VYSOKÉ UČENÍ TECHNICKÉ V BRNĚ
BRNO UNIVERSITY OF TECHNOLOGY

FAKULTA INFORMAČNÍCH TECHNOLOGIÍ
ÚSTAV POČÍTAČOVÉ GRAFIKY A MULTIMÉDIÍ

FACULTY OF INFORMATION TECHNOLOGY
DEPARTMENT OF COMPUTER GRAPHICS AND MULTIMEDIA

INTERPRETACE DAT Z BIOČIPŮ

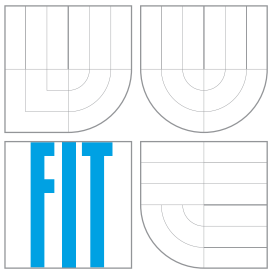
BAKALÁŘSKÁ PRÁCE
BACHELOR'S THESIS

AUTOR PRÁCE
AUTHOR

PETR LUDWIG

BRNO 2008

VYSOKÉ UČENÍ TECHNICKÉ V BRNĚ
BRNO UNIVERSITY OF TECHNOLOGY



FAKULTA INFORMAČNÍCH TECHNOLOGIÍ
ÚSTAV POČÍTAČOVÉ GRAFIKY A MULTIMÉDIÍ

FACULTY OF INFORMATION TECHNOLOGY
DEPARTMENT OF COMPUTER GRAPHICS AND MULTIMEDIA

INTERPRETACE DAT Z BIOČIPŮ

MICROARRAY DATA INTERPRETATION

BAKALÁŘSKÁ PRÁCE
BACHELOR'S THESIS

AUTOR PRÁCE — PETR LUDWIG
AUTHOR

VEDOUCÍ PRÁCE — doc. RNDr. PAVEL SMRŽ, Ph.D.
SUPERVISOR

BRNO 2008

Licenční smlouva je uvedena v archivním výtisku uloženém v knihovně FIT VUT v Brně.

Abstrakt

Tato práce se zabývá interpretací dat získaných pomocí technologie biočipů. Práce obsahuje také krátký úvod do problematiky genetické informace a jejího významu. Jádrem práce je sada skriptů provádějící analýzy nad testovacími daty. Jako vstupní data jsou použity výstupy biočipové analýzy tkání s rakovinou tlustého střeva. Dílčí výsledek je určení genových markerů rakoviny tlustého střeva. Finální výsledek určuje postavení markerů v kontextu objevených signálních drah, ty jsou nakonec seřazeny dle relevance.

Klíčová slova

biočip, genové signální dráhy, genová ontologie, Desmin, bioinformatika, genetika, geny, exprese genů, down regulace, up regulace, KEGG, BioRuby, Ruby, rakovina, karcinom, tumor, tlusté střevo, DNA, RNA, nukleové kyseliny.

Abstract

This Bachelor thesis explains the basics of biochip or microarray data interpretation, starting with short introduction to genetics, especially genetic information significance evaluating. The focus was set mainly on the set of scripts transforming and analyzing the sample data. The data used in this thesis are a result of biochip analysis of the Colon Tumor tissues. The secondary result represents disclosing the main marker for this particular type of cancer, the primary result is evaluation of marker significance in the context of signaling pathways. The resulting pathways are sorted by relevance.

Keywords

biochip, microarray, signaling pathways, gene ontology, Desmin, bioinformatics, genetics, genes expression, downregulation, upregulation, KEGG, BioRuby, Ruby, cancer, tumor, colon, DNA, RNA, nucleus acids

Citace

Petr Ludwig: Interpretace dat z biočipů, bakalářská práce, Brno, FIT VUT v Brně, 2008

Interpretace dat z biočipů

Prohlášení

Prohlašuji, že jsem tuto bakalářskou práci vypracoval samostatně pod vedením pana doc. RNDr. Pavla Smrže, Ph.D.
Uvedl jsem všechny literární prameny a publikace, ze kterých jsem čerpal.

.....
Petr Ludwig
13. května 2008

Poděkování

Rád bych poděkoval svému vedoucímu doc. RNDr. Pavlu Smržovi, Ph.D. za inspirativní konzultace, rady a odbornou pomoc v průběhu tvorby této práce. Osobní poděkování také patří mé rodině za stálou podporu.

Obsah

Kapitola 1

Úvod

Moderní technologie a nová poznání v genetice nabízí stálé další pohledy na aspekty života. Mnohé jevy, dříve považované za náhodné, nalézají v dnešní době jasná vysvětlení a jsou řízeny konkrétními pravidly. Veliký význam na tomto pokroku nese nejen rozvoj v oblasti biologie, chemie a medicíny, ale svou roli v tomto pokroku jistě přinesly i informační technologie, které umožňují se získanými znalostmi a daty v ostatních oborech efektivně, rychle a názorně nakládat. Spojením výše zmíněných oborů vzniklo mladé odvětví bioinformatika. Dá se konstatovat, že by se právě tato práce do tohoto odvětví dala zařadit. Hranice mezi moderními disciplínami však nejsou ostré, proto se často budeme úzce dotýkat i jiných odvětví, nejvíce genetiky.

Pokud se pozastavíme chvíli nad historií, vidíme, že kořeny bioinformatiky a genetiky nejsou zakořeněny na poli dějin hluboko. Pokud vezmeme základní milníky genetiky, tak za počátek můžeme považovat rok 1856. V tomto roce na základě pokusů s hrachem utváří brněnský mnich Johan Gregor Mendel první zákony genetiky. Mendel tím pokládá základní kámen ke klasické genetice. Ke zdokonalení poznatků však tehdejší doba neposkytovala dostatečný aparát, který by pomohl proniknout hlouběji k jádru věci.

Na dlouho dobu zůstávaly poznatky z genetiky na minimální úrovni. V roce 1909 používá poprvé slovo gen dánský botanik Wilhelm Ludwig Johannsen. V této době je gen chápán jako jednotka spojená s určitým znakem dědičnosti. Zásadním mezníkem se stává rok 1953, kdy Američan James Watson a Brit Francis Crick objevují molekulovou strukturu DNA. Oba za svůj objev dostávají Nobelovu cenu. Na základě tohoto objevu se mění směr budoucího výzkumu. Od této doby sice uplynulo více než půl století, ale z pohledu dějin je tento objev stále velice mladý. Další nové poznatky přináší až doba nedávná nebo současná. [14]

V roce 1990 začíná v Americe tým vědců s projektem rozluštěním lidského geonomu. Je nutné zmínit, že jedním z vědců, který tento experiment prosadil, byl právě James Watson. Kompletní výsledek těchto výzkumů byl známý v roce 2003. Tento objev je všeobecně považován za jeden ze zásadních milníků v pokroku lidstva. Všechny tyto objevy jsou velice důležité pro moderní medicínu, vývoj nových léků a celkovou analýzu nejrůznějších onemocnění. Při vývoji se využívá stále dokonalejších bioinformatických technologií. Genové terapie a léčení nádorových onemocnění pomocí znalostí genetiky budou jistě častá témata v 21. století. [9]

Tato bakalářská práce slouží nejen jako ucelené uvedení do problematiky, ale také jako stavební kámen pro další úžeji specifikované práce. Jádrem tohoto dokumentu je zpracování dat získaných z bio-čipů a ukázka některých dostupných nástrojů. Cílem práce je vlastní řešení, které jde nad rámec těchto nástrojů a ukazuje možnosti analýzy testovacích dat.

Kapitola 2

Základní pojmy genetiky

Jelikož se v průběhu této publikace budeme setkávat s mnoha specializovanými pojmy, je nutné všechny náležitě vysvětlit. V rozsahu bakalářské práce však není možné jít do naprostých detailů, ač by důkladná znalost všech problémů pomohla lépe porozumět všem myšlenkám. Pojmy obecně známé nebudou vysvětlovány vůbec. Anglické názvy budou nechány ve své anglické podobě, případně budou doplněny českým překladem. Použité názvosloví vychází z odborných textů, které se touto problematikou zabývají.

Na počátku se pozastavíme nad pojmem *genetická informace*. Dále rozvedeme, co v sobě tato informace nese a uvedeme několik základních metod, jak tyto genetické informace získávat. Jelikož jsou tyto informace pouze velkým množstvím dat, je nutné tato data dále interpretovat, zamyslet se nad jejich významem a se získanými znalostmi efektivně nakládat.

