8. DNA sequencing II. – nrDNA, low-copy markers

Tomáš Fér
tomas.fer@natur.cuni.cz
Nuclear genome

• many genes in more copies (*multiple-copy*)
  • homology problem – we do not know what we sequence/observe
  • e.g., genes for rRNA

• *low-copy* or *single-copy* genes
  • problem with primers for the studied group
  • genes for specific proteins – *Adh, Tpi, Pgi*, phytochrome *c*, *waxy* (GBSSI)…
rDNA

• genes for rDNA – commonly used marker in systematics
• hundreds to thousands of tandem repeats (250-2,500 in *Arabidopsis thaliana*, up to 22,000 in *Vicia faba*)
• about 5% of total DNA
• in one or few chromosomal loci
• 45S rDNA locus - transcribed region (ETS-18S-ITS1-5.8S-ITS2-26S) is separated by intergenic spacer (IGS)
• 5S rDNA locus – linked with 45S locus (L-type – streptophyte algae, bryophytes, lycophytes) or separated (S-type arrangement – gymnosperms, angiosperms)
• *concerted evolution* – creates intragenomic uniformity of repeat units
  – if proceeds slowly – several different ITS sequences within genome exists (paralogs) → phylogeny inference problematic, take care about hybridization and polyploidization
rDNA structure

ETS – external transcribed spacer

ITS – internal transcribed spacer

IGS – intergenic spacer (NTS – non-transcribed spacer)
rDNA on chromosomes

Organizations of 45S rDNA and 5S rDNA loci on metaphase chromosomes

Monkheang et al. 2016
• ITS1 (200-300 bp) – greater length variation than ITS2
• ITS2 (180-240 bp)
• frequently used to detect relationships among closely related genera and at the species level (but sometimes low variability)
• have certain function when forming ribosomal units
• i.e., some evolutionary constrain of structure and sequence does exits
  • 40% of ITS2 conserved among all angiosperms
  • 50% of ITS2 is possible to align across family and higher level
• much longer in gymnosperms, high length variability (1,550-3,125 bp in Pinaceae)
ETS – external transcribed spacer

- at least same evolutionary rate as ITS
- 258-635 bp
- problem with sequencing – missing conserved region at 5′-end of the spacer
- *long-distance* PCR for amplification of the whole IGS (using universal primers from 18S and 26S DNA)
- after sequencing of the product from 3′-end it is possible to design internal primers
18S rDNA
• about 1,800 bp
• length mutations often just 1 bp, at particular sites
• i.e., simple alignment

26S rDNA
• between 3,375 and 3,393 bp
• very conserved
• evolves 1.6-2.2x faster than 18S rDNA
• contains conserved regions and expansion segments
• *conserved core regions* – systematics at higher taxonomic level
• *expansion segments* – evolves up to $10\times$ faster
rRNA – model of secondary structure

- **loops and stems**
- Watson-Crick pairing (A-U and G-C, but often G-U as well)
- **CBC - compensatory base changes** (substitutions to hold stem structure)
- i.e., bases are not independent – different weight than *loop*
Secondary structure of RNA

- prediction – RNA folding (Mfold, Zucker 2003)
- databases
  - European ribosomal RNA database
  - ITS2
ITS secondary structure

rDNA markers, ITS

**pros**
- many copies – easily sequenced
- universal primers
- favorite marker – many sequences in databases
- variable (ITS)
- biparentally inherited (useful for parent identification of a hybrid)

**cons**
- multiple-copy marker
- more copies after hybridization/polyploidization (if concerted evolution is slow) – cloning necessary
- sometimes not enough variable (even ITS) in closely related species
Low-copy nuclear markers

- genes that are only in several copies in the genome
- × multiple copy – hundreds to ten thousands copies (nrDNA...)
- higher variability than ITS and non-coding cpDNA (at least some of them)
- homology problem
  paralogy × orthology × homeology
Paralogue, orthologue and homeologue genes

**Orthologue genes**
- originated by speciation

**Paralogue genes**
- originated by gene duplication

**Homeologue genes**
- originated by polyploidization

- orthologue genes
  1A-1B-1C
  1A-2A, 1B-2B, 1C-2C
  1A-2B, 1A-2C atd...

