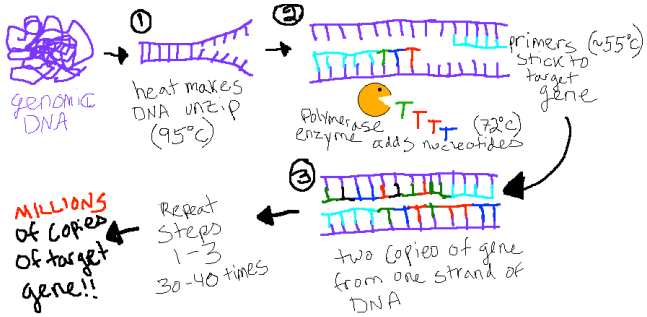
Pokročilé metody hodnocení sekvencí DNA a multilokusových dat

1. Analýza sekvenačních dat - I

- úprava alignmentu **[Mafft, BioEdit, MEGA,]**
- detekce rekombinantů **[Splitstree, GARD, RDP3]**
- testování vhodných modelů evoluce sekvencí **[PAUP, Modeltest, jModeltest, PartitionFinder]**
- testování fylogenet. signálu v datech a saturace sekvencí **[G-blocks, SiteStripper, SOAP, Tree-Puzzle]**

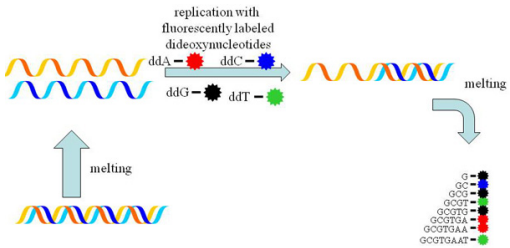
- **praktická část - příprava sekvenačních dat pro fylogenetickou analýzu** na příkladovém souboru dat

Sekvenování DNA

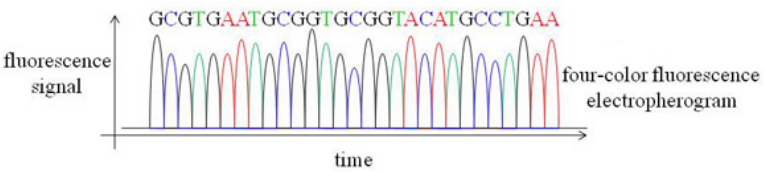


- určení pořadí nukleotidů v řetězci DNA
- potřeba **specifických primerů pro PCR** amplifikaci sekvenovaného úseku

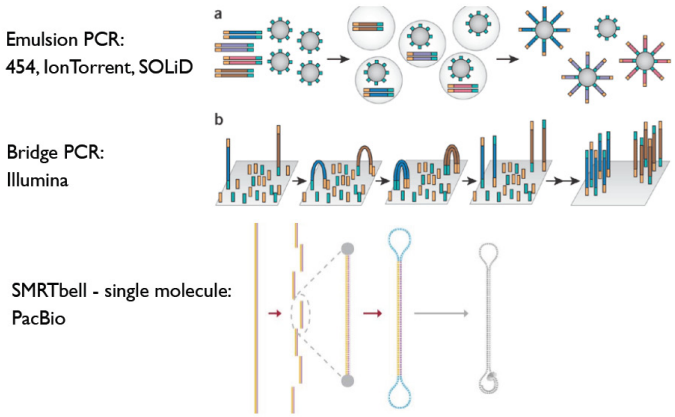
- **klasicky** - využití automatických sekvenátorů–fluorescenční značení bazí



ABI prism 3130



- **moderně** - různé metody tzv. “Next generation sequencing” [454, Illumina,...]
- velké množství dat, stává se výhodné cenově i časově



Illumina HiSeq

454 GS Junior

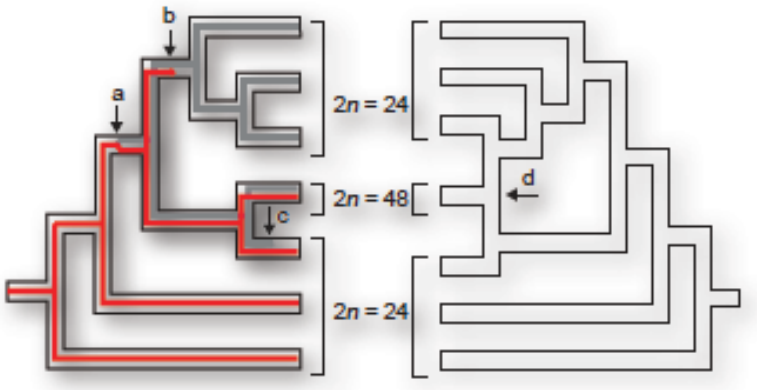
454 GS FLX+

PacBio RS

K čemu jsou dobrá sekvenační data?

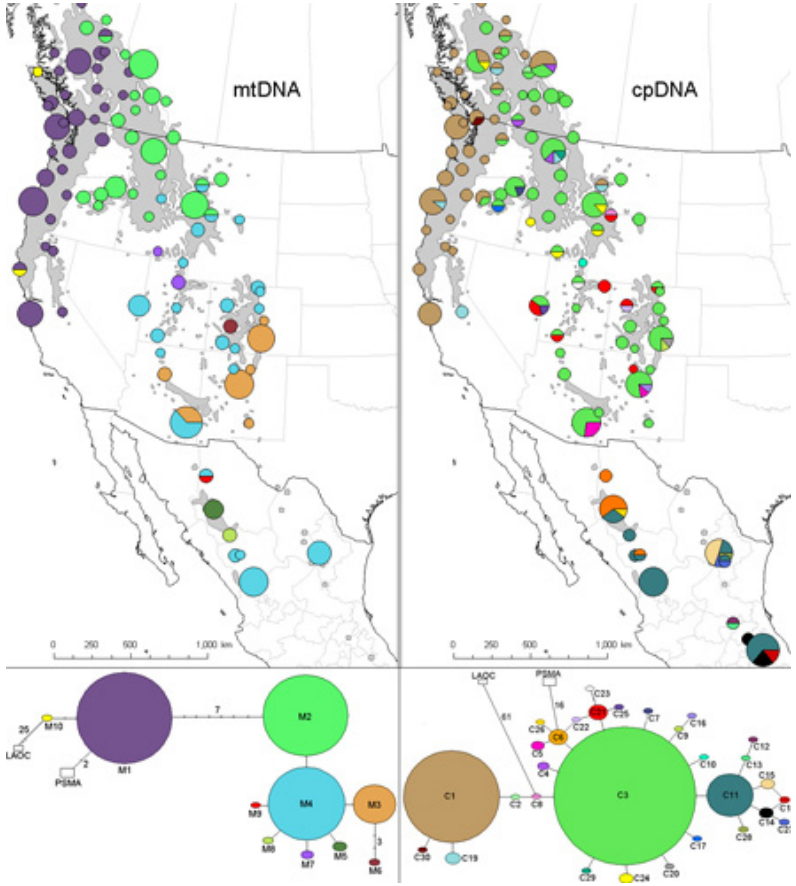
- rekonstrukce evoluce a systematika na různých úrovních (kódující vs. nekódující úseky, nDNA vs. cpDNA)
- mezidruhové vztahy v rámci rodu
- vnitrodruhová fylogeografie (definice haplotypů)
- hybridizace - zjištění mateřského/otcovského taxonu (cpDNA haploty vs. jaderné sekvence)

EVOLUCE



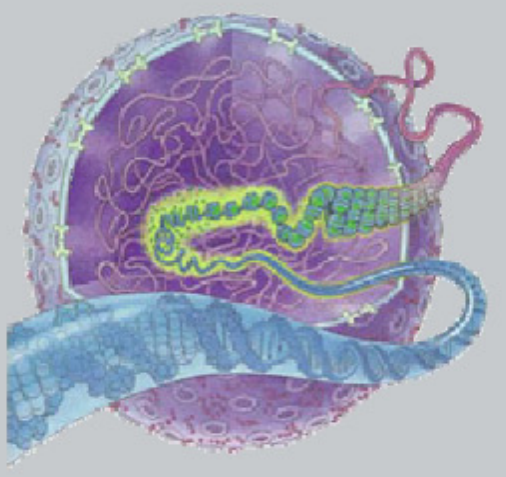
TAXONOMIE

FYLOGEOGRAFIE



Co a proč sekvenujeme?

jádro	plastidy	mitochondrie
• různá ploidie	• 1 kruhová molekula	• kruhová molekula
• jedna nebo více kopií genu		
• rekombince	• bez rekombince	• strukturní přestavby
• biparentální přenos	• uniparentální přenos	

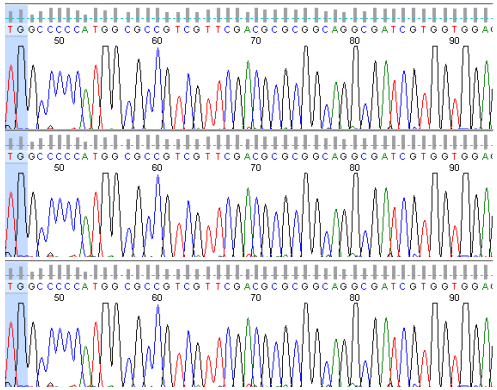


Charakteristika genomů

	nDNA zvířata	rostliny	cpDNA	mtDNA zvířata	rostliny
dědičnost	biparentální	biparentální	krytosemené maternální, konifery paternální	maternální	maternální
struktura	lineární	lineární	cirkulární	cirkulární	cirkulární, komplexní
velikost (kb)	$4.9 \times 10^4 – 7.0 \times 10^8$	$5.0 \times 10^4 – 3.0 \times 10^8$	71 – 214	15 – 20	200 – 2400
substituční rychlost	3.5×10^{-9}	$4.1 – 5.7 \times 10^{-9}$	$0.86 – 1.20 \times 10^{-9}$	56×10^{-9}	$0.36 – 0.50 \times 10^{-9}$
substituční rychlost relativně vůči rostlinné mtDNA	8.1	11.4	2.4	130.2	1.0
cizorodé sekvence	běžné	běžné	vzácné	vzácné	běžné
strukturní mutace	běžné	běžné	vzácné	vzácné	běžné
rekombinace	ano	ano	intramolekulární	ne	inter- a intramolekulární

Lowe et al., 2004

když už jsme se rozhodli, co budeme sekvenovat a máme data...



