

„Propojení výuky oborů Molekulární a buněčné biologie a Ochrany a tvorby životního prostředí“

Reg. č.: CZ.1.07/2.2.00/28.0032



evropský
sociální
fond v ČR



EVROPSKÁ UNIE



MINISTERSTVO ŠKOLSTVÍ,
MLÁDEŽE A TĚLOVÝCHOVY



OP Vzdělávání
pro konkurenceschopnost

INVESTICE DO ROZVOJE VZDĚLÁVÁNÍ

Molekulární markery I.

Isozymy, PCR, RAPD, AFLP

Miloslav Kitner
Katedra botaniky PŘF UP



evropský
sociální
fond v ČR



EVROPSKÁ UNIE



MINISTERSTVO ŠKOLSTVÍ,
MLÁDEŽE A TĚLOVÝCHOVY



OP Vzdělávání
pro konkurenceschopnost

INVESTICE DO ROZVOJE VZDĚLÁVÁNÍ

Markery

- identifikační znaky, rozlišovače
- cíleně či náhodně vybraná část celkové informace o studovaném objektu
- jejich pomocí dovedeme popsat různorodost = variabilitu organismů

Variabilita

- vlastnost organismů se navzájem lišit
- *nedědičně* podmíněná (ovlivněná prostředím) a
- *dědičně* (geneticky) podmíněná

Viditelné polymorfismy = *fenotypový projev*

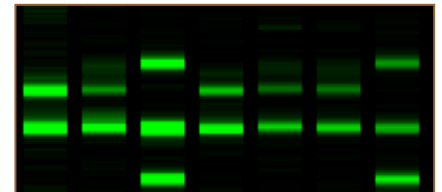
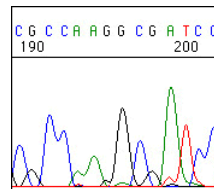


Viditelné polymorfismy - omezený počet variant (kriterií) pro charakterizaci jedince



Co jsou to molekulární markery?

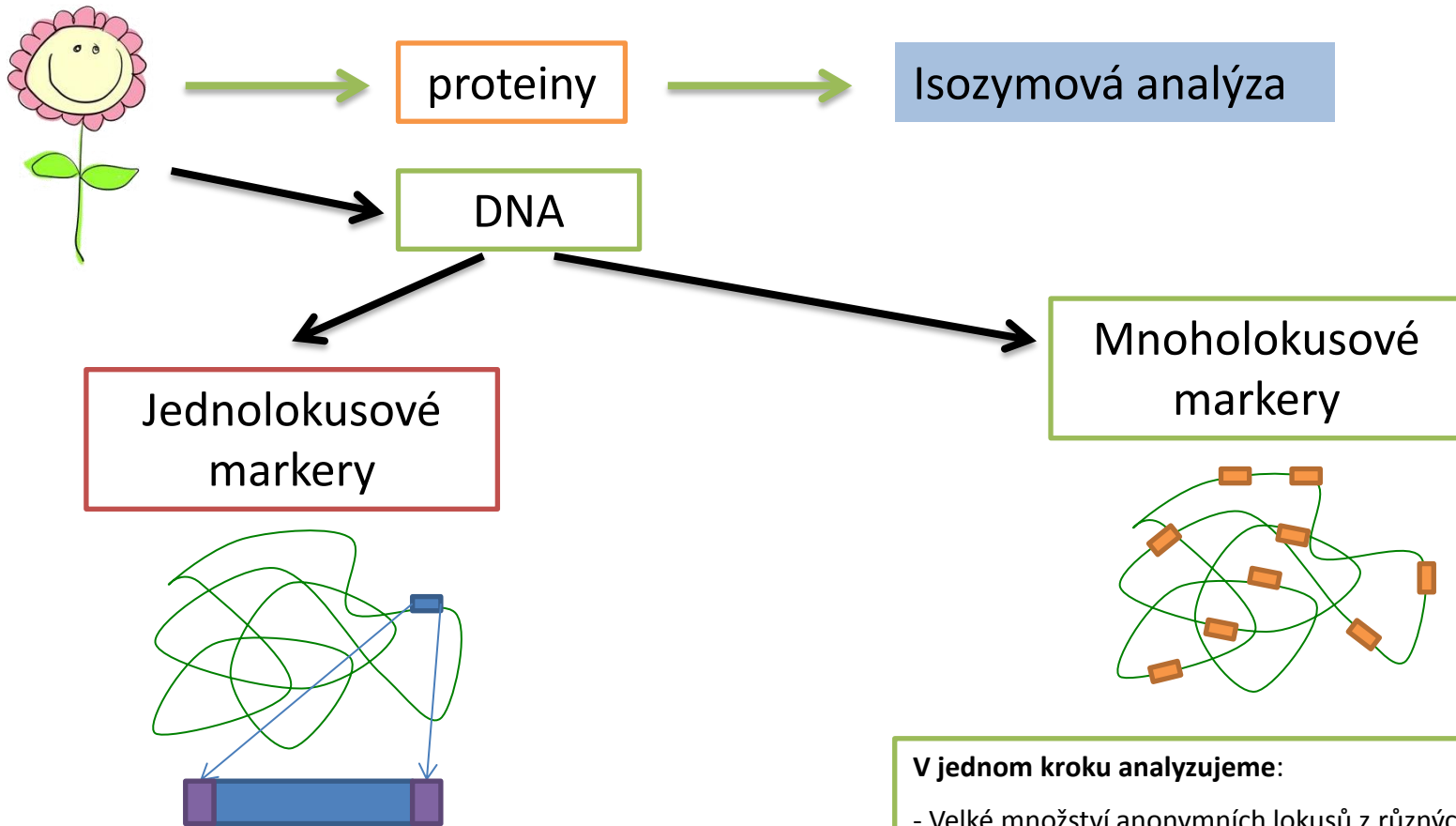
- informace o organismu získaná na základě analýzy jeho molekul – proteinů a/nebo DNA
- vypovídají o genetické podobnosti (příbuznosti) na úrovni jedinců, populací, druhů a vyšších taxonomických celků.
- doplňují informace o genetické proměnlivosti organismu získávané klasickými přístupy (srovnávací anatomie, morfologie, embryologie, fysiologie).



Vlastnosti molekulárních markerů

- musí vykazovat určitý stupeň polymorfismu (čím vyšší, tím lepší)
- jednoznačné určení alel, fragmentu nebo nukleotidu
- upřednostňovaný je kodominantní charakter markeru
- častý výskyt v genomu, distribuce v celém genomu
- poskytuje informaci nezávislou na podmínkách prostředí
- mění se při generativním rozmnožování – rekombinace
- lehce a rychle proveditelná metoda
- vysoká reproducibilita
- nízká cena analýzy, popř. vývoje nového markeru

Základní rozdělení “klasických” molekulárních markerů



V jednom kroku analyzujeme:

- pouze jedno, konkrétní místo (lokus), z přesně definované oblasti genomu.
- Získáme detailní informace o dané oblasti
- zpravidla kodominantní markery
- Klíčová je znalost primerových sekvencí pro provedení PCR
- Příklad markeru: sekvenace genů, mikrosatelity, PCR-RFLP

V jednom kroku analyzujeme:

- Velké množství anonymních lokusů z různých částí genomu
- Techniky využívají např.:
 - nespecificky nasedající primery (RAPD)
 - naštípání DNA na anonymních fragmentů a jejich následná úprava umožňující provedení PCR (př. AFLP)
- Příklad markeru: RAPD, AFLP

Isozymy

Isozymová analýza = allozomy

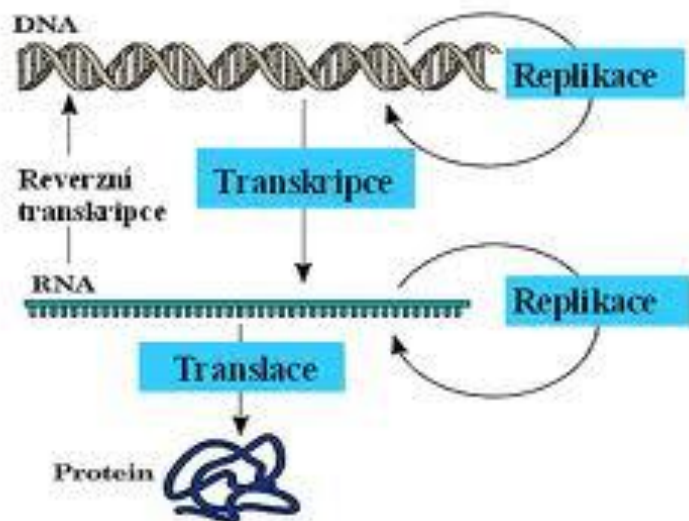
Enzymy: Jednoduché nebo složené bílkoviny katalyzující chemické reakce v živých organismech

Isoenzymy (isozomy) – enzymy se stejnou funkcí v metabolismu, katalyzující stejnou chemickou reakci, mají ale odlišnou (primární) strukturu (délka řetězce, záměna AMK)

- Analyzujeme a srovnáváme přítomnost různých isoform v daném organismu, populaci, druhu.

- **Obecně – vykazují nízkou variabilitu**

- Enzymy jsou nutné pro správnou funkci metabolických drah
 - Výraznější změna (mutace), která by pozměnila jejich funkčnost je eliminována (snížení životaschopnosti jedince)
- Fungují reparační mechanismy – např. oprava chybně začleněného nukleotidu během replikace
- Záměna nukleotidu v DNA nemusí vést k záměně AMK v proteinu (genetický kód je degenerovaný), popř. případná záměna AMK nemusí nutně vést ke změně náboje nebo délky proteinu a k detekci těchto změn elektroforézou



isozomy – katalyzují stejnou reakci

allozomy – produkty (alely) jednoho lokusu

Analýza isozymů

výhody

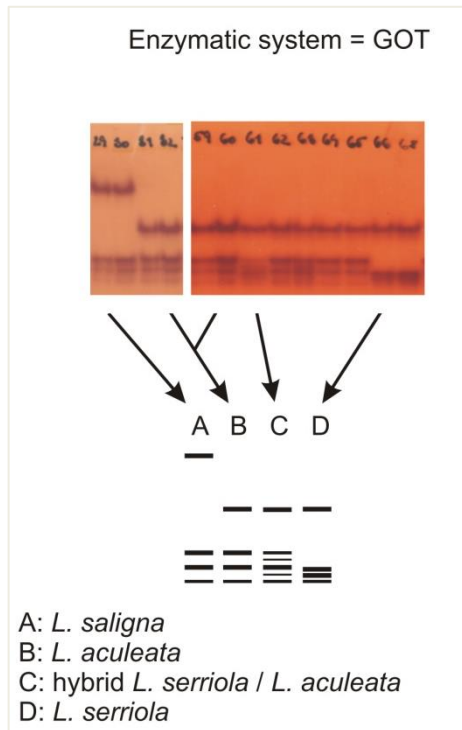
- rychlá metoda – možno analyzovat mnoho individuí najednou (?)
 - levná technika (v porovnání s DNA technikami) (?)
 - srovnatelná data mezi různými studiemi
 - kodominantní marker
-
- **Využití:**
 - Populační genetika
 - Studium druhové diverzity
 - Determinace druhů
 - Studium kryptických druhů
 - Identifikace mezidruhových hybridů

nevýhody

- potřeba čerstvého materiálu
- omezená variabilita – málo alel v lokusu – často jen 2-4
- Produkty genů = analyzujeme proteiny, nemusíme detekovat případnou variabilitu DNA, která je kóduje
- detekovaná variabilita
 - odráží pouze 10% variability na úrovni DNA (Nei 1987)
 - pouze 1/3 nukleotidových substitucí v DNA vede k záměně AMK a detekci těchto změn elektroforézou

- Dnes jsou isozymy nahrazovány modernějšími metodami.
- Můžete se setkat se studiemi, kde jsou výsledky (SSR, AFLP) srovnávány s pracemi popisujícími variabilitu populací/druhů získaných pomocí isozymů.

Isozymy = allozymy



Enzymatic system:	GOT			PGM			ME			LAP		
Phenotype:	A	B	C	A	B	C	A	B	C	A	B	C
Genotype:	<i>Got-1:</i> aa	aa	aa	<i>Pgm-1:</i> aa	aa	aa	<i>Me-1:</i> aa	aa	aa	<i>Lap-1:</i> aa	bb	cc
	<i>Got-2:</i> aa	ab	bb	<i>Pgm-2:</i> aa	aa	aa	<i>Me-2:</i> aa	ab	bb			
	<i>Got-3:</i> aa	ab	bb	<i>Pgm-3:</i> aa	ab	bb						
	<i>Got-4:</i> aa	aa	aa									
Species:	a	h	s	a	h	h, s	a	h	s	a, h	a, h	s

Fig. 2. Phenotypes of four enzymatic systems present in samples of *L. aculeata* (a), *L. serriola* (s) and *L. aculeata* × *L. serriola* (h), with putative loci, allele and genotype designation.

Fig. 1. Morphological traits of plants *L. aculeata* (acc. no. 376-15), putative hybrid *L. aculeata* × *L. serriola* (acc. no. 343-8A), and *L. serriola* f. *serriola* (acc. no. 377-8): nichomes on stems (a-c); stem leaves from abaxial (d) and adaxial sides (e); plants at a stage of beginning of bolting (f).

Postglacial recolonisation in *Fagus sylvatica* (389 pop, 18440 trees, 15 allozyme loci)

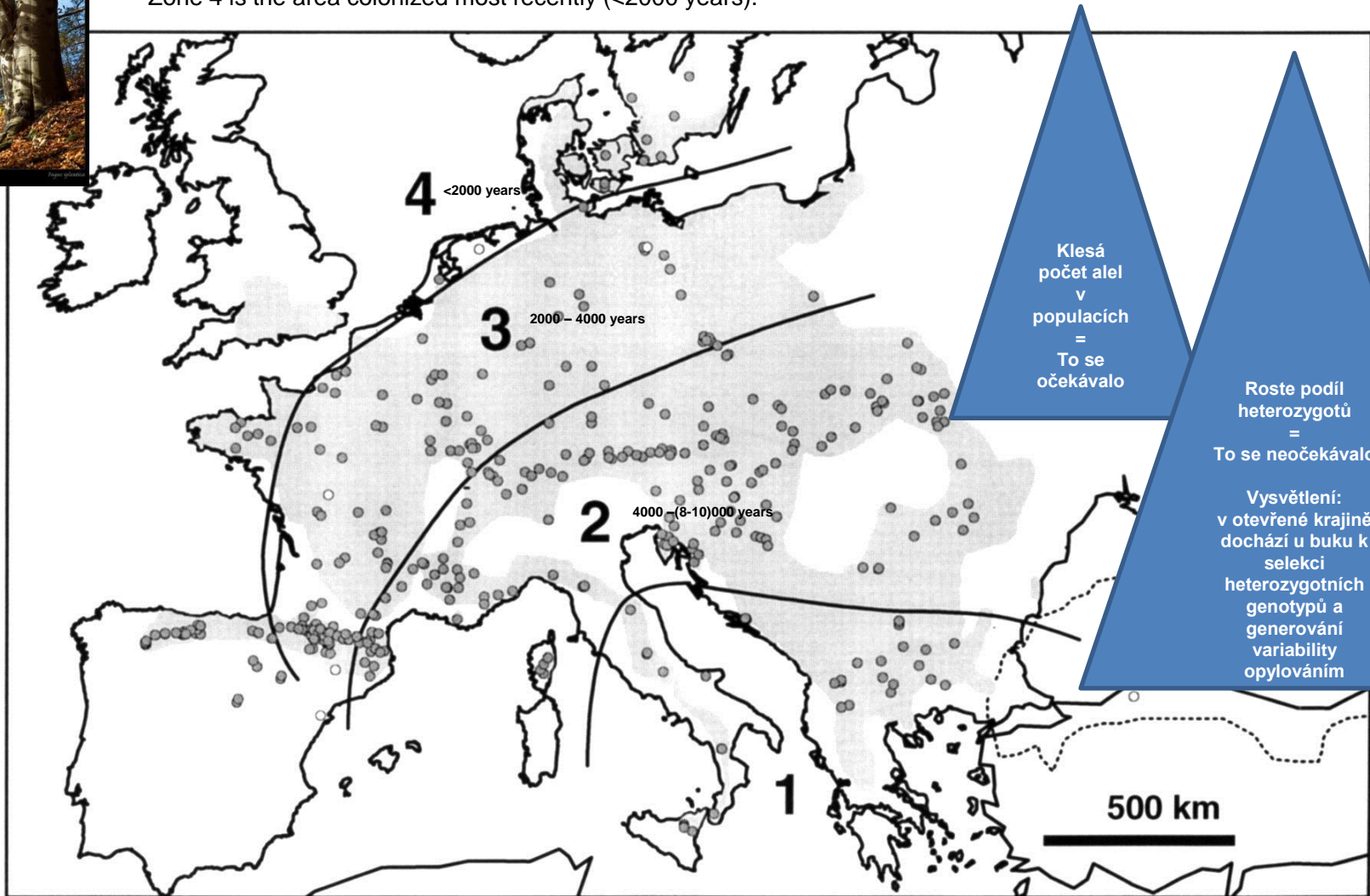


Four zones have been determined, corresponding to colonization history.

