DNA Dizi Hizalama Yöntemleri

Needleman Wunsch ve Smith Waterman

Biyoinformatik 2021 - 2022 Ders Notları

Giriş

Biyoinformatik dersinde iki temel DNA dizi hizalama algoritması gördük. Dizi hizalama yöntemlerinin temel amacı, DNA (veya Protein) dizilerini, homoloji gösteren bölgelerine göre tekrar düzenlemektir. Bu sayede farklı biyolojik problemler üzerine çözümler geliştirilebilir:

- DNA ve Protein dizilerinde bulunan korunmuş bölgelerin tespiti
- Bir referans veri tabanı üzerinde benzerlik araması gerçekleştirilmesi
    - Örnek olarak Blast aracı ile, NCBI Nükleotid veritabanı üzerinde yapılan aramalar
- Ortolog genler veya protein dizileri kullanılarak moleküler filogenetik analizler gerçekleştirilmesi

Şekil 1'i incelerseniz, iki farklı DNA dizisini hizaladığımızda elde ettiğimiz görüntüyü görebilirsiniz. Evrimsel süreci göz önüne aldığımızda, DNA dizileri üzerinde aşağıdaki değişimler gerçekleşebilir:

- Mutasyonlar sonucu oluşan transisyon ve transversiyon değişimleri
- Kromozomlar üzerinde oluşabilecek insersiyon ve delesyonlar
- Ya da herhangi bir değişim gerçekleşmez

```
a = AATCGT
     || ||
b = AA-CG-
```

*Şekil 1: İki DNA dizisinin hizalanmış hali. Bu şekilde, **a** ve **b** incelediğimiz dizileri temsil etsin. Nükleotidler arasındaki doğru eşleşme dik bir çizgi ile gösterilmiştir.*

Dolayısıyla, dizi hizalama algoritmaları bu üç ihtimali göz önünde bulundurarak çalışmaktadır. Eğer nükleotitler arasında bir değişim yoksa, bu **eşleşme (match)** olarak adlandırılır. Eğer nükleotidlerde bir değişim gözleniyorsa, bu **yanlış eşleşme (mismatch)** olarak isimlendirilir. İnsersiyon ve delesyon ihtimaller ise, **boşluk (gap)** karakteri ile ifade edilir.

Dizi hizalama algoritmaları, bu üç ihtimale farklı ağırlıklar atayarak en yüksek homoloji gösteren bölgeleri üst üste getirmeye çalışır. Bu işlem için puanlandırma tablosuna ihtiyacımız duyulur. Puanlandırma için yerleşmiş kriterler bulunmamaktadır ve ihtiyaç

dahilinde optimize edilmelidir. Bu puanlandırma işleminin ardından bir dizi hizalama skoru elde edilecektir.

Dikkat ederseniz eşleşme puanı pozitif iken, yanlış eşleşme ve boşluk puanları ise negatiftir. Bu, fazla yanlış eşleşme ve boşluk olması durumunda hizalama skorunun düşük olacağı anlamına gelmektedir. Düşük skor, aslında kötü bir hizalamayı işaret eder.

Çizelge 1: Örnek olarak verilen puanlama listesi. Bu liste, ders kapsamında verilen örnek için de kullanılmıştır

İhtimal	Puan
Eşleşme	1
Yanlış Eşleşme	-1
Boşluk	-2

Global hizalama yöntemi

Derste ağırlıklı olarak Needleman - Wunsch global hizalama yöntemini işledik. Bu yöntem, iki DNA veya Protein dizisini, uçtan uca hizalamak için kullanılabilecek tekniklerinden birisidir. Uçtan uca hizalama yönteminde kullanılacak diziler için bazı kıstaslar koymuştuk:

- Dizilerin yüksek homoloji göstermeleri gereklidir
- Dizi uzunluklarının da benzer olmaları gereklidir

Şimdi ise adım adım derste yaptığımız örneği tekrar edelim. Dizilerimiz aşağıdaki gibi olsun:

- Dizi 1 (a): AACG
- Dizi 2 (b): AATCGT

Bu iki diziyi hizalamak için aşağıdaki adımlar takip edilir:

1. Matris oluşturma
2. Matriste bulunan kutuların, Çizelge 1'deki puanlar uyarınca doldurulması
3. Geri takip yöntemi ile hizalama yapılması

Birinci adım: matris oluşturma

İlk olarak hizalanacak DNA dizileri bir matris üzerine yerleştirilir. Bu matris üzerinde bulunan her pozisyon, bir önceki tabloda verilen puanlandırma tablosuna göre doldurulacaktır. Puanlandırma her zaman boşluk ihtimallerinin yerleştirilmesiyle başlamaktadır. Şekil 2'de bu matrisi görebilirsiniz. Tire işareti boşluk olarak belirlenir ve boşluk karakterlerinin birleştiği nokta 0 olarak doldurulur (Şekil 2).

