Hough Dönüşümü Kullanılarak Protein Yapısal Bloklarının Karşılaştırılması

Protein Structural Block Comparison by Using Hough Transform

Özlem Özbudak[1], Zümray Dokur[1], Virginio Cantoni[2]

[1]Elektronik ve Haberleşme Mühendisliği Bölümü İstanbul Teknik Üniversitesi
ozbudak@itu.edu.tr, dokur@itu.edu.tr

[2]Department of Electrical and Computer Engineering University of Pavia
virginio.cantoni@unipv.it

Özet

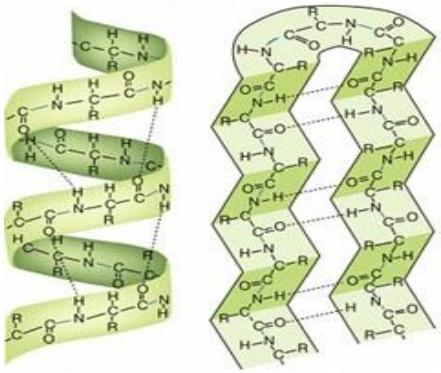
Bu çalışma motif çıkarılması için üç boyutlu protein sekonder yapılarının karşılaştırılmasında yeni bir yaklaşım sunmaktadır. Bu yaklaşım Genelleştirilmiş Hough Dönüşümüne (GHD) dayalı olarak iki ayrı şekilde test edilmektedir. Bunlardan birincisinde sekonder yapı ikilileri kullanılmaktadır ve bu ikililere ilişkin olarak, sekonder yapıların orta noktaları arasındaki mesafe, eksenleri arasındaki mesafe ve eksenler arasındaki açı karşılaştırma parametreleri olarak seçilmektedir. İkincisinde ise sekonder yapı üçlüleri kullanılmakta, sekonder yapıların orta noktaları birleştirilerek üçgenler oluşturulmakta ve bu üçgenlerin kenar uzunlukları karşılaştırma parametreleri olarak seçilmektedir. Her iki yöntemde de aranan motifin geometrik merkezi referans noktası (RN) olarak belirlenmektedir. Sonrasında birinci yöntem ve ikinci yöntem için sırasıyla motif ikilileri ve motif üçlüleri proteinlerdeki tüm ikili ve üçlülerle karşılaştırılır ve her bir eşleşme için özel bir haritalama kuralı ile belirlenen noktaya bir oy verilir. Oylama sonucunda en fazla oya sahip olan nokta aday RN olarak belirlenir. Bu çalışmada dört ve beş sekonder yapıdan oluşan motifler test edilmiştir. Test sonuçları motif RN'nin hatasız bir şekilde belirlendiğini ve motif çıkarılmasında bu iki yöntemin kolay uygulanabilir, hesaplama açısından etkin ve hızlı olduğunu göstermiştir.

Abstract

This paper presents a new approach for motif retrieval by comparing protein secondary structures. This approach is tested as two different methods based on Generalized Hough Transform (GHT). In the first one secondary structure couples are used, and midpoint distance, axis distance and axis angle, related to the couple, are considered comparison parameters. In the second one secondary structure triplets are used, the triangles are defined by joining the midpoints of secondary structures and the edge lengths of triangles are considered as comparison parameters. In both methods the barycenter of the motif is assigned as reference point (RP). Then motif couples and motif triplets are compared with protein couples and triplets respectively using the first and second methods, and for every correspondence a vote is given to the point which is defined with a special mapping rule. After the voting process, the point having the highest number of votes is defined as candidate RP. In this paper the motifs, formed by four and five secondary structures, are tested. Experimental results showed that the RP is determined precisely and both methods to retrieve the motif are simple to implement, computationally efficient and fast.

1. Giriş

Proteinler 20 standart amino asidin belirli türde, belirli sayıda ve belirli diziliş sırasında karakteristik düz zincirde birbirlerine kovalent bağlanmasıyla oluşan polipeptitlerdir. Her proteinin kendisine has özelliklerinin olmasını sağlayan özel amino asit dizilimleri vardır. Proteinler yapılarında karbon (C), hidrojen (H), oksijen (O) ve azot (N) elementlerini dolayısıyla bu elementlerden oluşan amino grubunu ve karboksil grubunu içerirler. Protein yapısı 4 ayrı şekilde incelenebilir. Belirli sayıda ve belirli türde amino asitlerin belirli diziliş sırası ile sıralanarak oluşturdukları peptid zinciri primer yapıyı ifade eder. *Sekonder yapı* peptid zincirinin yaptığı hidrojen bağlarına göre belirlenir. Bu yapıda başlıca iki tip tekrarlayan yapı vardır: α-heliks ve β-tabaka (bkz. Şekil 1). *Tersiyer yapı* sekonder yapının katlanmasıyla ve sekonder yapıdaki proteinlerin 3-boyutlu konformasyon oluşturmasıyla ortaya çıkarlar. *Kuaterner yapı*; primer, sekonder ve tersiyer yapılı tabakaların birleşmesiyle ortaya çıkan yapıdır.



Şekil 1. Sekonder yapı örnekleri: α-heliks ve β-tabaka

Proteinler tüm canlılardaki en temel birimlerdir ve organizmadaki bütün fonksiyonlar proteinler tarafından gerçekleştirilmektedir. Proteinlerin fonksiyonları onların yapıları tarafından belirlenmektedir. Protein yapılarının karşılaştırılması, onların fonksiyonlarını belirleme açısından oldukça önemlidir ve bu konuda çalışan araştırmacılara protein yapısının ve gelişiminin farklı yanlarını anlamada yardımcı olmaktadır. Bu sebeple yapısal karşılaştırma ve benzerlik analizi proteinin yaşam döngüsündeki rolünü anlamada büyük önem arz etmektedir.

