100.0	6E-42	8.6E-46	352.0	35.8	302	35-382	45-365 (438)
28	3ITXA	100.0	6.9E-42	1E-45	351.5	35.9	291	40-369	49-357 (438)
29	3M0MA	100.0	8.9E-42	1.3E-45	349.2	35.9	310	28-385	37-365 (421)
30	4GJIA	100.0	1.9E-41	2.7E-45	348.2	35.9	301	40-384	49-367 (438)
31	3ITLA	100.0	2.1E-41	3E-45	347.9	35.8	291	40-369	49-357 (438)
32	3M0VA	100.0	2.1E-41	3.1E-45	347.9	35.6	291	40-369	49-357 (438)
33	3KTCA	100.0	1.4E-32	2.2E-36	269.3	34.3	270	84-382	35-307 (330)
34	1D8WA	100.0	1.3E-28	1.7E-32	242.2	27.6	294	29-373	22-344 (402)
35	3KWSA	100.0	1.3E-27	2E-31	224.5	30.1	240	84-366	20-262 (265)
36	3TVAA	100.0	4.9E-27	7.9E-31	222.3	31.4	253	82-369	15-279 (282)
37	3VNIA	100.0	7.7E-27	1.2E-30	222.3	30.4	229	82-342	18-248 (289)
38	2QW5A	100.0	9.5E-27	1.5E-30	227.6	31.5	233	81-342	26-279 (327)
39	3CQJA	100.0	1.6E-26	2.5E-30	217.4	32.0	249	84-367	19-272 (276)
40	4OVXA	100.0	8.7E-27	1.4E-30	219.5	29.5	241	41-344	3-249 (270)
41	2ZVRA	100.0	1.9E-26	3E-30	215.0	30.5	221	84-343	18-241 (259)

Şekil 3.4. Örnek bir .hhr dosyasına ait ekran görüntüsü

3.3.3. Yapısal Profil Matrisi

HHblits programı kullanılarak elde edilen .hhr uzantılı dosyalar kullanılarak veri kümesindeki her bir amino asit için .struct uzantılı Nx3 boyutlu yapısal profil matrisleri elde edilmiştir. Buradaki N hedef proteindeki amino asitlerin sayısıdır ve her sütunda o amino asit için üç ikincil yapı durumundan birisinin gözlemlenme skoru bulunmaktadır. Örnek bir yapısal profil matrisi, Şekil 3.5.'te gösterilmiştir.

	H	E	L
M	0.07	0.03	0.90
S	0.09	0.11	0.80
N	0.60	0.10	0.30
T	0.70	0.10	0.20
T	0.80	0.15	0.05
W	0.20	0.70	0.10
...	...	...	...

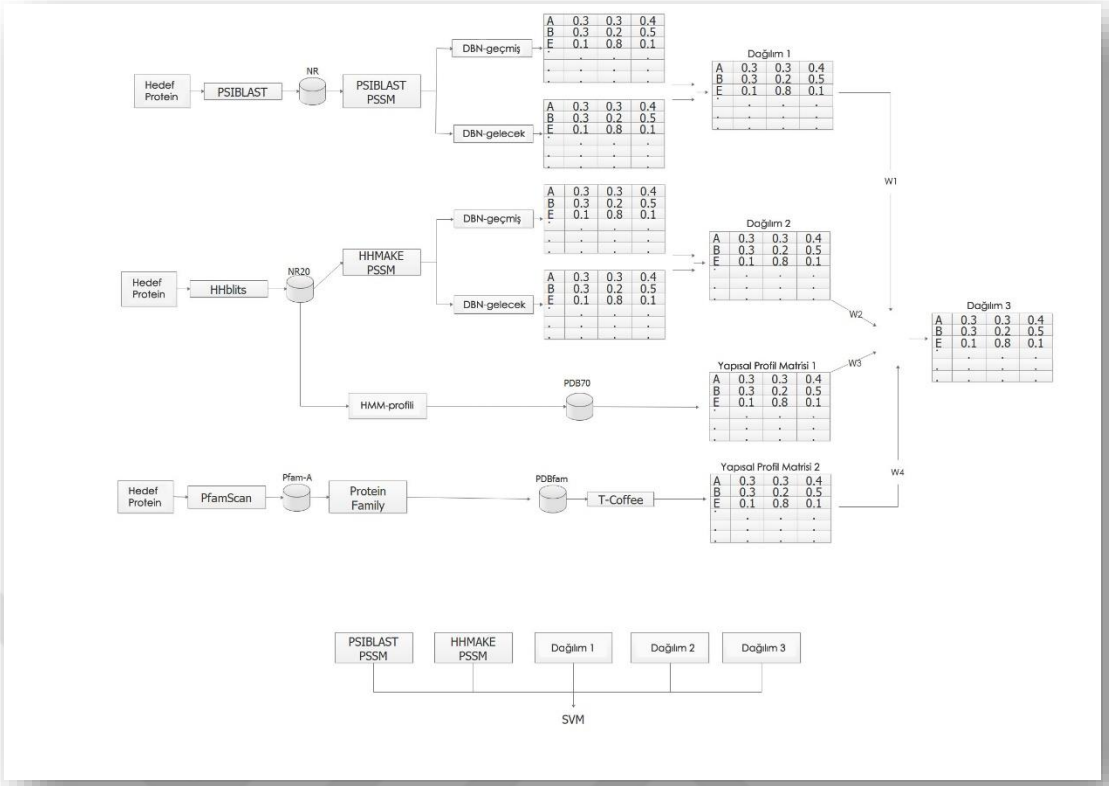
Şekil 3.5. Protein ikincil yapı tahmini için yapısal profil matrisi

3.4. DSPRED Metodu

DSPRED, DBN ve SVM sınıflandırıcılarını içeren iki aşamalı hibrit bir sınıflandırıcıdır. DSPRED metodu bu tez çalışmasında protein ikincil yapı tahmini öngörüsü için kullanılmaktadır. Şekil 3.6.'da görüleceği üzere, DSPRED metodunun ilk aşamasında, her bir amino asitin ikincil yapı sınıfını tahmin etmek için, PSIBLAST PSSM ve HHMAKE PSSM'leri için ayrı DBN sınıflandırıcıları kullanılmıştır. Her DBN modeli, giriş öznitelikleri verilen sınıf etiketlerinin olasılık dağılımını (Dağılım 1 ve 2 olarak adlandırılır) üretir. Her bir profil matrisi için DBN-geçmiş ve DBN-gelecek olmak üzere iki tip, toplamda 4 DBN modeli eğitilmiştir. DBN-geçmiş geçerli pozisyondaki profil vektörünün daha önce gelen komşu pozisyonlara bağlı olduğu modeli temsil ederken, DBN-gelecek ise verilen bir pozisyondaki profil vektörünün kendinden sonra gelen pozisyonlara bağlı olduğu modeli temsil etmektedir.

Şekil 3.6.'ya göre, PSIBLAST PSSM profil matrisleri için DBN-geçmiş ve DBN-gelecek olarak adlandırılan iki tip DBN sınıflandırıcısından elde edilen olasılık dağılımlarının ortalamaları alınarak Dağılım 1 elde edilir. Benzer şekilde, HHBlits'in ilk aşamasında üretilen HHMAKE PSSM profil matrisleri kullanılarak DBN-geçmiş ve DBN-gelecek olarak adlandırılan iki tip DBN sınıflandırıcısından elde edilen olasılık dağılımlarının ortalamaları alınarak da Dağılım 2 elde edilir. Elde edilen bu dağılımların (Dağılım 1 ve Dağılım 2) ortalaması alınarak, HHblits yönteminin ikinci aşamasından elde edilen Yapısal Profil Matrisi 1 ile birleştirilerek Dağılım 3 elde edilir [81].

Bu çalışmada, DBN sınıflandırıcılarının tek taraflı amino asit penceresi L_A = 5'e ve tek taraflı ikincil yapı geçmişi penceresi L_S = 4'e ayarlanmıştır. Bir sonraki aşama olan ikinci aşamada, PSI-BLAST PSSM, HHMAKE PSSM, Dağılım 1, 2 ve 3, SVM sınıflandırıcısının giriş öznitelikleri olarak kullanılır. İkincil yapı sınıfını tahmin etmek için, her bir amino asitin etrafına 11 boyutlu simetrik bir pencere alınır ve bu penceredeki öznitelikler toplam 539 öznitelik elde etmek için bir araya getirilir (PSI-BLAST PSSM: 20 × 11 = 220 öznitelik, HHMAKE PSSM: 20 × 11 = 220 öznitelik, Dağılımlar 1-3: 3 × 3 × 11 = 99 öznitelik). Mevcut çalışmada ikinci Yapısal Profil Matrisi 2 kullanılmayarak, W4=0 olarak ayarlanmıştır. DSPRED yönteminin adımları, Şekil 3.6.'da gösterilmiştir. DSPRED yönteminin detayları Aydın ve arkadaşlarının çalışmasında [45,81] ve Görmez'in [82] tez çalışmasında yer almaktadır.



