MASARYKOVA UNIVERZITA
FAKULTA INFORMATIKY

Zpracování experimentálních dat ve vazbě na modely FGFR signálních drah

DIPLOMOVÁ PRÁCE

Ing. Karel Sedlář

Brno, podzim 2014

Prohlášení

Prohlašuji, že tato diplomová práce je mým původním autorským dílem, které jsem vypracoval samostatně. Všechny zdroje, prameny a literaturu, které jsem při vypracování používal nebo z nich čerpal, v práci řádně cituji s uvedením úplného odkazu na příslušný zdroj.

Ing. Karel Sedlář

Vedoucí práce: RNDr. David Šafránek PhD.

Poděkování

Rád bych zde poděkoval několika lidem, bez kterých by tato práce nemohla vzniknout. Především RNDr. Davidu Šafránkovi PhD., vedoucímu diplomové práce a svému mentorovi, za velkou motivaci, rady a zkušenosti, které mi neustále předává. Dále Mgr. Pavlu Krejčímu PhD. za poskytnutí dat a Ing. Jiřímu Těthalovi a Mgr. Kateřině Hemalové za náměty při našich společných konzultacích.

Chci poděkovat také své rodině a přítelkyni, jejichž podpora v mém studiu je pro mě neocenitelná.

Abstrakt

FGFR3 signální dráha má zásadní význam při poruchách růstu kostí spojených s onemocněními jako je achondroplazie nebo nádorové bujení. Výzkum této dráhy pomocí nástrojů systémové biologie, jakými jsou například stále využívanější kvalitativní modely, je tak zcela zásadní pro pochopení mechanizmů těchto onemocnění. Laboratorní techniky jako Western blot poskytují možnost jak tuto signální dráhu pochopit. Naměřená data jsou ovšem spojitá a před použitím pro kvalitativní modelování vyžadují vhodné předzpracování a diskretizaci. Tato práce poskytuje rozsáhlý přehled technik, které je možné pro tyto účely využít, včetně jejich srovnání na datech FGFR signální dráhy. Techniky, které bylo potřeba pro použití na Western blot data modifikovat, stejně jako techniky nově navržené, jsou implementovány do přiloženého balíčku funkcí jazyka R.

Abstract

FGFR3 signaling pathway is essential for the bone growth disorders associated with diseases such as achondroplasia, or tumor proliferation. The research of this pathway using systems biology tools, such as more and more utilized qualitative modeling, is thus crucial for the understanding of these diseases and their mechanisms. Laboratory techniques such as Western blot provide a way to understand this signaling pathway. However, measured data are continuous and before using qualitative modeling, they require appropriate preprocessing and discretization. This work provides a comprehensive overview of techniques that can be used for these purposes, including their comparison provided on FGFR signaling pathway data. Techniques that required any modifications for ability to process Western blot data, as well as newly proposed techniques, are implemented in the enclosed package of functions for the R language.

Klíčová slova

FGFR3, Western blot, diskretizace, dynamický model, R

Keywords

FGFR3, Western blot, discretization, dynamic model, R

Obsah

1 Úvod

Mutace některých genů mohou vést ke vzniku různých onemocnění. Příkladem může být bodová mutace v genu FGFR3, která se projeví onemocněním růstu kostí označovaném jako achondroplazie. Tato bodová mutace zapříčiní změny v celé signální dráze, do které je tento gen zapojen. Systémová biologie pak může tyto změny vysvětlit pomocí testování různých modelů. Pochopení chování této signální dráhy pak může být klíčem k vytvoření cílené léčby.

Modely, které systémová biologie používá, mohou být kvantitativní nebo kvalitativní. Kvantitativní ODE (ordinary differential equation) modely nabízí možnost velmi přesné analýzy, která je ovšem závislá na kvalitě a přesnosti laboratorních dat. Proto se v poslední době stále častěji začínají využívat modely kvalitativní, jakými jsou boolovské sítě nebo PLA (piecewise linear aproximation) modely. Takové modely však potřebují, aby naměřená spojitá data byla vhodným způsobem převedena do diskrétní podoby.

V první části tohoto textu se chci zaměřit na popis laboratorní techniky Western blot. Analýzou naměřených dat chci zhodnotit přesnost této techniky a její využitelnost pro kvantitativní a kvalitativní studie.

Hlavním cílem práce je poskytnout přehled technik pro diskretizaci dat získaných Western blot experimenty. Protože využití takových dat pro diskretizaci není typické, chci shrnout všechny relevantní techniky a zjistit jejich využitelnost na poskytnutých datech. V rámci praktické části chci techniky, které bude nutné upravit, implementovat ve vhodném jazyce využívaném pro systémovou biologii. Stejně tak se chci pokusit o návrh a implementaci technik nových.

V poslední části práce se zaměřím na srovnání jednotlivých technik, které jsou pro testovaná data vhodné, a to jednak z hlediska věrnosti s jakou diskretizovaná data reprezentují původně spojitá data a z hlediska statistické úspěšnosti výsledků při srovnání s očekávaným výsledkem podle autorů experimentálních dat.

2 Základní pojmy

V následující kapitole si objasníme základní pojmy, které jsou nezbytné pro pochopení dalšího textu této práce. Korektní zpracování dat pomocí počítače totiž závisí na znalosti biologické podstaty problému a laboratorních postupů, kterými byla data získána.

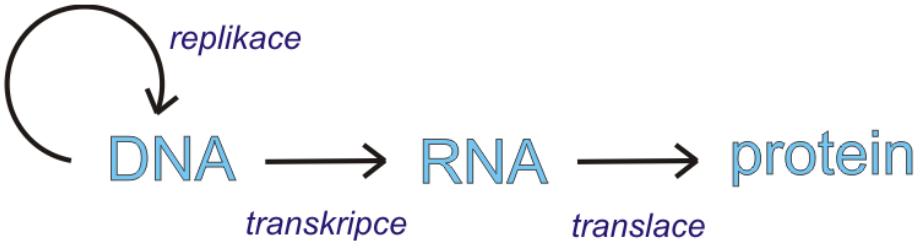
2.1 Omiky

Rozvoj bioinformatiky a systémové biologie jakožto vědních oborů zabývajících se zpracováním a intepretací experimentálních dat je vázán především na rozvoj oborů, v rámci kterých jsou tato data získána. Tyto můžeme souhrnně označit jako omiky, podle společné koncovky jejich názvů. Na prvním místě mezi těmito obory stojí genomika zabývající se studiem genomů, tj. souhrnem veškeré DNA, kterou organismus obsahuje. Její masivní rozvoj přinesla především první sekvenace lidského genomu realizovaná v rámci The Human Genome Project (HGP) [43]. Projekt byl dokončen v roce 2003 a přinesl odhalení zhruba 3,2 miliard párů bází a 25 000 genů lidského genomu. Znamenal také rychlý rozvoj bioinformatiky jakožto nástroje pro zpracování nově získaných dat.

Znalost kompletního genomu člověka znamenala posun do tzv. post genomické éry, která se vyznačuje dalším zkoumáním popsaných genů a jejich produktů. Zde se o další pokrok zasloužily především zbylé omiky. Transkriptomika zkoumá, jak jsou jednotlivé geny přepisovány do mRNA a následně exprimovány. Odhaluje tedy spojitosti mezi geny a změny v jejich expresi určitého typu buněk či tkáně například při chorobách. Navazuje na ni proteomika zabývající se interakcí mezi genovými produkty – proteiny. Metabolomika pak studuje metabolom, úplný soubor látek zapojených do metabolismu buněk a hledá změny jejich koncentrace v čase. Nástrojem pro zpracování dat z těchto vědních oborů se pak stala rychle se rozvíjející systémová biologie.

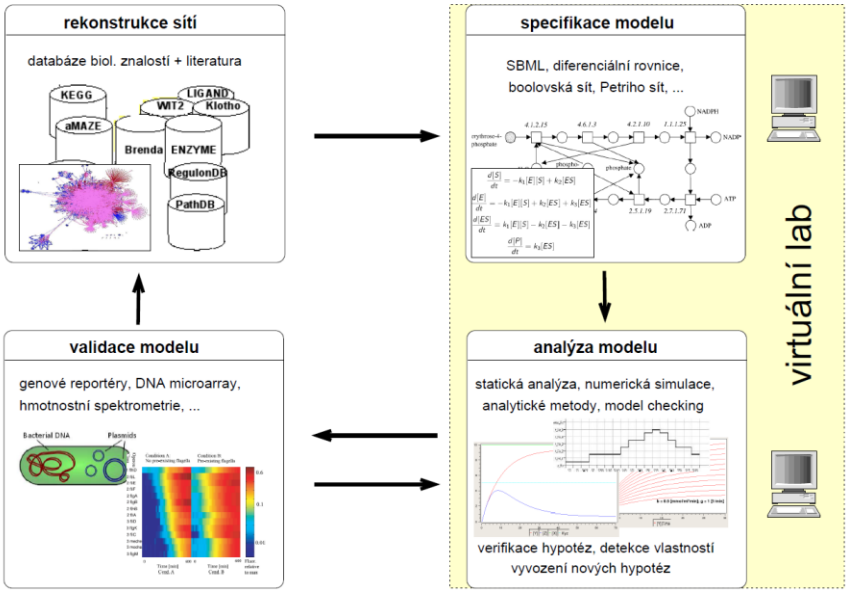
Souhrnně omiky studují, jak se určitý genotyp, tj. soubor veškeré genetické informace, kterou organismus obsahuje, společně s vlivem prostředí projeví ve fenotyp, tedy soubor všech vnějších pozorovatelných

vlastností a znaků organismu. Tuto cestu přenosu genetické informace popisuje centrální dogma molekulární biologie [11]. To se skládá ze tří základních kroků. Prvním krokem je *replikace* DNA zajišťující množení genetické informace. Dále je informace přepisována do mediátorové RNA (mRNA) v procesu *transkripce*. Následně je v procesu *translace* posloupnost tripletů dusíkatých bází přeložena do posloupnosti aminokyselin tvořící primární strukturu výsledného proteinu. Schéma centrálního dogmatu molekulární biologie ukazuje Obr. 2.1.



Obr. 2.1: Centrální dogma molekulární biologie.

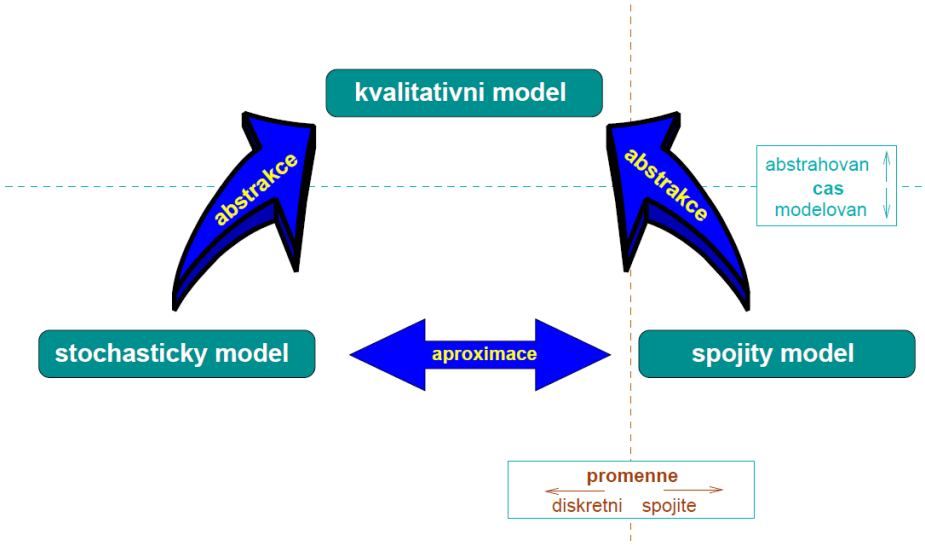
2.2 Systémová biologie



Obr. 2.2: Ilustrace systémového přístupu k biologii [32].

Systémová biologie je interdisciplinární nauka aplikovaná v biologickém a biomedicínském výzkumu. Její hlavní rozdíl oproti bioinformatice spočívá ve vlastní myšlence, se kterou pohlíží na molekulárně biologická data. Paradigmatem bioinformatiky je redukcionismus, tj. snaha zkoumat jednotlivé komponenty (molekuly) nebo jednotlivé interakce (vztah mezi molekulami) [37]. Systémová biologie je založená na holistickém přístupu, tj. snaze o komplexní pohled na systém, který je jako celek víc než pouhý součet jeho částí. Zkoumá tedy tzv. emergentní vlastnosti systému [34].