Literatürde protein yapısının analizi ve yapısal karşılaştırma ile ilgili olarak değişik çalışmalara rastlanmaktadır. Bunlardan Can ve ark. [1] protein yapı benzerliğini araştırmak için yeni bir yöntem sunarlar ve kavislenme, burulma, sekonder yapı tipleri gibi özellikleri çıkarmak için proteinin üç boyutlu yapısına farksal geometri bilgisini uygularlar. Çamoğlu ve ark. [2] protein veri tabanındaki proteinler arasındaki yapı benzerliğini bulmak amacıyla R-tree kullanarak üçlü sekonder yapılar üzerinde bir indeksleme yapısı oluştururlar. Chionh ve ark. [3] protein yapılarının karşılaştırılması için sekonder yapılar arasındaki açı ve mesafe matrislerini kullandıkları SCALE isimli bir algoritma önerirler. 2007 yılında Shuoyong ve ark. [4] katlama (fold) seviyesindeki yapısal benzerlikleri bulmak ve yapısal motiflerin varlığını araştırmak için ProSMos (Protein Structure Motif Search) isimli bir program geliştirmişlerdir. Bu program kullanıcı tarafından tanımlanmış üç boyutlu sekonder yapılardan oluşan örüntüyü, protein yapılarının bulunduğu bir veri tabanında aramaktadır. Bu çalışmada bir proteinin atomik koordinatlarının kümesi sekonder yapı elementlerinin bir kümesi olarak karakterize edilmektedir. Chi ve ark. [5] görüntü tabanlı mesafe matrislerini ve çok boyutlu indeksleri kullanarak protein yapısını belirleyen hızlı bir sistem tasarlamışlardır. Protein yapısının bir boyutlu dizi gösterimi yapısal bilgiyi belli bir ölçüde saklamaktadır. Bu tip bir gösterimin protein yapısının karşılaştırılması ve sınıflandırılması için faydalı olabileceğini ortaya koyarlar. Albretch ve ark. [6] proteinler arasındaki yapısal benzerlikleri ortaya çıkarmak amacıyla farklı bir yaklaşım sunarlar ve üç boyutlu veriyi iki boyutlu veriye dönüştürerek veri azaltma tekniklerini uygularlar. Bu da proteinler arasındaki yapısal karşılaştırmaları hızlandırmaktadır. Zotenko ve ark. [7] protein yapı karşılaştırmasını hızlandırmak amacıyla farklı bir yaklaşım önerirler. Bu yaklaşıma göre bir protein yapısı yüksek boyutlu bir vektöre haritalanmakta ve aynı vektörler arasındaki mesafeler kullanılarak yapısal benzerlik ortaya çıkarılmaktadır. Cantoni ve ark. [8] yapay sinir ağlarının bir alt modeli olan SOM'u (Öz-Düzenlemeli Harita, Self-Organizing Map) kullanarak proteinleri katlama seviyesinde sınıflandıran bir çalışma yapmışlardır. Bu çalışma sekonder yapıları vektör, proteinleri yönlü graf olarak kabul ederek ve SOM kullanarak proteinleri sınıflandırmaktadır. Ayrıca Hough Dönüşümü yöntemi kullanarak sekonder yapı karşılaştırması yapmışlardır [9]. 20 adet proteinin kullanıldığı çalışmada karşılaştırma yaparken 3, 4 ve 5 sekonder yapıdan oluşan ve protein içinden seçilen motifler kullanmışlardır. Bu motiflerin içindeki sekonder yapıları tekli, ikili, üçlü olarak ele alınıp protein içindeki tekli, ikili ve üçlü sekonder yapılarla karşılaştırdıkları gibi, ilgili motif protein içindeki olası bütün motiflerle de bire bir karşılaştırılmıştır.

Önerilen çalışma sekonder yapıların karşılaştırılması ve motif çıkarılmasında kullandığı algoritma ve öznitelikler nedeniyle diğer yöntemlerden farklılık göstermektedir. Bu çalışmada öznitelik olarak protein içinde bulunan sekonder yapıların başlangıç noktası, bitiş noktası ve doğrultu koordinatları kullanılmaktadır ve aranan motifin geometrik merkezinin lokasyonu % 100 doğruluk oranı ile belirlenmektedir.

Çalışmanın geri kalan kısmı şu şekilde organize edilmiştir: İkinci bölümde Hough dönüşümünden bahsedilecek ve GHT tabanlı ikili ve üçlü sekonder yapıları kullanan metotlar açıklanacaktır. Üçüncü bölümde yapılan testler ve sonuçları gösterilecektir. Sonuç bölümünde ise elde edilen sonuçlar değerlendirilecek ve gelecek çalışmalardan bahsedilecektir.

2. Metodoloji

Bu bölümde Hough dönüşümü ve GHD tabanlı ikili ve üçlü sekonder yapıları kullanan metotlar açıklanacaktır.

2.1. Genelleştirilmiş Hough Dönüşümü

Hough Dönüşümü (HD) görüntü analizi, bilgisayarla görü ve sayısal görüntü analizinde uygulanan bir öznitelik çıkartma tekniğidir. Koordinat dönüşümü ilkesine dayanır. Genellikle obje tanımada kullanılır. Bu dönüşümün amacı kümülatif oylama prosedürü ile test edilen objenin şeklini ortaya çıkartmaktır.

Klasik HD 1962'de P.V.C Hough tarafından tanımlanan bir koordinat dönüşümüdür [10]. Sayısal görüntülerdeki doğruların ve diğer şekillerin tespiti için kullanılan bir görüntü analiz yöntemi olarak da tanımlanabilir. 1972'de R.O. Duda ve P.E. Hart tarafından daire tespiti amacıyla [11] ve H. Wechsler ve J. Sklansky tarafından parabol tespiti amacıyla [12] geliştirilmiştir. 1981'de ise Ballard tarafından rastgele şekillerin tespiti için GHD olarak genelleştirilmiştir [13]. G(HD) temelde oylama işlemine dayanmaktadır ve bu işlem Hough uzayında olmaktadır. GHD'de rastgele şekiller değişmez hareket parametrelerini içeren Hough uzayında gösterilirler. İki boyutta gösterilen bu parametreler; dönüşüm eksenlerini gösteren x ve y, dönme açısını gösteren θ ve ölçekleme faktörü olan s'dir. Bu parametreler bir tabloda saklanmakta ve oylama işlemi için gerekli olan haritalama kuralı bu parametreler tarafından belirlenmektedir.