Zone 1 - last ice age, as inferred from fossil pollen

Zones 2 and 3 correspond to beech forests of intermediate age (older and younger than ~4000 years)

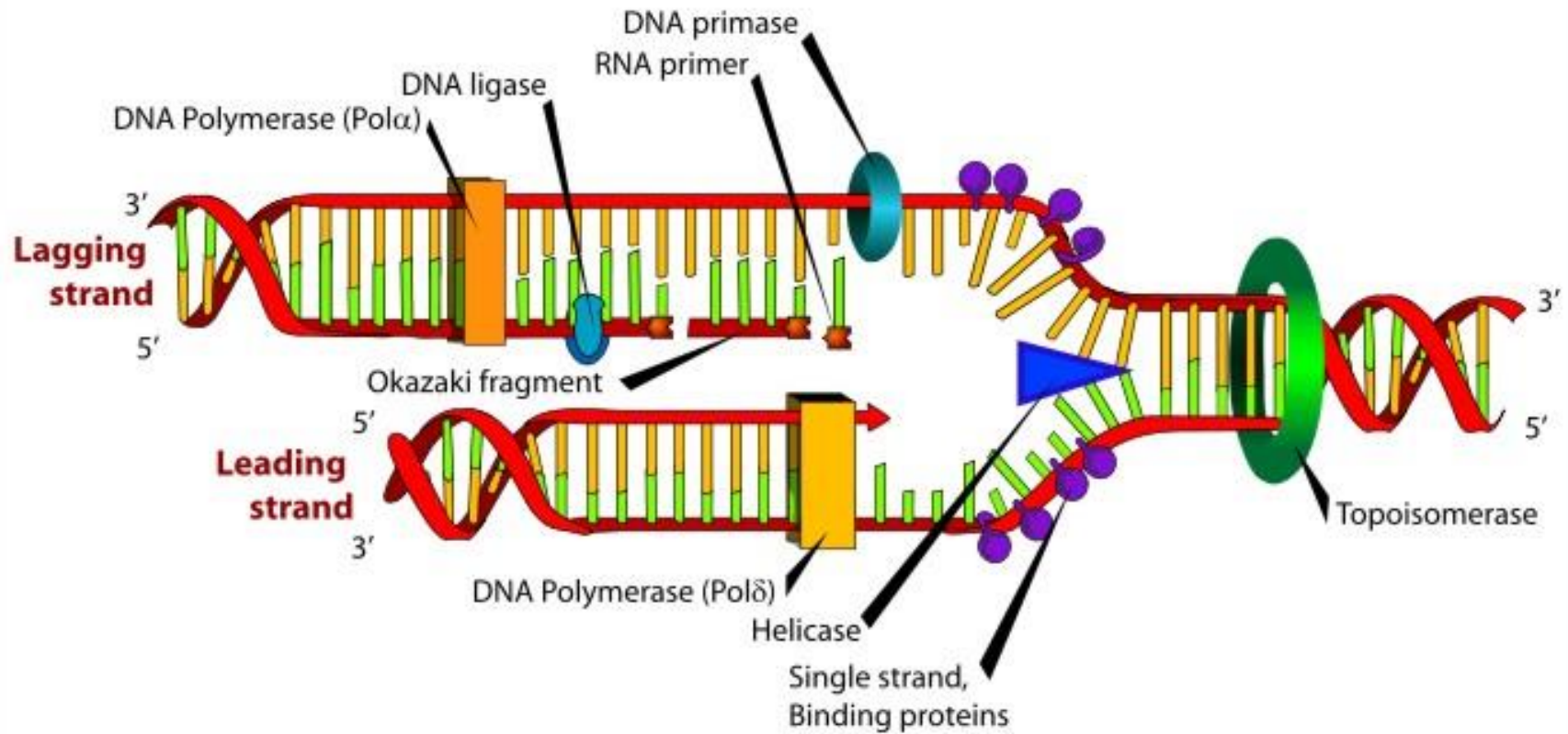
Zone 4 is the area colonized most recently (<2000 years).



PCR – Polymerase Chain Reaction

Polymerázová řetězcová reakce

Replikace „in vivo“



Dr. Kary Banks Mullis



<http://www.karymullis.com/>

PCR – replikace „*in vitro*“

PCR – Polymerase Chain Reaction

Polymerázová řetězová reakce

- 1955 - objev DNA polymerázy a první pokusy s umělou syntézou úseků DNA (amplifikace DNA)
- 1983 – **Karry Mullis** pracoval na projektu geneticky podmíněných chorob člověka v kalifornské firmě Cetus Corporation. Hledal novou metodu analýzy mutací DNA.

Modifikoval Sangerovu metodu sekvenace DNA – přidal zpětný primer pro syntézu krátkých úseků DNA.

„Pomocí PCR můžete z jedné molekuly DNA vytvořit miliardy kopií krátkého úseku DNA za jedno odpoledne. Potřebujete jenom zkumavku, několik jednoduchých reagensů a zdroj tepla“

- 1984 – rozpracování metody s matematikem Fredem Faloonou, patentování (Mullis et al. 1986; Mullis a Faloona 1987)
- 1993 – Nobelova cena za chemii
- V devadesátých letech Cetus Corp. prodává patenty „PCR a Taq“ firmě Hoffman–La Roche za 330 miliónů \$



Amplifikace fragmentů DNA v *in vitro* podmínkách

Polymerázová řetězová reakce PCR (Polymerase Chain Reaction):

- Metoda založená na principu replikace nukleových kyselin, umožňující rychlé a snadné namnožení úseku DNA, který studujeme. Předpokladem je znalost koncových sekvencí, které studovaný úsek DNA ohraničují.

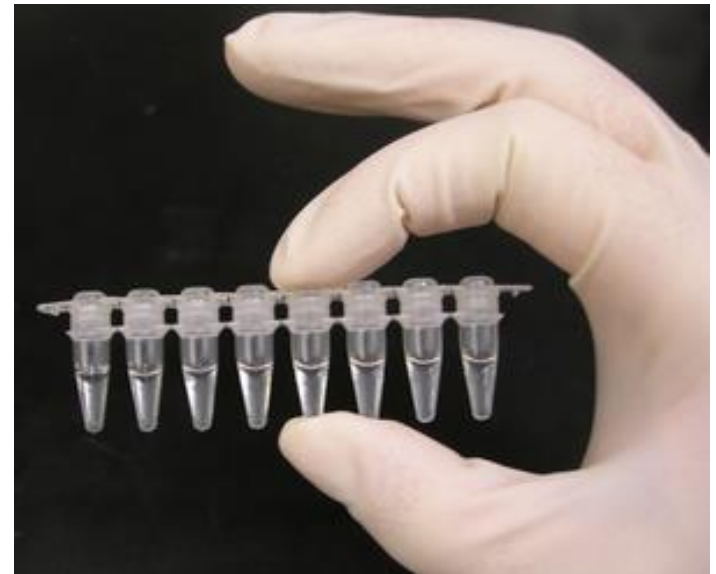
Při PCR dochází k periodickému opakování tří kroků:

1. denaturace analyzované templátové DNA
2. nasedání dvojice primerů – annealing
3. prodlužováním fragmentů DNA.

Zmnožení fragmentu DNA je prováděno v termocykleru.

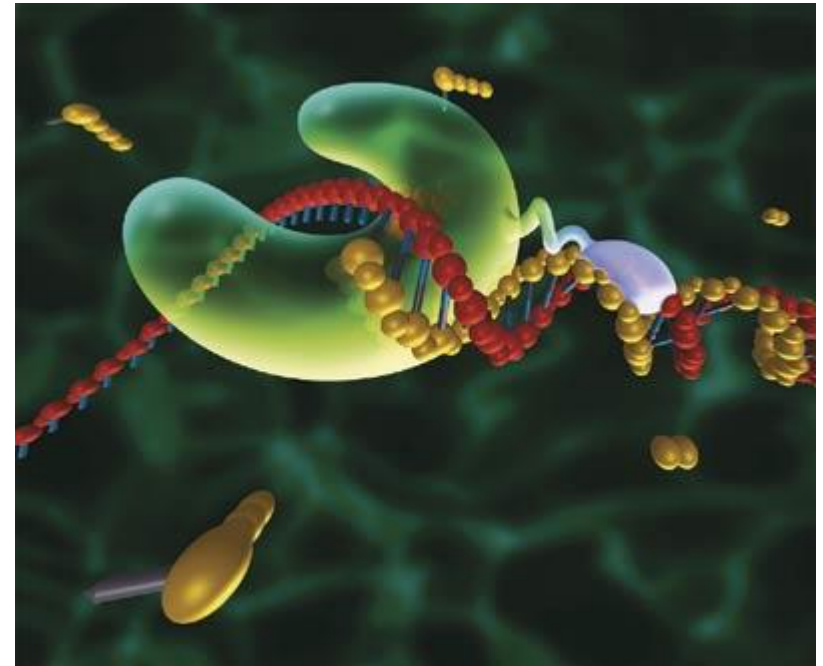
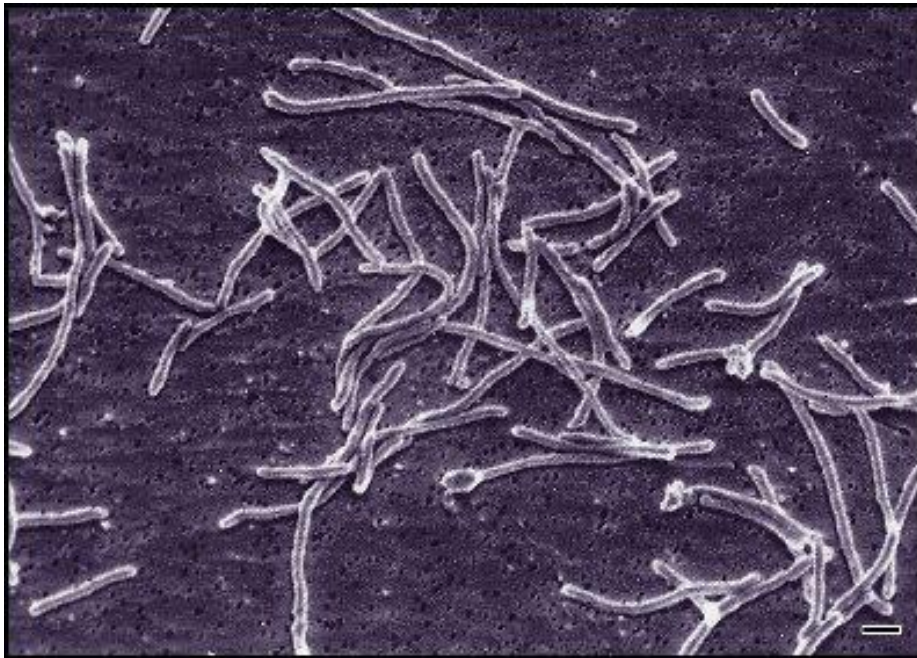


PCR – replikace „in vitro“



PCR – Polymerase Chain Reaction

- další rozvoj metody s použitím termostabilní polymerázy z termofilní bakterie *Thermus aquaticus* = zkratka **Taq Pol** nebo "**Taq**")



PCR – hlavní faktory ovlivňující průběh reakce

Koncentrace MgCl₂

- optimum je 1,5-2mM,
- nad 6mM klesá aktivita *Taq* Polymerázy
- nízká koncentrace – nízké výtěžky PCR
- příliš vysoké koncentrace – rozmazané proužky na gelu nebo zmnožení nespecifických produktů

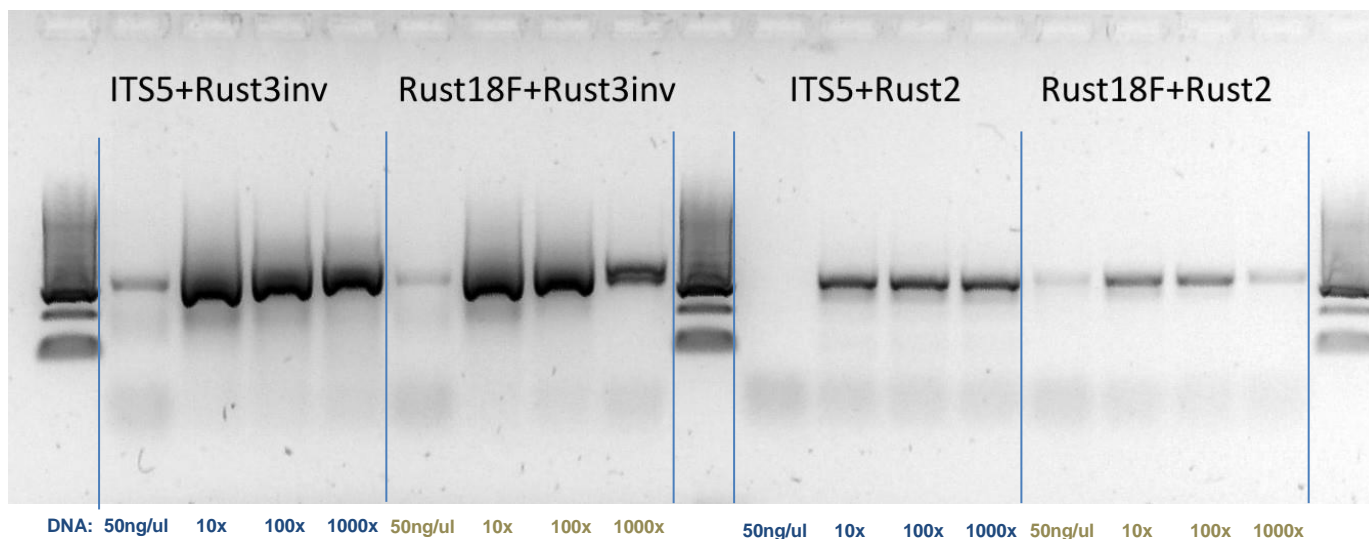
Teplota annealingu – nasednutí primerů

- **vyšší teplota** = vyšší specifičnost ke komplementární sekvenci = **primer nasedá na komplementární místo s vyšší přesností** = přesnější výsledek PCR
- **nižší teplota** – vhodné pro studie, kdy se používají primery odvozené od příbuzných druhů; **nebezpečí vzniku nespecifických produktů** (jiné úseky DNA, než které jsme chtěli získat)
- zvýšení specifičnosti např. tzv. **Touch down PCR** – v prvních cyklech vyšší teplota = specifické produkty, ale v s nižším výtěžkem = v dalších cyklech vyšší teplota annealingu = zvýší se výtěžek PCR

- Koncentrace templátové DNA – příliš vysoká konc. DNA může bránit PCR, ředěním se sníží koncentrace inhibitory PCR ve směsi i (např. chemikálie použité při izolaci)

PCR – hlavní faktory ovlivňující průběh reakce

Koncentrace templátové DNA – příliš vysoká konc. DNA může bránit PCR, ředěním se sníží koncentrace inhibitory PCR ve směsi



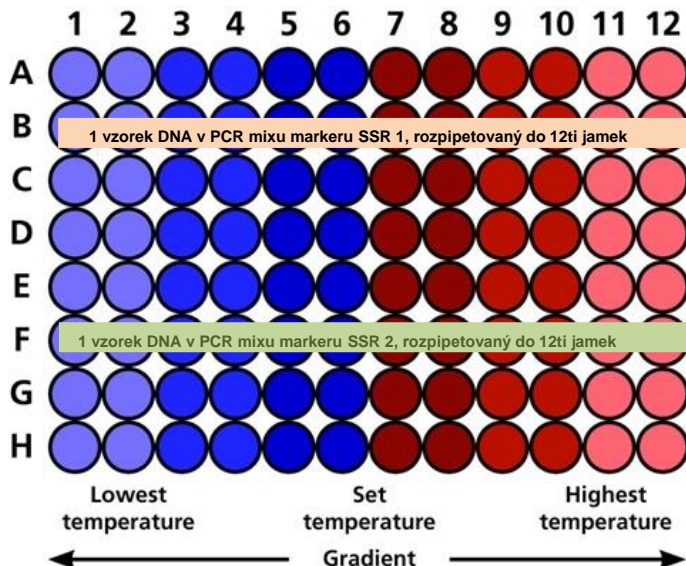
PCR – hlavní faktory ovlivňující průběh reakce

Annealingová teplota

- Teplota nasedání primerů na templátovou DNA
 - Vyšší teplota – vyšší přesnost nasedání odladit
 - Nižší teplota – vyšší pravděpodobnost vzniku nespecifických produktů
- Optimalizace pomocí gradientové PCR
 - PCR na cycleru s gradientovým blokem = umožňuje vytvoření teplotního gradientu
- Např. cycler *Eppendorf Mastercycler Pro S*
 - Nastavení gradientu 1-24°C v rozsahu teplot 30-99°C

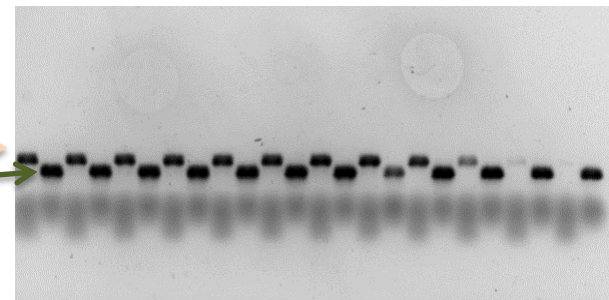


<http://eshop.eppendorfna.com/>



48°C → 64°C

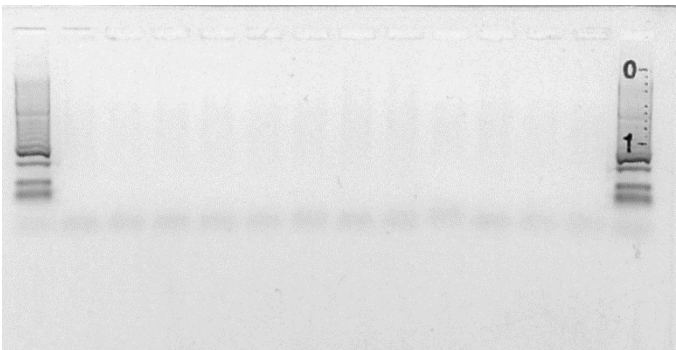
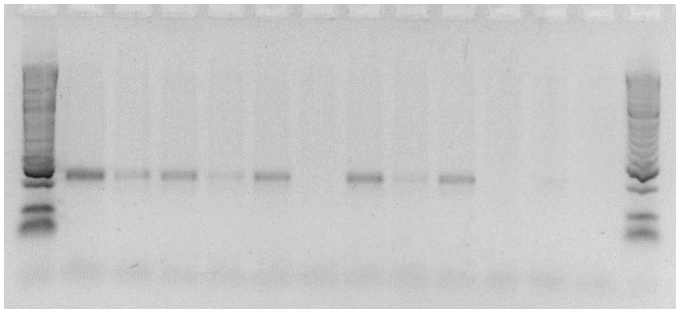
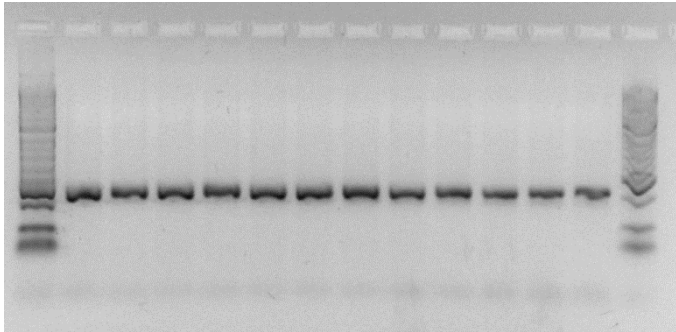
SSR1
SSR2



PCR – hlavní faktory ovlivňující průběh reakce

Polymeráza

– stejné koncentrace dNTP's, MgCl₂, primerů, stejné vzorky DNA, „pouze jiná polymeráza“



