VYSOKÉ UČENÍ TECHNICKÉ V BRNĚ

Fakulta elektrotechniky
a komunikačních technologií

BAKALÁŘSKÁ PRÁCE

Brno, 2020 Hana Šandová

VYSOKÉ UČENÍ TECHNICKÉ V BRNĚ
BRNO UNIVERSITY OF TECHNOLOGY

FAKULTA ELEKTROTECHNIKY A KOMUNIKAČNÍCH TECHNOLOGIÍ
FACULTY OF ELECTRICAL ENGINEERING AND COMMUNICATION

ÚSTAV BIOMEDICÍNSKÉHO INŽENÝRSTVÍ
DEPARTMENT OF BIOMEDICAL ENGINEERING

GENOTYPIZACE BAKTERIÍ NA ZÁKLADĚ TEPLOTY TÁNÍ OLIGONUKLEOTIDŮ
BACTERIAL GENOTYPING BASED ON THE OLIGONUCLEOTIDE MELTING TEMPERATURE

BAKALÁŘSKÁ PRÁCE
BACHELOR'S THESIS

AUTOR PRÁCE — **Hana Šandová**
AUTHOR

VEDOUCÍ PRÁCE — **Ing. Markéta Nykrýnová**
SUPERVISOR

BRNO 2020

VYSOKÉ UČENÍ TECHNICKÉ V BRNĚ
FAKULTA ELEKTROTECHNIKY A KOMUNIKAČNÍCH TECHNOLOGIÍ

Bakalářská práce

bakalářský studijní obor **Biomedicínská technika a bioinformatika**

Ústav biomedicínského inženýrství

Studentka: Hana Šandová ***ID:*** 203204

Ročník: 3 ***Akademický rok:*** 2019/20

NÁZEV TÉMATU:

Genotypizace bakterií na základě teploty tání oligonukleotidů

POKYNY PRO VYPRACOVÁNÍ:

1) Seznamte se s výpočetními a laboratorními technikami pro stanovení teploty tání krátkých DNA sekvencí. 2) Navrhněte metodiku klasifikace sekvenčních typů bakterií na základě výpočetně stanovené teploty tání pro účely optimalizace genotypizační techniky Minim-typing. 3) V programovém prostředí Matlab realizujte funkce pro stanovení teploty tání DNA sekvencí na základě alespoň tří metod. 4) Srovnejte výpočetně stanovené hodnoty teplot tání s reálně změřenými hodnotami bakteriálních izolátů poskytnutých FN Brno. Zaměřte se na možnosti metod zohlednit různé pořadí stejných nukleotidů v sekvenci a chemické příměsi v měřeném vzorku DNA. 5) Proveďte klasifikaci bakterií rodu Klebsiella na základě teploty tání variabilních úseků genů infB, mdh, phoE, rpoB a tonB a srovnejte s laboratorními výsledky genotypizace metodou Minim-typing. 6) Vyhodnoťte diskriminační schopnosti použité metodiky a na základě výsledků navrhněte optimalizaci laboratorního protokolu pro Minim-typing.

DOPORUČENÁ LITERATURA:

[1] ANDERSSON, Patiyan, Steven Y. C. TONG, Jan M. BELL, John D. TURNIDGE, Philip M. GIFFARD a Igor MOKROUSOV. Minim Typing – A Rapid and Low Cost MLST Based Typing Tool for Klebsiella pneumoniae. PLOS ONE. 2012, 7(3), e33530. DOI: 10.1371/journal.pone.0033530. ISBN 1932-6203.

[2] PANJKOVICH, A. a F. MELO. Comparison of different melting temperature calculation methods for short DNA sequences. Bioinformatics. 2005, 21(6), 711–722. DOI: 10.1093/bioinformatics/bti066. ISBN 1367-4803.

Termín zadání: 3.2.2020 ***Termín odevzdání:*** 5.6.2020

Vedoucí práce: Ing. Markéta Nykrýnová

prof. Ing. Ivo Provazník, Ph.D.
předseda oborové rady

ABSTRAKT

Tato bakalářská práce se zabývá genotypizačními metodami, které využívají teplotu tání oligonukleotidů. V teoretické části je popsána DNA, je vysvětlen pojem typizace a jsou zde popsány laboratorní i výpočetní metody genotypizace. V praktické části byly v programovacím prostředí MATLAB navrženy funkce pro výpočet teploty tání ze sekvence a pro jejich spuštění bylo vytvořeno grafické uživatelské prostředí. Dále byla zpracovávána reálná data změřená pro 52 izolátů bakterie *Klebsiella pneumoniae* poskytnutá FN Brno. Byly porovnávány laboratorně určené teploty tání s teoretickými výpočty. Nakonec bylo zjišťováno, zda je možná klasifikace bakterií shlukovacími metodami na základě teoreticky vypočtených teplot tání oligonukleotidů.

KLÍČOVÁ SLOVA

Teplota tání, genotypizace, DNA, mini-multilokusová sekvenční typizace, vysokorozlišovací analýza křivek tání

ABSTRACT

This bachelor's thesis deals with genotyping based on the oligonucleotide melting temperatures. In the theoretical part, DNA, typing, experimental and theoretical genotyping methods are described. The practical part of the thesis deals with a program that is designed to calculate the melting temperatures of sequences based on different methods, and the graphical user interface was created. In the following part, the measured melting temperatures of bacterial isolates of *Klebsiella pneumoniae* are compared to the theoretically calculated melting temperatures. In the last part of the thesis, the possibility of genotyping using cluster analysis based on the theoretical melting temperatures is explored.

KEYWORDS

Melting temperature, genotyping, DNA, minim typing, High Resolution Melting analysis

ŠANDOVÁ, Hana. *Genotypizace bakterií na základě teploty tání oligonukleotidů*. Brno, Rok, 53 s. Bakalářská práce. Vysoké učení technické v Brně, Fakulta elektrotechniky a komunikačních technologií, Ústav biomedicínského inženýrství. Vedoucí práce: Ing. Markéta Nykrýnová

PROHLÁŠENÍ

Prohlašuji, že svou bakalářskou práci na téma „Genotypizace bakterií na základě teploty tání oligonukleotidů“ jsem vypracovala samostatně pod vedením vedoucího bakalářské práce a s použitím odborné literatury a dalších informačních zdrojů, které jsou všechny citovány v práci a uvedeny v seznamu literatury na konci práce.

Jako autorka uvedené bakalářské práce dále prohlašuji, že v souvislosti s vytvořením této bakalářské práce jsem neporušila autorská práva třetích osob, zejména jsem nezasáhla nedovoleným způsobem do cizích autorských práv osobnostních a/nebo majetkových a jsem si plně vědoma následků porušení ustanovení § 11 a následujících autorského zákona č. 121/2000 Sb., o právu autorském, o právech souvisejících s právem autorským a o změně některých zákonů (autorský zákon), ve znění pozdějších předpisů, včetně možných trestněprávních důsledků vyplývajících z ustanovení části druhé, hlavy VI. díl 4 Trestního zákoníku č. 40/2009 Sb.

Brno ....

podpis autorky

PODĚKOVÁNÍ

Ráda bych poděkovala vedoucí bakalářské práce paní Ing. Markétě Nykrýnové za konzultace, trpělivost a podnětné návrhy k práci.

Obsah

Úvod **10**

1 DNA **11**
1.1 Struktura DNA . . . 11
1.2 Amplifikace DNA . . . 11
1.2.1 Polymerázová řetězová reakce . . . 12

2 Typizace bakterií **14**
2.1 Multilokusová sekvenční typizace . . . 14

3 Techniky pro stanovení teploty tání oligonukleotidů **16**
3.1 Laboratorní techniky . . . 16
3.1.1 Vysokorozlišovací analýza křivek tání . . . 16
3.1.2 Minim typing . . . 19
3.2 Výpočetní techniky . . . 20
3.2.1 Elementární metoda . . . 20
3.2.2 Úprava soli . . . 21
3.2.3 Metoda nejbližšího souseda . . . 21
3.2.4 Fenomenologický model . . . 22

4 Realizace funkcí pro stanovení teploty tání DNA sekvencí **23**
4.1 Realizace pomocné funkce *pocty_bazi.m* . . . 23
4.2 Funkce *elementarni.m* a *uprava_soli.m* . . . 23
4.3 Funkce *NN.m* . . . 23
4.4 Funkce *fenomenologicka.m* . . . 24
4.5 Grafické uživatelské prostředí . . . 24

5 Návrh metodiky klasifikace sekvenčních typů na základě výpočetně stanovené teploty tání **26**
5.1 Metodika shlukování . . . 26
5.1.1 Hierarchické shlukování . . . 26
5.1.2 Nehierarchické shlukování - metoda k-means . . . 27
5.2 Použitá data . . . 28
5.3 Výsledky shlukování . . . 29

6 Srovnání laboratorně změřených a výpočetně stanovených teplot tání **32**
6.1 Popis reálných dat . . . 32

6.1.1 Surová data 32
6.1.2 Normalizovaná data 32
6.1.3 Diferenční data 32
6.1.4 Derivační data a určení teplot tání 33
6.2 Úprava dat, získání teplot tání 34
6.3 Výsledky srovnání teplot tání 34
6.4 Vliv různého pořadí stejných nukleotidů na vypočtenou teplotu tání . 39

7 Klasifikace bakterií na základě teploty tání variabilních úseků genů 41
7.1 Klasifikace 41
7.2 Vyhodnocení výsledků klasifikace 41
7.3 Návrh optimalizace 43

Závěr 47

Literatura 49

Seznam symbolů, veličin a zkratek 52

A Tabulky 53

Seznam obrázků

1.1 Struktura DNA 11
1.2 Schéma PCR 13
2.1 Schéma MLST 15
3.1 Křivka tání s normalizovanou fluorescencí, derivační křivka 17
3.2 Křivky tání pro různé kombinace A a C 18
3.3 Porovnání křivek heterozygota s homozygotem, zobrazení diference mezi jednotlivými vzorky 19
3.4 Schéma mini-MLST 20
4.1 Grafické uživatelské prostředí sloužící pro výpočet teplot tání zadaných sekvencí různými numerickými metodami 25
5.1 Dendogram vytvořený metodou UPGMA a euklidovskou vzdáleností . 30
5.2 Analýza siluet a výsledky shlukování pro kosinovou vzdálenost pro 7 shluků 31
6.1 Zobrazení závislostí různě upravené fluorescence na teplotě změřených pro 52 izolátů bakterie *Klebsiella pneumoniae* pro fragment genu tonB28 33
6.2 Určení T_m z derivační křivky pro fragment genu tonB28 vzorku KP1231 34
6.3 Určení průměrné T_m alel pro gen tonB28 35
6.4 Vznik různých (A) nebo stejných (B) dvojic sousedících bází v závislosti na místě začlenění vyštěpeného řetězce 40
7.1 Výsledek hierarchického shlukování (dendrogram) pro elementární metodu 42
7.2 Výsledek hierarchického shlukování (dendrogram) pro metodu nejbližšího souseda 44
7.3 Výsledek k-means shlukování teplot tání analyzovaných vzorků elementární metodu 45
7.4 Výsledek k-means shlukování teplot tání analyzovaných vzorků metodu nejbližšího souseda 46

Seznam tabulek

3.1 Parametry síly (St) pro jednotlivé dinukleotidy 22
5.1 18 náhodně vybraných izolovaných vzorků bakterie *Klebsiella pneumoniae* poskytnutých FN Brno s laboratorně stanovenými teplotami tání a sekvenčním typem použité pro návrh metodiky shlukování . . . 28
5.2 Porovnání průměrných siluet při různém počtu shluků (k) pro vybraných 18 vzorků, pro čtverec euklidovské vzdálenosti, kosinovou a manhattanskou metriku 31
6.1 Průměrné teploty tání určené laboratorně a čtyřmi různými numerickými metodami 36
6.2 Diference teplot tání mezi laboratorně určenými a vypočtenými . . . 37
6.3 Porovnání délek analyzovaných sekvencí genů (fragmentů genů používaných pro mini-MLST) 38
6.4 Teploty tání tří různých řetězců o stejném nukleotidovém zastoupení z obrázku 6.4 vypočítané metodou nejbližšího souseda a fenomenologickou metodou 40
A.1 Nukleotidový kód IUPAC 53
A.2 Hodnoty změn entalpie a entropie pro výpočet teploty tání metodou nejbližšího souseda 53

Úvod

Jedním z nejrozšířenějších mikroorganismů na této planetě jsou bakterie. Mohou být člověku prospěšné, ať už přímo usídlené v našem organismu v symbiotickém vztahu (např. bakterie střevní mikroflóry), nebo jejich využitím v průmyslu (např. k výrobě kyseliny mléčné). Na druhou stranu existuje řada patogenních bakterií, způsobující mnohá onemocnění. A právě tyto bakterie je často potřeba identifikovat (tzv. typizovat), protože jejich správné zařazení následně může odhalit zdroj nákazy nebo i určovat léčbu.

Typizace je metoda, která slouží k bližší identifikaci bakterií jednoho druhu, jejich rozdělení do jednotlivých kmenů. Dříve probíhala klasifikace bakterií převážně podle pozorovatelných vlastností a znaků (fenotypická analýza). Dnes se díky novým technologiím přidala identifikace na základě genetického kódu (genotypizace), která je mnohem účinnější a je schopna mnohdy rozlišit i linie, které jsou mezi sebou blízce příbuzné.

Typizační metody se využívají především v epidemiologii, kdy se sledují zdroje a cesty šíření původců onemocnění. Příkladem je bakterie *Klebsiella pneumoniae*, která bývá častou příčinou nozokomiálních infekcí, šířících se po nemocnicích. Včasné odhalení zdroje nákazy nebezpečného kmenu této bakterie může zabránit pandemii.

Genotypizací lze odhalit i rezistenci bakterie vůči určitému antibiotiku. Informace o antibiotické rezistenci bakterie může mít i životně důležitou roli pro pacienta. O tom svědčí například nemoc tuberkulóza způsobená bakterií *Mycobacterium tuberculosis*. Tato nemoc se dnes vyskytuje už převážně jen v rozvojových zemích, kde zdaleka není potřebné technické vybavení, a to ani pro určení rezistence. Proto se nemocní léčí primárními antibiotiky a rezistence je většinou odhalena, až když u pacienta primární léčba nezabírá. V této chvíli však může být na podání účinnějších antibiotik už pozdě.

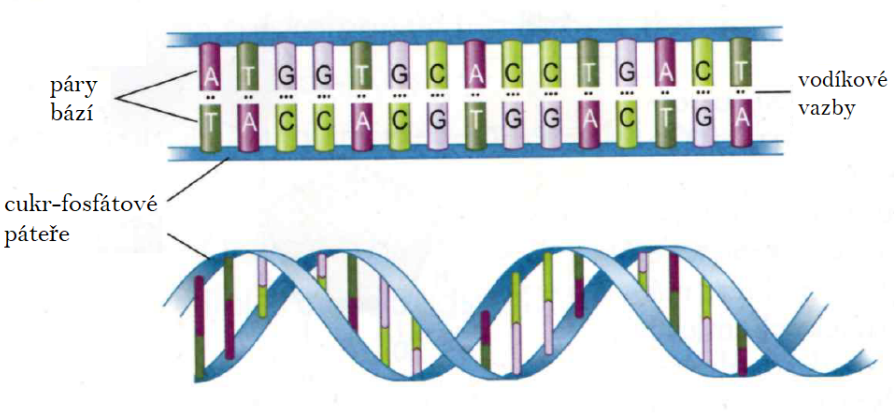
Pro oba výše uvedené příklady by bylo potřeba zajištění rychlé, přesné, jednoduché, a z důvodů velkého množství prováděných testů, i cenově nenáročné typizační metody. Tato metoda by se uplatnila i v určitých případech epidemií, kdy jsou laboratoře molekulární biologie a genetiky denně zahlcovány desítkami, stovkami, případně až tisícovkami vzorků vyžadujících co nejrychlejší analýzu.

