Sınıflandırma için Protein Dizilerinin Özniteliklerinin Çıkarılmasında Model Tabanlı Yeni Bir Yöntem

A Novel Model-based Method for Feature Extraction from Protein Sequences for Classification

Ömer Sinan Saraç[1], *Volkan Atalay*[1], *Rengül Çetin Atalay*[2]

[1]Bilgisayar Mühendisliği Bölümü, Orta Doğu Teknik Üniversitesi, Ankara
[2]Moleküler Biyoloji ve Genetik Bölümü, Bilkent Üniversitesi, Ankara
{sarac, volkan}@ceng.metu.edu.tr, rengul@bilkent.edu.tr

Özetçe

Proteinlerin işlevsel ve yapısal sınıflara ayrılmasında en önemli nokta amino asit dizilerinin gösterimidir. Gösterim, proteinin birincil dizisinde saklı biyolojik olarak anlamlı bilgiyi içermeli ve temsil edebilmelidir. Korunmuş veya benzer altdiziler işlevsel ve yapısal benzerlik için kuvvetli belirtilerdir. Bu çalışmada, protein dizilerindeki altdizilerinin modellerini hesaba katan bir öznitelik eşlemesi sunulmaktadır. Saklı Markov karışım modeli ile birlikte bir Tahmin-Azami algoritması verilen bir küme proteinin altdizilerini demetlemek ve modellerini öğrenmek için kullanılmıştır.

Abstract

Representation of amino-acid sequences constitutes the key point in classification of proteins into functional or structural classes. The representation should contain the biologically meaningful information hidden in the primary sequence of the protein. Conserved or similar subsequences are strong indicators of functional and structural similarity. In this study we present a feature mapping that takes into account the models of the subsequences of protein sequences. An expectation-maximization algorithm along with an HMM mixture model is used to cluster and learn the models of subsequences of a given set of proteins.

1. Giriş

İşlemsel biyolojide en önemli sorunlardan birisi proteinlerin, birincil dizilerine dayanarak işlevsel ve yapısal sınıflarına ayrıştırılmasıdır. Doğada 20 değişik amino asit vardır ve dolayısıyla protein dizileri 20 harflik bir alfabeden üretilmiş değişik uzunluklardaki kelime dizileri (string) olarak görülebilirler. Proteinin işlev ve yapısı hakkında önemli bilgi birincil dizide saklıdır. Protein sınıflamadaki en büyük sorun bu fark gözetici özniteliklerin çıkartılmasıdır. Farklı protein dizileri arasında korunmuş altdiziler işlevsel ve yapısal benzerlik için kuvvetli belirtilerdir, ancak bunların farkedilmesi çok güçtür; çünkü korunmuş altdiziler farklı uzunluklarda olabilir ve dizinin farklı yerlerinde bulunabilirler. Yazında bu sorunu çözmek için gösterilen gayretler dört sınıfta toplanabilir: hizalama (alignment) tabanlı yöntemler, model tabanlı yöntemler, örge (motif) tabanlı yöntemler ve ayırtedici (discriminative) yöntemler.

Hizalama tabanlı yöntemler [1, 2, 3], özellikle de çoklu dizi hizalama (ÇDH) [3], çokça kullanılmaktadır ve dizi benzerliği yüksek olduğunda türdeşlik (homology) bulmakta iyi çalışmaktadır. Bununla birlikte ÇDH NP-zor bir sorundur ve uzak türdeşlik sözkonusu olduğunda yani dizi benzerliği %40'ın altında olduğunda iyi sonuç alınamamaktadır. Bu durumda, ÇDH yöntemleri en uygun hizalamanın daha düşüğünü bulabilmektedirler. Model tabanlı yöntemlerde, örneğin bir protein ailesinin profilinin olasılıklsal modelini tanımlamak için bir saklı Markov model (SMM) kullanılmıştır [4,5]. Örge tabanlı yöntemler işlevsel ve yapısal özelliklerle ilintili olduğu bilinen yüksek derecede korunmuş kısa altdiziler olarak tanımlanan örgelerden yararlanırlar [6]. Örgeler sınıfı bilinen bir grup protein dizisinden çıkartıldıktan sonra sınıfları bilinmeyen dizilerde bu çıkartılmış olan örgelerin olup olmadığı sınanır. Bununla birlikte birçok protein sınıfı için o sınıfı tanımlayan örgeler henüz belirlenmemiştir ve hatta bazı sınıflar için hiç örge de yoktur. Ayrıca, örge çıkartmak da kolay bir iş değildir. Örgeler mütasyon veya eksik kalıntı (residue) içerdiğinden genellikle tam doğru değildir [7]. Bunun da ötesinde, tek bir örge bir proteinin işlevini belirlemek için yeterli de olmayabilir [6].

