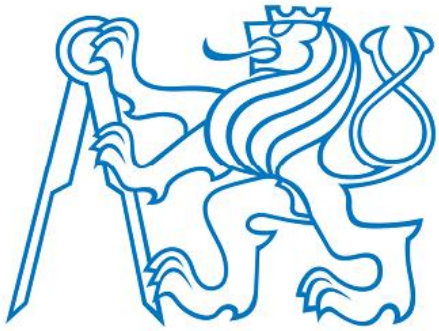
ČESKÉ VYSOKÉ UČENÍ TECHNICKÉ V PRAZE

FAKULTA ELEKTROTECHNICKÁ, KATEDRA KYBERNETIKY



Diplomová práce

HLA GENETICKÁ PŘÍBUZNOST ČECHŮ S OSTATNÍMI NÁRODY

Rok: 2009

Autor: Lukáš Kábrt (lukas@kabrt.cz)

Vedoucí diplomové práce: Doc. Ing. Lhotská Lenka CSc. (lhotska@fel.cvut.cz)

Poděkování

Na tomto místě bych rád poděkoval následujícím registrům dárců krvetvorných buněk za poskytnutí potřebných dat. Český registr dárců krvetvorných buněk, Český národní registr dárců dřeně a Banka pupečníkové krve ČR poskytli data pro českou populaci, Österreichisches Stammzell-Register poskytl data pro rakouskou populaci, Národný register darcov kostnej drene SR pro slovenskou populaci a Against Leukemia Foundation Marrow Donor Registry, Warsaw pro polskou populaci.

Velký dík též patří Mgr. Ing. Davidovi Steinerovi za cenné postřehy a konzultace k této diplomové práci.

Prohlášení

Prohlašuji, že jsem svou diplomovou práci vypracoval samostatně a použil jsem pouze podklady (literaturu, projekty, SW atd.) uvedené v přiloženém seznamu.

V Praze dne 20. května 2009

Lukáš Kábrt

České vysoké učení technické v Praze
Fakulta elektrotechnická

Katedra kybernetiky

ZADÁNÍ DIPLOMOVÉ PRÁCE

Student: Bc. Lukáš K á b r t

Studijní program: Elektrotechnika a informatika (magisterský), strukturovaný

Obor: Kybernetika a měření, blok KM2 – Umělá inteligence

Název tématu: HLA genetická příbuznost Čechů s ostatními národy

Pokyny pro vypracování:

1. Nastudujte existující principy a postupy určování HLA genetické shody dvou jedinců (viz [1]).
2. Nastudujte metody mapování genetické příbuznosti populací (např. fylogenetické stromy).
3. Nastudujte vlastnosti celosvětové databáze dárců krvetvorných buněk [2].
4. Navrhněte a implementujte metodu pro HLA genetické porovnání populací. Tuto metodu aplikujte na porovnání české populace vůči ostatním populacím.
5. Pokuste se vysvětlit Vaše výsledky, např. na základě historických nebo geopolitických souvislostí.

Seznam odborné literatury:

[1] Steiner, D.: Diplomová práce - Hledání nepříbuzenských dárců kostní dřeně. ČVUT v Praze, Fakulta elektrotechnická, katedra počítačů, 2007.
[2] Bone Marrow Donors Worldwide - http://www.bmdw.org/ [online]

Vedoucí diplomové práce: doc. Ing. Lenka Lhotská, CSc.

Platnost zadání: do konce zimního semestru 2009/2010

ČESKÉ VYSOKÉ UČENÍ TECHNICKÉ V PRAZE Fakulta elektrotechnická L.S.

prof. Ing. Vladimír Mařík, CSc.
vedoucí katedry

doc. Ing. Boris Šimák, CSc.
děkan

V Praze dne 3. 9. 2008

Abstrakt

Náplní této diplomové práce je nalezení metody, pomocí které lze na základě znalosti HLA určit genetickou vzdálenost mezi populacemi, tuto metodu implementovat, aplikovat na českou populaci a výsledky publikovat v systému HLA Explorer.

Práci lze rozdělit do třech hlavních částí – teoretická, implementační a experimentální část. V teoretické části jsou rozebrány základy populační genetiky, vlastnosti a význam HLA a podrobeny testování různé metody výpočtu genetické vzdálenosti. V části věnované implementaci popisuji zvolená technická řešení a zdůvodňuji výběr použitých technologií. Experimentální část obsahuje výsledky provedených výpočtů a jejich rozbor z hlediska historických a geopolitických souvislostí.

Důležitou součástí diplomové práce je elektronická příloha s implementací a zdrojovými kódy zvolené metody výpočtu genetické vzdálenosti.

Summary

The goal of this thesis is search for a technique, which can calculate genetic distance between two populations based on the knowledge of their HLA, its implementation and using this technique for calculation genetic distance between Czechs and other nations. The results will be published on the HLA Explorer web page.

This thesis can be divided into three the most important parts - theoretical part, implementation part and experimental part. The basics of population genetics and properties of HLA along with various methods for genetic distance calculation are described in theoretical part. In the Implementation chapter are explained used technologies and algorithms. Experimental part contains results and their analysis from the historical and geopolitical point of view.

Important part of this thesis is electronic attachment with implementation and source code for selected genetic distance calculation method.

Obsah

1. Úvod

Hlavním cílem této diplomové práce je nalezení metody, pomocí které lze na základě znalosti HLA určit míru příbuznosti, tzv. genetickou vzdálenost, mezi populacemi a následně aplikovat tuto metodu na dostupná data hlavně se zaměřením na českou populaci.

Informace o příbuznosti populací mohou být užitečné nejen pro zkoumání různých historických souvislostí či sledování původu a migrace populací, ale i pro klinickou praxi. HLA, pomocí které budeme genetickou vzdálenost počítat, je totiž hlavním parametrem sledovaným při hledání dárců kostní dřeně.

Informace o potenciálních dárcích je často neúplná nebo získaná metodami s nižším rozlišením než je nutné pro provedení transplantace. Proto, pokud máme více vhodných kandidátů, musíme se rozhodnout, u kterých provedeme podrobnější testy. Na základě znalosti genetické vzdálenosti mezi populací pacienta a populacemi dárců můžeme provést kvalifikovaný odhad a určit, u kterého dárce je pravděpodobnost shody vyšší. Pro něj se by se provedly další, podrobnější testy.

Vzhledem k cenám testů HLA, které podle [1] činí v závislosti na použité metodě $80 až $600, může mít výběr správného potenciálního dárce k dalším testům velmi pozitivní ekonomický efekt.

Aby bylo možné výsledky této diplomové práce opravdu využít, tak jsou publikovány v systému HLA Explorer [2], který vznikl na katedře kybernetiky a soustřeďuje výsledky dalších projektů studujících HLA.

1.1. Struktura diplomové práce

Práci lze rozdělit do třech hlavních částí – teoretická, implementační a experimentální část. V teoretické části – v kapitolách 2 až 4 jsou rozebrány základy populační genetiky, vlastnosti a význam HLA a podrobeny testování různé metody výpočtu genetické vzdálenosti. V kapitole 5 věnované implementaci popisuji zvolená technická řešení a zdůvodňuji výběr použitých technologií. Kapitola 6 popisuje získané výsledky a obsahuje jejich rozbor z hlediska historických a geopolitických souvislostí.

2. Populační genetika

Pro plné pochopení této práce je nejprve nutné vysvětlit několik základních pojmů a principů z genetiky a medicíny.

2.1. Základní pojmy z genetiky

Gen

Jednotka dědičné informace. Jedná o úsek nukleové kyseliny se specifickým pořadím nukleotidů, která podmiňuje strukturu a funkci výsledného produktu [3].

Lokus

Místo na určitém chromozomu, kde je daný gen uložen.

Alela

Konkrétní formu genu.

Haplotyp

Skupina alel na jednom chromozomu, které pocházejí od jednoho rodiče a dědí se společně (nedochází mezi nimi k rekombinaci)

Genotyp

Genová výbava jedince, případně genový fond celé populace.

Fenotyp

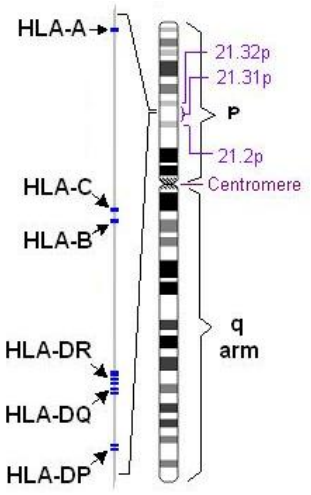
Soubor všech pozorovatelných vlastností a znaků živého organismu. Představuje výsledek spolupůsobení genotypu a prostředí.

2.2. HLA – Human leukocyte antigen

Na šestém chromozomu se nachází úsek zvaný Major histocompatibility comlex (MHC), který hraje velmi důležitou roli v imunitním systému a autoimunitě organizmu. V tomto úseku se vyskytuje na 140 genů, přičemž více než polovina z nich má imunologickou funkci. Geny obsažené v MHC patří mezi nejvíce různorodé v celém genomu obratlovců. [4]

U člověka tyto geny označujeme jako Human Leukocyte Antigen (HLA). V literatuře se někdy označení HLA používá pro výsledné syntetizované proteiny, kdežto pro oblast genomu se používá označení MHC, ale není to vždy pravidlem.

Nejdůležitější jsou geny označované jako HLA-A, HLA-B HLA-C a skupiny genů označované jako HLA-DP, HLA-DQ a HLA-DR. Geny z posledních třech zmíněných skupin se vyskytují na více lokusech a jednotlivé geny se pak značí DPA1, DPB1, DQA1, DQB1, DRA1, DRB1, DRB3, DRB4 a DRB5 [4].



Obr. 1 – Rozložení alel HLA na 6. chromozomu [4]

Geny HLA-E, HLA-F a HLA-G spolu se skupinami genů HLA-DM a HLA-DO se označují jako vedlejší, jsou méně prozkoumané a v běžně se nezjišťují.

MHC je studován především imunology, pro svoji klíčovou roli v imunitním systému. Vzhledem k vysoké diverzitě se ale stal předmětem zájmu i mezi evolučními biology. Ti mohou získat studiem zastoupení jednotlivých genů v různých populacích data pro svoje studie.

Gen	Počet alel
HLA-A	733
HLA-B	1115
HLA-C	392
HLA-E	9
HLA-F	21
HLA-G	42

Gen		Počet alel
HLA-DP	A1	27
	B1	132
HLA-DQ	A1	34
	B1	95
HLA-DR	A1	3
	B1	608
	B3, B4, B5	81
HLA-DM	A1	4
	B1	7
HLA-DO	A1	12
	B1	9

Tab. 1 – Diverzita jednotlivých genů HLA [5]

Snadno můžeme spočítat, že celkem existuje více než 10^{20} možných kombinací genů HLA-A, HLA-B, HLA-C a skupin genů HLA-DP, HLA-DQ a HLA-DR. Možných fenotypů je řádově méně – okolo 10^{12}, stále je to ale mnohem více než je populace celé planety. Neznamená to ale, že bychom na zemi nenašli dva jedince se stejným HLA

fenotypem, naopak. Lidé jsou mezi sebou více příbuzní, než by se mohlo z výše uvedených čísel zdát. Představu o celkovém počtu fenotypů ve světě nám může dát výzkum prováděný ve Francii. [6] Podle něho je ve francouzské populaci o 60 000 000 jedincích 488 000 fenotypů.

2.3. Zjišťování HLA

Pro určení HLA u jedince se využívají speciální vyšetření krve. Vyšetření se provádí buď sérologickou metodou nebo metodami analýzy DNA. [7]

2.3.1. Sérologické metody

Při vyšetření sérologickými metodami se využívá sledování reakce protilátek na membráně buňky. Sérologické vyšetření je méně přesné a nedokáže identifikovat všechny alely HLA. Výhodou této metody je její jednoduchost a nižší nároky na vybavení laboratoře. Pro svoje nižší rozlišení se v současné době od jejího používání upouští.

2.3.2. Analýza DNA

Analýza DNA nám dává přesnější výsledky, protože nesledujeme reakci buňky na protilátky, ale zkoumáme přímo DNA dvoušroubovici. Díky tomu jsme schopni rozlišit všechny alely HLA. Pro analýzu se používají následující postupy [8]:

- PCR-SSP – polymeric chain reaction - sequentially specific primers
- PCR-SSO - polymeric chain reaction - sequentially specific oligonucleotids
- Testování DNA sekvencí

Výsledky získané analýzou DNA umíme ve většině případů převést na výsledky získané sérologickým vyšetřením. Obráceně to možné není. Převodní tabulku je možné najít například v The HLA dictionary 2008. [9]

	A	A	B	B	DRB1	DRB1
DNA metoda	A*0219	A*2402	B*1309	B*1801	DRB1*1117	DRB1*1405
Sérologická metoda	A19	A24	-----	B18	-----	DR14

Tab. 2 – Příklad převodu dat na úrovni DNA na sérologickou úroveň

Na příkladu vidíme srovnání výsledků sérologické metody a DNA metody. Pro alely A*0219 a DRB1*1117 získané vyšetřením DNA není jednoznačně určen žádný ekvivalent pro sérologickou metodu vyšetření. [9]

Alel, pro které není určen sérologický ekvivalent, je minimum [9].

3. HLA v praxi

Transplantace kostní dřeně je výměna kostní dřeně v případě špatně fungující krvetvorby uvnitř kostí za novou krvetvornou tkáň od zdravého dárce. Nejčastější diagnózou, při které je indikována transplantace kostní dřeně, je leukémie.

Metoda transplantace se liší od transplantací jiných orgánů. Krvetvorné buňky získané od dárce se totiž podají pacientovi jako kterákoliv jiná transfúze do krevního řečiště. Takto vpravené buňky do těla pacienta se samy uchytí uvnitř kostí, kde začnou opět růst, množit se a po určité době také obnoví tvorbu zdravých krvinek a celkovou obranyschopnost organismu [10] [11].

Provést transplantaci kostní dřeně není jednoduchá záležitost, tělo se přirozeně brání určitým druhem bílých krvinek proti všemu, co do něho nepatří. Tato schopnost je umožněna proteiny na povrchu buňky syntetizovanými právě podle genů HLA. Tyto bílkoviny se musí při transplantaci co možná nejvíce shodovat, jinak může dojít k tomu, že tělo tuto dřeň vůbec nepřijme. Při menší neshodě se tyto buňky sice uchytí, ale postupně začnou ničit tělo pacienta, protože je vnímají jako cizí.

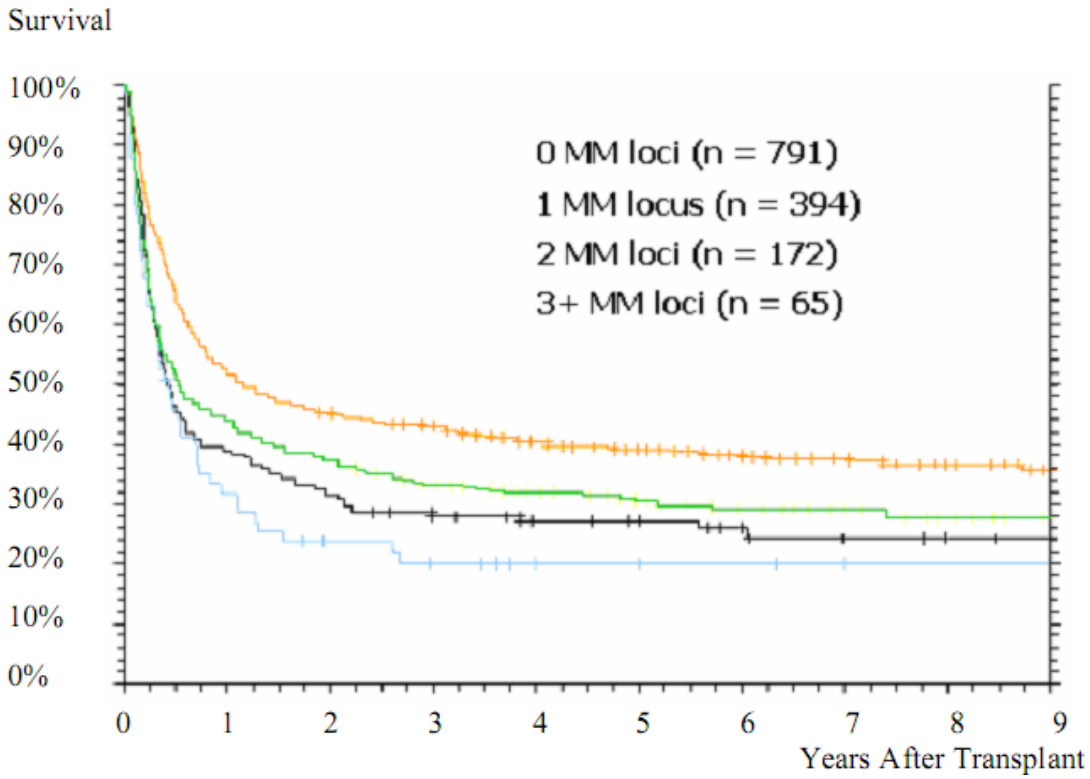
Okolo jedné třetiny pacientů najde příbuzenského dárce – to je ideální stav, protože rizika při transplantaci kostní dřeně od příbuzného dárce jsou nižší. U ostatních pacientů se hledá nepříbuzenský dárce kostní dřeně. Pro tyto účely jsou zřízeny registry, které shromažďují data o HLA potenciálních dárců. V případě, kdy je potřeba kostní dřeň pro pacienta, lékařské zařízení kontaktuje registr dárců kostní dřeně a ten se ve svých záznamech pokusí nalézt vhodného dárce. Pokud se to podaří, registr dárce kontaktuje. Specializované pracoviště provede odběr kostní dřeně u dárce a transplantaci u pacienta. Proces je pro dárce i pacienta anonymní – tzn. pacient neví, kdo je dárce a dárce neví, kdo je pacient.

V České republice existují dva registry – Český registr dárců krvetvorných buněk [10] v Praze a Český národní registr dárců dřeně [12] v Plzni a Banka pupečníkové krve ČR [13]. U nás je evidováno asi 50 000 dobrovolných dárců. Naše země je v přepočtu na počet obyvatel nejlepší ze zemí bývalého východního bloku a na 15. příčce ve světě [14]. Pokud se pro pacienta nenajde dárce v českých registrech, musí se hledat v zahraničí.

3.1. HLA genetická shoda

Proces hledání HLA genetické shody je vzhledem k vyššímu počtu sledovaných znaků a jejich vysoké rozmanitosti mnohem složitější než například hledání vhodného dárce krve.

Bylo provedeno několik různých vědeckých studií, které sledovaly šance na přežití pacienta v závislosti na kvalitě shody jeho HLA s HLA dárce. Výsledky jedné z těchto studií jsou zobrazeny na obr. 2.



