VYSOKÉ UČENÍ TECHNICKÉ V BRNĚ
BRNO UNIVERSITY OF TECHNOLOGY

FAKULTA INFORMAČNÍCH TECHNOLOGIÍ
ÚSTAV INTELIGENTNÍCH SYSTÉMŮ

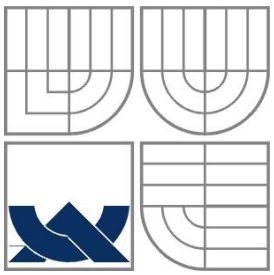
FACULTY OF INFORMATION TECHNOLOGY
DEPARTMENT OF INTELLIGENT SYSTEMS

PREDIKCE VAZEBNÍCH MÍST PROTEINU P53

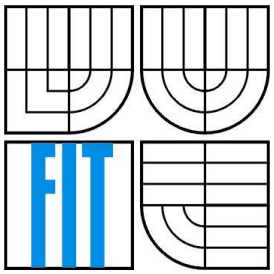
DIPLOMOVÁ PRÁCE
MASTER'S THESIS

AUTOR PRÁCE Bc. Jozef Radakovič
AUTHOR

BRNO 2015



VYSOKÉ UČENÍ TECHNICKÉ V BRNĚ
BRNO UNIVERSITY OF TECHNOLOGY



FAKULTA INFORMAČNÍCH TECHNOLOGIÍ
ÚSTAV INTELIGENTNÍCH SYSTÉMŮ

FACULTY OF INFORMATION TECHNOLOGY
DEPARTMENT OF INTELLIGENT SYSTEMS

PREDIKCE VAZEBNÍCH MÍST PROTEINU P53

PREDICTION OF P53 PROTEIN BINDING SITES

DIPLOMOVÁ PRÁCE
MASTER'S THESIS

AUTOR PRÁCE
AUTHOR

Bc. Jozef Radakovič

VEDOUCÍ PRÁCE
SUPERVISOR

Ing. Martínek Tomáš, Ph.D.

BRNO 2015

Abstrakt

Proteín p53, ktorý je kódovaný génom TP53 zohráva významnú úlohu v bunečnom cykle, ako regulátor transkripcie génov pri reakcii bunky na stresové podnety, čím funguje ako potláčateľ rakoviny. Pochopenie spôsobu jeho regulácie ako aj jeho väzby na regulovaný gén je jedným z hlavných záujmov moderného výskumu v genetike a bioinformatike. V prvej časti tejto práce predstavujeme nevyhnutné poznatky z molekulárnej biológie nutné k pochopeniu spôsobu regulácie proteínu p53 a úvod do analýzy a predikcie väzobných miest transkripčných faktorov. V druhej časti sa venujeme implementovaniu a testovaniu nami vytvoreného nástroja, ktorý bude schopný tieto väzobné miesta pre proteín p53 predikovať.

Klíčová slova

DNA, proteín p53, transkripčné faktory, Skryté Markovské Modely, HMM

Abstract

Protein p53 which is encoded by gene TP53 plays crucial role in cell cycle as a regulator of transcription of genes in cases when cell is under stress. Therefore p53 acts like tumor suppressor. Understanding the pathway of p53 regulation as well as predicting its binding sites on p53 regulated genes is one of the major concerns of modern research in genetics and bioinformatics. In first part of this project we aim to introduce basics from molecular biology to better understand the p53 protein pathway in gene transcription and introduction to analysis of prediction of p53 binding sites. Second part is about implementation and testing of tool which would be able to predict transcription factor binding sites for protein p53.

Keywords

DNA, protein p53, transcription factors, Hidden Markov Models, HMM

Predikce vazebních míst proteinu p53

Prohlášení

Prohlašuji, že jsem tuto diplomovou práci vypracoval samostatně pod vedením Ing. Tomáša Martínka, Ph.D.
Uvedl jsem všechny literární prameny a publikace, ze kterých jsem čerpal.

……
Jozef Radakovič
27.5.2015

Poděkování

Rád by som sa poďakoval prof. Jiřímu Doškařovi za uvedenie do problematiky molekulárnej genetiky v rámci absolvovanému predmetu na FIT. Takisto ďakujem Ing. Tomášovi Martínkovi, Ph.D, za poskytnutie značného množstva zdrojovej literatúry a možnosti "pasívnemu" zorientovaniu sa v problematike vďaka kvalitným materiálom v predmete Bioinformatika.

Obsah

1 Úvod

Proteín p53, tiež známy ako bunečný nádorový antigén p53, alebo supresor nádoru p53, hrá kľúčovú úlohu v bunečnom cykle, kde ako supresor nádoru zabraňuje vlastne vzniku rakoviny.

Cieľom tejto práce je oboznámiť sa so základmi molekulárnej biológie, od jednotlivých stavebných elementov, ktoré tvoria genetickú informáciu organizmu, až po vytváranie proteínov na základe génov uložených v DNA. Tejto problematike sa venuje prvá časť tejto práce.

Druhá časť predstavuje niektoré často využívané metódy analýzy a predikcie miest väzby proteínov ako transkripčných faktorov na DNA pri expresii (vytváraní, resp. syntéze) proteínu z génu. Časť práce, ktorá je venovaná tejto problematike je ukončená predstavením algoritmu p53HMM a jeho Markovského modelu použitom autormi v [1].

Tretia časť popisuje samotný návrh a implementačné detaily nástroja, ktorý je schopný spracovať vstupné dáta a pomocou použitia externých nástrojov vytvoriť model, ktorý bude schopný predikovať väzobné miesta transkripčného faktoru proteínu p53.

Vstupné data, a teda samotné väzobné miesta sú popísané v štvrtej časti. Popisujeme tu ich zdroj, formát a spôsob ako sme ich spracovali.

Externé nástroje ktoré používame sú podrobné predstavené a rozobraté v nasledujúcej kapitole.

Práca je ukončená samotnými testami vytvoreného produktu. Testujeme tu schopnosť vytvoriť vhodný predikátor z celej vstupnej množiny dát, rôznych zaujímavých podmnožín a taktiež testujeme citlivosť aplikácie na rôzne vstupné parametre.

2 Základy z molekulárnej genetiky

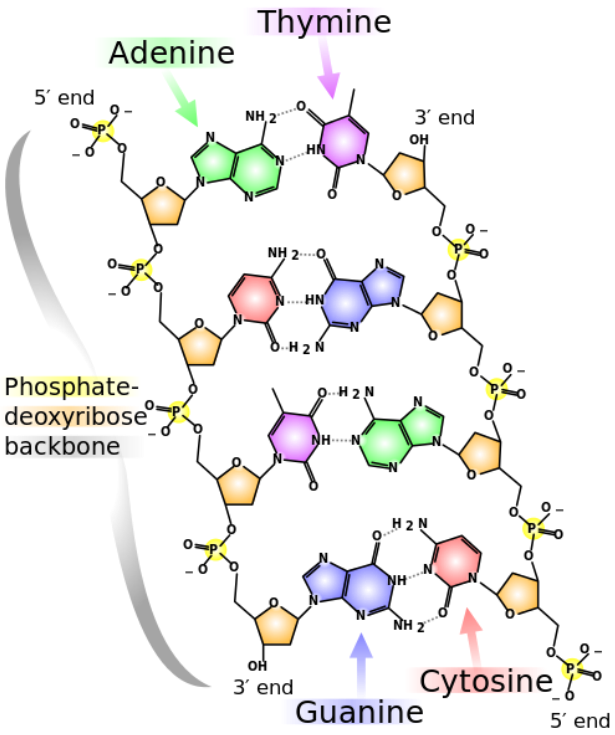
2.1 DNA

Deoxyribonukleová kyselina je prírodný polymér zložený z deoxiribonukleitodov. Patrí spolu s RNA medzi nukleové kyseliny a je nositeľkou genetickej informácie bunky. Riadi jej rast, delenie a regeneráciu. DNA je zložená z jednotiek zvaných monoméry - nukleotidy, ktoré sú navzájom pospájané esterovou väzbou. Nukleotid je zložený z:

- zvyšku kyseliny trihydrogenfosforečnej
- deoxyribonukleozid, tvorený dvoma zložkami
    - cukor ribonukleáza
    - dusíkatá báza

Postupnosť, alebo sekvencia nukleotidov tvorí reťazec zvaný vlákno DNA, ktoré ak je v nespojenom tvare, tak na 5´konci je ukončené fosfátom a na 3´ konci je ukončené hydroxilovou -OH skupinou. Ribonukleáza a kyselina fosforečná slúžia k tomu, aby udržali dusíkaté bázy vo vhodných polohách a vzdialenostiach a tvoria takzvanú pentózafosfátovú kostru DNA. Pretože kyselina fosforečná a ribonukleáza sú spoločné pre všetky nukleotidy, tieto sa líšia v dusíkatých bázach. Dusíkaté bázy, ktoré sa vyskytujú v DNA sú:

- adenín (A)
- cytozín (C)
- guanín (G)
- tymín (T)



Obrázok 1 Chemická štruktúra DNA (Prevzaté z en.wikipedia.org)

DNA je tvorená dvoma vláknami, ktoré sú k sebe komplementárne (adenín a tymín - sú spojené tromi vodíkovými väzbami, cytozín a guanín - dve vodíkové väzby). Je to dané tým, že adenín a guanín patria medzi puríny, zatiaľ čo cytozín a týmín medzi pyrimdíny. Podľa pravidla, ktoré sa nazýva podľa objaviteľov Watson-Crickovo pravidlo, sa páruje vždy purínová a pyrimidínová báza, nakoľko takéto párovanie je najstabilnejšie. Reťazce DNA sú voči sebe postavené v opačnom smere, jeden v smere 5´- 3´ a druhý v smere 3´- 5´, pričom poradie nukleotidov sa číta vždy v smere od 5´ konca ku 3´ koncu. Toto vlákno sa nazýva aj kódujúce, sense a naopak smer 3´- 5´ sa nazýva nekódujúce, anti-sense, prípadné templátové (angl. template strand).

DNA vytvára komplexnejšie štruktúry, ktoré sa delia podľa úrovne zložitosti (resp. pohľadu):

- primárna - je daná fyzickým poradím jednotlivých nukleotidov
- sekundárna - priestorové usporiadanie polynukleotidového reťazca - dvojvlákno má najčastejšie tvar pravotočivej dvojzávitnice (alfa helix)
- terciálna - priestorové usporiadanie dvojzávitnice - môže sa stočiť do superhelixu v tzv. nadzávitnicovom vinutí, ktoré je zabezpečené špeciálnymi enzýmami

2.2 Gén

Gén je informačná a funkčná jednotka obsahujúca genetickú informáciu o primárnej štruktúre funkčnej molekuly translačného produktu (tým je proteín), alebo funkčnej molekuly produktu transkripcie RNA (napr. tRNA, rRNA, snRNA), ktorá nepodliehajúcej translácii.

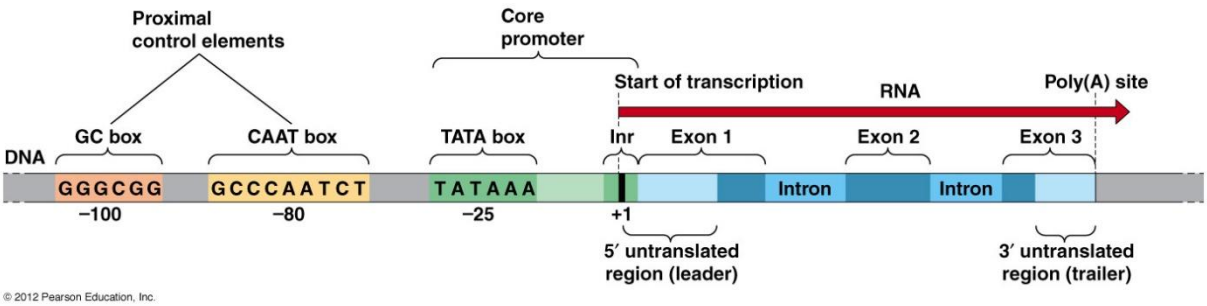
Konkrétne formy génu sú:

- štruktúrny gén - kóduje polypeptid
- gén pre funkčné typy RNA - prepisuje sa do tRNA, rRNA, sRNA (a ďalších funkčných typov)
- gén ako regulačná oblasť - úsek DNA, na ktorú sa viaže regulačný proteín

V rámci práce nás zaujímajú hlavne štruktúrne gény a regulačné oblasti.

Štruktúrny gén je zložený z promotóru, kódujúcej oblasti a sekvencií označujúcich začiatok a koniec transkripcie z tripletov (kodónov). Promotórová oblasť sa nachádza na začiatku génu, na ktorú sa naviaže transkripčný aparát (RNA-polymeráza) a ďalšie transkripčné regulačné signály. Regulačná oblasť génu obsahuje určité sekvencie nukleotidov, pričom promotóry rôznych génov sa navzájom líšia počtom, kombináciou a umiestnením týchto sekvencií. Začiatok a koniec transkripcie ďalej obsahuje špecifické sekvencie, ktoré sú rozpoznávané transkripčným aparátom. Tieto úseky teda nekódujú aminokyseliny, ale nesú informácie, ktoré sú nutné k celkovému priebehu proteosyntézy. Proteosyntéza je proces v bunke, počas ktorého sa tvoria bielkoviny.

Kódujúca oblasť sa skladá z exónov a intrónov. Exóny (z angl. expressed region), tiež označované aj ako CDS (Coding segment), kódujú samotné aminokyseliny. Každý triplet, čo sú tri po sebe idúce bázy, predstavuje informáciu o zaradení jednej aminokyseliny. To, ktoré triplety prislúchajú ktorej aminokyseline určuje genetický kód. Intróny sú nekódujúce úseky, ktoré aj keď sú z DNA prepisované, nakoniec sa z RNA vystrihnú v procese zvanom zostrih. Intróny môžu prerušiť, resp. byť vložený do exónu v ktoromkoľvek mieste. Exón potom pokračuje za koncom intrónu.



Obrázok 2 Eukaryotický gén (Prevzaté z en.wikipedia.org)

Na DNA sa tiež nachádzajú úseky, ktoré sú rôzne vzdialené od promotóru génu, ale ovplyvňujú transkripciu génu. Ide o väzobné miesta transkripčných faktorov (TFBS - transcription factor binding sites). Sú definované určitým vzorom poradia nukleotidov, nazývaným tiež logo, alebo konsensus sekvencia. Na tieto oblasti sa viažu regulačné proteíny pôsobiace ako transkripčné faktory, a teda ovplyvňujúce transkripciu. Eukariotické gény sú charakteristické prítomnosťou viacerých TFBS.

