Západočeská univerzita v Plzni
Fakulta aplikovaných věd
Katedra kybernetiky

BAKALÁŘSKÁ PRÁCE

PLZEŇ, 2014 **ADAM VONÁŠEK**

ZÁPADOČESKÁ UNIVERZITA V PLZNI
Fakulta aplikovaných věd
Akademický rok: 2013/2014

ZADÁNÍ BAKALÁŘSKÉ PRÁCE

(PROJEKTU, UMĚLECKÉHO DÍLA, UMĚLECKÉHO VÝKONU)

Jméno a příjmení:	**Adam VONÁŠEK**
Osobní číslo:	**A11B0730P**
Studijní program:	**B3918 Aplikované vědy a informatika**
Studijní obor:	**Kybernetika a řídicí technika**
Název tématu:	**Citlivostní analýza metod identifikace haplotypů**
Zadávající katedra:	**Katedra kybernetiky**

Zásady pro vypracování:

1. Seznamte se s problematikou HLA systému.
2. Seznamte se s metodami identifikace haplotypů z fenotypových dat užívanými v biomedicínském výzkumu.
3. Implementujte vybrané metody a proveďte citlivostní analýzu (zaměřte se na současné možnosti a limity jednotlivých metod).
4. Proveďte zhodnocení výsledků.

Rozsah grafických prací: **dle potřeby**

Rozsah pracovní zprávy: **30-40 stránek A4**

Forma zpracování bakalářské práce: **tištěná**

Seznam odborné literatury:

Dodá vedoucí bakalářské práce.

Vedoucí bakalářské práce: **Ing. Lucie Houdová**
Katedra kybernetiky

Datum zadání bakalářské práce: **1. listopadu 2013**

Termín odevzdání bakalářské práce: **16. května 2014**

Doc. Ing. František Vávra, CSc.
děkan

L.S.

Prof. Ing. Josef Psutka, CSc.
vedoucí katedry

V Plzni dne 1. listopadu 2013

P R O H L Á Š E N Í

Předkládám tímto k posouzení a obhajobě bakalářskou práci zpracovanou na závěr studia na Fakultě aplikovaných věd Západočeské univerzity v Plzni.
Prohlašuji, že jsem bakalářskou práci vypracoval samostatně a výhradně s použitím odborné literatury a pramenů, jejichž úplný seznam je její součástí.

V Plzni dne 15. května 2014

.....
vlastnoruční podpis

PODĚKOVÁNÍ

Děkuji vedoucí bakalářské práce Ing. Lucii Houdové, Ph.D. a Ing. Miloši Fetterovi za odborné vedení, pomoc a podporu při vypracování bakalářské práce.

Abstrakt

Práce se zabývá citlivostní analýzou metod pro identifikaci haplotypů. V teoretické části jsou obsaženy základní poznatky o transplantaci krvetvorných buněk, reakcích imunitního systému na ni, informace o HLA ("Human leucocyte antigens") znacích, způsobech typizace HLA znaků, a teoretický rozbor metod identifikace haplotypů, tj. Clarkova, EM ("Expectation-Maximization") a Bayesova algoritmu. Praktická část je věnována implementaci dvou zvolených algoritmů v prostředí programovacího jazyka Java, jejich výsledkům a zhodnocení, co se týče předností a nedostatků, které jsme stanovili na základě implementace. Na konci práce se nachází výkladový slovník pojmů.

Klíčová slova

transplantace kostní dřeně, krvetvorné buňky, HLA, HLA typizace, alela, haplotyp, genotyp, Clarkův algoritmus, EM algoritmus.

Abstract

The work deals with a haplotype identification methods sensitivity analysis. A theoretical section should provide basic facts about haematopoietic stem cells transplantation, following reactions of immune system, HLA ("Human leucocyte antigens") types, methods for typing and a theoretical analysis of the haplotype identification methods, i.e. Clark, Expectation-Maximization and Bayes algorithm. A practical section focuses on two elected methods implementation using the computer programming language Java, results and an evaluation on advantages and disadvantages derived by the implementation. There is a monolingual dictionary at the end of the work.

Keywords

bone marrow transplantation, haematopoietic stem cells, HLA, HLA typing, allele, haplotype, genotype, Clark algorithm, EM algorithm.

Obsah

Seznam obrázků

Seznam tabulek

1 Úvod

Během různých chorob dochází v lidském těle ke snížení krvetvorby. Současná medicína je schopna odstranit mnoho nemocí, ale pokud nastane porušení krvetvorby natolik vážné, že jej nelze vyléčit běžným způsobem, přichází na řadu možnost vyměnit celou krvetvornou tkáň za novou pomocí transplantace. Transplantací kostní dřeně rozumíme postup, kdy do těla nemocného umisťujeme zdravé krvetvorné buňky, které substituují krvetvorbu buněk pacientových. Samotná transplantace představuje poměrně jednoduchý lékařský zákrok. Její problém spočívá v reakci imunitního systému na transplantované buňky a nutnosti brát v potaz tkáňové HLA znaky, které s sebou krvetvorné buňky nesou. Z důvodu velkého množství kombinací tkáňových znaků se tak hledání vhodného dárce velmi komplikuje.

Všechny zděděné znaky jedince tvoří dohromady fenotyp. Ten je nutno identifikovat. Tomuto procesu se říká HLA typizace. Rozlišujeme sérologickou typizaci, kdy je pozorována reakce protilátek na buněčné membráně a modernější variantu, molekulárně genetickou (DNA) typizaci, kdy se HLA alely určují přímo. Molekulárně genetická typizace může probíhat na nízké, střední a vysoké úrovni rozlišení. K určování HLA znaků používáme názvosloví, tzv. nomenklaturu.

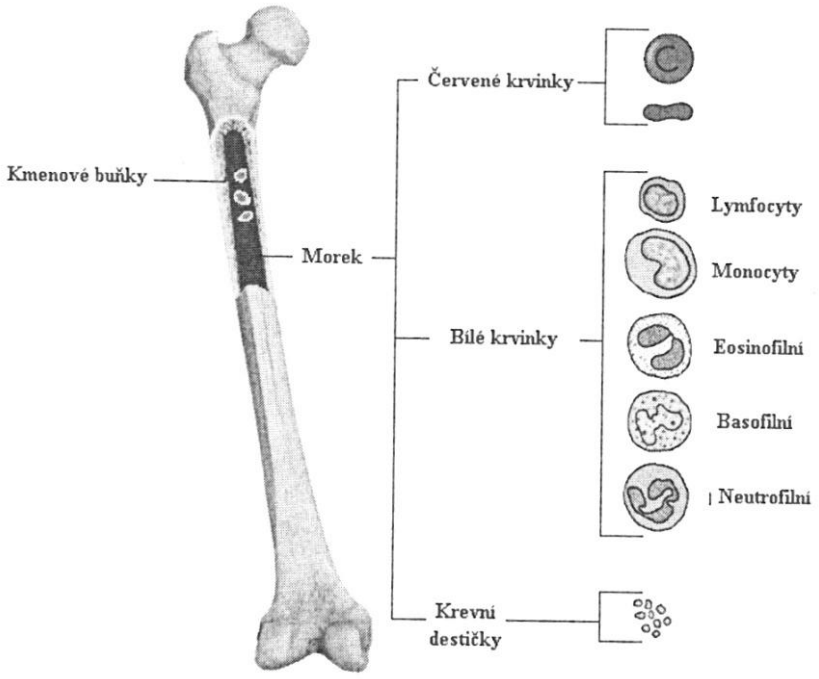
V rámci všech dědičných znaků u člověka definujeme jednotky zvané „haplotypy“. Haplotyp je sada alel, které se dědí společně a jsou důležité pro imunitní systém. [1] Není možné je laboratorně získat a musíme je proto odhadnout z fenotypových dat za pomoci speciálních metod. Cílem práce je zhodnotit funkčnost těchto metod. Uveďme tři nejpoužívanější metody: Clarkův, EM a Bayesův algoritmus. Aby bylo možné tyto algoritmy spouštět, máme k dispozici fenotypová data několika tisíců dárců z České republiky.

2 Transplantační imunologie

Hlavními tématy této kapitoly jsou: jak se imunitní systém vypořádá s transplantátem (neboli štěpem), jak může imunitní systém odmítnout transplantát a jaké existují možnosti, že bude zabráněno odmítnutí transplantátu. [2]

2.1 Kostní dřeň

Kostní dřeň je krvetvorný orgán. Váží přibližně 2,6 kg. Jedná se o rosolovitou tkáň tvořící výplň dřeňové dutiny kostní. Její nejdůležitější činností je krvetvorba („hematopoéza“). Ta se realizuje pomocí tzv. hematopoetických kmenových buněk HSC ("Haematopoietic stem cells"), které pocházejí z jater lidského embrya. S koncem embryonálního vývoje se HSC přesouvají do nově vznikající kostní dřeně. Ze zárodečných HSC buněk vznikají tři hlavní typy krvinek - červené, bílé a krevní destičky. Mozková centra provádějí kontrolu počtu krevních buněk vyskytujících se v těle. Při ztrátě některého z typů posílá nervové centrum signál do kostní dřeně a ta je připravena krvetvorbu zvýšit. [3]



Obr. 1: Kostní dřeň. [1]

2.2 Transplantace kostní dřeně

Zdravé lidské tělo pravidelně reaguje na ztrátu nebo zvýšenou potřebu kteréhokoli z typů krvinek. Použitím speciálních mezibuněčných působků bílkovinné povahy, které jsou nazývány „růstové faktory“, vyšle organismus signál do kostní dřeně a ta ihned nastartuje zrychlenou tvorbu. Nedospělé krvinky se ještě několik dní zdrží v kostní dřeni, než úplně dozrají. Pak hotové a zralé se přesouvají do krevního oběhu, aby tam plnily své funkce. Pokud se v těle nachází dostatek krvinek, jejich množení se opět utlumí.

Samotné kmenové buňky se objevují v krvi za běžného stavu jen ojediněle. Jejich vyplavení do krevního oběhu se může ovšem za určitých okolností dočasně zvýšit, např. v době rychlé regenerace krvetvorby po útlumu, který je obvyklým jevem u pacientů krátce po odeznění účinků proti-nádorových léků. Vyplavení kmenových buněk je možné vyvolat i uměle a to podáváním růstových faktorů ve formě léků.

Přirozenou regulaci krvetvorby přerušují různé choroby. Současná medicína umožňuje úspěšně bojovat s řadou dříve neodstranitelných nemocí, ovšem pokud je základní porucha krvetvorby natolik vážná, že nemůže být vyléčena jiným, jednodušším způsobem, nabízí se volba vyměnit kompletně krvetvornou tkáň za zdravou pomocí transplantace. Její technické provedení je relativně jednoduché. Krvetvorné buňky jsou tekuté, tím pádem mají schopnost volně se pohybovat v krvi. To umožňuje dostat je do organismu příjemce velmi snadno. [4]

Transplantace kostní dřeně je postup léčení dané diagnostikované nemoci. Jde o léčebný zákrok, kdy do těla nemocného vpravujeme zdravé krvetvorné buňky, které nahrazují produkci buněk pacientových. Krvetvorné buňky je zaprvé možno odebírat přímo z kostí, tzn. zapuštěním jehly do horních lopat kostí kyčelních do úrovně v okolí dřeňové dutiny. Druhou variantou je odsátí krvetvorných buněk z periferní krve, tzn. odběr přímo z krevního oběhu. Třetí varianta je odběr pupečníkové krve z pupečníku během porodu. Tento druh transplantace je velmi výhodný v tom, že je kostní dřeň odebírána v tekutém stavu a k úspěšné transplantaci postačuje jen nepatrné množství kmenových buněk. [1]

2.2.1 Druhy transplantací:

- **Autologní transplantace kmenových buněk -** principem je získání kmenových buněk od samotného pacienta. Ty jsou pak transplantovány. Tento způsob je prováděn většinou jako pojistka následně po vysoko-dávkové chemoterapii různých nádorů. S podpůrnou transplantací vlastními kmenovými buňkami se skýtá možnost podat mnohem vyšší dávky účinných léků než bez ní. Podmínkou však je, aby byly pacientovy základní krvetvorné buňky před odběrem zdravé.
- **Alogenní transplantace dřeně -** kmenové buňky jsou získávány od jiného, zdravého člověka. Je indikována hlavně tehdy, pokud se u pacienta prokáže kritické poškození základních kmenových buněk a není možné je vyléčit jiným, bezpečnějším způsobem. V porovnání s autologní transplantací má však tato daleko širší účinky. Kostní dřeň v tomto případě slouží jako náhrada nemocné krvetvorby a zároveň může za určitých okolností vyvolat proti-nádorový efekt, který souvisí s reakcí bílých krvinek dárcovského původu proti nedoléčeným zbytkům původní nemoci u pacienta. Nejčastějšími důvody použití tohoto druhu transplantace jsou: prognosticky nepříznivé typy leukémií, vážné krvetvorné útlumy, těžké vrozené poruchy imunity a látkové přeměny u dětí. Nicméně okruh nemocí se stále rozšiřuje i na další. Základní podmínkou je dostupnost vhodného dárce, jehož buňky jsou schopné se s tělem pacienta trvale spojit. To je možné pouze při úplné či velmi blízké shodě tzv. tkáňových (HLA) znaků (definice HLA v kapitole 3.1) dárce a příjemce krvetvorných buněk. Přivykání na cizí dřeň je dlouhodobý proces trvající v řádech měsíců. Přes všechny potíže pro pacienta, které s sebou nese, znamená transplantace dřeně od zdravého dárce u řady nemocí nejspolehlivější způsob léčby. [4]
- **Syngenní transplantace -** od jednovaječného dvojčete.
- **Xenogenní transplantace** - od příslušníka jiného živočišného druhu. [1]

2.3 Výběr vhodného dárce

Složitost procesu transplantace kostní dřeně nespočívá v samotném vpravení krvetvorných buněk do těla, ale v následující reakci imunitního systému. [1] Účel bílých krvinek je takový, že brání svůj organismus proti všemu, co do těla nepatří. Mají receptor a při průchodu organismem jím prověřují podezřelé buňky. Pokud se taková objeví, ihned se ji snaží zneškodnit. Při transplantaci se tedy musíme řídit tkáňovými (HLA) znaky, které se nacházejí ve všech buňkách. Bílá krvinka ověřuje podezřelé buňky na zmíněné znaky. Pokud tyto znaky s pacientovými do určité míry odpovídají, buňku ignoruje. V opačném případě tělo transplantované buňky vůbec nepřijme anebo (pokud dojde k menší neshodě) se transplantované krvinky uchytí, ale protože považují tělo svého nositele za cizí, rozhodnou se jej ničit.