Optimalizace reakčních podmínek PCR

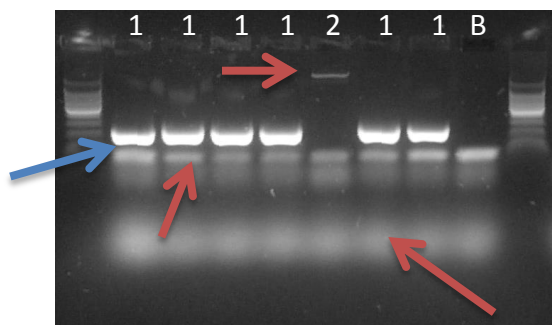
- Změny v protokolu vedoucí k odstranění nežádoucích nespecifit a ke zvýšení množství PCR produktu.
- počet cyklů
- annealingová teplota
- koncentrace Mg^{2+} , polymerázy, templátové DNA
- aditiv (např. DMSO, formamid)

Příklad optimalizace PCR

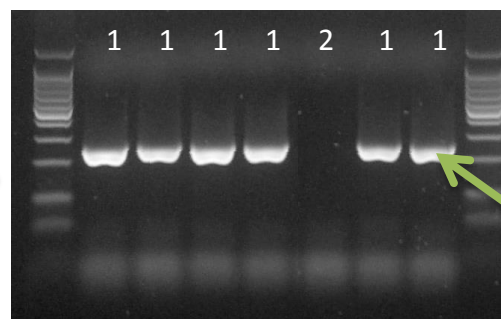
Podmínky podle literatury

PCR po optimalizaci

Marker 1
- produkt 240bp

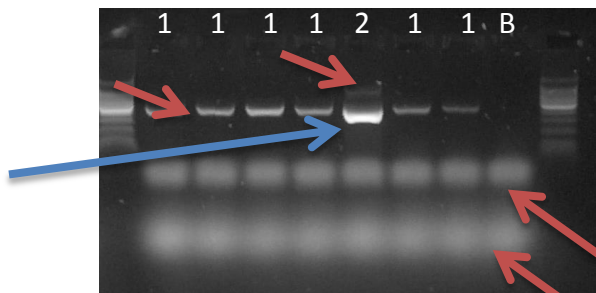


Ta = 62°C, 30 cyklů
20uM primery

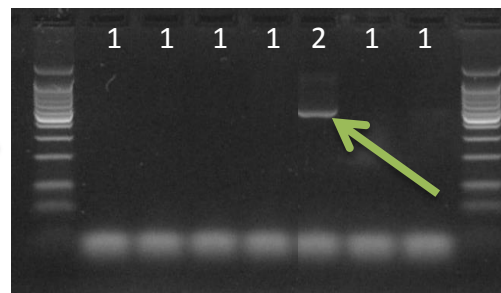


Ta = 67°C, 25 cyklů
10uM primery

Marker 2
- produkt 460bp

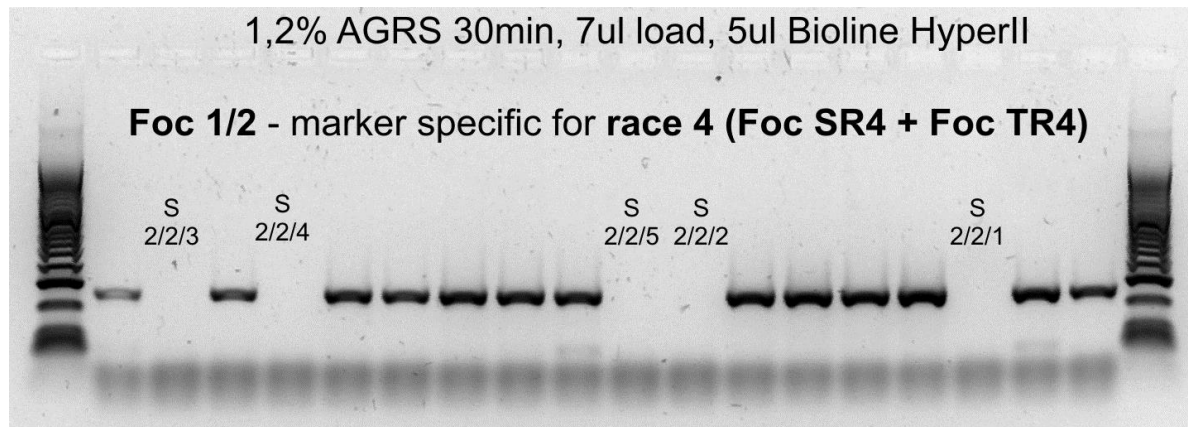


Ta = 66°C, 30 cyklů
20uM primery



Ta = 66°C, 25 cyklů
10uM primery

- Molecular characterization of *Fusarium oxysporum* f. sp. *cubense* race 1 and 4 isolates from Mauritius
- Využití rasově specifických PCR markerů pro determinaci ras fytopatogenních hub



V tomto případě:

Přítomnost/nepřítomnost PCR produktu vypovídá o vlastnostech studovaného organismu (okruhu kultivarů banánovníku, které jsou schopny infikovat).

RAPD – Random Amplified Polymorphic DNA

Variabilita délek náhodně amplifikované DNA

RAPD

Random Amplified Polymorphic DNA

„Variabilita délek náhodně amplifikované DNA“

- PCR s jedním primerem
 - krátká délka, nasedají velice snadno na DNA
- Primer dosedne na různých místech a v různých směrech amplifikuje DNA fragmenty

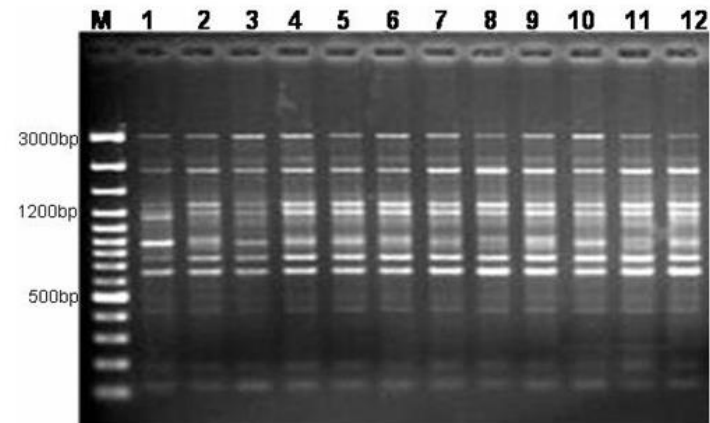
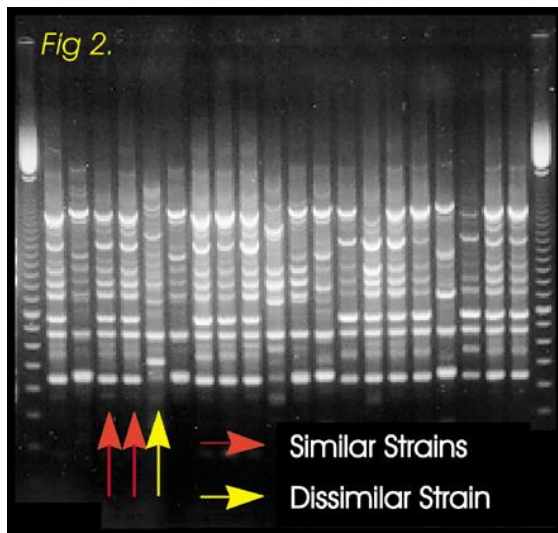
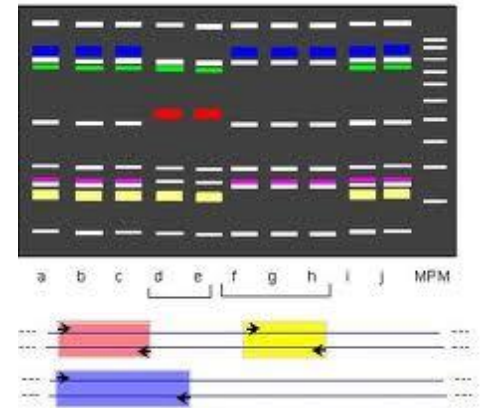
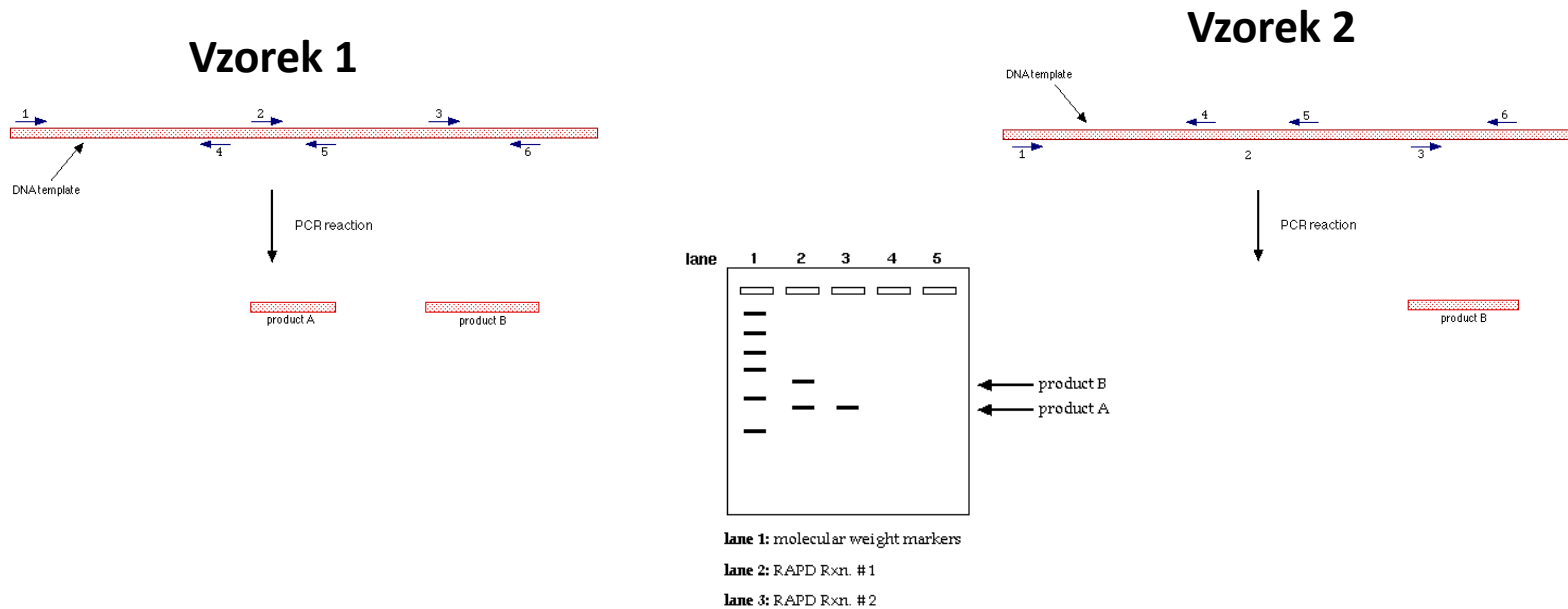


Figure 3. Randomly Amplified Polymorphic DNA (RAPD) amplification pattern obtained for mother plant (lane 1) and long-term micropropagated shoot cultures (lanes 2-12) generated by primer OPN 10. M: GeneRuler™ 100 bp DNA Ladder Plus.

Identifikace např. kmenů bakterií popř. ověřování homogenity linií

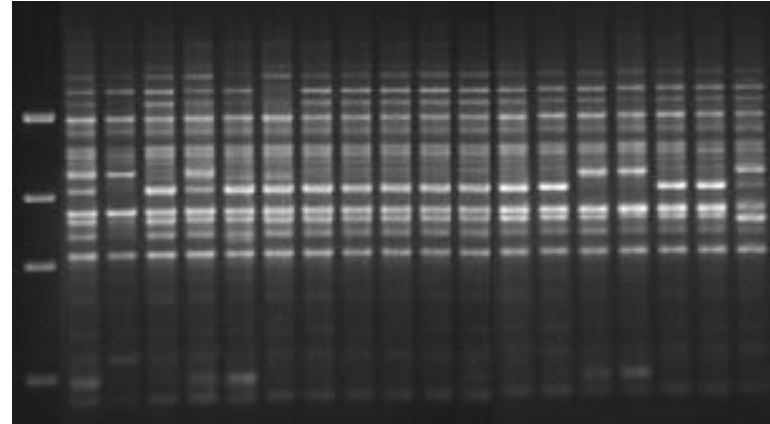
- **RAPD: PCR s jedním arbitrárním (nespecifickým) primerem**



- The primers must anneal in a particular orientation (such that they point towards each other).
- The primers must anneal within a reasonable distance of one another.
- **In this example, only 2 RAPD PCR products are formed:**
- **1)** Product A is produced by PCR amplification of the DNA sequence which lies in between the primers bound at positions 2 and 5.
- **2)** Product B is the produced by PCR amplification of the DNA sequence which lies in between the primers bound at positions 3 and 6.
- Note that no PCR product is produced by the primers bound at positions 1 and 4 because these primers are too far apart to allow completion of the PCR reaction.
- Note that no PCR products are produced by the primers bound at positions 4 and 2 or positions 5 and 3 because these primer pairs are not oriented towards each other.

Výhody

- Velmi levné
- Rychlé a snadné = není třeba znát sekvenci primeru, stačí velmi málo DNA, poměrně dost fragmentů...



Nevýhody převažují:

- Omezena hodnota informace (dominantní marker)
- Nereproducibilní

DNES TREND NEUŽÍVAT **RAPD** MARKERY

GENETIC VARIATION AMONG STRAINS OF THE TOXIC DINOFLAGELLATE *GYMNODINIUM CATENATUM* (DINOPHYCEAE)¹

Christopher J. S. Bolch²

School of Plant Science, University of Tasmania, GPO Box 252-55, Hobart, Tasmania, 7001, Australia

Susan I. Blackburn

CSIRO Marine Research, GPO Box 1538, Hobart, Tasmania, 7001, Australia

Gustaaf M. Hallegraeff and René E. Vaillancourt

School of Plant Science, University of Tasmania, GPO Box 252-55, Hobart, Tasmania, 7001, Australia



skepticwonder.fieldofscience.com

Toxická obrněnka způsobující vodní květy u JV Tasmánii

Předpoklad

- zavlečení toxických kmenů v odpadní (vyvažovací) vodě tankerů plujících z Japonska nebo Španělska a následné šíření podél pobřeží Tasmánie a do vnitrozemí

Metoda: RAPD

Výsledek:

- tři nepřekrývající se clustery: **Austrálie**, **Japonsko**, **Španělsko/Portugalsko**
- vzorky z Austrálie stejně podobné vzorkům z JPN jako ze ESP/PRT
- původ Australských vzorků zůstává nejasný

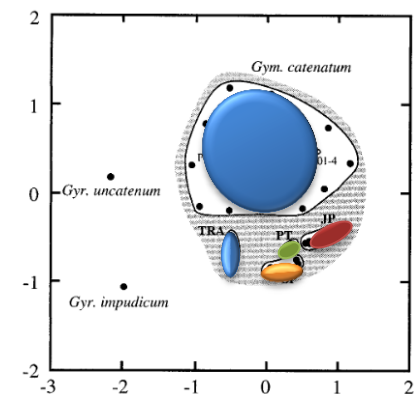
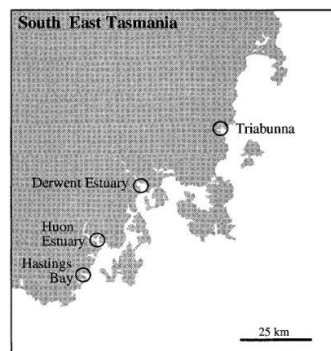
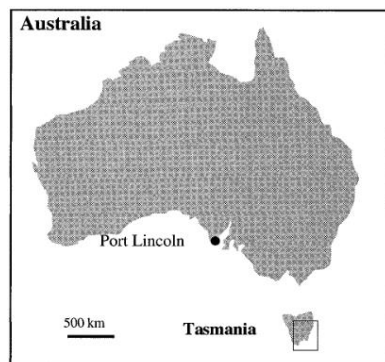


FIG. 2. Two-dimensional MDS plot of RAPD-PCR data for *G. catenatum* strains (shaded area) and *G. uncatenatum* and *G. impudicum* (outgroup species). *G. catenatum* clusters (bounded by solid lines): Derwent Estuary, Huon Estuary, and Hastings Bay (DE/HU); Triabunna (TRA); Japan (JP); Spain (SP); and Portugal (PT). Kruskal stress = 0.249.