Obr. 2 – Závislost šancí na přežití v závislosti na počtu lokusů, ve kterých se dárce neshodoval s pacientem [15]

Tzv. shoda 6/6 znamená HLA shodu mezi pacientem a dárcem na lokusech HLA-A, -B a -DRB1. Shoda 10/10 znamená HLA shodu mezi pacientem a dárcem na lokusech HLA-A, -B, -C, -DRB1 a –DQB1. [15]

Každé transplantační centrum má jiné požadavky na kvalitu shody mezi dárcem a pacientem. Např. některá transplantační centra v USA mají jako požadavek minimální shodu 8/10 (pacient a dárce se mohou lišit až dvou HLA znacích) [16], v České republice se při transplantacích kostní dřeně požaduje shoda 10/10. [15]

Algoritmy hledající potenciální dárce v registrech můžeme rozdělit do několika kategorií [15]:

a) HLA je porovnáváno na sérologické úrovni, data na úrovni DNA jsou ignorována
b) HLA je porovnáváno na sérologické úrovni, data na úrovni DNA jsou převedeny na sérologickou úroveň
c) Pokud jsou jak pro pacienta, tak pro dárce data dostupná na úrovni DNA, provádí se porovnání na úrovni DNA, jinak se převedou na sérologickou úroveň

Vstupem těchto algoritmů jsou data z registrů dárců kostní dřeně a data o pacientově HLA. Výstupem je seznam potenciálních dárců pro pacienta a kvalita HLA shody mezi nimi.

Vzhledem k charakteru výstupu se tyto algoritmy příliš nehodí pro určení genetické vzdálenosti mezi populacemi, proto jsem k jejímu výpočtu použil jiné metody, jak je popsáno v kapitole 4.

3.2. **Bone Marrow Donors Worldwide**

Bone Marrow Donors Worldwide (BMDW) je dobrovolné sdružení registrů dárců kostní dřeně z mnoha zemí světa. V současné době (2009) tato organizace sdružuje 102 registrů a bank pupečníkové krve z 44 zemí světa. V databázi této organizace je ke dni 28. dubna 2009 celkem 13 131 966 dárců [17].

Cílem této organizace je zajistit výměnu informací mezi jednotlivými registry a zajistit tak pacientům lepší šance na nalezení kompatibilního dárce.

Data jsou vyměňována mezi registry a BMDW ve standardizovaném formátu. Ukázka z takového souboru dat je uvedena ve výpisu 1.

```
A1  A2  B1  B2  D1  D2  ID          RB11  RB12 QB11  QB12  NVC TNC
1   2   35  51  13  8   CBB3-00197  08    13   02AB  03XX  60  121
DRB1*1301/1303/1306/1310 DRB1*0801/0802/0803/0805
1   2   51      3   17   CBB7-00201                        65  110
###2
```

Výpis 1- Příklad datového souboru z databáze BMDW [18]

Přesný popis formátu souboru a význam jednotlivých polí je uveden ve specifikaci na webu BMDW [18].

Pro účely této diplomové práce by se jednalo o ideální zdroj dat pro výpočet genetických vzdáleností mezi populacemi. Přístup do tohoto registru ale mají možnost získat pouze akreditovaná transplantační centra. Obrovskou výhodou použití dat z registrů dárců kostní dřeně je vysoký počet dostupných vzorků a kvalita dat. Jedná se o data, která jsou používána registry a transplantačními centry v praxi, na jejich správnosti závisí životy pacientů, tudíž jsou velmi přísně kontrolována.

Určitou nevýhodou může být skutečnost, že struktura dárců nemusí odpovídat struktuře populace. To je zapříčiněno způsobem výběru dárců – jako dárce se preferuje mladý, zdravý muž. Navíc registry se snaží o co nejvyšší rozmanitost dárců – není nutné mít v registru dárce s určitým HLA stokrát, ale je mnohem cennější mít sto dárců s různým HLA. Právě proto je snaha získat do registrů dárce z různých národnostních menšin a minoritních etnik. Jejich zastoupení v registrech je pak rozdílné než zastoupení v populaci.

Dalším faktorem ovlivňujícím složení dárců v registrech je ochota jednotlivých skupin obyvatelstva k darování kostní dřeně. Příkladem může být Jihoafrická republika, která

se svými 43 miliony obyvatel a 13% [19] zastoupením bělochů v populaci má v registru dárců kostní dřeně podíl bělochů vyšší než 50% (David Steiner – osobní konzultace).

3.3. Vědecké studie

Další oblastí, ve které se využívá analýzy HLA jsou různé vědecké studie zaměřené na vztahy populací. Pro svoji vysokou rozmanitost je HLA pro tyto účely velmi vhodná.

Pokud budeme srovnávat data získaná speciálně pro vědecké studie s daty z registrů dárců kostní dřeně, najdeme řadu odlišností. Ty nejpodstatnější jsou shrnuty v následujících bodech.

- malý počet vzorků (řádově stovky)
- jasně a úzce definovaná populace (geograficky, etnicky, apod.)
- HLA v naprosté většině případů získáno analýzou DNA
- všechny vzorky získány stejnou technikou, HLA testováno u všech vzorků na stejných lokusech

Data z různých vědeckých studií jsou shromažďována na webovém portálu www.allelefrequencies.net. V současné době (duben 2009) jsou v této databázi uloženy HLA data o 287 populacích ze 70 zemí světa [20]. Často se jedná o, pro nás exotické, populace a různé kmeny z jižní Ameriky a různých oblastí Asie. Nalezneme zde ale i některé evropské populace.

Pro účely této diplomové práce se ale nejedná o příliš vhodný zdroj dat z následujících důvodů:

- pouze 65 z 287 populací v databázi má uvedeny data pro lokusy A, B a DRB1, které jsou použity pro výpočet genetické vzdálenosti
- u všech populací mimo jedné jsou uvedeny pouze haplotypy s pravděpodobností výskytu vyšší než 1%. V důsledku to znamená, že součet pravděpodobností všech haplotypů u dané populace je nižší než 100%. Hodnoty součtů pravděpodobností se pohybují v rozmezí 75% a 95%. Podle testu metody výpočtu popsané v kapitole 4.1.2 může v takovém případě dojít k podstatnému ovlivnění výsledku.

4. Metody výpočtu genetické vzdálenosti

Člověk je, stejně jako naprostá většina savců, tvor diploidní tzn., že má dvě kopie každého chromozomu – jednu od matky a jednu od otce. Kopii genů, které získáme od každého z rodičů, nazýváme haplotyp. Běžnými vyšetřeními nejsme schopni rozlišit zda, daný gen patří do haplotypu získaného od matky nebo do haplotypu získaného od otce. Nejsme tedy schopni jednoznačně rozdělit geny jedince do 2 haplotypů. Pokud máme k dispozici údaje od větší skupiny jedinců, tak jsme ale schopni s využitím statistických metod odhadnout haplotypy a jejich frekvence v celé pozorované skupině [21].

Haplotypy se dědí celé, nedochází mezi nimi k rekombinaci. To má pro sledování příbuznosti veliký význam – pokud budeme pracovat s celými haplotypy a ne s jednotlivými geny získáme přesnější obraz o příbuznosti mezi populacemi.

V literatuře jsem nalezl několik metod pro výpočet genetické vzdálenosti mezi populacemi. Většina z nich předpokládá použití dat na úrovni genů. Z výše uvedených důvodů jsem se pro výpočet genetické vzdálenosti mezi populacemi rozhodl vycházet z frekvencí haplotypů a ne z frekvencí alel. Výpočet se v tom případě mírně zjednoduší. Vzorce představující tento zjednodušený výpočet jsou uvedeny níže.

Euklidovská vzdálenost [22]

$$D_E = \sqrt{\sum_u (X_u - Y_u)^2} \qquad \textbf{(1)}$$

Cavali-Sforza [23]

$$D_C = \frac{2}{\pi}\sqrt{2 - 2\sum_u \sqrt{X_u Y_u}} \qquad \textbf{(2)}$$

Bhattacharyya [24]

$$D_B = \left(\arccos\sum_u \sqrt{X_u Y_u}\right)^2 \qquad \textbf{(3)}$$

Sanghavi [25]

$$D_S = 2\sum_u \frac{(X_u - Y_u)^2}{X_u + Y_u} \qquad \textbf{(4)}$$

Prevosti [26]

$$D_P = \sum_u \frac{|X_u - Y_u|}{2} \quad \textbf{(5)}$$

Nei [27]

$$D_N = 1 - \sum_u \sqrt{X_u Y_u} \quad \textbf{(6)}$$

Ve vzorcích označuje X_u frekvenci haplotypu u v první populaci Y_u a frekvenci haplotypu u v druhé populaci. Protože X_u a Y_u představují frekvence výskytu haplotypů, tak se přepokládá, že jejich hodnoty jsou z intervalu $\langle 0,1 \rangle$. Na základě tohoto předpokladu můžeme určit obory hodnot pro jednotlivé vzdálenosti.

Metoda výpočtu	minimum	maximum
Euklidovská vzdálenost	0	$\sqrt{2}$
Cavali-Sforza	0	$\frac{2\sqrt{2}}{\pi}$
Bhattacharyya	0	$\left(\frac{\pi}{2}\right)^2$
Sanghavi	0	4
Prevosti	0	1
Nei	0	1

Tab. 3 – Minimální a maximální teoretické vzdálenosti

Rozdíly mezi jednotlivými metodami jsou v tom, že některé z nich berou v potaz hlavní evoluční síly, které hrály roli při vývoji populace. Pokud testované populace daný předpoklad splňují, má vzdálenost určitou biologickou reprezentaci. Například Nei distance předpokládá převládající vliv mutace – pokud je tento předpoklad splněn, pak je vzdálenost přímo úměrná akumulovanému počtu změn na jednom lokusu [27].

4.1. Testování jednotlivých metod výpočtu

Všechny metody výpočtu jsem testoval s použitím umělé databáze, která obsahovala 10 000 000 jedinců s údaji na sérologické úrovni ve třech HLA znacích - A, B a DRB. Tato databáze obsahovala přibližně 10 000 haplotypů [28].

Prováděl následující testy:

- sledování vlivu vzorkování
- sledování vlivu počtu haplotypů použitých pro výpočet vzdálenosti

- sledování vzdáleností při změně frekvencí haplotypů
- sledování vzdáleností při přidaní dalších haplotypů

Cílem prvního testu bylo vyzkoušet chování jednotlivých metod z hlediska daného omezeným množstvím vstupních dat. Cílem druhého testu bylo zjistit, zda je možné snížit časovou náročnost výpočtu vzdálenosti eliminací málo frekventovaných haplotypů. Cílem třetího a čtvrtého testu bylo pozorování chování vývoje vzdáleností při řízené změně vstupních dat.

4.1.1. **Sledování vlivu vzorkování**

V databázích dárců kostní dřeně je ve většině států pouhý zlomek populace. Proto je pro nás důležité zjistit, jak jednotlivé metody aproximují genetickou charakteristiku populace na základě výběru menší skupiny jedinců.

Z celé populace o 10 000 000 jedincích bylo provedeno 100 náhodných výběrů po 100 000 jedincích a pro každý výběr byly odhadnuty frekvence haplotypů. Data s touto strukturou pochází od pracovní skupiny WMDA - Haplotype Tools Group [28]. S těmito daty jsem provedl výpočet vzdáleností mezi všemi náhodnými výběry navzájem. Pokud by metoda výpočtu byla dokonale odolná proti chybě vzorkování, tak by genetická vzdálenost mezi všemi výběry navzájem měla být nulová.

Pro každou ze sledovaných vzdáleností bylo nutno provést $\binom{100}{2} + 100 = 5050$ výpočtů. Výsledky jsem poté statisticky zpracoval a vyhodnotil střední hodnotu vzdálenosti $\overline{D}$ (7), její směrodatnou odchylku σ a relativní směrodatnou odchylku $\sigma/\overline{D}$.

$$\overline{D} = \frac{\sum_{i=1}^{100} \sum_{j=i}^{100} D_{ij}}{\binom{100}{2} + 100} \qquad \textbf{(7)}$$

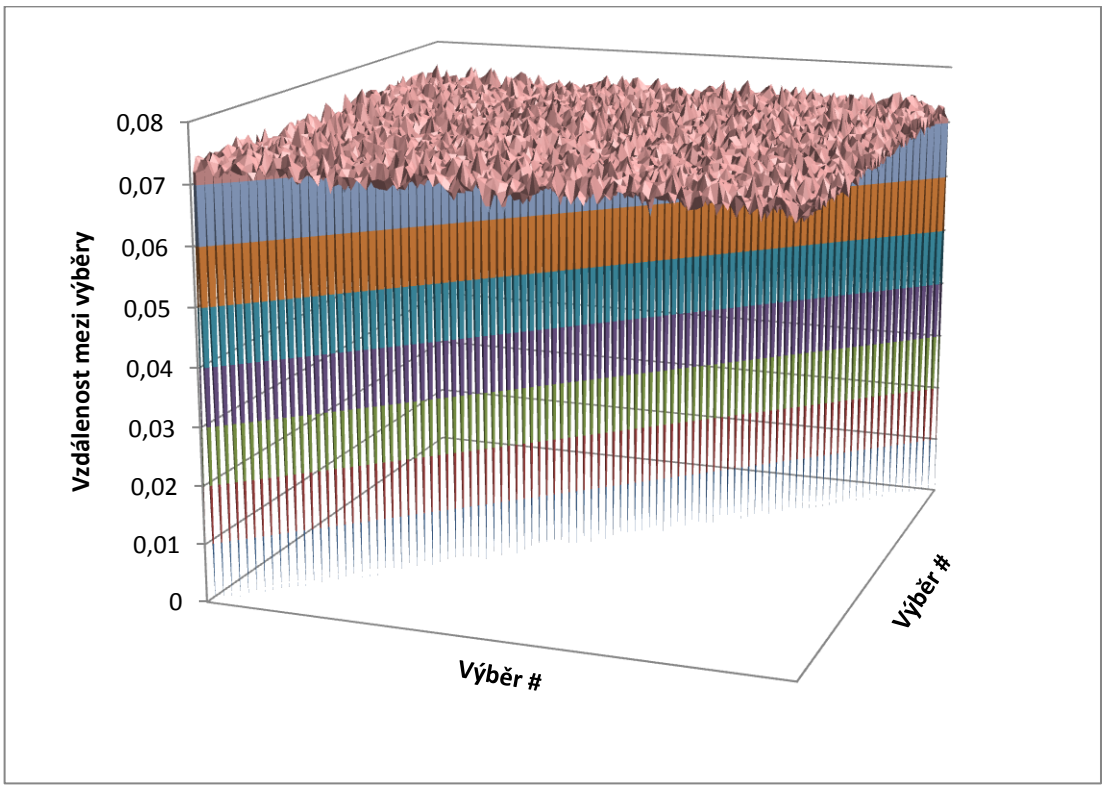
Výsledky pro všechny typy vzdáleností jsou uvedeny v Tab. 4.

Metoda výpočtu	$\overline{D}$	σ	$\sigma/\overline{D}$
Bhattacharyya	0,029931	0,000652	0,021786
Cavali-Sforze	0,109994	0,001196	0,010869
Euclidean	0,004113	0,000160	0,038937
Nei	0,014928	0,000320	0,021731
Prevosti	0,072934	0,001232	0,016896
Sanghavi	0,085469	0,001657	0,019388

Tab. 4 – Porovnání výsledků jednotlivých metod výpočtu

Jednotlivé metody samozřejmě ideální nejsou a vzdálenost mezi různými výběry vyšla nenulová. Pro ilustraci uvádím na obr. 3 graf, který znázorňuje spočítané hodnoty vzdáleností mezi jednotlivými výběry pro Prevosti distance.

Z grafu je jasně vidět, že při porovnávání stejných výběrů (body na diagonále grafu) je genetická vzdálenost mezi výběry nulová. Na výpočtech genetické vzdálenosti mezi ostatními výběry vidíme chybu vzniklou vzorkováním.



Obr. 3 – Vzdálenosti mezi jednotlivými výběry – Prevosti distance

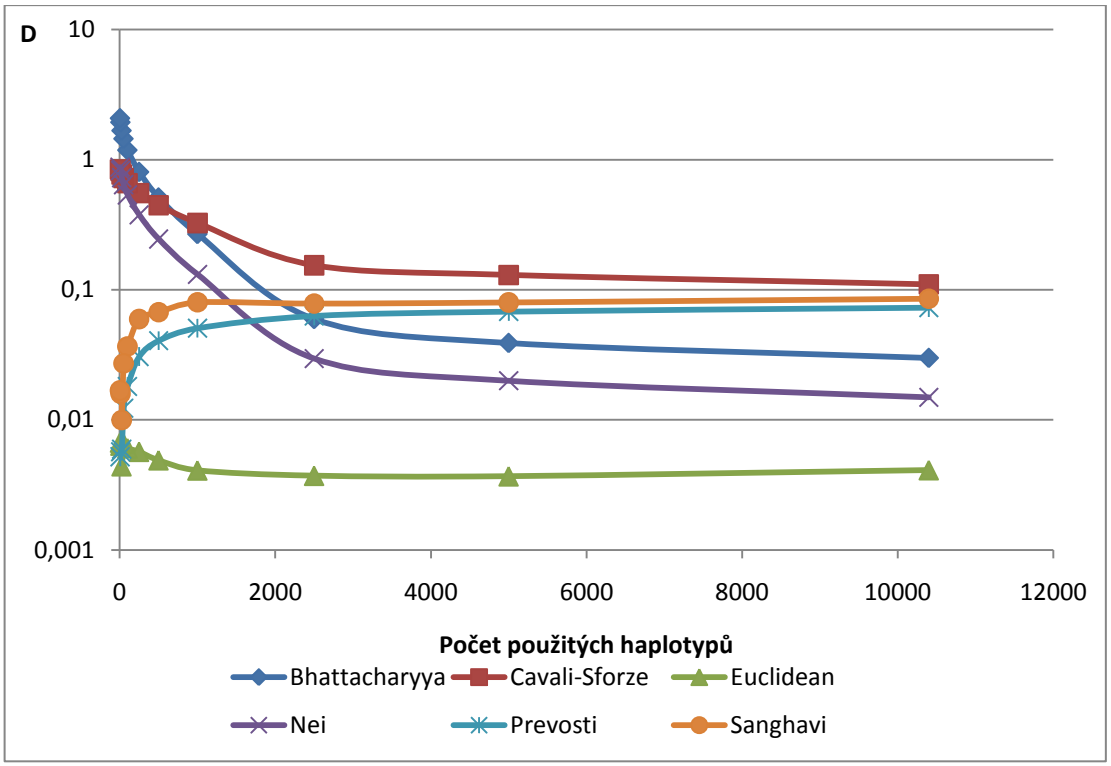
Jak je z výsledků vidět, tak střední hodnota vzdáleností mezi výběry je nejnižší u Euclidean distance, nejvyšší u Cavali-Sforze distance. Je zajímavé, že nejnižší relativní odchylku má naopak Cavali-Sforze distance a nejvyšší Euclidean distance.