Systémová biologie pracuje s virtuální laboratoří, ve které je na základě již zjištěných dat sestaven model, který je podroben analýze. Výsledky analýz jsou pak ověřovány pomocí laboratorních postupů. Srovnáním dat z reálných a virtuálních experimentů lze model dále upravit. Celý postup je tak uzavřeným cyklem, který se může několikrát opakovat. Schéma tohoto postupu ukazuje Obr. 2.2.



Obr. 2.3: Dynamické modely v systémové biologii [22].

Kvantitativní modely

Modely kvantitativní se snaží o přesné vyčíslení množství jednotlivých látek v čase. Toto množství může být popsáno buďto koncentrací nebo počtem molekul zkoumané látky. Předpokladem pro popis modelu pomocí

koncentrací je dostatečně vysoká molární koncentrace látek ve zkoumaném systému. Vytvořený spojitý model je deterministický, protože poskytuje pohled na celou populaci molekul. Za daných podmínek generuje pouze jedno chování. Oproti tomu model pracující s interakcemi mezi jednotlivými molekulami poskytuje pohled na jednotlivé molekuly. Je proto modelem stochastickým, který za daných podmínek generuje více různých chování [22].

Kvantitativní modely byly dříve považovány za jediný dostatečně informativní nástroj pro popis modelů v systémové biologii. Standardem se tedy stal především popis pomocí diferenciálních rovnic tzv. ODE (ordinary differential equation) modely [1]. Validace takových modelů je založena na porovnání s laboratorně naměřenými daty [50]. Přesnost laboratorních metod je ovšem velmi omezená, často se používané metody, např. Western blot, přímo označují pouze jako metody semikvantitativní. Vyvstává tak otázka, jestli je možné vytvořit a validovat dostatečně přesný kvantitativní model.

Kvalitativní modely

Modely kvalitativní popisují jednotlivé komponenty systému v různých stavech. Není přitom modelován čas, zajímá nás pouze, jak změna stavu jedné komponenty ovlivní stav ostatních komponent, a tedy stav celého systému [22].

Kvalitativní modely byly dlouho považovány za nedostatečný nástroj pro popis biologických modelů. V posledním desetiletí se však i tyto techniky začaly velmi rychle rozvíjet pro použití v systémové biologii [38]. Pro validaci takovýchto modelů je potřeba laboratorně naměřená data nejprve diskretizovat. Mezi nejrozšířenější nástroj pro tvorbu kvalitativních modelů patří Boolovské sítě [48]. Ty se vyznačují použitím pouze 2 úrovní diskretizovaných dat, a to 1 a 0 (ON a OFF). Současné diskretizační techniky jsou proto zaměřené na binarizaci naměřených dat. V poslední době se však rozvíjí i další nástroje pro popis kvalitativních modelů např. pomocí po částech lineárních aproximací tzv. PLA (piecewise linear aproximation) modely [3]. Tyto modely lze chápat jako nadaproximace

ODE modelů. Mohou pracovat s více než 2 diskrétními úrovněmi. Nástroje pro takovou diskretizaci naměřených dat ovšem zatím chybí.

Systémová biologie a FGFR signální dráha

Z pohledu systémové biologie nás tedy nezajímají změny na jednotlivých molekulách FGFR, ke kterým dochází při jejich fosforylaci, tj. změna hmotnosti, změna konformace atd. Zajímá nás, jak tyto pozměněné molekuly budou interagovat s dalšími molekulami ve smyslu změny šíření určitého signálu napříč celou signální drahou a jak se tato změna projeví navenek, tj. ve fenotypu organismu. Poměrně nový a stále populárnější nástroj pro sledování těchto změn *in silico* tvoří kvalitativní modely. Spolu s tím však musí systémová biologie nabídnout metody pro převod naměřených kvantitativních dat do diskrétní kvalitativní podoby, aby bylo možné vytvořený model validovat. Technik pro binarizaci existuje více a mohou se lišit ve výsledcích, které pro daná data poskytují. Techniky pro diskretizaci na více úrovní zatím chybí. Pro správnou validaci vytvořeného kvalitativního modelu je proto stěžejní buď použít vhodnou stávající diskretizační techniku nebo vyvinout zcela novou.

2.3 FGFR signální dráha

Receptory fibroblastových růstových faktorů (FGFR) jsou receptory s vysokou afinitou k ligandům z rodiny fibroblastových růstových faktorů (FGF) [10]. Samotné FGFR jsou transmembránové proteiny skládající se z 3 hlavních domén. Na extracelulární doménu se vážou ligandy FGF přenášející signál mezi buňkami. Následuje transmembránová doména přenášející signál do buňky. Uvnitř buňky je intracelulární doména, která šíří signál dál díky tyrosin kinázové aktivitě [5]. Existuje více typů FGRF, přičemž některé z nich se mohou uplatňovat při vzniku patologických vlastností. Takovým příkladem může být FGFR3, jehož bodová mutace genu se projeví jako onemocnění achondroplazie.

FRS2 a ERK

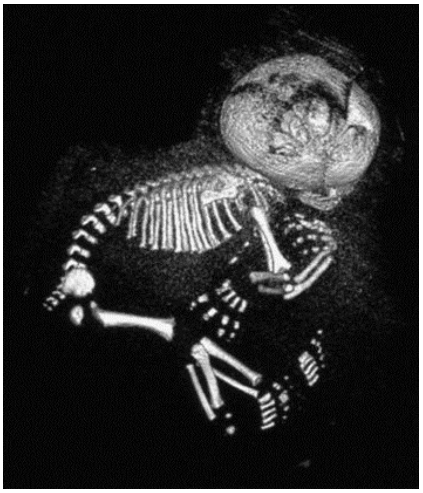
Uvnitř buňky je signální dráha tvořena dvěma hlavními komponentami. První komponentou je FRS2, neboli receptor fibroblastových růstových

faktorů substrát 2, který hraje klíčovou roli v přenosu extracelulárního signálu z intracelulární domény FGFR na další komponentu. Tu tvoří rodina extracelulárním signálem regulovaných kináz (ERK). ERK se pak uplatňují při aktivaci velkého počtu transkripčních faktorů [7].

Achondroplazie

Achondroplazie je charakterizována jako disproporcionální trpaslictví. Nejčastěji postihuje kosti končetin, ty jsou sice standardně široké, ale výrazně kratší a zakřivenější než u zdravého jedince. Vzrůst postiženého v dospělosti dosahuje v průměru 125 cm. Intelektuální schopnosti jedince bývají často nadprůměrné [13]. Onemocnění je autozomálně dominantně dědičné, ale naprostá většina postižených dětí se rodí z důvodu sporadických mutací FGFR3. Tyto mutace mají za následek zvýšení aktivity receptoru, které vede právě k poruchám kostí, neboť inhibuje růst chondrocytů, tj. hlavních buněk chrupavky [15].

Zvýšená aktivita receptoru se v signální dráze projeví trvalou aktivací FRS2 při působení FGF, tedy koncentrace FRS2 se zvýší a v čase neklesá nebo klesá pouze velmi pomalu. Tento stav se označuje jako stav vytrvalý (sustained). Při normální aktivitě receptoru dochází pouze k dočasné aktivaci FRS2, jeho koncentrace začne zhruba po 45 minutách působení FGF zase prudce klesat. Tento stav se označuje jako přechodný (transient) [24].



Obr. 2.4: Obrázek plodu postiženého achondroplazií z počítačové tomografie [4].

3 Laboratorní techniky

Validace vytvořeného modelu přímo závisí na porovnání modelových dat s reálnými daty získanými experimentem. Ve třetí kapitole si proto přiblížíme experimentální techniky. Zaměříme se především na techniku Western Blot, protože zpracování dat, které tato technika produkuje, tvoří hlavní zaměření této práce.

3.1 Laboratorní techniky v systémové biologii

Pro systémovou biologii lze využít široké spektrum laboratorních technik. Ty můžeme rozdělit buďto podle charakteru látky, kterou zjišťují, tj. na transkriptomické, proteomické a metabolomické nebo podle charakteru laboratorní technologie, kterou využívají.

qPCR a sekvenace mRNA

Jednou z nejjednodušších technik na měření genové exprese je kvantitativní PCR (qPCR). Tato technika je založena na cílené amplifikaci transkriptomu pomocí vhodně zvolených primerů. Kvantitativní data jsou získána z amplifikačních křivek pro jednotlivé typy molekul mRNA. Výhodou je vcelku příznivá cena a nenáročnost provedení. Do jedné analýzy je však možné zahrnout pouze omezený počet vzorků [29].

Sekvenace mRNA je moderní a stále častěji využívaná metoda, díky výraznému poklesu ceny sekvenace. Data mohou být kvantifikována pomocí určení počtu jednotlivých čtení připadajících na určitý sekvenovaný amplikon [35]. Výhodou je, že známe přímo i sekvenci mRNA a můžeme tedy sledovat i mutace pro jednotlivé geny.

DNA microarrays

DNA microarrays, česky též DNA čipy, jsou jedním z nejpoužívanějších nástrojů pro experimentální měření v systémové biologii. Název DNA čip je odvozen z toho, že na destičce čipu jsou navázány desetitisíce krátkých oligonukleotidů, ke kterým se komplementárně vážou zkoumané molekuly [42]. Samotnou mRNA, kterou chceme při zjišťovaní exprese analyzovat, je potřeba nejprve převést na cDNA (komplementární DNA),

která má mnohem větší stabilitu než mRNA (to platí i pro qPCR a sekvenaci). Tato cDNA je fluorescenčně značena. Při hybridizaci na komplementární oligonukleotid na čipu lze tak místa, kde došlo k navázání, snadno identifikovat. Výhodou je masivní paralelizace a s tím spojená možnost sledovat velké množství genů najednou. Při použití pro kvantitativní analýzu je však potřeba naměřená data normalizovat a vhodně předzpracovat.

Protein microarrays

Neboli proteinové čipy pracují na obdobném principu jako DNA čipy. Na destičce čipu jsou imobilizovány protilátky, pro které jsou analyzované proteiny antigeny. Tyto proteiny jsou opět, nejčastěji fluorescenčně, značeny, a proto je možné zjistit, kde došlo k navázání antigenu na imobilizovanou protilátku [46]. Hlavní motivací pro vznik této technologie byl fakt, že množství mRNA nemusí vždy reflektovat expresi proteinu. Výhodou proteinových čipů oproti DNA microarrays je také fakt, že proteinové čipy umí zachytit i post-translační modifikace proteinů, které jsou velmi důležité pro určení jejich funkce. Nevýhodou je pak složitější práce s proteiny a s tím spojená technologická náročnost přípravy proteinového čipu. Tato technika proto není komerčně tolik rozšířená.

Hmotnostní spektrometrie

Hmotnostní spektrometrie (MS, mass spectrometry) je chemická analytická metoda sloužící ke zjišťování poměru náboje k hmotnosti analyzovaných molekul. Pomocí ní je tedy možné identifikovat jednotlivé molekuly proteinů [40]. Lze ji také použít přímo k sekvenaci proteinu, tedy odhalení primární struktury tvořené posloupností jednotlivých aminokyselin. Taková analytická souprava je tvořena izoelektrickou fokusací, která separuje jednotlivé proteiny (či spíše peptidy). Poté následuje vysokotlaká kapalinová chromatografie (HPLC) spojená s tandemovou hmotnostní spektrometrií, která dokáže identifikovat jednotlivé aminokyseliny [47]. Na rozdíl od sekvenace mRNA je tato metoda ovšem finančně velmi nákladná, což zabraňuje jejímu použití pro kvantitativní studie.

Mikroskopické techniky

S rozvojem superrezoluční mikroskopie v posledních letech došlo na aplikaci těchto technik i v systémové biologii. Problémem pro sledování proteinů v buňce pomocí optické mikroskopie byl dlouhou dobu difrakční limit. Pomocí technik optické mikroskopie, které umí tento limit překonat, např. konfokální mikroskopie, PALM, STORM [16], je tak možné sledovat proteiny přímo uvnitř buňky *in vivo*. Lze tak provádět experimenty, které jsou ostatními technikami neproveditelné, jako například přímé sledování proteinových interakcí, které může být klíčové pro metabolomické studie [2]. Navíc díky použití fluorescenčních značek lze měření i kvantifikovat a využít tak získaná data pro analýzu dynamických modelů.

