Birincil Dizi Veri Temelli Protein Hücre İçi Yer Belirleme Tahmini

Prediction of Protein Subcellular Localization based on Primary Sequence Data

Mert Özarar [1], Volkan Atalay [1], Rengül Çetin Atalay [2]

[1] ODTÜ Bilgisayar Mühendisliği Bölümü {ozarar@ceng.metu.edu.tr}

[1] ODTÜ Bilgisayar Mühendisliği Bölümü {volkan@ceng.metu.edu.tr}

[2] Bilkent Üniversitesi Moleküler Biyoloji ve Genetik Bölümü {rengul@bilkent.edu.tr}

Özetçe

Proteinlerin işlevlerinin belirlenmesinde hücre içi yer belirleme çok önemlidir. Bu çalışmada, ökaryotik canlılarda, amino asit sırası kullanılarak amino asit birincil dizi içeriği temelli, protein hücre içi yer belirlenmesi için, P2SL adında, yeni bir sistem tasarlanmıştır. Tahmin yaklaşımı, öz düzenlemeli haritalara dayanarak verilen bir sınıfta her protein için, en yaygın motifleri bulmak ve bunları, öznitelik olarak kullanarak çok katmanlı perseptronların yardımıyla sınıflandırmaktır. Bu yaklaşım dizi uzunluğundan bağımsız bir sınıflandırmaya izin vermektedir. Bunlara ek olarak, kabul edilebilir nokta mutasyon (PAM) değiştirme matrisi temelli, biyolojik işlevi muhafaza eden, yeni bir kodlama planı kullanımı tarif edilmektedir. Dört sınıflı bir problemde, sistemin istatistiksel test sonuçları sunulmaktadır. P2SL, benzer çalışmalardan biraz daha yüksek tahmin doğruluğuna ulaşmıştır.

Abstract

Subcellular localization is crucial for determining the functions of proteins. A system called prediction of protein subcellular localization (P2SL) that predicts the subcellular localization of proteins in eukaryotic organisms based on the amino acid content of primary sequences using amino acid order is designed. The approach for prediction is to find the most frequent motifs for each protein in a given class based on clustering via self organizing maps and then to use these most frequent motifs as features for classification by the help of multi layer perceptrons. This approach allows a classification independent of the length of the sequence. In addition to these, the use of a new encoding scheme is described for the amino acids that conserves biological function based on point of accepted mutations (PAM) substitution matrix. The statistical test results of the system is presented on a four class problem. P2SL achieves slightly higher prediction accuracy than the similar studies.

1. Giriş

Ökaryotik hücreler fonksiyonel olarak zarla kaplı kompartımanlara ayrılmıştır. Proteinler hücre içinde, etkin olabilmek için belli bir bölgede bulunmalıdır. Geniş çaplı genom analizi sayesinde oldukça fazla genler olduğu tahmin edilmektedir. Bundan dolayı, protein hücre içi yer belirleme tahmini fonksiyon özelliği bulma ve yapay protein tasarımı açısından önem kazanır. Tam otomatik ve doğru çalışan bir tahmin sistemi çok yararlı olacaktır.

Bu makalede, birincil dizilerin amino asit içeriği baz alınarak protein hücre içi yer belirleme tahmini için P2SL adında bir sistemin gelişme sonuçları verilmektedir. Tam veya yarı dizilerin amino asit içeriği global bir öznitelik, amino asit sırası da yerel bir öznitelik olarak ele alınır. Biz sadece yerel olanlarla ilgilendik. Bizim yaklaşımımızda, kümelendirme yardımıyla en sık görülen motifleri bulma ve bunu sınıflandırma esastır. Böylece dizi uzunluğundan bağımsız bir tasarım elde edilir. Geri analiz ve uzman sistem için gerekli kurallar içinde altyapı görevi görülür. Bunlara ilave olarak ve daha da önemlisi P2SL, kabul edilebilir nokta mutasyon değiştirme matrisi (PAM) temelli, biyolojik işlevi muhafaza eden, yeni bir kodlama planı kullanımı tarif eder. PAM sıralanmış peptid dizilerine bakarak amino asit benzerliklerini kullanır. Amino asitlerin evrimsel yakınlıkları hakkında fikir vermesi bakımından önemlidir. Bu çalışmada, iki sınıflı bir sınıflandırıcı üzerinde ön sonuçlar verilir.