4.1.2. **Sledování vlivu počtu haplotypů použitých pro výpočet**

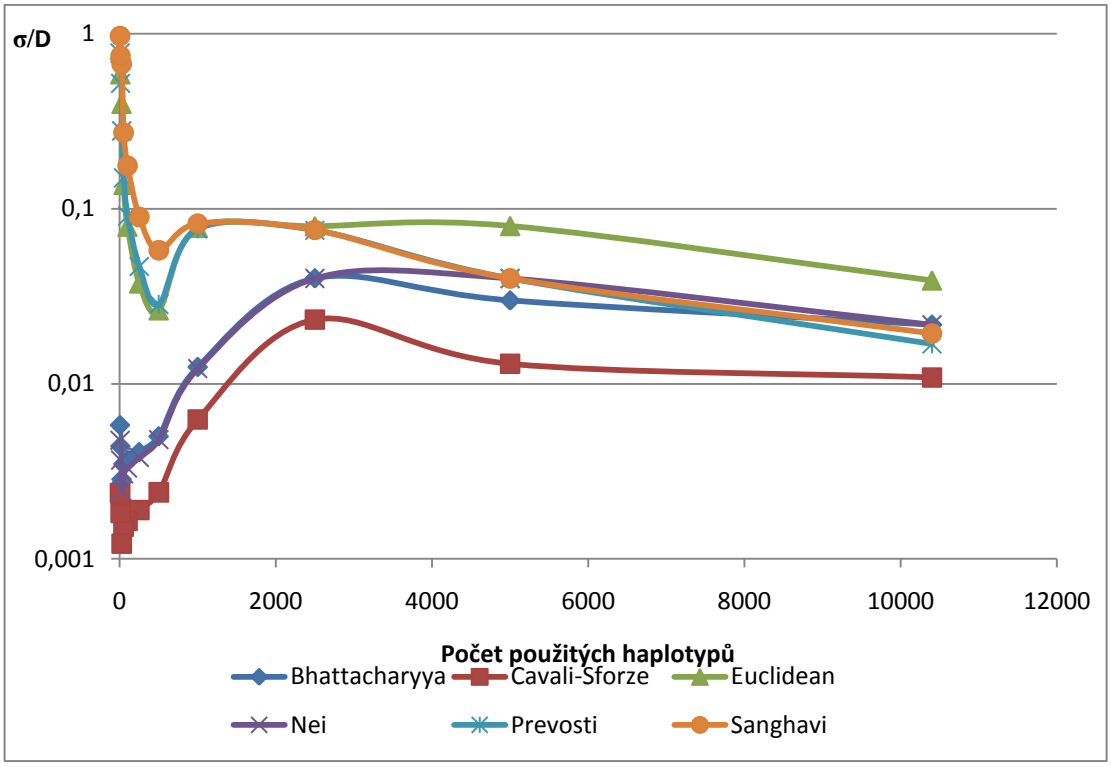
V tomto pokusu jsem se zaměřil na to, zda by nešlo výpočet vzdálenosti urychlit použitím pouze několika haplotypů s nejvyšší frekvencí výskytu.

Pro výpočet vzdáleností a vyhodnocení výsledků byla použita stejná metodika jako v předchozím případě. Pro výpočet vzdáleností mezi jednotlivými výběry bylo vždy použito jen N nejčastějších haplotypů z každého výběru.

Výsledky přehledně znázorňují grafy na obrazcích 4 a 5.



Obr. 4 – Střední hodnota vzdálenost mezi výběry z populace v závislosti na počtu použitých haplotypů



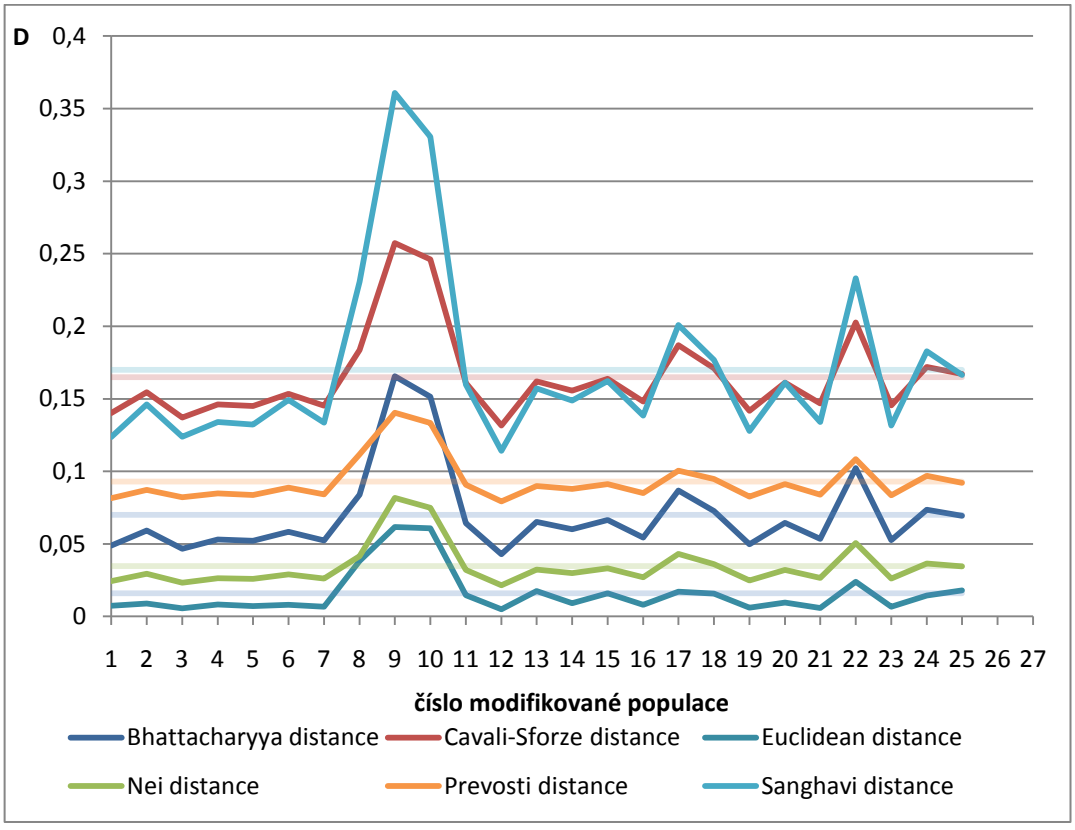
Obr. 5 – Relativní směrodatné odchylky vzdáleností mezi výběry z populace v závislosti na počtu použitých haplotypů

4.1.3. **Sledování vzdáleností při změně frekvencí haplotypů**

Pro skutečné měření vzdáleností mezi populacemi je důležité, jak se budou jednotlivé metody chovat, pokud populace shodné nebudou. V tomto testu jsem simuloval situaci, kdy populace obsahují stejné haplotypy, ale s různými frekvencemi.

Data z původního testovacího souboru byla modifikována tak, že frekvence haplotypů byly náhodně navzájem zaměňovány. Těchto záměn bylo provedeno *N*, pak byla počítána vzdálenost mezi původní a modifikovanou populací.

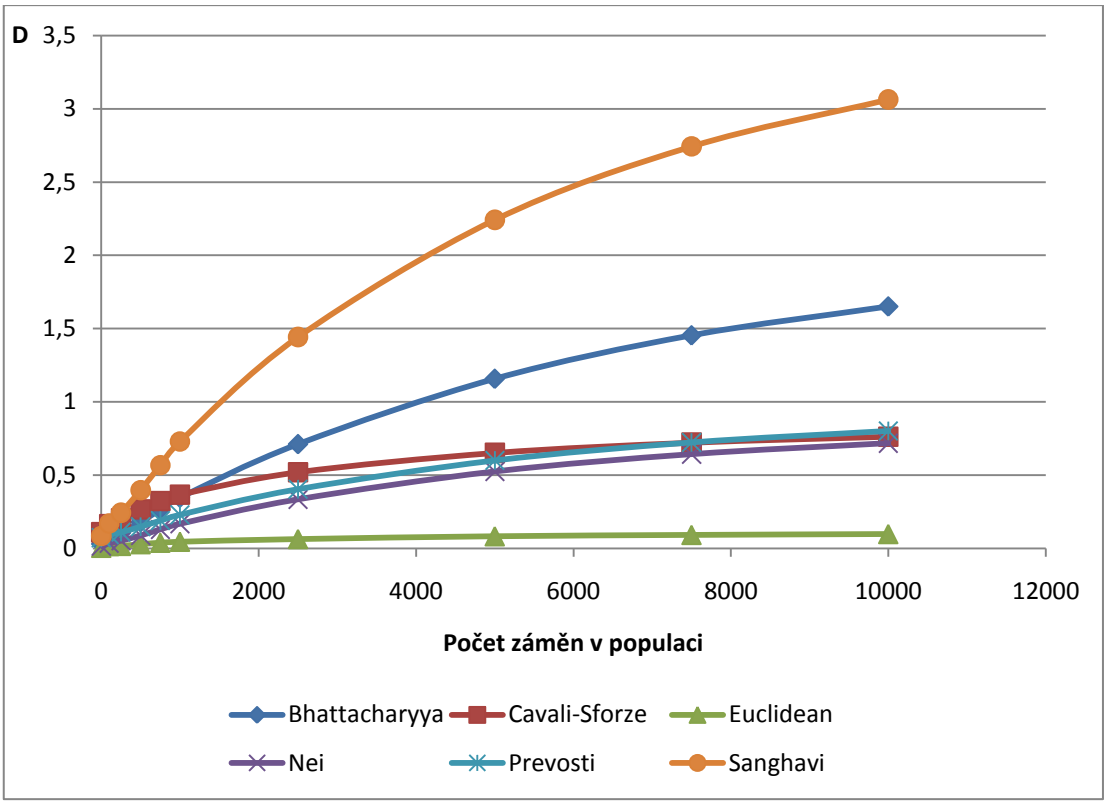
Pro každé N bylo vytvořeno 25 modifikovaných populací a výsledky poté statisticky zpracovány. K tomuto kroku jsem se rozhodl proto, aby se zprůměrovaly odchylky způsobené tím, že někdy docházelo k záměnám u haplotypů s vyššími frekvencemi a někdy u haplotyp s nižšími frekvencemi. Tento stav ilustrují např. vzorky 9 a 10 na obrázku 6. Díky tomu je možné použít počet záměn jako určité relativní měřítko mezi původní a modifikovanou populací.



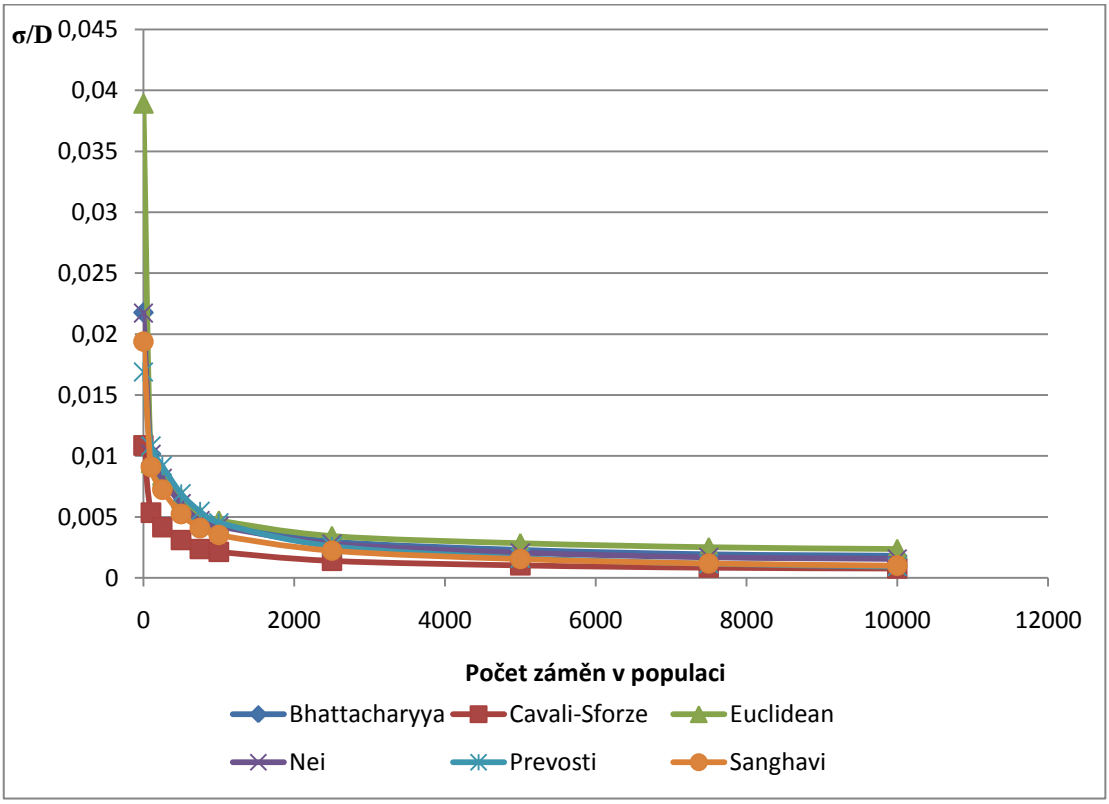
Obr. 6 – Vzdálenosti mezi původní a modifikovanou populací – 100 záměn

Graf na obrázku 2 zobrazuje vzdálenosti mezi původní a 25 modifikovanými populacemi při provedení 100 záměn mezi haplotypy. Na grafu je navíc světlou barvou znázorněna střední hodnota získaná zprůměrováním 25 měření.

Stejná měření jsem provedl pro 250, 500, 750, 1000, 2500, 5000, 7500 a 10000 záměn mezi haplotypy. Výsledky jsou zobrazeny na obrázcích 7 a 8.



Obr. 7 – Střední hodnoty vzdáleností mezi původní a modifikovanými populacemi



Obr. 8 – Relativní směrodatná odchylka mezi původní a modifikovanými populacemi

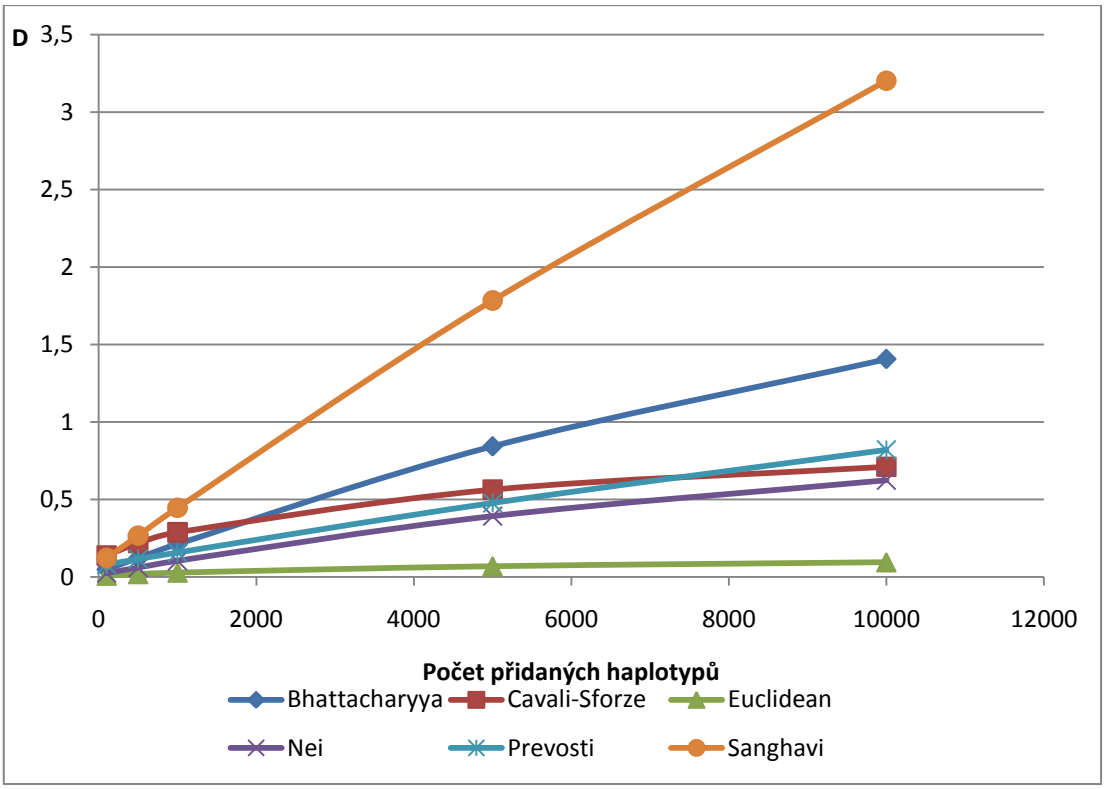
4.1.4. **Sledování vzdáleností při přidání dalších haplotypů**

Jedná se o podobný test jako v předchozím případě, pouze způsob generování modifikovaných dat z původní populace jsem zvolil jiný. Z původní populace byly náhodně vybírány haplotypy a jejich frekvence přiřazeny novým haplotypům v modifikované populaci. Přehledně tento postup znázorňuje tabulka 5.

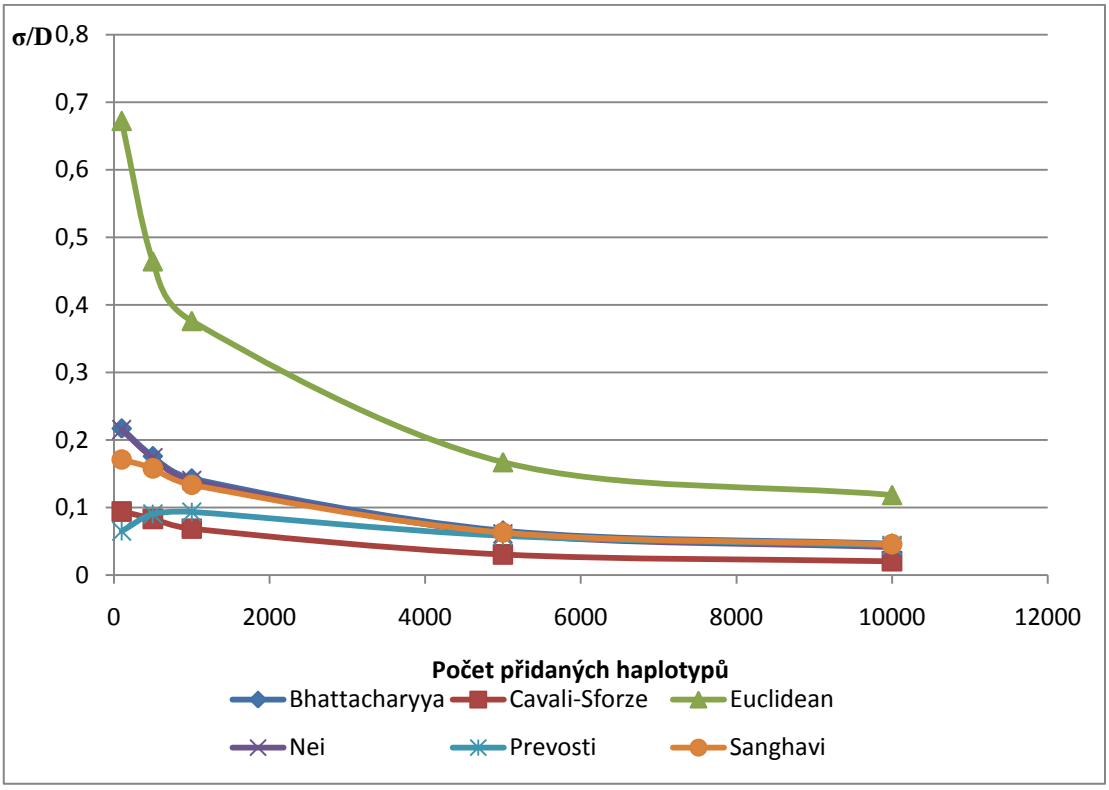
	Frekvence v původní populaci	Frekvence v modifikované populaci
Haplotyp 1	A	A
Haplotyp 2	B	0
Haplotyp 3	C	C
Haplotyp 4	D	0
…		
Haplotyp M	E	E
Haplotyp M + 1	0	B
Haplotyp M + N	0	D

Tab. 5 – Postup generování modifikované populace

Stejně jako v předchozím případě jsem pro každých *N* provedených úprav provedl 25 měření. Provedl jsem výpočty pro 100, 500, 1000, 5000 a 10000 záměn. Výsledky jsou zobrazeny na obrázcích 7 a 8.



