YETİM PROTEİNLERDE İKİNCİL YAPI ÖNGÖRÜSÜ İÇİN EĞİTİM KÜMESİ İNDİRGEME YÖNTEMLERİ

TRAINING SET REDUCTION METHODS FOR SINGLE SEQUENCE PROTEIN SECONDARY STRUCTURE PREDICTION

İsa Kemal Pakatcı[1], Zafer Aydın[2], Hakan Erdoğan[1], Yücel Altunbaşak[2]

[1]Mühendislik ve Doğa Bilimleri Fakültesi
Sabancı Üniversitesi, Tuzla 34956 İstanbul
[2]School of Electrical and Computer Engineering
Georgia Institute of Technology, Atlanta, GA 30332-0250, USA

isakemal@su.sabanciuniv.edu, aydinz@ece.gatech.edu,
haerdogan@sabanciuniv.edu, yucel@ece.gatech.edu

Özetçe

Yetim proteinler yaygın olarak kullanılan ve amino asit dizisi seviyesindeki benzerlikleri araştıran yöntemlerle incelendiğinde veritabanındaki proteinlerin hemen hiçbirine benzememektedir. Bu proteinlerin işlevlerinin daha iyi anlaşılması için yapısal bilgileri de kullanan yöntemlere ihtiyaç vardır. Bu amaçla yapı öngörüsü önem kazanmaktadır. Tek-dizi algoritmalar olarak da bilinen ve yetim proteinlerin ikincil yapısını öngörmek için geliştirilen algoritmalar, çoklu hizalama ve hizalama profilleri gibi benzer proteinlerden türetilen bilgileri kullanamamaktadırlar. Bu durum, algoritmaların başarı düzeylerinde sınırlayıcı bir etken olmaktadır. Tek-dizi algoritmaların başarılarını arttırmak için araştırılan tekniklerden biri de yeniden eğitim yöntemidir. Bu yöntemde ilk önce algoritmanın kullandığı model seçkin proteinlerden oluşan bir kümeyle eğitilir ve ikincil yapı öngörüsü hesaplanır. Daha sonra belli bir uzaklık ölçütü kullanılarak eğitim kümesinden, sınanacak proteine benzemeyen proteinler çıkartılır ve model indirgenmiş küme ile tekrar eğitilerek ikincil yapı öngörüsü hesaplanır. Bu çalışmada IPSSP [1] algoritmasında kullanılan saklı yarı Markov modellerini yeniden eğitmek için eğitim kümesi indirgeme yöntemleri karşılaştırılmış ve karşılaştırılan yöntemler arasında en iyisinin bileşim tabanlı indirgeme yöntemi olduğu bulunmuştur. Ayrıca eşik değer konarak yapılan eğitim kümesi indirgeme yönteminin, ilk %80 proteini seçme yöntemine göre daha iyi sonuç verdiği görülmüştür.

Abstract

Orphan proteins are characterized by the lack of significant sequence similarity to almost all proteins in the database. To infer the functional properties of the orphans, more elaborate techniques that utilize structural information are required. In this regard, the protein structure prediction gains considerable importance. Secondary structure prediction algorithms designed for orphan proteins (also known as single-sequence algorithms) cannot utilize multiple alignments or aligment profiles, which are derived from similar proteins. This is a limiting factor for the prediction accuracy. One way to improve the performance of a single-sequence algorithm is to perform re-training. In this approach, first, the models used by the algorithm are trained by a representative set of proteins and a secondary structure prediction is computed. Then, using a distance measure, the original training set is refined by removing proteins that are dissimilar to the initial prediction. This step is followed by the re-estimation of the model parameters and the prediction of the secondary structure. In this paper, we compare training set reduction methods that are used to re-train the hidden semi-Markov models employed by the IPSSP algorithm. We found that the composition based reduction method has the highest performance compared to the other reduction methods. In addition, threshold-based reduction performed better than the reduction technique that selects the first 80% of the dataset proteins.