AFLP - Amplified Fragment Length Polymorphism

Polymorfismus délky amplifikovaných fragmentů

AFLP – princip, 1. fáze = restrikce

Princip - metoda založená na:

- restrikci DNA dvěma enzymy
- ligaci adaptorů na vzniklé „lepivé“ konce fragmentů DNA
- selektivní namnožení jen některých fragmentů
- vizualizace proužků na gelu

Postup – relativně složitý (4 fáze)

1. RESTRIKCE

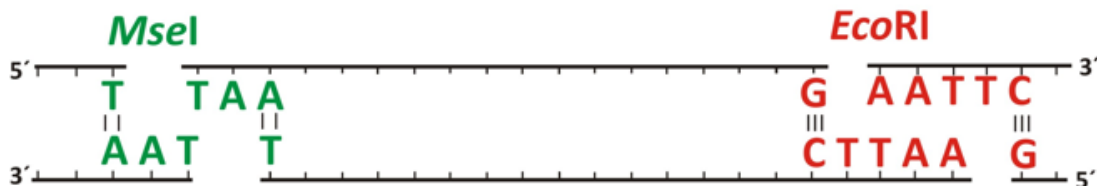
- specifické rozštěpení DNA restrikčními endonukleázami

MseI - rozpoznává 4bp dlouhou sekvenci (TTAA)

EcoRI - rozpoznává 6bp dlouhou sekvenci (GAATTC)

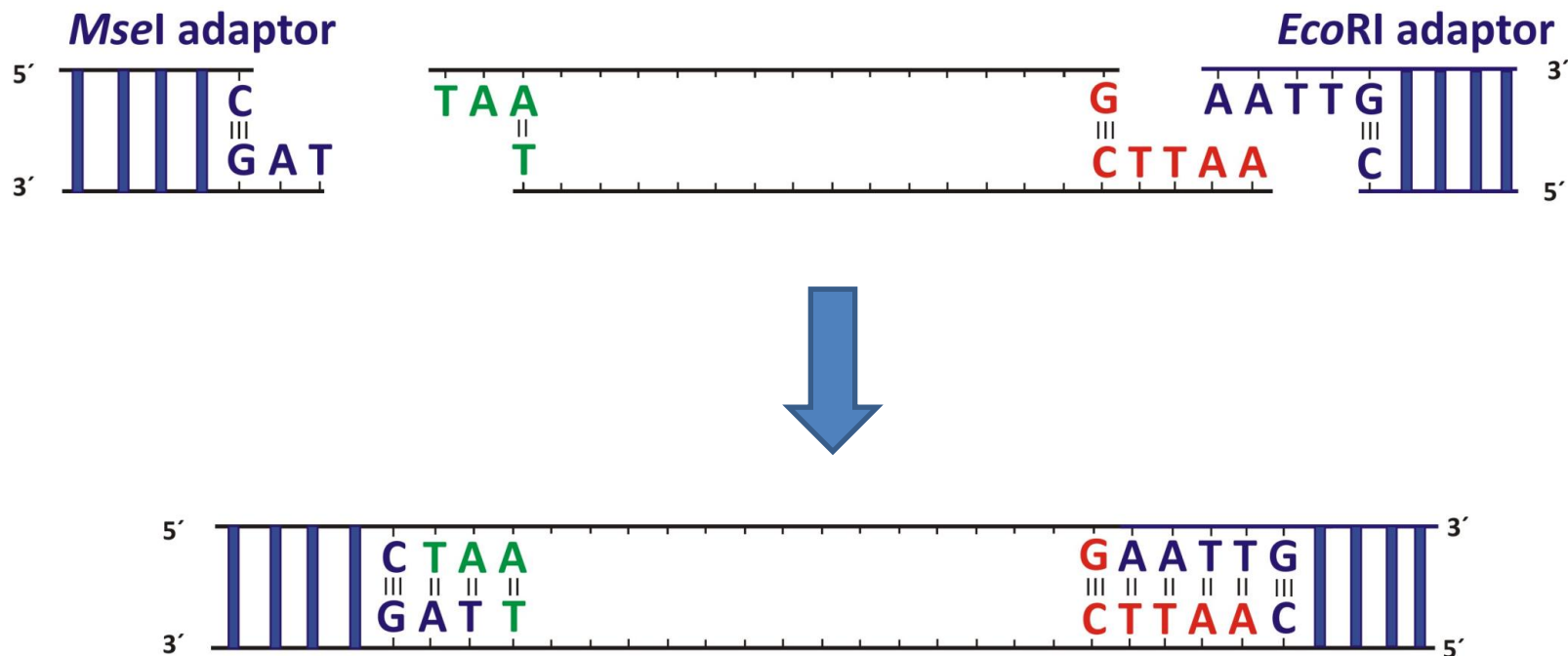
(Variabilita je dána mutací v restrikčním místě insercí - delecí mezi restrikčními místy)

- získáme množství fragmentů s „lepivými konci“ = „sticky ends“



AFLP – 2. ligace

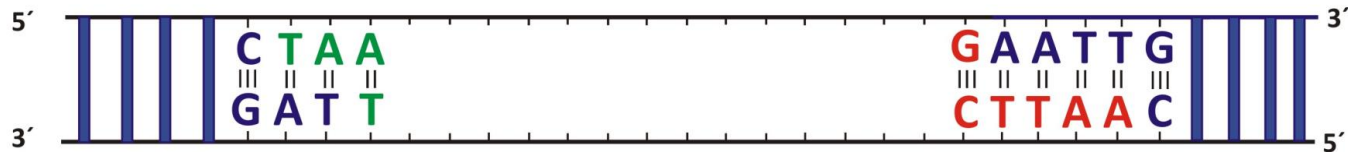
- pomocí T4 ligázy k fragmentům přidány adaporty = synteticky připravené kousky DNA, jejich sekvenci známe
- známe sekvence konců všech fragmentů ... můžeme je amplifikovat PCR



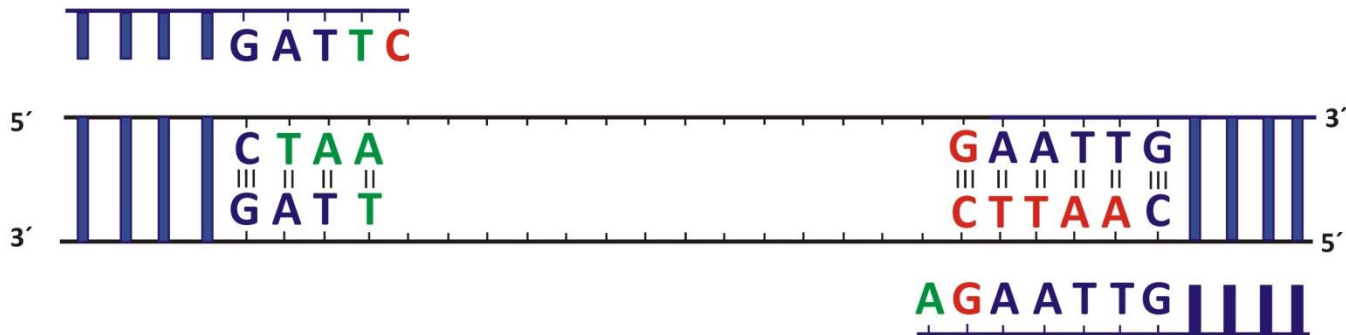
AFLP – 3. preselektivní amplifikace

→ klasická PCR s primery komplementárními k sekvenci adaptorů, navíc 1 nukleotid směrem dovnitř amplifikovaného úseku

= selekce jen ca. 1/4 fragmentů (fragmenty komplementární k primerům + fragmenty *EcoRI*- *MseI*)



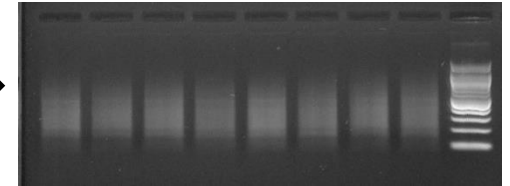
MseI primer



EcoRI primer

AFLP – 4. selektivní amplifikace

-fragmentů je stále mnoho pro objektivní vyhodnocení →

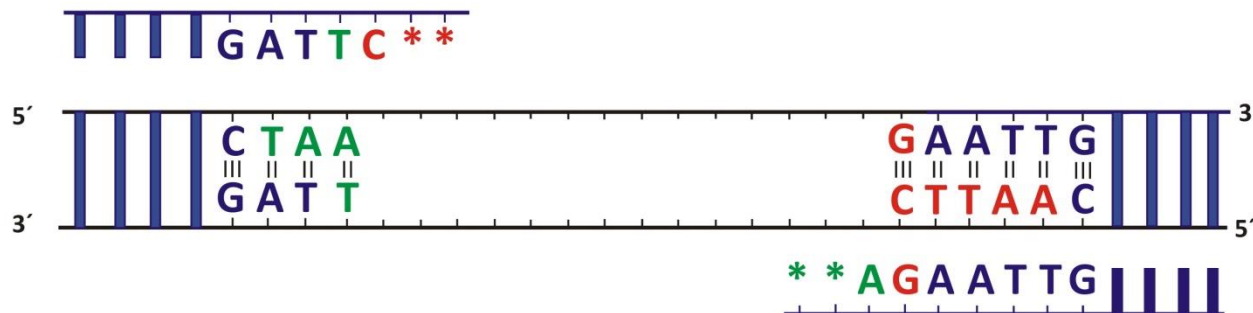


- další redukce jejich počtu použitím primerů se 3mi selektivními nukleotidy (s přesahem "dovnitř" amplifikovaných fragmentů)

- tento krok se provádí opakovaně s různými primerovými kombinacemi (primery se liší počtem a typem použitých nukleotidů)

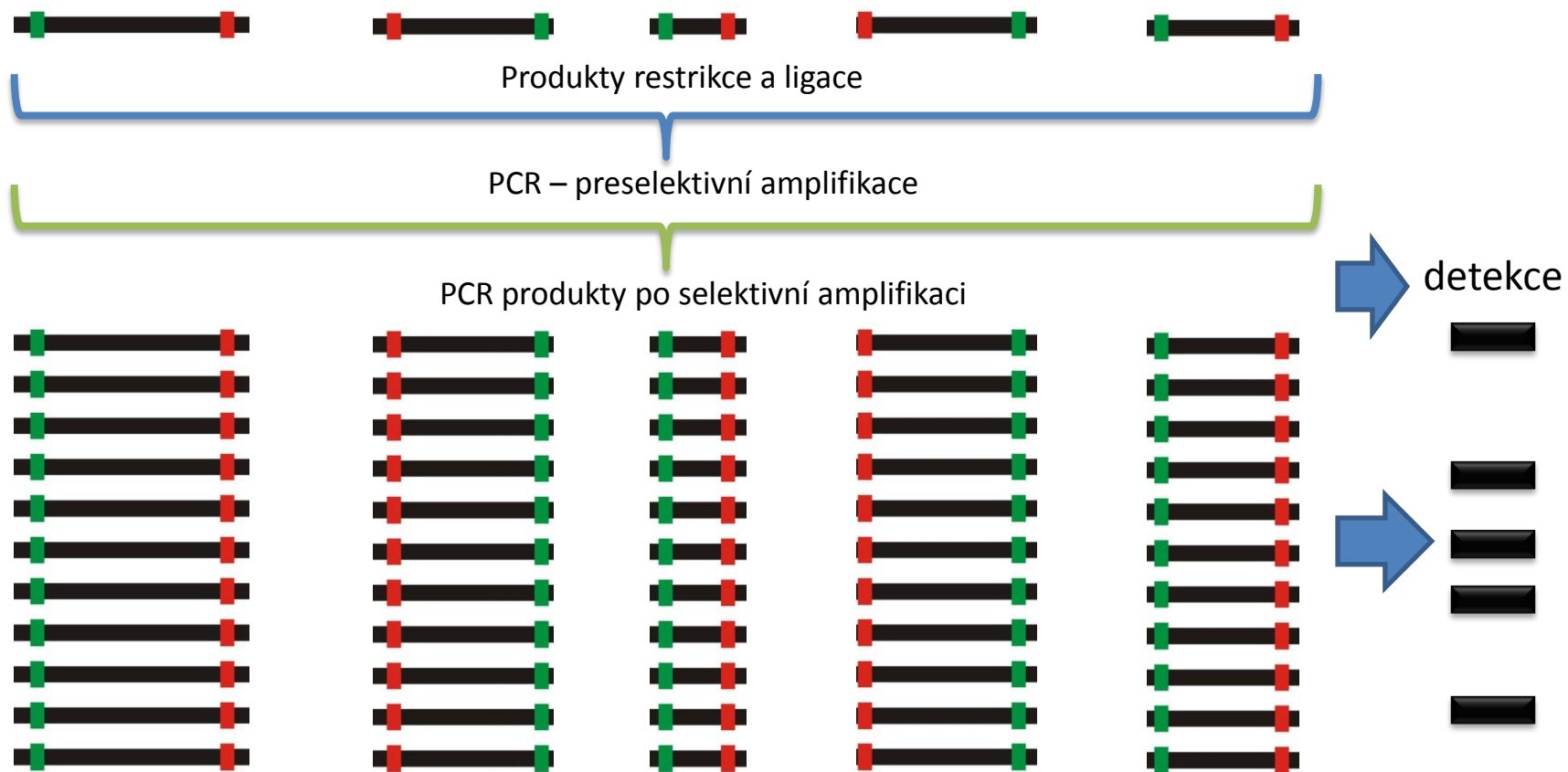
- Typ použitých primerů se musí otestovat – výsledkem by měl být spolehlivě hodnotitelný záznam (optimální počet bandů, pokud možno polymorfních)

MseI primer

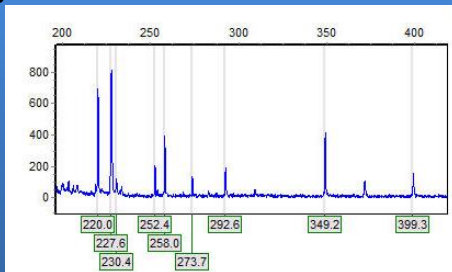


EcoRI primer

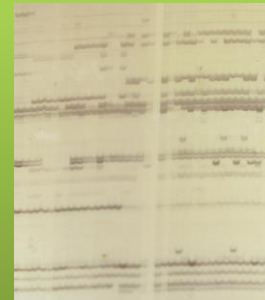
AFLP – separace a detekce fragmentů



- primery jsou fluorescenčně značeny → sekvenátor



- primery bez značení → PAGE elektroforéza

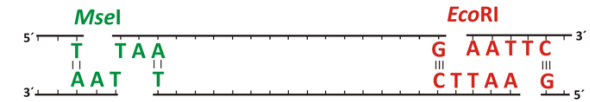


AFLP - shrnutí

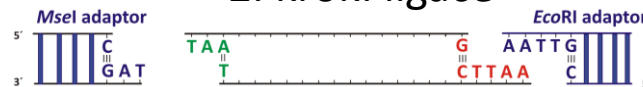
Princip

- metoda založená na restrikci DNA dvěma enzymy
- selektivní namnožení jen některých fragmentů pomocí PCR
- vizualizace proužků na gelu/sekvenátoru

1. krok: restrikce

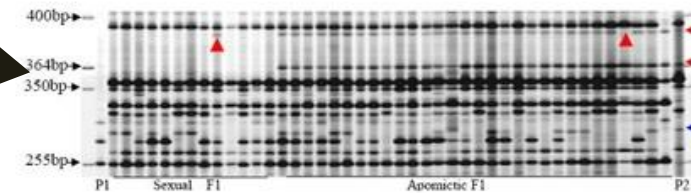
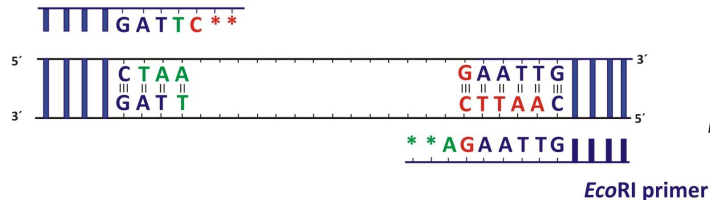


2. krok: ligace



3. & 4. krok: PCR – preselektivní a následně selektivní amplifikace

MseI primer



Statistické zpracování dat

presence/absence bandů (1-0)

AFLP – realizace



1: Restrikce

- příprava restrikčního mixu (EcoI, MseI, buffer, voda)
- přidat k DNA
- inkubovat 2 hod při 37°C

2: Ligace

- příprava ligačního mixu (LIGAZA, BUFF, ADAPTORY, VODA)
- přidat k roztoku z předchozího kroku
- inkubovat 2,5 hod při 37°C



3: Preamplifikace

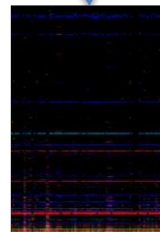
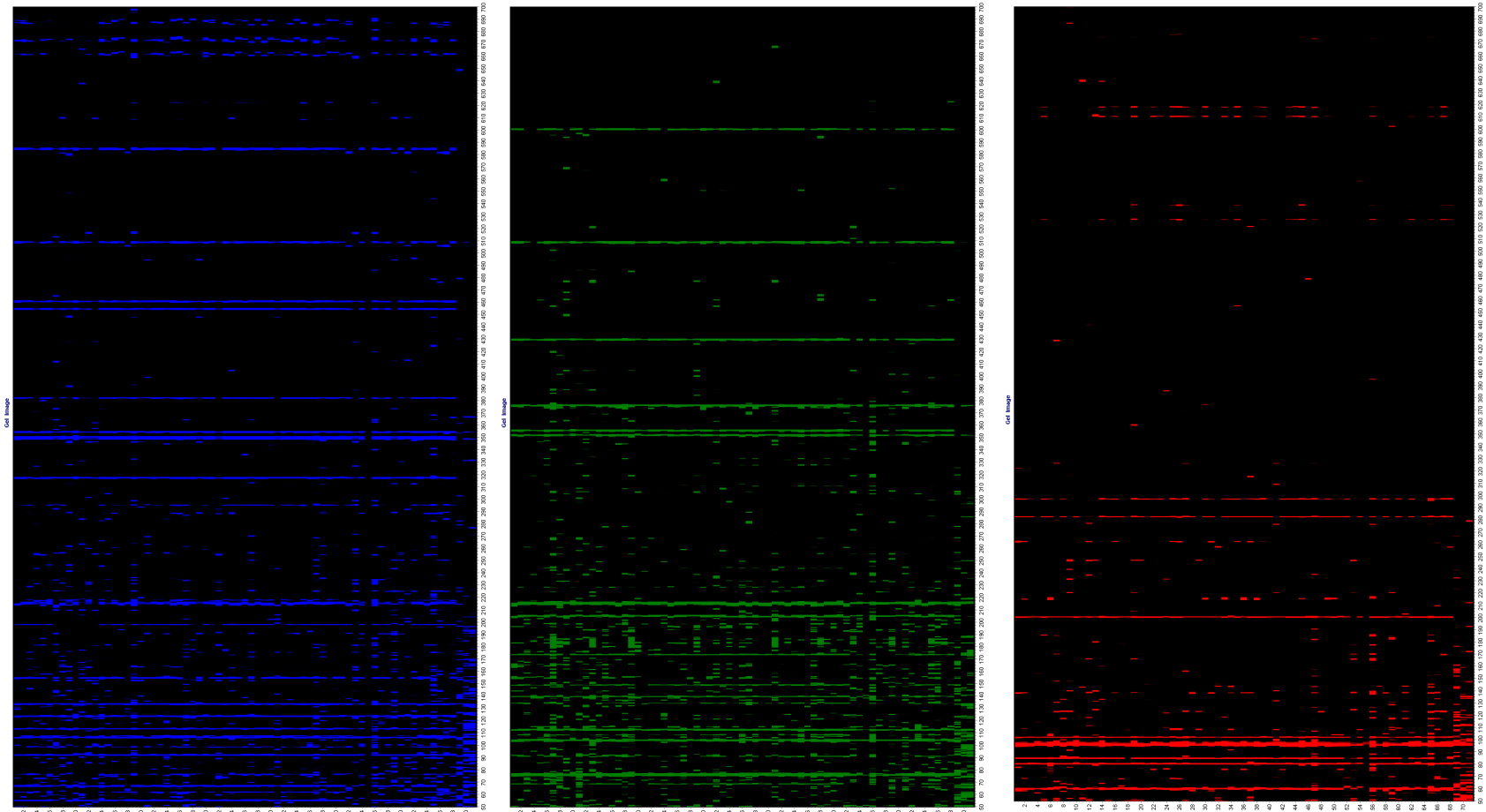
- klasická PCR
- ligační roztok jako matrice