Blotovací techniky

Blotovací techniky jsou biochemické metody založené na přenosu zkoumaných fragmentů z gelu pro elektroforézu, ať už agarózového nebo polyakrylamidového, na nitrocelulózovou nebo nylonovou membránu. Na této membráně mohou být zkoumané látky detekovány pomocí radioaktivních nebo i jiných sond. Zkoumané fragmenty vytvoří skvrny (anglicky blot). Na základě velikosti a absorbance těchto skvrn lze měření kvantifikovat. Jako první byla tato metoda popsána v roce 1975 pro analýzu fragmentů DNA a podle svého objevitele byla pojmenována jako Southern blot [39]. Metodologicky podobné techniky pro zpracování RNA a proteinů pak byly pojmenovány přesmyčkou tohoto názvu jako Northern a Western Blot. Výhodou těchto metod je vcelku dobře zvládnutelný technologický postup s poměrně příznivou cenou a možnost kvantitativního měření. Na druhou stranu přesnost měření může být ovlivněna několika faktory, jak si ukážeme v následující kapitole o Western blottingu.

3.2 Western Blot

Western blot je kvantitativní nebo spíše semikvantitativní biochemická metoda pro zpracování proteinů vyvinutá v roce 1979 Harrym Towbinem a jeho kolegy [44]. Někdy se označuje také pojmem imunoblot, protože

k detekci specifických proteinů ze vzorku se používají specifické imunoglobuliny, které jsou protilátkami pro analyzované antigeny. V době svého vzniku byla metoda popsána jako kvalitativní s cílem pouze detekovat přítomnost specifických antigenů. I v dnešní době je takto metoda stále využívána pro diagnostiku řady infekčních onemocnění díky své vysoké specificitě. Současně s vývojem zařízení pro digitální akvizici obrazů našla metoda uplatnění i pro kvantitativní měření v nasnímaných blotech při využití denzitometrického měření. Takové kvantitativní měření je ovšem zatíženo mnoha nepřesnostmi, způsobenými například přidáním různých chemických látek, použitím různých snímacích zařízení a v neposlední řadě i použitím jiného software pro zpracování nasnímaného obrazu [17]. Z tohoto důvodu je metoda Western blot spíše semikvantitativní. Než pracovat s přesně naměřenými daty, je často vhodnější sledovat spíše průběhy jednotlivých měření v čase nebo rozdíly mezi hodnotami v jednotlivých časech měření. Pro diskretizaci dat se pak jeví jako vhodnější použít více opakování stejného měření, než snaha o nalezení konkrétní hladiny, kterou jsou naměřená spojitá data diskretizována. Takové metody diskretizace ovšem zatím nebyly popsány.

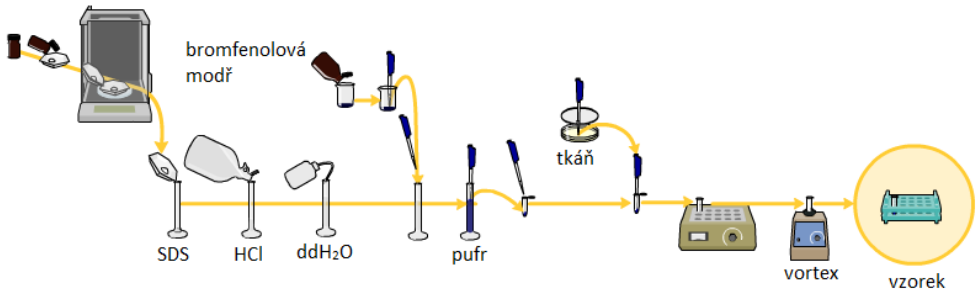
Samotný Western blot pokus lze rozdělit do několika částí. Prvním krokem je izolace proteinů z tkáně nebo buněk. Následuje elektroforetická separace proteinů, v určitých případech i s využitím 2D elektroforézy. Rozdělené proteiny jsou poté přeneseny na membránu, na které jsou pouze požadované proteiny vizualizovány. Nakonec je membrána nasnímána a data jsou kvantifikována. Celý postup si podrobněji popíšeme v následujících podkapitolách.

Izolace proteinů

Izolaci proteinů je možné provádět přímo z tkání nebo z kultivovaných buněčných linií. V prvním případě je potřeba tkáň mechanicky rozrušit, například v třecí misce. Následně je potřeba buňky lyzovat, aby došlo k uvolnění proteinů. K tomu slouží směs rozpouštědel, solí a pufrů. Často se přidávají inhibitory proteáz a fosfatáz, aby nedošlo ke zničení proteinů vlastními enzymy buňky, které jsou při lyzaci také uvolněny. Při celém

procesu je potřeba udržovat nízkou teplotu, aby nedošlo k nežádoucí denaturaci proteinů.

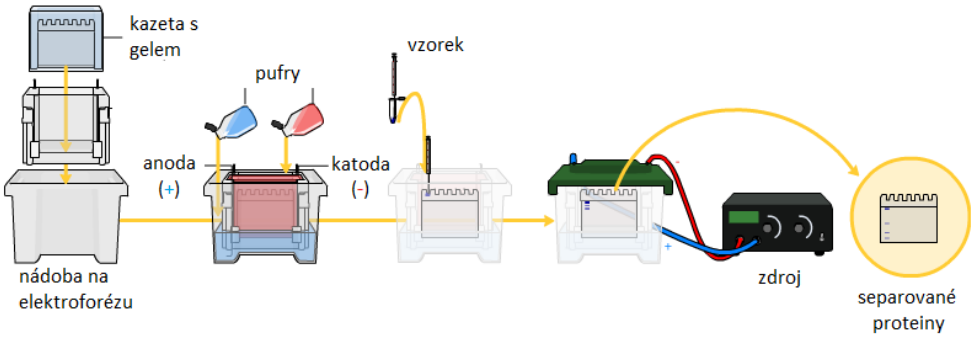
Před samotnou elektroforézou, zpravidla SDS-PAGE (sodium dodecyl sulfát - polyakrylamidová gelová elektroforéza), je potřeba izolované proteiny smíchat s nanášecím pufrem. Ten obsahuje bromfenolovou modř, aby bylo možné sledovat průběh elektroforézy a včas ji zastavit. Dále obsahuje SDS, neboli sodium dodecyl sulfát. Úkolem SDS v gelu je denaturace proteinů. Proteinům působením SDS zůstává pouze primární struktura. SDS se naváže na protein v poměru 1,4 g SDS na 1 g proteinu a odstíní jeho náboj, který překryje svým vlastním záporným nábojem. Relativní náboj vztažený na velikost proteinu je tak u všech fragmentů stejný a při jejich rozdělování v gelu tak závisí pouze na velikosti molekuly. Postup izolace je ukázán na Obr. 3.1.



Obr. 3.1: Izolace proteinů pro Western blot. [31]

Elektroforetická separace proteinů

Jak již bylo řečeno v předchozí kapitole, pro Western blot se používá především SDS-PAGE elektroforéza. Jedná se o separaci proteinů v polyakrylamidovém gelu na základě jejich hmotnosti, protože proteiny mají pouze záporný náboj navázaného SDS, který je úměrný jejich hmotnosti. Pohyb proteinů tak probíhá směrem od katody k anodě. Postup při elektroforéze je ukázán na Obr. 3.2. Pokud do některé z jamek v gelu navíc naneseme proteinový ladder (žebříček), je možné určit hmotnost molekuly v kDa.



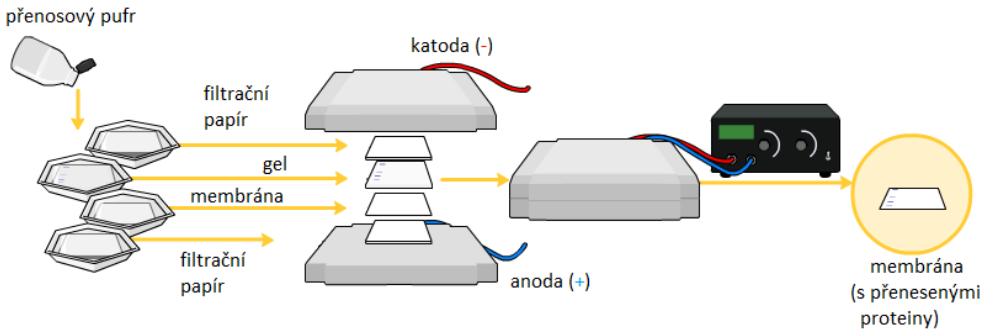
Obr. 3.2: Elektroforéza proteinů. [31]

Při větších nárocích na přesnost separace proteinů je možné provést 2D elektroforézu. V takovém případě se v prvním kroku provede izoelektrická fokusace, což je gelová elektroforéza, při níž jsou fragmenty rozděleny na základě izoelektrického bodu (pH při kterém mají nulový náboj). Následně se provede klasická elektroforéza ve směru kolmém na tu první. Pro kvantitativní Western blot je ale tato metoda využívána pouze okrajově.

Přenos na membránu, blokování a detekce

Pro samotný blotting je potřeba přenést proteiny z gelu na membránu. Používá se buďto membrána nitrocelulózová nebo PVDF (polyvinylidendifluoridová). Je možné použít celý gel, nebo vyříznout pouze tu část gelu s fragmenty o požadované délce, kterou odečteme pomocí proteinového ladderu. U kvantitativního Western blotu chceme totiž měřit látky, které známe, respektive, jejichž hmotnost známe.

Přenos lze provést dvěma způsoby. Gel položíme na membránu a seshora i zdola přiložíme filtrační papír. Celou soustavu zatížíme. Kapilárním vzlínáním následně dojde k přenosu proteinů z gelu na membránu. Tento proces je však zbytečně zdlouhavý a proto se již moc nepoužívá. Separované proteiny mají stále záporný náboj, místo kapilárního vzlínání je tak možné opět použít elektrické pole, které přenos proteinů výrazně urychlí. Taková soustava je ukázána na Obr. 3.3.



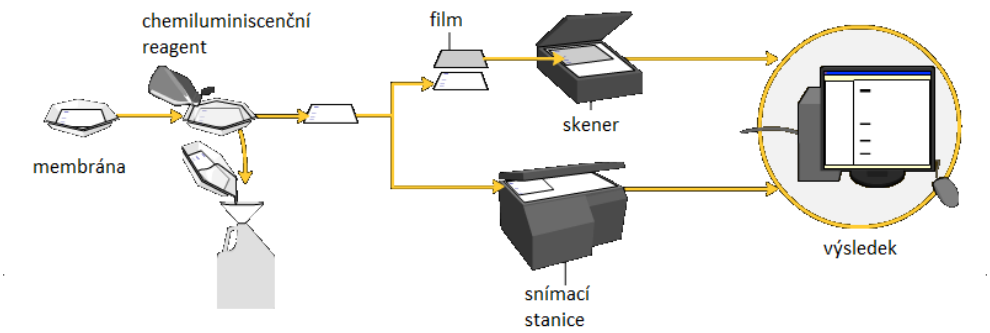
Obr. 3.3: Přenos proteinů na membránu. [31]

Samotné rozdělení proteinů dle jejich hmotnosti není dostatečné, vizualizovat je nutné pouze specifické proteiny, které nás zajímají. Jak již bylo řečeno v úvodu kapitoly, lze toho dosáhnout použitím specifických protilátek. Membrána ale váže nespecificky každý protein a je proto potřeba volná místa blokovat, jinak by se při vizualizaci zabarvila celá membrána. To lze provést velmi jednoduše umístěním membrány do zředěného roztoku nějakého levného proteinu, např. netučného mléka.

Detekci proteinu už následně provedeme použitím specifických protilátek, pro které je zjišťovaný protein antigenem. Takové protilátky jsou navíc značeny vhodnou značkou. Existují přitom 2 metody detekce. Při jednokrokové metodě se na antigen vážou přímo specifické značené protilátky. Protože vytvoření značených protilátek je poměrně náročné, je častější vizualizace pomocí metody dvoukrokové. V prvním kroku se navážou specifické protilátky a až v druhém kroku se na tyto protilátky navážou další protilátky se značkou. Díky tomu, že se používají protilátky se značkou pouze jednoho druhu, je metoda vhodnější pro komerční použití. Navázání značené protilátky je navíc možné zlepšit přidáním heparinu. To však opět mění výsledek při kvantifikaci naměřených dat.