Bu ana kadar bahsedilen yöntemlerde sadece bir sınıftan olumlu örnek diziler kullanılmıştır ve yeni bir dizinin sınıflandırılması, olumlu örnek dizilerden yaratılan modele benzerliğine göre yapılır. Önceki üç yöntemin tersine, ayrıtedici sınıflandırıcılar hem olumlu hem de olumsuz olarak etiketlenmiş dizileri girdi olarak kullanır ve sonuç olarak bir karar sınırı belirlerler. Destek vektör makina (DVM) sınıflandırıcıları uygun öznitelik gösterimi veya çekirdekle birleştirildiğinde uzak türdeşlik belirlemede günümüzde en iyi başarıma ulaşmaktadırlar [8, 9]. Fisher-DVM yöntemi [8], çekirdek tanımlamak için profil SMM kullanmaktadır. İlk olarak hedef aile için profil bulan SMM inşa edilir. Ardından profil SMM'nin olasılıksal modelinin parametreleri vektör olarak düzenlenir. Bir dizi girdi olarak verildiğinde, bu vektörü doğrudan kullanmak yerine Fisher skoru denilen model parametre değerlerinden farklılığı hesaplanmaktadır ve bu Gausyen çekirdeğine öznitelik vektörü olarak sürülmektedir. Bu gösterim, verilen ailenin olasılıksal profil modeline hizalama bilgisini içerir. Bu yöntemle önemli bir iyileştirme elde edilmiştir ve SCOP veritabanında tanımlı ailelerin uzak türdeşlerinin belirlerlenmesinde en iyi başarıma ulaşılmıştır [10].

Leslie vd. DVM için uyumsuz eşlemeli dizi eşleştirme çekirdeği önermektedir [9]. Uyumsuz eşleştirme çekirdeği, sabit uzunluktaki tüm olası amino asit altdizilerini gösteren vektörleri kapsayan öznitelik uzayında tanımlanır. *K*-mer adı verilen herbir k uzunluğundaki altdizi buradan en fazla *m* mevki farkeden-yani *m* yanlış eşleştirme kordinatına katkıda bulunur. Bu seyrek öznitelik vektörlerini bulmak yerine Leslie vd. Verilen iki dizi arasındaki ortak (*k,m*) uyumsuz eşleştirme altdizilerini sayarak bu çekirderği hesaplar. Küçük k ve m

1-4244-0239-5/06/$20.00 ©2006 IEEE

değerleri-tipik olarak *k* için 5 ve *m* için 1 için işlemsel olarak verimli olan bu yöntem, Fisher DVM'in başarısına yakın bir başarı elde etmektedir.