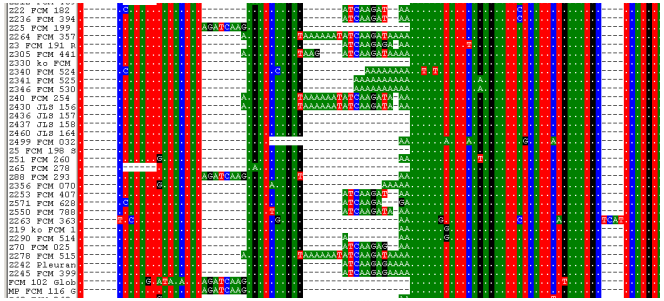
FASTA file

*.fas

```
>73477recli4 _ A

CATTGTTGAGAGAGCACAGAATA-
ATGGATGATTGTGAATGTGT-
GAACGTGACCCTTTCGTTCGGTC-
GAAGAGCGGGTAGTCGGTAATCGTC-
GAGCACGATGGACGTTGGTCGTCGC-
GAAC

>73477recli5 _ A

CATTGTTGAGAGAGCACAGAATA-
ATGGATGATTGTGAATGTGT-
GAACGTGACCCTTTCGTTCGGTC-
GAAGAGCGGGTAGTCGGTAATCGTC-
GAGCACGATGGACGTTGGTCGTCGC-
GAAC

>73477recli3 _ R

CATTGTTGAGAGAGCACAGAAT-
GATGGATGATTGTGAATGTGTG-
GAATCAAATGACTCTCGGCAATG-
GATATCTCGGCTCTTGCATCGAT-
GAAGAACGTAGTG
```

ALIGNMENT

*.fas, *.aln



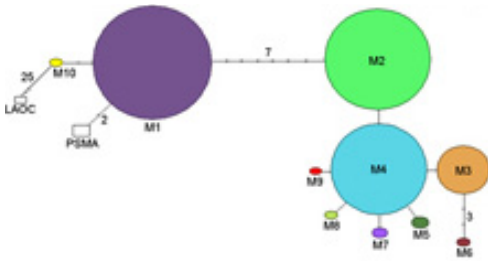
manuální editace ALIGNMENTu

detekce REKOMBINANTů

*.nexus, *.phy, *.fas, ...

haplotypové sítě

Maximum parsimony



BLAST

struktura studovaného úseku - exon vs. intron

testování modelů evoluce DNA

testování fylogenetického signálu

saturace sekvencí

rekonstrukce fylogenetických vztahů pomocí

ML a Bayesovských metod

Alignment a jeho editace

- **tvorba alignmentu** = “zarovnání” primárních sekvenačních dat uložených (nejběžněji) ve FASTA formátu
  - různé programy - **mafft**, ClustalX, Muscle
- **editace** alignmentu - PROČ ?
  - odstranění “šumu” v primárních sekvencích (chyby polymerázy, kvalita sekvencí, vícenásobný signál)
  - struktura sekvenovaného úseku (např. detekce hranic exonů a intronů)
  - kódování indelů
- **editace** alignmentu - JAK?
  - zodpovědně :)
  - porovnání s primárními daty ze sekvenátoru

```
>73477recli4 _ A

CATTGTTGAGAGAGCACAGAATAATGGATGATTGTGAATGTGTGAACGT-
GACCCTTTCGTTCGGTCGAAGAGCGGGTAGTCGGTAATCGTCGAGCAC-
GATGGACGTTGGTCGTCGCGAAC

>73477recli5 _ A

CATTGTTGAGAGAGCACAGAATAATGGATGATTGTGAATGTGTGAACGT-
GACCCTTTCGTTCGGTCGAAGAGCGGGTAGTCGGTAATCGTCGAGCAC-
GATGGACGTTGGTCGTCGCGAAC

>73477recli3 _ R

CATTGTTGAGAGAGCACAGAATGATGGATGATTGTGAATGTGTGGAAT-
CAAATGACTCTCGGCAATGGATATCTCGGCTCTTGCATCGATGAAGAAC-
GTAGTG
```

Literatura

- Popp & al., 2005
- Simmons & Ochoterena, 2000

PROGRAMY

- Seed - odkazy na web??
- Mafft
- BioEdit, MEGA
- SeqState

Manuální editace alignmentu

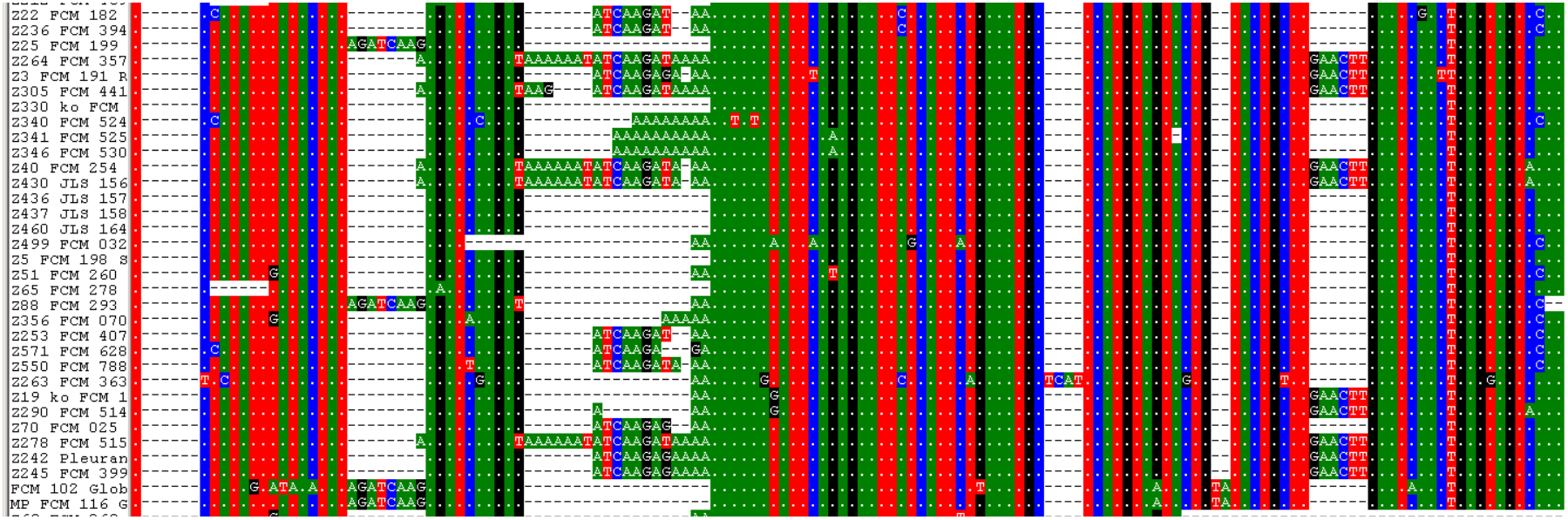
- **inserce-delece (tzv. indel)**

-> gap “-” v ML a MrBayes = missing data

-> pro MP lze kódovat jako pátý znak, nebo podle jiné zákonitosti (např. simple-indel coding)

- **chyba polymerázy nebo autapomorfie?**

-> může vnášet zbytečný šum

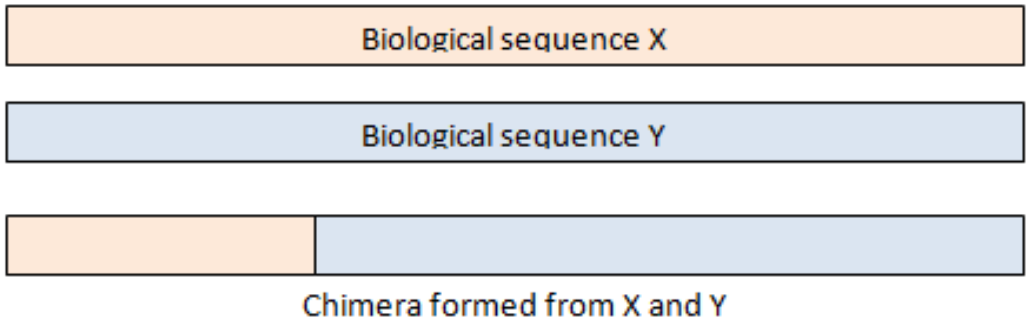


- **hypervariabilní a poly úseky**

-> lepší odstranit celý blok

Detekce rekombinantů

- kde a proč vznikají rekombinanti?
  - **In vitro** - PCR rekombinace - formování chimeických sekvencí z rozdílných templátů DNA
  - **In vivo** - intergenomické interakce po sjednocení odlišných genomů ve společném jádře
  - nejčastěji ve vícekopiových genech s nedokončenou concerted evolution (např. ITS)
- proč nám vadí? - vnáší šum do analyzovaného datasetu, podobně jako hybridi
- jak je najít? - vizuální inspekce alignmentu, programy
- co s nimi? - odstranit, případně analyzovat dva separátní datasety (bez a s rekombinanty)