Najít pro pacienta vhodného dárce je velice obtížné, protože různých kombinací HLA znaků jsou tisíce. [1]

Není-li k dispozici jakýkoli vhodný dárce z rodiny a existuje-li indikace k nepříbuzenské transplantaci dřeně, je započato hledání dobrovolného dárce dřeně z registru. Aby transplantace proběhla úspěšně, musí mít dárce a pacient co nejpodobnější tkáňový typ (soubor HLA znaků). Čím větší je v nich rozdíl, tím je u pacienta vyšší riziko, že nastanou pozdější imunitní reakce. Existuje stupeň HLA neshody, při kterém již nelze transplantaci provést. Nejsnazší možnost z důvodu dědičnosti tkáňových znaků, je hledat dárce mezi pokrevním příbuzenstvem. Z technického hlediska považujeme za vhodné, aby dárce a příjemce měli také stejnou krevní skupinu. Nezbytně nutná podmínka transplantace to však není. Krevní skupina má totiž význam jen u zralých červených krvinek. V případě, že by převod dřeně s krví jiné skupiny byl pro pacienta riskantní (podobně jako nestejnorodá transfúze), červené krvinky se z odebrané dřeně jednoduše oddělí a nemocnému je podáván zbývající koncentrát kmenových buněk. Častěji jsou kmenové buňky od dárce odebírány použitím separátoru, kde příměs červených krvinek je v minimálním množství. Výsledkem transplantace potom je, že pacientovi po určité době nastane trvalá změna jeho krevní skupiny. [4]

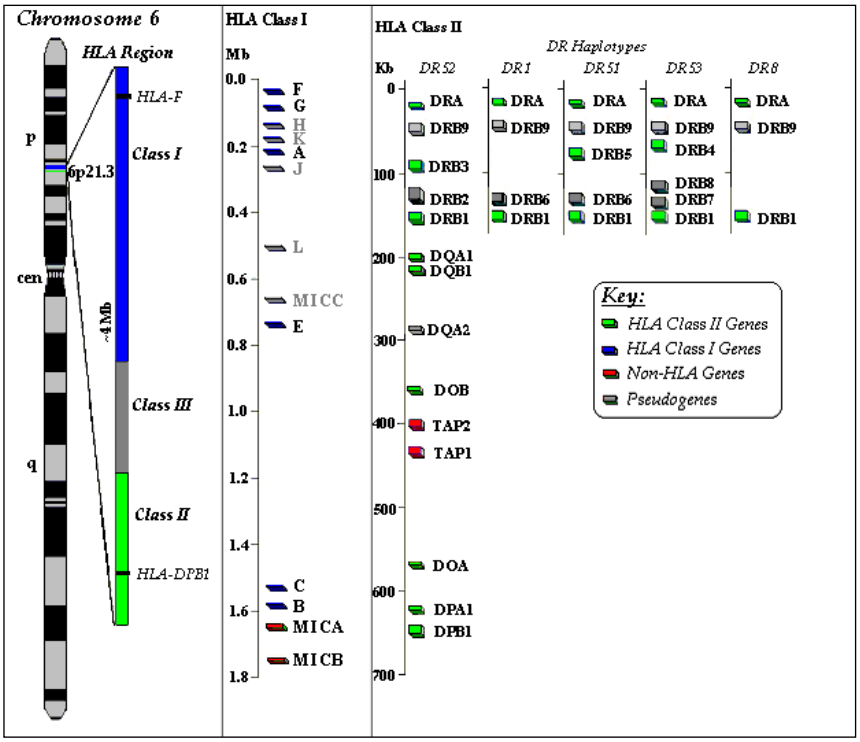
3 HLA, MHC, HLA typizace, haplotyp, fenotyp, genotyp

Tématy další kapitoly jsou HLA antigeny, hlavní histokompatibilní komplex MHC, způsoby typizace HLA znaků, úvod do nomenklatury a definice pojmů haplotyp, fenotyp a genotyp.

3.1 HLA antigeny

Zkratka HLA má význam "Human leucocyte antigens" neboli „lidské leukocytové antigeny“. Tyto antigeny byly poprvé objeveny na leukocytech, ale jinak se nacházejí na povrchu všech jaderných buněk. [5]

Jedná se o skupinu bílkovin, která má klíčovou funkci v imunitním systému člověka. [1] Tyto bílkoviny jsou lokalizovány jako antigeny na povrchu buněk. Jsou to polypeptidové produkty skupiny genů nazývaných „hlavní histokompatibilní komplex“ (MHC). Byly dosud definovány pouze u obratlovců. Pro každý druh obratlovců se pak geny tohoto komplexu nazývají druhově specifickým jménem. Tzn. pro lidi jde o HLA a např. u psů mluvíme o DLA ("Dog leucocyte antigens"). Co se však týče funkčního hlediska, jde stále o stejnou skupinu genů. [6]



Obr. 2: Genetická organizace HLA systému. [6]

HLA molekuly spolu s imunoglobuliny a receptory T-lymfocytů tvoří základní obranyschopnost člověka. Jejich funkce v antigen-specifické imunitní odpovědi spočívá

v tom, že T-lymfocyty, jakožto klíčová regulační a ejektorová složka imunitního systému, jsou schopny poznávat antigenní fragmenty a reagovat na ně pouze v rámci HLA molekul, tedy vážou li-se tyto antigeny na molekuly HLA sytému. [6]

3.2 Funkční a strukturální rozdělení HLA antigenů

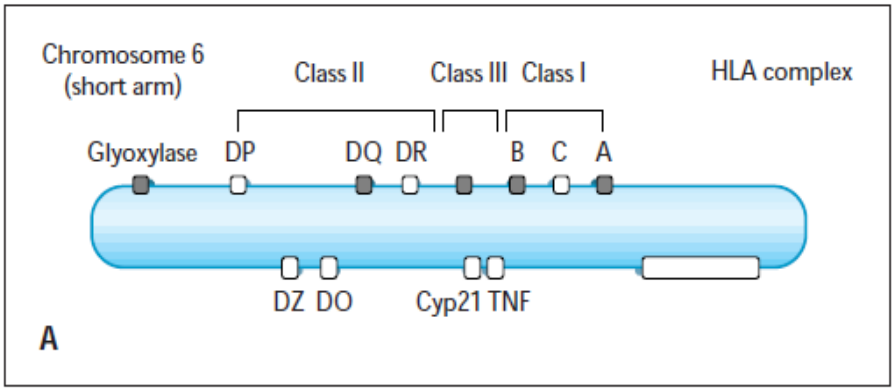
Ze strukturálního a funkčního hlediska lze HLA u genů a molekul (antigenů) rozlišit dvě základní skupiny: HLA (anti)geny I. a II. třídy. [6] Dle poslední aktualizace nomenklatury provedené v roce 2012 WHO je evidováno již více než 5500 alel I. třídy a 1600 alel II. třídy. [1]

3.3 MHC

MHC ("Major histocompatibility complex") neboli hlavní histokompatibilní komplex u lidí je skupina 40 – 50 genů, které jsou seřazené v dlouhém úseku DNA na krátkém raménku 6. chromozomu. Právě pouze u lidí tento soubor nazýváme HLA komplex. MHC geny jsou organizovány do oblastí, které kódují tři třídy MHC molekul. První a druhá oblast zahrnují geny, které kódují HLA antigeny. I. a II. třídy, III. třída obsahuje produkty související s imunitními procesy. [1] Důležité je uvést, že pojmy HLA a MHC bývají používány synonymně a nekonzistentně. [6]

Izotopy v jednotlivých třídách MHC:

- MHC I : HLA-A, -B, -C, klasické izotopy
  HLA -E, -F, -G: neklasické izotopy mající ve srovnání s klasickými geny omezenou buněčnou distribuci a nižší stupeň polymorfismu, z těch HLA-E a HLA-G: pomáhají k toleranci plodu v děloze
- MHC II: HLA -DR, -DQ, -DP, klasické izotopy
  LMP2, LMP7: bílkoviny štěpící proteiny pro HLA I. třídy
  TAP: membránové transportéry
- MHC III: C4, C2, TNF (faktor nekrotizující tumory), ... [1]



Obr. 3: Krátké raménko 6. chromozomu. [1]

3.4 Polymorfismus antigenů

Antigeny jsou velmi polymorfní (neboli variabilní), to znamená, že v lokusu (definice viz výkladový slovník) určitého genu na daném místě DNA, kde se uvedený gen nachází, je možné najít jeho různé varianty. Těm se říká „alely“. K tomu aby měl gen možnost se fenotypově projevit (tzn. aby se přepsal do stanoveného produktu), musí jeho lokus obsahovat alespoň dvě alely - jednu původem od matky a jednu otcovu. [1]

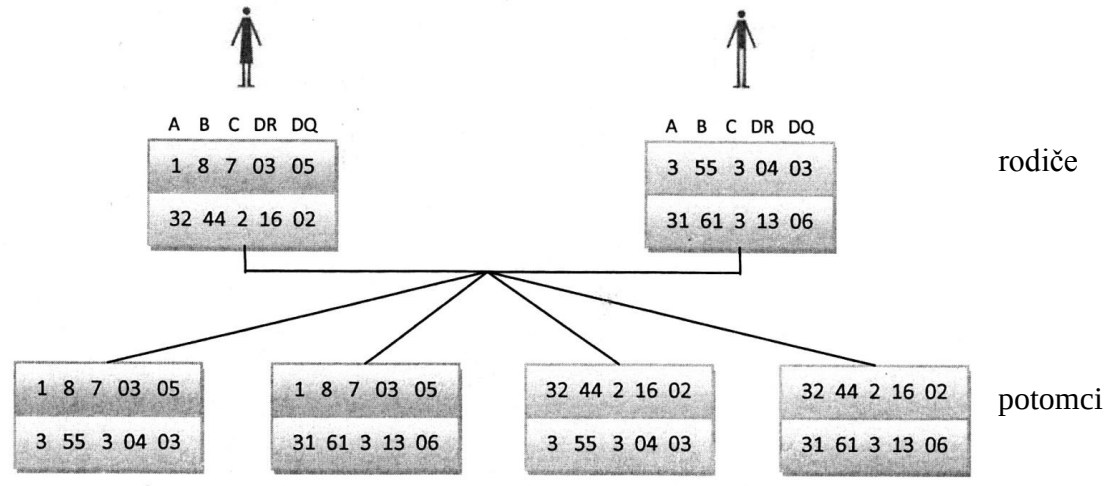
MHC lokusy obecně jsou geneticky nejvariabilnějšími kódujícími lokusy u savců, a lidské HLA lokusy nejsou výjimkou. [6]

3.5 Dědičnost HLA znaků

Všechny geny HLA systému na 6. chromozomu se dědí vcelku jako blok - tzv. haplotyp (definice v kapitole 3.10). Geny HLA systému se vyznačují poměrně těsnou genetickou vazbou, takže se většinou přenáší jako bloky genů, mezi kterými pouze výjimečně nastává rekombinace ("Crossing-over"). Dle rodinných studií je relativní absence rekombinací velmi typická pro některé regiony uvnitř HLA komplexu. Hovoříme především o subregionech HLA-B k HLA-C a HLA-DQA1 k HLA-DRB1.

Každý člověk má tzv. fenotyp (definice v kapitole 3.11), tj. soubor HLA znaků skládající se ze dvou haplotypů. Každý haplotyp je tvořen ze sady antigenů obsahujících konkrétní alely. Pro transplantaci se v současnosti pokládají za nejdůležitější HLA antigeny I. třídy A, B, C a antigeny II. třídy DRB1 a DQB1. [4] V povědomí je také mnoho dalších "minoritních" antigenů, jejich efekt na průběh transplantaci je však zatím předmětem výzkumu. Současný požadavek na míru shody je 10/10, což znamená v pěti HLA antigenech (konkrétně v HLA -A, -B, -C, -DRB1 a -DQB1). [1] V závislosti na klinické situaci pacienta a dostupnosti dárce můžeme akceptovat shodu 9/10, velmi ojediněle pak 8/10. [6]

Jako nejmenší možnou shodu lze přijmout 6/10 v genech HLA -A,-B, -DRB1, ale pacient zde čelí smrtelnému riziku odvržení štěpu. Počet teoreticky možných kombinací HLA znaků je u člověka několik miliard. Některé tkáňové typy (kombinace znaků) se vyskytují v určitém národě (oblasti) častěji, jiné téměř nikdy. Z důvodu dědění jednotlivých znaků je nejsnazší nalézt shodu mezi dvěma jedinci v pokrevním příbuzenstvu. Od rodičů na potomky je příslušná polovina znaků přenesena většinou v kompletní sadě. Co se týče širšího okruhu členů rodiny, zde existuje pouze malá pravděpodobnost úplné shody. Je sice šance, že jednu sadu budou navzájem sdílet se společnými předky, ale druhá pochází náhodně z jiné příbuzenské větve a tam už je míra shody minimální. [1]



Obr. 4: Dědičnost tkáňových znaků. [1]

[1]

Je tedy obtížný úkol vyhledat mezi lidmi někoho se shodnými či velmi podobnými tkáňovými znaky. Tato možnost klesá či stoupá v závislosti na tom, zda je pacientův fenotyp složen ze znaků vyskytujících se v dané populaci častěji nebo jestli je kombinace jeho HLA znaků nestandardní.

