GA – TP – AULA 4:
Sequence Analyses

SEQUENCE STATS

SUBSTITUTIONS

GENETIC DISTANCES

MEGA

EXERCISE

Quantitative methods for analysing and comparing sequences

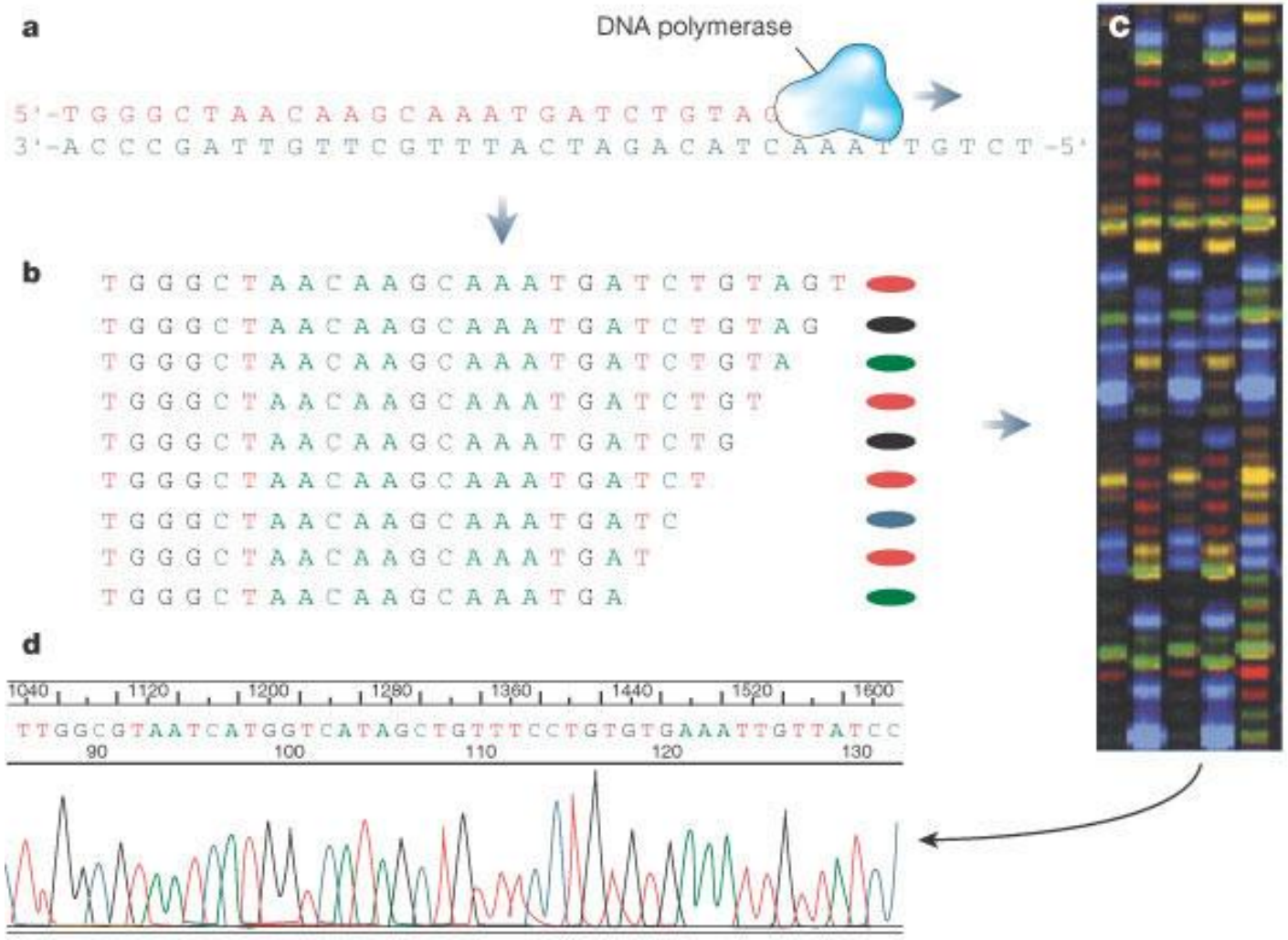
Marta Antunes
mmaantunes@ciencias.ulisboa.pt

GA – TP – AULA 4:
Sequence Analyses

- SEQUENCE STATS
- SUBSTITUTIONS
- GENETIC DISTANCES
- MEGA
- EXERCISE

Marta Antunes
mmaantunes@ciencias.ulisboa.pt

Quantitative sequence analyses



GA – TP – AULA 4:
Sequence Analyses

SEQUENCE STATS

SUBSTITUTIONS

GENETIC DISTANCES

MEGA

EXERCISE

Marta Antunes
mmaantunes@ciencias.ulisboa.pt

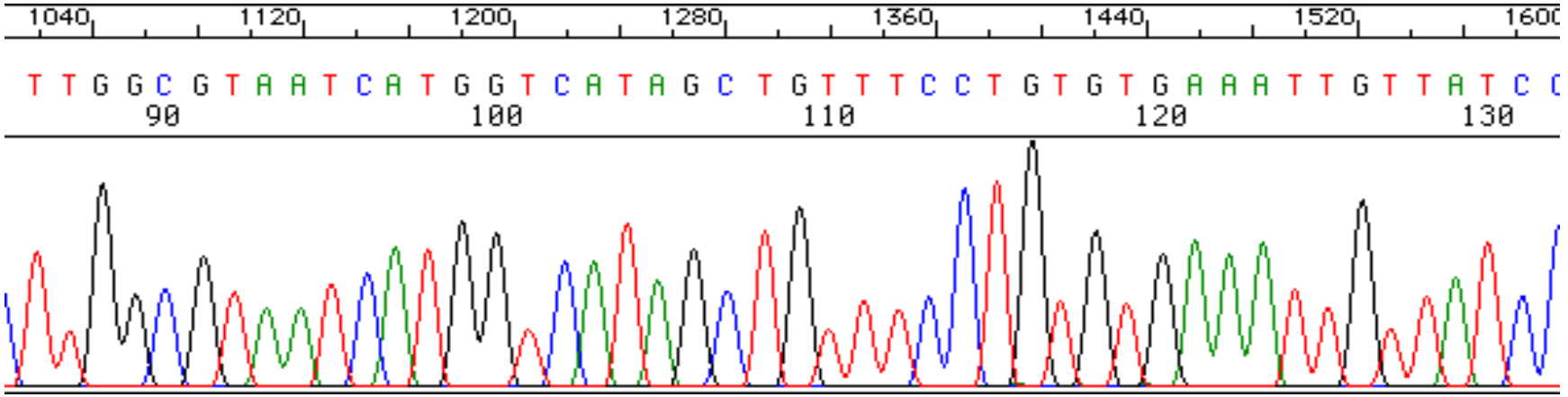
Sequence Characteristics / Stats

INDIVIDUAL SEQUENCES

. Length
. CG content
. Complexity / repeatitiveness
. ORF

. What gene?
. What species?

> sequence 1
TTGGCGTAATCATGGTCATAGCTGTTTCCTGTGTGAAATTGTTATCC



GA – TP – AULA 4:
Sequence Analyses

SEQUENCE STATS

SUBSTITUTIONS

GENETIC DISTANCES

MEGA

EXERCISE

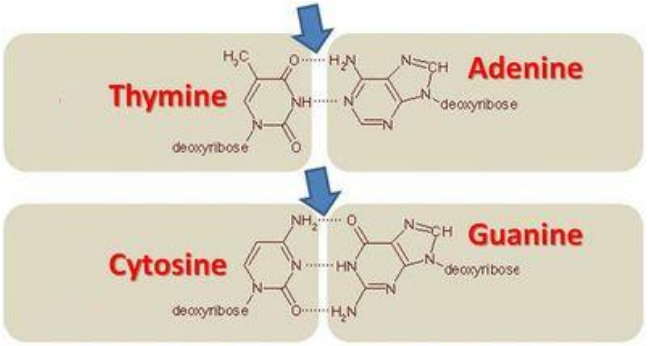
Marta Antunes
mmaantunes@ciencias.ulisboa.pt

Sequence Characteristics / Stats

INDIVIDUAL SEQUENCES

. Length
. CG content (thermal stability)
. Complexity (information content)
. ORF

. What gene?
. What species?



CG content: G≡C *vs* A=T or A=U
DNA stability: high-CG content, higher stability
(but not because of nr of hydrogen bonds)

Bicyclus anynana distal-less (Dll) mRNA, complete cds

GenBank: AF404825.1
FASTA Graphics

Go to:

```
LOCUS       AF404825               4171 bp    mRNA    linear   INV 30-JAN-2002
DEFINITION  Bicyclus anynana distal-less (Dll) mRNA, complete cds.
ACCESSION   AF404825
VERSION     AF404825.1
KEYWORDS    .
SOURCE      Bicyclus anynana (squinting bush brown)
  ORGANISM  Bicyclus anynana
            Eukaryota; Metazoa; Ecdysozoa; Arthropoda; Hexapoda; Insecta;
            Pterygota; Neoptera; Holometabola; Lepidoptera; Glossata; Ditrysia;
            Papilionoidea; Nymphalidae; Satyrinae; Satyrini; Mycalesina;
            Bicyclus.
REFERENCE   1  (bases 1 to 4171)
  AUTHORS   Beldade,P., Brakefield,P.M. and Long,A.D.
  TITLE     Contribution of Distal-less to quantitative variation in butterfly
            eyespots
  JOURNAL   Nature 415 (6869), 315-318 (2002)
   PUBMED   11797007
REFERENCE   2  (bases 1 to 4171)
  AUTHORS   Beldade,P., Brakefield,P.M. and Long,A.D.
  TITLE     Direct Submission
  JOURNAL   Submitted (02-AUG-2001) Evolutionary Biology - EEW, Leiden
            University, 63 Kaiserstraat, Leiden 2311 GP, NL
FEATURES             Location/Qualifiers
     source          1..4171
                     /organism="Bicyclus anynana"
                     /mol_type="mRNA"
                     /db_xref="taxon:110368"
     gene            1..4171
                     /gene="Dll"
     CDS             873..1949
                     /gene="Dll"
                     /note="homeobox transcription factor"
                     /codon_start=1
                     /product="distal-less"
                     /protein_id="AAL69325.1"
                     /translation="MTTQELDHQHHHLGGSQTPHDISNSTNSTPTNVSSKSAFIELQQ
                     HGYGPFKGGYQHPHHFGSPGGQQNPHEASGFPSPRSLGYPFPPMHQNTYGYHIGSYAP
                     QCASPPKDEKCGLSDDPGLRVNGKGKKMRKPRTIYSSLQLQQLNRRFQRTQYLALPER
                     AELAASLGLTQTQVKIWFQNRRSKYKKMMKAAQVGAVPPGLGLPPGSPPNNNQLLHGG
                     GGSSSGSQHSPSGYAGGPAQAHSPTPSSTPVSELSPGLSPTATPWDVKQPPQPSWDVK
                     VGYPTAGRSPDGTSCDVKPPHQQSWDPRVGYGAPPGPWDMKGAHAHALHHQGAPQPHP
                     PYVPQYSWYQADANPGLLTVWPAV"
```

GA – TP – AULA 4:
Sequence Analyses

- SEQUENCE STATS
- SUBSTITUTIONS
- GENETIC DISTANCES
- MEGA
- EXERCISE

Marta Antunes
mmaantunes@ciencias.ulisboa.pt

Sequence Characteristics / Stats

INDIVIDUAL SEQUENCES

. Length
. CG content
. Complexity (~repetition)
. ORF

. What gene?
. What species?

MULTIPLE SEQUENCES

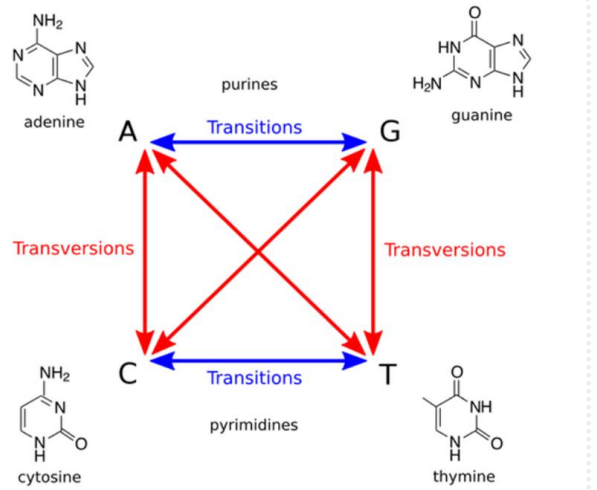
. Sequence polymorphism:
    . SNPs, indels, STRs
. Nr of substitutions:
    . transitions *versus* transversions
    . synonymous *versus* non-synonymous
. Frequencies of variants

Types of substitutions

Nonsynonymous / Synonymous

TCCGATATATGGCAACCCGACAAA
S D I W Q P D K

TCAGATCTATGGCAGCCCCACAAA
S D L W Q P R K



GA – TP – AULA 4:
Sequence Analyses

SEQUENCE STATS

SUBSTITUTIONS

GENETIC DISTANCES

MEGA

EXERCISE

Marta Antunes
mmaantunes@ciencias.ulisboa.pt

Sequence Comparisons

. Same gene within species

individuals from same or different populations

. Paralogs (homologs within one species)

. Orthologs (homologs in different species)

GA – TP – AULA 4:
Sequence Analyses

SEQUENCE STATS

SUBSTITUTIONS

GENETIC DISTANCES

MEGA

EXERCISE

Marta Antunes
mmaantunes@ciencias.ulisboa.pt

Sequence Comparisons

. Same gene within species

individuals from same or different populations

. Paralogs (homologs within one species)

. Orthologs (homologs in different species)

genes with common ancestor = **HOMOLOGS**

PARALOGS
= homologs within one species
i.e.
separated by DUPLICATION event

ORTHOLOGS
= homologs in different species
i.e.
separated by SPECIATION event



GA – TP – AULA 4:
Sequence Analyses

SEQUENCE STATS

SUBSTITUTIONS

GENETIC DISTANCES

MEGA

EXERCISE

Marta Antunes
mmaantunes@ciencias.ulisboa.pt

Sequence Comparisons

. Same gene within species

individuals from same or different populations

. Paralogs (homologs within one species)

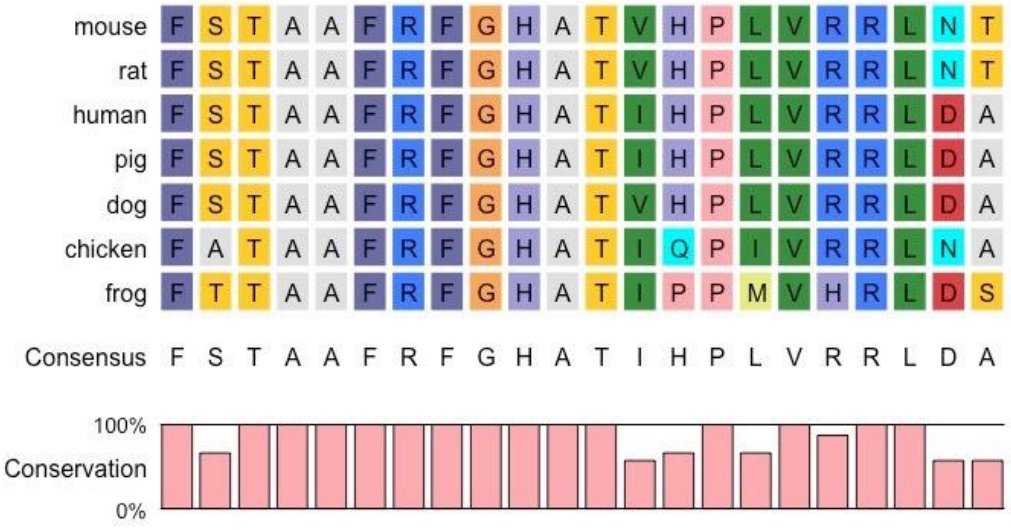
. Orthologs (homologs in different species)

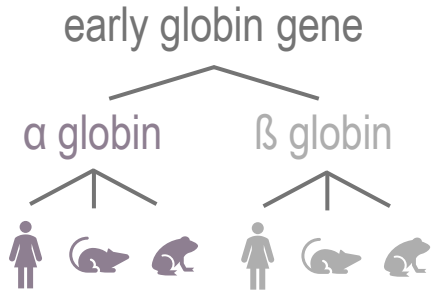
genes with common ancestor = **HOMOLOGS**

PARALOGS
= homologs within one species
i.e.
separated by DUPLICATION event

ORTHOLOGS
= homologs in different species
i.e.
separated by SPECIATION event

. Nucleotide or aminoacid sequences





GA – TP – AULA 4:
Sequence Analyses


SEQUENCE STATS


SUBSTITUTIONS


GENETIC DISTANCES


MEGA


EXERCISE

Marta Antunes
mmaantunes@ciencias.ulisboa.pt

Genetic Distance / Divergence

Between sequences

? What types of sequence differences are there?
What is the frequency of variants
How distinct are pairs of sequences?
How “evolutionarily different” are pairs of sequences?

