Biyoenformatik

Yapay Zekâ ve Veri Dersleri Rehberi · Kısım VI: Sistem ve Uygulama

Biyolojik veri | Dizi hizalama | Genom birleştirme | Gen ifadesi | Protein yapısı | Komşu dersler

Sercan KÜLCÜ · Giresun Üniversitesi · Bilgisayar Mühendisliği

© Sercan Külcü. Tüm hakları saklıdır. · Sunumun etkileşimli HTML sürümü: https://sercankulcu.github.io/files/ai_courses/sunumlar/Bolum_39_Biyoenformatik.html

GİRİŞ

Ders Hangi Soruyu Sorar?

> **Biyolojik diziler ve moleküler veriler hesaplamayla nasıl çözümlenir?**

Biyoenformatik, DNA, RNA ve protein dizileri gibi biyolojik verileri algoritma ve istatistikle çözümler. Ders **dizi hizalama**, genom birleştirme, gen ifadesi analizi ve protein yapısı tahmini gibi problemleri, bu verinin büyüklüğü ve gürültüsüyle başa çıkan yöntemlerle ele alır.

Bir bakışta

YAKLAŞIM biyoloji, algoritma ve istatistik	**GİRDİ** DNA, protein dizileri, ifade verisi	**ÇIKTI** hizalama, yapı, işlev tahmini
BAŞARI ÖLÇÜTÜ biyolojik geçerlilik	**MATEMATİK YÜKÜ** orta-yüksek	**VERİ İHTİYACI** çok büyük dizi verisi

GİRİŞ

Kısa Tarihçe

1953
Watson ve Crick, Franklin'in verilerinden de yararlanarak DNA'nın çift sarmal yapısını önerir.

1970
Needleman ve Wunsch dinamik programlamayla global dizi hizalama algoritmasını yayımlar.

1981
Smith ve Waterman yerel dizi hizalama algoritmasını yayımlar.

1990
BLAST yayımlanır: büyük veritabanlarında hızlı, sezgisel benzerlik araması.

2003
İnsan Genom Projesi tamamlanır; dizileme maliyeti hızla düşmeye başlar.

2020
AlphaFold 2, CASP14'te protein yapısını deneysel yöntemlere yakın doğrulukla tahmin eder.

DERS

Temel Kavramlar

Dizi hizalama
Global ve yerel hizalama; puanlama matrisleri.

Dizi veritabanları
GenBank, UniProt, PDB; benzerlik araması.

Saklı Markov modeli
Gen bulma ve protein ailesi profilleri.

Genom birleştirme
Kısa okumalardan genomu kurmak; de Bruijn çizgesi.

Gen ifadesi
RNA-seq ile hangi genin ne kadar çalıştığını ölçmek.

Filogenetik
Dizilerden türlerin akrabalık ağacını kurmak.

Protein yapısı
Amino asit dizisinden üç boyutlu biçimi tahmin.

Çoklu test
Binlerce gen sınanınca yanlış keşif oranını denetlemek.

DERS

Tipik İş Akışı



Örnek: Hastadan alınan bir bakteri dizilenir, kısa okumalar bir genoma birleştirilir ve genler BLAST ile veritabanındaki bilinen dizilerle karşılaştırılır. Antibiyotik direnci genleri bulunursa hekime hangi ilacın işe yaramayacağı bildirilir.

DERS

Derste Neler İşlenir?

1. Moleküler biyoloji özeti ve biyolojik veritabanları
2. İkili hizalama: Needleman-Wunsch, Smith-Waterman
3. Puanlama matrisleri ve BLAST ile benzerlik araması
4. Çoklu dizi hizalama
5. Saklı Markov modelleri ve gen bulma
6. Motif bulma
7. Genom birleştirme: de Bruijn çizgeleri
8. Okuma eşleme ve varyant çağırma
9. Filogenetik ağaçlar
10. RNA-seq ile gen ifadesi ve çoklu test
11. Gen ifadesinde kümeleme ve sınıflandırma
12. Protein yapısı tahmini ve AlphaFold

Kazandırdığı beceriler

Biopython ile dizi işlemek | BLAST ile benzerlik araması | BWA ve SAMtools ile okuma eşleme | Bioconductor ile ifade analizi | hizalama algoritması yazmak

Tipik bir dönemin akışıdır; sıra ve kapsam üniversiteden üniversiteye değişir.