2.3 Transkripcia a translácia génu

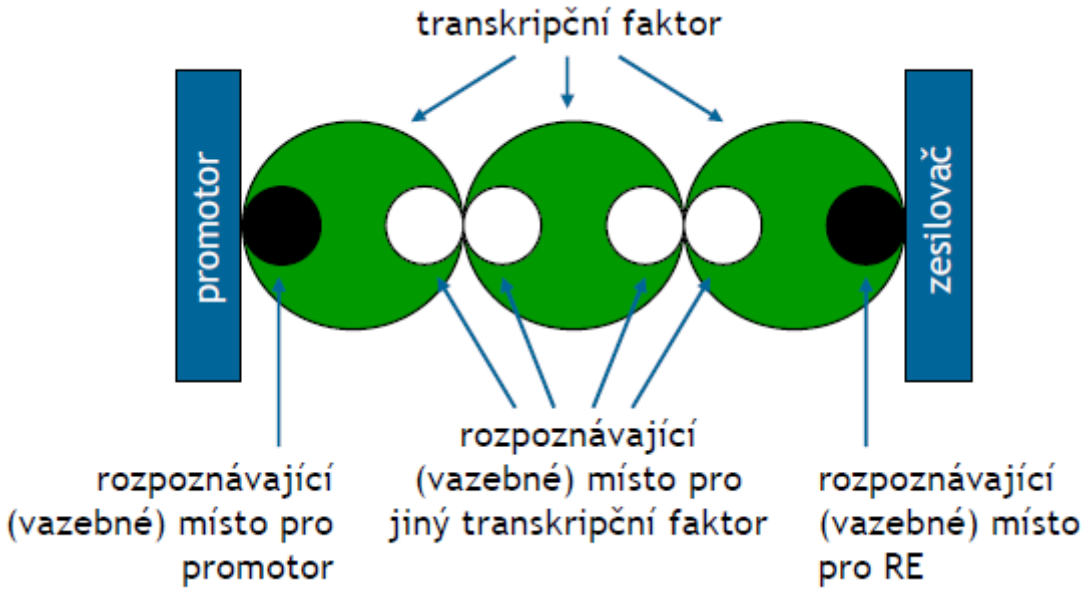
Expresia génu v eukaryotických bunkách sa vykonáva na niekoľkých úrovniach, kde najzaujímavejšie z nášho pohľadu sú transkripcia a translácia.

Transkripcia je proces, kedy sa poradie nukleotidov z DNA prepisuje do poradia nukleotidov v RNA. Prebieha pomocou párovania komplementárnych báz a pozostáva z troch častí:

- iniciácie
- elongácie
- terminácie

Pri iniciácii sa naviažu transkripčné faktory a RNA-polymeráza, ktorá je kľúčový enzým nutný pri syntéze RNA, na promotór génu. Počas procesu elongácie sa RNA-polymeráza posúva po DNA v smere 5´-3´ a syntetizuje RNA. Pri terminácii sa RNA-polymeráza spolu s nasyntetizovaným RNA reťazcom (transkriptom) a transkripčnými faktormi odpojí (disociuje) z templátu DNA. RNA transkript, ktorý je syntetizovaný z DNA, tiež nazývaný primárny transkript musí podstúpiť ďalšie postranskripčné úpravy (ukončenie reťazca, zostrih exónov), čím sa vytvorí mRNA.

Transkripcia je regulovaná regulátormi transkripcie a to sú buď **proteíny, ktoré riadia zahájenie transkripcie** - a teda pozitívne, vtedy ide o pozitívnu reguláciu (naviazaním pozitívneho regulačného signálu na operátorovú oblasť promotóra , čím sa zvýši hladina transkripcie), alebo negatívne (naviazaním represora na operátor, čím sa zabráni aby RNA-polymeráza iniciovala transkripciu z promotóra). Druhý typ regulačných proteínov sú tzv. **zosilovače** (enhancery) transkripcie, resp. v prípade potláčania transkripcie **silencery**. Tieto sa viažu na rozpoznané sekvenčné motívy, ktoré môžu byť jednak v oblasti promotóru, alebo v rôznych vzdialenostiach od promotóru, prípadne na iné transkripčné faktory.



Obrázok 3 Väzobné miesta transkripčných faktorov a ich vzájomná interakcia (Prevzaté z [11])

Pri **translácii** dôjde k prekladu poradia nukleotidov z mRNA do poradia aminokyselín v polypeptidovom reťazci.

Aj keď je genóm rovnaký v každej bunke, v organizme sa nachádzajú rôzne druhy buniek. Množina aktivovaných a deaktivovaných génov je teda rozdielna pre každú bunku, čo je dosiahnuté regulačnými mechanizmami zahrňujúcimi transkripčné faktory a ich sekvencií väzieb na DNA, štruktúry chromatínu (chromatín je komplex DNA a niektorých proteínov v jadre) a histónovou modifikáciou. Je známe, že expresia génu je regulovaná v každej fáze od RNA syntézy až po RNA preklad do proteínu. Rôzne procesy a faktory, ktoré sú zahrnuté v regulácii sú navzájom prepojené.

2.4 Proteíny ako transkripčné faktory

Proteíny sú veľké biologické molekuly pozostávajúce z jedného, alebo viacerých reťazcov aminokyselín, ktoré sú pospájané peptidovou väzbou medzi karboxylovou a amino skupinou.

Sekvencia aminokyselín v proteíne je definovaná sekvenciou génu, ktorý je zakódovaný v genetickom kóde. Genetický kód popisuje 20 štandardných aminokyselín.

Funkcia proteínu je daná jeho štruktúrou. Tá má štyri úrovne:

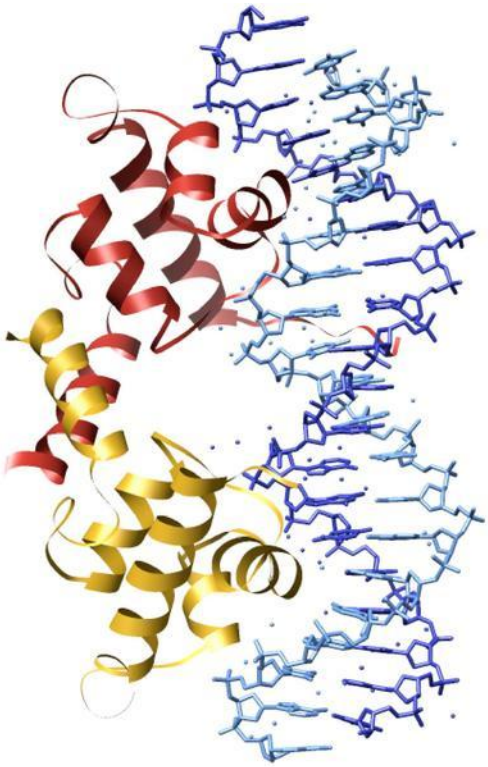
- **primárna** - sekvencia aminokyselín, ktoré tvoria molekulu
- **sekundárna** - označenie pre priestorové usporiadanie spojených aminokyselín, ktoré sa vďaka vodíkovo mostíkovým väzbám rôzne skladajú a tvoria štruktúry ako alfa-hélix (pravotočivá závitnica), beta-štruktúra (skladaný list beta) či iné zatočenia
- **terciálna** - niekoľko sekundárnych štruktúr vytvára zložitejší 3D objekt
- **kvartérna** - je tvorené niekoľkými stabilne spojenými proteínovými reťazcami, sú to napr. enzým, alebo vírus

Niektoré proteíny majú schopnosť viazať sa na DNA. Bez týchto proteínov by DNA nebola schopná sa replikovať, či vykonávať svoju funkciu. Schopnosť proteínu viazať sa je vďaka úsekom, ktoré obsahujú vhodný motív, ktorý dokáže rozpoznať sekvenciu DNA (tzv. rozpoznávacia sekvencia), alebo má istú silu priľnavosti k DNA. Dôvod viazania sa k DNA môže byť štruktúrny (spolupodieľajú sa u replikácii, oprave, ukladaní a modifikácii DNA), alebo regulujúca transkripciu génov. Proteíny, ktoré regulujú transkripciu génov sa nazývajú **transkripčné faktory.**

Proteíny sa môžu viazať na DNA buď na základe poradia nukleotidov v DNA, alebo voľne. Je známe, že aj toto voľné viazanie zahrňuje čiastočnú molekulárnu komplementaritu medzi DNA a proteínom. Pri rozpoznávaní DNA na základe pripojovacej domény (z angl. binding domain) sa proteín môže prichytiť k veľkému, alebo malému žľabu (veľký a malý žľab sú medzery medzi závitmi vytvorené vďaka dvojzávitnicovému usporiadaniu DNA) vytvorenému na dvojzávitnici DNA, alebo k cukor-fosfátovej kostre.

2.5 Kvartérne štruktúry proteínov

Ako už bolo spomenuté v predchádzajúcej podkapitole, jednotlivé proteíny sa môžu navzájom spájať a vytvárať kvartérne štruktúry. Spojenie je najčastejšie umožnené kovalentnou, alebo medzimolekulárnou väzbou. V prípade transkripčných faktorov často regulačný proteín vytvára kvartérnu štruktúru a viaže sa na niekoľko väzobných miest na DNA naraz. Ak takúto štruktúru tvoria rovnaké proteíny, potom má tento objekt predponu *homo.* V prípade p53 sa stretneme s dvoma typmi týchto zoskupení, ide o homodimér a homotetramér. Ako je už z názvov zrejmé, v prípade homodiméru sú spojené dva proteíny a v prípade homotetraméru štyri identické proteíny.



Obrázok 4 Proteín vo forme homodiméru naviazaný na DNA (Prevzaté z en.wikipedia.org)

2.6 Proteín p53

Proteín p53 hrá veľkú úlohu v prevencii rozvoja nádoru a teda rakoviny. Je kódovaný v DNA v TP53 géne, ktorý sa nazýva gén potláčajúci nádory. Reaguje na množstvo onkogénových stresov (onkogén - gén, ktorý ma potenciál spôsobiť rakovinu, v rakovinových bunkách sú často zmutované) aktivovaním ochranných mechanizmov. Medzi hlavné patria uzatvorenie bunečného cyklu a apoptóza (apoptóza je programovaná bunečná smrť). Jeho dôležitosť v potláčaní rakoviny sa odráža vo vysokej frekvencii mutácii, kde viac ako 50% ľudských tumorov je spojených s inaktivačnými mutáciami, alebo deléciami (delécia je odstránenie, resp. zmazanie, jedného, alebo viacerých nukleotidov z DNA) v TP53 géne. Pri mnohých nádoroch, kde p53 sa chová prirodzene, smery jeho pôsobenia môžu byť pozmenené inými onkogénovými faktormi a preto sa predpokladá, že odpovede, resp. reakcie p53 sú nefungujúce pravdepodobne u väčšiny nádorov.

2.7 Motív p53

Sekvencia, ku ktorej sa proteín p53 pripája s vysokou afinitou zodpovedá reťazcu **5´-RRRCWWGYYY-3´,** kde R je purín (purínové bázy sú adenín a guanín), Y je pyrimidín (pyrimidínové bázy sú tymín a cytozín), W môže byť A (adenín), alebo T (tymín), G je guanín a C je cytozín.

Oblasť pripojenia p53 v genóme mnohých organizmov je zložená z polo-úseku (half-site) **RRRCWWGYYY**, ktorý je nasledovaný medzerou, alebo výplňou - ide o sekvenciu nukleotidov, zvyčajne zloženou z 0-21 párov báz. Tieto sú nasledované druhým polo-úsekom **RRRCWWGYYY**. Ak označíme každý štvrť-úsek RRRCW ako → a WGYYY ako ←, potom môžeme graficky označiť miesto väzby napr. ako →← medzera →←. Táto konfigurácia štyroch štvrť-úsekov sa často označuje ako orientácia hlava k hlave (head-to-head - skr. HH). Ďalšie možné orientácie štvrť-úsekov sú chvost k chvostu (tail-to-tail, TT, ←→ medzera ←→) a hlava k chvostu (head-to-tail, HT, →→ medzera →→). Orientácia chvost k hlave sa nepoužíva, pretože komplementárne DNA vlákno zákonite musí byť tvorené orientáciou hlava k chvostu (HT).

V takmer všetkých prirodzených p53 úsekoch spojenia (response element RE) dva pol-úseky zdieľajú rovnakú štvrť-úsekovú orientáciu. Experimenty ukázali, že tetramér p53 proteín sa môže pripojiť k všetkým trom (HH, TT a HT) štvrť-úsekom s rovnakou afinitou, aj keď len niekoľko z experimentálne potvrdených p53 úsekoch spojenia nemajú HH orientáciu. Pretože sú povolené vloženia a vymazania jednotlivých nukleotidov v rámci väzobného úseku p53, dĺžka pol-úseku sa pohybuje medzi 8 a 12 pármi báz, najčastejšie 10. Niektoré väzobné oblasti p53 majú viac ako dva pol-úseky, v takomto prípade sa označujú ako klastrované oblasti (cluster sites). Rôzne experimenty ukázali, že úroveň vôle pripojenia p53 rastie lineárne s počtom pol-úsekov. Niektoré gény dokonca obsahujú niekoľko p53 úsekov spojenia v rôznych miestach v rámci génu, alebo v oblasti promotéru a každý jeden úsek môže prispieť k ovplyvnenie expresie génu proteínom p53. Napr. úsek →→→←→ sa vyskytuje v promotéry génu CDKN1A (proteín p21) nasledovaný zhruba 900 pármi báz k ďalšiemu klasickému miestu väzby p53 →←medzera→← a obidva miesta prispievajú k transkripcii génu CDKN1A.

2.8 Faktory ovplyvňujúce reguláciu p53

Výskumy ukázali, že veľa faktorov môže ovplyvniť spôsob a stupeň do akého p53 reguluje transkripciu génov. Tieto faktory môžu byť kofaktory, rôzne dĺžky medzier, orientácia štvrť-úsekov, nukleosómy (globulárna častica, ktorá je súčasťou chromozómu - guľatý tvar zložený z 8 molekúl histónov obmotaných vláknom DNA) a post translačné modifikácie p53. Niektoré faktory ovplyvňujúce reguláciu p53 môžu byť zaujímavé pre túto prácu, preto si ich bližšie popíšeme.

2.8.1 Flexibilná CATG sekvencia

Experimentálne bolo ukázané, že v orientácii hlava k hlave (HH), p53 preferuje opakovaný RRRCATGYYY motív. Podrobnejšou štúdiou bolo ukázané, že najdôležitejšie bázy pre interakciu s proteínom p53 je centrálna časť RCWWGY, kde dochádza k úzkemu kontaktu s aminokyselinami z centrálnej domény p53. V nadväznosti na túto skutočnosť, po zarovnaní všetkých experimentálne

potvrdených funkčných p53 reakčných oblastí, najviac zakonzervované (najstabilnejšie) sú centrálne CWWG nukleotidy v každom pol-úseku, hlavne C a G. To znamená, že zmeny, ktoré sa vykonajú na týchto centrálnych pozíciách, by mali najviac ovplyvniť schopnosť spojenia DNA s p53. Výskum potvrdil, že viac ako 50% miest s vysokou priľnavosťou p53 obsahovalo CATG sekvenciu v centre oboch oblastí.

Pretože CATG sekvencia sa vykazuje neobvyklou schopnosťou ohybu u veľa DNA-proteínových komplexov, usudzuje sa, že táto ohybnosť taktiež ovplyvňuje schopnosť afinity p53. Experimenty ukázali, že schopnosť ohybu p53 a DNA v CATG sekvencii, vysoko koinciduje s vyššou priľnavosťou k DNA oblastiam obsahujúcim gény ovplyvňujúce bunkový cyklus, než úsekmi DNA obsahujúce gény využívané pri apoptóze.

