



Bioinformatické nástroje pro hodnocení variability genomu živočišných genetických zdrojů

Modul č. 2: Zachování a udržitelné využívání živočišných genetických zdrojů

Tomáš Urban

Mendelova univerzita v Brně

Agronomická fakulta



Co-funded by
the European Union

Úvod – význam bioinformatiky

- Bioinformatika je interdisciplinární obor, který kombinuje biologii, informatiku, statistiku a matematiku s cílem analyzovat a interpretovat biologická data pomocí výpočetních nástrojů a algoritmů.
- V genetice zvířat hraje klíčovou roli při získávání a analýze rozsáhlých genomických dat, jako jsou sekvence DNA, s cílem získat poznatky o genetické výbavě a biologických procesech zvířat.
- Bioinformatika v genetice zvířat pomáhá při identifikaci a charakterizaci genů odpovědných za specifické znaky, nemoci nebo odchylky u zvířat. Může například pomoci identifikovat geny spojené se zbarvením srsti, produkcí mléka, rychlostí růstu nebo náchylností k nemocem.



Co-funded by
the European Union

- Umožňuje výzkumníkům porovnávat a analyzovat genomy různých živočišných druhů, pochopit jejich evoluční vztahy a identifikovat společné genetické prvky. To může přinést poznatky o rozmanitosti a evoluci živočišných druhů.
- B. pomáhá při vývoji nových šlechtitelských strategií a zlepšování chovatelských postupů. Analýzou genetických dat pomáhá identifikovat zvířata s žádoucími vlastnostmi pro šlechtitelské programy, zlepšovat vlastnosti, jako je produktivita, odolnost vůči chorobám nebo přizpůsobivost.
- B. podporuje ochranu a zachování ohrožených druhů zvířat studiem jejich genomů a identifikací genetických markerů pro monitorování populací, hodnocení genetické rozmanitosti a pomoc při programech chovu v zajetí.



Data a databáze

- **Co je potřeba: internet, PC, programy, vlastní data**
- **Data**
 - Sekvence DNA, RNA a proteinů
 - Genomická, transkriptomická, proteomická, metabolomická, fylogenetická, strukturální data
- **Databáze**
 - Na internetu je k dispozici **1959** databází v lednu 2024 (evidováno v Nucleic Acids Research, doi.org/10.1093/nar/gkad1173)
 - Seznam kategorií databází www.oxfordjournals.org/nar/database/c
 - Významné sekvenční databáze:
 - GenBank (NCBI www.ncbi.nlm.nih.gov/genbank/)
 - ENA (EMBL-EBI www.ebi.ac.uk/),
 - UniProt (www.uniprot.org/)
 - Genomická databáze Ensembl (www.ensembl.org/)



Co-funded by
the European Union

Bioinformatické nástroje

Na serverech, kde jsou databáze, jsou i nástroje pro hledání, přirovnávání (alignment) a analýzu bioinformatických dat (sekvencí)

Párový sekvenční alignment se používá k identifikaci oblastí podobnosti, které mohou naznačovat funkční, strukturní a/nebo evoluční vztahy mezi dvěma biologickými sekvencemi (proteiny nebo nukleovými kyselinami).

```

1 ATGAGTCTCTCT-----GATAAG-----GACAAGGCTGC--TGTGAAA    36
   ||.|||      ||.|||      |.|||||.||  .|.|.||
1 -----CTGTCTCCTGCCGACAAGACCAACGTCAAGGCCGCCTGGGGTAA    44
  
```

Vícečetný sekvenční alignment (MSA) je zarovnání tří nebo více biologických sekvencí podobné délky. Z výstupu aplikací MSA lze odvodit homologii a studovat evoluční vztah mezi sekvencemi.

```

ATGAGTCTCTCTGATAAGGACAAGGCTGCTGTGAAAGCCCTATGG----- 45
-----CTGTCTCCTGCCGACAAGACCAACGTCAAGGCCGCCTGGGGTAAG----- 45
-----ACAAAAGCAACATCAAGGCTGCCTGGGGGAAGATTGGTGGC 41
          ****          * * * *          ***
  