! Different metrics

Between populations

? How genetically distinct are “different” populations?
Differences in allelic frequencies

! Many metrics

GA – TP – AULA 4:
Sequence Analyses

- SEQUENCE STATS
- SUBSTITUTIONS
- GENETIC DISTANCES
- MEGA
- EXERCISE

Marta Antunes
mmaantunes@ciencias.ulisboa.pt

Genetic Distance between populations

Measures of genetic distance [edit]

Although it is simple to define genetic distance as a measure of genetic divergence, there are several different statistical measures that have been proposed. This has happened because different authors considered different evolutionary models. The most commonly used are Nei's genetic distance,[5] Cavalli-Sforza and Edwards measure,[6] and Reynolds, Weir and Cockerham's genetic distance,[7] listed below.

In all the formulae in this section, X and Y represent two different populations for which L loci have been studied. Let X_u represent the u^{th} allele frequency at the l^{th} locus.

Nei's standard genetic distance [edit]

In 1972, Masatoshi Nei published what came to be known as Nei's standard genetic distance. This distance has the nice property that if the rate of genetic change (amino acid substitution) is constant per year or generation then Nei's standard genetic distance (D) increases in proportion to divergence time. This measure assumes that genetic differences are caused by mutation and genetic drift.[5]

$$D = -\ln \frac{\sum_l \sum_u X_u Y_u}{\sqrt{\left(\sum_u X_u^2\right)\left(\sum_u Y_u^2\right)}}$$

This distance can also be expressed in terms of the arithmetic mean of gene identity. Let j_X be the probability for the two members of population X having the same allele at a particular locus and j_Y be the corresponding probability in population Y. Also, let j_{XY} be the probability for a member of X and a member of Y having the same allele. Now let J_X, J_Y and J_{XY} represent the arithmetic mean of j_X, j_Y and j_{XY} over all loci, respectively. In other words,

$$J_X = \sum_u \frac{X_u^2}{L}$$

$$J_Y = \sum_u \frac{Y_u^2}{L}$$

$$J_{XY} = \sum_l \sum_u \frac{X_u Y_u}{L}$$

where L is the total number of loci examined.[8]

Nei's standard distance can then be written as[5]

$$D = -\ln \frac{J_{XY}}{\sqrt{J_X J_Y}}$$

Cavalli-Sforza chord distance [edit]

In 1967 Luigi Luca Cavalli-Sforza and A. W. F. Edwards published this measure. It assumes that genetic differences arise due to genetic drift only. One major advantage of this measure is that the populations are represented in a hypersphere, the scale of which is one unit per gene substitution. The chord distance in the hyperdimensional sphere is given by:[2][8]

$$D_{CH} = \frac{2}{\pi}\sqrt{2\left(1 - \sum_l \sum_u \sqrt{X_u Y_u}\right)}$$

Some authors drop the factor $\frac{2}{\pi}$ to simplify the formula at the cost of losing the property that the scale is one unit per gene substitution.

Reynolds, Weir, and Cockerham's genetic distance [edit]

In 1983, this measure was published by John Reynolds, Bruce Weir and C. Clark Cockerham. This measure assumes that genetic differentiation occurs only by genetic drift without mutations. It estimates the coancestry coefficient Θ which provides a measure of the genetic divergence by:[7]

$$\Theta_w = \frac{\sum_l \sum_u (X_u - Y_u)^2}{2\sum_l (1 - \sum_u X_u Y_u)}$$

Other measures of genetic distance [edit]

Many other measures of genetic distance have been proposed with varying success.

Nei's D_A distance 1983 [edit]

This distance assumes that genetic differences arise due to mutation and genetic drift, but this distance measure is known to give more reliable population trees than other distances particularly for microsatellite DNA data.[9][10]

$$D_A = 1 - \sum_l \sum_u \sqrt{X_u Y_u}/L$$

Euclidean distance [edit]

Main article: Euclidean distance

$$D_{EU} = \sqrt{\sum_u (X_u - Y_u)^2}$$ [2]

Goldstein distance 1995 [edit]

It was specifically developed for microsatellite markers and is based on the stepwise-mutation model (SMM). μ_X and μ_Y are the means of the allele sizes in population X and Y.[11]

$$(\delta\mu)^2 = \sum_l (\mu_X - \mu_Y)^2/L$$

Nei's minimum genetic distance 1973 [edit]

This measure assumes that genetic differences arise due to mutation and genetic drift.[3]

$$D_m = \frac{(J_X + J_Y)}{2} - J_{XY}$$

Roger's distance 1972 [edit]

$$1\sqrt{\sum_u (X_u - Y_u)^2}$$ [12]

Fst et al.

Fixation index [edit]

Main article: Fixation index

A commonly used measure of genetic distance is the fixation index which varies between 0 and 1. A value of 0 indicates that two populations are genetically identical (minimal or no genetic diversity between the two populations) whereas a value of 1 indicates that two populations are genetically different (maximum genetic diversity between the two populations). No mutation is assumed. Large populations between which there is much migration, for example, tend to be little differentiated whereas small populations between which there is little migration tend to be greatly differentiated. Fst is a convenient measure of this differentiation, and as a result Fst and related statistics are among the most widely used descriptive statistics in population and evolutionary genetics. But Fst is more than a descriptive statistic and measure of genetic differentiation. Fst is directly related to the *Variance* in allele frequency among populations and conversely to the degree of resemblance among individuals within populations. If Fst is small, it means that allele frequencies within each population are very similar; if it is large, it means that allele frequencies are very different.

GA – TP – AULA 4:
Sequence Analyses


SEQUENCE STATS


SUBSTITUTIONS


GENETIC DISTANCES


MEGA


EXERCISE

Marta Antunes
mmaantunes@ciencias.ulisboa.pt

Genetic Distance / Divergence

Between sequences

TODAY'S TP

? What types of sequence differences are there?
What is the frequency of variants
How distinct are pairs of sequences?
How "evolutionarily different" are pairs of sequences?

! Different metrics

Between populations

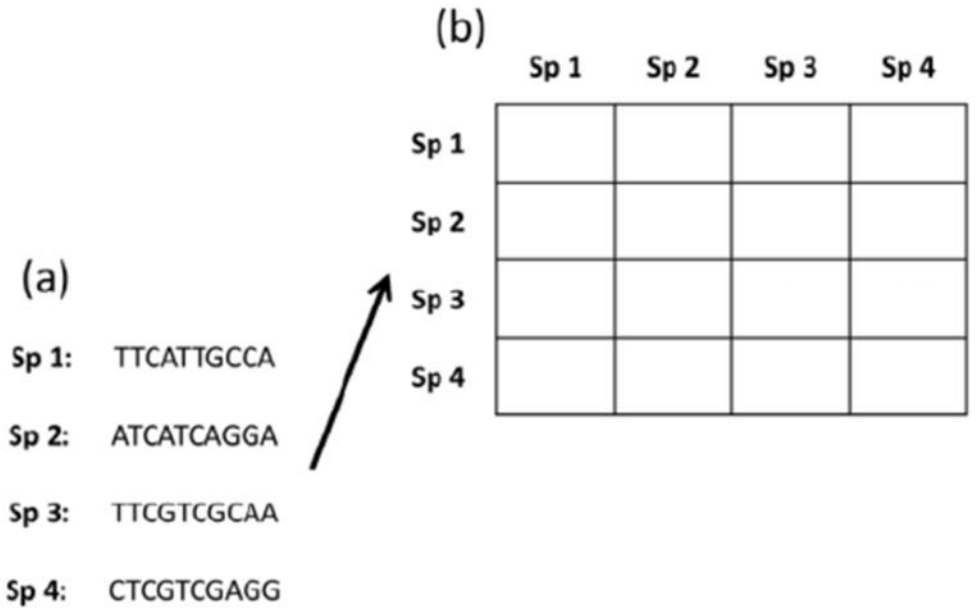
? How genetically distinct are "different" populations?
Differences in allelic frequencies

! Many metrics

GA – TP – AULA 4:
Sequence Analyses

- SEQUENCE STATS
- SUBSTITUTIONS
- GENETIC DISTANCES
- MEGA
- EXERCISE

Genetic Distances: Sequence dissimilarities



Marta Antunes
mmaantunes@ciencias.ulisboa.pt

GA – TP – AULA 4:
Sequence Analyses

SEQUENCE STATS

SUBSTITUTIONS

GENETIC DISTANCES

MEGA

EXERCISE

Marta Antunes
mmaantunes@ciencias.ulisboa.pt

Genetic Distances: Sequence dissimilarities

(a)

Sp 1: TTCATTGCCA

Sp 2: ATCATCAGGA

Sp 3: TTCGTCGCAA

Sp 4: CTCGTCGAGG

(b) proportion of bases that are different

	Sp 1	Sp 2	Sp 3	Sp 4
Sp 1	0	0.5		
Sp 2				
Sp 3				
Sp 4				

GA – TP – AULA 4:
Sequence Analyses

- SEQUENCE STATS
- SUBSTITUTIONS
- GENETIC DISTANCES
- MEGA
- EXERCISE

Marta Antunes
mmaantunes@ciencias.ulisboa.pt

Genetic Distances: Sequence dissimilarities

(a)

Sp 1: TTCATTGCCA
Sp 2: ATCATCAGGA
Sp 3: TTCGTCGCAA
Sp 4: CTCGTCGAGG

(b) proportion of bases that are different

	Sp 1	Sp 2	Sp 3	Sp 4
Sp 1	0	0.5	0.3	0.6
Sp 2	0.5	0	0.5	0.5
Sp 3	0.3	0.5	0	0.4
Sp 4	0.6	0.5	0.4	0

GA – TP – AULA 4:
Sequence Analyses

- SEQUENCE STATS
- SUBSTITUTIONS
- GENETIC DISTANCES
- MEGA
- EXERCISE

Marta Antunes
mmaantunes@ciencias.ulisboa.pt

Genetic Distances: Substitution Models

(a)

Sp 1: TTCATTGCCA
Sp 2: ATCATCAGGA
Sp 3: TTCGTCGCAA
Sp 4: CTCGTCGAGG

(b) p-distance

	Sp 1	Sp 2	Sp 3	Sp 4
Sp 1	0	0.5	0.3	0.6
Sp 2	0.5	0	0.5	0.5
Sp 3	0.3	0.5	0	0.4
Sp 4	0.6	0.5	0.4	0

(c) Jukes-Cantor

	Sp 1	Sp 2	Sp 3	Sp 4
Sp 1	0	0.824	0.383	1.207
Sp 2	0.824	0	0.824	0.824
Sp 3	0.383	0.824	0	0.572
Sp 4	1.207	0.824	0.572	0

Simple example of distance matrices computed from the sequence alignment of four species. (b) Standard p-distance or the proportion of dissimilarity between the sequences; (c) Jukes-Cantor distance matrix.

Conversion of dissimilarity into evolutionary distance by correcting for multiple events per site – different models exist for nucleotide substitutions

Transition matrix

$$P = \begin{array}{c} \\ a \\ g \\ c \\ t \end{array} \begin{array}{c} \begin{array}{cccc} a & g & c & t \end{array} \\ \begin{bmatrix} p_{aa} & p_{ag} & p_{ac} & p_{at} \\ p_{ga} & p_{gg} & p_{gc} & p_{gt} \\ p_{ca} & p_{cg} & p_{cc} & p_{ct} \\ p_{ta} & p_{tg} & p_{tc} & p_{tt} \end{bmatrix} \end{array}$$

GA – TP – AULA 4:
Sequence Analyses

SEQUENCE STATS

SUBSTITUTIONS

GENETIC DISTANCES

MEGA

EXERCISE

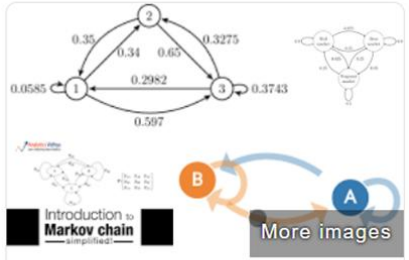
Marta Antunes
mmaantunes@ciencias.ulisboa.pt

Genetic Distances: Substitution Models

A number of different Markov **models of DNA sequence evolution** have been proposed. These substitution models differ in terms of the parameters used to describe the rates at which one nucleotide replaces another during evolution. These models are frequently used in molecular phylogenetic analyses. In particular, they are used during the calculation of likelihood of a tree (in Bayesian and maximum likelihood approaches to tree estimation) and they are used to estimate the evolutionary distance between sequences from the observed differences between the sequences.

Contents [hide]

1 Introduction
2 DNA evolution as a continuous-time Markov chain
2.1 Continuous-time Markov chains
2.2 Deriving the dynamics of substitution
2.3 Ergodicity
2.4 Time reversibility
2.5 Scaling of branch lengths
3 Most common models of DNA evolution
3.1 JC69 model (Jukes and Cantor 1969)
3.2 K80 model (Kimura 1980)
3.3 F81 model (Felsenstein 1981)
3.4 HKY85 model (Hasegawa, Kishino and Yano 1985)
3.5 T92 model (Tamura 1992)
3.6 TN93 model (Tamura and Nei 1993)
3.7 GTR model (Tavaré 1986)
4 See also
5 References
5.1 Further reading
6 External links



Markov chain

A Markov chain is a stochastic model describing a sequence of possible events in which the probability of each event depends only on the state attained in the previous event.
Wikipedia

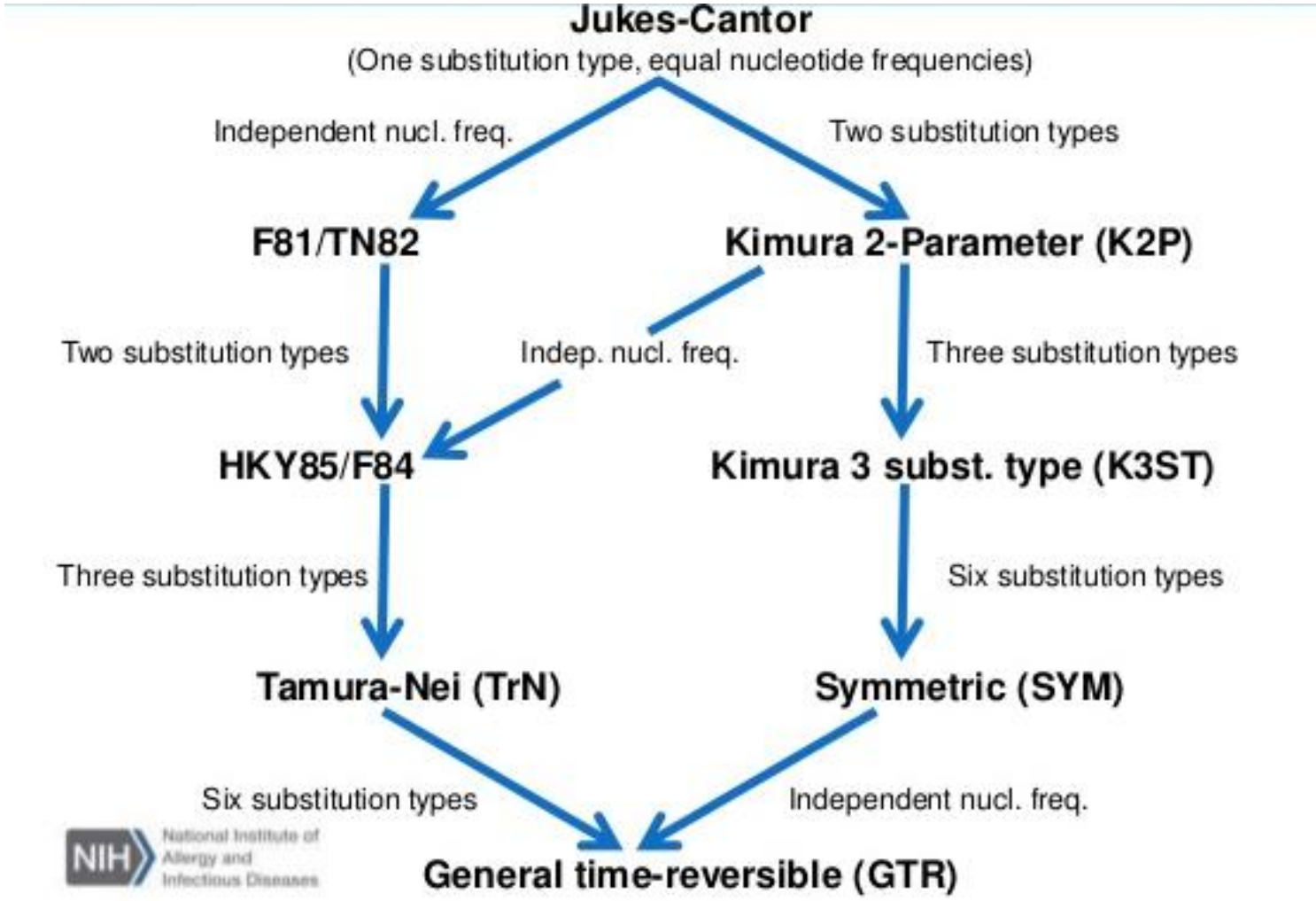
Substitution model	Transition/ transversion bias	Base frequency bas	Rate variation among sites	Heterogeneous patterns among lineages
Jukes–Cantor[28]			Yes	
Kimura[24]	Yes		Yes	
Tamura[26]	Yes	Yes (G+C)	Yes	Yes
Tajima–Nei[27]		Yes	Yes	Yes
Tamura–Nei[25]	Yes	Yes	Yes	Yes
Log-Det[23,29,30]	Yes	Yes		Yes