Dikkat ederseniz bu matris üzerinde üç farklı yönde hareket edebileceğimiz görürsünüz. Bu yönler aslında, hangi ihtimalin (eşleşme, yanlış eşleşme, ya da boşluk) puanlandırılacağını belirler:

- Matris üzerinde yatay ve düşey eksendeki hareket, dizilerden bir tanesine boşluk ekleme ihtimalini gösterir
- Köşegen yönünde hareket ise eşleşme veya yanlış eşleşme ihtimalini belirtir.

d_{ij}	-	A_1	A_2	C_3	G_4
-	0				
A_1					
A_2					
T_3					
C_4					
G_5					
T_6					

Şekil 2: Puanlandırma matrisinin başlangıcı. Dikkat ederseniz nükleotidlerin altında hangi pozisyonlarda olduğu yazmaktadır. Bu bilgi işimize yarayacak.

İkinci adım: boşluk puanlarını ekleme

Bu adımda ise, yatay ve düşey eksende boşluk olma ihtimallerini dolduracağız. Bu aslında yapılacak en basit hesaplamadır. Yatay ve düşey eksende hareket ettiğimiz için sadece boşluk olma ihtimallerini ekleyeceğiz.

Bu adımdan sonra skorlama matrisimizi doldurmaya başlayacağız. Artık her boş kutu için bir puan hesabı yapmalıyız.

d_{ij}	-	A_1	A_2	C_3	G_4
-	0	-2	-4	-6	-8
A_1	-2				
A_2	-4				
T_3	-6				
C_4	-8				
G_5	-10				
T_6	-12				

Şekil 2: Yatay ve düşey eksende ilk olarak boşluk skorlarının belirlenmesi. Bu hem ilk hem de yapılması en kolay adım. Sadece Çizelge 1'de verilen boşluk skorunu ekliyoruz.

Üçüncü adım: birinci kutucuğu doldurma

Bu noktada sonra, Şekil 2'de görülen matristeki bütün kutucukları doldurmamız gerekecektir. Her kutucuktan elde edeceğimiz sonuç, bir önceki kutucuktaki sonuca bağımlı olacaktır. Burada biraz matris işlemlerine başlayacağız. Aşağıdaki gördüğünüz formül, bize herhangi bir kutucuktaki değerin nasıl hesaplanacağını göstermektedir:

$$di,j = \max \begin{cases} d_{i-1,j-1} + s(a_i, b_j) \\ d_{i-1,j} + s(a_i, -) \\ d_{i,j-1} + s(-, b_j) \end{cases}$$

Bu formülü şu şekilde ifade edebiliriz. Matris üzerinde üç farklı yönde hareket edebileceğimizden bahsetmiştik:

- Eğer düşey ve yatay eksende hareket ediliyorsa, bu boşluk anlamına gelmektedir. Bir önceki kutudaki değere, boşluk skorunu eklememiz gerekir.
- Ancak köşegen üzerinde hareket ediliyorsa, bu eşleşme ya da yanlış eşleşme ihtimalini değerlendirdiğimiz anlamına gelir.
    - Eğer incelenen nükleotidler eşleşme gösteriyorsa, bir önceki kutudaki değere eşleşme skoru eklememiz gerekir
    - Eğer eşleşme yoksa, bu sefer de bir önceki kutudaki değere yanlış eşleşme skorunu eklememiz gerekir.

Daha sonra bu üç değeri inceleyerek, en yüksek puanı o kutucuğa yazacağız ve o değerin hangi kutudan geldiğini işaretleyeceğiz.