Obr. 9 – Střední hodnoty vzdáleností mezi původní a modifikovanými populacemi



Obr. 10 – Relativní směrodatná odchylka mezi původní a modifikovanými populacemi

4.1.5. **Vyhodnocení výsledků testů**

První test měl prokázat odolnost jednotlivých metod proti chybě vzorkování. Upřednostnil bych ty metody, které vykazovaly menší směrodatné odchylky od střední hodnoty. Podle mého názoru je důležitější, že známe výsledek s malou průměrnou odchylkou od střední hodnoty, než to, že vzdálenost mezi dvěmi různými výběry ze stejné populace by měla být nulová. Při interpretaci výsledků totiž lépe zohledníme fakt, že vzdálenost mezi dvěma různými výběry je nenulová, než skutečnost, že námi vypočítaná hodnota vzdálenosti může mít vysokou chybu.

Druhý test prokázal, že je možné snížit náročnost výpočtu použitím pouze nejčastějších haplotypů. Bez obav můžeme vypustit haplotypy s frekvencí výskytu nižší než 10^{-6}.

I u dalších dvou testů bych preferoval ty metody, které vykazovaly menší směrodatné odchylky od střední hodnoty. V těchto dvou případech nižší směrodatná odchylka, především u populací s nižším počtem modifikací, ukazuje na schopnost lépe zachytit celkovou strukturu populace. Vysoká směrodatná odchylka naopak ukazuje, že vysoký rozdíl ve frekvencích výskytu u malého počtu haplotypů může výrazným způsobem ovlivnit výslednou vzdálenost mezi populacemi.

Ve všech případech jsem sledoval relativní směrodatné odchylky $\sigma/\overline{D}$, protože vzdálenosti nejsou normovány a absolutní chyba např. 0,1 u jedné metody, může představovat celou hodnotu vzdálenosti u jiné metody. Relativní směrodatná odchylka tedy dobře reprezentuje odstup chyby od hodnoty vzdálenosti.

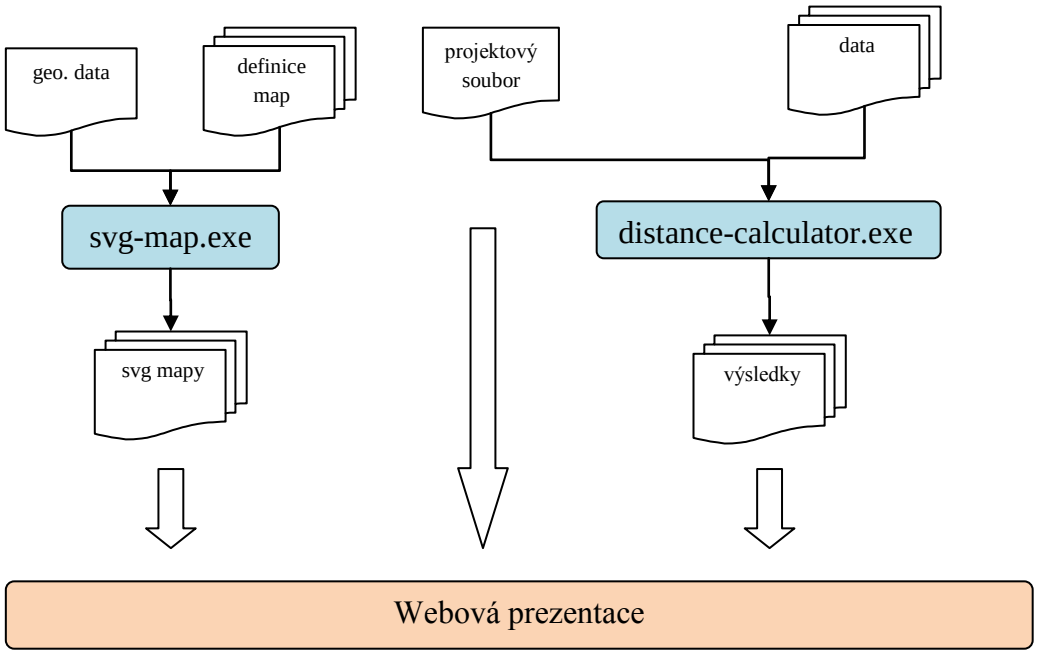
Na základě dosažených výsledků se jako nejlepší jeví metoda Cavali-Sforze.

5. Implementace

Na základě provedených testů jsem vybral metodu výpočtu Cavali-Sforze a přistoupil jsem k implementaci.

Celý proces výpočtu genetických vzdáleností mezi populacemi a prezentace výsledků lze rozdělit do několika logických celků – tvorba podkladů pro prezentaci, samotný výpočet a prezentace výsledků. Implementace respektuje rozdělní do výše uvedených logických celků.

Celkovou strukturu systému a jeho jednotlivé celky znázorňuje diagram na obrázku 11. Programy svg-map.exe a distance-calculator.exe jsou implementovány na platformě .NET v jazyce C# a spouštějí se na klientském počítači. Webová prezentace je implementována na platformě Apache, MySQL, PHP a běží na webovém serveru.



Obr. 11 – Struktura implementace projektu

Všechny části jsem se snažil navrhnout tak, aby byla do budoucna umožněna rozšiřitelnost systému a aby obsluha byla komfortní jak pro uživatele, tak pro jeho administrátory.

5.1. Výpočetní program

Výpočetní program je implementován jako konzolová aplikace v jazyce C#, který jsem zvolil jako kompromis mezi možností optimalizace na rychlost a náročností implementace.

Program na svém vstupu přijímá projektový soubor ve formátu XML a soubory s frekvencemi haplotypů v jednotlivých populacích ve formátu CSV. Výstupem programu je CSV soubor obsahující genetické vzdálenosti mezi všemi populacemi navzájem a soubor ve formátu PNG s fylogenetickým stromem. Popis struktury projektového souboru je uveden v příloze A, formát souboru s frekvencemi haplotypů v příloze B a formát výstupního CSV souboru v příloze C.

Program vyžaduje následující parametry z příkazové řádky.

```
genetic-distance.exe [project] [output-directory] [algorithm]
```

`[project]`	adresa projektového souboru
`[output-directory]`	cesta, kam se mají uložit výstupní soubory
`[algorithm]`	metoda použitá pro výpočet genetické vzdálenosti

V programu jsou implementovány následující metody výpočtu genetické vzdálenosti:

Název metody	Parametr příkazové řádky
Euklidovská vzdálenost	euclidean
Cavali-Sforze	cavalisforze
Bhattacharyya	bhattacharyya
Sanghavi	sanghavi
Prevosti	prevosti
Nei	nei

Tab. 6 – Implementované metody výpočtu genetické vzdálenosti

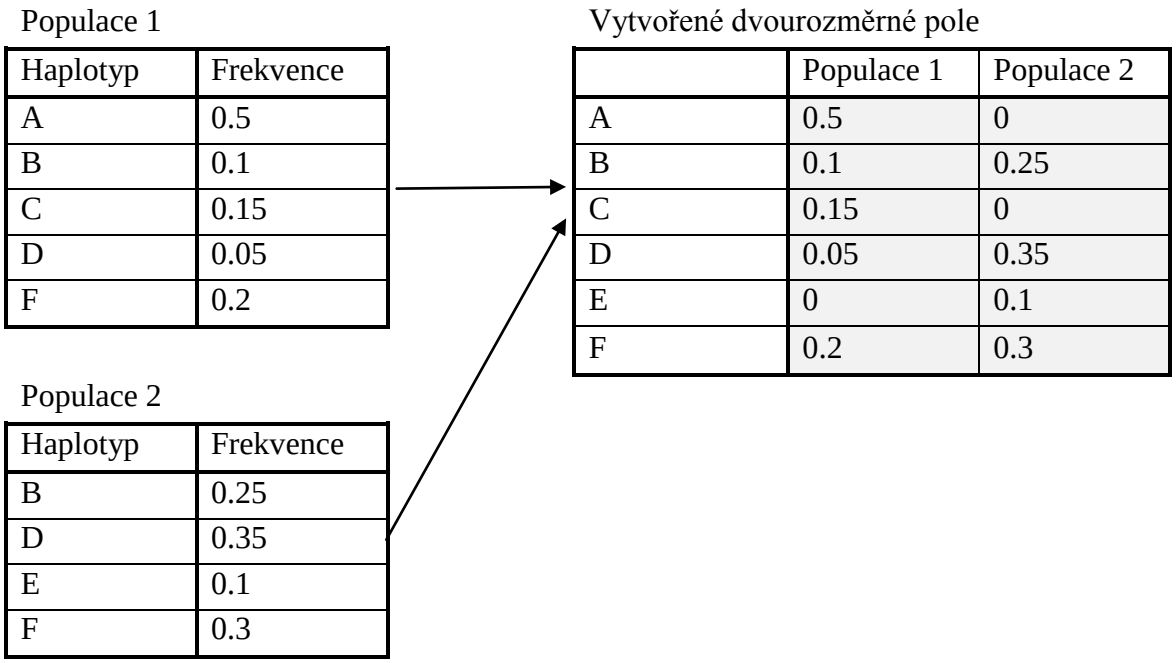
Všechny použité metody pro výpočet genetické vzdálenosti pracují na podobném principu – nějakým způsobem se porovnávají frekvence haplotypů u jednotlivých populací. Přesný popis jednotlivých metod výpočtu je uveden v kapitole 4.

K výpočtu vzdálenosti mezi dvěma populacemi je tedy třeba porovnat frekvence všech haplotypů ve všech populacích. To znamená nalézt odpovídající si haplotypy v jednotlivých populacích a s jejich frekvencemi provést matematické operace podle vzorců (1) – (6).

Původní implementace, která k vyhledávání korespondujících haplotypů využívala hash tabulku byla velmi pomalá – výpočet genetických vzdáleností mezi 10 populacemi s průměrně 3000 haplotypy trval více než deset minut. Rychlost programu jsem se proto snažil optimalizovat.

Využil jsem skutečnosti, že všechna data máme k dispozici před zahájením výpočtu a vytvořil jsem dvourozměrné pole, ve kterém jsou uloženy frekvence všech haplotypů

pro všechny populace. Při výpočtu se data nemusí žádným způsobem vyhledávat ať už v seznamu nebo v hash tabulce a máme k nim přístup přímou indexací pole. Tvorbu tohoto pole znázorňuje diagram na obrázku 12, v paměti se uchovává pouze šedě podbarvená část pole.



Obr. 12 – Tvorba dvourozměrného pole s daty pro urychlení výpočtu

S použitím této tabulky se podařilo výpočet značně urychlit - výpočet genetických vzdáleností mezi 10 populacemi s průměrně 3000 haplotypy trvá řádově jednotky sekund (počítač s procesorem Intel Core2Duo 2GHz).

5.1.1. **Generování fylogenetického stromu**

Neighbor-joining [29] je metoda shlukové analýzy používaná ke konstrukci fylogenetických stromů. Ke konstrukci stromu využívá matici vzdáleností mezi všemi sledovanými populacemi navzájem. Metoda konstrukce stromu pomocí algoritmu neighbor-joining je vyjádřena následujícím postupem:

- na základě matice vzdáleností obsahující r populací spočti matici Q

$$Q(i,j) = (r-2)d(i,j) - \sum_{k=1}^{r} d(i,k) - \sum_{k=1}^{r} d(j,k) \qquad \textbf{(8)}$$

  kde $d(i,j)$ je vzdálenost mezi populacemi i a j
- najdi dvojici populací, které přísluší v matici Q nejnižší hodnota a pro tuto dvojici vytvoř uzel, který spojuje tyto dvě populace
- spočítej vzdálenost obou populací k nově vytvořenému uzlu (f a g představují populace s nejnižší hodnotou v matici Q a u představuje nově vytvořený uzel)

$$d(f,u) = \frac{1}{2}(f,g) + \frac{1}{2(r-2)}\left[\sum_{k=1}^{r} d(f,k) - \sum_{k=1}^{r} d(g,k)\right] \qquad \textbf{(9)}$$

- spočítej vzdálenost všech ostatních populací k nově vytvořenému uzlu

$$d(u,k) = \frac{1}{2}[d(f,k) - d(f,u)] + \frac{1}{2}[d(g,k) - d(g,u)] \qquad \textbf{(10)}$$

  k představuje populaci, pro kterou vzdálenost počítáme, *u* nově vytvořený uzel, *g* a *f* původní populace
- spusť algoritmus od začátku a nově vytvořený uzel považuj za jednu populaci

Algoritmus neighbor-joining pracuje s kritériem minimální evoluce pro fylogenetické stromy – preferuje topologii s nejmenší celkovou délkou všech větví, tedy takovou, které odpovídá minimální počet evolučních změn. Protože se jedná o tzv. hladový algoritmus, tak nemusí vždy nalézt optimální strom. Byl ale rozsáhle testován a většinou najde strom, který se optimálnímu stromu velmi blíží [30] [31].

Největší výhodou tohoto algoritmu je jeho rychlost. Na rozdíl od ostatních (maximum parsimony, maximum likelihood nebo Bayesian inference) pracuje s polynomiální složitostí, takže je možné jej aplikovat na rozsáhlé soubory dat. Jiný polynomiální algoritmus UPGMA zase má jako vstupní požadavek konstantní rychlost evoluce [32], což nejsme schopni pro naše data dokázat.

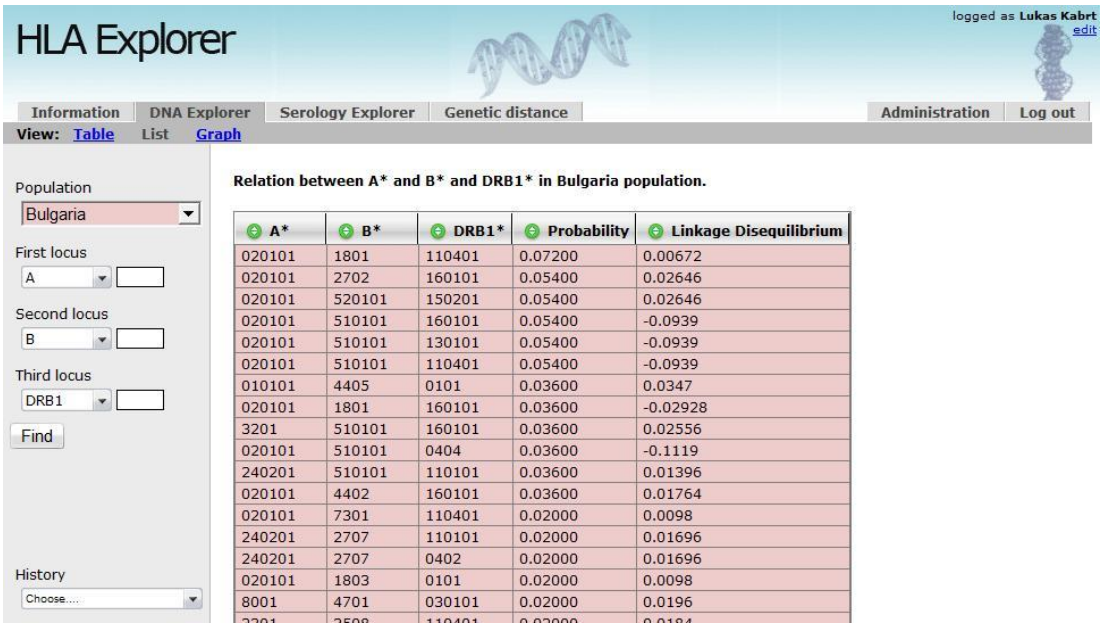
Z výše uvedených důvodů jsem pro konstrukci fylogenetického stromu zvolil algoritmus neighbor-joining. Jedná se o dobrý kompromis mezi optimálními výsledky a výpočetní náročností.

Výsledný fylogenetický strom je vygenerován jako obrázek ve formátu PNG.

5.2. Webová prezentace

Prezentaci výsledků jsem integroval do již existující stránek HLA Explorer, které vznikají na Katedře kybernetiky a které soustřeďují výsledky dalších projektů studujících HLA.

Webové stránky HLA Explorer v současné době obsahují nástroj pro analýzu závislostí výskytu alel a haplotypů HLA. Stránky podporují autorizaci uživatelů na základě uživatelského jména a hesla a obsahují administrační rozhraní pro aktualizaci dat.



Obr. 13 – HLA Explorer – ukázka uživatelského rozhraní

Protože webové stránky HLA Explorer již existují, tak byla technologie předem dána – jako skriptovací jazyk je použito PHP a jako databáze je použito MySQL. Prezentace výsledků této práce využívá moderních technologií jako je JavaScript a SVG (Scalable Vector Graphics) [33].

