SLOVENSKÁ TECHNICKÁ UNIVERZITA V BRATISLAVE

FAKULTA ELEKTROTECHNIKY A INFORMATIKY

Evidenčné číslo: FEI-5384-5958

Určovanie genetických predispozícií pomocou regulárnych výrazov

diplomová práca

SLOVENSKÁ TECHNICKÁ UNIVERZITA V BRATISLAVE

FAKULTA ELEKTROTECHNIKY A INFORMATIKY

Evidenčné číslo: FEI-5384-5958

Určovanie genetických predispozícií pomocou regulárnych výrazov

diplomová práca

|  |  |
| --- | --- |
| Študijný program : | Aplikovaná informatika |
| Číslo študijného odboru: | 2511 |
| Názov študijného odboru: | 9.2.9 Aplikovaná informatika |
| Školiace pracovisko: | Ústav informatiky a matematiky |
| Vedúci záverečnej práce: | Mgr. Zuzana Ševčíková |

Sem vložte zadanie z AIS

SÚHRN

SLOVENSKÁ TECHNICKÁ UNIVERZITA V BRATISLAVE

FAKULTA ELEKTROTECHNIKY A INFORMATIKY

|  |  |
| --- | --- |
| Študijný program : | Aplikovaná informatika |
| Vyberte typ práce | Určovanie genetických predispozícií pomocou regulárnych výrazov |
| Autor: | Bc. Jakub Kanitra |
| Vedúci záverečnej práce: | Mgr. Zuzana Ševčíková |
| Miesto a rok predloženia práce: | Bratislava 2015 |

Vložte text súhrnu, ktorý obsahuje informáciu o cieľoch práce, jej stručnom obsahu a v závere abstraktu sa charakterizuje splnenie cieľa, výsledky a význam celej práce. Píše sa súvisle ako jeden odsek a jeho rozsah je spravidla 100 až 500 slov

Kľúčové slová: Sem vložte 3 - 5 kľúčových slov

ABSTRACT

SLOVAK UNIVERSITY OF TECHNOLOGY IN BRATISLAVA

FACULTY OF ELECTRICAL ENGINEERING AND INFORMATION TECHNOLOGY

|  |  |
| --- | --- |
| Study Programme: | Applied Informatics |
| Bachelor Thesis: | Vložte názov práce. |
| Autor: | Bc. Jakub Kanitra |
| Supervisor: | Mgr. Zuzana Ševčíková |
| Place and year of submission: | Bratislava 2015 |

Vložte text súhrnu, ktorý obsahuje informáciu o cieľoch práce, jej stručnom obsahu a v závere abstraktu sa charakterizuje splnenie cieľa, výsledky a význam celej práce. Píše sa súvisle ako jeden odsek a jeho rozsah je spravidla 100 až 500 slov

Key words: Sem vložte 3 - 5 kľúčových slov

Vyhlásenie autora

Podpísaný Bc. Jakub Kanitra čestne vyhlasujem, že som diplomovú prácu Určovanie genetických predispozícií pomocou regulárnych výrazov vypracoval na základe poznatkov získaných počas štúdia a informácií z dostupnej literatúry uvedenej v práci.

Uvedenú prácu som vypracoval pod vedením Mgr. Zuzany Ševčíkovej.

V Bratislave dňa 05.01.2015

..................................................

podpis autora

Poďakovanie

Obsah

Úvod 1

1 Analýza problému 2

1.1 Nejaký podtext 2

2 Opis riešenia 3

3 Zhodnotenie 4

Záver 5

Zoznam použitej literatúry 6

Prílohy I

Príloha A: Nadpis II

Zoznam obrázkov a tabuliek

Figure 1 Obory potrebné pre riešenie diplomovej práce 3

Figure 2 Schéma eukaryotickej bunky 4

**No table of figures entries found.**  
In your document, select the words to include in the table of contents, and then on the Home tab, under Styles, click a heading style. Repeat for each heading that you want to include, and then insert the table of contents in your document. To manually create a table of contents, on the Document Elements tab, under Table of Contents, point to a style and then click the down arrow button. Click one of the styles under Manual Table of Contents, and then type the entries manually.