Literatura

- Kosakovsky Pond et al., 2006,
- Martin et al., 2005
- Anthony et al. 2007
- Russell et al., 2010
- Posada and Crandall, 2001

PROGRAMY a užitečné odkazy

- Splitstree http://www.splitstree.org/
- GARD http://www.datamonkey.org/
- RDP3 http://web.cbio.uct.ac.za/~darren/rdp.html
- http://sequenceconversion.bugaco.com/converter/biology/sequences/fasta_to_nexus.php

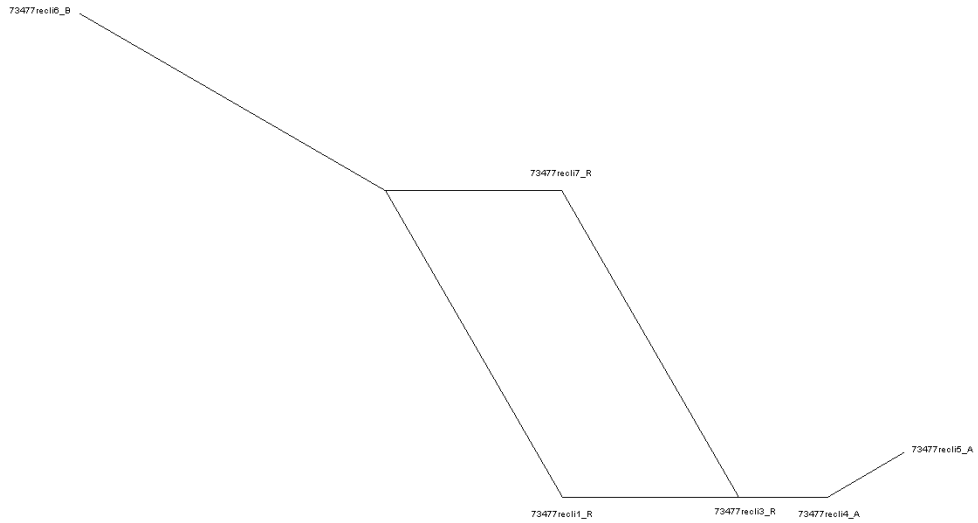
Detekce rekombinantů - “od oka” + Splitstree

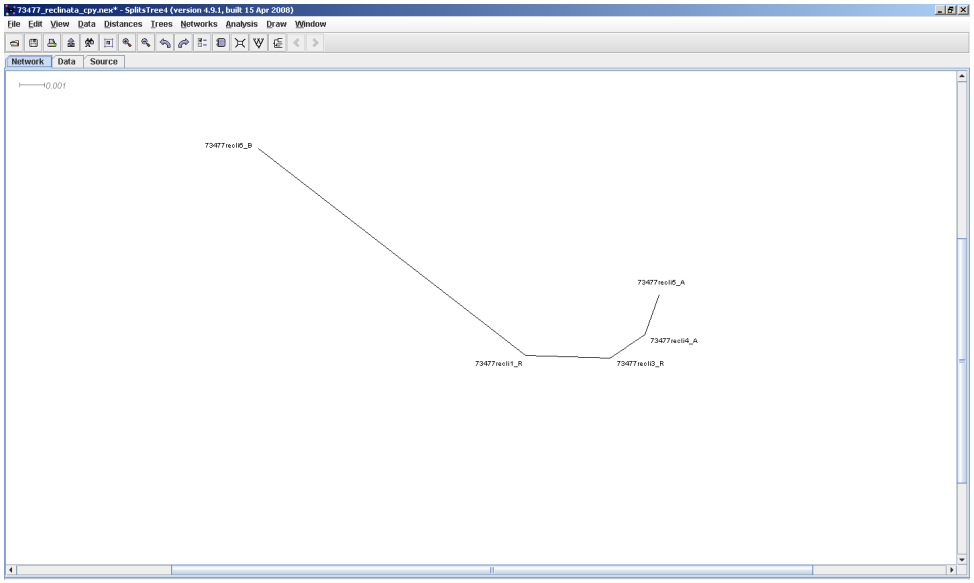
- modelová data - sekvence ITS (multi-copy charakter) - jedinci s intraindividuální variabilitou sekvencí
- nutné klonování - během opakovaných PCR i připrozeně v rámci genomu - vznikají rekombinace
- např. intraindividuální variabilita v rámci 6x jedince 73477 - pozorujeme 3-5 alel - které jsou původní?

```
73477recli4  A  ACGGA-GATGCGG
73477recli5  A  ....T-.......
73477recli3  R  G....-.......
73477recli1  R  G....-....TA.
73477recli6  B  GAAA.CAGCATAA
73477recli7  R  G.AA.CAG.....
```

Splitstree - odhaluje protichůdné informace v rámci datasetu

- “hybrid” vizualizován jako vrchol kosočtverce
- po odstranění “hybrida” - lineární struktura/strom