Bütün yöntemler öyle yada böyle protein sınıfları arasında korunmuş, ayırtedici altdizilerin çıkartılmasına odaklanmıştır. Bu bildiride anlatılan yöntem altdizileri değil de onları üretmiş olabilecek modelleri bulmaya çalışır. Ana düşünce, ortak bir özelliği paylaştığı bilinen bir protein dizisi kümesi verildiğinde, bu ortak özellikle ilgili özel altdiziler varsa ve eğer verilen dizileri uygun bir şekilde bölebilirsek, bu özel altdizilere çokça rastlanması gerektiğidir. Tahmin-azami (expectation-mazimization-TA) algoritmasını kullanarak [11], tamamen gözetimsiz bir şekilde böylesi altdizilerin modelleri belirlenmektedir. Modellerin gösterimi için 20 durumlu, birbiriyle tamamen bağlı, ergodik bir küme SMM kullanılmaktadır. Gözlenme olasılıkları amino asit benzerlik matrisi kullanarak sabitlenmiştir ve bu nedenle bir altdizinin modeli amino asit durumları arasında izlediği yoldur. Böylece, bir protein dizisinin öznitelik uzayındaki gösterimi altdizilerinin SMM karışım modeli üzerindeki dağılımı şeklinde tanımlanabilir. Buradaki ana sorun dizilerin uygun bölünmesidir. Bu sorun, alt parçalarının değişik modeller tarafından üretilmiş bir işaretin altparçalarına bölütlenmesi olarak düşünülebilir. Başlangıçta modeller bilinmemektedir ve TA algoritması ile öğrenilmesi beklenmektedir. SMM'lerin ergodik olmasıyla sorun daha da karmaşık hale gelmekte ve bu durumda SMM'ler her uzunluktaki altdizileri üretebilmekte veya kabul edebilmektedir. Bu bildiride bölütleme işlemi için de yöntem önermektedyiz.

Burada sunulan yaklaşımın pek çok yararı vardır. Dizi benzerliği düşük olduğunda sorunlu olduğu ve en iyinin altını verdiği bilinen çoklu dizi hizalaması gerekmemektedir. Bu anlamda, Jakkola vd. çalışması en iyi olmayan hizalamaya meyillidir. Bundan da önemlisi, olası altdiziler için herhangi bir sabit uzunluk kısıntısı getirilmemiştir ve yöntem biyolojik olarak makul kabul edilen yanlış eşleştirme ve mütasyonlara izin vermektedir. Örneğin, örge dizilerinde sıkça rastlanan ancak yanlış eşleştirme çekirdeği için ümitsiz bir vaka olan 20 amino asitlik bir altdizide 5 yanlış eşlemeye izin vermektedir. Bunlara ek olarak da, altdizi modellerimizi kullanarak en ayırtedici altdizileri bulma ve herbir uzay boyutunun (karışım modelindeki herbir SMM'nin) DVM sınıflandırıcısına katkısını inceleyerek örgeleri keşfetmek olasılığı vardır. Bu da biyoloji araştırmacılarının sistemi daha iyi anlamasına yardımcı olabilir.

2. Yöntem

Üretici modellerden oluşan vektör öznitelik haritası olarak kullanılmaktadır. Öznitelik uzayında herbir dizi altdizilerinin bu üretici modeller üzerindeki dağılımı ile temsil edilmektedir. Altdizilerin modellerini temsil edebilecek özel SMM'ler tasarlanmıştır. Herbir SMM birbiriyle tamamen bağlı 20 durumdan oluşmaktadır. Doğada 20 amino asit bulunmaktadır; bu nedenle herbir durum bir amino asiti temsil etmektedir. Bir durumun simge salma olasılığı amino asitlerin birbirlerinin yerine konma (substitution) matrisi ile sabitlenmiştir [12]. Amino asitlerin birbirlerinin yerine konma matrisi, bir amino asitin zaman içinde diğer herbir amino asite değişme hızının olasılığını tahmin eder. Bu da modelimizde herhangi bir yanlış eşlemeye değil de biyolojik olarak makul olanlarına izin verir. Durum geçiş olasılıkları başlangıçta rastgele değerler olarak verilmiştir ve TA algoritması ile öğrenileceklerdir. Bir altdizinin modeli aslında SMM durumları arasında izlediği yoldur. SMM'ler hakkında ayrıntlı bilgi için [13]'e basvurulabilir.

Karışım modelinde SMM'lerin sayısı olan C algoritma çalışmaya başlamadan önce belirlenmelidir. Bu sayı, verilen bir küme protein dizilerinin altdizilerinin oluşturacağı demet sayısı olarak düşünülebilir. Eğer C küçük seçilirse, bazı farklı ayırtedici altdiziler aynı SMM'ye atanmaya zorlanacaklardır. İşleme zamanı C ile doğrudan orantılı olarak artar. TA algoritmasının ana hatları aşağıda verilmektedir.