Aby mohlo začít vyhledávání dárce, existují po celém světě registry dárců kostní dřeně. Nezřídka se však stává, že složení pacientova HLA je natolik unikátní, že pro něho téměř nelze vyhledat vhodného dárce. [1]

Celá problematika vlivu míry HLA shody na výsledek transplantace krvetvorných buněk je velice komplikovaná, jelikož na transplantaci má kromě HLA shody efekt i řada dalších faktorů, za všechny: stav nemoci příjemce, CMV shoda, pohlaví dárce apod. [6]

[1] Dále budeme znak DR vyjadřovat jako DRB1.

3.6 Metody typizace

HLA typizace znamená rozbor HLA znaků, tj. určení zděděného fenotypu. Typizace neumožňuje určit konkrétní haplotypy (k identifikaci haplotypů slouží řada jiných k tomu vytvořených složitých metod).

Jsou dva způsoby, jak typizaci provádět. První, starší je sérologické vyšetření. Touto metodou typizace je umožněno definovat molekuly HLA exprimované (tzn. vyloučené) na buněčných membránách. Modernější metoda je molekulárně genetická typizace (DNA typizace). Tou lze určit přímo HLA alely (tj. sekvence nukleotidů kódující HLA antigeny). [1]

3.6.1 Sérologická typizace HLA

Sérologická typizace se také nazývá "mikrolymfocytotoxický test". Je to metoda typizace postavená na zkoumání reakci protilátek působících na buněčné membráně. [1]

Během této metody vyhledáváme antigenní rozdíly, které jsou podmíněné unikátním složením aminokyselin polymorfní části HLA. Jde o historicky první techniku identifikace HLA antigenů zavedenou profesorem Paulem Terasakim v průběhu 60. let 20. století. [6]

Její značnou výhodou je jednoduchost a to, že k její realizaci není nutné pořizovat drahé vybavení. Test je poměrně krátkodobý (cca tří hodiny). Nicméně u sérologické metody jsou známy také některé nevýhody. K typizaci je potřeba velké množství krve, životaschopnost lymfocytů a bývá také obtížné nalézt vhodné antisérum, které je nezbytné pro reakci se vzácnými antigeny různých populací. Největší nevýhodou je pravděpodobně nepřesnost HLA typizace. To odůvodňuje, proč ji začínají nahrazovat jiné a přesnější molekulárně genetické metody. [1]

3.6.2 Molekulárně genetická typizace HLA-PCR

Metoda PCR či DNA typizace byla vyvinuta v kalifornské Cetus Corporation. [1] Jedná se o stanovení polymorfismu na úrovni nukleotidů. [6] Umožňuje zmnožení určitého úseku DNA na několik kopií v cyklické reakci o třech teplotních fázích. Za účelem replikace slouží malé množství DNA, což může být například i jediná molekula. Možností, jak vymezit místo na řetězci, které chceme namnožit, je osadit jej po obou stranách tzv. primery. Primer je prvek nukleové kyseliny fungující jako počáteční bod replikace DNA.

Zařízení pro spuštění PCR se nazývá „termocykler“ a dokáže, poté co je naprogramováno, během několika sekund změnit teplotu o desítky °C. [1]

Molekulárně genetická typizace HLA molekul („genotypizace“), tedy stanovení polymorfismu na úrovni nukleotidů, znamená v současnosti standardní techniku, která minimálně pro transplantaci krvetvorných buněk úplně nahradila sérologii. PCR se používá v různých modifikacích. V současné praxi jsou pro molekulárně genetickou typizaci HLA systému uplatňovány výhradně tři základní technologie a jejich různé modifikace. [6]

1.7.2.1 Přehled molekulárně-genetických HLA typizačních metod

1) PCR-SSP=PCR se sekvenčně specifickými primery. Stanovení HLA alel zahrnuje řadu PCR reakcí, které obsahují různé HLA (sekvenčně) specifické primery. Metoda se také vyskytuje pod zkratkou ARMS ("Amplification Refractory Mutation System").

2) PCR-SSO = PCR se sekvenčně specifickými oligonukleotidy. Nejprve pomocí PCR amplifikujeme (namnožíme) celou polymorfní oblast HLA genu a poté hybridizujeme s HLA specifickými próbami (oligonukleotidy).

3) PCR-SBT ("Sequencing Based Typing"), přímá sekvenace. Udává nejspolehlivější a nejpřesnější informaci o DNA sekvenci HLA genu. Na rozdíl od dvou výše uvedených metod není informace získaná touto limitována pouze na známá místa sekvenčního polymorfismu hypervariabilních regionů, ale zahrnuje prakticky kompletní místa polymorfismu (exon 2 u II. třídy a exony 2 a 3 u HLA molekul I. třídy). SBT navíc pomocí group-specifické amplifikace (GSAP) či sekvenace (GSSP) umožňuje přesně definovat cis/trans vazbu sekvenčních motivů. [6]

Fáze metody PCR:

1) Denaturace - rozpojení řetězců DNA na dva jedno-vláknové řetězce.
2) Hybridizace - osazení vybraných úseků DNA primery.
3) Syntéza DNA - opětovné spojování řetězce DNA.

Body 1.-3. se v cyklu opakují. Abychom dosáhli dostatečného rozmnožení molekuly DNA, postačuje většinou 30 cyklů. Pokud se na začátku nacházela ve vzorku pouze jedna molekula DNA, po 32 cyklech může vzniknout až miliarda nasyntetizovaných molekul. [1]

2.7.2.1 Úrovně molekulárně-genetické typizace HLA systému

Pomocí molekulárně genetických metod můžeme typizovat až na úroveň přesného určení sekvence nukleotidů, ovšem tato přesnost se v řadě případů ukazuje jako zbytečná a neefektivní. Proto jsou uplatňovány dva přístupy genotypizace:

Nízké rozlišení (low resolution neboli 2-digits typizace)

Typizace na nízké úrovni. Tou získáme rozlišení na úrovni skupin alel HLA genu sdílejících stejné základní polymorfní sekvence. Tato rozlišení jsou ekvivalentní typizaci na úrovni sérologie, ovšem je zde zaručena mnohem větší spolehlivost. Výsledek dostáváme jako první dvě číslice DNA nomenklatury (např. A*02, DRB1*04 apod.) a úroveň této genotypizace označujeme jako 2-digits HLA typizaci. Tento přístup je používán v oblasti populační genetiky, pro účely orgánových transplantací, pro základní typizaci dárců v registrech dárců krvetvorných buněk nebo pro typizaci členů úzké rodiny, kdy chceme definovat základní haplotypy.

Vysoké rozlišení (high resolution neboli 4-digits typizace)

Přesná typizace. Identifikace přímo jednotlivých alel, dle HLA nomenklatury, tzn. alespoň první 4 čísla (např. A*02:01, DRB1*04:02 apod.). Tato úroveň genotypizace je naprosto nezbytná, pokud provádíme transplantaci do nepříbuzenského organismu, kdy i malé strukturní rozdíly jsou znatelné. Protože počet alel se neustále zvyšuje, tato úroveň rozlišení je těžko dosažitelná, a to i za použití přímé sekvenace. [6]

3.7 Požadovaný stupeň HLA shody

Kompletní HLA shoda je zajištěna pouze u HLA genoidentických sourozenců, kteří zdědili stejné sady HLA molekul. Stále zůstává riziko aloreaktivity (definice viz výkladový slovník), která je však podmíněna inkompatibilitou v jiných imunitu regulujících genech mimo HLA systém, jako jsou KIR geny, mHA geny apod.

U nepříbuzenských transplantací je situace složitější. Zde musíme vyhledávat co největší shodu na úrovni molekulární - tj. na úrovni shody sekvencí DNA kódující HLA molekuly. Jakýkoli rozdíl v nukleotidové sekvenci, který má za následek změnu proteinové sekvence příslušné HLA molekuly, může být příčinou aloreaktivity. Molekulární typizací je totiž možno objevit variace v sekvenci DNA i u sérologicky identických HLA antigenů a haplotypů.

U příbuzenských dárců hledáme hlavně shodu ve zděděných haplotypech, která pak generuje i shodu v HLA systému. Můžeme-li jednoznačně oddělit haplotypy, pak nám postačuje sérologická úroveň molekulární typizace ("low resolution"). U nepříbuzného páru dárce - příjemce hledáme primárně shodu v jednotlivých alelách každého HLA genu. Pro hodnocení shody a uspokojující výsledek požadujeme výhradně typizaci na úrovni 4-digits ("high resolution").

Co se týče výsledků transplantace, je nutno uvést, že negativní vliv má každá neshoda v HLA molekule. Nemůžeme považovat neshodu na úrovni alely za méně znepokojivou než neshodu v HLA antigenech. [6]

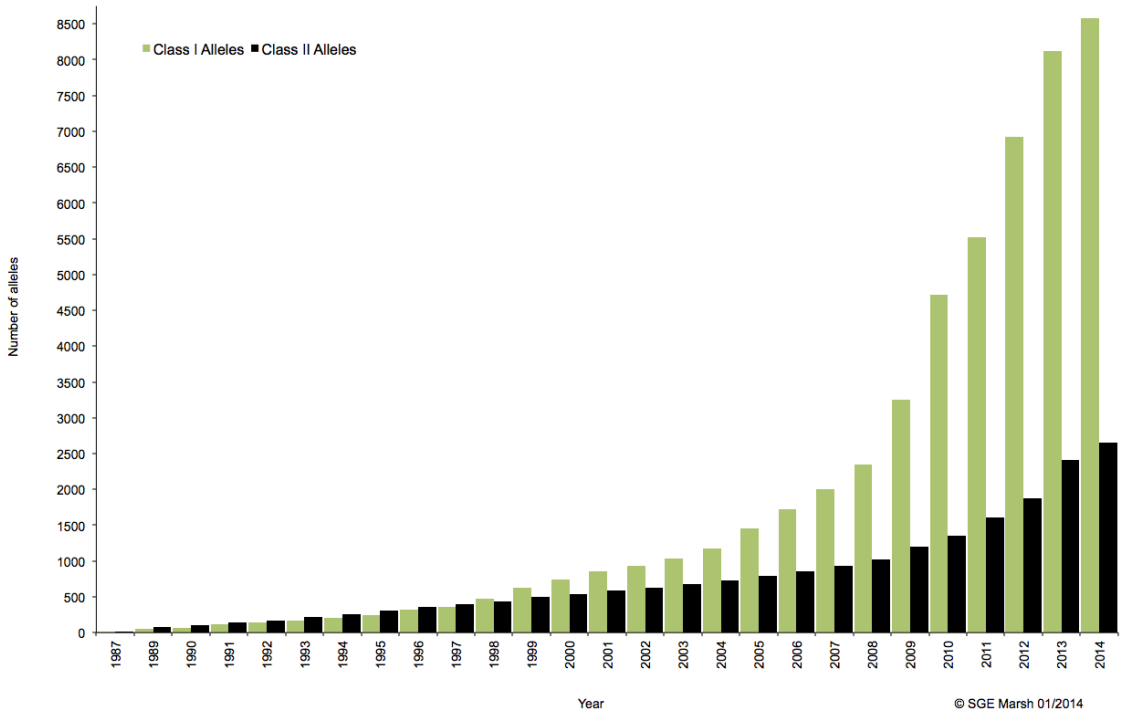
3.8 Nomenklatura

Nomenklatura je univerzální názvosloví k určování HLA znaků. Oddělení specifických oblastí antigenů je provedeno dvojtečkou (A*02:101). Dále nomenklatura umožňuje pokrývat HLA alely, které mají identickou sekvenci domén (např. A*02:01:01G) nebo také alely, které kódují identické vazebné domény HLA molekuly (např. A*02:01P). [1]

Tab. 1: *Nomenklatura platná od 1. dubna 2010.* [6]

Nomenklatura	**Označuje:**
HLA	HLA region a předpona pro HLA gen
HLA-DRB1	Specifický HLA lokus, např. zde DRB1
HLA-DRB1*13	Skupina alel, které kódují DR13 antigen
HLA-DRB1*13:01	Specifická HLA alela
HLA-DRB1*13:01N	"Null" alela (alela, která není exprimována)
HLA-DRB1*13:01:02	Alela, která se liší synonymní mutací
HLA-DRB1*13:01:01:02	Alela s mutací mimo kódující region
HLA-A*24:09N	"Null" alela
HLA-A*30:14L	Alela kódující protein se signifikantně redukovanou nebo "nízkou" expresí na buňkách
HLA-A*24:02:01:02L	Alela kódující protein se signifikantně redukovanou nebo "nízkou" expresí na buňkách
HLA-A*44:02:01:02S	Alela kódující protein

Z důvodu historického vývoje typizačních technik je nomenklatura HLA genů poměrně komplikovaná, nepřehledná a navíc často nesprávně a nekonzistentně používaná. Tento jev je zapříčiněn dvěma paralelně používanými systémy HLA nomenklatury: 1) sérologická nomenklatura - značí reagující antigeny. Jedná se o historický systém, který byl postaven na sérologické (proti-látkové) reaktivitě HLA molekul. V pořadí, jak byly všechny specificity objevovány, přiřazovala se jim písmena označující kódující gen, a číslo (např. B27). 2) DNA nomenklatura - označuje HLA geny a jejich alely. Oba systémy jsou v klinické praxi bohužel chybně používány promiskuitně. Přitom v současné době se celá HLA nomenklatura řídí mezinárodními standardy. [6]



Obr. 5: Počet nových alel od roku 1987 do konce prosince roku 2013. [7]

3.8.1 Přesnost rozlišení typizace

HLA typizace umožňuje rozlišit konkrétní antigen v sérologii nebo antigen včetně jeho specifické alely v případě DNA typizace. Sérologická typizace dokáže zachytit "Broady" a "Splity". "Broad" je soubor antigenů, tj. jednotlivých splitů. Jako "Split" je označován daný antigen.