1. Giriş

Bir gen dizisindeki bütün protein kodlayan genlerin ve bu genler tarafından kodlanan proteinlerin hücre içi görevlerinin belirlenmesi modern biyolojide önemli bir problemdir. Protein yapıları yani üç boyutlu katlanmaları bilgisi de özellikle yetim proteinler için protein işlevini belirlemeye yarayabilir. Bütün genlerin ve proteinlerin deneysel olarak çalışılabileceğini düşünmek gerçekçi değildir. Bununla beraber nispeten ucuz, hızlı ve hesaba dayalı yaklaşımlarla, DNA'daki protein kodlayan bölgeleri güvenilir bir şekilde belirlemek ve bu kodlanmış proteinlerin yapısal ve işlevsel gruplarını tahmin etmek mümkündür [2]. Her geçen gün hesaba dayalı metotlarla protein kodlayan genler keşfedilmektedir. Bu yeni proteinlerin yapılarının belirlenmesi için güvenilir hesaba dayalı yöntemlerin geliştirilmesi önem ve aciliyet kazanmıştır.

Proteinin işlevinin tahmini için kullanılan tipik yaklaşım, proteinin amino asit dizisi ile veri tabanındaki fonksiyonu bilinen dizileri karşılaştırmaktır. Bu işlem Smith-Waterman metodu [3] ve bu metodun verimli yakınsamaları olan (örneğin, BLAST [4] ve FASTA [5] gibi) ikili hizalama algoritmaları ile yapılabilmektedir. Dizi seviyesinde bilinen bir proteine önemli bir benzerlik söz konusuysa, bu onların yapısal yakınlıklarını göstermektedir. Böylece eldeki proteinin işlevi yüksek bir güvenilirlik derecesi ile tahmin

edilebilir [6]. Bununla beraber, dizi tabanlı tahmin yöntemlerinin başarısız olduğu binlerce yeni keşfedilen protein bulunmaktadır. Çünkü bunlara veri tabanı proteinleri içinde yeterli bir dizi benzerliği bulunan hiçbir protein bulunamamaktadır. Ayrıca, bu proteinlerin çoğu için profil tabanlı metotlar [7,8], iteratif araştırma metotları (PSIBLAST [9],SAM [10]) da uygulanamaz. Böyle durumlarda fonksiyon tahmini sadece dizi seviyesindeki karşılaştırmalara dayandırılamaz. Protein fonksiyonunun belirlenmesi için bir alternatif, protein yapılarının karşılaştırılmasıdır. Bunun sebebi de bir proteinin işlevinin ana olarak üç boyutlu şekli tarafından belirlenmesidir. Protein yapılarının amino asit dizilerine göre mutasyona karşı daha iyi korunduğu bilinmektedir. Amino asit dizilerinin önemli ölçüde farklılaştığı durumlarda bile, protein yapıları mutasyon sırasında genellikle korunmaktadır. Bu sebeple, proteinin veri tabanındaki proteinlere dizi seviyesinde benzer olmadığı tek-dizi durumlarında (yetim proteinler), yapı öngörüsü algoritmaları proteinin işlevini tahmin etmeye yardımcı olabilir.

Protein yapısının dört unsuru vardır fakat biz sadece ikincil yapıyı öngörmeye yoğunlaşacağız. İkincil yapının en önemli üç elemanı şunlardır: α-helezon (α-helix) {H}, β-lif (β-strand) {E} ve ilmik (loop) {L}.

```
K D N Q … N T V Y Amino Asitler
| | | | | | | | |
L E E E … H H L L İkincil Yapı Durumları
```

Şekil 1: İkincil yapı öngörüsü

Şekil 1'de görüldüğü gibi ikincil yapı öngörüsü, her amino aside 3 harfli {H, E, L} bir alfabeden yapısal bir durum atar. İkincil yapı öngörüsü için geliştirilmiş yöntemlerde genellikle ilk önce bir model kurulup bu modelin, ikincil yapısı bilinen proteinlerle eğitilmesi gerekmektedir. Daha sonra, eğitilmiş model, yapısı bilinmeyen bir proteinin ikincil yapısını öngörmek için kullanılır. Eğitim kümesinin niteliği çok önemlidir. Eğitim kümesinin, ikincil yapısı öngörülecek proteine benzer proteinleri içerecek şekilde indirgenmesinin hem başarıyı arttırdığı hem de hesaplama zamanını azalttığı gözlemlenmiştir [11]. Eğitim kümesinin indirgenmesindeki amaç sınanacak proteine benzerlik yönünden uzak olan proteinlerin modeli "bozmasını" engellemektir. Bu çalışmada üç farklı eğitim kümesi indirgeme yöntemi karşılaştırılmıştır.