1. Durma ölçütüne kadar yap:
   a. SMM'lerden elde edilen benzerlik bilgisine göre her diziyi altdizilerine böl.
   b. Herbir altdiziyi onu en yüksek olasılıkla yaratmış olabilecek SMM'e ata.
   c. Atanmış altdizilerle SMM'leri eğit (yakınsamaya kadar değil de dizilerin üstünden sadece bir kez geçerek).

1.c adımı iyi bilinen Baum-Welch algoritması ile yerine getirilmektedir [14]. Nazik adım, 1.a.'da belirtilen dizilerin bölütlenmesidir. Verilen bir dizinin nitelik uzayı gösterimi herbir SMM'e düşen altdizilerin benzerlik değerlerinin toplamı olarak tanımlandı. Sonuç olarak, C karışım modelindeki SMM sayısı olmak üzere C boyutlu bir öznitelik uzayı yaratılmış oldu.

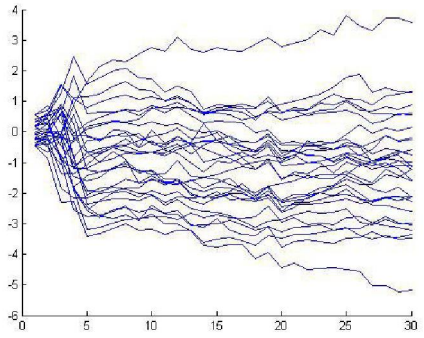
3. Dizilerin Bölütlenmesi

Tanımlanmış olan SMM'lerin davranışlarını çözümlemek için bir seri deney yapılmıştır. En iyi bölütlemenin verilmiş olduğunu varsayarak, SMM karışım modelinin demetleme başarımı incelenmiştir. Rastgele oluşturulan 4000 dizinin yanısıra 30 değişik proteinden 12 değişik metabotropik glutamat GPCR imza örgesinden oluşturulmuş 360 dizi ile bir veri kümesi hazırlanmıştır. Rastgele dizilerin uzunlukları 5 ile 30 arasında değiştirilmiştir. TA algoritması, 1.a adımı atlanarak 1.b adımındaki atamalardaki değişim belli bir eşiğin altına düşene kadar C=20 ile koşturduk. Sonuçlar beklendiği gibi oldu: rastgele dizilerin SMM'ler üzerindeki dağılımı oldukça muntazamdı; oysa aynı örgeden gelen diziler birlikte öbeklenmişlerdi. Bölütlenmenin nasıl olabileceği hakkında fikir edinmek için rastgele üretilmiş dizleri az sayıda örgelere ilave edip karışım modelindeki 20 SMM ile üretilme log-benzemelerini (log-likelihood) inceledik. Şekil 1 örnekleme ilave edilmiş dizilerin rastgele bir yerinden başladığında elde edilen log-benzerliklerini göstermektedir. Öte yandan, Şekil 2 ilave edilmiş dizilerde bir örgenin başlangıcından başlandığında elde edilen log-benzerliklerini göstermektedir. Dikkat edilmelidir ki log-benzeme değerleri uzunluğa göre rastgele model kullanılarak normalleştirilmiştir. Log-benzeme aşağıdaki şekilde tanımlanmıştır.

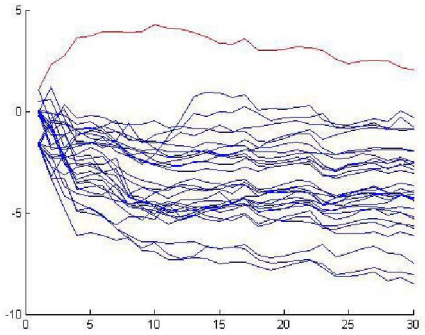
$$\log \frac{P(X \mid H_m)}{P(X \mid H_0)} \qquad (1)$$

Formül 1'de X altdiziyi, H_m HMM modeli ve H_0 rastgele modeli temsil etmektedir. Normalleştirme terimi olarak rastgele model altdizinin o anda incelenen SMM'ye uygunluğunu incelemek için bir başvuru değeri vermektedir. Daha da önemlisi, bu yaklaşım değişik uzunluklardaki

altdizilerin benzerlik değerlerinin karşılaştırılmasına olanak sağlamaktadır. Kullanılan rastgele model herbir amino asitin eğitim kümesindeki sıklıklarıdır.