Makalenin organizasyonu şöyledir. Kısım 2'de ilgili çalışmalar, kısım 3'de kullanılan veri ve hesaplama yöntemleri, kısım 4'de deneyler ve sonuçlara ilişkin yorumlar, kısım 5'de de vargılar ve geleceğe yönelik planlar anlatılmaktadır.

2. İlgili Çalışmalar

Hücre içi yer belirleme tahmini için çeşitli çalışmalar vardır. Bunlardan en önemlileri, tahmin yöntemleri açısından, N-uçlu sıralama sinyallerine dayananlar ve amino asit içeriği temelli olanlar diye ikiye ayrılır.

PSORT bilinen protein sıralama sinyallerini öznitelik olarak kullanarak çıkarımlara ulaşmak için geliştirilmiştir. iPSORT amino asit indis kuralı tabanlı karar ağacı çıkartır. TargetP, yapay sinir ağları kullanır ve iyi bir başarı yüzdesi vardır. Belli kompartımanları tahmin için birkaç çalışma da vardır. MitoProt ve MTS mitokondriye ait olan proteinleri incelerler. MTS saklı Markov modeli kullanır. SignalP ve ChloroP yapay sinir ağları yardımıyla sırasıyla endoplazmik retikuluma ve kloroplasta yönelen proteinleri tahmin etmeye çalışırlar. SortPred hem yapay sinir ağları hem de saklı Markov modeli kullanarak dört sınıflı bir problemde hücre içi yer belirleme tahmininde bulunur. Doğruluk yüzdeleri açısından SortPred bitki hücrelerinde %86, hayvan hücrelerinde %91 ile bu alanda lider konumundadır. TargetP için aynı değerler sırasıyla %90 ve %88, iPSORT için de %85

0-7803-8318-4104/$20.00©2004 IEEE

ve %84'tür. PSORT ve iPSORT sadece global, SortPred hem yerel hem global, diğerleri de sadece yerel öznitelikleri esas alırlar.

Bu çalışmanın en önemli yanı, kümelendirme arkası sınıflandırma birleşimi gibi görülebilir. P2SL'nin iki sınıflı tahmin sonuçları geçmişteki çalışmalara kıyasla gayet yeterlidir. En son gelmek istenilen nokta, TargetP ve SortPred ile başa baş tahmin doğruluğu veren ve sadece insana ait proteinleri baz alan bir sistem yaratmaktır.

Geçmişteki çalışmalarla karşılaştırmak gerekirse;

- TargetP'e benzer pencere (motif) kullanılır,
- SortPred'e benzer şekilde öz düzenleyen haritalar (SOM) yer alır,
- Yeni bir kodlama biçimi olan PAM matrisi vardır.

3. Yöntemler

Bir protein sınıfı içinde sık ama diğerlerinde nadir olarak bulunan ortak alt dizgiler bulmak, yerel özniteliklerin yer aldığı protein hücre içi yer belirleme tahmini için temel fikirdir. Bu ortak alt dizgilere motif denir. Bilinmeyen bir girdi dizgisi için, varolan motifler belirlenir ve sınıflandırma bunlardan yaralanılarak yapılır. Sisteme verilen girdi, amino asit dizileridir. Bunlar verilerden çıkartılır. Birincil dizi alt dizgilere ayrıştırılır ve PAM250 değiştirme matrisi ile kodlanır. Kodlanmış alt dizgi üstünde kümelendirme, öz düzenleyen haritalar yardımıyla olur. Eğitim fazında, her sınıf için motifler bulunur. Sınama fazında, bilinmeyen bir girdi dizisindeki alt dizgiler verildiğinde, öz düzenleyen haritadaki kazanan düğümlere göre, belirli bir sınıftaki motiflerin varlığını belirten ikili bir vektör oluşturulur. Vektörün boyu sabit olup önceden belirlenir. k-en yakın komşuluk (kNN) sınıflandırması bu ikili vektöre uygulanır ve bilinmeyen protein dizisine bir etiket verilir.