Şekil 3.6. DSPRED metodunun aşamaları

3.5. Büyük Veri Kümeleriyle SVM Eğitimi

SVM, Vapnik tarafından önerilen, sınıflandırma ve regresyon problemlerinde kullanılan başarılı bir makine öğrenmesi yöntemidir [83,84]. Yüksek doğruluk oranı, vektörel olmayan verileri işleyebilmesi ve çeşitli veri kaynaklarının modellenmesinde esneklik gibi nedenlerle sinyal işleme, görüntü işleme ve biyoinformatik de dahil olmak üzere birçok gerçek dünya problemine başarıyla uygulanmıştır [85]. Her ne kadar SVM karmaşık tahmin görevlerinde iyi performans gösterse de, model eğitimi sırasında büyük veri kümeleri ile çalışırken eğitim süresi maliyetli olmakta ve sınıflandırma süreci uzamaktadır [86]. Örneğin, içerisinde bir milyon veri kaydı barındıran birçok özelliğe sahip bir veri kümesi ile SVM'yi eğitmek yıllar sürebilmektedir [41,87]. Veri toplama, depolama ve işleme teknolojilerindeki gelişmelere dayanarak, biyoinformatik de dahil olmak üzere birçok disiplinde veri

tabanlarının büyüklüğü hızla artmaktadır [88]. Bu nedenle, SVM'nin eğitim aşamasını hızlandırmak için etkili yöntemler geliştirilmelidir.

Çalışmamızın bundan sonraki aşaması iki ana başlık altında detaylandırılacaktır. İlkinde CB513 veri kümesi için tabakalı örnekleme yöntemi kullanarak, ikincisinde de hiyerarşik kümeleme yöntemi uygulayarak DSPRED metodunun SVM sınıflandırıcısının eğitim süresini iyileştirme ve protein ikincil yapı tahmin başarısı hesaplama ile ilgili detaylar açıklanacaktır.

3.5.1. Tabakalı Rastgele Seçim Yöntemi ile Örnek İndirgeme

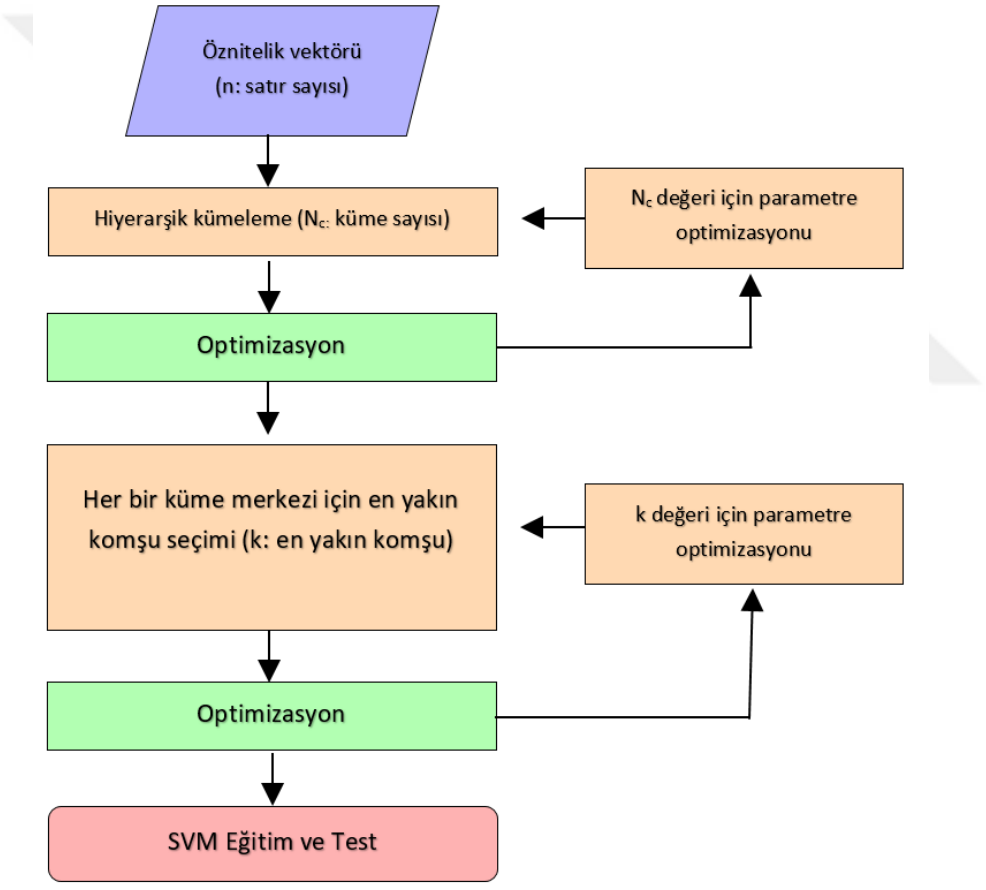
CB513 veri kümesi aşağıdaki aşamalardan geçirilerek SVM yöntemi ile tahmin başarısı hesaplanmıştır: Tabakalı rastgele örneklemede, eğitim setindeki örneklerin (yani amino asitlerin), her bir sınıf etiketinin (sarmal, beta iplik ve döngü) eğitim setindeki oranı korunarak %10, %20, … , %100 gibi %10'luk artışlarla elde edilen sabit bir yüzdesi rastgele ve eşsiz bir şekilde seçilmiştir. Seçilen bu indirgenmiş veri kümeleri için SVM modeli eğitilip, tahmin doğruluğu test setlerinde hesaplanmıştır. Özetle yüzde parametresi %10'luk artışlarla %10'dan %100'e yükseltilmiştir. Örneğin, eğer bu parametre %10'a ayarlanmışsa, sonuçta elde edilen eğitim seti orijinal eğitim setindeki amino asitlerin yaklaşık %10'unu içermektedir ve %100'e ayarlanırsa tüm veri örneklerini içermektedir.

3.5.2. Hiyerarşik Kümeleme Yöntemi ile Örnek İndirgeme

Python tabanlı Scipy [89] paketi kullanılarak ward (en küçük varyans) yöntemi ile 7 kat çapraz doğrulama uygulanmış CB513'e ait eğitim veri kümeleri üzerinde hiyerarşik kümeleme algoritması kullanılarak kümeleme işlemi gerçekleştirilmiştir. Eğitim veri kümelerindeki veri örnekleri, daha önce validasyon veri kümesi kullanılarak optimize edilen küme sayısı ve bu küme merkezlerine en yakın komşu parametre değerleri kullanılarak, indirgenmiştir. Şekil 3.7., hiyerarşik kümeleme kullanılarak eğitim veri kümesi örnek indirgeme aşamalarını özetlemektedir. Hiper parametreler Nc (küme sayısı) ve k (küme merkezine en yakın komşu sayısı), bir

sonraki bölümde açıklandığı gibi validasyon kümelerindeki tahmin doğruluğu hesaplanarak optimize edilmiştir. Hiyerarşik kümeleme için farklı yöntemler kullanılmaktadır ve bunlar arasında bu çalışmada ward yönteminin en iyi sonuçları verdiği gözlenmiştir. Ward yöntemi, küme varyansındaki toplamı en aza indiren minimum bir sapma kriteri uygulamaktadır. Her adımda, birleşme sonrası toplam küme varyansında minimum artışa neden olan küme çiftini bulur [90,91].

Scipy dokümanına göre, Ward metodu için, zaman karmaşıklığı O (n^2) olan en yakın komşular zinciri (nearest-neighbors chain) olarak adlandırılan bir algoritma uygulanmaktadır [92,93].



Şekil 3.7. Hiyerarşik kümeleme yöntemi ile örnek indirgeme aşamaları

3.6. Kümeleme İçin Çapraz Doğrulama ve Hiper Parametre Optimizasyonu

Veri indirgeme stratejilerinin doğruluğu daha güvenilir sonuçlar elde etmek için çapraz doğrulama (cross-validation) ortamında değerlendirilmektedir. Bu amaç için CB513'teki proteinler rastgele yedi parçaya ayrılır ve eğitim / test bölmeleri buna göre oluşturulur. Bu işlem toplam yedi eğitim test seti çiftiyle sonuçlanır. Örneğin, ilk eğitim setinde % 34,70'i (25,544) sarmal, % 22,26'sı (16,387) beta iplik ve % 43,04'ü (31,691) döngü olan toplam 73.622 amino asit örneği vardır. 0 değeri sarmal yapısını, 1 değeri beta iplik yapısını ve 2 değeri ise döngü yapısını temsil etmektedir (Çizelge 3.2.). İlk test setinde toplam 10.497 amino asit yer almaktadır. Eğitim ve test veri kümelerinde, her bir amino asit toplam 539 öznitelik (feature) ile temsil edilmektedir.

Hiyerarşik kümelemeyle eğitim veri örneği sayısı azaltma yaklaşımının hiper parametresi olan küme sayısı ve en yakın komşu sayısı, grid search tekniği kullanılarak optimize edilmiştir. Küme sayısını temsil eden N_c ilk hiper parametredir. Bu parametreyi optimize etmek için 500 ila 1500 arasında değişen değerler göz önünde bulundurulmuştur. İkinci hiper parametre olan k ise, 1'den 19'a kadar olan değerlerin seçilmesiyle optimize edilen en yakın komşuların sayısıdır. Bu amaçla, her eğitim setinden proteinlerin yaklaşık %10'u rastgele seçilmiş ve toplam yedi validasyon kümesi oluşturulmuştur. Validasyon kümeleri, hiper parametreleri optimize etmek için ikincil test setleri olarak kullanılmıştır. Eğitim setinin %10'unun seçilmesinin nedeni, eğitim setinde mümkün olduğunca çok sayıda örneğe izin vermektir. Validasyon kümeleri oluşturulduktan sonra, kalan örnekler SVM modellerini eğitmek için kullanılır ve tahmin doğruluğu hiper parametrelerin farklı değerleri için validasyon kümelerinde hesaplanmıştır. Ardından, çapraz doğrulama deneyinin her yinelemesi için en iyi validasyon kümesi tahminine sahip parametreler seçilmiştir. Optimum hiper parametreler bulunduktan sonra (toplam yedi optimum parametre çifti), SVM orijinal eğitim setlerinde bulunan parametre değerleri kullanılarak eğitilip, tahminler test setlerinde hesaplanmıştır.