Şekil 1: Rastele bir noktadan başlandığında oluşan log-benzerliklerin dizi uzunluğuna göre değerleri.



Şekil 2: Belirli bir örgünün başından başlandığında oluşan log-benzerliklerin dizi uzunluğuna göre değerleri

Şekil 1 ve 2'de görülebileceği gibi SMM'ler önemli altdizilerin başlangıçlarına hassas olmalarına rağmen rastgele üretilmiş altdizilerin başlangıçlarına duyarlı hale gelmemişlerdir. SMM'ler ne yazık ki, altdizilerin sonlarına da duyarlı değillerdir. Gözlemlerimizden birisi kazanan SMM'nin diğer SMM'lerden farkının en yükseğe yaklaşık 5 amino asit civarında çıkmasıdır. Bu gözlemlere dayanarak, bölütlemenin olası başlangıç noktalarının yerlerini saptayarak gerçekleştirebileceği söylenebilir. Dizideki tüm konumlar için o konumdan başlayarak 5 uzunluğundaki altdizilerin log-benzerliklerini ürettik ve kazanan SMM'in benzerlik değerinin diğer SMM'lerin benzerlik değerlerinden farkının toplamının tepe yaptığı yerleri işaretledik. Diziler işaretlenmiş yerlerden bölündüler ve altdizi uzunluları 5 ile 30 arasına sınırlandırıldı.

4. Deneyler

Bahsedilen yöntemin başarımını ölçmek için iki farklı sınıfa ait 200 dizi içeren bir yapay veri kümesi oluşturduk. Öncelikle, herbir sınıf için bir tane olmak üzere iki küme örge yararttık. Herbir küme iki farklı tipte örgeye sahipti. Bir sınıfın üyeleri rastgele yerlere konuşlandırılmış ilintili bir veya iki örge ile birlikte rastgele üretilmiş amino asit dizilerini içermektedir. Herbir örnekte, 0'dan 8'e değişen uzunluklarda olası mütasyon vardır ve örge uzunluğu 13'ten 30'a değişmektedir. Sonuçta ortaya çıkan dizilerin uzunlukları 130 ile 220 arasındadır.

5 SMM karşım modeli ve 20 SMM karışım modeli olmak üzere iki karışım modeli eğitilmiştir. Eğitimden sonra her dizi 5 ve 20 boyutlu uzaya eşlenmiştir. Çok hızlıca yapılan k-orta algoritması uygulaması sonucunda aynı sınıfın elemanları 5-SMM ile %70 oranında birlikte öbeklenmiştir. Aynı oran 20-SMM için %90'a ulaşmıştır.

5. Tartışma

Bir proteinin işlevi ve yapısı hakkında önemli bilgi birincil dizisinde saklıdır. Dizi benzerliğinin düşük olduğu uzak türdeş bulma durumunda, bazı korunmuş altdiziler proteinlerin özellikleri hakkında önemli ipuçları verir. Protein sınıflandırmada en önemli sorun saklı öznitelikleri ve korunmuş altdizileri bulmaktır. Bu amaçla, bu çalışmada altdizileri üreten modelleri öğrenmeye dayanan ortak özellik taşıyan bir sınıf proteinin paylaşılan özniteliklerini bulmaya yarayan bir yöntem anlatılmaktadır. Bu çalışmada anlatılan özel SMM yapıları farklı uzunluklardaki altdizileri modelleyebilecek kabiliyettedir. Buna ek olarak, bir altdizide biyolojik olarak olası mütasyonlara izin vermektedir ve bu da çok nazik örgelerin bulunmasına olanak vermektedir.