Kroky 1-3 se provedou pouze 1x, dál pracujeme s preamplifikátem

4: Amplifikace

- klasická PCR
- matrice = preamplifikát
- měníme kombinace primerů = detekujeme jiné fragmenty

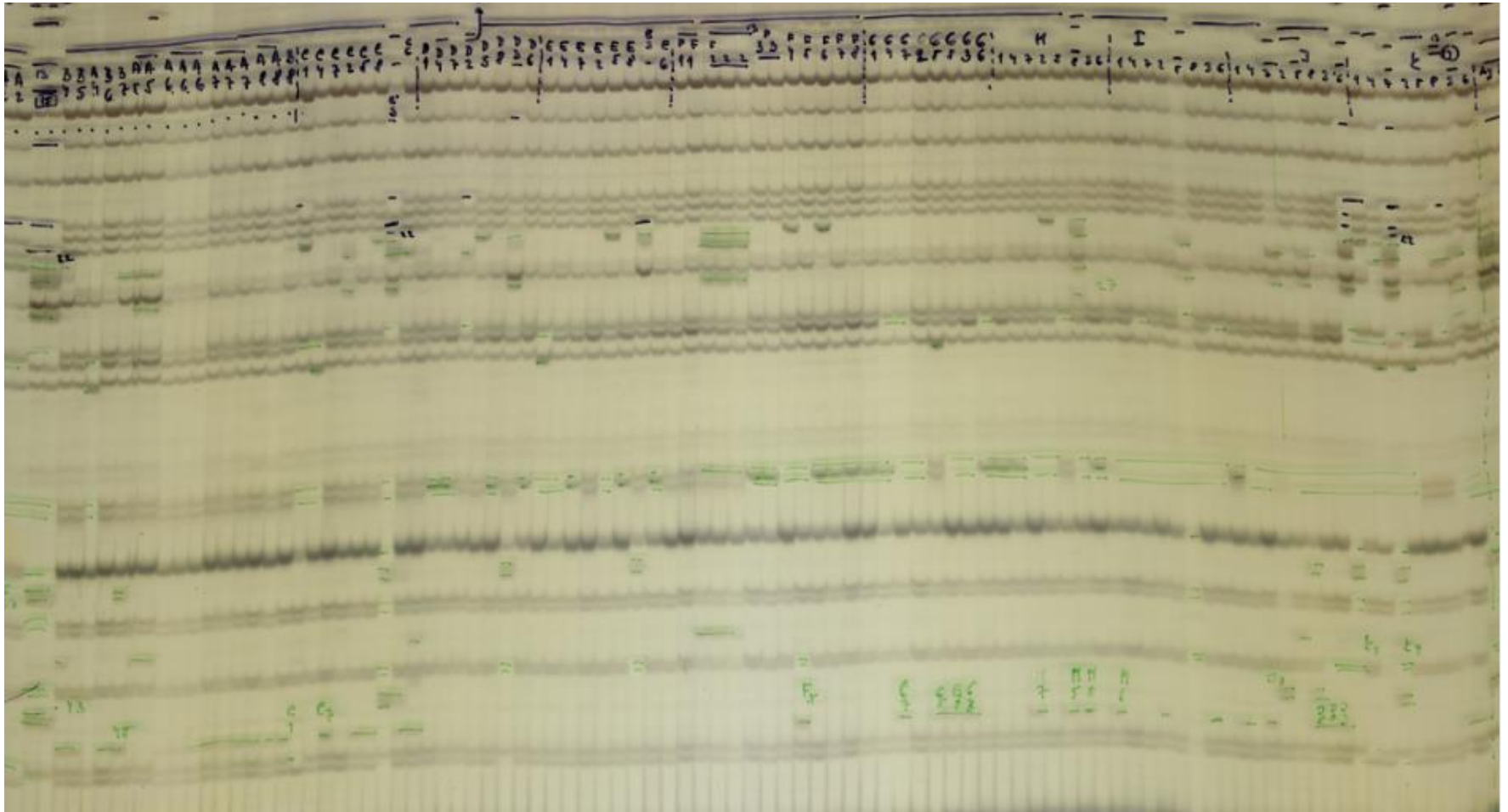
AFLP – fluorescenčně značené primery, separace na sekvenátoru



Výhoda:

- multiplex – separace více PCR v jednom kroku
- všechny vzorky separovat 2x (2 nezávislé PCR)
- předejít posunu fragmentů při vyhodnocování
- duplikáty, blanky

AFLP – separace na PAA ELFO a detekce barvením stříbrem



Výhody

x

Nevýhody

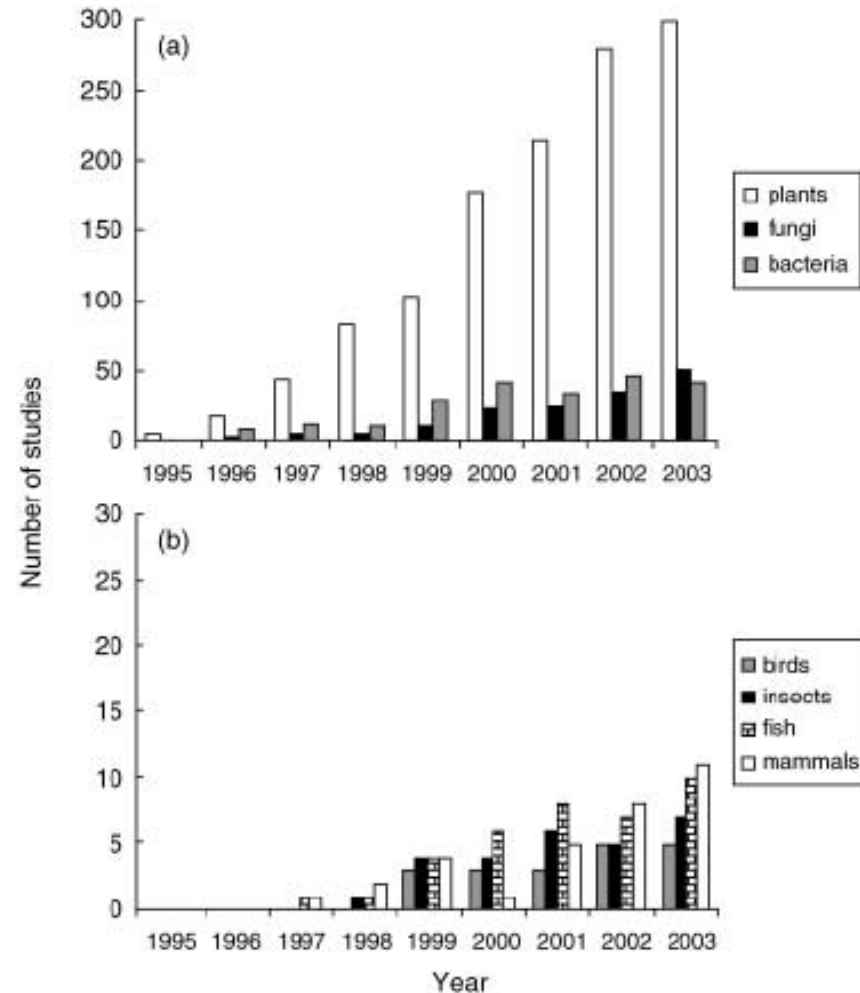
- Rychlé získání polymorfismu
 - Proto relativně levná
 - Citlivá (na bodové mutace)
 - Velmi snadno opakovatelná
 - Celý genom
 - 50–100 fragmentů najednou
 - Není třeba znát sekvenci primeru
 - Pokud je metodika zavedená na pracovišti a máme DNA v dostatečné kvalitě – AFLP lze aplikovat na jakýkoliv organismus
- DOMINANTNÍ MARKER
 - Problematická je homologie fragmentů mezi vzdálenějšími taxony
(předejít: snížit výsledný počet bandů)
 - Nemožnost srovnání časově a prostorově oddělených AFLP analýz
 - Správně: připravit celý soubor vzorků a vše zpracovat naráz.

Předpokladem správně provedené analýzy je použití kvalitní, nefragmentované DNA

Hlavní oblasti využití AFLP

- Analýza struktury a genetické variability populací
- Taxonomické a evoluční vztahy blízce příbuzných taxonů
- Studium hybridizace a polyploidie
- Fylogeografie

Počet „AFLP“ článků
u různých skupin organismů
Bensch & Åkesson 2005



AFLP – zpracování dat



1. Vytvoření „binární matice“
2. Provedení výpočtu hodnot indexů genetické variability
3. Provedení vizualizace vzájemných vztahů mezi analyzovanými jedinci
 - Fylogenetické stromy (UPGMA, NJ)
 - Mnohorozměrné analýzy (PCoA)
 - Bayesiánské analýzy (BAPS, STRUCTURE)

AFLP – zpracování dat

1. Vytvoření „binární batice“

Přepis získané informace z elektroforetogramu do digitální podoby




















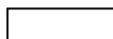
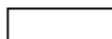
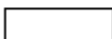

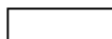









přítomnost bandu = 1

absence bandu = 0

jedinec A: genotyp 11010110000

jedinec B: genotyp 10100101000

jedinec C: genotyp 10011100111

A	B	C		A	B	C
			→	1	1	1
			→	1	0	0
			→	0	1	0
			→	1	0	1
			→	0	0	1
			→	1	1	1
			→	1	0	0
			→	0	1	0
			→	0	0	1
			→	0	0	1
			→	0	0	1

2. Nejčastěji využívané přístupy hodnocení genetické variability populací

a) Výpočty hodnot indexů genetické variability:

- Polymorfismus = podíl polymorfních lokusů P (nebo PLP)
- Vážená očekávaná heterozygotnost H_e
 - různé přístupy výpočtu (viz dál)
- Fixační index F_{st}
- AMOVA (analýza molekulární variance)

AFLP – zpracování dat

Podíl polymorfních lokusů P (nebo PLP) *Které bandy jsou polymorfní?????*

Monomorfní lokus u všech jedinců stejná alela (band)



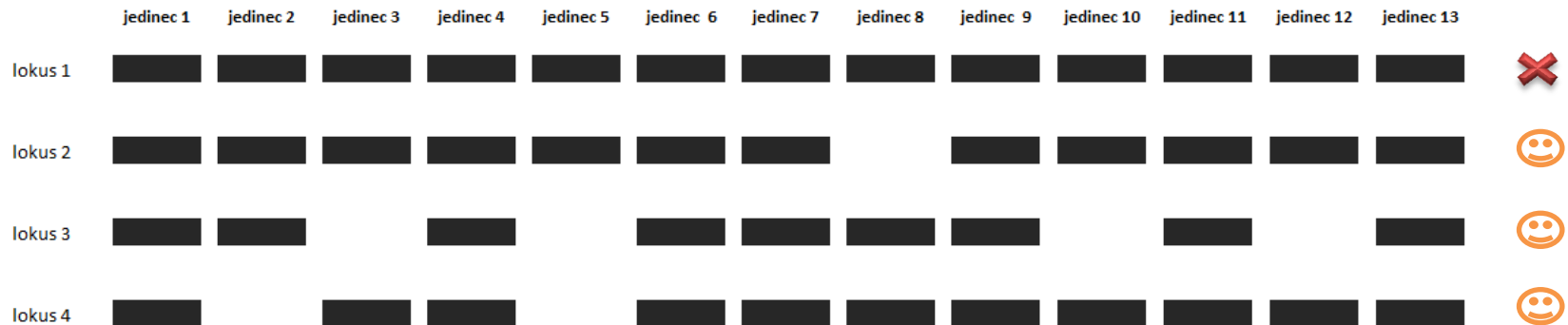
AFLP – zpracování dat

Podíl polymorfních lokusů P (nebo PLP) Které bandy jsou polymorfní?????

Polymorfní lokus na hladině 1%

..... bandy, kdy alespoň u jednoho jednice fragment chybí

př: u populace 100 jedinců = lokus přítomný u 99 a méně jedinců
- někdy uváděno jako „**segregující bandy**“



Podíl polymorfních lokusů P (nebo PLP)

Polymorfní lokus na hladině 5%

- Výpočet z frekvence alel: **alely s frekvencí $0,95 < f > 0,05$**
- **Polymorfní lokusy podle metody Lynch & Milligan:**
- za polymorfní jsou považovány **lokusy, kde aspoň u 4 jedinců **band** chybí**



AFLP – zpracování dat

Podíl polymorfních lokusů P (nebo PLP) – nač to je dobré

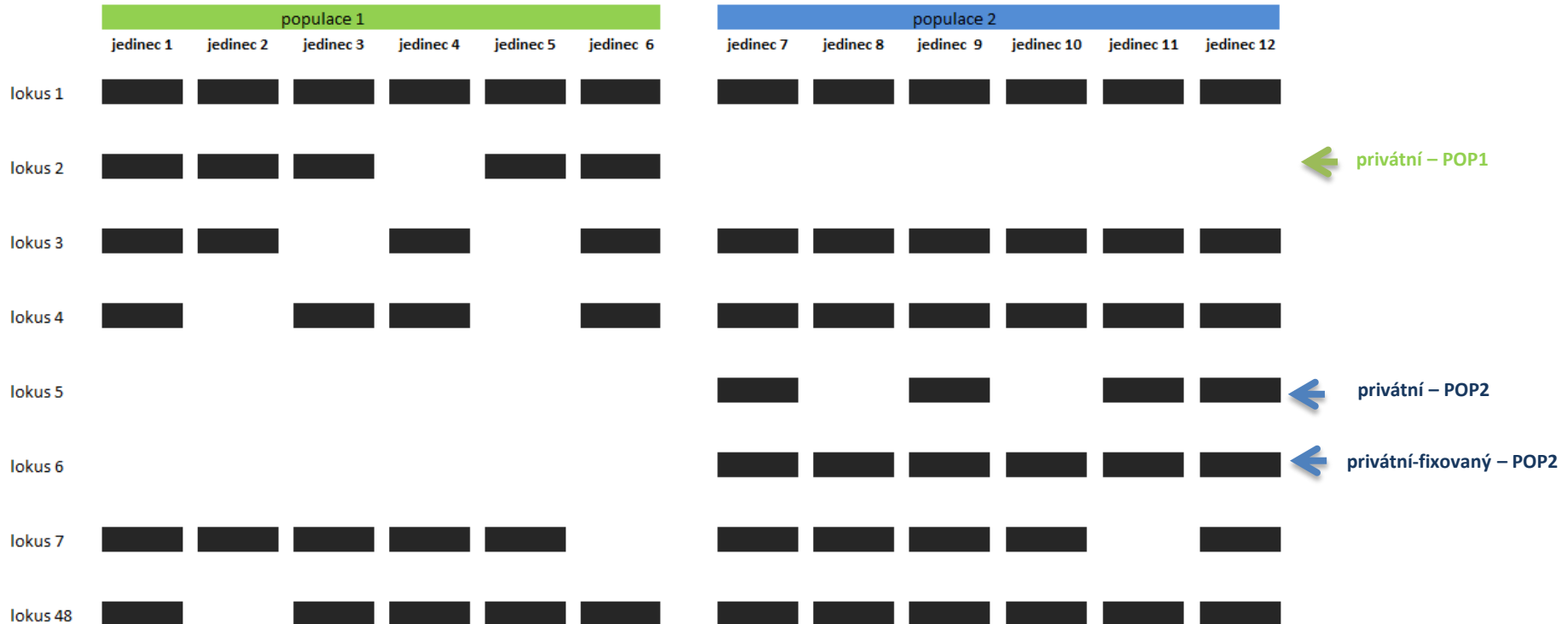
- a) Základní informace o variabilitě používaných primerových kombinací
- b) Důležitější:
 - Informace o celkové variabilitě studovaného souboru vzorků
 - Informace o variabilitě uvnitř dílčích populací/srovnávaných skupin



AFLP – zpracování dat

Privátní a privátní-fixované bandy

- **privátní bandy** – charakteristické pro část vzorků z určité skupiny/populace, u ostatních skupin/populací nejsou
- **privátní-fixované** – charakteristické pro všechny vzorky z určité skupiny/populace, u ostatních skupin/populací nejsou





Godfrey Hardy



Wilhelm Weinberg

Odhad heterozygotnosti H_e z AFLP dat

- AFLP nedovede odlišit heterozygota (1/0) od homozygota (1/1)

Z prezence a absence dat lze odvodit hodnoty heterozygotnosti **nepřímo**, pokud připustíme platnost Hardy-Weinbergova zákona.

- Celkové genotypové složení populace: $p^2 + 2pq + q^2 = 1$
- Zároveň $2pq = H_e$... a ... $p + q = 1$... tedy: **$H_e = 1 - p^2 - q^2 = 1 - p^2 - (1 - p^2)$**

- Tj: pokud známe frekvenci dominantní alely p , můžeme dopočítat frekvenci recesivní alely q a potažmo i hodnotu heterozygotnosti H_e

- Nutno si uvědomit, že **maximální hodnota heterozygotnosti**, takto nepřímo odvozená, je **$H_e = 0,5$**

Proč?: $H_e = 2pq$ $p + q = 1$

např. při	$p = 0.9$	$2 \times 0,9 \times (1-0,9) = H_e = 0,18$
	$p = 0.5$	$2 \times 0,5 \times (1-0,5) = H_e = 0,50$
	$p = 0.1$	$2 \times 0,1 \times (1-0,1) = H_e = 0,18$

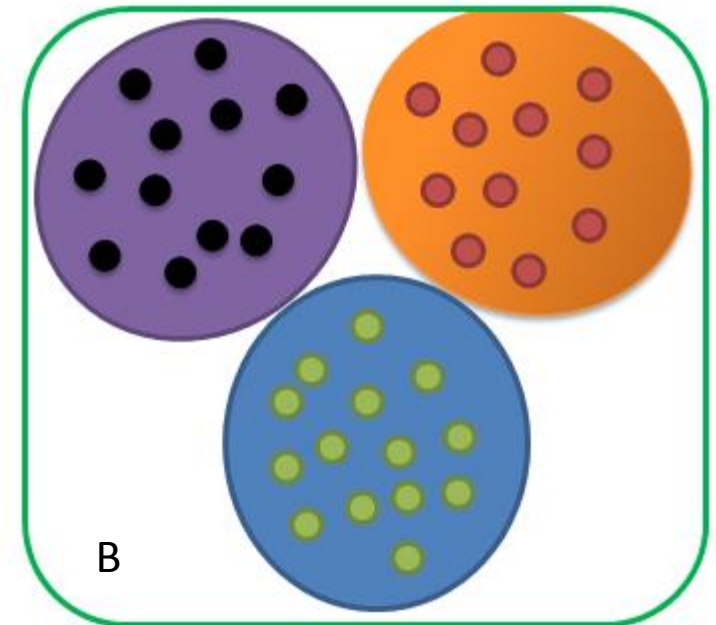
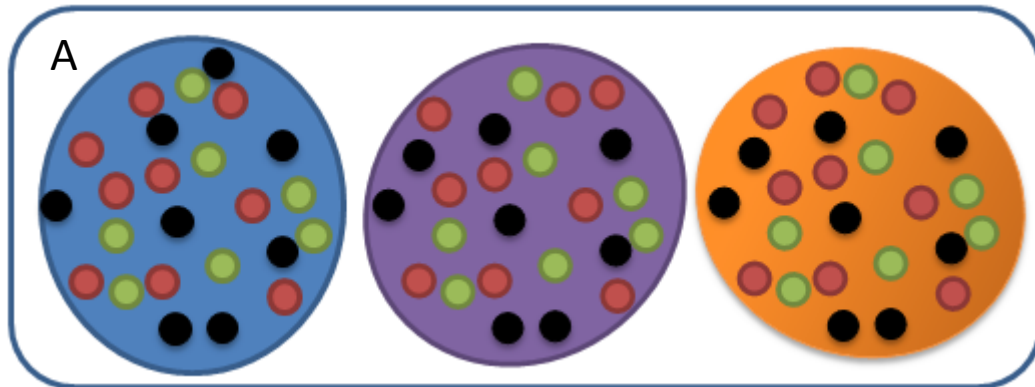
Zprůměrováním H_e ze všech lokusů získáme

PRŮMĚRNOU GENETICKOU DIVERZITU populace, druhu.