- paralogue genes
  1A-2B, 1A-2C atd...

- homeologue genes
  1A˚-1C˚, 2A˚-2C˚
Low-copy markers

**pros**
- higher evolutionary rate than organelar sequences
- possibility to use many independent (unrelated) loci
- biparental inheritance

**cons**
- complex genetic structure (gene duplication)
- difficult identification of orthologous loci
- within-species, within-population and within-individual variability (heterozygosity)
Evolutionary rate variation

- synonymous substitutions – 5× faster than cpDNA genes and 20× faster than mtDNA
- e.g., relationships within *Gossypium* – cotton (Small et al. 1998)
  - 7,000 bases of non-coding cpDNA provided incomplete and poorly supported resolution
  - 1,650 bases of *AdhC* – complete and robust resolution
- great differences in variability (mutation rate) among different genes – up to 7× differences
- i.e., it is necessary to test more markers for each group and select the variable loci
Evolutionary rate variation

Total gene tree depth in synonymous (dS) and non-synonymous (dN) substitutions per site for protein-coding genes in three genomic compartments across Bryophyta.

Parsimony informative sites

Structure of eucaryote genes

- **5′ UTR (untranslated region)** – promoters for gene regulation (conserved), sometimes includes highly variable introns
- **exons** – more conserved at non-synonymous positions (first and second codon position), at synonymous (third codon) positions similar to non-coding regions
- **introns** – fewer functional constraints at the sequence level, often limited length
- **3′ UTR (untranslated region)** – controls maturation of mRNA and addition of poly-A signal, but often highly variable as well
- different functions, various evolutionary constraints
Multiple unlinked loci

- independent reconstructions of evolution
- markers at different chromosomes (or distant enough from each other at a single chromosome) – evolutionary independent
- incongruence among different markers – possibility to detect, i.e., hybridization, introgression, polyploidization, ILS (incomplete lineage sorting)
Biparental inheritance

- Low-copy markers are less often a subject of concerted evolution

- i.e., they are ideal candidates for identification of parental genomes in putative hybrids or polyploids
Gene families

• multiple copies of homologous genes originating by duplication
• gene families differs in copy number
  • single copy – GBSSI in diploid *Poaceae*
  • hundreds of copies – actine, small heat-shock proteins
• gene and whole-genome duplication (and consequent loss of genes) – dynamic and ongoing process
• characteristics of a gene family – taxon (group) specific
• gene family characteristics in one group need not to be applied in another group
  • *Adh* generally – 1 to 3 loci
  • in *Gossypium* or *Pinus* – up to 7 loci
• wrong characterisation of gene family leads to erroneous phylogeny reconstruction (it is always necessary to compare orthologous copies!)
Study of ortholog sequences

1. design of *universal primers* – amplification of more PCR products (of different length) → characterisation of the gene family (identification the number of loci?)

2. design of *locus-specific primers* – only orthologous sequences are amplified

- evidence of orthology
  - overall sequence similarity (orthologs are mutually more similar than paralogs)
  - *expression pattern* – orthologous sequences share the same pattern
- great differences in the variability among genes and loci – preliminary study for detection of sufficient variability is necessary
Intraspecific variability

- allelic variability within and among populations
- *coalescence within species* – alleles evolved within a single species – it does not violate correct phylogeny inference, i.e., this is a useful variability for within-species studies – population, phylogeographic etc.
- *deep coalescence* – allelic variability exceeds species boundary, i.e., some alleles are more related to alleles from a different species rather than to other alleles from the same species – more probable in species with high population sizes
- loci under balanced selection (maintain high allelic variability) – unsuitable for phylogeny reconstruction
  - e.g., self-incompatibility genes in *Solanaceae* – allelic variability exceeds species and even generic boundaries
- also due to hybridization and introgression
Recombination