Kolay bir sınıflandırma problemi üzerinde yapılan deneyler, bu gözetimsiz yöntemin, verilen bir küme proteinde istatistiksel olarak önemli (çok sayıda) örüntüyü bulmaya kabiliyetli olduğunu göstermiştir. Bir sonraki adım olarak, yöntem SCOP değerlendirme veritabanı gibi gerçek biyolojik veri üzerinde sınanmalıdır [10]. Örge bulmakta kullanıp kullanılamayacağını görmek için eğitimden sonra SMM'ler incelenmelidir. Kullanılan bölütleme yöntemi buluşsaldır. Bu bağlamda, en iyi bölütlemeyi bulabilecek bir matematik çerçevebulmak iyi bir ilerleme olacaktır. TA algoritması en düşük yerel değere takılmaya eğilimli olduğundan dolayı her TA tabanlı algoritmasında olduğu gibi, başlangıç değerleri ve durumu başarımı çok etkilemektedir. Bunu aşmanın bir yöntemi BLAST yada PSI-BLAST [4, 15] benzeri bir hızlı hizalama yöntemi kullanarak karışım modelindeki bazı SMMleri başlatmak olabilir. Bu da modelin bazı kolayca hizalanabilen altdizilere duyarlı olmasını güvence altına alacaktır. Geri kalan SMM'ler de, hizalama yöntemi ile hizalanmamış altdizilerin oluşturduğu uzayda arama yapmaya yarayacak rastgele başlatılmış bazı SMM'ler olacaktır.

6. Kaynakça

[1] Altschul, S. F., Gish, W., Miller, W., Myers, E. W., and Lipman, D. J, "A basic local alignment search tool", *Journal of Molecular Biology,* 215:403-410, 1990.

[2] Smith, T. and Waterman, M., "Identification of common molecular subsequences", *Journal of Molecular Biology,* 147:195-197, 1981.

[3] Park, J., Karplus, K., Barrett, C., Hughey, R., Haussler, D., Hubbard, T., and Chothia, C., "Sequence comparisons

using multiple sequences detect twice as many remote homologues as pairwise methods", *Journal of Molecular Biology*, 1998, 284(4):1201-1210.

[4] Baldi, P., Chauvin, Y., Hunkapiller, T., and McClure, M. A., "Hidden Markov models of biological primary sequence information", *Proc. Natl. Acad. Sci.* USA, 91:1059-1063, 1994.

[5] Krogh, A., Brown, M., Mian, I. S., Sjolander, K., and Haussler, D., "Hidden Markov models in computational biology: Applications to protein modeling", *Journal of Molecular Biology*, 235:1501:1531, 1994.

[6] Ben-Hur, A. and Brutlag, D., "Remote homology detection: a motif based approach", *Bioinformatics*, 19:26-33, 2003.

[7] Yang, J., Deogun, J. S, Sun, Z., "A New Scheme for Protein Sequence Motif Extraction", *Proc. Of the 38th Hawaii Intl. Conf. on System Sciences,* 9:280.1, 2005.

[8] Jaakkola, T., Diekhans, M., and Haussler, D., "A discriminative framework for detecting remote protein homologies", *Journal of Computational Biology,* 7(1-2):95-114, Feb 2000.

[9] Leslie, C. S., Eskin, E., Cohen, A., Weston, J., and Noble, W. S., "Mismatch string kernels for discriminative protein classification", *Bioinformatics*, 20(4):467-476, 2005.

[10] Murzin, A. G., Brenner, S. E., Hubbard, T., and Chotia, C., "SCOP: A structural classification of proteins database for the investigation of sequences and structures", *Journal of Molecular Biology*, 247:536-540, 1995.

[11] Dempster, A. P., Laird, N. M., and Bubin, D. B., "Maximum likelihood from imcomplete data via EM algorithm", *Journall of the Royal Stat. Soc.*, pp. 1-38, 1977.

[12] Henikoff S., and Henikoff, J. G., "Amino acid substitution matrices from protein blocks", *Proc. Natl. Acad. Sci. USA,* pp.10915-10919, 1992.

[13] Rabiner, L. R., "A tutorial on hidden Markov models and selected applications in speech recognition", *Proc. IEEE*, 77:257-258, 1989.

[14] Baum, L. E., Peterie, T., Souled, G., and Weiss, N., "A maximization technique occurring in the statistical analysis of probabilistic functions of Markov chains", *Ann. Math. Statist.*, 41:164-171, 1970.

[15] Altschul, S. F., Madden, T. L., Schaffer, 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 Res.,* 25(17):3389-3402, 1997.