

CZECH TECHNICAL UNIVERSITY IN PRAGUE
FACULTY OF ELECTRICAL ENGINEERING



DIPLOMA THESIS

Geographical Analysis of Databases of Stem
Cell Donor Registries

Praha, 2011

Author: Jiří Těhník

ZADÁNÍ DIPLOMOVÉ PRÁCE

Student: Bc. Jiří Těhník

Studijní program: Elektrotechnika a informatika (magisterský), strukturovaný

Obor: Kybernetika a měření, blok KM2 – Umělá inteligence

Název tématu: Geografická analýza dat registrů dárců krvetvorných buněk

Pokyny pro vypracování:

Registr dárců krvetvorných buněk eviduje potenciální dárce krvetvorných buněk (kostní dřeň, kmenové buňky z periferní krve nebo pupečnicková krev) a vyhledává HLA shodné dárce pro české i zahraniční pacienty, kteří potřebují transplantaci. Cílem této práce je analýza a srovnání databází registrů dárců krvetvorných buněk odlišných populací, vizualizace geografické distribuce HLA-genetických faktorů dané populace uvnitř státu a vytvoření nástroje pro odhad původu konkrétního člověka, tedy odkud pochází jeho předci.

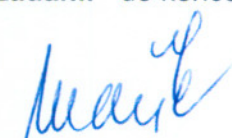
1. Nastudujte existující principy a postupy určování HLA genetické shody dvou jedinců.
2. Nastudujte problematiku registrů dárců krvetvorných buněk, cíle jejich činnosti.
3. Navrhněte metody pro vizualizaci geografické distribuce HLA v subregionech daného státu.
4. Navrhněte metody pro vizualizaci geografického rozložení geneticky příbuzných dárců určitého jedince, a to jak na úrovni daného státu, tak na mezinárodní úrovni.
5. Tyto metody aplikujte na databázi Českého (Finského, Jihoafrického, Argentinského) registru dárců krvetvorných buněk.

Seznam odborné literatury:


- [1] Steiner, D.: Hledání nepříbuzenských dárců kostní dřeně, diplomová práce, FEL ČVUT, 2007
- [2] Bone Marrow Donors Worldwide, User Guide
http://www.bmdw.org/fileadmin/templates/main/Downloads/BMDW_User_Guide.pdf
- [3] Hurley, C.K. et al.: Hematopoietic stem cell donor registry strategies for assigning search determinants and matching relationships. Bone Marrow Transplantation (2004) 33, 443-450.

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

Platnost zadání: do konce letního semestru 2010/2011


prof. Ing. Vladimír Mařík, CSc.
vedoucí katedry




doc. Ing. Boris Šimák, CSc.
děkan

V Praze dne 1. 2. 2010

DIPLOMA THESIS ASSIGNMENT

Student: Bc. Jiří Těhník

Study programme: Electrical Engineering and Information Technology

Specialisation: Cybernetics and Measurement – Artificial Intelligence

Title of Diploma Thesis: Geographical Analysis of Databases of Stem Cell Donor Registries

Guidelines:

A stem cell donor registry is an institution that maintains database of potential donors of haematopoietic stem cells and searches HLA identical donors for both national and foreign patients that need stem cell transplantation. The goal of this work is analysis and comparison of databases of different registries, visualization of the geographical distribution of HLA factors of the population inside the country and creation of a tool for estimation of family origin of a human.

1. Study existing principles of HLA typing and HLA matching of two individuals.
2. Study operation of the haematopoietic stem cells donor registry.
3. Propose methods for visualization of geographical distribution of HLA in regions of a country.
4. Propose methods for visualization of geographical distribution of relatives of an individual, on both national and international level.
5. Apply suggested methods to the Czech (Finnish, South African, Argentinean) registry database.

Bibliography/Sources:


- [1] Steiner, D.: Hledání nepříbuzenských dárců kostní dřeně. Diplomová práce, FEL ČVUT, 2007.
- [2] Bone Marrow Donors Worldwide, User Guide;
http://www.bmdw.org/fileadmin/templates/main/Downloads/BMDW_User_Guide.pdf
- [3] Hurley, C.K. et al.: Hematopoietic Stem Cell Donor Registry Strategies for Assigning Search Determinants and Matching Relationships. Bone Marrow Transplantation, Vol. 33, Page 443-450, 2004.

Diploma Thesis Supervisor: doc. Ing. Lenka Lhotská, CSc.

Valid until: the end of the summer semester of academic year 2010/2011


prof. Ing. Vladimír Mařík, CSc.
Head of Department





doc. Ing. Boris Šimák, CSc.
Head

Prague, February 1, 2010

Prohlášení

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

V Praze dne 4.1.2011


.....
podpis

-

Poděkování

Děkuji především panu Ing. Davidu Steinerovi za pravidelné a přínosné konzultace a za zprostředkování anonymizovaných dat z národních registrů dárců krevtovorných buněk, bez kterých by tato práce nemohla vzniknout.

Abstrakt

Tato práce se zabývá analýzou národních registrů dárců krvetvorných buněk se zaměřením na geografickou strukturu. Registry obsahují anonymizované údaje o HLA fenotypu dárců a další atributy. Práce se zaměřuje na porovnání genetické podobnosti jedince se skupinou jedinců v regionu. Jedním z atributů u každého jedince je poštovní směrovací číslo, podle kterého jsou jedinci klasifikováni do regionů a poté je podle navržené metriky měřena vzájemná genetická vzdálenost. Vzniklé matice jsou pak vizualizovány pomocí navržené webové aplikace *GeoRelatives* v podobě přehledných geografických map s barevně odlišenými regiony. Navržený způsob předzpracování dat registru umožňuje zobrazovat výsledky na dotazy v reálném čase. Pomocí experimentů nad českým, finským a švédským registrem je dokázána hypotéza stanovená v úvodu a to že jedinci žijící uvnitř regionů jsou si navzájem HLA geneticky bližší než ve srovnání s jedinci z ostatních regionů. U každého ze zkoumaných registrů byla stanovena míra HLA genetické diverzity. Tento ukazatel přináší informaci o interregionální genetické diverzitě národního registru a také o míře důvěryhodnosti výsledků. Specialistům pracujícím s registrem umožňuje tato aplikace zobrazit k libovolnému HLA fenotypu distribuci genetických vzdáleností přes všechny regiony a vytvořit seznam regionů seřazených od nejbližšího. Při hledání vhodného dárce je tedy pak možné se zaměřit přímo na regiony s menší genetickou vzdáleností od potřebného pacienta a ušetřit čas i peníze celoplošným hledáním.

Abstract

This work analyses anonymised databases of unrelated stem cell donor registries with a focus on its geographical structure. It examines genetic distance of individuals living in geopolitical regions in particular countries. Dataset contains zip codes according to which the individuals are classified to subregions and then they are measured all to all using an own-suggested high resolution metric. Matrices of genetic distances are visualised by the means of suggested web application in the form of well-arranged geographical maps with colour defined regions. The way of data preprocessing enables to show the results in real time. On basis of the experiments with Czech, Finnish and Swedish national registries I prove the hypothesis introduced at the beginning of the thesis which states that individuals living in a region are mutually HLA closer to each other than to individuals from other regions. In each of the examined national registry I introduce the factor of genetic diversity as a percentual ratio between an average genetic distance of individuals in and throughout the regions. This indicator provides information about regional genetic diversity of the examined country and gives us as well information about the credibility of the outcomes. Web application *GeoRelatives* enables to employees of a registry to see geographical distribution of genetic distances for the given HLA phenotype throughout all the regions and also list of regions ordered from the nearest one. Managers of national registries when searching for a suitable donor can by the means of this application focus on the region with the nearest genetical distance from the patient and thus save money and time by full-area searching.

Contents

List of figures	xii
List of tables	xiii
1 Introduction	1
1.1 Motivation	1
1.2 Hypothesis	1
1.3 Plan and structure	3
2 Research	5
2.1 Unrelated Stem Cells Donor Registries	5
2.1.1 Czech Bone Marrow Donor Registry	5
2.1.2 BMDW	6
2.2 Current state of national stem cell registries worldwide	7
2.2.1 Analysis of HLA geographical structure	9
2.3 Genetic distance measuring on HLA	9
2.3.1 Allele Frequencies	10
2.3.2 Estimating Haplotype Frequencies	10
2.3.3 HLA diversity in United Kingdom	10
2.3.4 HLA search for donors	10
2.4 Allele variation in HLA	11
3 Teoretical part	13
3.1 Antigen equivalents	13
3.2 Genetic distance function	13
3.3 Distance matrix methods	14
3.4 Neighbour joining	15
3.5 Unweighted pair group method with arithmetic mean	15

3.6	Construction of phylogenetic tree	16
3.7	Geographical vizualization	17
3.7.1	SVG maps	17
3.7.2	GIS	17
3.7.3	Google maps API	17
3.7.4	GPS	18
3.7.5	Geocoding	18
3.8	Implementation tools	18
3.8.1	PHP	18
3.8.2	JavaScript	19
3.8.3	Googlemaps overlays	19
3.8.4	CSV files	19
4	Designing of HLA metric	21
4.1	Definition of HLA resolution	21
4.2	Low resolution antigens mismatch metric	22
4.2.1	Mismatch counter	23
4.2.2	Unphased chromosomes problem solving	24
4.2.3	Handling with missing values	25
4.3	High resolution metric	25
4.3.1	Principle schema of the metric	25
4.3.2	Definition of the resolution DNA-4	26
4.3.3	Probability of possible allele match	26
4.3.4	Downgrading higher resolutions	27
4.3.5	Decomposition of alleles with lower resolution	27
4.3.6	Allele frequency assignment and probability normalization	27
4.3.7	Probabilistic intersection of weighted sets	28
4.3.8	Phasing the gens	29
4.3.9	Combination of particulat loci results	30
5	Implementation	31
5.1	Schema of the project	31
5.2	Dataset adjustment	32
5.2.1	Raw HLA data overview	32
5.2.2	Specification of problems	33

5.2.3	Redundancy elimination	34
5.2.4	Homozigode identification	34
5.2.5	Plotting of regional maps	35
5.2.6	Individuals localization	35
5.2.7	Zip code classification	35
5.3	Preprocessing	36
5.3.1	Reduction of the string	36
5.3.2	Indexing and locus distance preprocessing	37
5.3.3	Getting genetic distances from precomputed data	37
5.4	Definition of the Country	38
5.4.1	National registry import	38
5.4.2	Dataset structure definition	39
5.4.3	Names and zip code intervals of regions	40
5.4.4	GPS borders of subregions	41
5.4.5	Preprocessed allele frequencies	41
5.5	User interface	42
5.5.1	Maps of genetic relativeness	42
5.5.2	Distance function check	42
5.5.3	Distance matrix	44
5.5.4	Q matrix	45
5.6	Applicability	45
5.6.1	Direct applicability	45
5.6.2	Libraries and functions	45
5.6.3	Iframe objects	46
5.7	Expandability	46
6	Experiments	47
6.1	Mutual HLA genetic similarity of geopolitical regions	47
6.1.1	Distance matrix computation	47
6.2	Visualization of the triangular distance matrix	48
6.2.1	Genetic diversity of Czech Republic regions	50
6.2.2	Genetic diversity of regions in Finland	51
6.2.3	Genetic diversity of Swedish regions	52
6.3	Fast method for comparison of metrics	52
6.4	Interesting outcomes	56

7 Conclusion	57
7.1 Results summarization	57
7.2 Corrolary	58
Literature	61
A <i>GeoRelatives</i> user manual	I
A.1 Layout	I
A.2 Menu	I
A.2.1 Mode	I
A.2.2 Country	III
A.2.3 Metric	IV
B Attached CD	V
C Abbreviations and terms	VII

List of Figures

2.1	Number of donors in CBMD (source [9])	6
2.2	Long term linear trend of increasing donors worldwide (source [4]	7
2.3	Number of HLA donors worldwide (22.11.2010, source [4])	8
2.4	Number of different HLA alleles and antigens (source [4])	9
3.1	Example of phylogenetic tree	16
3.2	Example of Googlemaps API overlay	19
4.1	DNA allele to serology antigen transformation [13]	22
4.2	Principle schema of high resolution metric	26
4.3	Higher resolution downgrading	27
4.4	Decomposition of lower resolutions	27
4.5	Alleles and their frequencies assignment	28
4.6	Probabilities after normalization	28
4.7	Example of intersection of weighted sets	28
4.8	Unphased phenotypes elimination	29
5.1	Schema of the project	32
5.2	Classification of individuals into the geopolitical regions	36
5.3	The diagram of getting genetic distances from precomputed data	38
5.4	Map of genetic distance distribution in regions of Czech Republic	42
5.5	Two phenotype input form	43
5.6	Reduction check	43
5.7	Phasing process of unphased chromosomes	43
5.8	Example of breakdown alleles probabilities distribution	44
5.9	Example of distance matrix	44
5.10	Example of Q-matrix	45
6.1	Distance matrix with diagonal element and related neighbours	49

6.2	Mutual HLA genetic similarity between Czech regions	49
6.3	HLA genetic diversity of regions of The Czech Republic	50
6.4	HLA genetic diversity of regions in Finland	52
6.5	HLA genetic diversity of regions in Sweden	53
6.6	Iterations of HLA genetic diversity using low resolution metric	54
6.7	Iterations of HLA genetic diversity using high resolution metric	55
6.8	Searching for suitable donor	56
A.1	GeoRelatives visual layout and control panels	II
B.1	directory tree	V

List of Tables

4.1	List of different levels of resolution	22
4.2	Low resolution downgraded antigens	23
4.3	Low resolution examples	24
5.1	Examples of HLA dataset	33
5.2	Example of reduced dataset	34
5.3	Homozigod cloning	35
5.4	Example of anonymized national stem cell registry file	39
5.5	Example of file with zip code intervals	40
5.6	Example of file with zipcode intervals	40
5.7	Example of GPS regions border definition	41
5.8	Allele frequencies	41
6.1	Regions and zip-code intervals in Czech Republic	51
6.2	Regions and zip-code intervals in Finland	53
6.3	Regions and zip-code intervals in Sweden	54
7.1	Comparison of low and high metric results	57

Chapter 1

Introduction

1.1 Motivation

National registry of stem cell donors [4] is the database of specific genetic information which was established for searching suitable unrelated donors for patients which need to transplant organs, bone marrow or blood haematopoietic stem cells. The database collects such set of HLA gens which ensure the best transplantant immunology reaction. Apart fom genetical information there are also additional attributes, like a zip code by each record. It can be used for geographical localization of individuals and classification to the particular geographical region. This thesis introduces the new approach of analyzing stem cells registries focused on the geographical distribution of genetic similarity or distances from the particular geopolitical province. Also it brings the unique extensible web tool which allows registry experts to work with the actual database and provide geographical view of their data.

1.2 Hypothesis

The objective of this work is to analyse the given genetic data in terms of geographical distribution. Let us introduce hypothesis which connects both, genetics and geographics.

General hypothesis of geographical-genetic relativeness

Hypothesis 1.1: *Individuals living geographically closer to each other are also genetically closer to each other.*

This hypothesis is not valid in general, but it is supposed that it holds true on average. Let us introduce countries as a super-regions and their geopolitical provinces as sub-regions and adjust previous hypothesis:

Hypothesis 1.2: *Individuals living in the particular sub-region are on average genetically more similar with each other than with the individuals living in different regions.*

To be able to prove the introduced hypothesis, there is need to have sufficiently representative and variant genetic information about all the individuals living in the super region, attributes for their classification into the sub-regions as well as suitable metric for measuring genetic distances between any two of them.

One of the main aims of this thesis is to come up with the suitable distance function and implement it in accordance with the assignment.

We look for such a metric which meets:

1. Introduced geo-genetic hypothesis
2. Natural meaning of genetic similarity and heritage
3. HLA immunological experts knowledges
4. Invariancy to the size of regions
5. Invariancy to the size of sample
6. Invariancy to the quality of data
7. Significantly better results than random metric
8. Distance between any two individuals has to be comparable with any distance of other two individuals

If such a metric is found, then:

- Each individual can be assigned to their most similar region
- Regions can be genetically comparable with each other like averaged individuals

- Regions can be sorted and compared with regard to any individual
- Results of such a measurement can be used to plot the maps of genetic distance distribution

1.3 Plan and structure

This work introduces the reader the whole process of the realization of the project which was based on the following initial plan:

1. Learn how HLA genetic distance is measured for the purpose of donor-patient search (chapter 2).
2. Suggest the way of individuals classification to geographical regions
3. Suggest the way of measuring clusters of individuals and their comparison
4. Suggest metric and compare results with the current metric using suitable experiments
5. Create extendable vizualization tool

The points of the plan more or less follow the chapters in this document. There are discussed suggested approaches, the problems which occurred during the work and the way how they were solved.