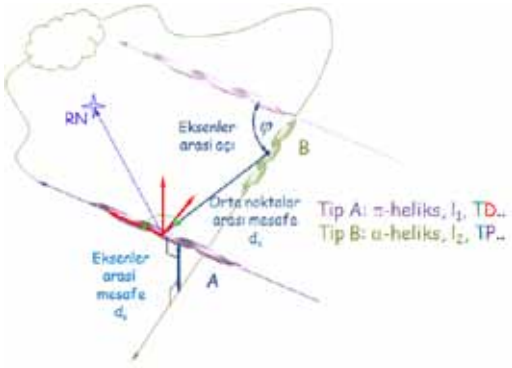
Bu çalışmada GHD, önceden belirlenmiş bir yapısal blok (motif, domain ya da protein) ile PDB (Protein Data Bank) [14] gibi bir veritabanına ait proteinlerdeki yapısal blokların karşılaştırılması amacıyla kullanılmaktadır. Bu veri bankasında protein, DNA ve RNA gibi biyolojik makromoleküllerin üç boyutlu modelleri saklanmaktadır. Bütün veriler internet üzerinden erişime açıktır. Dünyanın her tarafından biyologlar ve biyokimyacılar bu veri bankasına veri yükleyebilmektedir ve şu anda (8 Kasım 2013) itibarıyla da 95280 yapı bu veri bankasında bulunmaktadır.

2.2. İkili Metot

Bu yöntemde sekonder yapı ikilileri yerel bir referans sistemi kurarlar. Örneğin bu sistemde birinci sekonder yapının orta noktası orijin olarak kabul edilir, y-ekseni birinci sekonder yapı üzerinde, x-ekseni ise y-ekseni ve ikinci sekonder yapının orta noktası tarafından oluşturulan düzlem üzerinde ve son olarak z-ekseni de bu iki eksene dik olarak belirlenir (bkz. Şekil 2). Bu sistemde motif referans noktası (RN), onu oluşturan sekonder yapıların ağırlık merkezi olarak belirlenir.

RN hesaplanırken, motif içerisindeki her bir sekonder yapının orta noktası bulunur ve bulunan tüm orta noktaların geometrik merkezi RN olarak kabul edilir. Motif içerisindeki her bir sekonder yapı ikilileri için RN'yi tanımlayan parametreler hesaplanır. Bu parametreler açı ve mesafe bilgisinden oluşmaktadır ve Referans Tablosu'na (RT) kaydedilmektedir. Bu parametreler şöyle hesaplanır: Birinci sekonder yapının orta noktası $s_{1o}(x,y,z)$, ikinci sekonder yapının orta noktası $s_{2o}(x,y,z)$ ve uç noktası $s_{2u}(x,y,z)$ tarafından bir düzlem oluşturulur. RN'nin bu düzlem üzerindeki izdüşümü noktası $i_{RN}(x,y,z)$ hesaplanır. Sonrasında s_{1o} ile i_{RN} arasındaki ve i_{RN} ile RN arasındaki açı ve uzaklık hesaplanır: $\theta_{s_{1o}-i_{RN}}$, $\rho_{s_{1o}-i_{RN}}$, $\theta_{i_{RN-RN}}$, $\rho_{i_{RN-RN}}$. Motif içerisindeki her bir sekonder yapı ikilisi için bu parametreler hesaplanır. Bu parametreler RT'nin elemanları olup aynı zamanda haritalama kuralını oluştururlar. Bunun dışında motif içerisindeki her bir ikili sekonder yapı için üç parametre hesaplanır: Sekonder yapıların orta noktaları arasındaki mesafe d_o, sekonder yapı eksenleri arasındaki en kısa mesafe d_e ve eksenler arasındaki açı φ. Bu parametreler karşılaştırma parametreleridir. Aynı karşılaştırma parametreleri d_o, d_e, φ protein sekonder yapı ikilileri için de hesaplanır. Motif ikilileri protein ikilileri ile karşılaştırılır, eğer eşleşme var ise protein sekonder yapı ikilisinin birinci sekonder yapısının orta noktasından: $\theta_{s_{1o}-i_{RN}}$ açısı ile $\rho_{s_{1o}-i_{RN}}$ kadar mesafe gidilip i'_{RN} noktasına buradan da $\theta_{i_{RN-RN}}$ açı ile $\rho_{i_{RN-RN}}$ kadar mesafe gidilip RN' noktasına ulaşılır ve RN' aday RN olarak belirlenir ve bu noktaya bir oy verilir. Bu metoda ilişkin akış diyagramı Çizelge 1'de gösterilmiştir.

Cantoni ve ark. [15] ikili metot olarak adlandırılan yöntemi kullanarak, 1FNB ve 4GCR proteini içinde mevcut olan Greek Key motifinin çıkarılmasına ilişkin temel düzeyde bir çalışma yapmışlardır. Yapılan çalışmada sadece bir adet motif 2 farklı protein içinde aranmıştır. Bu çalışma protein sayısı ve motif sayısı arttırılarak önerilen çalışmada geliştirilmiştir. Bahsedilen çalışmada oylama süreci önerilen çalışmadaki oylama sürecinden farklıdır. Önceki çalışmada oylama sonucunda motifin lokasyonunu gösteren yerde oy yoğunluğu görülürken farklı noktalarda da oylara rastlanmaktadır. Ancak önerilen çalışmadaki oylama uzayında oylar sadece bir noktada toplanmaktadır, bu nokta da % 0.00 hata oranı ile motifin geometrik merkezinin koordinatlarını göstermektedir. Böylelikle hem lokasyonu belirlemedeki doğruluk oranı hem de arama süresi açısından önerilen çalışmanın literatürde büyük bir öneme sahip olacağı öngörülmektedir.