2.8.2 Vzdialenosť a zatočenie DNA

Je známe, že vzdialenosť pripojovacej oblasti cis-elementu (cis-element je oblasť DNA, alebo RNA, ktorá ovplyvňuje expresiu génu nachádzajúcom sa na rovnakej molekule DNA) a oblasti začiatku transkripcie (transcription start site - TSS) môže významne ovplyvniť stupeň regulácie génu. Výskumom bolo preukázané, že vložením 200 párov bází medzi TATA box a p53 RE eliminovalo 45-preloženie, ktoré bolo indukované p53. Je tiež známe, že eukaryotické bunky obsahujú proteíny pripájajúce sa ako transkripčný faktor, ktorý spôsobujú zatočenie DNA a spájajú takéto proteíny k sebe. Vďaka tomu je možné dostať vzdialený transkripčný faktor blízko k TATA boxu natoľko, že môže prispieť k regulácii. Napr. pomocou elektrónového mikroskopu sa podarilo zistiť, že p53 tetraméry naukladané na seba pripojené k DNA boli takto schopné vďaka zatočeniu DNA priblížiť vzdialené miesta väzby p53. Taktiež sa ukázalo, že osamelé vzdialené miesta väzby p53 majú slabú schopnosť indukovať transkripciu, ale že blízkosť miesta pripojenia p53 k TSS spôsobí 25-ohyb, čo zapríčiní priblíženie vzdialených p53 k TSS a tým zvýši ich koncentráciu v okolí TSS. V prípade, že chýba blízky p53, môže byť nahradený iným "lepkavým" proteínom, ak ich miesta pripojenia sú blízko k TSS a distálne k miestu pripojenia p53.

2.8.3 Medzery

Experimentálne bolo ukázané, že medzery, ktoré oddeľujú pol-úseky, môžu významne ovplyvniť schopnosť pripojenia proteínu p53. Napr. série experimentov ukázali bimodálnu distribúciu schopnosti priľnavosti p53 na DNA s vrcholmi v 0 a 10 (dĺžka medzery). Hypotéza vyslovená autormi týchto experimentov znela, že optimálna schopnosť priľnavosti nastane, keď obe pol-úseky sú spolu na rovnakej strane dvoj závitnice DNA, alebo keď sú oddelená jedným závitom (10 párov báz). Iní výskumníci ukázali, že v rámci istých podmienok, medzery dĺžky 4, 13 a 14 významne znižujú schopnosť priľnavosti p53 v porovnaní so žiadnou medzerou (medzera 10 bp nebola

testovaná). Zaujímavé je, že databáza 160 funkčných miest pripojenia p53 neukazuje bimodálne rozdelenie dĺžky medzier. Je zrejmé, že vplyv veľkosti medzier je významný na schopnosť pripojenia sa p53, avšak nemusí mať až taký vplyv na samotnú reguláciu p53. Aj keď je teda zrejmé, že schopnosť pripojenia je ovplyvnená veľkosťou medzier, nie je možné túto schopnosť, alebo efekt kvantifikovať.

3 Výpočtové metódy predikcie väzobných miest p53

Konkrétna štruktúra a pozícia väzobných miest proteínu p53 sa získava experimentálne na živých bunkách. Dáta získané a potvrdené týmto spôsobom môžu byť využívané k ďalšej predikcii väzobných miest p53. V tejto oblasti existuje niekoľko metód, alebo výpočtových modelov, ktoré sa k predikcii využívajú. Ich aplikácia je založená na schopnosti vyhľadávať známe vzory sekvencií nukleotidov - motívy.

Hlavné a najčastejšie používané metódy sú Position-Specific Scoring Matrix (PSSM) a metóda založená na skrytých Markových modeloch (Hidden Markov Model - HMM). Ďalšie metódy sú založené na komparatívnej genomike, teórii informácie a metódy strojového učenia.

3.1 PSSM

Position-specific scoring matrix (PSSM), tiež známe ako position-specific weight matrix (PSWM), alebo position weight matrix (PWM) sa dá voľne preložiť ako pozičná váhová matica. Často sa používa k reprezentácii vzorov sekvencie nukleotidov.

PSSM je daná obdĺžnikovou maticou, kde riadky predstavujú jednotlivé nukleotidy (celkovo má teda štyri riadky) a stĺpce predstavujú pozíciu v danej sekvencii, resp. vzore. Základná PSSM používajúca relatívne frekvencie sa dá pre každú pozíciu vzoru vypočítať ako normalizovaný pomer frekvencie nukleotidu na danej pozícii. Teda ak máme množinu *X N* zarovnaných sekvencií dĺžky *l*, prvky matice sa vypočítajú ako:

$$\boldsymbol{M_{k,j} = \frac{1}{N}\sum_{i=1}^{N} I(X_{i,j} = k),} \qquad \textbf{(1)}$$

kde $i \in (1, ..., N)$, $j \in (1, ..., l)$, k je symbol nukleotidu (A, G, C, T) a $I(a = k)$ je funkcia, ktorá vráti *1* ak $a = k$, inak vráti *0*. Je zrejmé, že suma pravdepodobností symbolov na jednotlivých pozíciách bude 1. Preto je jednoduché vypočítať pravdepodobnosť nejakej sekvencie vynásobením pravdepodobností na jednotlivých pozíciách.

V praxi sa často využíva log-pravdepodobnosť, to znamená, že sa matica transformuje pomocou modelu pozadia (background model), ktorý môže hovoriť, že niektoré symboly sa vyskytujú v skúmanej sekvencii častejšie než iné. Výsledná hodnota prvku matice bude teda:

$$M_{k,j} = ln(M_{k,j} \,/\, b_k), \tag{2}$$

kde b_k je pravdepodobnosť symbolu v modeli a *ln* prirodzený logaritmus. V prípade log-pravdepodobnosti sa mení výpočet celkovej pravdepodobnosti nejakej sekvencie, kde sa pravdepodobnosti jednotlivých pozícií sčítajú (na rozdiel od použitia jednotlivých sekvencií, kde sa pravdepodobnosti násobia).

3.2 Skryté Markovské Modely (HMM)

3.2.1 Markovský proces

Markovský proces je náhodný proces [11], ktorého budúce pravdepodobnosti sú určené jeho najposlednejšími hodnotami. Táto vlastnosť sa často označuje ako *Markov property* a dá sa vyjadriť vzťahom:

$$P[q_t = S_j | q_{t-1} = S_i, q_{t-2} = S_k, \dots] = P[q_t = S_j | q_{t-1} = S_i\ ], \tag{3}$$

kde q_t znamená aktuálny stav v čase *t*, ktorý sa môže v každom okamžiku rovnať jednému zo stavov $S_1, S_2, \dots, S_n$.

3.2.2 Definícia HMM

Skrytý Markovský model je stochastický automat HMM = (N, M, A, B, π), kde:

- N označuje počet stavov HMM v množine S = {S_1, ..., S_N} s hodnotou q_t v čase t
- M označuje počet navzájom rôznych pozorovaných symbolov $v_1, \dots, v_M$ s hodnotou O_t v čase *t*. Pozorované symboly odpovedajú fyzickému výstupu automatu.
- $A = \{a_{ij}\}$ označuje pravdepodobnostné rozdelenie prechodov medzi jednotlivými stavmi. Pravdepodobnosť

$$a_{ij} = P\left[q_{t+1} = S_j | q_t = S_i\right], \quad 1 \le i, j \le N \tag{4}$$

znamená, že automat nachádzajúci sa v čase q_t v stave S_i, prejde s pravdepodobnosťou $a_{i,j}$ do stavu S_j. *A* je štvorcová matica rádu *N*.

- B = {b_j(k)} označuje pravdepodobnostné rozdelenie pozorovaných symbolov v jednotlivých stavoch. Pravdepodobnosť

$$b_j = P[v_k \text{ v čase } t | q_t = S_i], \quad 1 \le i \le N, 1 \le k \le N \tag{5}$$

znamená, že automat generuje v čase q_t pozorovaný symbol v_k a nachádza sa v stave S_i. *B* je matica s rozmermi *N x M.*

- *π = {πi}* označuje počiatočné rozdelenie pravdepodobností jednotlivých stavov.

$$\boldsymbol{\pi_i = P[q_1 = S_i],} \quad \mathbf{1} \leq \boldsymbol{i} \leq \boldsymbol{N} \tag{6}$$

znamená pravdepodobnosť, že sa automat v čase q_1 nachádza v stave S_i. *π* je *N*-rozmerný vektor.

Pretože PSSM je daný maticou s pevnými rozmermi, nie je schopný tento systém zachytiť (modelovať) situáciu, kedy v motíve väzobného miesta dôjde k odmazaniu, alebo pridaniu nukleotidu. Pri analýze väzobného motívu uvidíme, že takáto situácia nastáva pomerne často.

3.3 Komparatívna genomika

Táto metóda skúma evolučne zakonzervované pozície väzby p53 medzi človekom a inými druhmi organizmov. Z výsledkov výskumu s genómom myší, zajacov, potkanov a psov sa ukázalo, že mnoho väzobných miest zostalo zachovaných. Na druhú stranu sa zistilo, že mnohé nielen väzobné miesta, ale aj činnosť p53 sa evolúciou zmenila [2].

3.4 Metódy strojového učenia

V rámci metód strojového učenia boli v niekoľkých prácach použité support vector machines (SVM). SVM patrí medzi metódy učenia sa s učiteľom, ktorej princíp je optimálne rozdelenie vstupnej množiny dát takou nadrovinou, od jednotlivé body ležia v opačných polpriestoroch a hodnota vzdialenosti týchto bodov od rozdeľujúcej roviny je čo najväčšia. Na popis nadroviny stačí len pár blízkych bodov - tieto sa nazývajú podporné vektory (angl. support vectors). SVM je teda binárny lineárny klasifikátor. Výsledky klasifikácie záležia od použitého jadra - to je funkcia, ktorá počíta podobnosť dvoch príznakových vektorov. V základe sa používa lineárne binárne jadro, avšak výmenou jadra za také, ktoré efektívne mapuje vstupy do viacrozmerných priestorov, je SVM schopné aj nelineárnej klasifikácie [2].

V rámci výskumu použitia SVM pre predpovedanie väzobných miest p53, bolo skúmané okolie experimentálne potvrdených väzobných miest p53, tak aby sa našli často sa vyskytujúce funkčné motívy iných transkripčných faktorov. Na základe týchto informácii sa vytrénoval SVM klasifikátor, ktorý potom predpovedal väzobné miesta p53.

3.5 Analýza p53HMM

3.5.1 Analýza motívu

3.5.1.1 Inzercie a delécie nukleotidov

Pri analýze motívu budeme vychádzať z konkrétnych väzobných miest proteínu p53 získaných experimentálne. V podkapitole Motív p53 sme popísali najčastejšie motívy, a ich orientáciu, na ktoré sa viaže p53. Ide o motív RRRCWWGYYY, kde R je purín, Y pyrimidín, W môže byť A, alebo T (adenín, alebo tymín) a nakoniec G je guanín a C je cytozín. Štvrť-úsek RRRCW sme označili ako → a štvrť-úsek WGYYY ako ←.

Clone	5′ Region	1^{st} Half-site R R R C W W G Y Y Y	*Spacer*	2^{nd} Half-site R R R C W W G Y Y Y	3′ Region
s57	CGACCTGTCA caccg	G G G C C T G T C A		C A G C A T GaC C T	acctgtcacaccggg
N22	atttt CACCATGCTT	C T G C A T G T C T		A G G C A A G T C A	ccttctc CACTGGCC
11A2	ccccatcctccatcc	A A A C AaT G C C C		A G A C T T G T C T	ct CCGCCTGAAT ga
W211	tttgtcctaccatcc	A G G C A T G C C T		- - - - T T G C C T	CACTCGTTA tttcct
W7B2	tatct GTGCAGCTG t	G G G C A T G T T T	t	A G G C A A G C T T	cct GTGCTAGTTC cc
3H	AACTAGATC cttttc	A G A C A T G T T A		T A A C A A G T C A	GTACAAGTTT atttt
8A	gctggt GCACAAGAG	T G A C A T G T C C		C G A C G T G T T T	tgtc
532	CATCATGCCA cctgc	A G G C A T G T T C	tggat	G G G C - T G T C T	t GTGCTTTGTTG ttt
64A2	c AAACCAGGGT gtct	T G A C T T G C C T	atcctgggaggt	T G A C A T G T T C	ctccccttcccctc
W7A1	gccaaacataaccac	C A G C - T G C C A		A G G C A T G C A G	tacc ACGCTCAGCCC
s61	c	C A A C T T G T C T	attctgtgttgat	G G A C A T G T T C	ccgtttttggctatt
11B3	actgttgatgatgaa	A G A C A A G C C T	a	G G G C A G G T C C	tgggggtgggg
N42	gcagtgtggtggagg	A A A C A A G C C C	a	G G A T G T G C C C	a GGGCAGGCTG ggac
s201	tgttc ATACCTGTCC	A C A C T T G T C T		A T A C C T G C C T	ACACCTGTCT tgttt
s1583	ctttaattcagttgt	A A A C A T GaC T T	gttcattata	T G A C A T G T T C	aattacaattcgatt
s592I	ctcagttctcagctg	G G A C T T G C C C		T G G C C A G C C C	tgg GGTCACTGCTG c
s592II	tgcctcagcacctcc	A G G T TcT G C C -		G G G C T T G T T C	ctttcctttcagcat
2NB	gcctttgttgtgccc	T G A C T T G C C C		A G A C A T G T T T	gggaa TGTCTTGTGC
9H	gtattctcttttcct	A A G C A T G C C T		T G A C T T G T T C	tttcatctcctctga
CBE10d	tgaaagcaggtagat	T G C C T T G C C T		G G A C T T G C C T	GGCCTTGCCT tttct

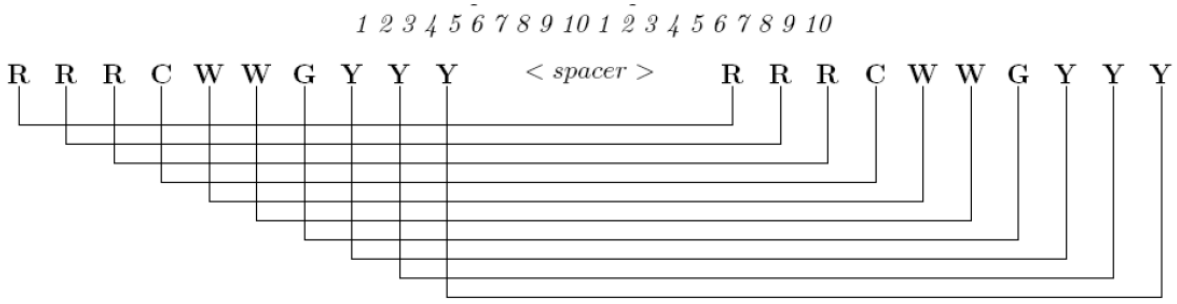
Obrázok 5 Väzobné miesta p53 u 20 vybraných génov s HH (hlava-k-hlave) orientáciou [1]

Ako vidno na obrázku [Obrázok 5], kde žlté sekvencie predstavujú jednotlivé pol-úseky u vybraných 20 génov, na základe ktorých bol stanovený motív väzby p53, u 7 z 20 génov je zaznamenaný vo svojom motíve vloženie (zelená farba), zmazanie (červená), alebo vloženie aj zmazanie (fialová) bázy, resp. nukeotidu. To odpovedá 35%. Keď vezmeme súbor všetkých v sučasnosti zhruba 160 experimentálne potvrdených väzobných miest p53, zistíme, že takmer 30% obsahuje vloženie, alebo zmazanie nukleotidu. Z toho výplýva, že motív väzby p53 je značne degeneratívny, či dosť diskvalifikuje použitie PSSM.

