# Bioinformatika

Adrian Klimaševski

2025-11-11

## 1 *Lab1*

### 1.1 Atstumo funkcijos skaičiavimas

Atstumo funkcija apskaičiuota taikant *Euklido atstumo* formulę tarp normalizuotų kodonų ir dikodonų dažnių vektorių.

Yra 2 sekos: A ir B. Kiekvienai jų apskaičiuojami normalizuoti dažniai:

$$f_{A,i} = \frac{n_{A,i}}{\sum_{j=1}^{N} n_{A,j}}, \quad f_{B,i} = \frac{n_{B,i}}{\sum_{j=1}^{N} n_{B,j}}$$

kur:

- $n_{A,i}$ - kiek kartų kodonas arba dikodonas $i$ pasirodo sekoje A
- $N = 4^3 = 64$ kodonų atveju arba $N = 4^6 = 4096$ dikodonų atveju (4, nes *A, T, G, C*, bei 3 ir 6, nes kodonų ir dikondonų ilgiai atitinkamai)
- $f_{A,i}$ ir $f_{B,i}$ - atitinkamai normalizuoti dažniai

Ir tada tarp sekų A bei B apskaičiuojamas atstumas:

$$d(A, B) = \sqrt{\sum_{i=1}^{N} (f_{A,i} - f_{B,i})^2}$$

Gauta reikšmė $d(A, B)$ parodo, kiek skiriasi ir iš jų suformuojama matrica.

## 1.2 Virusų sekų identifikavimas

Table 1: Virusų sekų atitikmenys

| Kodas | Tikras pavadinimas | Ilgis | Forward ORFs | Reverse ORFs |
|---|---|---|---|---|
| B1 | Lactococcus_phage | 29305 | 28 | 35 |
| B2 | KM389305.1 | 28906 | 38 | 50 |
| B3 | NC_028697.1 | 33900 | 41 | 43 |
| B4 | KC821626.1 | 28760 | 32 | 11 |
| M1 | coronavirus | 29903 | 15 | 29 |
| M2 | adenovirus | 34745 | 54 | 51 |
| M3 | U18337.1 | 29309 | 21 | 20 |
| M4 | herpesvirus | 25674 | 46 | 33 |

## 1.3 Atstumų matricos

### 1.3.1 Kodonų atstumų matrica

Table 2: Kodonų dažnių atstumų matrica

| | B1 | B2 | B3 | B4 | M1 | M2 | M3 | M4 |
|---|---|---|---|---|---|---|---|---|
| **B1** | 0.0000 | 0.0610 | 0.0446 | 0.0573 | 0.0581 | 0.0700 | 0.0516 | 0.0928 |
| **B2** | 0.0610 | 0.0000 | 0.0545 | 0.0919 | 0.0617 | 0.0394 | 0.0736 | 0.0534 |
| **B3** | 0.0446 | 0.0545 | 0.0000 | 0.0635 | 0.0579 | 0.0591 | 0.0719 | 0.0836 |
| **B4** | 0.0573 | 0.0919 | 0.0635 | 0.0000 | 0.0782 | 0.0985 | 0.0672 | 0.1174 |
| **M1** | 0.0581 | 0.0617 | 0.0579 | 0.0782 | 0.0000 | 0.0628 | 0.0716 | 0.0849 |
| **M2** | 0.0700 | 0.0394 | 0.0591 | 0.0985 | 0.0628 | 0.0000 | 0.0899 | 0.0415 |
| **M3** | 0.0516 | 0.0736 | 0.0719 | 0.0672 | 0.0716 | 0.0899 | 0.0000 | 0.1038 |
| **M4** | 0.0928 | 0.0534 | 0.0836 | 0.1174 | 0.0849 | 0.0415 | 0.1038 | 0.0000 |