Podle toho jaké charakteru je značka, je provedena samotná detekce, a to kolorimetrická, chemiluminiscenční, radioaktivní nebo fluorescenční. Při kvantitativním měření se nejčastěji používá metoda chemiluminiscenční. Pro takové zviditelnění se přidá k membráně chemiluminiscenční reagent, který při kontaktu se značkou na protilátce začne vyzařovat viditelné světlo. To je buď přímo detekováno CCD čipem, nebo je membrána

přiložena k fotografickému filmu, který je následně digitalizován pomocí skeneru. Tato metoda přináší do kvantitativního měření další nepřesnosti kvůli nelinearitám v obrázku na filmu. Postup detekce ukazuje Obr. 3.4.



Obr. 3.4: Vizualizace měřených proteinů. [31]

3.3 Kvantifikace naměřených dat

Naměřená data ve formě obrázku, ať již získaného přímo snímací stanicí nebo skenem fotografického filmu, je potřeba kvantifikovat s využitím denzitometrického měření. Denzitometrií lze zjistit intenzitu skvrny. V kombinaci s velikostí skvrny pak můžeme definovat integrovanou optickou hustotu IOD (integrated optical density) jako

$$IOD = S \cdot D, \tag{3.1}$$

kde S je velikostí skvrny, často v mm^2 (čtvereční milimetry) a D je její průměrnou optickou hustotou (bezrozměrná veličina). Průměrnou optickou hustotu lze zjistit ze vztahu

$$D = \log\left(\frac{1}{T}\right), \tag{3.2}$$

kde T je transmitance skvrny. Protože pro zjištění transmitance vzorku je třeba měřit světlo prošlé vzorkem, není možné tuto hodnotu skenem fotografie v běžném kancelářském skeneru zjistit. Optickou hustotu však můžeme zjistit i s využitím světla odraženého jako

$$D = \log\left(\frac{1}{R}\right), \tag{3.3}$$

kde *R* je remise vzorku, tedy rozptýlená část odraženého světla od vzorku.

Pokud pomineme omezený dynamický rozsah běžného kancelářského skeneru, snaží se ještě tato zařízení o maximální informační zisk automatickou úpravou osvícení a kontrastu. Tedy optická hustota jedné části snímku je závislá na denzitách sousedících částí [17]. Přes tato úskalí jsou běžné skenery nejpoužívanějšími nástroji pro snímání Western blot experimentů díky své finanční dostupnosti.

Dalším problémem pro přesnou kvantifikaci je použití software pro výpočet *IOD*. Jednotlivé programy, jako Photoshop, ImageJ, QuantityOne atd., totiž produkují různé výsledky. Je tak velmi problematické srovnávat data z jednotlivých experimentů [17]. Srovnání těchto rozdílů je ovšem nad rámec této práce.

Shrnutím všech těchto problémů dohromady pak logicky dospějeme k tomu, že Western blot není kvantitativní metodou, neboť přesné určení množství sledovaného proteinu na základě *IOD* pozorovaných skvrn je, přes množství nepřesností vzniklých při měření, nemožné. Jako zdroj dat pro vytváření přesných kvantitativních, především ODE, modelů, by tak měl být Western blot brán s velkou rezervou. Jedná se o metodu semikvantitativní, která dokáže věrně popsat vývoj množství proteinu při různých pokusech. Taková data však mohou být dostatečná pro diskretizaci a využití při vytváření PLA modelů, které jsou, jak již bylo řečeno, nadaproximací ODE modelů. Na rozdíl od jiných laboratorních technik, např. DNA microarrays, nelze pro porovnání Western blot dat z různých experimentů plně použít ani normalizaci dat, neboť výše popsané chyby nemusí mít lineární charakter.

3.4 Experimentální data

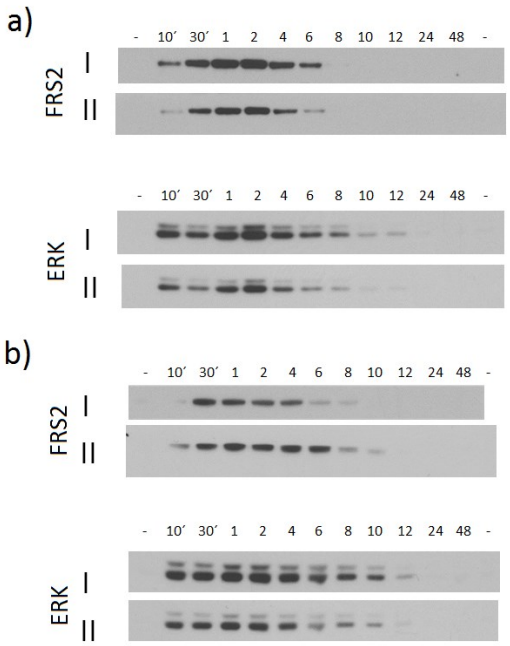
Cílem této práce je zpracování experimentálních dat FGFR signální dráhy. Pro analýzu byla kolegy z LF a PřF poskytnuta data z kvantitativního Western blot experimentu zachycující časový vývoj dvou základních složek signální dráhy, a to FRS2 a ERK v několika časových řadách pro různé buněčné linie. Měření dalších složek signální dráhy nebylo zatím provedeno.

Popis experimentu

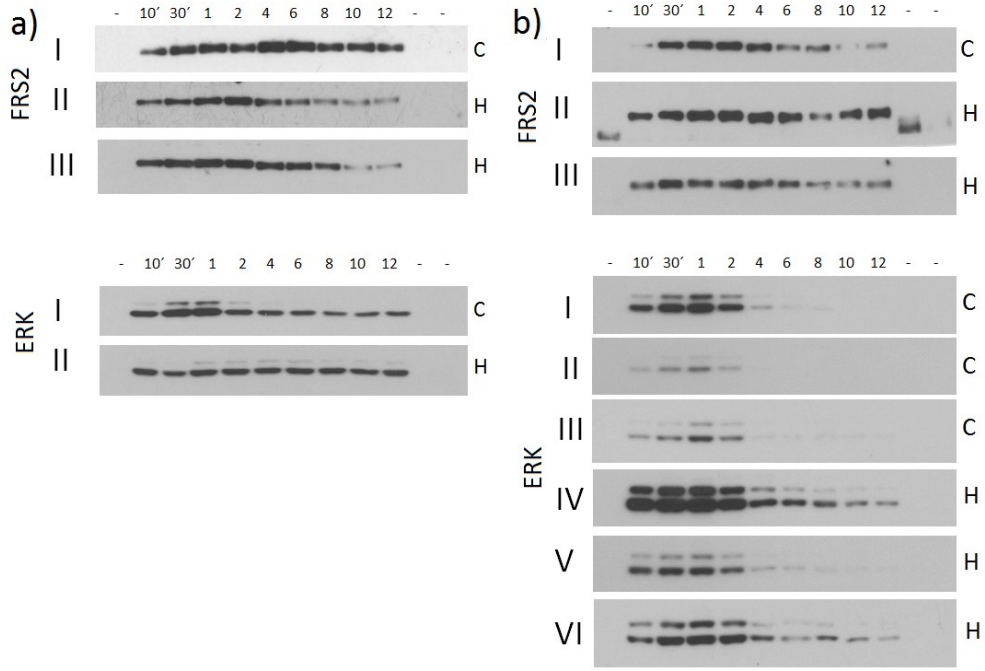
Měření byla uskutečněna pro 4 skupiny různých buněčných linií. Jednou skupinou jsou potkaní chondrocyty postižené chondrosarkomem (RCS – rat chondrosarcoma chondrocytes). Zbylé tři skupiny tvoří lidské kmenové buňky (hESC - human embryonic stem cells) označené jako hESC, hESC CCTL12 a hESC CCTL14. Všechny tyto buňky trpí zvýšenou aktivitou receptorů vedoucí k trvalé aktivaci FSR2 a ERK při přítomnosti ligandu (sustained stav).

Měření množství RFS2 a ERK bylo provedeno Western blot experimentem. Aktivace dráhy byla provedena přidáním ligandu FGF. Pro změření jedné časové řady buněčných linií hESC CCTL12 a CCTL14 bylo vždy provedeno 11 měření v různých časových okamžicích od 10 minut do 48 hodin po přidání ligandu FGF. Navíc před a po zastavení působení ligandu byla provedena kontrolní měření pro zjištění bazální hladiny proteinů FRS2 a ERK. Pro zbylé buněčné linie byla časová řada vytvořena 9 měřeními v časech od 10 minut do 12 hodin při působení ligandu. Navíc bylo provedeno jedno kontrolní měření před přidáním FGF a dvě po odebrání FGF.

Pro každou z buněčných linií bylo naměřeno více časových řad průběhu aktivace signální dráhy. Tato jednotlivá měření jsou označena římskými číslicemi. Pro jednotlivé skupiny byl proveden různý počet měření, dle toho jak moc se jednotlivá měření od sebe lišila (hodnoceno pouze pohledem experimentátora). Pro skupiny hESC a RCS byla navíc provedena měření bez přidání heparinu (označena C) a s přidáním heparinu (označena H). Výsledné snímky pro jednotlivá měření ukazují Obr. 3.5 a Obr. 3.6.



Obr. 3.5: Snímky Western blotů pro buněčné linie a) hECS CCTL12 a b) hECS CCTL12.



Obr. 3.6: Snímky Western blotů pro buněčné linie a) hECS a b) RCS.

Naměřená data

Tab. 3.1: Naměřené *IOD* pro buněčné linie CCTL12 a CCTL14.

	FRS2				ERK			
	CCTL12		CCTL14		CCTL12		CCTL14	
	I	II	I	II	I	II	I	II
CTRL	334	80	376	491	572	178	1198	118
10′	12799	6189	1068	15365	40517	24530	63094	32362
30′	27097	46588	29521	41722	30023	14583	66915	34480
1h	38134	71919	25440	53485	48347	31821	77988	43866
2h	39027	74796	20350	48048	65960	41532	83996	43234
4h	25756	38292	18252	48985	37484	19980	67324	32285
6h	14903	9010	4527	46212	23531	9784	49976	16426
8h	293	435	1872	13408	15365	5499	41203	16063
10h	98	327	696	5500	4849	1091	28188	9068
12h	276	276	292	489	3451	702	7458	1178
24h	183	379	459	409	353	303	604	480
48h	662	982	499	308	299	149	1026	381
CTRL	448	740	523	226	184	542	795	97

Cílem práce je zpracování dat v číselné podobě, proto byla použita již kvantifikovaná data ve formě zjištěné *IOD* přímo na pracovišti, kde byly experimenty prováděny. Pro měření *IOD* byl dle informací použit program Adobe Photoshop. Integrovaná optická hustota není vztažena na jednotku plochy, ale pouze vůči počtu pixelů. Protože byla všechna měření provedena na stejném pracovišti s použitím stejného skeneru (pixel má v každém měření stejnou plochu), je možné pro diskretizaci použít i tato data. Navíc pro diskretizaci nejsou důležité samotné hodnoty, ale spíš průběh celých časových řad. Naměřené hodnoty *IOD* pro buněčné linie CCTL12 a CCTL14 jsou dostupné v Tab. 3.1. Hodnoty *IOD* pro hESC a RCS jsou shrnuty v Tab. 3.2 a Tab. 3.3.