3.5.1.2 Zhoda medzi väzobnými miestami p53

Pri úvahách nad zlepšením schopnosti predikcie väzobného miesta p53 je možné využiť znalosť, že p53 sa často viaže na DNA v homodimérnej, alebo homotetramérnej forme. Z toho vyplýva, že zodpovedajúce si miesta väzieb majú často formát palindrómu, opakovania, alebo

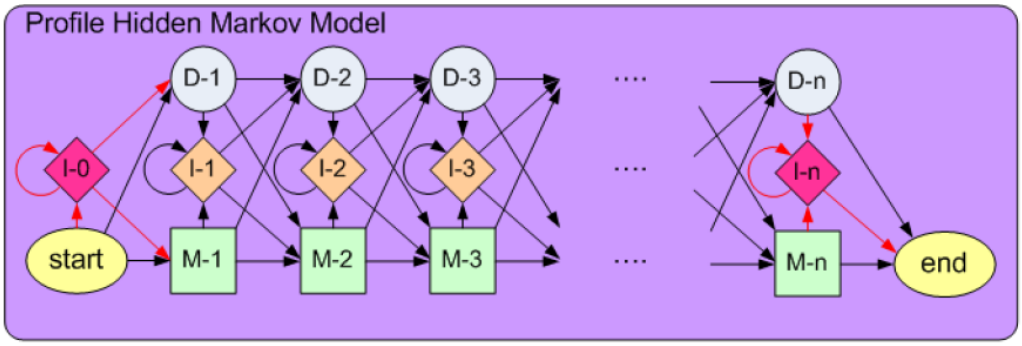
vzájomného negatívneho doplnku. Toto je možné využiť pri tvorbe modelu, kde je možné previazať odpovedajúce si miesta.



Obrázok 6 Opakovací sa motív väzby a vzájomne prepojené miesta [1]

3.5.2 Analýza HMM

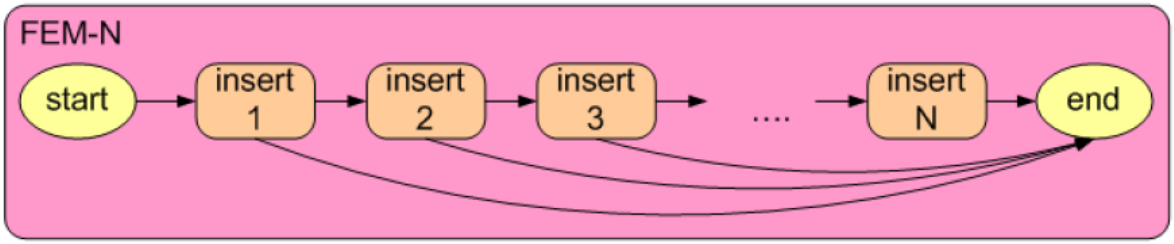
Skrytý Markovský Model (HMM) je popísaný v [1]. Na obrázku [Obrázok 7] je zobrazená štandardná architektúra PHMM. Tento obsahuje tri skryté stavy pre každú pozíciu v sekvencii o dĺžke *n*. Zelené štvorce predstavujú zhodu (match), oranžové objekty predstavujú vloženie bázy (insertion) a biele kruhy predstavujú zmazanie bázy. Stavy sú prepojené šípkami, ktoré budú mať priradené pravdepodobnosti prechodov. Stavy zhoda (match) a vloženie (insertion) majú navyše emisné pravdepodobnosti pre jednotlivé nukleotidy.



Obrázok 7 Architektúra štandardného Profile HMM [1]

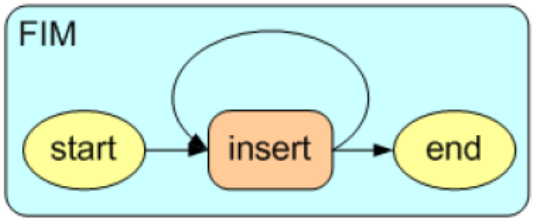
HMM sa trénuje pre jedno miesto väzby a preto je potrebné spojiť niekoľko takýchto modelov do sekvencie, tak aby zodpovedali motívu väzby p53. Pred, za a medzi jednotlivými miestami väzieb môžu byť vložené medzery.

V práci autori predstavujú dva markovské modely, ktoré riešia problematiku medzier. Prvý je Finite Emission Module (FEM) [Obrázok 8]. Tento slúži na modelovanie medzier medzi dvoma väzobnými miestami tak aby bolo možné nastaviť pravdepodobnosť emisie pre jednotlivé pozície medzery rôzne.



Obrázok 8 Finite Emission Module (FEM)
modelovanie medzery z rôznymi pravdepodobnosťami miest [1]

Druhý model je Free Insertion Model (FIM) [Obrázok 9]. Tento modeluje, ako je už z názvu zrejmé ľubovolný počet medzier a v prípade uniformného nastavenia pravdepodobností medzi nukleotidy neovplyvňuje výsledné skóre celého systému.



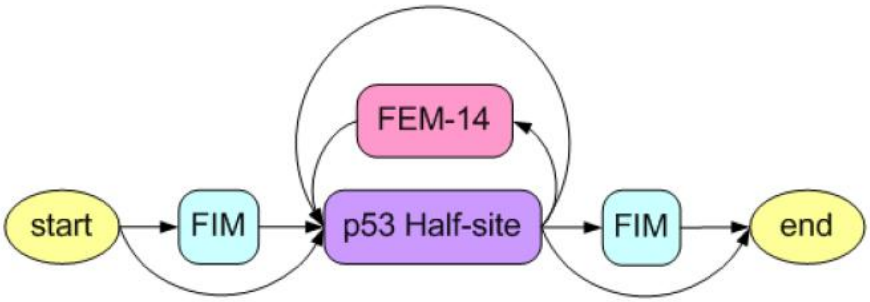
Obrázok 9 Free Insertion Model (FIM)
modelovanie volného počtu a pravdepodobnosti medzier [1]

V prípade jednoúsekového väzobného miesta (single site) sú tieto modely spojené do single site modelu [Obrázok 10].

Obrázok 10 Single-site model [1]

Tento model zodpovedá rôznym formám väzobného motívu, ktorý je rozdelený na dve pol-oblasti, medzi ktorými môže byť medzera o dĺžke až 20 nukleotidov. Pred a za pol-oblasťami sú umiestnené FIM modely, aby bolo možné nájsť čo najlepšie motívy v vstupnej sekvencii.

Ak uvažuje klastrovanú väzobnú oblasť (cluster site), potom sa model zmení následovne [Obrázok 11].



Obrázok 11 Cluster-site model [1]

Pri klastrovanej väzobnej oblasti prichádza do úvahy niekoľko väzobných pol-úsekov a teda je nutné tieto za seba "naukladať" a medzi ne vložiť prípadnú medzeru. Tú predstavuje FEM o dĺžke 14 nukleotidov. Rovnako ako v prípade single-site modelu, je celá sekvencia obalená do FIM modelov.

4 Návrh a implementácia

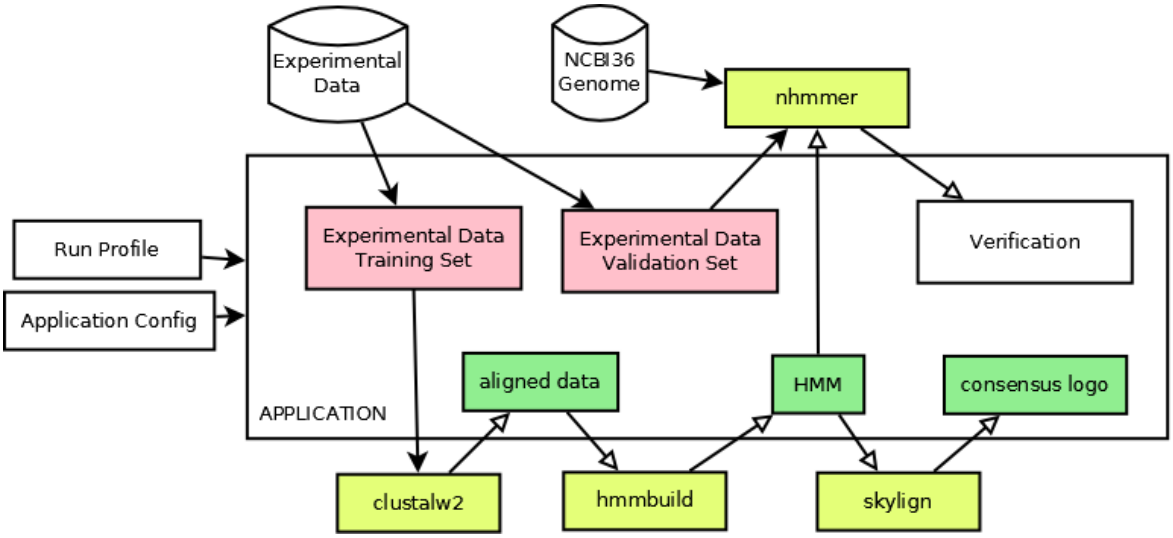
Pri implementácii tejto práce sme identifikovali dve cesty, ktorými je možné splniť jej ciele. Prvý je vytvoriť aplikáciu schopnú modelovať, nastavovať a trénovať skrytý Markov model (HMM). Pri tejto ceste by bolo kľúčovým bodom porozumenie detailom definície, návrhu a učenia Markových modelov. Výsledný produkt by sme potom použili na trénovaciu množinu a boli schopný verifikovať správnosť návrhu a pochopenia modelu. Výhodou by bola značná možnosť kontroly nad výstupným modelom. Nevýhodou je čas nutný k správnej a optimálnej implementácii HMM.

Druhá možnosť je použiť už vytvorené nástroje, ktoré sú dostupné. Nevýhoda prvej možnosti sa tak presunie na bedrá tvorcov nástroja, ktorý budeme používať a my získame čas venovať sa spracovaniu vstupov a výstupov trénovania HMM. Naopak sa vytvára nevýhode v podobe obmedzenej možnosti kontroly tvorby a optimalizácii HMM, čo sa môže nakoniec ukázať ako problematické.

V rámci tejto práce sme sa rozhodli ísť druhou cestou a to použitím už existujúcich nástrojov.

4.1 Návrh Aplikácie

Cieľom aplikácie je možnosť pohodlne pracovať s nástrojmi, ktoré zabezpečujú tvorbu HMM a schopnosť validácie vytvoreného modelu. Na obrázku [Obrázok 12] je zobrazená schéma aplikácie. Biele objekty predstavujú vstupné dáta aplikácie. Žlté sú externé nástroje, ktoré aplikácia používa. Tieto budú popísané neskôr. Červenou farbou sú označené vstupné dáta, ktoré aplikácia spracovala a je schopná ich predávať externým nástrojom. Zelenou sú zas výstupy externých nástrojov, ktoré sú spracované aplikáciou pripravené k predaniu do ďalšieho kroku.



Obrázok 12 Schéma aplikácie

Základom bude možnosť nastavenia aplikácie a to z dvoch pohľadov. Je treba definovať vstupné dáta, jednotlivé nástroje, ich všeobecnú konfiguráciu a pod. Tieto nastavenia sú myslené pod položkou "Application Config". Druhý typ nastavenia sú nastavenia, ktoré ovplyvňujú samotný spôsob behu aplikácie a manipulácie s dátami. Tieto nastavenia nazvime termínom "Profile".

4.2 Platforma

Platforma v ktorej je aplikácia implementovaná je Java8. Výber tejto platformy ponúka výhody v podobe natívnej prenositeľnosti aplikácie medzi operačnými systémami. Jednotlivé použité nástroje majú svoje verzie pre všetky známe operačné systémy, preto by nemal byť problém aplikáciu spustiť aj mimo Windows, i keď nebola takto testované.

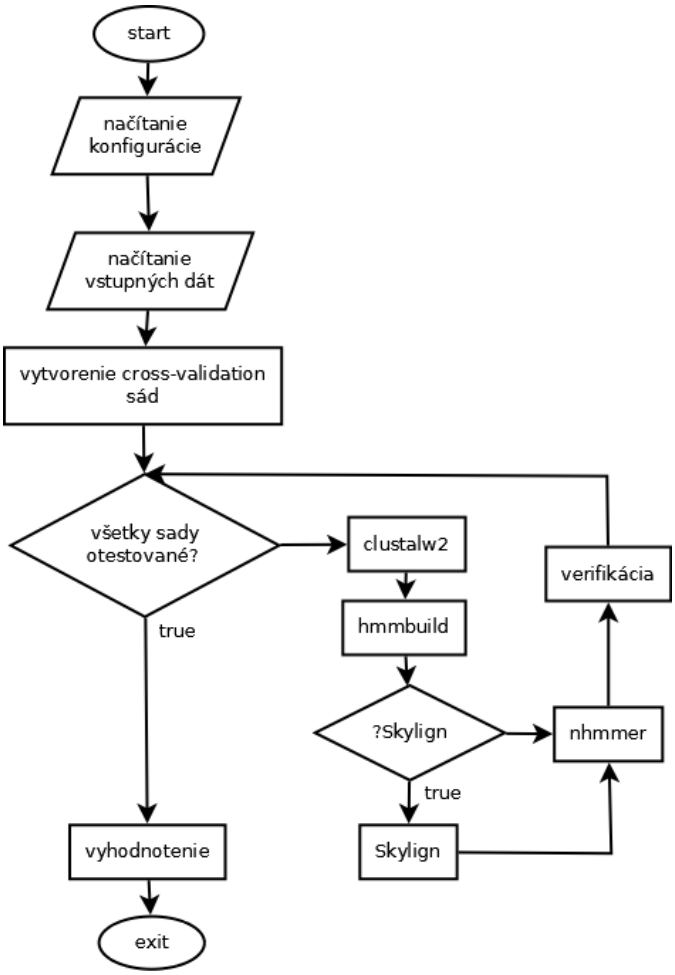
Okrem štandardnej Javy používa aplikácia ďalšie knižnice voľne dostupné v Maven repozitári.

knižnica	použitie
commons-lang	utility knižnica pre prácu s reťazcami (String)
commons-io	utility knižnica pre prácu so vstupmi a výstupmi
jackson-annotation jackson-core jackson-databind	konverzia výstupu z Skylign z JSON formátu do Java POJO
http-client http-core http-mime	komunikácia so REST API serveru Skylign
commons-logging commons-codec	používané "http" knižnicami

Tabuľka 1 Zoznam použitých knižníc

4.3 Beh aplikácie

Beh aplikácie by mal byť zrejmý z vývojového diagramu zobrazenom na obrázku [Obrázok 13].