Mesela $i = 1$ ve $j = 1$ kutucuğundaki değeri hesaplamaya çalışalım. Burada i değeri satır numaralarını ifade etmekte, j ise sütun numaralarını göstermektedir. Yani $d_{1,1}$ kutucuğundan bahsettiğimizde, aslında bu *a* dizisindeki A_1 nükleotidi ve *b* dizisindeki A_1 nükleotidlerinin kesişim noktasını ifade etmektedir. Yukarıda verilen denklem uyarınca üç farklı ihtimali hesaplıyoruz:

Birinci yol köşegenden gelen puan (eşleşme veya yanlış eşleşme). Dikkat ederseniz, burada nükleotidler birbirleriyle uyuştuğu için eşleşme puanı işliyor. Eğer eşleşme olmasaydı, yanlış eşleşme puanı gelecekti:

$d_{i-1,j-1} + s(a_i, b_j)$

$$= d_{0,0} + s(a_1, b_1)$$
$$= 0 + 1$$
$$= 1$$

İkinci yol düşey eksenden gelen puan (boşluk):

$d_{i-1,j} + s(a_i, -)$

$$= d_{0,1} + s(a_1, -)$$
$$= -2 + (-2)$$
$$-4$$

Üçüncü yol düşey eksenden gelen puan (boşluk):

$d_{i,j-1} + s(-, b_j)$

$$= d_{1,0} + s(-, b_j)$$
$$= -2 + (-2)$$
$$-4$$

Sonuç olarak:

$$d_{1,1} = max\{1, -4, -4\}$$
$$d_{1,1} = 1$$

d_{ij}	-	A_1	A_2	C_3	G_4
-	0	-2	-4	-6	-8
A_1	-2	1			
A_2	-4				
T_3	-6				
C_4	-8				
G_5	-10				
T_6	-12				

Şekil 3: Matris üzerindeki $d_{1,1}$ kutusundaki değerin hesaplanması. Bu kutu maksimum değeri $d_{0,0}$ kutusundan aldığı için, bu yönü işaretledik.

Dördüncü adım: iterasyon, bütün kutucukların doldurulması

Artık bu denklemi her kutuya uygulayarak bütün matrisi dolduruyoruz. Örnek vermek gerekirse, $d_{1,2}$ kutusu için aşağıdaki işlemi uygulayacağız:

$$d_{1,2} = \max \begin{cases} d_{0,1} + s(a_1, b_2) \\ d_{0,2} + s(a_1, -) \\ d_{1,1} + s(-, b_2) \end{cases}$$

Bütün kutuları doldurduğumuzda Şekil 4'ü elde edeceğiz. Dikkat ederseniz, maksimum değerleri elde ettiğimiz yönlerin hepsini işaretledik.

d_{ij}		A_1	A_2	C_3	G_4
	0	-2	-4	-6	-8
A_1	-2	1	-1	-3	-5
A_2	-4	-1	2	0	-2
T_3	-6	-3	0	1	-1
C_4	-8	-5	-2	1	0
G_5	-10	-7	-4	-1	2
T_6	-12	-9	-6	-3	0

Şekil 4: Matris üzerindeki bütün kutucuklar puanlandıktan sonra durum.

Son adım: geri takip

Bu adımda matrisimizin dolmuş olması gerekir artık. Doldurduğumuz son kutucuk bizim dizi hizalama skorumuzu temsil etmektedir. Bu kutu bizim sonlanma kutumuz olarak da adlandırılır.

Artık bu kutucuktan geriye giderek, dizi hizalama işlemini sonlandıracağız. Sağ alt kutudan işleme başlayacağız:

1. Sağ alt kutuya git
2. Bu kutudan hangi kutuya geri gideceğini takip et
    a. Eğer köşegenden geriye doğru gidiyorsan bu hareket eşleşme veya yanlış eşleşme olduğunu gösterir
    b. Eğer düşey veya yatay eksende hareket ediyorsak, bu bize dizilerden bir tanesine boşluk verileceğini gösterir
        i. Eğer düşey eksende hareket varsa, sütunda bulunan diziye bir boşluk ekle
        ii. Eğer yatay eksende hareket varsa, satırda belirtilen diziye bir boşluk ekle
3. Bu şekilde sol üst kutucuğa kadar devam et

Geri takip ve DNA hizalaması

Şimdi adım adım takip edelim:

1. Sağ alt kutucuktan ($d_{6,4}$), düşey eksende bir hareket gerçekleştiriyoruz. Dolayısıyla, sütunda bulunan diziye bir boşluk ekleyeceğiz. Yani b dizisinde bulunan T'nin karşısına bir boşluk gelmiş oldu:

```
a = -

b = T
```