DENE

İki Gen Parçasını Hizala

İki türün aynı genindeki DNA parçaları ne kadar benziyor? Needleman-Wunsch tablosu en iyi hizalamayı bulur. Dizileri ve puanları değiştirin, bir hücreye tıklayıp hesabına bakın.

	–	G	C	A	T	G	C	A
–	**0**	−1	−2	−3	−4	−5	−6	−7
G	−1	**1**	**0**	−1	−2	−3	−4	−5
A	−2	0	0	**1**	0	−1	−2	−3
T	−3	−1	−1	**0**	2	1	0	−1
T	−4	−2	−2	−1	**1**	1	0	−1
A	−5	−3	−3	−1	0	**0**	0	1
C	−6	−4	−2	−2	−1	−1	**1**	0
A	−7	−5	−3	−1	−2	−2	0	**2**

Dizi 1 `GATTACA` Dizi 2 `GCATGCA`

Eşleşme **+1**

Uyumsuzluk **−1**

Boşluk **−1**

```
G-ATTACA
| | |·||
GCA-TGCA
```

Puan **2**: 5 eşleşme, 1 uyumsuzluk, 2 boşluk (turuncu yol).

Biyoenformatik, evrimi bir **puanlama modeline** çevirir: tabloyu doldurmak m×n adım sürer, her hizalamayı denemekse üstel. Aynı fikir BLAST gibi araçların temelidir.

KONUM

Ön Koşullar

Bu dersten önce bilinmesi yararlı olan dersler ve neden gerektikleri.

Ders	Neden gerekir
Makine Öğrenmesi	gen ifadesi sınıflandırma, protein yapısı
Veri Madenciliği	dizi motifleri, kümeleme
Algoritmalar	dinamik programlama: dizi hizalamanın temelindeki yöntem
Veri Yapıları	karma tablolar, ağaçlar ve çizgeler: k-mer sayımı ve genom birleştirme
Biyoloji	DNA, RNA, protein ve gen ifadesinin temel işleyişi

KONUM

Bu Ders Sana Göre mi?

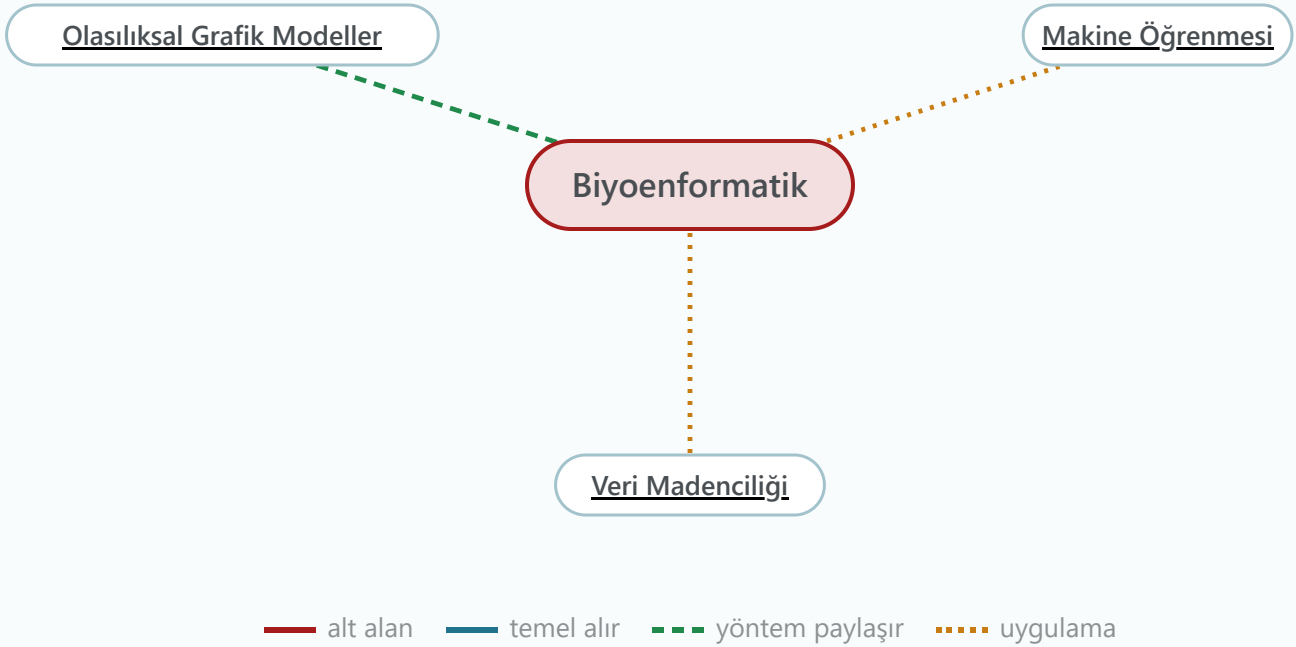
Kendine sor; "hayır" dediklerin için sağdaki sütun nereden başlayacağını gösterir.

