DNA barcoding

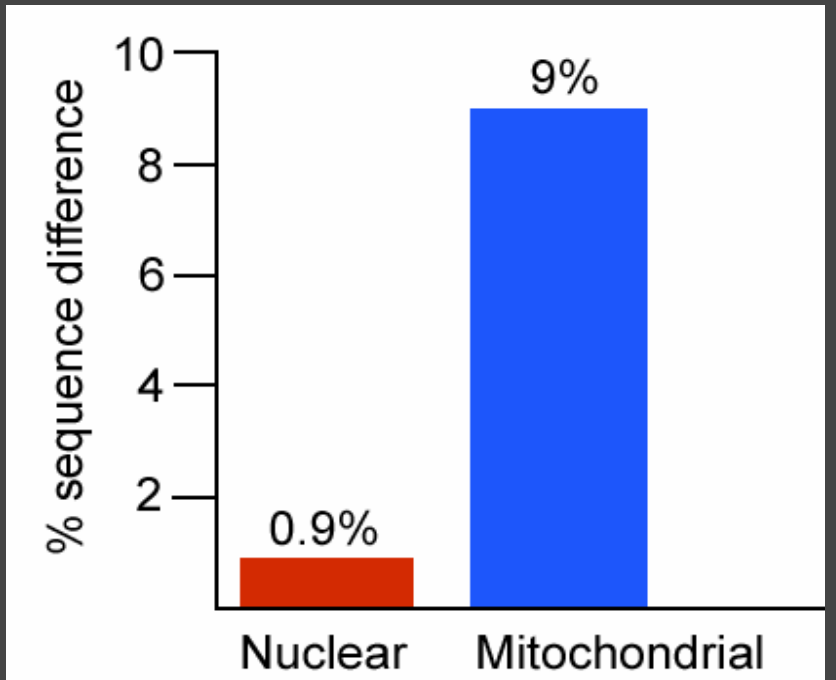
M. Krkavec, M. Štalmašek

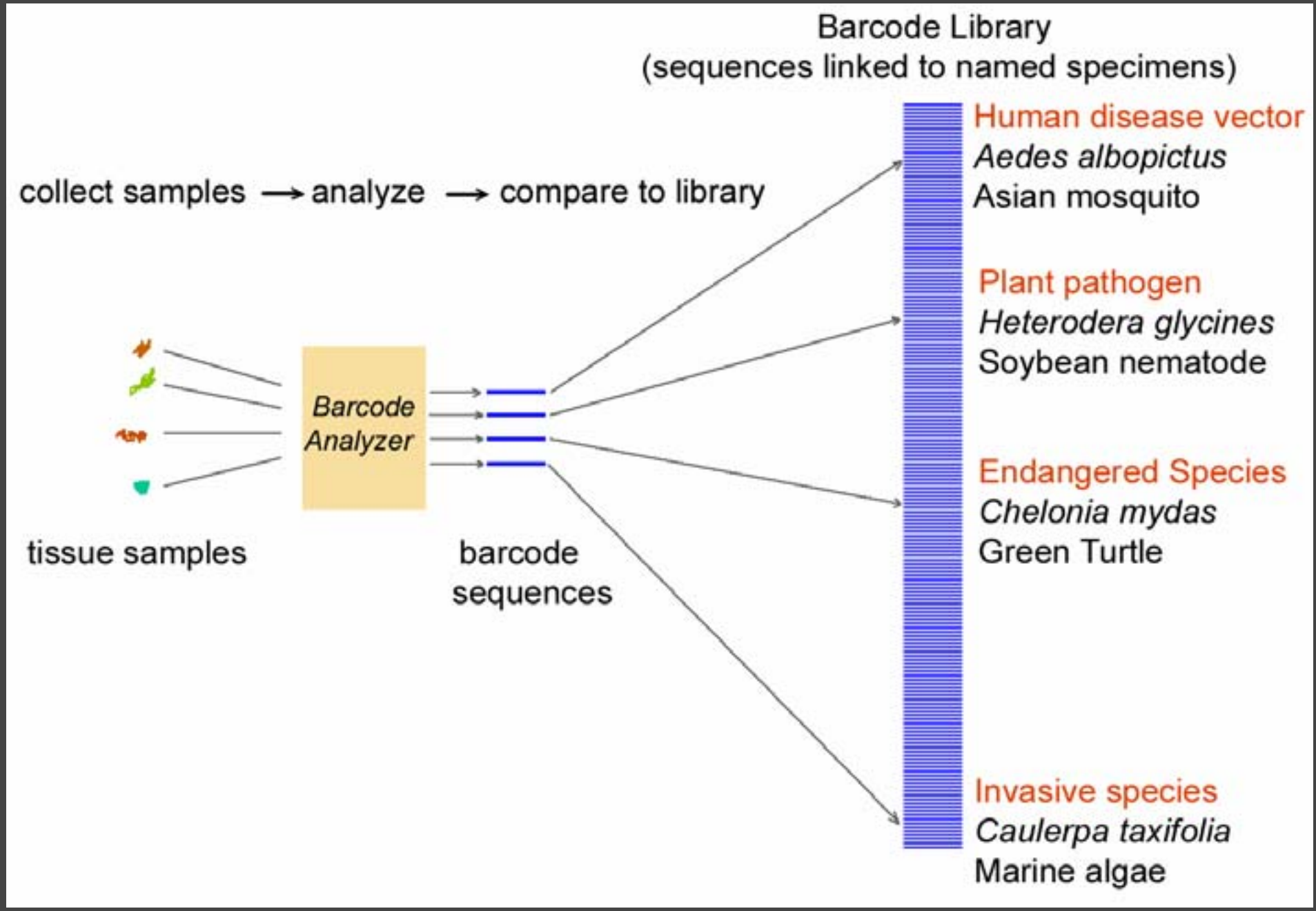
DNA barcoding

V biologii odpovídá různým metodám:

1. zanesení krátkých sekvencí ("otisků") do vzorků DNA
   - cílem je pozdější identifikace vzorku DNA podle otisku
   - je potřeba nalézt vhodné otisky, abychom dokázali identifikovat vzorky i po případných chybách v DNA
   - chyby vznikají mutací nebo při sekvenování

2. taxonomická klasifikace
   - podle úseku DNA zařadit vzorek do existující taxonomie
   - užívá se mitochondriální DNA kódující COI (pouze 648bp)