Chapter 2

Research

The objective of this chapter is to search in articles, books and websites for similar works and to discuss the current state and latest knowledge of all main topics we intend to work on.

2.1 Unrelated Stem Cells Donor Registries

Only one out of three patients will find a suitable donor within their family, the rest must search for an unrelated donor. Therefore volunteer stem cells donor registries have been established in many countries. In the Czech republic there are two registries: Czech Bone Marrow Donor Registry in Prague [9] and Czech National Marrow Donor Registry in Plzeň [6] and one public cord blood bank [5]. About half of Czech patients who do not have suitable donor within their family, will find suitable donor in Czech registries [8]. The rest must search for a donor in foreign registries.

2.1.1 Czech Bone Marrow Donor Registry

The Czech Bone Marrow Donor Registry in Prague (CBMD) [9], organized under Department of Immunology, is a stem cell donor registry. The Department of Immunology is located at the Institute for Clinical and Experimental Medicine (IKEM) that is a major research institute controlled by the Czech Ministry of Health. CBMD was established in 1991, as the first in Central Europe to join an international registry BMDW.

At the end of December 2009 [9] - annual report, 23 332 donors were registered in

CBMD (see figure 2.1).

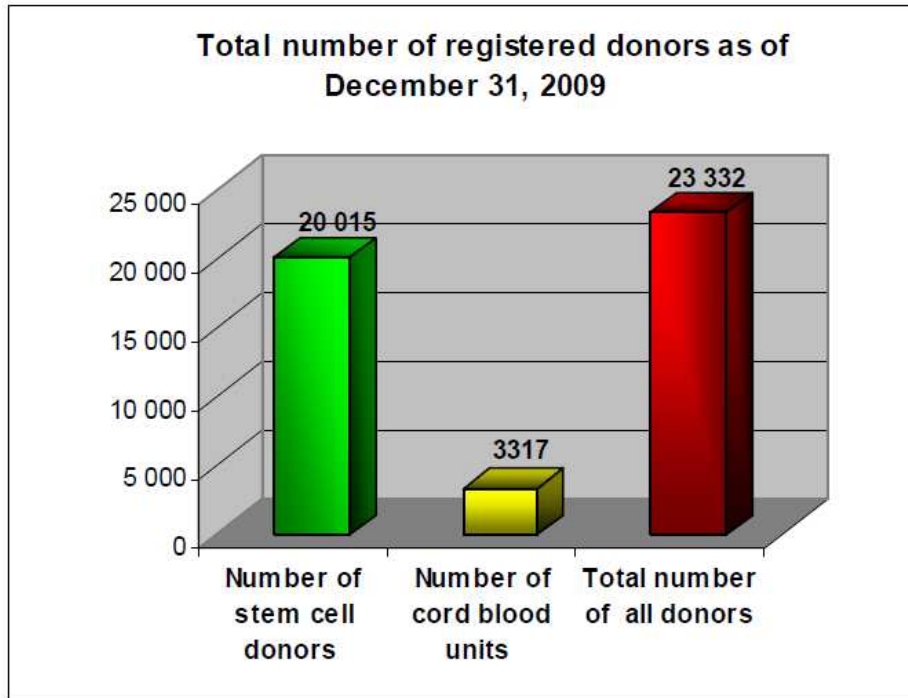


Figure 2.1: Number of donors in CBMD (source [9])

Main goal of the CBMD is to search for unrelated bone marrow donors for patients for whom no donor can be identified within their family, with the best possible HLA match between donor and patient. The CBMD Registry is responsible for maintaining a database of HLA typed volunteer donors, for performing national and international searches in the files of international registers and coordinating the communication between participating centers.

2.1.2 BMDW

BMDW [4] is a voluntary collaborative effort of stem cell donor registries and cord blood banks whose goal is to provide centralised information on the HLA phenotypes and other relevant data of unrelated stem cell donors and cord blood units and to make this information easily accessible to the physicians of patients in need of a hematopoietic stem cell transplant. The original goal to collect the HLA phenotypes of volunteer stem cell donors and cord blood units, and to co-ordinate their world-wide distribution remain our primary goals. But new initiatives have been added:

2.2. CURRENT STATE OF NATIONAL STEM CELL REGISTRIES WORLDWIDE⁷

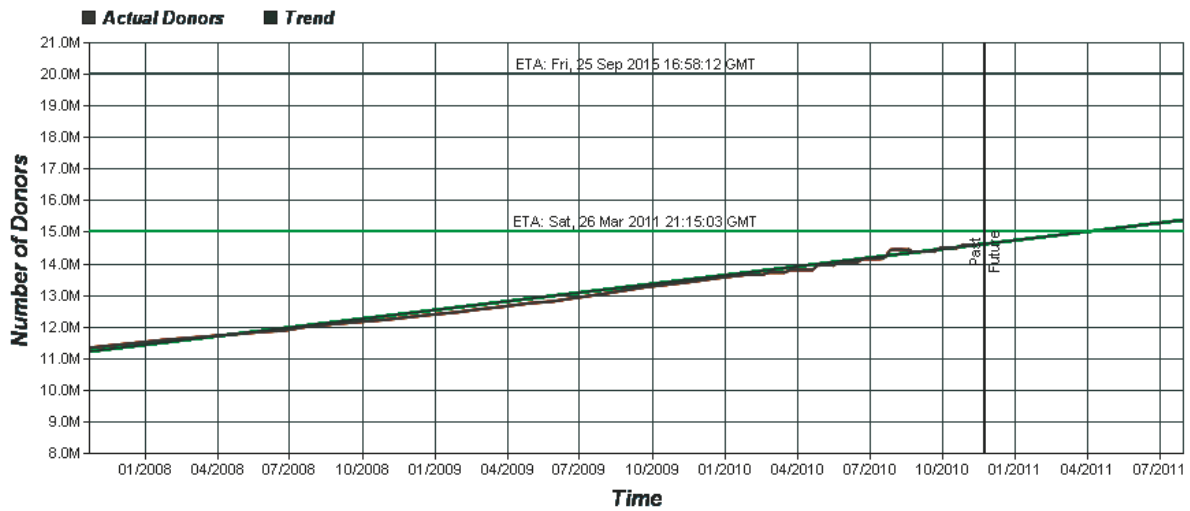


Figure 2.2: Long term linear trend of increasing donors worldwide (source [4])

- To maximise the chance of finding a stem cell donor or cord blood unit by providing access to all stem cell donors and cord blood units available in the world.
- To minimise the effort required for stem cell donor or cord blood unit searches: only registries with potential stem cell donors or cord blood units need to be contacted
- To facilitate search advice requests via the Internet

2.2 Current state of national stem cell registries worldwide

The current (22.11.2010) number of donors in the BMDW database is: 14631714

There are 748 users from 471 organisations authorized to access the on-line BMDW services. Table 2.4 summarize the number of named alleles for each locus. Loci A,B and DRB1 are the most important loci for imunological match. More in section 2.4.

Registry	Registry Code	Total	ABDR	%ABDR Typed	DNA Class I	DNA Class II	Date Last File
Argentina	AR	28,806	24,540	85.2	28,806	24,429	2010-09-24
Argentina CORD	ARCB	1,094	1,093	99.9	1,094	1,092	2010-03-19
Armenia	AM	15,631	15,629	100.0	15,631	15,630	2010-10-26
Australia and New Zealand	AUS	182,643	136,696	74.8	58,359	125,339	2010-10-26
Australia CORD #	AUCB	23,089	23,089	100.0	17,737	23,089	2010-10-26
Austria	A	61,303	26,268	42.8	17,601	17,126	2010-10-22
Austria CORD	ACB	1,079	1,079	100.0	1,079	1,079	2010-10-22
Belgium	B	49,419	40,401	81.8	7,132	26,497	2010-10-26
Belgium CORD ##	BCB	7,567	7,567	100.0	5,244	7,567	2010-10-26
Belgium-Leuven CORD	LVCB	7,057	7,057	100.0	7,048	7,048	2010-10-27
Bulgaria	BG	431	422	97.9	323	422	2010-10-15
Canada-OneMatch	CND	265,948	200,408	75.4	111,387	191,462	2010-10-23
China-Sunshine	CN1	2,368	2,215	93.5	2,368	2,215	2010-07-13
Croatia	HR	17,315	15,926	92.0	15,712	15,710	2010-10-22
Croatia CORD	HRCB	969	969	100.0	969	969	2010-10-22
Cyprus	CY	6,061	856	14.1	3,425	373	2010-10-25
Cyprus BMDR	CY2	109,904	61,712	56.2	87,200	49,368	2010-11-17
Czechia CORD	CSCB	3,557	3,557	100.0	3,439	3,555	2010-10-22
Czechia-Central BMDR	CS2	35,237	30,152	85.6	6,078	30,099	2010-10-25
Czechia-Czech BMDR	CS	20,182	8,519	42.2	1,422	8,461	2010-10-22
Denmark	DK	24,641	22,973	93.2	13,133	22,972	2010-10-19
Denmark-BMDC	DK2	12,817	12,038	93.9	0	11,807	2010-09-25
Duesseldorf CORD #	DUCB	15,460	15,460	100.0	11,365	15,455	2010-10-11
Finland	FI	19,892	19,377	97.4	3,026	4,550	2010-10-26
Finland CORD	FICB	3,086	3,086	100.0	1,297	3,086	2010-09-29
France	F	185,785	169,643	91.3	93,402	157,729	2010-11-15
France CORD ##	FCB	10,292	10,292	100.0	4,828	9,779	2010-11-15
Germany	D	3,992,853	2,867,407	71.8	2,541,487	2,785,066	2010-10-25

Figure 2.3: Number of HLA donors worldwide (22.11.2010, source [4])

HLA Class I									
Gene	A	B	C	E	F	G			
Alleles	1,381	1,927	960	9	21	46			
Proteins	1,024	1,505	709	3	4	15			
Nulls	72	61	23	0	0	2			
HLA Class II - DRB Alleles									
Gene	DRB1	DRB2	DRB3	DRB4	DRB5	DRB6	DRB7	DRB8	DRB9
Alleles	831	1	52	14	19	3	2	1	1
Proteins	639	0	42	8	16	0	0	0	0
Nulls	8	0	0	3	2	0	0	0	0

Figure 2.4: Number of different HLA alleles and antigens (source [4])

2.2.1 Analysis of HLA geographical structure

To the best of our knowledge, there is only project *www.allelefreuencies.net* [18]. The main purpose of the website is to provide one central source, freely available to all, for the storage of allele frequencies from different polymorphic areas in the HUMAN genome. Users can contribute the results of their work into one common database, and can perform database searches on information already available. This approach is suitable for scientific purpose and is quite different from the way we are going to present data from stem cell registries in this work.

2.3 Genetic distance measuring on HLA

There are existing methods to compute genetic distance between two HLA phenotypes. Mr. Kalbrt [21] showed in his thesis why Cavali-Sforze method is optimal for genetic distance computation in HLA. Personal communication with David Steiner [30] and parts about search strategies in his thesis was used for designing algorithms suitable for the purpose of geographical analysis of the given dataset.

2.3.1 Allele Frequencies

Project *Bioinformatics* [1] provides database of allele frequencies estimated also for the european individuals, which can be used in case of lower resolution records decomposition (section 4.3.5). It is recommended to change the set of frequencies for countries out of europe. Different philosophy of handling with allele frequencies is described in article [25].

2.3.2 Estimating Haplotype Frequencies

A commonly used tool in disease association studies is the search for discrepancies between the haplotype distribution in the case and control populations. In order to find this discrepancy, the haplotypes frequency in each of the populations is estimated from the genotypes [19]. If the haplotypes are not phased, finding the maximum value of the likelihood function is NP-hard. The given data are not phased, thus it would be unreasonably difficult and time demanding to use this method for the purpose of this work.

2.3.3 HLA diversity in United Kingdom

The aim of [23] large study of Anthony Nolan Trust HLA registry was to understand and increase HLA diversity in the registry, because during past 30 years the methods for HLA typing and the level of resolution obtained has changed.

In this thesis we improved this method using high resolution typing data instead of low resolution which is used in [23]

2.3.4 HLA search for donors

Locus match grades

To design a suitable metric we started with [30] where we found so called *match grades* used in the process of donor-patient search in HLA registry searching platforms.

Locus *match grade* is a level of match between two individuals in one locus. The algorithm uses seven hierarchical levels of locus match grades covering all cases that can happen:

- **Uncountable** - Unable to compute the value, i.e. DRB1 locus match grade of AB typed donors
- **Allele Match** - Exact match at HR level
- **Allele/Antigen Potential Match** - Potential allele match
- **Broad Matched** - Potential match
- **Allele Mismatch** - The same broad group and allele group, but different allele
- **Split antigen/Allele Mismatch** - The same broad group, but different serology antigen
- **Antigen (Broad) Mismatch** - Different broad group

2.4 Allele variation in HLA

There are many studies that evaluate the role of HLA matching in outcome. We have chosen to focus on large, contemporary studies from 3 groups [15] that have evaluated most of the HLA loci by using DNA testing to resolve alleles.

1. Japanese Marrow Donor Program [27]
2. Fred Hutchinson Cancer Research Center [26]
3. NMDP: an abstract on this study by Flomenberg et al. [17]

These studies prove that there exist three most important loci for immunological match in HLA region:

- HLA-A
- HLA-B
- HLA-DRB1

This proof is based on statistical outcomes of observation of patients after stem cell transplantation. Patients and their unrelated donors were both HLA typed. It was clearly found out that there is much higher probability of the patients' survival in case

of matching alleles on loci A, B and DRB1 with donor. Other loci in HLA region (C,DRB2,DQ1,DQ2) contribute to the ratio of human patients' survival much less [15]. Fortunately, there is also a high level of genetic diversity at the three most important loci we are going to work with.

Chapter 3

Theoretical part

3.1 Antigen equivalents

Antigen equivalents are used in this work to transform DNA records to potential serology antigens. The 2008 report of the human leukocyte antigen [28] data dictionary presents serologic equivalents of HLA-A, -B, -C, -DRB1, -DRB3, -DRB4, -DRB5, and -DQB1 alleles. The dictionary is an update of the one published in 2004. The data summarize equivalents obtained by the World Health Organization Nomenclature Committee for Factors of the HLA System, the International Cell Exchange, UCLA, the National Marrow Donor Program, recent publications, and individual laboratories. The 2008 edition includes information on 832 new alleles (685 class I and 147 class II) and updated information on 766 previously listed alleles (577 class I and 189 class II). The tables list the alleles with remarks on the serologic patterns and the equivalents. The serological equivalents are listed as expert assigned types, and the data are useful for identifying potential stem cell donors who were typed by either serology or DNA-based methods. The tables with HLA equivalents are available as a searchable form on the IMGT/HLA database Web site (<http://www.ebi.ac.uk/imgt/hla/dictionary.html>).

3.2 Genetic distance function

When trying to reconstruct a deep family tree, some measure of genetic distance is required. In an ideal world, we could directly count the total number of mutations at which two chromosomes differ. This is not be quite as straightforward as it sounds, as it makes

assumptions about the underlying mutation process. I work with two main approaches to determine final genetic distance between two phenotypes: *Antigen mismatch counter* and the *Probability of allele match*. Each of them is described in following paragraphs.

Antigen mismatch counter

This metric takes two phenotypes, separate it into the particular loci. If necessary it assigns all loci to antigens using antigen equivalents (see chapter 3.1) and finally counts the number of mismatches with phasing approach. Implementation is described in chapter 4.

Probability of allele match

This metric work with preprocessed allele frequencies and it express the probability of allele match at the locus. Finally probabilities of loci contributions are merge together either with averaging or using euklidian distance. It provides quite accurate genetic distance function between two phenotypes. Implementation is described in chapter 4.

3.3 Distance matrix methods

Distance matrix methods [20] of phylogenetic analysis explicitly rely on a measure of genetic distance between the phenotypes being classified, and therefore they require an multiple sequence alignment as an input. Distance is often defined as the fraction of mismatches at aligned positions, with gaps either ignored or counted as mismatches. Distance methods attempt to construct an all to all matrix from the particular distance between each phenotype pair. From this is constructed a phylogenetic tree that places closely related sequences under the same interior node and whose branch lengths closely reproduce the observed distances between sequences. Distance matrix methods may produce either rooted or unrooted trees, depending on the algorithm used to calculate them. The main disadvantage of distance matrix methods is their inability to efficiently use information about local high variation regions that appear across multiple subtrees.