1. *allelic recombination*
   - recombination at the individual locus level
   - generates allelic variability
     - do not violate the assumption of bifurcate relationships among alleles
     - introduce a reticulate evolution
     - do not violate correct phylogeny reconstruction as far as alleles are monophyletic within a species

2. *non-homologous recombination*
   - recombination among paralogous loci
   - can be rare or specific for particular gene
Concerted evolution

- common in highly repetitive loci (nrDNA)
- exists in low-copy markers as well – non-homologous recombination

- *do not occur* → sequencing of all genes from the gene family produces orthology-paralogy tree (OP-tree)
- *completed* (assumed in nrDNA) → sequencing of whichever gene from the gene family produces the correct phylogenetic tree
- *incomplete* → mixture of orthologous and incompletely homogenised paralagous sequences, i.e., correct phylogenetic reconstruction is practically impossible
Concerted evolution

- common in highly repetitive loci (nrDNA)
- exists in low-copy markers – non-homologous recombination
- do not occur → sequencing of all genes from the gene family produces orthology-paralogy tree (OP-tree)
- completed (assumed in nrDNA) → sequencing of whichever gene from the gene family produces the correct phylogenetic tree
- incomplete → mixture of orthologous and incompletely homogenised paralogous sequences, i.e., correct phylogenetic reconstruction is practically impossible
PCR-mediated recombination

- *in vitro* non-homologous recombination
- sources
  - 1. template interchange during PCR
  - 2. incompletely amplified copies of one locus serve as primers for amplification of paralogous locus
- degree depends on
  - degree of sequence similarity among paralogous loci
  - universality/specificity of primers
  - PCR conditions (optimization of annealing temperature, product length and extension time)
How to determine suitable nuclear markers for phylogenetic analyses

1. candidate gene selection (and representative taxa) for preliminary study
2. candidate gene isolation from representative taxa
3. assessment of orthology among isolated sequences
4. determination of relative rate of sequence evolution – selection of suitable locus
5. generating sequences from all studied taxa for the particular locus
Candidate gene selection

- no general assumptions of universality for any particular gene
- not necessary to use frequently sequenced genes (*Adh*...)
- little known or even anonymous nuclear loci is possible to use
- where to start?
  - previous studies within the group
  - literature search for the utility of a gene at given taxonomical level in different groups
  - GenBank, EMBL... – taxon × gene name combination search
  - BLAST search – search for similar sequences → primer design, gene structure identification (exons, introns)...
  - NGS sources – transcriptome, genome skimming

Incongruencies among markers: gene trees vs species tree

- gene duplications and losses (orthology problem)
- incomplete lineage sorting/deep coalescence
- hybridization
- polyploidization
- recombination
Species tree estimation

- concatenation

- multispecies coalescence
  - coestimation of gene trees and species tree
  - summary methods

(more in the Hyb-Seq lesson...)
Application of single-copy genes

- phylogenetic studies – can provide enough variability for the full resolution at lower taxonomical level (e.g., relationships among closely related species)
- study of polyploids – ‘picking’ individual parental sequences from the polyploid genome and identification of complex allopolyploid pattern
- phylogeography – variability within species
Relationships among closely related species

Gaertnera
PepC, Tpi
Malcomber 2002
Allohexaploid origin

Elymus repens

GBSSI (single-copy)
Mason-Gamer 2008
Parental genomes within allopolyploids

Cerastium
4x, 8x, 12x

RPB2 tree, network
Brysting et al. 2007
Parental genomes within allopolyploids

Asplenium
4x, 8x
LFY, 2\textsuperscript{nd} intron, cpDNA
Shepherd et al. 2008
Phylogeography with low-copy gene

Armeria pungens
GapC
Pineiro et al. 2008
Systematic study

Literature

- Wu F. et al. (2006): *Combining bioinformatics and phylogenetics to identify large sets of single-copy orthologous genes (COSII) for comparative, evolutionary and systematic studies: a test case in the Euasterid plant clade*. Genetics 174: 1407-1420