Çizelge 3.2. Üç durumlu DSSP etiketlerinin DSPRED metoduna ait çıkış kodlarının sayısallaştırılmış gösterimi

Sınıf Adı	3 Durumlu DSSP Etiketi	Çıkış Kodu
Sınıf 0	Sarmal (H)	0
Sınıf 1	Beta iplik (E)	1
Sınıf 2	Döngü (L)	2

3.7. Sistem Mimarisi ve SVM'nin Hiper Parametreleri

Protein ikincil yapı tahmini için tatmin edici sonuçlar veren Radyal Temelli Çekirdek Fonksiyonunlu (RBF) SVM, libSVM yazılımı (sürüm 3.21) kullanılarak uygulanmıştır. SVM'nin hiper parametreleri CB513 veri kümesi için Aydın ve arkadaşları [45] tarafından optimize edilmiştir. Bu iki hiper parametre için optimum değer olarak elde edilen gama parametresi → $\alpha = 0.00781$ ve C parametresi → $C = 1.0$ olarak ayarlanmıştır.

Bu yöntemlerin çalıştırıldığı bilgisayar sistemlerinin özellikleri şu şekildedir:

- TRUBA-levrek hesaplama sunucusu : Centos Enterprise Linux 7.3 işletim sistemi, Intel (R) Xeon® E5-2690 işlemci, 2.90 GHz CPU ve 256GB RAM.

- Ubuntu 16.04.2 LTS (Xenial Xerus) işletim sistemi, Intel (R) Xeon (R) CPU E5-2650 v2 @ 2.60 GHz ve 64 GB RAM.

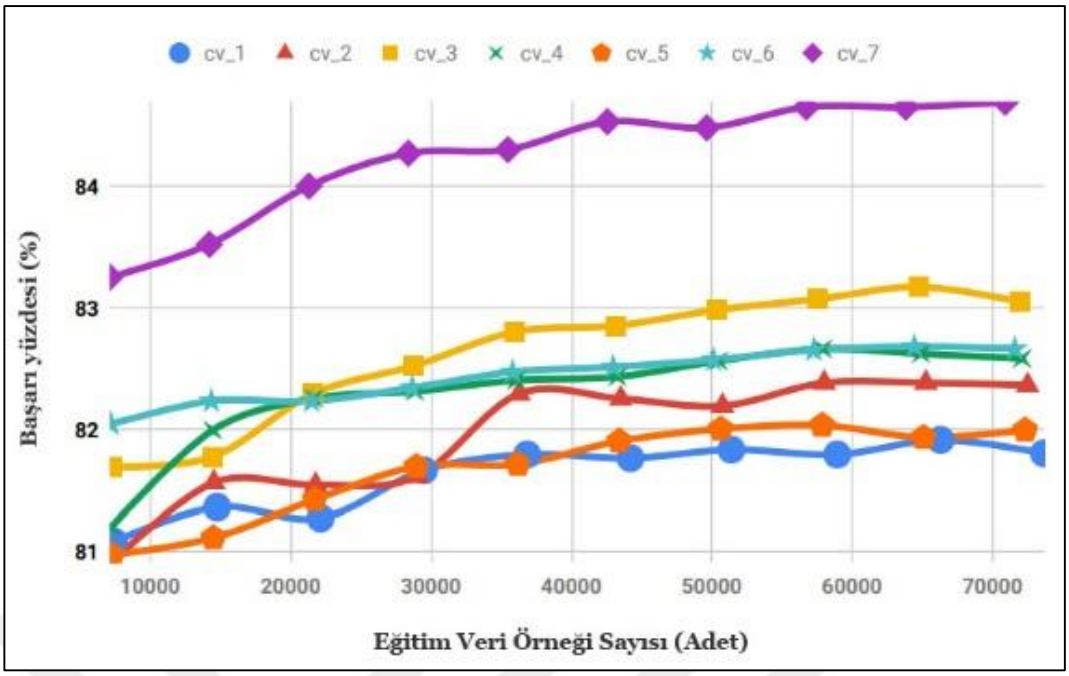
4. SONUÇLAR VE TARTIŞMA

4.1. Sonuçlar

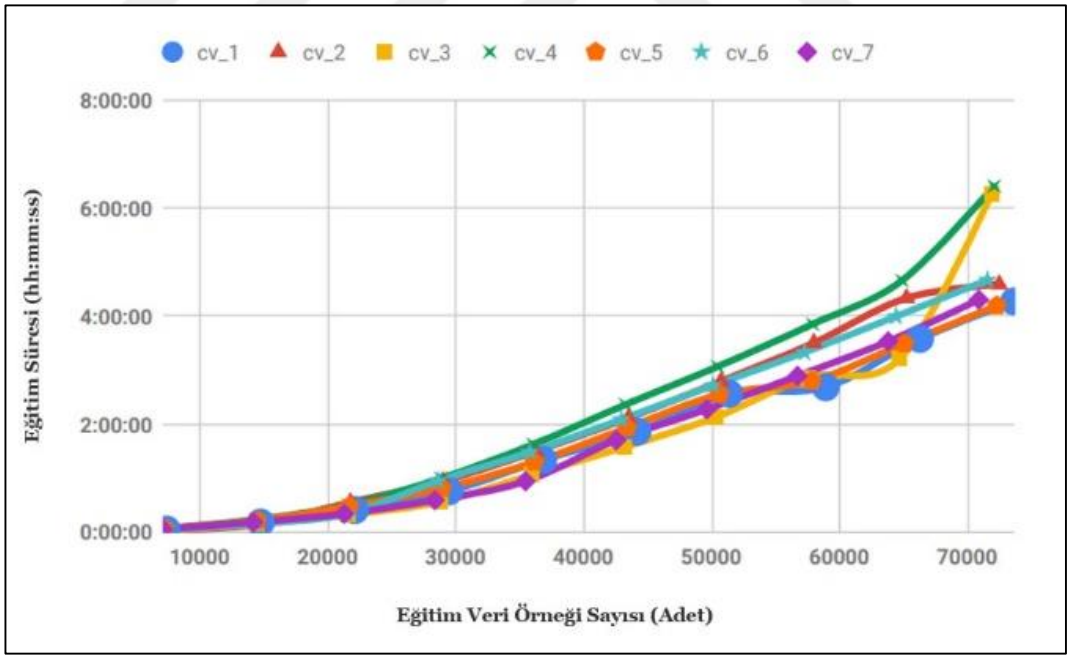
Bu tez çalışmasında, DSPRED olarak adlandırılan iki aşamalı hibrit bir sınıflandırıcının, ikinci aşamasında kullanılan SVM sınıflandırıcısının model eğitme süresini iyileştirmek için iki örnek indirgeme stratejisi önerilmiştir. Önerilen çözümler veri kümesi boyutunu %26-50 oranında azaltarak yaklaşık 36.000 amino asit örneğine kadar indirgeyebilmektedir. Doğruluk değerlendirmeleri CB513 veri kümesi üzerinde çapraz doğrulama deneyleri yapılarak gerçekleştirilmiştir. Daha büyük veri kümeleri ile çalışırken, tatmin edici bir tahmin doğruluğu elde etmek için eğitim veri kümesinde yaklaşık 36.000 veri örneğini tutmak yeterli olmaktadır.

4.1.1. Tabakalı Rastgele Seçim Yöntemi ile Örnek İndirgeme

Tabakalı örnekleme kullanılarak 7 kat çapraz doğrulama uygulanmış eğitim setlerinden, %10 ile %100 arasında %10 artımla değişen bir yüzdeye sahip veri örnekleri rastgele ve eşsiz bir biçimde seçilerek indirgenmiş veri kümeleri elde edilmiştir. Bu indirgenmiş veri kümeleri ile SVM modeli eğitilerek test setleri kullanılıp tahmin başarısı hesaplanmıştır. Şekil 4.1. ve Şekil 4.2., SVM sınıflandırıcısının ikincil yapı tahmin doğruluğunu ve ayrıca çapraz doğrulamanın tüm katları için model eğitim sürelerini göstermektedir. Elde edilen sonuçlara göre, tahmin doğruluğunu önemli ölçüde düşürmeden veri örneklerinin yaklaşık %50'sini azaltmanın mümkün olduğu gözlenmiştir. 7 kat çapraz doğrulama uygulanmış bütün veri kümelerinden elde edilen değerler göz önüne alındığında ortalama olarak %50 oranında veri kümeleri azaltıldığında, SVM'nin model eğitim süresi %73.38’lik bir iyileşme göstermiştir (Çizelge 4.1., 4.2., 4.3., 4.4., 4.5., 4.6., 4.7.).