3.4 Neighbour joining

Neighbor-joining is a bottom-up clustering method [29] used for the construction of phylogenetic trees. Usually used for trees based on DNA or protein sequence data, but we will use HLA phenotypes in this work. The algorithm requires knowledge of the distance between each pair of taxa (e.g., groups or regions) in the tree. Neighbor-joining is an iterative algorithm. Each iteration consists of the following steps:

1. Based on the current distance matrix calculate the matrix Q
2. Find the pair of taxa in Q with the lowest value. Create a node on the tree that joins these two taxa
3. Calculate the distance of each of the taxa in the pair to this new node
4. Calculate the distance of all taxa outside of this pair to the new node
5. Start the algorithm again, considering the pair of joined neighbors as a single taxon and using the distances calculated in the previous step

Based on a distance matrix relating the r taxa, calculate Q as follows:

$$Q(i, j) = (r - 2)d(i, j) - \sum_{k=1}^r d(i, k) - \sum_{k=1}^r d(j, k) \quad (3.1)$$

where $d(i, j)$ is the distance between taxa i and j .

3.5 Unweighted pair group method with arithmetic mean

UPGMA is a agglomerative or hierarchical clustering method used in bioinformatics for the creation of phenetic phylogenetic trees (phenograms). UPGMA assumes a constant rate of evolution (molecular clock hypothesis), and is not a well-regarded method for inferring phylogenetic trees unless this assumption has been tested and justified for the data set being used. UPGMA was initially designed for use in protein electrophoresis studies, but is currently most often used to produce guide trees for more sophisticated phylogenetic reconstruction algorithms.

The algorithm examines the structure present in a pairwise distance matrix (or a similarity matrix) to then construct a rooted tree (dendrogram).

At each step, the nearest two clusters are combined into a higher-level cluster. The distance between any two clusters A and B is taken to be the average of all distances between pairs of objects "x" in A and "y" in B, that is, the mean distance between elements of each cluster:

$$\frac{1}{|A| \cdot |B|} \sum_{x \in A} \sum_{y \in B} d(x, y) \quad (3.2)$$

3.6 Construction of phylogenetic tree

Phylogenetic trees among a nontrivial number of input sequences are constructed using computational phylogenetics methods. Distance-matrix methods such as neighbor-joining or UPGMA, which calculate genetic distance from multiple sequence alignments, are simplest to implement, but do not invoke an evolutionary model. Many sequence alignment methods such as ClustalW also create trees by using the simpler algorithms of tree construction.

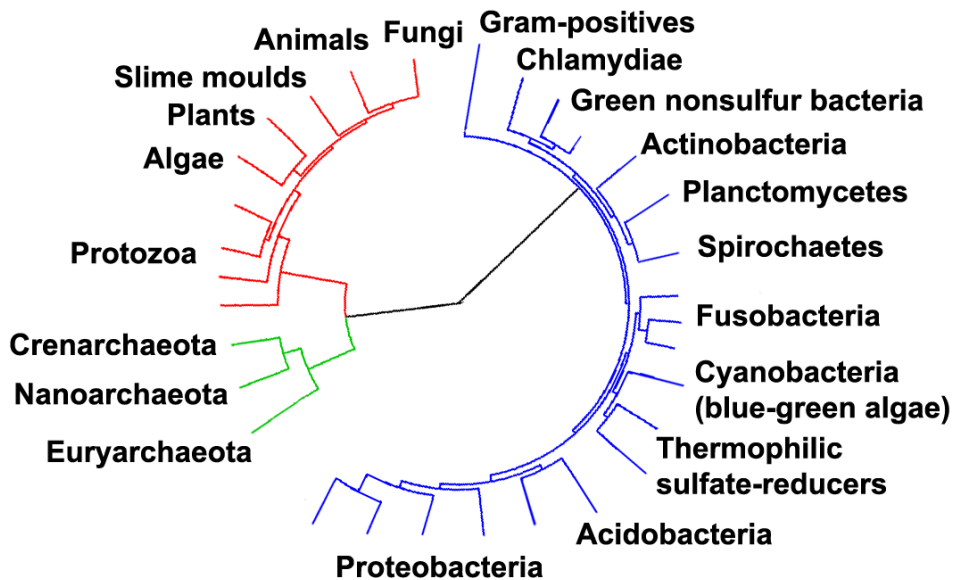


Figure 3.1: Example of phylogenetic tree

Maximum parsimony is another simple method of estimating phylogenetic trees, but

implies an implicit model of evolution. More advanced methods use the optimality criterion of maximum likelihood, often within a Bayesian Framework, and apply an explicit model of evolution to phylogenetic tree estimation. Identifying the optimal tree using many of these techniques is NP-hard, so heuristic search and optimization methods are used in combination with tree-scoring functions to identify a reasonably good tree that fits the data.

3.7 Geographical vizualization

3.7.1 SVG maps

Scalable Vector Graphics is a text-based graphics language that describes images with vector shapes, text, and embedded raster graphics. It is a royalty-free vendor-neutral open standard developed under the W3C [3] Process.

3.7.2 GIS

Geographic information systems [16] or geospatial information systems is a set of tools that captures, stores, analyzes, manages, and presents data that are linked to location(s). In the simplest terms, GIS is the merging of cartography, statistical analysis, and database technology. GIS may be used in geography, cartography, remote sensing, land surveying, public utility management, natural resource management, precision agriculture, photogrammetry, urban planning, emergency management, navigation, aerial video, and localized search engines.

3.7.3 Google maps API

Google maps API [11] is a free service, available for any web site that is free to consumers. I choose this last option because SVG is difficult to extend and GIS is too big, commercial and complicated tool for our purpous.

3.7.4 GPS

The Global Positioning System [32] is a space-based global navigation satellite system that provides reliable location and time information in all weather and at all times and anywhere on or near the Earth when and where there is an unobstructed line of sight to four or more GPS satellites. It is maintained by the United States government and is freely accessible by anyone with a GPS receiver. In addition to GPS other systems are in use or under development. The Russian GLObal NAvigation Satellite System was for use by the Russian military only until 2007. There are also the planned Chinese Compass navigation system and Galileo positioning system of the European Union (EU). GPS was created and realized by the U.S. Department of Defense and was originally run with 24 satellites. It was established in 1973 to overcome the limitations of previous navigation systems.

3.7.5 Geocoding

Geocoding is automatic, intelligent, data mining process of converting street addresses or other locations (ZIP codes, address, city, state, street, etc.) to latitude and longitude.

3.8 Implementation tools

3.8.1 PHP

Hypertext Preprocessor is a widely used, general-purpose scripting language that was originally designed for web development to produce dynamic web pages. For this purpose, PHP code is embedded into the HTML source document and interpreted by a web server with a PHP processor module, which generates the web page document. As a general-purpose programming language, PHP code is processed by an interpreter application in command-line mode performing desired operating system operations and producing program output on its standard output channel. It may also function as a graphical application. PHP is available as a processor for most modern web servers and as a standalone interpreter on most operating systems and computing platforms.

3.8.2 JavaScript

JavaScript is primarily used in the form of client-side JavaScript, implemented as part of a web browser in order to provide enhanced user interfaces and dynamic websites. However, its use in applications outside web pages—for example in PDF-documents, site-specific browsers and desktop widgets—is also significant.

3.8.3 Googlemaps overlays

Wide array of APIs that let the user embed the robust functionality and everyday usefulness of Google Maps into the website and applications, and overlay [11] the data on top of the underlying maps.

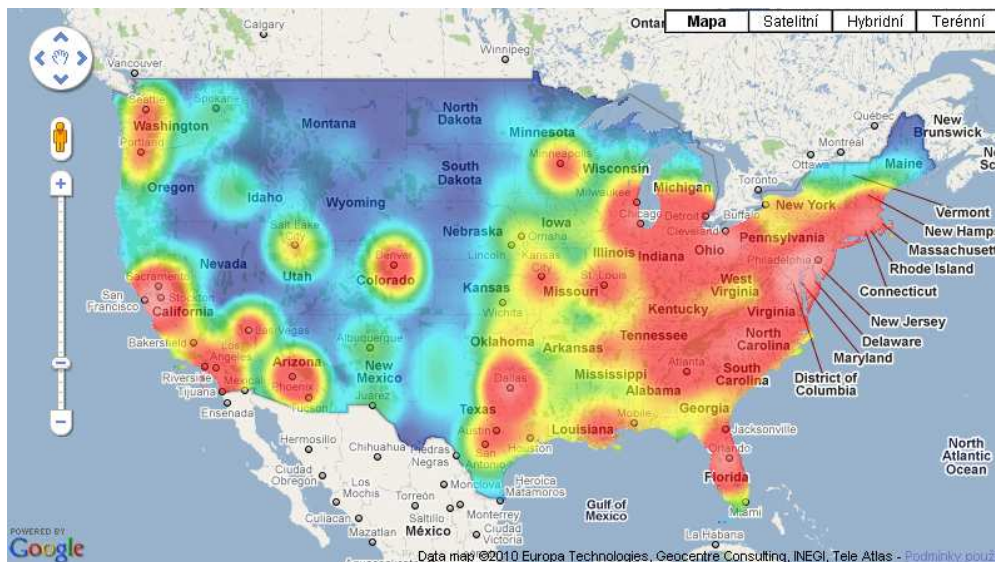


Figure 3.2: Example of Googlemaps API overlay (source [11])

Overlays are objects on the map that are tied to latitude/longitude coordinates. It is possible to expose the set of coordinates (usually XML file) to the API and let the API core to display results on the top of the map with no loss of other map functions like zooming or sliding.

3.8.4 CSV files

A comma-separated values [10] or character-separated values file is a simple text format for a database table. Each record in the table is one line of the text file. Each field

value of a record is separated from the next by a character (typically a comma, but some European countries use a semi-colon as a value separator instead). Implementations of CSV can often handle field values with embedded line breaks or separator characters by using quotation marks or escape sequences. CSV is a simple file format that is widely supported, so it is often used to move tabular data between different computer programs that support the format. For example, a CSV file might be used to transfer information from a database program to a spreadsheet.

Chapter 4

Designing of HLA metric

Important part of this thesis is to design a suitable metric to measure how two individuals are related to each other with respect to their HLA phenotype. The meaning of the term relativeness does not necessarily refer to family relativeness, but it generally represents genetical distance according to the accessible HLA information.

I work with two approaches how to measure the distance between two HLA phenotypes. First approach is based on simple antigen locus match (LR). Second, more advanced approach is based on conditional probability of allele match (HR). Both approaches compare particular loci and then combine its results, regardless interlocus linkage equilibrium [24].

4.1 Definition of HLA resolution

Depending on the particular typing method there are can be find different resolutions in the HLA registry.

Important resolutions are highlighted using bold text (table 4.1). Note that records with * denote DNA and without * it denotes serology antigens. In following text two different metrics are design.

First one computes simple HLA match based on LOW RESOLUTION. This distance function is called *Antigen mismatch* because low resolution alleles are converted to antigens using antigen equivalents (see chapter 3.1) and the other one is based on computing probability of allele match.

example	resolution		description
	Serology	DNA	
xxxx	X	-	missing value or homozygot
xxxx	-	X	missing value
A9	Broad	-	uncertain antigen - generally set of splits
A3	Split	-	certain antigen
$A^*02 : XX$ OR A^*02	-	LOW	possible to decomposed to the set alleles
A^*42AB	-	Intermediate	set of high resolution alleles
A^*0201	-	HIGH	certain allele - the best outcome
$A^*02 : 01$	-	DNA-4	new nomenclature
$A^*020101$	-	DNA-6	certain allele - seldom present
$A^*02 : 01 : 01$	-	DNA-6	new nomenclature
$A^*02010101$	-	DNA-8	certain allele - very seldom present

Table 4.1: List of different levels of resolution

4.2 Low resolution antigens mismatch metric

Antigen equivalents method is used (see chapter 3.1) to express serology equivalents and to downgrade all values with higher resolution. There [12] is file with DNA \rightarrow SEROLOGY relation and with the following structure:

HLA Locus	HLA Antigen/ Allele name	Unambiguous Serology	Possible Serology	Assumed Serology	Expert Assigned Exceptions
A*	01:01:01	1			
A*	01:04N	0			
A*	01:10			1	
A*	02:01:01:02L		0/2		
A*	02:03:01	203			
B*	13:04		15/21		13
B*	83:01	?			

Figure 4.1: DNA allele to serology antigen transformation [13]

Online database of serologically defined antigens [13] contains many details of all current HLA alleles where known their unambiguous, possible or assumed serologically equivalent antigens.

Details of the unambiguous serology is defined from submissions to the WHO Nomenclature Committee for Factors of the HLA System [22] at the time an allele is submitted for naming, or from the WMDA HLA Dictionary 2008 [28]. For Null alleles a value of zero (0) is given and for alleles with no corresponding antigen a question mark ? is given.

In cases where an allele has been shown to be associated with more than one serologically defined antigen, these are indicated in the 'Possible Serology' field. Multiple values are separated by a forward slash /. In cases where there is currently no information about the serological equivalent of an allele, the 'Assumed Serology' field contains the antigen equivalent as expected by the first two digits of the allele name.

The file of equivalents contains details of all current HLA antigens and alleles, and is sorted by locus and allele number.

Table 4.2 shows downgraded dataset from table 5.1.

	Locus A		Locus B		Locus DRB1	
id	DNA-LOW		DNA-LOW		DNA-LOW	
a	3	30	7	13	xx	7
b	2	30	18	62	17	8
c	3	11	7	35	1	8
d	2	11	55	62	4	13
e	11	xx	7	35	xx	xx
f	3	11	27	44	11	13
g	2	11	35	51	1	9
h	2	68	39	42	18	14

Table 4.2: Low resolution downgraded antigens

4.2.1 Mismatch counter

Let us introduce notation X_m^n to handle with the orientation in phenotypes (antigen equivalents) of both given individuals (tab.4.3).

n	Locus A		Locus B		Locus DRB1	
m	1	2	1	2	1	2
X	3	30	7	13	xx	7
Y	2	30	18	62	17	8

Table 4.3: Low resolution examples

Mismatching counter $MM_c(X, Y)$ decrements the number of mismatching loci. Two unphased phenotypes X, Y are the input of this function. Each one of them represents one row in the reduced dataset (table 5.1). Output $d = MM_c(X, Y)$ is the distance (number of mismatched antigens) between given inputs. $d = \langle 0, 1, 2, 3 \rangle$, where 0 means no loci mismatch and 3 means all loci mismatch. d is playing the role of the genetic distance for the purpose of this metric.

Implementation of mismatching counter:

$d \leftarrow 3$

for all loci n **do**

if $X_1^n == Y_1^n$ AND $X_2^n == Y_2^n$ **then**

$d = d - 1$

else if $X_1^n == Y_2^n$ AND $X_2^n == Y_1^n$ **then**

$d = d - 1$

end if

end for

Output d is finally the number of matching loci representing distance between two phenotypes. Despite the fact that this approach is notably inaccurate for two individuals it is fast enough to be use in many to many algorithms which is the objective of this work.

4.2.2 Unphased chromosomes problem solving

Note that at each locus, there are always two unphased alleles (each comes from one ascendant). Therefore we have to compare both order combinations. If pairs of alleles are equal at least in one of two possible configurations, then these two individuals MATCH at the entire locus. Two individuals were chosen as an example for the further study of this problem (table 4.3).

4.2.3 Handling with missing values

Missing values are automatically evaluated as mismatched for the purpose of this simple distance function. High resolution metric in next section handle with missing values more sophisticated way.

4.3 High resolution metric

The distance function of this metric answers the question: "How great is probability that two given individuals are equal at the high resolution level". Preprocessed alleles frequencies (more in section 5.4.5) were used to decompose antigens and alleles at lower resolutions. Following chapter describes the design of this metric in detail.

There are two issues which have to be taken into account to design more accurate metric:

1. Use aprior knowledge about entire population - alleles frequencies
2. Use as much HLA information as possible from both input individuals

Let us define two different phenotypes A and B . Result of the HR distance funtions is the probability P determining that phenotype A is the same as phenotype B regarding the high level of resolution. Let p_i is probability that the gen a_i of the phenotype A is equal to the gen b_i of the phenotype B , considering high resolution. Final probability P is geometric combination in the form of euklidian distance $P = \sqrt{p_1^2 + p_2^2 + \dots + p_n^2}$. Apparently, probabilities of loci equivalence are measured separately, without respect to the linkage equilibriums [24].

4.3.1 Principle schema of the metric

This schema (figure 4.2) helps to understand the overall principle of the suggested metric. HLA phenotypes of two individuals come in to the measuring process, then particular gens undergo adjustment to become a set of alleles with probabilities.