Obrázok 13 Vývojový diagram behu aplikácie

Algoritmus aplikácie pracuje v cykloch. V každom cykle použije inú cross-validation sadu, ktorej trénovaciu časť dá spracovať externým nástrojom a validačnú časť použije na validovanie výsledkov nálezu *nhmmer*a. Po vyčerpaní všetkých sád vyhodnotí výsledky a ukončí sa. Počas behu predáva dáta na vstupy externým nástrojom, ktoré sám spúšťa a čaká na ich výsledky, ktoré následne spracuje. Každý medzivýsledok sa ukladá na disk do pracovného adresára, a teda je možné si tieto výsledky prezrieť (trénovacie sekvencie, ich zarovnanie, vygenerovaný HMM, consensus logo, výsledok validovania každého prvku z validačnej sady).

4.4 Krížová validácia

Krížová validácia alebo bežnejšie angl. používané cross-validation je metóda zisťovania ako veľmi bude daný model štatistickej analýzy ovplyvňovať nezávislé vzorky dát. Tento postup je dôležitý pre predikciu neznámych vzorkov po predchádzajúcej klasifikácií známych vzorkov.

Princíp fungovania je pomerne jednoduchý. Vstupný množinu dát rozdelíme na podmnožiny, pričom jednu podmnožinu používame na natrénovanie klasifikátora a zostávajúce podmnožiny na overenie schopnosti klasifikácie neznámej vzorky dát, podobne ako je zobrazené na obrázku [Obrázok 14]. Tieto kroky sa niekoľkokrát opakujú, pričom trénovacia množina sa obmieňa.

Obrázok 14 Rozdelenie sady TFBS na trénovaciu a validačnú

Obrázok ukazuje, že naša aplikácia delí vstupnú množinu na dve podmnožiny, trénovaciu a validačnú, resp. testovaciu. Výber prvkov do trénovacej sady je možné uskutočniť viacerými spôsobmi. Pre potreby našej aplikácie používame jednoduché rozdelenie podľa zadaného pomeru v profilovom súbore behu (popísaným nižšie). Tieto množiny sa použijú na natrénovanie HMM a jeho validáciu. V ďalšom cykle sa posunie prvý prvok z trénovacej množiny na koniec testovacej a prvý prvok z testovacej na koniec trénovacej a znovu sa spustí trénovanie a validácia. Ak veľkosť vstupnej množiny je N, tak toto sa opakuje N-krát, čo znamená, že každý prvok vstupnej množiny sa ocitne v trénovacej aj vo validačnej množine.

Pre štatistické zhodnotenie výsledkov predikcie slúžia hodnoty True Positive, False Positive, True Negative a False Negative. Ak máme binárny predikátor, tak True Positive situácia nastane ak na výstupe predikcie očakávame 1 a pozoruje taktiež 1. Ak pozorujeme 0, tak ide o situáciu False Positive. Analogicky je to pri očakávaní 0. Pri pozorovaní 1 máme False Negative a pri pozorovaní 0 dostávame True Negative. Vzájomnými pomermi týchto premenných je možné získať predstavu o vlastnostiach a schopnostiach predikcie.

4.5 Výstupy

Aplikácia počas behu vytvára veľké množstvo výstupov. Jednak ide o výstupy samotných použitých nástrojov a potom o výstupy aplikácie samotnej. Pretože pracuje v cykle, vďaka trénovaniu

a validovaniu cross-validáciou, ak N je veľkosť vstupnej množiny, V je veľkosť validačnej množiny a R je počet výstupov počas jedného behu, tak potom N*V*R je celkový počet výstupov.

Z hľadiska analýzy výsledkov je nutné po ukončení aplikácie zistiť, ktorý HMM sa javil ako najlepší, resp. zhrnúť celkové výsledky. K tomu vytvorí aplikácia súbor *prefix*-results.txt, ktorý má nasledovný CSV formát:

```
run;total;true positive;false negative;ratio;genomeSearch;
1;4;2;2;0.5;1;
```

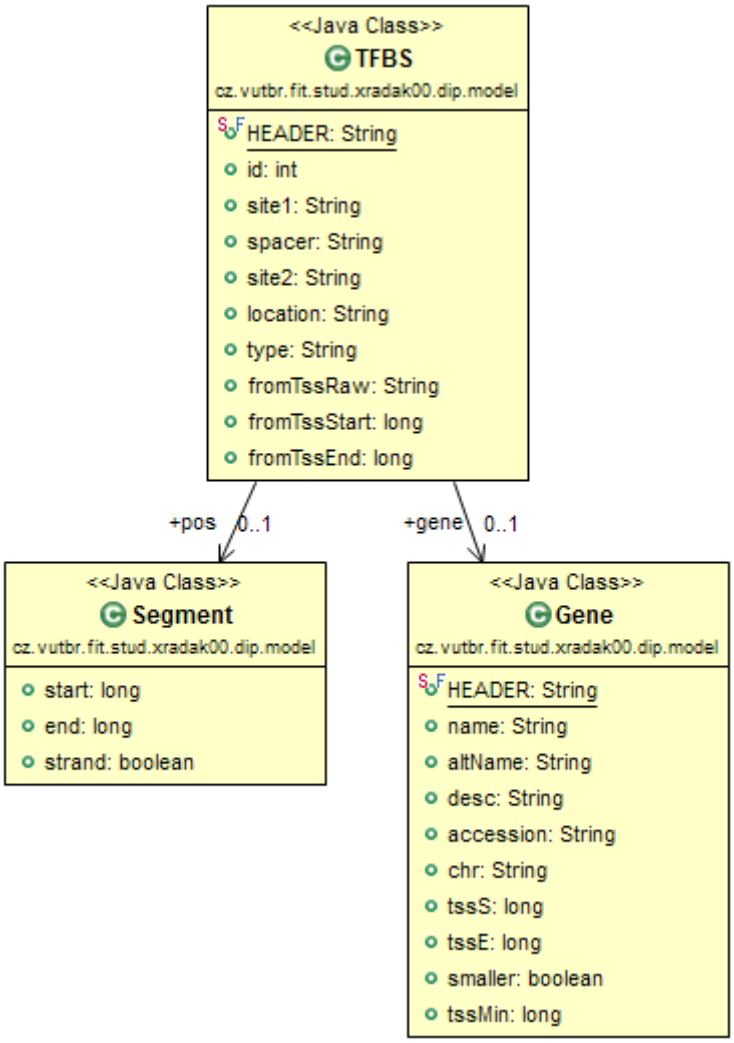
Prvý riadok je hlavička CSV súboru. Hodnota *run* predstavuje ID jedeného cyklu cross-validácie. V rámci nej bolo validované niekoľko TFBS voči nejakému HMM. Tento HMM je jednoznačne identifikovateľný v pracovnom adresári aplikácie podľa ID behu. Celkový počet validovaných TFBS udáva druhý stĺpec. Tie, ktoré sa podarilo nájsť sú sčítané v stĺpci *true positive*, tie ktoré sa nepodarilo nájst v stĺpci *false negative*. Posledný stĺpec vyjadruje, pre lepšiu orientáciu pomer *true positive* k celkovému počtu validovaných TFBS. Týmito výstupmi je možné rýchlo vizuálne zhodnotiť schopnosti daného HMM, prípadne importovať do tabuľkového procesora na zobrazenie v grafe, alebo ďalšiu analýzu. Ak výsledok HMM patril medzi najlepšie (s najvyšším počtom *true positive* zásahov), tak sa pre tento HMM prehľadá celý genóm a výsledok tohto hľadania sa uloží do *genomeSearch* stĺpca. Tým bude možné porovnať tie najlepšie výsledky navzájom. Ostatné výsledky budú mať v tomto stĺpci 0.

5 Dáta a ich spracovanie

Práca je založená, resp. vychádza z článku [1]. V k tomuto článku sú priložené dva dokumenty s dvoma tabuľkami. Oba popisujú popisuje gény a ich väzobné miesta transkripčných faktorov (TFBS) pre p53. V prvom sa nachádza zoznam génov, ich popis, tvar a sekvencia TFBS a to prvý half-site, medzera a druhý half-site. U každého génu je uvedený zdroj, ktorý toto miesto pre p53 popisuje. Druhý dokument spresňuje údaje z prvého, pripája pozíciu voči génu, typ regulácie, funkciu a skóre HMM, ktoré autori získali. Dáta získané z týchto dokumentov sú v aplikácii označované ako "Riley".

5.1 Model dát

Dáta, s ktorými aplikácia pracuje sú modelované niekoĺkými triedami, ktoré sú zobrazené na obrázku Obrázok 15].



Obrázok 15 Data model väzobných miest transkripčných faktorov reprezentovaný class diagramom

Trieda *TFBS* je v hierarchii najvyššie a reprezentuje samotné väzobné miesto. K nemu náleží gén, ktorý toto väzobné miesto reguluje, reprezentovaný triedou *Gene*. Na spresnenie pozície voči tomuto génu je použitá trieda *Segment,* ktorá popisuje kde presne sa *TFBS* nachádza a na ktorom vlákne. Tento model zodpovedá dátam prezentovaným v [1] a získaných ďalším spracovaním.

5.2 Čistenie a predspracovanie dát

V prvej fáze bolo nutné zoznam týchto elementov dostať do spracovateľnej podoby. K tomu bola vytvorená trieda *RileyParserMain*, ktorá načíta a spracuje vstupný dokument, vyfiltruje ho a prekonvertuje na objekty *TFBS*.

V druhej fáze bolo nutné overiť polohu väzobných miest voči samotným génom. Pôvodné dáta boli vytvorené voči zostaveniu genómu človeka verzii NCBI36 z roku 2007. Síce sú existujúce mapovania medzi týmto a najnovším zostavením, avšak z dôvodu zložitejšej manipulácie s týmto mapovaním aplikácia používa priamo dáta zo zostavenia genómu NCBI36 verziu 43. Anotácie génov sú dostupné v GTF súbore, ktorý patrí k danému genómu. Aplikácia obsahuje teda nástroj, ktorý je schopný vytiahnuť všetky dostupné gény z GTF súboru a k nim prislúchajúce polohy miest začiatkov transkripcie. Následne sme tieto dáta spojili s už vopred získaným zoznamom väzobných miest. Tu nastali komplikácie v podobe nie vždy zodpovedajúcemu väzobného miesta voči génu. Preto bolo nutné prejsť k istej heuristike a to nasledovným algoritmom:

1. Ak pozícia TFBS zodpovedá polohe génu tak priraď gén k tomuto TFBS
2. Ak nie, tak vyhľadaj príslušnom chromozóme všetky mieste zodpovedajúce vzoru pre tento TFBS
3. Ak počet nájdených pozícií = 1 tak priraď gén k tomuto TFBS
4. Inak nájdi v zozname nájdených pozícií tú najbližšiu ku génu a túto priraď TFBS

Týmto spôsobom sa každému TFBS priradí gén a pozícia voči tomuto génu, ktorú je možné nasledovne verifikovať pri testovaní správnosti nájdenia TFBS, čo pre účely tejto aplikácie je dostačujúce.

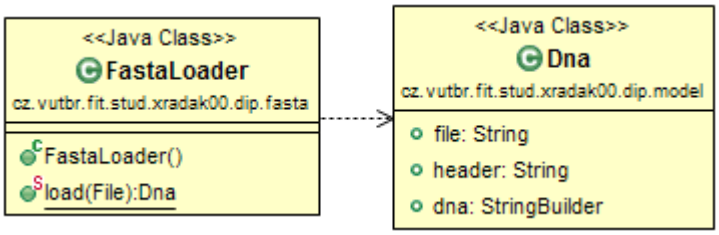
Pri čistení a spracovaní boli identifikované niektoré gény, ktoré nemajú priradené väzobné miesto. U tých ktoré priradené miesto mali sa vždy podarilo priradiť TFBS ku génu. Problémom tejto sady dát je ich rozsah, ide iba o 151 validných TFBS.

5.3 Genóm

Ako bolo spomenuté vyššie aplikácia pracuje so zostavením genómu NCBI36 verzia 43. Je dostupná online na FTP Ensembl projektu a takisto uložená na priloženom DVD. Dáta celého genómu človeka sú rozdelené na jednotlivé chromozómy, tzn., že máme súbor 23 sekvencií. Aplikácia prístup k týmto sekvenciám modeluje v triede *Chromosome*. Tá je schopná zistiť pre každý chromozóm (1-22, X) súbor, v ktorom sa nachádza DNA sekvencia tohto chromozómu.

Okrem jednotlivých chromozómov bol vytvorený aj súbor s celým genómom pri potrebe vyhľadávať na celom genóme. Ide o spojenie súborov všetkých chromozómov použitím známeho nástroja *cat*.

Samotná sekvencia, keď je načítaná v aplikácia, tak je reprezentované triedou *DNA*. Tá je načítavaná triedou *FastaLoader* tak ako je zobrazené na obr. Obrázok 16.



Obrázok 16 Triedy použité na načítanie sekvencie DNA

Pretože sekvencie jednotlivých chromozómov zaberá často stovky MB, tak napriek snahám aplikácie nemanipulovať okrem prvotného načítania so sekvenciou inak ako *read* prístupom, pre potrebu načítania celého chromozómu do pamäti je nutné zvýšiť aplikácii, a teda jvm, pridelenú pamäť.

6 Nástroje

6.1 HMMER

Vytvorenie aplikácie schopnej konštruovať a trénovať skrytý markov model (Hidden Markov Model - HMM) by trvalo strašne dlho s nejasným výsledkom.

HMMER je súbor nástrojov pre hľadanie homológov proteinových, alebo DNA sekvencií v databázach sekvencií a vytváranie zarovnávania sekvencií, založených na pravdepodobnostnom modeli skrytých profilovaných Markových modeloch (angl. profile HMM). Snaha autorov nástroja je porovnávať sa, resp. nahradiť klasické metódy používané k zarovnávaniu, resp. prehľadávaniu databáz sekvencií, snahou byť viac presnejší a schopnejší detegovať homológy, opierajúc sa predovšetkým na HMM. HMMER je poskytovaný pod licenciou GNU GPLv3 a HMMER je chránený ochrannou známkou Howard Huges Medical Institute.

HMMER je oficiálne distribuovaný a podporovaný len pre POSIX operačné systémy. Pretože táto práca je implementovaná v prostredí Windows, bolo nutné stiahnuť a skompilovať zdrojové kódy HMMERu v prostredí Cygwin. Preto je nutné pri spustení vytvorených spustiteľných súborov, ktoré sú nas DVD prílohe, mimo prostredia Cygwin, mať v dispozícii v ceste dynamickú knižnicu Cygwinu *cygwin1.dll* (taktiež priložená).

Pre naše potreby potrebujeme nástroj, ktorý vie vygenerovať HMM zo zarovnaných sekvencií a potom nástroj, ktorý je schopný hľadať v nejakej sekvencii reťazce, ktoré zodpovedajú HMM.