### 1.3.2 Dikodonų atstumų matrica

Table 3: Dikodonų dažnių atstumų matrica

| | B1 | B2 | B3 | B4 | M1 | M2 | M3 | M4 |
|---|---|---|---|---|---|---|---|---|
| **B1** | 0.0000 | 0.0198 | 0.0179 | 0.0228 | 0.0202 | 0.0198 | 0.0194 | 0.0235 |
| **B2** | 0.0198 | 0.0000 | 0.0189 | 0.0271 | 0.0203 | 0.0164 | 0.0219 | 0.0185 |
| **B3** | 0.0179 | 0.0189 | 0.0000 | 0.0228 | 0.0196 | 0.0185 | 0.0218 | 0.0222 |
| **B4** | 0.0228 | 0.0271 | 0.0228 | 0.0000 | 0.0250 | 0.0265 | 0.0237 | 0.0297 |
| **M1** | 0.0202 | 0.0203 | 0.0196 | 0.0250 | 0.0000 | 0.0193 | 0.0226 | 0.0228 |
| **M2** | 0.0198 | 0.0164 | 0.0185 | 0.0265 | 0.0193 | 0.0000 | 0.0229 | 0.0165 |
| **M3** | 0.0194 | 0.0219 | 0.0218 | 0.0237 | 0.0226 | 0.0229 | 0.0000 | 0.0252 |
| **M4** | 0.0235 | 0.0185 | 0.0222 | 0.0297 | 0.0228 | 0.0165 | 0.0252 | 0.0000 |

## 1.4 Matricų analizė

### 1.4.1 Artimiausios ir tolimiausios poros

Table 4: Artimiausių ir tolimiausių virusų porų palyginimas

| Poros tipas | Virusų pora | Kodonų atstumas | Dikodonų atstumas |
|---|---|---|---|
| **Artimiausios** | B2-M2 | 0.0394 | 0.0164 |
| | B1-B3 | 0.0446 | 0.0179 |
| | M2-M4 | 0.0415 | 0.0165 |
| **Tolimiausios** | B4-M4 | 0.1174 | 0.0297 |
| | B4-M2 | 0.0985 | 0.0265 |
| | B4-M1 | 0.0782 | 0.0250 |

## Stebimi modeliai

- Dikodonų atstumai yra žymiai mažesni (3-4 kartus) nei kodonų atstumai
- Artimiausios poros išlieka panašios abiejuose lygmenyse:
  - B2(KM389305.1) - M2(adenovirus)
  - B1(Lactococcus_phage) - B3(NC_028697.1)
  - M2(adenovirus) - M4(herpesvirus)
- Tolimiausios poros visada apima B4(KC821626.1) virusą