Tématem této práce je rozvíjející se genotypizační metoda, která tento potenciál vykazuje. Jedná se o metodu mini-multilokusové sekvenční typizace, která je založená na určování teploty tání krátkých sekvencí DNA. Cílem práce je srovnání laboratorně určených teplot tání bakteriálních izolátů *Klebsiella pneumoniae* poskytnutých FN Brno s výpočetně stanovenými a následně určit, zda je pomocí těchto vypočtených teplot možná klasifikace poskytnutých vzorků.

1 DNA

1.1 Struktura DNA

V polovině 20. století bylo prokázáno, že DNA (z angl. deoxyribonucleic acid) je nositelkou dědičné informace. Její základní jednotkou je nukleotid, který se skládá z molekuly cukru (konkrétně deoxyribózy), molekuly fosfátu a dusíkaté báze. Genetická informace je uchována v uspořádání různých kombinací čtyř druhů dusíkatých bází, adeninu (A), cytosinu (C), guaninu (G) a thyminu (T). V roce 1953 se James Watson a Francis Crick dopracovali k závěru, že se DNA vyskytuje jako pravotočivá dvojitá šroubovice, ve které se dva polynukleotidové řetězce obtáčejí navzájem kolem sebe ve spirále. Polynukleotidové řetězce se skládají ze sekvence nukleotidů, které jsou navzájem spojeny kovalentními fosfodiesterovými vazbami. Tato dvě vlákna, tvořící páteř DNA, jsou držena pohromadě poměrně slabými vodíkovými vazbami, které vznikají mezi bázemi opačných vláken. Párování bází je specifické, A se páruje s T za vzniku dvou vodíkových vazeb a C se páruje s G za vzniku tří vodíkových vazeb. Struktura je zobrazena na obrázku 1.1. Proto jsou řetězce DNA označovány jako komplementární. [1]



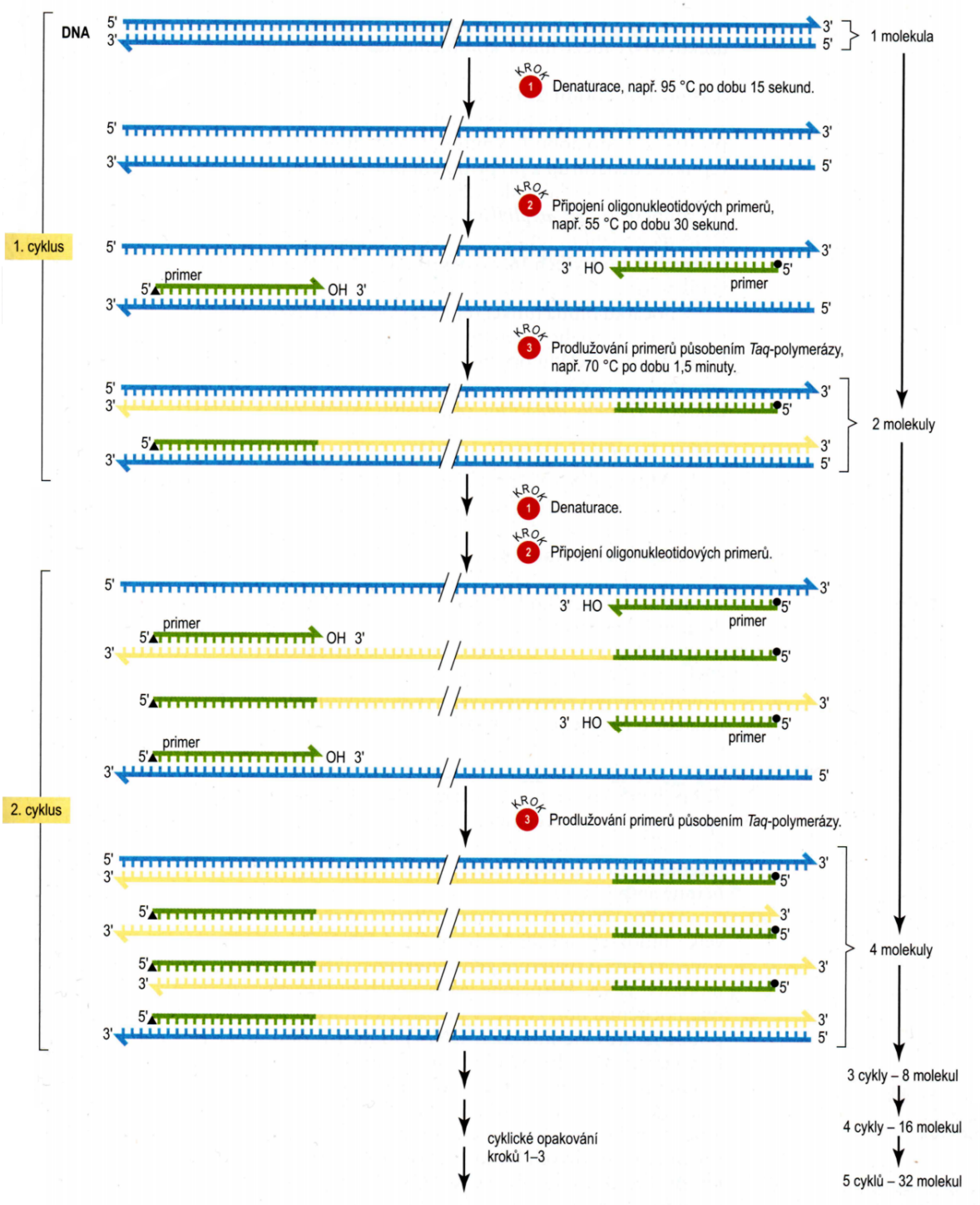
Obr. 1.1: Struktura DNA, převzato z [1].

1.2 Amplifikace DNA

Pro většinu laboratorních genetických analýz je potřeba větší množství DNA, než je v odebraném vzorku. Z toho důvodu se provádí namnožení zájmových sekvencí DNA *in vitro*. Nejčastější používanou metodou je polymerázová řetězová reakce (PCR).

1.2.1 Polymerázová řetězová reakce

Metoda PCR začíná vytvořením směsi reagencií. Tato směs obsahuje templátovou DNA, jejíž část chceme namnožit, enzym DNA polymerázu, který buduje nový řetězec podle templátu, nukleotidy, primery, což jsou krátké syntetické sekvence, které jsou komplementární ke známým sekvencím ohraničující oblast zájmu, a pufr, který vytváří vhodné prostředí pro reakce. Tento mix následně prochází třemi kroky, které se mnohokrát opakují. Nejprve je asi na 15 sekund směs zahřáta na 92 - 95 °C. Při této teplotě řetězec denaturuje a rozplete se. V dalším kroku dochází při teplotách okolo 50 - 60 °C k navázání oligonukleotidových primerů. Poslední krok probíhá 1 - 3 minuty při teplotách okolo 70 - 72 °C, kdy DNA-polymeráza za navázanými primery syntetizuje podle templátového řetězce nové komplementární vlákno. Tyto tři kroky se opakují, dokud nedosáhneme požadovaného stupně amplifikace. Schéma je na obrázku 1.2. [1], [2]



Obr. 1.2: Schéma PCR, převzato z [1].

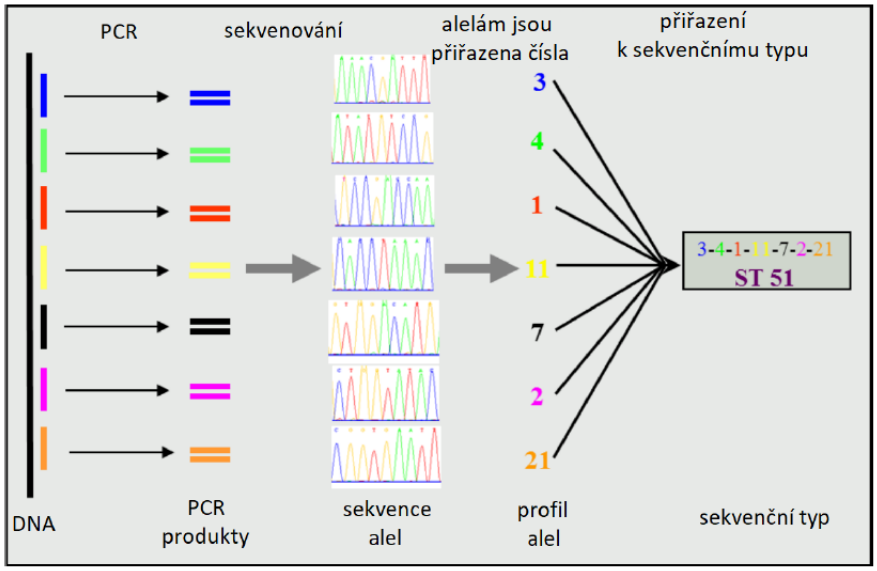
2 Typizace bakterií

Typizace nám slouží k bližší identifikaci bakterií stejného druhu, které je nezbytné především pro epidemiologii. Pro klasifikaci je používána řada metod. První metody spočívaly pouze v pozorování vlastností a znaků (fenotypu) organizmu. Například sledování tvaru a velikosti (tedy morfologie) buňky, nebo zařazení podle Gramova barvení (tj. podle stavby buněčné stěny). Dále se dá zaměřit na kultivační vlastnosti bakterií. Jaké velikosti a tvaru nabývají kolonie, jaké podmínky k růstu vyžadují, a jak reagují na jejich změny. Tyto metody jsou stále zlatým standardem v rutinní lékařské mikrobiologii a to především u druhů mikroorganismů, které nemají zvláštní požadavky na růst. Tento postup totiž poskytuje čistou kulturu potřebnou pro testování citlivosti na antibiotika.

V dnešní době se čím dál častěji používají i molekulárně-genetické techniky (genotypizace), které klasifikují bakterie na základě jejich genetické informace. Těchto metod je velké množství v nejrůznějších modifikacích. Patří sem například pulzní gelová elektroforéza (PFGE), celogenomové sekvenování nebo multilokusová sekvenční typizace (MLST) a mini-multilokusová sekvenční typizace (mini-MLST), které budou popsány dále. Pro tyto metody je často nutné předem DNA amplifikovat. [1], [3]

2.1 Multilokusová sekvenční typizace

Metoda multilokusová sekvenční typizace byla poprvé popsána roku 1998. Postup této metody začíná amplifikováním několika standardizovaných fragmentů z genů (tzv. housekeeping genes), kódujících proteiny, které jsou nezbytné pro chod buňky. Vybrány jsou právě tyto typy genů, protože jejich sekvence jsou poměrně stabilní. Tyto úseky o délce okolo 450 bp jsou osekvenovány a každá unikátní alela je číslována chronologicky dle objevu. Sekvenční typ (ST) je pak dán souborem čísel, kdy každé definuje alelu konkrétního genu. Schéma metody je znázorněno na obrázku 2.1. Všechny alely a sekvenční typy jsou pod tímto kódováním přístupné na internetu. Tato metoda má vysokou rozlišovací schopnost a je reprodukovatelná. Mezi její nevýhody patří finanční náročnost způsobená především sekvenováním jednotlivých fragmentů. Tento problém řeší odvozená metoda mini-MLST, která je popsána v kapitole 3.1.2. [4], [5]



Obr. 2.1: Schéma MLST, převzato z [6].

3 Techniky pro stanovení teploty tání oligonukleotidů

Při pokojové teplotě je DNA stabilní a zachovává si dvouřetězcovou podobu (dsDNA, z angl. double-stranded DNA). Při zvyšování teploty však dochází k denaturaci DNA a dvoušroubovice je postupně rozplétána až do podoby dvou jednořetězových DNA (ssDNA, z angl. single-stranded). Teplota, při níž je rozpleteno 50 %, se nazývá teplota tání T_m (z angl. melting temperature). Výše teploty tání molekuly DNA závisí především na dvou faktorech, na její délce a zastoupení párů bází guanin cytosin. S rostoucí délkou molekuly a také s vyšším procentuálním zastoupením GC párů je potřeba více tepelné energie k rozpletení řetězce kvůli většímu množství vodíkových vazeb. [3]

3.1 Laboratorní techniky

V roce 2002 byla představena nová metoda genotypizace, pracující na základě analýzy teploty tání DNA řetězců, tzv. vysokorozlišovací analýza křivek tání (HRM z angl. High Resolution Melting). Tato metoda našla uplatnění například při genotypizaci bakterií nebo při skenování mutací. Popularita této metody značně roste, protože disponuje několika výhodami oproti ostatním genotypizačním a sekvenačním metodám. Mezi hlavní klady patří jednoduchost a rychlost provedení, finanční nenáročnost a navíc je i poměrně citlivá a spolehlivá. [7]

Pro určování jednonukleotidových polymorfismů (SNP) u některých druhů bakterií (např. *Klebsiella pneumoniae, Streptococcus pyogenes, Staphylococcus aureus* aj.) byl zaveden nový přístup zvaný minim typing nebo také mini-MLST, který je modifikací metod MLST a HRM. [5]

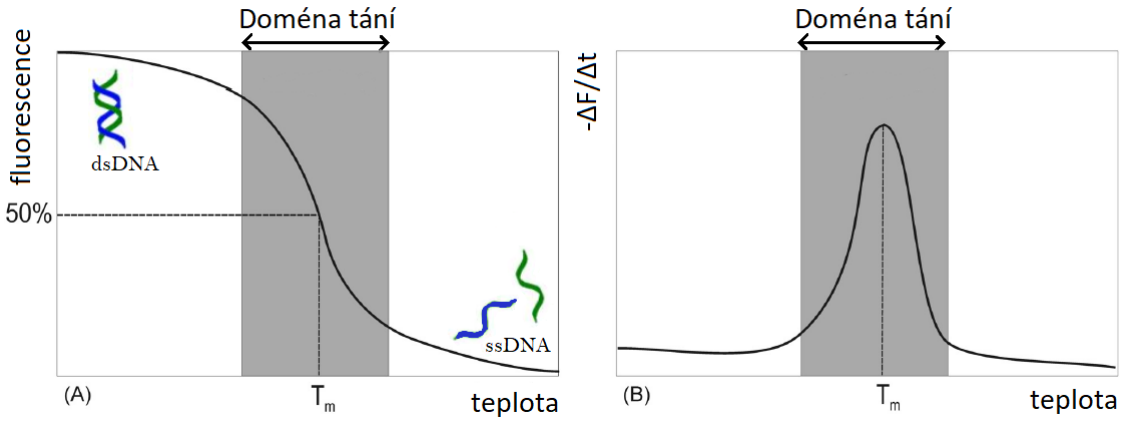
3.1.1 Vysokorozlišovací analýza křivek tání

Dříve byla teplota tání monitorována měřením absorbance UV záření (u denaturované DNA dochází k vyšší absorpci než u nativní). Aby touto metodou vznikly kvalitní křivky tání, bylo potřeba velké množství izolované DNA (v řádech μg) a kompletní měření často trvalo až hodiny. Tyto dva problémy odezněly poté, co byly na trh přivedeny první real-time PCR kombinované s termocyklerem a fluorimetrem. S touto technologií trvá analýza teploty tání pár minut a stačí nanogramy DNA.

Dnešní metody snímají fluorescenci, proto je nutné do vzorku přidat saturační barviva, která mají vysokou afinitu k dsDNA. Většina používaných barviv navíc neovlivňuje průběh PCR, a proto se mohou přidávat před amplifikací, což šetří čas

a snižuje pravděpodobnost kontaminace vzorku. Je ale nutné zmínit, že koncentrace barviva ovlivňuje tání produktu, konkrétně může způsobit posun křivek tání k vyšším teplotám. Proto je pro reprodukovatelnost výsledků určitých testů nutná optimalizace a dodržování protokolů. Obarvený produkt PCR je vystaven postupnému a řízenému zvedání teploty rychlostí 0,1 - 1 °C/s. Se zvyšující se teplotou dochází k rozvolňování („tání“) dvouřetězce, tím se z něj uvolňuje navázané barvivo, a přístroje zaznamenávají snižování fluorescence. Vykreslená závislost fluorescence na teplotě se označuje jako křivka tání. Nejedná se o lineární proces. Křivka pro krátké řetězce kopíruje tvar sigmoidy, pro delší sekvence může nabývat komplexnějších tvarů.