2. İkincil Yapı Öngörü Algoritması

Bu çalışmada [1]'de belirtilen ikincil yapı öngörü algoritması (IPSSP) baz alınmıştır. Bu algoritmada saklı yarı Markov modelleri kullanılmıştır. Saklı Markov modelleri gen öngörüsü, protein profilleri, ve ikincil yapı öngörüsü gibi bir çok alanda başarıyla uygulanmıştır. Bir saklı Markov modelin temel bileşenleri durumlar ve durumlarda üretilen sembollerdir (çıktılardır). Herhangi bir anda tek bir durum aktiftir ve aktif olan durum bir sembol üretir. Durumlar arasındaki geçişler ve sembollerin üretilmesi olasılık dağılımlarıyla modellenir12.

Saklı yarı Markov modeli, saklı Markov modelinin bir türevidir. Saklı Markov modeliinde her bir durum bir tek sembol üretirken, saklı yarı Markov modelinde bir dizi sembol üretilmektedir. Ayrıca saklı Markov modelde bir durum bir sonraki adımda kendine dönüş yapabilirken, saklı yarı Markov modelinde kendine dönüş yoktur.

İkincil yapı öngörüsünde kullanılan bir saklı yarı Markov modelinde her bir ikincil yapı (H, E, L) için bir durum tanımlanır ve her durumda o yapıyı oluşturan amino asit grubu üretilir. İkincil yapılardaki amino asitlerin üretilmesinde kullanılan olasılık dağılımları modellenirken protein veri bankalarındaki veri sayısı göz önüne alınarak sadece ilintisi yüksek pozisyonlar kullanılır. Örneğin, bir α-helezon yapısının *i*'ninci pozisyonundaki amino asidin, *i-2, i-3, i-4, i+2* ve *i+4*'üncü pozisyonlardaki amino asitlerle yüksek ilinti gösterdiği bulunmuştur. Dolayısıyla *i*'ninci pozisyondaki amino asidin görülme olasılığı sadece bu beş pozisyondaki amino asitlerin çeşidine (ya da özelliklerine) bağlı olmaktadır.

IPSSP algoritmasında üç saklı yarı Markov modeli kullanılmaktadır. Birinci modelde, ikincil yapılardaki amino asitler sadece önceki pozisyonlara bağlıdır. Örneğin, bir α-helezon yapısının *i*'ninci pozisyonundaki amino asit, *i-2, i-3* ve *i-4*'üncü pozisyondaki amino asitlerin çeşidine (ya da özelliklerine) bağlıdır. İkinci modelde hem önceki hem de sonraki pozisyonlar kullanılırken, üçüncü modelde sadece sonraki pozisyonlar (*i+2* ve *i+4* v.s.) kullanılır. Bu sayede farklı pozisyonlara özgü ilintiler daha özgül olarak modellenmektedir. Ayrıca IPSSP algoritmasında her bir ikincil yapı bölgesinin baştaki, ortadaki ve sondaki amino asitleri ayrı ayrı modellenmektedir. Bunun sebebi baştaki ve sondaki amino asitlerin, ortadakilerden farklı bir dağılıma sahip olmasıdır.

IPSSP algoritması, sınanacak proteinin ikincil yapısını öngörmek için şu yöntemi kullanır:

(1) Her bir saklı yarı Markov model için, sınanacak proteindeki bir amino asidin belli bir ikincil yapıda olma olasılığını veren bir dağılım hesaplanır. Bu dağılımların hesaplanmasında posterior decoding algoritması (forward-backward ya da BCJR MAP) kullanılır [1].

(2) Bu dağılımlar kullanılarak her modelden bir ikincil yapı öngörüsü hesaplanır. Bunun için, her amino asite en yüksek olasılığa sahip ikincil yapı verilir.

(3) Her model için eğitim kümesi indirgenir. Bunun için belli bir yakınlık kriteri kullanılarak ikincil yapısı bir önceki aşamada öngörülen ikincil yapıya benzer olan proteinler seçilir. Daha sonra modeller indirgenmiş eğitim kümelerine göre tekrar eğitilir ve her model için ikincil yapı olasılık dağılımları ve ikincil yapılar tekrar hesaplanır.