Zoznam skratiek a značiek

DNA – Deoxyribonukleová kyselina

RNA – Ribonukleová kyselina

HGP – Human Genome Project

API – Application Programming Interface

UI – User Interface

BRE – Basic Regular Expression

ERE – Extended Regular Expression

POSIX – Portable Operating System Interface

IEEE – Institute of Electrical and Electronics Engineers

RFC – Request For Comments

**Úvod**

Genetickým poruchám a chorobám sa veľmi ťažko dá predísť a v posledných desaťročiach ich rapídne pribúda. Podľa štatistík 3-4% všetkých novorodencov trpí určitým genetickým defektom a spôsobuje 20% všetkých umrtí novorodencov. O vážnosti týchto ochorení značí aj fakt, že 10% všetkých dospelých a 30% detských hospitalizovaných pacientov má geneticky ovplyvnené choroby. [[1](#Ann101)]

Diagnostika týchto ochorení nieje jednoduchá, no vďaka rozsiahlemu štúdiu a analýze ľudskej DNA a následnom skúmaní génov vznikajú rozsiahle databázy, ktoré opisujú gény a ich mutácie, na ich základe sa časť z týchto chorôb dá diagnostikovať v počiatkoch a teda umožniť včasnú liečbu.

Rozmachu takejto diagnostiky ako jednej zo základných bráni finančná náročnosť a rýchlosť DNA sekvencizátora, teda prístroja ktorý zosekvencuje DNA na sekvenciu dusíkových báz ktoré opíšem v 1.1 . Tento prístroj je v relatívnych počiatkoch (prvý vytvoril Lloyd M. Smith v roku 1987) a už teraz vidíme rapídnu evolúciu, a teda sa môže očakávať nárast rýchlosti a dostupnosti tohto prístroja.

Ďalším dôležitým faktom je výpočtová náročnosť analýzy sekvencie, keďže kompletná ľudská DNA sekvencia má približne 3 miliardy nukleotidových párov a 25 000 – 30 000 génov ich určenie vyžaduje značnú výpočtovú silu. Práve na tento fakt sa táto práca zameriava.

**Cieľom** **práce** je vytvorenie systému, ktorý by pre danú DNA sekvenciu určil genotyp jedinca a teda určil aj prítomnosť dostupných chorôb. Pre vytvorenie takéhoto systému sú potrebné základné poznatky z molekulárnej biológie a genetiky, ale aj poznatky z teórie regulárnych výrazov a ich spracovávania napríklad konečnými automatmi, tie sú opísané v 1.2. Táto práca sa neobmedzuje iba na DNA sekvencie, jej variácia sa môže použiť napríklad ako prostriedok big data analýzy napríklad na analyzovanie veľkých dát textov.

Rýchlosť a efektívnosť systému bude zabezpečovať jeho distributívna povaha. Tá zabezpečuje minimálne nároky na výpočtovú silu riadiaceho servera a spolieha sa na pripojené výpočtové zariadenia (uzly). Tie sú platformovo nezávislé, môže to byť tablet, smartfón alebo počítač. Bližšie informácie o tomto type systémov je možné nájsť v 1.3. Tento druh bol vybraný pre jeho malé rozšírenie, no enormný potenciál, keďže správne navrhnutý môže byť efektívnejší ako superpočítač a takisto rozšírením dostupných potenciálnych uzlov vďaka zrýchľovaniu zariadení a zväčšovaniu pokrytia vysokorýchlostným internetom.

Túto tému diplomovej práce som vymyslel nie z dôvodu, že som expert v nejakej zo spomínaných oblastí, no práve pre chuť a entuziazmus sa im venovať. Taktiež verím, že koncepty, algoritmy, výsledný systém a jeho podsystémy sa budú mocť uplatniť pri rôznych projektoch.

1. Analýza problému

Na Figure 1 sú v podobe diagramu ukázané potrebné znalosti na vyriešenie problému. V podkapitolách opíšem informácie a teóriu potrebné k splneniu zadania.



Figure 1 Obory potrebné pre riešenie diplomovej práce

* 1. Genetika

Táto časť je venovaná ozrejmeniu potrebných znalostí ohľadom genetiky a molekulárnej biológie, je nutné zdôrazniť, že práca je technického oboru a teda sa nebude opisovať do veľkej hĺbky, nebude tu spomenutá zložitá dedičnosť a replikácia jadra bunky a bunky samotnej. Tieto informácie sa dajú získať z knižných referencií [[2](#Har11)] [[3](#Rob05)].