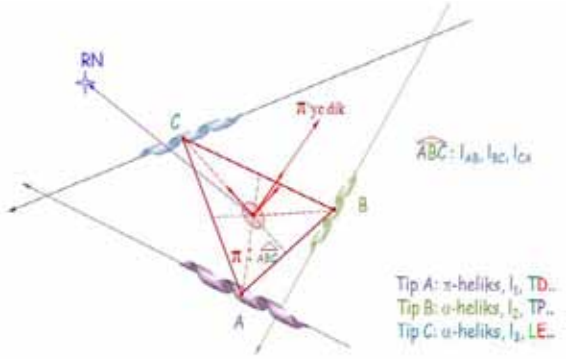


Şekil 2: İkili metot için referans tablosu (RT) parametreleri: d_o, d_e, φ.

2.3. Üçlü Metot

Bu metotta sekonder yapı üçlüleri kullanılır. Üç boyutta, üç tane sekonder yapının orta noktaları birleştirilerek hayali bir üçgen oluşturulur. Böylelikle sekonder yapı üçlüleri yerel ve sabit bir referans sistemi kurarlar. Bu sistemde üçgenin ağırlık merkezi, G, orijin olarak belirlenir; y-ekseni orijine en uzak köşe üzerinde, x-ekseni üçgen düzlemi üzerinde ve z-ekseni de bu iki eksene dik olarak belirlenir (bkz. Şekil 3). Önceki metottakine benzer şekilde RN motif içerisindeki sekonder yapıların ağırlık merkezi olarak belirlenir ve yine benzer şekilde motif içerisindeki sekonder yapı üçlüleri için RN'yi tanımlayan açı ve mesafe değerleri hesaplanır. Bunlar, G ve RN arasındaki açı ve mesafe değerleridir: θ_{G-RN}, ρ_{G-RN}. Bu parametreler haritalama kuralıdır ve RT'nin elemanlarını oluştururlar. Bunun dışında motif içerisindeki her bir sekonder yapı üçlüleri için karşılaştırma parametreleri hesaplanır. Bu karşılaştırma parametreleri sekonder yapı üçlülerinin oluşturduğu üçgenlerin kenar uzunluklarıdır: I_{AB}, I_{BC}, I_{CA}. Protein sekonder yapı üçlüleri içinde aynı karşılaştırma parametreleri hesaplanır ve motif üçgenleri ile protein üçgenleri bu parametreler kullanılarak karşılaştırılır. Eğer eşleşme var ise protein üçgeninin ağırlık merkezinden, G' noktasından, θ_{G-RN} açı ile ρ_{G-RN} kadar mesafe gidilerek RN' noktasına ulaşılır ve RN' noktası aday RN olarak belirlenerek bu noktaya bir oy verilir. Bu metoda ilişkin akış diyagramı Çizelge 1'de gösterilmiştir.

Cantoni ve ark. [16] üçlü metot olarak adlandırdığımız üç sekonder yapının hayali bir üçgen oluşturarak kullanıldığı yöntemi kullanarak yaptıkları çalışma, önerilen çalışmaya temel olmuştur. Önceki çalışmada sekonder yapıların başlangıç ve bitiş koordinatları kullanılarak, protein içindeki tüm sekonder yapıları içeren kübik bir uzay oluşturulmuştur. Oylama işlemi bittikten sonra 3x3x3 boyutunda kübik kafesler oluşturularak tüm alan taranmıştır. Böylelikle oy yoğunluğunun nerede olduğu ve motifin geometrik merkezinin lokasyonu tespit edilmiştir. Bu çalışmada kübik kafeslerle tarama yapıldığından dolayı arama süresi daha uzun zaman almıştır. Ancak önerilen çalışmada oylama süreci daha farklı olduğu için kübik kafeslere gerek olmadan oylama yapılmıştır, daha az sürede ve % 0.00 hata oranı ile motifin lokasyonu belirlenmiştir. [17]'de ise [16]'daki çalışma genişletilerek anlatılmaktadır. Bu çalışmada test amacıyla 20 protein kullanılmış ve proteinler içindeki olası bütün 3, 4 ve 5 sekonder yapıdan oluşan motifler test edilmiştir. Bu çalışmada ve önerilen çalışmada hata oranı % 0.00 iken arama süresi önerilen çalışmada daha düşüktür.



Şekil 3: Üçlü metot için RT parametreleri: I_{AB}, I_{BC}, I_{CA}.

Çizelge 1: İkili ve üçlü metoda ilişkin akış diyagramı. İkili metot için $v = 2, p1 = d_o, p2 = d_e,\ p3 = \varphi$ iken; üçlü metot için $v = 3,\ p1 = I_{AB},\ p2 = I_{BC},\ p3 = I_{CA}$'dır.

Giriş: Protein .nss dosyası; *N*:protein içindeki sekonder yapı sayısı; *m*: motif içindeki sekonder yapı sayısı
Çıkış: Parametre uzayını gösteren A_{RN} akümülatöründeki aday motiflerin lokasyonu
1 Motif RN'sini hesapla: $m(x, y, z)$
2 Motif üçlülerinin sayısını hesapla: $P_m = C(m, v)$
3 Protein üçlülerinin sayısını hesapla : $P_p = C(N, v)$
4 **for** $k = 1$ **to** P_m **do**
5 Karşılaştırma parametrelerini hesapla: $p1, p2, p3$
6 **for** $l = 1$ **to** P_p **do**
7 Karşılaştırma parametrelerini hesapla: $p1', p2', p3'$
8 **for** $k = 1$ **to** P_m **do**
9 **if** $p1, p2, p3$ **match** $p1', p2', p3'$ **then** $A_{RNl} = A_{RNl} + 1$
10 Hough uzayındaki en fazla oy alan noktayı belirle: $a(x, y, z)$