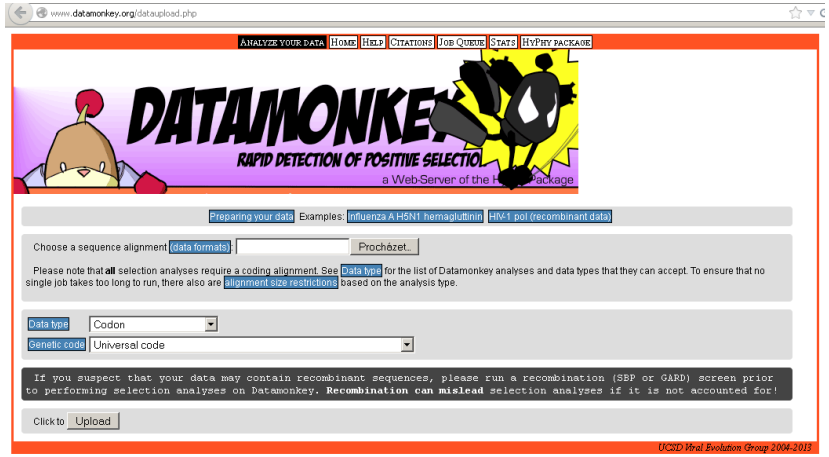


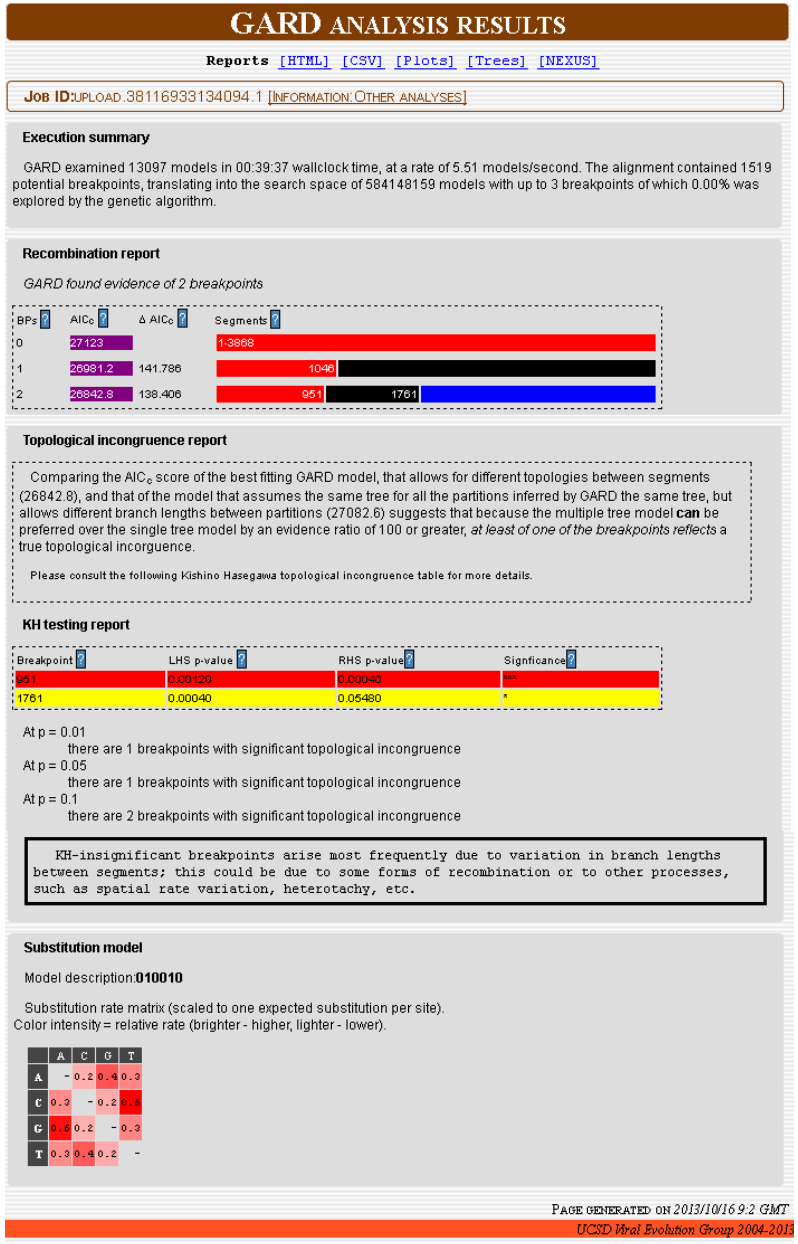


Detekce rekombinantů - GARD

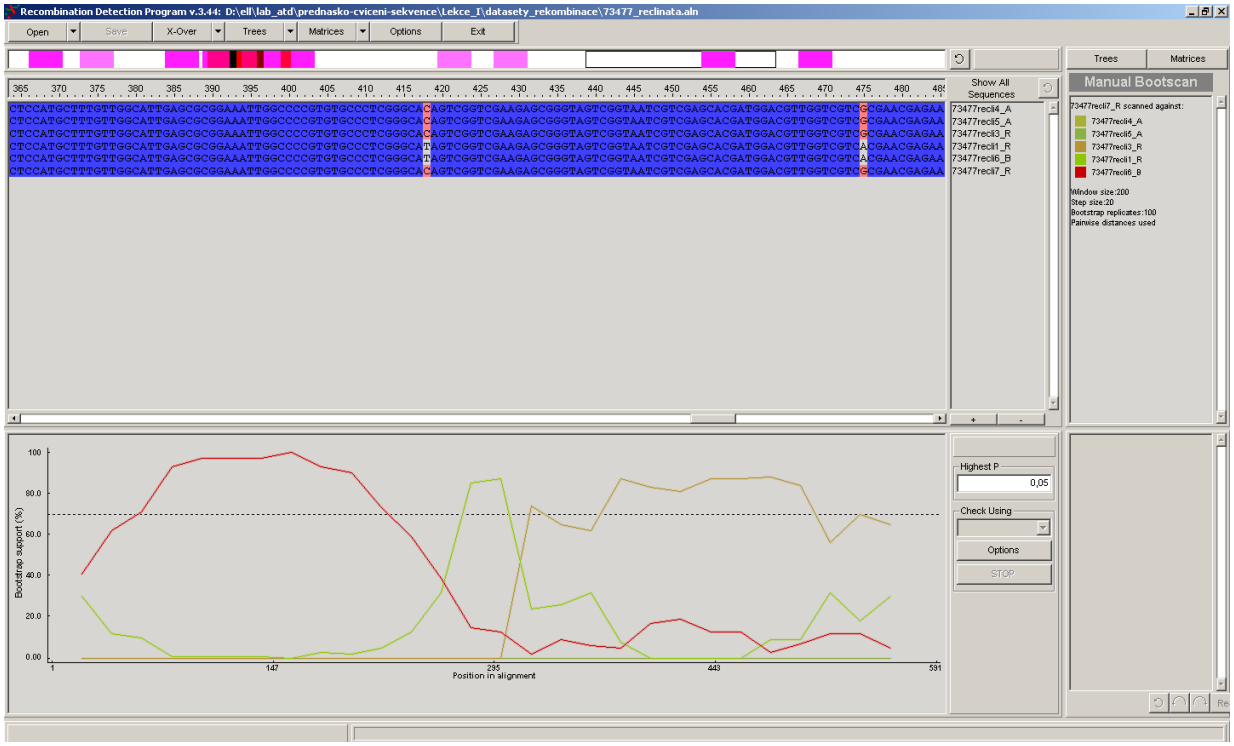
- součást balíku HyPhy (Hypothesis testing using Phylogenies)
- analýzy online - http://www.datamonkey.org/
- využívá srovnání topologie stromů
- vhodnější na delší a variabilnější úseky - schopnost detekovat rekombinace vzrůstá s mírou divergence sekvencí (Kosakovsky Pond et al. (2006)
- detekuje místo rekombinace, rozdělí dataset na X inkongruentních
- neoznačí jedince, kteří inkongruenci způsobují

- alternativní využití (!?) - test inkongruence datasetů (např. nDNA vs cpDNA)





Detekce rekombinantů - RDP3



- RDP3 - program zahrnující min 7 metod analýz rekombinantních dat
- analýza rekombinantů proti “rodičovským sekvencím”
- detekce rekombinantů de novo
- podobně jak GARD spolehlivější při vyšší divergenci v datech
- pro každý dataset nutno nastudovat vhodnou metodu dle dokumentace !

- Alternativní využití? Detekce hybridů v konkatenovaném datasetu nDNA a cpDNA úseků

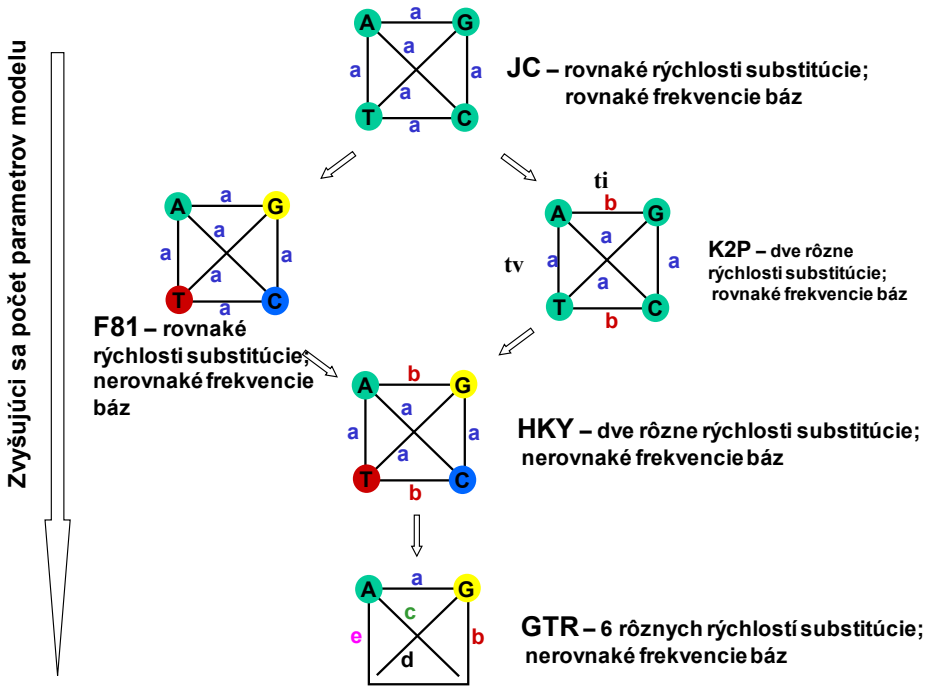
Modely evoluce DNA

- modely charakterizující evoluci DNA pomocí několika parametrů
  - frekvence bazí
  - typy substitucí (tranzice, tranzverze) a jejich rychlosti
  - heterogenita rychlosti substitucí na různých pozicích

- Vhodně zvolený model je klíčový při výpočtech věrohodností topologií fylogenetických stromů pomocí pravděpodobnostních metod (např. Maximum likelihood nebo Bayesovská analýza)

- ca 5 klasických modelů (JC, K2P, HKY,...GTR)
- až 56 různých modelů celkem

- jak zjistíme, který model vystihuje naše data nejlépe?
  - otestujeme jeden podruhém (získáme *log likelihood scores*)
  - porovnáme je pomocí AIC/hLRT, abychom dostali "nejoptimálnější" model



PROGRAMY

- Modeltest, jModeltest
- MrModeltest
- PAUP, MEGA
- PartitionFinder

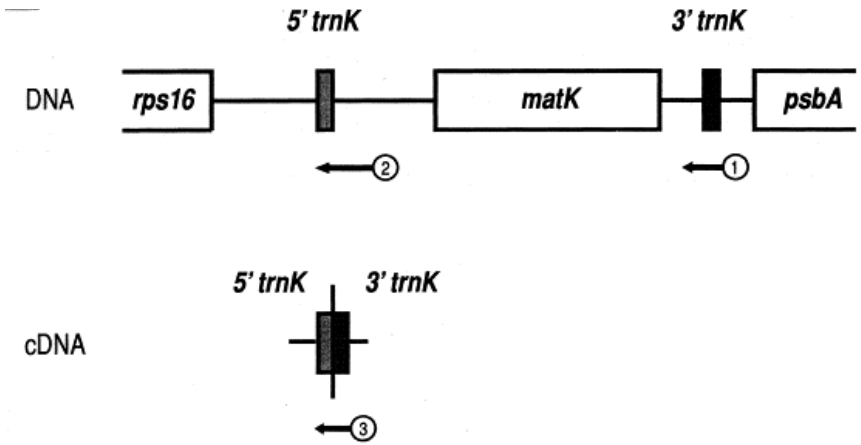
testování modelů evoluce DNA - partitions