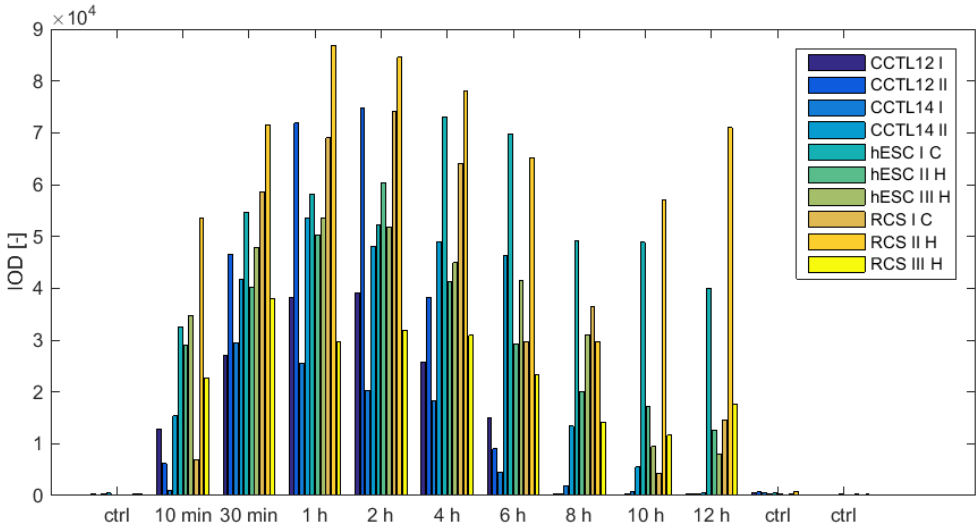
Pohledem do tabulek naměřených dat lze zjistit, že hodnoty pro stejnou dobu působení FGF se mohou mezi jednotlivými měřeními velmi podstatně lišit. Nelze tedy hovořit o kvantitativním měření, nicméně průběh měření celé časové řady zůstává velmi podobný mezi jednotlivými měřeními. Můžeme tak mluvit o semikvantitativním měření. Rozdíly mezi jednotlivými měřeními jsou dobře vidět na sloupcových grafech na Obr. 3.7, respektive Obr. 3.8, zobrazujících průběhy měření FRS2, respektive ERK.

Tab. 3.2: Naměřené *IOD* pro buněčnou linii hESC.

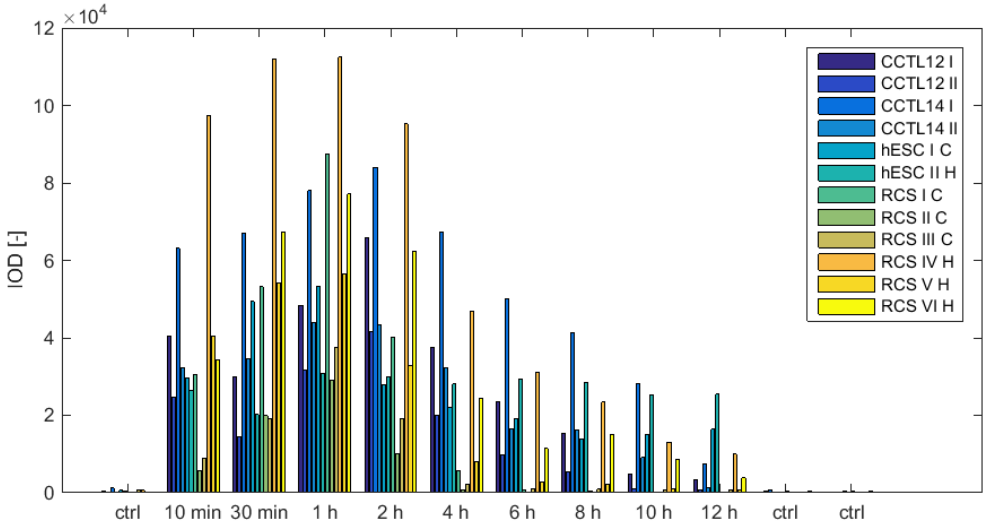
	FRS2			ERK	
	C	**H**	**H**	**C**	**H**
	I	**II**	**III**	**I**	**II**
CTRL	126	173	190	614	469
10´	32513	29123	34756	29629	26385
30´	54649	40264	47873	49371	20199
1h	58204	50250	53603	53369	30679
2h	52247	60358	51883	27907	29937
4h	73054	41213	44922	21933	28066
6h	69792	29275	41542	19132	29478
8h	49184	20077	30989	13975	28408
10h	48863	17294	9457	14962	25181
12h	39960	12621	8076	16404	25436
CTRL	447	336	90	93	190
CTRL	362	144	205	382	121

Tab. 3.3: Naměřené *IOD* pro buněčnou linii RSC.

	FRS2			ERK					
	C	**H**	**H**	**C**	**C**	**C**	**H**	**H**	**H**
	I	**II**	**III**	**I**	**II**	**III**	**IV**	**V**	**VI**
CTRL	178	388	221	183	135	817	611	220	284
10´	6884	53606	22769	30466	5803	8824	97485	40514	34432
30´	58525	71565	38053	53148	19978	19170	112038	54216	67232
1h	69014	86763	29651	87580	29169	37487	112436	56528	77098
2h	74081	84599	31801	40137	10131	19062	95226	32836	62389
4h	64104	77988	30883	5771	627	2234	46808	8172	24472
6h	29633	65187	23408	716	170	1163	31192	2858	11408
8h	36463	29637	14165	399	234	911	23410	2293	15053
10h	4282	57148	11645	275	225	772	13081	1078	8595
12h	14608	71007	17626	146	101	754	9968	669	3813
CTRL	249	715	157	401	183	113	55	273	433
CTRL	382	152	227	298	186	68	62	500	141



Obr. 3.7: Průběh *IOD* pro různá měření FRS2.



Obr. 3.8: Průběh *IOD* pro různá měření ERK.

Zhodnocení dat

Z grafů je patrné, že rozdíl v hodnotě *IOD* mezi jednotlivými měřeními může být i několikanásobný. To může mít za následek, že diskretizace jednotlivých časových řad může dosáhnout jiných výsledků, byť by všechny řady měli po diskretizaci vypadat shodně. Jak již bylo řečeno v teoretické úvodní kapitole, pokud se koncentrace sledovaných proteinů

po 45 minutách nezačne opět prudce snižovat k hodnotě bazální koncentrace, jedná se o stav vytrvalý. Po diskretizaci by tedy všechna měření při působení FGF měla mít jinou než nulovou hodnotu. Jako stav s nulovou hodnotou přitom budeme brát hodnotu bazální přítomnosti proteinů. Poločas rozpadu obou sledovaných proteinů je mnohem delší než naměřené časové řady, sledované hodnoty tak popisují výlučně děje spojené s funkcí sledované signální dráhy FGFR.

4 Diskretizační techniky

V následující kapitole se dostáváme k jádru práce, kterým je samotné zpracování dat ve smyslu jejich diskretizace pro použití při vytváření a zkoumání kvalitativních modelů. Kapitola obsahuje přehled již dostupných metod a jejich analýzu na experimentálních datech, stejně tak jako metody nově navržené v rámci této práce. Jedná se především o metody vhodné pro zkoumání PLA modelů, které mohou mít více než 2 kvalitativní úrovně. Diskretizace často využívá statistické výpočty, většina nástrojů je proto psaná v jazycích zaměřených na statistické zpracování dat jako R nebo Matlab. Nástroje realizované v rámci této práce jsou přiloženy jako funkce napsané v jazyce R kvůli lepší dostupnosti této otevřené platformy.

Stávající nástroje jsou často primárně vytvořeny pro diskretizaci dat z microarray experimentů, jakožto nejvíce používané experimentální techniky při práci s Boolovskými sítěmi. Takové nástroje je ale možné použít beze změny i na kvantitativní Western blot data.

Pro lepší práci při vyhodnocování dat budeme nadále používat pro všechny časové řady dobu měření do 12 h po působení FGF. Pokud to bude žádoucí, je navíc možné data normalizovat použitím přiložené R funkce *data.norm*. Ta upraví naměřená data tak, aby měření pro jednotlivé časové řady nabývaly hodnot $X \in \langle 0,1 \rangle$.

4.1 Metody řízené absolutními hodnotami

Jedná se o základní metody, které využívají určitou statistickou hodnotu jako práh, který je přechodem mezi kvalitativními stavy [28]. Tímto prahem může být průměr, medián, či jiný percentil, případně hodnota odvozená od maxima. Tyto nástroje nejsou přímo dostupné pro zpracování Western blot dat a budou proto realizovány jako funkce R balíčku, který je přílohou této práce. Byť se jedná o vcelku jednoduché techniky, lze diskretizaci provádět několika různými způsoby. Je také možné využít normalizaci dat nebo použít nejčastější výsledek z více časových řad.

Diskretizace s využitím průměru

Při diskretizaci s využitím průměru je jako práh použit průměr ze všech naměřených hodnot *IOD* v rámci časové řady [20]. Diskretizovaná data tak mohou nabývat dvou stavů. Pro PLA modely je pak třetím stavem samotný práh. Pro *i*-tý prvek časové řady X pak platí, že nabývá diskrétní hodnoty D_i dle vztahu:

$$D_i = \begin{cases} 1, \text{když } X_i \geq \bar{X} \\ 0, \text{jinak} \end{cases}, \tag{4.1}$$

kde $\bar{X}$ značí aritmetický průměr.

Pro tento typ diskretizace je možné využít funkci *abs.discretization* s parametrem *method* nastaveným na hodnotu „mean“. Výsledky se mezi jednotlivými časovými řadami mohou podstatně lišit, jak ukazují příklady diskretizaci pro FRS2 v Tab. 4.1 a pro ERK v Tab. 4.2.

Tab. 4.1: Příklady diskretizace časových řad FRS2 podle průměru

	ctrl 1	10′	30′	1 h	2 h	4 h	6 h	8 h	10 h	12 h	ctrl 2	ctrl 3	práh
CCTL14 I	0	0	1	1	1	1	0	0	0	0	0	0	8620
hESC I C	0	0	1	1	1	1	1	1	1	1	0	0	39950
RCS II H	0	1	1	1	1	1	1	0	1	1	0	0	49896

Tab. 4.2: Příklady diskretizace časových řad ERK podle průměru

	ctrl 1	10′	30′	1 h	2 h	4 h	6 h	8 h	10 h	12 h	ctrl 2	ctrl 3	práh
CCTL14 I	0	1	1	1	1	1	1	1	0	0	0	0	40744
hESC I C	0	1	1	1	1	1	0	0	0	0	0	0	20647
RCS II H	0	1	1	1	1	0	0	0	0	0	0	0	5578

Zjištěný práh pro jedno časové měření může nabývat i několikanásobné hodnoty prahu pro měření jiné. To samo o sobě ukazuje, že metoda Western blot není schopna absolutní kvantifikace. Samotná data po diskretizaci se mohou vzájemně také výrazně lišit. Mohou se blížit očekávanému výsledku, že všechna, kromě kontrolních, měření by měla nabývat jiné než nulové hodnoty, ale celkově takového výsledku nebylo dosaženo. Metoda diskretizace na základě průměru je totiž velmi ovlivněna

počtem kontrolních měření, která značně mění průměrnou hodnotu časové řady.

Přesné statistické hodnocení výsledků bude shrnuto až v příslušné kapitole, která bude následovat po představení všech diskretizačních technik.

Diskretizace s využitím mediánu

Metoda je v podstatě stejná jako předchozí, ale místo průměru využívá jako práh medián časové řady [20]. Pro diskrétní hodnotu D_i pak platí vztah:

$$D_i = \begin{cases} 1, \text{když } X_i \geq \tilde{X} \\ 0, \text{jinak} \end{cases}, \tag{4.2}$$

kde $\tilde{X}$ značí medián časové řady a X_i její i-tý prvek.

Pro tento typ diskretizace je opět možné využít funkci *abs.discretization*, avšak s parametrem *method* nastaveným na hodnotu „median“. Vybrané výsledky ukazují Tab. 4.3 pro FRS2 a Tab. 4.4 pro ERK.

Tab. 4.3: Příklady diskretizace časových řad FRS2 podle mediánu

	ctrl 1	10′	30′	1 h	2 h	4 h	6 h	8 h	10 h	12 h	ctrl 2	ctrl 3	práh
CCTL14 I	0	0	1	1	1	1	1	1	0	0	0	0	1470
hESC I C	0	0	1	1	1	1	1	1	0	0	0	0	49023
RCS II H	0	0	1	1	1	1	1	0	0	1	0	0	61167

Tab. 4.4: Příklady diskretizace časových řad ERK podle mediánu

	ctrl 1	10′	30′	1 h	2 h	4 h	6 h	8 h	10 h	12 h	ctrl 2	ctrl 3	práh
CCTL14 I	0	1	1	1	1	1	1	0	0	0	0	0	45589
hESC I C	0	1	1	1	1	1	1	0	0	0	0	0	17767
RCS II H	0	1	1	1	1	1	1	0	0	0	0	0	1036

Z výsledků je patrné, že diskretizace s využitím mediánu časové řady je mnohem robustnější než diskretizace založená na průměru, protože rozdíly mezi jednotlivými diskretizacemi jsou minimální. Na druhou stranu se výsledky opět liší od těch očekávaných. Je ale potřeba brát v potaz, že výsledky je nutné porovnávat s výsledky, které produkuje nějaký model, aby srovnání bylo objektivní. Srovnání s očekávanými výsledky, které vidí v datech člověk provádějící laboratorní experiment, je srovnáním zcela

subjektivním, podloženém pouze znalostí experta. Využití mediánu vždy vede na výsledek, kdy je polovina hodnot pod a polovina hodnot nad prahem.