Přesnější je, jak už bylo výše řečeno, DNA (PCR) typizace, u které lze nastavit několik úrovní rozlišení (nízkou, střední, vysokou). V praxi je možné se s nízkou a střední úrovní setkat při HLA typizaci dárců při vstupu do registrů. Vysoké rozlišení slouží k určení kompatibility pacienta a dárce. Užívá se však méně a to z důvodu finanční náročnosti. [1]

Tab. 2: Přesnost rozlišení pro jednotlivé typizace. [1]

Příklad	Rozlišení		Popis
	Sérologie	DNA	
A9	Broad	-	Nejistý antigen - obvykle soubor splitů
A3	Split	-	Jistý antigen
A*02:XX OR A*02	-	LOW	Možnost rozložení na soubor alel, antigenu
A*02:AB	-	Intermediate	Alely s vysokým rozlišením. AB zde představuje NMDP kód.
A * 0201	-	HIGH	Jistá alela - nejlepší výsledek

3.9 NMDP (MAC) kódy

NMDP ("National Marrow Donor Program") kódy, jinak označované jako MAC - Multiple Allel Codes, jsou „*kódy zahrnující skupinu alel získaných DNA typizací, ze kterých je právě jedna platná“*. [1] Existuje typizační přístroj k zobrazování těchto kódů.

Uveďme příklady MAC kódu a v něm obsažené množiny alel:

AB - {01,02}
XA - {02,21}
DKRA - {10,16,69}
HFNT - {01,08,11,13,14,17}

3.10 Haplotyp

„*Haplotyp je sada alel vázaných na jednom chromozomu, které mají tendenci dědit se společně“*. [8] Jinak řečeno, sada alel přenášených společně na potomka. Je-li identifikována asociace nějakého haplotypu s nemocí, lze tento haplotyp použít jako prediktor rizika vzniku nemoci u konkrétního jedince. U člověka pozorujeme diploidní genotyp, což znamená dva haplotypy. Nejčastěji typizované znaky jsou HLA-A, -B, -DRB1, tzn. v případě jednoho člověka: -A1, A2, -B1, -B2, -DRB1, -DRB2:
např. A*01,03-B*08,35-DRB1*01,03

Možné kombinace haplotypů:

A1-B1-DRB1 / A2-B2-DRB2,
A1-B1-DRB2 / A2-B2-DRB1,
A1-B2-DRB1 / A2-B1-DRB2,
A1-B2-DRB2 / A2-B1-DRB1

Fenotyp je různě ovlivňován tím, jestli jsou konkrétní alely SNPs (definice viz výkladový slovník pojmů) umístěny současně na jednom chromozomu nebo na dvou chromozomech zvlášť. Uveďme tři důvody, proč ve studiích dávat přednost haplotypům před jednotlivými SNPs [8]:

1) Proteinová sekvence je kódována v DNA na paternálním a maternálním chromozomu, a nestačí tedy znát pouze genotyp, ale důležitá je znalost samotných haplotypů. Překlad každého haplotypu, resp. sekvence DNA, kterou tento haplotyp reprezentuje, způsobí vznik samotného proteinu. Mohou se utvořit dva proteinové řetězce, přičemž jeden odpovídá maternálnímu, druhý paternálnímu haplotypu. Výskyt více mutací na jednom chromozomu současně mění výsledný proteinový řetězec více než při stejném počtu mutací, které jsou rozdělené mezi oba dva chromozomy a způsobují změny menšího rozsahu.

2) Variace v populaci se ukládá v haplotypech přenášených z jedné generace na další.

3) Pokud provádíme statistické hodnocení haplotypů, lze sledovat zvýšení síly testu oproti testování jednotlivých markerů separovaně. [8]

Na konci genotypizace genetických markerů máme většinou informaci **bez znalosti stavu** (genotyp) na jednotlivých chromozomech, zatímco námi žádaná informace je **znalost stavu** (haplotypy) na jednotlivých chromozomech. Haplotypy z genotypu jsou rozlišovány jedině v případě, že všechny markery jsou homozygotní - dostáváme dva stejné haplotypy, nebo je heterozygotní pouze jedna pozice - můžeme získat jen jednu kombinaci haplotypů, to znamená dva různé haplotypy. Tyto genotypy tak mají **jednoznačně rozluštitelné** haplotypy. Pokud se ale v genotypu nacházejí minimálně dvě heterozygotní pozice, pak jde o haplotypy **nejednoznačně rozluštitelné** a potom existuje 2^{s-1} variant, kde s je počet heterozygotních pozic a $s \geq 2$. To znamená, že haplotypy je možno snadným způsobem odvodit, pokud se genotyp skládá pouze z homozygotních SNPs nebo nanejvýš jednoho SNP, které je heterozygotní. Pro genotypy, které jsou nejednoznačně rozluštitelné, je potřeba použít další metody.

Haplotypy lze identifikovat experimentálně pomocí speciálních laboratorních technik, ty ale nejsou příliš praktikovány z časových a finančních důvodů. [1]

Jiný postup může být odvození z genotypů rodičů, čímž ale většinou není rozluštěn kompletní soubor, což práci komplikuje. Mimo to rodiče bývají dostupní jen v případě pacientů a ne kontrol (rodiče kontrol totiž většinou nechtějí pomáhat objasnění nemoci,

kterou netrpí jejich potomek), a při tomto přístupu je pak ztracena velká část kontrol, kterých je i jinak nedostatek. Problém se získáváním genetické informace rodičů se objevuje i u nemocí, které se projevují ve vyšším věku, např. diabetes mellitus 2. typu, kdy rodiče pacienta už ve většině případů nežijí.

Další způsob je vyloučení jedinců, kteří mají nejednoznačně rozluštitelný genotyp, z analýzy. Tento přístup se ale nejeví příliš použitelný a to kvůli ztrátě informace. Navíc je vhodný pouze u haplotypů, které jsou tvořeny jen malým počtem markerů, neboť s přibývajícím počtem markerů se snižuje pravděpodobnost výskytu osob, jež mají jednoznačně rozluštitelný genotyp.

Zaměříme se na metody in silico, které řeší problém vyloučení části souboru ze studie, popřípadě je možné použít je náhradou za náročné experimentální metody. Jedná se o odvození (rekonstrukci) haplotypů z genotypů nepříbuzných jedinců a díky tomu můžeme pracovat uvnitř celého souboru. [8]

Pokud provádíme genetické analýzy, pracujeme se vzorkem populace, kde na začátku této analýzy kontrolujeme, aby vzorek ve skladbě alel byl vyvážený. Ověřujeme, zda počty pozorovaných genotypů jsou rovny počtům, které bychom očekávali s ohledem na pozorované alelické počty ve vzorku. K takovému ověření se používá **Hardy-Weinbergův (HW) princip** zkoumající, jaké je rozložení alel jedinců v populaci. Jde o matematický aparát, kterým je možné vypočítat či odvodit genotypové četnosti z alelových četností, jestliže předpokládáme náhodné kombinace alel v populaci.

Pozorujme v populaci dvě alely určitého genu a označme je jako *A* a *a*. Můžeme sledovat tři různé genotypy *AA, Aa, aa* a jejich počty n_{AA}, n_{Aa} a n_{aa} (viz Tab. 3). Pomocí počtu genotypů pak jednoduše vypočteme četnost obou alel.

Tab. 3: Zastoupení genotypů.

Genotyp	Počet osob s daným genotypem
AA	n_{AA}
Aa	n_{Aa}
Aa	n_{aa}
Celkem	*N*

Určení četnosti alely *A*, *p*:

$$p = (2n_{AA} + n_{Aa})/N \,. \qquad (1)$$

Určení četnosti alely *a*, *q*:

$$q = (2n_{aa} + n_{Aa})/N \,. \qquad (2)$$

Platí, že:

$$p + q = 1. \qquad (3)$$

Z alelových četností poté můžeme odvodit očekávané četnosti genotypů (viz Tab. 4), a ty pak porovnáváme s pozorovanými genotypovými četnostmi (pozorované četnosti je možné snadno vypočítat, např. pro genotyp *AA* jako n_{AA}/N).

Tab. 4: Odvození genotypových četností na základě kombinování alel.

	A(p)	*A(q)*
A	*AA*	*Aa*
(p)	p^2	*pq*
A	*aA*	*aa*
(q)	*qp*	q^2

Tab. 5: Odvozená očekávaná četnost genotypů.

Genotyp	Očekávaná četnost
AA	p^2
Aa	*2pq*
Aa	q^2
Celkem	*1*

Očekávané četnosti určíme následujícím způsobem použitím binomického rozvoje:

$$(p + q)^2 = p^2+2pq+q^2. \qquad (4)$$

Konstanty použité v rovnici (4):
p... četnost alely *A*
q... četnost alely *a*

Využitím výše získaných očekávaných četností nakonec můžeme určit očekávané počty (např. pro genotyp AA vypočítáme p^2N), které porovnáváme s pozorovanými počty genotypů. K ověření rovnováhy slouží X^2 test dobré shody:

$$X^2 = \sum \frac{(pozorované\ počty\ genotypů - očekávané\ počty\ genotypů)^2}{očekávané\ počty\ genotypů}. \qquad (5)$$

V našem případě uvažujeme pouze jeden stupeň volnosti.

Rozbor HW rovnováhy jednotlivých SNPs může např. poukázat na fakt, že distribuce alel mezi jedinci s onemocněním je nerovnoměrná. [8]

3.11 Fenotyp

Fenotyp je soubor všech definovatelných znaků jedince. Zařazujeme sem znaky nejen pozorovatelné a definovatelné na úrovni organismů, ale také charakteristiku určité fyziologické nebo biologické funkce. Pokud pracujeme na buněčné úrovni, můžeme za odlišný fenotyp označit např. tvar různých typů buněk nebo jejich odlišnou funkci. Na biochemické úrovni lze odlišnost fenotypů demonstrovat např. na molekulách hemoglobinů. U zdravého člověka v dospělém věku je přítomen hemoglobin dospělého typu HbA (97 %) a HbA2 (2,5 %) a fetální hemoglobin HbF (0,5 %). Vznik variant hemoglobinu je důsledek mutace a tyto varianty bývají provázeny různými typy hemoglobinopatií. Fenotyp je podmíněn genotypem, epigenetickými změnami a vlivy prostředí. [9]

3.12 Genotyp

Genotyp charakterizuje buď informaci o genetické konstituci buňky, nebo organismu či jedince. Genotyp u daného jedince označuje veškerou jeho genetickou charakteristiku. U každého jedince téhož druhu se liší z hlediska genomických sekvencí v genových i mimogenových oblastech. Jako příklad velké variability genotypů u jedinců lidské populace lze uvést HLA lokus. [10]

4 Metody odvozování/rekonstrukce haplotypů

V této kapitole bylo čerpáno a citováno ze zdroje [8], není-li uvedeno jinak.

Kapitola se zabývá teorií metod pro rekonstrukci haplotypů z genotypů nepříbuzných osob. Máme k dispozici tři základní metody:

- **Clarkův algoritmus** - umožňuje rekonstruovat přímo haplotypy jedinců v souboru.
- **EM ("Expectation-Maximization") algoritmus** - odhaduje frekvence haplotypů v populaci, uvnitř populace pak odvodí haplotypy pro jedince v souboru**.**
- **Bayesovské algoritmy** - určují pravděpodobnosti haplotypových párů pro dané genotypy.

4.1 Clarkův algoritmus

Jedná se o první algoritmus pro rekonstrukci haplotypů, které mají délku větší než tři polymorfismy. Haplotypy tímto algoritmem odvozujeme z genotypů získaných z populace.

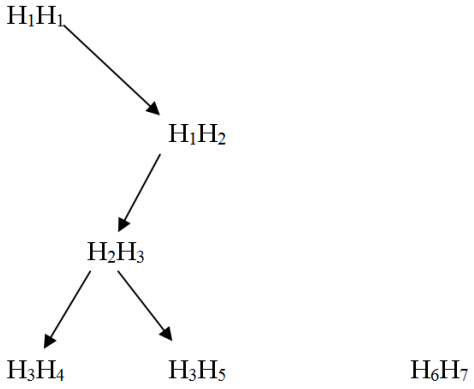
Při jeho použití očekáváme, že ve vzorku dostatečného počtu jedinců z populace s daným genotypem se určitě nachází nějaké genotypy zcela homozygotní nebo s nejvýše jednou heterozygotní pozicí (tzn. jednoznačně rozluštitelné genotypy).

Tab. 6: Postup odvozování haplotypů dle Clarkova algoritmu.

(1) Na počátku rozlušti všechny homozygotní genotypy a genotypy s jednou heterozygotní pozicí. Haplotypy takto získané jsou označeny jako známé.
(2) Tyto známé haplotypy se pokus přiřadit k zatím nerozluštěným genotypům. Pokud se některý haplotyp shoduje, potom jej k danému genotypu přiřaď a druhý haplotyp odvoď tak, aby haplotypový pár odpovídal genotypu.
(3) Takto je získán další známý haplotyp. V hledání pokračuj dále (bod(2)), dokud existují jakékoli rozluštitelné genotypy.
(4) Odvozování ukonči, pokud již není možné žádné genotypy pomocí známých haplotypů rozluštit nebo pokud se podařilo rozluštění všech genotypů.