AFLP – zpracování dat

- Fixační index *F_{st}*

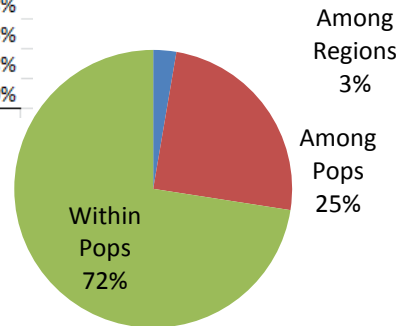
- posouzení izolovanosti jednotlivých (sub)populací
- Novější přístupy porovnávají rozdílné zastoupení alel v subpopulacích vůči průměrné hodnotě pro celý studovaný soubor
- Nabývá hodnot 0 – 1
- max. hodnota 1 = úplná diferenciace = subpopulace mají zcela odlišné, fixované alely, které jsou charakteristické pro danou subpopulaci
 - 0 – 0.05 malá diferenciace (zanedbatelná, nejsou rozdíly mezi subpopulacemi ve frekvenci alel)
 - 0.05 – 0.15 střední
 - 0.15 – 0.25 vysoká
 - > 0.25 velmi vysoká



AFLP – zpracování dat

Summary AMOVA Table

Source	df	SS	MS	Est. Var.	%
Among Regions	2	179,809	89,904	0,618	3%
Among Pops	9	518,864	57,652	5,704	25%
Within Pops	81	1353,327	16,708	16,708	73%
Total	92	2052,000		23,030	100%



AMOVA

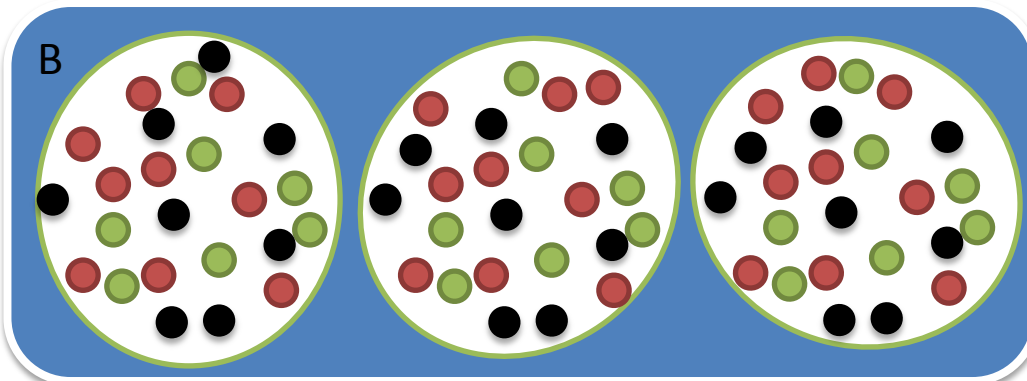
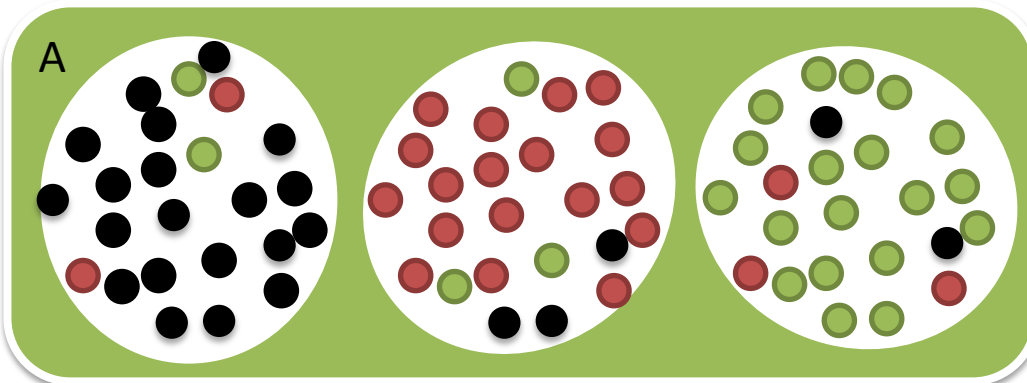
- Analýza Molekulární Variance
- Multivariační metoda popisující distribuci genetické variability do několika úrovní:
- variabilita mezi jedinci
- variabilita mezi populacemi
- Popř. (variabilita mezi regiony, druhy)

Kvantifikuje jaký podíl celkové variability studovaného souboru představuje variabilita uvnitř populací a mezi populacemi.

Výstupem je mj. i Φ_{ST} představující ekvivalent ***Fst***

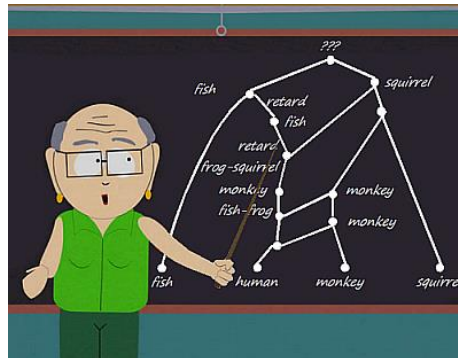
Software:

GeneAEx, FAMD, AFLP-surv,



Postup konstrukce stromu

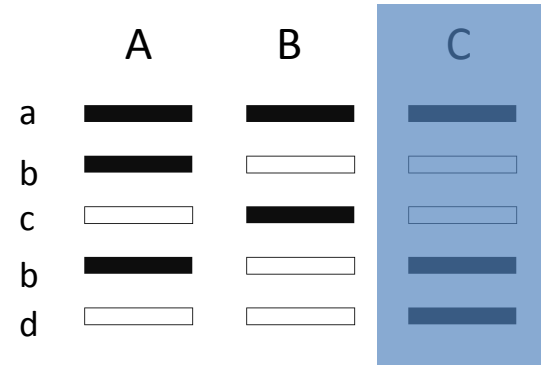
1. *Výpočet matice genetických vzdáleností*
2. *Vlastní konstrukce stromu (volba vhodné shlukovací metody)*
3. *Ověření správnosti konstrukce (bootstrapping)*



Indexy genetické podobnosti/vzdálenosti (binární data)

		Individual <i>i</i>	
		Band presence 1	Band absence 0
Individual <i>j</i>	Band presence 1	<i>a</i>	<i>b</i>
	Band absence 0	<i>c</i>	<i>d</i>

where $n = a + b + c + d$



Jaccard coefficient (Jaccard 1908)

$$\frac{a}{a + b + c}$$

The Jaccard coefficient only takes into account the bands present in at least one of the two individuals, and is therefore unaffected by homoplastic absent bands (when the absence of the same band is due to different mutations).

Dice coefficient (Dice 1945)

The Dice coefficient is equivalent to the Nei and Li coefficient (Nei & Li 1979) and the Sørensen coefficient (Sørensen 1948).

$$\frac{2a}{2a + b + c}$$

Comparable to the Jaccard coefficient, the Dice coefficient gives more weight to the bands present in both individuals. It thus lays the emphasis on the similarity between individuals, rather than on their dissimilarity.

Simple-matching coefficient (Sokal & Michener 1958)

$$\frac{a + d}{a + b + c + d} = \frac{a + d}{n}$$

The simple-matching coefficient maximizes the amount of information drawn from an AFLP profile by considering all scored loci. Double-band absence and double band presence are given the same biological importance, which may not be adequate in case of frequent band absence homoplasy. This coefficient has interesting Euclidean metric properties that allow its use in an analysis of molecular variance (AMOVA; Excoffier *et al.* 1992).

FreeTree - [C:\Users\kitner\Documents\= pracon\p\program\program\freetree\2011 Adenophora\2011-12-07_Adenophora_3PC.txt]

File Window Help

Data input Distance/similarity matrix Reference tree Trees Resampling methods

B - Jaccard

Tree-construction method:
☒ UPGMA ☐ Neighbor-joining

	SC1	SC4	SC7	2a	5a	8a	11a	14a	SC2	SC5	SC8	3a	6a	9a	12a	MP-1	SC3	SC6	1a	4a	7a	10a	13a	MP-2	MP-3
SC1		0.81818	0.85106	0.38889	0.48980	0.52174	0.48000	0.40000	0.89130	0.65957	0.66667	0.47826	0.44681	0.47917	0.48077	0.52000	0.86667	0.88636	0.46939	0.43137	0.45833	0.39623	0.50943	0.42857	0.42857
SC4	0.81818		0.76596	0.36538	0.56818	0.60976	0.59556	0.48980	0.76596	0.76190	0.68889	0.48837	0.52381	0.52273	0.48980	0.50007	0.86118	0.83721	0.54545	0.46809	0.43478	0.42857	0.52000	0.40741	0.46
SC7	0.85106	0.76596		0.36842	0.46154	0.52083	0.48077	0.40351	0.87500	0.62000	0.59615	0.44898	0.42000	0.45098	0.45455	0.49051	0.93333	0.86957	0.47059	0.38182	0.46000	0.40000	0.48214	0.43103	0.39
2a	0.38889	0.36538	0.36842		0.52381	0.52500	0.51163	0.54545	0.41818	0.43750	0.51064	0.55263	0.55263	0.51220	0.54545	0.59524	0.38889	0.42308	0.53659	0.64103	0.52500	0.41304	0.47917	0.51064	0.57
5a	0.48980	0.56818	0.46154	0.52381		0.73529	0.65789	0.60976	0.46154	0.45652	0.46809	0.58333	0.83871	0.62162	0.57143	0.58537	0.48980	0.50000	0.90625	0.72222	0.55263	0.50000	0.53333	0.46809	0.55
8a	0.52174	0.60976	0.52083	0.52500	0.73529		0.66667	0.57500	0.52083	0.48837	0.46667	0.54286	0.68750	0.72727	0.53659	0.5118	0.59556	0.56818	0.75758	0.59459	0.55556	0.57895	0.60976	0.53488	0.61
11a	0.48000	0.55556	0.48077	0.51163	0.65789	0.66667		0.55914	0.48077	0.51111	0.48936	0.48718	0.61111	0.52500	0.55914	0.50000	0.51020	0.52083	0.67568	0.46512	0.50000	0.56098	0.55556	0.52174	0.58
14a	0.40000	0.48980	0.40351	0.54545	0.60976	0.57500	0.55914		0.42857	0.44898	0.48980	0.45238	0.48780	0.52381	0.62791	0.60465	0.40000	0.43396	0.58537	0.65000	0.53659	0.48889	0.55319	0.58636	0.58
SC2	0.89130	0.76596	0.87500	0.41818	0.46154	0.52083	0.48077	0.42857		0.62000	0.66000	0.44898	0.42000	0.45098	0.45455	0.51923	0.85106	0.86957	0.47059	0.43396	0.46000	0.40000	0.48214	0.43103	0.39
SC5	0.65957	0.76190	0.62000	0.43750	0.45652	0.48837	0.51111	0.44898	0.62000		0.76190	0.44186	0.47619	0.47727	0.51064	0.45833	0.65957	0.71111	0.43478	0.42553	0.48837	0.38776	0.45098	0.42308	0.48
SC8	0.66667	0.68889	0.59615	0.51064	0.46809	0.46667	0.48936	0.48980	0.66000	0.76190		0.48837	0.48837	0.48889	0.48980	0.50000	0.63265	0.68085	0.41667	0.53333	0.43478	0.40000	0.46154	0.38182	0.50
3a	0.47826	0.48837	0.44898	0.55263	0.58333	0.																			

Metody tvorby fylogenetických stromů

	<u>typ dat</u>	
	vzdálenosti/podobnosti	DNA sekvence (nebo jiné znaky)
<u>shlukovací</u> algoritmus	UPGMA <small>Unweighted Pair Group Method with Arithmetic mean</small> neighbor-joining tree	
<u>optimalizační</u> kritérium	minimum evolution	maximum parsimony maximum likelihood Bayesova analýza

UPGMA - založena na předpokladu, že stromy jsou aditivní a všechny taxy jsou stejně vzdálené od kořene. Jedná se čistě o analýzu fenotypu, nebere v potaz např. rozdílnou délku evoluce.

neighbor-joining pracuje s kritériem minimální evoluce - hledá topologii s nejmenší celkovou délkou všech větví, tedy takovou, které odpovídá minimální počet evolučních změn.

minimum evolution – podobný princip jako NJ, vychází ze sekvenčních dat

maximum parsimony - nejvíce pravděpodobný strom je ten, který vyžaduje nejmenší počet evolučních změn k vysvětlení pozorovaných dat

maximum likelihood - prohledává všechny možné stromy a stanovuje pravděpodobnost, s jakou mohl evoluční scénář generovat soubory znaků, odpovídajících vloženým znakům

Metody konstrukce stromu u distančních dat

- *UPGMA* - nejjednodušší metoda konstrukce fylogenetických stromů. Jedná se čistě o analýzu fenotypu, nebere v potaz např. rozdílnou délku evoluce.
- *Neighbor-joining* pracuje s kritériem minimální evoluce - hledá topologii stromu s nejmenší celkovou délkou všech větví (odpovídá minimální počet evolučních změn).

Příklad konstrukce UPGMA dendrogramu

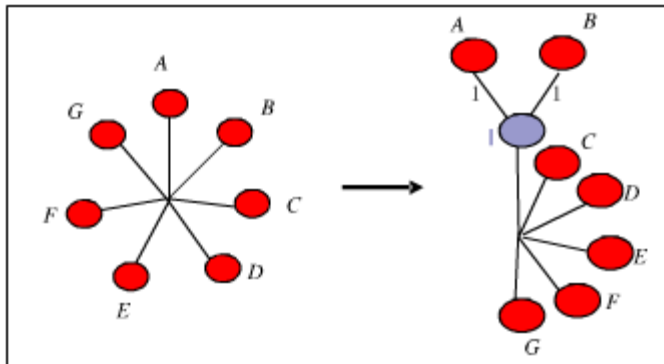


Step 1 – Find the smallest entry in the distance matrix

$d(i,j)$	A	B	C	D	E	F
A	-					
B	2	-				
C	4	4	-			
D	4	4	2	-		
E	7	7	7	7	-	
F	5	5	5	5	6	-
G	8	8	8	8	9	5

Step 2 - Cluster taxa A and B, form a new internal node I

Calculate the lengths of the new edges $d(A,I)=d(B,I)=1/2$ $d(A,B)=1$



Step 3 – Update the distance matrix

$$d(C,I) = \frac{1}{2}(d(A,C) + d(B,C)) = 4$$

etc...

1. V matici distancí najde dvojici s nejmenší distancí a shlukne ji dohromady.
2. V našem případě je to **AB**
3. Vypočítá vzdálenost této společné skupiny od ostatních.

Počítá to jako aritmetický průměr ze vzdáleností všech dvojic jednoduchých OTU (druhů), kde jeden člen dvojice pochází z jedné porovnávané OTU () a druhý z druhé OTU (v našem případě pouze A)

4. Z vypočtených genetických vzdáleností vytvoří novou (menší) matici.....

OTU = operational taxonomic unit
operačně taxonomická jednotka - druh nebo skupina druhů vytvářející jednu větev na fenogramu či kladogramu.

Literatura:

- Felsenstein J. (2004): Inferring phylogenies, Sinauer associates Inc, USA, 664 pp.
- <http://www.docstoc.com/docs/83830686/Distance-Based-Methods-in-Phylogenetic-Tree-Building>

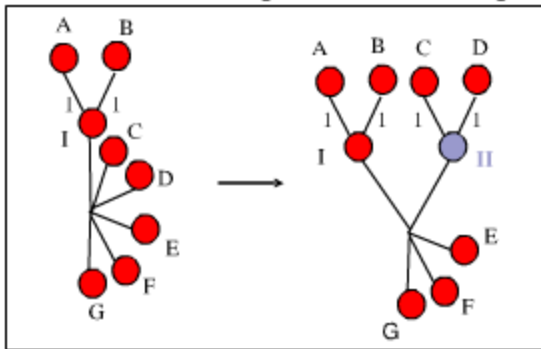
Příklad konstrukce UPGMA dendrogramu

Step 1 – Find the smallest entry in the distance matrix

$d(i,j)$	$I(A+B)$	C	D	E	F
$I(A+B)$	-				
C	4	-			
D	4	2	-		
E	7	7	7	-	
F	5	5	5	6	-
G	8	8	8	9	5

Step 2 - Cluster taxa C and D , form a new internal node II

Calculate the lengths of the new edges $d(C,II)=d(D,II)=1/2 d(C,D)=1$



Step 3 – Update the distance matrix

$$d(I,II)=1/2(d(I,C)+d(I,D))$$

$$= 4$$

$$d(E,II)=1/2(d(E,C)+d(E,D))$$

$$= 7$$

etc...

4. Z vypočtených genetických vzdáleností vytvoří novou (menší) matici.

5. Opakujeme předchozí kroky –

- Nalezneme dvojici s nejmenší distancí a vytvoříme další skupinu

- Provedeme výpočet vzdálenosti této dvojice od ostatních....