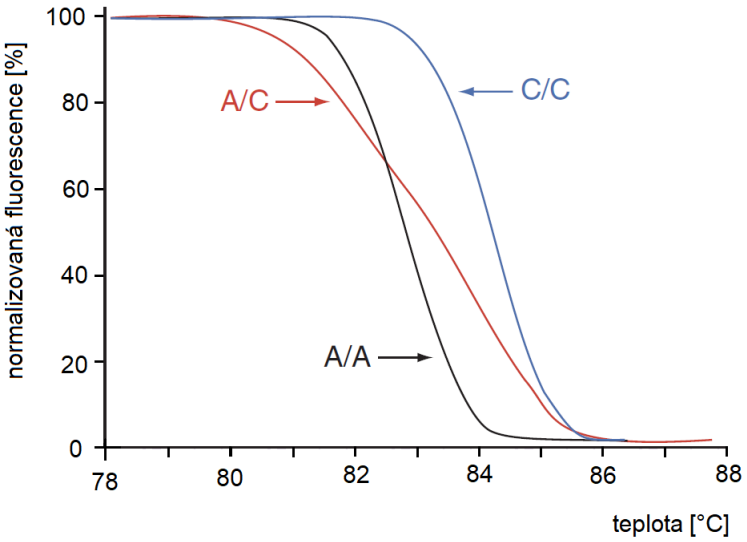
Aby bylo možné surová data, kdy se zaznamenává závislost fluorescence na teplotě, porovnat, je nutné data normalizovat. Tím budou všechny křivky začínat ve stejném místě na ose y. Normalizace mezi hodnotami 0 až 1 probíhá v takzvané doméně tání. Doména tání se nachází v úseku, kde dochází k prudkému poklesu fluorescence, což odpovídá tání analyzovaného řetězce. Pak 0 % fluorescence souhlasí s plně „roztátou“ DNA (ssDNA) a 100 % odpovídá dsDNA. Inflexní bod křivky představuje 50 % a určuje teplotu tání (T_m) (obrázek 3.1 (A)). Pro snadnější získání teploty tání může být zobrazena také derivační křivka (obrázek 3.1 (B)), která bývá zkonstruována jako závislost první záporně vzaté derivace fluorescence podle času ($-\Delta F/\Delta t$) na teplotě. V této křivce pak odpovídá teplotě tání vrchol. [3], [7]



Obr. 3.1: Křivka tání s normalizovanou fluorescencí (A), derivační křivka, která byla vypočítána z normalizovaných dat jako první záporně vzatá derivace fluorescence podle času vykreslená v závislosti na teplotě (B), převzato z [3].

Teplota tání představuje sice velmi vhodnou porovnávací jednotku, ale jedná se pouze o jeden bod celé křivky tání, která jako celek obsahuje mnohem více informací, a proto bývá často využíván pro porovnávání tvar celé křivky.

Metoda HRM dokáže zaznamenat změnu i v jedné bázi. Na obrázku 3.2 je možné vidět křivky pro různé záměny adeninu a cytosinu, konkrétně varianty AA, AC a CC. Z obrázku je patrné, že homozygoti AA a CC mají stejný tvar křivky, pouze křivka homozygota CC je posunuta k vyšší teplotě, protože obsahuje trojnou vazbu. Oproti tomu křivka heterozygota vykazuje rozdílný tvar. Ke snižování fluorescence zde dochází už při nižších teplotách a křivka nedosahuje takového spádu jako u homozygotů.

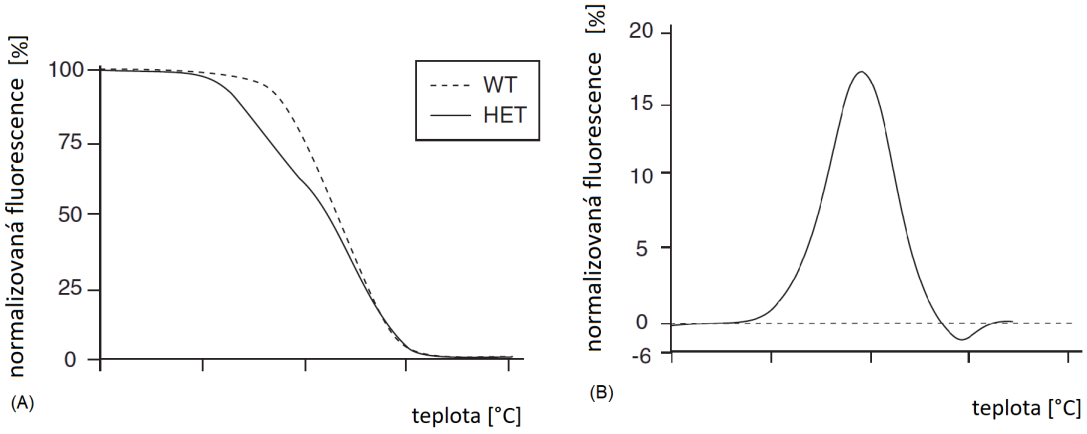


Obr. 3.2: Křivky tání pro různé kombinace A a C, převzato z [7].

Někdy není potřeba ani znát kompletní genotyp, ale pouze nás zajímá shoda určitých sekvencí. I pro takové případy je HRM velmi vhodná, protože stačí porovnat křivky tání obou sekvencí. Toho by se dalo velmi dobře využít v transplantační medicíně při zjišťování kompatibility mezi příjemcem transplantátu a dárcem, kdy by stačilo porovnat křivky tání genů HLA. Porovnávání křivek lze také využít pro skenování mutací, kdy porovnáváme normalizované křivky heterozygotního vzorku (s mutací) a homozygotní referenci (nebo průměr všech křivek běžně se vyskytujících typů). Aby bylo možné křivky porovnat, je třeba jejich základny posunout po ose x (teplotní osa) tak, aby se překrývaly. Příklad porovnání je na obrázku 3.3 (A). Rozdíly v tomto grafu se zdají nepatrné, proto se zobrazuje diference mezi jednotlivými vzorky viz obrázek 3.3 (B). [7]

Pro kvalitní analýzu je často nutné použít jen konkrétní krátké úseky DNA. U menších fragmentů je změna teploty v závislosti na obsahu GC párů poměrně dobře předvídatelná a detekovatelná, ale pro úseky už s malým navýšením nad 100 bp

může analýzu značně zkreslovat. Kratší produkty PCR obvykle také tají pouze s jednou doménou tání a tedy s jediným vrcholem v derivační křivce. Více vrcholů v křivce se také objeví, pokud je v analyzovaném vzorku více oligonukleotidů, jako nespecifické produkty DNA nebo dimery. Křivka tedy může sloužit jako ukazatel čistoty výstupů z PCR. Tím pádem je také nutné navrhnout přesný primer, který bude amplifikovat pouze požadovaný oligonukleotid. [3],[8],[9]

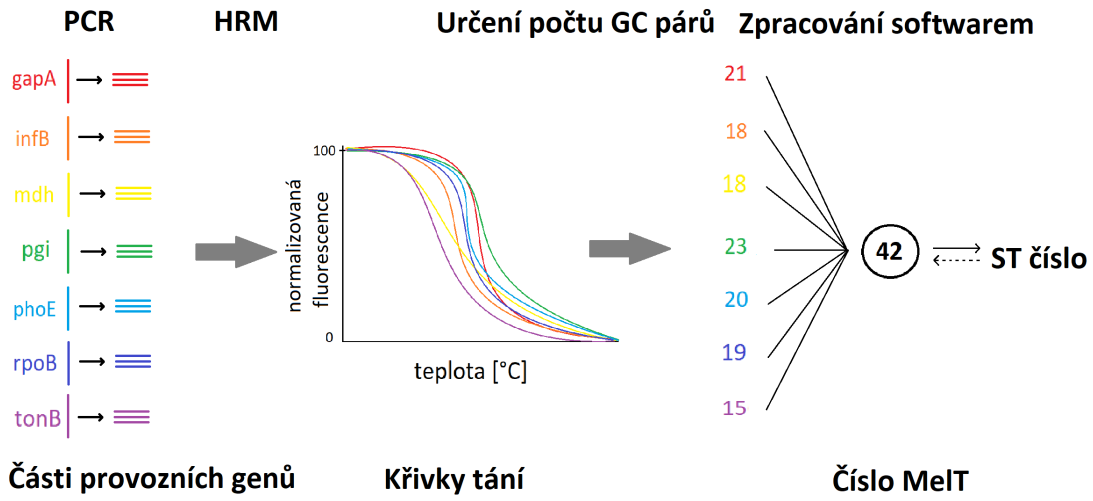


Obr. 3.3: Porovnání křivek heterozygota (HET) s homozygotem (WT) (A), zobrazení diference mezi jednotlivými vzorky (B), převzato z [7].

3.1.2 Minim typing

Minim typing, nebo také mini-MLST, je poměrně nová typizační metoda využívána především pro některé bakterie. Stejně jako u metody MLST jsou nejprve amplifikovány určité části stabilních provozních genů. Místo finančně nákladné sekvenace, ale poté probíhá analýza křivek tání.

Pomocí křivek tání je následně určován mimo jiné i počet GC párů. Křivky bývají porovnávány s referenční sekvencí, u níž je počet GC párů znám. Hodnoty GC párů a další analýza křivek identifikuje alelu. Tyto informace se dále pomocí softwaru zpracovávají a na jeho výstupu je definováno číslo označované jako MelT. Podle převodního klíče je dále možné převádět ST na MelT a do jisté míry i naopak. Schéma metody je zobrazeno na obrázku 3.4. [5], [10], [11]



Obr. 3.4: Schéma mini-MLST.

3.2 Výpočetní techniky

Pro některé techniky molekulární biologie je predikce přesné hodnoty teploty tání často zásadní. Jsou to především metody, u kterých se reakce účastní více oligonukleotidů současně, jako je kvantitativní nebo multiplexní PCR. U těchto metod může špatně stanovenou hodnotou T_m dojít k znehodnocení požadovaných výstupů reakcí. Pro stanovení teploty tání existuje několik výpočetních metod, které lze dále různě modifikovat (např. změnou parametrů). Na internetu je řada volně přístupných softwarů, které tyto metody pro výpočet T_m využívají. Použitím různých metod ale bohužel ve většině případů obdržíme rozdílné hodnoty T_m, pak je velmi složité se rozhodnout, která je ta přesná. Teplota tání totiž závisí na více faktorech, nejen na obsahu GC párů, ale i na uspořádání bází v sekvenci, na koncentraci solí v roztoku (resp. na iontové síle), na pH roztoku aj. Každá metoda pak bere v potaz tyto faktory různou mírou. [12], [13]

3.2.1 Elementární metoda

Elementární metoda (z angl. basic method) byla jedna z prvních implementovaných metod. Předpokládá, že proces probíhá v pufrovaném roztoku, s koncentrací sodných iontů Na^+ 50 mM, 50 nM koncentrací oligonukleotidů a hodnotou pH blízké 7. Jedinými proměnnými tu jsou počty jednotlivých bází. Výpočet T_m probíhá podle rovnice:

$$T_m = 64,9 + 41,0 \cdot \frac{G + C - 16,4}{A + T + G + C}, \tag{3.1}$$

nebo v jednodušším tvaru pro sekvence kratší než 14 bází:

$$T_m = 2 \cdot (A + T) + 4 \cdot (C + G) - 7, \tag{3.2}$$

kde T_m je teplota tání ve stupních Celsia a G, C, A a T jsou počty bází. [12], [14]

3.2.2 Úprava soli

Koncentrace solí v roztoku také ovlivňuje teplotu tání. A přestože je stále poměrně málo známá přesná závislost mezi T_m krátkých DNA řetězců a koncentrací iontů, především těch dvojmocných, byla vytvořena řada korekčních faktorů v rovnicích pro výpočet T_m. Rovnice pro výpočet:

$$T_m = 100,5 + 41,0 \cdot \frac{G + C - 16,4}{A + T + G + C} - \frac{820,0}{A + T + G + C} + 16,6 \cdot log\left[Na^+\right], \tag{3.3}$$

kde T_m je teplota tání ve stupních Celsia. G, C, A a T jsou počty bází a $[Na^+]$ je koncentrace sodných iontů v $mol \cdot l^{-1}$. Druhý člen rovnice upravuje závislost na počtu GC párů a třetí člen délku řetězce. [12]

3.2.3 Metoda nejbližšího souseda

Metoda nejbližšího souseda (NN, z angl. Nearest Neighbor) je označována za jednu z nejpřesnějších a nejpoužívanějších metod. Tento model předpokládá, že potřebná energie k rozpletení duplexu nezávisí pouze na jednotlivém zastoupení bází, ale, jak už název napovídá, pro oddělení jednoho nukleotidového páru hraje roli i nukleotidový pár, který po něm následuje. Vzorec pro výpočet je následující:

$$T_m = \frac{\Delta H}{A + \Delta S + R \cdot ln(\frac{C}{4})} - 273,15 + 16,6 \cdot log\left[Na^+\right], \tag{3.4}$$

kde T_m je teplota tání ve stupních Celsia, ΔH značí změnu entalpie s jednotkou $kcal \cdot mol^{-1}$, A je konstanta (A = -0,0108 $kcal\ K^{-1} \cdot mol^{-1}$), ΔS určuje změnu entropie v $kcal \cdot K^{-1} \cdot mol^{-1}$, R je plynová konstanta (R = 0,00199 $kcal\ K^{-1} \cdot mol^{-1}$), C stanovuje koncentraci oligonukleotidů v $mol \cdot l^{-1}$, hodnota -273,15 je konverzní faktor, který převádí výslednou teplotu tání z K na $°C$ a $[Na^+]$ je koncentrace sodných iontů v $mol \cdot l^{-1}$.

Ale i při použití této jedné metody pro výpočet teploty tání, je možné získat více rozdílných výsledků. Termodynamické parametry ΔH a ΔS jsou totiž pro jednotlivé dvojice sousedících bází určovány experimentálně. Byla publikována řada tabulek

těchto parametrů, které se od sebe liší, protože byly odvozovány z jiných datových souborů. [12], [15]

3.2.4 Fenomenologický model

Tato metoda zohledňuje fyzikálně-chemické události probíhající při procesu tání. Zaměřuje se na nutnost přerušit vodíkové vazby mezi komplementárními bázemi, účinek solí a koncentraci nukleotidových řetězců. Počítá i s nutností přerušení vazeb, které vznikají interakcemi mezi sousedícími bázemi, je tedy do jisté míry podobná metodě nejbližšího souseda. U tohoto modelu se však nepočítá s entropií a entalpií, ale s malými celými čísly, které byly stanoveny na základě experimentů. Nejprve byla pomocí simulací určena síla interakce mezi dinukleotidy, tedy dvojicí sousedících nukleotidů. Dvojice byly rozděleny na purinové (R), kam patří A a G, a pyrimidinové (Y), obsahující T a C. Vznikly čtyři skupiny možných interakcí RR, RY, YR a YY. Pro tyto skupiny byly vyzkoušeny náhodné kombinace hodnot a nakonec jim byly přiřazeny hodnoty RR = 3, RY = 5, YR = 2 a YY = 3. K těmto číslům se ještě přičítají hodnoty odvozené na základě počtu vodíkových vazeb mezi komplementárními nukleotidy. Pár GC byl spojen s hodnotou 4 a pár AT s hodnotou 1. Počítají se hodnoty pro každý dinukleotid a vzniká 16 různých kombinací, které jsou vypsané i s hodnotami v tabulce 3.1.