Výhodou diskretizace pomocí průměru a mediánu může být právě fakt, že výsledek nelze ovlivnit nastavením žádného parametru, čímž eliminuje chybu lidského faktoru. Z biologického hlediska však nelze potvrdit, proč by zrovna medián nebo průměr měl být brán jako prahová hodnota.

Diskretizace s využitím maxima

Další možností jak provést diskretizaci dat řízenou absolutní hodnotou je s využitím maxima časové řady. Pro diskrétní hodnotu D_i pak můžeme definovat vztah:

$$D_i = \begin{cases} 1, \text{když } X_i \geq max(X) \cdot p \\ 0, \text{jinak} \end{cases}, \tag{4.3}$$

kde $max(X)$ značí maximální hodnotu časové řady X a p určuje procento z maxima, které se má považovat za prahovou hodnotu, tedy $p \epsilon \langle 0,1 \rangle$.

Výsledek diskretizace tedy může být ovlivněn uživatelem právě nastavením parametru p. Diskretizaci je možné provést s využitím funkce *abs.discretization* s parametrem *method* nastaveným na hodnotu „max“. Vybrané výsledky pro různé hodnoty parametru p pro časové řady FRS2 ukazuje Tab. 4.5.

Tab. 4.5: Příklady diskretizace časových řad FRS2 podle maxima

	p	ctrl 1	10′	30′	1 h	2 h	4 h	6 h	8 h	10 h	12 h	ctrl 2	ctrl 3	práh
CCTL14 I	0,7	0	1	1	1	1	1	0	0	0	0	0	0	58797
hESC I C	0,7	0	0	1	1	0	0	0	0	0	0	0	0	37358
RCS II H	0,7	0	0	0	1	0	0	0	0	0	0	0	0	26241
CCTL14 I	0,5	0	1	1	1	1	1	1	0	0	0	0	0	41997
hESC I C	0,5	0	1	1	1	1	0	0	0	0	0	0	0	26684
RCS II H	0,5	0	0	1	1	1	0	0	0	0	0	0	0	18743
CCTL14 I	0,1	0	1	1	1	1	1	1	1	1	0	0	0	8399
hESC I C	0,1	0	1	1	1	1	1	1	1	1	1	0	0	5336
RCS II H	0,1	0	1	1	1	1	0	0	0	0	0	0	0	3748

Metodou diskretizace s využitím maxima lze dosáhnout použitím různého vstupního parametru velmi odlišných výsledků. Je možné

parametr nastavit tak, aby diskretizace dosáhla požadovaného výsledku, kdy všechna, kromě kontrolních, měření jsou v nenulovém stavu. Ovšem při daném nastavení je tohoto výsledku dosaženo pouze u některých časových řad, což ukazuje na to, že metoda není moc robustní. To je způsobeno především tím, že je metoda vázána pouze na jednu hodnotu, navíc hodnotu maximální, což může být mnohdy odlehlá hodnota, která vznikla chybou v měření.

Diskretizace s využitím percentilu

Z metod pro diskretizaci dat je nejčastěji využívanou metodou diskretizace dle zvoleného percentilu, který je prahovou hodnotou [12], [21]. Pro diskrétní hodnotu D_i pak platí vztah:

$$D_i = \begin{cases} 1, \text{když } X_i \geq X_p \\ 0, \text{jinak} \end{cases}, \tag{4.4}$$

kde X_p je p percentil časové řady X, tedy $p \epsilon \langle 0,1 \rangle$

I u této metody je tedy možné ovlivnit výsledek uživatelem. Nastavením parametru p na hodnotu 0,2 můžeme na našich datech dosáhnout nejlepší výsledků, jak ukazuje Tab. 4.6.

Tab. 4.6: Příklady diskretizace časových řad FRS2 podle maxima

	p	ctr l1	10′	30′	1 h	2 h	4 h	6 h	8 h	10 h	12 h	ctr l2	ctr l3	práh
CCTL14 I	0,2	0	1	1	1	1	1	1	1	1	0	0	0	523
hESC I C	0,2	0	1	1	1	1	1	1	1	1	1	0	0	6859
RCS II H	0,2	0	1	1	1	1	1	1	1	1	1	0	0	6499

Tohoto typu diskretizace je možné je možné docílit využitím funkce *abs.discretization* s parametrem *method* nastaveným na hodnotu „percentile“. Při nastavení parametru p na hodnotu 0,5 je metoda shodná s metodou mediánu. Volbou tohoto parametru lze velmi dobře volit, kolik hodnot má být po diskretizaci nenulových. Proto ze všech metod řízených absolutní hodnotou poskytuje nejlepší výsledek, jak bude ještě ukázáno v kapitole zabývající se statistickým vyhodnocením úspěšnosti jednotlivých technik.

4.2 Konsenzuální metody řízené absolutními hodnotami

Z představení jednotlivých technik v minulé kapitole je jasně patrné, že výsledky diskretizace pro jednotlivé časové řady se mohou lišit, byť by měl být výsledek stejný. Může za to charakter naměřených dat, která vykazují mnoho nepřesností vznikajících při laboratorním experimentu. Tyto nepřesnosti se dají minimalizovat několikanásobným opakováním experimentu a využitím naměřených dat pro získání pouze jedné, konsenzuální diskretizované časové řady.

Z důvodu přílišného spoléhání na přesnost Western blot experimentů nejsou tyto techniky dostupné. Jejich realizace je však nenáročná a po objektivním zhodnocení přesnosti techniky Western blot i zcela logická. Konsenzuální výsledek lze získat dalším zpracováním tabulky diskretizovaných dat tak, že z každého sloupce je určen nejčastěji zastoupený stav, který je použit jako výsledný konsenzuální stav K_i pro i-tý okamžik časové řady

$$K_i = Mod(\{D_{i,1}, D_{i,2}, .., D_{i,n}\}), \tag{4.5}$$

kde $Mod()$ značí modus z množiny diskrétních stavů pro n měření v i-tém okamžiku.

Konsenzuální diskretizaci je možné provést pomocí funkce *cons.discretization* z přiloženého balíčku R funkcí. Z charakteru diskretizace řízené absolutní hodnotou není nutné před analýzou data normalizovat, protože normalizace na výsledek nemá vliv. Výsledky diskretizace pro FRS2 a ERK ukazují Tab. 4.7 a Tab. 4.8.

Tab. 4.7: Příklady konsenzuální diskretizace časových řad FRS2

	p	ctrl 1	10′	30′	1 h	2 h	4 h	6 h	8 h	10h	12h	ctrl 2	ctrl 3
průměr	-	0	0	1	1	1	1	1	0	0	0	0	0
medián	-	0	1	1	1	1	1	1	0	0	0	0	0
maximum	0,1	0	1	1	1	1	1	1	1	1	1	0	0
percentil	0,2	0	1	1	1	1	1	1	1	1	1	0	0

Tab. 4.8: Příklady konsenzuální diskretizace časových řad ERK

	p	ctrl 1	10′	30′	1 h	2 h	4 h	6 h	8 h	10h	12h	ctrl 2	ctrl 3
průměr	-	0	1	1	1	1	1	0	0	0	0	0	0
medián	-	0	1	1	1	1	1	1	0	0	0	0	0
maximum	0,1	0	1	1	1	1	1	1	1	0	0	0	0
percentil	0,2	0	1	1	1	1	1	1	1	1	1	0	0

Z výsledků vyplývá, že metody, u kterých je možné výsledek ovlivnit vstupním parametrem se více přibližují očekávaným hodnotám, na rozdíl od metod, které nijak ovlivnit nelze. Nejlepší výsledek poskytuje metoda percentilu nastaveného na 20 %, kdy u obou komponent signální dráhy došlo k takové diskretizaci, kdy pouze kontrolní měření byla označena za stav 0.

Také není možné určit prahovou hodnotu. I když PLA modely berou prahovou hodnotu jako jeden ze stavů, není ji nutné diskretizací určit. V kvalitativním modelu totiž není důležité znát číselnou hodnotu prahu, stačí určit jeho pozici v časové řadě. Tu můžeme určit na intervalu, který je určen dvěma sousedícími časovými okamžiky měření, které po diskretizaci nabývají rozdílných stavů. Přesné časové určení také není nutné, protože v PLA modelu je čas abstrahován. Navíc okamžik prahové hodnoty je limitním okamžikem, tedy ve skutečnosti nekonečně krátkým okamžikem.

4.3 BoolNet

BoolNet je balíček funkcí napsaných v jazyce R pro generování, upravování a analýzu boolovských sítí [33]. Jedná se přímo o balíček napsaný pro systémovou biologii, který je komplexním nástrojem pro práci s genovými regulačními sítěmi. Obsahuje i funkci *binarizeTimeSeries* na diskretizaci dat, která v sobě implementuje 3 různé algoritmy. Jelikož je zaměření balíčku na regulační genové sítě, předpokládá použití především na data získaná z microarray experimentů. Diskretizační funkci je ale možné bez úprav použít i na western blot data a signální dráhy. Protože však funkce předpokládá zpracování microarray dat, jednotlivé řádky matice nebere jako opakování téhož měření, ale jako časové řady pro různé geny získané v rámci měření na jediném DNA čipu. Poskytuje tak výsledek stejného

charakteru jako funkce *abs.discretization*, kdy je každá časová řada zpracována zvlášť. Z takto binarizované matice je poté opět možné získat konsenzuální výsledek. Protože je balíček zaměřen na boolovské sítě, umí data pouze binarizovat, diskretizace na více úrovní možná není. Navíc pro jednotlivá měření určuje i práh, pro který nastává změna stavu.

K-means shlukování

Prvním algoritmem, který funkce nabízí pro zpracování dat je diskretizace založená na k-means shlukování. Jedná se o algoritmus nehierarchického shlukování [27], kdy je na začátku určen počet shluků, do kterého mají být objekty, tedy jednotlivá měření v čase, rozděleny. Takový algoritmus by tedy bylo možné použít i na diskretizaci do více než dvou úrovní, avšak funkce má tento parametr nastaven na 2 a nedovoluje změnu. Samotný algoritmus je iterativní, kdy se v každém kroku přepočítává nový střed shluku, a jednotlivé objekty jsou zařazeny do shluku, jehož středu jsou blíže. Algoritmus končí, když se shluky přestávají měnit nebo při dosažení maximálního povoleného počtu iterací. Hodnoty, které patří do shluku s menším centroidem jsou označeny jako 0 a hodnoty shluku s větším centroidem jako 1. Navíc je možné nastavit počet opakování, kolikrát se bude algoritmus opakovat, přičemž výsledek je získán jako konsenzus získaný opakováním pro každou časovou řadu zvlášť. Ukázka diskretizace při 100 opakování a maximální počet iterací 250 pro FRS2 je v Tab. 4.9.

Tab. 4.9: Příklady diskretizace časových řad RFS2 podle k-means

	ctrl 1	10′	30′	1 h	2 h	4 h	6 h	8 h	10 h	12 h	ctrl 2	ctrl 3	práh
CCTL14 I	0	0	1	1	1	1	0	0	0	0	0	0	12312
hESC I C	0	1	1	1	1	1	1	1	1	1	0	0	26737
RCS II H	0	1	1	1	1	1	1	0	1	1	0	0	39352

Výsledky pro jednotlivé časové řady se mohou výrazně lišit, jak je vidět z tabulky výše. Nevýhodou této implementace k-means algoritmu je práce 1D daty, kdy je každá časová řada brána zvlášť. Pro potřeby Western blot dat a opakované řešení by bylo lepší brát každý časový bod jako multidimenzionální, kde je počet dimenzí určen počtem opakování

experimentu. Takový algoritmus by pak vedl k jedné diskretizované řadě dat.