2. Bir yukarı hareketten sonra geldiğimiz kutucuk $d_{5,4}$ oldu. Bu kutudan geriye gittiğimizde ise köşegen üzerinden $d_{4,3}$kutusuna geldik. Bu ya eşleşme ya da yanlış eşleşme hareketidir:

```
a = G-
     |
b = GT
```

3. Geldiğimiz kutudan ($d_{4,3}$) ise yine köşegen üzerinden $d_{3,2}$ kutusuna geldik.

```
a = CG-
     ||
b = CGT
```

4. Şimdi ise düşey eksende hareket ederek $d_{3,2}$ kutusundan $d_{2,2}$ kutusuna geldik. Bu yüzden sütunda bulunan diziye bir boşluk ekleyeceğiz.

```
a = -CG-
      ||
b = TCGT
```

5. Geldiğimiz kutudan $d_{2,2}$ geriye doğru gittiğimizde ise yine köşegen hareketi var. Bu da bir eşleşme ($d_{1,1}$ kutusuna geldik):

```
a = A-CG-
     | ||
b = ATCGT
```

6. Son olarak, $d_{1,1}$ kutusundan $d_{0,0}$ kutusuna gelerek geri takip işlemini sonlandırıyoruz. Bu da bir köşegen hareketi, ve eşleşme var:

```
a = AA-CG-

    || ||

b = AATCGT
```

Bölgesel hizalama

Derste gördüğümüz ikinci yöntem ise bölgesel hizalama tekniğiydi. Bu teknikte amaç, iki DNA dizisinin sahip olduğu bölgesel homolojileri açığa çıkartmak olarak tanımlanabilir. Global hizalamada yüksek homoloji ihtiyacından bahsetmiştik. Bu teknikte ise incelediğimiz diziler birbirleri ile düşük homoloji gösterebilirler. Temel amaç ise, homoloji gösteren kısa nükleotid dizilerini açığa çıkartmak.

Bu sayede çok uzun bir dizi ile, kısa bir DNA dizisini karşılaştırarak, kısa DNA dizisini, uzun dizinin en uygun bölgesine yerleştirebiliriz.

Burada bahsedilen tekniği aslında Genomik ve Proteomik derslerinde sıkça anlatıyorum. Fastq dosyalarındaki DNA okumalarını, insan genomundaki uygun bölgelere yerleştirerek varyantların tespit edilmesi bölgesel hizalamaya en uygun örneklerden verilebilir.

Bölgesel hizalama tekniği, global hizalama ile genel olarak aynı şekilde yapılmaktadır. **Global hizalamadan farklı olarak geri takip işlemine her zaman sağ alttaki kutucuktan başlamıyoruz. En yüksek skora sahip kutucuktan geri takip işlemine başlıyoruz.** Kutucuklardaki puanlandırma aynı şekilde gerçekleşmektedir. Ancak bu yöntemde bölgesel bir eşleşme arandığı için, puanlama parametrelerinde ufak bir değişiklik yapılmaktadır. Eğer, kutucuklardan alınan değer 0'ın altına indiyse, bu değer 0 olarak kabul edilir. Yani skorlama matrisinde 0'ın altında bir değer görme imkanımız olmaz. Aşağıdaki denklemi incelerseniz, parametrelere 0 değerinin eklendiğini görebilirsiniz:

$$di,j = \max \begin{cases} d_{i-1,j-1} + s(a_i, b_j) \\ d_{i-1,j} + s(a_i, -) \\ d_{i,j-1} + s(-, b_j) \\ 0 \end{cases}$$

Artık $d_{i,j}$ kutusunun alabileceği maksimum değerlere bir de 0 eklenmekte. Bu sayede, her kutunun alabileceği en yüksek skorun, sıfırın altına düşmesi engellenmiş olur. Geri kalan skorlama işlemleri şey global hizalama ile aynı şekilde işlemektedir. Bu teknik ile elde edilen skorlama matrisindeki kutucukların büyük bir kısmı 0 olacaktır.

Şimdi ise derste yaptığımız örneği adım adım tekrar edelim. Dizilerimiz aşağıdaki gibi olsun:

- Dizi 1 (a): AACG
- Dizi 2 (b): GTAAACG

Çalışma kapsamında kullanacağımız puanlama yöntemi Çizelge 1'deki gibi olsun tekrar.