```

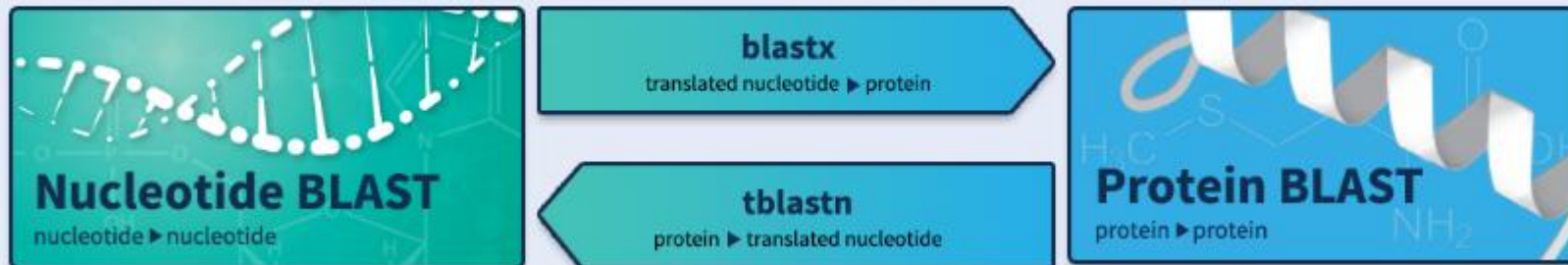
Přirovnávání dvojice sekvencí (párový alignment)

• BLAST - Basic Local Alignment Search Tool- na NCBI

blast.ncbi.nlm.nih.gov/Blast.cgi

- BLAST vyhledává oblasti podobnosti mezi biologickými sekvencemi. Program porovnává nukleotidové nebo proteinové sekvence s databázemi sekvencí a vypočítává statistickou významnost.
- Lze přirovnat vlastní sekvenci s databázovými (v GenBank)
- Lze přirovnat konkrétní dvě sekvence
- Primárně jde o lokální přiřazení (je k dispozici i pro globální přiřazení)

Web BLAST



Co-funded by
the European Union

EMBOS na EMBL-EBI

- Nástroj pro párový sekvenční alignment na serveru EMBL-EBI
- www.ebi.ac.uk/jdispatcher/psa

Kdy použít lokální nebo globální alignment?

- **Lokální** (např. Smith-Watermanův algoritmus)
 - u sekvencí odlišnějších, evolučně vzdálenějších
 - Omezí se na přiřazování jednoznačných úseků a skončí tam, kde se sekvence rozcházejí nad únosnou míru (**lokální přiřazení**)
- **Globální** (např. Needleman-Wunschův algoritmus)
 - **Nejvhodnější pro sekvence, které jsou podobné a přibližně stejně dlouhé**
 - pokusíme se sekvence přiřadit po celé délce i za cenu vnášení mezer do jedné z nich či do obou (**globální přiřazení**)



Co-funded by
the European Union

Modelový postup alignmentu

- stanovení odpovědi na otázku, zda jsou si vzájemně podobnější sekvence DNA A a B nebo C a D.

