Molecular markers in plant systematics and population biology

7. DNA sequencing (cpDNA)

Tomáš Fér
tomas.fer@natur.cuni.cz
DNA sequencing

• detection the order of nucleotides in a DNA strand

...ATATATAGGCAAGGAATCTCTATTATTAAATCATT...

• use the information to model evolutionary and population genetic processes

• make hypothesis about similarity and relationships among taxa
Sequencing principle

• PCR with a primer pair
  • amplification of the target region
• cycle sequencing (dideoxy, Sanger)
  • use of one primer only
  • dNTP as well as ddNTP are present in the mixture
  • produce fragments differing exactly by one base
• electrophoretic separation of fragments in the gel
  • automated sequencer
2′, 3′- dideoxy NTPs

3′-CTGGACTGCA-5′
5′-GACCT
Cycle sequencing

3´-TACG-5´
5´-ATGCATGC-3´

primer

template

ddGTP

ddCTP

ddATP

ddTTP

GTACG
ATGCATGC

CGTACG
ATGCATGC

ACGTACG
ATGCATGC

TACGTACG
ATGCATGC

[Graphical representation of the sequencing process with nucleotides and colors]

[Genetic sequence analysis results graph]
Automated sequencer

ABI (Applied Biosystems) – gel, capillary systems (up to 96)
Genome structure

• genetic information – order of nucleotides (ACGT)

• coding regions – exons – *conserved*

• non-coding regions – introns, spacers – *variable*

• nuclear, chloroplast and mitochondrial genome
Sequence evolutionary rate
Types of variability in DNA sequences

- 5bp indel
- Point mutations (SNPs)
Descriptive statistics for DNA sequences

<p>| | | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>ATTGCCACCCCTAGGCTA</td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>-----G------------</td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>-----G---------A---</td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>-------------A---GA-</td>
<td></td>
</tr>
<tr>
<td>5</td>
<td>-------------AT--GA-</td>
<td></td>
</tr>
</tbody>
</table>

L  sequence length  18

Π  average number of nucleotide difference  2.8

S  number of segregating sites  5

π  nucleotide diversity  0.155

s  number of segregating sites per site  0.277
Synonymous vs non-synonymous mutations

1. ATT GCC ACC CCT AGG CTA
   Ile Ala Thr Pro Arg Leu

2. --- G --- --- --- ---
   Cys

3. --- C --- --- A ---
   Cys Pro

4. --- --- --- A --- GA-
   Pro Glu

5. --- --- --- A T GA-
   Pro Trp Glu

synonymous (silent)  non-synonymous
Chloroplast genome

- many genes are *single-copy* (only 1 copy in the whole genome)
- conserved evolution of the chloroplast genome
  - disadvantage when studying intraspecific or population variability
  - many conserved regions can be used as *priming sites*
- structural rearrangements of chloroplast genome
  - mainly on larger evolutionary scale
  - inversion – e.g., 30kb inversion differentiates bryophytes and higher plants
  - extensive deletions
  - loss of specific genes and intrones
- *chloroplast capture*
  - chloroplast transfer from one species to another by introgression
  - can influence phylogeny in a wrong way (when not recognized)
Chloroplast genome

- 4 rRNAs
- 30-11 tRNAs
- 21 ribosomal proteins (rps)
- 4 RNA polymerase subunits (rpo)
- 28 thylakoid proteins (ps)
- rbcL (large RuBisCO subunit)
- 11 proteins similar to NADH (ndh)
Chloroplast genome

Genome alignment highlighting diagnostic changes among land plant plastomes. Coloured boxes – genome homology segments, horizontal white box – a copy of IR.
Frequently sequenced cpDNA regions

+ many others...
• gene for large subunit of ribuloso-1,5-bisphosphate-carboxylase/oxygenase (RUBISCO)
• 1,428, 1,431 or 1,434 bp in length – indels are extremely rare
• one of the first sequenced genes
• very conserved, systematics at family or generic level, in some groups at species level
**atpB**

- gene coding beta subunit of ATP synthase
- 1,497 bp in length, indels not found
- similar use as *rbcL*

**ndhF**

- codes a subunit of chloroplast NADH-dehydrogenase
- 2,233 bp in length (tobacco)
- about 2x more substitutions then *rbcL*
- for generic level
\textbf{matK}