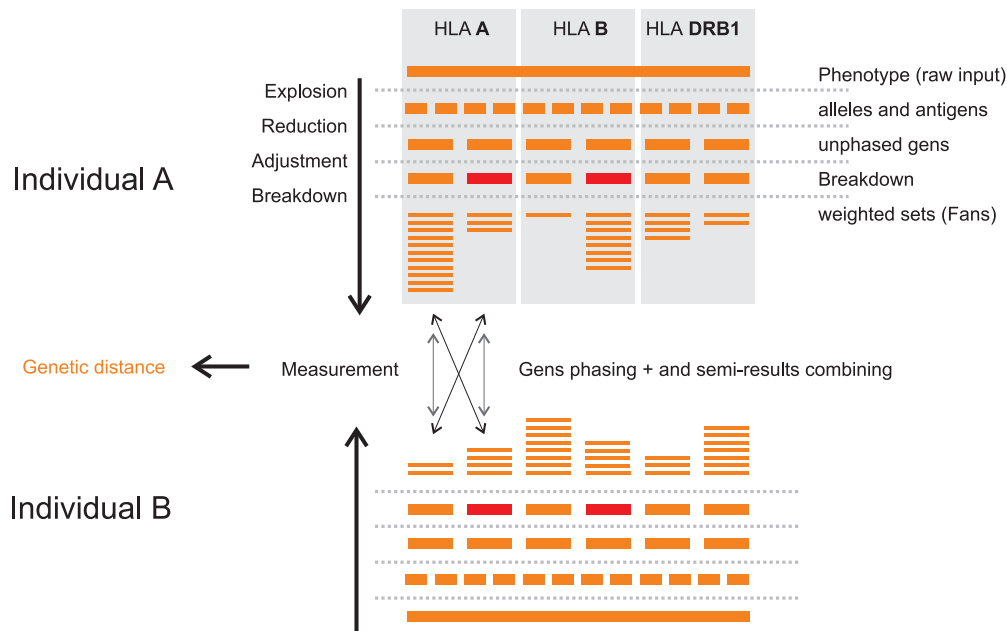


Figure 4.2: Principle schema of high resolution metric

4.3.2 Definition of the resolution DNA-4

Resolution DNA-4 was introduced for any description of allele with first four digits regarding the old HLA nomenclature (e.g. A*0101) and two first blocks regarding the new HLA nomenclature (e.g. A*01:01).

4.3.3 Probability of possible allele match

There are two input phenotypes of unphased values at each locus (gen). Each gen is handled with separately, and after computation particular results are combined to get final phenotype distance.

1. DNA/Serology redundancy elimination (more in section 5.2.3)
2. Downgrade all higher resolutions than DNA-4 to DNA-4
3. Fan brakedown of the records with the resolution DNA-2 and lower
4. Allele frequency assignment and normalization
5. Intersection of weighted sets of alleles

Each of these points is described further.

4.3.4 Downgrading higher resolutions

Downgrading of all higher resolutions than DNA-4 consist in decisive degradation of higher resolution levels. The outcome after downgrading is substring of the first 4 digits. (see figure 4.3).

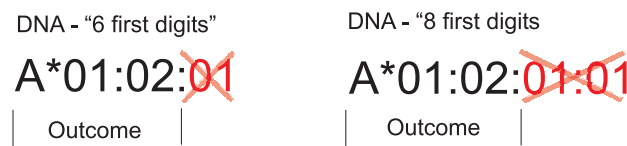


Figure 4.3: Higher resolution downgrading

4.3.5 Decomposition of alleles with lower resolution

Each lower resolution allele is expressed as the set of alleles of the DNA-4 resolution. Figure 4.4 shows why the *FAN* keyword is used.

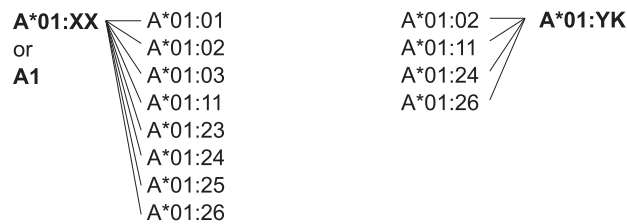


Figure 4.4: Decomposition of lower resolutions

4.3.6 Allele frequency assignment and probability normalization

Probability for each allele in the set is weighted using the preprocessed allele frequencies. After assigning the frequencies, normalized probability distribution, of entire set is computed. The sum of all the elements in any probability distribution has to be 1.

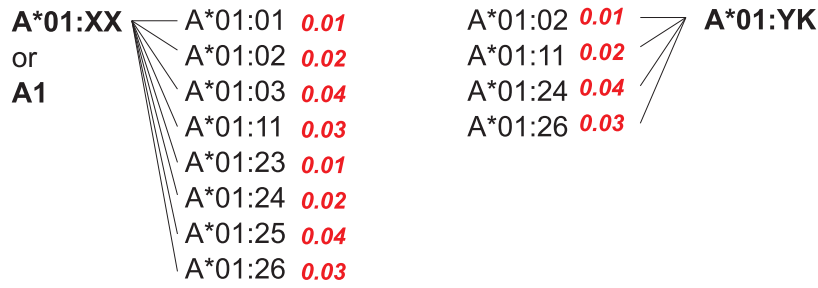


Figure 4.5: Alleles and their frequencies assignment

This is the way how I got *weighted set of alleles* for any input. If input is an empty value, then we get weighted set of possible alleles.

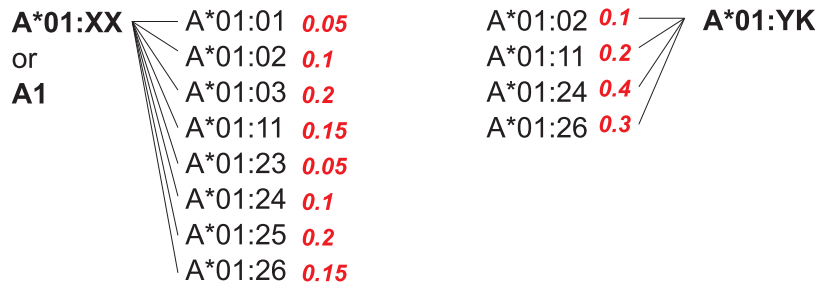


Figure 4.6: Probabilities after normalization

4.3.7 Probabilistic intersection of weighted sets

Genetic distance of two gens is always compute using probability intersection of their weighted sets in the case of this metric.

Figure 4.7 shows the own process of two different weighted sets comparison.

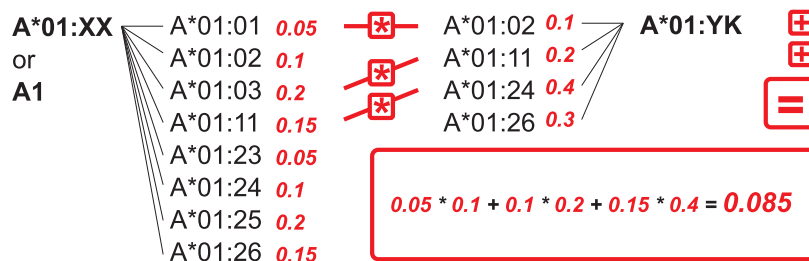


Figure 4.7: Example of intersection of weighted sets

Two sets of alleles (set_A and set_B) are input of the algorithm and the output is overall

probability of their high resolution equity. Function $P_{AB} = \text{intersect}(\text{set}_A, \text{set}_B)$ works in accordance with the following algorithm:

```

 $P_{res} = 0$ 
for all alleles  $a$  in  $\text{set}_A$  do
  for all alleles  $b$  in  $\text{set}_B$  do
    if  $a == b$  then
       $P_{res} += P(a) \cdot P(b)$ 
    end if
  end for
end for

```

where $P(a)$ is probability of allele a in set_A and $P(b)$ is probability of allele b in set_B

4.3.8 Phasing the gens

There are two unphased chromosomes X and Y at each locus and it is necessary to choose the right constalation. HLA experts realized [2] that a closer genetic combination of unphased chromosomes is preferred in the nature and that is the reason why this approach was used for this metric. Fortunately, there are only two possible order combinations from two individuals (X_1X_2, Y_1Y_2 and X_1Y_2, Y_2X_1) and we are looking for the one with the higher probability of match. The following approach was designed to find the most suitable order of the pair elements for the purpose of phenotype relativity computation.

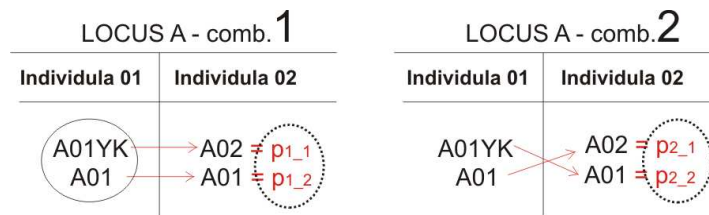


Figure 4.8: Unphased phenotypes elimination

In case of antigen equivalents is easy to choose preferred combination of chromosomes. Preferred combination is just the one where both antigens match. We are looking for the way how to choose preferred combination of decomposed alleles. Let us introduce the formula which takes the maximum of the minimal chromosome combination.

$$p_i = \max(\min(P_X^1 P_X^2, P_Y^1 P_Y^2), \min(P_X^1 P_Y^2, P_Y^1 P_X^2))$$

where p_i is probability of locus i match at DNA-4 resolution. It was empirically derived that *maxi-min* approach provides higher HLA genetic diversity than *maxi-mean* and also than *mean-mean*.

4.3.9 Combination of particular loci results

To get the final relativeness of the two phenotypes it is necessary to combine all the contributive probabilities of particular loci. It can be done using averaging $\sqrt{\sum p_i^2}$ or using euclidian distance $\frac{1}{N} \sum p_i^2$ discussed in chapter 6.

Chapter 5

Implementation

This chapter explains how we designed tool for geographical analysis of stem cell donors registry. The main goal of this online application is to compute genetic distribution of HLA relativeness to the given HLA phenotype and visualize the results as the map with proportionally colored regions.

To facilitate the maintenance and retrieval of information, the back-end is based on a relational database model utilizing MySQL as the database management system. The database can be accessed utilizing any of the most common web browsers. The use of a web browser as a front-end gives the facility to users to access data without the necessity of installing a package. Web pages were implemented using the active server pages scripting environment for the development of dynamic pages, with the assistance of the JavaScript language for data entry validation.

The graphical display of this software was developed using HTML and CSS to guarantee a standard visualization in the majority of common browsers.

5.1 Schema of the project

In the following schema the project is depict as the black box with inputs and outputs.

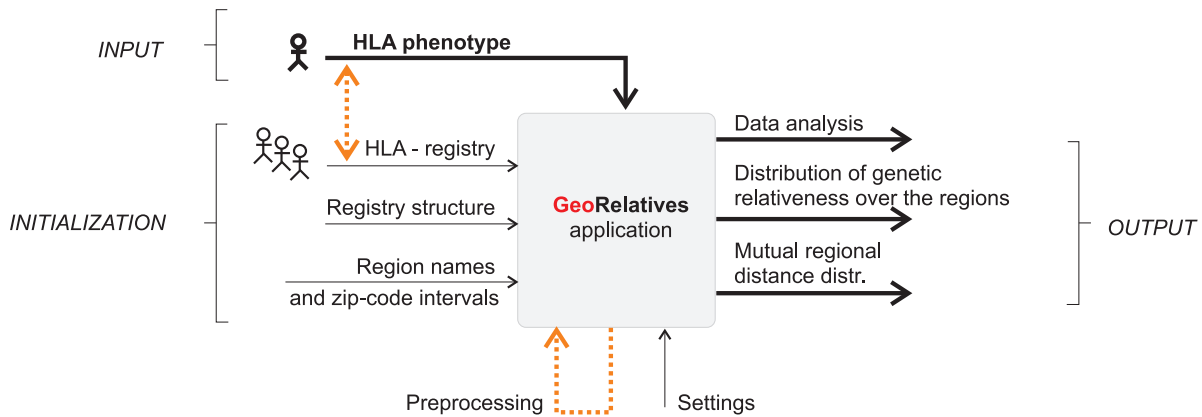


Figure 5.1: Schema of the project

The process of genetic distance computation consists of five steps:

1. System initialization
2. Registry preprocessing
3. Introduction of input HLA phenotype
4. Measuring of the HLA genetic distance over all regions
5. Visualization of outcomes

These steps will be discussed more extensively in following chapters.

5.2 Dataset adjustment

This section follows the natural flow of entire algorithm. It starts with the raw HLA given data and it ends with the preprocessed internal data structure which allows real time computation of genetic distance distributions.

5.2.1 Raw HLA data overview

Let us foreshadow HLA dataset to imagine the complexity of the problem.

id	Locus A				Locus B				Locus DRB1			
	serology		DNA		serology		DNA		serology		DNA	
a	3	30	xxxx	xxxx	7	13	xxxx	xxxx	xx	7	xxxx	xxxx
b	2	30	0201	3002	18	62	xxxx	xxxx	17	08	0301	0801
c	3	11	xxxx	xxxx	7	35	xxxx	xxxx	1	08	01XX	08XX
d	2	11	xxxx	xxxx	55	62	xxxx	xxxx	04	13	04DAX	13XX
e	11	xx	xxxx	xxxx	7	35	xxxx	xxxx	xx	xx	xxxx	xxxx
f	3	11	xxxx	xxxx	27	44	xxxx	xxxx	11	13	xxxx	xxxx
g	2	11	02XX	1101	35	51	3501	5101	1	9	0101	0901
h	2	68	02XX	68XX	39	42	3905	42AB	18	14	0302	14AYWB

Table 5.1: Examples of HLA dataset

Each row in this dataset represents one individual and we are looking for objective indicator of the distance between any two of them. Individuals are indexed using characters a, b, c, \dots, h in the first column. For each loci there are two pairs of two columns. In the first pair, there are two unphased outcomes from serology typization. In the second pair of columns there are outcomes from the DNA screening. For most of individuals we don't have data in DNA columns, because the technique of obtaining these data is expensive and the equipment for this technique is available only for some of registries.

5.2.2 Specification of problems

Given HLA dataset includes thousands of donors which have been typed at the different level of resolution and accuracy. Let me summarize the main problems of the raw HLA data:

- redundancy
- missing values
- uncertain values - different resolution
- unphased chromosomes

In the further process of the data preparing all these problems are taken into account (see section 5.2.3). Due to the data inconsistency we have to reorganize them into the

internal database. We would like to get smaller dataset to get algorithms faster and to be able to work with the data in accordance with the assignment.

5.2.3 Redundancy elimination

Each individual has 4 possible records at each loci. Two serological and two DNA. One of these pairs are always redundant. When DNA pair is present and then serological results are irrelevant or DNA is empty and then serology has to be used. There remain only two rows out of four with each locus after elimination of redundancy. There remains 6 rows for 3 loci (A,B,DRB1). This process is also called *reduction*.

	Locus A		Locus B		Locus DRB1	
id	DNA-LOW		DNA-LOW		DNA-LOW	
a	3	30	7	13	xx	7
b	0201	3002	18	62	0301	0801
c	3	11	7	35	1	8
d	2	11	55	62	04DA	13
e	11	xx	7	35	xx	xx
f	3	11	27	44	11	13
g	02XX	1101	3501	5101	0101	0901
h	02	68	3905	42AB	0302	14AYWB

Table 5.2: Example of reduced dataset

5.2.4 Homozigode identification

If only one of the chromosome pair information is present, the other one chromosome is considered to be the same and the individual is homozigode at the particular locus.

	Locus A		Locus B		Locus DRB1	
id	DNA-LOW		DNA-LOW		DNA-LOW	
a	3	30	7	13	7	7
e	11	11	7	35	xx	xx

Table 5.3: Homozigod cloning

5.2.5 Plotting of regional maps

It is possible to create own arbitrarily maps and use them like a googleMap vector overlay. Not every province is directly accessible in suitable format, thus *Digitizer tool* [7] can be used for creating own polygons via GPS coordinates. User defined regions can be draw and import to the *GeoRelatives* application in accordance with the manual.

5.2.6 Individuals localization

Each province (region) usually contents the set of zip codes, but it is almost imposible to retrieve suitable list from the post office or any other office, so individuals have to be classified manually using suggested zip code intervals.

5.2.7 Zip code classification

Fortunately, it is possible to interpret sets of discreet values like a continuous interval or a union of more continuous intervals.