GA – TP – AULA 4:
Sequence Analyses

- SEQUENCE STATS
- SUBSTITUTIONS
- GENETIC DISTANCES
- MEGA
- EXERCISE

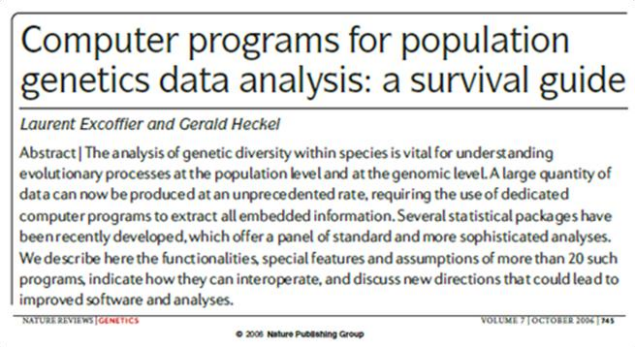
Marta Antunes
mmaantunes@ciencias.ulisboa.pt

Genetic Distances: Substitution Models



GA – TP – AULA 4:
Sequence Analyses

- SEQUENCE STATS
- SUBSTITUTIONS
- GENETIC DISTANCES
- MEGA
- EXERCISE



Computer programs for population genetics data analysis: a survival guide

Laurent Excoffier and Gerald Heckel

Abstract | The analysis of genetic diversity within species is vital for understanding evolutionary processes at the population level and at the genomic level. A large quantity of data can now be produced at an unprecedented rate, requiring the use of dedicated computer programs to extract all embedded information. Several statistical packages have been recently developed, which offer a panel of standard and more sophisticated analyses. We describe here the functionalities, special features and assumptions of more than 20 such programs, indicate how they can interoperate, and discuss new directions that could lead to improved software and analyses.

NATURE REVIEWS | GENETICS VOLUME 7 | OCTOBER 2006 | 745

© 2006 Nature Publishing Group

MARKER ANALYSIS SOFTWARE

Choice

Type of marker
Type of data
Functions
Accessibility

Table 2 | **Program functionalities and assumptions: multi-purpose packages**

Name	Short description of functionalities	Special features	Inference framework	Assumptions and issues*
Arlequin	Computes indices of genetic diversity, *F*-statistics and genetic distances between populations; exact test of HWE, LD and population differentiation; tests selective neutrality within populations; Mantel test; estimates gametic phase from multilocus genotypes; estimates demographic parameters from mismatch distribution	Hierarchical analysis of genetic structure based on the AMOVA framework, with up to 3 levels; tests LD without specifying gametic phase; analysis of multiple files in batch mode	Moment, least-square, ML or Bayesian estimators; most tests are either non-parametric, exact or based on coalescent simulations	Recessive alleles only supported when estimating haplotype frequencies
DnaSP	DNA sequences only; computes LD indices and several pairwise distances between populations; estimates demographic parameters from mismatch distributions; tests selective neutrality at the intra- and interspecific level	Convenient DNA-sequence browser; definition of working subsets of sites; detection of recombination and gene conversion; possibility to compute several statistics in a sliding window	Moment or least-square estimators; coalescent simulation module for confidence intervals of statistics	
FSTAT	Computes basic indices of genetic diversity, allelic richness and *F*-statistics; test of HWE; multiple regression analysis; Mantel test	Test of genotypic disequilibrium between loci for unphased diploid data; compares various statistics among groups of samples; tests sex-biased dispersal based on summary statistics; batch mode	Moment estimators; tests and confidence intervals are based on resampling techniques	
GDA	Computes basic indices of genetic diversity, LD indices and *F*-statistics	Hierarchical analysis of genetic variance with up to 4 levels; detects private alleles	Moment estimators; most tests are exact or based on resampling techniques	
Genepop	Computes basic indices of genetic diversity and *F*-statistics; exact test of HWE and LD; Mantel test	Web interface for remote computations; estimates the number of migrants exchanged between populations based on rare alleles	Moment estimators; most tests are either based on MCMC or resampling techniques	Not updated for some time
GENETIX	Computes basic indices of genetic diversity, LD indices, *F*-statistics and genetic distances between populations; Mantel test	Factorial correspondence analysis; synthetic measure of LD between multi-allelic loci; partitions LD owing to population structure; convenient data editor	Moment-based estimators; tests and confidence intervals are based on resampling techniques	French version only
MEGA	DNA sequences only; computes basic indices of nucleotide diversity, evolutionary distances between sequences and populations; computes phylogenetic trees; tests selective neutrality	Can handle DNA or protein sequences and distance matrices; powerful sequence-data viewer; tree explorer for the visualization of phylogenetic trees with bootstrap support	Moment-based estimators; tests are based on resampling techniques	
MSA	Handles large STR data sets; computes standard genetic diversity indices and several genetic distances between populations	Computes $\delta\mu^2$; input file created in a spreadsheet, and conversion into several other formats	Moment-based estimators	Assumes diploid data
SPAGeDi	Computes genetic distances between populations, as well as several indices of inbreeding, kinship and relatedness at the individual level; focuses on relationships between genetic and geographical distances	Specific handling of spatially explicit genotypic data; estimates gene dispersal from patterns of isolation by distance; polyploid, diploid dominant or co-dominant markers and haploid markers	Moment-based estimators; tests are based on resampling techniques	

*Specific assumptions exist for all implemented methodologies. AMOVA, analysis of molecular variance (this is an inference framework for *F*-statistics that is derived from molecular data, and is based on conventional analysis of variance); HWE, Hardy–Weinberg equilibrium; LD, linkage disequilibrium; MCMC, Markov chain Monte Carlo method; ML, maximum likelihood; STR, short tandem repeats; $\delta\mu^2$, the genetic distance between populations that is specially designed to handle microsatellite (STR) data.

GA – TP – AULA 4:
Sequence Analyses

- SEQUENCE STATS
- SUBSTITUTIONS
- GENETIC DISTANCES
- MEGA
- EXERCISE

Marta Antunes
mmaantunes@ciencias.ulisboa.pt

MOLECULAR EVOLUTIONARY GENETIC ANALYSIS

Name	Short description of functionalities	Special features	Inference framework	Assumptions and issues*
MEGA	DNA sequences only; computes basic indices of nucleotide diversity, evolutionary distances between sequences and populations; computes phylogenetic trees; tests selective neutrality	Can handle DNA or protein sequences and distance matrices; powerful sequence-data viewer; tree explorer for the visualization of phylogenetic trees with bootstrap support	Moment-based estimators; tests are based on resampling techniques	

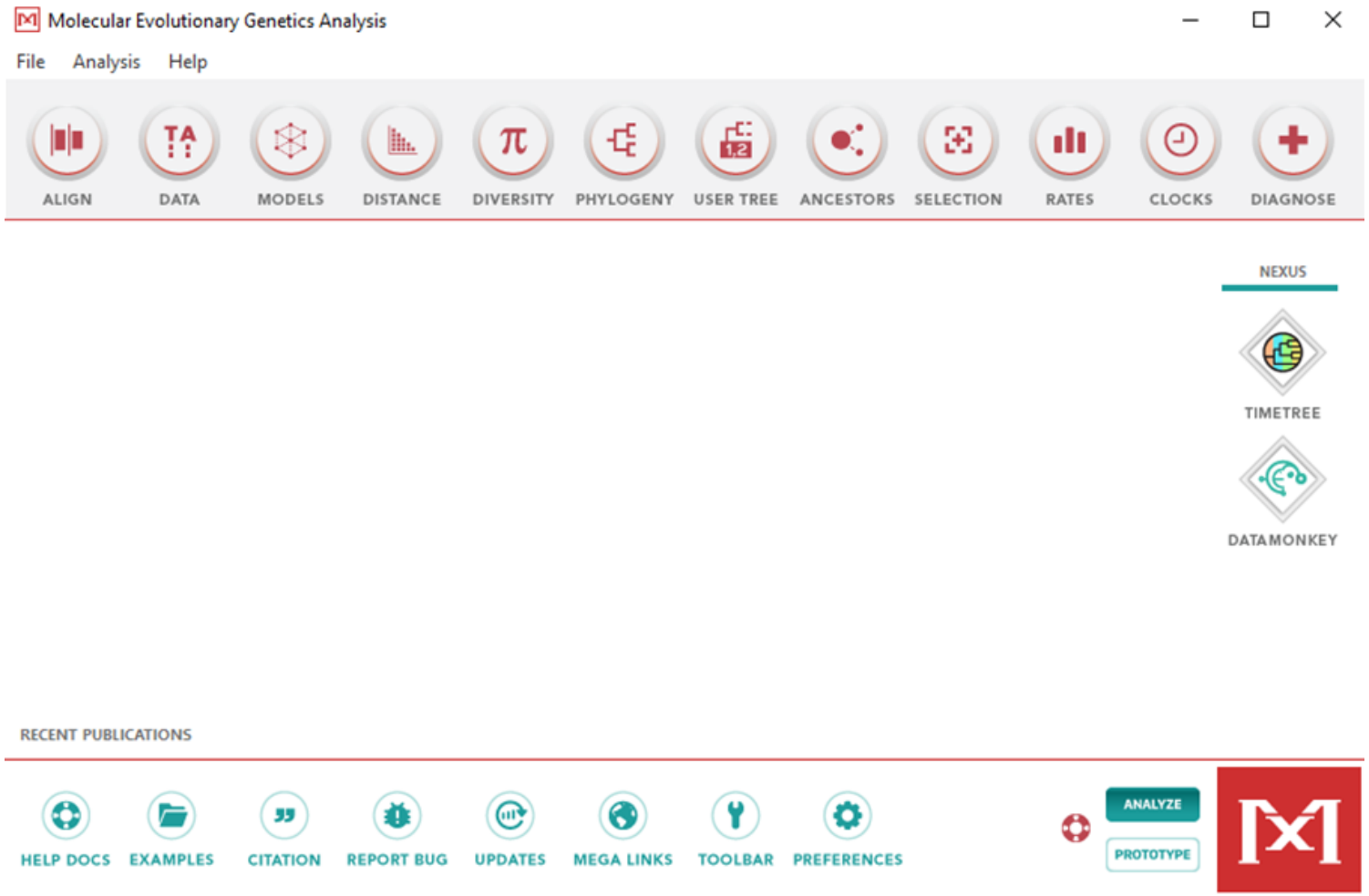


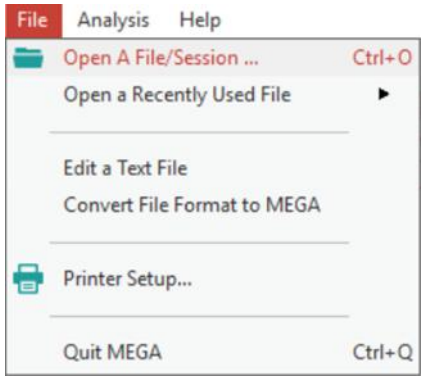
https://www.megasoftware.net/

MEGA Molecular Evolutionary Genetics Analysis

tutorial features documentation feedback









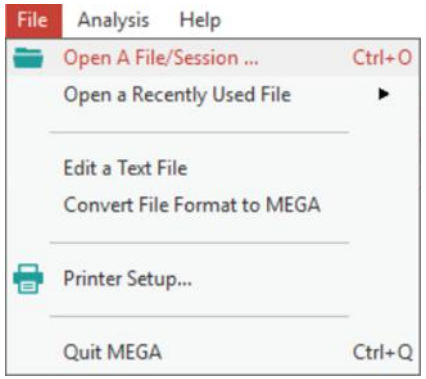
Input files:

nt/aa sequences

distance matrices

Samples:

same / different species





Input files:

nt/aa sequences

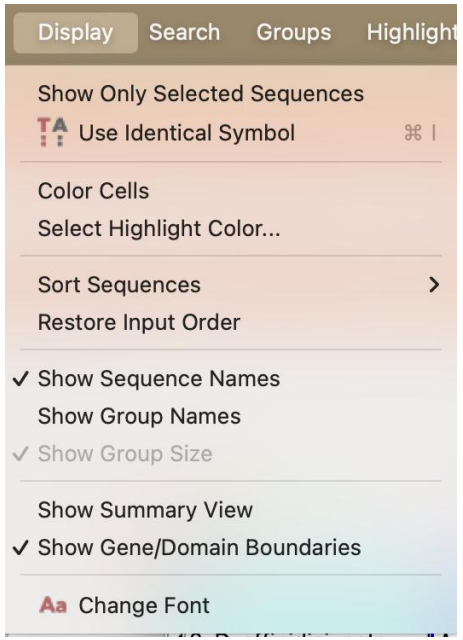
distance matrices

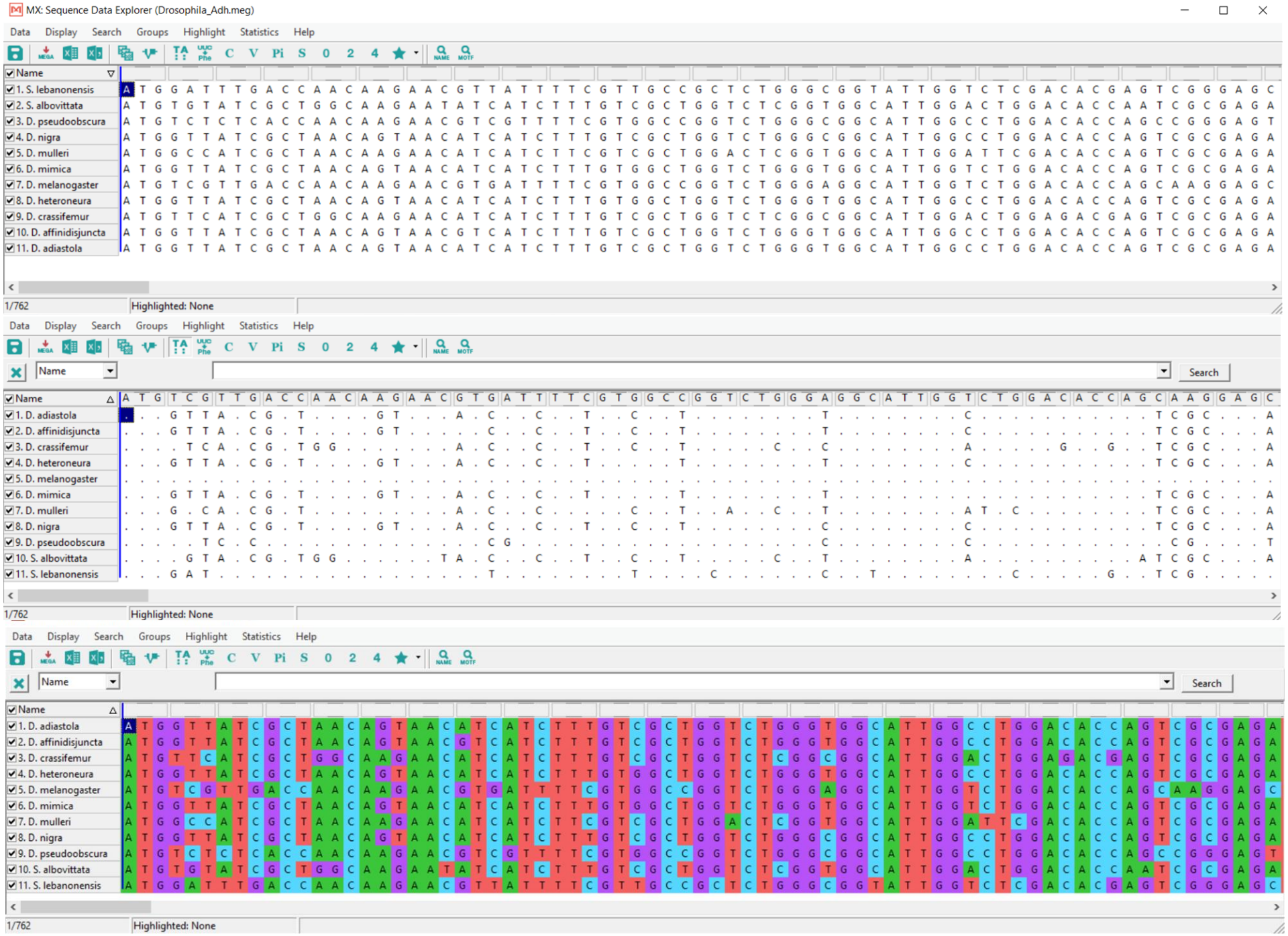
Samples:

same / different species

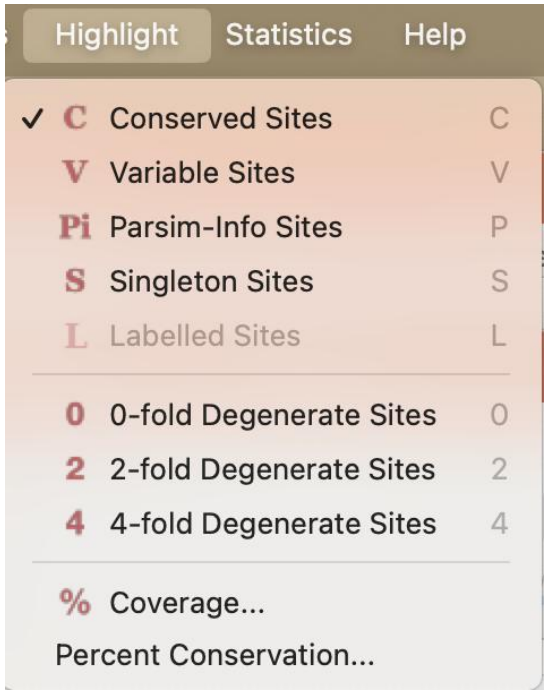
Alignment in different formats

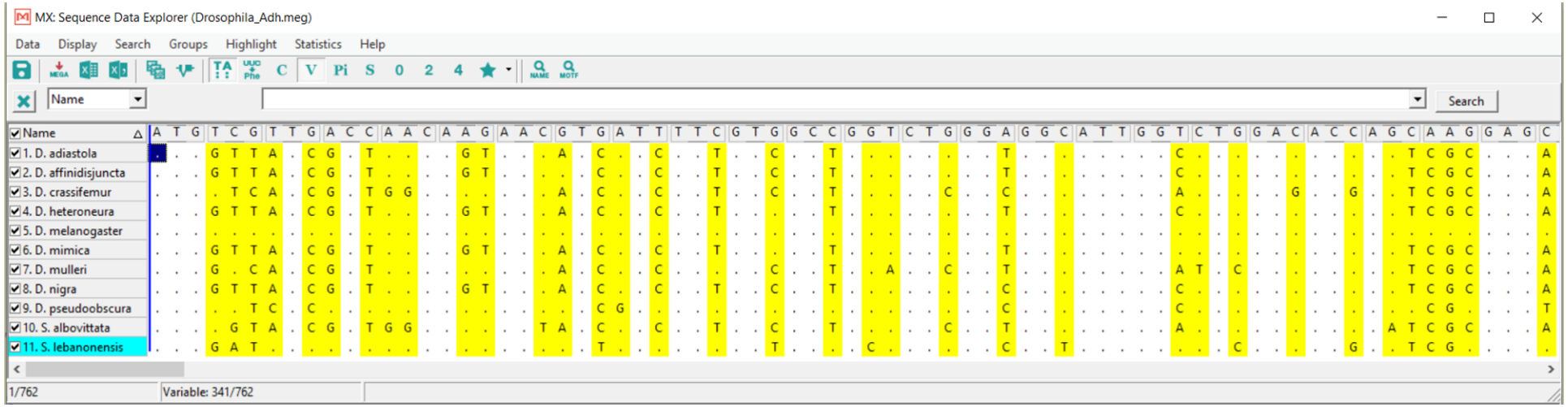


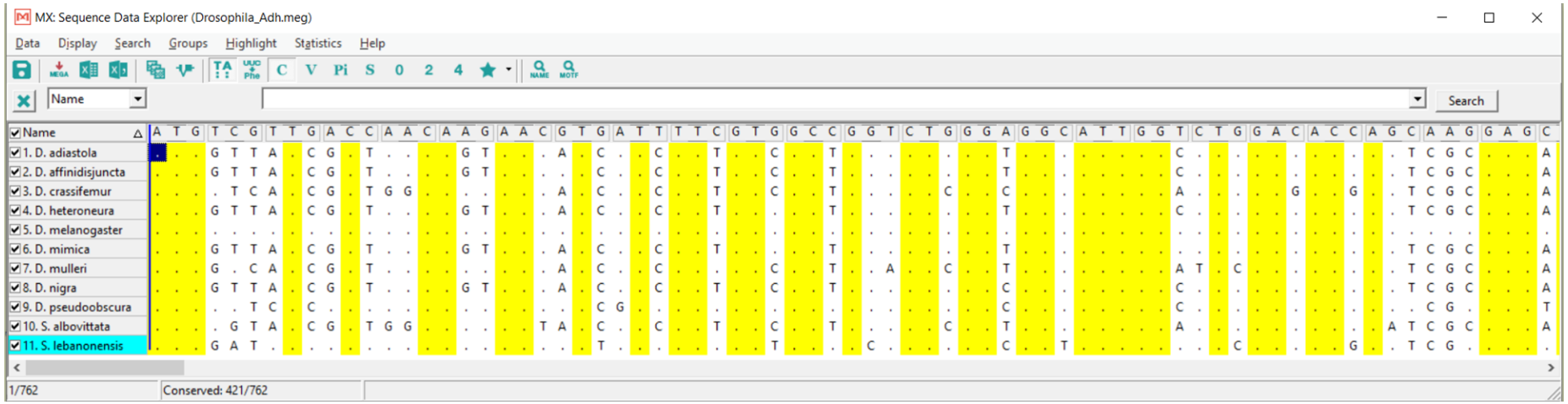


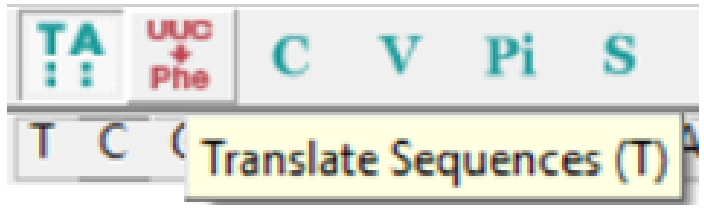


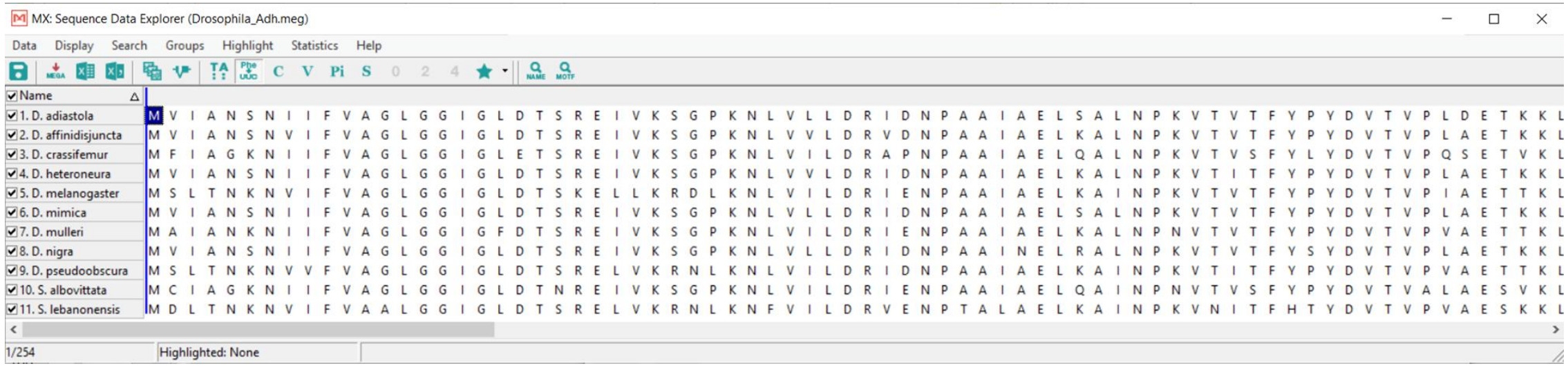
Alignment in different formats

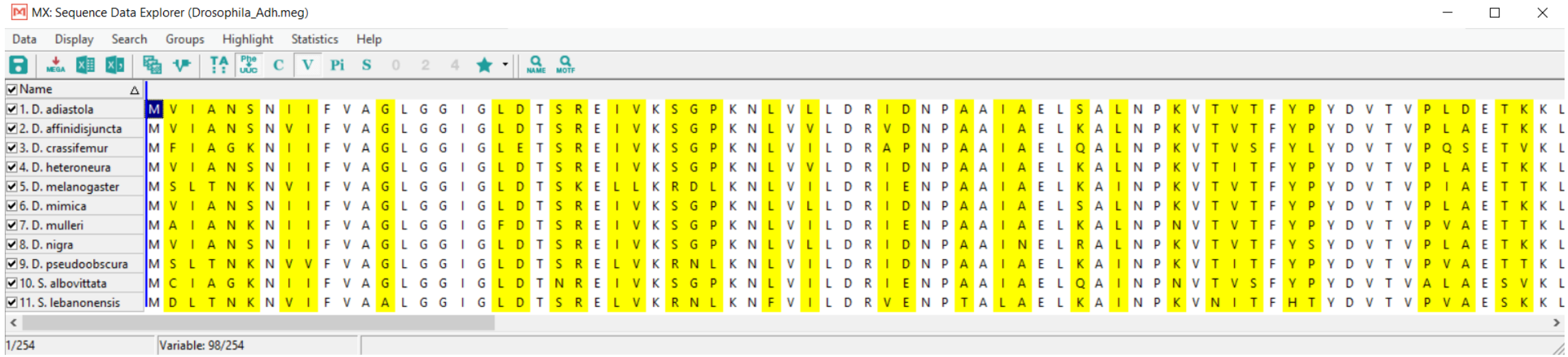




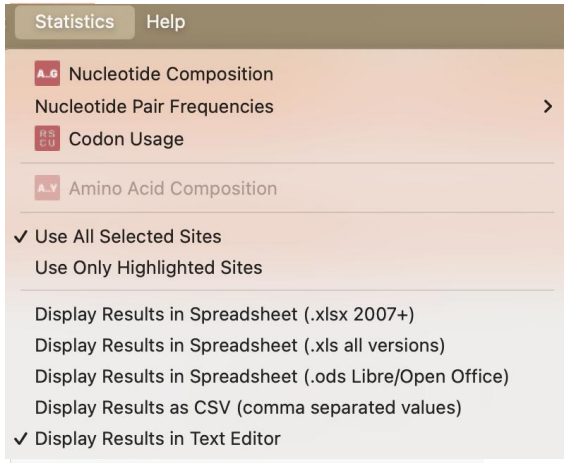



TA
UUC
Phe
C
V
Pi
S
T
C
Translate Sequences (T)


MX: Sequence Data Explorer (Drosophila_Adh.meg)
Data
Display
Search
Groups
Highlight
Statistics
Help
Name
1. D. adiastola M V I A N S N I I F V A G L G G I G L D T S R E I V K S G P K N L V L L D R I D N P A A I A E L S A L N P K V T V T F Y P Y D V T V P L D E T K K L
2. D. affinidisjuncta M V I A N S N V I F V A G L G G I G L D T S R E I V K S G P K N L V V L D R V D N P A A I A E L K A L N P K V T V T F Y P Y D V T V P L A E T K K L
3. D. crassifemur M F I A G K N I I F V A G L G G I G L E T S R E I V K S G P K N L V I L D R A P N P A A I A E L Q A L N P K V T V S F Y L Y D V T V P Q S E T V K L
4. D. heteroneura M V I A N S N I I F V A G L G G I G L D T S R E I V K S G P K N L V V L D R I D N P A A I A E L K A L N P K V T I T F Y P Y D V T V P L A E T K K L
5. D. melanogaster M S L T N K N V I F V A G L G G I G L D T S K E L L K R D L K N L V I L D R I E N P A A I A E L K A I N P K V T V T F Y P Y D V T V P I A E T T K L
6. D. mimica M V I A N S N I I F V A G L G G I G L D T S R E I V K S G P K N L V L L D R I D N P A A I A E L S A L N P K V T V T F Y P Y D V T V P L A E T K K L
7. D. mulleri M A I A N K N I I F V A G L G G I G F D T S R E I V K S G P K N L V I L D R I E N P A A I A E L K A L N P N V T V T F Y P Y D V T V P V A E T T K L
8. D. nigra M V I A N S N I I F V A G L G G I G L D T S R E I V K S G P K N L V L L D R I D N P A A I N E L R A L N P K V T V T F Y S Y D V T V P L A E T K K L
9. D. pseudoobscura M S L T N K N V V F V A G L G G I G L D T S R E L V K R N L K N L V I L D R I D N P A A I A E L K A I N P K V T I T F Y P Y D V T V P V A E T T K L
10. S. albovittata M C I A G K N I I F V A G L G G I G L D T N R E I V K S G P K N L V I L D R I E N P A A I A E L Q A I N P N V T V S F Y P Y D V T V A L A E S V K L
11. S. lebanonensis M D L T N K N V I F V A A L G G I G L D T S R E L V K R N L K N F V I L D R V E N P T A L A E L K A I N P K V N I T F H T Y D V T V P V A E S K K L
1/254
Highlighted: None


MX: Sequence Data Explorer (Drosophila_Adh.meg)
Data
Display
Search
Groups
Highlight
Statistics
Help
Name
1. D. adiastola M V I A N S N I I F V A G L G G I G L D T S R E I V K S G P K N L V L L D R I D N P A A I A E L S A L N P K V T V T F Y P Y D V T V P L D E T K K L
2. D. affinidisjuncta M V I A N S N V I F V A G L G G I G L D T S R E I V K S G P K N L V V L D R V D N P A A I A E L K A L N P K V T V T F Y P Y D V T V P L A E T K K L
3. D. crassifemur M F I A G K N I I F V A G L G G I G L E T S R E I V K S G P K N L V I L D R A P N P A A I A E L Q A L N P K V T V S F Y L Y D V T V P Q S E T V K L
4. D. heteroneura M V I A N S N I I F V A G L G G I G L D T S R E I V K S G P K N L V V L D R I D N P A A I A E L K A L N P K V T I T F Y P Y D V T V P L A E T K K L
5. D. melanogaster M S L T N K N V I F V A G L G G I G L D T S K E L L K R D L K N L V I L D R I E N P A A I A E L K A I N P K V T V T F Y P Y D V T V P I A E T T K L
6. D. mimica M V I A N S N I I F V A G L G G I G L D T S R E I V K S G P K N L V L L D R I D N P A A I A E L S A L N P K V T V T F Y P Y D V T V P L A E T K K L
7. D. mulleri M A I A N K N I I F V A G L G G I G F D T S R E I V K S G P K N L V I L D R I E N P A A I A E L K A L N P N V T V T F Y P Y D V T V P V A E T T K L
8. D. nigra M V I A N S N I I F V A G L G G I G L D T S R E I V K S G P K N L V L L D R I D N P A A I N E L R A L N P K V T V T F Y S Y D V T V P L A E T K K L
9. D. pseudoobscura M S L T N K N V V F V A G L G G I G L D T S R E L V K R N L K N L V I L D R I D N P A A I A E L K A I N P K V T I T F Y P Y D V T V P V A E T T K L
10. S. albovittata M C I A G K N I I F V A G L G G I G L D T N R E I V K S G P K N L V I L D R I E N P A A I A E L Q A I N P N V T V S F Y P Y D V T V A L A E S V K L
11. S. lebanonensis M D L T N K N V I F V A A L G G I G L D T S R E L V K R N L K N F V I L D R V E N P T A L A E L K A I N P K V N I T F H T Y D V T V P V A E S K K L
1/254
Variable: 98/254