Celá genetická informace v sobě nenese pouze lineární data s vlastním významem, ale důležitou roli hrají vyšší struktury, které se z primární informace dají odvodit. Nejnižší informace se nachází na atomární úrovni. Její význam však ovlivňuje díky *mezibuněčné signalizaci* (vzájemnému ovlivňování buněk) nejen syntézu jednotlivých bílkovin (proteinů), ale ve výsledku i podobu celého jedince. Vzájemné interakce různých buněčných struktur se snaží popsat *genová ontologie* (angl.: Gen Ontology). Celé struktury se poté dají sestavit od genových *signálních drah* (angl.: Signaling Pathways).

Při psaní této kapitoly jsem vycházel především z informací uvedených v [9], [15], [1], [12], [10] a [13].

2.1 Genetická informace

Genetická informace je uložena v genetickém kódu, který v sobě nese každý organismus. Dokonce většina živých organismů, až na drobné výjimky, má podobu tohoto kódu totožnou – zde hovoříme o *standardním genetickém kódu*.

Genetická informace je uložena v nukleových kyselinách, u většiny živočichů v molekule *DNA*. U některých virů je tato životně důležitá informace zakódována v *RNA*. Soubor všech informací jednoho konkrétního organismu nese název *genom*. Tento genom v sobě nese jak všechny geny jedince, tak i nekódující sekvence. Pojem nekódující sekvence bude objasněn dále v kapitole 2.1.2.

2.1.1 DNA a RNA – nositelé informace

Deoxyribonukleová kyselina (zkratkou DNA) a *ribonukleová kyselina* (RNA) jsou nositelé genetické informace pro všechny buněčné organismy. Jelikož se budeme zabývat převážně lid-

skou genetickou informací, budeme dále hovořit převážně o DNA, ve které je právě u člověka kódována podstata celého lidského života, od jeho zrodu v podobě oplozeného vajíčka, až po jeho postupný vývoj k dospělému jedinci.

Často bývá DNA označována jako stavební plán organismu, který předurčuje jeho vzhled, vlastnosti, sklon k nemocem a mnohé další kvalitativní i kvantitativní znaky. V buněčných DNA jsou zakódovány veškeré informace, které určují druh a vlastnosti buňky, řídí její růst a dělení i *syntézu enzymů* a ostatních proteinů nezbytných pro funkci buňky. Všechny tyto vlastnosti a obsah, které v sobě DNA nese, jsou při pohlavním rozmnožování předávány na další generace.

2.1.2 Chemická struktura nukleových kyselin

Jak jsou tyto genetické informace v DNA zakódovány, poodhalí chemická struktura tohoto biopolymeru. Celý biopolymer je tvořen stavebními jednotkami, nukleotidy, které vytvářejí dva dlouhé řetězce, spojené do takzvané dvoušroubovice. Oba řetězce jsou v DNA orientovány *antiparalerně* (proti sobě). Každý řetězec je tvořen mnoha *nukleotidy*, které jsou navázané na fosfátové skupiny a každý nukleotid se skládá z aldopentosy vázané na heterocyklickou purinovou nebo pirimidinovou bázi.

Cukernou složkou v DNA je 2–deoxyribosa (oproti tomu v RNA je cukernou složkou ribosa). Předpona 2–deoxy- udává, že ribosa nemá v poloze 2– hydroxylovou skupinu.

V deoxyribonukleotidech, zastoupených právě v DNA, se vyskytují čtyři *heterocyklické báze*. Dvě z nich jsou substituční deriváty *purinu*, hovoříme o purinových bázích. Těmito purinovými bázemi jsou *adenin* (A) a *guanin* (G). Další dvě báze jsou deriváty *pyrimidinu* a jsou jimi *cytosin* (C) a *thymin* (T). Rozdíl RNA od DNA je u pyrimidinových bázích, kde se v případě RNA nachází místo thyminu *uracil* (U).

Molekuly DNA jsou uloženy hlavně v buněčném jádru a jsou obrovské. Jejich molekulová hmotnost dosahuje až 150 miliard hmotnostních jednotek a délka molekuly by při natažení dosahovala přibližně 12 cm.

Vzorky DNA izolované z různých tkání téhož biologického druhu mají heterocyklické báze zastoupeny vždy ve stejných poměrech. U různých biologických druhů se tento poměr může lišit. V lidské DNA je zastoupení A – 30 %, T – 30 %, G – 20 %, C – 20 %.

Dva navzájem stočené řetězce DNA nejsou identické, ale komplementární. Spojení dvou komplementárních vláken do dvoušroubovice umožňuje slabá mezi-molekulová interakce ve formě vodíkových můstků mezi dvěma komplementárními bázemi. Mezi A a T jsou dvě vazby vodíkovými můstky, mezi G a C jsou vodíkové můstky tři.

Molekula DNA je tedy chemickým nositelem genetické informace organismů. Informace je v molekule zakódována pořadím nukleotidů, které tvoří řetězec DNA. Jádro lidské buňky obsahuje 23 párů chromozómů, každý chromozóm se skládá z jedné molekuly DNA, která je složena z několika tisíc úseků. Tyto úseky se nazývají *geny*. Souhrn všech genů v jedné buňce (*genom*) poté obsahuje na 3 miliardy párů bází. Geny u lidské DNA kódují pouze malá část bázových dvojic, zbylé části se nazývají *nekódující sekvence*. Tyto sekvence nemají zatím žádný známý význam, ale je možné, že právě tyto části poodhalí další nové skutečnosti. Jedna z hypotéz uvádí, že tyto části jsou pouze jakýsi evoluční artefakt.

2.1.3 Geny a jejich význam

Genetické informace nejsou v buňce dostupné okamžitě a stále. Použití instrukce pro tvorbu určitého proteinu vyžaduje *nabuzení* určitého genu. Tento úsek DNA je nabuzen specifickým požadavkem přineseným ze vzdálených buněk, například pomocí určitého hormonu. Signál

může případně přijít i z buňky sousední nebo požadavek může vzniknout uvnitř buňky vlastní. Tato cesta signálu je často velice složitá a vede přes mnohé mezičlánky.

Když příslušný signál dorazí až k jádru buňky, je aktivována příslušná část DNA obsahující požadovaný gen. Tím je gen spuštěn (zvýšena jeho *exprese*) a může být zahájena jeho *transkripce* (přepis) do mRNA, která se účastní syntézy výsledného proteinu. Genetická instrukce pro tvorbu proteinů obsažená v 30 000 – 35 000 lidských genech může vytvořit až 500 000 různých proteinů. Pravidlo jeden gen na jeden protein platí pouze asi jen u 2 procent genů.

Kapitola 3

Získávání informací, zdroje a nástroje

Na počátku této kapitoly jsou uvedeny dva základní postupy získávání genetických dat. V druhé části (3.2) se zamyslíme nad významem těchto dat, protože správné pochopení jejich významu je pro další zpracování to nejpodstatnější. V další části (3.3) zkráceně analyzujeme dostupné nástroje, které genetická data zpracovávají. Důležitou podkapitolou je zde pojednání o *BioRuby*, knihovně skriptovacího jazyku Ruby, ve kterém je implementována praktická část této práce.

3.1 Nástroje na získání základní genetické informace analýzou nukleových kyselin

Jelikož jsou požadovaná data uložena v nukleových kyselinách, jejich získávání se primárně zaměřuje právě na analýzu těchto kyselin. Protože jsou tyto molekuly velmi malé, je analýza poměrně časově a finančně náročná. Mezi dvě základní metody patří *sekvenování* DNA a analýza pomocí *biočipů*. Obě tyto metody mají velmi odlišný přístup i význam výsledku.

3.1.1 Sekvenování DNA

Žádný z velkých pokroků na poli genetiky by nebyl možný, kdyby se v roce 1977 nepodařila objevit metoda určování pořadí nukleotidů – *sekvenování DNA*.

Prvním krokem při sekvenování DNA je rozštěpení této makromolekuly v přesně definovaných místech, aby bylo možné získat menší, lépe zpracovatelné části. Molekula je štěpena *restrikčními enzymy* (restrikční endonulkeázy). Tyto enzymy umožňují přesné a předem definované štěpení v místě výskytu určitých sekvencí bází. Jestliže se původní molekula štěpí dalším restrikčním enzymem, získávají se nové fragmenty, jejichž sekvence se částečně překrývají s fragmenty získanými štěpením prvním enzymem. Toto umožňuje určit pořadí nukleotidů v celé DNA.

Tato metoda zpracování je však velice nákladná a laboratorně složitá. Výsledkem sekvenování DNA je kompletní analýza včetně nekódujících sekvencí. Výstup analýzy však nezahrnuje aktuální stav aktivních genů v buňce, toto odhaluje metoda biočipů (viz. kapitola 3.1.2).