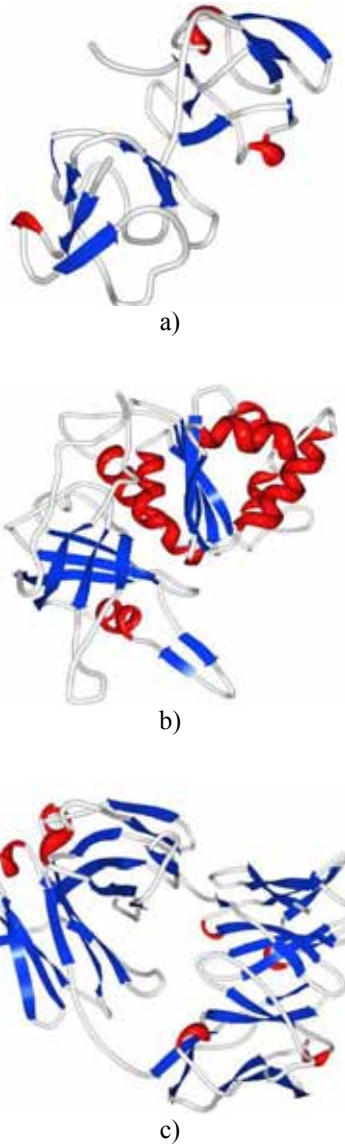
3. Deneysel Karşılaştırmalar

Proteinin sekonder yapısını tanımlamak için kullanılan birçok metot vardır. Bunlardan en sık kullanılanı DSSP'dir (Dictionary of Protein Secondary Structures) [18]. DSSP 8 tip sekonder yapı tanımlar fakat ikincil tahmin metotları bunu 3 baskın tipe indirger: Heliks, tabaka ve dönüşler (coil). Bu çalışmada testler sekonder yapıların DSSP'deki tanımları kullanılarak yapılmıştır. Çalışmalar iki kısımdan oluşmaktadır. Birinci kısımda dört sekonder yapıdan oluşan ve protein sekonder yapıları içerisinden rastgele olarak belirlenen bir motif, ilgili protein içerisinde ikili ve üçlü yöntemler kullanılarak aranmaktadır. Testler PDB'den seçilen üç ayrı protein üzerinde gerçekleştirilmiştir: 4GCR, 1FNB, 7FAB (bkz. Şekil 4). Bu proteinlere ilişkin bazı parametreler Çizelge 2'de gösterilmiştir.

Çizelge 2: Birinci kısım testlerde kullanılan proteinlere ilişkin bazı özellikler

PDB kodu	Molekül Tanımı	#Sekonder yapı
4GCR	GAMMA-B CRYSTALLIN	18
1FNB	FERREDOXIN-NADP-REDUCTASE	22
7FAB	IGG1-LAMBDA NEW FAB	46

Bu gruptaki testlerde öncelikle bilinen bir protein içinden rasgele bir motif seçilmiştir. Burada motif, 4GCR proteininde 6., 9., 13. ve 17. sekonder yapılardan; 1FNB proteininde 2., 8., 15. ve 20. sekonder yapılardan; 7FAB proteininde 9., 17., 32. ve 40. sekonder yapılardan oluşmaktadır. Öncelikle ikili metot kullanılarak testler yapılmıştır İlk olarak aranacak motifin ağırlık merkezi RN olarak belirlenir ve bu RN lokasyonu motif içerisindeki her bir ikili için tanımlanır. Motif içerisindeki ikililer, protein ikilileri ile d_o, d_e, φ parametreleri dikkate alınarak bağıl hata ε=%1 olacak şekilde karşılaştırılır. Her bir eşleşme için haritalama kuralı ile belirlenen noktaya bir oy verilir. Burada motif içerisindeki ikili sayısı $C(4,2) = 6$ olarak hesaplanır. Bu sayı aynı zamanda oylama uzayında aday RN için beklenen oy sayısını göstermektedir. Bu yönteme ilişkin test sonuçları Çizelge 3'te gösterilmektedir.



a) b) c)

Şekil 4: PDB'den alınan proteinlerin sekonder yapı düzeyindeki üç boyutlu görüntüleri. Kırmızı çizgiler α-heliksi, mavi çizgiler β-tabakaları göstermektedir. a) 4GCR proteinine ilişkin sekonder yapılar, b) 1FNB proteinine ilişkin sekonder yapılar, c) 7FAB proteinine ilişkin sekonder yapılar.

Çizelge 3: İkili metot kullanılarak dört sekonder yapıdan oluşan motifin çıkarılmasına ilişkin sonuçlar

PDB kodu	4GCR	1FNB	7FAB
Motif RN	x:14.30 y:13.46 z:22.43	x:23.55 y:8.10 z:15.57	x:-30.17 y:14.39 z:13.23
Aday RN	x:14.30 y:13.46 z:22.43	x:23.55 y:8.10 z:15.57	x:-30.17 y:14.39 z:13.23
#Maksimum oy	6	6	6
Arama süresi (sn)	0.004	0.006	0.009

İkinci olarak üçlü yöntem test edilmiştir. Önceki yöntemde olduğu gibi motif ağırlık merkezi RN olarak belirlenir ve

motif içerisindeki her bir üçlü için RN lokasyonu tanımlanır. Motif üçlüleri ve protein üçlülerinden üçgenler oluşturulur ve bu üçlüler üçgen kenar uzunlukları kullanılarak bağıl hata ε=%1 olacak şekilde karşılaştırılır. Her bir eşleşme için haritalama kuralı ile belirlenen noktaya bir oy verilir. Burada motif içerisindeki üçlü sayısı $C(4,3) = 4$ olarak hesaplanır. Bu sayı aynı zamanda oylama uzayında aday RN için beklenen oy sayısını göstermektedir. Bu yönteme ilişkin test sonuçları Çizelge 4'te gösterilmektedir. Bu teste göre maksimum sayıda oy alan noktalar beklenen oy sayısı kadar oy almış ve motif RN %100 başarım oranı ile hatasız bir şekilde belirlenmiştir.