Sequence stats: nucleotide composition / aminoacid composition



Data Filename: Drosophila_Adh.meg
Data Title: Adh-sequence data from 11 species
Nucleotide Frequencies
Sites Used: All selected
All frequencies are given in percent.
Domain: Adh data (Gene: Adh)

	T(U)	C	A	G	Total	T-1	C-1	A-1	G-1	Pos #1	T-2	C-2	A-2	G-2	Pos #2	T-3	C-3	A-3	G-3	Pos #3
D. adiastola	25.5	25.6	24.0	24.9	762	13.8	18.1	33.9	34.3	254	31.9	24.8	27.6	15.7	254	30.7	33.9	10.6	24.8	254
D. affinidisjuncta	24.4	25.7	24.8	25.1	762	13.4	17.7	33.9	35.0	254	31.9	26.0	28.0	14.2	254	28.0	33.5	12.6	26.0	254
D. crassifemur	23.9	26.5	23.1	26.5	762	15.4	16.5	33.9	34.3	254	31.9	27.2	27.2	13.8	254	24.4	35.8	8.3	31.5	254
D. heteroneura	24.9	26.0	24.5	24.5	762	13.8	17.7	34.6	33.9	254	32.3	26.0	28.0	13.8	254	28.7	34.3	11.0	26.0	254
D. melanogaster	19.4	32.4	21.9	26.2	762	12.6	19.7	33.9	33.9	254	31.9	26.8	28.3	13.0	254	13.8	50.8	3.5	31.9	254
D. mimica	24.9	26.4	23.6	25.1	762	13.0	18.9	34.3	33.9	254	31.9	25.6	27.6	15.0	254	29.9	34.6	9.1	26.4	254
D. mulleri	21.0	29.9	23.9	25.2	762	13.8	18.9	33.9	33.5	254	30.7	27.6	28.7	13.0	254	18.5	43.3	9.1	29.1	254
D. nigra	24.9	26.8	24.0	24.3	762	14.2	17.7	34.3	33.9	254	31.9	25.6	28.0	14.6	254	28.7	37.0	9.8	24.4	254
D. pseudoobscura	21.7	30.6	22.0	25.7	762	13.4	18.9	34.3	33.5	254	31.5	26.4	29.1	13.0	254	20.1	46.5	2.8	30.7	254
S. albovittata	26.5	24.1	24.3	25.1	762	15.4	15.7	33.9	35.0	254	31.9	26.4	26.8	15.0	254	32.3	30.3	12.2	25.2	254
S. lebanonensis	24.9	27.3	23.5	24.3	762	14.2	18.5	34.6	32.7	254	31.9	25.6	29.9	12.6	254	28.7	37.8	5.9	27.6	254
Avg.	23.8	27.4	23.6	25.2	762	13.9	18.0	34.1	34.0	254	31.8	26.2	28.1	14.0	254	25.8	38.0	8.6	27.6	254



Data Filename: Drosophila_Adh.meg
Data Title: Adh-sequence data from 11 species
Amino acid Frequencies
Sites Used: All selected
All frequencies are given in percent.
Domain: Adh data (Gene: Adh)

	Ala	Cys	Asp	Glu	Phe	Gly	His	Ile	Lys	Leu	Met	Asn	Pro	Gln	Arg	Ser	Thr	Val	Trp	Tyr	Total
D. adiastola	8.27	0.79	5.51	3.54	3.94	8.27	1.18	8.66	6.69	9.84	0.79	6.30	3.94	2.36	2.76	5.12	9.45	8.66	1.97	1.97	254
D. affinidisjuncta	9.45	0.79	5.12	3.54	3.54	7.09	1.57	8.27	7.09	9.45	0.79	6.30	3.94	2.36	2.36	5.12	9.45	9.84	1.97	1.97	254
D. crassifemur	9.84	0.79	4.33	4.33	4.33	7.48	1.57	9.45	7.48	9.06	0.79	5.51	3.94	1.97	1.97	5.91	9.06	8.27	1.97	1.97	254
D. heteroneura	9.06	0.79	5.12	3.54	3.54	7.09	1.57	9.06	7.09	9.84	0.79	6.30	3.94	2.36	2.36	4.72	9.84	9.06	1.97	1.97	254
D. melanogaster	9.06	0.79	4.72	3.94	3.15	7.48	1.57	9.06	6.69	10.63	0.39	6.30	4.33	2.76	1.97	3.54	10.63	8.66	1.97	2.36	254
D. mimica	9.06	0.79	5.12	3.54	3.54	7.09	1.57	8.66	5.91	9.84	0.79	7.09	4.33	2.36	2.76	5.51	9.06	9.06	1.97	1.97	254
D. mulleri	10.24	0.79	4.33	4.33	3.54	6.69	1.57	9.45	6.30	9.06	0.79	6.69	4.72	3.15	1.97	5.12	9.06	7.87	1.97	2.36	254
D. nigra	9.06	0.79	5.12	3.94	3.54	7.09	1.57	9.45	6.30	9.45	0.79	6.69	3.94	2.36	3.15	5.12	9.06	8.66	1.97	1.97	254
D. pseudoobscura	7.87	0.79	4.72	3.94	3.15	7.48	1.57	9.06	7.09	9.45	0.39	6.69	4.33	2.76	2.36	3.94	10.63	9.45	1.97	2.36	254
S. albovittata	10.24	1.18	4.72	4.33	3.94	7.09	1.18	9.45	6.30	9.06	0.79	6.69	3.94	1.57	3.15	5.51	8.27	8.66	1.97	1.97	254
S. lebanonensis	8.66	0.79	5.12	4.33	3.94	6.30	2.76	9.45	7.09	9.84	0.39	7.09	3.94	2.36	2.76	4.72	9.06	8.27	1.97	1.18	254
Avg	9.16	0.82	4.90	3.94	3.65	7.19	1.61	9.09	6.73	9.59	0.68	6.51	4.12	2.40	2.51	4.94	9.41	8.77	1.97	2.00	254.0

Sequence stats: codon usage

In MEGA, Relative Synonymous Codon Usage (RSCU) is a statistical measure that indicates the non-random usage of synonymous codons for a specific amino acid within a gene sequence.

MEGA calculates RSCU by dividing the observed frequency of a particular codon by its expected frequency, revealing positive (RSCU > 1), negative (RSCU < 1), or random (RSCU = 1) codon usage bias.

```
Codon Usage
Code Table: Standard
Codons Used: All selected
No. of sequences used: 11
Sequences used
  D. adiastola
  D. affinidisjuncta
  D. crassifemur
  D. heteroneura
  D. melanogaster
  D. mimica
  D. mulleri
  D. nigra
  D. pseudoobscura
  S. albovittata
  S. lebanonensis
All frequencies are averages over all taxa.
Relative synonymous codon usage is given in parentheses
following the codon frequency

Domain: Adh data (Gene: Adh)
UUU(F)    3.0(0.65)  | UCU(S)    1.2(0.57)  | UAU(Y)    2.4(0.93)  | UGU(C)    0.3(0.26)  |
UUC(F)    6.3(1.35)  | UCC(S)    4.5(2.17)  | UAC(Y)    2.7(1.07)  | UGC(C)    1.8(1.74)  |
UUA(L)    0.0(0.00)  | UCA(S)    1.5(0.70)  | UAA(*)    0.0(0.00)  | UGA(*)    0.0(0.00)  |
UUG(L)    5.0(1.23)  | UCG(S)    1.6(0.78)  | UAG(*)    0.0(0.00)  | UGG(W)    5.0(1.00)  |

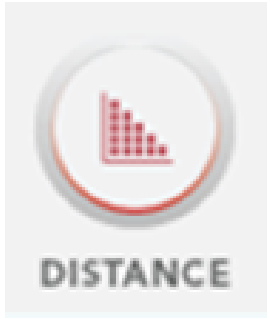
CUU(L)    2.9(0.72)  | CCU(P)    1.5(0.59)  | CAU(H)    2.0(0.98)  | CGU(R)    1.8(1.71)  |
CUC(L)    2.9(0.72)  | CCC(P)    5.3(2.02)  | CAC(H)    2.1(1.02)  | CGC(R)    3.6(3.43)  |
CUA(L)    1.1(0.27)  | CCA(P)    1.9(0.73)  | CAA(Q)    0.5(0.15)  | CGA(R)    0.0(0.00)  |
CUG(L)   12.5(3.07)  | CCG(P)    1.7(0.66)  | CAG(Q)    5.6(1.85)  | CGG(R)    0.4(0.34)  |

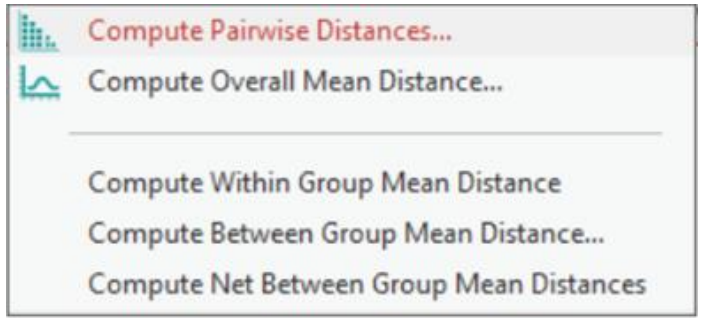
AUU(I)   10.9(1.42)  | ACU(T)    5.1(0.85)  | AAU(N)    6.7(0.81)  | AGU(S)    2.2(1.04)  |
AUC(I)   11.8(1.54)  | ACC(T)   11.5(1.93)  | AAC(N)    9.8(1.19)  | AGC(S)    1.5(0.74)  |
AUA(I)    0.4(0.05)  | ACA(T)    5.0(0.84)  | AAA(K)    3.5(0.41)  | AGA(R)    0.4(0.34)  |
AUG(M)    1.7(1.00)  | ACG(T)    2.3(0.38)  | AAG(K)   13.5(1.59)  | AGG(R)    0.2(0.17)  |

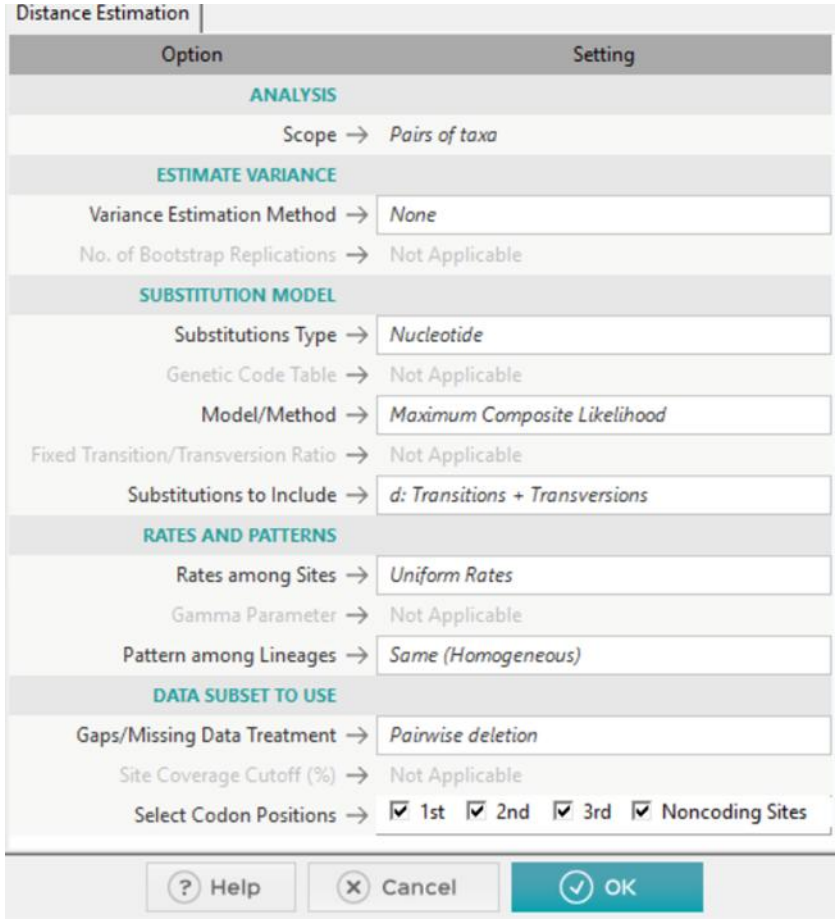
GUU(V)    6.5(1.16)  | GCU(A)    6.6(1.14)  | GAU(D)    6.2(0.99)  | GGU(G)    6.3(1.37)  |
GUC(V)    5.0(0.90)  | GCC(A)   10.8(1.86)  | GAC(D)    6.3(1.01)  | GGC(G)   10.4(2.27)  |
GUA(V)    1.9(0.34)  | GCA(A)    2.1(0.36)  | GAA(E)    2.1(0.42)  | GGA(G)    1.6(0.36)  |
GUG(V)    8.9(1.60)  | GCG(A)    3.7(0.64)  | GAG(E)    7.9(1.58)  | GGG(G)    0.0(0.00)  |

Average# codons=   254
```

Compute distances between sequences







	1	2	3	4	5	6	7	8	9	10
1. D. adiastola										
2. D. affinidisjuncta	0.0633									
3. D. crassifemur	0.2219	0.2116								
4. D. heteroneura	0.0617	0.0377	0.2244							
5. D. melanogaster	0.3147	0.3066	0.2924	0.3049						
6. D. mimica	0.0591	0.0634	0.2151	0.0547	0.2878					
7. D. mulleri	0.2082	0.2053	0.2558	0.2050	0.2563	0.2053				
8. D. nigra	0.0972	0.0970	0.2121	0.0880	0.3126	0.0792	0.2077			
9. D. pseudoobscura	0.2996	0.2880	0.2988	0.2827	0.1540	0.2762	0.2330	0.2767		
10. S. albovittata	0.2067	0.2091	0.0988	0.2041	0.3075	0.1922	0.2536	0.2006	0.2984	
11. S. lebanonensis	0.3036	0.3031	0.3264	0.3064	0.2356	0.2939	0.2520	0.2858	0.2339	0.3238

Table. Estimates of Evolutionary Divergence between Sequences

The number of base substitutions per site from between sequences are shown. Analyses were conducted using the Maximum Composite Likelihood model [1]. This analysis involved 11 nucleotide sequences. Codon positions included were 1st+2nd+3rd+Noncoding. All ambiguous positions were removed for each sequence pair (pairwise deletion option). There were a total of 762 positions in the final dataset. Evolutionary analyses were conducted in MEGA X [2].

1. Tamura K., Nei M., and Kumar S. (**2004**). Prospects for inferring very large phylogenies by using the neighbor-joining method. *Proceedings of the National Academy of Sciences (USA)* **101**:11030-11035.
2. Kumar S., Stecher G., Li M., Knyaz C., and Tamura K. (**2018**). MEGA X: Molecular Evolutionary Genetics Analysis across computing platforms. *Molecular Biology and Evolution* **35**:1547-1549.