Kendine sor	Hayırsa önce
İki dizinin en uzun ortak alt dizisini dinamik programlamayla bulabiliyor musun?	Algoritmalar dersindeki dinamik programlama
Bilginin DNA'dan RNA'ya, oradan proteine nasıl aktarıldığını açıklayabiliyor musun?	Lise biyolojisi ya da moleküler biyolojiye kısa bir giriş
Bir metindeki tüm k uzunluklu alt dizileri karma tabloyla sayabiliyor musun?	Veri Yapıları dersindeki karma tablolar
Bir sınıflandırıcıyı eğitip bir kümeleme sonucunu yorumlayabiliyor musun?	Makine Öğrenmesi

Çoğuna "evet" diyebiliyorsan derse hazırsın; bir ya da iki "hayır" dönem içinde kapanabilir.

KONUM

Dersin Komşuları



ÖRTÜŞME

Örtüştüğü Dersler

İlişki	Ders	Ortak konu
uygulama	Bu ders Makine Öğrenmesi yöntemlerini uygular	gen ifadesi sınıflandırma, protein yapısı
uygulama	Bu ders Veri Madenciliği yöntemlerini uygular	dizi motifleri, kümeleme
yöntem paylaşır	Olasılıksal Grafik Modeller ile yöntem paylaşır	dizi hizalamada saklı Markov modeli

AYRIŞMA

Ayrıştığı Noktalar (1/2)

Ölçüt	Biyoenformatik	Makine Öğrenmesi	Veri Madenciliği
Temel soru	Biyolojik diziler ve moleküler veriler hesaplamayla nasıl çözümlenir?	Bir program örneklerden nasıl genelleme yapar?	Büyük veri yığınlarındaki gizli ve işe yarar örüntüler nasıl bulunur?
Yaklaşım	biyoloji, algoritma ve istatistik	istatistiksel öğrenme	keşfe dayalı çözümleme
Girdi	DNA, protein dizileri, ifade verisi	etiketli ya da etiketsiz örnekler	büyük tablo ya da işlem verisi
Çıktı	hizalama, yapı, işlev tahmini	tahmin yapan model	kural, küme, aykırı değer
Başarı ölçütü	biyolojik geçerlilik	görülmemiş veride hata	ilginçlik, destek, güven
Matematik yükü	orta-yüksek	yüksek: doğrusal cebir, olasılık, optimizasyon	orta
Veri ihtiyacı	çok büyük dizi verisi	orta ve çok	çok

- **Makine Öğrenmesi:** MÖ genel amaçlı öğrenme yöntemleri geliştirir; bu ders onları dizi ve ifade verisine biyolojik bilgiyle birlikte uygular.
- **Veri Madenciliği:** Veri madenciliği her türlü veride örüntü arar; bu ders DNA ve protein dizilerine özgü hizalama ve motif yöntemlerini kullanır.

Ayrıştığı Noktalar (2/2)

Ölçüt	Biyoenformatik	Olasılıksal Grafik Modeller	Tıbbi Görüntü İşleme
Temel soru	Biyolojik diziler ve moleküler veriler hesaplamayla nasıl çözümlenir?	Birbirine bağlı, belirsiz çok sayıda değişken nasıl modellenir ve sorgulanır?	Röntgen, BT ve MR görüntüleri tanıya ve tedaviye nasıl yardım eder?
Yaklaşım	biyoloji, algoritma ve istatistik	olasılık ve çizge bir arada	görüntü işleme ve klinik bilgi
Girdi	DNA, protein dizileri, ifade verisi	değişkenler, bağımlılık yapısı, veri	BT, MR, röntgen, ultrason
Çıktı	hizalama, yapı, işlev tahmini	koşullu olasılık, en olası açıklama	bölütleme, ölçüm, tanı önerisi
Başarı ölçütü	biyolojik geçerlilik	olabilirlik, çıkarım doğruluğu	duyarlılık, özgüllük, Dice
Matematik yükü	orta-yüksek	yüksek: olasılık	orta-yüksek
Veri ihtiyacı	çok büyük dizi verisi	orta	az; etiket pahalı

- **Olasılıksal Grafik Modeller:** Olasılıksal grafik modeller genel bir çerçevedir; bu ders saklı Markov modelini gen bulma ve protein ailelerinde kullanır.
- **Tıbbi Görüntü İşleme:** Tıbbi görüntü işleme organ ve dokunun görüntüsünü çözümler; bu ders moleküler düzeyde dizi ve ifade verisini çözümler.