(4) Üçüncü aşama belli bir sayıda tekrarlanır. Her seferinde en başta kullanılan eğitim kümesinden başlanır ve bu küme indirgenir.

(5) Son aşamada modellerden gelen dağılımların ortalaması alınarak ikincil yapı öngörüsü hesaplanır.

Deneylerimizde eğitim kümeleri sadece bir kez indirgenmiş ve modeller tekrar eğitilip ikincil yapı hesaplanmıştır.

3. Eğitim Kümesi İndirgeme

Uygulanan eğitim kümesi indirgeme yöntemlerinin hepsi tanımlanan bir çeşit benzerlik (ya da uzaklık) ölçütünü kullanır. Bu ölçütü kullanarak eleme yapmak için iki farklı yaklaşım kullanılmıştır. İlk yaklaşımda, sınanacak proteine benzeyen veya yakın olan ilk %80 protein alınır. İkinci yaklaşımda ise belirlenen bir eşik değerin (benzerlik ya da uzaklık ölçütü olmasına göre) altında veya üstünde kalan proteinler alınır.

3.1. Bileşim tabanlı eğitim kümesi indirgeme

Bu yöntemde, sınanacak proteinin ikincil yapısı öngörüldükten sonra öngörülen ikincil yapının veri kümesindeki proteinlerin ikincil yapılarına uzaklıkları hesaplanır. Bu uzaklık değeri şu şekilde hesaplanır:

$$D = \max(|H_s - H_{vt}|, |E_s - E_{vt}|, |L_s - L_{vt}|) \quad (1)$$

Burada H_s, E_s, L_s, sınanacak proteinin öngörülen ikincil yapısındaki α-helezon, β-lif ve ilmik oranları ve H_{vt}, E_{vt}, L_{vt} veritabanındaki proteinin sırasıyla α-helezon, β-lif ve ilmik oranlarıdır. Bu uzaklık ölçütü kullanılarak veritabanındaki bütün proteinler sıralanır ve bu değeri en düşük olan ilk %80 protein eğitim kümesi olarak kullanılarak bir daha öngörü yapılır. Bu yöntemin farklı bir sürümü olarak eşik değer kullanılmıştır. Veritabanındaki proteinlerinden, öngörülen ikincil yapıya olan uzaklıkları 0.35'in altında olanlar ($D < 0.35$) eğitim kümesine dahil edilmiştir. 0.35 değeri deneysel olarak bulunmuştur.

3.2. Hizalama tabanlı eğitim kümesi indirgeme

Bu yöntemde, sınanacak proteinin ikincil yapısı öngörüldükten sonra sınanacak protein ile veritabanındaki proteinler hizalanır. Daha sonra hizalanma skoru düşük olan proteinler eğitim kümesinden çıkartılır. Bu çalışmada yine önceki bölümde bahsedildiği gibi iki farklı indirgeme yöntemi denenmiştir. İlkinde hizalama skoru en yüksek %80 protein eğitim kümesine dahil edilirken ikincisinde eşik değer konarak eğitim kümesi oluşturulmuştur. Eşik değer seçmek için bileşim tabanlı yöntemdeki 0.35 değerinin hizalama yöntemlerinde hangi skor değerine karşılık geldiği hesaplanmış ve eşik değer olarak bu skor değeri alınmıştır.

Hizalama Yöntemleri*:* Proteinleri hizalamak için birbirine benzeyen semboller karşılıklı olarak eşleştirilirirken benzeri bulunamayan semboller '-' sembolü ile eşleştirilmektedir. Hizalama üç farklı şekilde yapılabilir:

1. Sadece ikincil yapıları hizalama (İY)
2. Sadece amino asit zincirlerini hizalama (AA)
3. Hem ikincil yapıları hem de aminoasit zincirlerini hizalama (İY+AA)

Birinci durumda semboller ikincil yapı çeşitleridir ve H, E, ya da L değerlerinden birini alır. İkinci durumda semboller amino asit çeşitleridir ve yirmi farklı değerden birini alır. Üçüncü durumda ise semboller ikincil yapı ve amino asit ikilileridir. Bu sayede ikincil yapı ve amino asit dizileri birlikte hizalanmış olur.