Příklad použití Clarkova algoritmu

Obr. 6: Příklad odvozování haplotypů Clarkovým algoritmem.



Nejprve rozluštíme haplotypy homozygotního genotypu. Máme známý haplotyp H_1 a pomocí něho nalezneme další známý haplotyp z genotypu H_1H_2 a dostáváme nový známý haplotyp H_2. Takto pokračujeme, dokud můžeme ze známých haplotypů získat nějaké haplotypy dosud nerozluštěných genotypů. Genotyp H_6H_7 je na konci nerozluštěn, neboť k rozluštění haplotypů H_6 a H_7 nelze použít žádný ze známých haplotypů H_1 - H_5.

Je nezbytné zopakovat postup s různým počátečním pořadím genotypů. Pokaždé se totiž mohou výsledky lišit.

Výhody Clarkova algoritmu

- Tento algoritmus funguje velice dobře pro menší vzorky s častými haplotypy.
- Je velmi jednoduše proveditelný.

Nevýhody Clarkova algoritmu

- V případě, že se ve vzorku osob nevyskytuje jakýkoli jednoznačně rozluštitelný genotyp, nelze algoritmus spustit.
- Algoritmus může být ukončen, i když se nepodařilo rozluštit všechny genotypy.
- Může dojít k chybnému rozluštění haplotypu, pokud nastal crossing-over a genotyp odpovídá jinému známému haplotypu v souboru, než je skutečný haplotyp.
- Nejpravděpodobnější správná metoda rekonstrukce haplotypů je ta, která může rozluštit co největší počet genotypů.
- Pokud algoritmus pracuje ve větších vzorcích, roste pravděpodobnost výskytu málo četných haplotypů a takové je pak obtížné tímto algoritmem získat.
- Je nutné několikrát zopakovat postup s různým počátečním pořadím genotypů.

4.2 EM algoritmus

Jde o iterační metodu, při které se hledají maximální důvěryhodné odhady haplotypových četností, přičemž je předpokládána Hardy-Weinbergova rovnováha.

Nejlépe pracuje pro velké vzorky z populace a není závislý na míře rekombinace. Je považováno za vhodné spouštět jej s různými počátečními podmínkami, ale na druhou stranu u něj nezáleží na pořadí vstupních dat jako u Clarkova algoritmu.

S rostoucím počtem znaků metoda vykazuje exponenciální vzrůst nároků na výpočetní čas. Tento algoritmus je použitelný na řešení problémů typu:

- nalezení haplotypů pro daný soubor.
- označení haplotypů s největší četností.
- odhad haplotypových populačních četností.
- určení nejpravděpodobnějšího haplotypového páru pro každého jedince v souboru.

Algoritmus stojí na předpokladu, že pokud bychom znali haplotypové páry pro dané genotypy, pak z nich je možno odhadovat četnost haplotypů, i naopak pokud bychom znali četnosti, pak nám umožňují odvodit z nich haplotypy, které odpovídají příslušným genotypům.

Tab. 7: Postup odvozování haplotypů pomocí EM algoritmu popsaný pomocí pseudokódu.

(1) Nastav konvergenční kritérium $\varepsilon < 10^{-v}$ pro přesnost odhadu četností haplotypů. Obvykle: $4 \leq v \leq 8$. (6) Nastav maximální počet iterací N. Obvykle: $25 \leq N \leq 100$. (7) Nastav hodnotu věrohodnostní funkce pro první srovnání, např. $L^{(-1)} = 1$.
(2) Zahaj iteraci s hodnotou $s = 0$
(3) Nastav počáteční podmínky, tedy haplotypové četnosti $p_k^{(0)}$, $k = 1, ..., M$.
(4) Dokud $s \leq N$, opakuj následující kroky:

(I) Počítej pravděpodobnosti všech možných kombinací haplotypů pro dané genotypy:

$$p^{(s)}(h_k,h_l)\begin{cases} (p_k^{(s)})^2 \; k = l \\ 2p_k^{(s)}p_l^{(s)} \; k \neq l \, . \end{cases} \tag{8}$$

(II) Odhadni pravděpodobnosti genotypů g_1, ..., g_r jako součet pravděpodobností kombinací haplotypů kompatibilních s daným genotypem spočítaných v předchozím bodě

$$p^{(s)}(g_i) = \sum_{k=1}^{M} \sum_{l=1}^{M} \hat{p}^{(s)}(h_k, h_l), \text{ kde } i = 1, \ldots, r \; a \; k \leq l \, . \tag{9}$$

(III) Spočítej algoritmus věrohodnostní funkce

$$\ln L^{(s)} = \sum_{s=1}^{r} n_{gi} \ln \hat{p}^{(s)}(g_i) \, . \tag{10}$$

a pokud $\left|L^{(s-1)} - L^{(s)}\right| < \varepsilon$, pokračuj na bod (6).
(IV) Odhadni očekávaný počet haplotypu *k, kde k = 1, ..., M :*

$$\hat{\hat{E}}^{(s)}(n_{hk}) = \sum_{i=1}^{r} \hat{p}^{(s)}\left(h_k, \frac{h_l}{g_i}\right) , \tag{11}$$

Kde

$$\hat{p}^{(s)}\left(h_k, \frac{h_l}{g_i}\right) = \frac{2.p_k^{(s)}.p_l^{(s)}}{\hat{p}^{(s)}(g_i)} \; pro \; h_k \neq h_l \; i \; h_k = h_l \, . \tag{12}$$

(V) Aktualizuj haplotypové četnosti:

$$p_k^{(s)} = n_{h_k}^{(s)} / 2n \, . \tag{13}$$

(5) Ukonči bez konvergence.

(6) Ukonči, odhady haplotypových četností jsou $p_k^{(s)}$.

Příklad použití EM algoritmu

(volně přeloženo ze zdroje [11])

Uvažujeme dva lokusy. Pro jednoduchost předpokládáme, že jsou dvoualelové. Lokus 1 má alely *A, a,* lokus 2 má alely *B, b* . Počty genotypů pozorujeme u dvou lokusů *N* jednotlivců z populace. Počty homozygotů pro oba lokusy získané EM algoritmem označíme jako: n_{AABB}; n_{aaBB} ; n_{AAbb}; n_{aabb} . Obdobně počty homozygotů pro jeden lokus jsou: n_{AABb}; n_{aaBb}; n_{AaBB}; n_{Aabb} . Dále heterozygoty pro oba lokusy mají četnost n_{AaBb} . Stanovme závěr o haplotypech. Jaké jsou odhady frekvenci haplotypů? Jaké jsou možné haplotypy pro každou genotypovou kombinaci dvou lokusů? Pro jakou genotypovou konfiguraci nelze odvodit fáze?

1) n_{AaBb} je četnost spojení dvou haplotypových párů: $n_{AB=ab}$, $n_{Ab=aB}$, $n_{AB}=_{ab}$ a $n_{Ab}=_{aB}$ jsou naše chybějící data od fáze, kdy tyto haplotypy nemohou být rekonstruovány z genotypových dat. Pokud by byly tyto fáze známy pro všechny haplotypy, potom můžeme snadno zapsat pravděpodobnostní funkci pro n_{AB}, n_{Ab}, n_{aB} a n_{ab} z hlediska četnosti haplotypů p_{AB}, p_{Ab}, p_{aB} a p_{ab}. Jaká má být tato pravděpodobnostní funkce v případě kompletních dat?

2) Vzorek jednotlivců *N* obsahuje *2N* haplotypů. n_{AB} v případě kompletních dat je:

$$n_{AB} = 2n_{AB|AB} + n_{AB|Ab} + n_{aB|AB} + n_{AB|ab} . \quad (14)$$

Pravděpodobnost kompletních dat se vypočítá jako:

$$L(n_{AB};\, n_{Ab};\, n_{aB};\, n_{ab}) = \frac{2N!}{n_{AB}!\, n_{Ab}!\, n_{aB}!\, n_{ab}!} p_{AB}^{n_{AB}} p_{Ab}^{n_{Ab}} p_{aB}^{n_{aB}} p_{ab}^{n_{ab}} . \quad (15)$$

Maximální pravděpodobnost (MLE) pro p_{AB} je:

$$\hat{p}_{AB} = \frac{n_{AB}}{2N} . \quad (16)$$

Totéž lze provést pro $\hat{p}_{Ab}, \hat{p}_{Ba}, \hat{p}_{ab}$.

Nebudeme sledovat kompletní data, ale můžeme použít EM algoritmus. Sledovaná data jsou Y = (n_{AABB}; n_{aaBB}; n_{AAbb}; n_{aabb}; n_{AABb}; n_{aaBb}; n_{AaBB}; n_{Aabb}; n_{AaBb}).

3) Krok E (Expectation = očekávání) tohoto algoritmu zahrnuje výpočet *Q*, což je očekávaná logaritmická pravděpodobnost kompletních dat. Musíme získat následující podmíněná očekávání:

$$n_{AB}^0 = E\left[n_{AB} \middle| Y, p_{AB}^0, p_{Ab,}^0 p_{aB,}^0 p_{ab,}^0\right] \quad (17)$$
$$n_{AB}^0 = E[n_{Ab} | Y, p_{AB}^0, p_{Ab,}^0 p_{aB,}^0 p_{ab,}^0]$$

$$n_{AB}^0 = E[n_{aB}|Y, p_{AB}^0, p_{Ab,}^0 p_{aB,}^0 p_{ab,}^0]$$
$$n_{AB}^0 = E[n_{ab}|Y, p_{AB}^0, p_{Ab,}^0 p_{aB,}^0 p_{ab,}^0] .$$

Jak vypočítáme n_{AB}^0 ?

$$E\left[n_{AB}\middle|Y, p_{AB}^0, p_{Ab,}^0 p_{aB,}^0 p_{ab,}^0\right] = 2n_{\mathrm{AABB}} + n_{\mathrm{AABb}} + n_{\mathrm{AaBB}} + E\left[n_{\frac{AB}{ab}}\middle|Y, p_{AB}^0, p_{Ab,}^0 p_{aB,}^0 p_{ab,}^0\right] . \quad (18)$$

Jak vypočítáme $E[n_{AB/ab}|Y, p_{AB}^0, p_{Ab,}^0 p_{aB,}^0 p_{ab,}^0]$?

$$E\left[n_{\frac{AB}{ab}}\middle|Y, p_{AB}^0, p_{Ab,}^0 p_{aB,}^0 p_{ab,}^0\right] = \mathrm{n}_{\mathrm{AaBb}} \frac{p_{AB}^0 p_{ab}^0}{p_{AB}^0 p_{ab}^0 + p_{Ab}^0 p_{aB}^0} . \quad (19)$$

Tedy:

$$n_{AB}^0 = 2n_{\mathrm{AABB}} + n_{\mathrm{AABb}} + n_{\mathrm{AaBB}} + \mathrm{n}_{\mathrm{AaBb}} \frac{p_{AB}^0 p_{ab}^0}{p_{AB}^0 p_{ab}^0 + p_{Ab}^0 p_{aB}^0} . \quad (20)$$

Podobně vypočítáme n_{Ab}^0, n_{aB}^0 a n_{ab}^0 .

4) Krok M (Maximization = maximalizace) zahrnuje maximalizaci *Q*, očekávané hodnoty logaritmické pravděpodobnosti (získané krokem E) s ohledem na p_{AB}, p_{Ab}, p_{aB} a p_{ab}.

Maximální pravděpodobnosti jsou

$$\hat{p}_{AB} = \frac{n^0{}_{AB}}{2N}, \hat{p}_{Ab} = \frac{n^0{}_{Ab}}{2N}, \hat{p}_{aB} = \frac{n^0{}_{aB}}{2N}, \hat{p}_{ab} = \frac{n^0{}_{ab}}{2N} .$$

5) V dalším kroku stanovíme

$$p_{AB}^1 = \hat{p}_{AB}, p_{Ab}^1 = \hat{p}_{Ab}, p_{aB}^1 = \hat{p}_{aB}, p_{ab}^1 = \hat{p}_{ab} .$$

6) Nyní se vrátíme do kroku E tohoto algoritmu a vypočítáme znovu *Q*. Pokračujeme v iteracích mezi kroky E a M, dokud parametry konvergují.

Výhody EM algoritmu

- Funguje nejlépe pro velké soubory bez ohledu na míru rekombinace mezi markery.
- Vhodný k použití ve větších souborech, protože zde pravděpodobněji dosahuje HW rovnováhy.
- Nemá zde roli počáteční pořadí genotypů.

Nevýhody EM algoritmu

- Je šance, že získáme haplotypy, které v souboru ve skutečnosti nejsou.
- Na druhou stranu může také dojít k opomenutí skutečných haplotypů.
- Algoritmus má citlivost na počáteční podmínky nastavení haplotypových četností.
- Pokud spouštíme algoritmus na velkém souboru, tak dochází k chybám pro haplotypy s malou četností (< 5 %) neboli řídké haplotypy.
- Přesnost výpočtu se zhoršuje pro stejně četné haplotypy.
- Algoritmus je založen na předpokladu HW rovnováhy, tzn. její velké porušení může způsobit chyby v odhadování pravděpodobností.

4.3 Bayesův algoritmus

Jde o algoritmy, jejichž prostřednictvím se na neznámé haplotypy díváme jako na neznámou náhodnou veličinu, a náš cíl je stanovit podmíněné rozdělení pravděpodobnosti v závislosti na pozorovaných genotypech.