Genetika, veda o dedičnosti, je vo svojom základe štúdium biologickej informácie. Všetky živé organizmy, od jednobunkových baktérií, rastlín a zvierat, musia uchovať, replikovať a preniesť na potomkov mnoho informácií o vývoji, reprodukcii a prežití vo svojom prostredí. Genetici skúmajú ako organizmy odovzdávajú biologické informácie vo forme DNA na svojich potomkov a ako ich využívajú počas života. [[2](#Har11)]

* + 1. Biológia bunky

Aby sa mohla pochopiť súvislosť a neuveriteľná prepracovanosť živých tvorov, je nutné aby sa opísala základná stavebná jednotka organizmu, bunka.

Existujú 2 typy buniek určené podľa ich zloženia:

* Prokaryotické – neobsahujúce jadro, a preto sa DNA voľne pohybuje v cytoplazme bunky
* **Eukaryotické** (zobrazená na Figure 2) – obsahujúce jadro, DNA je pevne uložené vo vnútri a za žiadnych okolností ho neopúšťa, informácie sa prenášajú iba pomocou RNA, ľudské bunky sú tohto typu, a preto je tento typ v našom centre záujmu

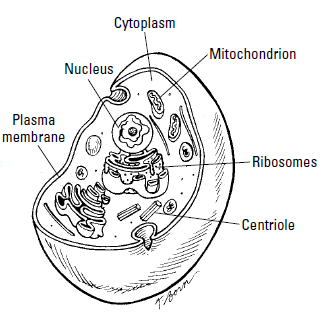


Figure 2 Schéma eukaryotickej bunky. Zdroj: [[3](#Rob05)]

Vo vnútri jadra sa nachádza genetická informácia v podobe chromozómov. Chromozóm je stužkovitý útvar pozostávajúci z DNA a pomocných naviazaných bielkovín. Nieje viditeľný ani pomocou mikroskopu, viditeľným sa stáva iba pri procese delenia bunky (mitózy). Počet chromozómov v jadre sa líši od druhu organizmu, človek má 46 chromozómov v jadre. Tie sa delia na 22 identických párov zhodujúcich sa v tvare a dĺžke a 1 pár pohlavných chromozómov, ktoré môžu byť zhodné (XX pre ženu) alebo rozdielne (XY pre muža). Ich unikátny tvar umožňuje presné zadefinovanie poradia chromozómov, čo je veľmi výhodné, keďže môžme sekvenciu DNA zapísať ako nepretržitý celok v definovanom poradí. Toto poradie sa nazýva aj karyotyp organizmu. Ukážka ľudského karyotypu je na Figure 3.

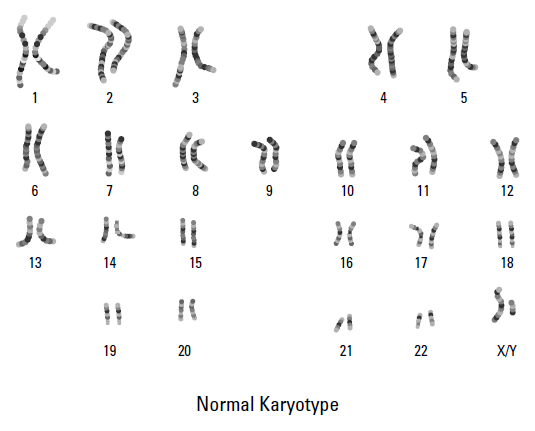


Figure Karyotyp človeka. Zdroj: [[3](#Rob05)]

Je nutné poznamenať, že nie každý organizmus obsahuje páry chromozómov, ale iba takzvané diploidné organizmy, napríklad osy sú haploidné čo znamená, že nemajú páry, ale sú organizmy, ktoré obsahujú až šesťnásť kópií toho istého chromozómu.

* + 1. DNA

Ako bolo spomenuté, chromozómy sa skladajú z molekúl deoxyribonukleonej kyseliny (DNA). Tieto makromolekuly sú veľmi odolné a vďaka tomu je možné ich neporušenú extrakciu zo skamenelých kostí alebo zvierat umrznutých v ľadovcoch.