Tab. 3.1: Parametry síly (St) pro jednotlivé dinukleotidy, převzato z [13].

	sousedící báze			
vodíkové můstky	**RY = 5**	**YY = 3**	**RR = 3**	**YR = 2**
4+4	GC = 13	CC = 11	GG = 11	CG = 10
1+4	AC = 10	TC = 8	AG = 8	TG = 7
4+1	GT = 10	CT = 8	GA = 8	CA = 7
1+1	AT = 7	TT = 5	AA = 5	TA = 4

Konečná rovnice je pak ve tvaru:

$$T_m = 7,35 \cdot \frac{St}{l} + 17,34 \cdot ln(l) + 4,96 \cdot ln([Na^+]) + 0,89 \cdot ln(C) - 25,42, \qquad (3.5)$$

kde T_m značí teplotu tání ve stupních Celsia, St je parametr síly vazby a je dán součtem hodnot z tabulky, l popisuje počet nukleotidů v sekvenci, $[Na^+]$ určuje koncentraci sodných iontů a C koncentraci DNA, obě koncentrace jsou udávány v $mol \cdot l^{-1}$. [13]

4 Realizace funkcí pro stanovení teploty tání DNA sekvencí

V programovacím prostředí MATLAB R2018a byly realizovány čtyři funkce pro stanovení teploty tání DNA sekvencí na základě metod popsaných v kapitole 3.2.

4.1 Realizace pomocné funkce *pocty_bazi.m*

Vstupem do pomocné funkce pro výpočty *pocty_bazi.m* je sekvence DNA a výstupem jsou počty bází A, C, G a T, které se ve vstupní sekvenci nachází. Nepočítají se zde pouze symboly A, C, G a T, ale jsou brány v úvahu i nedokonalosti při sekvenaci, kdy se ve vstupní sekvenci může objevit některý ze symbolů W, S, M, K, R, Y, B, D, H, V nebo N. Každý z těchto symbolů reprezentuje více než jednu bázi. Seznam a vysvětlení jednotlivých symbolů je v tabulce A.1 v příloze. Například znak W reprezentuje A nebo T (slabá vazba z angl. weak). Pokud je tento znak ve vstupní sekvenci přítomen, tak se počet bází A navýší o 0,5, stejně jako počet bází T. Tato funkce je volána v rámci funkcí *elementarni.m* a *uprava_soli.m*, které realizují výpočet na základě elementární metody a metody pro úpravu soli.

4.2 Funkce *elementarni.m* a *uprava_soli.m*

Funkce *elementarni.m* realizuje výpočet uvedený rovnicí 3.1. Pro určení počtu bází C a G si volá funkci *pocty_bazi.m*. Vstupem je sekvence a výstupem je teplota tání určená elementární metodou.

Pro výpočet teploty tání metodou úpravy soli slouží naprogramovaná funkce *uprava_soli.m*, jejímž vstupem je sekvence DNA a koncentrace sodných iontů. Teplota je počítána dle rovnice 3.3 a pro určení počtů bází opět využívá funkce *pocty_bazi.m*.

4.3 Funkce *NN.m*

Výpočet teploty tání na základě metody nejbližšího souseda dle rovnice 3.4 je realizována funkcí *NN.m*. Vstupem funkce je sekvence a koncentrace oligonukleotidů a sodných iontů. Pro tento výpočet je potřeba zjistit změnu entropie a entalpie, které jsou pro jednotlivé sousedící dvojice dány tabulkově. Podle tabulky A.2 v příloze byla v MATLABu vytvořena struktura *tab_NN.mat*, která je v rámci funkce načítána a z níž jsou vybírány odpovídající hodnoty změn entropie a entalpie pro jednotlivé nukleotidy na základě jejich sousedící báze. Opět jsou brány v potaz i symboly, které

odpovídají více než jedné bázi. Pro názornost je uveden příklad. Pokud je v sekvenci znak M (značí amino skupinu), který se stejnou pravděpodobností předpokládá bázi A nebo C, a po tomto znaku následuje R (značí purin), tedy možnosti A nebo G, tak tato dvojice MR umožňuje 4 různé kombinace. Konkrétně:

AA (ΔH = -9,1 kcal·mol^{-1}, ΔS = -0,024 kcal·K^{-1}·mol^{-1}),
AG (ΔH = -7,8 kcal·mol^{-1}, ΔS = -0,0208 kcal·K^{-1}·mol^{-1}),
CA (ΔH = -5,8 kcal·mol^{-1}, ΔS = -0,0129 kcal·K^{-1}·mol^{-1})
a CG (ΔH = -11,9 kcal·mol^{-1}, ΔS = -0,0278 kcal·K^{-1}·mol^{-1}).

Funkce všechny tyto možné kombinace vyhledá a jejich změny entropie a entalpie zprůměruje. Do konečného výpočtu jdou sumy všech změn entropií a entalpií pro jednotlivé dvojice.

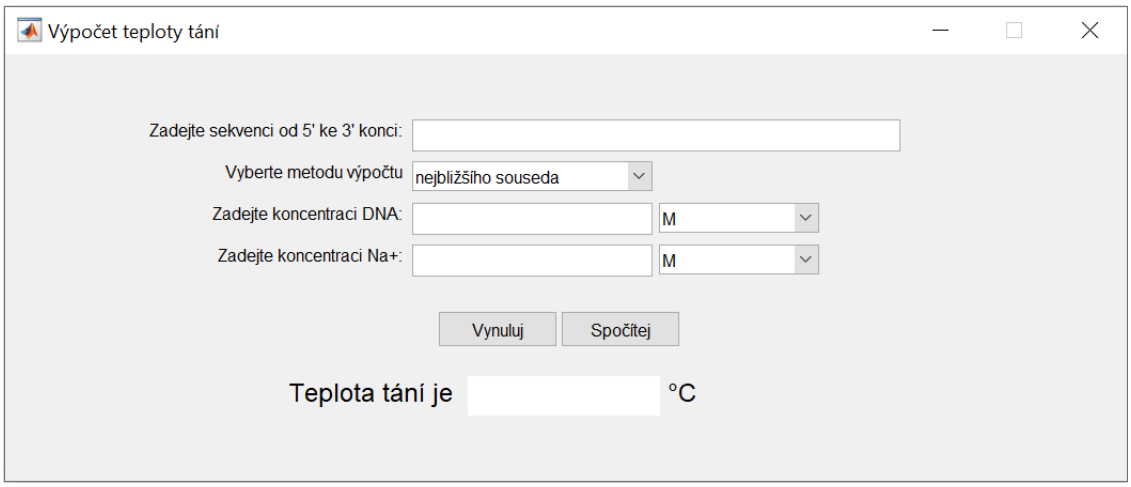
4.4 Funkce *fenomenologicka.m*

Poslední realizovaná funkce *fenomenologicka.m* vypočítává teplotu tání vstupující sekvence na základě fenomenologické metody a rovnice 3.5. Dalšími vstupy do funkce mimo sekvence jsou koncentrace oligonukleotidů a sodných iontů. Pro výpočet je nutné znát parametry síly, které jsou opět dány tabulkově (tabulka 3.1). Tato tabulka je v podobě struktury do funkce načítána (*tab_phen.mat*). Pokud jsou opět v sekvenci přítomny znaky jiné než pro základní báze, jsou dopočítávány průměry parametrů sil všech kombinací.

4.5 Grafické uživatelské prostředí

Pro snadnější a přehlednější přístup k naprogramovaným funkcím byl vytvořen skript *Graficke_prostredi.m*, sloužící jako grafické uživatelské prostředí (GUI z angl. graphical user interface). Vzhled prostředí je ukázán na obrázku 4.1. V prvním řádku je uživatel nabádán, aby zadal sekvenci DNA, pro níž chce vypočítat teplotu tání. Pokud je v sekvenci jiný než povolený znak (znaky z tabulky A.1 v příloze), tak výpočet není realizován a místo toho se pole podbarví červeně a vyskočí varovná tabulka, že byl v sekvenci zadán nepovolený znak. V dalším řádku uživatel pomocí vyskakovacího okna volí metodu, kterou si přeje teplotu tání spočítat. Podle zvolené metody se v dalších dvou řádcích buď objeví nebo zmizí políčka pro zadání koncentrací oligonukleotidů a sodných iontů. Ve vyskakovacím okně může uživatel zvolit v jakých jednotkách koncentraci zadává. Pokud nejsou do políček zadána čísla nebo je zadána nula, pole se podbarví červeně a uživatel je varovnou tabulkou vyzván, aby zadání změnil. Ošetřeno je, i pokud je místo desetinné tečky, kterou MATLAB používá, zadána desetinná čárka. V dalším řádku jsou tlačítka Vynuluj, které vrátí

program do původního, tedy prázdného nastavení, a tlačítko Vypočítej, které, pokud je vše zadáno ve správném formátu, vypočítá výsledek, který se vypíše v posledním řádku.



Obr. 4.1: Grafické uživatelské prostředí sloužící pro výpočet teplot tání zadaných sekvencí různými numerickými metodami.

5 Návrh metodiky klasifikace sekvenčních typů na základě výpočetně stanovené teploty tání

5.1 Metodika shlukování

Shluková analýza je postup, pomocí kterého dochází k rozdělení objektů do skupin na základě jejich vzájemných podobností. V našem případě objekty představují vzorky bakteriálních izolátů. Pro každý vzorek jsou pak určeny teploty tání jednotlivých částí provozních genů, konkrétně jsou to geny infB, mdh, phoE, rpoB, tonB26 a tonB28. Tyto teploty tání představují příznaky pro vzorek, na základě kterých se různými metrikami mezi vzorky určuje podobnost, podle které je pak daný objekt různými metodami zařazen do skupin. Tato seskupení by pak měla odpovídat sekvenčním typům.

5.1.1 Hierarchické shlukování

Hierarchické shlukování bylo realizováno pomocí funkce *linkage* v programovacím prostředí MATLAB R2018a. Jedná se o aglomerační metodu, kdy se vychází z jednotlivých objektů, které se spojují, a nabízí více alternativních řešení, která se vyjadřují pomocí dendrogramů. Výhodou této metody je, že není potřeba znát předem počet skupin a je reprodukovatelná. Jedná se však o výpočetně náročnou metodu, která je pro větší počet dat nepřehledná. Výsledek (dendrogram) bývá často silně závislý na zvolených metodách a metrikách, proto v rámci optimalizace byla vyzkoušena kombinace tří metod shlukování a tří metrik pro výpočet vzdáleností.

Pro výpočet vzdáleností mezi jednotlivými vzorky byla využita euklidova, manhattanská a kosinová vzdálenost. Stanovení vzdáleností mezi nově vzniklým shlukem a ostatními objekty je dáno metodou shlukování. Byly použity metody UPGMA (z angl. unweighted pair group method using arithmetic averages), SLINK (z angl. single linkage clustering method) a CLINK (z angl. complete linkage clustering method). Metoda UPGMA počítá vzdálenosti mezi aritmetickým průměrem objektů tvořících shluk a ostatními objekty. Shlukovací algoritmus SLINK odvozuje vzdálenost z nejbližších objektů. A metoda CLINK vzdálenost odvozuje z nejvzdálenějších objektů. [16]

5.1.2 Nehierarchické shlukování - metoda k-means

Metoda k-means, také občas překládána jako metoda k-středů, rozděluje data do k předem definovaných shluků. Všechny shluky jsou reprezentovány svými geometrickými středy tzv. centroidy, které jsou na začátku shlukování umístěné ve většině algoritmů zcela náhodně, proto se nejedná o reprodukovatelnou metodu.

Během procesu dochází k iterativní optimalizaci. Nejprve jsou spočítány vzdálenosti mezi všemi objekty a centroidy, každý objekt je přiřazen ke skupině, jejíž centroid je pro něj nejbližší. Následně jsou přepočítávány souřadnice centroidů jako průměry ze všech objektů, které byly k dané skupině přiřazeny. Poté dochází znovu k přerozdělení objektů do skupin, podle toho kterému nově vzniklému centroidu mají nejblíž, a znovu přepočítání poloh centroidů. Tento postup je opakován, dokud nedojde k dostatečnému snížení nějaké kriteriální funkce.

Ke shlukování byla využita funkce *kmeans* v programovacím prostředí MATLAB R2018a, která jako kriteriální funkci používá sumu všech distancí mezi objekty a skupinami, do kterých byly zařazeny. Pokud v následující iteraci není tato suma menší než v předešlém cyklu, algoritmus se ukončí s výsledkem z předešlého cyklu. Jelikož tento princip nemusí najít globální minimum, ale pouze lokální minimum kriteriální funkce, bude nastavován parametr funkce 'Replicates' na 10, kdy celé shlukování proběhne desetkrát, pokaždé s náhodně zvolenými prvotními souřadnicemi centroidů. Nakonec bude jako výsledek bráno shlukování s nejmenší sumou distancí. Tento přístup sice neposkytne jistotu nalezení globálního minima, ale alespoň vybere nejmenší z deseti lokálních minim.

Dalším potenciálním problémem je, že nehierarchické shlukování vyžaduje zadání počtu skupin. Však informaci o tom, kolik je mezi našimi vzorky sekvenčních typů, není možné zjistit. Proto bude algoritmus k-means spuštěn pro data vícekrát s různým nastavením počtu skupin. Vyhodnocení výsledků jednotlivých shlukování bude prováděno za pomoci tzv. analýzy siluet, při níž dochází k porovnání vzdáleností ve shluku se vzdálenostmi mezi shluky. Za optimální řešení lze považovat to s maximální hodnotou průměrné siluety, spočítané přes všechny objekty. Silueta nabývá hodnot mezi -1 až 1, kdy hodnoty mezi -1 až 0,25 ukazují na naprosto nevhodný výsledek shlukování, od 0,26 do 0,50 na slabou, nejspíše i náhodnou strukturu, 0,51 až 0,70 je přijatelná struktura a nad 0,71 se jedná o silnou strukturu. Pro analýzu siluet bude využita funkce *silhouette*. Pro optimalizaci byly pro výpočty distančních vzdáleností vyzkoušeny čtverec euklidovské vzdálenosti, manhattanská a kosinová metrika. [17]

5.2 Použitá data

Pro klasifikaci bylo náhodně vybráno 18 izolovaných vzorků bakterie *Klebsiella pneumoniae* poskytnutých Fakultní nemocnicí v Brně, obsahujících 8 různých sekvenčních typů, které byly laboratorně stanoveny. Kompletní data jsou popsána a klasifikována v kapitole 7. Hodnoty teplot tání byly na základě poskytnutých sekvencí spočítány pomocí elementární metody, která je popsaná v kapitole 3.2.1. Teploty tání pro jednotlivé geny a určený sekvenční typ je v tabulce 5.1.