3.1. Veri Gösterimi

Protein dizileri değişken boyutlu karakter katarları olarak, amino asitler de tek bir harf olarak gösterilirler. $\hat{W}$, boyu len($\hat{W}$) olan bir protein dizisini göstersin. $\hat{W}$ sabit boylu alt dizgilere ayrışabilir. Eğer $\kappa <$ len($\hat{W}$) ise, $\hat{W}$'de (len($\hat{W}$)- κ+1) tane alt dizgi vardır ve $\hat{W}$(j:m+j) de j. alt dizgiyi gösterir. Daha fazla hesaplama analizi için, amino asitleri kodlamak lazımdır. Bu işlev için PAM250 kullanacağımızı daha önce belirtmiştik. Bundan sonra, PAM ile kodlanmış diziyi W ile göstereceğiz.

3.2. Kümelendirme

SOM, yönetilmeyen yapay sinir ağı modeli olup, benzer düğümler (nöronlar) arasında yakınlığı da bakarak ilişkilendirme kurup bir harita oluşturur. Girdiyi topolojik olarak hizaya sokar. SOM sık sık yüksek boyutlu bir girdiyi, düşük boyutlu (genelde 2) uzaya indirgemek için kullanılır. Her girdi, çok boyutlu vektörden oluşur. Harita dikdörtgen ya da altıgen şeklinde olabilir. Her düğüm için, girdi öznitelik vektörleriyle aynı boyutta olan bir dayanak vektörü oluşturulur. Girdi vektörleri bu dayanak vektörleriyle kıyaslanır. Eğitim aşamasında her girdi, ağa sürülür ve ağırlık vektörleriyle karşılaştırılır. O andaki girdi ile Euclid metriğine göre en yakın ağırlık vektörünün nöronu, kazanan hücre olarak adlandırılır. Kazanan hücre ve onun belli komşuları, girdi yönünde güncellenir. Böylece, benzer girdi vektörleri haritada kümelenirler.

Sistemde, kümelendirme, eğitim fazında olur ve alt dizgiler topolojik olarak gruplaşır. Bir protein sınıf için belli motifleri bulma problemi SOM'daki o sınıfa ait hücreleri bulma problemine dönüşür. Eğitimin sonunda, kritik düğümler iki sınıftaki alt dizilerin sayılarının farkı alınarak tespit edilir.

C_i^X ve C_i^Y, sırasıyla, *i.* hücredeki X ve Y sınıfları için, alt dizgi kümesinin eleman sayıları olsun. Eğer haritanın boyutu *m*'ye *n* ise, toplam *m.n* tane düğüm vardır. *i.* hücrede her iki sınıf için, alt dizgilerin farkı,

$$\Delta C_i^X = (C_i^X - C_i^Y) \text{ ve } \Delta C_i^Y = -\Delta C_i^X = (C_i^Y - C_i^X)$$

olur. Eğitim sırasında, X ve Y sınıflarının motiflerinin atandığı SOM kritik düğüm kümeleri P^X ve P^Y olup, şöyle belirlenir. Eğer $\Delta C_i^X > \tau^X$ ise, $i \in P^X$ olur, aksi takdirde $i \notin P^X$ de değildir. τ^X önceden belirlenmiş eşik değeridir. Benzer şekilde, $\Delta C_j^Y > \tau^Y$ ise, $j \in P^X$ olur, aksi takdirde $j \notin P^X$ de değildir. P^X ve P^Y deki eleman sayısı eşit olup, *s* kadardır.

3.3. Sınıflandırma

k-en yakın komşuluk yöntemi sınıflandırma da kullanılır. Doğrusal ve karesel sınıflandırıcılarla karşılaştıracak olursak, özellikle karmaşık örnek dağılımları için kNN yöntemi daha etkindir.