- Gene coding maturase (splicing of plastid genes)
- About 1,550 bp in length – low number of indels
- Systematics at family and generic level
**trnL intron and spacer between trnL and trnF**

- tRNA genes – secondary structure
- accumulation of insertions/deletions with the same rate as nucleotide substitutions
- alignment problems, especially in distant organisms (sometimes already at family level)
- suitable for systematics of (closely) related species
atpB-rbcL

• spacer of about 900-1,000 bp in length
• systematics at family and generic level
Variable non-coding cpDNA regions

Variable non-coding cpDNA regions

• another 13 regions

• top 13 regions within each major evolutionary lineage
Use of chloroplast sequences

- phylogeny of large groups
- among-species relationships within a genus
- within-species phylogeography (haplotype definition)
- hybridization – inference of the maternal taxon (individual) – cpDNA maternally inherited in angiosperms
Viridiplantae plastid phylogeny

Gitzendanner et al. (2018)
78 coding plastid genes
1827 taxa + 52 outgroups
Relationships among species

Capsicum
atpB-rbcL spacer
Walsh & Hoot (2001)
Inter-specific hybridization

incongruence between cpDNA and nDNA

Persicaria
matK, psbA-trnH, trnL-trnF
versus ITS
Kim & Donoghue 2008
Data analysis

- **multiple alignment**

<table>
<thead>
<tr>
<th></th>
<th>ATATATATATAGGCAAGGAATCTCTATTATAATCATTTAGAATCCATA</th>
</tr>
</thead>
<tbody>
<tr>
<td>S206</td>
<td>ATATATATA--GGCAAGGAATCTCTATTATAATCATTTAGAATCCATA</td>
</tr>
<tr>
<td>S207</td>
<td>ATATATATA--GGCAAGGAATCTCTATTATAATCATTTAGAATCCATA</td>
</tr>
<tr>
<td>S208</td>
<td>ATATATATA--GGCAAGGAATCTCTATTATAATCATTTAGAATCCATA</td>
</tr>
<tr>
<td>S209</td>
<td>ATATATATA--GGCAAGGAATCTCTATTATAATCATTTAGAATCCATA</td>
</tr>
<tr>
<td>S210</td>
<td>ATATATATA--GGCAAGGAATCTCTATTATAATCATTTAGAATCCATA</td>
</tr>
<tr>
<td>S0G3</td>
<td>ATATATATA--GGCAAGGAATCTCTATTATAATCATTTAGAATCCATA</td>
</tr>
<tr>
<td>TL</td>
<td>ATATATATATAGCAGGAATCTCTATTATAATCATTTAGAATCCATA</td>
</tr>
</tbody>
</table>

- **construction of phylogenetic tree**
  - distance methods
  - maximum parsimony (MP)
  - maximum likelihood (ML)
  - Bayesian inference (BI)
Maximum parsimony (MP)

- cladistic method
- search for the simplest tree (most parsimonious tree)
- i.e., tree in which the evolution is explained by minimum number of substitutions
- software
  - PAUP *
    Phylogenetic Analysis Using Parsimony
    (* and other methods)
  - TNT
    Tree Analysis Using New Technology
Maximum likelihood (ML)

- search for tree with the highest probability (likelihood – L)
- probability that observed sequences evolved under given tree topology (and under given evolutionary model)
- software GARLI, PhyML, RAxML, PAML...
Evolutionary models for DNA sequences

- models for sequence changes

- parameters
  - base frequencies
  - substitution types (transitions, transversions)
  - heterogeneity in substitution rates (G)
  - proportion of invariant sites (I)
Substitution models

**JC** (Jukes-Cantor 1969)
- same substitution rates
- same base frequencies

**K2P** (Kimura 2 parameter 1980)
- two different substitution rates
- same base frequencies

**F81** (Felsenstein 1981)
- same substitution rates
- different base frequencies

**HKY** (Hasegawa, Kishino & Yano 1985)
- two different substitution rates
- different base frequencies

**GTR** (General time-reversible model) (Tavaré et al. 1986)
- six different substitution rates
- different base frequencies
Which model to select?