Tab. 5.1: 18 náhodně vybraných izolovaných vzorků bakterie *Klebsiella pneumoniae* poskytnutých FN Brno s laboratorně stanovenými teplotami tání a sekvenčním typem použité pro návrh metodiky shlukování.

vzorek	T_m **pro infB** [°C]	T_m **pro mdh** [°C]	T_m **pro phoE** [°C]	T_m **pro rpoB** [°C]	T_m **pro tonB26** [°C]	T_m **pro tonB28** [°C]	**ST**
KP1272	79,3320	74,5138	76,7795	79,9955	83,0050	85,2633	14
KP1278	79,3320	74,5138	76,7795	79,9955	83,0050	85,2633	14
KP1195	79,3320	74,5138	76,2538	80,3061	83,0050	85,2633	25
KP1204	79,3320	74,5138	77,3051	80,3061	83,0050	84,9217	111
KP1224	79,3320	74,5138	77,3051	80,3061	83,0050	84,9217	111
KP1176	79,3320	74,5138	76,7795	79,9955	83,0050	85,2633	253
KP1226	79,3320	73,8069	76,2538	80,3061	83,4109	84,9217	405
KP1231	79,3320	73,8069	76,2538	80,3061	83,4109	84,9217	405
KP1267	79,3320	73,8069	76,2538	80,3061	83,4109	84,9217	405
KP1193	79,3320	74,5138	76,7795	80,3061	83,4109	84,5800	551
KP1196	79,3320	74,5138	76,7795	80,3061	83,4109	84,5800	551
KP1205	79,3320	74,5138	76,7795	80,3061	83,4109	84,5800	551
KP1237	79,3320	74,5138	76,7795	80,3061	83,4109	84,5800	551
KP1210	78,5120	73,8069	76,7795	79,9955	83,0050	85,2633	628
KP1270	78,5120	73,8069	76,7795	79,9955	83,0050	85,2633	628
KP1241	79,3320	74,5138	76,7795	80,3061	83,0050	85,2633	950
KP1238	79,3320	74,5138	76,7795	80,3061	83,0050	85,2633	950

5.3 Výsledky shlukování

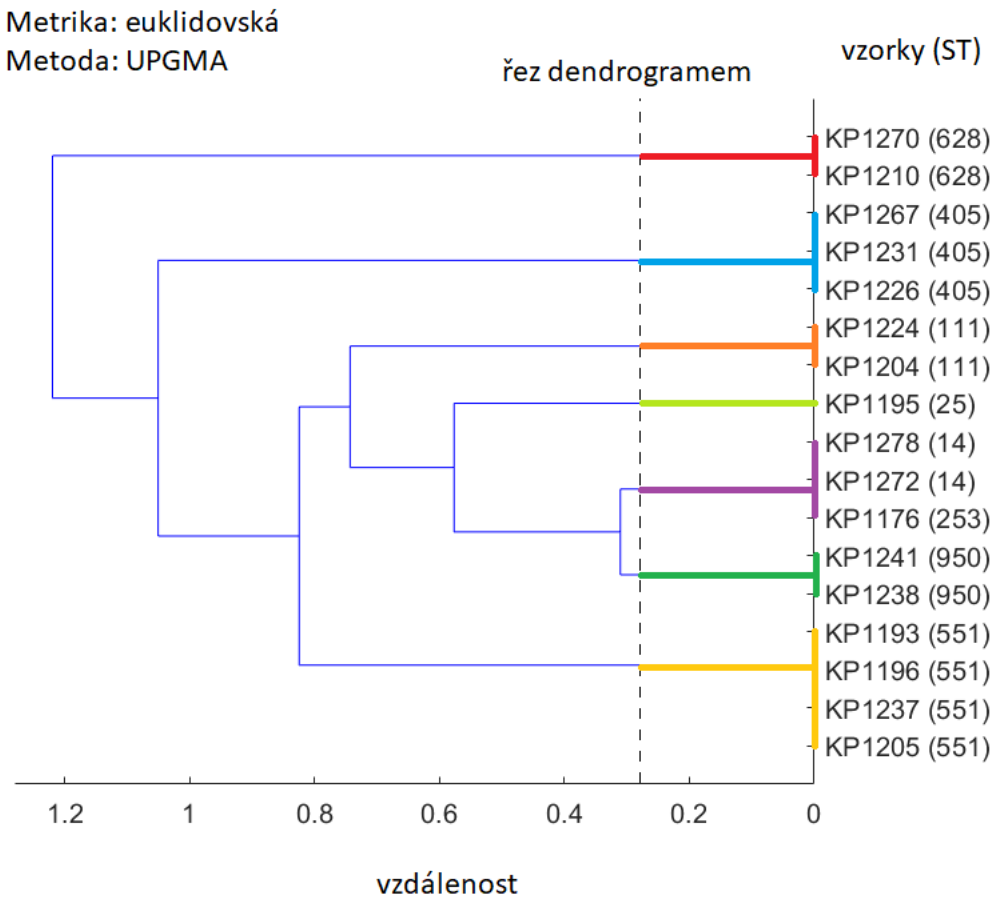
Přestože pro výpočet teplot tání byla použita jednoduchá elementární metoda, bylo shlukování jak hierarchickou, tak i nehierarchickou metodou poměrně úspěšné. Až na jeden případ byly rozpoznány všechny sekvenční typy. Nerozpoznány od sebe byly ST 253 a ST 14. Jejich jediná rozdílná sekvence byla v části genu tonB28, kde byl rozdíl jen v jedné bázi. Šlo o záměnu adeninu a thyminu (A v ST 14 a T v ST 253), tento rozdíl elementární metoda nepodchytí a pro oba sekvenční typy vyšel naprosto totožný soubor teplot tání, takže ani jedna shlukovací metoda nemohla najít rozdíl.

Hierarchická metoda

Protože všechny kombinace použitých metod a metrik byly schopny určit správné sekvenční typy až na jeden případ, byla jako optimální kombinace vybrána ta s nejpřijatelnější topologií, která byla hodnocena pouze opticky. Pokud bude bráno v úvahu, že řez dendrogramem pro oddělení shluků se nejčastěji nastavuje v místě s největší distanční vzdáleností mezi jednotlivými shluky, tak se jako nejvhodnější metrika jeví euklidovská vzdálenost s metodou UPGMA (obrázek 5.1). Proto byla vybrána tato kombinace pro shlukování v rámci všech dat v kapitole 7.

K-means

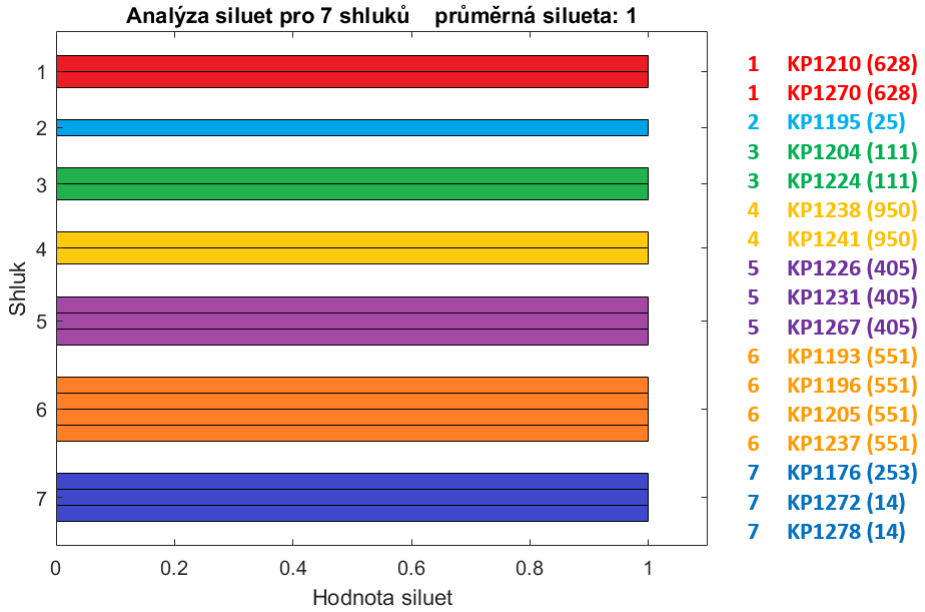
Metoda k-means byla na datech realizována s výpočtem distančních vzdáleností pomocí manhattanské metriky, čtverce euklidovské vzdálenosti a kosinové vzdálenosti. Pro každou z těchto metrik bylo provedeno shlukování do 2 až 12 skupin a pro každou vypočítána průměrná silueta. Bylo však nastaveno, že ve chvíli, kdy průměrná silueta dosáhne hodnoty 1, shlukování do více skupin už neproběhne, protože to znamená, že centroidy se přesně rovnají zařazeným objektům. V rámci všech metrik k tomuto došlo při zařazení do 7 shluků (viz tabulka 5.2). Jako metrika, která bude použita pro následné shlukování všech vzorků, byla vybrána kosinová vzdálenost, protože vykazovala větší rozdíly mezi jednotlivými průměrnými siluetami pro různý počet skupin. Ukázka analýzy siluet a výsledky shlukování pro kosinovou vzdálenost pro 7 shluků je na obrázku 5.2.



Obr. 5.1: Dendogram vytvořený metodou UPGMA a euklidovskou vzdáleností z dat z tabulky 5.1.

Tab. 5.2: Porovnání průměrných siluet při různém počtu shluků (k) pro vybraných 18 vzorků, pro čtverec euklidovské vzdálenosti (A), kosinovou (B) a manhattanskou (C) metriku.

(A)	čtverec euklidovské vzdálenosti	(B)	kosinová vzdálenost	(C)	manhattanská vzdálenost
2	0,558	2	0,495	2	0,456
3	0,596	3	0,604	3	0,609
4	0,762	4	0,766	4	0,757
5	0,900	5	0,905	5	0,887
6	0,946	6	0,950	6	0,863
7	1,000	7	1,000	7	1,000



Obr. 5.2: Analýza siluet a výsledky shlukování pro kosinovou vzdálenost pro 7 shluků.

6 Srovnání laboratorně změřených a výpočetně stanovených teplot tání

6.1 Popis reálných dat

Pro navazující práci byla Fakultní nemocnicí v Brně poskytnuta data změřená pro 52 izolátů bakterie *Klebsiella pneumoniae*. Pro těchto 52 vzorků byly zjištěny a očíslovány alely konkrétních úseků genů gapA, infB, mdh, pgi, phoE, rpoB a tonB nejprve metodou MLST, ze kterých se následně určil sekvenční typ. Totéž bylo provedeno metodou mini-MLST, kdy byly ale zpracovávány ještě kratší úseky genů infB, mdh, phoE, rpoB, tonB26 a tonB28. Nemocnicí byla poskytnuta surová, normalizovaná a diferenční data. Na základě těchto dat lze vykreslit 3 rozdílné křivky závislosti různě upravené fluorescence na teplotě a jsou z nich získávány další informace, které se dále softwarově zpracovávají pomocí převodního klíče, jehož výstupem je MelT (melt typ) obdoba ST (sekvenční typ). Aby bylo možné porovnání teploty tání stanovené laboratorně s numerickým výpočtem, byly poskytnuty i sekvence těchto krátkých analyzovaných genových úseků. Metoda mini-MLST byla prováděna na přístroji Biorad.

6.1.1 Surová data

Jako surová data se zaznamenává fluorescence vzorku v závislosti na teplotě. V rámci tohoto měření nejsou data srovnatelná, protože všechna nezačínají ve stejném místě na ose y (obrázek 6.1 (A)), což je způsobeno různou intenzitou fluorescence na začátku analýzy. Aby mohla být data dále porovnávána, je nutná úprava normalizací.

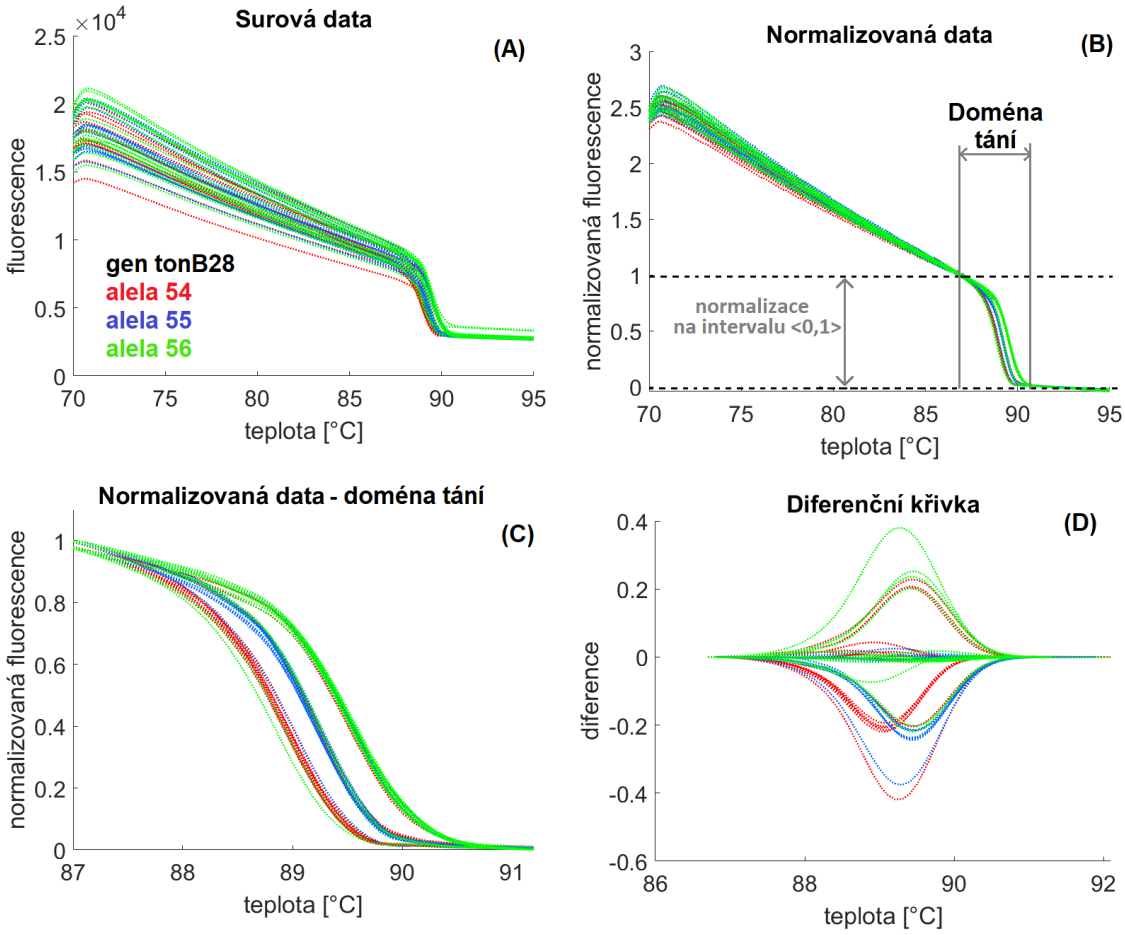
6.1.2 Normalizovaná data

Před samotnou normalizací musí software určit doménu, kde dochází k tání samotného řetězce. Normalizace od 0 k 1 se provádí pouze v rámci tohoto úseku (obrázek 6.1 (B), (C)).

6.1.3 Diferenční data

Tato data jsou používána pro zobrazení a určení odlišností v analyzovaných sekvencích. Oproti normalizované křivce se nehledají posuny na ose x (teplotní), ale posuny na ose y, na které je diference relativních fluorescenčních jednotek. Pro každý teplotní bod je spočítána průměrná referenční hodnota, tím vzniká referenční křivka.

Z referenční a každé analyzované křivky je vypočítána diference, která může být vynesena v podobě diferenčního grafu (obrázek 6.1 (D)), umožňujícího další analýzy sekvencí.

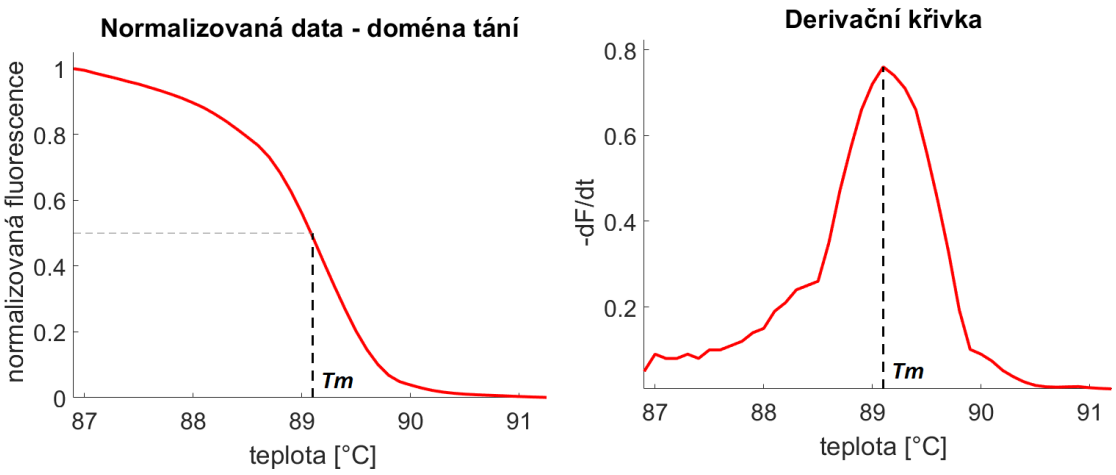


Obr. 6.1: Zobrazení závislostí různě upravené fluorescence na teplotě změřených pro 52 izolátů bakterie *Klebsiella pneumoniae* pro fragment genu tonB28.