- Řešíme další nově vzniklé matice, než jsou všechny vzorky uspořádány v dendrogramu.

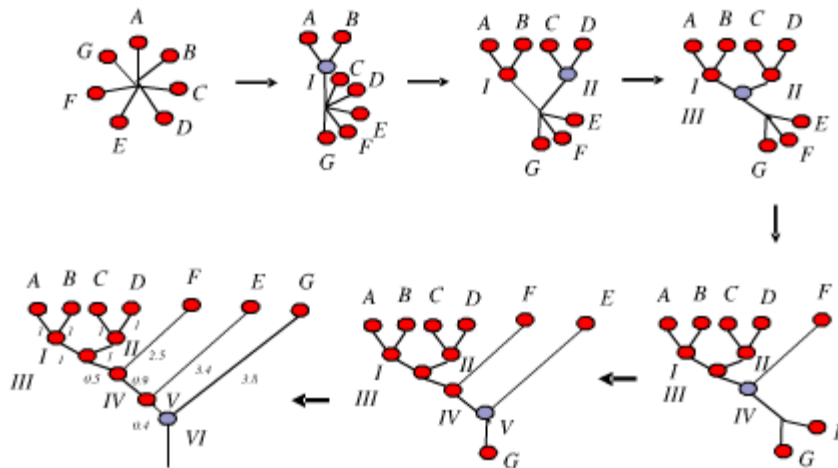
Literatura:

- Felsenstein J. (2004): Inferring phylogenies, Sinauer associates Inc, USA, 664 pp.

- <http://www.docstoc.com/docs/83830686/Distance-Based-Methods-in-Phylogenetic-Tree-Building>

Příklad konstrukce UPGMA dendrogramu

And so on...



...until we have a rooted tree.

4. Z vypočtených genetických vzdáleností vytvoří novou (menší) matici.

5. Opakujeme předchozí kroky –

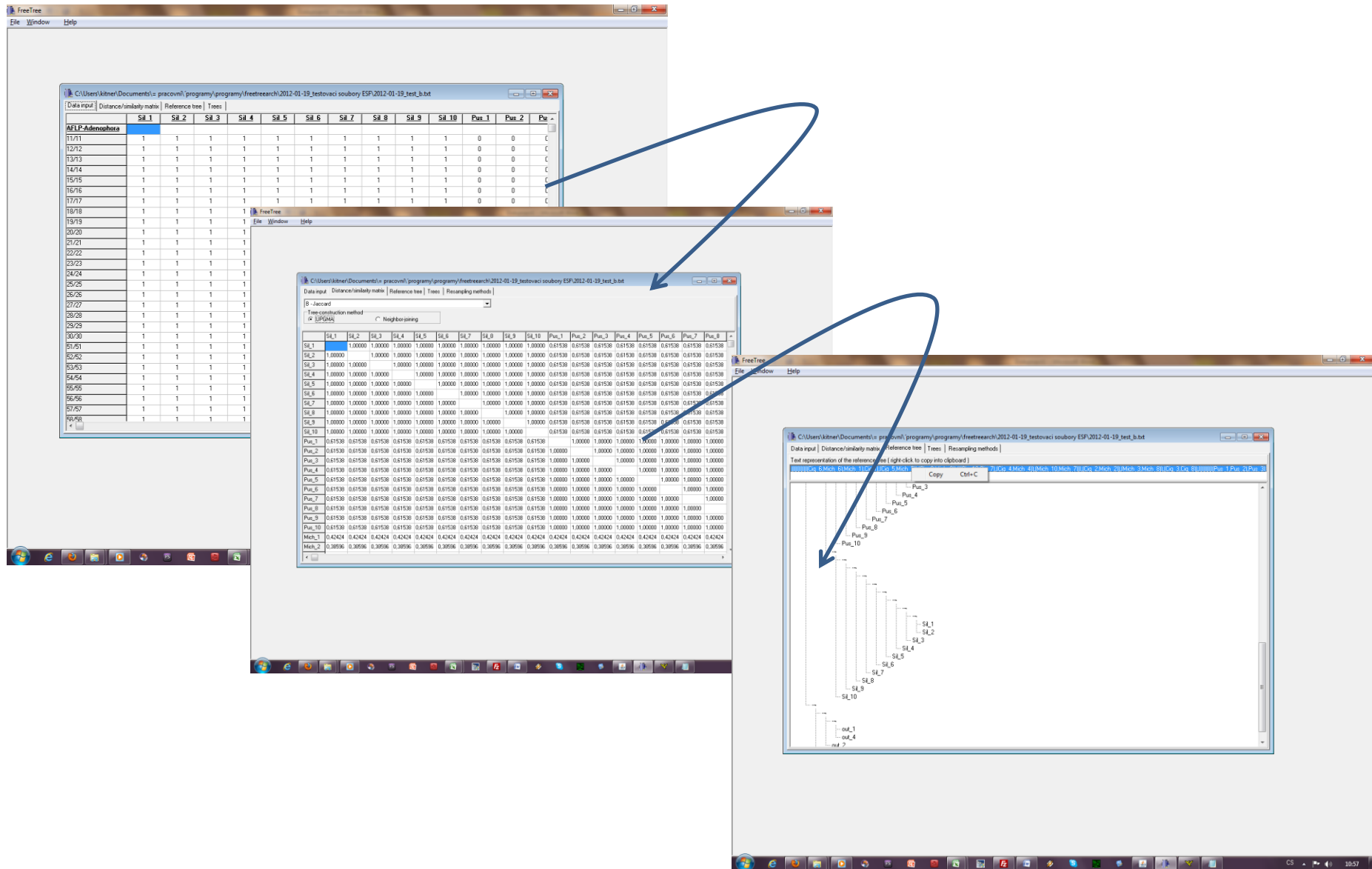
- Nalezneme dvojici s nejmenší distancí a vytvoříme další skupinu
- Provedeme výpočet vzdálenosti této dvojice od ostatních....
- Řešíme další nově vzniklé matice, než jsou všechny vzorky uspořádány v dendrogramu.

Literatura:

- Felsenstein J. (2004): Inferring phylogenies, Sinauer associates Inc, USA, 664 pp.
- <http://www.docstoc.com/docs/83830686/Distance-Based-Methods-in-Phylogenetic-Tree-Building>

Program FreeTree

<http://web.natur.cuni.cz/flegr/programs/freetree.htm>



Software pro vizualizaci a základní grafickou editaci fylogenet. stromů

program **TreeView**

<http://taxonomy.zoology.gla.ac.uk/rod/treeview.html>

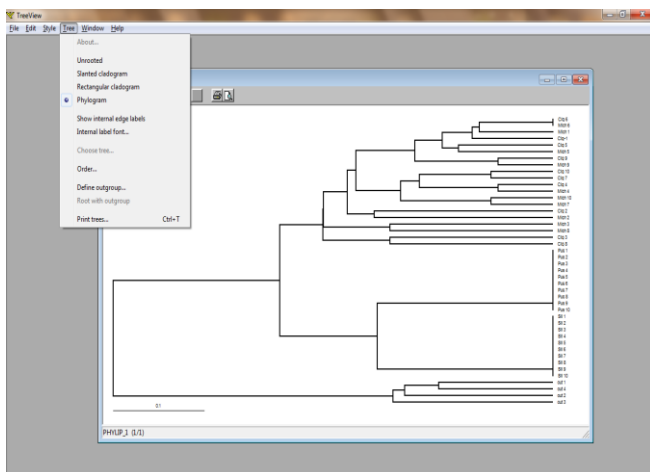
Page, R. D. M. 1996. TREEVIEW: An application to display phylogenetic trees on personal computers. Computer Applications in the Biosciences 12: 357-358.

Taxonomy and Systematics at Glasgow

TreeView

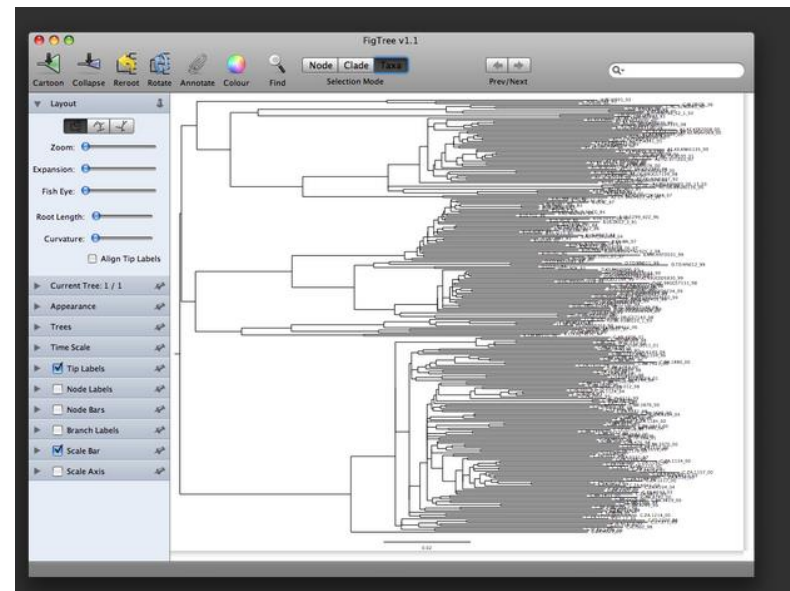
Tree drawing software for Apple Macintosh and Windows

 (and now [Linux and Unix](#))



program **FigTree**

<http://tree.bio.ed.ac.uk/software/figtree/>

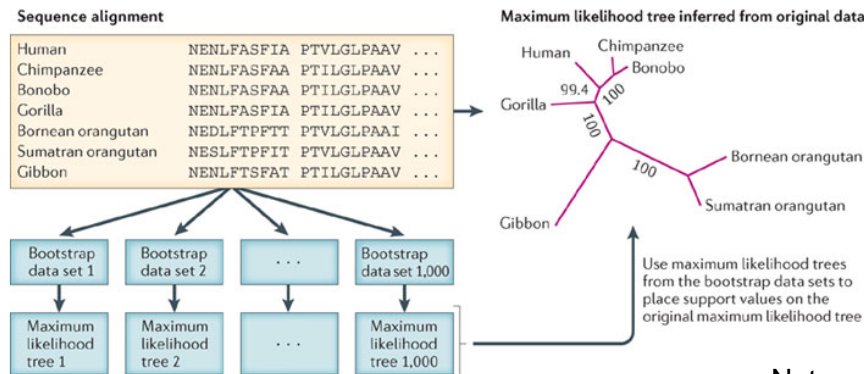


Bootstrapping

Ověřování správnosti topologie výsledného stromu

- Z datové matice je vypuštěn určitý znak/marker a jeho místo v matici je nahrazeno jiným markerem.
- Proveďte konstrukce nového stromu a srovnává se topologie nově vzniklého stromu s původním.
- Postup se opakuje 100-1000x.
- Čím častěji dostáváme stejnou topologii, tím vyšší je výsledná statistická podpora.
- Statisticky průkazné hodnoty se zobrazují na uzlech větvení.

In traditional parameter estimation, we attach a confidence interval to indicate the uncertainty involved in the point estimate of the parameter. This has not been possible in molecular phylogenetics, as concepts such as the variance and confidence interval are not meaningful when applied to trees. For distance, parsimony and likelihood methods, the most commonly used procedure to assess the confidence in a tree topology estimate is the bootstrap analysis⁷³. In this approach, the sites in the sequence alignment are resampled with replacement as many times as the sequence length, generating a bootstrap pseudo-sample that is of the same size as the original data set. Typically, 100 or 1,000 bootstrap samples are generated in this way, and each one is analysed in the same way as the original sequence alignment. An example that uses the maximum likelihood method is illustrated in the figure. The inferred trees from those bootstrap samples are then tabulated to calculate the bootstrap support values. For every clade in the estimated tree, its bootstrap support value is simply the proportion of bootstrap trees that include that clade^{24, 65, 133}. The commonly used but less satisfactory approach is to use the bootstrap trees to generate a majority-rule consensus tree, which shows a clade if — and only if — it occurs in more than half of the bootstrap trees.



Co nám řeknou stromečky:

Diverzita, divergence vs. populační historie

Divergence~
strukturovanost

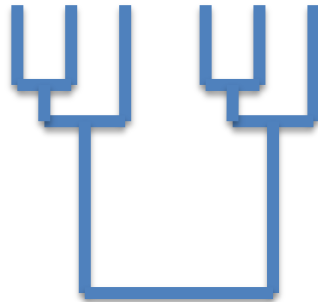
Vysoká

(přítomnost privátních,
privátních fixovaných
markerů)

Nízká

genetická diverzita

vysoká

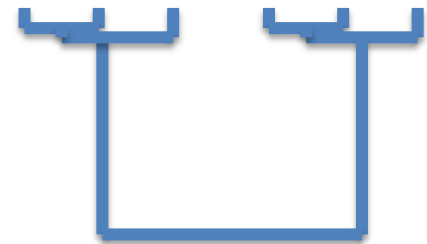


- omezený genový tok mezi populacemi, které jsou jinak relativně variabilní



Mezi populacemi dochází k výměně
alel: vysoký genový tok, imigrace

nízká



- malé, izolované populace, přežívající dlouhodobě v refugiálních stanovištích
- prakticky absence gene-flow
- absence sex. rozmnožování



- Málo variabilní populace, rozdíly pravděpodobně souvisí s sub-recentním rozšířením na nové stanoviště, postupně se začíná odlišovat od mateřské populace

Fer & Kolář

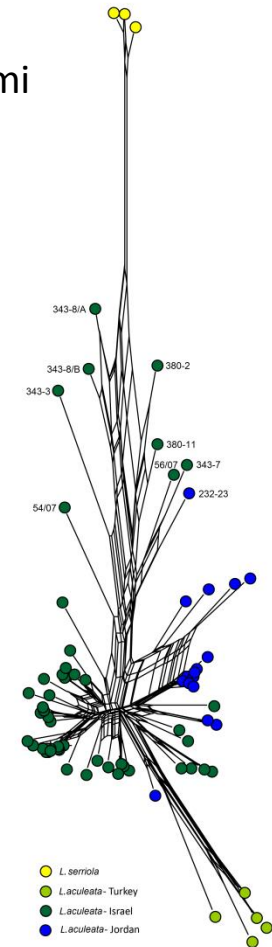
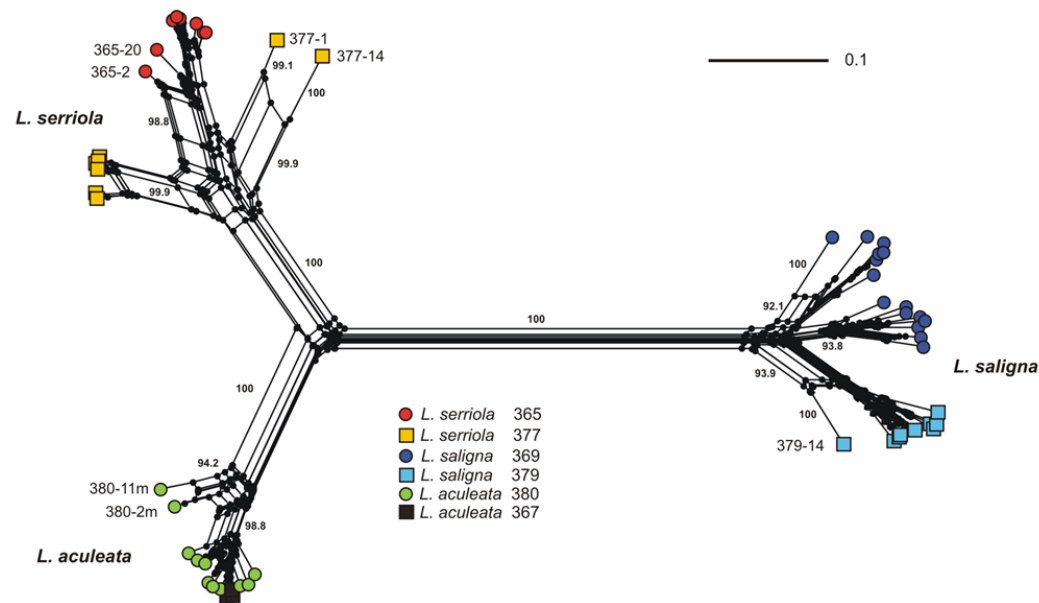
Neighbour-net

<http://www.splitstree.org/>

SplitsTree4 

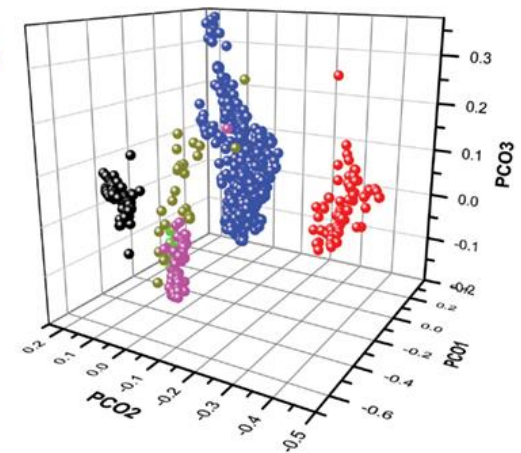
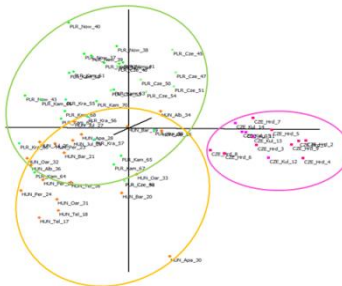
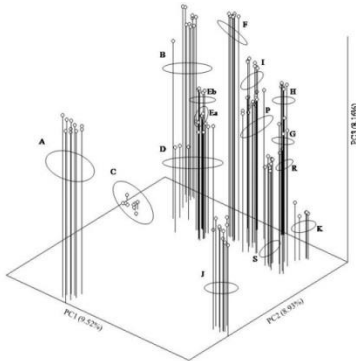
- Zhlukovací metoda založená na N-J algoritmu
- Konstrukce nezakořeněných fylogenetických sítí
- Někdy přináší detailnější a přehlednější informaci o vztazích mezi populacemi
 - např. interspecifická hybridizace = mezidruhové křížení

Figure 2. Neighbor-Network cluster analysis of the three wild *Lactuca* species based on EST-SSR and AFLP data (Dice coefficient of similarity, 1000 bootstraps)



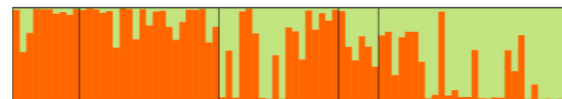
Principal coordinate(s) analysis – PCoA

- pro analýzu binárních znaků
- počet znaků může být větší než počet objektů
 - tj. typicky pro RAPD, AFLP data...
- nehierarchická vizualizace struktury dat
- výpočet matice vzdáleností mezi objekty (Jaccard)
- ordinační diagram – převedení objektů do nového prostoru na základě vzdálenosti mezi nimi a maximalizace vysvětlené variability
 - software – NTSys, FAMD, PAST, R, CANOCO, SYNTAX, ...