Şekil 4.1. Yedi kat çapraz doğrulama yapılan CB513 veri kümesi için tabakalı rastgele seçim yöntemi uygulanarak elde edilen Q3 doğruluk yüzdeleri



Şekil 4.2. Yedi kat çapraz doğrulama yapılan CB513 veri kümesi için tabakalı rastgele seçim yöntemi uygulanarak elde edilen model eğitim süreleri

Çizelge 4.1. Tabakalı rastgele seçim yöntemi ile örnek indirgemeye ait sonuçlar: k-kat (k=7), cv1

İndirgenmiş veri kümesi	Rasgele ve Eşsiz Seçilen Satır Sayısı	Tahmin Başarısı (%)	Model Eğitim Süresi (hh-mm-ss)	Tahmin Süresi (mm-ss)
10	7362	81,0803	00:02:55	00:46
20	14724	81,3566	00:10:32	01:22
30	22087	81,2708	00:24:26	02:31
40	29449	81,6614	00:45:17	03:17
50	36811	81,7853	01:20:09	03:24
60	44173	81,7567	01:51:31	04:07
70	51353	81,8329	02:33:51	04:00
80	58898	81,7948	02:41:01	04:26
90	66260	81,9091	03:34:47	04:50
100	73622	81,8043	04:16:21	04:59

** Model Eğitim Süresi, h : saat, m: dakika, s:saniye*

Çizelge 4.2. Tabakalı rastgele seçim yöntemi ile örnek indirgemeye ait sonuçlar: k-kat (k=7), cv2

İndirgenmiş veri kümesi	Rasgele ve Eşsiz Seçilen Satır Sayısı	Tahmin Başarısı (%)	Model Eğitim Süresi (hh-mm-ss)	Tahmin Süresi (mm-ss)
10	7246	80,9144	00:02:19	00:46
20	14492	81,5577	00:08:26	01:20
30	21738	81,5406	00:33:08	03:08
40	28984	81,6092	00:56:39	03:37
50	36231	82,2868	01:29:45	04:04
60	43477	82,2525	02:06:47	04:48
70	50723	82,1925	02:47:47	05:14
80	57969	82,3812	03:30:18	05:41
90	65215	82,3812	04:20:07	06:11
100	72461	82,3555	04:35:17	06:44

** Model Eğitim Süresi, h : saat, m: dakika, s:saniye*

Çizelge 4.3. Tabakalı rastgele seçim yöntemi ile örnek indirgemeye ait sonuçlar: k-kat (k=7), cv3

İndirgenmiş veri kümesi	Rasgele ve Eşsiz Seçilen Satır Sayısı	Tahmin Başarısı (%)	Model Eğitme Süresi (hh-mm-ss)	Tahmin Süresi (mm-ss)
10	7188	81,6860	00:02:20	00:48
20	14375	81,7677	00:08:57	01:28
30	21563	82,2905	00:19:09	01:55
40	28751	82,5192	00:33:35	02:25
50	35939	82,7969	01:03:47	03:32
60	43126	82,8459	01:34:15	04:04
70	50314	82,9848	02:07:46	04:52
80	57502	83,0665	02:52:04	05:36
90	64689	83,1727	03:12:45	05:33
100	71877	83,0502	06:16:19	09:59

** Model Eğitim Süresi, h : saat, m: dakika, s:saniye*

Çizelge 4.4. Tabakalı rastgele seçim yöntemi ile örnek indirgemeye ait sonuçlar: k-kat (k=7), cv4

İndirgenmiş veri kümesi	Rasgele ve Eşsiz Seçilen Satır Sayısı	Tahmin Başarısı (%)	Model Eğitim Süresi (hh-mm-ss)	Tahmin Süresi (mm-ss)
10	7207	81,2044	00:04:03	01:15
20	14413	81,9924	00:14:00	02:04
30	21619	82,2495	00:30:18	02:52
40	28825	82,3076	00:59:44	04:49
50	36032	82,3988	01:37:28	05:09
60	43238	82,4320	02:21:37	05:05
70	50444	82,5647	03:03:54	05:32
80	57872	82,6642	03:51:04	06:05
90	64857	82,6228	04:39:29	06:34
100	72063	82,5813	06:24:23	08:42

** Model Eğitim Süresi, h : saat, m: dakika, s:saniye*

Çizelge 4.5. Tabakalı rastgele seçim yöntemi ile örnek indirgemeye ait sonuçlar: k-kat (k=7), cv5

İndirgenmiş veri kümesi	Rasgele ve Eşsiz Seçilen Satır Sayısı	Tahmin Başarısı (%)	Model Eğitim Süresi (hh-mm-ss)	Tahmin Süresi (mm-ss)
10	7228	80,9697	00:03:41	01:09
20	14456	81,1133	00:12:58	01:56
30	21684	81,4173	00:28:58	02:38
40	28912	81,6876	00:48:29	03:16
50	36140	81,7130	01:17:51	04:00
60	43368	81,8988	01:55:50	04:34
70	50596	82,0002	02:33:00	05:08
80	57824	82,0340	02:47:52	05:19
90	65052	81,9326	03:29:32	05:19
100	72280	81,9917	04:11:39	05:44

** Model Eğitim Süresi, h : saat, m: dakika, s:saniye*

Çizelge 4.6. Tabakalı rastgele seçim yöntemi ile örnek indirgemeye ait sonuçlar: k-kat (k=7), cv6

İndirgenmiş veri kümesi	Rasgele ve Eşsiz Seçilen Satır Sayısı	Tahmin Başarısı (%)	Model Eğitim Süresi (hh-mm-ss)	Tahmin Süresi (mm-ss)
10	7154	82,0372	00:02:33	00:51
20	14309	82,2281	00:09:09	01:33
30	21463	82,2281	00:20:14	02:08
40	28618	82,3394	00:57:00	04:15
50	35772	82,4666	01:28:50	04:30
60	42925	82,5143	02:05:09	04:43
70	50081	82,5700	02:42:56	05:01
80	57234	82,6495	03:19:41	05:16
90	64390	82,6813	03:59:41	05:37
100	71543	82,6574	04:40:00	05:48

** Model Eğitim Süresi, h : saat, m: dakika, s:saniye*

Çizelge 4.7. Tabakalı rastgele seçim yöntemi ile örnek indirgemeye ait sonuçlar: k-kat (k=7), cv7

İndirgenmiş veri kümesi	Rasgele ve Eşsiz Seçilen Satır Sayısı	Tahmin Başarısı (%)	Model Eğitim Süresi (hh-mm-ss)	Tahmin Süresi (mm-ss)
10	7087	83,2541	00:03:16	01:08
20	14173	83,5182	00:10:53	01:47
30	21260	84,0012	00:20:23	02:20
40	28347	84,2729	00:35:28	02:53
50	35435	84,2955	00:56:28	03:16
60	42521	84,5295	01:42:06	04:42
70	49608	84,4842	02:16:27	05:14
80	56695	84,6502	02:53:34	05:39
90	63781	84,6502	03:31:57	06:36
100	70868	84,6879	04:18:13	07:17

** Model Eğitim Süresi, h : saat, m: dakika, s:saniye*

4.1.2. Hiyerarşik Kümeleme Yöntemi ile Örnek İndirgeme

7 kat çapraz doğrulama uygulanmış CB513 eğitim kümesindeki örnekler hiyerarşik bir kümeleme algoritmasıyla kümelenip, bu örnekler küme merkezlerinin en yakın komşularıyla değiştirilerek eğitim veri kümelerinin içerdiği örnek sayısı indirgenerek üç durumlu ikincil yapı tahmin başarısı hesaplanmıştır. İlk önce küme sayısı ve küme merkezine en yakın komşu sayısı optimize edilmiştir. Çizelgeler (4.8., 4.9., 4.10., 4.11., 4.12., 4.13., 4.14.) hiyerarşik kümeleme analizi ward yöntemiyle elde edilen deney sonuçlarını özetlemektedir. Çizelgelerdeki CV-Kat hangi çapraz doğrulama parçası olduğunu, k : küme merkezine en yakın komşu sayısı, N_c : küme sayısını, N_{tr} : eğitim seti örnek sayısını, Acc_v (%) : validasyon seti kullanılarak eğitilmiş SVM'den elde edilen yüzde cinsinden genel doğruluk oranını (yani Q3), Acc_t (%) ise test setleri kullanılarak elde edilen genel doğruluk oranını karşılık gelmektedir.

Çizelge 4.15. diğer çizelgelerde (4.8., 4.9., 4.10., 4.11., 4.12., 4.13., 4.14.) verilen deney sonuçlarının optimum değerlerini içermektedir. Buna göre, üçüncü ve dördüncü

çapraz doğrulama deneyleri hariç her çapraz doğrulama parçası için optimum küme sayısı 1500'dür. Tipik olarak, küme merkezinden uzaklığa bağlı olarak her kümeden en yakın 17 örnek seçilmiştir. Birinci çapraz doğrulama deneyi için, en yakın komşuların optimum sayısı 13 olmuştur. Literatürdeki sonuçlarla kıyaslanabilecek düzeyde elde edilmiş olan test seti tahmin doğruluğu, çapraz doğrulama deneyinin her bir parçası için hem indirgenmiş hem de bütün veri setleri kullanıldığında hemen hemen aynı oranda elde edilmiştir. Özetle, kullandığımız bu ikinci yöntemde tahmin doğruluğunu azaltmadan eğitim veri kümesinin %26 oranında azaltılabileceği sonucu elde edilmiştir (Çizelge 5.8., 5.9., 5.10., 5.11., 5.12., 5.13., 5.14., 5.15.). Kullanılan hiyerarşik kümeleme teknikleri arasında ward yönteminin en iyi kümeleme sonucunu sağladığı gözlenmiştir.