- odvození struktury studovaného úseku - kódující a nekódující úseky mutují jinak, potřeba jiný model
- alignment našeho datasetu **s anotovanou sekvencí** z GB => anotace našeho datasetu

cpDNA, gen: maturase K

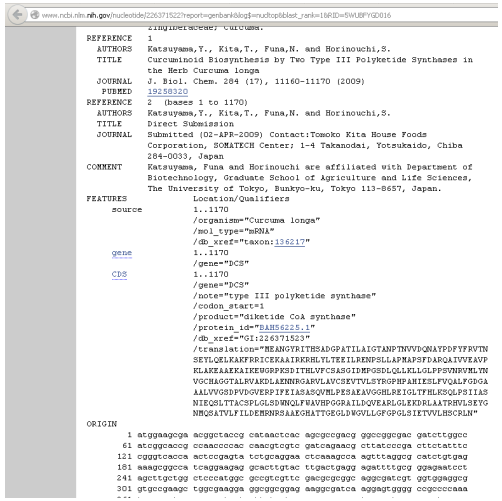
exon (811-2358 bp)

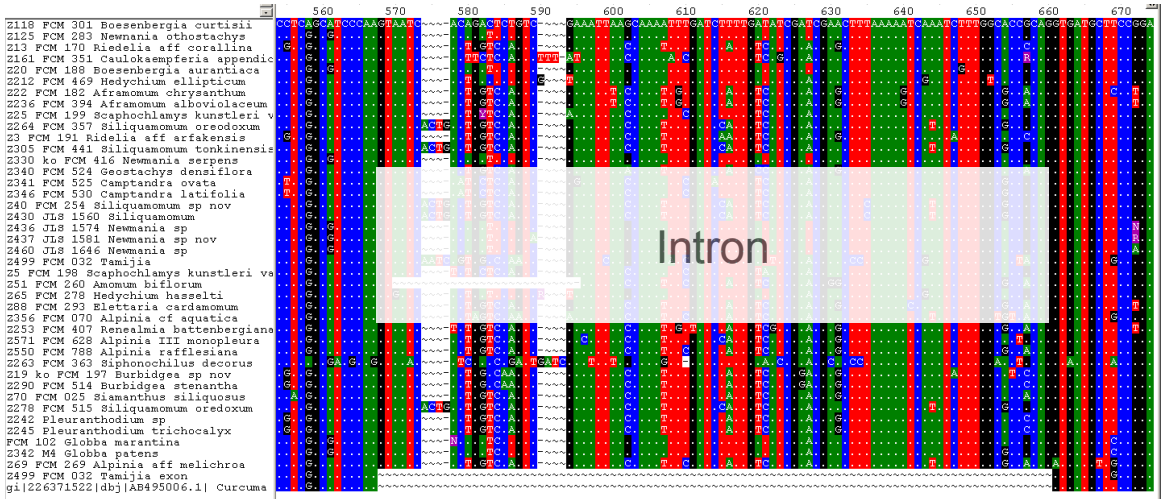
intron (1-810 & 2359-2651 bp)




```
FEATURES             Location/Qualifiers
     source          1..2651
                     /organism="Plagiostachys sp. LMP-2002-1"
                     /organelle="plastid:chloroplast"
                     /mol_type="genomic DNA"
                     /db_xref="taxon:200879"
                     /tissue_type="leaf"
     gene            <1..>2651
                     /gene="trnK"
                     /note="tRNA-Lys"
     intron          <1..>2651
                     /gene="trnK"
     gene            811..2358
                     /gene="matK"
     CDS             811..2358
                     /gene="matK"
                     /codon_start=1
                     /transl_table=11
                     /product="maturase K"
                     /protein_id="AAN63230.1"
                     /db_xref="GI:24634874"
                     /translation="MEELQGYLEEYRSRQQQFLYPLLFQEYIYVFAYDHGLNSSIFYE
                     PKNSLGYDNKFSSVLVKRLIIRMYQKNYWIYSVNDIYQNIFVGHNNYFYFHFFSQILF
                     EGFAVIVEIPFSLQLISSLEEKEIPKSHNLQSSHSIFPFLEDKLLHLNYLSDILIPYP
                     AHMEILVQMLQSWIQDALSLHLLQFLLHEYYNWNSLIIPNKSIYVFSKDNKHLFCFLY
                     NLYIYEYEFLLVFPCKQSSFLRLISSGVLLERIHFYVKIEHLGVCRIFCQKTLWIFKD
                     PFIHYIRYQGKSILGSRGTHFLMKKWKYHLVNFWQYYFHFWSQPYRIDTKKLSNYSFY
                     FLGYFSSVQMNSSMVRNQMLENSFLIDTLTKKLDTRIPIIPLIRSLSKAQFCTVSGYP
                     ISKPIWTDLADCDIINRFGQICRKLSHYHSGSSKKQSLYRMKYILRLSCARTLARKHK
                     SSARSFLQRLSSGLLEEFFTEEEQVISLIFPKRTSFYLYGSYRERIWYLDIIRINDLV
                     NSVLVTT"
ORIGIN
        1 gtgcgactat gatcttttac acatttggat gaagcaataa attcgtccag actattggta
       61 gagtctataa gaccacgact gatcctcaaa ggtaatgaat ggaaaaagta gcatgtcgta
      121 atacgtaata taataaaata cgaaatttcg tattttatta taattgaaaa aatttcaaat
      181 tgaaatgaaa gtaaaattaa gaacttttcc ttactcacat tcttacaatt tttttgtagg
```

nDNA, gen: DCS (CHS)





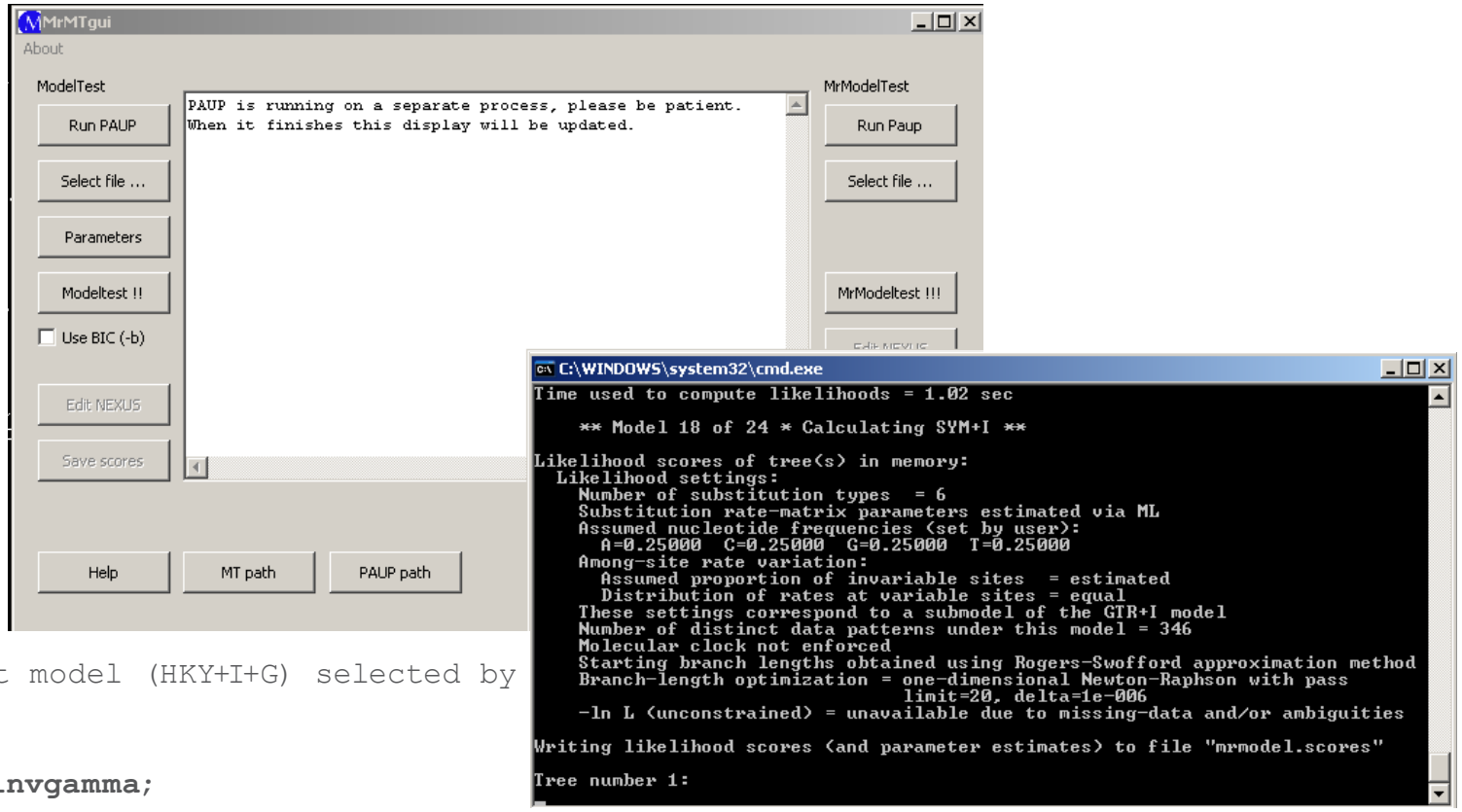
exon (1-567 & 661-1063 bp)