## 1.5 Filogenetiniai medžiai

### 1.5.1 Kodonų medis

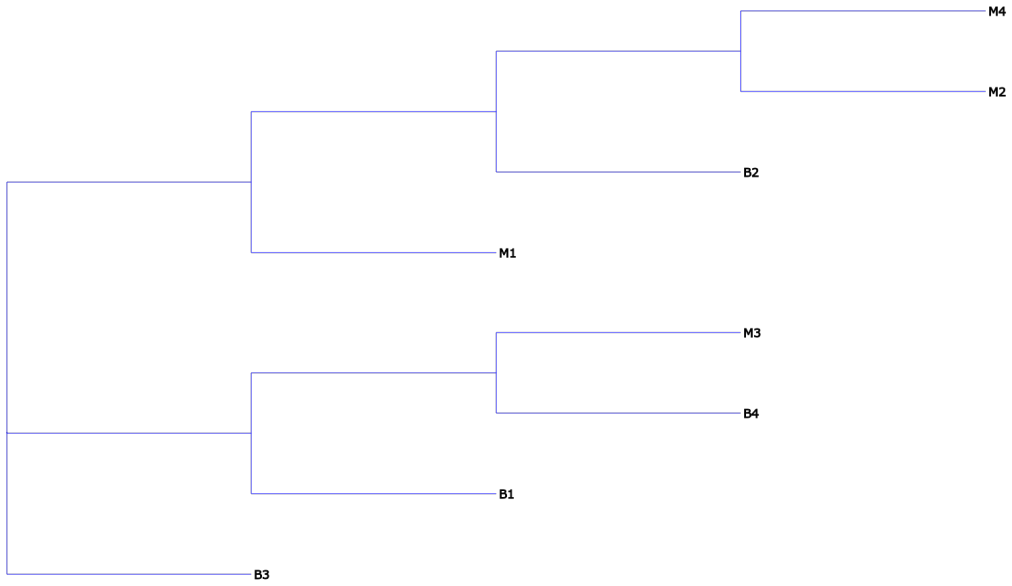


Figure 1: Filogenetinis medis pagal kodonų dažnius

### 1.5.2 Dikodonų medis

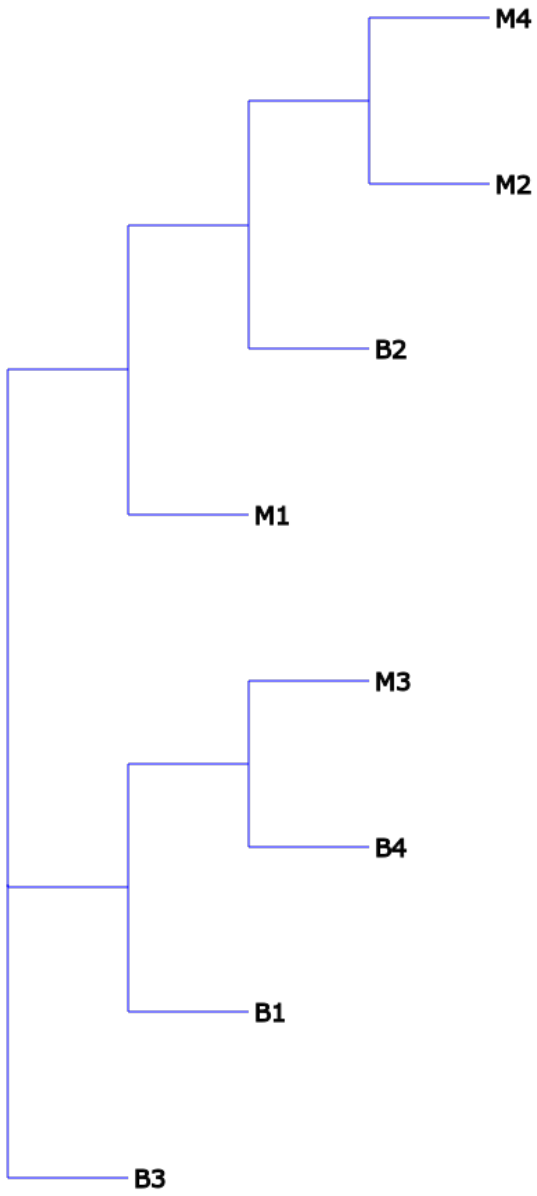


Figure 2: Filogenetinis medis pagal dikodonų dažnius

## 1.6 Extra informacija

### 1.6.1 Kodonų medžio rezultatai

```
(M1:0.2662,(B3:0.1994,(B1:0.1687,(B4:0.3729,M3:0.2990):0.0401):0.0937):0.0479,
(B2:0.1433,(M2:0.1107,M4:0.3044):0.1131):0.1704);
```

- Weight matrix: $W = 1/D^{0.000000}$
- Tree reconstruction method: Weighted least-squares method MW (global optimization)

**Tree Metric (Additive Distance) Matrix:**

```
B1    0.000000 0.062391 0.046181 0.058174 0.057651 0.070446 0.050778 0.089810
B2    0.062391 0.000000 0.056091 0.086822 0.057988 0.036709 0.079426 0.056073
B3    0.046181 0.056091 0.000000 0.070613 0.051352 0.064147 0.063217 0.083511
B4    0.058174 0.086822 0.070613 0.000000 0.082083 0.094878 0.067192 0.114242
M1    0.057651 0.057988 0.051352 0.082083 0.000000 0.066043 0.074687 0.085407
M2    0.070446 0.036709 0.064147 0.094878 0.066043 0.000000 0.087482 0.041510
M3    0.050778 0.079426 0.063217 0.067192 0.074687 0.087482 0.000000 0.106846
M4    0.089810 0.056073 0.083511 0.114242 0.085407 0.041510 0.106846 0.000000
```

**Statistics:**

- Least-squares coefficient $\sum_{i<j}(Dij - ADij)^2 = 0.0003735881$
- Average absolute difference $\sum_{i<j} |Dij - ADij|/(n(n-1)/2) = 0.0028781161$
- Maximum absolute difference $\max_{i,j} |Dij - ADij| = 0.0086831878$
- Total length of the tree $L = 0.2329761219$

**Tree Edge Lengths:**

```
9--4   0.037294
10--9  0.004009
9--7   0.029898
11--3  0.019941
12--5  0.026624
13--2  0.014327
14--6  0.011073
1--10  0.016871
10--11 0.009369
11--12 0.004787
12--13 0.017037
13--14 0.011309
14--8  0.030437
```