3.1.2 Biočipy - DNA microarrays

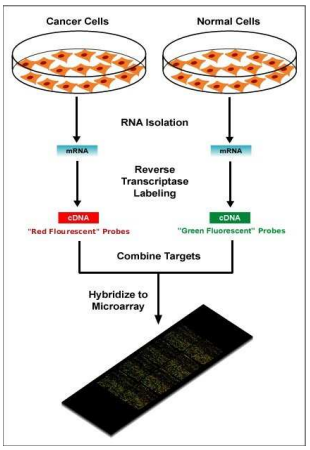
Důležitou technikou, která je levnější a efektivnější, než klasické sekvenování DNA, je dnes technologie DNA biočipů (angl.: DNA microarrays). Tato metoda přechází z výzkumu do laboratorní praxe s přímým klinickým využitím. DNA biočip je složen z mikroskopických částí DNA, běžně reprezentujících právě jeden určitý gen. Tato část DNA se nazývá *prob*. Čip, tedy obsahuje mnoho zafixovaných probů, reprezentujících geny, které chceme analyzovat. Tímto postupem je umožněno sledovat naráz *exprese* (aktivitu) několik tisíců genů současně.

Analýza exprese genu se užívá primárně právě proto, že ne každý gen, který je součástí genomu určitého organizmu, je vždy činný. Naopak v každé z buněk téhož organizmu jsou činné jen některé geny, které jsou právě v oné buňce a v daný vývojový moment potřebné. Jaký konkrétní gen je právě činný, podléhá složité regulaci a je ovlivněno buněčnými signály. Analýza pomocí biočipu probíhá vložením čipu do roztoku vzorku DNA, který chceme analyzovat. Vzorek obyčejně obsahuje extrahované mRNA, přepsané díky transkripci z analyzované DNA, kódující námi hledané geny. Tímto postupem tedy neanalyzujeme přímo DNA, ale aktuálně přepsané RNA, které se v buňce aktuálně nachází.

Takto zhotovený vzorek je v průběhu analýzy obarven fluorescenčním roztokem, který v závěru poslouží ke zvýraznění výsledku experimentu. Jelikož jsou na čipu umístěny fragmenty DNA (takzvané cDNA), které jsou komplementární k částem mRNA obsaženým v roztoku, začnou se komplementární části spojovat – *hybridizovat*.

Místa, na kterých mRNA hybridizuje, jsou díky zvýraznění fluorescenční vrstvou opticky výraznější, než místa, kde se analyzovaná mRNA s probem nespojuje. Proto nám tento experiment umožňuje analyzovat více výroků naráz (obecně zdravou a nemocnou tkáň), můžeme velice snadno a rychle tyto vzorky porovnat a určit, které geny jsou v příslušných buňkách aktivní. Pomocí této analýzy lze identifikovat geny, jejichž exprese je při určité nemoci pozměněna.

Pro zkoumání dynamiky imunitní odpovědi jsou k dispozici biočipy, které umožňují sledovat expresi genů kódujících např. cytokiny, pro a protizánětlivé faktory, nitrobuněčné signální molekuly apod. Schéma analýzy je uvedeno na obrázku 3.1.



Obrázek 3.1: Analýza pomocí biočipů

Metoda analýzy pomocí biočipů našla své uplatnění při zkoumání nádorových onemocnění. Tato metodika, zakládající se na detekci velkého množství genů a jejich mutací

objevených v různých jednotlivých nádorech pomocí genových sond, umožňuje rozpoznat přítomnost alterovaných (poškozených) genů ve vzorku vyšetřovaného nádoru. Pomocí vybraných cDNA těchto genů lze připravit *biomarkery*, jimiž lze rychle a spolehlivě rozlišit podskupiny pacientů s různými histopatologickými a klinickými projevy. Výrazně se tím zlepší diferenciální diagnostika, určení prognózy, nastavení a monitorování optimální terapie. [16] [11]

Data zpracovávaná v této práci jsou získaná právě metodou biočipů. Více o testovacích datech, která byla použita v analýzách této práce naleznete v kapitole 4.

3.2 Význam dat

Data zpracovaná pomocí metody biočipů udávájí aktuální expresi genů. Význam těchto genů se dá určit pouze na základě konkrétních aplikovaných výzkumů. Tyto výzkumy přinesou určitý pohled na to, co určitá aktivita genů znamená. Formalizací těchto významů se zabývá *genová ontologie* (angl.: Gen Ontology). Nad rámec genové ontologie jdou *signální dráhy* (angl.: Signaling Pathways), které specifikují postupné aktivace genů k finálnímu projevu (například rakovině).

3.2.1 Genové ontologie

Při interpretaci genových pojmů je nutné využívat jednotné popisné nástroje, které umožní mít jednoznačnou vypovídající hodnotu. O toto se pokouší *Gene Ontology Project* (zkráceně GO), který poskytuje slovník vhodně volených termínů pro popis jejich chování (atributů) genů a procesů týkajících se genů. Jedná se o souborný slovník druhově nezávislých biologických termínů. GO popisuje struktury z pohledu buněčných komponent, biologických procesů a molekulárních funkcí.

Molekulární funkce popisují různé aktivity genových produktů na molekulární úrovni. U molekulárních funkcí však není popsáno, kdy a kde se tyto děje odehrávají. Biologické procesy popisují řadu prováděných akcí složených z více molekulárních funkcí. Buněčné komponenty jsou části buňky, organely nebo skupiny genových produktů.

3.2.2 Signální dráhy

Buněčné signální dráhy, též *signální kaskády*, jsou sekvence událostí, které umožňují buňce přijmout signál a biologicky na něj reagovat. Jedná se o biochemickou interakci popsanou nejčastěji grafem. Zde je názorná ukázka popisující možnou funkci signální dráhy. Receptorový protein na povrchu buňky převádí obdržený mimobuněčný (extracelulární) signál na signál vnitrobuněčný (intercelulární), přičemž zahajuje signální kaskádu, ve které se štafetou různých molekul dostane signál postupně až do nitra buňky. Tam se kaskáda rozdělí do několika směrů. Řada kroků kaskády může být ovlivněna (modulována) jinými pochody v buňce. Ukázkou možných efektů může být regulace metabolické dráhy, regulace exprese určitého genu nebo změny v *cytoskeletu* (buněčné kostře). [1]

3.3 Nástroje na zpracování dat z biočipů

Existuje velká řada dostupných nástrojů na zpracovaní dat z biočipů. Vybrané nástroje umožňují analýzu dat pomocí genové ontologie a signálních drah. Na konci této části jsou popsány základní vlastnosti jazyka Ruby a jeho knihovna BioRuby.

3.3.1 KEGG

KEGG - Kyoto Encyclopedia of Genes and Genomes. On-line encyklopedie a skupina nástrojů genové analýzy nabízející široké spektrum dat. Jeden z nejucelenějších pramenů informací dostupných přes Internet.

Hlavním plánem a ambicí této encyklopedie je vytvořit v *post-genomického éře* (období po rozlušetí lidského genomu) kompletní počítačovou reprezentaci buňky, organizmu a celé biosféry (viz. [7]). Tyto počítačové modely by měli umožnit predikci buněčných procesů a chování organizmů na základě genetické a molekulární informace. V rámci naplňování tohoto cíle vznikly nástroje, které KEGG nabízí. Toto vše vzniká jako část projektů univerzit Kyoto a Tokyo.

Veškeré záznamy v KEGG jsou postaveny na jednotné struktuře a umožňují počítačové zpracování. KEGG nepoužívá GO, ale vlastní genové ontologie, označené jako *KO* (KEGG orthology), ontologie z GO je však možné do jisté míry na KO převádět. [7] [5] [6]

- KEGG obsahuje informace o genech a proteinech (KEGG GENES);
- chemických informací o endogenních a exogenních substancích (KEGG LIGAND);
- molekulárních schémat interakcí a reakčních sítí (KEGG PATHWAY);
- hierarchií a vztahy mezi různými biologickými objekty (KEGG BRITE).

Sama encyklopedie nabízí rozhraní *KEGG API*, které umožňuje vzdáleně s daty nakládat. Uživatelé mohou k KEGG API serveru přistupovat pomocí SOAP (Simple Object Access Protocol) technologie přes HTTP protokol. Tento SOAP server používá rozhraní WSDL (Web Services Description Language), který umožňuje snadno vytvořit na straně klienta knihovny pro některé programovací jazyky (např.: Java, Pyhon, Perl, Ruby). Toto vše dává vývojářům silný nástroj, jak přistupovat k rozsáhlé databázi KEGG ze svých vlastních programů.