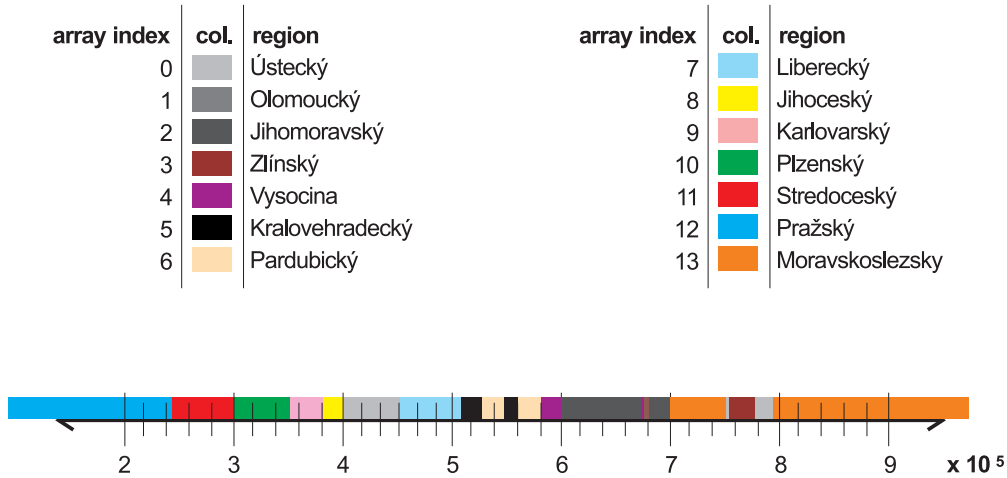


Figure 5.2: Classification of individuals into the geopolitical regions

Region names and their zip code intervals have to be uploaded to the system in format which specifies the name of the region and its set of continuous zip-code intervals.

5.3 Preprocessing

Each locus is described with two pairs of two unphased chromosomes, serology and DNA. The list of four descriptors in HLA dataset and locus A: $A1_{SER}|A1_{DNA}|A2_{SER}|A2_{DNA}$. Let us call this entire representation of locus "A" *The string for locus A*. Generally $*1_{SER}|*1_{DNA}|*2_{SER}|*2_{DNA}$ is string for locus X. This denotation is used in the following text.

5.3.1 Reduction of the string

The goal of this operation is to choose the correct representant of the locus. It takes either serology or DNA pair. DNA pair is used if both are present because it is supposed to be better. Then chosen elements are sorted in descending order and delimited with $|$. Finally there is a string with only two elements $*1_{RED}|*2_{RED}$, where $*$ stands for any locus, *RED* for *reduced* and number is distinctive index for unphased chromosomes.

5.3.2 Indexing and locus distance preprocessing

There are many redundant records in the given dataset and it doesn't make any sense to repeat the same computation more than once, which is the reason we suggested the following algorithm for each locus:

1. Sort elements $*1_{RED}$ and $*2_{RED}$ in descending order and optionally adjust the string
2. Make a list of all different reduced strings $*1_{RED}|*2_{RED}$
3. Assigning indexes
4. Sort the list in descending order
5. Create all to all triangular distance matrix

These matrices are then used for mutual phenotype distance measuring. Algorithm checks if preprocessed value is present. If so, then preprocessed distance is used, otherwise distance is measured directly. Direct computation of HR distance takes much longer time, especially when there are many alleles in sets going to the intersection process. It increases the efficiency of the measuring process more than 100 times.

5.3.3 Getting genetic distances from precomputed data

For the proper generating of genetic maps in real time, there must be following tables created in the internal database.

1. List of all different reduced strings
2. Preproceed values of all string combinations

Then entire process of computation can be modify the way shown in figure 5.3.

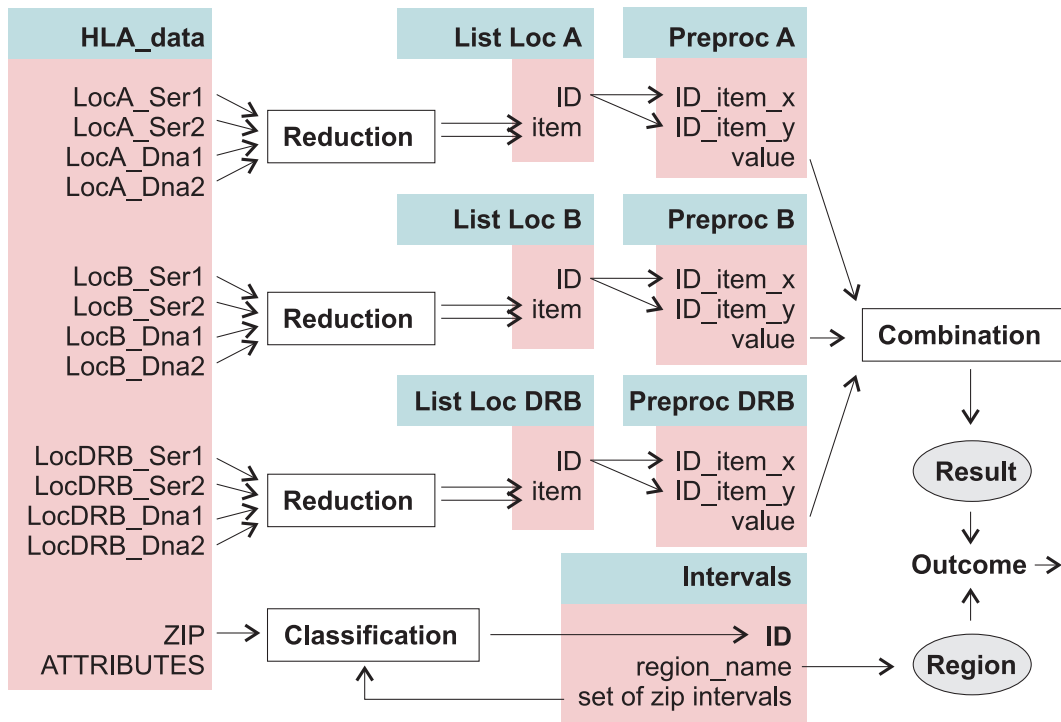


Figure 5.3: The diagram of getting genetic distances from precomputed data

Each computation of genetic distance can be done either directly or using preprocessed data.

5.4 Definition of the Country

User can define new country and edit existing country using CSV files with specific structure. It is sufficient to create a new folder in *Country * with the name of the country. Inside of this folder there are generally ASCII text files where each line is one record (row in DB) and columns are separated using *delimiter*.

5.4.1 National registry import

File *HLA.csv* must be in *Country \ GEN *. This CSV file includes all available HLA phenotypes of individuals in the country. There are mandatory columns A, B, DRB1,

ZIP and optional columns, ATTRIBUTES and other loci. Structure of this file is strictly defined by file *HLA-structure*. More in section 5.4.2.

<i>MD</i> 2 13 39 7 12 07XX 12XX 75661
<i>MD</i> 10 30 13 44 75661
<i>MD</i> 1 11 8 51 3 4 03XX 04XX 35735
<i>MD</i> 2 11 60 62 14700
<i>MD</i> 1 24 39 57 7 13 0701 13XX 27201
<i>MD</i> 1 24 0101 24XX 7 57 0702 5701 4 15 04BK 1501 12000
<i>MD</i> 2 35 60 28201
...

Table 5.4: Example of anonymized national stem cell registry file

5.4.2 Dataset structure definition

File *HLA-structure.csv* has to be placed in *Country \ GEN * and each line of this file consist of three columns:

- Index of the column in *HLA.csv*
- Name
- Description

Index is the order of the column in the *HLA.csv* beginning with zero value. *Name* has to be one of these: A1_SER, A1_DNA, A2_SER, A2_DNA, B1_SER, B1_DNA, B2_SER, B2_DNA, DRB1_SER, DRB1_DNA, DRB2_SER, DRB2_DNA, ATTRIBUTE, ZIP. This file determine the structure of *HLA.csv* file. It assigns HLA data columns to their meaning. See example in figure 5.5

0;DonType;Donor Type
1;A1.SER;Locus A 1 Serology
2;A2.SER;Locus A 2 Serology
3;A1.DNA;Locus A 1 DNA
4;A2.DNA;Locus A 2 DNA
5;B1.SER;Locus B 1 Serology
..., ...

Table 5.5: Example of file with zip code intervals

5.4.3 Names and zip code intervals of regions

File *regionsZIP.csv* has to be placed in *Country \ GEO * with attributes:

- Index of the region
- Name
- Zipcode intervals

Each Zip code interval is separated by | and consists of two or three parts. First is the sign $< or > or ><$, second or second and third are numerical interval extremes. For example *Moravskoslezsky kraj* is defined this way 13; *Moravskoslezskykraj*; $>, 792| ><, 700, 750$. The file should contain as many lines as there are regions is in the country. See example:

0;Ústecký;><,400,450
1;Olomoucký;><,750,753 ><,770,793 ><,796,798
2;Jihomoravský;><,600,674 ><,676,686 ><,689,700
3;Zlínský;><,686,689 ><,753,770
4;Vysočina;><,580,600 ><,674,676
5;Kralovehradecký;><,515,530 ><,542,560
... ; ... ; ...

Table 5.6: Example of file with zipcode intervals

5.4.4 GPS borders of subregions

Geographical borders of regions are determined by the list of GPS coordinates. Files *X.gps* have to be placed in *Country \ GEO \ GPS-region-borders \ . X* is index of the region and contents each coordinate on the line in decimal format *latitude, longitude*. See example:

49.274633220,17.160206083
49.276968797,17.162732143
49.278188519,17.162391767
..., ...

Table 5.7: Example of GPS regions border definition

5.4.5 Preprocessed allele frequencies

Files *Freq-A.csv*, *FreqB.csv* and *FreqDRB1.csv* have to be placed in *Country \ GEN \ Freq * directory and each line of this file has to start with the name of allele and its frequency in population. Other optional columns can be added as a frequency

See example:

0101g 0,17181 2 0,04742 8 0,05082 5 0,06702 4
0102 0,00006 52 0,00645 25 0,00000 NA 0,00301 30
0103 0,00013 35 0,00021 47 0,00000 NA 0,00000 NA
0116N 0,00006 43 0,00000 NA 0,00000 NA 0,00000 NA
0201g 0,29604 1 0,12458 1 0,09458 3 0,19403 1
0202 0,00083 24 0,04201 10 0,00028 4 —0,00678 22
...

Table 5.8: Allele frequencies

5.5 User interface

Web tool *GeoRelatives* is written in PHP language. The outcomes of genetic distance computation are visualized using Googlemaps API with colored polygonal overlayers.

5.5.1 Maps of genetic relativeness

Let us select the Czech Republic and fill in patients' phenotype to the form at the top. After button *measure* is pressed, application compute genetic distances to the all individuals in the country and show the map of HLA genetic distance distribution. On the right side of the map, there is a list of regions sorted from the most similar one.

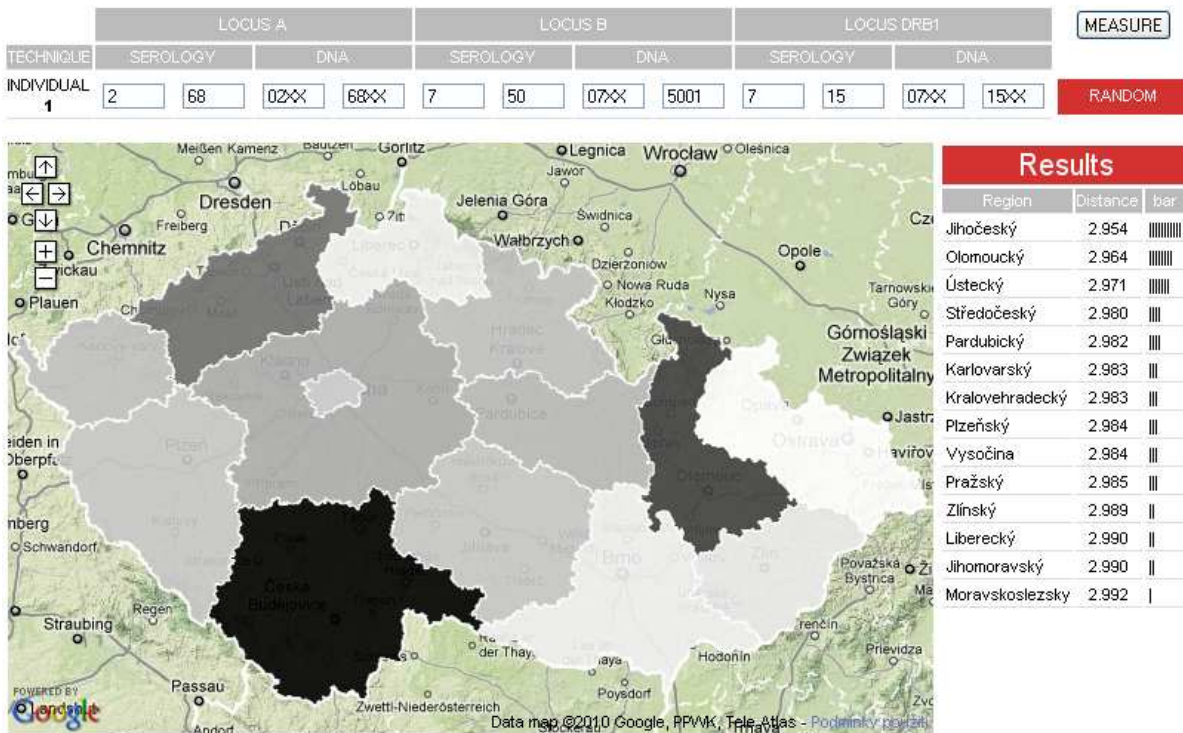


Figure 5.4: Map of genetic distance distribution in regions of Czech Republic

5.5.2 Distance function check

This tool allows to check the overall process of genetic distance computation. It consists of the following blocks:

Input form

Input form for two phenotypes to be measured in figure 5.5. DNA alleles, serology antigens or both can be filled in or phenotypes can be load using identification of individual in the registry.

TECHNIQUE	LOCUS A				LOCUS B				LOCUS DRB1				MEASURE
	SEROLOGY		DNA		SEROLOGY		DNA		SEROLOGY		DNA		
INDIVIDUAL 1	<input type="text" value="2"/>	<input type="text" value="26"/>	<input type="text"/>	<input type="text"/>	<input type="text" value="5"/>	<input type="text" value="12"/>	<input type="text"/>	<input type="text"/>	<input type="text" value="7"/>	<input type="text"/>	<input type="text"/>	<input type="text"/>	ID: <input type="text" value="332"/> <input type="button" value="set"/>
INDIVIDUAL 2	<input type="text" value="2"/>	<input type="text" value="29"/>	<input type="text"/>	<input type="text"/>	<input type="text" value="8"/>	<input type="text" value="44"/>	<input type="text"/>	<input type="text"/>	<input type="text" value="15"/>	<input type="text" value="3"/>	<input type="text" value="15XX"/>	<input type="text" value="03XX"/>	ID: <input type="text" value="1234"/> <input type="button" value="set"/>

Figure 5.5: Two phenotype input form

Reduction

Reduction check (figure 5.6) shows how the redundancy of full unphased HLA input record is eliminated.

Reduction							
	LOCUS A		LOCUS B		LOCUS DRB1		
INDIVIDUAL 1	2	3	7	38	1101	1501	
	Split	Split	Split	Split	DNA-4 digits	DNA-4 digits	
INDIVIDUAL 2	1	2	X	X	11	15	
	Split	Split	Missing	Missing	Split	Split	
Prob. of loc. equity	0.00000		0.00298		0.56473		

Figure 5.6: Reduction check

Phasing process of unphased chromosomes

Chromosome phasing check in figure 5.7. More about chromosome phasing is in section 4.3.8.

FAN INTERSECTIONS LOCUS A		FAN INTERSECTIONS LOCUS B		FAN INTERSECTIONS LOCUS DRB	
I1X --> I2X :	I1X --> I2Y :	I1X --> I2X :	I1X --> I2Y :	I1X --> I2X :	I1X --> I2Y :
0.00000000	0.92421381	0.13712181	0.13712181	0.59578504	0.00000000
I1Y --> I2Y :	I1Y --> I2X :	I1Y --> I2Y :	I1Y --> I2X :	I1Y --> I2Y :	I1Y --> I2X :
0.00000000	0.00000000	0.02174229	0.02174229	0.94788316	0.00000000
product :	product :	product :	product :	product :	product :
0.000000000000	0.000000000000	0.002981341692	0.002981341692	0.564734605684	0.000000000000
winner: 0.00000000		winner: 0.00298134		winner: 0.56473461	

Figure 5.7: Phasing process of unphased chromosomes

Breakdown sets (fans)

Breakdown allele sets of both chromosomes and three loci A,B and DRB1 in figure 5.8.