İki sembolün birbiriyle benzerlik derecesini belirlemek için bir fonksiyon (benzerlik tablosu) kullanılmaktadır. İki protein arasındaki herhangi bir hizalamanın skor değerini hesaplamak için birbiriyle eşleşen sembollerin benzerlik değerleri toplanır. Bu değere, karşılığı olmayan bölgelerin (açık bölgeler) toplam skoru eklenir. Bu şu şekilde formüle edilir:

$$S = \sum_{k=1}^{r} \left(\alpha M_{aa}(a_k, b_k) + \beta M_{iy}(c_k, d_k) \right) + G \quad (2)$$

Yukarıdaki denklemde S hizalama skoru, r birbiriyle eşleşen sembol sayısı, toplam sembolünün içindeki değer eşleşen sembollerden gelen skor, G eşi bulunamayan sembollerden (açık bölgeler) gelen toplam skor, a_k ve b_k sırasıyla birinci ve ikinci proteinin k'nıncı eşleşen amino asitleri, c_k ve d_k sırasıyla birinci ve ikinci proteinin k'nıncı eşleşen ikincil yapıları, $M_{aa}()$ ve $M_{iy}()$ sırasıyla amino asit ve ikincil yapı benzerlik tablosu, α ve β sırasıyla amino asit ve ikincil yapı eşleşme skorlarının katsayılarıdır. Sadece ikincil yapıları hizalanmak için α=0, β=1 alınır. Sadece amino asit dizileri hizalanmak için α=1, β=0 alınır. Amino asit ve ikincil yapıları birlikte hizalamak içinse α=1, β=1 alınmaktadır.

İki protein arasındaki en yüksek skora sahip hizalamanın bulunabilmesi için bütün hizalamaların olduğu uzayın taranması gerekmektedir. Bunu hızlı bir şekilde yapan algoritmalar tasarlanmıştır. Bu çalışmada, hizalama algoritması olarak Smith-Waterman [3] yerel hizalama algoritması kullanılmıştır. Bu algoritmada proteinlerin bütününü kullanma zorunluluğu yoktur. Onun yerine proteinlerin yerel olarak en çok benzeyen kısımları hizalanmaktadır.

Bu çalışmada, üç hizalama yöntemi de değerlendirilmiştir. Amino asit zincirlerini hizalamak için Blosum62 [13] tablosu kullanılmış, ikincil yapıları hizalamak için ise İkincil Yapı Benzerlik Matrisi (Secondary Structure Similarity Matrix – SSSM) [14] kullanılmıştır.

Standartlaştırma Yöntemleri: Hizalama skoru hesaplandıktan sonra belli bir katsayı ile çarpılır. Bu sayede uzunluk farklarından gelen etkiler standartlaştırılmış olur. Bu çalışmada, skorları standartlarştırmak için hizalama skorları iki proteinin ortalama uzunluğuna bölünmüştür. Bu yöntemin proteinlerin katlanma sınıflarının tespitinde başarılı olduğu gösterilmiştir.

3.3. Chou-Fasman uzaklık ölçütü kullanılarak eğitim kümesi indirgeme

Bu yöntemde Chou-Fasman [15] uzaklık skorları hesaplanmış, en düşük skora sahip ilk %80 protein eğitim kümesine alınmıştır. Chou-Fasman uzaklık skoru şöyle tanımlanmıştır:

$$\sum_{k \in \{H,E,L\}} \left(\frac{1}{l_t} \sum_{j=1}^{l_t} f_k(q(j)) - \frac{1}{l_{vt}} \sum_{j=1}^{l_{vt}} f_k(h(j)) \right)^2 \quad (3)$$

Burada l_t ve l_{vt} sınanacak proteinin ve veritabanındaki proteinin uzunluğu, $q(j)$ ve $h(j)$ sınanacak proteinin ve veritabanındaki proteinin j pozisyonundaki amino asit, $f_k(z)$ fonksiyonu ise z amino asitinin k ikincil yapısında bulunma yatkınlığını gösteren Chou-Fasman katsayısı. Bu formülde her amino asit için üç farklı ikincil yapıda olma ihtimali düşünülmüştür. Bu yöntemin başka bir sürümü de ikincil yapı bilgisi kullanarak bu skoru hesaplamaktır. Burada sınanacak proteinin öngörülen ikincil yapısını kullanılırken, veritabanındaki proteinin bilinen ikincil yapısı kullanılır. İkincil yapı bilgisi kullanan Chou-Fasman uzaklık fonksiyonu şu şekildedir:

$$\left(\frac{1}{l_t} \sum_{j=1}^{lt} f_{k(q(j))}(q(j)) - \frac{1}{l_{vt}} \sum_{j=1}^{l_{vt}} f_{k(h(j))}(h(j)) \right)^2 \quad (4)$$

Bu formülde $k(q(j))$, sınanacak proteinin j'ninci pozisyonundaki amino aside karşılık öngörülen ikincil yapıyı, $k(h(j))$ ise veritabanındaki proteinin j'ninci pozisyonundaki amino aside karşılık gelen ikincil yapıyı ifade etmektedir.