Çizelge 4: Üçlü metot kullanılarak dört sekonder yapıdan oluşan motifin çıkarılmasına ilişkin sonuçlar

	4GCR	1FNB	7FAB
Motif RN	x:14.30 y:13.46 z:22.43	x:23.55 y:8.10 z:15.57	x:-30.17 y:14.39 z:13.23
Aday RN	x:14.30 y:13.46 z:22.43	x:23.55 y:8.10 z:15.57	x:-30.17 y:14.39 z:13.23
#Maksimum oy	4	4	4
Arama süresi (sn)	0.007	0.007	0.011

İkinci kısımda yapılan testlerde birinci kısımda yapılan testlerden farklı olarak beş sekonder yapıdan oluşan motifler kullanılmıştır. Bu motifler Hough dönüşümü tabanlı ikili ve üçlü metotlar kullanılarak test edilmiştir. Bu motifler PDB'den alınan 2Z9B, 3C94 ve 3DHP proteinleri içinden rastgele seçilmiş ve böylelikle motif oluşturulmuştur (bkz. Şekil 5). Bu gruptaki testlerde motif, 2Z9B proteininde 3., 5., 7., 12. ve 15. sekonder yapılardan; 3C94 proteininde 10., 19., 21., 27. ve 32. sekonder yapılardan; 3DHP proteininde 6., 11., 22., 30. ve 39. sekonder yapılardan oluşmaktadır. Bu proteinlere ilişkin sekonder yapı sayıları ve molekül adları Çizelge 5'te gösterilmektedir.

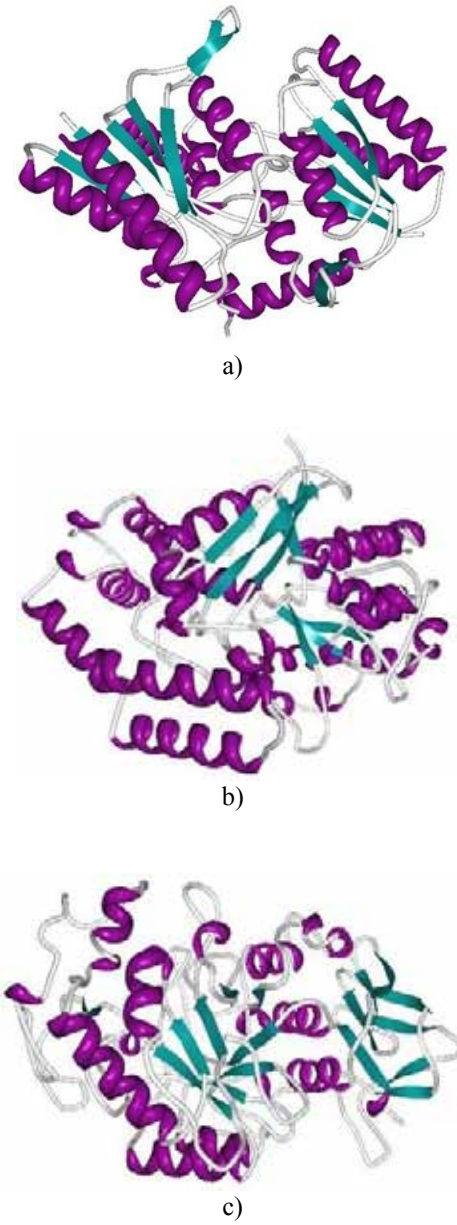
Çizelge 5: İkinci kısım testlerde kullanılan proteinlere ilişkin bazı özellikler

PDB kodu	Molekül Tanımı	#Sekonder yapı
2Z9B	FMN-dependent NADH-azoreductase	16
3C94	Exodeoxyribonuclease I	37
3DHP	Alpha-amylase 1	44

Burada öncelikle ikili metot uygulanmıştır. Motif ikilileri protein ikilileri ile karşılaştırılmıştır. Beklendiği gibi aday RN C(5,2) = 10 tane oy almıştır. Bu teste ilişkin sonuçlar Çizelge 6'da gösterilmektedir.

Çizelge 6: İkili metot kullanılarak beş sekonder yapıdan oluşan motifin çıkarılmasına ilişkin sonuçlar

PDB kodu	2Z9B	3C94	3DHP
Motif RN	x:7.38 y:26.97 z:6.14	x:14.03 y:28.67 z:26.88	x:3.62 y:55.52 z:19.69
Aday RN	x:7.38 y:26.97 z:6.14	x:14.03 y:28.67 z:26.88	x:3.62 y:55.52 z:19.69
#Maksimum oy	10	10	10
Arama süresi (sn)	0.006	0.011	0.008



Şekil 5: PDB'den alınan proteinlerin sekonder yapı düzeyindeki üç boyutlu görüntüleri. Mor çizgiler α-heliksi, turkuaz renkli çizgiler β-tabakaları göstermektedir. a) 2Z9B proteinine ilişkin sekonder yapılar, b) 3C94 proteinine ilişkin sekonder yapılar, c) 3DHP proteinine ilişkin sekonder yapılar

Daha sonra üç sekonder yapının hayali bir üçgen oluşturarak kullanıldığı üçlü metot test edilmiştir. Motif üçlüleri ile protein üçlüleri karşılaştırılmıştır. Her bir eşleşme sonucu belirlenen noktaya bri oy verilmiştir. Beklendiği gibi burada da aday RN C(5,3) = 10 tane oy almıştır. Böylelikle aranan motifin geometrik merkezi yani motif RN % 100 başarım ile belirlenmiştir. Bu teste ilişkin sonuçlar Çizelge 7'de gösterilmiştir.

Testler yapılırken donanım platformu olarak Intel Core 2 Duo 6600, 2.4 GHz, 2 GB RAM özelliklerine sahip bir masaüstü bilgisayar, yazılım platformu olarak da C ve Matlab programlama dilleri kullanılmıştır.

