INVESTICE DO ROZVOJE VZDĚLÁVÁNÍ

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

CZ.1.07/2.2.00/28.0032

MOLEKULÁRNÍ METODY V EKOLOGII MIKROORGANIZMŮ

(EKO/MMEM)

ZPRACOVÁNÍ DGGE VÝSTUPŮ A SEKVENAČNÍCH DAT

Zpracování výstupů z denaturační gradientové gelové elektroforézy (DGGE) – práce s programem Gel2k

Program Gel2k používáme především pro zpracování výstupů (fotografií) denaturační gradientové gelové elektroforézy (DGGE). Snímky DGGE nejprve upravíme v programu NIS-Elements. Fotografii gelu převedeme do černobílého spektra tak, aby jednotlivé bandy byly světlé a pozadí tmavé. Dále snímek překlopíme horizontálně. Ukládáme buď ve formátu TIFF nebo JPEG.

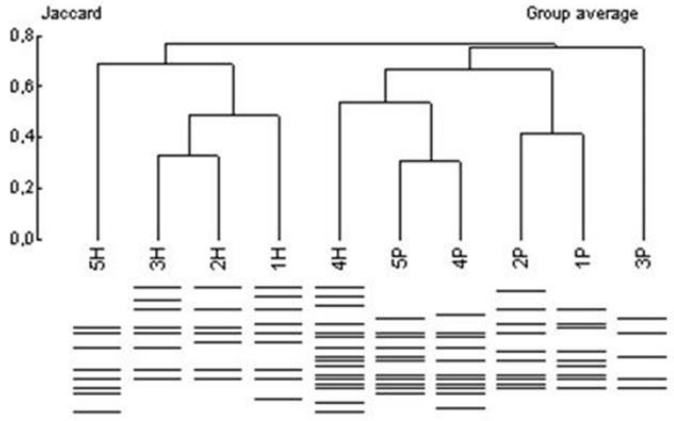
POSTUP:

Otevřeme si program Gel2k a zvolíme ikonu *File* a dále *Add image.* Vybereme požadovaný (předem upravený) snímek DGGE. Snímek je většinou zobrazován ve větší velikosti, proto si pro snazší práci se snímkem zvolíme funkci *Edit* a dále *Modify image,* v tabulce upravíme velikost obrazu např. na 60 %. Poté je potřeba označit část snímku, kterou chceme analyzovat, zpravidla označujeme k analýze část snímku, kde se nachází bandy, a okraje snímku z analýzy vynecháme.

Pro výběr analyzované části snímku zvolíme ikonu *Calibrate* a dále *Make frame.* Program na snímku vyznačí rámeček, se kterým bude dále pracovat. Poté si označíme jednotlivé „dráhy“ na snímku (tedy samotné vzorky) tak, že na vybranou dráhu klikneme myší. Objeví se okno, do něhož zapíšeme pojmenování vybrané dráhy (většinou stejně jako popisovaný vzorek např. S – sediment, K – kámen atd.). Program automaticky po popisu dráhy vzorek zanalyzuje a vloží nad oblast fotky histogram, kde označí ve vzorku jednotlivé bandy jako červené svislé úsečky. Každá dráha má svůj vlastní histogram, který se objeví při kliknutí na danou zanalyzovanou dráhu. Stává se, že někdy program označí band chybně, proto je možné při kliknutí na vrchol chybně označené červené úsečky band smazat, program se dotáže, zda chceme band smazat (*Delete peak?*), potvrdíme. Pro přidání bandu je potřeba v histogramu

kliknout myší na místo (černou svislo čáru histogramu), kde by měl být označen band a není, program se dotáže, zda chceme band přidat (*Add peak?*), potvrdíme.
Po této základní analýze obrázek uložíme jako soubor s příponou *.gel tak, že zvolíme funkci *File* a dále *Save as.* Při práci s dalšími funkcemi program někdy nahlásí error a zavře otevřené okno, proto je dobré uložit soubor hned po analyzování drah, abychom neztratili data. Soubor s příponou *.gel otevřeme pomocí funkce *File* a dále *Open gel-file* a v průběhu práce průběžně ukládáme změny pomocí funkce *Save gel-file.* Je dobré si také zkontrolovat rozložení a počet bandů pomocí funkce *View* a dále *Band pattern.* Po zadání tohoto příkazu se objeví okno s rozložením (růžově označených) bandů v jednotlivých vzorcích, kdy na spodním okraji můžeme vidět počet a v obrázku (zeleně označené) pozice chybně označených bandů (*errors*). Pro další práci se souborem je dobré, aby byl počet *errors* rovný 0. Chybné bandy můžeme odstranit pomocí změny nastavení šířky (tedy hranice intenzity) jednotlivých bandů. Nastavení parametrů pro bandy nalezneme pod funkcí *Setup* a dále *Bands,* kde si zvolíme optimální šířku bandu pro daný vzorek (tak aby počet *errors* byl 0 a pozice a počet bandů co nejvíce odpovídal reálnému snímku). Jakmile je snímek finálně zanalyzovaný, zvolíme funkci *File* a dále *Export Binary data.* Program nám tak uloží počty a pozice bandů ve všech označených vzorcích do souboru binárních dat s příponou *.bin (binární dat jsou souborem 1 a 0, kdy 1 značí přítomnost a 0 nepřítomnost bandu). Soubor s příponou *.bin dále vyhodnocujeme v programu Clust, který je součástí balíčku Gel2k.