Výsledky jsem se snažil zobrazit v takové formě, aby byly pro uživatele co nejnázornější. Uživatel si může zvolit jeden ze tří pohledů:

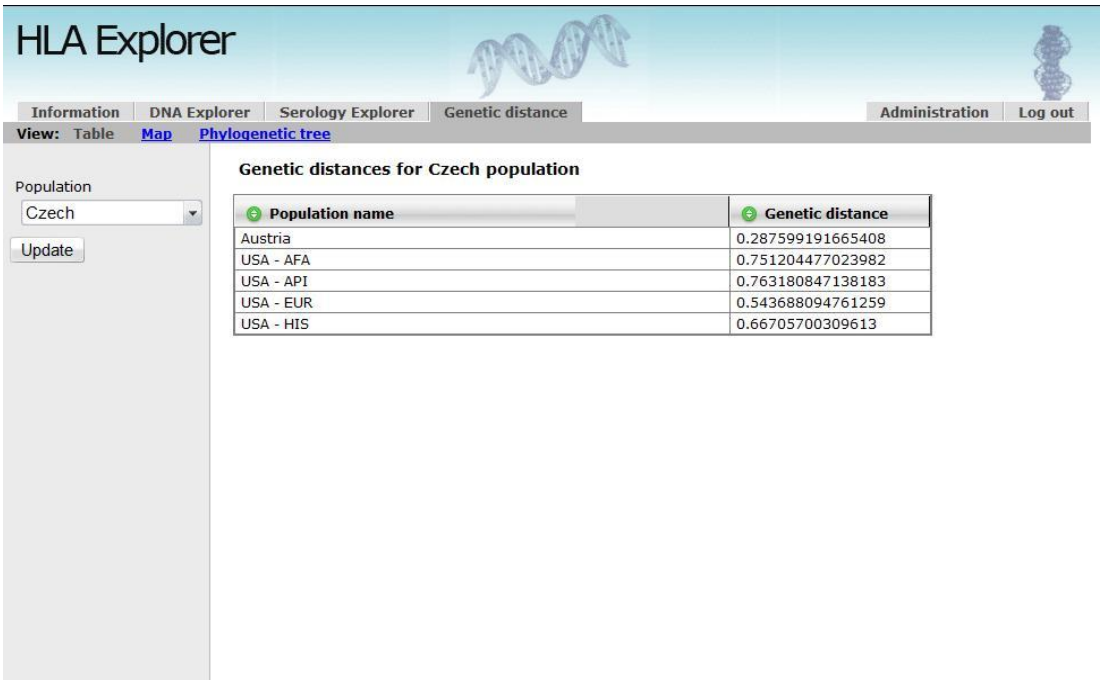
- tabulka
- mapa
- fylogenetický strom

5.2.1. **Zobrazení ve formě tabulky**

Zobrazení výsledků ve formě tabulky je základním prostředkem prezentace dat. Výhodou tabulkového zobrazení je množnost zobrazit velké množství dat najednou.

V tabulce se pro zvolenou populaci zobrazuje vzdálenost k ostatním populacím. Data je možné třídit jak podle názvu, tak podle genetické vzdálenosti.

Ukázka tabulkového zobrazení genetických vzdáleností v aplikaci HLA Explorer je na obrázku 14.

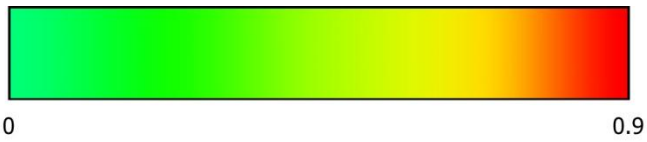


Obr. 14 – HLA Explorer – zobrazení genetických vzdáleností ve formě tabulky

5.2.2. **Zobrazení dat na mapě**

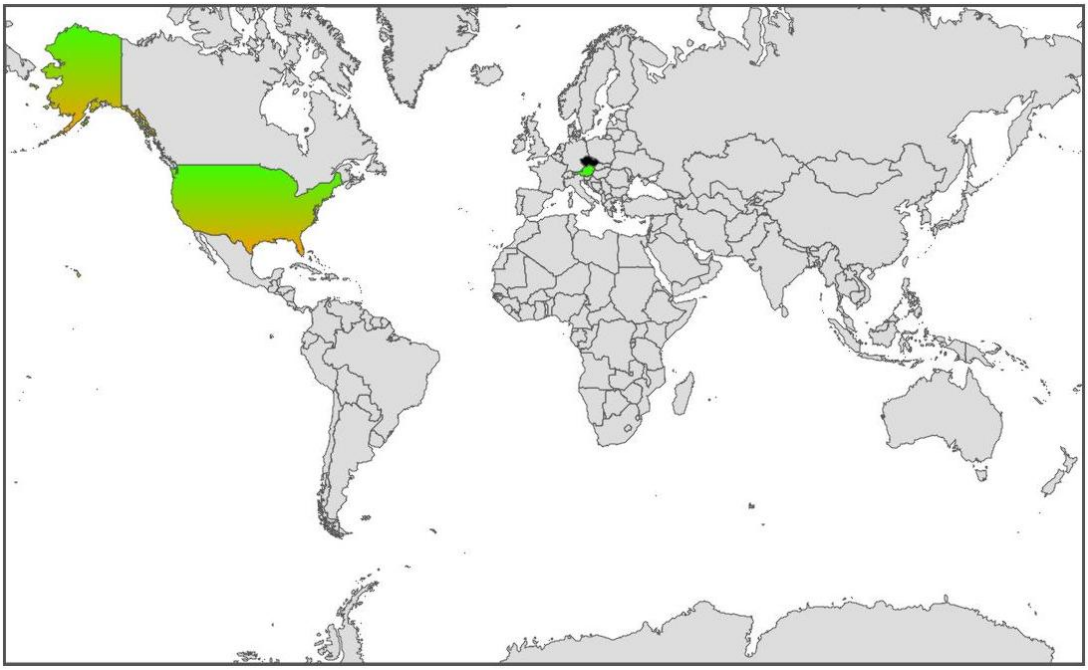
Zobrazení dat na mapě je velmi názorné. Umožňuje jednoduše pro vybranou populaci studovat závislost genetické vzdálenosti k ostatním populacím vzhledem ke geografické poloze.

Hodnota genetické vzdálenosti je vyjádřena barvou – nejnižší genetické vzdálenosti odpovídá zelená barva, nejvyšší genetické vzdálenosti odpovídá červená barva. Černá barva znázorňuje region, ve kterém se nachází vybraná populace. Stupnice je uvedena na obrázku 15.

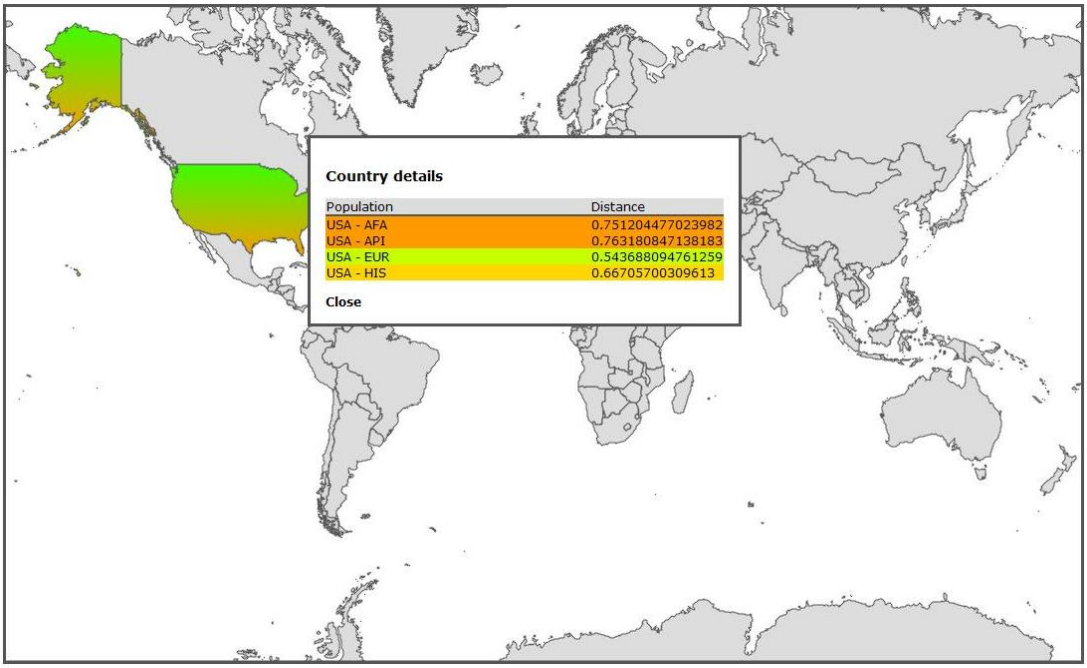


Obr. 15 – Stupnice pro vyjádření genetické vzdálenosti na mapě

Pokud se v jedné oblasti vyskytuje více populací, tak se daná oblast nevyplní jednolitou barvou ale barevným gradientem. Počáteční a konečná barva gradientu je určena minimální a maximální hodnotou genetické vzdálenosti mezi vybranou populací a populacemi v daném regionu. Na obrázku 16 je to případ Spojených států Amerických. Kliknutím na takovýto region se v prohlížeči zobrazí dialogové okno s detaily pro všechny populace v daném regionu, viz obrázek 17.



Obr. 16 – Znázornění regionu s více populacemi



Obr. 17 – Zobrazení detailu regionu s více populacemi

V současné verzi jsou k dispozici následující mapy:

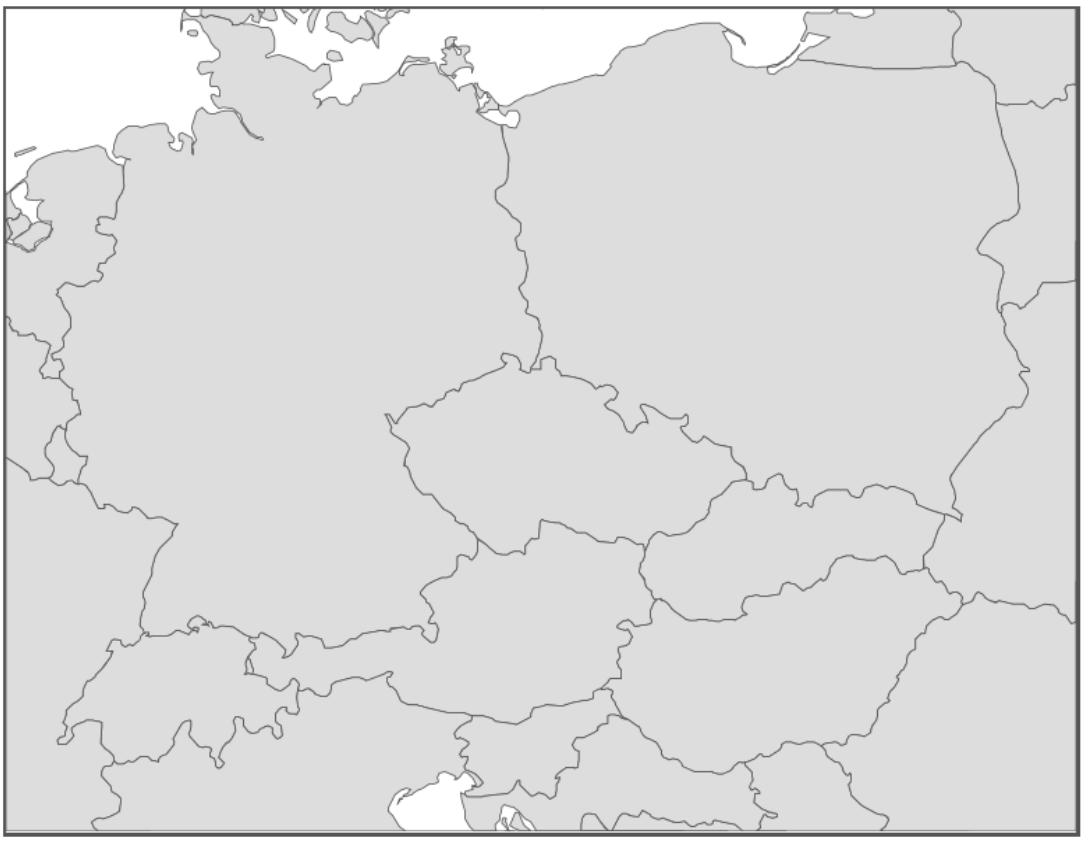
- svět
- Afrika
- Austrálie
- Evropa

- severní Evropa
- západní Evropa
- východní Evropa
- jižní Evropa
- střední Evropa
- severní Amerika
- střední Amerika
- jižní Amerika
- Asie
- jiho-východní Asie

Na všech mapách je území děleno na úrovni států.

Mapy byly vybrány tak, aby poskytly celkový přehled o genetických vzdálenostech mezi populacemi na celosvětové úrovni a na úrovni kontinentů. Podrobnější mapy byly doplněny pro oblasti s větším počtem malých států a pro území Evropy.

Ukázka podrobnější mapy střední Evropy je na obrázku 18.



Obr. 18 – Podrobná mapa střední Evropy

Pro zobrazení map se nabízelo několik různých technologií – staticky generované bitmapy, použití mapových serverů (Google Maps, MapQuest nebo jiné) nebo mapy ve formátu SVG.

Staticky generované bitmapy jsem zavrhnul hned v počátku – při každé aktualizaci dat by se totiž musely aktualizovat (přegenerovat) všechny mapy. Při současném počtu 14 map a 8 populací v databázi se jedná o více než 100 souborů. S ohledem na budoucí rozvoj a přidávání dalších populací do databáze se mi tato technologie nezdála perspektivní.

Další možností bylo použití mapových serverů jako je Google Maps nebo dalších. Toto řešení by poskytovalo komfortní uživatelské rozhraní s možností plynulého pohybu po mapě a zoomování, byla by zajištěna aktuálnost mapových podkladů i vývoj a testování platformy v nových internetových prohlížečích. Bohužel při testování této technologie jsem narazil na výkonnostní omezení, pro které jsem musel toto řešení opustit.

Google Maps umožňuje zobrazit uživatelská data dvojím způsobem – kreslením do základní mapy nebo vytvořením uživatelské mapové vrstvy.

První způsob funguje tak, že se pomocí skriptu na straně klienta do mapy přidávají uživatelské objekty – body, čáry, ikony či polygony. Při testování jsem zjistil, že v přiměřeném čase (jednotky sekund) lze zobrazit polygony, jejichž hranice jsou určeny řádově tisíci bodů. Mapa světa s přiměřeným počtem detailů (rozlišení cca 10km/pixel) je přitom tvořena stovkami polygonů určených přibližně 50 000 body.

Druhou možností je vytvoření uživatelské mapové vrstvy. V tomto případě se vrstva mapy dynamicky generuje jako bitmapa na takzvaném „tile-serveru“ a ve webovém prohlížeči se sloučí s daty z mapového serveru. Tile-server je serverová aplikace, která na základě požadavku z webového prohlížeče uživatele vygeneruje požadovaný úsek mapy (nebo v tomto případě jedné mapové vrstvy) v požadovaném měřítku a vrátí ho klientovi. Toto řešení by splňovalo výkonnostní požadavky, ale bohužel se mi nepodařilo nalézt žádný volně dostupný nebo cenově přijatelný tile-server. Implementace vlastního řešení by byla velmi náročná. Tento úkol může být cílem další práce zdokonalující HLA Explorer.

Třetí možností, kterou jsem nakonec zvolil, jsou mapy ve formátu SVG. SVG je zkratka pro Scalable Vector Graphics – vektorový formát založený na technologii XML navržený speciálně pro použití v internetových prohlížečích. Mapa ve formátu SVG je rozdělena na skupiny objektů (v tomto případě odpovídající státům světa) jejichž vlastnosti můžeme měnit pomocí javascriptu. To je veliká výhoda tohoto řešení – mapa se na uživatelův počítač stahuje pouze jednou a data se do ní přidávají dynamicky. SVG mapy jsou generovány programem popsaným v kapitole 5.3

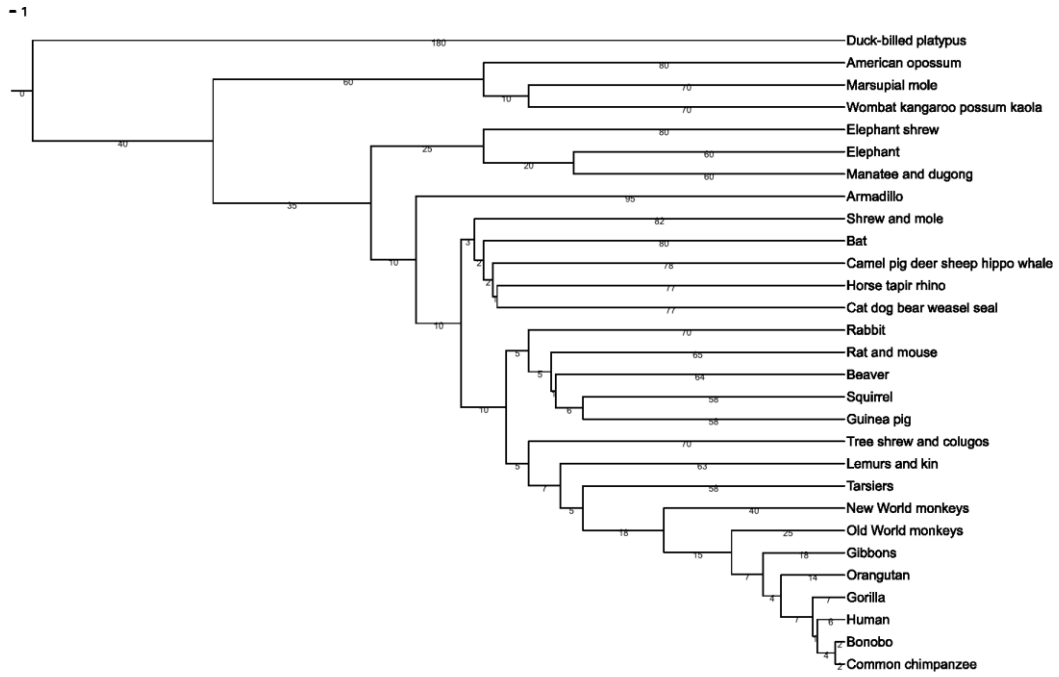
Rychlost zobrazení je dostatečná, mapa se stejnými parametry jako použitá při testování Google maps, se zobrazí v čase cca 2 sekund (počítač s procesorem Intel Core2Duo, prohlížeč Opera 9.52). Velikost map se pohybuje od 250 kB u mapy světa do 60 kB u mapy střední Ameriky.

Určitou nevýhodou tohoto řešení může být nedostatečná podpora SVG formátu ve starších internetových prohlížečích a nemožnost dynamické změny měřítka mapy. Nejrozšířenější současné webové prohlížeče ale SVG formát již podporují – Firefox, Mozzila, Opera, Safari a Google Chrome nativně, Internet Explorer formou ActiveX pluginu [34].

5.2.3. **Fylogenetický strom**

Fylogenetický strom je grafické zobrazení připomínající strom, jímž se znázorňují příbuzenské vztahy mezi různými biologickými druhy či jinými taxonomickými jednotkami, o nichž se předpokládá, že mají jednoho společného předka. Každé větvení představuje posledního hypotetického společného předka. Každá větev znázorňuje jednu evoluční linii, na jejímž konci jsou dané taxony [35]. Pokud budeme předpokládat, že se všechny populace vyvinuly z jedné prapůvodní – což je pravděpodobné, můžeme fylogenetický strom použít i pro znázornění vztahů mezi populacemi.