- **MODELTEST**: A tool to select the best-fit model of nucleotide substitution (Posada et al.)
- testing different models – selecting the simplest that sufficiently explain the data using
  - hierarchical likelihood ratio tests (hLRTs)
  - Akaike information criterion (AIC)
- **jModelTest2** (https://code.google.com/p/jmodeltest2/)
Saturation

- signal and noise in the data
- corrected versus uncorrected distance
- skewness (g₁-statistics), I_{SS}
Molecular clock

- **strict (global)**
  - *clocklike evolution*

- **local**

- **relaxed clocks**
  - autocorrelated (closely related taxa have similar mutation rates)
  - uncorrelated (lognormal, exponential)

- **calibration**
  - substitution rates from another study or generally assumed rate (e.g., for cpDNA)
  - fossils
  - biogeography

- **software**
  - BEAST (Bayesian), r8s (non-parametric rate smoothing, penalized likelihood), ...

Estimates of divergence times
(BEAST – Bayesian Evolutionary Analysis Sampling Trees)

Gene banks – databases of sequences

- **GenBank**
  National Centre for Biotechnology Information (NCBI)

- **EMBL**
  European Bioinformatics Institute (EBI)
  http://www.ebi.ac.uk/embl/
**LOCUS**: JQ409881  
**562 bp**  
**DNA**  
**linear**  
**PLN 31-DEC-2012**

**DEFINITION**  
Curcuma ecomata voucher JLS 73353 18S ribosomal RNA gene, partial sequence; internal transcribed spacer 1 and 5.8S ribosomal RNA gene, complete sequence; and internal transcribed spacer 2, partial sequence.

**ACCESSION**: JQ409881  
**VERSION**: JQ409881.1  
**KEYWORDS**: .

**SOURCE**  
Curcuma ecomata

**ORGANISM**  
Curcuma ecomata

Eukaryota; Viridiplantae; Streptophyta; Embryophyta; Tracheophyta; Spermatophyta; Magnoliophyta; Liliopsida; Zingiberales; Zingiberaceae; Curcuma.

**REFERENCE**  
1 (bases 1 to 562)

**AUTHORS**  
Zaveska, E., Fer, T., Sida, O., Krak, K., Man, Leong-Skornickova, J.

**JOURNAL**  
Unpublished

**REFERENCE**  
2 (bases 1 to 562)

**AUTHORS**  
Zaveska, E., Fer, T., Sida, O., Krak, K., Man, Leong-Skornickova, J.

**JOURNAL**  
Direct Submission

Submitted (17-JAN-2012) Department of Botany, Faculty of Science, Benatska 2, Prague, Faculty of Science, Benatska 2, Czech Republic

**FEATURES**  

<table>
<thead>
<tr>
<th>LOCATION/QUALIFIERS</th>
<th>DESCRIPTION</th>
</tr>
</thead>
<tbody>
<tr>
<td>1..562</td>
<td>source</td>
</tr>
<tr>
<td>/organism=&quot;Curcuma ecomata&quot;</td>
<td></td>
</tr>
<tr>
<td>/mol_type=&quot;genomic DNA&quot;</td>
<td></td>
</tr>
<tr>
<td>/specimen_voucher=&quot;JLS 73353&quot;</td>
<td></td>
</tr>
<tr>
<td>/db_xref=&quot;taxon:252240&quot;</td>
<td></td>
</tr>
<tr>
<td>1..8</td>
<td>rRNA</td>
</tr>
<tr>
<td>/note=&quot;authority: Curcuma ecomata Craib&quot;</td>
<td></td>
</tr>
<tr>
<td>9..179</td>
<td>misc_RNA</td>
</tr>
<tr>
<td>/product=&quot;18S ribosomal RNA&quot;</td>
<td></td>
</tr>
<tr>
<td>180..344</td>
<td>rRNA</td>
</tr>
<tr>
<td>/note=&quot;ITS 1&quot;</td>
<td></td>
</tr>
<tr>
<td>345..562</td>
<td>misc_RNA</td>
</tr>
<tr>
<td>/product=&quot;internal transcribed spacer 1&quot;</td>
<td></td>
</tr>
<tr>
<td>/product=&quot;internal transcribed spacer 2&quot;</td>
<td></td>
</tr>
</tbody>
</table>

**ORIGIN**

```
1 cattgttgg aagagataga atgtgatgat atgtgatgat tgtgaagcgc acccttttctgt
dg3 tggctggtgg cctgtactac ggctgcctgc actaaggaac aatgaactgt
gcagtacggt cgtgacgagct acgtggtgaag atctgctagc aatgctgtagt cttttttgttgc
```
Population study

Systematic study