Models for estimating distances

The evolutionary distance between a pair of sequences usually is measured by the number of nucleotide (or amino acid) substitutions occurring between them. Evolutionary distances are fundamental for the study of molecular evolution and are useful for phylogenetic reconstructions and the estimation of divergence times. Most of the widely used methods for distance estimation for nucleotide and amino acid sequences are included in *MEGA*. In the following three sections, we present a brief discussion of these methods: nucleotide substitutions, synonymous-nonsynonymous substitutions, and amino acid substitutions. Further details of these methods and general guidelines for the use of these methods are given in Nei and Kumar (2000). Note that in addition to the distance estimates, *MEGA* also computes the standard errors of the estimates using the analytical formulas and the bootstrap method.

Sequences are compared nucleotide-by-nucleotide. These distances can be computed for protein coding and non-coding nucleotide sequences.

- *No. of differences*
- *p-distance*
- ***Jukes-Cantor*** *Model*
  - *with Rate Uniformity Among Sites*
  - *with Rate Variation Among Sites*
- ***Tajima-Nei*** *Model*
  - *with Rate Uniformity and Pattern Homogeneity*
  - with Rate Variation Among Sites
  - with Pattern Heterogeneity Between Lineages
  - with Rate Variation and Pattern Heterogeneity
- ***Kimura 2-Parameter*** *Model*
  - *with Same Rate Among Sites*
  - *with Rate Variation Among Sites*
- ***Tamura 3-Parameter*** *Model*
  - *with Rate Uniformity and Pattern Homogeneity*
  - with Rate Variation Among Sites
  - with Pattern Heterogeneity Between Lineages
  - with Rate Variation and Pattern Heterogeneity
- ***Tamura-Nei*** *Model*
  - *With Rate Uniformity and Pattern Homogeneity*
  - *with Rate Variation Among Sites*
  - with Pattern Heterogeneity Between Lineages
  - with Rate Variation and Pattern Heterogeneity
- ***Log-Det*** *Method*
  - with Pattern Heterogeneity Between Lineages
- ***Maximum Composite Likelihood*** *Model*
  - with Rate Uniformity and Pattern Homogeneity
  - with Rate Variation Among Sites
  - with Pattern Heterogeneity Between Lineages
  - with Rate Variation and Pattern Heterogeneity

p-distance

This distance is the proportion (p) of nucleotide sites at which two sequences being compared are different. It is obtained by dividing the number of nucleotide differences by the total number of nucleotides compared. It does not make any correction for multiple substitutions at the same site, substitution rate, or differences in evolutionary rates among sites.

	1	2	3	4	5	6	7	8	9	10
1. D. adiastola										
2. D. affinidisjuncta	0.0604									
3. D. crassifemur	0.1903	0.1824								
4. D. heteroneura	0.0591	0.0367	0.1916							
5. D. melanogaster	0.2520	0.2454	0.2375	0.2454						
6. D. mimica	0.0564	0.0604	0.1850	0.0525	0.2336					
7. D. mulleri	0.1798	0.1772	0.2139	0.1772	0.2113	0.1772				
8. D. nigra	0.0906	0.0906	0.1824	0.0827	0.2507	0.0748	0.1798			
9. D. pseudoobscura	0.2428	0.2349	0.2415	0.2323	0.1378	0.2270	0.1969	0.2270		
10. S. albovittata	0.1785	0.1798	0.0919	0.1759	0.2480	0.1667	0.2126	0.1732	0.2402	
11. S. lebanonensis	0.2454	0.2441	0.2585	0.2467	0.1995	0.2388	0.2113	0.2336	0.1995	0.2559

Jukes-Cantor model

In the Jukes and Cantor (1969) model, the rate of nucleotide substitution is the same for all pairs of the four nucleotides. The multiple hit correction equation produces a maximum likelihood estimate of the number of nucleotide substitutions between two sequences. It assumes an equality of substitution rates among sites, equal nucleotide frequencies.

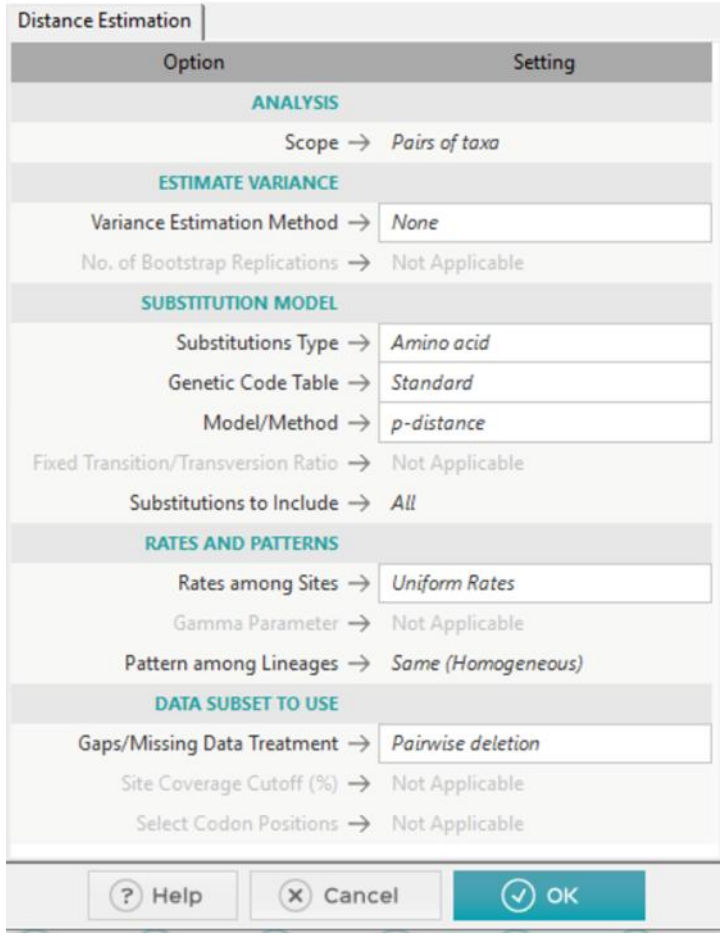
	1	2	3	4	5	6	7	8	9	10
1. D. adiastola										
2. D. affinidisjuncta	0.0629									
3. D. crassifemur	0.2195	0.2090								
4. D. heteroneura	0.0615	0.0377	0.2212							
5. D. melanogaster	0.3071	0.2972	0.2856	0.2972						
6. D. mimica	0.0587	0.0629	0.2125	0.0544	0.2799					
7. D. mulleri	0.2056	0.2021	0.2518	0.2021	0.2482	0.2021				
8. D. nigra	0.0965	0.0965	0.2090	0.0876	0.3051	0.0788	0.2056			
9. D. pseudoobscura	0.2933	0.2818	0.2914	0.2780	0.1523	0.2704	0.2283	0.2704		
10. S. albovittata	0.2038	0.2056	0.0980	0.2004	0.3012	0.1885	0.2500	0.1970	0.2895	
11. S. lebanonensis	0.2972	0.2953	0.3170	0.2992	0.2319	0.2876	0.2482	0.2799	0.2319	0.3130

Tamura-Nei model

The Tamura-Nei model (1993) corrects for multiple hits, taking into account the differences in substitution rate between nucleotides and the inequality of nucleotide frequencies. It distinguishes between transitional and transversional substitution rates. It also assumes equality of substitution rates among sites

	1	2	3	4	5	6	7	8	9	10
1. D. adiastola										
2. D. affinidisjuncta	0.0632									
3. D. crassifemur	0.2249	0.2131								
4. D. heteroneura	0.0618	0.0378	0.2256							
5. D. melanogaster	0.3171	0.3027	0.2933	0.3047						
6. D. mimica	0.0588	0.0632	0.2170	0.0546	0.2851					
7. D. mulleri	0.2110	0.2057	0.2595	0.2062	0.2514	0.2061				
8. D. nigra	0.0972	0.0974	0.2126	0.0883	0.3139	0.0793	0.2107			
9. D. pseudoobscura	0.3018	0.2886	0.2985	0.2866	0.1546	0.2762	0.2325	0.2753		
10. S. albovittata	0.2072	0.2082	0.0986	0.2025	0.3099	0.1901	0.2570	0.1990	0.2934	
11. S. lebanonensis	0.3061	0.3010	0.3239	0.3059	0.2382	0.2939	0.2555	0.2852	0.2400	0.3170

Compute distances between sequences
also for aminoacids

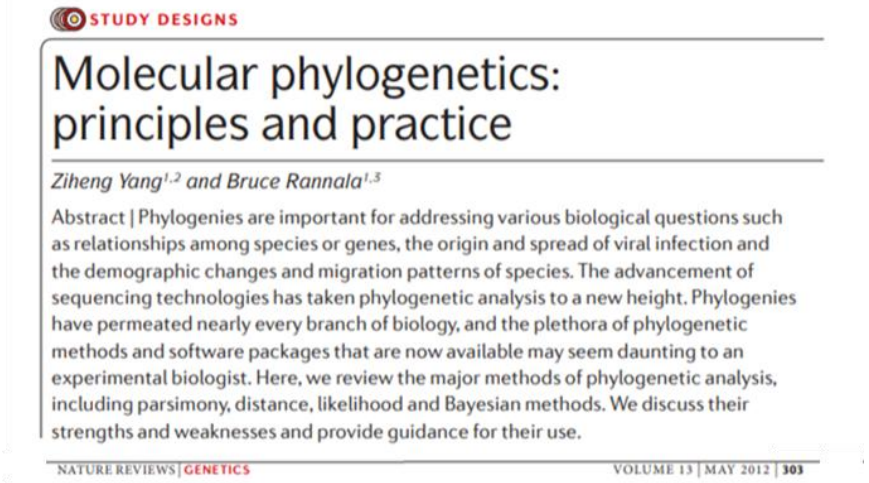


	1	2	3	4	5	6	7	8	9	10
1. D. adiastola										
2. D. affinidisjuncta	0.0394									
3. D. crassifemur	0.1378	0.1299								
4. D. heteroneura	0.0433	0.0276	0.1299							
5. D. melanogaster	0.2244	0.2087	0.2165	0.2126						
6. D. mimica	0.0551	0.0591	0.1457	0.0512	0.2205					
7. D. mulleri	0.1142	0.1063	0.1575	0.1024	0.1850	0.1102				
8. D. nigra	0.0709	0.0669	0.1457	0.0669	0.2205	0.0669	0.1220			
9. D. pseudoobscura	0.2165	0.2008	0.2165	0.1969	0.0827	0.2008	0.1811	0.2087		
10. S. albovittata	0.1535	0.1496	0.0827	0.1378	0.2087	0.1378	0.1457	0.1614	0.2165	
11. S. lebanonensis	0.2087	0.1929	0.2402	0.1969	0.1693	0.2047	0.1575	0.1969	0.1614	0.2283

Use distance estimates to establish phylogenetic relationships

Phylogenetic Inference

Reconstruction of the evolutionary history of genes and species is currently one of the most important subjects in molecular evolution. If reliable phylogenies are produced, they will shed light on the sequence of evolutionary events that generated the present day diversity of genes and species and help us to understand the mechanisms of evolution as well as the history of organisms. Phylogenetic relationships of genes or organisms usually are presented in a treelike form with a root, which is called a *rooted tree*. It also is possible to draw a tree without a root, which is called an *unrooted tree*. The branching pattern of a tree is called a *topology*. There are numerous methods for constructing phylogenetic trees from molecular data (Nei and Kumar 2000). They can be classified into *Distance methods*, *Parsimony methods*, and *Likelihood methods*. These methods are explained in Swofford et al. 1996, Li (1997), Page and Holmes (1998), and Nei and Kumar (2000).



STUDY DESIGNS

Molecular phylogenetics: principles and practice

Ziheng Yang[1,2] and Bruce Rannala[1,3]

Abstract | Phylogenies are important for addressing various biological questions such as relationships among species or genes, the origin and spread of viral infection and the demographic changes and migration patterns of species. The advancement of sequencing technologies has taken phylogenetic analysis to a new height. Phylogenies have permeated nearly every branch of biology, and the plethora of phylogenetic methods and software packages that are now available may seem daunting to an experimental biologist. Here, we review the major methods of phylogenetic analysis, including parsimony, distance, likelihood and Bayesian methods. We discuss their strengths and weaknesses and provide guidance for their use.

NATURE REVIEWS | GENETICS VOLUME 13 | MAY 2012 | 303

Phylogenetic reconstruction methods

. Distance-based (pairwise distances between sequences)

. Character-based (comparing all sequences for each letter)

- . Maximum parsimony (number of changes)
- . Maximum Likelihood (log-likelihood value)
- . Bayesian Inference (posterior probability)

Phylogenetic reconstruction methods

Table 2 | **A summary of strengths and weaknesses of different tree reconstruction methods**

Strengths	Weaknesses
Parsimony methods	
• Simplicity and intuitive appeal • The only framework appropriate for some data (such as SINES and LINES)	• Assumptions are implicit and poorly understood • Lack of a model makes it nearly impossible to incorporate our knowledge of sequence evolution • Branch lengths are substantially underestimated when substitution rates are high • Maximum parsimony may suffer from long-branch attraction
Distance methods	
• Fast computational speed • Can be applied to any type of data as long as a genetic distance can be defined • Models for distance calculation can be chosen to fit data	• Most distance methods, such as neighbour joining, do not consider variances of distance estimates • Distance calculation is problematic when sequences are divergent and involve many alignment gaps • Negative branch lengths are not meaningful
Likelihood methods	
• Can use complex substitution models to approach biological reality • Powerful framework for estimating parameters and testing hypotheses	• Maximum likelihood iteration involves heavy computation • The topology is not a parameter so that it is difficult to apply maximum likelihood theory for its estimation. Bootstrap proportions are hard to interpret
Bayesian methods	
• Can use realistic substitution models, as in maximum likelihood • Prior probability allows the incorporation of information or expert knowledge • Posterior probabilities for trees and clades have easy interpretations	• Markov chain Monte Carlo (MCMC) involves heavy computation • In large data sets, MCMC convergence and mixing problems can be hard to identify or rectify • Uninformative prior probabilities may be difficult to specify. Multidimensional priors may have undue influence on the posterior without the investigator's knowledge • Posterior probabilities often appear too high • Model selection involves challenging computation[138,139]

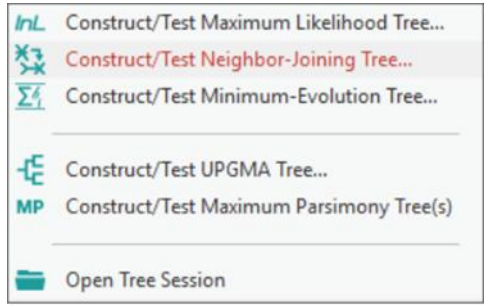
STUDY DESIGNS

Molecular phylogenetics: principles and practice

Ziheng Yang[1,2] and Bruce Rannala[1,3]

Abstract | Phylogenies are important for addressing various biological questions such as relationships among species or genes, the origin and spread of viral infection and the demographic changes and migration patterns of species. The advancement of sequencing technologies has taken phylogenetic analysis to a new height. Phylogenies have permeated nearly every branch of biology, and the plethora of phylogenetic methods and software packages that are now available may seem daunting to an experimental biologist. Here, we review the major methods of phylogenetic analysis, including parsimony, distance, likelihood and Bayesian methods. We discuss their strengths and weaknesses and provide guidance for their use.