Intron (568-660 bp)

testování modelů evoluce DNA - Modeltest, MrModeltest & jModeltest

- praktický pomocník pro výpočet likelihood jednotlivých modelů - MrMtGui, alternativa je **jModeltest**
- MrMtGui propojen s
  - PAUP - výpočet likelihood pro jednotlivé modely
  - Modeltest - vyhodnocení, který model je pro daná data nejvhodnější - výstup pro ML
  - MrModeltest - podobně jako Modeltest, ale porovnává jen vybrané modely - výstup pro MrBayes

- Run PAUP (výběr souboru *.nex)
- save scores
- select file (*.scores)
- (Mr)Modeltest!
- zkopíruj příkazy pro MrBayes (nebo ML)




```
[! MrBayes settings for the best-fit model (HKY+I+G) selected by
AIC in MrModeltest 2.3]

BEGIN MRBAYES; Lset nst=2 rates=invgamma;

  Prset statefreqpr=dirichlet(1,1,1,1);

END;
```

testování modelů evoluce DNA - PartitionFinder

- testuje, které úseky datasetu mají podobný model evoluce
- rozdělíme dataset na nejvíce možných “partition” - kódující sekvence (separátně 1., 2., 3. pozice), nekódující
- PartitionFinder otestuje, kolik z původních “partition” má smysl rozeznávat

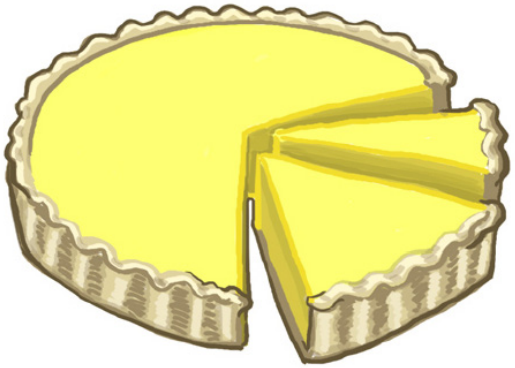
- PartitionFinder je python script - je potřeba mít instalovaný Python
- spouští se příkazem

```
python “<PartitionFinder.py>” “<InputFoldername>”
```

- vstupní soubory
  - sekvence ve PHYLIP formátu (*.phy)
  - definice “partition” v datasetu a příkazy pro PF (*.cgf)

hy | CHS_complete_aln_E_short_headers.phy | CHS_complete_aln_E_short_headers.phy | partition_finder.cfg | CHS_exon_only_MrBayes_partition.nex

```
## ALIGNMENT FILE ##
alignment = CHS_complete_aln_E_short_headers.phy;

## BRANCHLENGTHS: linked | unlinked ##
branchlengths = linked;

## MODELS OF EVOLUTION for PartitionFinder: all | raxml | mrbayes | beast | <list> ##
##              for PartitionFinderProtein: all_protein | <list> ##
models = all;

# MODEL SELECCTION: AIC | AICc | BIC #
model_selection = BIC;

## DATA BLOCKS: see manual for how to define ##
[data_blocks]
Gene1_pos1 = 1-567\3;
Gene1_pos2 = 2-567\3;
Gene1_pos3 = 3-567\3;
CHS_intron = 568-660
Gene2_pos1 = 661-1063\3;
Gene2_pos2 = 662-1063\3;
Gene2_pos3 = 663-1063\3;

## SCHEMES, search: all | greedy | rcluster | hcluster | user ##
[schemes]
search = greedy;

#user schemes go here if search=user. See manual for how to define.#
```

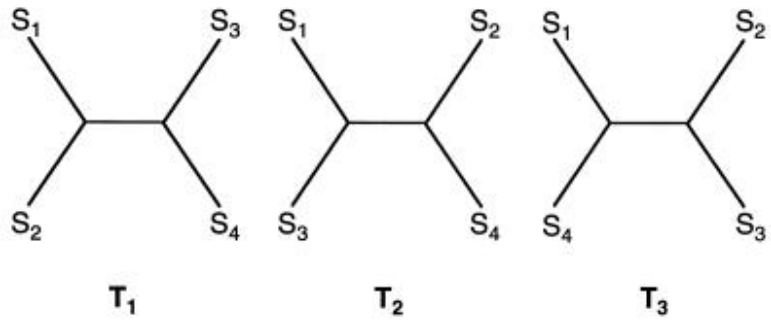


Testování fylogenetické struktury v datech

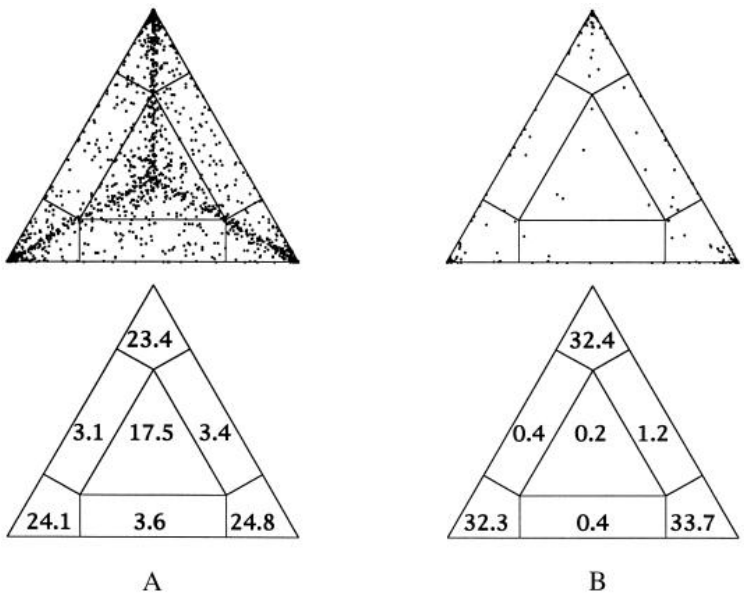
- Jaká je míra fylogenetické informace a šumu v datech?

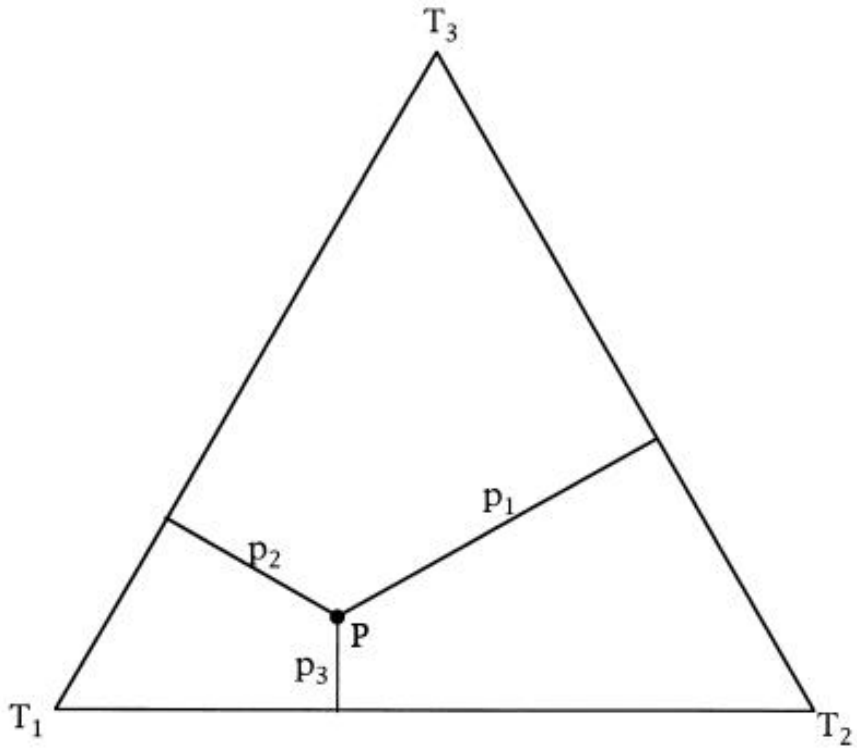
1) **Likelihood mapping**

- porovnání pravděpodobností ML topologií čtyř vybraných sekvencí (kvartetů)



- Rozdíl v pravděpodobnostech je zobrazen pomocí vektoru P uvnitř rovnostranného trojúhelníku





Testování fylogenetické struktury v datech

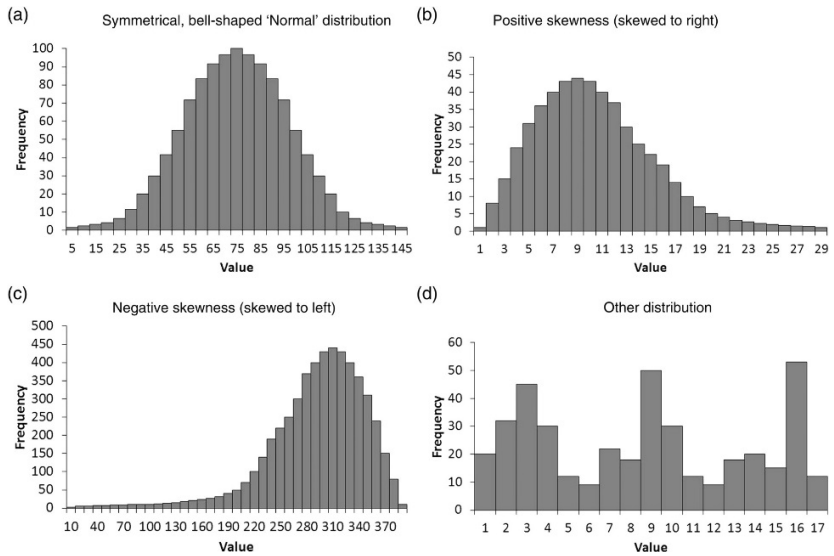
- Jaká je míra fylogenetické informace a šumu v datech?