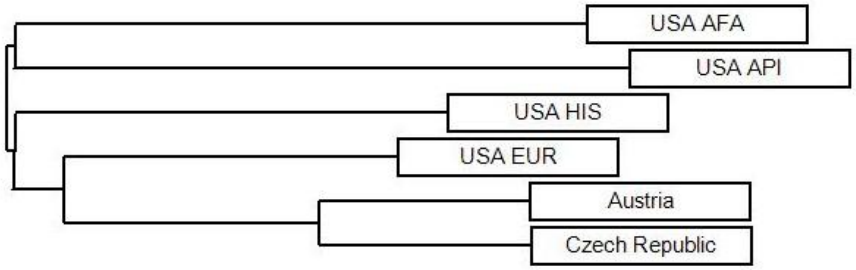
Ukázka komplexnějšího fylogenetického stromu znázorňujícího příbuznost mezi savci je na obr 19.



Obr. 19 – Fylogenetický strom znázorňující příbuznost savců [36]

Délka cesty mezi dvěma uzly, či listy ve fylogenetickém stromu (počítáno pouze ve vodorovném směru) je přímo úměrná genetické vzdálenosti mezi příslušnými druhy či v tomto případě populacemi.

Příklad fylogenetického stromu generovaného aplikací HLA Explorer je na obr. 20. Podrobnější informace o algoritmu použitém ke generování fylogenetického stromu jsou uvedeny v kapitole 5.1.1.



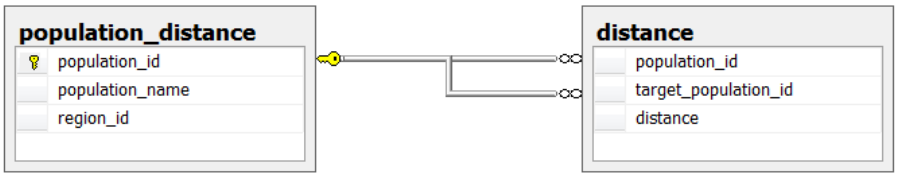
Obr. 20 – Příklad fylogenetického stromu zobrazovaného v aplikaci HLA Explorer

5.2.4. **Integrace do HLA Exploreru**

Při integraci výsledků do systému HLA Explorer byly využity stejné technologie jako při jeho tvorbě – PHP jako skriptovací jazyk a databáze MySQL pro uložení dat.

Modul pro zobrazení výsledků této diplomové práce využívá stávajících podpůrných funkcí již integrovaných v systému HLA Explorer. Jedná se především o funkce zajišťující autentifikaci a autorizaci uživatelů, přístup do databáze a funkce pro integraci do struktury stránek.

Schéma databázové struktury použité pro uložení dat je znázorněno na obrázku 21.



Obr. 21 – Schéma databázové struktury

Jak je vidět struktura tabulek je velmi jednoduchá – jedna tabulka slouží pro uložení informací o populacích a druhá pro uložení vzdáleností mezi nimi.

Aktualizace dat je prováděna přes administrační rozhraní. Uživatel s příslušnými oprávněními má možnost importovat soubory generované výpočetním programem a tak provést aktualizaci.

5.3. **Nástroj pro generování mapových podkladů**

Tento program slouží pro generování SVG map na základě geografických dat. Jedná se o konsolovou aplikaci vytvořenou v prostředí .NET a jazyce C#.

Program na svém vstupu přijímá soubor s geografickými daty ve formátu XML a soubory s definicemi map (také ve formátu XML). Výstupem programu je sada map ve formátu SVG. Formát souboru s geografickými daty je uveden v příloze D a formát definičního souboru mapy je uveden v příloze E.

Program vyžaduje následující parametry z příkazové řádky.

```
svg-maps.exe [geo-data] [maps]
```

`[geo-data]`	adresa souboru s geografickými daty
`[maps]`	cesta, k adresáři obsahujícímu soubory s definicemi map

Soubor s geografickými daty obsahuje hranice zobrazovaných regionů jako polygony určené body v geografických souřadnicích (zeměpisná délka a zeměpisná šířka). Při generování map pro webovou prezentaci jsem použil geografická data s regiony odpovídajícími státům světa. K dispozici jsou ale i geografická data s podrobnějším členěním [37].

Vlastnosti generované mapy – zobrazovaná oblast, rozlišení a zobrazené regiony jsou určeny jejím definičním souborem. Na jeho základě program vybere příslušná geografická data, provede jejich projekci do mapy a výslednou mapu uloží ve formátu SVG.

Při tvorbě mapy se snažíme zobrazit kulovou plochu do roviny, proto je výsledný obraz nějakým způsobem deformován. Při tvorbě map se tomuto zobrazení říká kartografická projekce.

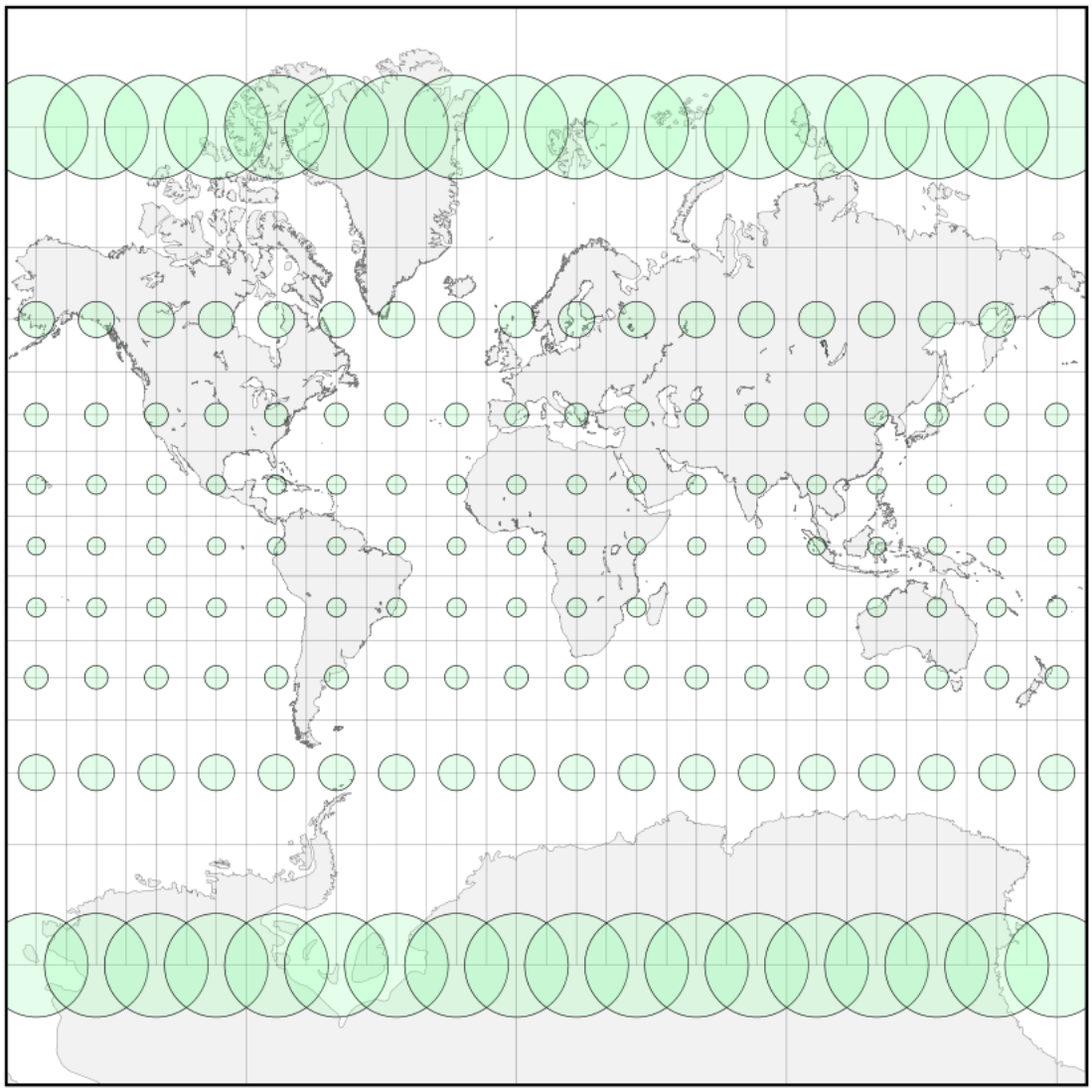
Program používá projekci Mercator [38] – jedná se o typ úhlojevné válcové projekce. Její výhodou je zachování úhlů a tvarů. K deformacím dochází v polárních oblastech, které se jeví větší. Zkreslení, ke kterému dochází je znázorněno na obrázku 22 pomocí takzvaného Tissotova indikatrixu [39]. V ideálním případě by všechny kruhy měly být stejně veliké.

Projekce Mercator je definována vzorci 11 a 12. [38]

$$x = \lambda - \lambda_0 \tag{11}$$

$$y = \frac{1}{2} ln\left(\frac{1 + \sin(\varphi)}{1 - \sin(\varphi)}\right) \tag{12}$$

kde x a y značí souřadnice v pravoúhlé síti na mapě, λ zeměpisnou délku a φ zeměpisnou šířku. λ_0 určuje střed mapy.



Obr. 22 – Zkreslení projekce Mercator znázorněné pomocí Tissotova indikatrixu [39]

Projekci Mercator jsem vybral, protože se jedná o úhlojevné zobrazení. Díky tomu mají všechny oblasti na všech vygenerovaných mapách stejný tvar. Takové chování je z hlediska ergonomie přínosem.

6. Výsledky

Úkolem této práce mělo být sledovat příbuznost české populace s ostatními. Při výběru metody výpočtu a implementaci jsem zjistil, že s minimálním úsilím mohu zároveň vypočítat i genetické vzdálenosti mezi ostatními populacemi navzájem. To by mohlo přinést zajímavé výsledky a umožní mi to zkonstruovat fylogenetický strom. Z těchto důvodů jsem se rozhodl určit genetické vzdálenosti i mezi ostatními populacemi. Při vyhodnocování výsledků jsem se ale zaměřil na českou populaci.

Největším problémem se ukázala dostupnost dat – podařilo se získat data o české, rakouské, slovenské a polské populaci. Jedná se o anonymizovaná data z registrů dárců kostní dřeně z příslušných zemí, se kterými má katedra kybernetiky navázanou spolupráci. Data o populacích z USA, rozdělená podle etnického původu, byla získána od organizace National Marrow Donor Program, která anonymní data o HLA potenciálních dárců kostní dřeně z USA poskytuje na svých webových stránkách [40].

Pro výpočet jsem původně chtěl použít i data z některých vědeckých studií publikovaná na webu allefrequencies.net, ale jak bylo řečeno v kapitole 3.2 tato data nakonec nešlo využít.

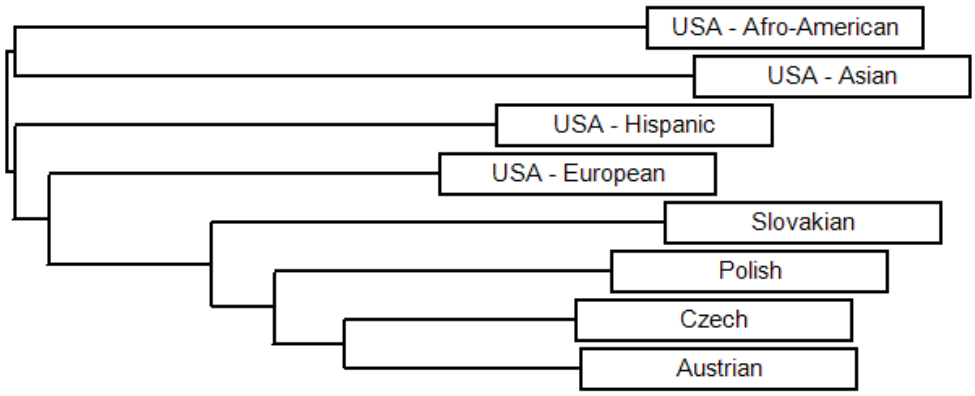
Vypočtené genetické vzdálenosti mezi všemi populacemi navzájem jsou uvedeny v následující tabulce.

	Česká	Rakouská	Slovenská	Polská	USA – Afro-americká	USA - Asijská	USA - Evropská	USA - Hispánská
Česká		0,29	0,51	0,39	0,75	0,76	0,54	0,67
Rakouská	0,29		0,52	0,40	0,75	0,76	0,54	0,67
Slovenská	0,51	0,52		0,52	0,79	0,80	0,63	0,72
Polská	0,39	0,40	0,52		0,77	0,78	0,58	0,69
USA – Afro-americká	0,75	0,75	0,79	0,77		0,81	0,68	0,69
USA - Asijská	0,76	0,76	0,80	0,78	0,81		0,70	0,73
USA - Evropská	0,54	0,54	0,63	0,58	0,68	0,70		0,54
USA - Hispánská	0,67	0,67	0,72	0,69	0,69	0,73	0,54	

Tab. 7 – Genetické vzdálenosti mezi všemi populacemi navzájem

I když z hlediska vyhodnocování výsledků je zajímavější sledovat relativní rozdíly mezi populacemi, tak je dobré připomenout, že minimální hodnota vzdálenosti pro použitou metodu výpočtu (Cavali-Sforza) je 0 a maximální hodnota vzdálenosti je 0,9.

Na základě vypočtených vzdáleností mezi všemi populacemi navzájem byl vygenerován následující fylogenetický strom.



Obr. 23 – Fylogenetický strom znázorňující vztahy mezi všemi populacemi

Při pohledu na výsledky vidíme, že populace z evropských zemí jsou nejvíce příbuzné mezi sebou, z populací v USA je k nim nejvíce příbuzná populace tvořená evropskými přistěhovalci. Na základě historických a geopolitických souvislostí jsou to výsledky pochopitelné.

Stejně tak jsou pochopitelné výsledky i pro populace žijící v USA. Populace tvořená evropskými přistěhovalci je stále nejvíce příbuzná s ostatními evropskými populacemi, i když se od nich křížením s ostatními poněkud vzdálila. Populace z USA mají mezi sebou nižší genetickou vzdálenost než k evropským, z toho vyplývá, že mezi nimi došlo k určitému promísení. To je ostatně zřejmé i z vygenerovaného fylogenetického stromu.

Za povšimnutí stojí vysoké hodnoty genetických vzdáleností pro asijskou populaci v USA ke všem ostatním. Ukazuje to na vyšší genetickou odlišnost asiatů a jen to potvrzuje statistiky z Japonska, podle kterých japonský pacient najde nejaponského dárce kostní dřeně jen v opravdu výjimečných případech [41].

Žádný z výsledků není vyloženě překvapivý, takový, který by nešel zdůvodnit historickými a geopolitickými souvislostmi. Na jejich základě můžeme prohlásit, že použitá metoda výpočtu genetické vzdálenosti funguje správně.

Výsledky z pohledu české populace jsou podrobněji rozebrány v následující kapitole 6.1.

6.1. Výsledky z pohledu české populace

Pokud chceme studovat příbuznost české populace s ostatními, je dobré připomenout historii osidlování našeho území a objasnit hlavní imigrační a emigrační vlny z české a do české populace.

6.1.1. Historie osídlení českých zemí

Dějiny osídlení českých zemí známými etniky začínají zhruba v polovině prvního tisíciletí před Kristem. Tehdy žily na území Čech a Moravy keltské kmeny, které se v době kolem přelomu letopočtu přesunuly dále na západ a byly vystřídány germánskými kmeny. Také Germáni migrovali zhruba po pěti staletích pobytu dále na západ. V období takzvaného stěhování národů se zde natrvalo usídlili Slované, přicházející z oblasti východně od Karpat [42].

V průběhu 13. a 14. století, především za vlády Karla IV., kdy byla Praha evropskou metropolí, se v Čechách usazovali skupiny jiných národností. Na počátku 15. století, v období husitských válek, naopak došlo k útěku většiny cizinců, hlavně Němců. Zbytky německé menšiny se udržely pouze v příhraničních oblastech.

V 16. století a na počátku 17. století do Prahy opět přicházejí vyslanci, obchodníci a řemeslníci různých národností z celé Evropy. Vrchol této éry nastává za vlády Rudolfa II.

Po bitvě na Bílé hoře v roce 1620 jsou desetitisíce nekatolických měšťanů a šlechticů [42] nuceni odejít ze země a nastává téměř tři století trvající područí v rámci Rakousko-Uherska. V této době docházelo k migraci v rámci monarchie a tak při vzniku samostatného státu je Československo značně národnostně roztříštěno. Nejpočetnějšími menšinami jsou Němci, Rakušané, Poláci, Maďaři a Rusíni.

V průběhu 2. světové války dochází k perzekuci Čechů (především židovského vyznání nebo romského původu) a útěkům do zámoří. Následně v poválečných letech dochází k odsunu sudetských Němců do Rakouska a Německa.

Jak se vidět, původ obyvatelstva na území ČR musí být po složitém historickém vývoji velmi pestrý.

Podle výzkumů společnosti Genomac, která se zabývá analýzou DNA je původ české populace následující: západoslovanský (40%), románský (25%), jihoslovanský (11%), Skandinávie a Německo (10%), Středomoří, Balkán, severní Afrika (5%), Středomoří, Blízký východ, Asie (4%), Pobaltí, Skandinávie, Sibiř (3%), Středomoří, Kavkaz (1%).

Genetické znaky typické pro Slovany má ve své DNA jen 51% Čechů a Moravanů. Zbytek má románské, germánské, židovské, ugrofinské či jihokavkazské předky [43].

6.1.2. **Češi v ostatních zemích**

K českému původu se v zahraničí hlásí téměř dva miliony lidí. Příčiny jejich emigrace lze podle dominantních příčin rozdělit do tří skupin – vystěhovalectví náboženské, sociální a politické.

Nejvýznamnější vlny emigrace spolu s místem, kam směřovaly [44], jsou shrnuty v následujících bodech:

- náboženský pobělohorský exil; Sasko a Prusko, USA
- sociální vystěhovalectví v 2. polovině 19. a na počátku 20. století; USA, později i jižní Amerika (Argentina, Brazílie) a západní Evropa
- emigrace za 2. světové války; USA, jižní Amerika
- politická emigrace v letech 1948 a 1968; západní Evropa, USA, Kanada, Austrálie, jižní Amerika a jižní Afrika

Tabulka 8 ukazuje přehled 10 zemí s největší českou menšinou a jejich percentuální zastoupení v tamní populaci.

Stát	Počet Čechů	Zastoupení v populaci
USA	1 599 000	0,53%
Kanada	139 910	0,43%
Rakousko	54 627	0,60%
SRN	cca 50 000	0,06%
Slovensko	46 801	0,85%
V. Británie	cca 40 000	0,07%
Argentina	cca 30 000	0,08%
Austrálie	27 196	0,14%
Francie	cca 30 000	0,05%
Švýcarsko	cca 15 000	0,18%

Tab. 8 – Počty Čechů v ostatních zemích a jejich zastoupení v tamní populaci [44]

Počty Čechů žijících v zahraničí jsou založeny na výsledcích sčítaní lidu v daných zemích. Představují ty obyvatele, kteří se přihlásili k české národnosti. Skutečné počty lidí pocházejících z ČR, nebo majících české předky, mohou být ještě vyšší.