Matris oluşturma

Skorlama matrisini, global hizalama ile aynı şekilde oluşturuyoruz. İlk olarak boşluk karakterlerini ekliyoruz, ardından satır ve sütunlara dizilerin sahip olduğu nükleotidleri ekliyoruz. Boşluk karakterlerinin sahip olduğu puanları yatay ve düşey eksenlere yazıyoruz. Ancak dikkat ederseniz, puanların 0'ın altına düşmesine izin vermiyoruz (Şekil 5).

d_{ij}	-	G_1	T_2	A_3	A_4	A_5	C_6	G_7
-	0	0	0	0	0	0	0	0
A_1	0							
A_2	0							
C_3	0							
G_4	0							

Şekil 5: Bölgesel hizalama yönteminde skorlama matrisinin ilk görünümü

Artık $d_{1,1}$ kutucuğunun alacağı değeri hesaplayabiliriz:

Birinci yol köşegenden gelen puan (eşleşme veya yanlış eşleşme). Burada A ve G nükleotidleri yanlış eşleşme gösterdikleri için -1 puan veriyoruz:

$$d_{i-1,j-1} + s(a_i, b_j)$$
$$= d_{0,0} + s(a_1, b_1)$$
$$= 0 + (-1)$$
$$= -1$$

İkinci yol düşey eksenden gelen puan (boşluk):

$$d_{i-1,j} + s(a_i, -)$$
$$= d_{0,1} + s(a_1, -)$$
$$= -2 + (-2)$$
$$-4$$

Üçüncü yol düşey eksenden gelen puan (boşluk):

$$d_{i,j-1} + s(-, b_j)$$

$$= d_{1,0} + s(-, b_j)$$
$$= -2 + (-2)$$
$$-4$$

Sonuç olarak:

$$d_{1,1} = max\{-1, -4, -4, 0\}$$
$$d_{1,1} = 0$$

Diğer kutucukları da bu puanlama yöntemi esasına göre doldurup, en yüksek gelen değerlerin hangi kutucuklardan geldiklerini hesaplıyoruz.

d_{ij}		G_1	T_2	A_3	A_4	A_5	C_6	G_7
	0	0	0	0	0	0	0	0
A_1	0	0	0	1	1	1	0	0
A_2	0	0	0	1	2	2	0	0
C_3	0	0	0	0	0	1	3	1
G_4	0	1	0	0	0	0	1	4

Şekil 6: Doldurulmuş bir skorlama tablosu. Skorumuz 4 olarak belirlenmiştir

Geri takip yöntemi ile hizalama

Artık geri takip yöntemi ile hizalamayı gerçekleştirebiliriz. **Global hizalamadan farklı olarak geri takip işlemine her zaman sağ alttaki kutucuktan başlamıyoruz. En yüksek skora sahip kutucuktan (bu örnekte $d_{4,7}$) geri takip işlemine başlıyoruz.** Bu pozisyonda hareket köşegen üzerinden gittiği için eşleşme ya da yanlış eşleşme ihtimali var:

```
a = G

    |

b = G
```

Bundan $d_{1,4}$ kutucuğuna kadar hep eşleşme ile devam edilmiş.

```
a = AACG

    | | | |

b = AACG
```

Bu pozisyondan sonra, yatay eksende hareket var. Yani *a* dizisine boşluk karakteri ekleyerek devam edip hizalamayı sonlandıracağız:

```
a  =  ---AACG

          ||||

b  =  GTAAACG
```

Şu animasyon bölgesel hizalamayı daha iyi anlatabilir:

https://upload.wikimedia.org/wikipedia/commons/9/92/Smith-Waterman-Algorithm-Example-En.gif

Initialize the scoring matrix

		T	G	T	T	A	C	G	G
	0	0	0	0	0	0	0	0	0
G	0								
G	0								
T	0								
T	0								
G	0								
A	0								
C	0								
T	0								
A	0								

Substitution matrix: $S(a_i, b_j) = \begin{cases} +3, & a_i = b_j \\ -3, & a_i \neq b_j \end{cases}$

Gap penalty: $W_k = kW_1$
$W_1 = 2$

Teşekkür

Metin ve slaytları hazırlanmasındaki katkılarından dolayı Tanya Beril Korkmaz, Merve Akküp ve Caner Doğan Üçgül'e teşekkür ederim.