Individual 1		Individual 2	
X (2)	Y (3)	X (1)	Y (2)
0201 -->	0301 -->	0101 -->	0201 -->
0.9610	0.9801	0.9985	0.9610
0202 -->	0302 -->	0102 -->	0202 -->
0.0027	0.0191	0.0003	0.0027
0203 -->	0305 -->	0103 -->	0203 -->
0.0000	0.0004	0.0008	0.0000
0204 -->	0307 -->	0116 -->	0204 -->
0.0000	0.0004	0.0003	0.0000
0205 -->		0205 -->	
0.0260		0.0260	
0206 -->		0206 -->	
0.0064		0.0064	
0207 -->		0207 -->	
0.0000		0.0000	
0210 -->		0210 -->	
0.0000		0.0000	

Individual 1		Individual 2	
X (7)	Y (38)	X (X)	Y (X)
0702 -->	3801 -->	0702 -->	0702 -->
0.9800	0.9973	0.1399	0.1399
0704 -->	3802 -->	0704 -->	0704 -->
0.0031	0.0000	0.0004	0.0004
0705 -->	3809 -->	0705 -->	0705 -->
0.0147	0.0027	0.0021	0.0021
0709 -->		0709 -->	0709 -->
0.0000		0.0000	0.0000
0710 -->		0710 -->	0710 -->
0.0013		0.0002	0.0002
0715 -->		0715 -->	0715 -->
0.0004		0.0001	0.0001
0721 -->		0721 -->	0721 -->
0.0004		0.0001	0.0001
		0801 -->	0801 -->
		0.1253	0.1253

Individual 1		Individual 2	
X (1101)	Y (1501)	X (11)	Y (15)
1101 -->	1501 -->	1101 -->	1501 -->
1.0000	1.0000	0.5958	0.9479
		1102 -->	1502 -->
		0.0160	0.0509
		1103 -->	1503 -->
		0.0509	0.0012
		1104 -->	1504 -->
		0.3360	0.0000
		1106 -->	1506 -->
		0.0000	0.0000
		1108 -->	1507 -->
		0.0000	0.0000
		1109 -->	1514 -->
		0.0006	0.0000
		1110 -->	
		0.0000	

Figure 5.8: Example of breakdown alleles probabilities distribution

5.5.3 Distance matrix

Example of triangular distance matrix, computed using high resolution metric in figure 5.9 There are 14 rows and 14 columns for each province in Czech Republic.

0.03970													
0.04061	0.03281												
0.03932	0.03109	0.03952											
0.04578	0.03907	0.03910	0.04566										
0.04321	0.03297	0.03384	0.04207	0.03585									
0.03857	0.04018	0.03829	0.04352	0.03966	0.03903								
0.03796	0.03854	0.03598	0.04509	0.03498	0.03543	0.04195							
0.04157	0.03144	0.03975	0.04184	0.03652	0.04028	0.03650	0.03390						
0.03223	0.03002	0.02956	0.02948	0.02707	0.03239	0.03393	0.03058	0.02774					
0.03568	0.03283	0.03178	0.04004	0.03717	0.03535	0.03801	0.03631	0.02830	0.02949				
0.03749	0.03216	0.03576	0.03904	0.03704	0.02900	0.03258	0.03058	0.02741	0.03490	0.03129			
0.03862	0.03817	0.03650	0.04621	0.04217	0.04202	0.03791	0.04469	0.02800	0.03672	0.03348	0.04378		
0.03943	0.04097	0.03167	0.04727	0.04565	0.03830	0.04107	0.04465	0.03296	0.03352	0.03471	0.03802	0.03902	
0.04338	0.03921	0.04104	0.04953	0.04284	0.04086	0.03860	0.03890	0.03030	0.03918	0.03318	0.04044	0.04264	0.04233

Figure 5.9: Example of distance matrix

5.5.4 Q matrix

There is special iterative algorithm which transform distance matrix to Q-matrix. How to compute Q-matrix is explained in section 3.4. See example of triangular Q-matrix in figure 5.10. There are also 14 rows and 14 columns for each province in Czech Republic. Q-matrix like this is prerequisite of phylogenetic tree.

-0.6366														
-0.55925	-0.58642													
-0.57744	-0.60977	-0.51132												
-0.60085	-0.61494	-0.61729	-0.6395											
-0.54216	-0.59861	-0.59088	-0.59305	-0.57816										
-0.64355	-0.5578	-0.58319	-0.62136	-0.57815	-0.63142									
-0.6401	-0.56671	-0.60014	-0.59175	-0.62354	-0.66385	-0.57484								
-0.5766	-0.63173	-0.53472	-0.61057	-0.58488	-0.58547	-0.62006	-0.63108							
-0.59293	-0.55302	-0.56125	-0.66314	-0.60253	-0.5844	-0.55515	-0.57517	-0.5135						
-0.61017	-0.57794	-0.59325	-0.59506	-0.53997	-0.60752	-0.56483	-0.56505	-0.56542	-0.60978					
-0.57545	-0.57298	-0.53249	-0.59406	-0.52853	-0.67072	-0.61699	-0.62081	-0.5631	-0.53186	-0.56218				
-0.64239	-0.58136	-0.60411	-0.58852	-0.54747	-0.59498	-0.63353	-0.53199	-0.63652	-0.59052	-0.6164	-0.5733			
-0.62984	-0.54493	-0.65924	-0.57297	-0.50288	-0.63679	-0.59278	-0.52964	-0.57417	-0.62609	-0.59881	-0.63959	-0.62476		
-0.59837	-0.58198	-0.56273	-0.56178	-0.55253	-0.622	-0.63835	-0.61457	-0.62202	-0.5741	-0.6331	-0.62648	-0.59725	-0.6169	

Figure 5.10: Example of Q-matrix

5.6 Applicability

This section shows how the application *GeoRelatives* can be used in profession life. Either directly typing `www.tehnik.cz/geoRelatives` to the browser or indirectly using libraries, functions or iframe objects.

5.6.1 Direct applicability

Simply typing in `www.tehnik.cz/geoRelatives` to the browser. The short manual is in appendix A.

5.6.2 Libraries and functions

Libraries of *GeoRelatives* functions can be included to any other project. The list and description of all functions is inside of attached CD in appendix B.

5.6.3 Iframe objects

There is a possibility to insert *GeoRelatives* iframe with the map to any online project calling the *url* with the specific structure. The specific structure of the url tag is defined with the following pattern: `iframeMap.php?phenotype=P,country=C,attributes=A`, where the phenotype P is a string: $A_{SER}^1 | A_{SER}^2 | A_{DNA}^1 | A_{DNA}^2 | B_{SER}^1 | B_{SER}^2 | B_{DNA}^1 | B_{DNA}^2 | DRB_{SER}^1 | DRB_{SER}^2 | DRB_{DNA}^1 | DRB_{DNA}^2$ phenotype separated by | delimiter. Attribute country C is simply the name of the country written with small caps (*czech, finland, sweden, ...*). Attribute A is a set of forming attributes determined by the string of values with | delimiter in the following order: map width (px), map height (px), zoom (1-9), central longitude (decimal), central latitude (decimal), color (RGB).

Few seconds after iframe url tag changes, the particular map will appear with the colored regions and the distribution of the HLA genetic distance to the given phenotype.

5.7 Expandability

There are many possibilities how expand *GeoRelatives* application in the future to offer additional services. See the list of some examples.

- Comparison of regions with each other
- Comparison of entire countries
- User friendly way of adding new registries
- Additional distance functions
- Realtime constructing of phylogenetic trees

Chapter 6

Experiments

6.1 Mutual HLA genetic similarity of geopolitical regions

6.1.1 Distance matrix computation

Individuals are classified to the particular regions and then mutual HLA genetic distances of all individuals are measured and results are averaged and sorted out into the triangular distance matrix of regions (table 6.1).

Measuring of HLA genetic distance is based on probability of match, that is the reason why P_x^y was used as an expression for HLA genetic distance between individual x and individual y . In this case *distance* is replaced with inversal term *similarity*.

Let S_A^B is the mutual genetic similarity between two regions A and B , then:

$$S_A^B = \frac{1}{I_A I_B} \sum_{x=1}^{I_A} \sum_{y=1}^{I_B} P_x^y \quad (6.1)$$

where I_A is number of individuals in region A and I_B is number of individuals in region B . Identical individuals are excluded from the computation. It has to be take into account when identical regions are comparing. Also it doesn't make any sense to measure two identical individuals in oposite order, because the result will be the same. The following equation 6.2 shows the way of computing self similarity of any region A .

$$S_A^A = \frac{2}{I_A^2 - I_A} \sum_{x=1}^{I_A} \sum_{y=x+1}^{I_A} P_x^y \quad (6.2)$$

Let S_f is *foreign* mutual HLA genetic similarity of all different regions:

$$S_f = \frac{2}{N^2 - N} \sum_{a=1}^N \sum_{b=a+1}^N S_a^b, \quad (6.3)$$

where N is the number of regions.

Then S_s is *self* similarity of all identical regions n :

$$S_s = \frac{1}{N} \sum_{n=1}^N S_n^n \quad (6.4)$$

Let us make a fraction of these two similarities to get the indicator of the national HLA genetic diversity.

$$I_{div}^N = \left(\frac{S_s}{S_f} - 1 \right) * 100 \quad (6.5)$$

where I_{div} is percentual indicator of genetic diversity among all regions in the whole country. The indicator grows when individuals within regions are more similar to each other than to individuals in other regions. The value of the indicator is influenced by three main factors:

- Used distance metric
- Real genetic spatial diversity of the country
- Shape, location and number of investigating regions

The diversity indicator approaches to zero in case of random classification. The greater indicator I_{div} , the more useful informatin of HLA genetic diversity in dataset is present and the better the chosen metric is able to mine it.

The following section introduces triangular matrices of mutual genetic similarity S_A^B among all the regions in The Czech Republic, Finland and Sweden. These results are used to prove the hypothesis defined at the beginning of this thesis (section 1.2).

6.2 Visualization of the triangular distance matrix

The mutual similarity of all regions in The Czech Republic is depict on the three dimensional triangular bar chart, where the height of each bar is proportional to the averaged similarity between all individuals in two particular regions.

0.03970																					
0.04061	0.03281																				
0.03932	0.03109	0.03952																			
0.04578	0.03907	0.03910	0.04566																		
0.04321	0.03297	0.03384	0.04207	0.03585																	
0.03857	0.04018	0.03829	0.04352	0.03966	0.03903																
0.03796	0.03854	0.03598	0.04509	0.03498	0.03543	0.04195															
0.04157	0.03144	0.03975	0.04184	0.03652	0.04028	0.03650	0.03390														
0.03223	0.03002	0.02956	0.02948	0.02707	0.03239	0.03393	0.03058	0.02774													
0.03568	0.03283	0.03178	0.04004	0.03717	0.03535	0.03801	0.03631	0.02830	0.02949												
0.03749	0.03216	0.03576	0.03904	0.03704	0.02900	0.03258	0.03058	0.02741	0.03490	0.03129											
0.03862	0.03817	0.03650	0.04621	0.04217	0.04202	0.03791	0.04469	0.02800	0.03672	0.03348	0.04378										
0.03943	0.04097	0.03167	0.04727	0.04565	0.03830	0.04107	0.04465	0.03296	0.03352	0.03471	0.03802	0.03902									
0.04338	0.03921	0.04104	0.04953	0.04284	0.04086	0.03860	0.03890	0.03030	0.03918	0.03318	0.04044	0.04264	0.04233								

Figure 6.1: Distance matrix with diagonal element and related neighbours

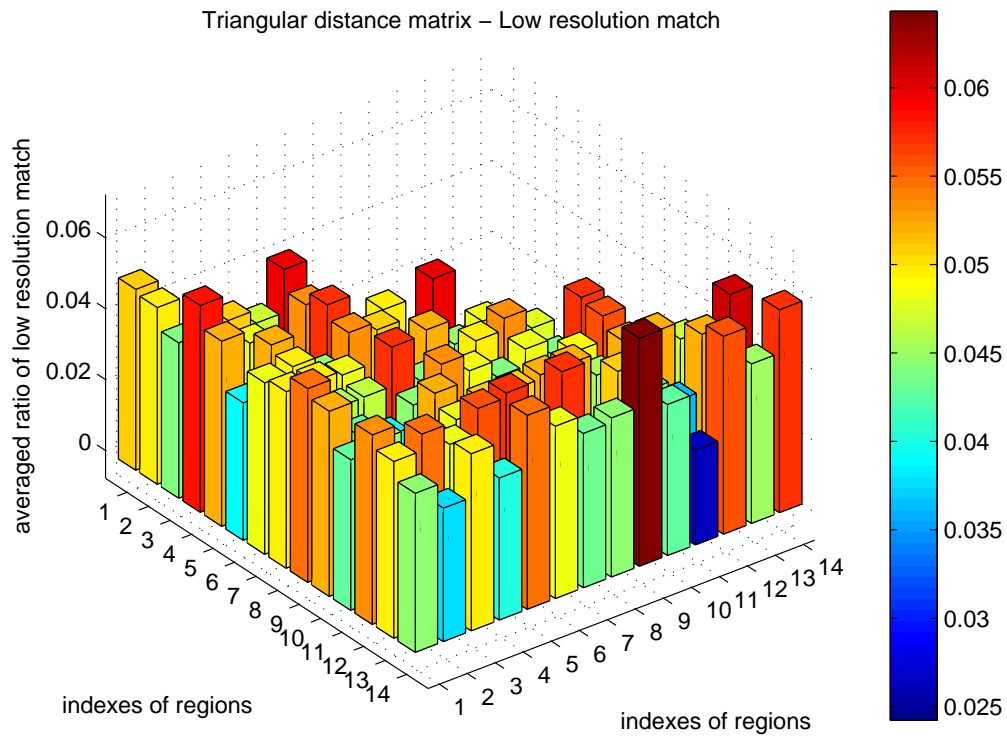


Figure 6.2: Mutual HLA genetic similarity between Czech regions

Regions are indexed on the base plane and each bar is placed at the intersection of two indexes in figure 6.2. Figure 6.2 is simple visualization of table in figure 6.1, but it

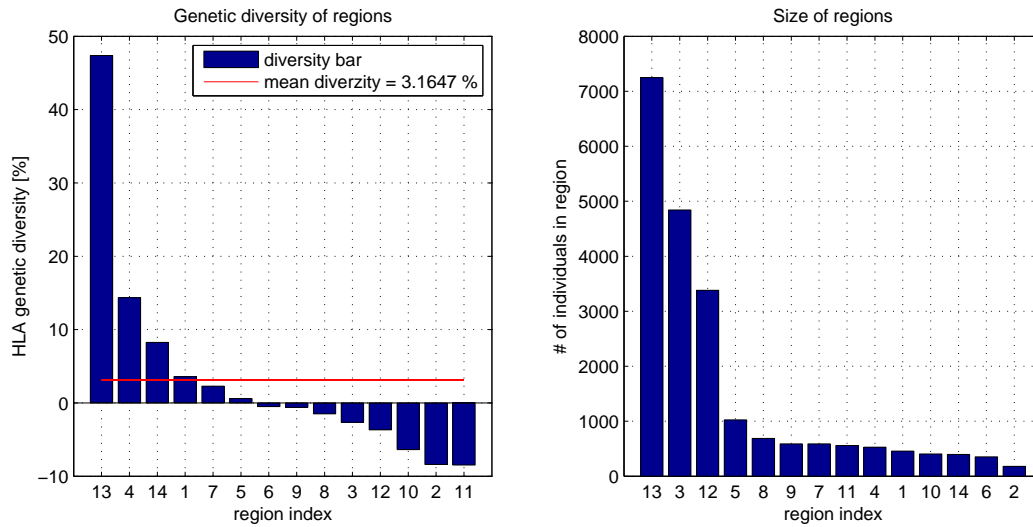


Figure 6.3: HLA genetic diversity of regions of The Czech Republic

is still difficult to interpret it properly, so additional adjustments will be done further to withdraw information from this fuzzy data. There is a list of Czech regions with, indexes, zip codes and number of individuals classified into the particular region in table 6.1.

6.2.1 Genetic diversity of Czech Republic regions

For each region, the value on the main diagonal of the distance matrix was taken and divided with the average of all other regional distances on the same row and column of selected region. Cells relating to the the diagonal item are indicated in figure 6.1. Diagonal value is dividing with the average of the relating values to get genetic diversity of the region. Results for all regions are sorted in figure 6.3.

