AUTOMATIC GENOTYPING OF BACTERIA BY REP-PCR

Veronika Pelikánová

Master Degree Programme (2.), FEEC BUT

E-mail: xpelik16@stud.feec.vutbr.cz

Supervised by: Helena Škutková

E-mail: skutkova@feec.vutbr.cz

Abstract: This paper deals with automatic genotyping of bacteria by rep-PCR. The main goal is to create a program to analyze bacterial type in the samples. The result of the algorithm is phylogenetic tree, which indicates the cluster of samples according to bacterial type. This program consists of two main parts, a transformation function to reduce distortion of data and a clustering apparatus to bacterial type classification.

Keywords: genotyping, rep-PCR, chip capillary electrophoresis

1 ÚVOD

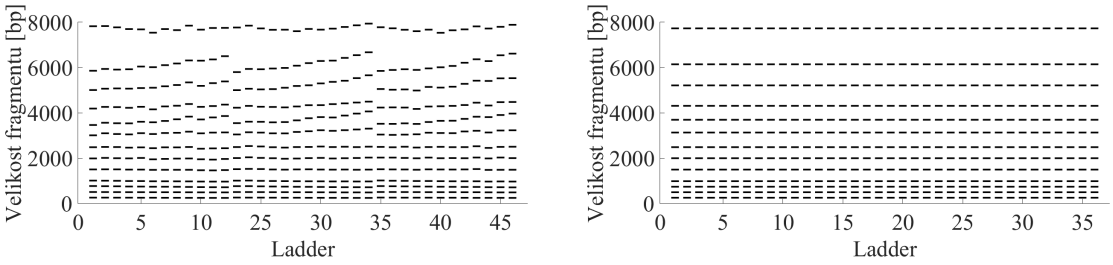
Genotypizace bakterií je pro klinickou praxi důležitá část laboratorního vyšetření, protože včasné zjištění původu bakteriální infekce může pomoci k rychlejšímu uzdravení pacientů a zamezení šíření v rámci oddělení či nemocnice.

K rozpoznání kmene bakterie je nutné ze vzorku získat potřebné části genomu a následně je kvantifikovat. Za tímto účelem se používá metoda rep-PCR, která je založena na předpokladu existenci hojně se opakujících repetitivních sekvencí v intergenních oblastí genomu [1]. Získané fragmenty bývají dále separovány elektroforetickou metodou, např. čipovou kapilární elektroforézou [2]. Data získaná kombinací těchto dvou metod jsou předložena k fylogenetické klasifikaci, díky níž jsou odlišeny vzorky dle genetické podobnosti.

Existuje řada softwarů umožňující fylogenetickou klasifikaci vzorků rep-PCR. Bohužel pro většinu klinických pracovišť jsou finančně nedostupné a jejich fungování je zastřeno „know-how“. Cílem práce je navrhnout algoritmus pro automatickou genotypizaci bakterií z výsledků čipové elektroforézy vzorků z rep-PCR, který by mohl tyto softwary nahradit.

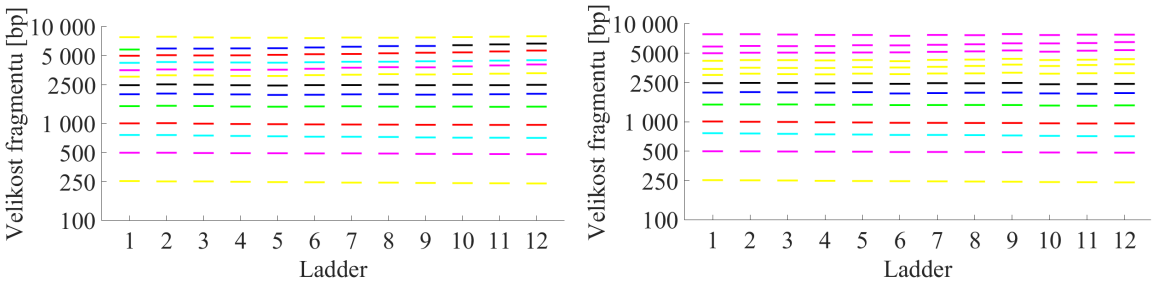
2 AUTOMATICKÁ GENOTYPIZACE BAKTERIÍ METODOU REP-PCR

Získaná surová data mají podobu, která není pro klasifikaci použitelná. Proto je nutná korekce pozic bandů, která připraví data pro následné operace sloužící přímo ke klasifikaci. Na Obr. 1 jsou ukázána surová data a data po předzpracování.