Detektor hran

Druhým algoritmem, který funkce implementuje, je binarizace dat založená na hranovém detektoru. Jedná se přitom o detektor navržený pro microarray data [38]. V prvním kroku jsou všechna X_i měření časové řady seřazena vzestupně. Následně jsou zkoumány diference mezi sousedícími hodnotami, zda splňují nastavenou podmínku.

Detektor „firstEdge" porovnává tyto diference s průměrným gradientem časové řady. První diference, která tento gradient převyšuje je označena jako hrana. Nižší z hodnot, ze kterých byla diference vypočítána, pak určuje maximální hodnotu pro skupinu 0 a vyšší z hodnot minimální hodnotu pro skupinu 1.

Detektor „maxEdge" hledá mezi diferencemi nejvyšší hodnotu. Skupiny jsou pak vytvořeny podle hodnot, ze kterých byla diference získána, stejným způsobem jako u „firstEdge" detektoru. Ukázky diskretizací pro oba typy detektorů ukazuje Tab. 4.10.

Tab. 4.10: Příklady diskretizace časových řad FRS2 podle detektoru hran

	det	ctrl1	10′	30′	1 h	2 h	4 h	6 h	8 h	10 h	12 h	ctrl2	ctrl3	práh
hESC I C	first	0	1	1	1	1	1	1	1	1	1	0	0	16479
hESC I C	max	0	1	1	1	1	1	1	1	1	1	0	0	16479
RCS I H	first	0	1	1	1	1	1	1	1	0	1	0	0	10746
RCS I H	max	0	0	1	1	1	1	0	0	0	0	0	0	47494

Ačkoliv na testovacích datech poskytují oba detektory většinou stejný výsledek, v tabulce výše je vidět, že pro některé časové řady se výsledek obou algoritmů může lišit.

Scan statistika

Třetí z algoritmů implementovaných ve funkci je binarizace pomocí scan statistics [18]. Jedná se o metodu, která zjišťuje, zda jsou měření pro jednotlivé časové řady uniformně a nezávisle rozložena na celém rozsahu časové řady. K tomu využívá plovoucího okna, ve kterém se statistickým testem rozhoduje, jestli je tato podmínka splněna nebo ne. Volbu velikosti

okna lze nastavit na vstupu algoritmu. Přitom si algoritmus pamatuje okno s nejmenší *p* hodnotou. Vstupním parametrem tohoto algoritmu je navíc i požadovaná hladina významnosti. Pokud je na zvolené hladině významnosti podmínka testu splněna, jsou hranicemi okna definovány prahy definující konec jedné a začátek druhé skupiny.

Oproti předchozím dvěma algoritmům umožňuje scan statistics i filtraci málo kvalitních měření právě nastavením požadované hladiny významnosti. Pokud této není dosaženo ani u okna s nejnižší *p* hodnotou, je binarizace sice provedena, ale zároveň algoritmus upozorní, že pro tento výsledek není splněna nastavená hladina významnosti. Taková měření mohou být z dalších analýz vypuštěna. Ukázky binarizace FRS2 pro relativní délku okna w a hladinu významnosti α jsou shrnuty v Tab. 4.11.

Tab. 4.11: Příklady diskretizace časových řad FRS2 podle scan statistiky

	w/α	ctrl 1	10′	30′	1 h	2 h	4 h	6 h	8 h	10 h	12 h	ctrl 2	ctrl 3	práh
hESC IIH	0,1/0,5	0	1	1	1	1	1	1	1	1	1	0	0	6478
hESC IIH	0,2/0,05	0	1	1	1	1	1	1	1	1	1	0	0	6478
RCS I C	0,1/0,5	0	0	1	1	1	1	1	1	0	1	0	0	10746
RCS I C	0,2/0,05	0	0	1	1	1	1	1	1	0	0	0	0	22120

Z uvedených výsledků byla testová statistika splněna pouze pro časovou řadu RCS I C s délkou okna w=0,1 na hladině významnosti α=0,5. To ukazuje především na fakt, že tento statistický test není optimalizován pro použití na Western blot data, protože na této velmi benevolentní hladině významnosti nebyla testová statistika splněna pro druhou zmíněnou časovou řadu. Jinak jsou výsledky binarizace vcelku dobré, tedy s úpravou podmínek hodnocení testu je možné tuto techniku použít i na Western blot data.

Konsenzuální výsledky

Pro použití na Western blot data s opakovaným měřením by bylo vhodné dosáhnout konsenzuálních výsledků pro jednotlivé techniky. K tomu lze využít funkci *get.consensus* z přiloženého balíčku funkcí. Vstupem funkce je matice binarizovaných časových řad, výstupem pak konsenzuální výsledek

napříč měřeními. Ukázky s uvedenými nastavenými parametry v závorce jsou zobrazeny v Tab. 4.12 a Tab. 4.13.

Tab. 4.12: Příklady konsenzuální výsledků diskretizace časových řad FRS2 dle nástrojů BoolNet

	ctrl 1	10′	30′	1 h	2 h	4 h	6 h	8 h	10h	12h	ctrl 2	ctrl 3
kmeans(100/250)	0	0	1	1	1	1	1	0	0	0	0	0
firstEdge	0	1	1	1	1	1	1	1	1	1	0	0
maxEdge	0	1	1	1	1	1	1	0	0	0	0	0
scan (0,1/0,5)	0	1	1	1	1	1	1	1	0	0	0	0

Tab. 4.13: Příklady konsenzuální výsledků diskretizace časových řad ERK dle nástrojů BoolNet

	ctrl 1	10′	30′	1 h	2 h	4 h	6 h	8 h	10h	12h	ctrl 2	ctrl 3
kmeans(100/250)	0	1	1	1	1	0	0	0	0	0	0	0
firstEdge	0	1	1	1	1	1	1	1	0	0	0	0
maxEdge	0	0	1	1	1	0	0	0	0	0	0	0
scan (0,1/0,5)	0	1	1	1	1	1	1	1	0	0	0	0

Nejlepších výsledků dosáhla binarizace založená na hranovém detektoru zachycujícím první hranu. Pro časové řady FRS2 a ERK ale neposkytuje stejný výsledek, dle očekávání experimentátorů. To však může ukazovat i na zatím neobjevený vztah těchto dvou komponent signální dráhy FGFR3.

4.4 Infotheo

Infotheo je další z balíčků jazyka R, který obsahuje nástroje na diskretizaci dat. Jedná se o balíček funkcí primárně určených pro aplikaci teorie informace na měření informačního obsahu dat pomocí entropie. Balíček však obsahuje i funkci na diskretizaci proměnných. I když není balíček primárně zaměřen na systémovou biologii, jeho autorem je Patrick E. Meyer, jehož disertační práce pojednává o zpracování microaray dat [30]. Zvláště funkce *discretize* je pak založena na poznatcích z této jeho práce. Protože je jedná o obecnou diskretizační funkci, nejsou předem nastaveny výstupní stavy na 0 a 1, ale je potřeba přímo specifikovat pro kolik

diskrétních stavů má transformace proběhnout. Ty jsou potom číslovány od 1 do n, kde n je zvolený počet stavů. Jako proměnnou, tj. časovou řadu, chápe funkce sloupec matice. Tabulky měření, které jsme doposud používali, je tak potřeba na vstup funkce poslat transponované. Celkem nabízí funkce 3 různé techniky na diskretizaci.

Shodné frekvence

Prvním z algoritmů, které funkce nabízí, je rozdělení dle principu „equal frequency". Vstupní časová řada X_i jejíž rozsah hodnot je definován intervalem $X_i \in \langle a, b \rangle$, kde a je nejnižší a b je nejvyšší naměřená hodnota, je rozdělena do $|\chi_i|$ zvoleného počtu skupin tak, že každá skupina obsahuje právě stejný počet $m/|\chi_i|$ naměřených hodnot. Rozsah hodnot pro každý z těchto intervalů tak může být odlišný.

Shodné délky

Dalším algoritmem je rozdělení na intervaly stejné délky neboli „equal width" algoritmus [30], který interval $\langle a, b \rangle$ rozděluje na $|\chi_i|$ zvolený počet subintervalů stejné délky dle principu:

$$\left\langle a, a + \frac{b-a}{|\chi_i|} \right), \left\langle a + \frac{b-a}{|\chi_i|}, a + 2\frac{b-a}{|\chi_i|} \right), \ldots, \left\langle a + (|\chi_i| - 1)\frac{b-a}{|\chi_i|},\ b + \varepsilon \right), \quad (4.6)$$

kde $\varepsilon > 0$ je konstanta přidaná do posledního intervalu proto, aby i nejvyšší hodnota b byla zahrnuta do tohoto intervalu.

Globální shodné délky

Poslední možností je rozdělení do intervalů s využitím principu „global equal width" [49]. Pravidla pro určení hranic jednotlivých subintervalů jsou tak stejná jako pro princip stejné délky (4.6). Rozdíl je však v rozsahu původního intervalu $\langle a, b \rangle$, kde pro rozdělení na základě globální shodné délky je a nejnižší hodnotou a b nejvyšší hodnotou ze všech použitých časových řad, a ne pouze z hodnot pro aktuální časovou řadu, pro niž se diskretizace provádí.

S tím jaké jsou rozdíly v oboru hodnot pro jednotlivé časové řady je použití této techniky na opakovaná měření Western blot nevhodné. Vždyť jen přidání heparinu při detekci zvolených proteinů výrazně zvyšuje *IOD* naměřená při kvantifikaci blotů.

Konsenzuální výsledky

Pro udržení rozumné míry ukázaných příkladů diskretizaci si uvedeme pouze konsenzuální výsledky, které nás zajímají více než dílčí diskretizace. Příklad diskretizace FRS2 do dvou skupin ukazuje Tab. 4.14

Tab. 4.14: Příklady konsenzuální výsledků diskretizace časových řad FRS2 dle infotheo

	ctrl 1	10′	30′	1 h	2 h	4 h	6 h	8 h	10h	12h	ctrl 2	ctrl 3
equalfreq	1	2	2	2	2	2	2	1	1	1	1	1
equalwidth	1	1	2	2	2	2	1	1	1	1	1	1
globalwidth	1	1	1	2	2	1	1	1	1	1	1	1

První z metod má velmi omezené použití, protože při diskretizaci do 2 úrovní rozděluje data vždy na 2 stejně velké skupiny, což není vhodné. Ani další metody nedosahují očekávaných výsledků. Pro zpracování ERK zvolíme pro ukázku v Tab. 4.15 diskretizaci do 3 stavů, konsenzuální zpracování je opět možné s využitím funkce *get.consensus*.

Tab. 4.15: Příklady konsenzuální výsledků diskretizace časových řad ERK dle infotheo

	ctrl 1	10′	30′	1 h	2 h	4 h	6 h	8 h	10h	12h	ctrl 2	ctrl 3
equalfreq	1	3	3	3	3	2	2	2	2	1	1	1
equalwidth	1	2	3	3	3	1	1	1	1	1	1	1
globalwidth	1	1	1	1	1	1	1	1	1	1	1	1

Pro diskretizaci do 3 úrovní ani jeden z výsledků neodpovídá očekávaným hodnotám, kdy ve stavu 1 by měla být pouze kontrolní měření. Pro algoritmus globálních délek jsou dokonce všechna měření zařazena do stavu 1, což ukazuje na fakt, že mezi měřeními je časová řada s výrazně vyššími hodnotami, která při klasifikaci do diskrétních skupin negativně ovlivňuje transformaci.

Zmíněné algoritmy jsou často využívání při strojovém učení a dobývání znalostí z dat. Patří mezi tzv. algoritmy bez učitele (unsupervised algorithms).

4.5 Discretization

Dalším souborem nástrojů pro diskretizaci je opět balíček funkcí v jazyce R nazvaný „discretization“. Balíček je zaměřen výhradně na implementaci několika různých obecných diskretizačních algoritmů, nenabízí žádné další funkce. 22 funkcí, které balíček obsahuje, implementuje 8 hlavních algoritmů pro diskretizaci. Těmito jsou Ameva, CACC (class-attribute contingenty coefficient), CAIM (class-attribute independence maximization), MDLP (minimum description length principle) a 4 algoritmy založené na Chi kvadrát testu. Tyto algoritmy se řadí mezi tzv, algoritmy s učitelem (supervised algoritms). Většina funkcí balíčku je použita v rámci funkcí hlavních pro pomocné výpočty.