Prvú požiadavku spĺňa nástroj *hmmbuild*. Vstupom je súbor obsahujúci viacnásobné zarovnanie sekvencií a výstupom potom HMM. Príklad použitia a popis parametrov v Tabuľka 2.

```
hmmbuild --dna output.hmm input.aln
```

parameter	popis
`--dna`	explicitné oznámenie aplikácii, že vstupné dáta sú DNA sekvencia
`output.hmm`	názov výstupného súboru, do ktorého sa uloží vygenerovaný HMM
`input.aln`	vstupný súbor so zarovnanými sekvenciami

Tabuľka 2 Parametre príkazového riadka *hmmbuild*

Pre druhú úlohu, ktorou je hľadanie vhodného reťazca v sekvencii DNA pomocou HMM slúži nástroj *nhmmer*. *nhmmer* je možné konfigurovať veľa parametrami, nás však z hľadiska tejto práce zaujíma predovšetkým, aké sú možnosti vstupov, výstupov a nastavení citlivosti pri hľadaní vhodných reťazcov. Najdôležitejšia je možnosť výstupu výsledkov hľadania vo strojovo spracovateľnom formáte, nakoľko naša aplikácia bude s týmto výstupom pracovať a ďalej ho analyzovať. *nhmmer* sa spúšťa nasledovne:

```
nhmmer -T 5.0 --dna --tblout output.tab model.hmm dnaSequence.fa
```

parameter	popis
`-T 5.0`	parameter T nastavuje nhmmer-u minimálne skóre pre analyzovanú sekvenciu, kedy ju má reportovať na výstup
`--dna`	explicitné oznámenie aplikácii, že vstupné dáta sú DNA sekvencia
`--tblout output.tab`	výstup hľadania bude uložený do daného súboru
`model.hmm`	vstupný súbor s HMM
`dnaSequence.fa`	vstupný súbor s analyzovanými sekvenciami vo FASTA formáte

Tabuľka 3 Parametre príkazového riadka *nhmmer*

Výstup *nhmmer* je v tabuľkovom formáte, kde sú položky oddelené medzerami. Aplikácia by mala byť schopná tento súbor načítať a spracovať. Na príklade výstupu vidíme:

```
# target hmm  hmm alifrom  ali to   envfrom  env to   strand E-value score bias
# name   from to
#------- ---- --- -------- -------- -------- -------- ------ ------- ----- ----
8           5  24 22982100 22982081 22982104 22982080      -      12  13.6  3.3
```

Ide o výstup orezaný o ďalej nespracovávané dáta, a teda pre nás nezaujímavé položky.

target name -identifikátor sekvencie, na ktorej bol reťazec nájdený. V tomto prípade vidíme číslo 8 a teda sa jedná o sekvenciu ôsmeho chromozómu.

hmm from/to -pozícia nájdeného reťazca v HMM

ali from/to -pozícia zarovnaného nájdeného reťazca v prehľadávanej sekvencii

env from/to -obálka pozície v rámci prehľadávanej sekvencie (obaluje zarovnanie)

strand -udáva, na ktorom vlákne sa reťazec našiel (sense +, antisense -)

E-value -udáva štatistickú významnosť výsledku hľadania. Počet nájdených reťazcov, ktoré by mali skóre tak vysoké ako nájdený reťazec, v sekvencii, ktorá by bola rovnako veľká ako aktuálne prehľadávaná a zložená z nehomologénnych náhodných sekvencií. Z toho vyplýva, že závisí na veľkosti prehľadávanej sekvencie.

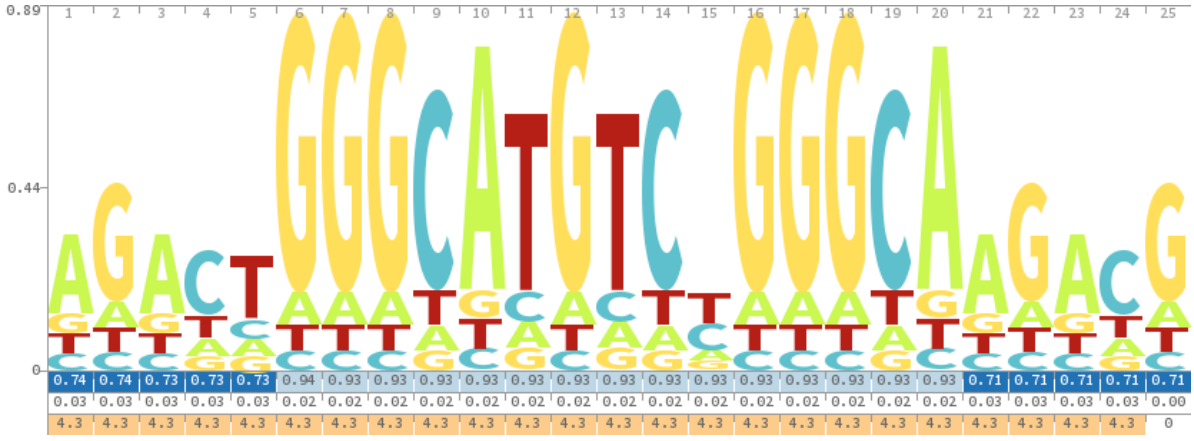
score -skóre nájdeného reťazca. *log-odd* skóre, nezávislé od veľkosti reťazca
bias -korekcia skóre, ktorá bola aplikované. Tvorcovia HMMERu v dobe písania práce nepublikovali ako tento parameter počítajú. Upozorňujú len na zvýšenú pozornosť, ked je *bias* moc vysoké.

6.2 Skylign

Výstup nástroja HMMER je skrytého Markov Model (Hidden Markov Model - HMM). Tento sa dá dobre vizualizovať ako logo. Namiesto generovania vlastného loga sme hľadali nástroj, ktorý je schopný tejto služby efektívne vo vhodnom formáte.

Skylign je nástroj (vo forme offline, alebo online služby) slúži na generovanie lóg, ktoré graficky reprezentujú zarovnanie sekvencií, alebo profile HMM. Vygenerované logo je interaktívne (vo webovom prehliadači) a je možné na neho klikať a zobrazovať podrobné informácie o jednotlivých pozíciách. Zaujímavosť je, že toto interaktívne logo je možné použiť kdekoľvek na inej web stránke, kde ho stačí vložiť ako Javascript. Pre naše účely postačuje vygenerovanie obrázka s logom podľa poskytnutého HMM. K tomuto ponúka Skylign REST API. Veľká výhoda je, že nástroj je schopný generovať logá priamo z výstupu nástroja HMMER. Služba Skylign je voľne dostupná, nie je nutné sa registrovať.

V prvom kroku je nutné pomocou príkazu POST odoslať vygenerovaný HMM do služby, ktorá vygeneruje logo a vráti ID tohto loga. To je potom možné pomocou príkazu GET stiahnuť a uložiť na lokálne. API je zdokumentované online na http://skylign.org/help/api. Príklad vygenerovaného loga môžeme vidieť na obrázku [Obrázok 17].



Obrázok 17 Logo vygenerované nástrojom Skylign

6.3 ClustalW2

Vzhľadom na nutnosť zarovnania použitých rôznych vstupných sekvencií počas testovania, bolo nutné nájsť jednoducho použiteľný nástroj, ponúkajúci výpočet viacnásobného zarovnania, pokiaľ možno offline, schopný pracovať v command line režime a byť prenositeľný na rôzne platformy. ClustalW2 tieto podmienky spĺňa. Ide nástroj na viacnásobné zarovnanie sekvencií DNA alebo proteínov. Je možné ho stiahnuť z webových stránok http://www.clustal.org/clustal2/. K dispozícii je pre všetky populárne operačné systémy vo forme command line nástroja alebo aplikácie s GUI, vydaný pod licenciou GNU Lesser GPL. V práci sme použili poslednú verziu pre Windows v command line forme.

Command line verzia je schopná bežať v interaktívnom aj pasívnom režime. Práca využíva pasívny režim pri ktorom sú ClustalW2 predložené sekvencie DNA vo FASTA formáte a výstupom je vypočítané viacnásobné zarovnanie v ALN formáte. ClustalW2 poskytuje mnoho parametrov, my však využívame len niektoré. Spustenie nástroja vyzerá nasledovne (parametre sú popísané v tabuľke Tabuľka 4):

```
clustalw2 sequences.fa -TYPE=DNA -QUIET -OUTFILE=multialignment.aln
```

Parameter	popis
`sequences.fa`	vstupný súbor s DNA sekvenciami vo FASTA formáte
`-TYPE=DNA`	explicitné oznámenie aplikácii, že vstupné dáta sú DNA sekvencia
`-QUIET`	potlačenie výstupov priebehu počítania zarovnania sekvencií
`-OUTFILE=multialignment.aln`	súbor, do ktorého budú zapísané výsledky viacnásobného zarovnania

Tabuľka 4 Parametre príkazového riadka ClustalW2

Vo vstupnom súbore musí mať každá sekvencia priradený jednoznačný identifikátor, čo u FASTA formátu je text bezprostredne za úvodným znakom ">" do najbližšej medzery. Toto je nutné dodržať, pretože ClustalW2 skončí s chybou. Výstupný súbor obsahuje všetky vstupné sekvencie označené svojim identifikátorom a priradené zarovnanie. Príklad výstupu je nasledovný:

```
s30             AGACTGGGCATGTCTGGGCA-----
s134            -----GGGCATGTCCGGGCAAGACG
                     ********* *****
```

Na výstupe vidíme zarovnané dve sekvencie. Riadok začína identifikátorom sekvencie a nasleduje samotnou sekvenciou priestorovo zarovnanou s ostatnými. Znak "-" označuje vloženie medzery.

Posledný riadok je špeciálny. Sú na ňom označené znakom hviezdička ("*") tie pozície na zarovnaných sekvenciách, na ktorých sa všetky sekvencie zhodujú. Tento súbor je automaticky spracovaný nástrojom NMMER, ktorý mu práca len predá ako parameter.

6.4 Nastavenie aplikácie

Aplikácia sa nastavuje dvoma konfiguračnými súbormi. Prvý je nastavenie samotnej aplikácie, aby mohla fungovať a vedela nájsť používané nástroje a dáta a druhý je nastavenie behu aplikácie pri spracovaní vstupných dát a výsledkov.

6.4.1 Základná konfigurácia

Konfiguácia popisuje kde sa nachádzajú vstupné dáta. Ide o dáta z publikácie [1] a dáta z NCBI36 databáze (cesta k súborom s DNA chromozómov, maska súboru s chromozómami a cesta k GTF súboru). Ďalej je nutné nastaviť cestu k HMMER nástroju, cestu k clustalw2 a cestu (hostname) k skylign. Príklad konfiguračného súboru s podrobným popisom je v prílohe.

6.4.2 Profil behu

Profil samotného behu aplikácie sa zadáva ako prvý a jediný parameter príkazového riadku. Obsahuje nasledovné položky:

konfigurácia	popis
filter	Slúži na možnosť vyfiltrovania vstupných TFBS určených k trénovaniu HMM. Je možné vybrať iba prvý site, druhý site, medzeru, prípadne ich kombináciu. U každého site-u je možné určiť jeho minimálnu a maximálnu dĺžku. Filtrom je navyše možné filtrovať len niektoré TFBS (vymenované zoznamom podľa ich id), alebo ich odfiltrovať (id uvedené znakom "!").
hmmer.score.min	Nastaví nástroju *nhmmer* minimálne skóre, od ktorého má reportovať výsledky. Znížením tejto hodnoty je možné zvýšiť citlivosť *nhmmer*-u.
crossvalidation.ratio	Nastavaní rozdelenie vstupnej množiny TFBS pre trénovanie cross-validáciou. Určuje pomer veľkosti množiny trénovacích dát k testovacím.
run.fileprefix	Všetky súbory vytvorené počas behu aplikácie budú mať daný prefix.

Tabuľka 5 Konfigurácia profilu behu aplikácie

6.4.3 Filtrovanie vstupných dát

V profile behu aplikácie je nutné nastaviť filter pre vstupné väzobné miesta. Tento filter je ponúka rôzne možnosti ako upraviť množinu vstupných dát. Jedná sa o jeden reťazec, ale s pomerne zložitým formátom. Ponúka možnosť nastaviť filtrovanie pre prvý site väzobného miesta, druhý site aj medzery medzi nimi. V prípade, že nie je možné odfiltrovať vhodné TFBS pomocou tohto, je možné zadať konkrétne id TFBS, ktoré sa má odfiltrovať.

Filtrovanie site1: Má formát *s1/<minSiteLen>/<maxSiteLen>*. Na konci je ešte možné zadať regulárny výraz, na ktoré sa má filtrovať motív tejto sekvencie. Potom bude filter vyzerať nasledovne: *s1/<minSiteLen>/<maxSiteLen>/<RE>*. RE predstavuje regulárny výraz.
Filtrovanie medzier, site2: Zodpovedá filtrovaniu site1, s tým rozdielom, že je uvedené reťazcom "spacer" pre medzeru a "s2" pre site2. Ak sa site2, alebo medzery neuvedú, potom nebudú vôbec použité.
Filtrovanie TFBS podľa ID: Toto filtrovanie sa uvádza za filtrami pre site-y. Ide o zoznam id oddelených čiarkou pre TFBS, ktoré chceme aby prešli filtrom. Zaujímavejšie je opačné použitie a teda označenie tých TFBS, ktoré nechceme aby filtrom prešli. V takom prípade sa uvedú znakom "!".

Príklad filtrovania:
`filter=s1/10/10/^GGG.*;spacer/0/0;s2/10/10;!30,!15`
Popis: Má sa filtrovať site1 na minimálnu a maximálnu dĺžku 10 (efektívne sa teda vyfiltrujú tie site1, ktoré majú dĺžku 10). Ďalej podľa regulárneho výrazu, musí site1 začínať prefixom GGG. Medzera sa má filtrovať na prázdny úsek a site2 na dĺžku 10, rovnako ako site1. Podľa zoznamu id, sú filtrovaním odfiltrované (pretože, sú uvedené "!") TFBS s id 30 a 15.

Filtrovaním sú automaticky odstránená tie TFBS, ktoré nemajú definovaný site1.