6.1.4 Derivační data a určení teplot tání

Jak bylo zmíněno v kapitole 3.1, je pro určení teploty tání vhodná derivační křivka. Hodnoty pro její zobrazení byly vypočteny z normalizovaných dat jako závislost první záporně vzaté derivace fluorescence podle času ($-\Delta F/\Delta t$). Teplota tání je pak určená pozicí vrcholu této křivky.

V programovacím prostředí MATLAB R2018a byly pro všechny vzorky a pro všechny jejich dostupné geny vyhledány teploty tání jakožto pozice maxim v derivační křivce. Ukázka pro vzorek KP1231 a gen tonB28 je na obrázku 6.2.



Obr. 6.2: Určení T_m z derivační křivky pro fragment genu tonB28 vzorku KP1231.

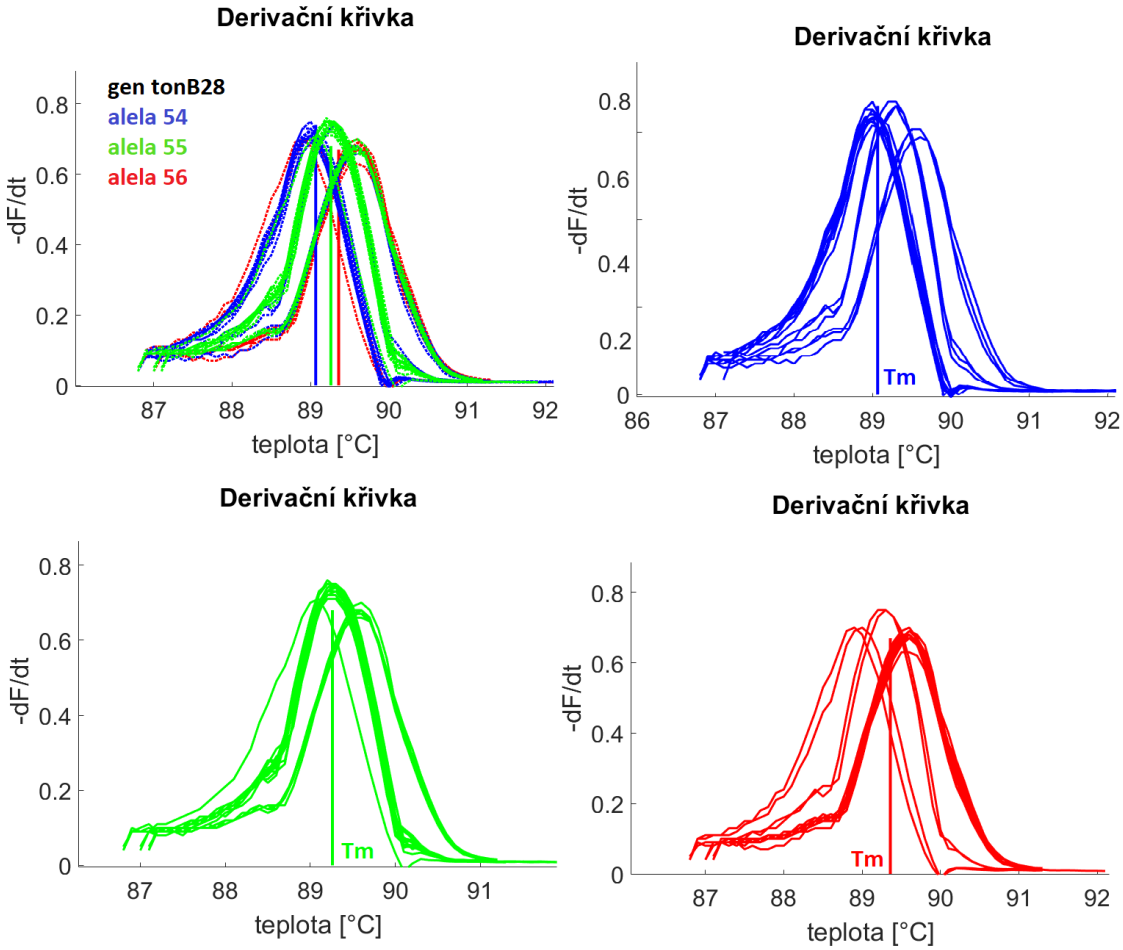
6.2 Úprava dat, získání teplot tání

Numericky určené teploty tání byly vypočítány čtyřmi popsanými metodami v kapitole 3.2 ze sekvencí alel, které byly poskytnuty v rámci reálných dat. Hodnoty koncentrace sodných iontů a DNA, které jsou pro výpočty potřebné, byly poskytnuty firmou Bioline, jejíž chemii (konkrétně mastermix SensiFAST™ HRM Kit) Fakultní nemocnice pro analýzu použila. Koncentrace jednomocných iontů (často nahrazována pouze koncentrací sodných iontů) je 30 mM. Určení přesné koncentrace DNA je složitější, protože PCR neproběhne vždy s naprosto totožným výsledkem, ale pokud jsou dodržovány přesně protokoly, měla by se koncentrace DNA pohybovat okolo 50 nM. Tato koncentrace byla tedy použita i při výpočtech. Laboratorně změřené teploty tání byly určovány z derivační křivky (kapitola 6.1.4).

Pro přehlednost nebyly porovnávány vypočtené a laboratorní (obrázek 6.3) teploty tání napříč všemi vzorky, ale byly porovnávány průměrné teploty tání všech unikátních alel určených metodou mini-MLST. Pro každou unikátní alelu tedy byly vyhledány všechny vzorky, ve kterých se daná alela vyskytovala, a jejich teploty tání, které byly zprůměrovány. Hodnoty průměrných teplot tání pro všechny unikátní alely jsou zaznamenány v tabulce 6.1.

6.3 Výsledky srovnání teplot tání

Aby bylo možné porovnávat laboratorně změřených teplot tání s vypočtenými, byl mezi každou T_m alel určenou numerickou metodou a výsledky z laboratoře spočítán rozdíl. Následně byla vypočítána průměrná diference pro každý gen. Konečné



Obr. 6.3: Určení průměrné T_m alel pro gen tonB28.

porovnání, o kolik se průměrně liší teploty tání určené jednotlivými numerickými metodami od laboratorně změřených, bylo vypočteno jako průměr absolutních hodnot rozdílů všech alel v rámci jedné numerické metody. Tyto výsledky jsou zaznamenány v tabulce 6.2. Celkové průměrné rozdíly teplot pro jednotlivé metody se až na metodu nejbližšího souseda pohybují okolo 4 °C. Metoda nejbližšího souseda vykazuje průměrný rozdíl přes 7 °C. Z tabulky je ale patrné, že diference mezi teplotami není pro všechny geny stejná. Jelikož procentuální zastoupení jednotlivých bází se napříč analyzovanými sekvencemi genů nijak extrémně neliší, důvodem bude pravděpodobně délka těchto fragmentů, která se pohybuje od 50 do 132 bp (viz tabulka 6.3).

Tab. 6.1: Průměrné teploty tání [°C] určené laboratorně a čtyřmi různými numerickými metodami.

gen	alela	laboratorně změřená	elementární	nejbližší soused	úprava soli	fenomenologická
infB	**12**	83,58	78,92	91,51	72,84	76,50
	13	83,87	79,33	91,91	73,25	76,94
mdh	**9**	79,77	73,92	83,33	70,10	72,31
	10	80,08	74,51	84,28	70,70	72,95
phoE	**19**	80,03	76,25	85,88	76,06	76,76
	20	80,17	76,79	86,14	76,60	77,34
	21	80,20	77,31	86,68	77,11	77,89
rpoB	**43**	83,70	80,00	91,11	84,10	86,81
	44	83,55	80,00	91,11	84,10	86,81
	45	83,71	80,31	91,44	84,41	87,14
tonB26	**40**	86,67	83,00	94,88	85,21	86,44
	41	86,89	83,36	95,34	85,56	86,82
tonB28	**54**	89,07	84,67	98,64	88,16	90,69
	55	89,26	84,92	98,83	88,41	90,96
	56	89,36	85,26	99,65	88,75	91,33

Elementární metoda

Průměrný rozdíl mezi teplotami tání změřenými a vypočítanými elementární metodou je 4,18 °C. Teploty vycházely pro všechny geny i alely nižší než změřené (to ukazují záporné hodnoty v tabulce 6.2). Z tabulky lze také vidět, že tato numerická metoda je nezávislá na délce sekvence, protože pouze poměrově počítá zastoupení AT a GC párů.

Metoda nejbližšího souseda

Přestože, jak bylo zmíněno v kapitole 3.2.3, by se mělo jednat o nejpoužívanější a nejpřesnější metodu určování teplot tání oligonukleotidů, vyšla průměrná diference ze všech metod nejvyšší, konkrétně 7,28 °C. Možná byla využita tabulka pro entropii a entalpii odvozená z nevhodného datového souboru. Díky tomu, že vyšly vyšší rozdíly teplot tání pro nejkratší fragment genu infB oproti delším sekvencím genů mdh a phoE, nedá se předpokládat, že chyba je způsobená přílišnou délkou sekvencí.

Tab. 6.2: Diference teplot tání [°C] mezi laboratorně určenými a vypočtenými.

gen	alela	elementární	nejbližší soused	úprava soli	fenomenologická
infB	**12**	-4,66	7,93	-10,74	-7,08
	13	-4,54	8,04	-10,62	-6,92
	průměrná dif. genu	**-4,60**	**7,98**	**-10,68**	**-7,00**
mdh	**9**	-5,86	3,55	-9,67	-7,46
	10	-5,56	4,21	-9,38	-7,12
	průměrná dif. genu	**-5,71**	**3,88**	**-9,53**	**-7,29**
phoE	**19**	-3,78	5,85	-3,97	-3,27
	20	-3,38	5,97	-3,57	-2,83
	21	-2,89	6,48	-3,09	-2,31
	průměrná dif. genu	**-3,35**	**6,10**	**-3,54**	**-2,81**
rpoB	**43**	-3,70	7,41	0,40	3,11
	44	-3,55	7,56	0,56	3,26
	45	-3,40	7,73	0,71	3,43
	průměrná dif. genu	**-3,55**	**7,57**	**0,56**	**3,27**
tonB26	**40**	-3,67	8,21	-1,47	-0,24
	41	-3,53	8,44	-1,33	-0,07
	průměrná dif. genu	**-3,60**	**8,33**	**-1,40**	**-0,16**
tonB28	**54**	-4,40	9,58	-0,91	1,63
	55	-4,34	9,57	-0,85	1,70
	56	-4,10	10,29	-0,61	1,97
	průměrná dif. genu	**-4,28**	**9,81**	**-0,79**	**1,77**
	celková dif. metody	**4,18**	**7,28**	**4,42**	**3,71**

Metoda úprava soli

Metoda úprava soli vykazuje průměrný rozdíl mezi teplotami 4,42 °C. V rámci této metody je patrná závislost diference mezi laboratorní a vypočítanou teplotou tání

Tab. 6.3: Porovnání délek analyzovaných sekvencí genů (fragmentů genů používaných pro mini-MLST).

Gen	infB	mdh	phoE	rpoB	tonB26	tonB28
délka sekvence	50	58	78	132	101	120

na délce řetězce. Čím delší je sekvence, tím menší diference mezi teplotami vyšla.

Fenomenologická metoda

Nejmenší průměrný rozdíl teplot vykazuje fenomenologická metoda, průměrně se teploty všech alel liší o 3,71 °C. Nejvyšší rozdíly mezi teplotami jsou pro krátké geny infB a mdh, kde vychází teploty až o 7 stupňů nižší než u laboratorních. Čím se dostáváme k delším úsekům genů, tím se rozdíl v teplotách zmenšuje až ke genu tonB26, kde je průměrný rozdíl teplot pouhých 0,16 °C. Přibližně u této délky 101 párů bází se zpřesňování metody láme, protože u delších fragmentů tonB28 a rpoB vychází teploty vyšší než laboratorní o 1,77 a 3,27 °C.

Vyhodnocení výsledků

Přestože nejnižší průměrná diference teplot vyšla u fenomenologické metody, pravděpodobně nebude tato metoda pro určování přesných teplot tání nejvhodnější. Důvodem je vysoké rozpětí těchto diferencí pro různě dlouhé geny. Pro geny o kratší délce než 100 bází vychází teploty nižší než laboratorní a to až o 7 °C, a naopak pro fragmenty genů delších než cca 100 párů bází vychází teploty až o 3 °C vyšší. Obdobný problém je i u metody úpravy soli, jejíž celková diference teplot všech genů byla druhá nejnižší. U této metody je také velmi patrná závislost rozdílu teplot na délce analyzovaného genu. Proto pravděpodobně nejlepší výsledky dává elementární metoda, která je už ze svého principu naprosto nezávislá na délce sekvence. Přestože pro metodu nejbližšího souseda vyšla nejvyšší průměrná diference, ani u této metody se neprojevila vyšší závislost rozdílu teploty na délce řetězce. Proto by i výpočty touto metodou mohly být vhodnější.

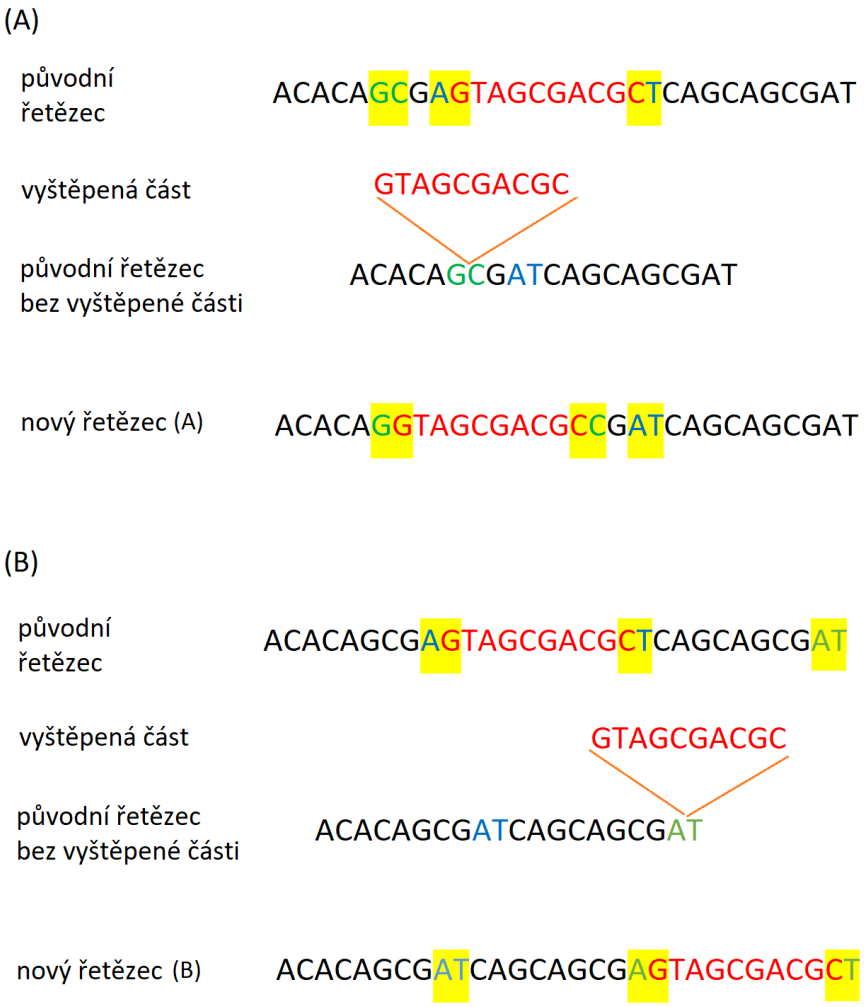
Poměrně velký rozdíl mezi vypočtenými a laboratorními teplotami mohl být způsoben i koncentracemi, které se pro výpočty používaly. Ve vzorcích se obvykle počítá pouze s koncentrací sodných iontů, která má nahrazovat koncentraci všech jednomocných iontů. Pro zjednodušení se koncentrace ostatních iontů opomíjí, ale rovnice mohly být kvůli tomuto opomíjení trochu poupraveny. Od firmy však z důvodů obchodního tajemství nemohla být prozrazená přesná koncentrace iontů sodíku, ale

pouze koncentrace všech jednomocných iontů. I koncentrace DNA nebyla přímo měřena, ale pouze se empiricky předpokládá její přibližná hodnota.