Otevřeme si program Clust a zvolíme ikonu *File* a dále *Open* a vybereme si náš soubor s příponou *.bin. Dále zvolíme ikonu *File* a *Run.* Po tomto příkazu se nám v okně objeví cluster podobnosti mezi zkoumanými vzorky. U jednotlivých vzorků lze v clusteru dále zobrazit počty a pozice bandů pomocí funkce *Options,* dále *Plot setup* a *Show binary data.* Dále je možné v programu měnit typ shlukování a typ podobnosti pomocí funkce *Options* a dále *Cluster setup* (z typů shlukování máme na výběr z možností „Single link“, „Complete link“ a „Group average“; z typů podobnosti jsou to podobnosti “Jaccard”, “Faith” a “Dice”).
Výsledný cluster lze pak exportovat např. do MS Word pomocí funkce *Edit* a dále *Copy to clipboard* (výsledný cluster viz Obr. 1).



Obr. 1 Cluster podobnosti vzorků hyporheického sedimentu (Sitka)

Zpracování sekvenačních dat – konstrukce fylogenetických stromů a základy bioinformatiky

Bioinformatika je vědní disciplína na pomezí biologie a informatiky, která se zabývá zpracováním, prohledáváním a analýzou dat o sekvenci, struktuře a popřípadě i funkci biologických makromolekul, především DNA a proteinů.
Vzhledem k tomu, že většinu environmentálních mikroorganismů nelze kultivovat *in vitro,* jsou metody založené na bázi analýzy nukleových kyselin běžně využívány k identifikaci a fylogenetickému zařazení mikroorganismu. Identifikace nekultivovatelných mikroorganismů se rutinně provádí pomocí kombinace PCR amplifikace, následované klonováním a sekvenační analýzou.

Sekvenačními daty, získanými sekvenační analýzou, rozumíme zápis lineární posloupnosti monomerů (bazí) v molekule biologické makromolekuly – obvykle DNA (proteinu, RNA). Na výsledek sekvenační analýzy se pak zpravidla díváme jako na digitální zápis posloupnosti znaků, jdoucích po sobě v posloupnosti odpovídající směru biosyntézy daného typu molekuly. V případě nukleových kyselin je to od 5‘ ke 3‘ konci. K zápisu sekvencí se

používají jednopísmenové kódy, stanovené UIPAC (Mezinárodní unie pro čistou a užitou chemii). Protože není vždy možno jednoznačně stanovit každý monomer v sekvenci pro účely digitalizace, kódy IUPAC obsahují také možnost zápisu nejednoznačných pozic a proto zápis DNA může obsahovat skoro celou abecedu.
V nejjednodušší digitalizované podobě je sekvence zapsána jako prostý řetězec IUPAC znaků v textovém souboru. Tento formát se označuje jako surová data. Nejběžněji používaný formát je formát FASTA. Datový soubor ve formátu FASTA je textový soubor, jehož první řádek začíná znakem > (větší než), na němž je uveden název sekvence. Na dalších řádcích pak následuje surová sekvence.
Surové výsledky sekvenační analýzy jsou následně upraveny – ořezány o okrajové části s nejednoznačnými monomery (např. program Sequencher 4.1.4) a porovnány s databázemi genových knihoven a dle míry podobnosti je zkonstruován fylogenetický strom. Volné sdílení sekvenačních dat se děje pomocí veřejně dostupných zdrojů dat, databází. Tzv. „ velká trojka“ primárních databází nukleotidových sekvencí, je představována americkou databází GenBank, evropskou EMBL a japonskou DDBJ. Tyto databáze si denně vyměňují změny v datech a prakticky se zálohují (jejich obsah je tedy v zásadě totožný). Původní data může do databází vložit kdokoli, pokud dodrží formální požadavky na formát datových souborů. K vyhledávání a stahování záznamů slouží webové uživatelské rozhraní, získávat data a analyzovat je může kdokoliv. Nejčastěji používaná rozhraní jsou SRS (Sequence Retrieval System) evropské databáze EMBL a NCBI Entrez databáze GenBank.