7 Testy

7.1 Test s malým počtom TFBS

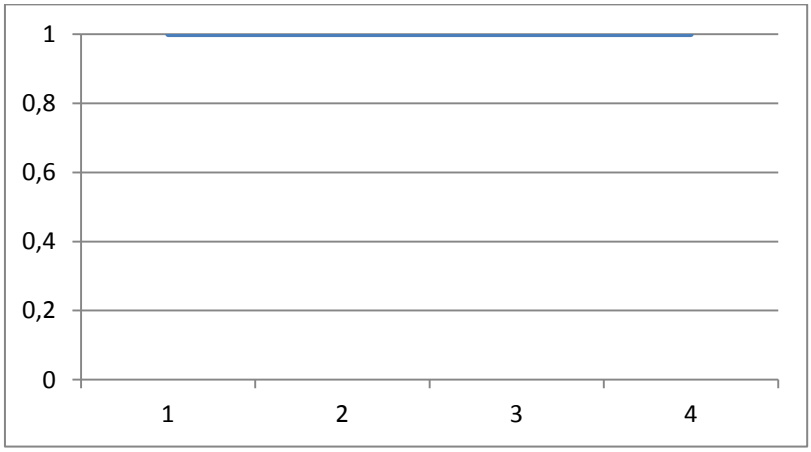
7.1.1 Konfigurácia

Ide o prvý test, ktorý má verifikovať úspešnosť trénovania, verifikovania a aplikovania hľadania väzobného motívu, ak tento motív je pomerne konzervatívne definovaný. Vstupná množina je teda niekoľko TFBS, ktoré sú si navzájom veľmi podobné.

konfigurácia	
konfiguračný súbor	smallProfile.properties
crossvalidation.ratio	.75
hmmer.score.min	10
filter	s1/10/10;spacer/0/0;s2/10/10;133,134,135,136
počet TFBS v test	4
cross-validation rozdelenie (train/test)	3/1
počet behov	4

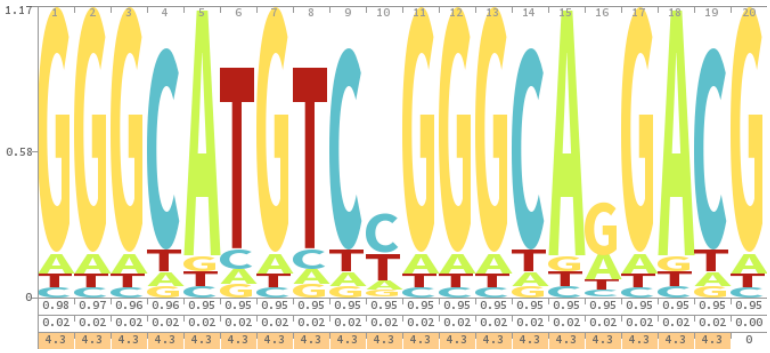
7.1.2 Výsledky

V grafe vidíme rozloženie úspechu validácie jednotlivých HMM vytvorených s každým behom.



Vidíme, že každý vytvorený HMM bol úspešný a validný. Tento test teda môžeme pokladať za úspešný. Najlepšie logo našlo v celom genóme 224 potenciálnych väzobných miest. Výsledky sú zhrnuté v nasledovnej tabuľke.

výsledky	
počet najlepších HMM	4
úspešnosť nájdenia HMM	100%
počet TFBS v genóme pre najlepší HMM	224



Obrázok 18 Logo pre najlepší HMM v teste

7.2 Test pre všetky TFBS

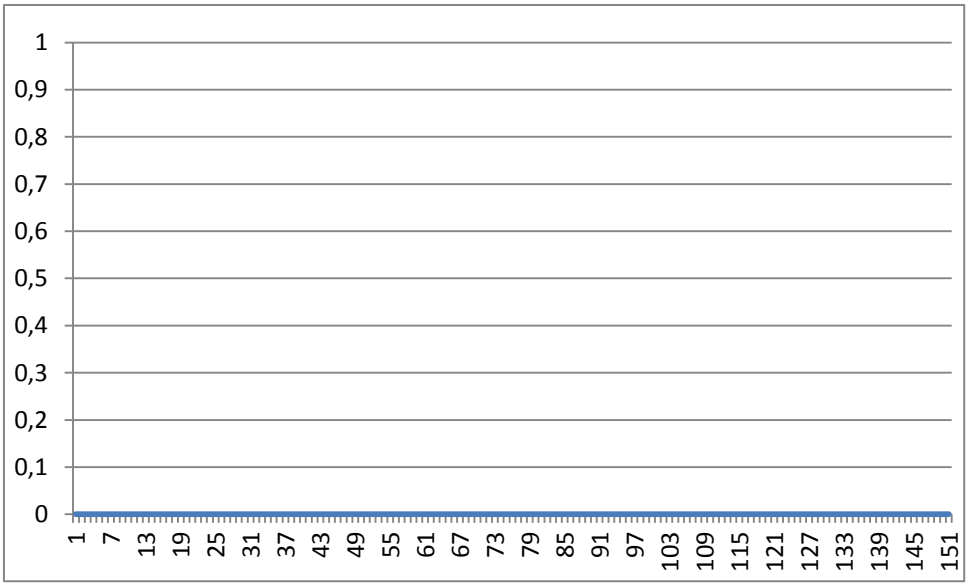
7.2.1 Konfigurácia

Ide o prvý test, ktorý má verifikovať úspešnosť trénovania, verifikovania a aplikovania hľadania väzobného motívu, ak tento motív je pomerne konzervatívne definovaný. Vstupná množina je teda niekoľko TFBS, ktoré sú si navzájom veľmi podobné.

konfigurácia	
konfiguračný súbor	testFull.properties
crossvalidation.ratio	.95
hmmer.score.min	10
filter	s1/0/100;spacer/0/100;s2/0/100;
počet TFBS v test	151
cross-validation rozdelenie (train/test)	3/1
počet behov	151

7.2.2 Výsledky

V grafe vidíme rozloženie úspechu validácie jednotlivých HMM vytvorených s každým behom.



Vidíme, že každý vytvorený HMM bol úspešný a validný. Tento test teda môžeme pokladať za úspešný. Najlepšie logo našlo v celom genóme 224 potenciálnych väzobných miest. Výsledky sú zhrnuté v nasledovnej tabuľke.

výsledky	
počet najlepších HMM	0
úspešnosť nájdenia HMM	0%
počet TFBS v genóme pre najlepší HMM	0

Najlepší HMM sa nevyhodnocoval, pretože žiadny neprešiel validátorom.

7.3 Test s prefixom GGG

7.3.1 Konfigurácia

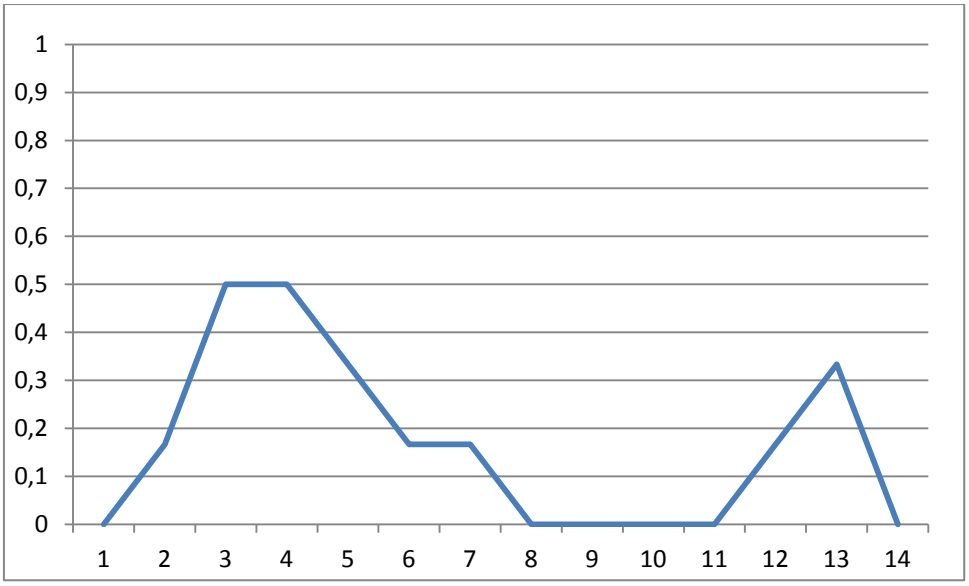
Ide o test kde obmedzíme vstupné TFBS na prefix GGG. Cieľom je dosiahnuť čo najväčší konsensus medzi jednotlivými TFBS a tým zvýšiť šancu na úspešnú validáciu HMM. Rozdiel medzi týmto testom a základným prvým je, že tento test má väčšiu vstupnú množinu.

konfigurácia	
konfiguračný súbor	testGGG.properties
crossvalidation.ratio	.6
hmmer.score.min	5
filter	s1/10/10/^GGG.*;spacer/0/0;s2/10/10;!30,!15

počet TFBS v test	14
cross-validation rozdelenie (train/test)	8/6
počet behov	14

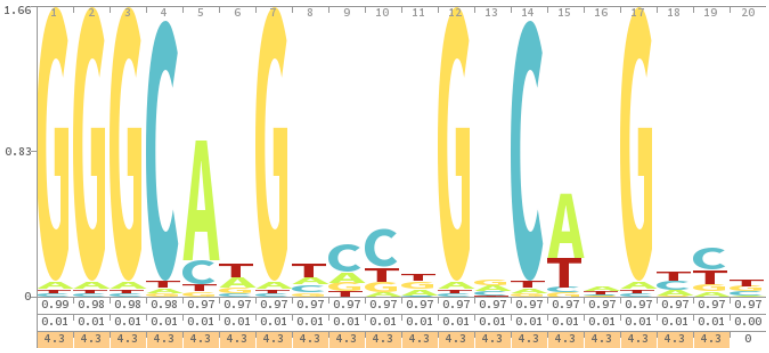
7.3.2 Výsledky

V grafe vidíme rozloženie úspechu validácie jednotlivých HMM vytvorených s každým behom.



Väčšina vytvorených HMM prešlo čiastočne validáciou, avšak žiadny na 100%. Maximálna úspešnosť bola 50%, ktorú dosiahli dva HMM. Z nich ten lepší bol potom schopný nájsť v genóme 801 potenciálnych väzobných miest.

výsledky	
počet najlepších HMM	2
úspešnosť nájdenia HMM	50%
počet TFBS v genóme pre najlepší HMM	801



Obrázok 19 Logo pre najlepší HMM v teste

7.4 Test s citlivosti na parameter cross-validation ratio

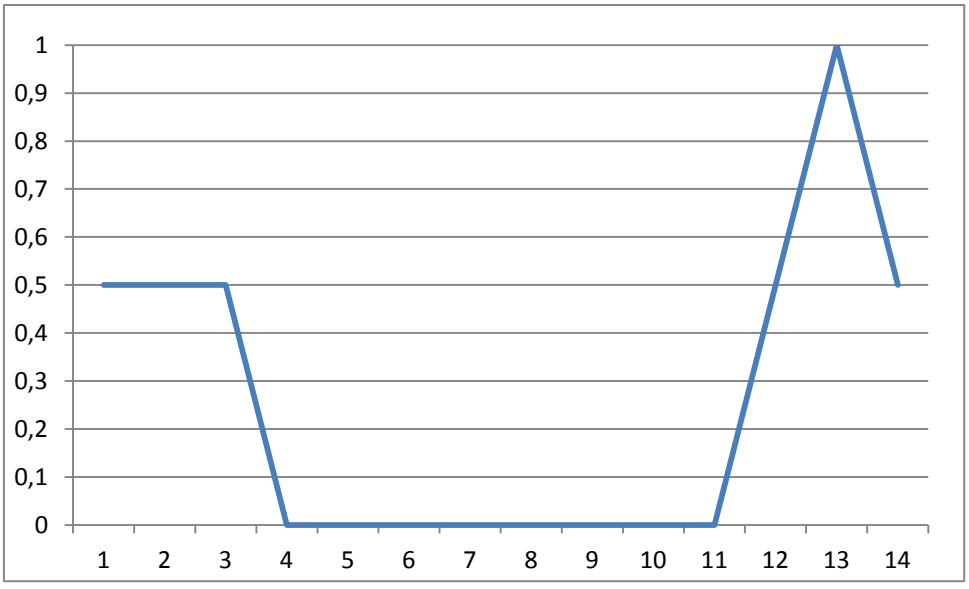
7.4.1 Konfigurácia

Tento test zodpovedá testu s prefixom GGG s jedným rozdielom. Tým je zmenený cross-validation pomer rozdelenia sád. V tomto prípade bude testovacia sada oveľa menšia. Cieľom je zistiť citlivosť aplikácie na tento parameter.

konfigurácia	
konfiguračný súbor	testGGG90.properties
crossvalidation.ratio	.9
hmmer.score.min	5
filter	s1/10/10/^GGG.*;spacer/0/0;s2/10/10;!30,!15
počet TFBS v test	14
cross-validation rozdelenie (train/test)	12/2
počet behov	12

7.4.2 Výsledky

V grafe vidíme rozloženie úspechu validácie jednotlivých HMM vytvorených s každým behom.

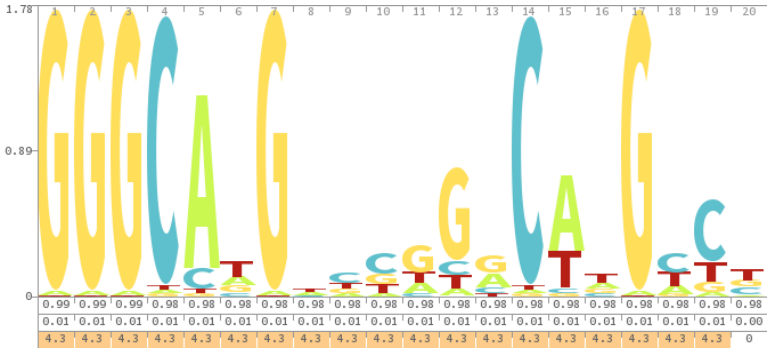


Z grafu vidíme, že výsledok trénovania je iný v porovnaní s GGG testom 1. V tomto prípade bol najúspešnejší HMM s úspešnosťou 100%. Táto úspešnosť je v skutočnosti nižšia v porovnaním s

predchádzajúcim testom, pretože ten validoval 6 miest s max 50%, zatiaľ čo tento len 2 miesta s úspešnosťou 100%. Pri následnom sputení na celý genóm, bol schopný nájsť 1911 potenciálnych väzobných miest. Z výsledkov teda vyplýva, že aplikácia je citlivá na tento parameter. A to pomerne výrazne, pretože napriek podobnosti loga, dostávame odlišné výsledky. Výsledky sú zhrnuté v nasledovnej tabuľke.

výsledky	
počet najlepších HMM	1
úspešnosť nájdenia HMM	100%
počet TFBS v genóme pre najlepší HMM	1911

Vizuálnym porovnaním loga vidíme rozdiely v strede loga, kde je v predošlom teste veľmi výrazné G, ale v tomto sa jeho výraznosť znížila, zato sa však zvýšila výraznosť sprava aj zľava susediacich G.