Her bir eğitim girdisi *W* için, *2s* boyundaki ikili vektör *Z* şu şekilde oluşturulur. *0,1, ... , s-1* nolu elemanlar *X* sınıfını, geriye kalanlar ise *Y* sınıfını temsil eder. Eğer $\hat{W}(j:j+\kappa)$ alt dizgisi için kazanan düğüm P^X in *m.* elemanı ise *Z(m)=1* olur. Benzer şekilde, eğer $\hat{W}(j:j+\kappa)$ alt dizgisi için kazanan düğüm P^Y nin *m.* elemanı ise *Z(s+m)=1* olur. Bir protein dizisindeki, tüm alt dizgiler işlendikten sonra, *Z* son halini alır.

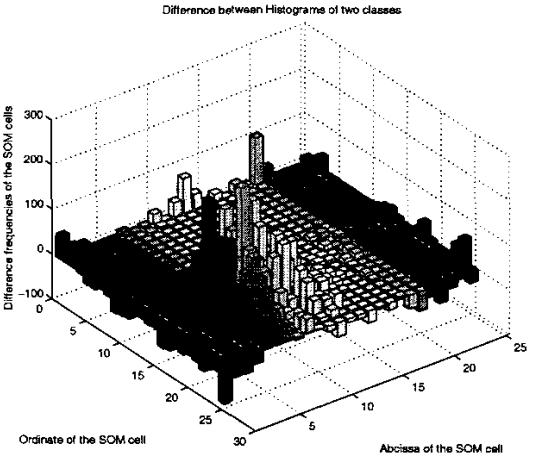
Farz edelim ki, Z" eğitim kümesindeki ikili vektörü, *Z* de sınama kümesindekini temsil etsin. Sınama kümesindeki her bir protein *i* için, Z_i *ve* $Z"_j$ (*j* eğitim kümesindeki her eleman için) arasındaki Hamming uzaklığı hesaplanır. Önceden belirlenmiş bir tek *k* değeri için, Z_i ile en az olan *k* tane eğitim kümesi proteini kontrol edilir. Bu *k* protein içinden, *q* tanesinin X sınıfına, *r* tanesinin de Y sınıfına ait olduğunu sanalım. Buradan *k=q+r* olur. Bundan sonra, bir oylama düzeneği devreye girer. Eğer $q > r$ ise, sınama kümesinin *i.* elemanı X sınıfı ile etiketlenir, değilse Y sınıfı ile etiketlenir. *k* sayısı tek olduğundan, $q=r$ ihtimali yoktur.

4. Malzeme ve Neticeler

Deneylerimizde, daha evvelden yayınlanmış bir veri kümesine kullandık. Bu veri kümesindeki "signal peptide" (SP) proteinleri ile "nuclear" (NP) proteinleri birbirinden ayırmak için bir sınama düzeneği tasarlandı. Her iki sınıftaki proteinler de "Fasta" dosya biçimindeydi.

İki sınıf için de, girdi verisi oluşturmak üzere 40 protein rasgele seçildi. Birbirini dışlayacak şekilde, 20 tanesi

eğitim kümesine, 20 tanesi de sınama kümesine dahil edildi. $\kappa=30$ olmak şartıyla, alt dizgiler çıkartıldı. SOM için, SOM-PAK adlı hesaplama programı kullanıldı. Değişik boyutlu, topolojili ve komşuluk işlevli haritalar denendi. Deneylerden elde ettiğimiz sonuçlara göre, rasgele başlangıç durumlu, 25x25 ebadında, dikdörtgen topolojili, Gausian komşuluk işlevli haritalar daha iyi neticeler verdi. Bir deneydeki eğitim sonrası SOM düğümleri histogram farkları şekil 1 de verilir. Her iki sınıf için de tepeler aşikardır. Fakat önemli düğümler, bir sınıf için içinde oldukça fazla örneklerin olduğu ama diğer sınıf için anlamlı olmayan düğümlerdir. k değeri 5 alınıp, sınama neticeleri 100% doğruluk vermiştir. Sınamanın eğitime oranı ¼ olup, bir örüntü algılama deneyi için oldukça yeterlidir.