Regions are again indexed at the bottom and each bar shows HLA genetic divergence of the particular region. The left chart plots just only number of individuals in regions to prove that there is no direct correlation and that distance function is invariant to the size of particular regions.

Intervals are represent with first three digits of zip code, because it was found out sufficient enough.

index	name	zip-code intervals	number of individuals
1	Ústecký	(400, 450)	456
2	Olomoucký	(750, 753) \cup (770, 793) \cup (796, 798)	181
3	Jihomoravský	(600, 674) \cup (676, 686) \cup (689, 700)	4841
4	Zlínský	(686, 689) \cup (753, 770)	528
5	Vysočina	(580, 600) \cup (674, 676)	1025
6	Kralovehradecký	(515, 530) \cup (542, 560)	353
7	Pardubický	(530, 542) \cup (560, 580)	589
8	Liberecký	(450, 515)	688
9	Jihočeský	(370, 400)	590
10	Karlovarský	(350, 370)	405
11	Plzeňský	(300, 350)	558
12	Středočeský	(255, 300)	3381
13	Pražský	< 255	7249
14	Moravskoslezský	(792) \cup (700, 750)	6477

Table 6.1: Regions and zip-code intervals in Czech Republic

Negative values of genetic diversity

For some regions, self similarity is smaller than the foreign similarity. Then, the value of regional genetic diversity is pointless and is shown only for comparison with positive values and also it has to be included to the final averaged national diversity computation.

The red horizontal line is arithmetical average of all regional divergences I_{div}^i and it is what is called HLA genetic diversity of whole country or national stem cell registry $I_{div}^N = \frac{1}{N} \sum_{i=1}^N I_{div}^i$. It approaches to zero when individuals are randomly distributed in regions. The size of this factor indicates the rate of geographical distribution of HLA phenotypes in the given registry.

6.2.2 Genetic diversity of regions in Finland

Let us use previous method for other countries as well, to prove that it is possible to mine useful information with different national stem cell donors registries.

There is very small population of individuals in the region *Ahvenanmaa* and it can influence the diversity, but see the contrast to the Czech regions where the highest diversity

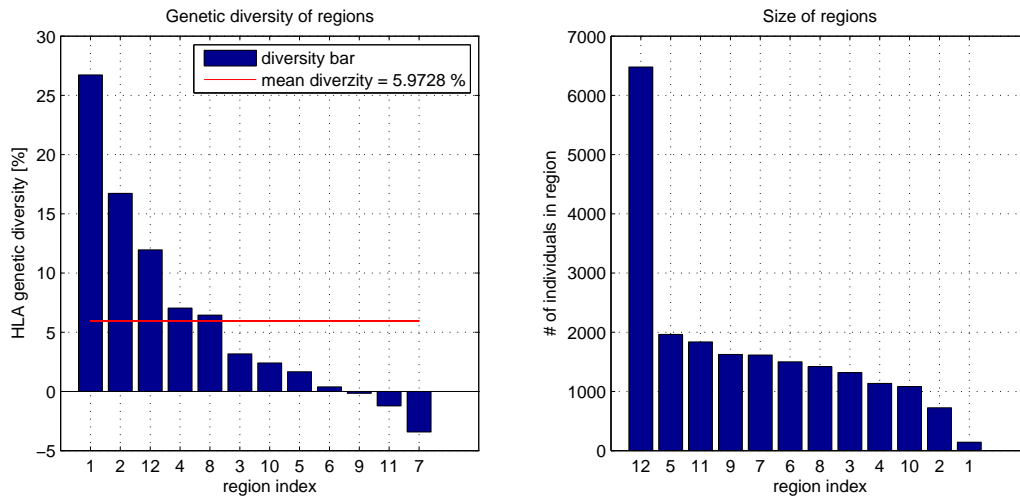


Figure 6.4: HLA genetic diversity of regions in Finland

has the region with the greatest population *Pražský*.

6.2.3 Genetic diversity of Swedish regions

Sweden registry has the lowest HLA genetic diversity from all three investigated national registries, but it is still positive number. In this national registry only regions *South-East*, *South-Central* and *Southern Norrland* are notably different from others and so only results from these three regions are confidential.

6.3 Fast method for comparison of metrics

These experiments was performed to find out capabilities of mining information about HLA genetic diversity from the given data. Iterative method of genetic diversity computation was used for testing, adjusting and setting distance functions.

The reason why I used iterative method is that using *all to all* method, computation with high resolution metric is time demanding and at the initial adjustments of metrics it was important to know properties of the metric as soon as possible. In first iteration algorithm randomly chooses N individuals in each country and then mutual distance matrix is built from these data.

Next iterations keep doing still the same all round, but results are averaged into the

index	name	zip-code intervals	number of individuals
1	Ahvenanmaa	(220, 229)	143
2	Lappi	(930, 999)	724
3	Keski-Suomi	(350, 369) \cup (380, 449)	1318
4	Kuopio	(700, 739) \cup (760, 789)	1135
5	Oulu	(840, 929) \cup (740, 759)	1962
6	Vaasa	(620, 699) \cup (390, 399)	1500
7	Hame	(300, 349) \cup (370, 379)	1613
8	Kymi	(450, 559)	1420
9	Mikkeli	(100, 199)	1623
10	Pohjois-Karjala	(560, 619) \cup (790, 839)	1082
11	Turku Ja Pori	(200, 219) \cup (230, 299) \cup (380, 389)	1835
12	Uusimaa	< 100	6477

Table 6.2: Regions and zip-code intervals in Finland

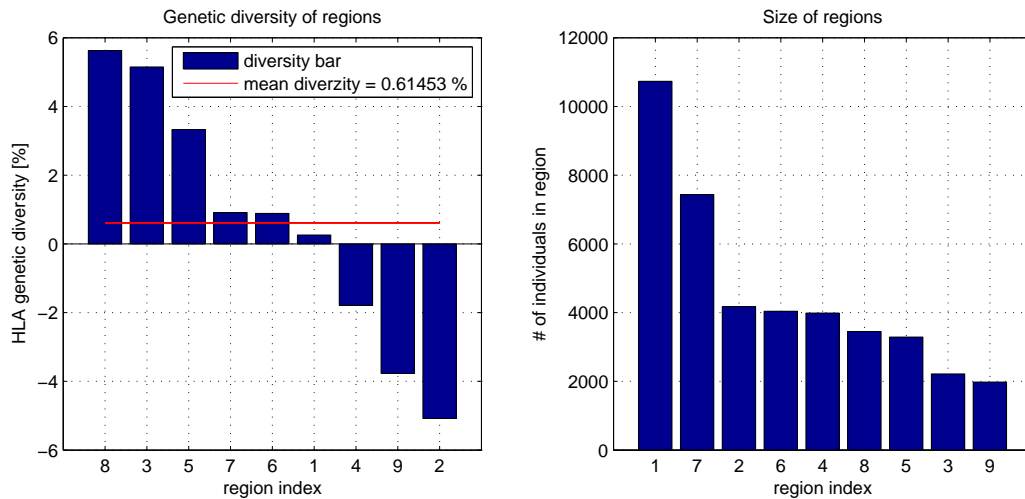


Figure 6.5: HLA genetic diversity of regions in Sweden

index	name	zip-code intervals	number of individuals
1	Greater Stockholm	(< 199)	10734
2	Skane	(200, 299)	4170
3	South-East	(300, 399)	2216
4	South-West	(400, 499)	3986
5	South-Central	(500, 599)	3290
6	West	(600, 699)	4040
7	North Central	(700, 799)	7432
8	Southern Norrland	(800, 899)	3452
9	Northern Norrland	(900, 999)	1981

Table 6.3: Regions and zip-code intervals in Sweden



Figure 6.6: Iterations of HLA genetic diversity using low resolution metric

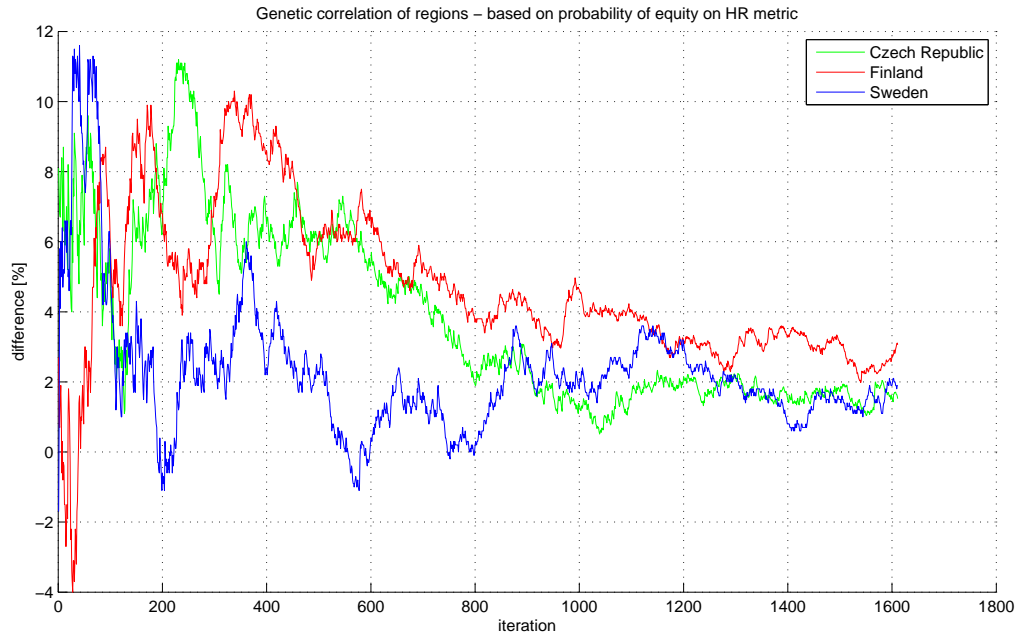


Figure 6.7: Iterations of HLA genetic diversity using high resolution metric

previous matrix. After infinity number of iterations matrix should converge to the stable configuration. After each iteration I_{div} was computed the results can be seen in figures 6.6 and figure 6.7. Advantage of this method is that it is possible to predict the final value in advance, stop the process and decide about the quality of tested metric.

As can be seen in figure 6.6, the value of all three national genetic diversities decrease below 0.2%. It is a weak indicator of diversity and outcomes using these metric are not representative enough.

Values from high resolution metric fluctuate from 1 to 4 percent of diversity which is much better and it is only matter of time when this fluctuation calms down. It was achieved approximately ten times higher ratio of HLA genetic diversity using metric based on high resolution probability of allele match.

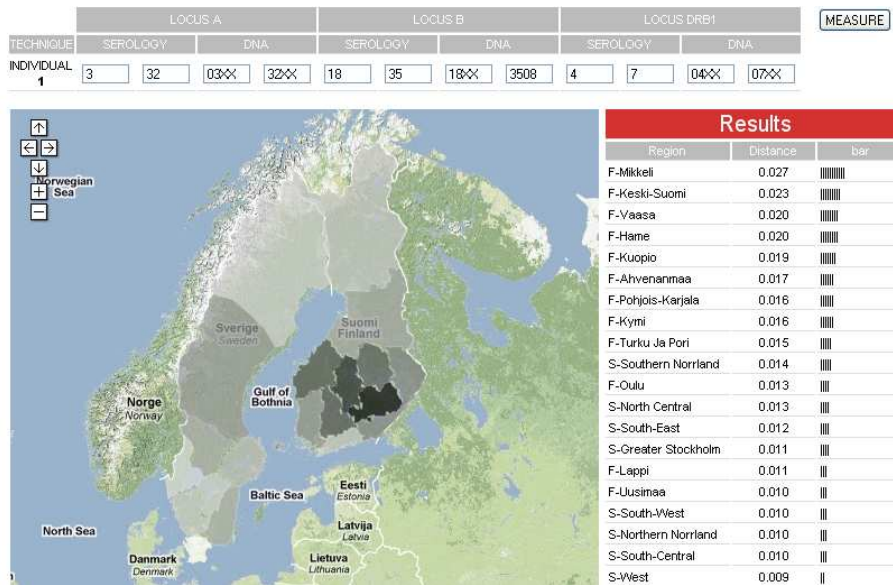


Figure 6.8: Searching for suitable donor

6.4 Interesting outcomes

The main goal of *GeoRelatives* application is providing easy way of HLA national registries exploration based on the geographical point of view. Experts need to know as much information as possible when they are deciding the right way of searching suitable donor for the patient and also the suitable way of registry extension. Let us foreshadow specific example now.

There is a patient in Finland with the given HLA phenotype living in *Keski-suomi* region (see figure 6.8). The expert will use *GeoRelatives* application to see the distribution of genetic relatedness with the patient.

As can be seen in figure 6.8 the gradient of HLA genetic relatedness spatially decreasing down out of the *Mikkeli* region. Note that the middle part of Sweden is involved by this phenotype as well. Now, focus your attention to the sorted list of Finnish and Swedish regions on the right side of the map. Almost all Finnish regions are ahead of Swedish regions which is expected, but not obvious. Various interpretations can be done using these geo-relative maps. It depends only on the particular assignment and application.

Chapter 7

Conclusion

Summarization of goals:

1. Study existing principles of HLA typing and matching
2. Design HLA metric suitable for geographical research of the data from stem cell donors registries
3. Create tool for easy access to the HLA phenotype genetic distribution over the geographical regions of national registries
4. Apply suggested methods to the Czech(Finnish, Swedish) registry database

First three points are discussed in previous chapters. Results of experiments applied to the particular registries are summarized in following section.

7.1 Results summarization

Summarization of final genetic diversities of all three investigated countries with their uncertainties.

metric/country	Czech Rep.	Finland	Sweden
LR - antigens mismatch	$0.07 \pm 0.06 \%$	$0.54 \pm 0.13 \%$	$0.12 \pm 0.07 \%$
HR - probability of match	$3.16 \pm 0.7 \%$	$5.97 \pm 0.4 \%$	$0.61 \pm 0.3 \%$

Table 7.1: Comparison of low and high metric results

Details of these results are in section 6.

Apparently finish registry is the most HLA diversified over the given geographical regions. Most likely it is caused due to the specific island *Ahvenanmaa* with very small population and with highly HLA diversified population.

7.2 Corrolary

The stated results prove that national stem cell donor registries are suitable for interpretation not only from the international point of view but it is also possible and useful to focus on the genetical relations of individuals living in particular regions. The experiments show that by the means of a suitable metrics it is possible to find correlations in relatively strongly imprecise data which enables to prove the presence of the searched genetical information.

For a given phenotype we can thus count genetical distances to particular groups of individuals and it can be further visualised in the form of geographical maps.

It is possible to suggest new metrics and try to maximalize the amount of the mined information in future. Successive possible specifications of the results lead to an even more concrete geographical localization of the suitable donor and to more detailed maps of geographical relativeness.

Currently, the system becomes a part of the international information system for donors of hematopotetic stem cells. After logging in to the system, each donor will have the possibility to see the map of HLA genetical relativeness for his unique phenotype. Among others, this service aims at rewarding the current and motivating the new donors of the registry.