6.1.3. **Cizinci v ČR**

Podle výsledků statistických výzkumů se 94,2% obyvatel žijících v ČR hlásí k české národnosti. To je velmi vysoký podíl. Zastoupení jiných národností je uvedeno v tabulce 9.

Národnost	Počet	Zastoupení v populaci
slovenská	314 877	1,9%
polská	51 968	0,5%
německá	39 106	0,4%
ukrajinská	22 112	0,2%
vietnamská	17 402	0,2%
maďarská	14 672	0,1%
ruská	12 369	0,1%
romská	11 746	0,1%

Tab. 9 – Zastoupení ostatních národností v české populaci [45]

Výše uvedená data o národnosti cizinců legálně žijících na území ČR jsou založena na výsledcích sčítání lidu z roku 2001. Skutečné počty cizinců, kteří žijí v ČR, budou mnohem vyšší. Mnoho dočasně žijících cizinců u nás pobývá nelegálně, a tudíž nemohou být ve výsledcích zahrnuty. Dalším problémem jsou cizinci, kteří se v průzkumech nehlásí ke své národnosti. Například při sčítání lidu v roce 2001 se k romské národnosti přihlásilo necelých 12 tisíc obyvatel, přitom romštinu jako svůj rodný jazyk uvedlo okolo 40 tisíc obyvatel. [45]

Při zjišťování původu obyvatel tak může být mimo národnosti důležitým faktorem i rodný jazyk nebo náboženství.

Počty Čechů žijících v zahraničí vzhledem k velikostem tamních populací nejsou příliš vysoké – v žádné zemi nedosahují ani jednoho procenta. Stejně tak počty cizinců žijících v ČR nepřesahují 1% (mimo Slováků, kteří mají 1,8% zastoupení). Jsou to tak nízká čísla, že celkovou příbuznost české populace s ostatními nemohou příliš ovlivnit.

Počty a směry emigrace Čechů nám ale mohou poskytnout vodítko, jak se situace vyvíjela i v jiných zemích, protože historický vývoj v ostatních státech střední Evropy (Polsko, Slovensko, Maďarsko) byl z hlediska emigrace v 20. století podobný.

6.1.4. **Vypočtené genetické vzdálenosti**

Genetické vzdálenosti české a ostatních populací shrnuje následující tabulka.

Populace	Genetická vzdálenost
rakouská	0,29
slovenská	0,51
polská	0,39
USA – afro-americká	0,75
USA - asijská	0,76
USA - evropská	0,54
USA - hispánská	0,67

Tab. 10 – Genetická vzdálenost české populace k ostatním populacím

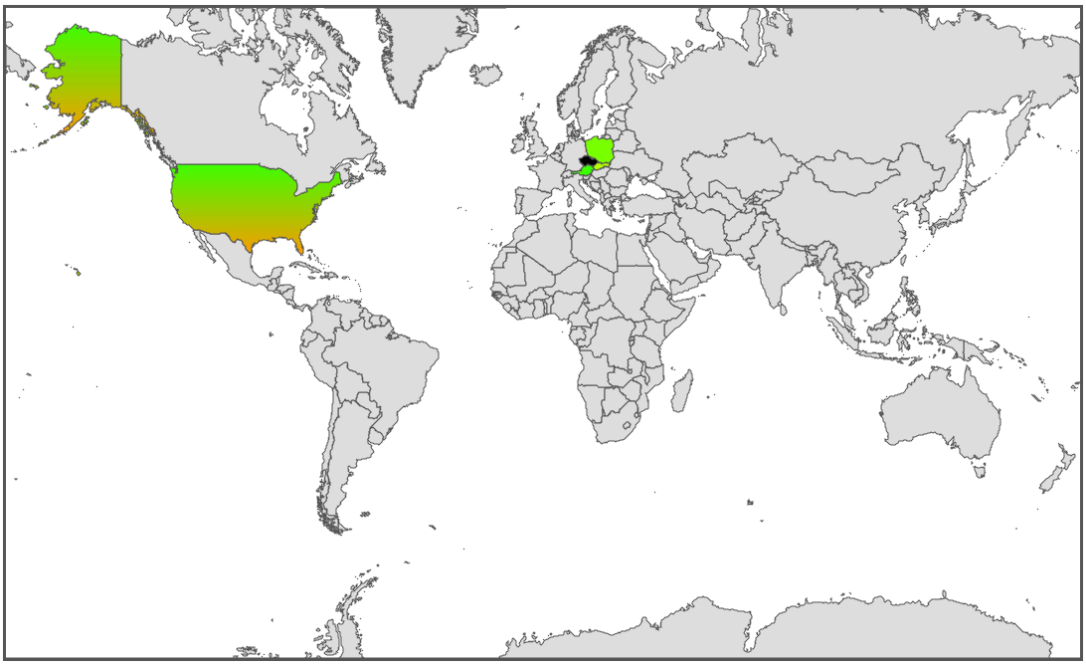
Vzhledem k populacím z USA (obr. 24) má česká populace nejnižší genetickou vzdálenost k populaci tvořené evropskými přistěhovalci. Následuje hispánská – jedná se o etnikum vzniklé smísením španělských přistěhovalců a místních obyvatel v střední a jižní Americe. Původ Hispánců je alespoň z části také evropský, takže výsledek to není překvapivý.

Další dvě populace z USA - afro-americká a asijská mají vzhledem k české nejvyšší hodnotu genetické vzdálenosti 0,75 a 0,76. To jsou čísla blízká maximální možné hodnotě vzdálenosti.

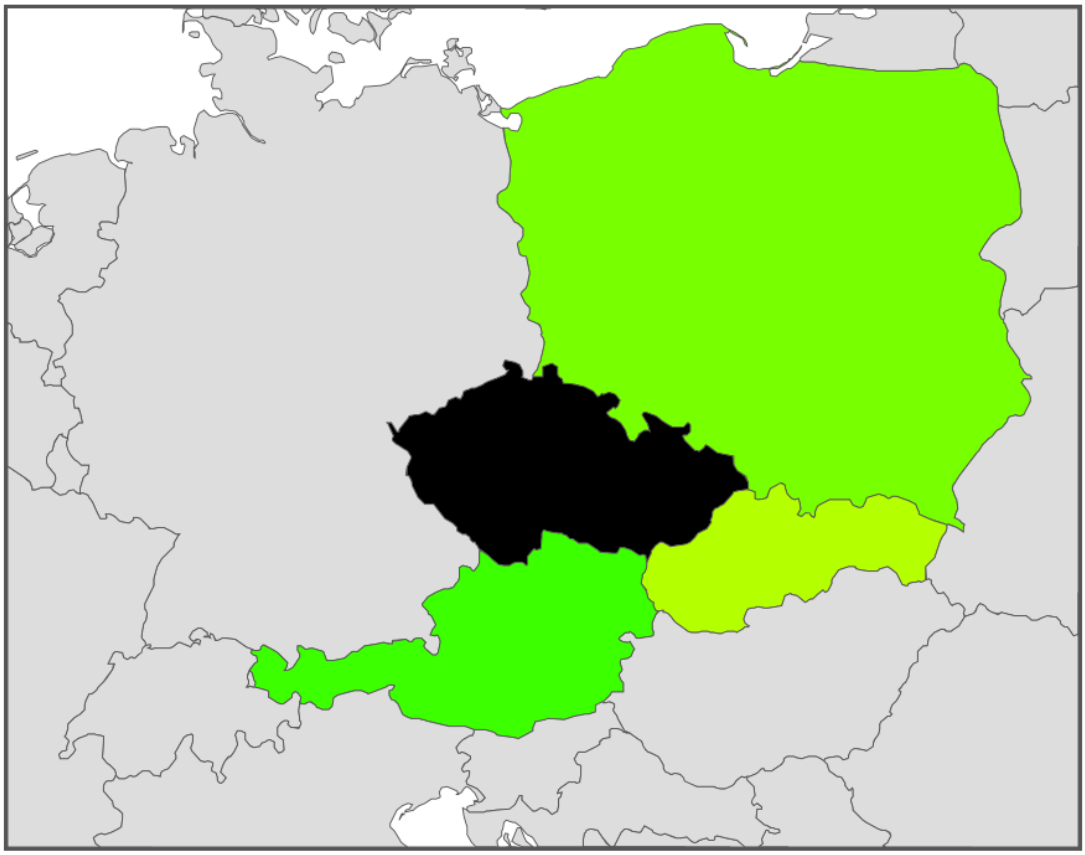
Z evropských zemí byla zkoumána příbuznost české populace s rakouskou, polskou a slovenskou. Detailní zobrazení výsledků v regionu střední Evropy je zobrazeno na obrázku 25.

Nejnižší genetická vzdálenost byla vypočtena vzhledem v rakouské populaci. Je to trochu překvapivý výsledek. Můžeme ho ale zdůvodnit úzkými historickými vazbami.

Vzdálenost polské populace je 0,39. V příbuznosti k české to řadí polskou populaci mezi slovenskou a rakouskou.



Obr. 24 – Genetická vzdálenost mezi českou a ostatními populacemi znázorněná na mapě



Obr. 25 – Genetická vzdálenost mezi českou a ostatními populacemi znázorněná na mapě střední Evropy

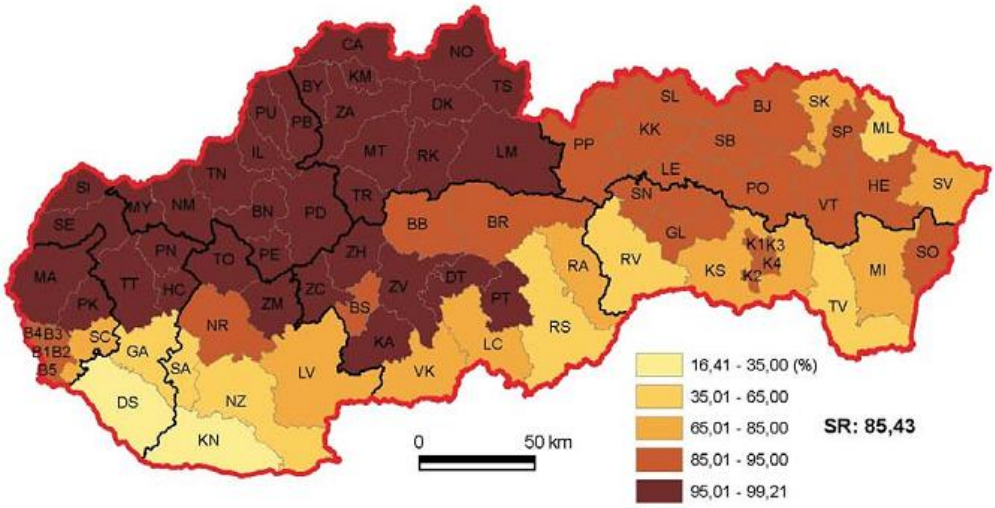
Kupodivu ze zemí střední Evropy vyšla nejvyšší genetická vzdálenost české populace k populaci slovenské. Příčin může být několik.

Slovensko má podle výsledků sčítání lidu vysoké zastoupení ostatních národů v populaci. Přesná čísla jsou uvedena v tabulce 11. Jak vyplývá z mapy na obrázku 26, nejvyšší zastoupení menšin je na jihu a východě Slovenska.

Národnost	Počet	Zastoupení v populaci
maďarská	520 528	9,7%
romská	89 920	1,7%
česká	44 620	0,8%
rusínská	24 201	0,4%
ukrajinská	10 814	0,2%
ostatní	74 518	1,4%

Tab. 11 – Zastoupení ostatních národností v slovenské populaci

..



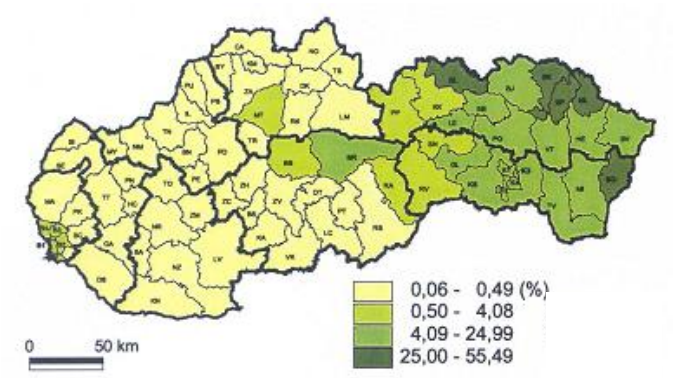
Obr. 26 – Zastoupení slovenské národnosti v jednotlivých okresech [46]

Nejvíce zastoupená je maďarská národnost s téměř 10% zastoupením. Původ Maďarů lze přitom vysledovat až ke kmenům v centrální Asii cca v 5. století našeho letopočtu. Jedná se o etnikum s úplně jiným původem, než mají ostatní populace ve střední Evropě. Jejich 10% menšina tedy může genetickou vzdálenost mezi českou a slovenskou populací ovlivnit.

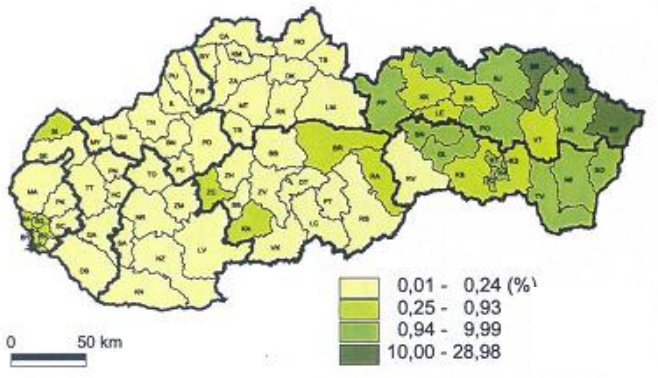
Při sčítání lidu se v případě romské populace na Slovensku projevily stejné problémy jako v ČR, kdy mnoho lidí nechtělo romskou národnost uvést. [47]

Na rozdílný kulturní a etnický původ obyvatel především východní části Slovenska ukazuje i zastoupení řeckokatolické a pravoslavné církve (mapy na obrázcích 27 a 28). Pokud porovnáme zastoupení těchto církví se zastoupením obyvatel hlásících se k slovenské národnosti na východě země, zjistíme, že podíl těchto minoritních

náboženství je nižší než podíl obyvatelstva ostatních národností. To může dokazovat ještě větší podíl ostatních národností nebo alespoň obyvatel s jiným kulturním původem, než ukazují statistiky sčítání lidu.



Obr. 27 – Zastoupení řecko-katolické církve na území SR [48]



Obr. 28 – Zastoupení pravoslavné církve v SR [48]

Všechny výše uvedené příčiny zdůvodňují proč genetická vzdálenost české a slovenské populace je tak vysoká.

7. Závěr

Cílem této diplomové práce bylo nalézt a implementovat metodu výpočtu genetické vzdálenosti mezi populacemi. Tuto metodu aplikovat na dostupná data o výskytu HLA v populacích a pokusit se vysvětlit dosažené výsledky.

V literatuře jsem nalezl několik různých metod výpočtu genetické vzdálenosti – Euklidovská vzdálenost, Cavali-Sforza, Bhattacharyya, Sanghavi, Prevosti a Nei. Tyto metody jsem podrobil několika testům a na jejich základě vybral jako nejvhodnější metodu Cavali-Sforza.

Vybranou metodu jsem aplikoval na anonymizovaná data z několika registrů dárců kostní dřeně a vypočetl jsem vzdálenosti mezi českou, slovenskou, rakouskou, polskou populací a čtyřmi populacemi z USA (asijská, hispánská, afroamerická a evropská). Mezi vypočtenými výsledky jsem nenašel žádnou vyloženě chybnou, hodnotu, která by se nedala vysvětlit historickými a geopolitickými souvislostmi. Určitým překvapením může být pouze vyšší genetická vzdálenost mezi českou a slovenskou populací.

Podstatnou částí této diplomové práce byla implementace prezentační části, jež je integrována do stávající webové aplikace HLA Explorer. Zobrazení výsledků na mapě, které je podle mého názoru ideální pro zkoumání historických a geopolitických souvislostí mezi populacemi, si vyžádalo použití neobvyklého technického řešení. Vzhledem k omezením stávajících mapových serverů jsem byl nucen implementovat vlastní řešení na bázi SVG a javascriptu.

Tato práce je zajímavá tím, že na rozdíl od jiných používá pro výpočet genetické vzdálenosti mezi populacemi data z registrů dárců kostní dřeně. Díky tomu by výsledky této práce šly aplikovat v procesu hledání dárců kostní dřeně.

Jako dílčí neúspěch této práce vidím nižší počet populací, pro které se podařilo získat data a provést výpočet genetické vzdálenosti. Přesto je na dosažených výsledcích vidět, že výpočet funguje správně a výsledky jsou slibné.

Pro rozvoj aplikace vidím jako hlavní cíl získat další data z registrů dárců kostní dřeně. A to jak pro ostatní populace, tak i pro populace stávající s rozlišením na nižší celky než je stát.

8. Zdroje

8.1. Literatura

[1]. ClinImmune Labs. [Online] [11. květen 2009.] http://www.uchsc.edu/clinimmune/documents/histocompatability/Price_List_072408.pdf.

[2]. HLA Explorer. [Online] [13. květen 2009.] http://hlaexplorer.net.

[3]. Genetika - Základy genetiky, dědičnosti a evoluce. [Online] [9. květen 2009.] http://genetika.wz.cz/.

[4]. Wikipedia - Human Leukocyte Antigen. [Online] [18. duben 2009.] http://en.wikipedia.org/wiki/Human_leukocyte_antigen.

[5]. IMGT/HLA Database. [Online] [30. Leden 2009.] http://www.ebi.ac.uk/imgt/hla/stats.html.

[6]. Fève, Frédérique a Florens, Jean-Pierre. How many diferent HLA phenotypes exist in a population? [Online] 2005. [12. květen 2009.] http://ideas.repec.org/p/ide/wpaper/1404.html.

[7]. HLA laboratoř. Hematologicko-onkologické oddělení, FN Plzeň. [Online] [19. duben 2009.] http://www.fnplzen.cz/data/prac/Lochotin/hoo/odborna-verejnost/metody/6-hla.html.

[8]. A., Gerlach J. Human Lymphocyte Antigen Molecular Typing. Archives of Pathology and Laboratory Medicine: Vol. 126, No. 3, pp. 281–284. 2001.

[9]. Holdsworth, R, a další. The HLA dictionary 2008: a summary of HLA-A, -B, -C, -DRB1/3/4/5, and -DQB1 alleles and their association with serologically defined HLA-A, -B, -C, -DR, and -DQ antigens.: John Wiley & Sons A/S, 2009.

[10]. Český registr dárců krvetvorných buněk. [Online] [28. duben 2009.] http://www.czechbmd.cz/.