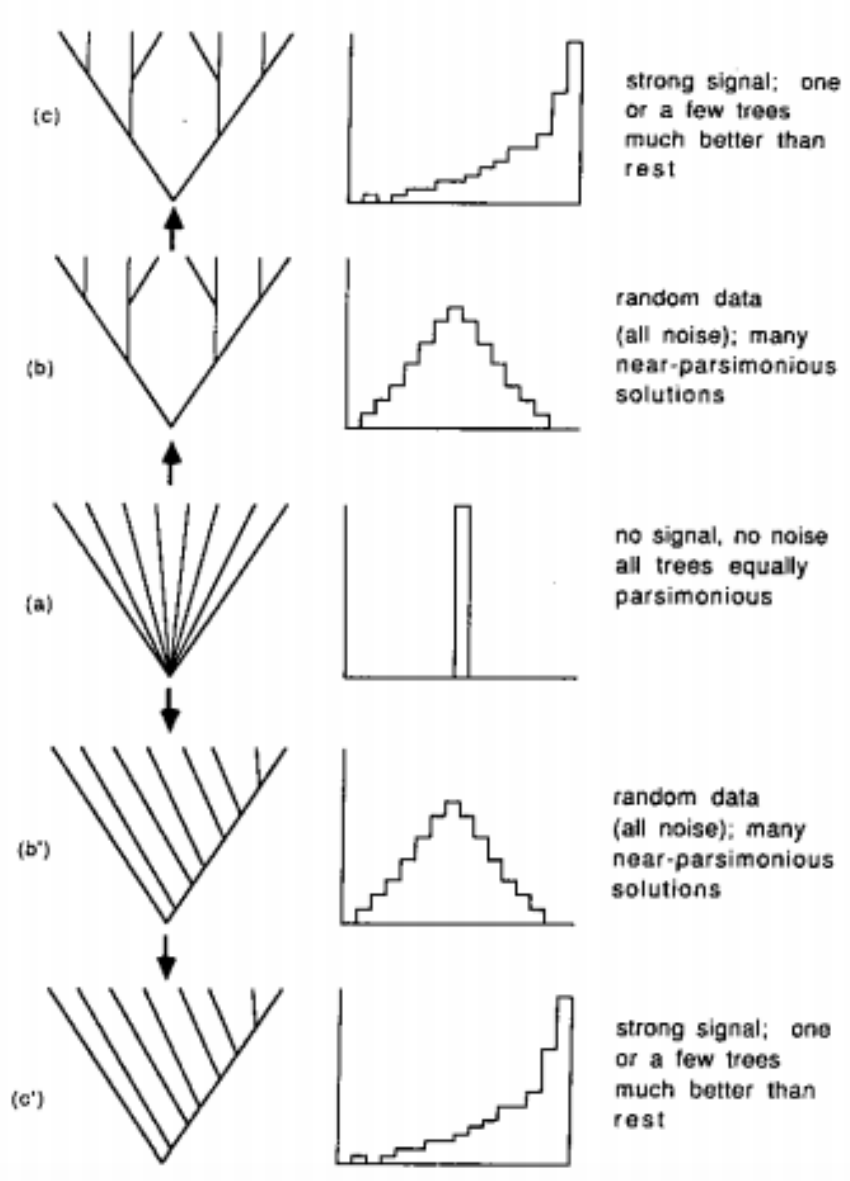
2) **g1 statistika**

- Distribuce délek MP stromů u náhodně vygenerovaných sekvencí je symetrická

- U fylogeneticky strukturovaných dat je distribuce délek MP stromů doleva zkosená

- g1 statistics of skewness – vypočtená hodnota udává směr (-/+) a míru zkosení

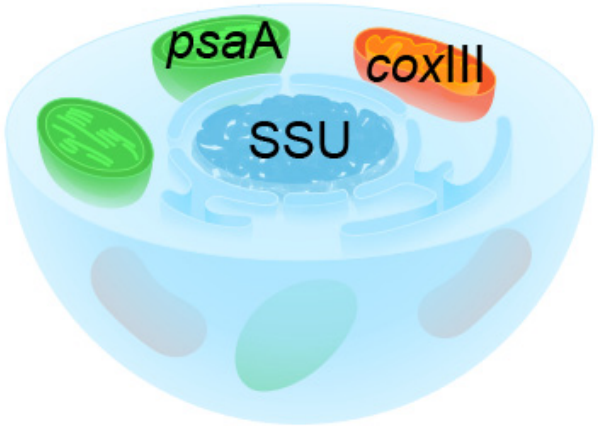




Testování fylogenetické struktury v datech

1) **Likelihood mapping**

- Porovnání fylogenetické struktury u tří vybraných genů (SSU, psaA, coxIII)
- Sekvence ve formátu Phylip
- program Tree Puzzle



2) **g1 statistika**

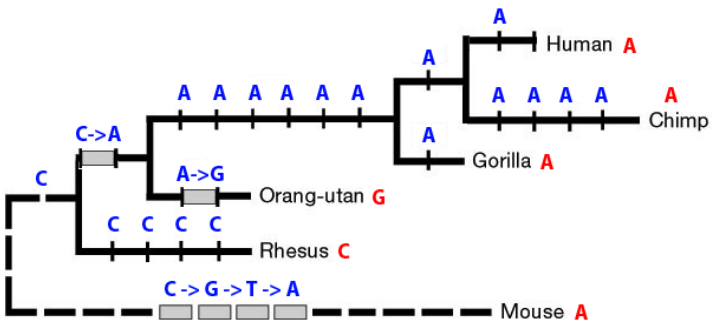
- Porovnání distribuce délek stromů u tří vybraných genů (SSU, psaA, coxIII)
- Sekvence ve formátu Nexus
- program PAUP na generování stromů
- R, případně Excel na vypočtení hodnoty g1
- Hodnoty g1 menší než -0.09 poukazují na statisticky významné levé zešikmení distribuce délek MP stromů (P = 0.01)

PROGRAMY

- Tree Puzzle
- PAUP
- R
- Excel

Substituční saturace sekvencí

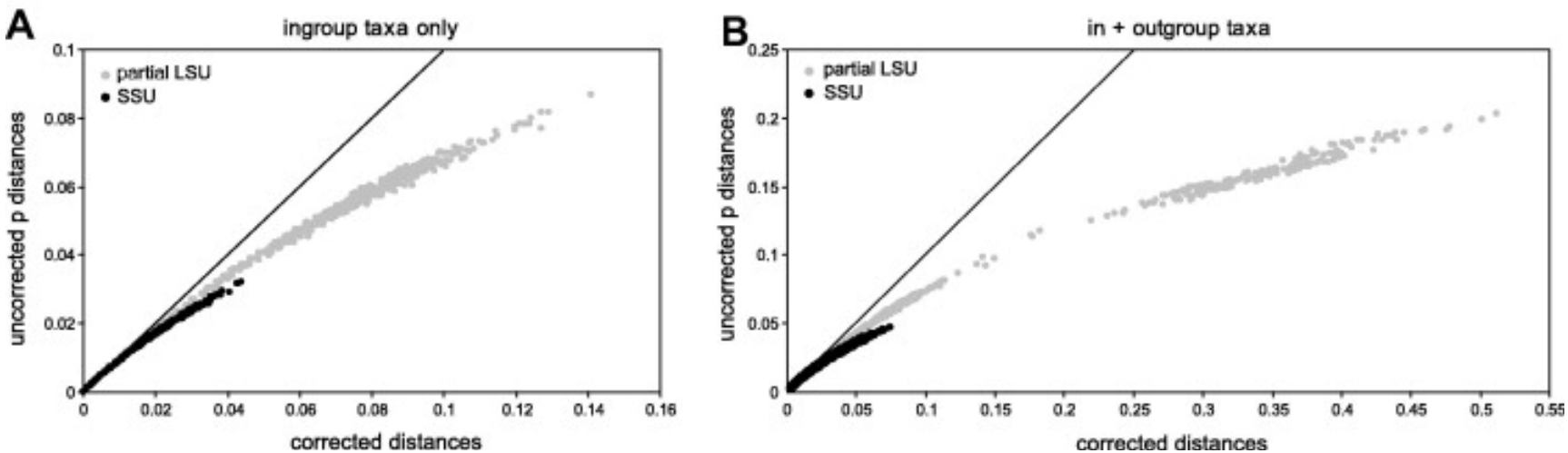
- Jaká je míra šumu v datech, způsobená substituční saturací?