Máme množinu neznámých haplotypových párů $H = (H_1, ...,H_n)$ pro n osob, resp. známých genotypů $G = (G_1, ...,G_n)$. K výpočtu podmíněné pravděpodobnosti $P(H|G)$ a jejího rozložení je potřeba použít Gibbsův vzorkovač, což je typ algoritmu Markovských řetězců - Monte Carlo, které vytváří řetězec odhadů $H^{(0)}$, $H^{(1)}$, $H^{(2)}$, ... Hodnota $P(H|G)$ vyjadřuje, s jakou pravděpodobností haplotypový pár odpovídá genotypu vzhledem k celkovému souboru genotypů. Např. pokud máme jednoznačně rozluštitelný genotyp $g_1 = (11,11,22)$, tak mu odpovídají dva haplotypy $h_1 = (1,1,2)$. Je to jediné možné řešení, takže pravděpodobnost této rekonstrukce se rovná jedné nezávisle na tom, jak vypadá zbytek souboru. Pro nejednoznačně rozluštitelné genotypy nastává řešení obtížnější. S pomocí Gibbsova vzorkovače můžeme procházet všechna možná řešení a přiřazovat jim příslušnou pravděpodobnost.

4.3.1 Základní Bayesův algoritmus

Základní Bayesův algoritmus je zahájen počátečním odhadem $H^{(0)}$ pro H. Následně se opakovaně odvozuje $H^{(t+1)}$ z $H^{(t)}$ pro $t=0,1,2,$ ... dle následujících kroků v Tab. 8.

Tab. 8: Postup odvozování haplotypů dle základního Bayesova algoritmu.

(1) Náhodně vyber jedince *i* ze všech osob s nejednoznačným genotypem. Tyto opakované výběry proveď rovnoměrně.
(2) Odvoď $H_i^{(t+1)}$ z podmíněné pravděpodobnosti $P(Hi\|G,H_{-i}^{(t)})$, kde H_{-i} jsou všechny haplotypové páry mimo páru jedince *i*. Předpokládej, že tito ostatní nevybraní jedinci mají správně rekonstruované haplotypy.
(3) Proveď úpravu $H_j^{(t+1)} = H_j^{(t)}$ pro $j=1, ..., n\ j \neq i$.

Pokud uděláme dostatečný počet opakování těchto kroků, získáme odhad $P(H|G)$. Během každé iterace aktualizujeme hodnotu $P(H_i|G, H\text{-}_i)$, kde haplotypový pár H_i je ekvivalentní kombinaci dvou haplotypů h_{i1},h_{i2}, jež jsou kompatibilní s genotypem G_i. Určení této podmíněné pravděpodobnosti s sebou nese problémy, jelikož výpočet pravděpodobnosti závisí na předpokladech o genetických a demografických modelech. Navíc daná pravděpodobnost pro modely nám není známa, proto ji tedy upravujeme následujícím způsobem: $P(H_i|G,H\text{-}i) \propto P(H_i|H\text{-}_i) \propto \pi(h_{i1}|H_{-i})\ \pi(h_{i2}|H_i,h_{i1})$, kde $\pi(.|\boldsymbol{H})$ znamená podmíněnou hustotu příštího vybraného haplotypu na základě haplotypů, které byly již vybrány. Tato hustota však není obecně známá a zároveň používáme tzv. „naivní Gibbsův vzorkovač“. Nutno předpokládat, že výskyt potomka s mutantním haplotypem *h* s pravděpodobností v_h nezávisí na genotypech rodičů. Pak $\pi(h|\boldsymbol{H})$ je určena jako:

$$\pi(h|\boldsymbol{H}) = (r_h + \theta v_h)/(r + \theta)\ . \qquad (21)$$

Konstanty použité v rovnici (21) :
r_h... počet haplotypů *h* v souboru haplotypů $\boldsymbol{H}$
r... celkový počet haplotypů v $\boldsymbol{H}$
v_h ... pravděpodobnost haplotypu *h*
θ... mutační míra čili šance vzniku mutace v každé generaci, s pomocí které můžeme definovat variabilitu odvozených haplotypů

Určíme pravděpodobnost haplotypu *h*, $v_h = 1/M$ pro všechny *h*, kde *M* označuje celkový počet haplotypů, které bychom mohli pozorovat v dané populaci. Platí, že $\theta v_h = 1$.

Použitím této základní podoby algoritmu dojdeme ke přibližně stejným výsledkům jako u EM algoritmu. Výhodu zde představuje možnost použití pro haplotypy delší, co se týče počtu sledovaných markerů, a navíc je ve výpočtu zahrnuta nejistota odhadu samotných haplotypů. Tato verze Bayesova algoritmu má další vylepšení.

4.3.2 Bayesův algoritmus na základě koalescenční teorie

Bayesův algoritmus na základě koalescenční teorie je základní přístup rozšířený o aplikaci koalescenční teorie Ta zní, že každý následující vybraný haplotyp bude tím více podobný těm předchozím podle toho, čím více osob jsme již sledovali a čím je mutační míra menší. Při rekonstrukci haplotypů je tedy počítáno s poznatkem, že haplotypy sledovaného genotypu budou stejné nebo podobné těm haplotypům vyskytujícím se ve výběru.

Pro názornost aplikace koalescenční teorie v odvozování haplotypů je zde uveden příklad. Přístup v něm praktikovaný využívá při rekonstrukci znalost některých haplotypů a jejich četností.

Tab. 9: Příklad odvozování haplotypů dle Bayesova algoritmu s aplikací koalescenční teorie.

Známé haplotypy: 12111 12111 12111 12111 21121 21121 11222 22122	**Jedinec č. 1 s neznámými haplotypy:** 00101 → 12111 21121 **Jedinec č. 2 s neznámými haplotypy:** 00201 → 12211 21221	**Legenda haplotypů:** 1 - alela majoritní 2 - alela minoritní	**Legenda genotypů:** 0 - heterozygot 1 - homozygot majoritní 2 - homozygot minoritní
		Známé haplotypy zobrazují doposud pozorované haplotypy i jejich četnost. Potom jedinci č. 1 přiřadíme nejpravděpodobnější kombinaci haplotypů, to znamená dva známé haplotypy s nejvyšší četností. Jedinci č. 2 již ale nemůžeme přiřadit žádný ze známých haplotypů a volíme proto ty nejvíce podobné haplotypy, které mají vysokou pozorovatelnou četnost se záměnou na třetí pozici.	

U tohoto algoritmu jde konkrétně o tzv. „pseudo-Gibbsův vzorkovač“. Podmíněnou hustotu $\pi(.|H)$ je možné aproximovat jako:

$$\pi(h|\boldsymbol{H}) = \sum_{a\in E}\sum_{s=0}^{\infty}\frac{r_\alpha}{r}\left(\frac{\theta}{r+\theta}\right)^s\frac{r}{r+\theta}(P^s)_{\alpha h}\,. \qquad (22)$$

Konstanty použité v rovnici (22):
r_α... počet haplotypů α v souboru haplotypů $\boldsymbol{H}$ z množiny E obecného mutačního modelu s mutační maticí P
r... celkový počet haplotypů v $\boldsymbol{H}$
θ... mutační míra
s... náhodný počet mutací náhodně zvoleného haplotypu α

Mutační míru odhadujeme podle vztahu:

$$\theta = S/log(2n) . \qquad (23)$$

Konstanty použité v rovnici (23):
S... počet pozorovaných heterozygotních pozic
n... počet známých genotypů

Algoritmus znovu začíná počátečním odhadem $\boldsymbol{H}^{(0)}$. Pokračuje opakovaným odvozováním $\boldsymbol{H}^{(t+1)}$ *z* $\boldsymbol{H}^{(t)}$ pro *t=0,1,2,* ... tzn. použitím kroků v Tab. 10.

Tab. 10: Postup odvozování haplotypů dle Bayesova algoritmu s aplikací koalescenční teorie.

(1) Náhodně vyber jedince *i (i=1, ...,n)* ze všech osob s nejednoznačným genotypem.
(2) Vyber podsoubor *S* heterozygotních markerů jedince *i*. Nechť *H(S)* označuje haplotyp jedince *i* tvořený pouze heterozygotními pozicemi a nechť *H(-S)* jsou stejným způsobem vzniklé haplotypy všech jedinců souboru mimo jedince *i*. (Pak $H(S) \cup H(-S) = \boldsymbol{H}$.). Odvoď $H^{(t+1)}(S)$ z podmíněné pravděpodobnosti $P(H(S)\|G, H^{(t)}(-S))$.
(3)Uprav $H^{(t+1)}(-S) = H^{(t)}(-S)$.

Použitím takovéto úpravy pak můžeme Bayesův algoritmus učinit přesnějším než Clarkův a EM algoritmus, nabízejí se i rozsáhlejší možnosti jeho aplikace a ve výpočtu je zahrnuta nejistota související s odhadováním haplotypů. Nutné je ovšem podotknout, že aplikace tohoto postupu je použitelná jen na takové datové soubory, které vyhovují předpokladům koalescenční teorie. Takovým souborem rozumíme vývoj populace po dlouhé časové období, kdy v této populaci neprobíhá ani neproběhl genový tok, stratifikace a nedošlo zde k tzv. bottlenecku neboli situaci, že nastala významná redukce části reprodukce-schopné populace, která by měla za následek nerovnoměrnou distribuci genotypů napříč populací. Důležitá je myšlenka, že haplotypy, které chceme určit pro některého jedince se podobají nebo odpovídají již známým haplotypům.

Stejně jako u EM algoritmu i zde je předpokládána platnost HW rovnováhy. Výpočet nejlépe funguje pro případ těsně vázaných markerů. V rámci implementace této metody tak bude nutno znát genetické vzdálenosti sledovaných markerů a použití je pak možné i na markery s velkou vzdáleností. Výpočet umožňuje pracovat s více markery než standardní EM algoritmus. Při rekonstrukci haplotypů lze také do výpočtu zahrnout znalost některých haplotypů, které jsme předtím získali experimentálně a takto odhady

zpřesnit. Dále se také naskytuje možnost metodu rozšířit o aplikaci na neúplné genotypy některých jedinců souboru.

4.3.3 Rozdíl mezi základním BA a BA s aplikací koalescenční teorie

Rozdíl mezi dvěma výše uvedenými variantami Bayesova algoritmu tkví v tom, že při využití koalescenční teorie předpokládáme odvozování haplotypových párů. V základním BA pracujeme s Dirichletovým rozdělením. To znamená, že při rekonstrukci genotypu, ke kterému neexistují žádné známé odpovídající haplotypy, algoritmus přiřazuje náhodně vybraný haplotypový pár, který se každý vyznačuje stejnou váhou. Nicméně jestliže využijeme předpokladů z koalescenční teorie, potom takovému genotypu vyhledáváme haplotypy podobné již dříve pozorovaným haplotypům. To znamená, že uděláme záměnu jedné nebo malého množství jednotlivých bází (SNPs), a dostaneme haplotypový pár, jenž je tvořen haplotypy podobnými těm již pozorovaným.

Odvozování haplotypů na základě podobnosti k již známým haplotypům sice v některých případech vyvolává u některých odborníků odpor, nicméně znamená zajisté realističtější volbu než náhodný výběr haplotypových párů, což bylo dokázáno i experimentálně. Celkově je tak vhodnější použít takové Bayesovy algoritmy, které rekonstruují haplotypy na základě podobnosti s těmi již známými v rámci souboru, ovšem jen tehdy, kdy nám to umožní typ sledované populace.

Vlastnosti Bayesovských algoritmů

- Nejlépe jsou použitelné na těsně vázané markery.
- Podobně jako EM algoritmus i Bayesovy stojí na předpokladu HW rovnováhy.
- Implementace algoritmu na základě koalescenční teorie může vyžadovat také znalost genetické vzdálenosti sledovaných markerů. Toto rozšíření lze pak použít i na vzdálené markery.
- Algoritmy na základě koalescenční teorie jsou optimální a hodí se k aplikaci na souborech jedinců z populace, kteří mohou splnit předpoklady koalescenční teorie.
- Nadstavbou EM a BA je metoda Partition-Ligation. [8]

Společná vlastnost Clarkova, EM a Bayesova algoritmu

Jeden z hlavních faktorů, který rozhoduje úspěšnost těchto metod, je porušení HW rovnováhy. Je-li rovnováha porušena zvýšením četnosti heterozygotů, stává se rekonstrukce problematická pro všechny tři algoritmy.

5 Implementace zvolených algoritmů

Praktická část je věnována implementaci Clarkova a EM algoritmu. Klade si za cíl ukázat, jak lze algoritmy řešit použitím programovacího jazyka. Dále porovnává výsledky obou metod a řeší jejich jednotlivé výhody a nevýhody.

Clarkův a EM algoritmus jsem implementoval pomocí programovacího jazyka Java. Jde právě o tyto dva algoritmy, protože bývají používány nejčastěji a také protože Bayesův algoritmus je implementačně velmi náročný. Mám k dispozici fenotypová data 17 219 dárců kostní dřeně. Uvažuji pouze znaky -A,-B,-DRB1. Na nich pracuji jen s alelickou skupinou (tzn. první dvě číslice před dvojtečkou).

5.1 Implementace Clarkova algoritmu

Clarkův algoritmus využiji pro rekonstrukci haplotypů v daném souboru jedinců. Jako první krok provedu rozdělení genotypů na homozygoty a heterozygoty. Následně provádím rekonstrukci podle Tab. 7. Způsob implementace s použitím pseudokódu je zapsán v Tab. 11.

Tab. 11: Postup počítačové implementace Clarkova algoritmu popsaný pomocí pseudokódu.