Bibliography

- [1] Hla resources. Website, 2007. http://bioinformatics.nmdp.org/HLA/Haplotype_Frequencies/
- [2] Unrelated selection strategies. Website, 2007. http://bioinformatics.nmdp.org/STRATEGIES/Unrelated_Search_Selection/
- [3] Adobe - svg. Website, 2011. <http://www.adobe.com/svg/>
- [4] Bone marrow donors worldwide. Website, 2011. <http://www.bmdw.org>
- [5] Cord blood bank czech republic. Website, 2011. <http://www.bpk.cz>
- [6] Czech national marrow donors registry. Website, 2011. <http://www.kostnidren.cz>
- [7] Digitizer tool. Website, 2011. <http://www.birdtheme.org/useful/googletool.html>
- [8] Český národní registr dárců dřeně: Hledání dárců pro konkrétní nemocné a realizované transplantace 1992 – 2005. Website, 2011. <http://www.kostnidren.cz/registr/vysledky/uspesnost.php>
- [9] Český registr dárců krvetvorných buněk. Website, 2011. <http://www.czechbmd.cz/pro-odborniky.php>
- [10] FileInfo.com. Csv file extension. Website, 2011. <http://www.fileinfo.com/extension/csv>
- [11] Google. Google maps api family. Website, 2011. <http://code.google.com>
- [12] Anthony Nolan. marrow.org. Website, 2011. <http://hla.alleles.org/nomenclature/naming.html>
- [13] Anthony Nolan. Wmda directiry. Website, 2011. <http://hla.alleles.org/wmda/index.html>

- [14] Medline plus trusted health information. Serology. Website, 2011. <http://www.nlm.nih.gov/medlineplus/ency/article/003511.htm>.
- [15] BRAY, R, HURLEY, C., KAMANI, N., WOOLFREY, A., and MULLER, C. National marrow donor program hla matching guidelines for unrelated adult donor hematopoietic cell transplants. *Biology of Blood and Marrow Transplantation*, 14:45–53, 2008.
- [16] BURROUGH, P and MCDONNELL, D. *Principles of Geographical Information Systems*, volume 9. 2010.
- [17] FLOMENBERG, N, BAXTER-LOWE, LA, and CONFER, DL. Impact of hla-class i and class ii high resolution matching on outcomes of unrelated donor bmt. *Blood*, 98:813–815, 2001.
- [18] GONZALEZ-GALARZA, FF, CHRISTMAS, S, and MIDDLETON, D. Allele frequency net: a database and online repository for immune gene frequencies in worldwide populations. *Nucleic Acid Research 2011*, 39:913–D919, 2011. <http://www.allelefrequencies.net/>.
- [19] HALPERIN, ERAN and HAZAN, ELAD. Haplofreq - estimating haplotype frequencies efficiently. pages 2–4, 2006.
- [20] JIN, LI and MASATOSHI, NEI. Relative efficiencies of the maximum-parsimony and distance-matrix methods of phylogeny construction for restriction data. *Molecular Biology Evolution*, (8(3)):356–365, 1991.
- [21] KÁBRT, LUKÁŠ. Master thesis - hla genetická příbuznost česků s ostatními národy, 2009.
- [22] MARSH, SGE, ALBERT, ED, and BODMER, WF. Nomenclature for factors of the hla system. *International Journal of Immunogenetics*, 66:571–636, 2005.
- [23] MARSH, STEVE. Evaluating registry diversity - presentation. *Anthony Nolan Research Institute*, pages 3–45, 2007. Attached CD / reference / marsh.pdf.
- [24] MAYO, O. A century of hardy–weinberg equilibrium. *Twin Research and Human Genetics*, 11:249–256, 2008.
- [25] MIDDLETON, D, MENCHACA, H, and ROOD, R. A new allele frequency database. *Scientific communication*, 1:112–113, 2010.

- [26] PETERSDORF, EW, GOOLEY, TA, and ANASETTI, C. Optimizing outcome after unrelated marrow transplantation by comprehensive matching of hla class i and ii alleles in the donor and recipient. *Blood*, 92:3515–3520, 1998.
- [27] SASAZUKI, T, JULI, T., MORISHIMA, Y., WOOLFREY, A., and MULLER, C. Effect of matching of class i hla alleles on clinical outcome after transplantation. *N Engl J Med.*, 339:1177–1185, 1998.
- [28] SCHREUDER, GMTH, HURLEY, CK, and MARSH, SGE. The hla dictionary 2008. *Tissue Antigens*, (73):95–170, 2009.
- [29] SIMONSEN, M, MAILUND, T, and PEDERSEN, C. Rapid neighbour-joining. *Springer-Verlag Berlin Heidelberg*, pages 113–122, 2008.
- [30] STEINER, DAVID. Master thesis - search for unrelated bone marrow donors, 2007.
- [31] Wikipedia. Dna profiling. Website, 2011. http://en.wikipedia.org/wiki/DNA_profiling.
- [32] Wikipedia. Global positioning system. Website, 2011. http://en.wikipedia.org/wiki/Global_Positioning_System.

Appendix A

GeoRelatives user manual

A.1 Layout

Layout of the page consists of three blocks. *Menu* on the left, *Form* at the top and *Content* in the center. See figure A.1

A.2 Menu

Menu consists of three blocks. *Mode* at the top, *Country* in the middle and *Metric* at the bottom.

A.2.1 Mode

Distance function check

In this mode, a user can check results from the measurements of any two HLA phenotypes. It is possible to type both phenotypes manually or insert ID of the individual and let the record be loaded from the database. After both lines are filled properly, button *measure* can be pressed and application will show the entire process of distance computation. In case of high resolution metric it shows reduction, downgrading, breakdown, intersection and recombination of particular loci results. For low resolution metric only reduction and counter of antigens match is shown. These outcomes are used for constructing maps and distance matrices. Check mode can be used for testing or additional improving of suggested metric.

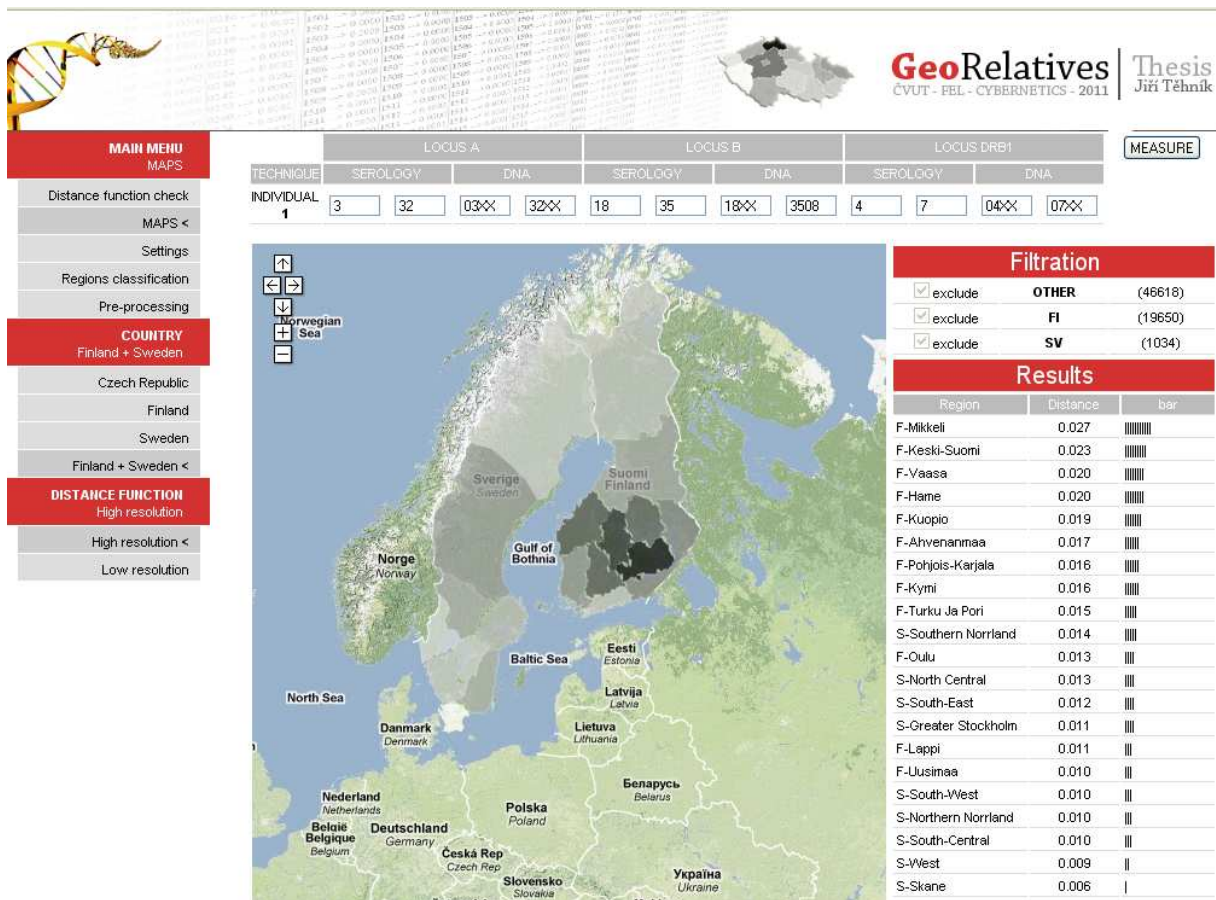


Figure A.1: GeoRelatives visual layout and control panels

Maps

The most important part of *GeoRelatives* application. User can see geographical map of selected country with HLA genetic distribution to the given phenotype. The map is divided to the particular regions and these areas are proportionally colored according to the genetic distance to the given phenotype. Phenotype can be typed manually or using numerical identifier (ID) of the individual in selected country. On the right side of the map there is a sorted list of regions with the averaged HLA genetic distance from the given phenotype/individual. In case that ID is used, the region from which the individual comes is highlighted.

Settings

Basic settings for the visualization of the map and for the zip-intervals of regions. There is a small blue icon with help and additional information of zip-code intervals adjustment.

Check of regions classification

The list of all the regions with index, size and the set of IDs. These sets content individuals which was classified into the region due to the given zip-code intervals.

Preprocessing

The run of preprocessing consists of two phases. First, the accumulative list of different records at each loci is created. Accumulative means that it holds the number of frequency and these records are sorted from the most frequent one. Then triangular mutual distance matrix of x most frequent records is created in phase 2. Preprocessed are only the most often alleles and the rest is compute in real time. For alleles/records which are present seldom, preprocessing doesn't make any sense.

A.2.2 Country

In this block of menu, user can select the country to be investigated. The name of each country is attached with the national HLA registry inside of the program. When the country is selected ones, phenotype is then measured with all the individuals of the relevant national registry. Individuals are classified to the regions, eventually distance matrix is built.

A.2.3 Metric

There are two metrics (distance functions) we handle with in this work. All the computation depending of the selected metric. Both metrics are described in chapter 4.

Appendix B

Attached CD

Source codes to *GeoRelatives* online application and the whole text of the thesis in PDF format can be found on the CD attached. Files are sorted in following directories.

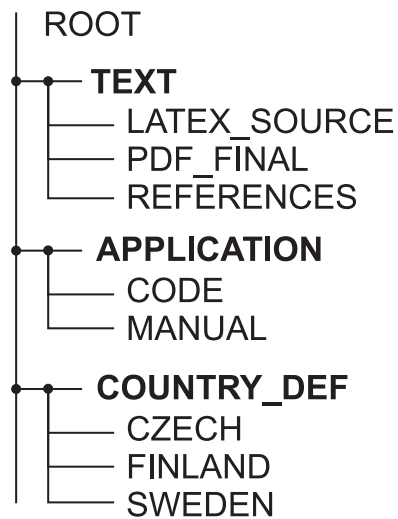


Figure B.1: directory tree

All the cited sources in this work are to be found in directory *references*.

Appendix C

Abbreviations and terms

National stem cell donor registries

In this thesis the anonymised databases of unrelated stem cell donor registries is analyzed. It contains HLA data, zip-codes and optionally some additional attributes. Exact structure of the data is shown in section 5.2.1.

Chromosome

A chromosome is an organized building of DNA and protein that is found in cells. It is a single piece of coiled DNA containing many genes, regulatory elements and other nucleotide sequences. Chromosomes also contain DNA-bound proteins, which serve to package the DNA and control its functions. The word chromosome comes from the Greek (chroma - color) and (soma - body) due to their property of being very strongly stained by particular dyes. Chromosomes vary widely between different organisms. In a series of experiments, Theodor Boveri gave the definitive demonstration that chromosomes are the vectors of heredity.

For the purpose of this work is sufficient to notify that every individual and also record in database have always two alleles at each locus, because two connected chromosomes are present. This is why problem with unphasing have to be solved.

Locus

Specific location of a gene or DNA sequence on a chromosome.

Allele

A variant of the DNA sequence at a given locus. One of two or more forms of the DNA sequence of a particular gene.

Allele frequency

Occurance rate of the allele in entire population compared to all other alleles.

Homozygotes vs. heterozygotes

Most organisms have two sets of chromosomes, that is, they are diploid. Diploid organisms have one copy of each gene (and one allele) on each chromosome. If both alleles are the same, they are homozygotes. If the alleles are different, they are heterozygotes.

Antigens

Molecule recognized by the immune system. Originally the term came from antibody generator and was a molecule that binds specifically to an antibody, but the term now also refers to any molecule or molecular fragment that can be bound by a major histocompatibility complex (MHC) and presented to a T-cell receptor.

Match and mismatch

These terms are used to express equity resp. difference of two values. Refferential resolution must be allways set to be able to decide whether inputs *match* or *missmatch*. It is basically binary operator with inputs I_1, I_2 and *resolution* and binary output *MATCH* or *MISMATCH*. For example alleles $A * 01 : 01$ and $A * 01 : 02$ don't match (mismatch) at the high resolution, but they match at the low resolution, when $01 = 01$.

Probability of match

The probability of two alleles match at the particular level of resolution. It calculates with decomposition of lower resolutions to the probabilistic distribution of its finer subsets.

To get the result intersections of two distributions are combine. (For more see section 4.3.7).

Distance

The rate of dissimilarity between two phenotypes, two loci. Depends on particular metric how it can be interpreted or which units are used. For the purpose of this work, the number of loci mismatches is used as a distance between two phenotypes.

Similarity

The rate of similarity between two phenotypes, two loci. Depends on particular metric how it can be interpreted or which units are used. For the purpose of this work, the probability of phenotype high resolution match is used as a genetic similarity between two individuals.

Relativeness

Relativeness and *Similarity* are equal terms in this work. Term *Relativeness* only point out that genetic similarity has something to do with genetic relativeness and that is why word *Relativeness* is sometimes preferred.

Divergence

Genetic divergence is property of the region comparing to all other regions. It express if individuals in particular region are somehow different from the rest of the country. In the cases of this work the higher divergence is, the closer are phenotypes to each other in the particular region in compare to the other regions.

Diversity

Genetic diversity can be also called *global divergence* and it si indicator for the entire country or registry. The higher *Diversity* is, the better are similar individuals clusterd

into the regions. The *Diversity* approaches to the zero when individuals are randomly classified.

Genotype

The genotype is the genetic constitution of an individual.

Phenotype *in biology*

A phenotype is any observable characteristic or trait of an organism such as its morphology, development, biochemical or physiological properties or behavior. Phenotypes result from the expression of an organism's genes as well as the influence of environmental factors.

Phenotype *in this work*

Complete description of individual represents by record in given HLA database consists of locus A, B and DRB1.

Individual

In this work both terms *individual* and *phenotype* have the same meaning, because there is only information about phenotype of the individual in the given data.

Haplotype

Combination of alleles at different places on the chromosome that are transmitted together. A haplotype may be one locus, several loci, or an entire chromosome depending on the number of recombination events that have occurred between a given set of loci.

Haplotype frequency

The probability of occurrence of particular combination of alleles at the different loci in the population.

DNA

Deoxyribonucleic acid - is a nucleic acid that contains the genetic instructions used in the development and functioning of all known living organisms with the exception of some viruses. The main role of DNA molecules is the long-term storage of information. DNA is often compared to a set of blueprints, like a recipe or a code, since it contains the instructions needed to construct other components of cells, such as proteins and RNA molecules.

DNA sequence

Succession of letters representing a real or hypothetical DNA molecule or strand. The sequence has capacity to carry information and can be read from the biological raw material through DNA sequencing methods.

MHC

The major histocompatibility complex is a large genomic region or gene family found in most vertebrates that encodes MHC molecules. MHC molecules play an important role in the immune system and autoimmunity.

HLA

Human leukocyte antigen - the name of the major histocompatibility complex (MHC) in humans. The super locus contains a large number of genes related to immune system function in humans. This group of genes reside on chromosome 6, and encode cell-surface antigen-presenting proteins and many other genes. The HLA genes are the human versions of the MHC genes that are found in most vertebrates (and thus are the most studied of the MHC genes).

Typing

The process of acquiring genetic information and its description in form of alleles (DNA) or antigens (Serology).

Methods for HLA typing

There are two main techniques how to gain MHC/HLA genetic sequence. Identification of antibodies in blood serum called serology [14] and more accurately DNA analysis [31].

Classification

For the purpose of this thesis term classification means allways the way of sorting individuals to the particular region using zip-codes.

Serology

Serology is the scientific study of blood serum and other bodily fluids. In practice, the term usually refers to the diagnostic identification of antibodies in the serum. Such antibodies are typically formed in response to an infection, against other foreign proteins, or to one's own proteins.

Linkage equilibrium

Non-random association of alleles at two or more loci, not necessarily on the same chromosome.

Patient

Someone, who have been HLA typized to find suitable donor. Patient is usually screend using DNA techniques to get as many HLA information as possible. If ones someone needs a transplantation, time and accuracy plays the main role since that time. Thats the reason why patient are screend with the latest and the most expensive techniques.

Donor

Volounteer, who have been HLA typed and who offer onw cells transplantation for potential patient in the future. Donors have been usually typized serologically, but nowadays DNA screening becomes available for donors as well.

Individual

Someone, whose phenotype is known (patient or donor). Each one record in the main HLA database is one individual.

Region

Closed geographical area. Most often province of the nation. Also called sub-region to point out that it is a part of bigger complex, called super-region.