4. Deneyler ve Tartışma

Veri kümesi olarak birbirine benzemeyen proteinlerden oluşan EVA kümesi kullanılmıştır [16]. Bu kümeden 30 aminoasitten küçük proteinler çıkartılıp 2720 protein içeren bir veri kümesine ulaşılmıştır. Bu kümedeki proteinlerin

ikincil yapıları PDB (Protein Data Bank)'tan alınmıştır [17]. Bu ikincil yapıları 8 durum alfabesinden 3 durum alfabesine indirgemek için şu yöntem kullanılmıştır: H, G→ H, E, B→ E, S, T, I, ' ' → L. Chou-Fasman uzaklık ölçütünde kullanılan f fonksiyonunu hesaplamak için PDB_SELECT kümesi kullanılmıştır [18]. Bu kümedeki proteinlerin ikincil yapıları da 8 durum alfabesinden 3 durum alfabesine indirgenmiştir. Buna paralel olarak f fonksiyonu amino-asitlerin bu 3 durumda bulunma yatkınlığını göstermektedir. Yöntemlerin başarı oranlarını test etmek için veri kümesindeki ilk 600 protein üzerinden çapraz geçerlilik (cross validation) sınaması uygulanmıştır. Bunun için her seferinde veri kümesinden bir sınama proteini seçilir. Kalan proteinler eğitim kümesini oluşturur. Seçilen proteininin ikincil yapısı tahmin edilip kaydedilir ve sınama proteini tekrar veri kümesine eklenir. Bu işlem veri kümesindeki ilk 600 protein için tekrarlandıktan sonra başarı ölçütü hesaplanır. Test kümesini 600 proteine sınırlamaktaki amaç, hesaplamalarda tasarruf etmektir. Yapılan deneylerde ilk 600 proteinde alınan sonucun, veri kümesindeki tüm proteinler test edilerek alınan sonuca yakınsadığı gözlemlenmiştir. Başarı ölçütü olarak şu ölçüt (duyarlılık) kullanılmış-tır: $Q_3 = N_c / N$. Burada N_c doğru tahmin edilen aminoasit sayısı iken N ise veritabanındaki toplam aminoasit sayısıdır.

Deney sonuçları Tablo I ve II'de özetlenmiştir. Tablo I'de ilk %80 protein alınarak yöntemlerin başarı oranları verilmiştir. Bu tabloya göre en iyi üç yöntemin bileşim tabanlı, ikincil yapıyı kullanan hizalama tabanlı ve hem ikincil yapıyı hem aminoasit dizisini kullanan hizalama tabanlı yöntemler olduğu görülmektedir. Bu üç yöntem için eşik değer koyarak yapılan deneylerin sonuçları Tablo II'de verilmiştir. Tablo I ve II'de görüldüğü gibi en iyi sonucu veren bileşim tabanlı yöntemdir. Eşik değer konması uygulanan üç yöntem için de başarı oranını arttırmıştır. Sanıyoruz bunun sebebi, eşik değer konulduğunda her sınama için eğitim kümesindeki protein sayısının dinamik olarak değişebilmesidir. Ayrıca en iyi sonucu veren eşik değer koyarak bileşim tabanlı eğitim kümesi indirgeme yönteminin başarıyı %0.6 oranında arttırdığı gözlemlenmiş-tir. Bileşim tabanlı yöntemin diğerlerine göre daha hızlı çalışması bu yöntemin ek avantajıdır.