Tahmin doğruluğuna ek olarak, bu tez çalışmasında optimize ettiğimiz parametrelerden olan küme sayısı N_c=1000 ve en yakın komşu sayısı k=13 için, SVM sınıflandırıcısının çalışma zamanı kümeleme uygulanmış ve uygulanmamış olarak analiz edilmiştir. Bu amaçla, yedi kat çapraz doğrulama uygulanmış CB513 veri kümesinin ilk katı için aşağıda sonuçları verilen analiz gerçekleştirilmiştir. Küme sayısı N_c=1000 ve en yakın komşu sayısı k=13 olarak seçildiğinde, SVM modelinde eğitilecek veri örneği sayısı=36.622 olmaktadır. Her bir adım için çalışma süreleri aşağıdaki gibi elde edilmiştir.

- Hiyerarşik kümeleme (N_c=1000, küme sayısı) için geçen zaman : 14.51 saniye,
- Her bir küme merkezine en yakın k komşunun (k = 13) bulunması için geçen zaman : 59.25 saniye,
- 36.622 eğitim verisi kullanarak eğitilen SVM'nin eğitimi için geçen zaman : 6 saat, 16 dakika, 43 saniye sürmüştür.

Seçilen bu parametreler için hiyerarşik kümeleme yaklaşımıyla indirgenmiş eğitim verisi kullanarak eğitilen SVM'nin eğitim süresi 6 saat, 17 dakika, 56 saniye olarak elde edilmiştir. SVM eğitim süresi, CB513 veri kümesinin ilk katındaki tüm veriler kullanılarak eğitildiğinde ise 14 saat, 10 dakika, 22 saniye sürmüştür. Bu sonuçlara dayanarak, hiyerarşik kümeleme yoluyla indirgenmiş eğitim verisi kullanıldığında

elde edilen SVM çalışma süresinin, bütün eğitim veri kümesi kullanılarak eğitilen SVM eğitim süresinden daha düşük olduğu görülmektedir.

Çizelge 4.8. Hiyerarşik kümeleme yöntemi ile örnek indirgemeye ait sonuçlar: k-kat (k=7), cv1

CV-Kat	k	N_c	N_{tr}	Acc_v(%)	Acc_t(%)
cv_1	Küme ort	500	1500	79,3900	80,4706
cv_1	Küme ort	550	1650	79,3900	80,4516
cv_1	En Yakın 1	1500	4500	81,5117	80,8040
cv_1	En Yakın 3	1500	13497	81,6932	81,0708
cv_1	En Yakın 5	1500	22293	82,0563	81,2804
cv_1	En Yakın 7	1500	30578	82,5489	81,5566
cv_1	En Yakın 9	1500	37845	82,886	81,7757
cv_1	En Yakın 11	1500	43952	83,1972	81,7281
cv_1	En Yakın 13	1500	48928	83,5473	81,9567
cv_1	En Yakın 15	1500	52904	83,7417	82,0806
cv_1	En Yakın 17	1500	56049	83,8714	81,8424
cv_1	En Yakın 19	1200	53370	83,8325	81,8805
cv_1	**Verilerin tümü**		**65903**	**83,8700**	**81,7186**

Çizelge 4.9. Hiyerarşik kümeleme yöntemi ile örnek indirgemeye ait sonuçlar: k-kat (k=7), cv2

CV-Kat	k	N_c	N_{tr}	Acc_v(%)	Acc_t(%)
cv_2	Küme ort	520	1560	77,6100	78,9501
cv_2	En Yakın 1	1500	4500	79,2903	80,1939
cv_2	En Yakın 3	1500	13499	80,1463	80,7943
cv_2	En Yakın 5	1500	22500	80,4225	81,1717
cv_2	En Yakın 7	1500	30473	80,6710	81,3776
cv_2	En Yakın 9	900	23944	81,5377	80,8629
cv_2	En Yakın 11	1500	43750	80,9747	81,9952
cv_2	En Yakın 13	1500	48632	81,2371	82,1067
cv_2	En Yakın 15	1500	52477	81,3889	82,0124
cv_2	En Yakın 17	1500	55458	81,3475	82,1839
cv_2	En Yakın 19	1500	57761	81,3613	82,1067
cv_2	**Verilerin tümü**		**65212**	**81,1100**	**82,1839**

Çizelge 4.10. Hiyerarşik kümeleme yöntemi ile örnek indirgemeye ait sonuçlar: k-kat (k=7), cv3

CV-Kat	k	N_c	N_{tr}	Acc_v(%)	Acc_t(%)
cv_3	Küme ort	540	1620	81,5004	80,3872
cv_3	En Yakın 1	1200	3600	82,2677	81,0325
cv_3	En Yakın 3	1300	11698	82,8929	81,5308
cv_3	En Yakın 5	1200	17947	83,2765	81,8494
cv_3	En Yakın 7	1500	30577	83,4328	82,2251
cv_3	En Yakın 9	1500	37787	83,8733	82,6009
cv_3	En Yakın 11	1500	43833	83,7738	82,7397
cv_3	En Yakın 13	1500	48717	84,1006	82,944
cv_3	En Yakın 15	1300	48860	84,0722	82,9113
cv_3	En Yakın 17	1100	47607	84,0580	82,9848
cv_3	En Yakın 19	1100	50701	84,1432	82,8459
cv_3	**Verilerin tümü**		**64833**	**83,8733**	**83,0583**

Çizelge 4.11. Hiyerarşik kümeleme yöntemi ile örnek indirgemeye ait sonuçlar: k-kat (k=7), cv4

CV-Kat	k	N_c	N_{tr}	Acc_v(%)	Acc_t(%)
cv_4	Küme ort	570	1710	79,1585	79,3713
cv_4	En Yakın 1	1200	3600	81,6927	80,5408
cv_4	En Yakın 3	1300	11700	81,9851	81,1961
cv_4	En Yakın 5	1500	22307	82,4561	81,4117
cv_4	En Yakın 7	1500	30574	82,6673	81,8016
cv_4	En Yakın 9	1500	37828	82,7973	82,0504
cv_4	En Yakın 11	1500	43900	82,9922	82,3490
cv_4	En Yakın 13	1500	48883	83,3008	82,5066
cv_4	En Yakın 15	1500	52856	83,4308	82,6145
cv_4	En Yakın 17	1400	54249	83,3821	82,6808
cv_4	En Yakın 19	1100	50875	83,4470	82,5730
cv_4	**Verilerin tümü**		**65901**	**83,3983**	**82,5315**

Çizelge 4.12. Hiyerarşik kümeleme yöntemi ile örnek indirgemeye ait sonuçlar: k-kat (k=7), cv5

CV-Kat	k	N_c	N_{tr}	Acc_v(%)	Acc_t(%)
cv_5	Küme ort	510	1530	81,8018	79,9138
cv_5	En Yakın 1	1500	4500	81,2898	79,5591
cv_5	En Yakın 3	1400	12599	81,9817	79,9476
cv_5	En Yakın 5	1200	17908	82,2447	80,2264
cv_5	En Yakın 7	1500	30549	82,7567	80,4544
cv_5	En Yakın 9	1500	37737	83,3795	81,0119
cv_5	En Yakın 11	1500	43816	83,8085	81,2822
cv_5	En Yakın 13	1200	41981	83,6839	81,1977
cv_5	En Yakın 15	1400	50687	83,9607	81,6285
cv_5	En Yakın 17	1500	55628	84,1406	81,8397
cv_5	En Yakın 19	1500	57957	84,1544	81,8228
cv_5	**Verilerin tümü**		**65048**	**84,2652**	**82,0593**

Çizelge 4.13. Hiyerarşik kümeleme yöntemi ile örnek indirgemeye ait sonuçlar: k-kat (k=7), cv6

CV-Kat	k	N_c	N_{tr}	Acc_v(%)	Acc_t(%)
cv_6	Küme ort	580	1740	79,9852	79,6040
cv_6	En Yakın 1	1000	3000	79,9852	80,3276
cv_6	En Yakın 3	1400	12597	80,7621	80,7809
cv_6	En Yakın 5	1500	22330	80,6635	81,0035
cv_6	En Yakın 7	1500	30539	80,8608	81,2261
cv_6	En Yakın 9	1500	37617	81,2431	81,5124
cv_6	En Yakın 11	1500	43464	81,613	81,6953
cv_6	En Yakın 13	1500	48200	81,7857	81,8464
cv_6	En Yakın 15	1500	51923	81,7733	82,1485
cv_6	En Yakın 17	1500	54810	81,9583	82,4427
cv_6	En Yakın 19	1200	52368	81,9583	82,1326
cv_6	**Verilerin tümü**		**63428**	**82,1063**	**82,6733**

Çizelge 4.14. Hiyerarşik kümeleme yöntemi ile örnek indirgemeye ait sonuçlar: k-kat (k=7), cv7