6.4 Vliv různého pořadí stejných nukleotidů na vypočtenou teplotu tání

Různé alely mohou mít stejné zastoupení nukleotidů, které se však budou vyskytovat v jiném pořadí. Během evoluce mohlo totiž například dojít k vyštěpení části řetězce a jeho následnému začlenění do jiné části DNA. Numerické metody jako elementární nebo metoda úpravy soli tyto změny nezohledňují, počítají totiž pouze změny v počtech jednotlivých bází, které jsou ale v těchto případech zachovány. Naopak fenomenologická metoda a metoda nejbližšího souseda tyto změny při výpočtu teplot zohledňují, protože u těchto výpočetních postupů záleží i na pořadí nukleotidů, jak bylo popsáno v kapitole 3.2.

Tyto dvě metody pro každou dvojici sousedících bází počítají určitý parametr. Pro metodu nejbližšího souseda jsou to změny entropie a entalpie a pro fenomenologickou je to parametr síly. Pokud při změně pořadí nukleotidů způsobem popsaným výše dojde alespoň k jedné změně dvojice sousedících bází, tyto dvě metody to zaznamenají na základě změny parametrů, což se většinou projeví i na vypočtené teplotě tání. Z obrázku 6.4 (A) je patrné, že když dojde k vyštěpení části řetězce a jeho opětovnému navázání na jinou pozici, změní se tři sousedící dvojice bází (žlutě zvýrazněno). V tomto konkrétním případě metoda nejbližšího souseda rozdíl zaznamená a výsledná teplota tání se pro obě sekvence liší (viz tabulka 6.4). Teplota tání vypočtená fenomenologickou metodou ale vyšla pro oba řetězce stejná (viz tabulka 6.4), přestože jsou v obou řetězcích rozdílné sousedící dvojice. Důvodem je, že přestože pro dvojice vychází rozdílné parametry sil, tak jejich součet je totožný. Pro původní řetězec GC = 13, AG = 8, CT = 8 a pro nový řetězec GG = 11, CC = 11, AT = 7 (hodnoty jsou převzaty z tabulky 3.1). I z tohoto důvodu se může jevit metoda nejbližšího souseda jako přesnější. Ale jak je patrné z obrázku 6.4 (B), pokud se vyštěpený řetězec začlení mezi totožné dva nukleotidy, mezi kterými se nacházel před vyštěpením (na obrázku se jedná o nukleotidy A a T), tak přestože se změnilo pořadí nukleotidů, mají oba řetězce stejné sousedící dvojice. Tuto záměnu není tedy schopná detekovat ani metoda nejbližšího souseda.



Obr. 6.4: Vznik různých (A) nebo stejných (B) dvojic sousedících bází v závislosti na místě začlenění vyštěpeného řetězce.

Tab. 6.4: Teploty tání [°C] tří různých řetězců o stejném nukleotidovém zastoupení z obrázku 6.4 vypočítané metodou nejbližšího souseda a fenomenologickou metodou.

	metoda výpočtu *Tm* [°C]	
řetězec	**nejbližší soused**	**fenomenologická**
původní řetězec	75,0793	84,3259
nový řetězec (A)	76,0360	84,3259
nový řetězec (B)	75,0793	84,3259

7 Klasifikace bakterií na základě teploty tání variabilních úseků genů

7.1 Klasifikace

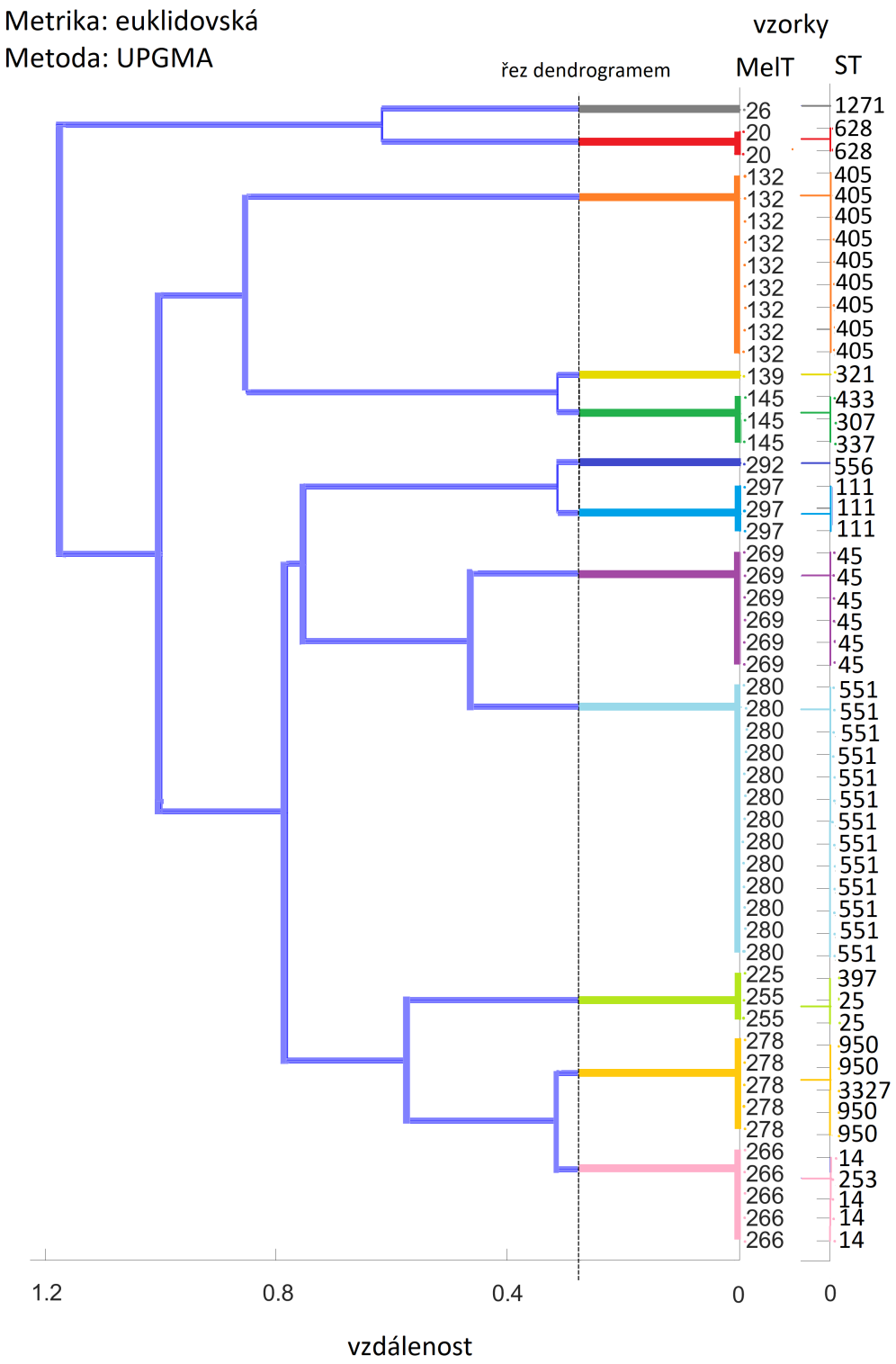
Pro klasifikaci bakterií na základě teplot tání byly vybrány shlukovací metody. Byla použita jak hierarchická metoda za pomoci funkce *linkage* v prostředí MATLAB, tak i nehierarchická metoda realizovaná funkcí *kmeans*. Hierarchické shlukování proběhlo použitím euklidovské vzdálenosti s metodou UPGMA. Pro metodu k-means byla použita kosinová vzdálenost. Podrobnější popis vybraných metod a jejich nastavení je v kapitole 5.

Shlukování bylo realizováno pro všechny vzorky, jejichž teplota byla vypočítána pomocí všech čtyř numerických metod, popsaných v kapitole 3.2. Celkem tedy bylo provedeno osm shlukování. Výsledek všech shlukování až na jeden případ dopadl totožně. U hierarchických metod se pro teploty tání vypočítané různými vzorci vždy nabízel řez dendrogramem vytvářející dvanáct shluků odpovídající MelT číslům. Tato metoda shlukování na základě vypočtených teplot nebyla schopná od sebe odlišit jednotlivé sekvenční typy. Výsledné dendrogramy pro elementární metodu a metodu nejbližšího souseda jsou na obrázcích 7.1 a 7.2. Pro větší přehlednost v dendrogramech již nejsou uvedeny názvy vzorků, ale pouze čísla MelT a ST.

U metody k-means byla nastavena možnost zařazení vzorků do 2 až 52 skupin. Pro každý počet shluků byla počítána průměrná silueta. Shlukování bylo zastaveno ve chvíli, kdy průměrná silueta dosáhla hodnoty 1. Pro shlukování vzorků s vypočtenými teplotami tání pomocí elementární a fenomenologické metody a metody úpravy soli vyšla průměrná silueta 1 při vytvoření 12 skupin odpovídajících, stejně jako v případě dendrogramů, číslu MelT. Výsledek pro elementární metodu je na obrázku 7.3. Jenom v rámci vypočtených teplot metodou nejbližšího souseda vyšla průměrná silueta 1 pro 13 shluků. V tomto případě mimo správného shluknutí všech MelT typů byl rozdělen shluk 3 vzorků odpovídajících stejnému MelT číslu 145. Vzorky s Melt 145 jsou ale všechny označeny jako jiný sekvenční typ (307, 337, 433) a právě jeden z těchto sekvenčních typů (ST 433) se při tomto shlukování oddělil. Výsledek je na obrázku 7.4, rozdělení MelT typů zde odpovídá shlukům 2 a 13.

7.2 Vyhodnocení výsledků klasifikace

Shlukování vzorků na základě vypočtených teplot tání fragmentů sekvencí genů dokázalo odlišit pouze čísla MelT. Ani hierarchická metoda, ani metoda k-means při výpočtech teplot tání různými metodami, nebyla schopná rozřadit a klasifikovat



Obr. 7.1: Výsledek hierarchického shlukování (dendrogram) pro elementární metodu.

bakteriální vzorky podle sekvenčních typů. Jen při výpočtu teplot za použití metody nejbližšího souseda a shlukování k-means byl odlišen jeden sekvenční typ v rámci jednoho stejného čísla MelT. Pro čísla MelT je tedy diskriminační schopnost stoprocentní, ale v rámci sekvenčních typů je výrazně malá.

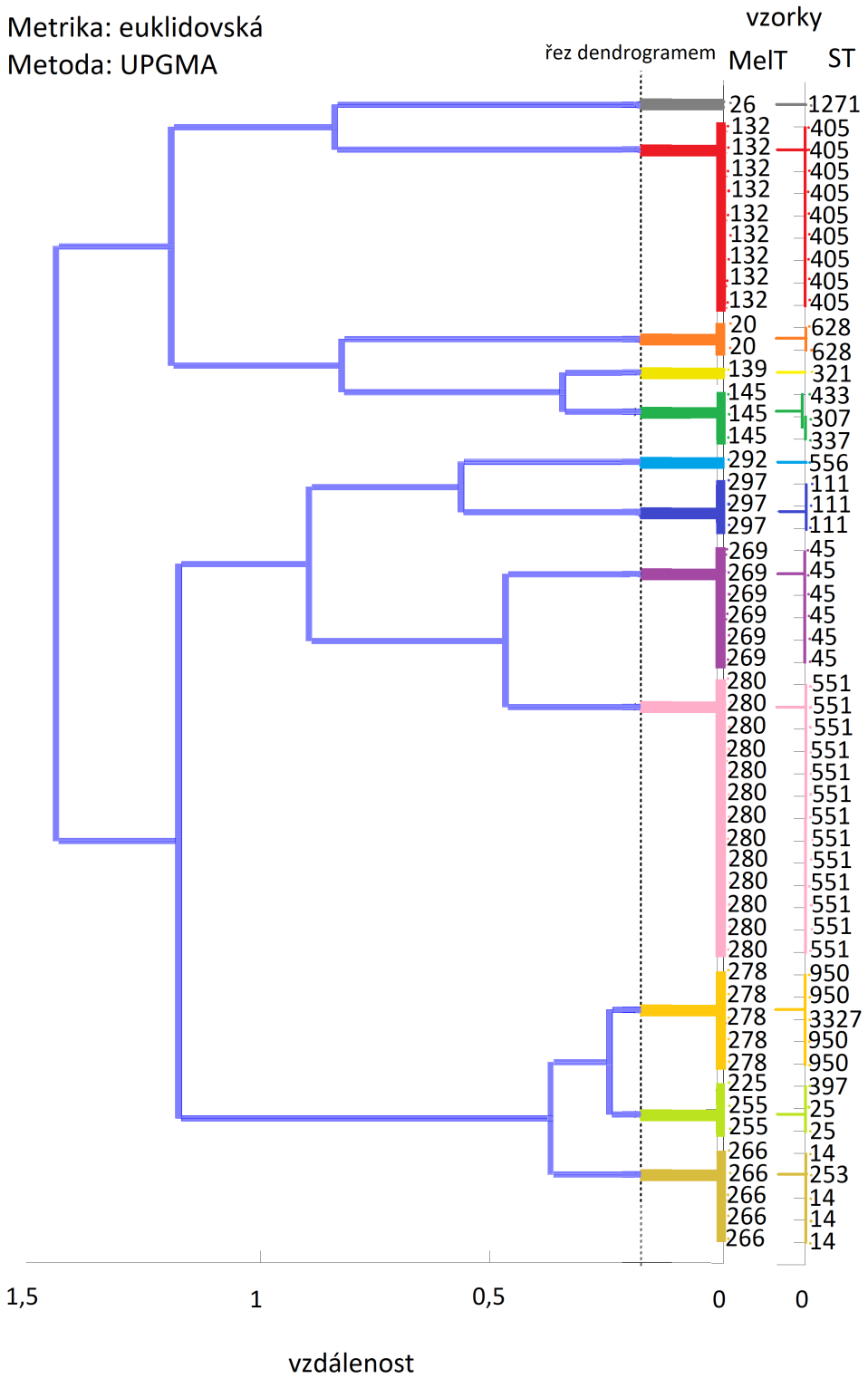
7.3 Návrh optimalizace

Optimalizace metody může být navržena prostřednictvím změn koncentrací. Proto byly v prostředí MATLAB pro různé koncentrace vypočítány podle metody nejbližšího souseda teploty tání vzorků a ty byly opět průměrovány podle alel. Jako nejvhodnější koncentrace se jeví takové, které pro různé alely určí co nejrozdílnější teplotu tání. Proto byl jako ukazatel vhodnosti koncentrací počítán průměrný rozdíl teplot mezi všemi alelami genů. Nejprve, ať byly zadány jakékoliv rozsahy koncentrací DNA, vždy jako nejvhodnější vyšla ta nejmenší, tato hodnota se láme přibližně na $10^{-263}M$, což je samozřejmě naprosto nereálná koncentrace, která by nešla ani analyzovat. Proto se vycházelo ze zadané hodnoty 50 nM a rozsah byl nastaven v rozmezí o dva řády nižší a o dva řády vyšší koncentrace. Stejným způsobem byl nastaven rozsah koncentrace sodných iontů podle hodnoty 30 mM. V tomto rozsahu vyšla jako nejlepší dvojice koncentrací pro DNA 500 pM a pro sodné ionty 2,14 M.