Şekil 1. SP ve NP sınıfları için SOM hücreleri histogram farkı

5. Sonuçlar

Amino asit sırası kullanarak, protein hücre içi yer belirleme tahmini için tasarlanan bir sistem olan P2SL'i tanıttık. Diğer iki sınıflı çalışmalara göre daha iyi doğruluk sonuçlarına ulaşıldı. Kümelendirme için öz düzenleyen haritalar, sınıflandırma için k-en yakın komşuluk yöntemleri kullanıldı. Amino asit dizilerinden, öznitelik çıkartmak için PAM250 değiştirme matrisi ile kodlama yapıldı. Bu kodlama, protein hücre içi hedef dizi motiflerindeki her farklı amino asitin biyolojik işlevini muhafaza etmesini sağlar. Kümelendirmeden çıkartılan baskın vektörleri seçmek, sınıflandırma stratejimizin temelini oluşturur. Bir sonraki aşamada, iki sınıftan dört sınıfa çıkmayı planlıyoruz. Sitoplazmik ve mitokondriye ait sınıflar eklenecektir. SOM eğitimden daha fazla örnekler kullanılmalıdır. Çok katmanlı perseptronlar, kNN yerine veya beraber yer alabilir. SOM kümelendirme işlemi ters analiz için faydalı olabilir.

6. Kaynakça

[1] van Vliet C., Thomas E.C., Merino-Trigo A., Teasdale R.D., Gleeson P.A. and Smith, J. O., "Intracellular sorting and transport of proteins", *Prog. Biophys. Mol Biology*, 83(1):1-45, 2003.

[2] Corpet F., Servant F., Gouzy J. and Kahn D., "ProDom and ProDom-CG: tools for protein domain analysis and whole genome comparisons", *Nucleic Acids Research*, 28:267-269, 2000.

[3] Dayhoff M.O., Schwartz R.M. and Orcutt B.C., "A model of evolutionary change in proteins", *Atlas of protein sequence and structure. Vol. 5, Suppl. 3*:345-352, 1979.

[4] Nakai K. and Kanehisa M., "A knowledge base for predicting protein localization sites in the eukaryotic cells", *Genomics*, 14:897-991, 1992.

[5] http://hypothesiscreator.net/iPSORT

[6] Emanuelsson O., Nielsen H., Brunak S. and von Heijne G., "Predicting subcellular localization of proteins based on their N-terminal amino acid sequence", *Journal of Molecular Biology*, 300:1005-1016, 2000.

[7] Claros M.G., "MitoProt: a Macintosh application for studying mitochondrial proteins", *Computer Applications in the Biosciences*, 11(4):441-447, 1995.

[8] Fujiwara Y., Asogawa H. and Nakai K., "Prediction of mitochondrial targeting signals using hidden Markov models", *Genome Informatics*, 8:53-60, 1997.

[9] Nielsen H., Engelbrecht J., Brunak S., von Heijne G., "A neural network method for identification of prokaryotic and eukaryotic signal peptides and prediction of their cleavage sites", *International Journal of Neural Systems*, 8(5-6):581-599, 1997.

[10] Emanuelsson O., Nielsen H. and von Heijne G., " ChloroP, a neural network-based method for predicting chloroplast transit peptides and their cleavage sites", *Protein Science*, 8:978-984, 1999.

[11] Fujiwara Y. and Asogawa M., "Prediction of Subcellular Localization Using Amino Acid Composition and Order", *Genome Informatics*, 12:103-112, 2001.

[12] Cai Y., Liu X., Chou K., "Artificial neural network model for predicting protein subcellular location", *Computers and Chemistry*, 26:179-182, 2002.

[13] Altschul S.F., "Amino acid substitution matrices from an information theoretic perspective", *Journal of Molecular Biology*, 219:555-565, 1991.

[14] Kohonen T., "The self-organizing map", *Proceedings of the IEEE*, 78(9):1464-1480, 1990.

[15] http://www.cis.hut.fi/nnrc/papers