(1) Deklaruji: číselnou proměnnou „*a" = 0*, *„prozkoumáno" = 0*, pole „*homozygoti“*, pole „*heterozygoti“*, kde každý heterozygot má dvouúrovňový stav, dále seznam „*rekonstruované"* a haplotyp „*poslední“*.
(2) *„prozkoumáno“ = 0*. Rekonstrukce haplotypu na *a*-té pozici pole *„homozygoti“*. Uložení haplotypu do seznamu "rekonstruované". Haplotyp „*poslední“* = prázdný. „*poslední“* reprezentuje haplotyp, který byl rekonstruován jako poslední v pořadí. Na začátku je tedy prázdný.
(3) Pro každý prvek seznamu heterozygotů ověřuji: **Pokud stav aktuálního heterozygotu *!*= hotov a lze jej rozluštit s odpovídajícím haplotypem,** odvodím pomocí tohoto haplotypu druhý, který pak uložím do seznamu „*rekonstruované",* stav aktuálního heterozygotu = *„hotov"*. **Pokud aktuální heterozygot == poslední v seznamu**, přechod na bod (4).
(4) **Pokud *„prozkoumáno“ == 0,*** pak *„prozkoumáno“ = 1* a proměnná „*poslední“* = haplotyp, který byl rekonstruován jako poslední v pořadí a návrat na bod (3). **Jinak** přechod na bod (5).
(5) Porovnávám poslední rekonstruovaný haplotyp, s haplotypem „*poslední“*. Nyní ověřuji podmínku: **Pokud poslední rekonstruovaný haplotyp == *„poslední“,*** *a = a+1*, návrat k bodu (2). **Jinak** návrat k bodu (3).

Příklad použití implementace Clarkova algoritmu

Mám k dispozici fiktivní fenotypová data uvedená v Tab. 12.

Tab. 12: Ilustrační příklad souboru dárců s heterozygotními genotypy (každý řádek představuje jednoho dárce).

H_2	H_3
H_1	H_2
H_1	H_4
H_4	H_6
H_3	H_5
H_7	H_8

Vytvořím seznam homozygotních genotypů a dále seznam heterozygotních. Pro jeden z homozygotů, H_1H_1, provádím rekonstrukci:

1) Protože se jedná o homozygot, rekonstruuji jeho haplotyp H_1 a uložím ho do seznamu známých haplotypů.

2) Postupuji seznamem heterozygotů. Genotyp H_2H_3 nelze rekonstruovat, protože neobsahuje žádný ze známých haplotypů. Proto pokračuji. Z genotypu H_1H_2 lze pomocí již známého haplotypu H_1 rekonstruovat H_2. H_2 uložím do seznamu známých haplotypů. Genotyp označím. Pokračuji. Z H_1H_4 je možno rekonstruovat haplotyp H_4 opět ze znalosti H_1. Uložím H_4 do seznamu známých haplotypů. Genotyp označím. Nyní následuje H_4H_6. Protože v seznamu známých haplotypů se nachází H_4, rozluštím H_6. Uložím. Genotyp označím. K rozluštění genotypu H_3H_5 nelze použít žádný ze známých haplotypů. Totéž lze tvrdit o genotypu H_7H_8.

3) Program se nachází na konci seznamu heterozygotů. Uložím do určité proměnné poslední haplotyp, který byl rozluštěn, tj. H_6.

4) H_2H_3 již nyní lze rekonstruovat, protože v seznamu rekonstruovaných haplotypů se nachází H_2. Mohu tedy rozluštit haplotyp H_3 a uložit ho do seznamu. Genotyp označím. H_1H_2 byl již označen a tudíž pokračuji. H_1H_4 byl již označen. H_4H_6 byl již označen. H_3H_5 označen není. Haplotyp H_5 mohu odvodit přes H_3, který se již nachází v seznamu rekonstruovaných haplotypů. Ukládám H_5. Genotyp označím. K rozluštění genotypu H_7H_8 nelze použít žádný ze známých haplotypů.

5) Nyní se program opět nachází na konci seznamu heterozygotů. Poslední uložený haplotyp je H_5. Ten není totožný s haplotypem H_6 v proměnné deklarované po minulém průchodu. Z toho lze usoudit, že od minulého průchodu koncem byly nalezeny další haplotypy a tudíž má význam prozkoumávat seznam heterozygotů znovu. Ukládám H_5 do proměnné a cyklus opakuji.

6) H_2H_3 byl již označen. Přeskakuji. H_1H_2 byl již označen. H_1H_4 byl již označen. H_4H_6 byl již označen. H_3H_5 byl již označen. H_7H_8 označen není. K jeho rekonstrukci ale nelze použít žádný ze známých haplotypů.

7) Opět jsem na konci seznamu. Poslední rekonstruovaný haplotyp je H_5 a proměnná je též rovna H_5. Je tedy zřejmé, že program již žádné další haplotypy rozluštit nemůže a pro homozygotní genotyp H_1H_1 je nutné algoritmus ukončit, i když nejsou rozluštěny všechny haplotypy.

Ukázka výpisu aplikace vytvořené pomocí jazyka Java

Řetězec alelických skupin haplotypu je v aplikaci zapsán v pořadí A,B,DRB1, tzn. např. haplotyp A*01,B*08,DRB1*03 je označen jako 010803).

Po spuštění programu se pro každý homozygot zobrazí proces rekonstrukce. Oba výpisy vznikly pro případ výběru prvních 1 000 genotypů v pořadí. Počáteční homozygoty ale vybírám z celého souboru nezávisle na tom, jaký je zvolen podsoubor (z důvodu velkého rozsahu výpisu je zde uveden jen jeho krátký úsek), např. :

```
homozygot: 010803
genotyp:023511 010803
známý:010803
odvozeno:023511
genotyp:293511 010803
známý:010803
odvozeno:293511
genotyp:020803 010803
známý:010803
odvozeno:020803
```

V případě nemožnosti najít jakékoli jednoznačně rozluštitelné genotypy se zobrazí výpis, např. :

```
homozygot: 014014
1 rozluštěných haplotypů
nenalezen žádný jednoznačně rozluštitelný genotyp
nerozluštěných: 756
```

Tab. 13: Seznam 20 haplotypů s nejvyšší četností při použití Clarkova algoritmu (skutečný počet výskytů haplotypů v souboru).

Haplotyp A,B,DRB1	Četnost	Haplotyp A,B,DRB1	Četnost
010803	1 418	021504	228
020715	817	024416	220
030715	627	020803	211
010715	422	021307	183
024404	385	030701	179
010801	268	025115	172
020701	259	024415	167
024411	236	025101	163
010815	235	021811	161
024401	231	234407	159

Haplotypů, které se ve skutečnosti v souboru vyskytují pouze jednou, bylo nalezeno 878.

Homozygoti, kteří jako počáteční genotypy identifikovaly nejvíce haplotypů (Podsoubor byl zvolen jako 600. - 900. dárce v souboru):

```
homozygot: 030715: 168 rozluštěných haplotypů, nerozluštěných: 191
homozygot: 010803: 167 rozluštěných haplotypů, nerozluštěných: 192
Dále: 021307, 033501, 024416, 025707, 021511, 251815, 020715, 023813, 243511,
022701, 021504, 024404, 301307, 021811, 015707, 234407, 021513, 294407,
113501, 033511, 113504
```

Homozygoti, pomocí nichž nebyly identifikovány žádné haplotypy:

```
homozygot: 024412: 1 rozluštěný haplotyp, nenalezen žádný jednoznačně
rozluštitelný genotyp, nerozluštěných: 358
Dále: 014014, 020803, 024103, 024015, 234911, 683504, 021407, 244407, 241504,
021508, 241307
```

5.2 Implementace EM algoritmu

Postup počítačové implementace EM algoritmu je popsán v Tab. 14.

Tab. 14: Postup počítačové implementace EM algoritmu.

(1) V souboru genotypů si seřadím odpovídající alelické skupiny podle velikostí čísel. Například mám dárce s genotypem daným 1. haplotypem A1, B1, DRB1 a 2. haplotypem A2, B2, DRB2. Je-li A2 < A1, definuji 1. haplotyp jako A2, B1, DRB1 a 2. jako A1, B2, DRB2. Výpočty pak provádím s četnostmi takto zobrazených genotypů. Dále budu značit lokus DRB1 a DRB2 jako "D" a "d".
(2) Označím si četnost haplotypu jako *n*. Generuji všechny haplotypové kombinace (ABD, aBD, AbD, ABd, abD, aBd, Abd, abd) pro každého dárce a ukládám je do seznamu (vytvářím tzv. model existujících haplotypů). Každé přiradím počáteční odhad pravděpodobnosti $\hat{p}_{AB} = 0{,}125$. Pokud již byla daná kombinace vytvořena v předchozích krocích, nepřidávám novou kombinaci, ale připočítávám k parametru *n* očekávanou logaritmickou pravděpodobnost kompletních dat Q pomocí rovnice (19).
(3) Poté, co takto naleznu všechny kombinace všech dárců, počítám nové odhady pravděpodobnosti pro všechny vygenerované kombinace modelu. U každé parametr *n* vydělím dvojnásobkem počtu všech dárců, tzn. podle rovnice (16). Parametr *n* vynuluji.
(4) Nyní pro každého dárce počítám pravděpodobnosti pro navzájem se doplňující haplotypové páry, viz rovnice (19). Každé kombinaci tohoto dárce zvolím pravděpodobnost z pravděpodobností párů *p1*, *p2*, *p3*, *p4*, kde je tato kombinace součásti páru. Příslušnou pravděpodobnost páru přičtu k parametru *n* dané kombinace.
(5) Na konci průchodu všech opět vypočtu nové odhady pravděpodobností pomocí rovnice (16). Parametr *n* každé kombinace opět vynuluji. Jestliže pravděpodobnost výskytu některé kombinace klesne pod 10^{-10}, tuto pravděpodobnost zaokrouhlím na nulu.
(6) Průchody všech dárců, výpočty pravděpodobností párů a nových odhadů pravděpodobností haplotypů neustále opakuji, dokud není splněna zastavovací podmínka (součet všech rozdílů starých a nových odhadů pravděpodobností je menší než 0.1).

Ukázka konečných výsledků EM algoritmu po dosažení zastavovací podmínky:

Tab. 15: Seznam 20 haplotypů s nejvyšší pravděpodobností výskytu při použití EM algoritmu.

Haplotyp A,B,DRB1	Pravděpodobnost výskytu
010803	0,093 737
030715	0,043 525
020715	0,028 812
024404	0,0195 57
021307	0,017 988
234407	0,0173 30
021504	0,013 398
301307	0,012 969
033501	0,012 379
021811	0,012 196
251815	0,011 066
024416	0,011 004
015707	0,010 622
294407	0,010 479
021513	0,010 068
263813	0,009 564
240715	0,009 562
241307	0,008 961
022701	0,008 835
024411	0,008 471

EM algoritmus přiděluje nejvyšší pravděpodobnosti těm haplotypům, které tvoří genotyp s jiným velmi četným haplotypem. Nejčetnější haplotyp s nejvyšší pravděpodobností výskytu je A*01 B*08 DRB1*03. Řídkých haplotypů, tzn. takových, jejichž odhad pravděpodobnosti konverguje k nule, bylo při použití EM algoritmu nalezeno 209.

5.3 Porovnání obou algoritmů po implementaci

Výstupem Clarkova algoritmu je posloupnost haplotypů odvozená z jednoho určitého haplotypu, který je součástí homozygotního genotypu. Výstup EM algoritmu je pravděpodobnost výskytu haplotypu. Těmito dvěma způsoby je možné rozdělit genotypy na haplotypy dle jejich největšího výskytu.

Pokud spouštím Clarkův algoritmus pro příliš velký počet dárců, mizí zde postupně rozdíl mezi počty haplotypů rozluštěných při jednotlivých startovacích homozygotech. Pokud jej provádím pro kompletní soubor dárců, výsledek je pro všechny homozygoty stejný: 2 909 rozluštěných a zároveň 50 nerozluštěných haplotypů. Tato metoda je proto vhodná pro malý počet dárců na rozdíl od EM algoritmu, který vyžaduje spouštět pro kompletní soubor.

Při použití EM algoritmu mohou vzniknout kombinace, které v souboru ani neexistují a zbytečně tak zabírají paměť. Je také nutno stanovit zaokrouhlení odhadu pravděpodobnosti k nule, pokud má příliš malou hodnotu. Některé existující haplotypy mohou mít nižší pravděpodobnost než jiné vygenerované ale neexistující. U Clarkova algoritmu k tomu dojít nemůže, ovšem tam je nutné spouštět algoritmus pro různé počáteční homozygoty, protože bývají vždy rozluštěny jiné haplotypy.

EM algoritmus umožňuje identifikovat řídké haplotypy, tzn. ty které se vyskytují jen v minimální míře nebo netvoří genotyp s příliš četným haplotypem. Naproti tomu u Clarkova v této vlastnosti jistota není.

Pokud porovnávám nutnost vhodného pořadí dárců, u EM algoritmu na něm nezáleží, protože prochází vždy stejný počet a počítá četnost. Co se týče Clarkova algoritmu, zde je pořadí důležitý faktor ovlivňující rychlost běhu procesu. Pro některé homozygoty je potřeba procházet soubor dárců mnohokrát, protože zdaleka ne vždy je možné na první průchod rozluštit všechny haplotypy. U zvolené implementace nehrozí, že by došlo k opomenutí haplotypů, které by byly mohly být rekonstruovány, protože soubor je prozkoumáván, dokud zde jsou nějaké k nalezení.

Algoritmy jsem nespouštěl na rozlišení high resolution, protože nebyla k dispozici data dostatečného počtu dárců. To znamená, že by vzniklo příliš velké množství haplotypů a odhady jejich pravděpodobností by měly velmi malé hodnoty. Aby bylo vhodné spustit algoritmus s rozlišením na alely, musel bych použít data dárců hrubým odhadem 1 700 násobného počtu než nyní (kdy jich je 17 219).