**Chemické zloženie** DNA molekúl je pomerne jednoduché. Dusíková báza, deoxyribózový cukor a fosfát sa spojí za vzniku nukleotidu. Nukleotidy sa označujú podľa použitej dusíkovej bázy. Zistilo sa, že v celej DNA sa nachádzajú iba 4 druhy dusíkových báz, konkrétne sú to adenín, guanín, tymín a cytozín. Tie sa komplementárne dopĺňajú, a vznikajú nukleotidové páry adenínu s tymínom a cytozínu s guanínom. Tisíce takýchto párov sa skladajú do dvojitej závitnice zobrazenej na Figure 4. Táto závitnica je kľúčom k ochrane a veľkej odolnosti celej molekuly.

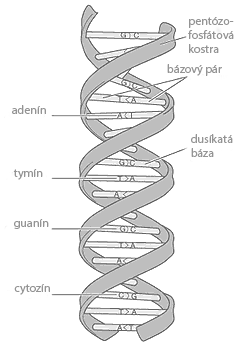


Figure 4 Štruktúra DNA makromolekuly

DNA molekuly zo všetkých ľudských chromozómov obsahujú približne 3 miliardy nukleotidových párov a kebyže sa tieto závitnice rozložia do radu, bol by dlhý zhruba 185cm [[3](#Rob05)]. Pre porovnanie baktérie majú zhruba 4 milióny nukleotidových párov.

Zápis celej alebo časti ľudskej DNA ako poradie nukleotidov sa nazýva **sekvencovanie**. Keďže nukleotidový pár je komplementárny stačí zápis iba jednej strany závitnice. Sú vyvinuté automatizované prístroje, ktoré používajú viacero techník na sekvencizáciu DNA, najznámejšia a najpoužívanejšia je Sangerova metóda.

Je nutné poznamenať, že zosekvencovanie celej ľudskej DNA, takzvané sekvencovanie genómu je časovo a finančne náročné. Avšak trend rapídneho klesania ceny a tým súvisiaceho času je vidno napríklad pri porovnaní ceny zosekvencovania celého genómu jednotlivca. V roku 2001 bola cena 100 miliónov dolárov no v roku 2014 to bolo menej ako 10 tisíc dolárov pri zachovaní menej ako 1% miery chybovosti, ktorá je komunitou akceptovaná [[4](#Wet14)].

* + 1. Gén a mutácia

**Gén** je základná fyzická a funkčná jednotka dedičnosti. Gény môžu byť takzvané non-coding RNA a gény ktoré kódujú syntézu určitého druhu bielkoviny. Bielkoviny sú základné stavebné jednotky organizmu a zabezpečujú všetky fyzické a funkčné vlastnosti jedinca, napríklad farbu vlasov, očí, výšku no aj základné fyziologické vlastnosti ako trávenie a dýchanie. Všetky vlastnosti zákódované v DNA sa označujú ako fenotyp jedinca.

Vieme, že ľudský genotyp sa skladá z približne 21 000 génov. Tie sa líšia v dĺžke, a dosahujú až veľkosť 2,3 milióna nukleotidových párov. Niektoré vlastnosti sú ovplyvnené jediným génom (monogenické) a ovplyvnené skupinov génov (polygenické). Vedci odhadujú, že je 10 000 monogenických genetických ochorení, ako príklady uvedieme cystickú fibrózu alebo Huntingtonovu chorobu.

??Kódovanie bielkovín pomocou RNA – ak nebude dosť teórie??

**Mutácia** je permanentná zmena nukleotidovej sekvencie v DNA. Väčšina mutácií prebehne bez postrehnutia, pretože prebehne v takzvanom „junk DNA“, teda časti sekvencie, ktoré nepatrí do žiadneho génu. No ak prebehne v génovej sekvencii, dôsledky môžu byť fatálne. Napríklad na treťom chromozóme, ktorý nesie 1000-2000 génov sa nachádza jeden ktorý zabezpečuje syntézu rhodopsínu, svetlocitlivej bielkoviny nachádzajúcej sa na sietnici. Je zaznamenaných až 30 mutácií tohto génu, ktoré ovplyvňujú korektné videnie.