CV-Kat	k	N_c	N_{tr}	Acc_v(%)	Acc_t(%)
cv_7	Küme ort	570	1710	80,5053	81,9259
cv_7	En Yakın 1	1400	4200	80,7328	82,5447
cv_7	En Yakın 3	1500	13498	80,9843	84,3106
cv_7	En Yakın 5	1400	20850	81,2957	83,5635
cv_7	En Yakın 7	1300	26782	81,9303	83,7597
cv_7	En Yakın 9	1500	37558	82,1937	84,3106
cv_7	En Yakın 11	1500	43373	82,5769	84,5144
cv_7	En Yakın 13	1500	47993	82,8643	84,6502
cv_7	En Yakın 15	1300	47987	82,9601	84,5446
cv_7	En Yakın 17	1500	54330	83,0679	84,8011
cv_7	En Yakın 19	1500	56419	83,2116	84,6804
cv_7	**Verilerin tümü**		**62517**	**83,0439**	**84,7257**

Çizelge 4.15. Hiyerarşik kümeleme yöntemi ile örnek indirgemeye ait sonuçlar: k-kat (k=7), cv1- cv7

CV-Kat	k	N_c	N_{tr}	Acc_v(%)	Acc_t(%)
cv_1	13	1500	48928	83.5473	81.9567
cv_1		Bütün veri	65903	83.8700	81.7186
cv_2	17	1500	55458	81.3475	82.1839
cv_2		Bütün veri	65212	81.1100	82.1839
cv_3	17	1100	47607	84.0580	82.9848
cv_3		Bütün veri	64833	83.8733	83.0583
cv_4	17	1400	54249	83.3821	82.6808
cv_4		Bütün veri	65901	83.3983	82.5315
cv_5	17	1500	55628	84.1406	81.8397
cv_5		Bütün veri	65048	84.2652	82.0593
cv_6	17	1500	54810	81.9583	82.4427
cv_6		Bütün veri	63428	82.1063	82.6733
cv_7	17	1500	54330	83.0679	84.8011
cv_7		Bütün veri	62517	83.0439	84.7257

4.1.3. Ortalama Doğruluk Oranı

Çizelge 4.16., CB513 veri kümesi üzerinde gerçekleştirilen 7 kat çapraz doğrulama deneyinin 7 katından elde edilen doğruluk oranlarının ortalamasını ve standart sapmasını özetlemektedir. Çizelge 4.16.'daki ilk iki satır, eğitim veri kümesi örnek azaltma yöntemlerinden elde edilen sonuçları içerir. Son satır, tüm örneklerin SVM sınıflandırıcısını eğitmek için kullanıldığı durumu temsil eder. Düşük standart sapma değerlerine sahip olmak, doğruluk değerlendirmelerinin sağlam olduğunu ve modellerin yeterince büyük veri örnekleri ile eğitildiğini göstermektedir. Örnek indirgeme yöntemlerinin doğruluk değerleri ile tüm örnekleri kullanan yöntem arasındaki farkın istatistiksel olarak anlamlı olup olmadığını değerlendirmek için, iki yönlü Z testi (two-tailed Z test), %95'lik bir güven aralığı ile gerçekleştirilmiştir. Bu teste dayanarak, tabakalı rastgele seçim ile veri örneği indirgemesi ile tüm örnekleri kullanan yöntem arasındaki doğruluk farkının, -0.0217 Z-puanı ve 0.98404 p değeri ile istatistiksel olarak anlamlı olmadığı bulunmuştur. Öte yandan, hiyerarşik kümelemeyle veri örneği indirgemesi ile tüm örnekleri kullanan yöntem arasındaki doğruluk farkı, Z-puanı -4.8713 ve p değeri $<1 \times 10^{-5}$ değerleri ile istatiksel olarak anlamlı bulunmuştur. Bu sonuçlara dayanarak, protein ikincil yapı tahmini problemi için tabakalı rastgele örneklemeyle veri örneği indirgemenin hiyerarşik kümeleme yaklaşımı ile veri örneği indirgemeye göre daha iyi ortalama doğruluk oranına sahip olduğu elde edilmiştir.

Çizelge 4.16. CB513 veri kümesi için 7 kat çapraz doğrulama deneyinden elde edilen test doğruluğunun ortalama ve standart sapması

Yöntem	Acc_t (%)	std(Acc_t)
1. yöntem : tabakalı rastgele örnekleme	82.728	0.947
2. yöntem: hiyerarşik kümeleme	81.825	1.285
Bütün veriler	82.732	0.958

4.2. Tartışma

Protein ikincil yapısı, 1951'de tanımlanmasından günümüze kadar uzanan uzun bir tarihçeye sahiptir. Ve bugün sağlık sektöründe yeni ilaçların tasarım süreci gibi alanlarda uygulama imkanına sahip olan, amino asit dizilimi kullanılarak üçüncül yapının tahmin karmaşıklığını basitleştirecek bir ara aşama sunan ve yıllardır üzerinde çalışılan bilimsel bir çalışma alanı haline gelmiştir. Konunun 1990'larda ilgi çekmeye başlaması ve dizi hizalama çalışmalarının yaygınlaşmasıyla birlikte araştırmacılar doğruluk oranının iyileştirilmesi amacıyla protein ikincil yapı tahmini için çeşitli metot ve algoritmalar kullanmışlardır. Yapısı bilinen binlerce proteinin varlığına karşın, deneysel olarak yapısı çözülememiş pek çok protein de mevcuttur. Bununla birlikte deneysel yöntemlerin emek yoğun ve zaman alıcı olması nedeniyle, yapısı bilinmeyen proteinlerin yapı tahmininde çeşitli hesaplama yöntemlerinin ve makine öğrenmesi yöntemlerinin kullanımı kaçınılmaz olmuştur. Protein ikincil yapı tahmininde kullanılan hesaplama yöntemlerinden elde edilen doğruluk oranları günümüze kadar sürekli bir iyileşme göstermiştir.

Bu tez çalışmasında DBN ve SVM kullanan iki aşamalı hibrit bir sınıflandırıcının protein ikincil yapı tahmini için gelişmiş tahmin doğruluğu sağladığı gösterilmiştir. Veri azaltma teknikleriyle elde edilen sonuçlar, orijinal veri kümesiyle elde edilen doğruluk oranları ile kıyaslandığında benzer başarı oranlarına ulaşılmıştır. Bununla birlikte SVM'nin model eğitim süresi karesel optimizasyon nedeniyle büyük veri kümeleri için verimli olamamaktadır. Bu çalışmada, SVM'nin eğitim setindeki örnek sayısını azaltmak için CB513 veri kümesi üzerinde iki farklı yöntem uygulanmıştır. Elde edilen sonuçların SVM'nin eğitim süresinin iyileştirilmesine katkıda bulunduğu gösterildiğinden gelecekte yapılacak olan protein yapı tahmini çalışmaları için faydalı olacağı düşünülmektedir.

4.2.1. Gelecekteki Çalışmalar

Gelecekte planlanan çalışmalar aşağıda maddeler halinde sıralanan konuları içerecek şekilde planlanmaktadır.

- Bundan sonraki süreçte, kümeleme çalışmalarıyla benzerlik gösteren parçalama (declustering) tabanlı yaklaşımlar denenecektir. Veri uzayını alt uzaylara parçalayan bu yöntem ile, hiyerarşik kümelemenin üst seviyelerinde bir yerden başlayıp SVM modeli eğitip daha sonra support vektörlere karşılık gelen kümeleri bir seviye açıp sonra tekrar model eğitilecektir. Dolayısıyla hiyerarşik kümeleme ağacını tek bir seviyede kesmek yerine support vektörlere karşılık gelen veri örneklerini homojen olmayan bir şekilde seçmek hedeflenecektir.

- CB513 veri kümesi için farklı kümeleme yaklaşımlarının denenmesi : kümeleme topluluğu metodu (ensemble method) : CB513 veri kümesi için farklı kümeleme yöntemleri kullanılıp bu kümeleme yöntemlerinden elde edilen sonuçlar bireysel kümeleme algoritmalarına göre daha verimli, tutarlı ve güvenilir bir kümeleme topluluğu oluşturacak şekilde çoklu kümeleme modelleri birleştirilecektir. Böylece tek bir kümeleme yaklaşımından elde edilen sonuçlardan daha etkili bir ortak çözümün bulunması hedeflenmektedir.

- Aynı yöntemler, diğer bir boyutlu tahmin problemlerinden olan çözücü erişilirlik ve bükülme açısı tahmini için uygulanarak, sınıflandırma yönteminin doğruluk ve eğitim süresindeki iyileşmesi analiz edilecektir.

KAYNAKLAR

[1] Nelson, D.L., Lehninger Biyokimyanın İlkeleri. Palme Yayıncılık, 2016.

[2] Berg, J.M., Tymoczko, J.L., Stryer, L., Biochemistry, W H Freeman, New York 2002.

[3] Bujnicki, J.M., Prediction of protein structures, functions, and interactions. Wiley Online Library, 2009.

[4] Alberts, B., Johnson, A., Lewis, J., Raff, M., Roberts, K., Walter, P., Molecular Biology of the Cell, Garland Science, New York, 2002.

[5] Anfinsen, C.B., The formation and stabilization of protein structure. Biochemical Journal. 128(4): 737, 1972.