Dalším důvodem, proč high resolution nepoužívám, je vysoká složitost a obtížná zpracovatelnost NMDP kódů. Např. kód:

24:02/24:02L/24:03/24:04/24:05/24:06/24:08/24:09N/24:10/24:11N/24:13/24:14/24:15/24:17/24:18/24:20/24:21/24:22/24:23/24:25/24:26/24:27/24:28/24:29/24:30/24:31/24:32/24:33/24:34/24:35/24:36N. [12]

Clarkův a EM algoritmus spouštím pouze na úrovni alelické skupiny, tedy podle nomenklatury jedné z nejnižších úrovní DNA rozlišení. Neusiluji však o přesnost rozlišení, ale o analýzu funkčnosti algoritmů pro identifikaci haplotypů. Není proto nutno spouštět algoritmy na vyšší úroveň. Počet dárců, který používám, je k těmto účelům také velmi postačující.

V případě EM algoritmu je vhodné odpovídající alelické skupiny seřadit podle velikosti čísla v označení (např. 02 < 03). Jedná se o předzpracování dat, kterým je možné učinit algoritmus výpočetně jednodušším.

Data dárců jsou pro Clarkův algoritmus načítána z textového souboru, zpracovávána a ukládána do tabulky podle jednotlivých dárců. Vytvořím dva jednorozměrné seznamy s odkazy na indexy v tabulce, kde se nacházejí heterozygoti a kde homozygoti. Pro samotný algoritmus je možno nastavit si podsoubor dárců, ve kterém budou haplotypy identifikovány. Tito dárci však musí následovat po sobě, tzn. volím jeden interval.

Ukládání do tabulek namísto objektové reprezentace vzniklo jako prvotní řešení problému, které bylo nakonec použito. Postupným vývojem programu vyplývá, že ukládat dárce jako instance objektů by bylo vhodnější a přehlednější. Algoritmická složitost by se nicméně nezměnila.

Při implementaci EM algoritmu opět načítám data dárců z textového souboru, zpracovávám je do tabulky, ale pro jednotlivé haplotypové kombinace používám objektovou reprezentaci, abych měl lepší přístup k atributům pravděpodobnost a četnost.

EM algoritmus je oproti Clarkovu implementačně náročnější a zároveň má mnohem vyšší algoritmickou složitost. Tu stanovuji jako počet operací bezprostředně se týkajících běhu algoritmu. To znamená, že do ní nezahrnuji operace spojené s načítáním dat, umisťováním do tabulky, textovým výpisem a řazením alel. Za operaci považuji deklaraci instance objektu, testování proměnných typu boolean (logický datový typ) a změnu hodnoty proměnné.

Algoritmická složitost Clarkova algoritmu

Provedl jsem součet operací při použití Clarkova algoritmu při spuštění pro různý počet dárců. Výsledky jsou obsaženy v Tab. 16**.**

Tab. 16: Vývoj algoritmické složitosti při použití Clarkova algoritmu.

Počet dárců	Počet operací
200	2 821 957
300	7 881 608
400	12 454 119
500	26 117 607
600	37 014 524
1 000	94 656 988
1 500	207 078 781

Algoritmus spouštím se všemi dostupnými homozygoty, kteří takto znamenají počáteční jednoznačně rozluštitelné genotypy. Lze vidět, že algoritmická složitost roste přibližně exponenciálně se zvětšujícím se počtem dárců ve zkoumaném podsouboru. Nejvhodnější je spouštět algoritmus pro menší celky, tzn. cca 300 dárců. Při takovém počtu funguje nejlépe.

Algoritmická složitost EM algoritmu

Algoritmická složitost při použití EM algoritmu je 3 951 252 342. U tohoto algoritmu neměním počáteční podmínky (podsoubor dat, se kterým pracuji), tato hodnota je pro tyto předem určené dárce konstantní. Jde však o mnohem vyšší hodnotu než při použití Clarkova algoritmu s podsouborem 300 dárců.

5.4 Problémy při implementaci

Komplikace nastaly po převedení tabulky s genotypy dárců do textového souboru a jeho následném čtení. Instance třídy *FileReader* určená ke čtení nerozpoznávala mezery. Dále jsem v prvotní fázi používal chybná data, kde většina znaků nebyla definována. Bylo tak možné použít jen malou část těchto dat. Velmi netriviální se ukázal problém řazení alel podle velikosti čísla označení. Obtížnost spočívala v nutnosti převedení datového typu řetězec na datový typ číslo a zpět.

6 Závěr

Motivací bakalářské práce bylo přinést základní přehled o transplantaci kostní dřeně, o genetických komplikacích s dárcovstvím a matematických metodách potřebných při řešení těchto komplikací.

Teoretická část v první (mimo úvod) kapitole definovala pojmy kostní dřeň a její transplantace. Bylo zde popsáno, jaké situace mohou nastat při vpravení krvetvorné buňky do těla příjemce. Náplní další kapitoly byly mj. HLA znaky. Podle míry shody HLA tkáňových znaků dárce a příjemce imunitní systém rozhodne, zda buňku přijme či odmítne. V práci bylo popsáno, jak se přenášejí znaky z rodiče na potomka. Další kapitola podávala informace, jakými způsoby lze HLA znaky určit (typizovat). Bylo zde rozebráno, že existují hlavní dvě úrovně rozlišení: nízká a vysoká, ze kterých je vybráno podle toho, co je cílem činnosti a jaká je požadovaná přesnost. Názvosloví pro HLA znaky se nazývá „nomenklatura“. Ta je používána mimo jiné ve fenotypových datech dárců kostní dřeně. Důležitým pojmem je haplotyp, což znamená sada alel (variant určitých genů) děděných společně na potomka. Je-li k dispozici soubor dárců, určuje se v něm četnost haplotypů a pravděpodobnost jejich výskytu. Důvodem je to, že s haplotypy se pracuje mnohem snáze a rychleji než s celými genotypy. Jednou z hlavních priorit provedení transplantace je krátký čas, protože pokud je provedena přibližně do doby dvou měsíců po stanovení diagnózy, zvyšuje se šance na přežití pacienta. Za tímto účelem existují matematické algoritmy: Clarkův, EM a Bayesův. Ty se staly také tématy třetí kapitoly teoretické části, kde je možno nalézt rozbor jejich aplikace, výhod a nevýhod každého z nich.

Praktická část se zabývala programovou implementací Clarkova a EM algoritmů na datech dárců kostní dřeně a posléze porovnáním funkčnosti těchto algoritmů. Byl zde uveden obecný postup, s jehož znalostí lze algoritmus vyřešit pomocí programovacího jazyka a spustit jej pro daná data. Následovala analýza výsledků, kde byly zpracovány tabulky s nejvyššími četnostmi a odhady pravděpodobností vzešlými z obou algoritmů, a porovnání předností a nevýhod té které metody oproti druhé. Součástí byla krátká zmínka o způsobu reprezentace fenotypových dat v programu, a problémech, které při tvorbě kódu nastaly.

7 Výkladový slovník pojmů

Definice jsou převzaty z [13], není-li uvedeno jinak.

Aloreaktivita - reakce štěpu na hostitele během alogenní transplantace. [14]

Alela - jedna z konkrétních forem genu.

Antisérum - sérum obsahující protilátky „protijed“ proti určitým jedům.

CMV - angl. zkr. controlled mechanical ventilation; regulovaná mechanická ventilace, způsob objemem limitované ventilace, přístroj dodá určený počet vdechů bez ohledu na dýchání pacienta.

Crossing-over - překřížení odpovídajících částí chromatid analogických chromozomů během meiózy srov. chiasma. Může dojít k výměně příslušných částí chromatid s následnou novou kombinací dědičných vlastností na jednom chromozomu rekombinace. Umožňuje vznik nových kombinací vlastností u potomků.

Diabetes mellitus - chronické onemocnění s vysokou morbiditou a mortalitou.

Ejekce - vypuzení krve ze srdeční komory při její systole.

Exon - část genu eukaryontů, která obsahuje vlastní dědičnou informaci.

Exprese - vytlačení.

Exprimace - vyloučení.

Fragment - zlomek.

Genotyp - souhrn všech dědičných vloh jedince uložený v genech.

Genom - soubor všech struktur nesoucích genetickou informaci ve formě DNA. Je tvořen chromozomy uloženými v buněčném jádře.

HbA - angl. zkr. hemoglobin dospělých adult.

Hemoglobinopatie - nemoc, jejíž podstatou je tvorba vadného krevního barviva hemoglobinu v důsledku dědičné poruchy.

Homozygot - jedinec, jehož obě alely na sledovaném lokusu (pro příslušný gen) jsou stejné. [15]

Heterozygot - jedinec, jehož obě alely na sledovaném lokusu (pro příslušný gen) jsou různé. [16]

Chromozom - vláknitá struktura buněčného jádra, v níž je v podobě DNA obsažena dědičná informace.

Indikace - rozhodný důvod či soubor okolností, vyžadující určitý léčebný nebo diagnostický postup.

Inkompatibilita – neslučitelnost.

In silico - spojení má význam: „s použitím počítačové simulace".

Izotopy - soubory atomů téhož prvku lišících se hmotnostním číslem, tj. počtem neutronů.

Konstituce – základní utváření jedince, tj. jeho tělesné a duševní vlastnosti jako celek.

Lokus - místo. V genetice místo na chromozomu, kde je lokalizován určitý gen.

Lymfocyt - druh bílé krvinky, který se významně podílí na specifické imunitě organismu.

Marker - angl. znak, který je typický pro určité buňky a jehož prokázáním lze tyto buňky v těle odhalit nádorové m. či spočítat m. různých druhů bílých krvinek.

Maternální - mateřský.

Membrána - blána.

Nukleotid - sloučenina skládající se z cukru ribosy pyrimidinové či purinové báze. N. jsou stavebními kameny nukleových kyselin.

Oligonukleotid - dva a více nukleotidů spojených specifickou chemickou vazbou.

Paternální - otcovský.

Polypeptid - peptid tvořený mnoha aminokyselinami.

Primer - malá molekula potřebná k zahájení syntézy makromolekuly, např. oligonukleotid specifické sekvence k syntéze DNA.

Próba - pomůcka umožňující průnik do hlouběji uložených nebo obtížněji přístupných částí těla k vyšetření nebo léčbě.

Působek - faktor, látka vznikající v organismu a působící na jeho různé části, buňky apod.

Replikace - proces zdvojení DNA, který předchází rozdělení buňky na dvě buňky dceřiné.

Signifikantní - významný.

SNP - Jednonukleotidový polymorfizmus ("Single nucleotide polymorphism") je záměna individuálního nukleotidového páru v sekvenci DNA, jejíž frekvence je v populaci častější jak 1 %.

Stratifikace - rozvrstvení, vrstevnatost.

Syntéza - spojování. Tvorba složitějších látek z látek jednodušších předchůdců.

Transportér - přenašeč. Obvykle molekula přenášející určitou sloučeninu např. z buňky nebo do buňky, v krvi.

Tumor - nádor.

Variabilita - proměnlivost, rozmanitost.

WHO - World Health Organization neboli Světová zdravotnická organizace.

8 Literatura a použité zdroje

1. CHVOJKOVÁ, M. *Podpora tvorby aplikace pro určení kompatibility příznaků pacient-dárce*. 2012. 54 s. Bakalářská práce na Fakultě aplikovaných věd Západočeské univerzity na Katedře kybernetiky. Vedoucí bakalářské práce Ing. Lucie Houdová.

2. KALINA, T**.** *Transplantační imunologie*. [online] [cit. 2014 04-27] <http://imunologie.lf2.cuni.cz/soubory_vyuka/cz_medici3_7.ppt>.

3. *Wikiskripta: Hematopoetická kmenová buňka*. [online] [cit. 2011-11-29] <http://cs.wikipedia.org/wiki/Hematopoetická_kmenová_buňka>.

4. ŠVOJGROVÁ, M., KOZA, V., HAMPLOVÁ, A. *Transplantace kostní dřeně: Průvodce Vaší léčbou.* Sv. 1. vyd. Plzeň : F. S. Publishing, 2006. 66 s. ISNB 80-903560-2-8..

5. *Lékařské slovníky: HLA antigen.* [online] [cit. 2014-05-06] <http://lekarske.slovniky.cz/lexikon-pojem/hla-antigeny-hla-system>.

6. POSPÍŠILOVÁ, Š., DVOŘÁKOVÁ, D., MAYER, J. *Molekulární hematologie.* Praha : Galén, 2013. 316 s. 9788072629428.

7. *Nomenclature.* [online] [cit. 2014-05-10] <http://hla.alleles.org/nomenclature/>.

8. CVANOVÁ, M. *Matematické metody hodnocení genových polymorfizmů v biomedicínském výzkumu,* 2011, 95 s. Diplomová práce na Institutu biostatiky a analýzy Masarykovy univerzity v Brně v Centru pro výzkum toxických látek v prostředí MU. Vedoucí diplomové práce doc. RNDr. Ladislav Dušek, Dr. .

9. *Wikiskripta: Fenotyp.* [online] [cit. 2013-03-01] <http://www.wikiskripta.eu/index.php/Fenotyp>.

10. *Wikiskripta: Genotyp.* [online] [cit. 2014-05-07] <http://www.wikiskripta.eu/index.php/Genotyp>.

11. THORNTON, T**.***: Haplotype Frequency Estimation: EM Algorithm.* Výukový materiál Washington University k předmětu Statistical Methods in Genetic Epidemiology.[online] [cit. 2013-03-08] <http://courses.washington.edu/b516/lectures_2010/EM_Algorithm_Haplotype_Frequency_2010.pdf >.

12. *Bioinformatics: Allele code lists*. [online] [cit. 2014-05-12] <http://bioinformatics.nmdp.org/HLA/alpha.v3.pdf>.

13. *Lékařské slovníky* [online] [cit. 2014-05-07] <http://lekarske.slovniky.cz/>.