KEGG obsahuje v době vzniku této práce 361 zpracovaných referenčních signálních drah. Jedná se kolekci ručně sestavených map reprezentujících současnou znalost molekulární buněčné interakce.

KEGG PATHWAYS obsahuje 6 druhů drah:

- Metabolické dráhy (Metabolism);
- zpracování genetické informace (Genetic Information Processing);
- zpracování signálů z okolí buňky (Environmental Information Processing);
- buněčné procesy (Cellular Processes);
- lidské nemoci (Human Diseases);
- vývoj léčiv (Drug Development).

V této práci jsou právě KEGG API a KEGG PATHWAYS použity k analýze genetických dat ve spojení s knihovnou BioRuby skriptovacího jazyka Ruby. Referenční manuál k KEGG API naleznete v [8].

3.3.2 APP - Advanced Pathway Painter

Tento program je určen pro vizualizaci signálních drah (KEGG, GenMAPP, BioCarta a další). Program zpracovává data z genových a proteinových experimentů. Tato data jsou přímo barevně zobrazena v signálních drahách. Přes propojení různých webových služeb program umožňuje rychlý přístup k dalším informacím ohledně genů či proteinů.

Program je zajímavý pro vizualizaci dat, neprovádí však žádnou hlubší analýzu těchto dat. Jeho použití je vhodné pro optické zdůraznění aktivních genů v signální dráze.

3.3.3 FatiGO

FatiGO je online nástroj dostupný přes webové rozhraní. V době vzniku práce byla dostupná verze 2.0, která umožňovala porovnávání sady genů v kontextu Genové Ontologie. FatiGO+ je druhý z nástrojů vytvořených k analýze skupiny genů. Tento nástroj umožňuje vyhledávání zadaných genů v signálních drahách KEGG a BioCarta.

3.3.4 BioRuby

Pro tuto práci byl vybrán skriptovací jazyk Ruby. Hlavními výhodami je dobrá čitelnost kódu a jeho přehlednost. Čtenář na první pohled vidí, co skripty dělají. Programovaní v Ruby je velmi názorné a popisné. Dalším důvodem použití skriptovacího jazyka Ruby je fakt, že podobně jako databáze genů a signálních drah KEGG pochází z Japonska. Existují tedy dostupné modelové příklady používající KEGG přímo pro Ruby. Nejdůležitější důvod je však integrace knihovny BioRuby, která je sama o sobě velice silným nástrojem pro práci s bioinformatickými daty.

Snahou projektu BioRuby je vytvořit integrované prostředí pro bioinformatiku s použitím skriptovacího jazyka Ruby. Velký důraz je kladen na jednoduchost a popisnost kódů napsaných v BioRuby tak, aby bylo možné BioRuby používat i lidmi bez hlubokých znalostí programování jako takového. Tento open-source projekt začal v roce 2000 v Japonsku podporován University of Tokyo (Human Genome Center), University of Kyoto (Bioinformatics Center) a Open Bio Foundation.

Veškeré dostupné materiály týkající se BioRuby jsou kompletní v Japonštině, většina důležitých věcích je dostupná v Angličtině. V době psaní této práce není znám žádný český zdroj zabývající se BioRuby. [3]

Následující příklad demonstruje výhody a snadnost užití BioRuby.

```
# definice sekvence nukleové kyseliny
seq = Bio::Sequence::NA.new("atgcatgcaaaa")
# ==> "atgcatgcaaaa"

# komplementární řetězec
seq.complement
# ==> "ttttgcatgcat"

# zastoupení nukleotidů
seq.composition
# ==> {"a"=>6, "c"=>2, "g"=>2, "t"=>2}

# vrátí jaké aminokyseliny řetězec kóduje
```

```
seq.translate
# ==> "MHAK"
seq.translate.codes
# ==> ["Met", "His", "Ala", "Lys"]
seq.translate.names
# ==> ["methionine", "histidine", "alanine", "lysine"]

# vrátí molekulovou hmotnost
seq.translate.molecular_weight
# ==> 485.605

# aminokyseliny, které kóduje komplementární řetězec
seq.complement.translate
# ==> "FCMH"
```

Kapitola 4

Řešení

Cílem práce dle zadání je zhotovit a implementovat systém, který půjde nad rámec současných řešení. Jako současná řešení jsou v kontextu této práce brány takové programy, které jsou veřejně dostupné a dokumentované. Obsahem této práce však není analýza těchto současných řešení, ačkoli bylo nutné se při tvorbě výsledného systému jejich obsahem a funkčností zabývat. Pokud existují řešení, která nejsou veřejně přístupná, není na ně v této práci brán zřetel.

Zhotovený systém bude testován na reálných vstupních datech. Pro analýzu, od které si slibujeme reálné výsledky, musíme použít kvalitní vstupní data. Od výstupů implementovaného systému této bakalářské práce jsou očekávány relevantní data. Tyto data mohou sloužit k dalším analýzám a výzkumům. Program práce bude zhotoven tak, aby byla možná jeho snadná a maximální možnost konfigurace. Tím bude zajištěna znovupoužitelnost této technologie při analýze dalších dat.

Aby bylo možné tyto nároky splnit, je nutné přesně chápat veškerou problematiku týkající se genové exprese a signálních drah (viz. kapitola 3.2). Navrhované řešení bude právě tyto jevy zkoumat a na základě analýzy také vyvodí o datech konkrétní závěry.

4.1 Návrh řešení

Hlavní výsledek a cíl práce, by měl odhalit odchylky v genové expresi mezi zdravou a nemocnou tkání. Konkrétně by se mělo jednat o rozdíl mezi lidskou tkání s karcinomem a bez karcinomu. Celkovou ambicí je nalézt ukazatele (*markery*) na základě kterých se u dané tkáně dá mikro-čipovou analýzou s určitou pravděpodobností rozhodnout, zda se jedná o rakovinu, či nikoli. Analýzou dat by mělo být možné najít spojitost mezi těmito ukazateli na úrovni genových signálních drah a na úrovni mezibuněčné komunikace.

Za úspěšné řešení bude považováno nalezení sady ukazatelů, které u dané tkáně budou schopny rozhodovat, zda se jedná o tkáň nemocnou či zdravou s minimální úspěšností 75 procent.

Pokud analýzy neodhalí výrazné a prokazatelné spojitosti, bude i toto považováno za dílčí úspěch, protože v moderní genetické analýze je také velice důležité určit slepé uličky.

Použité algoritmy by měly být přehledné v tom smyslu, že jejich použití a porozumění by mělo být určeno i biologům, bez hlubších znalostí informatiky. Proto je jako jazyk práce vybrán skriptovací jazyk Ruby, který svou přehledností přesně tato kritéria splňuje. Pro snadnou analýzu genových drah byla vybrána knihovna BioRuby.

4.2 Vstupní data

Samotnému návrhu analytického řešení předcházela rozsáhlá analýza potenciálních vstupních dat. Výběr kvalitních a dokumentovaných dat se ukázal jako stěžejní pro další postup práce. Pro další zpracování nejlépe vyhovovaly sady dat, na kterých byly prováděny jiné ukázkové analýzy (takzvané *Testbench data*). Pro analýzu lidské rakovinné tkáně je velké množství dat dostupné v [4].

Velké množství nejasností však z nemalé části praktickému využití těchto dat brání. Přesto se tato databáze zdá být nejucelenějším zdrojem výsledků micro-array analýz u různých druhů rakovin. Data, která byla pro tuto práci použita jako referenční, pochází z [2].

4.2.1 Formát, obsah dat a jejich význam

Použitá data byla použita jako příloha článku zabývajícím se *klastrováním* (angl.: Clustering analysis) rakovinných a normálních dat. Klastrování je jeden z přístupů moderní analýzy těchto dat. Vzorek obsahuje výsledky analýzy exprese 2000 genů pro 62 případů. Dokumentace těchto dat uvádí názvy jednotlivých genů (několik odlišných označení – *UMGAP*, *HSAC07*). Každý případ je dále jasně označen zda se jedná o rakovinou nebo o zdravou tkáň.

Popis obsahuje informace, jestli dané vzorky pochází od jednoho zkoumaného člověka. Část dat obsahuje páry zdravé tkáně a nemocné tkáně pro jednoho jedince. Zbytek dat tyto dvě varianty neobsahuje. Zkoumaná rakovina v tomto případe je rakovina tlustého střeva (angl.: *Colon tumor*).