Obrázok 20 Logo pre najlepší HMM v teste

7.5 Test citlivosti na parameter HMM score min

7.5.1 Konfigurácia

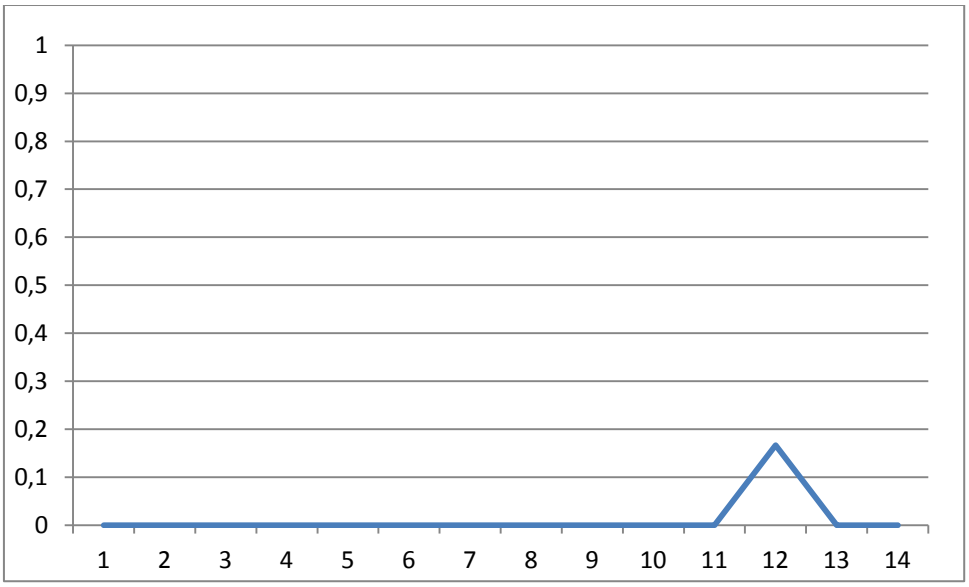
Cieľom tohto testu je zistenie citlivosti na parameter, ktorý hovorí *nhmmer*u, pri akom skóre má reportovať skúmaný reťazec ako zodpovedajúci HMM. Test je rovnaký v porovnaní s GGG testom, líši sa teda len score parametri. Tento sa zvýšil z pôvodnej pomerne benevolentnej hodnoty 5 na hodnotu 10.

konfigurácia	
konfiguračný súbor	testGGGscore10.properties
crossvalidation.ratio	.6
hmmer.score.min	10

filter	s1/10/10/^GGG.*;spacer/0/0;s2/10/10;!30,!15
počet TFBS v test	14
cross-validation rozdelenie (train/test)	8/6
počet behov	14

7.5.2 Výsledky

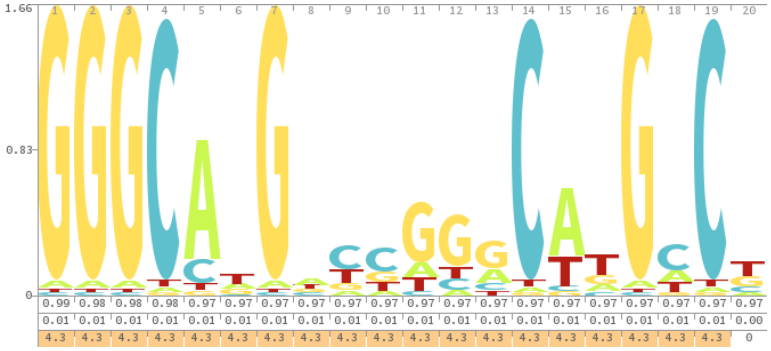
V grafe vidíme rozloženie úspechu validácie jednotlivých HMM vytvorených s každým behom.



Výsledok je úplne odlišný od výsledkov GGG testu. Počet úspešných HMM v cross-validácii klesol na jediný HMM. Je to dané tým, že sa ostatné HMM odfiltrovali. Tento HMM sa vyskytoval aj v GGG teste, avšak nebol vybraný ako úspešný, aj keď mal najvyššie skóre, pretože podľa výsledkov cross-validácie boli iné HMM úspešnejšie, čo je v našej aplikácii rozhodujúce kritérium. Z tohto testu sa teda ukázalo, že HMM s vysokým skóre môžu byť neúspešné v cross-validácii a naopak, HMM úspešné v cross-validácii nemusia mať nutne vysoké skóre. Tieto hodnoty teda zjavne spolu nekorelujú. Výsledok je možné uvažovať do budúcnosti nad zmenou ohodnotenia úspešnosti HMM založeným na kombinácii výsledkov cross-validácie a skóre. Výsledky sú zhrnuté v nasledovnej tabuľke.

výsledky	
počet najlepších HMM	1
úspešnosť nájdenia HMM	16%
počet TFBS v genóme pre najlepší HMM	984

Samotné logo je odlišné od GGG testu, rovnako ako v teste na citlivosť cross-validation rozdelenia líši v znížením výraznosti G zhruba v strede loga a na rozdiel od cross-validation testu pribudlo výrazné C v pravej časti loga.



Obrázok 21 Logo pre najlepší HMM v teste

7.6 Test TFBS bez medzery

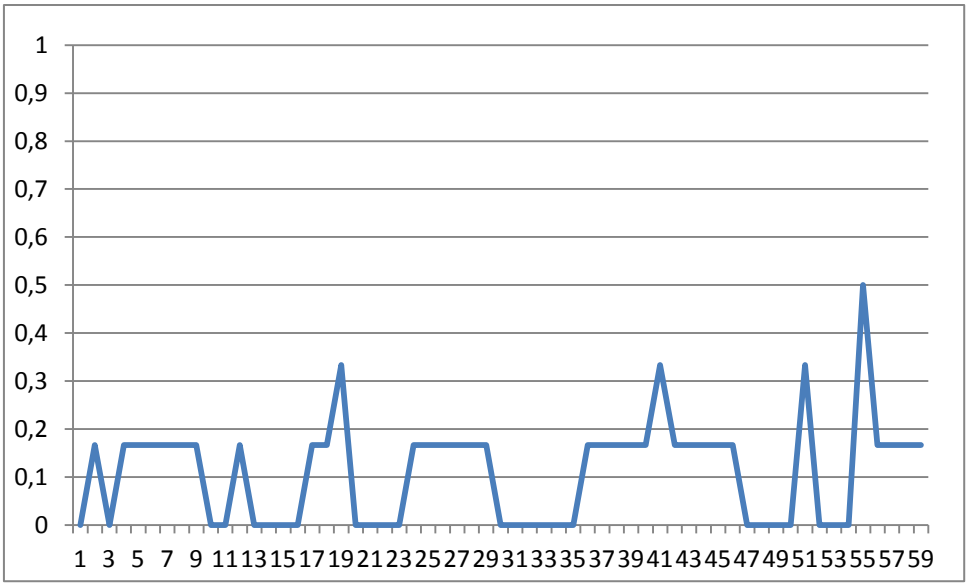
7.6.1 Konfigurácia

Tento test skúsi vytvoriť HMM pre všetky väzobné miesta, ktoré neobsahujú medzery. Filtrom prejde 59 väzobných miest, ktoré spĺňajú túto požiadavku.

konfigurácia	
konfiguračný súbor	testNoSpacer.properties
crossvalidation.ratio	.9
hmmer.score.min	5
filter	filter=s1/10/10;spacer/0/0;s2/10/10;
počet TFBS v test	59
cross-validation rozdelenie (train/test)	53/6
počet behov	59

7.6.2 Výsledky

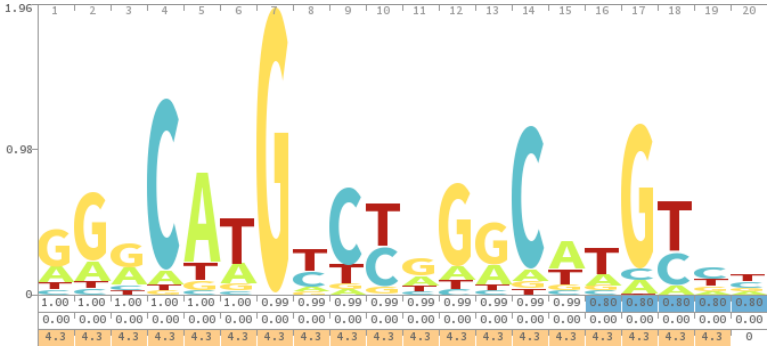
V grafe vidíme rozloženie úspechu validácie jednotlivých HMM vytvorených s každým behom.



Výsledok je oveľa úspešnejší než pri teste so všetkými väzobnými miestami (aj s medzerami). Dôvod by mohol byť ten, že medzery generujú zložitejšie viacnásobné zarovnania, čo sa prejaví na výslednom HMM. V tomto teste najlepšie uspel len jeden HMM s úspešnosťou cross-validácie 50%. Počet nájdených potenciálych miest v genóme je 1452.

výsledky	
počet najlepších HMM	1
úspešnosť nájdenia HMM	50%
počet TFBS v genóme pre najlepší HMM	1452

Samotné logo je odlišné od GGG testu, rovnako ako v teste na citlivosť cross-validation rozdelenia líši v znížením výraznosti G zhruba v strede loga a na rozdiel od cross-validation testu pribudlo výrazné C v pravej časti loga.



Obrázok 22 Logo pre najlepší HMM v teste

8 Záver

Práca sa snaží pokryť vyhľadávanie väzobných miest transkripčných faktorov využitím profilovaných skrytých Markových modelov (HMM). K tejto problematike sa postavila štýlom využitia už vytvorených nástrojov a analýzou ich nastavení a výstupov sa snaží získať optimálny HMM.

Práca je implementovaná ako command-line aplikácia. Pôvodne uvažované GUI nakoniec nebolo nutné vytvárať, jednak sa s command-line aplikáciou pracuje dobre a všetky potrebné nastavenia sú aplikácii dodané v konfiguračnom súbore a súbore profilu a po druhé ušetrili sme pri vývoji čas, ktorý by sme museli venovať vytvoreniu vhodného užívateľského rozhrania. V ďalšom vývoji pri potrebe vytvorenia UI, navrhujeme najprv začať vytvorením UI pre nastavenie konfiguračných súborov a neskôr prípadne kompletné UI.

Nevýhodou výstupnej aplikácie je, že pracuje s obmedzeným súborom vstupných dát. Ide len o 151 validných väzobných miest transkripčných faktorov, ktoré boli získané z [1] a analyzované. V rámci práce sme sa zaoberali aj inými zdrojmi a to [13] a [14], ale nakoniec sme ich nepoužili. To dáva priestor ich analýze a využitiu do budúcnosti.

Ďalšia nevýhoda sa javí použitie nástroja HMMER pre vytvorenie HMM. Pretože neposkytuje možnosť zasiahnuť do tvorby HMM, nie sme schopní implementovať detaily vyplývajúce z analýzy väzobných miest p53. Ide hlavne o vytvorenie nezávislých modelov, ktoré by boli navrhnuté podľa výsledkov analýzy väzobných polo a štvrť úsekov, medzier a ich vzájomné prepájanie do klastrov. Gro tejto práce tým spadlo len do pripravenia dát pre HMMER a sledovanie výsledkov jeho výstupu. O samotných HMM sme sa veľa nedozvedeli. HMMER má taktiež problém s hľadaním krátkych sekvencií (v nedávnych verziách bola dĺžka vstupných úsekov obmedzená dokonca na 30). Toto obchádzame znižovaním limitu skóre, kedy HMMER reportuje výsledky. Tým sme síce dosiahli isté výsledky, ale aj ich kvalita bola nižšia.

Jedno z mnohých vylepšení by mohla byť podpora multiprocessingu. V prípade veľkého súboru validačných dát trvá prehľadávanie jednotlivých chromozómov pomerne dlho. Toto je možno vykonávať pre validované TBFS paralelne. Ďalšia optimalizácia je spojiť hľadanie validovaných TFBS pomocou HMM podľa chromozómov, na ktorých sa vyskytujú.

Literatúra

[1] Riley T., Yu X., Sontag E., Levine A. *The p53HMM algorithm: using profile hidden markov models to detect p53-responsive genes.* BMC Bioinformatics 2009, 10:111

[2] Lim J. *A computational approach to discovering p53 binding sites in the human genome*. PhD Thesis. University of St Andrews, UK. 2012

[3] Szabóová A. *Prediction of DNA-binding propensity of proteins using machine learning*. PhD Thesis. České vysoké učení technické v Praze, Praha. 2013.

[4] Riley T., Sontag E., Chen P., Levine A. *Transcriptional control of human p53-regulated genes*. Nature Rev Mol Cell Biol 9(5). 2008. 402-412.

[5] Wikipedia: Hidden Markov model – Wikipedia, free encyclopedia. [online], [cit.2014-01-13]. URL http://en.wikipedia.org/wiki/Hidden_Markov_model

[6] Wikipedia: Support vector machine – Wikipedia, free encyclopedia. [online], [cit.2014-01-13]. URL http://en.wikipedia.org/wiki/Support_vector_machine

[7] Wikipedia: Transcription (genetics) – Wikipedia, free encyclopedia. [online], [cit.2014-01-13]. URL http://en.wikipedia.org/wiki/Transcription_%28genetics%29

[8] Wikipedia: Protein – Wikipedia, free encyclopedia. [online], [cit.2014-01-13]. URL http://en.wikipedia.org/wiki/Protein

[9] Wikipedia: Position weight matrix – Wikipedia, free encyclopedia. [online], [cit.2014-01-13]. URL http://en.wikipedia.org/wiki/Position-Specific_Scoring_Matrix

[10] Wikipedia: p53 – Wikipedia, free encyclopedia. [online], [cit.2014-01-13]. URL http://en.wikipedia.org/wiki/P53

[11] Kovář J. *Skryté Markovské modely a neuronové sítě.* Diplomová práce. České vysoké učení technické v Praze, Praha. 2008.

[12] Alberts, B., Bray, D., Johnson, A., Lewis, J., Raff, M., Roberts, K., Walter, P. *Základy buněčné biologie: Úvod do molekulární biologie buňky*. Espero Publishing, 2005. 2. vydání. ISBN 80-902906-2-0.

[13] Wei et al. *A GlobalMap of p53 Transcription-Factor Binding Sites in the Human Genome*. Cell, 2013, 124, 10:111

[14] Smeenk et al. *Characterization of genome-wide p53-binding sites upon stress response*. Nucleic Acids Research, 2008, Vol. 36, No. 11 3639–3654

Zoznam príloh

Príloha 1. DVD

Popis adresárovej štruktúry na DVD:

\app			adresár so samotnou aplikáciou
	_workDir		pracovný adresár , defaultne nastavený v config.properties
		___cleanUp.cmd	dávkový súbor, ktorý spustením premaže pracovný adresár
		riley-data.dump.enriched	predspracované vstupné dáta nutné pre beh aplikácie
	\profiles		obsahuje jednotlivé profily behu použité pri testoch
	\src		zdrojové kódy aplikácie
	\jar		obsahuje spustiteľný jar file
	\lib		knižnice použité v aplikácii
	_BUILD.cmd		skript zostaví aplikáciu
	_RUN.cmd		skript otvorí príkazovú riadku s nastavenými cestami a príkladom spustenia aplikácie
\doc			adresár s textom práce
\tools			adresár obsahuje aplikácie nutné k zostaveniu projektu (java, ant) a potom externé nástroje, ktoré projekt používa (clustalw2, hmmbuild, nhmmer)
data.zip			zdrojové data genómu človeka

Inštalácia:

1. Celé DVD je nutné skopírovať do adresára, kde je povolenie k zápisu (aplikácia vytvára počas behu veľa súborov)
2. *data.zip* je nutné rozbaliť v koreňovej zložke, čím sa vytvorí adresár *data*
3. v prípade potreby generovania loga je nutný prístup na internet (http://skylign.org)
4. v adresári *app*\ spustiť skript *_RUN.cmd*, čím sa otvorí príkazový riadok
5. spustiť aplikáciu napr:

```
java -jar jar\DIP15xradak00.jar profiles\smallProfile.properties
```