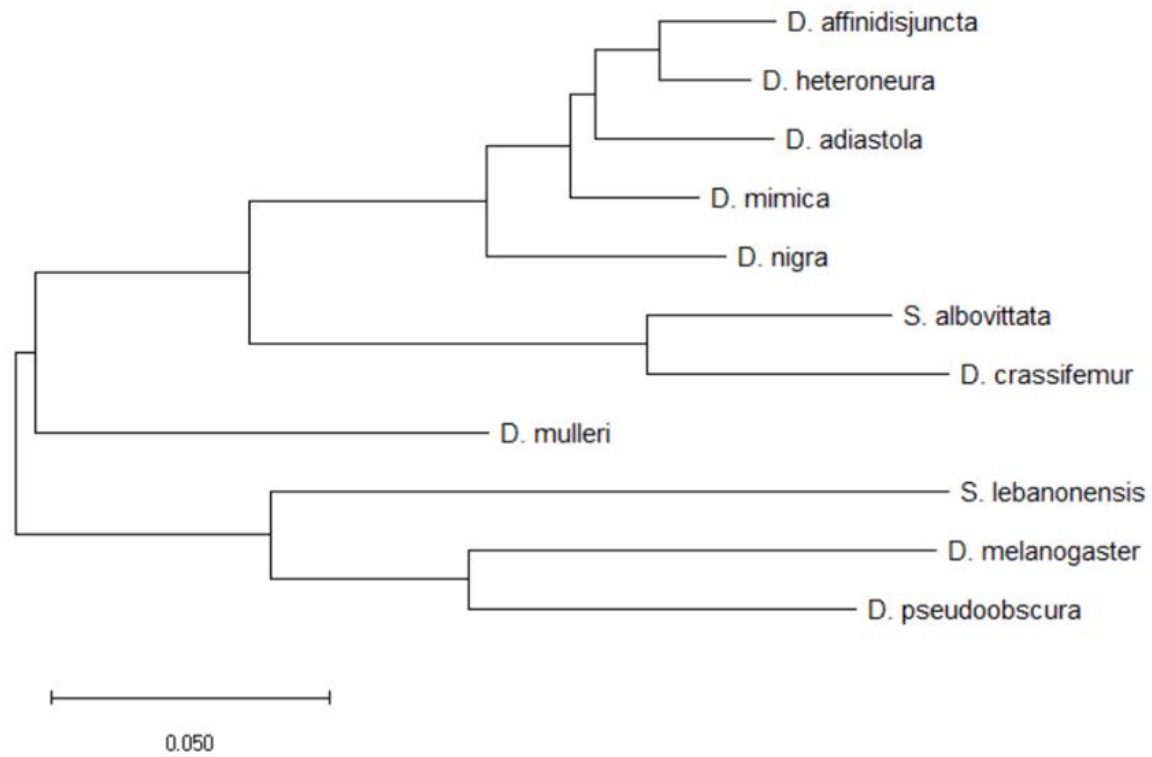




Phylogeny Reconstruction

Option	Setting
ANALYSIS	
Scope →	*All Selected Taxa*
Statistical Method →	*Neighbor-joining*
PHYLOGENY TEST	
Test of Phylogeny →	*None*
No. of Bootstrap Replications →	Not Applicable
SUBSTITUTION MODEL	
Substitutions Type →	*Nucleotide*
Genetic Code Table →	Not Applicable
Model/Method →	*Maximum Composite Likelihood*
Fixed Transition/Transversion Ratio →	Not Applicable
Substitutions to Include →	*d: Transitions + Transversions*
RATES AND PATTERNS	
Rates among Sites →	*Uniform Rates*
Gamma Parameter →	Not Applicable
Pattern among Lineages →	*Same (Homogeneous)*
DATA SUBSET TO USE	
Gaps/Missing Data Treatment →	*Pairwise deletion*
Site Coverage Cutoff (%) →	Not Applicable
Select Codon Positions →	☑ 1st ☑ 2nd ☑ 3rd ☑ Noncoding Sites
SYSTEM RESOURCE USAGE	
Number of Threads →	Not Applicable

Help Cancel OK



Evolutionary relationships of taxa

The evolutionary history was inferred using the Neighbor-Joining method [1]. The optimal tree with the sum of branch length = 0.86146990 is shown. The tree is drawn to scale, with branch lengths in the same units as those of the evolutionary distances used to infer the phylogenetic tree. The evolutionary distances were computed using the Maximum Composite Likelihood method [2] and are in the units of the number of base substitutions per site. This analysis involved 11 nucleotide sequences. Codon positions included were 1st+2nd+3rd+Noncoding. All ambiguous positions were removed for each sequence pair (pairwise deletion option). There were a total of 762 positions in the final dataset. Evolutionary analyses were conducted in MEGA X [3].

1. Saitou N. and Nei M. (**1987**). The neighbor-joining method: A new method for reconstructing phylogenetic trees. *Molecular Biology and Evolution* **4**:406-425.
2. Tamura K., Nei M., and Kumar S. (**2004**). Prospects for inferring very large phylogenies by using the neighbor-joining method. *Proceedings of the National Academy of Sciences (USA)* **101**:11030-11035.
3. Kumar S., Stecher G., Li M., Knyaz C., and Tamura K. (**2018**). MEGA X: Molecular Evolutionary Genetics Analysis across computing platforms. *Molecular Biology and Evolution* **35**:1547-1549.

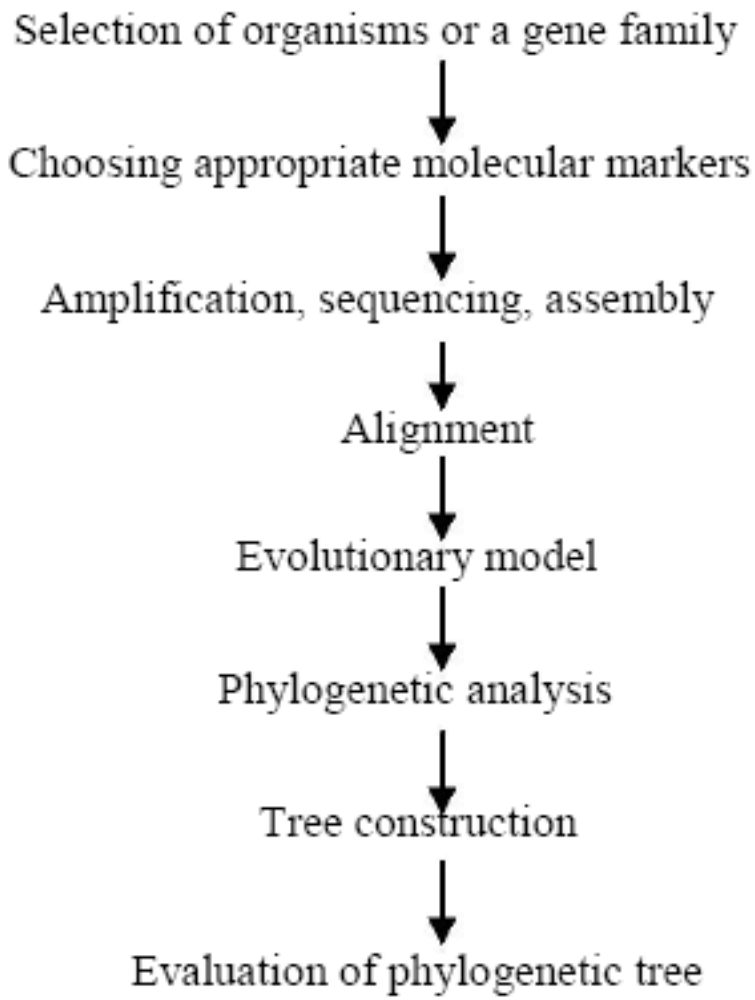


Figure 1: General steps in studying molecular phylogeny.

GA – TP – AULA 4:
Sequence Analyses

- SEQUENCE STATS
- SUBSTITUTIONS
- GENETIC DISTANCES
- MEGA
- EXERCISE

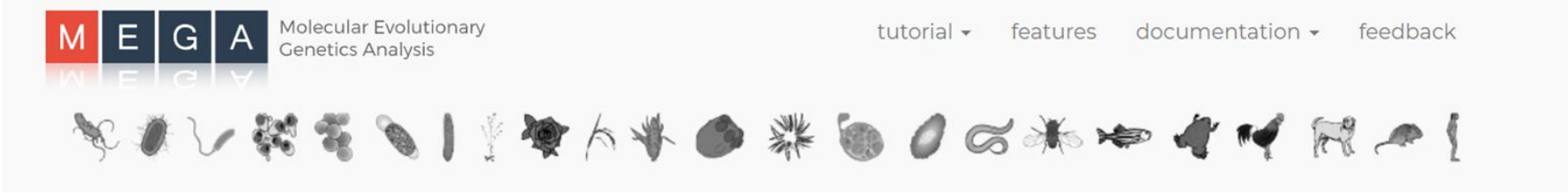
Marta Antunes
mmaantunes@ciencias.ulisboa.pt

MOLECULAR EVOLUTIONARY GENETIC ANALYSIS

Input files: nucleotide/aminoacid sequences & distance matrices

Samples: same or diferente populations/species

Analysis: stats, distances, trees, tests of selection



Sequence Analyses	Statistical Methods	Powerful Visual Tools
Phylogeny Inference	Maximum Likelihood	Alignment/Trace Editor
Model Selection	Distance Methods	Tree Explorer
Dating and Clocks	Ordinary Least Squares	Data Explorers
Ancestral States	Maximum Parsimony	Legend Generator
Selection and Tests	Composite Likelihood	Gene Duplication Wizard
Sequence Alignment	Bayesian	Timetree Wizard

https://www.megasoftware.net/

GA – TP – AULA 4:
Sequence Analyses

SEQUENCE STATS

SUBSTITUTIONS

GENETIC DISTANCES

MEGA

EXERCISE

Marta Antunes
mmaantunes@ciencias.ulisboa.pt

Sequence Analyses in MegaX

Moodle > GA > TP > TP04_exercise

1. Visualization of sequence data
2. Sequence statistics
3. Evolutionary distances



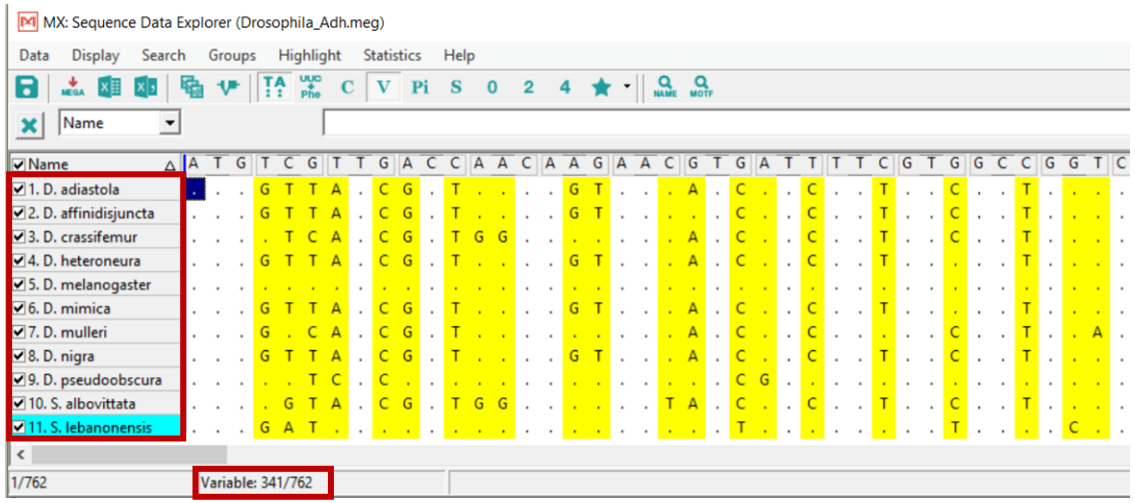
https://www.megasoftware.net/

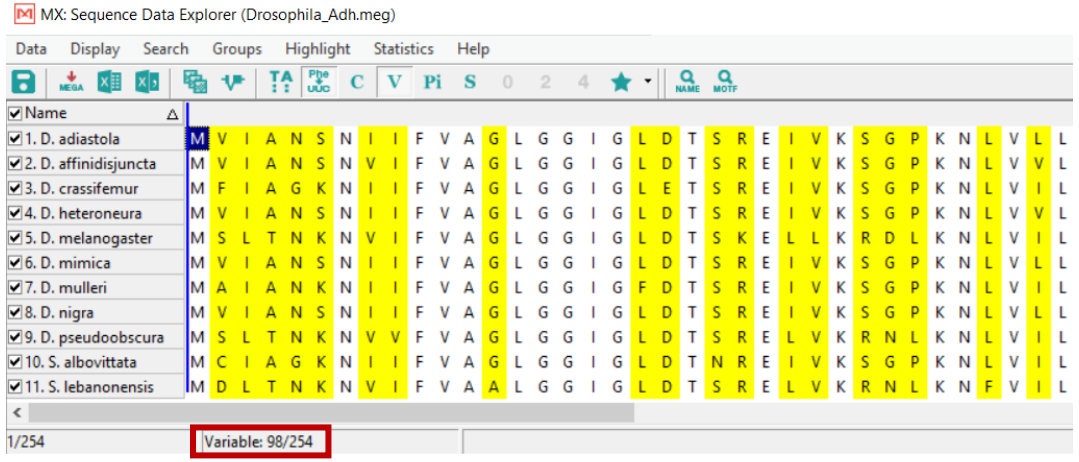
Questions:

1.1. How many sequences? What are they?

1.2. How long are they in nr of nucleotides? How long is the peptide translation?

1.3. What proportion of nucleotides is variable? What proportion of aminoacids is variable?





1.1. There are 11 sequences, corresponding to the Adh gene of each of 11 Drosophila/Sophophora species

1.2. For each species, the sequence is 762 nucleotides long and its translation corresponds to 254 aminoacids.

1.3. There are 341/762 variable nucleotides and 98/254 variable aminoacids.

Questions:

2.1. Which species have the sequences with the highest and lowest % of Ts? What are those frequencies?

2.2. What are the average nucleotide frequencies for this group of sequences?

2.3. What is the most and least common aminoacids, on average?

2.4. What species have the highest and lowest frequencies of Asparagine (Asn)? What are those frequencies?

2.5. Of the alternative codons for proline (P), which are the most and least common for these 11 sequences?

2.1. Highest % T: *D. albovittata* (26.5%). Lowest % T: *D. melanogaster* (19.4%).

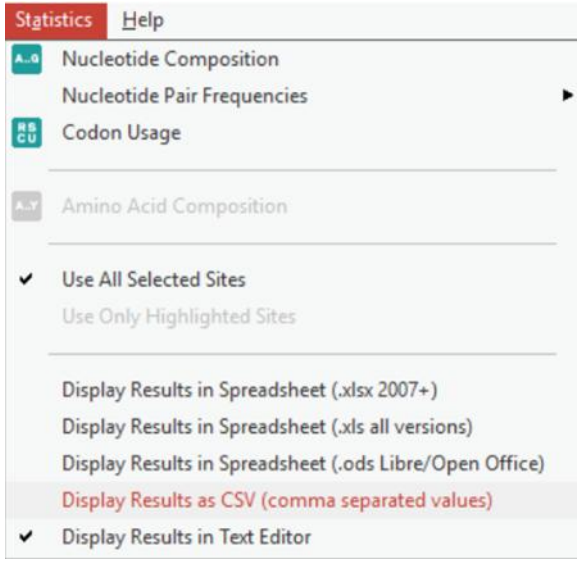
2.2. Average nucleotide frequencies: 23.8% T, 27.4% C, 23.6% A, 25.2% G.

2.3. Least common: Met (0.68%). More common: Leu (9.59%).

2.4. Highest Asn: *D. mimica* = *S. lebanonensis* (7.09%). Lowest Asn: *D. crassifemur* (5.51%).

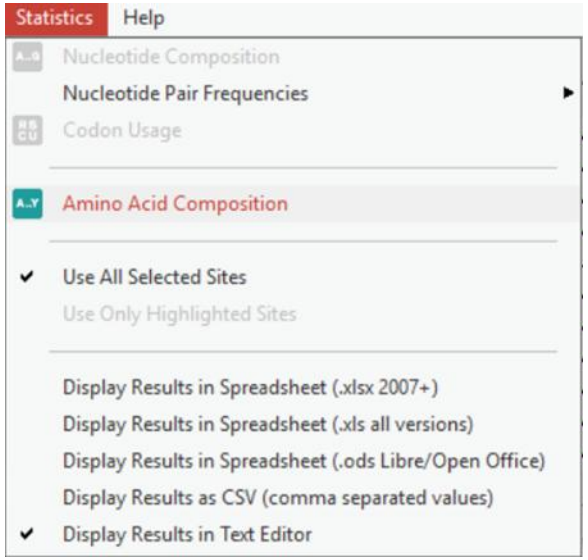
2.5. Most common: CCC (5.3%). Least common: CCU (1.5%).