Centrálním tématem bioinformatiky je porovnávání dvou či více sekvencí a zjišťování jejich vzájemné podobnosti. Obecně se dá říci, že stanovení podobnosti dvou sekvencí spočívá v jejich přiložení po celé délce do dvou řádků tak, aby identické pozice ležely pod sebou. Takový zápis se nazývá přiřazení – *alignment*. Poté se vypočte celková hodnota podobnosti (*score*). Zjednodušeně lze říci, že dojde ke stanovení poměru identických párů a nepárů mezi dvou sekvencí nukleotidů. K tomu lze využít počítačové programy, které používají algoritmy s různými variantami tohoto postupu (např. program CLUSTAL). Alignment lze dělat i ručně, například v programu BioEdit lze přesouvat i celistvé bloky sekvencí.
Nejrozšířenějším vyhledávacím heuristickým* algoritmem současnosti je algoritmus BLAST (*Basic Local Alignment Search Tool*), který má přehledné grafické zpracování, jednoduchý

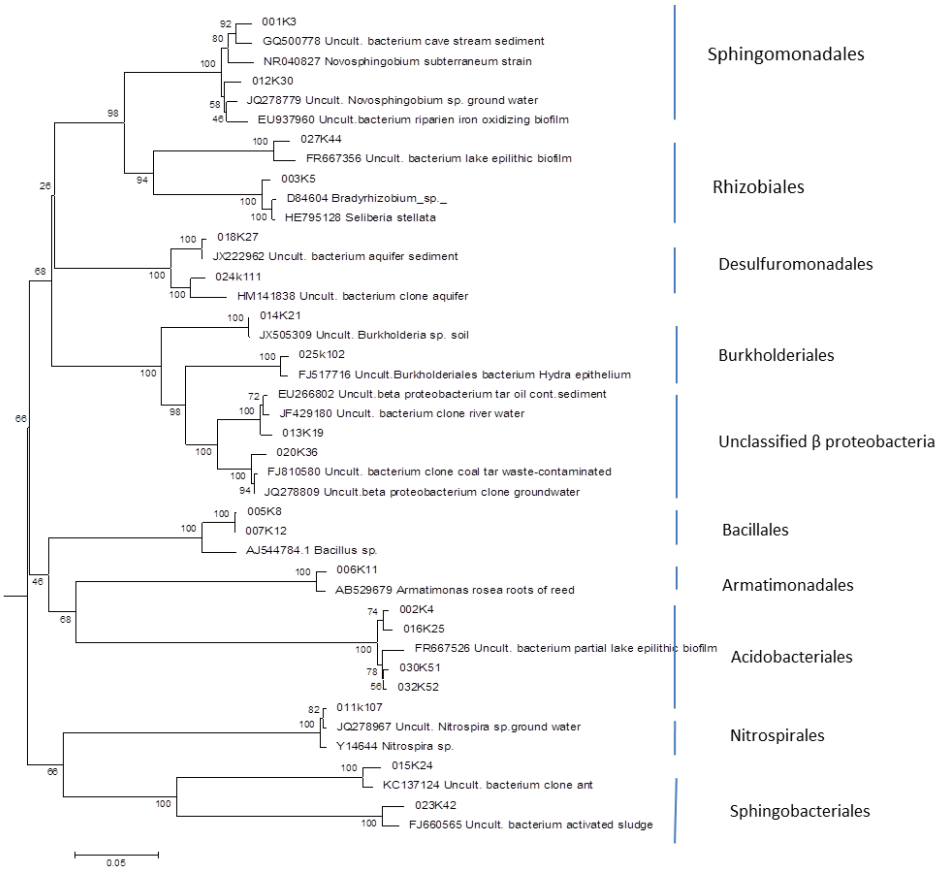
výběr varianty programu vhodného pro konkrétní typ dotazu a data. Nalezené nejbližší sekvence, seřazené podle míry podobnosti, lze přímo stáhnout z databáze pomocí webových odkazů.