[6] Liu, W., Chou, K.C., Prediction of protein secondary structure content. Protein Engineering. 12(12): 1041-1050, 1999.

[7] Zhou, T., Shu, N., Hovmöller, S., A novel method for accurate one-dimensional protein structure prediction based on fragment matching. Bioinformatics. 26(4): 470-477, 2009.

[8] Cooper, C., Packer, N., Williams, K., Amino acid analysis protocols, Springer Science & Business Media, 2001.

[9] Lodish, H., Berk, A., Zipursky, S.L., Matsudaira, P., Baltimore, D., Darnell, J., Molecular cell biology, National Center for Biotechnology Information, 2000.

[10] Yaseen, A., Li, Y., Template-based prediction of protein 8-state secondary structures, 2013 IEEE 3rd International Conference on Computational Advances in Bio and medical Sciences (ICCABS). pp. 1-2, 2013.

[11] Kosloff, M., Kolodny, R., Sequence-similar, structure-dissimilar protein pairs in the PDB. Proteins: Structure, Function, and Bioinformatics. 71(2): 891-902, 2008.

[12] E. Aygün, Protein İşlev Kestiriminde Yapısal Bilginin Ve Dizi Geçiş Olasılıkları İle Peptit Sınıflandırma. Yüksek Lisans Tezi, İstanbul Teknik Üniversitesi, İstanbul, 2009.

[13] Kaiser, C.M., Liu, K., Folding up and moving on-nascent protein folding on the ribosome. Journal of molecular biology. 430(22): 4580-4591, 2018.

[14] Mukherjee, A., Morales-Scheihing, D., Butler, P.C., Soto, C., Type 2 diabetes as a protein misfolding disease. Trends in molecular medicine. 21(7): 439-449, 2015.

[15] Chaudhuri, T.K., Paul, S., Protein-misfolding diseases and chaperone-based therapeutic approaches. The FEBS journal. 273(7): 1331-1349, 2006.

[16] Jankovic, B.G., Polovic, N.D., The protein folding problem. Biologia Serbica. 39(1), 2017.

[17] Zvelebil, M., Baum, J., Understanding bioinformatics. Garland Science, 2007.

[18] Jonic, S., Vénien-Bryan, C., Protein structure determination by electron cryo-microscopy. Current opinion in pharmacology. 9(5): 636-642, 2009.

[19] Belviso, B.D., Caliandro, R., Salehi, S.M., Di Profio, G., Caliandro, R., Protein Crystallization in Ionic-Liquid Hydrogel Composite Membranes. Crystals. 9(5): 253, 2019.

[20] Tsihrintzis, G.A., Sotiropoulos, D.N., Jain, L.C., Machine Learning Paradigms: Advances in Data Analytics. In Machine Learning Paradigms. Springer, 2019.

[21] Langlois, R.E., Lu, H., Machine Learning for Protein Structure and Function Prediction. In Annual Reports in Computational Chemistry. Vol. 4: 41-66, Elsevier, 2008.

[22] Kessel, A., Ben-Tal, N., Introduction to proteins: structure, function, and motion. CRC Press, London, 2010.

[23] Abdulganiyu, Y.A., Zahraddeen, S., Mamman, K.Y., Suleiman, A.U., Comparison of Popular Bioinformatics Databases. International Journal of Applied and Advanced Scientific Research. 1(1):19-28, 2016.

[24] Li, Y., Chen, L., Big biological data: challenges and opportunities. Genomics, proteomics & bioinformatics. 12(5):187, 2014.

[25] Mehmood, M.A., Sehar, U., Ahmad, N., Use of bioinformatics tools in different spheres of life sciences. Journal of Data Mining in Genomics & Proteomics. 5(2):1, 2014.

[26] Rocha, M., Ferreira, P.G., Bioinformatics Algorithms: Design and Implementation in Python, Academic Press, 2018.

[27] NCBI Statistics, https://www.ncbi.nlm.nih.gov/genbank/statistics (Erişim Tarihi: 19.08.2019).

[28] Koonin, E.V., Galperin, M.Y., Sequence-Evolution-Function: Computational Approaches in Comparative Genomics. Kluwer Academic, Boston, 2003.

[29] Tong, J.C., Ranganathan, S., Computer-aided vaccine design, Elsevier, 2013.

[30] RCSB Protein Data Bank - PDB , https://www.rcsb.org (Erişim Tarihi: 10.08.2017).

[31] PDB Statistics: Protein-only Structures Released Per Year, https://www.rcsb.org/stats/growth/protein (Erişim Tarihi: 20.08.2019).

[32] Babu, M.M., Biological databases and protein sequence analysis. Center for Biotechnology. Anna University, Chenna, 1997.

[33] Yang, Y., Gao, J., Wang, J., Heffernan, R., Hanson, J., Paliwal, K., Zhou, Y., Sixty-five years of the long march in protein secondary structure prediction: the final stretch?. Briefings in bioinformatics. 19(3): 482-494, 2016.

[34] Chazelle, B., Kingsford, C., Singh, M., A semidefinite programming approach to side chain positioning with new rounding strategies, INFORMS Journal on Computing. 16(4): 380-392, 2004.

[35] Hanson, J., Paliwal, K., Litfin, T., Yang, Y., Zhou, Y., Improving prediction of protein secondary structure, backbone angles, solvent accessibility and contact numbers by using predicted contact maps and an ensemble of recurrent and residual convolutional neural networks. Bioinformatics. 35(14): 2403-2410, 2018.

[36] Lin, L., Yang, S., Zuo, R., Protein secondary structure prediction based on multi-SVM ensemble. In 2010 International Conference on Intelligent Control and Information Processing. pp. 356-358, 2010.

[37] Hua, S., Sun, Z., A novel method of protein secondary structure prediction with high segment overlap measure: support vector machine approach. Journal of molecular biology. 308 (2): 397-407, 2001.

[38] Jun, S.H., Support Vector Machine based on Stratified Sampling, 9(2):141-146, 2009.

[39] Hens, A.B., Tiwari, M.K., Computational time reduction for credit scoring: An integrated approach based on support vector machine and stratified sampling method. Expert Systems with Applications. 39(8): 6774-6781, 2012.

[40] Awad, M., Khan, L., Bastani, F., Yen, I.L., An effective support vector machines (SVMs) performance using hierarchical clustering. In 16th IEEE International Conference on Tools with Artificial Intelligence. pp. 663-667, 2004.

[41] Yu, H., Yang, J., Han, J., Classifying large data sets using SVMs with hierarchical clusters. In Proceedings of the ninth ACM SIGKDD international conference on Knowledge discovery and data mining, pp. 306-315, 2003.

[42] Aydin, Z., Altunbasak, Y., Borodovsky, M., Protein secondary structure prediction for a single-sequence using hidden semi-Markov models, BMC Bioinformatics. 7 (1): 178, 2006.

[43] Rashid, S., Saraswathi, S., Kloczkowski, A., Sundaram, S., Kolinski, A., Protein secondary structure prediction using a small training set (compact model) combined with a Complex-valued neural network approach. BMC Bioinformatics. 17 (1): 362, 2016.

[44] Wang, L.H., Liu, J., Li, Y.F., Zhou, H.B., Predicting protein secondary structure by a support vector machine based on a new coding scheme, Genome Informatics. 15 (2): 181-190, 2004.

[45] Aydin, Z., Singh, A., Bilmes, J., Noble, W.S., Learning sparse models for a dynamic Bayesian network classifier of protein secondary structure. BMC bioinformatics. 12(1): 154, 2011.

[46] Atasever, S., Aydın, Z., Erbay, H., Protein Secondary Structure Prediction Using Support Vector Machine and Hierarchical Clustering. 3rd World Conference on Big Data, BIGDATA-2018, Kuşadası, Turkey, 2018.

[47] Atasever, S., Aydın, Z., Erbay, H., Sabzekar, M., Sample Reduction Strategies for Protein Secondary Structure Prediction. Applied Sciences. 9(20): 4429, 2019.

[48] Francis, J.W., Studying amino acid sequence using word processing programs. The American Biology Teacher. 56(8): 484-487, 1994.

[49] Branden, C.I., Tooze, J., Introduction to protein structure. Garland Science, 2012.

[50] Webster, D.M., Protein structure prediction: methods and protocols. Vol. 143, Springer Science & Business Media, 2000.

[51] Oren, E.E., Tamerler, C., Sahin, D., Hnilova, M., Seker, U.O.S., Sarikaya, M. Samudrala, R., A novel knowledge-based approach to design inorganic-binding peptides, Bioinformatics. 23 (21): 2816-2822, 2007.

[52] Chakraborty, A., Bandyopadhyay, S., FOGSAA: Fast optimal global sequence alignment algorithm, Scientific reports. 3:1746, 2013.

[53] Daugelaite, J., O'Driscoll, A., Sleator, R.D., An overview of multiple sequence alignments and cloud computing in bioinformatics. ISRN Biomathematics, 2013.

[54] Rosenberg, M.S., Sequence alignment: methods, models, concepts, and strategies, Univ of California Press, 2009.

[55] Wang, L., Jiang, T., On the complexity of multiple sequence alignment, Journal of computational biology. 1 (4): 337-348, 1994.

[56] Just, W., Computational complexity of multiple sequence alignment with SP-score. Journal of computational biology. 8 (6): 615-623, 2001.