Tablo 1: İlk %80 Protein Alınarak Yöntemlerin Başarı Oranları

Yöntem	Başarı(%)
Bileşim tabanlı	67.01
Hizalama (İY)	67.00
Hizalama (İY+AA)	66.92
Hizalama (AA)	66.69
Chou Fasman	66.65
Yeniden eğitilmeden	66.59
Chou Fasman (İY)	66.50

Tablo I1: Eşik Değeri Konarak Yöntemlerin Başarı Oranları

Yöntem	Başarı(%)
Bileşim tabanlı	67.17
Hizalama (İY)	67.12
Hizalama (İY+AA)	67.06

Başarı oranları arasındaki farklar az gibi görünse de biyolojik olarak bakıldığında bu farklar önemli olabilmektedir. Bu deneyde %0.01'lik bir fark yaklaşık 12 daha fazla aminoasidin doğru işaretlenmesine karşılık gelmektedir.

5. Sonuçlar ve Gelecek Çalışmalar

Eşik değer koyarak bileşim tabanlı eğitim kümesi indirgemenin protein ikincil yapı öngörüsünün başarı oranını arttırdığı gözlemlenmiştir. Gelecekte, öngörü başarısını en iyileyecek bir eşik değerinin bulunması için çalışacağız. Bu çalışmada literatürde rastladığımız ve bu problemde manalı olarak görünen uzaklık ölçütleri kullanılmıştır. İlerdeki çalışmalarda diğer olası uzaklık ölçütleri de denenebilir.

6. Kaynakça

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

[2] E. Koonin and M. Galperin, Sequence-Evolution Function, Kluwer Academic Publishers, 2002.

[3] T. Smith and M. Waterman, "Identification of common molecular subsequences," J. Mol. Biol., vol. 147,pp. 195–197, 1981.

[4] S. F. Altschul, W. Gish, W. Miller, E. Y. Myers, and D. J. Lipman, "A basic local alignment search tool," J. Mol. Biol., vol. 215, pp. 403–410, 1990.

[5] W. R. Pearson, "Rapid and sensitive sequence comparisons with FASTP and FASTA," Methods inEnzymology, vol. 183, pp. 63–98, 1990.

[6] B. Rost, "Twilight zone of protein sequence alignments," Protein Eng., vol. 12, pp. 85-94, 1999.

[7] M. Gribskov, A. McLachlan, and D. Eisenberg, "Profile analysis: Detection of distantly related proteins," PNAS, USA, vol. 84, pp. 4355–4358, 1987.

[8] A. Krogh, M. Brown, I. S. Mian, K. Sjolander, and D. Haussler, "Hidden Markov models in computational biology: Applications to protein modeling," J. Mol. Biol., vol. 235, pp. 1501–1531, 1993.

[9] S. F. Altschul, T. L. Madden, A. A. Schaffer, J. Zhang, Z. Zhang, W. Miller, and D. J. Lipman, "Gapped BLAST and PSI-BLAST: A new generation of protein database search programs," Nucleic Acids Research, vol. 25, pp. 3389–3402, 1997.

[10] K. Karplus, C. Barrett, and R. Hugley, "Hidden Markov models for detecting remote protein homologies," Bioinformatics, vol. 14, pp. 846–856, 1998.

[11] A. A Salamov and V. V. Solovyev "Prediction of protein secondary structure by combining nearest neighbor algorithms and multiple sequence alignments," J. Mol. Biol. 247, pp. 11-15, 1995.

[12] Rabiner, L.R. "A tutorial on Hidden Markov Models and selected applications in speech recognition," Proceedings of the IEEE, vol. 77(2), pp. 257-286, 1989.

[13] S. Henikoff and J.G. Henikoff, "Amino acid substitution matrices from protein blocks," P.N.A.S. USA, vol. 89, pp. 10915-10919, 1992.

[14] A. Wallqvist, Y. Fukunishi, L. R. Murphy, A. Fadel, and R. M. Levy, "Iterative sequence/secondary structure search for protein homologs: comparison with amino acid sequence alignments and applition to fold recognition in genome databases," Bioinformatics, vol. 16, pp. 998-1002, 2000.

[15] P. Chou and G. Fasman, "Empirical predictions of protein conformation," Annu. Rev. Biochem., vol. 47, pp. 251-276, 1978.

[16] EVA: secondary structure (intro): http://cubic.bioc.columbia.edu/eva/doc/into_sec.html.

[17] The Protein Data Bank: http://www.rcsb.org/pdb

[18] PDB_SELECT: http://bioinfo.tg.fh-giesen.de/pdbselect