Çizelge 7: Üçlü metot kullanılarak beş sekonder yapıdan oluşan motifin çıkarılmasına ilişkin sonuçlar

PDB kodu	2Z9B	3C94	3DHP
Motif RN	x:7.38 y:26.97 z:6.14	x:14.03 y:28.67 z:26.88	x:3.62 y:55.52 z:19.69
Aday RN	x:7.38 y:26.97 z:6.14	x:14.03 y:28.67 z:26.88	x:3.62 y:55.52 z:19.69
#Maksimum oy	10	10	10
Arama süresi (sn)	0.007	0.011	0.012

4. Sonuçlar

Canlılardaki en temel birim olan proteinler işlevleri açısından büyük öneme sahiptirler. Proteinlerin işlevleri onların yapıları tarafından belirlenmektedir. Bu nedenle protein yapılarının karşılaştırılması ve motif çıkarılması gibi protein yapılarına ilişkin çalışmalar, yapısal biyolojide gittikçe önem kazanmaktadır. Bu çalışmada Hough dönüşümü tabanlı ikili metot ve üçlü metot olarak adlandırdığımız iki yöntem kullanılmıştır. Bu yöntemlerde sırasıyla sekonder yapı ikilileri ve sekonder yapı üçlüleri kullanılarak protein içerisindeki daha küçük boyutlu yapısal bloklar karşılaştırılmıştır. Protein içinden rastgele seçilen dört ve beş sekonder yapıdan oluşan motifler yine ilgili protein içerisinde aranmıştır. Sonuç olarak her iki yöntemde de motifin protein içindeki varlığı saptanmış ve RN'nin lokasyonu % 100 başarım ile hatasız olarak belirlenmiştir. Süreler açısından karşılaştırıldığında bu iki metot arasında fazla fark gözlenmemektedir. Ancak proteindeki sekonder yapı sayısı çok fazla olursa sekonder yapı ikililerinin sayısı üçlülerinin sayısından daha az olacağından ikili metot daha avantajlı olacaktır. Literatürdeki diğer yöntemlerle karşılaştırıldığında ise önerilen yöntem daha kısa sürede arama yapması açısından avantajlıdır. Sonuç olarak GHD protein motif eşleme açısından iyi bir metottur, uygulanması kolay ve hızlıdır. Sonraki çalışmalarda bu yöntemin geliştirilerek protein domainleri ve proteinin tamamının karşılaştırılmasında kullanılması planlanmaktadır.

5. Kaynaklar

[1] Can, T. ve Wang, Y.F., "CTSS:A Robust and Efficient Method for Protein Structure Alignment Based on Local Geometrical and Biological Features", *IEEE Computer Society Conference on Bioinformatics*, pp. 169-179, 2003.

[2] Camoğlu, O., Kahveci, T. ve Singh, A. "PSI: Indexing Protein Structures for Fast Similarity Search", Bioinformatics, vol.19, suppl.1, pp.81-83, 2003.

[3] Chionh, C.H., Huang, Z., Tan, K.L. ve Yao, Z., "Augmenting SSEs with Structural Properties for Rapid Protein Structure Comparison", *IEEE Symposium on Bioinformatics and Bioengineering*, pp.341-348, 2003.

[4] Shuoyong, S., Zhong, Y., Majumdar, I., Krishna S.S. ve Grishin, N.V., "Searching for Three-Dimensional Secondary Structural Patterns in Proteins with ProSMoS", *Bioinformatics*, vol.23, no.11, pp. 1331-1338, 2007.

[5] Chi, P.H., Scott, G. ve Shyu, C.R., "A Fast Protein Structure Retrieval System Using Image Based Distance Matrices and Multidimensional Index", *International Journal of Software Engineering and Knowledge Engineering, Special Issue on Software and Knowledge Engineering Support in Bioinformatics*, pp.522-532, 2004.

[6] Albrecht, B., Grant, G.H., Sisu, C. ve Richards, W.G., "Classification of Proteins Based on Similarity of Two-Dimensional Protein Maps", *Biophysical Chemistry*, pp.11-22, 2008.

[7] Zotenko, E., Dogan, R.I., Wilbur, W.J., O'Leary, D.P. ve Przytycka, T.M., "Structural Footprinting in Protein Structure Comparison: The Impact of Structural Fragments", *BMC Structural Biology*, vol.7, 7:53, 2007.

[8] Cantoni, V., Ferone, A., Ozbudak, O., Petrosino, A., "Protein Structural Blocks Representation and Search Through Unsupervised NN", *ICANN 2012*, September 11-14, Lausanne, Switzerland, vol. 7553, p. 515-522.

[9] A. Ferone, O. Ozbudak, "Comparison of GHT-based Approaches to Structural Motif Retrieval", A. Petrosino, L. Maddalena, P. Pala (Eds.): ICIAP 2013 Workshops, LNCS 8158, pp. 356-362, 2013.

[10] Hough, P.V.C., "Methods and Means for Recognizing Complex Patterns", *US Patent 3069654*, 1962.

[11] Duda, R.O. ve Hart, P.E., "Use of the Hough Transformation to Detect Lines and Curves in Pictures", *Comm. ACM*, vol.15, no.1, pp. 11-15, 1972.

[12] Wechsler, H. ve Sklansky, J., "Automatic Detection of Ribs in Chest Radiographs", *Pattern Recognition*, vol.9, pp.21-30, 1977.

[13] Ballard, D.H., "Generalizing the Hough Transform to Detect Arbitrary Shapes", *Pattern Recognition*, vol.13, no.2, pp. 111-122, 1981.

[14] http://www.rcsb.org/pdb/.

[15] Cantoni, V., Ferone, A. ve Petrosino, A., "Protein Motif Retrieval through Secondary Structure Spatial Co- occurrences", *New Tools and Methods for Pattern Recognition in Complex Biological Systems*, vol.35, no.5, suppl.1, 2012.

[16] V. Cantoni, A. Ferone, O. Ozbudak, A. Petrosino, "Search of Protein Structural Blocks Through Secondary Structure Triplets", 3rd International Conference Image Processing Theory, Tools and Applications, IPTA'12, October 15-18, Istanbul, Turkey, ISBN:978-1-4673-2584-4, pp.222-226, 2012.

[17] V. Cantoni, A. Ferone, O. Ozbudak, A. Petrosino, "Protein Motifs Retrieval By SS Terns Occurrences", Journal of Pattern Recognition Letters, Elsevier, vol.34, p.559-563, ISSN:0167-8655, 2013.

[18] Kabsach, W. and Sander, C., "Dictionary of Protein Secondary Structure: Pattern Recognition of Hydrogen Bonded and Geometrical Features", *Biopolymers*, 22(12), pp.2577-2637, 1983.