Nevýhodou těchto dat je malý počet vzorků. Toto však bývá obecný problém u většiny zdrojů, protože je technologie analýzy pomocí mikročipů stále ještě poměrně drahá. Druhou nevýhodou dat je jejich nejasná normalizace (data nebyla normalizována střední intenzitou každého experimentu). Jelikož byla tato data používána pro jiné experimenty a jejich střední intenzita se výrazně neliší, budeme tato data považovat za dostatečně normalizovaná. Všechny vstupní data jsou umístěna v elektronické CD příloze této práce a jsou popsána v příloze A.

4.3 Návrh experimentů

4.3.1 Analýza 1 – prosté porovnání vzorků

Plán první analýzy nejprve spočívá v porovnání hladiny exprese 2000 genů dvou konkrétních vzorků proti sobě. Tento experiment je proveden prostým rozdílem exprese analyzovaných genů. Výsledek je v případě požadavku vykreslen do grafu. Druhá část prvního experimentu seřadí výsledný vektor hodnot rozdílu exprese genů. Tato data jsou opět použita jako základ grafu a případně vykreslena.

Analýza bude pracovat buď s konkrétními hodnotami exprese, nebo s hodnotami, které budou pomocí prahování omezeny na 1 (gen se projevuje) a 0 (gen se neprojevuje). Hodnotu prahu bude umožněno měnit.

Výsledek této analýzy ukazuje míru rozdílnosti mezi vzorky. Je tedy možné určit odlišnosti v projvech genů zdravé tkáně jednoho člověka oproti tkáni s karcinomem nebo oproti jiné zdravé tkáni, pocházející od jiného zkoumaného subjektu.

4.3.2 Analýza 2 – porovnání skupiny vzorků

Druhá analýza provádí složitější srovnání jednotlivých skupin (*Normal* = zdravá tkáň a *Cancer* = tkáň s karcinomem). Srovnání probíhá podobně jako v prvním experimentu s tím rozdílem, že je porovnáván každý subjekt z první porovnávané skupiny s každým subjektem ze skupiny druhé. Pokud děláme analýzu napříč jedné skupiny, do celkového skóre se nepočítají analýzy identických vzorků mezi sebou a také jsou odstraněny duplicitní výpočty. Touto analýzou můžeme vyšetřit tři možné hlavní varianty:

- *Normal* - *Normal* – analyzuje vzájemnou rozdílnost zdravých tkání na úrovni genové exprese;
- *Normal* - *Cancer* – analyzuje vzájemnou rozdílnost zdravých tkání oproti tkáním zasažených rakovinou;
- *Cancer* - *Cancer* – výsledkem poslední analýzy je určení vzájemné rozdílnosti mezi rakovinnými tkáněmi jednotlivých případů.

Od této analýzy si slibujeme prokázání, nebo vyvrácení tvrzení, že existuje rozdíl mezi expresí konkrétních genů mezi zdravou a nemocnou tkání. Tento rozdíl, pokud má prokázat skutečnou existenci odlišností, by měl být výrazně větší než výsledky analýzy napříč stejnými skupinami.

Výsledek experimentu bude opět možné vykreslit do grafu. Analyzující skript umožní provést naráz všechny tři analýzy nebo dovolí pouze porovnání pro dvě konkrétní skupiny. Jelikož je tento výpočet poměrně náročný, ostatní analýzy budou vycházet z dat zpracovaných touto analýzou.

4.3.3 Analýza 3 – označení markerů

Náplní tohoto experimentu je prokázat souvislosti mezi rakovinou a nejvíce odlišnými *up-regulovanými* a *down-regulovanými* geny (u kterých je nejvíce změněna hodnota jejich exprese – *up* nahoru, *down* dolů). Analyzující skript umožní brát v úvahu určitý počet genů, které se nejvíce projevili při druhé analýze (Cancer - Normal). Skript umožní několik úhlů pohledů. Analýza umožní:

- Na základě zadaného prahu rozdělit vstupní data (podobně jako v experimentu 1, skóre bude vypočteno jako suma zadaného počtu up-regulovaných genů mínus suma stejného počtu genů down-regulovaných.
- Bez použití prahování, skóre bude vypočteno jako suma expresí zadaného počtu up-regulovaných genů mínus suma expresí genů down-regulovaných.
- Obdobné výpočty pro zadaný počet nejvíce změněných genů.
- V neposlední řadě analýza umožní vypočítat skóre pro libovolnou sadu genů. Toho bude použito v experimentu 4.

Analyzující skript bude provádět interní korekce výsledků tak, aby poměr up-regulovaných a genů down-regulovaných vyjádřil mezní hranici, od které bude možné určit zastoupení rakoviny. Hodnoty nad touto mezí budou znamenat zdravou tkáň, hodnoty pod touto hladinou tkáň s karcinomem.

Výsledky této analýzy pomocí vstupních markerů jsou porovnány se skutečným stavem tkáně. Výstup skriptu bude určovat nakolik vstupní geny dokáží odlišit rakovinnou tkáň od tkáně zdravé.

4.3.4 Analýza 4 – kombinace genů

Test pomocí experimentu 4 generuje všechny kombinace vstupních genů a postupně na ně aplikuje analýzu 3. Jelikož vypočítání všech kombinací pro 2000 genů je nad rámec výpočetních možností, spokojíme se s kombinacemi zadaného počtu nejrozdílnějších genů dle analýzy 2. Analýza opět umožní zadání libovolné sady genů, ze kterých budou generovány kombinace.

Výsledek skriptu seřadí skupiny genů podle relevance. Která skupina genů má větší vliv na prokázání rakoviny dostane větší ohodnocení (*skóre*).

4.3.5 Analýza 5 – propojení výsledků analýz 1-4 se signálními drahami

Tato poslední analýza hledá vztahy mezi geny tím, že se pokusí najít jejich zastoupení v známých signálních drahách. Jelikož však zatím neexistuje signální dráha pro rakovinu tlustého střeva, budeme hledat vztahy mezi geny napříč ostatními známými signálními drahami.

Od této poslední analýzy si slibujeme nalezení jistého vztahu mezi geny, které skutečně analyticky ovlivňují výskyt rakoviny tlustého střeva.

Tato poslední analýza vyžaduje připojení k Internetu, protože přistupuje vzdáleně k KEGG API serveru (viz. 3.3.1).

4.4 Nastavení skriptů

Veškeré věci týkající se nastavení parametrů skriptů jsou uvedeny v CD příloze v souboru *readme.txt*, případně každý skript obsahuje nápovědu (při použití parametru `--help`).

Kapitola 5

Výsledky experimentů

Tato kapitola uvádí výsledky, kterých bylo dosaženo analýzami vstupních dat pomocí experimentů uvedených v kapitole 4. Výsledkem bylo získáno mnoho výstupů ve formě textových souborů, tak i mnoho grafů. V této části bude uvedeno, co je od analýzy očekáváno, parametry vstupu skriptů a výsledky ve formě textového popisu či grafů, ty jsou pro demonstraci výsledků nejefektivnějším nástrojem.

5.1 Analýza 1

5.1.1 Porovnání dvou vzorků tkání

První experiment porovnává vzorky nemocné tkáně subjektu 1 (označen *vzorek 0*) proti zdravé tkáni subjektu 1 (*vzorek 1*). Experimentem chceme určit, které geny mají odlišnou expresi při porovnání dvou konkrétních vzorků tkáně jedné osoby.

Parametry skriptu

```
./analyze1.rb -a 0 -b 1 -g
```

Výsledky

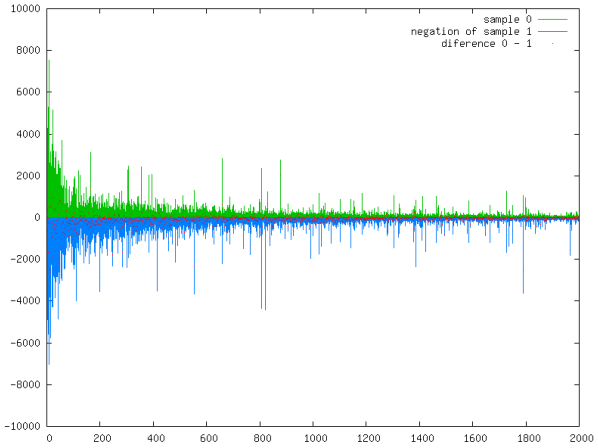
V grafu 5.1 je zelenou barvou jsou vyneseny diskrétní hodnoty vzorku 0. Modrou barvou pak hodnoty vzorku 1. Červenými body jsou označeny rozdíly těchto hodnot. Graf 5.2 obsahuje stejná data jako 5.1, ale pro větší názornost jsou tato data seřazena podle hodnoty rozdílu. Nalevo tedy nalezneme geny, které se projevují více ve zdravé tkáni, vpravo ty, které se projevují v tkáni s rakovinou.