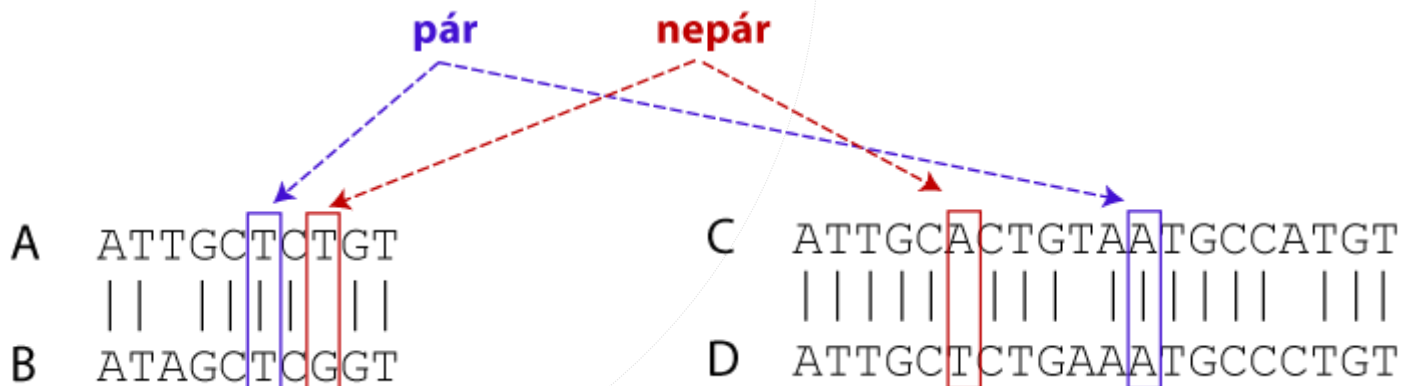
A **ATTGCTCTGT**

C **ATTGCACTGTAATGCCATGT**

B **ATAGCTCGGT**

D **ATTGCTCTGAAATGCCCTGT**

- **Přiřazení (alignment):** Sekvence po celé délce přiložíme k sobě, tj. zapíšeme je do dvou pod sebou umístěných řádků tak, aby identické pozice (báze či aminokyseliny) ležely pod sebou



N párů (match)	8	17
N nepárů (mismatch)	2	3



Co-funded by
the European Union

Výpočet normalizované hodnoty podobnosti (score):

- Hodnota páru (*match*): **1** (podle zvolených kritérií)
- Hodnota nepáru (*mismatch*): **0**

$$S_{AB} = (8 \times 1 + 2 \times 0) / 10 = 0,80$$

$$S_{CD} = (17 \times 1 + 3 \times 0) / 20 = 0,85$$

Počet párů Počet nepárů Počet pozic

→ Podobnější si jsou sekvence C a D

- **porovnávat míru podobnosti různě dlouhých** dvojic sekvencí -> **vydělíme celkovou hodnotu podobnosti délkou přiřazení** (normalizace hodnoty podobnosti) – skóre sekvenční podobnosti (***Sequence alignment score***).



Co-funded by
the European Union

Alignment dvou krátkých sekvencí

- Nepřiřazené sekvence

Sequence 1 (query)

AGGVLI IQVG

score = 6

| | | | |

Sequence 2 (subject)

AGGVLIQVG

- Přiřazené sekvence

Sequence 1 (query)

AGGVLI IQVG

score = 9

| | | | | | |

Sequence 2 (subject)

AGGVLI-QVG

Score se zvyšuje vložení mezer. Mezera zvyšuje počet přiřazených identických reziduí.