Typy mutácií na sekvenčnej úrovni sú [[2](#Har11)]:

* Substitučné, nukleotidový pár sa posunie v sekvencii alebo sa obrátia jeho strany
* Odstránenie, odstránenie jedného alebo viacerých nukleotidových párov
* Vloženie, vloženie jedného alebo viacerých nukleotidových párov
* Inverzia, obrátenie poradia časti sekvencie
  + 1. Projekty

Human Genome Project

Najväčšiemu skoku vo výskume genetiky vďačíme práve HGP [[5](#Tri)]. Začal v roku 1990 a bol označený za úspešne ukončený v roku 2003. Jeho cieľom bolo zosekvencovanie celého ľudského genómu a určenie všetkých génov. Pri začiatkoch sa predpokladalo, že existuje cca. 100 000 génov kódujúcich bielkoviny, no HGP potvrdilo,   
že ich je cca. 21 000 a zdokumentované ich uložilo vo verejných databázach.

Práve tento projekt odštartoval takzvanú genomickú revolúciu [[5](#Tri)]. Vytvoril 310 000 pracovných pozícií a považuje sa za jeden z najväčších vedeckých prínosov v histórií ľudstva. Ovplyvnil mnoho technologických a vedeckých odvetví od zdravotníctva, biotechnológií, poľnohospodárstva, veterinárstva, forénznych vied a mnohých iných.

Genome Browsers

V bioinformatike je nutný rýchly a spoľahlivý prístup k biologickým databázam za účelom získania genomických dát. Za týmto účelom bolo vytvorených viacero „genome browsers“, ktoré poskytujú API alebo UI pre získanie týchto dát. Väčšina obsahuje dáta z tých istých zdrojov a líšia sa iba vo forme.

Práve tieto dáta sú potrebné pre riešenie zadaného problému tejto Diplomovej práce a budú podrobnejšie opísané v Kapitole 2.

* 1. Regulárne výrazy

V Kapitole 1.1.2 bolo ukázané, že ľudská DNA a teda všetky vlastnosti jedinca sú zakódované do postupnosti 4 druhov nukleotidov, konkrétne sú to adenín (A), cytozín (C), tymín (T) a guanín (G). Preto sa ľudská DNA dá zapísať ako postupnosť týchto 4 znakov o veľkosti tri miliardy. V kapitole 1.1.3 zas boli opísané gény a mutácie vo vzťahu práve k tejto postupnosti. Z týchto poznatkov môžme vyvodiť, že určenie génu je vlastne zistenie prítomnosti daného vzoru v reťazci. Tento vzor môže zakomponovávať rôzne variácie reťazca, v tejto implementácií sú to mutácie. Práve na takýto účel boli vytvorené regulárne výrazy.

„Regulárny výraz je zápis popisujúci množinu znakových reťazcov. Keď konkrétny reťazec je v množine popísanej regulárnym výrazom, tak sa hovorí, že reťazec vyhovuje vzoru.“ [[6](#Cox07)] Je široko používaný v teoretickej počítačovej vede a teórií formálnych výrazov.

* + 1. Zápis

V najjednoduchšej forme môže regulárny výraz definovať konkrétne slovo alebo znak, no môže definovať aj zložité vzory napríklad validnú e-mailovú adresu, telefónne číslo, dátum alebo aj cystickú fibrózu v ľudskej DNA.

Časť IEEE POSIX štandardu, BRE (basic regular expression) a ERE (extended regular expression) zjednocuje zápis regulárnych výrazov. Využíva konvenčnú znakovú sadu a definuje metaznaky s kontrolnými funkciami.

Takýmito metaznakmi sú:

* **.** – ľuboľný znak
* **[ ]** - označuje ľubovoľný znak z množiny vo vnútri, napríklad výraz [abc] značí prítomnosť znaku a alebo b alebo c, môže sa použiť aj skrátený zápis   
  [a-c]
* **[^ ]** – negácia množiny vo vnútri zátvoriek
* **^** - začiatok textu alebo riadku
* **$** - koniec textu alebo riadku
* **( )** – definovanie podvýrazu
* **\n** – vloženie n-tého podvýrazu definovaného spôsobom popísaným vyššie
* **\*** - označuje výskyt predchádzajúceho elementu nula alebo viackrát, môže sa použiť s podvýrazom, napríklad (abc)\* vyhovuje “”, “abc”, “abcabc” atď.
* **{m, n}** – označuje výskyt predchádzajúceho elementu minimálne m-krát vrátane a maximálne n-krát vrátane, napríklad a{2,3} vyhovuje slovám “aa” alebo “aaa”
* **?** – označuje výskyt predošlého elementu nula alebo jedenkrát
* **+** – označuje výskyt predošlého elementu jeden alebo viackrát
* **|** – logické alebo

