Maximum likelihood modelleme kullanarak genotip çağırma

Giriş

Bu belgede, maximum likelihood modelleme tekniği kullanarak, varyant çağırma konusu işlenecektir.

Phred skoru

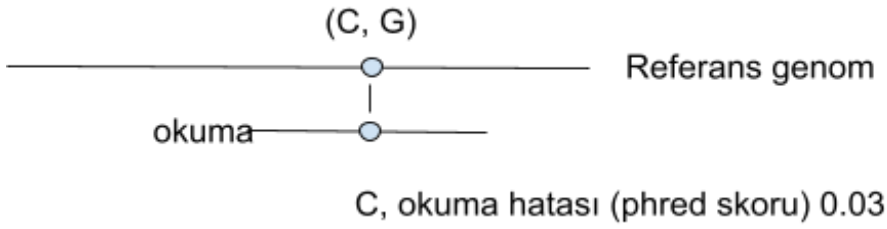
Kalite skoru, Phred skoru olarak da bilinen bu değer, DNA okunurken bir bazın ne kadar hata oranı ile tespit edildiğini belirten bir değerdir.

Phred kalite skoru	Yanlış baz olasılığı	Baz tespit hassasiyeti
10	1/10	90%
20	1/100	99%
30	1/1000	99.9%
40	1/10000	99.99%
50	1/100000	99.999%

Metodun örneklendirilmesi

Tek bir okuma üzerinden inceleme

Genotip çağırma işleminde aslında bu phred kalite değerlerini kullanacağız. Bunu bir örnek ile, tek okuma üzerinden gösterelim:



Şekil 1: Referans genoma hizalanmış bir okuma. Bu okumada gösterilen pozisyona, 0.03 phred kalite skoru ile C nükleotidi okunmuş.

Örnek verelim, elimizde bir okuma olsun. Bu okuma, insan referans genomu üzerine belirli bir pozisyonda hizalandığını düşünelim. Acaba bu pozisyonda, ilgilendiğimiz canlıda hangi genotipi görebiliriz? Bunu hesaplamak için öncelikle o pozisyonda hangi genotiplerin olabileceğini gösteren bir parametre uzayı tespit etmemiz gerek.

Düşünelim ki, o pozisyonda C veya G nükleotidi olabileceğini biliyoruz. O zaman 2n (diploid) bir organizmada, o pozisyonda kaç genotip gözlenebilir?

CC, GC; GG

Şimdilik biz sadece bu genotiplerin oluşma olasılığını tespit edeceğiz.

Adımlar:

1. Parametre uzayı belirle: GC; CC; GG
2. Gözlemini belirle: Tek bir okuma, C, okuma hatası 0.03
3. Parametre ile gözlem arasında olasılıksal bir bağlantı kur
4. Bunu bütün parametreler için tekrarla
5. En olası parametreyi seç

Örneğimiz için aslında 3 farklı genotipin oluşma olasılığını hesaplayacağız:

$$p(Genotip = CG \mid Gözlem = C)$$
$$p(Genotip = CC \mid Gözlem = C)$$
$$p(Genotip = GG \mid Gözlem = C)$$

$$p(Genotip = CG \mid Gözlem = C) = (\frac{1}{2} \times (p(C)) + (\frac{1}{2} \times (p(G) \times \frac{1}{3})$$
$$p(Genotip = CG \mid Gözlem = C) = (\frac{1}{2} \times 0.97) + (\frac{1}{2} \times 0.03 \times \frac{1}{3})$$
$$p(Genotip = CG \mid Gözlem = C) = 0.49$$

- İşlemleri yaptığımızda sonuç 0.49 çıkacaktır.
- ½= Anne veya babadan gelme olasılığıdır. İnsan genomun üzerinde çalıştığımız için genler diploittir yani 2n bu sebeple bir anneden bir babadan gibi düşünebiliriz.
- C C uyuşması olduğunda bu doğru okumadır. bu sebeple 0.97= doğru okuma oranıdır.
- Kırmızı olarak gösterilen yerin ⅓ olma nedeni ise C nükleotidi hariç diğer nükleotidlerle eşleşme olasılığıdır (A,T,G).

$$p(Genotip = CC \mid Gözlem = C) = (\frac{1}{2} \times (p(C)) + (\frac{1}{2} \times p(C))$$
$$p(Genotip = CC \mid Gözlem = C) = (\frac{1}{2} \times 0.97) + (\frac{1}{2} \times 0.97)$$
$$p(Genotip = CC \mid Gözlem = C) = 0.97$$