UYGULAMA

Uygulama Alanları

Kişiye özel tıp
Genetik varyantlara göre tedavi seçimi.

Salgın takibi
Virüs genomlarıyla yayılma yolunu izlemek.

İlaç keşfi
Hedef protein ve molekül etkileşimini öngörmek.

Kanser genomiği
Tümöre özgü mutasyonları bulmak.

Tarım
Verimli ve dayanıklı bitki çeşidi ıslahı.

Evrim çalışmaları
Türler arası akrabalığı dizilerle kurmak.

UYGULAMA

Riskler ve Etik

Genetik mahremiyet

Genom kişiyi ve akrabalarını tanımlar; anonim sayılan genetik veriden kimliğin yeniden çıkarılabildiği gösterildi.

Kökene göre yanlı veri

Genom veritabanları ağırlıkla Avrupa kökenli bireylerden oluşur; risk skorları başka kökenlerde daha az isabetlidir.

Toplu işleme etkisi

Örnekler farklı gün ya da cihazda işlenince oluşan teknik fark, biyolojik bir bulgu gibi görünebilir.

Klinik aşırı yorum

Anlamı belirsiz bir varyantı hastalık nedeni saymak, hastayı gereksiz tedaviye ya da kaygıya yöneltebilir.

Ayrıntılı tartışma: Açıklanabilir Yapay Zekâ ve Yapay Zekâ Etiği.

UYGULAMA

Araçlar ve Örnek Veriler

Araçlar ve kütüphaneler

BLAST | Biopython | Bioconductor (R) | BWA | SAMtools | HMMER | PyMOL

Tanınmış veri setleri

GenBank | UniProt | Protein Data Bank (PDB) | Gene Expression Omnibus (GEO) | 1000 Genom Projesi | The Cancer Genome Atlas (TCGA)

UYGULAMA

Sık Karıştırılanlar

Yaygın düşünce	Doğrusu
Bir gen, bir özellik demektir.	Çoğu özellik birçok genin ve çevrenin ortak etkisidir; tek genin belirlediği özellikler azdır.
Genom dizilenince işlevi de bilinir.	Dizi yalnızca bir harf dizisidir; genlerin ne yaptığını bulmak ayrı ve zor bir çözümlemedir.
Biyoenformatik yalnız biyologlar içindir.	Alanın çekirdeği algoritma, veri yapısı ve istatistiktir; bilgisayar mühendisleri merkezdedir.
Binlerce gen sınanınca anlamlı çıkanlar gerçektir.	Çoklu test yanlış pozitifleri çoğaltır; yanlış keşif oranı gibi düzeltmeler gerekir.

SONRASI

Bu Dersten Sonra

Bu dersi temel alan ya da genişleten dersler

Bu ders çoğunlukla uç noktadır; öteki dersleri temel alır.

İlgili meslekler

Biyoenformatik uzmanı

Hesaplamalı biyolog

Genomik veri bilimci

Biyoistatistikçi

İlaç keşfinde YZ araştırmacısı

Biyoenformatik makine öğrenmesini, veri madenciliğini ve olasılıksal modelleri tek bir uygulama alanında buluşturur. Bu dersten sonra hesaplamalı genomik, yapısal biyoloji ve protein dil modelleri ya da tıbbi görüntü işleme gibi sağlık alanındaki komşu dersler doğal devamdır.

SONRASI

Kaynaklar

- Durbin, Eddy, Krogh ve Mitchison, *Biological Sequence Analysis*, Cambridge University Press, 1998.
- Jones ve Pevzner, *An Introduction to Bioinformatics Algorithms*, MIT Press, 2004.
- Compeau ve Pevzner, *Bioinformatics Algorithms: An Active Learning Approach*, 3. baskı, Active Learning Publishers, 2018.
- Altschul ve ark., "Basic Local Alignment Search Tool", *Journal of Molecular Biology* 215(3), 1990.

Son

Bölüm 39: Biyoenformatik

← Önceki bölüm · Alanın haritası · Sonraki bölüm →