Obr. 1: Data před předzpracováním (vlevo) a po předzpracování (vpravo)

Obr. 2 ukazuje dva největší problémy, s nimiž je nutno se vypořádat. Vlevo je znázorněn problém rozdělení jedné skupiny fragmentů na více (na tři cca 500 bp, měly by být označeny jednou barvou). Vpravo jsou naopak dvě skupiny fragmentů určené jako jedna (růžová a žlutá nad 2500 bp). Nejvrchnější růžová skupina fragmentů by měla být rozlišena na tři.



Obr. 2: Ukázka problému při korekci bandů

2.1 NÁVRH ALGORITMU PRO AUTOMATICKOU KLASIFIKACI VZORKŮ

Fylogenetická analýza dat z automatické čipové elektroforézy se skládá ze dvou základních částí, předzpracování dat a samotné klasifikace.

Předzpracování dat z čipové elektroforézy je založeno na transformační funkci, ta byla sestavena na základě velkého množství hodnot naměřených standardizovaných DNA markerů. Analýza rozptylu těchto markerů ukázala, že rozptyl není na celém rozsahu stejný a ani lineárně rostoucí. Transformační funkce byla sestavena jako funkce po částech lineární. Upravená data jsou shlukovou analýzou (metoda nejbližšího souseda) zařazena do skupin velikostí fragmentů. Jak je ukázáno na Obr. 2 vpravo, některé bandy jsou tak blízké, že algoritmus klasifikuje 2 bandy (příp. více) jednoho vzorku do téže skupiny. Aby k tomuto jevu nedocházelo, je zařazena podmínková část.

Samotná fylogenetická klasifikace je umožněna ohodnocením vzorků asociační maticí. Při sestavování asociační matice jsou vždy dva vzorky porovnávány a ohodnoceny podílem neshod a celkovým počtem možných skupin fragmentů vyskytujících se v porovnávaných vzorcích. K vytvoření fylogenetického stromu je použito shlukování pomocí nejbližšího souseda.

Tab. 1: Správná a špatná klasifikace předzpracování dat

Soubor ladderů	L2+L3	L2+L3	L2+L3	L2+L3	L1+L3	L4	L4	L4	L1+L4	L1-4	Součet
Správně	168	168	168	180	168	156	152	150	156	1615	3013
Špatně	0	0	0	0	0	0	4	6	0	5	15
Procento chyb	0	0	0	0	0	0	0,026	0,038	0	0,003	0,005