5.1.2 Porovnání dvou vzorků tkání s použitím prahování

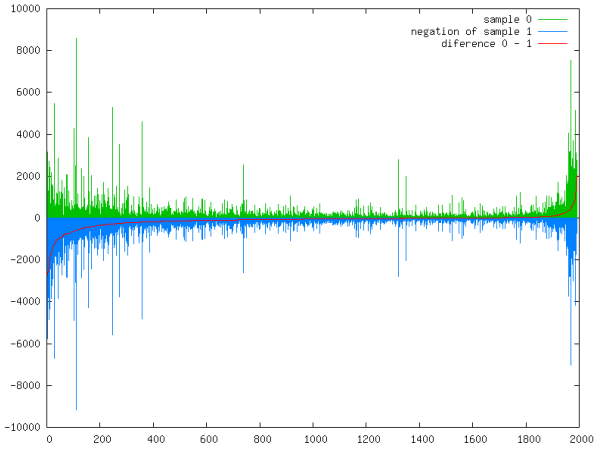
Druhý experiment porovná dva vzorky a použije metodu prahování. Hodnoty exprese budou omezeny na hodnotu 1 (gen se projevuje) a 0 (gen se neprojevuje), hranicí bude hodnota exprese 1000.

Parametry skriptu

```
./analyze1.rb -a 0 -b 1 -g -t 1000
```



Obrázek 5.1: Porovnání dvou vzorků tkání

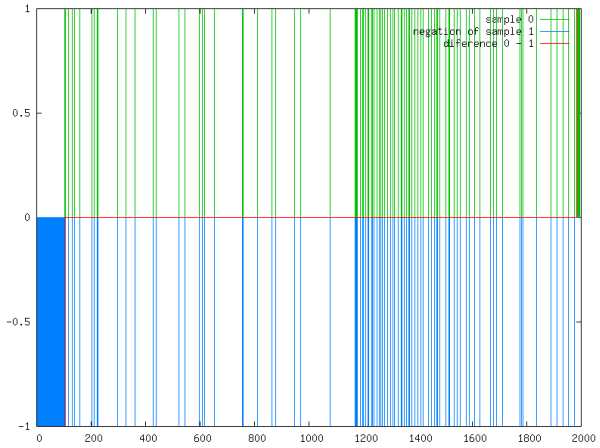


Obrázek 5.2: Porovnání dvou vzorků tkání – seřazeno

Výsledky

Graf 5.3 obsahuje hodnoty s použitím prahování (práh nastaven na 1000). Vzorky jsou nejprve oprahovány, pokud je hodnota exprese větší než práh, gen je považován za aktivní. Rozdílem takto oprahovaných vzorků (*samplů*) vidíme, že geny se dají rozdělit do čtyř skupin.

- aktivní u obou vzorků
- neaktivní obou vzorků
- aktivní u 0, neaktivní u 1 – tyto geny budeme nazývat *up-regulované*
- aktivní u 1, neaktivní u 0 – tyto geny budeme nazývat *down-regulované*

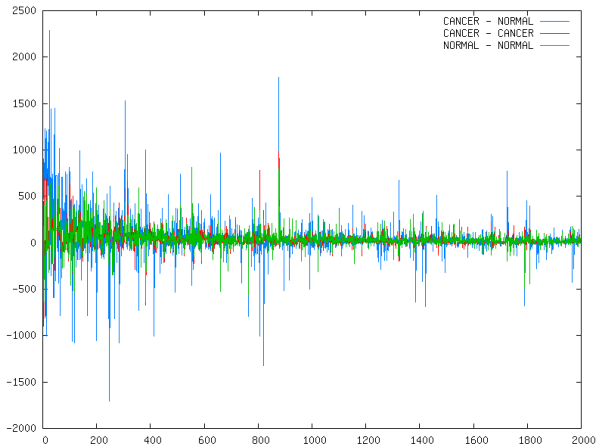


Obrázek 5.3: Porovnání dvou vzorků tkání – s použitím prahování (práh = 1000)

Závěr

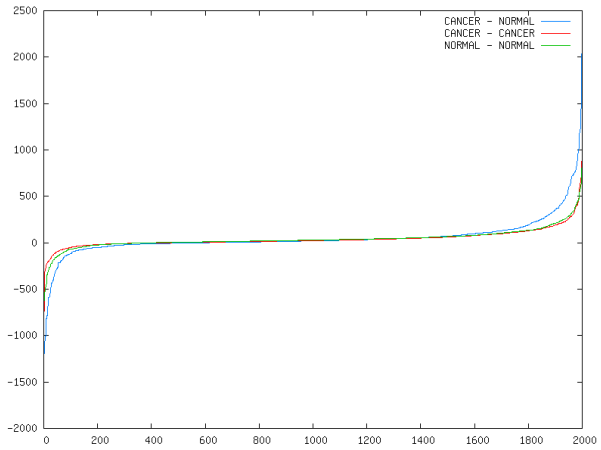
Podobné výsledky a podobné grafy byly získány i porovnáním dvou různých zdravých tkání nebo dvou odlišných tkání s rakovinou tlustého střeva. Počet up-regulovaných a down-regulovaných genů byl v průměru velmi podobný. Analýzou textových výstupů bylo zjištěno, že se jednalo o různé geny. Odlišnosti aktivaci konkrétních genů u nádorů oproti aktivaci genů u normální tkáně by měla odhalit analýza 2.

5.2 Analýza 2



Obrázek 5.4: Porovnání skupin – všechny varianty

Druhou analýzou byly porovnány všechny kombinace subjektů skupin *Normal* a *Cancer*. Roztřídění do těchto skupiny bylo učiněno na základě informací ze vstupních dat (soubor: *tissues.txt*).



Obrázek 5.5: Porovnání skupin – všechny varianty – seřazeno

Parametry skriptu

```
./analyze2.rb -a -g
```

Výsledky

V grafu 5.5 modrá křivka vyznačuje seřazené hodnoty sum rozdílů pro vzorky ze skupiny Cancer oproti zástupcům skupiny Normal. Červená vyjadřuje rozdíly v expresi genů mezi zástupci skupiny Cancer navzájem. Zelená nakonec rozdíly mezi členy skupiny Normal. Graf 5.4 obsahuje diskrétní hodnoty sum rozdílů pro jednotlivé geny.

Z grafu 5.5 je jasně vidět, že rozdíly mezi tkáněmi nemocnými a zdravými jsou výrazně větší, než u ostatních porovnání. Tomu odpovídá i výrazné zastoupení modrých vrcholů (angl.: *peaků*) v grafu22.

5.3 Analýza 3

Třetí analýza vychází z dat získaných v analýze druhé, konkrétně ze vztahu mezi skupinami Cancer a Normal. Snaží se o rozlišení rakovinné tkáně od zdravé na základě exprese zadaných vstupních genů.

5.3.1 Rozpoznání rakoviny na základě 5 genů s největší změnou exprese

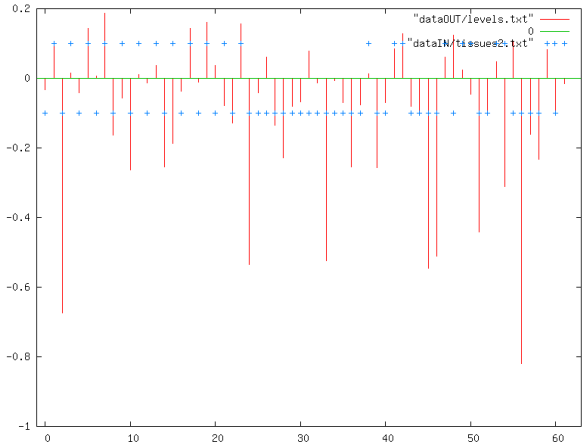
Tento experiment se snaží rozpoznat rakovinu na základě pěti genů s největší změnou exprese.