V praxi sa používajú aj takzvané **znakové triedy**. Existuje viacero znakových tried, napríklad malé písmená, všetky alfanumberické znaky, číselné znaky a ďalšie. Zápis týchto tried nieje ustálený a preto tu spomeniem iba najpoužívanejšiu implementáciu pomocou ASCII znakov najdôležitejších tried. Sú to:

* [a-z] – malé písmená
* [A-Z] – veľké písmená
* [0-9] – číselné znaky
* [A-Za-z0-9] – všetky alfanumerické znaky
* [A-Fa-f0-9] – znaky hexadecimálneho čísla

Je nutné poznamenať, že regulárne výrazy niesú obmedzené iba na ASCII znaky a môžu sa použiť aj znakové triedy unicode znakov použitím /u a unicode kódu znaku. Napríklad [/u00C1-/u01C4] obsahuje všetky špeciálne slovenské písmená.

**Príklad**

Demonštrácia zostrojenia regulárneho výrazu je možná na príklade overenia korektnosti zadanej e-mailovej adresy.

Štandardizovaná syntax emailovej adresy je definovaná v RFC 5322. Keďže emailová adresa je case-insensitive, je možné celý vstup pretransformovať na malé písmená. Syntax podľa RFC 5322 sa môže definovať ako *local-part@domain*, kde doménová časť musí obsahovať minimálne dve doménové úrovne a musí sa ukončiť top-level doménov, ktorá sa skladá z dvoch až šiestich ascii znakov.

Najintuitívnejší výraz, zabezpečujúci validitu e-mailovej adresy by mohol byť:   
*^[a-z0-9.]+@[a-z0-9.]+.[a-z]{2,6}$*

Tento výraz má zopár nedostatkov, prvým je nepodporovanie všetkých povolených znakov aj keď všetky sa v praxi nepoužívajú. Dokonca emailová služba spoločnosti google nepodporuje všetky znaky a lokálna časť sa môže skladať iba z [a-z0-9-\_’.].

Takisto adresa sa nemôže začínať ani končiť bodkov a nemôžu sa v nej nachádzať za sebou stojace bodky.

Kompletný regulárny výraz napísaný v javascriptovej syntaxe [[6](#Goy12)]:  
*^[\w!#$%&'\*+/=?`{|}~^-]+(?:\.[\w!#$%&'\*+/=?`{|}~^-]+)\*@(?:[A-Z0-9-]+\.)+*

*[A-Z]{2,6}$* [[7](#Cox07)]

* + 1. Konečné automaty

Najefektívnejším a najfundamentálnejším testovaním regulárnych výrazov je pomocou konečných automatov.

**Automat je ...**

Zostrojenie konečného automatu na určenie validnej emailovej adresy.

Výpočtová zložitosť...

* + 1. Použitie
  1. Distributívne systémy

1. Opis riešenia
2. Zhodnotenie

**Záver**

**Zoznam použitej literatúry**

x

|  |  |
| --- | --- |
| [1] | RN, PhD Anne Matthews. (2010, June) http://www.netwellness.org/. [Online]. <http://www.netwellness.org/healthtopics/idbd/2.cfm> |
| [2] | Leland Hartwell, Leroy Hood, Michael Goldberg, Ann Reynolds, and Lee Silver, *Genetics: from genes to genomes*, 4th ed. New York, USA: McGraw-Hill, 2011. |
| [3] | Tara ,PhD. Robinson, *Genetics for dummies*, 1st ed. Hoboken, USA: Wiley Publishing, Inc., 2005. |
| [4] | Kris, M.S. Wetterstrand. (2014, Oct.) www.genome.gov. [Online]. <http://www.genome.gov/sequencingcosts/> |

x

**Prílohy**

Príloha A: Nadpis. . . . . . . . . . . . . . . . . . . . . . . . . II

Príloha A: Nadpis