6 Evaluation of alignment scores

6.1 Statistical analysis

Statistical tests are performed to give an explanation to observed alignment scores.

Hypothesis testing

- Alternative hypothesis
- Null hypothesis

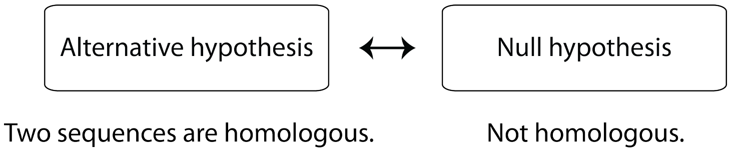

Figure 6.1: The null hypothesis and the alternative hypothesis

P-value

"The p-value is defined as the probability of obtaining a result equal to or more extreme than what was actually observed, assuming that the null hypothesis is true."

– the p-value page on Wikipedia (`https://en.wikipedia.org/wiki/P-value`)

Significance level (α)

The significance level should be chosen to indicate strong/weak evidence against the null hypothesis.

Significance levels 0.05 and 0.01 are often used in life sciences.

- Statistically significant: $\alpha = 0.05$
- Statistically highly significant: $\alpha = 0.01$

Common misunderstandings of p-value

"The p-value is not the probability that the null hypothesis is true or the probability that the alternative hypothesis is false."

– the p-value page on Wikipedia (`https://en.wikipedia.org/wiki/P-value`)

Underlying (background) score distributions

Table 6.1: Alignment methods and distributions

Method	Underlying distribution
Global alignment	Unknown
Local alignment (ungapped)	Gumbel

6.2 Evaluation of global alignment

The underlying distribution of global alignment scores is unknown.

Random generation of sequences

One needs to consider using the appropriate length and compositions of amino acids or nucleotides needs when creating randomised sequences.

Example
Input sequences

```
q: ACGT
d: AGTACC
```

Frequencies: $f_A = 0.2$, $f_C = 0.4$, $f_G = 0.1$, $f_T = 0.3$
Length: 6

```
d1: CCAGTC
d2: TCACCG
d3: CTTGAA
...
```

Frequency distributions

- Universal (e.g. the whole protein database)
- Global (e.g. protein super families)
- Local (e.g. query and database sequences)

Additional constrains

Constrains on sequences generation are often considered.

- Di-amino acid frequencies
- Sub-region specific frequencies

Non-parametric test and p-value

The simplest non-parametric test is calculating the rank of the score for the original alignment as the p-value.

$$p = (b + 1)/(n + 1)$$

where b is the number of randomly generated scores above the score of the original alignment, and n is the sample size.

N.B. n should be sufficiently large (e.g. >1000) to estimate an accurate p-value.

Example

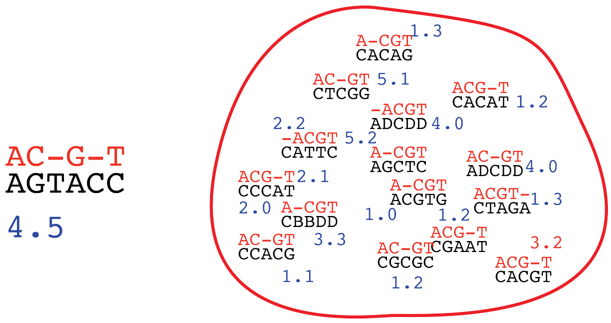

Figure 6.2: Randomly generated sequences and alignment scores

1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
1	1.1	1.3	1.4	1.7	2.1	2.2	2.2	2.3	2.5	2.8	3	3.2	3.3	3.4	3.6	4.2	4.4	4.7	5.2

4.5

p-value: $(2 + 1)/(20 + 1) = 0.1429$

- Significance level $\alpha = 0.2$: reject the null hypothesis
- Significance level $\alpha = 0.05$: the null hypothesis is not rejected

Exercise 6.1

1. Calculate the frequencies of nucleotides from the four sequences below.

```
d1: CCAGC
d2: TCACG
d3: CTTAA
d4: AACAA
```

   Frequencies: $\{f_A = \quad, f_C = \quad, f_G = \quad, f_T = \quad\}$

2. Calculate the p-value of the alignment below.

```
q: AACG
d: A-CG
Score: 40
```

   Assume that the scores are pre-calculated for the alignments of the query sequence and nine randomly generated sequences as follows. Use them for the p-value calculation.