   - u rostlin kombinace rbcL a matK genů chloroplastu





DNA barcoding - využití (© Barcoding life illus.)

DNA barcoding - klasifikátor

Formálně:

- vstup: sekvence $s_1, \ldots, s_n$
- výstup: k-tice řetězců $T = (t_1, \ldots, t_k)$ splňující
  - pro každé dvě různé sekvence $s_i <> s_j$ existuje řetězec t_l, který je podřetězcem s_i a není podřetězcem s_j

Chceme minimalizovat velikost k-tice T.

Barcode sekvence s:

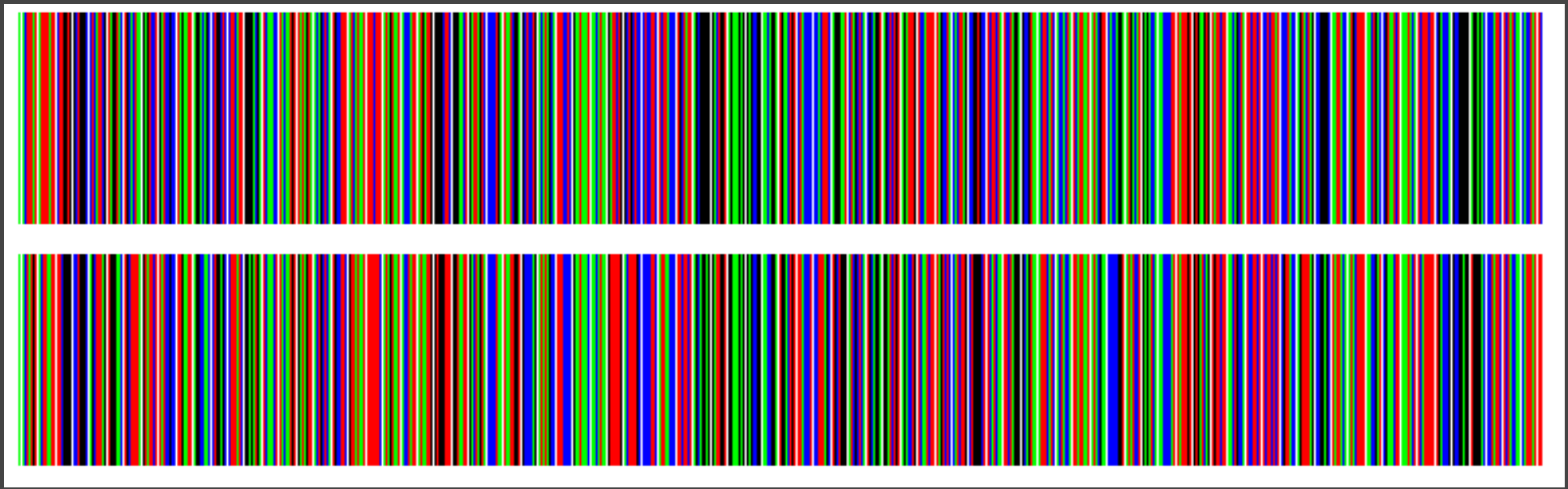
- posloupnost nad abecedou $\{0, 1\}$
- na pozici i je 0, pokud t_i není podřetězcem s
- na pozici i je 1 pokud je podřetězcem