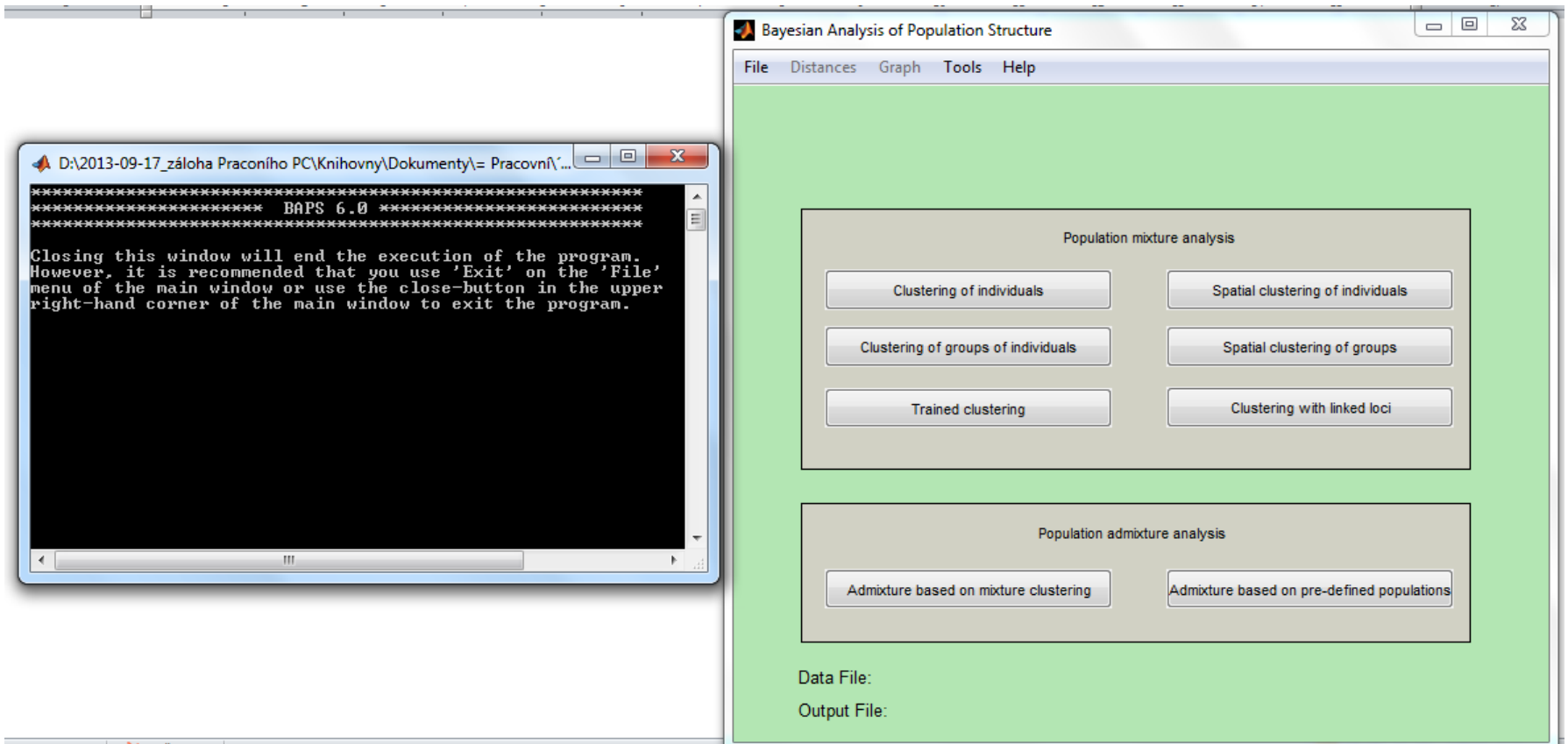
Bayesovské clusterování

- hledání takového rozdělení individuů do K clusterů, které je nejvíce pravděpodobné (resp. má maximální záporný logaritmus marginální pravděpodobnosti)
- výsledkem je zjištění optimálního počtu clusterů, tj. „reálných populací“ a rozřazení všech individuů
- Může probíhat ve dvou módech:
- mixture – každý jedinec do právě jednoho clusteru/populace
- admixture – pravděpodobnostní rozřazení jedince do více clusterů/populací = tj. připouštíme, že se jedinec svým genotypovým profilem může z části podobat jedincům z různých skupin



- **Software:** BAPS 6, STRUCTURE, Geneland

BAPS 6



Před během analýzy zadáme do kolika skupin se má program pokusit rozdělit náš dataset

Výstup z programu BAPS 6

```
D:\2013-09-17_záloha Praconího PC\Knihovny\Dokumenty\= Pracovní\programy\BAPS\BAPS6\b...
RESULTS OF INDIVIDUAL LEVEL MIXTURE ANALYSIS:
Data file: baps-input.txt
Model: independent
Number of clustered individuals: 75
Number of groups in optimal partition: 5
Log(marginal likelihood) of optimal partition: -4387.8875

Best Partition:
Cluster 1: {9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20,
21, 28, 29, 30, 31, 32, 33, 34, 35, 36, 37, 38,
39, 40, 41, 42, 43, 44, 45, 46, 47, 48, 49, 50,
51, 52, 53, 54, 55, 56, 57, 58, 59, 60, 61, 62,
63, 64, 65, 66, 67, 68, 69}
Cluster 2: {22, 23, 24, 25, 70, 71, 72, 73, 74, 75}
Cluster 3: {26, 27}
Cluster 4: {1, 2, 3, 4, 5}
Cluster 5: {6, 7, 8}

Changes in log(marginal likelihood) if individual i is moved to group j:
  ind      1      2      3      4      5
  1:    -121.2   -111.1   -117.5     .0   -210.7
  2:     -97.5   -104.4   -118.8     .0   -217.4
  3:     -31.2    -32.4    -73.0     .0   -172.4
  4:    -115.9    -89.9   -109.0     .0   -196.3
  5:    -120.9   -114.0   -122.3     .0   -217.0
  6:    -428.7   -174.0   -105.4   -251.1     .0
  7:    -431.3   -182.7   -104.5   -251.5     .0
  8:    -406.9   -178.1    -97.7   -219.5     .0
 73:    -108.6     .0    -59.1    -96.9   -124.2
 74:    -133.2     .0    -36.3    -88.5    -71.8
 75:     -98.4     .0    -62.2   -101.4   -148.5

KL-divergence matrix in PHYLIP format:
5
Cluster_1  0.000  0.340  0.741  0.487  1.483
Cluster_2  0.340  0.000  0.321  0.439  0.717
Cluster_3  0.741  0.321  0.000  0.535  0.434
Cluster_4  0.487  0.439  0.535  0.000  0.949
Cluster_5  1.483  0.717  0.434  0.949  0.000

List of sizes of 10 best visited partitions and corresponding log<ml> values
5 -4387.8875
5 -4389.5355
6 -4390.7572
5 -4392.9559
5 -4404.6106
6 -4406.854
4 -4415.8574
4 -4416.491
6 -4423.4995
6 -4424.7482

Probabilities for number of clusters
5  0.95482
6  0.045177
```

rozdělení jedinců do skupin

změna likelihood modelu,
pokud přesuneme jedince *i*
do skupiny *j*

matice nepodobnosti mezi skupinami

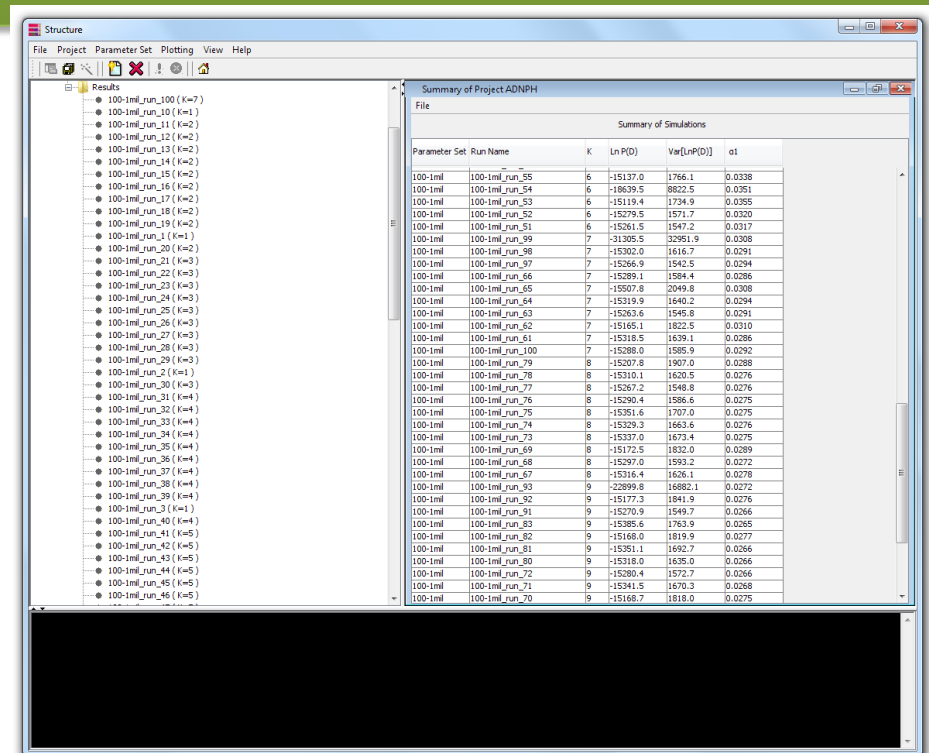


Pravděpodobnost rozdělení do clusterů
- zde je 95,48% pravděpodobnost rozdělení pro K=5 a
0,45% pro rozdělení do 6ti skupin

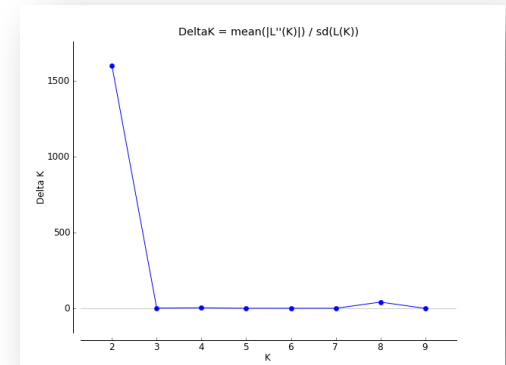
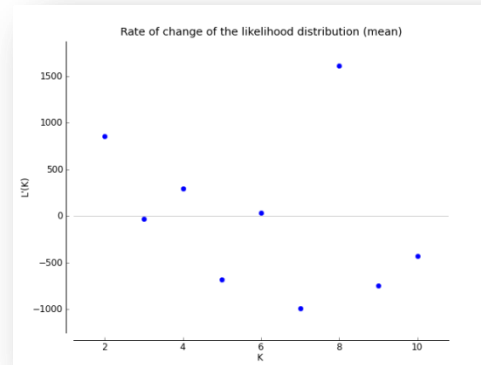
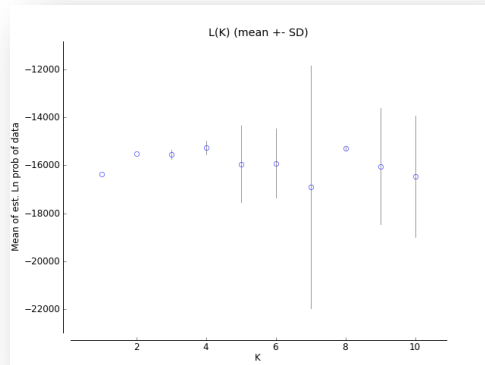
Structure

Structure

- uživatelský rozdíl oproti BAPS
 - Structure opakovaně testuje rozdělení analyzovaného souboru vzorků do jednotlivých skupin Zpravidla 10x pro každé K od K=1 až K=n (1-10)
- Podle nastavených parametrů a objemu zpracovávaných dat program běží několik dnů-týdnů



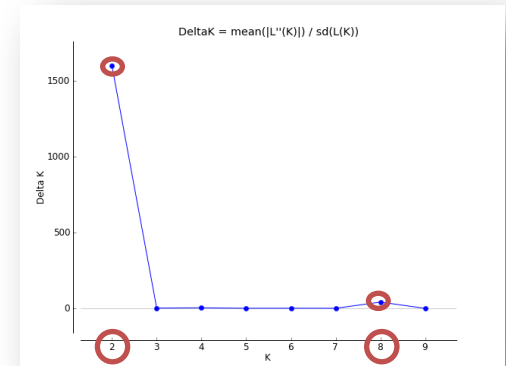
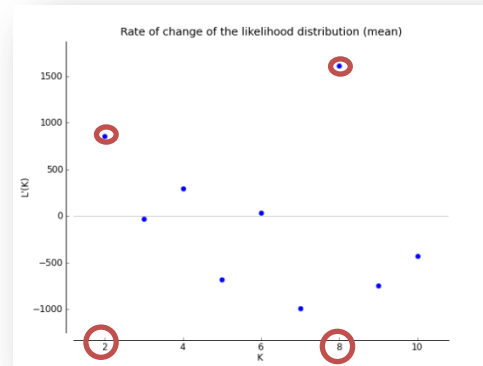
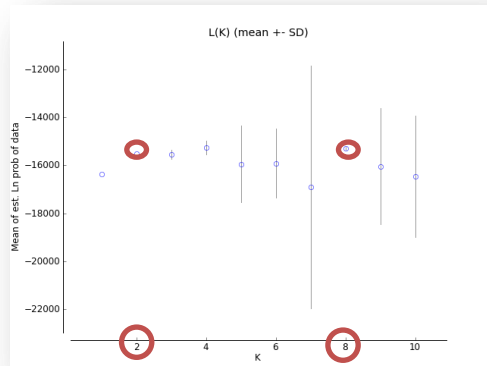
Structure -> StructureHarvester



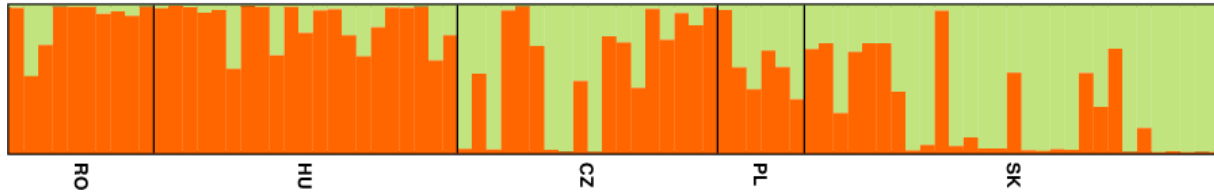
StructureHarvester

- provede vyhodnocení rozptylu výsledků běhů programu **Structure** pro jednotlivá K
- grafické výstupy nám pomohou při rozhodnutí pro jaké K se rozhodneme při interpretaci výsledků
- je běžné, že se neuvede výsledek pouze pro neoptimálnější K doporučené Structure, pokud je k tomu racionální zdůvodnění (např. znalosti o biologii studovaného druhu)
- Důležitý je i vygenerovaný „clummp“ soubor pro dané K, který se používá pro sestrojení finálního grafického výstupu Další software **Distruct**

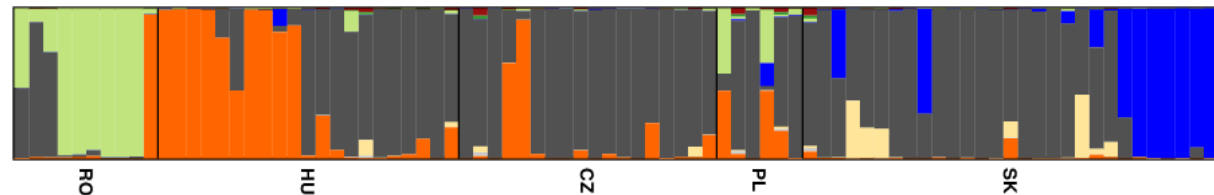
Structure -> StructureHarvester

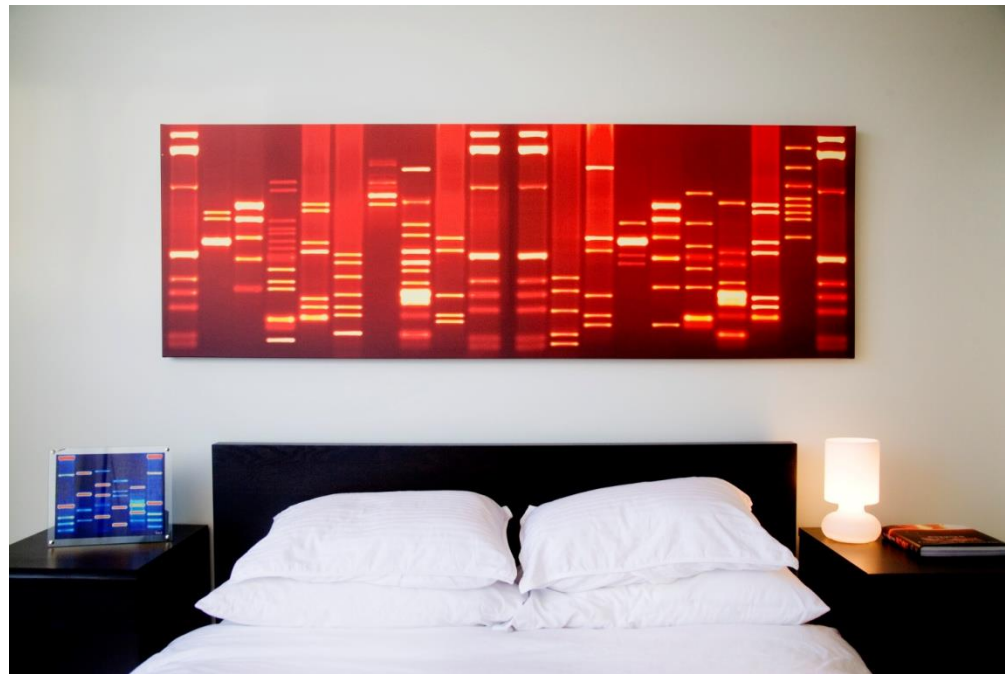


K = 2



K = 8





That's all folks!