[57] Mount, D.W., Sequence and genome analysis, Bioinformatics: Cold Spring Harbour Laboratory Press: Cold Spring Harbour, 2, 2004.

[58] Remmert, M., Biegert, A., Hauser, A., Söding, J., HHblits : lightning-fast iterative protein sequence searching by HMM-HMM alignment. Nature methods. 9(2):173, 2012.

[59] Bioinformatics Toolkit, http://toolkit.tuebingen.mpg.de/hhblits/help_ov (Erişim Tarihi: 10.08.2015).

[60] Rana, S., Jasola, S., Kumar, R., A hybrid sequential approach for data clustering using K-Means and particle swarm optimization algorithm. International Journal of Engineering, Science and Technology. 2(6), 2010.

[61] Alashwal, H., El Halaby, M., Crouse, J.J., Abdalla, A., Moustafa, A.A., The Application of Unsupervised Clustering Methods to Alzheimer's Disease. Frontiers in computational neuroscience. 13, 2019.

[62] Madhulatha, T.S., An overview on clustering methods. arXiv preprint arXiv: 1205.1117, 2012.

[63] Karabulut, E., Karacaoğlu, E., Biyoinformatik ve Biyoistatistik. Hacettepe Tıp Dergisi. 41:162-170, 2010.

[64] Everitt, B.S., Landau, S., Leese, M., Stahl, D., Cluster Analysis. John Wiley & Sons, London, 2011.

[65] Song, J., Nicolae, D.L., A sequential clustering algorithm with applications to gene expression data. Journal of the Korean Statistical Society. 38 (2): 175-184, 2009.

[66] Oyelade, J., Isewon, I., Oladipupo, F., Aromolaran, O., Uwoghiren, E., Ameh, F., Achas, M., Adebiyi, E., Clustering algorithms: Their application to gene expression data. Bioinformatics and Biology insights. 10, BBI-S38316, 2016.

[67] Flynn, J.M., Brown, E.A., Chain, F.J., MacIsaac, H.J., Cristescu, M.E., Toward accurate molecular identification of species in complex environmental samples: testing the performance of sequence filtering and clustering methods. Ecology and evolution. 5(11): 2252-2266, 2015.

[68] Coutand, O., A framework for contextual personalised applications, Kassel university press GmbH, 2009.

[69] Dias, J.G., Cortinhal, M.J., The skm algorithm: A k-means algorithm for clustering sequential data. In Ibero-American Conference on Artificial Intelligence (pp. 173-182). Springer, Berlin, Heidelberg, 2008.

[70] Rafsanjani, M.K., Varzaneh, Z.A., Chukanlo, N.E., A survey of hierarchical clustering algorithms. The Journal of Mathematics and Computer Science. 5 (3): 229-240, 2012.

[71] Aggarwal, C.C., Reddy, C.K., Data clustering. Algorithms and Application. Boca Rat. CRC Press, 2014.

[72] Kabsch, W., Sander, C., Dictionary of protein secondary structure: pattern recognition of hydrogen-bonded and geometrical features. Biopolymers: Original Research on Biomolecules. 22 (12): 2577-2637, 1983.

[73] Yan, P.V., Bioinformatics: new research. Nova Publishers, 2005.

[74] Kifer, I., Nussinov, R., Wolfson, H.J., Protein structure prediction using a docking-based hierarchical folding scheme. Proteins: Structure, Function, and Bioinformatics. 79 (6): 1759-1773, 2011.

[75] Cuff, J.A., Barton, G.J., Evaluation and improvement of multiple sequence methods for protein secondary structure prediction. Proteins: Structure, Function, and Bioinformatics. 34 (4): 508-519, 1999.

[76] Altschul, S.F., Madden, T.L., Schäffer, A.A., Zhang, J., Zhang, Z., Miller, W., Lipman, D.J, Gapped BLAST and PSI-BLAST : A new generation of protein database search programs. Nucleic acids research. 25 (17): 3389-3402, 1997.

[77] Zhu, X.J., Feng, C.Q., Lai, H.Y., Chen, W., Hao, L., Predicting protein structural classes for low-similarity sequences by evaluating different features. Knowledge-Based Systems. 163: 787-793, 2019.

[78] NCBI: National Center for Biotechnology Information, ncbi.nlm.nih.gov (Erişim Tarihi: 05.04.2018).

[79] Hu, H., Li, Z., Elofsson, A., Xie, S., A Bi-LSTM Based Ensemble Algorithm for Prediction of Protein Secondary Structure. Applied Sciences. 9 (17): 3538, 2019.

[80] PDB70 Database, wwwuser.gwdg.de/~compbiol/data/hhsuite/databases/hhsuite_dbs/old-releases/ (Erişim Tarihi: 04.05.2016).

[81] Aydin, Z., Kaynar, O., Görmez, Y., Dimensionality reduction for protein secondary structure and solvent accesibility prediction. Journal of Bioinformatics and Computational Biology. 16 (5) 1850020, 2018.

[82] Görmez, Y., Dimensionality Reduction for protein secondary structure prediction. Yüksek Lisans Tezi. Abdullah Gül Üniversitesi, Kayseri, 2017.

[83] Vapnik, V.N., Statistical Learning Theory. Adaptive and Learning Systems for Signal Processing, Communications, and Control. John Wiley & Sons, 1998.

[84] Gunn, S.R., Support vector machines for classification and regression. ISIS technical report. 14 (1): 5–16, 1998.

[85] Vert, J.P., Kernel methods in genomics and computational biology. arXiv preprint q-bio/0510032, 2005.

[86] Cortes, C., Vapnik, V., Support-Vector Networks. Machine learning, 20(3): 273-297, 1995.

[87] Cortes, C., Prediction of Generalization Ability in Learning Machines. Doktora Tezi, University of Rochester, New York, 1995.

[88] Zhang, Q., Wang, H., Yoon, S.W., A Hierarchical Feature Selection Model using Clustering and Recursive Elimination Methods. In IIE Annual Conference. Proceedings (pp. 416-421). Institute of Industrial and Systems Engineers (IISE), 2017.

[89] Scientific Python, https://www.scipy.org/ (Erişim Tarihi: 12.04.2017).

[90] Ward Jr, J.H., Hierarchical grouping to optimize an objective function. Journal of the American statistical association. 58 (301): 236-244, 1963.

[91] Ward's method, https://en.wikipedia.org/wiki/Ward%27s_method (Erişim Tarihi: 03.10.2019).

[92] Müllner, D., Modern hierarchical, agglomerative clustering algorithms. arXiv preprint arXiv:1109.2378, 2011.

[93] SciPy, https://docs.scipy.org/doc/scipy/reference/generated/scipy.cluster.hierarchy.linkage.html (Erişim Tarihi: 23.09.2019).

ÖZGEÇMİŞ

Adı Soyadı : Sema ATASEVER

Doğum Tarihi : 01.06.1980

Yabancı Dil : İngilizce (YDS:73.75, YÖKDİL:83.75)

Eğitim Durumu : (Kurum ve Yıl)

Lisans : Mersin Üniversitesi, Mühendislik Fakültesi, Bilgisayar Mühendisliği, 2001.

Yüksek Lisans : Gazi Üniversitesi, Eğitim Bilimleri Enstitüsü, Bilgisayar ve Öğretim Teknolojileri Eğitimi, 2007.

Çalıştığı Kurum/Kurumlar ve Yıl/Yıllar:

Mersin Üniversitesi Bilgi İşlem Araştırma ve Uygulama Merkezi, Kısmi Zamanlı Öğrenci, Mersin, 1998-2001.

Teclinn Teknoloji Hizmetleri Ltd.Şti, Programcı, Mersin, 2001-2002.

Gazi Üniversitesi, Öğrenci İşleri Dairesi Başkanlığı, Programcı, Ankara, 2002-2009.

Nevşehir Hacı Bektaş Veli Üniversitesi, Bilgi İşlem Dairesi Başkanlığı, Uzman, Nevşehir, 2009-2010.

Nevşehir Hacı Bektaş Veli Üniversitesi, Avanos Meslek Yüksekokulu, Öğretim Görevlisi, Nevşehir, 2010 – Devam Ediyor.

Yayınları (SCI/SCIE) :

1. Atasever, S., Aydın, Z., Erbay, H., & Sabzekar, M. (2019). Sample Reduction Strategies for Protein Secondary Structure Prediction. Applied Sciences, 9(20), 4429.

Yayınları (Uluslararası Bildiriler) :

1. Uluslararası Özet Bildiri, Sözlü Sunum, Atasever Sema, Aydın Zafer, Erbay Hasan, Protein Secondary Structure Prediction Using Support Vector Machine and Hierarchical Clustering, BIGDATA 2018, Kuşadası, Türkiye, 2018.

Projeleri (TÜBİTAK Projesi 3501) :

1. Zenginleştirilmiş Öznitelikler ve Makine Öğrenmesi Yöntemleriyle Protein Yerel Yapı Tahmini, 113E550, 18.05.2015, Bursiyer.

Projeleri (Avrupa Birliği, Leonardo Da Vinci Partnership) :

2. Microcontroller Aplications in Vocational Education, 2011-1-TR1-LE004-27425-10, Contact Person.

Araştırma Alanları : Biyoinformatik, Makine Öğrenmesi, Kümeleme