	AC	C	GA
AGGT	0	0	0
ACCTGA	1	1	1
TGGAT	0	0	1
GCA	0	1	0
CGCGATT	0	1	1
GTTAC	1	1	0

Nástroje & Algoritmy

SPIDER: Species identity and evolution in R

- obsahuje ukázkové datasety `dolomedes` a `anoteropsis`
- umožňuje získat GenBank sekvence přímo z prostředí R
- umožňuje získat BOLD sekvence z R (nezdařilo se)
- nástroje pro analýzu krátkých sekvencí (tedy barcode)
- **G**, **A**, **C**, **T**



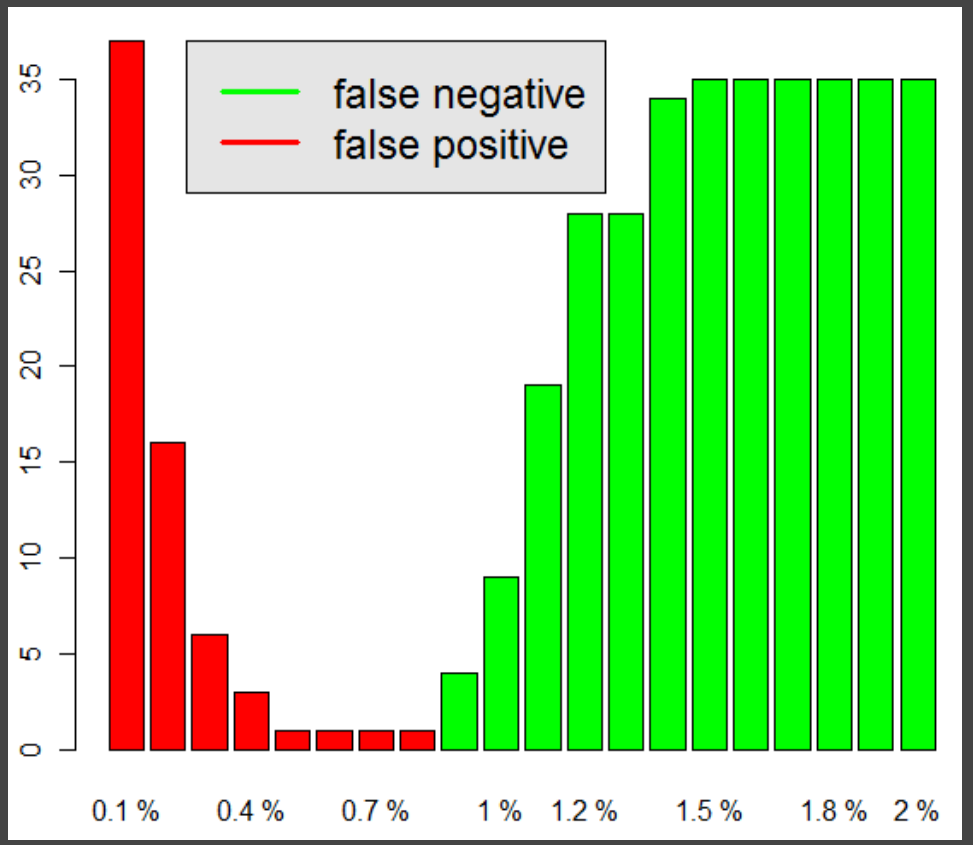
SpideR

Metriky identifikace:

- Kimurův dvouparametrový model:
  - předpokládá stejné zastoupení všech čtyř bázi
  - zohledňuje purinové (A, G) a pyrimidinové (C, T) báze

Klasifikace:

- nearest neighbour
- threshold identification
- Meier's best close match (kombinace předchozích dvou)
- optimalizace prahu (threshold)



Použitá data

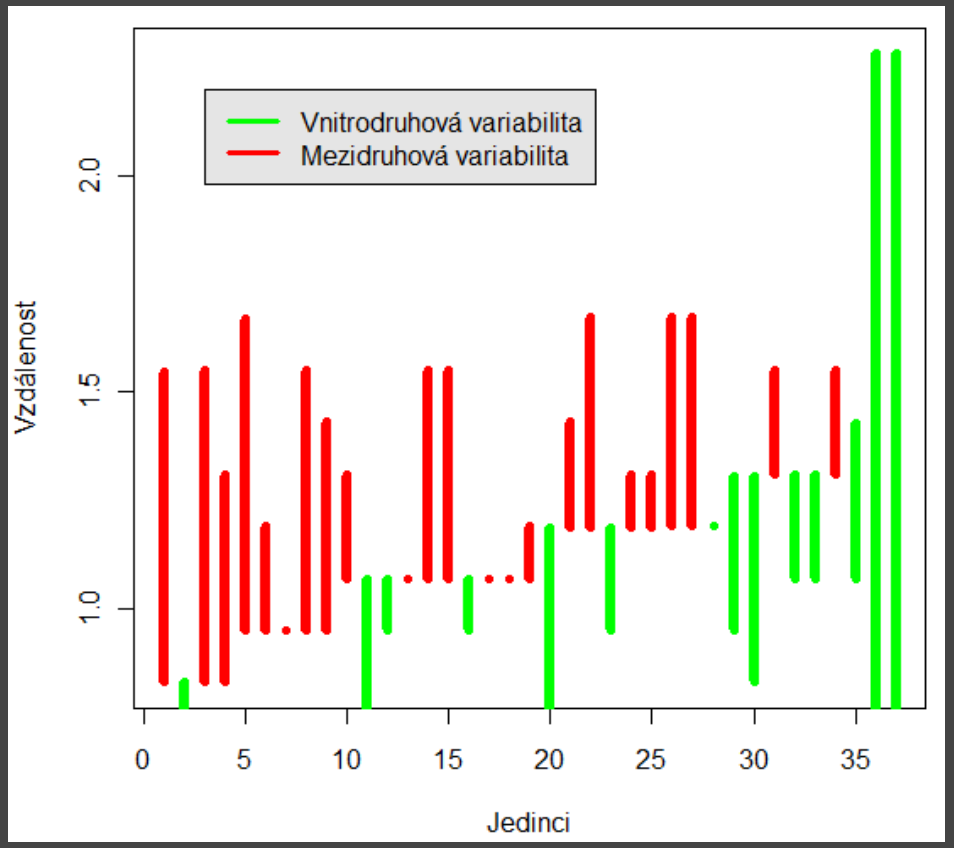
1. data z projektu The Barcode of Life (www.barcodinglife.org)
    - cca 600 000 sekvencí veřejně ke stažení ve formátu csv
    - trochu chaotické, nekompletní údaje

2. Spider
    - testováno na DNA pavouků lovčíků
    - GenBank GQ337328 až GQ337385

Výsledky

SpideR:

- identifikace po předchozím zanesení do referenční db
- pro obecné vyhledávání zatím nevhodné
- $O(nc^s)$, pro n délku sekvencí,
  s počet sekvencí
  a konstantu $c > 1$



Zdroje

Spider: Species identity and evolution in R:
http://spider.r-forge.r-project.org/SpiderWebSite/spider.html

The Barcode of Life Data Systems (BOLD):
http://www.boldsystems.org/views/datarelease.php

Barcoding life, illustrated: http://phe.rockefeller.edu/PDF_FILES/BLIllustrated26jan04print%20v1-3.pdf

Data sets: http://dimacs.rutgers.edu/Workshops/BarcodeResearchChallenges2007/

DNA-BAR: http://dna.engr.uconn.edu/?page_id=23

Zdroje - neúspěšné...

TaxI: a software tool for DNA barcoding using distance methods: http://www.mvences.de/TaxI.zip

BPSI2.0: a C/C++ interface program for species identification via DNA barcoding with a BP-neural network by calling the Matlab engine

DNA Barcode Linker: www.dnabarcodelinker.com

Statistical Assignment Package (SAP): http://users-cs.au.dk/kmt/StatisticalAssignmentPackage.html