No.	1	2	3	4	5	6	7	8	9
Score	4	14	33	45	74	76	82	83	94

6.3 Evaluation of local alignment

The underlying distribution of local alignment scores is an extreme value distribution.

Gumbel distribution

The Gumbel distribution is a member of the extreme value distribution family.

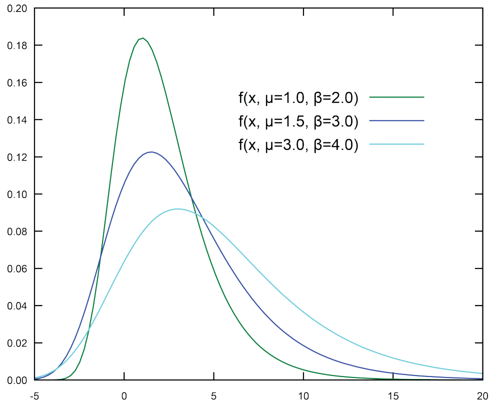

Figure 6.3: Gumbel distribution (source: Herr blaschke, Wikimedia Commons)

The cumulative distribution function (CDF) of the Gumbel distribution:
$F_Y(y) = exp[-e^{-\lambda(y-\mu)}]$

Parameters

- μ: the modal value of the distribution, characteristic value
- λ: a measure of the variance, decay constant

Extreme value distribution

An extreme value distribution is a limiting distribution for the minimum or the maximum of a sufficiently large sample. Ungapped alignments with large sequence lengths are known to have this type of distribution.

Example (m and n are not large in this example)

		A	C	G	C	A	C	G
	0	0	0	0	0	0	0	0
C	0	0	0.5	0	0.5	0	0.5	0
G	0	0	0	1	0.5	0.2	0	1
A	0	0.5	0	0.5	0.7	0.2	0	0.5

CG CG	CG CG	GC GA	...	AC CG	...	...
1	1	0.2		-0.6		

Parameter estimation

The p-value of the Gumbel distribution can be calculated as:

$$P[Y > y] = 1 - F_Y(y) = 1 - exp[-e^{-\lambda(y-\mu)}]$$

The parameters μ and λ can be estimated from the arithmetic mean m_Y and the variance σ_Y^2 of the observed sample.

$$\lambda \approx 1.282/\sigma_Y$$

$$\mu \approx m_Y - 0.577/\lambda$$

Example of parameter estimation

Below is the optimal local alignment with the score between q:ACAGACTACTA and d:TCAGACTGGGAACCE.

```
CAGACT
CAGACT
Score: 6
```

The mean and the variance of the alignment scores are estimated as follows from randomly generated sequences.

m_Y: 1.7221

σ_Y: 1.6025

Then, λ and μ are estimated from m_Y and σ_Y.

$$\lambda \approx 1.282/1.6025 = 0.8$$

$$\mu \approx 1.7221 - 0.577/0.8 \approx 1$$

The p-value is approximately 0.0181 when $\lambda = 0.8$ and $\mu = 1$. The test result is statistically significant (α= 0.05), and therefore, the null hypothesis is rejected.

Conclusion: The query and the database sequences are homologous (p-value: 0.0181).

6.4 Evaluation of database search

BLAST reports bit scores and e-values as search result. Bit score are calculated from raw scores, and e-values represent the expected numbers of database hits.

Example of BLAST output

- q: HSBGPG Human gene for bone gla protein (BGP)

- d: osteocalcin [Felis catus]
- Sequence ID: XP_003999760.1

Score	Expect	Identities	Positives	Gaps
38.5 bits (88)	3.5	19/25 (76%)	20/25 (80%)	0/25 (0%)

```
Query      677      TAFVSKQEGSEVVKRPRRYLYQWLG    751
                     AFVSKQEGSEVV+R RRYL   LG
Sbjct       36      AAFVSKQEGSEVVRRLRRYLAPGLG     60
```

Karlin-Altschul statistics

- λ is a scalar parameter for score matrix
- K is a scalar parameter for search space size

BLAST pre-calculates both parameters in a search space independent manner.

Example of Karlin-Altschul statistics

- Matrix: BLOSUM62
- Lambda: 0.267
- K: 0.041

Sequence databases

The NCBI site provides several databases for BLAST search.

- Nucleotide collection (nr/nt)
- Non-redundant protein sequences (nr)

Example of database statitics

- Database: nr
- Number of letters: 41,667,927,126
- Number of sequences: 113,671,629