Sequence stats: nucleotide composition / aminoacid composition



Data Filename: Drosophila_Adh.meg
Data Title: Adh-sequence data from 11 species
Nucleotide Frequencies
Sites Used: All selected
All frequencies are given in percent.
Domain: Adh data (Gene: Adh)

	T(U)	C	A	G	Total	T-1	C-1	A-1	G-1	Pos #1	T-2	C-2	A-2	G-2	Pos #2	T-3	C-3	A-3	G-3	Pos #3
D. adiastola	25.5	25.6	24.0	24.9	762	13.8	18.1	33.9	34.3	254	31.9	24.8	27.6	15.7	254	30.7	33.9	10.6	24.8	254
D. affinidisjuncta	24.4	25.7	24.8	25.1	762	13.4	17.7	33.9	35.0	254	31.9	26.0	28.0	14.2	254	28.0	33.5	12.6	26.0	254
D. crassifemur	23.9	26.5	23.1	26.5	762	15.4	16.5	33.9	34.3	254	31.9	27.2	27.2	13.8	254	24.4	35.8	8.3	31.5	254
D. heteroneura	24.9	26.0	24.5	24.5	762	13.8	17.7	34.6	33.9	254	32.3	26.0	28.0	13.8	254	28.7	34.3	11.0	26.0	254
D. melanogaster	19.4	32.4	21.9	26.2	762	12.6	19.7	33.9	33.9	254	31.9	26.8	28.3	13.0	254	13.8	50.8	3.5	31.9	254
D. mimica	24.9	26.4	23.6	25.1	762	13.0	18.9	34.3	33.9	254	31.9	25.6	27.6	15.0	254	29.9	34.6	9.1	26.4	254
D. mulleri	21.0	29.9	23.9	25.2	762	13.8	18.9	33.9	33.5	254	30.7	27.6	28.7	13.0	254	18.5	43.3	9.1	29.1	254
D. nigra	24.9	26.8	24.0	24.3	762	14.2	17.7	34.3	33.9	254	31.9	25.6	28.0	14.6	254	28.7	37.0	9.8	24.4	254
D. pseudoobscura	21.7	30.6	22.0	25.7	762	13.4	18.9	34.3	33.5	254	31.5	26.4	29.1	13.0	254	20.1	46.5	2.8	30.7	254
S. albovittata	26.5	24.1	24.3	25.1	762	15.4	15.7	33.9	35.0	254	31.9	26.4	26.8	15.0	254	32.3	30.3	12.2	25.2	254
S. lebanonensis	24.9	27.3	23.5	24.3	762	14.2	18.5	34.6	32.7	254	31.9	25.6	29.9	12.6	254	28.7	37.8	5.9	27.6	254
Avg.	23.8	27.4	23.6	25.2	762	13.9	18.0	34.1	34.0	254	31.8	26.2	28.1	14.0	254	25.8	38.0	8.6	27.6	254



Data Filename: Drosophila_Adh.meg
Data Title: Adh-sequence data from 11 species
Amino acid Frequencies
Sites Used: All selected
All frequencies are given in percent.
Domain: Adh data (Gene: Adh)

	Ala	Cys	Asp	Glu	Phe	Gly	His	Ile	Lys	Leu	Met	Asn	Pro	Gln	Arg	Ser	Thr	Val	Trp	Tyr	Total
D. adiastola	8.27	0.79	5.51	3.54	3.94	8.27	1.18	8.66	6.69	9.84	0.79	6.30	3.94	2.36	2.76	5.12	9.45	8.66	1.97	1.97	254
D. affinidisjuncta	9.45	0.79	5.12	3.54	3.54	7.09	1.57	8.27	7.09	9.45	0.79	6.30	3.94	2.36	2.36	5.12	9.45	9.84	1.97	1.97	254
D. crassifemur	9.84	0.79	4.33	4.33	4.33	7.48	1.57	9.45	7.48	9.06	0.79	5.51	3.94	1.97	1.97	5.91	9.06	8.27	1.97	1.97	254
D. heteroneura	9.06	0.79	5.12	3.54	3.54	7.09	1.57	9.06	7.09	9.84	0.79	6.30	3.94	2.36	2.36	4.72	9.84	9.06	1.97	1.97	254
D. melanogaster	9.06	0.79	4.72	3.94	3.15	7.48	1.57	9.06	6.69	10.63	0.39	6.30	4.33	2.76	1.97	3.54	10.63	8.66	1.97	2.36	254
D. mimica	9.06	0.79	5.12	3.54	3.54	7.09	1.57	8.66	5.91	9.84	0.79	7.09	4.33	2.36	2.76	5.51	9.06	9.06	1.97	1.97	254
D. mulleri	10.24	0.79	4.33	4.33	3.54	6.69	1.57	9.45	6.30	9.06	0.79	6.69	4.72	3.15	1.97	5.12	9.06	7.87	1.97	2.36	254
D. nigra	9.06	0.79	5.12	3.94	3.54	7.09	1.57	9.45	6.30	9.45	0.79	6.69	3.94	2.36	3.15	5.12	9.06	8.66	1.97	1.97	254
D. pseudoobscura	7.87	0.79	4.72	3.94	3.15	7.48	1.57	9.06	7.09	9.45	0.39	6.69	4.33	2.76	2.36	3.94	10.63	9.45	1.97	2.36	254
S. albovittata	10.24	1.18	4.72	4.33	3.94	7.09	1.18	9.45	6.30	9.06	0.79	6.69	3.94	1.57	3.15	5.51	8.27	8.66	1.97	1.97	254
S. lebanonensis	8.66	0.79	5.12	4.33	3.94	6.30	2.76	9.45	7.09	9.84	0.39	7.09	3.94	2.36	2.76	4.72	9.06	8.27	1.97	1.18	254
Avg	9.16	0.82	4.90	3.94	3.65	7.19	1.61	9.09	6.73	9.59	0.68	6.51	4.12	2.40	2.51	4.94	9.41	8.77	1.97	2.00	254.0

Sequence stats: codon usage

Relative Synonymous Codon Usage is a measure of codon bias, indicating whether some synonymous codons are preferred over others in a given gene or genome.

=1.0 all synonymous codons for an aa are used equally

>1.0 overrepresentation of that codon

< 1.0 underrepresentation of that codon

```
Codon Usage
Code Table: Standard
Codons Used: All selected
No. of sequences used: 11
Sequences used
  D. adiastola
  D. affinidisjuncta
  D. crassifemur
  D. heteroneura
  D. melanogaster
  D. mimica
  D. mulleri
  D. nigra
  D. pseudoobscura
  S. albovittata
  S. lebanonensis
All frequencies are averages over all taxa.
Relative synonymous codon usage is given in parentheses
following the codon frequency

Domain: Adh data (Gene: Adh)
UUU(F)   3.0(0.65) | UCU(S)   1.2(0.57) | UAU(Y)   2.4(0.93) | UGU(C)   0.3(0.26) |
UUC(F)   6.3(1.35) | UCC(S)   4.5(2.17) | UAC(Y)   2.7(1.07) | UGC(C)   1.8(1.74) |
UUA(L)   0.0(0.00) | UCA(S)   1.5(0.70) | UAA(*)   0.0(0.00) | UGA(*)   0.0(0.00) |
UUG(L)   5.0(1.23) | UCG(S)   1.6(0.78) | UAG(*)   0.0(0.00) | UGG(W)   5.0(1.00) |

CUU(L)   2.9(0.72)   CCU(P)   1.5(0.59)   CAU(H)   2.0(0.98) | CGU(R)   1.8(1.71) |
CUC(L)   2.9(0.72)   CCC(P)   5.3(2.02)   CAC(H)   2.1(1.02) | CGC(R)   3.6(3.43) |
CUA(L)   1.1(0.27)   CCA(P)   1.9(0.73)   CAA(Q)   0.5(0.15) | CGA(R)   0.0(0.00) |
CUG(L)  12.5(3.07)   CCG(P)   1.7(0.66)   CAG(Q)   5.6(1.85) | CGG(R)   0.4(0.34) |

AUU(I)  10.9(1.42) | ACU(T)   5.1(0.85) | AAU(N)   6.7(0.81) | AGU(S)   2.2(1.04) |
AUC(I)  11.8(1.54) | ACC(T)  11.5(1.93) | AAC(N)   9.8(1.19) | AGC(S)   1.5(0.74) |
AUA(I)   0.4(0.05) | ACA(T)   5.0(0.84) | AAA(K)   3.5(0.41) | AGA(R)   0.4(0.34) |
AUG(M)   1.7(1.00) | ACG(T)   2.3(0.38) | AAG(K)  13.5(1.59) | AGG(R)   0.2(0.17) |

GUU(V)   6.5(1.16) | GCU(A)   6.6(1.14) | GAU(D)   6.2(0.99) | GGU(G)   6.3(1.37) |
GUC(V)   5.0(0.90) | GCC(A)  10.8(1.86) | GAC(D)   6.3(1.01) | GGC(G)  10.4(2.27) |
GUA(V)   1.9(0.34) | GCA(A)   2.1(0.36) | GAA(E)   2.1(0.42) | GGA(G)   1.6(0.36) |
GUG(V)   8.9(1.60) | GCG(A)   3.7(0.64) | GAG(E)   7.9(1.58) | GGG(G)   0.0(0.00) |

Average# codons=   254
```

Questions:

3.1. Which species pairs have the sequences that are closest and furthest away based on each of the three models for nucleotide substitutions and on the aminoacid distance estimates?

3.2. In what general way do the distinct distance estimates differ?

3.1. For all cases, highest distance between *S. lebanonensis* and *D. crassifemur*, and lowest distance between *D. heteroneura* and *D. affinidisjuncta*.

3.2. No change in terms of the pairs that are closest and more distance (which would mean same species grouping in phylogenetic tree), but change in the actual distance values (which would impact size of branches in a phylogenetic tree): Tamura-Nei > Jukes-Cantor > p-distances.

p-distance

This distance is the proportion (p) of nucleotide sites at which two sequences being compared are different. It is obtained by dividing the number of nucleotide differences by the total number of nucleotides compared. It does not make any correction for multiple substitutions at the same site, substitution rate, or differences in evolutionary rates among sites.

	1	2	3	4	5	6	7	8	9	10
1. D. adiastola										
2. D. affinidisjuncta	0.0604									
3. D. crassifemur	0.1903	0.1824								
4. D. heteroneura	0.0591	0.0367	0.1916							
5. D. melanogaster	0.2520	0.2454	0.2375	0.2454						
6. D. mimica	0.0564	0.0604	0.1850	0.0525	0.2336					
7. D. mulleri	0.1798	0.1772	0.2139	0.1772	0.2113	0.1772				
8. D. nigra	0.0906	0.0906	0.1824	0.0827	0.2507	0.0748	0.1798			
9. D. pseudoobscura	0.2428	0.2349	0.2415	0.2323	0.1378	0.2270	0.1969	0.2270		
10. S. albovittata	0.1785	0.1798	0.0919	0.1759	0.2480	0.1667	0.2126	0.1732	0.2402	
11. S. lebanonensis	0.2454	0.2441	0.2585	0.2467	0.1995	0.2388	0.2113	0.2336	0.1995	0.2559

Jukes-Cantor model

In the Jukes and Cantor (1969) model, the rate of nucleotide substitution is the same for all pairs of the four nucleotides. The multiple hit correction equation produces a maximum likelihood estimate of the number of nucleotide substitutions between two sequences. It assumes an equality of substitution rates among sites, equal nucleotide frequencies.

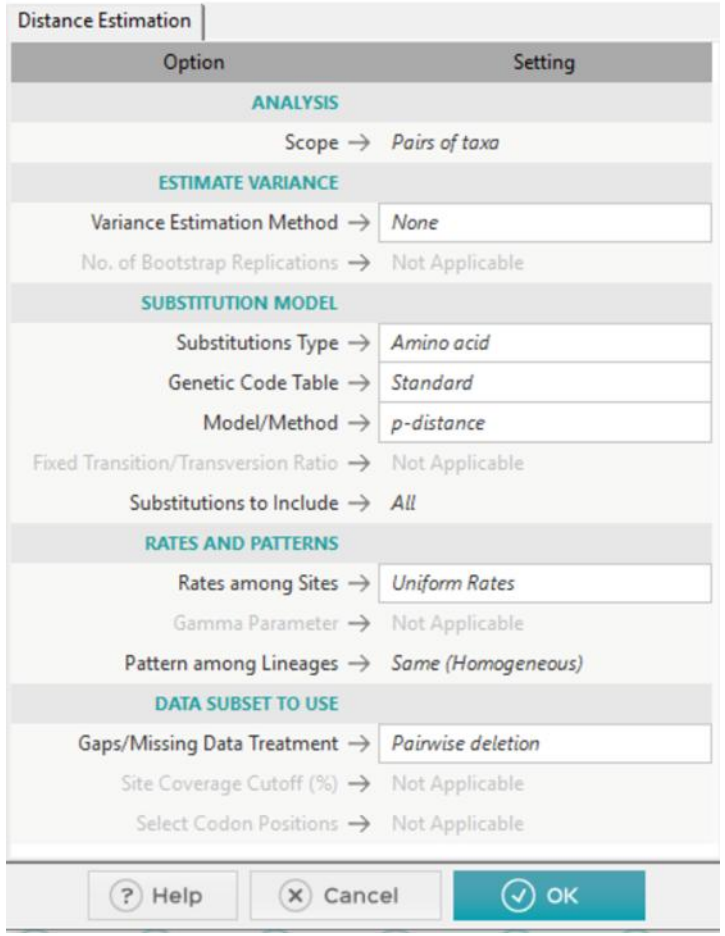
	1	2	3	4	5	6	7	8	9	10
1. D. adiastola										
2. D. affinidisjuncta	0.0629									
3. D. crassifemur	0.2195	0.2090								
4. D. heteroneura	0.0615	0.0377	0.2212							
5. D. melanogaster	0.3071	0.2972	0.2856	0.2972						
6. D. mimica	0.0587	0.0629	0.2125	0.0544	0.2799					
7. D. mulleri	0.2056	0.2021	0.2518	0.2021	0.2482	0.2021				
8. D. nigra	0.0965	0.0965	0.2090	0.0876	0.3051	0.0788	0.2056			
9. D. pseudoobscura	0.2933	0.2818	0.2914	0.2780	0.1523	0.2704	0.2283	0.2704		
10. S. albovittata	0.2038	0.2056	0.0980	0.2004	0.3012	0.1885	0.2500	0.1970	0.2895	
11. S. lebanonensis	0.2972	0.2953	0.3170	0.2992	0.2319	0.2876	0.2482	0.2799	0.2319	0.3130

Tamura-Nei model

The Tamura-Nei model (1993) corrects for multiple hits, taking into account the differences in substitution rate between nucleotides and the inequality of nucleotide frequencies. It distinguishes between transitional and transversional substitution rates. It also assumes equality of substitution rates among sites

	1	2	3	4	5	6	7	8	9	10
1. D. adiastola										
2. D. affinidisjuncta	0.0632									
3. D. crassifemur	0.2249	0.2131								
4. D. heteroneura	0.0618	0.0378	0.2256							
5. D. melanogaster	0.3171	0.3027	0.2933	0.3047						
6. D. mimica	0.0588	0.0632	0.2170	0.0546	0.2851					
7. D. mulleri	0.2110	0.2057	0.2595	0.2062	0.2514	0.2061				
8. D. nigra	0.0972	0.0974	0.2126	0.0883	0.3139	0.0793	0.2107			
9. D. pseudoobscura	0.3018	0.2886	0.2985	0.2866	0.1546	0.2762	0.2325	0.2753		
10. S. albovittata	0.2072	0.2082	0.0986	0.2025	0.3099	0.1901	0.2570	0.1990	0.2934	
11. S. lebanonensis	0.3061	0.3010	0.3239	0.3059	0.2382	0.2939	0.2555	0.2852	0.2400	0.3170

Compute distances between sequences
also for aminoacids



	1	2	3	4	5	6	7	8	9	10
1. D. adiastola										
2. D. affinidisjuncta	0.0394									
3. D. crassifemur	0.1378	0.1299								
4. D. heteroneura	0.0433	0.0276	0.1299							
5. D. melanogaster	0.2244	0.2087	0.2165	0.2126						
6. D. mimica	0.0551	0.0591	0.1457	0.0512	0.2205					
7. D. mulleri	0.1142	0.1063	0.1575	0.1024	0.1850	0.1102				
8. D. nigra	0.0709	0.0669	0.1457	0.0669	0.2205	0.0669	0.1220			
9. D. pseudoobscura	0.2165	0.2008	0.2165	0.1969	0.0827	0.2008	0.1811	0.2087		
10. S. albovittata	0.1535	0.1496	0.0827	0.1378	0.2087	0.1378	0.1457	0.1614	0.2165	
11. S. lebanonensis	0.2087	0.1929	0.2402	0.1969	0.1693	0.2047	0.1575	0.1969	0.1614	0.2283