Parametry skriptu

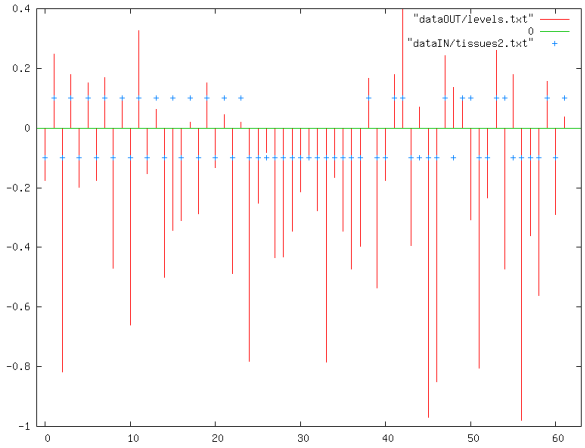
```
./analyze3.rb -x -n 5 -g
```

Výsledky

Pokud jsou červené hodnoty v grafu 5.6 menší než 0, daný vzorek má na základě exprese vybraných genů pravděpodobně rakovinu. Modré křížky vyjadřují reálný stav tkáně (záporné



Obrázek 5.6: Rozpoznání rakoviny na základě 5 genů s největší změnou exprese



Obrázek 5.7: Rozpoznání rakoviny na základě 5 genů up regulovaných a 5 down regulovaných

pro karcinom).

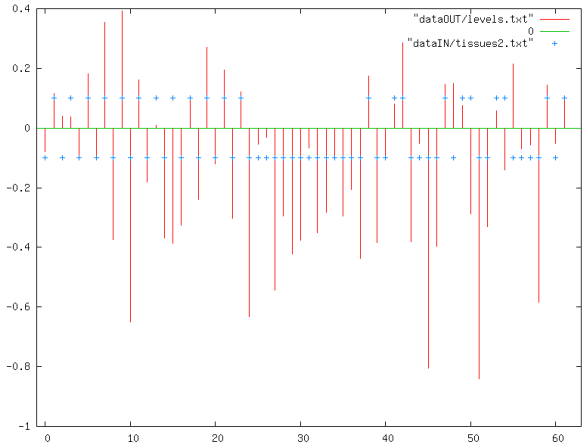
Procentuální úspěšnost: 80,65 %.

5.3.2 Rozpoznání rakoviny na základě genů up regulovaných a down regulovaných

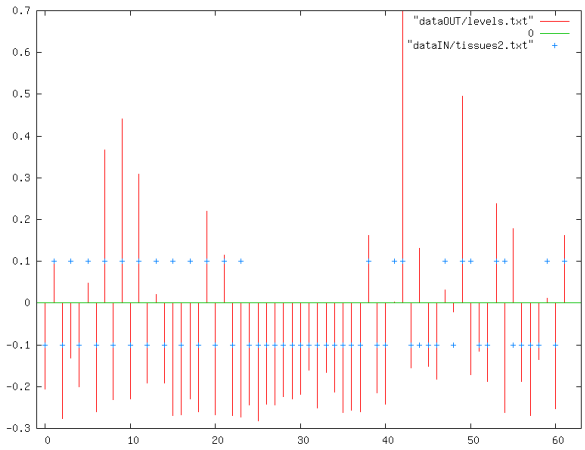
Takto nastavený skript rozpoznává rakovinu na základě pěti genů up regulovaných a pěti genů down regulovaných.

Parametry skriptu

```
parametry: ./analyze3.rb -u -n 5 -g
```



Obrázek 5.8: Rozpoznání rakoviny na základě nejvíce up regulovaného a down regulovaného genu



Obrázek 5.9: Rozpoznání rakoviny na základě jednoho genu – pořadové číslo 248

Výsledky

Z grafu 5.7 je na první pohled znát, že analýza na základě těchto 10 genů má lepší výsledky, než v případě 5.3.1. Rozdíly mezi zdravou a nemocnou tkání jsou výraznější.

Procentuální úspěšnost: 90,32 %.

5.3.3 Rozpoznání rakoviny na základě nejvíce up regulovaného a down regulovaného genu

K této analýze byly použity pouze dva markery. Nejvíce up regulovaný gen a nejvíce down regulovaný gen.

Parametry skriptu

```
parametry: ./analyze3.rb -u -n 1 -g
```

Výsledky

Získané výsledky demonstruje graf 5.8.

Procentuální úspěšnost: 90,32 %.

5.3.4 Rozpoznání rakoviny na základě jednoho genu

V posledním případě byl použit jako marker jediný gen (pořadové číslo 248, název: *Desmin*) a to nejvíce down regulovaný gen. Tento gen byl nejvíce potlačen u rakovinné tkáně narozdíl od tkáni zdravé.

Parametry skriptu

```
echo '248 -1' | ./analyze3.rb -m -x -g
```

Výsledky

Z grafu 5.9 je vidět, že se samotný gen *Desmin* (pořadové číslo 248) ukázal být sám o sobě velice kvalitním markerem.

Procentuální úspěšnost: 87,10 %.

Závěr

Tato analýza ukázala nejzajímavější výsledky. Podařilo se u vzorku 62 případů určit pravděpodobnost výskytu rakoviny v neznáme tkáni tlustého střeva na základě dvou genů s přesností větší než 90 procent.

5.4 Analýza 4

Výstupem analýzy 4 je procentuální ohodnocení úspěšnosti určení rakoviny pro skupiny genů, které jsou kombinací zadaných genů na vstupu.

Parametry skriptu

```
parametry: ./analyze4.rb -n 10
```

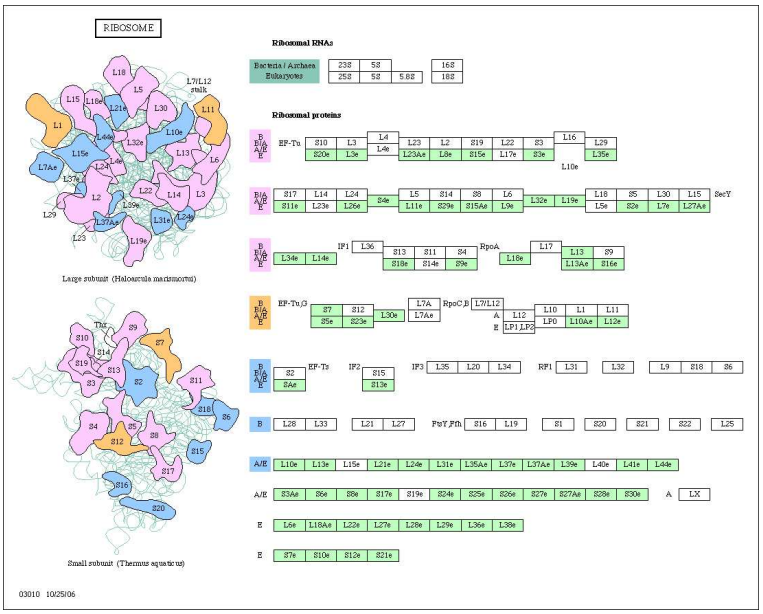
Výstup (prvních 5 nejlepších markerů):

pravděpodobnost	skupina genů
0.9032	25 877 **248** 821
0.9032	25 877 **248** 30 821
0.9032	25 **248** 821
0.9032	25 **248**
0.9032	**248** 46

Závěr

Z této analýzy vyplývá, že dobré markery z první desítky nejvýraznějších genů mohou být ve více kombinacích. Gen s číslem 248 je obsažen ve všech pěti kombinacích. Dále se často objevují geny s pořadovými čísly 25 a 821.

5.5 Analýza 5



Obrázek 5.10: hsa03010 – *Ribosome - Homo sapiens*

Před uskutečněním páté analýzy bylo nutné převést názvy genů na KEGG formát.

```
parametry: ./analyze5.rb -n 10
```

Umístění vstupních genů v genových drahách:

č. genu	KEGG formát	genová dráha
25	hsa:6222	path:hsa03010
248	hsa:1674	path:hsa01430
46	hsa:6156	path:hsa03010
30	hsa:3921	path:hsa03010
30	hsa:3921	path:hsa05060
0	hsa:6171	path:hsa03010
22	hsa:6161	path:hsa03010

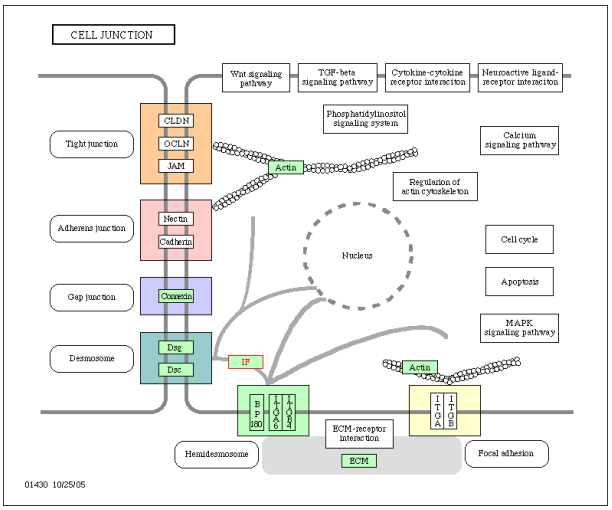
Závěr

Pět z deseti nejvýznamnějších genů se nacházejí v genové signální dráze hsa03010.