[11]. Transplantace kostní dřeně. Wikipedia. [Online] [8. květen 2009.] http://cs.wikipedia.org/wiki/Transplantace_kostní_dřeně.

[12]. Český národní registr dárců dřeně. [Online] [25. duben 2009.] http://www.kostnidren.cz/registr/.

[13]. Banka pupečníkové krve ČR. [Online] [25. duben 2009.] http://www.bpk.cz/.

[14]. Number of donors/CBU's per registry in BMDW. Bone Marrow Donors Worldwide. [Online] [1. květen 2009.] http://www.bmdw.org/index.php?id=number_donors0.

[15]. Steiner, David. Search for Unrelated Bone Marrow Donors. 2007. ČVUT, FEL. Diplomová práce.

[16]. NMDP - Be the match. [Online] [12. květen 2009.] http://www.marrow.org.

[17]. Bone Marrow Donors Worldwide. [Online] [1. květen 2009.] http://www.bmdw.org.

[18]. BMDW - File Format for Data Delivery. Bone Marrow Donors Worldwide. [Online] [10. leden 2009.] http://www.bmdw.org/index.php?id=file_formats.

[19]. Jihoafrická republika. Wikipedia. [Online] [13. květen 2009.] http://cs.wikipedia.org/wiki/Jihoafrická_republika.

[20]. Middleton D., Menchaca L., Rood H., Komerofsky R. New Allele Frequency Database: http://www.allelefrequencies.net. Belfast : Tissue Antigens 2003, 61, 403-407.

[21]. Steiner, D. Comparing Different Programs for HLA Haplotype Frequency Estimation in a Controlled Data Environment. New York : In: Tissue Antigens Imune Response Genetics., 2008. ISSN 0001-2815.

[22]. Smith, A. B. Cedric. A note on genetic distance. London : Galton Laboratory, University College London, 1991.

[23]. Ruzzante, Daniel E. A comparison of several measures of genetic distance and population structure with microsatelite data.: NRC Canada, 1998.

[24]. Chattopadhyay, Aparana, Chattopadhyay, Asis Kumar a B-Rao, Chandrika. Bhattacharyya's distance measure as a precursor of genetic distance measures. Dehli : Institute of Genomics and Integrative Biology, 2004.

[25]. Sanghvi, L.D. Comparsion of genetical and morphological methods for a study of biological difference.: American Journal of Physical Antropology, 1953.

[26]. Prevosti, A., a další. Genetic differentation between populations of Drosophila subobscura in Western Mediterrianean area with respect to chromosomal variation. 1983.

[27]. Nei, M. Molecular population genetics and evolution.: North-Holland publishing company, 1975.

[28]. Haplotype Tools Group. WMDA. [Online] [12. květen 2009.] http://www.worldmarrow.org/fileadmin/WorkingGroups_Subcommittees/ITWG/ITWG.pdf.

[29]. Neighbor-joining. Wikipedia. [Online] [13. květen 2009.] http://en.wikipedia.org/wiki/Neighbor-joining.

[30]. Saitou, N. a Nei, M. The neighbor-joining method: a new method for reconstructing phylogenetic trees.: Mol Biol Evol 4 (4): 406-425., 1987.

[31]. Mihaescu, R., Levy, D. a Pachter, L. Why neighbor-joining works. 2007.

[32]. Michener, C.D. a Sokal, R.R. A quantitative approach to a problem of classification: Evolution, 11:490–499., 1957.

[33]. Scalable Vector Graphics. W3C. [Online] [20. únor 2009.] http://www.w3.org/Graphics/SVG/.

[34]. Adobe SVG Viewer. [Online] [20. únor 2009.] http://www.adobe.com/svg/viewer/install/main.html.

[35]. Fylogenetický strom. Wikipedia. [Online] [13. květen 2009.] http://cs.wikipedia.org/wiki/Fylogenetický_strom.

[36]. Wikimedia Commons. [Online] [4. květen 2009.] http://commons.wikimedia.org/wiki/File:The_Ancestors_Tale_Mammals_Phylogenetic_Tree_in_mya.png.

[37]. Centers for Disease Control and Prevention. [Online] [20. únor 2009.] http://www.cdc.gov/epiinfo/shape/Admin00.zip.

[38]. Snyder, John P. Map Projections: A Working Manual.: Geological Survey (U.S.), 1987.

[39]. Furuti, Carlos A. Distortion Pattern. Cartographical Map Projections. [Online] [10. květen 2009.] http://www.progonos.com/furuti/MapProj/Dither/CartProp/Distort/distort.html.

[40]. Haplotype frequencies. National Marrow Donor Program. [Online] [18. leden 2009.] http://bioinformatics.nmdp.org/HLA/Haplotype_Frequencies/index.html.

[41]. Availability of bone marrow transplants between unrelated people. Japan marrow donor program. [Online] [15. květen 2009.] http://www.jmdp.or.jp/documents/file/08_data/hiketsuensya.pdf.

[42]. Křesťanské dějiny českých zemí. NICM. [Online] [9. květen 2009.] http://www.icm.cz/nabozenstvi-historie-nabozenstvi-v-cr.

[43]. Češi jsou Slované jen z půlky, říkají testy DNA. iDNES. [Online] [14. procinec 2008.] http://zpravy.idnes.cz/cesi-jsou-slovane-jen-z-pulky-rikaji-testy-dna-f4z-/vedatech.asp?c=A070613_215321_vedatech_ost.

[44]. Ministerstvo zahraničních věcí ČR. Češi v zahraničí. Czech Republic. [Online] [14. květen 2009.] http://www.czech.cz/cz/kultura/cesi-v-zahranici/cesi-v-zahranici?i=?i=.

[45]. Národnostní složení obyvatelstva. Český statistický úřád. [Online] [8. květen 2009.] http://www.czso.cz/csu/2003edicniplan.nsf/p/4114-03.

[46]. Obyvateľstvo. Štatistický úrad SR. [Online] [13. květen 2009.] http://portal.statistics.sk/files/Sekcie/sek_600/Demografia/Obyvatelstvo/2008/mapy_Narodnosti.pdf.

[47]. Štatistický úrad SR. Zahraničné sťahovanie a cudzinci v Slovenskej republike v roku 2006. Bratislava , 2007.

[48]. Podiely obyvateľov podľa náboženského vyznania v okresoch Slovenskej republiky. Štatistický úrad SR. [Online] [9. květen 2009.] http://portal.statistics.sk/files/Sekcie/sek_600/Demografia/SODB/grafy/sj/09.pdf.

[49]. English country names and code elements. ISO. [Online] [24. únor 2009.] http://www.iso.org/iso/english_country_names_and_code_elements.

[50]. ISO 3166-2. Wikipedia. [Online] [24. únor 2009.] http://en.wikipedia.org/wiki/ISO_3166-2.

[51]. Nei, M. Mathematical Models of Speciation and Genetic Distance. s.l. : Academic Press Inc., 1987.

8.2. Použitý software a knihovny

Microsoft Visual Studio 2008 Team Suite
Shp2KML Converter (http://www.reimers.dk/files/folders/google_maps/entry328.aspx)
Apache HTTP Server 2.2 (http://httpd.apache.org/)
PHP 5.2 (http://www.php.net)
MySQL 5.1 (http://www.mysql.com/)

JQuery 1.3.2 (http://jquery.com/)
JQuery SVG (http://plugins.jquery.com/project/svg)
JQuery Tablesorter 2.0 (tablesorter.com)

JQuery Simplemodal	(www.ericmmartin.com/projects/simplemodal/)
JQuery URL Parser	(http://plugins.jquery.com/project/url_parser)

9. Seznam příloh

9.1. Tištěné přílohy

A) Formát projektového souboru

B) Formát souboru s frekvencemi haplotypů

C) Formát souboru s genetickými vzdálenostmi

D) Formát souboru s geografickými daty

E) Formát definičního souboru mapy

9.2. Elektronické přílohy

Přiložené CD obsahuje následující přílohy v elektronické podobě

\bin	Zkompilovaný výpočetní program a nástroj pro generování mapových podkladů pro platformu Win32
\redist	Instalační verze Microsoft .NET Framework 3.0 Redistributable Package knihovny potřebné pro běh zkompilovaných programů.
\source	Zdrojové kódy
\www	modul do serverové aplikace HLA Explorer
\docs	Elektronická verze textu diplomové práce

Příloha A – formát projektového souboru

Projektový soubor je ve formátu XML a odpovídá následujícího schématu.

```
<?xml version="1.0" encoding="utf-8"?>
<xs:schema attributeFormDefault="unqualified"
elementFormDefault="qualified"
xmlns:xs="http://www.w3.org/2001/XMLSchema">
  <xs:element name="project">
    <xs:complexType>
      <xs:sequence>
        <xs:element maxOccurs="unbounded" name="population">
          <xs:complexType>
            <xs:attribute name="id" type="xs:unsignedByte"
             use="required" />
            <xs:attribute name="region-id" type="xs:string"
             use="required" />
            <xs:attribute name="name" type="xs:string"
             use="required" />
            <xs:attribute name="filename" type="xs:string"
             use="required" />
          </xs:complexType>
        </xs:element>
      </xs:sequence>
    </xs:complexType>
  </xs:element>
</xs:schema>
```

Výpis 2- XML schéma pro projektový soubor

Význam jednotlivých položek je přímočarý. Element `population` definuje jednu populaci a má 4 povinné atributy.

`id` jednoznačný identifikátor populace

`region-id` identifikátor oblasti, tak jak je popsán v kapitole o generování map

`name` název populace

`filename` cesta k souboru s frekvencemi haplotypů pro danou populaci

Ukázka jednoduchého projektového souboru je uvedena ve výpisu 3.

```
<?xml version="1.0" encoding="utf-8"?>
<project>
  <population id="1" region-id="AT" name="Austrian"
   filename="AT.csv" />
  <population id="2" region-id="CZ" name="Czech" filename="CZ.csv" />
  <population id="3" region-id="SK" name="Slovakian"
   filename="SK.csv" />
</project>
```

Výpis 3- Příklad projektového souboru

Příloha B – formát souboru s frekvencemi haplotypů

Soubor je ve formátu CSV (Comma separated value). Každý řádek reprezentuje jeden haplotyp, přičemž jednotlivé sloupce jsou odděleny čárkou.

Sloupce jsou definovány v následujícím pořadí

Název sloupce	Datový typ	Význam
A	integer	HLA-A na sérologické úrovni
B	integer	HLA-B na sérologické úrovni
DR	integer	HLA-DR na sérologické úrovni
frequency	double	frekvence výskytu

Tab. 12 – Přehled sloupců v souboru s frekvencemi haplotypů

Jako oddělovač desetinné části čísla se používá tečka '.'.

První řádek souboru obsahuje hlavičku s názvy sloupců „`A,B,DR,frequency`".

Pro správnou funkci programu nemusí být haplotypy žádným způsobem tříděny. Pro přehlednost při manuální práci se surovými daty je ale vhodné, aby byly haplotypy setříděny sestupně podle frekvence výskytu.

Příklad části datového souboru pro českou populaci je uveden ve výpisu 4.

```
A,B,DR,frequency
1,8,17,0.0566344963371
3,7,15,0.0271844554727
2,7,15,0.0176026311759
2,13,7,0.0159386741007
2,44,4,0.0137398412004
23,44,7,0.0125769621658
3,35,1,0.0120540859118
2,62,4,0.0101982922144
25,18,15,0.00977918878428
2,18,11,0.00942050516556
1,57,7,0.00842983904897
29,44,7,0.00749608726963
24,7,15,0.0073447616666
30,13,7,0.00711597895983
2,57,7,0.00658383861138
2,27,1,0.006495862733
11,35,1,0.00640940246503
```

Výpis 4- Příklad datového souboru s frekvencemi haplotypů

Příloha C – formát souboru s genetickými vzdálenostmi

Soubor je ve formátu CSV (Comma separated value). Každý řádek reprezentuje vzdálenost mezi dvěma populacemi A a B.

Sloupce jsou definovány v následujícím pořadí

Název sloupce	Datový typ	Význam
population-A	integer	ID populace A
population-B	integer	ID populace B
distance	double	genetická vzdálenost

Tab. 13 – Přehled sloupců v souboru s genetickými vzdálenostmi

Jako oddělovač desetinné části čísla se používá tečka.

ID populace odpovídá jejímu jednoznačnému identifikátoru, tak jak je definován v projektovém souboru.

První řádek souboru obsahuje hlavičku s názvy sloupců „`population-A, population-B, distance`“

Pro všechny metody výpočtu, které jsem testoval, platí, že genetická vzdálenost mezi populací A a B je stejná jako mezi B a A. V obecném případě, by to nemusela být pravda, proto se v souboru ukládá jak vzdálenost mezi A a B, tak mezi B a A.

Příklad datového souboru pro tři populace je uveden ve výpisu 5.

```
population-A, population-B, distance
1,2,0.287599191665408
1,3,0.517293754246923
2,1,0.287599191665408
2,3,0.51002840134778
3,1,0.517293754246923
3,2,0.51002840134778
```

Výpis 5- Příklad datového souboru s genetickými vzdálenostmi

Příloha D – formát souboru s geografickými daty

Soubor je ve formátu XML a odpovídá následujícímu schématu.

```
<?xml version="1.0" encoding="utf-8"?>
<xs:schema attributeFormDefault="unqualified"
           elementFormDefault="qualified"
           xmlns:xs="http://www.w3.org/2001/XMLSchema">
  <xs:element name="regions">
    <xs:complexType><xs:sequence>
      <xs:element maxOccurs="unbounded" name="region">
        <xs:complexType>
          <xs:sequence>
            <xs:element name="borders">
              <xs:complexType><xs:sequence>
                <xs:element maxOccurs="unbounded" name="border"
                            type="xs:string" />
              </xs:sequence></xs:complexType>
            </xs:element>
            <xs:element name="regions" />
          </xs:sequence>
          <xs:attribute name="id" type="xs:string" use="required" />
          <xs:attribute name="name" type="xs:string" use="required" />
        </xs:complexType>
      </xs:element>
    </xs:sequence>
  </xs:complexType></xs:element>
</xs:schema>
```

Výpis 6- XML schéma souboru s geografickými daty

Soubor definuje regiony zobrazované na mapě a jejich hranice. Každý region je definován elementem `region`. Jeho povinné atributy `id` a `name` definují jednoznačný identifikátor a jméno regionu.

Pokud je regionem stát, tak je jako ID použije dvoupísmenné označení státu podle ISO 3166-1 [49], v případě že se jedná o vyšší územně-správní celky, použije se dvoupísmenné označení státu a označení územně správního celku podle ISO 3166-2 [50] oddělené pomlčkou. Pro podrobnější členění je možné použít uživatelsky definované ID, jejich formát musí začínat dvoupísmenném označením státu následovanému pomlčkou a uživatelsky definovanou částí. V rámci státu musí být uživatelsky definovaná část ID unikátní.

ID	Název regionu
CZ	Česká republika
CZ-LI	CŘ, Liberecký kraj
AT	Rakousko
AT-7	Rakousko-Tyrolsko
CZ-TRUTNOV	ČR, město Trutnov

Tab. 14 – Příklady validních ID regionů spolu s jejich názvy

Hranice regionu si můžeme přestavit jako polygon (nebo několik polygonů). Tyto polygony jsou definovány v elementech `border` a jsou uloženy v následujícím formátu

`Longitude:Latitude|Longitude:Latitude|…`

kde každá dvojice `Longitude:Latitude` určuje jeden vrchol polygonu, přičemž `Longitude` představuje zeměpisnou délku a `Latitude` zeměpisnou šířku. Severní zeměpisná šířka je představována kladným číslem, jižní záporným číslem. Západní zeměpisná délka je představována záporným číslem, východní kladným číslem.

Regiony mohou být definovány ve stromové struktuře:

CZ
 CZ-LI
 CZ-KA
 CZ-KR
  CZ-TRUTNOV

Ukázka souboru s geografickými daty je uvedena ve výpisu 7.

```
<?xml version="1.0" encoding="utf-8"?>
<regions>
  <region id="SK" name="Slovakia">
    <borders>
  <border>
19.0012702941895:48.0689544677734|18.9897994995117:48.0671005249023|18
.9656753540039:48.0597991943359|18.9072856903076:48.057975769043|18.84
53369140625:48.049129486084|18.8273658752441:48.0361251831055|18.77166
36657715:47.9655532836914|18.759651184082:47.9147186279297|18.78350448
6084:47.8720779418945|18.8199996948242:47.8555526733398|18.85472106933
59:47.8316650390625
</border>
    </borders>
    <regions />
  </region>
</regions>
```

Výpis 7- Část souboru s geografickými daty

Příloha E – formát definičního souboru mapy

Definiční soubor mapy je ve formátu XML a odpovídá následujícímu schématu.

```
<?xml version="1.0" encoding="utf-8"?>
<xs:schema attributeFormDefault="unqualified"
           elementFormDefault="qualified"
           xmlns:xs="http://www.w3.org/2001/XMLSchema">
  <xs:element name="map">
    <xs:complexType><xs:sequence>
      <xs:element name="viewport"><xs:complexType>
        <xs:attribute name="min-longitude" type="xs:short"
                      use="required" />
        <xs:attribute name="min-latitude" type="xs:short"
                      use="required" />
        <xs:attribute name="max-longitude" type="xs:short"
                      use="required" />
        <xs:attribute name="max-latitude" type="xs:short"
                      use="required" />
      </xs:complexType></xs:element>
      <xs:element name="output-filename" type="xs:string" />
      <xs:element name="map-resolution" type="xs:unsignedShort" />
      <xs:element name="map-name" type="xs:string" />
      <xs:element name="regions" type="xs:string" />
    </xs:sequence></xs:complexType>
  </xs:element>
</xs:schema>
```

Výpis 8- XML schéma definičního souboru mapy

Element `viewport` určuje oblast zobrazenou na mapě – pro údaje o zeměpicné délce a šířce platí stejná pravidla jako v souboru s geografickými daty - severní zeměpisná šířka je představována kladným číslem, jižní záporným číslem. Západní zeměpisná délka je představována záporným číslem, východní kladným číslem.

Měřítko mapy je určeno elementem `map-resolution`. Jeho hodnota představuje šířku vygenerované mapy v pixelech. Výška mapy je automaticky dopočítána podle zobranované oblasti.

Element `regions` určuje, které oblasti budou na mapě zobrazeny. Obsahem elementu může být středníkem oddělený seznam ID regionů nebo *, což znamená všechny regiony.

Elementy `map-name` a `output-filename` určují název mapy resp. jméno výstupního souboru.

Příklad definičního souboru pro mapu světa použitou v aplikaci HLA Explorer je uveren ve výpisu 9.

```
<?xml version="1.0" encoding="utf-8"?>
<map>
  <viewport min-longitude="-180" min-latitude="-73"
            max-longitude ="180" max-latitude="73" />
  <output-filename>world.svg</output-filename>
  <map-resolution>1140</map-resolution>
  <map-name>World</map-name>
  <regions>*</regions>
</map>
```

Výpis 9- Příklad definičního souboru mapy