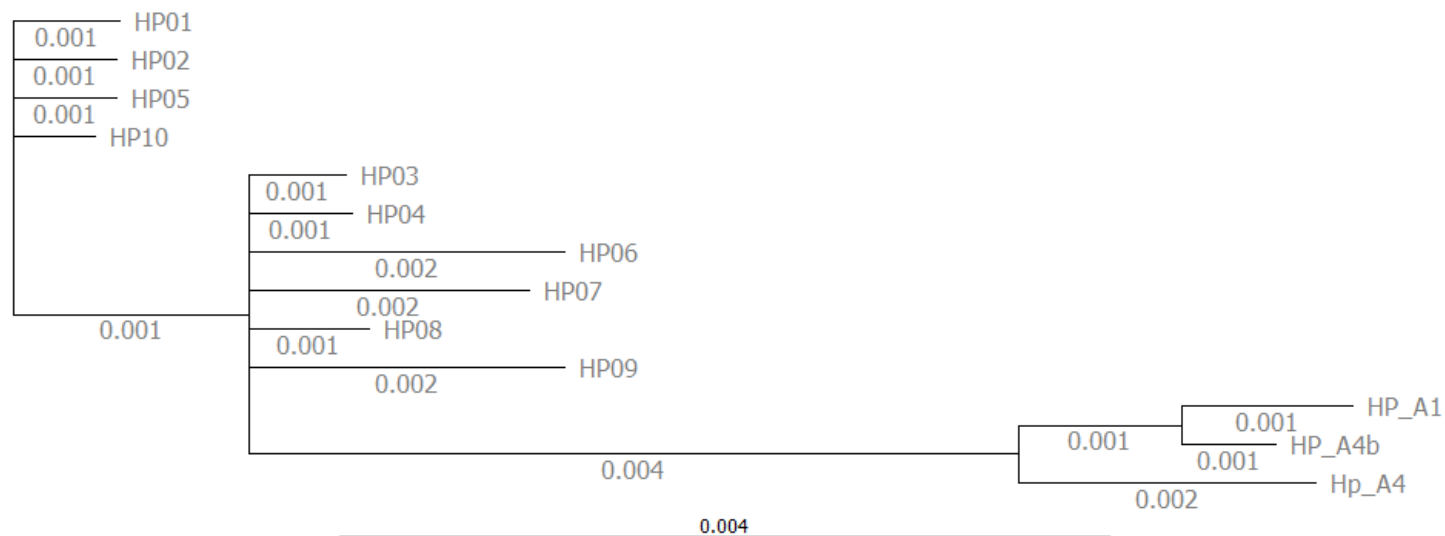
On-line nástroje pro MSA

- www.clustal.org > www.ebi.ac.uk/jdispatcher/msa
- **Clustal Omega**
 - www.ebi.ac.uk/jdispatcher/msa/clustalo
- **MAFFT**– používá Fast Fourier Transforms. Pro středně dlouhé alignmenty
 - www.ebi.ac.uk/jdispatcher/msa/mafft
- **T-Coffee** - vhodné pro menší alignmenty
 - <https://www.ebi.ac.uk/jdispatcher/msa/tcoffee>
 - tcoffee.crg.eu/
- **MUSCLE**– **M**Ultiple **S**equences **C**omparison by **L**og- **E**xpectation
 - Lepší přesnost a vyšší rychlost než [ClustalW2](http://www.ebi.ac.uk/jdispatcher/msa/mafft) nebo [T-Coffee](https://www.ebi.ac.uk/jdispatcher/msa/tcoffee), závisí na vybraných možnostech (options)
 - www.ebi.ac.uk/jdispatcher/msa/muscle
 - www.drive5.com/muscle/
- **Kalign** (EMBL EBI) – velmi rychlý MSA, na lokální oblasti. Pro dlouhé alignmenty
 - msa.sbc.su.se/cgi-bin/msa.cgi
- **COBALT** (NCBI) - **C**onstraint-**b**ased **M**ultiple **A**lignment **T**ool
 - (www.ncbi.nlm.nih.gov/tools/cobalt/)
 - Počítá mnohočetný sekvenční alignment **proteinů** použitím konzervovaných domén a lokální sekvenční podobnost.



Využití MSA pro fylogenetické analýzy – konstrukce fylogenetického stromu

ISAGREED





Děkuji za pozornost!

Partners:



Mendel
University
in Brno



Siedlce University
of Natural Sciences
and Humanities



Czech University
of Life Sciences Prague

This presentation has been supported by the Erasmus+ KA2 Cooperation Partnerships grant no. 2021-1-SK01-KA220-HED-000032068 "Innovation of the structure and content of study programs in the field of animal genetic and food resources management with the use of digitalisation - Inovácia obsahu a štruktúry študijných programov v oblasti manažmentu živočíšnych genetických a potravinových zdrojov s využitím digitalizácie". The European Commission support for the production of this presentation does not constitute an endorsement of the contents which reflects the views only of the authors, and the Commission cannot be held responsible for any use which may be made of the information contained therein.



Tomáš Urban



urban@mendelu.cz