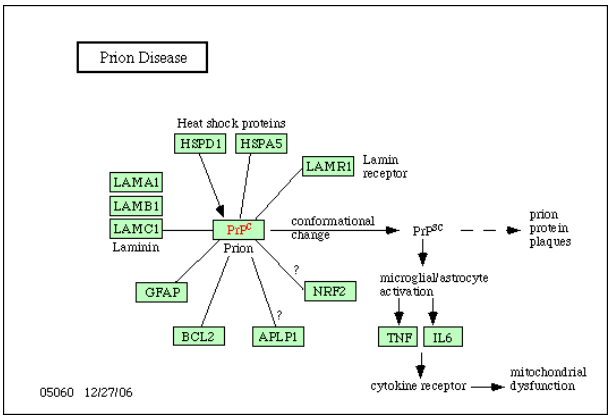
5.6 Celkové zhodnocení

Nalezené dráhy byly seřazeny dle četnosti výskytu takto:

1. hsa03010
2. hsa01430
3. hsa05060



Obrázek 5.11: hsa01430 – *Cell Communication - Homo sapiens*



Obrázek 5.12: hsa05060 – *Prion disease - Homo sapiens*

5.6.1 Popis a ukázky nalezených signálních drah

Genová dráha hsa03010 (*Ribosome - Homo sapiens*) je na obrázku 5.10.

Genová dráha hsa01430 (*Cell Communication - Homo sapiens*) je na obrázku 5.11. Červeně je zdůrazněno místo, kde se projevuje gen 248. Jak je vidět jeho uplatnění je velmi blízko jádra. U rakovinných buněk je tento gen silně potlačen.

Genová dráha hsa05060 (*Prion disease - Homo sapiens*) je na obrázku 5.12.

Kapitola 6

Závěr

V této poslední kapitole jsou uvedeny a shrnuty hlavní výsledky této bakalářské práce. V praktické části této publikace byly prováděny analýzy nad vstupními testovacími daty, které obsahovaly vzorky z tkáně rakoviny tlustého střeva. Řada experimentů přinesla zajímavé výsledky. Porovnáním zdravé tkáně a tkáně nemocné se ukázalo, že v obou tkáních byly aktivní jiné geny. Průzkum napříč skupin demonstroval, že rozdíly mezi všemi nemocnými tkáněmi a tkáněmi zdravými byly výrazně vyšší než vzájemné rozdíly mezi tkáněmi s rakovinou a bez rakoviny.

Pro další výsledky zkoumání byly brány geny s nejodlišnější expresí (mezi rakovinnou tkání a zdravou tkání). Tyto geny byly podrobeny analýze, jejíž výsledek měl za cíl určit s jakou úspěšností tyto geny mohou označovat přítomnost rakovinného bujení. Jako dobrý výsledek se dá považovat určení rakoviny na základě exprese pouze dvou genů, a to s více než devadesáti procentní úspěšností. Velmi překvapující byl výsledek, že ve zkoumané skupině 62 vzorků dokáže samotný jeden gen (*Desmin*, označení KEGG – hsa:1674) na základě své exprese rozlišovat s 87 % úspěšností tkáň rakovinnou od tkáně zdravé.

Deset genů, projevujících největší změny, bylo podrobeno zkoumání v kontextu signálních drah. Pět z deseti genů se nacházelo v jedné dráze (KEGG označení path:hsa03010 s názvem *Ribosome - Homo sapiens*). *Desmin*, gen s nejlepšími výsledky markování rakoviny, se objevil v dráze (KEGG označení path:hsa01430 nesoucí název *Cell Communication - Homo sapiens*). Za povšimnutí stojí fakt, že se funkce genu projevuje velmi blízko jádra buňky. Je tedy skutečně možné, že exprese genu může mít v případě rakoviny jistý význam.

Omezující pro další testování bude získání dostatečného počtu vstupních dat. Výsledek analýzy by byl o mnoho více hodnotnější v případě, že by byly testovány tisíce vzorků. Bude tedy otázkou dalšího bádání a zkoumání, zdali výsledky této práce skutečně nalezly kvalitní markery rakoviny tlustého střeva a zdali kontext signálních drah má význam, který se na první pohled nabízí.

Velkým úspěchem bude, pokud na tuto práci bude v budoucnu navázáno dalším výzkumem či výsledky této práce dále pomohou při laboratorním vyšetřování.

Literatura

[1] Alberts, B.; Bray, D.; aj.: *Základy buněčné biologie*. Espero Publishing, 2005, ISBN 10: 80-902906-2-0.

[2] Alon, U.; Barkai, N.; aj.: Broad patterns of gene expression revealed by clustering of tumor and normal colon tissues probed by oligonucleotide arrays. [online], [cit. 2008-5-12].
URL `http://microarray.princeton.edu/oncology/affydata/index.html`

[3] BioRuby: BioRuby project. [online], [cit. 2008-5-12].
URL `http://bioruby.org`

[4] Institute, B.: Cancer Program Data Sets. [online], [cit. 2008-5-12].
URL `http://www.broad.mit.edu/cgi-bin/cancer/datasets.cgi`

[5] Kanehisa, M.; Araki, M.; aj.: KEGG for linking genomes to life and the environment. [online], [cit. 2008-5-12].
URL `http://nar.oxfordjournals.org/cgi/reprint/36/suppl_1/D480`

[6] Kanehisa, M.; Goto, S.; aj.: From genomics to chemical genomics: new developments in KEGG. [online], [cit. 2008-5-12].
URL `http://nar.oupjournals.org/cgi/reprint/28/1/27.pdf`

[7] KEGG: Kyoto Encyclopedia of Genes and Genomes. [online], [cit. 2008-5-12].
URL `http://www.genome.jp/kegg`

[8] KEGG: Kyoto Encyclopedia of Genes and Genomes: KEGG API. [online], [cit. 2008-5-12].
URL `http://www.genome.jp/kegg/soap/doc/keggapi_manual.html`

[9] Masopust, J.: Rozluštěný lidský genom – a co dál. [online], [cit. 2008-5-12].
URL `http://web.telecom.cz/dotdiag/dokument/patobio/genom.pdf`

[10] McMURRY, J.: *Organická chemie*. Nakladatelství VUTIUM, 2007, ISBN 978-80-214-3291-8.

[11] Pavlík, E.: Molekulární biologické techniky pro mikrobiologickou diagnostiku. [online], [cit. 2008-5-12].
URL `www.roche-diagnostics.cz/download/la/0403/pcr.pdf`

[12] Rumlová, M.; Pačes, V.; Ruml, T.: Základní metody genového inženýrství. [online], [cit. 2008-5-12].
URL `http://teacher.vscht.cz/dokumenty/ManualFinal.pdf`

[13] Trna, M.: Klasifikace s apriorní znalostí. [online], [cit. 2008-5-12].
URL https://dip.felk.cvut.cz/browse/pdfcache/trnam1_2007bach.pdf

[14] Šárka Vondrášková: Milníky nových poznatků v genetice, jen geny na dědičnost nestačí. [online], [cit. 2008-5-12].
URL
http://www.agronavigator.cz/default.asp?ch=1&typ=1&val=69548&ids=1461

[15] Wikipedia: DNA. [online], [cit. 2008-5-12].
URL http://en.wikipedia.org/wiki/DNA

[16] Wikipedia: DNA microarray. [online], [cit. 2008-5-12].
URL http://en.wikipedia.org/wiki/DNA_microarray

Dodatek A

Obsah souborů se vstupními daty

Vstupní data jsou umístěna na CD příloze této práce. Stručný popis jejich významu naleznete v této příloze.

- *data.txt* – obsahuje matici naměřených dat (hodnoty exprese genů). Každý řádek odpovídá příslušnému genu, sloupec matice konkrétnímu vzorku.
- *tissues.txt* – obsahuje popis všech 62 případů. Každý řádek odpovídá jednomu vzorku tkáně. Pokud číslo, jež označuje vzorek, je záporné, jedná se o tkáň postiženou rakovinou. Pokud je číslo kladné, jedná se o zdravou tkáň. Tento soubor ještě obsahuje vazby mezi jedinci a to takové, že každé číslo označuje jednoho jedince. Příkladem budiž označení vzorků -1 a 1. Oba tyto vzorky tkání pochází od jedné osoby, první je postižena karcinomem a druhá nikoliv.
- *names.txt* – tento soubor obsahuje názvy genů. Každý řádek odpovídá jednomu genu v souboru *data.txt*.