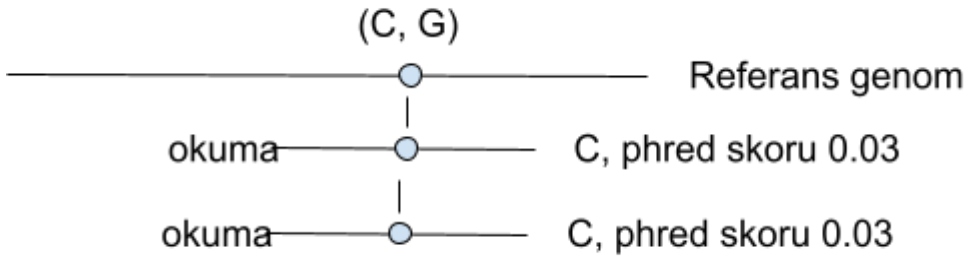
- ½= anne veya babadan gelme olasılığı. C çıkma olasılığı doğru okuma olduğu için 0.97'dir.
- Tekrar C çıkma olasılığı yeniden 0.97'dir.
- İşlemleri yaptığımızda sonuç 0.97 çıkacaktır.

$$p(Genotip = GG \mid Gözlem = C) = (\frac{1}{2} \times (p(G) \times \frac{1}{3}) + (\frac{1}{2} \times (p(G) \times \frac{1}{3})$$
$$p(Genotip = GG \mid Gözlem = C) = (\frac{1}{2} \times 0.03 \times \frac{1}{3}) + (\frac{1}{2} \times 0.03 \times \frac{1}{3})$$
$$p(Genotip = GG \mid Gözlem = C) = 0.01$$

- ½= anne veya babadan gelme olasılığı.G çıkarsa bu referans genomdaki C ile eşleşmediği için yanlış okumadır. 0.03=yanlış okuma oranıdır.
- ⅓ diğer nükleotidlerin gelme olasılığıdır bu sebeple ⅓ ile çarparız.
- İşlemleri yaptığımızda sonuç 0.01 çıkacaktır.

CG=GC çıkma olasılığı= 0.49'tur.
CC çıkma olasılığı = 0.97'dir.
GG çıkma olasılığı = 0.01'dir.

İki okuma üzerinden inceleme



Şimdi ise her okumadan elde ettiğimiz gözlemleri, parametrelerle eşleştireceğiz

$$LH(Genotip = GC \mid x = \{C, C\}) = p(GC|C) \times p(GC|C)$$
$$LH(Genotip = GC \mid x = \{C, C\}) = 0.49 \times 0.49$$
$$LH(Genotip = GC \mid x = \{C, C\}) = 0.24$$

- Burada genotipte GC ile C eşleşmesi (0.49) ve GC ile C eşleşmesi (0.49) olasılıklarını çarparak 0.24 değerini buluruz.

$$LH(Genotip = CC \mid x = \{C, C\}) = p(CC|C) \times p(CC|C)$$
$$LH(Genotip = CC \mid x = \{C, C\}) = 0.97 \times 0.97$$
$$LH(Genotip = CC \mid x = \{C, C\}) = 0.94$$

- Burada genotipte CC ile C eşleşmesi (0.97) ve CC ile C eşleşmesi (0.97) olasılıklarını çarparak 0.94 değerini buluruz.

$$LH(Genotip = GG \mid x = \{C, C\}) = p(GG|C) \times p(GG|C)$$
$$LH(Genotip = GG \mid x = \{C, C\}) = 0.01 \times 0.01$$
$$LH(Genotip = GG \mid x = \{C, C\}) = 0.001$$

- Burada genotipte GC ile C eşleşmesi (0.01) ve GC ile C eşleşmesi (0.01) olasılıklarını çarparak 0.001 değerini buluruz.

Bütün genotipler için

Bu iki örnek sadece bir pozisyonda G veya C olma olasılığı üzerinden işlendi. Aslında elimizde daha fazla olasılık var.

$x = \{A, T, G, C\}$

Bu dört nükleotidin oluşturabileceği 10 farklı ikili kombinasyon vardır.

Bunlar:

1. AA
2. AT
3. AC
4. AG
5. TT
6. TG
7. TC
8. CC
9. CG
10. GG

A

———————————

———

T

———

T

———

T

$$LH(Genotip = AA \mid x = \{A, T, T, T\}) = p(AA|A) \times p(AA|T) \times p(AA|T) \times p(AA|T)$$

= İşleminin sonucun -(log)'sını alıyoruz.
Yukarıda bulunan örnekteki gibi sonucun negatif çıkabileceği işlemlerde -(log)'ya alarak sonucu pozitif yapıyoruz.

Hazırlayanlar: Halit Kemal Aydın, Serhat Kaçan, Gülseven Aleyna Yatmaz