2.2 VÝSLEDKY NA TESTOVACÍ MNOŽINĚ DAT

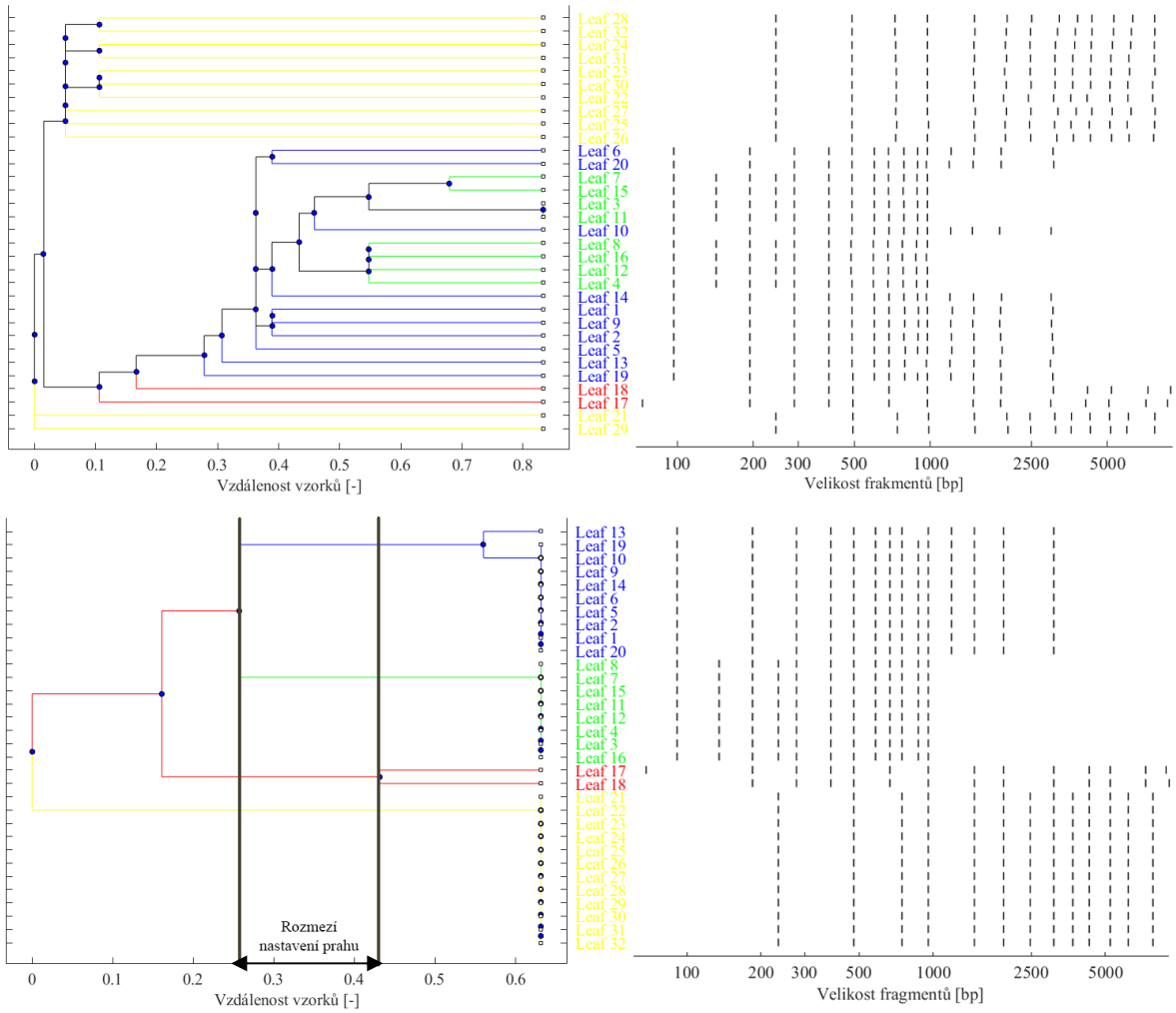
V Dětské nemocnici bylo speciálně naměřeno několik standardizovaných měření za účelem sestavení algoritmu. Na těchto datech lze algoritmus kvantitativně ohodnotit, na reálných datech by kvantifikace nebyla možná. K dispozici bylo 108 měření 4 ladderů (L1-GeneRuler 1 kb DNA Ladder,L2- GeneRuler 100 bp Plus DNA Ladder-GRP,L3- GeneRuler 50 bp DNA Ladder- GR 50 bp a L4 O'GeneRuler 1 kb DNA Ladder-OGR), což dohromady dalo přes 3000 bandů. V Tab. 1 jsou uvedeny hodnoty testování předzpracování dat. Při hodnocení zařazení bandů ke skupině fragmentů bylo špatně zařazeno nejvíce 0,038 % bandů, což hodnotím za dobrý výsledek.

Na Obr. 3 je vyobrazena v levé části klasifikační strom a v pravé části bandy odpovídající dané větvi. Nahoře je zobrazena klasifikace bez úpravy dat, v dolní části je klasifikace po úpravě dat. Zásadním problémem u vrchního obrázku je, obtížnost zvolení prahu pro rozdělení kmenů. V tomto případě pro neupravená data neexistuje práh, který by mohl vhodně oddělit kmeny. Po mnou navržené úpravě dat takový práh existuje a je možné jej volit ve větším rozmezí hodnot (konkrétně v tomto případě od hodnoty 0,25 do 0,43).

3 ZÁVĚR

Pro nemocniční praxi byl navržen algoritmus pro automatickou fylogenetickou klasifikaci bakteriálních kmenů z dat získaných z čipové kapilární elektroforézy a přecházející rep-PCR. Algoritmus byl navržen na 108 měřeních ladderů (více než 3000 bandů). Přestože se i tato standardizovaná měření ukázala značně zkreslená, chybovost předzpracování je menší než 0,04 %. Výsledná genotypizace na ladderech je velmi úspěšná, standardizované vzorky jsou vždy zařazeny do správné skupiny.

K algoritmu bylo sestaveno uživatelské rozhraní, díky němuž je možné program aplikovat v Dětské nemocnici v Brně, která rep-PCR s následnou čipovou kapilární elektroforézou využívá, a která poskytla data k sestavení algoritmu.



Obr. 3: Ukázka klasifikace

REFERENCE

[1] VERSALOVIC, James, Thearith KOEUTH a James R. LUPSKI. Distribution of repetitive DNA sequences in eubacteria and application to fingerprinting of bacterial genomes. Nucleic Acids Research. 1991, 19(24), 6823-6831.

[2] LAUER, Henk H. a Gerard P. ROZING, 2014. *High Performance Capillary Electrophoresis*. 2014. Germany: Agilent Technologies, 174 s. 5990-3777EN.