- Substituční saturace
  - některé pozice v alignmentu prošly během evoluce několika substitučními změnami
  - protože sekvence mají pouze 4 stavy, časem u nich dochází ke stochastickému hromadění šumu.
  - saturované pozice mohou tvořit většinu variability v datech
  - velký problém obzvlášť pro MP analýzy!

1) **Saturační křivky**

- Porovnání jednoduchých sekvenčních distancí a distancí spočítaných na základě substitučních evolučních modelů

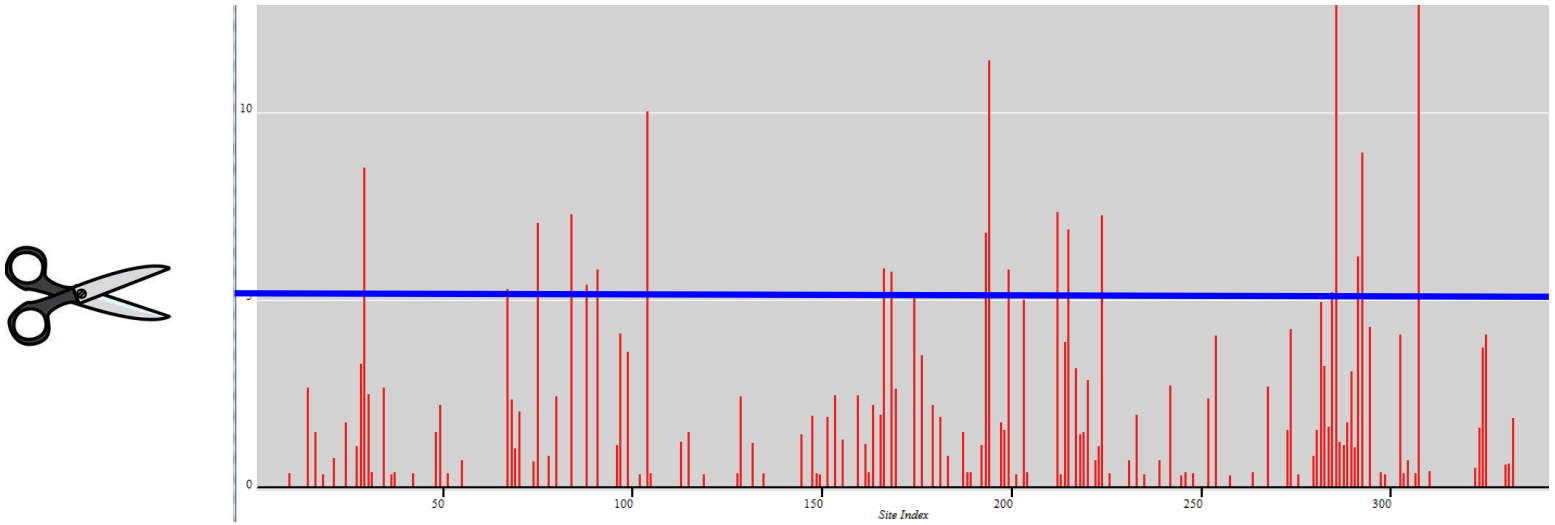


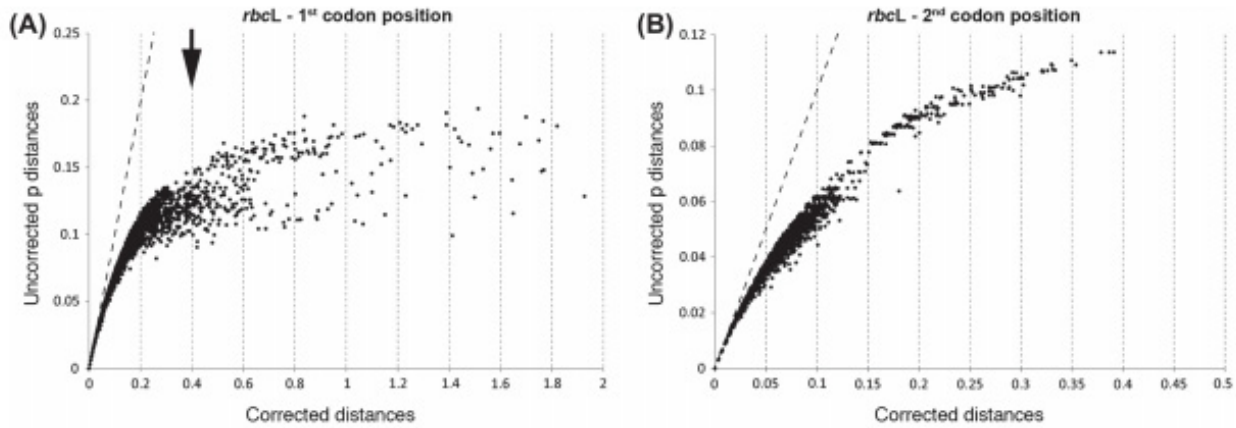
Substituční saturace sekvencí

- Jaká je míra šumu v datech, způsobená substituční saturací?

2) **Site stripping**

- odstranění saturovaných pozic z alignmentu sekvencí

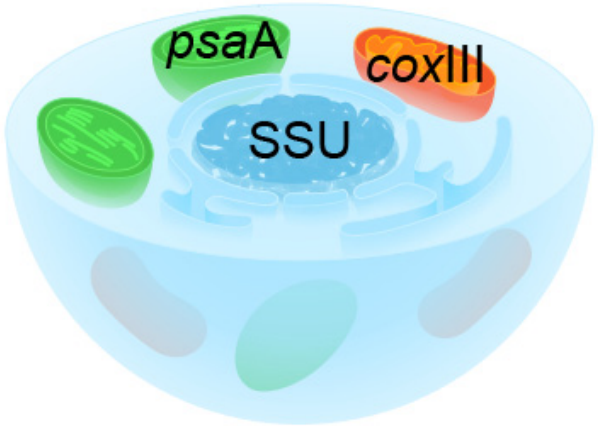




Substituční saturace sekvencí

1) **Saturační křivky**

- porovnání saturací různých kodónových pozic v genu rbcL
- alignment ve formátu Nexus
- program PAUP pro vypočtení distancí



2) **Site stripping**

- odstranění saturovaných pozic
- alignment ve formátu Fasta
- program MEGA na vypočtení rychlého MP stromu
- program HyPhy na spočítání substitučních rychlostí
- prostředí Perl a skript „sitestripper.pl“ pro odstranění saturovaných pozic

PROGRAMY

- PAUP
- R (Excel)
- MEGA
- HyPhy
- Perl
- SiteStripper

Praktické cvičení

- **Cílem** - připravit sekvenační data pro fylogenetickou analýzu některým z programů pro ML nebo Bayes (i MP)

1) **Editace alignmentu**

- vytvořte a manuálně upravte alignment (porovnejte s raw data), uložte jako *.fas (přejmenujte)

- určete a zaznamenejte strukturu strudovaného úseku (stačí kódující vs. nekódující). Využijte BLASTu k nalezení nejpodobnější anotované sekvence, kterou přidejte do svého datasetu, znovu alignujte a podle anotované sekvence odvoďte strukturu vašich sekvencí.

- původní alignment konvertujte do formátu NEXUS a Phylip pomocí webové aplikace https://app.bugaco.com/converter/biology/sequences/

2) **testování modelů evoluce**

- použijte vytvořený Nexus soubor a zanalyzujte ho pomocí jModeltest (jediná partition pro kódující i nekódující oblast). Vytvořte dva další soubory Nexus rozdělením původního alignmentu na kódující a nekódující oblast a znovu analyzujte v jModeltest.

- upravte soubor “partition_finder.cfg” (nejlepe v nějakém textovém editoru, např. NotePad) pro vaše data a zanalyzujte formát Phylip programem PartitionFinder

- porovnejte navržené modely pro celkový dataset a dataset kódujících a nekódujících oblastí.

3) **testování fylogenetické struktury v datech**

- pomocí likelihood mapping a g1 statistiky otestujte míru fylogenetického signálu v datech

4) **otestovat a zhodnotit míru saturace sekvenačních dat**

- pomocí saturačních křivek a site stripping určete míru substituční saturovanosti sekvencí

příkladové DATASETY

<- CHS_complete.fas (nDNA)

<-složka “CHS_raw_data”

<- matK_Zingiberaceae.fas

<-CHS_complete_outgroup_PKS.fas

<-matK_Zingiberaceae_aln_s_GB_sekvenci.fas

<- CHS_exon_only.phy

<- partition_finder.cfg

<-chryso_rbcl1(2,3).fas

<-chryso_rbcl1(2,3).nex

<-Micrasterias_cox.nex

<-Micrasterias_cox.phy

<-Micrasterias_psa.nex

Praktické cvičení - porovnání výsledků - diskuze

- testování modelů evoluce -jak se liší modely evoluce pro datasety z různých kompartmentů (nDNA, cpDNA, mtDNA)?
- jsou navržené modely pro tyto úseky stejné z programů jModeltest, MrModeltest a PartitionFinder?
- u kterých datasetů je vysoká míra saturace sekvencí?
- ....

- Na příště:
- - uschovat si alignované soubory (kódované vs. nekódované; s IUPAC vs. bez IUPAC, s missing data vs. bez)
- - vytvořit si vstupní soubory pro ML, MrBayes a MP