Özlem Özbudak

1981 yılında Sivas'ta doğan Özlem Özbudak ilk ve orta öğrenimini İstanbul'da tamamlamıştır. 2000 yılında Ankara S. Demirel Sağlık Meslek Lisesi'nden hemşire unvanı alarak mezun olmuştur. 2001 yılında Yıldız Teknik Üniversitesi Jeodezi ve Fotogrametri Mühendisliği Bölümü'nü kazanarak üniversiteye giriş yapmış, 2002 yılında iç transfer ile aynı üniversitenin Elektronik ve Haberleşme Mühendisliği Bölümü'ne 3.88/4.00 ortalama ile geçmiş ve 2005 yılında "ADuC814 Mikrodenetleyicisi ile Dijital Termometre Tasarımı" isimli tez ile mezun olmuştur. Yüksek lisansta 2009 yılında İstanbul Teknik Üniversitesi Elektronik Mühendisliği programını "Yüz Resimlerinden Cinsiyet Tanıma" isimli tez ile bitirmiş olup aynı yıl yine aynı üniversite ve programda doktoraya başlamıştır. 2011-2012 yılları arasında 12 ay süreyle İtalya'da Pavia Üniversitesi'nde Prof. Dr. Virginio Cantoni danışmanlığında araştırmalar yapan Özlem Özbudak; şuan Prof. Dr. Zümray Dokur'un danışmanlığında "Mikro ve Makro Yapılar Kullanılarak Proteinlerin Eşleştirilmesi ve Sınıflandırılması" başlıklı tez çalışmasıyla İTÜ'de doktoraya devam etmektedir. 2006 yılında İTÜ Elektrik-Elektronik Fakültesi'nde araştırma görevlisi olarak çalışmaya başlamış ve Aralık 2013'de bu görevinden ayrılmıştır. Şuan evli olan ve Sivas'ta yaşayan Özlem Özbudak Polat, Aralık 2013'de Cumhuriyet Üniversitesi Teknoloji Fakültesi Mekatronik Mühendisliği Bölümü'nde araştırma görevlisi olarak çalışmaya başlamıştır ve halen devam etmektedir.

Zümray Dokur

Zümray Dokur, orta öğrenimini 1988 yılında Elazığ Anadolu Lisesi'nde birincilikle tamamladı. Lisans öğrenimini 1992 yılında İTÜ Elektronik ve Haberleşme Mühendisliği Bölümü'nde, yüksek lisans öğrenimini 1995 yılında İTÜ Elektronik ve Haberleşme Mühendisliği Anabilim Dalı, Biyomedikal Mühendisliği Programı'nda tamamladı. 1992 yılında İTÜ Elektronik Anabilim Dalında araştırma görevlisi olarak göreve başladı. 2000 yılında İTÜ Fen Bilimleri Enstitüsünden "Yapay Sinir Ağları ve Genetik Algoritmalar Kullanılarak EKG Vurularının Sınıflandırılması" isimli teziyle "Doktor" ünvanını aldı. 2001 yılında İTÜ Elektronik Anabilim Dalına "Y. Doçent" olarak atandı. 2003 yılında "Elektrik-Elektronik Mühendisliği" bilim alanında "Doçent" ünvan ve yetkisi aldı. 2004 yılında İTÜ Elektronik Anabilim Dalına "Doçent" olarak atandı. 2009 yılından itibaren İTÜ Elektronik ve Haberleşme Mühendisliği Bölümü'nde "Profesör" olarak görev yapmaktadır. 52 tanesi uluslararası olmak üzere toplam 76 adet bilimsel yayını bulunmaktadır. 2008'den beri Neural Processing Letters isimli dergide yardımcı editör olarak görev yapmaktadır.

Örüntü tanıma, biyolojik işaretlerin analizi ve tanınması, medikal görüntülerin analizi, bölütlenmesi, yapay sinir ağ modellerinin geliştirilmesi, genetik eğitim, bulanık mantık, bulanık sınıflayıcılar, medikal enstrumantasyon, biyoinformatik ve beyin-bilgisayar arayüzü tasarımları konularında çalışmalarını sürdürmektedir.

Virginio Cantoni

Virginio Cantoni lisans derecesini 1972 yılında İtalya Pavia Üniversitesi'nden almıştır. 1975-1983 yılları arasında İtalyan Ulusal Araştırma komisyonunda araştırmacı olarak çalışmış olup, 1985-1990 yılları arasında International Association for Pattern Recognition (IAPR) İtalyan grubunun başkanlığını yapmıştır. 1987 yılı bahar döneminde New Jersey Rutgers Üniversitesinde ziyaretçi profesör olarak çalışmış, 1989-1995 yılları arasında Pavia Üniversitesi Bilgisayar ve Sistem Mühendisliği bölüm başkanlığı görevini yürütmüştür. 1994-2001 yılları arasında Paris XI Üniversitesi'ne sekiz defa ziyaretçi profesör olarak davet edilmiştir. 1995-2001 yılları arasında Pavia Üniversitesi Bilgisayar Merkezi'nin başkanlığı görevini yürütmüş; 1995-2002 yılları arasında Elektronik ve Bilgisayar Mühendisliği doktora programlarının koordinatörlüğünü yapmıştır. Pavia Üniversitesi'nde Avrupa Media Bilimi ve Teknolojilerinde İleri Çalışmalar Okulu'nun kurucusu olup 1997 yılından bu yana aynı okulun direktörlüğünü yapmaktadır. 270'den fazla makale, bildiri ve kitabın yazarı ya da yazarlarından biri olup, 30 tane kitabın editörü ya da editörlerinden biri olmuştur. Yazarı olduğu 3 kitabı vardır. Image Processing ve Computer Vision üzerine çok sayıda konferans organize etmiştir. IAPR'de (International Association for Pattern Recognition) ve IEEE'de (Institute of Electrical and Electronic Engineers) fellow olup, IEEE Bilgisayar Topluluğu'ndan (IEEE Computer Society) Takdir Belgesi (Certificate of Appreciation) almıştır. Evli ve 2 çocuklu olan Virginio Cantoni şuan İtalya'nın Pavia şehrinde yaşıyor olup Pavia Üniversitesi Bilgisayar Mühendisliği Bölümü'nde tam zamanlı profesör ve "Computer Vision and Multimedi Lab." da yürütücü olarak çalışmaktadır.