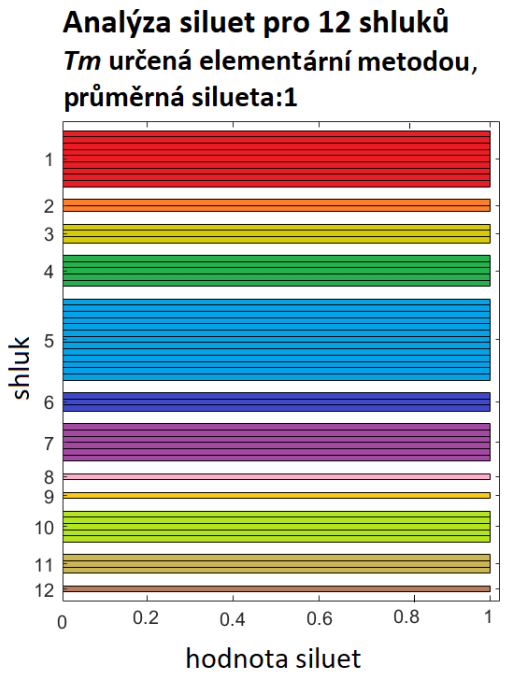
Pomocí těchto nových koncentrací byla znovu vypočítána teplota tání vzorků metodou nejbližšího souseda a znovu vyzkoušeno shlukování pomocí metody k-means. Výsledek shlukování však vyšel totožně jako je na obrázku 7.4.

Z tohoto důvodu lze vyvodit, že metoda shlukování teplot tání úseků genů není pro klasifikaci bakterií podle sekvenčního typu dostatečná. Na základě teplot tání lze ale přesně určit číslo MelT, které, jak bylo zmíněno v kapitole 3.1.2, lze podle převodního klíče převádět na ST.

Optimalizace laboratorního protokolu by však nebyla pro většinu laboratoří možná. Pro metodu minim typing jsou běžně využívány mastermixy dodávané externími firmami, pro které je jejich přesné složení obchodním tajemstvím. Přesná koncentrace nejen sodných, ale i všech ostatních iontů tedy není pro laboratoře známá. Aby laboratoře mohly optimalizovat protokol, musely by si míchat vlastní mastermixy o požadovaných koncentracích. Změněna by však mohla být koncentrace DNA přenastavením PCR reakce.

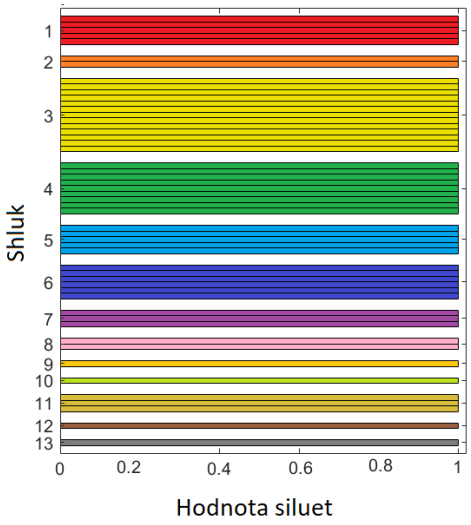


Obr. 7.2: Výsledek hierarchického shlukování (dendrogram) pro metodu nejbližšího souseda.



shluk	MelT	ST	shluk	MelT	ST
1	132	405	5	280	551
1	132	405	5	280	551
1	132	405	5	280	551
1	132	405	5	280	551
1	132	405	5	280	551
1	132	405	5	280	551
1	132	405	6	255	25
1	132	405	6	255	25
1	132	405	6	225	397
2	20	628	7	269	45
2	20	628	7	269	45
3	297	111	7	269	45
3	297	111	7	269	45
3	297	111	7	269	45
4	278	950	7	269	45
4	278	950	8	26	1271
4	278	950	9	139	321
4	278	3327	10	266	253
4	278	950	10	266	14
5	280	551	10	266	14
5	280	551	10	266	14
5	280	551	10	266	14
5	280	551	11	145	337
5	280	551	11	145	433
5	280	551	11	145	307
5	280	551	12	292	556

Obr. 7.3: Výsledek k-means shlukování teplot tání analyzovaných vzorků elementární metodu.



shluk	MelT	ST	shluk	MelT	ST
1	266	253	4	132	405
1	266	14	4	132	405
1	266	14	4	132	405
1	266	14	5	278	950
1	266	14	5	278	950
2	145	337	5	278	950
2	145	307	5	278	3327
3	280	551	5	278	950
3	280	551	6	269	45
3	280	551	6	269	45
3	280	551	6	269	45
3	280	551	6	269	45
3	280	551	6	269	45
3	280	551	6	269	45
3	280	551	7	297	111
3	280	551	7	297	111
3	280	551	7	297	111
3	280	551	8	20	628
3	280	551	8	20	628
3	280	551	9	292	556
4	132	405	10	26	1271
4	132	405	11	255	25
4	132	405	11	255	25
4	132	405	11	225	397
4	132	405	12	139	321
4	132	405	13	145	433

Obr. 7.4: Výsledek k-means shlukování teplot tání analyzovaných vzorků metodu nejbližšího souseda.

Závěr

V první části práce byla popsána struktura DNA a postup, jakým ji lze amplifikovat. V další kapitole je vysvětlen pojem typizace bakterií a jsou zde uvedeny i příklady některých genotypizačních metod. Hlavní část práce se následně zabývá technikami pro stanovení teploty tání laboratorními i výpočetními metodami. V rámci laboratorních technik je popsána vysokorozlišovací analýza křivek tání a mini-multilokusová sekvenční typizace. Výpočetní techniky jsou zastoupeny elementární a fenomenologickou metodou, metodou nejbližšího souseda a úpravou soli.

V programovacím prostředí MATLAB R2018a byly vytvořeny funkce pro výpočty teplot tání popsanými numerickými metodami. Pro jednodušší přístup k funkcím pro ně bylo i navrženo grafické uživatelské prostředí. Tyto funkce byly následně využívány v praktické části.

V rámci praktické části byly zpracovávány reálně naměřené izoláty bakterie *Klebsiella pneumoniae* poskytnuté FN Brno. Byly poskytnuty závislosti naměřené fluorescence na teplotě. Z normalizovaných dat byla vytvořena derivační křivka, z jejíž maxima byla určena teplota tání. Cílem práce bylo porovnat tato laboratorně naměřená data s teoreticky vypočítanými teplotami tání. Nejprve byly porovnávány laboratorní teploty tání s teplotami, které byly vypočítány pomocí všech čtyř popsaných metod. V rámci 52 poskytnutých vzorků nakonec byl nejpodobnější průběh a velikost teplot vypočítán elementární metodou.

V další části byla posuzována diskriminační schopnost metodiky výpočtů teplot tání. Bylo zjišťováno, jestli na základě vypočtených teplot tání je možná klasifikace bakteriálních vzorků. Nejprve byly vyzkoušeny na menším souboru dat různé metody a metriky hierarchického i nehierarchického shlukování. Jako nejvhodnější byla pro hierarchické shlukování vybrána euklidovská vzdálenost a metoda UPGMA a pro metodu k-means, zastupující nehierarchické shlukování, se jevila jako nejvhodnější metrika kosinová vzdálenost. Tyto metody a metriky byly následně použity pro shlukování všech 52 vzorků. Pro všechny metody výpočtů byla stoprocentní schopnost klasifikace izolátů na úrovni melt typů. Na úroveň sekvenčních typů byla ale klasifikace neúspěšná. Jediný sekvenční typ ST 433, který byl při použití všech ostatních shlukovacích metod zařazen spolu s ST 337 a ST 307 do jednoho shluku na základě shodnosti melt typu 145, byl odlišen při shlukování za pomoci k-means při vypočtených teplotách metodou nejbližšího souseda.

Přestože na úrovni melt typů byla schopnost klasifikace izolátů stoprocentní, není tato metoda pro reálné použití klasifikace bakterií vhodná. Pro výpočet teplot jsou totiž zapotřebí sekvence fragmentů genů, jejichž získání je finančně náročné. Na základě určování sekvence alel je založena metoda MLST, která pracuje přímo se sekvencemi, čímž je její diskriminační schopnost ještě vyšší, než když je ze sek-

vence určováno pouze jedno číslo v podobě teploty tání, jejíž hodnota nemusí zaznamenat například jednonukleotidové změny. Hlavní výhodou genotypizačních metod na základě teplot tání je právě absence finančně náročného sekvenování.

Z dostupných zdrojů je poměrně zřejmý potenciál metody minim typing, alespoň v oblasti mikrobiologie. Jedná se o levnou, rychlou a jednoduchou metodu, která se pravděpodobně bude v příštích pár letech ještě rozvíjet.

Literatura

[1] SNUSTAD, D. Peter a Michael J. SIMMONS, RELICHOVÁ, Jiřina, ed. *Genetika.* Druhé, aktualizované vydání. Brno: Masarykova univerzita, 2017. ISBN 978-80-210-8613-5.

[2] GARIBYAN, Lilit a Nidhi AVASHIA. *Polymerase Chain Reaction.* Journal of Investigative Dermatology. 2013, 133(3), 1-4. DOI: 10.1038/jid.2013.1. ISSN 0022202X.

[3] RUSKOVA, Lenka a Vladislav RACLAVSKY. *The Potential of High Resolution Melting Analysis (HRMA) to Streamline, Facilitateand Enrich Routine Diagnostics in Medical Microbiology.* Biomedical Papers. 2011, 155(3), 239-252. DOI: 10.5507/bp.2011.045. ISSN 12138118.

[4] MAIDEN, Martin C.J., Jane A. BYGRAVES, Edward FEIL, et al.*Multilocus sequence typing: a portable approach to the identification of clones within populations of pathogenic microorganisms.* Proc Natl Acad Sci U S A. 1998, 95, 3140–3145.

[5] ANDERSSON, Patiyan, Steven Y. C. TONG, Jan M. BELL, John D. TURNIDGE, Philip M. GIFFARD a Igor MOKROUSOV. *Minim Typing – A Rapid and Low Cost MLST Based Typing Tool for Klebsiella pneumoniae.* PLOS ONE. 2012, 7(3), e33530. DOI: 10.1371/journal.pone.0033530. ISBN 1932-6203.s

[6] RUPPITSCH, Werner. *Molecular typing of bacteria for epidemiological surveillance and outbreak investigation.* Die Bodenkultur: Journal of Land Management, Food and Environment. 2016, 67(4). DOI: 10.1515/boku-2016-0017. ISBN 10.1515/boku-2016-0017. ISSN 0006- 5471.

[7] REED, Gudrun H, Jana O KENT a Carl T WITTWER. *High-resolution DNA melting analysis for simple and efficient molecular diagnostics.* Pharmacogenomics. 2007, 8(6), 597-608. DOI: 10.2217/14622416.8.6.597. ISSN 1462-2416.

[8] TONG, S. Y. C. a P. M. GIFFARD. *Microbiological Applications of High-Resolution Melting Analysis.* Journal of Clinical Microbiology [online]. 2012, 50(11), 3418-3421 [cit. 2020-04-04]. DOI: 10.1128/JCM.01709-12. ISSN 0095-1137.

[9] RUIZ-VILLALBA, Adrián, Elizabeth VAN PELT-VERKUIL, Quinn D GUNST, Jan M RUIJTER a Maurice JB VAN DEN HOFF. *Amplification of*

nonspecific products in quantitative polymerase chain reactions (qPCR). Biomolecular Detection and Quantification [online]. 2017, 14, 7-18 [cit. 2020-04-04]. DOI: 10.1016/j.bdq.2017.10.001. ISSN 22147535.

[10] RICHARDSON, L.J., S.Y.C. TONG, R.J. TOWERS, et al. *Preliminary validation of a novel high-resolution melt-based typing method based on the multilocus sequence typing scheme of Streptococcus pyogenes.* Clinical Microbiology and Infection. 2011, 17(9), 1426-1434. DOI: 10.1111/j.1469-0691.2010.03433.x. ISSN 1198743X.

[11] BRHELOVA, Eva, Iva KOCMANOVA, Zdenek RACIL, Marketa HANSLIANOVA, Mariya ANTONOVA, Jiri MAYER a Martina LENGEROVA. *Validation of Minim typing for fast and accurate discrimination of extended-spectrum, beta-lactamase-producing Klebsiella pneumoniae isolates in tertiary care hospital.* Diagnostic Microbiology and Infectious Disease. 2016, 86(1), 44-49. DOI: 10.1016/j.diagmicrobio.2016.03.010. ISSN 07328893.

[12] PANJKOVICH, A. a F. MELO. *Comparison of different melting temperature calculation methods for short DNA sequences.* Bioinformatics. 2005, 21(6), 711-722. DOI: 10.1093/bioinformatics/bti066. ISSN 1367-4803.

[13] KHANDELWAL, Garima, Jayaram BHYRAVABHOTLA a Sudhindra GADAGKAR. *A Phenomenological Model for Predicting Melting Temperatures of DNA Sequences.* PLoS ONE. 2010, 5(8). DOI: 10.1371/journal.pone.0012433. ISSN 1932-6203.

[14] MARMUR, J. a P. DOTY. *Determination of the base composition of deoxyribonucleic acid from its thermal denaturation temperature.* Journal of Molecular Biology. 1962, 5(1), 109-118. DOI: 10.1016/S0022-2836(62)80066-7. ISSN 00222836.

[15] FREIER, S. M., R. KIERZEK, J. A. JAEGER, N. SUGIMOTO, M. H. CARUTHERS, T. NEILSON a D. H. TURNER. *Improved free-energy parameters for predictions of RNA duplex stability.* Proceedings of the National Academy of Sciences. 1986, 83(24), 9373-9377. DOI: 10.1073/pnas.83.24.9373. ISSN 0027-8424.

[16] RONZHINA, Marina. *Umělá inteligence v medicíně: Shluková analýza-Základní principy, hierarchické metody* [přednáška]. Brno: Fakulta elektrotechniky a komunikačních technologií VUT. 1. 10. 2019

[17] RONZHINA, Marina. *Umělá inteligence v medicíně: Shluková analýza-Nehierarchické metody algoritmus k means* [přednáška]. Brno: Fakulta elektrotechniky a komunikačních technologií VUT. 8. 10. 2019

Seznam symbolů, veličin a zkratek

A	adenin
bp	páru bazí
C	cytosin
CLINK	metoda sdružování odvozená z nejvzdálenějších objektů
DNA	deoxyribonukleová kyselina
dsDNA	dvouřetězcová DNA
G	guanin
HLA	hlavní histokompatibilní systém
HRM	vysokorozlišovací analýza křivek tání
MLST	multilokusová sekvenční typizace
PCR	polymerázová řetězová reakce
R	purinové báze
SLINK	metoda sdružování odvozená z nejbližších objektů
SNP	jednonukleotidový polymorfismus
ssDNA	jednořetězcová DNA
ST	sekvenční typ
T	thymin
$\mathbf{T_m}$	teplota tání
UPGMA	metoda užívající aritmetický průměr
Y	pyrimidinové báze

A Tabulky

Tab. A.1: Nukleotidový kód IUPAC.

Symbol	Význam	Reprezentovaná báze			
A	adenin	A			
C	cytosin		C		
G	guanin			G	
T	thymin				T
W	slabá vazba	A			T
S	silná vazba		C	G	
M	amino skupina	A	C		
K	keto skupina			G	T
R	purin	A		G	
Y	pyrimidin		C		T
B	není adenin		C	G	T
D	není cytosin	A		G	T
H	není guanin	A	C		T
V	není thymin	A	C	G	
N	jakákoliv báze	A	C	G	T

Tab. A.2: Hodnoty změn entalpie (ΔH [$kcal{\cdot}mol^{-1}$]) a entropie (ΔS [$kcal{\cdot}K^{-1}{\cdot}mol^{-1}$]) pro výpočet teploty tání metodou nejbližšího souseda, převzato z [15].

dinukleotid	**ΔH [$kcal{\cdot}mol^{-1}$]**	**ΔS [$kcal{\cdot}K^{-1}{\cdot}mol^{-1}$]**
AA/TT	-9,1	-0,0240
AT/TA	-8,6	-0,0239
TA/AT	-6,0	-0,0169
CA/GT	-5,8	-0,0129
GT/CA	-6,5	-0,0173
CT/GA	-7,8	-0,0208
GA/CT	-5,6	-0,0135
CG/GC	-11,9	-0,0278
GC/CG	-11,1	-0,0267
GG/CC	-11,0	-0,0266