### 1.6.2 Dikodonų medžio rezultatai

```
(M1:0.9825,(B3:0.8475,(B1:0.8395,(B4:1.3523,M3:1.0172):0.0858):0.1091):0.0638,
(B2:0.8005,(M2:0.6709,M4:0.9792):0.1194):0.2010);
```

- Weight matrix: $W = 1/D^{0.000000}$
- Tree reconstruction method: Weighted least-squares method MW (global optimization)

**Tree Metric (Additive Distance) Matrix:**

```
B1    0.000000 0.020138 0.017960 0.022775 0.019948 0.020036 0.019424 0.023119
B2    0.020138 0.000000 0.019128 0.026124 0.019840 0.015908 0.022773 0.018991
B3    0.017960 0.019128 0.000000 0.023946 0.018937 0.019026 0.020595 0.022109
B4    0.022775 0.026124 0.023946 0.000000 0.025933 0.026022 0.023695 0.029105
M1    0.019948 0.019840 0.018937 0.025933 0.000000 0.019738 0.022583 0.022821
M2    0.020036 0.015908 0.019026 0.026022 0.019738 0.000000 0.022671 0.016502
M3    0.019424 0.022773 0.020595 0.023695 0.022583 0.022671 0.000000 0.025754
M4    0.023119 0.018991 0.022109 0.029105 0.022821 0.016502 0.025754 0.000000
```

**Statistics:**

- Least-squares coefficient $\sum_{i<j}(Dij - ADij)^2 = 0.0000083402$
- Average absolute difference $\sum_{i<j} |Dij - ADij|/(n(n-1)/2) = 0.0004192914$
- Maximum absolute difference $\max_{i,j} |Dij - ADij| = 0.0012051365$
- Total length of the tree $L = 0.0806854125$

**Tree Edge Lengths:**

```
9--7   0.010172
10--9  0.000858
9--4   0.013523
11--3  0.008475
12--5  0.009825
13--2  0.008005
14--6  0.006709
1--10  0.008395
10--11 0.001091
11--12 0.000638
12--13 0.002010
13--14 0.001194
14--8  0.009792
```

## 1.7 Rezultatai ir išvados

**Ar skiriasi kodonų ir dikodonų dažnis tarp žinduolių ir bakterijų virusų?**

**Atsakymas:** Taip, skiriasi.

**Kaip klasterizuojasi virusai?**

**Atsakymas:** Abiejuose medžiuose grupuojasi panašiai.
**Artimiausios poros**:

- B1(Lactococcus_phage) - B3(NC_028697.1) (atstumas: 0.0462 kodonų, 0.0180 dikodonų)
- M2(adenovirus) - M4(herpesvirus) (atstumas: 0.0415 kodonų, 0.0165 dikodonų)

**Kiti pastebėjimai**:

- Stipriausias klasteris: B1(Lactococcus_phage) - B3(NC_028697.1) - B4(KC821626.1) su M3(U18337.1)
- M3(U18337.1) virusas nuolat grupuojasi su bakteriniais virusais
- Neaiški atskirtis tarp bakterinių ir žinduolių virusų

**Kuris virusas labiausiai išsiskyrė?**

**Atsakymas:** išsiskyrė labiausiai M3(U18337.1) virusas

- Abiejuose medžiuose yra artimesnis bakteriniams virusams B4(KC821626.1)
- Atstumas nuo kitų žinduolių virusų:
  - Iki M1(coronavirus): 0.0747 (kodonai), 0.0226 (dikodonai)
  - Iki M2(adenovirus): 0.0875 (kodonai), 0.0227 (dikodonai)
  - Iki M4(herpesvirus): 0.1068 (kodonai), 0.0258 (dikodonai)
- Atstumas iki B4(KC821626.1): 0.0672 (kodonai), 0.0237 (dikodonai)

**Kokie kodonai/dikodonai labiausiai varijuoja?**

- **Didžiausi atstumai**:
  - B4(KC821626.1) - M4(herpesvirus)
  - B4(KC821626.1) - M2(adenovirus)
  - B4(KC821626.1) - M1(coronavirus)
- **Mažiausi atstumai**:
  - B1(Lactococcus_phage) - B3(NC_028697.1)
  - M2(adenovirus) - M4(herpesvirus)
  - B2(KM389305.1) - M2(adenovirus)
- B4(KC821626.1) virusas nuolat pasirodo tolimiausiose porose, rodant didžiausią kodonų/dikodonų naudojimo skirtumą