*** Heuristika** (z řečtiny *heuriskó, εύρίσκω* – nalézt, objevit) znamená zkusmé řešení problémů, pro něž neznáme algoritmus nebo přesnější metodu. Heuristické řešení je často jen přibližné, založené na poučeném odhadu, intuici, zkušenosti nebo prostě na zdravém rozumu. První odhad se může postupně zlepšovat, i když heuristika nikdy nezaručuje nejlepší řešení. Zato je univerzálně použitelná, jednoduchá a rychlá.

Konstrukce fylogenetického stromu

Pomocí vybraných příbuzných sekvencí z databáze a z nich sestrojeného genealogického stromu lze vyčíst informace o příbuzenských vztazích mezi organismy a jejich geny. Předmětem studia ve vztahu ke konstrukci fylogenetických stromů, jsou zpravidla biologické druhy či jinak definované populace organismů – tzv. „ OTU - operační taxonomické jednotky“. OTU jsou představovány sekvencemi vybraných vzájemně ortologních genů (tj. homologní geny, vyskytující se v genomu v jedné kopii, která u všech zkoumaných organismů vykonává tutéž funkci - např. gen pro expresi koenzymu mcrA, klíčového pro produkci metanu u metanogenů). Cílem snažení je tedy konstrukce genealogického stromu studovaných sekvencí, nebo-li dendrogramu (viz. Obr. 2). Dendrogram je graf, který spojuje OTU pomocí větví (branches) s uzly (nodes), které reprezentují hypotetické společné předky propojených OTU. Délka větví představuje „evoluční vzdálenost“ uzlů, kterou si lze představit jako počet mutací, který odděluje uzly propojené danou mutací. Sestrojený strom může být „zakořeněný“, znamená to začlenění do analýzy tzv. *outgroup* – sekvenci, která je homologní, ale zároveň vzdálená studovaným vzorkům. Není-li k dispozici vhodná sekvence, sestrojí se strom nezakořeněný. K vlastnímu sestrojení stromu lze využít mnoho metod. Jednou z významných je např. metoda minimální evoluce, která se snaží minimalizovat součet délky větví, tzv. *neighbor-joining* (NJ), kterou využívá např. softwarový balíček PHYLIP. Metody „*maximum parsimony*“ pak hledají dendrogram odpovídající minimálnímu počtu mutací nezbytných k dosažení pozorovaného stavu. Tyto metody jsou využívány méně, protože jsou náchylné k artefaktům v případě, že od evolučního oddělení sekvencí uběhla dlouhá doba a tudíž se nahromadilo více mutací. Z hlediska spolehlivosti je asi nejlepší metoda *maximum likehood*, která prohledává všechny možné dendrogramy a stanovuje pravděpodobnost, s jakou mohl evoluční scénář generovat soubory znaků, odpovídajících vloženým znakům.

Po sestrojení dendrogramu je vhodné pokusit se vyhodnotit míru jeho důvěryhodnosti. Nejčastěji se k tomu používají vzorkovací metody, testující odolnost (robustnost) jednotlivých větví dendrogramu vůči změně (poškození) vstupních dat. Nejběžnější metodou je tzv. *bootstrapping*. Hodnoty bootstrappové podpory vypovídají o robustnosti struktury dendrogramu, čili jak je snadné tuto strukturu změnit přidáním nebo odebráním dat. Bootstrapping však nelze přímo interpretovat jako pravděpodobnost, že daná větev je správná. Za spolehlivé se považují zpravidla větve s bootstrapovou hodnotou 90 – 100 %, hodnoty pod 50 % by se neměly brát příliš vážně. Hlavním pravidlem pro konstrukci dendrogramů je, že jeho kvalita je podmíněna kvalitou přiřazení a ta zase kvalitou sekvenčních dat. A také platí, že při konstrukci dendrogramů je vždy lepší studovaná data zpracovat více metodami.



Obr. 2 Dendrogram fylogenetické příbuznosti bakteriálních klonů (16S rRNA) z epilitického biofilmu (Bystřice)

Použitá literatura a zdroje:

- Norland S. 2004. Department of Biology, University of Bergen, Norway http://folk.uib.no/nimsn/gel2k/
- Cvrčková F. 2006. Úvod do praktické bioinformatiky. Academia
- Stackebrandt E. 2006. Molecular identification, systematics, and population structure of procaryotes. Springer. Germany
- Sborn A. M., Smith C. J. 2005. Molecular microbial ekology. Tailor and Francis Group. UK
- www.wikipedie.com