Jako funkce balíčku infotheo, i funkce balíčku discretization předpokládají na vstupu tabulku objektů s různými naměřenými proměnnými, které mají být diskretizovány. Proto je nutné na vstup poslat naše testovací tabulky opět transponované. Na rozdíl od balíčku infotheo se zde projevuje fakt, že tyto algoritmy nejsou vhodné na diskretizaci časových dat. Na vstupu není možné určit počet hladin, do kterých mají být data diskretizována, funkce se snaží přímo najít jejich optimální počet. To může být vhodné pro diskretizaci různých proměnných, jejichž hodnoty jsou naměřeny u několika zkoumaných objektů, nikoliv pro časový vývoj u jediného objektu. Díky tomu mají tyto funkce tendenci volit vyšší počet hladin, aby data po diskretizaci lépe odpovídala původní časové řadě. Taková diskretizace je nesmyslná, protože reálně je počet hladin určený modelem, vůči kterému chceme data srovnávat. Pro ukázku si však jednotlivé diskretizace ukážeme.

Diskretizace shora dolů

Zmíněné algoritmy lze rozdělit do 2 skupin podle toho, jak diskretizace probíhá. První skupinou jsou algoritmy založené na principu ze shora dolů, kdy dochází k divizivnímu rozdělování kvantitativních dat do jednotlivých kvalitativních skupin. Mezi tyto algoritmy řadíme CAIM, CACC a Amevu.

Učitelem pro CAIM algoritmus [25] je vyhodnocení závislosti mezi třídou a hodnotou atributu, přičemž se snaží o dosažení co nejvyšší nezávislosti. Má tendenci rozdělovat data do méně intervalů než ostatní algoritmy. Algoritmus CACC [45] používá jako učitele kontingenční koeficient tabulek, které sestavuje při rozdělování dat. Oproti CAIM se nesnaží o minimalizaci skupin, ale o lepší pochopení dat. Princip algoritmu Ameva [19] je velmi podobný algoritmu CACC. Je také založen na maximalizaci kontingenčního koeficientu se zaměřením na minimální počet intervalů, do kterých data rozdělí. Příklady diskretizací pro FRS2 buněčnou, linii hESC, měření I C ukazuje Tab. 4.16.

Tab. 4.16: Příklady výsledků pro top-down algoritmy

	ctrl 1	10′	30′	1 h	2 h	4 h	6 h	8 h	10h	12h	ctrl 2	ctrl 3
CAIM	1	4	9	10	8	12	11	7	6	5	3	2
CACC	1	4	9	10	8	12	11	7	6	5	3	2
Ameva	1	4	9	10	8	12	11	7	6	5	3	2

Mezi algoritmy není ve výsledku rozdíl. Bohužel všechny rozdělily 12 měření do 12 skupin, což ukazuje, že algoritmy nejsou vhodné pro použití na diskretizaci časových řad z Western blot experimentů.

Diskretizace zdola nahoru

Druhou skupinou jsou algoritmy založené na opačném principu, tedy zdola nahoru, jedná se tedy o aglomerativní postup. Mezi tyto patří MDLP, Chi², ChiMerge, Extended Chi² a Modified Chi².

Algoritmus MDLP [14] je heuristickým algoritmem založeným na minimalizaci entropie. Jako všechny předcházející i tento odhaduje počet diskrétních skupin bez možnosti nastavení. Algoritmy založené na Chí kvadrát statistice v prvním kroku seřadí data vzestupně. Následně počítají Chi² test pro dvojice po sobě následujících hodnot. Chi² [26] algoritmus přitom dvojice spojuje, dokud test vyhovuje nastavené hladině významnosti, jeho vylepšení pak přidávají ještě další podmínky kontroly konzistence skupin. ChiMerge [23] pracuje obdobně s frekvencemi výskytu v jednotlivých třídách. Příklady diskretizací pro FRS2 buněčnou, linii hESC, měření I C ukazuje Tab. 4.17.

Tab. 4.17: Příklady výsledků pro down-top algoritmy

	ctrl 1	10′	30′	1 h	2 h	4 h	6 h	8 h	10h	12h	ctrl 2	ctrl 3
MDLP	1	2	3	4	3	4	4	3	2	2	1	1
Chi²	1	1	2	2	1	2	2	1	1	1	1	1
ChiMerge	1	1	2	2	1	2	2	1	1	1	1	1
Extended Chi²	1	1	2	2	1	2	2	1	1	1	1	1
Modified Chi²	1	1	1	1	1	1	1	1	1	1	1	1

Výsledky down-top algoritmů jsou mnohem lepší. I když u algoritmu MDLP nelze ovlivnit počet skupin, do kterých má data rozdělit, po diskretizaci mají data 4 úrovně, přičemž pouze kontrolní měření jsou v nejnižším stavu 1. Chi kvadrát algoritmy umožňují ovlivnit počet diskrétních hladin pomocí parametru α, ten byl použit při hodnotě 0,5. Až na modifikovaný algoritmus, poskytují všechny ostatní shodný výsledek, kdy byla data rozdělena do dvou skupin, diskretizace ale není ideální ve srovnání s očekávanými hodnotami.

4.6 Další algoritmy a nástroje

R je velmi často využívaným jazykem pro implementaci algoritmů v systémové biologii a strojovém učení, proto je přehled zaměřen především na nástroje napsané v R. Samozřejmě existují i jiné nástroje napsané v jiných jazycích, či zcela jiné diskretizační techniky. Tyto si však uvedeme již zkráceně z několika důvodů.

Dalším často využívaným jazykem pro systémovou biologii je Matlab. Jeho nevýhodou je, že se jedná o nástroj placený. I tak lze ale napsaný kód zveřejňovat na (http://www.mathworks.com/matlabcentral/fileexchange/) Matlab exchange. Je tak možné najít diskretizační techniky i pro tento jazyk, jedná se však o implementace zde již zmíněných algoritmů. Výhodou Matlabu je potom možnost vytvářet aplikace s grafickým prostředím. Příkladem takového nástroje je WellReader [8]. Tento nástroj je však uzpůsoben přímo pro zpracování microarray dat a kvůli jeho charakteru spustitelné aplikace, nelze využívat jeho dílčí funkce.

Pro práci s boolovskými sítěmi lze najít i nástroje v jazyce Python, jako například TS2B [6] napsaný pro Python 2.7. Bohužel i tento nástroj je psaný

s minimálním využitím funkcionálního paradigmatu jako kompletní pipeline pro zpracování vstupních dat. Nelze tak využít pouze diskretizační nástroje. Na druhou stranu nástroj implementuje diskretizační techniky již zmíněné v předchozích kapitolách.

Dále je teoreticky možné pro diskretizaci využít další obecné techniky diskretizace využívané v technikách strojového učení, např. CLIP algoritmy [9] atd. Tyto techniky jsou ale velmi podobné již prezentovaným technikám top-down algoritmů s učitelem, které nejsou prakticky schopné zpracování dat z Western blot experimentů.

Dalším teoreticky vhodným nástrojem by mohl být SSD (short series discretization) [12]. Tato technika dle autorů umožňuje spolehlivou diskretizaci krátkých časových řad a je vhodná jak na zpracování microarray dat, tak na zpracování Western blot dat. Bohužel implementace této techniky, byť je v článku [12] uvedena, není dostupná a nástroj tak není možné využít. Jedná se matematicky propracovanou techniku, která bohužel ve zmiňovaném článku není přepsána do pseudokódu ani v nejdůležitějších částech. Není ji tak možné zařadit mezi fungující a využívané techniky tohoto přehledu.

Přímo pro diskretizaci Western blot dat tedy neexistuje žádný přímo uzpůsobený nástroj. Jak jsme si již vysvětlili, je to podmíněno především užitím Western blot dat pro kvantitativní studie, které diskretizaci nepotřebují. Techniky, které zpracují i Western blot data pak nemají možnost analyzovat více časových řad toho stejného měření pro diskretizaci do jediného výsledku. Kromě již prezentovaných konsenzuálních vyhodnocení se tak v následujících kapitolách ještě podívejme na nově navržené implementace některých statistických technik.

4.7 Multidimenzionální k-means

V kapitole 4.3 u diskretizace pomocí k-means jsme navrhli, že by algoritmus bylo možné použít i na více časových řad najednou. Každý časový okamžik by pak byl reprezentován ne jednou, ale více hodnotami, které byly získány při měření různých časových řad, tedy ne v 1D ale nD prostoru, kde n je počet časových řad. Takový algoritmus nebyl dosud pro

diskretizaci dat v systémové biologii využit. Jeho implementaci obsahuje funkce *kmeans.discretization* z přiloženého balíčku R funkcí. Vstupní parametry jsou stejné jako u obdobné funkce z balíčku BoolNet, kterou je tato funkce inspirovaná, tedy počet opakování a maximální počet iterací. Navíc je ještě možné nastavit, do jakého počtu diskrétních úrovní mají být data rozřazena. Ty jsou označeny číslicemi 1 až *l*, kde *l* je nastavený počet úrovní. Přitom skupiny s vyšší průměrnou IOD mají přiřazeny vyšší číslice. Příklad diskretizace pro 100 opakování při maximálním počtu iterací ukazuje Tab. 4.18.

Tab. 4.18: Příklady multidimenzionální k-means diskretizace FRS2

	1	**ctrl 1**	**10′**	**30′**	**1 h**	**2 h**	**4 h**	**6 h**	**8 h**	**10h**	**12h**	**ctrl 2**	**ctrl 3**
kmeans	2	1	1	2	2	2	2	2	1	1	1	1	1
kmeans	3	1	2	3	3	3	3	2	2	2	2	1	1
kmeans	4	1	2	4	4	4	4	3	2	2	2	1	1

Pro rozdělení do 2 úrovní poskytuje multidimenzionální k-means diskretizace stejný výsledek jako konsenzuální 1D k-means diskretizace realizované prostřednictvím balíčku BoolNet.

4.8 Diskretizace hierarchickým shlukováním

Stejně jako funguje diskretizace s využitím nehierarchického shlukování, lze definovat i diskretizaci založenou na hierarchickém shlukování. Je samozřejmě možné postavit mnoho různých algoritmů podle použitých technik výpočtu vzdálenosti a shlukovacích algoritmů. Funkce *hier.discretization* z přiloženého balíčku v sobě implementuje použití euklidovské vzdálenosti a shlukovacího algoritmu UPGMA (Unweighted Pair Group Method with Arithmetic Mean). Euklidovská vzdálenost d je definována jako:

$$d = \sqrt{\sum_{i=1}^{n} [x(i) - y(i)]^2}, \tag{4.7}$$

kde x a y jsou vektory naměřených hodnot IOD v daných okamžicích pro všech n měření časové řady.

Diskrétní hladiny jsou označeny číslicemi 1 až *l*, kde *l* je nastavený počet úrovní. Přitom skupiny s vyšší průměrnou IOD mají opět přiřazeny vyšší číslice.

Tab. 4.19: Příklady hierarchické diskretizace FRS2

	1	**ctrl 1**	**10′**	**30′**	**1 h**	**2 h**	**4 h**	**6 h**	**8 h**	**10h**	**12h**	**ctrl 2**	**ctrl 3**
hierar.	2	1	1	2	2	2	2	1	1	1	1	1	1
hierar.	3	1	2	3	3	3	3	2	2	2	2	1	1
hierar.	4	1	2	4	4	4	4	3	2	2	2	1	1

Výsledky, které tato metoda produkuje, jsou velmi podobné výsledkům k-means diskretizace, což není překvapující, neboť se jedná o příbuzné techniky.

5 Vyhodnocení

V poslední kapitole této práce se podíváme na srovnání jednotlivých metod. Už na první pohled lze z jednotlivých příkladů vidět, že se některé techniky mohou ve výsledcích velmi lišit. U diskretizace do 2 úrovní pak můžeme jednotlivé techniky srovnat s očekávanými daty pomocí statistických testů. Věrnost s jakou diskretizovaná data reprezentují data původní, lze odhalit pomocí neparametrické korelace. Nejčastější technikou srovnání, je porovnání s vybraným modelem. Protože se ale zabýváme PLA modely, které doposud pro zkoumání signálních drah nebyly použity, je toto srovnání problematické.

5.1 Shrnutí modelů

Jak přehled nástrojů pro diskretizaci ukázal, v současné době neexistuje nástroj přímo uzpůsobený na diskretizaci časových řad získaných Western blot experimenty, který by navíc uměl zpracovat více měření stejného proteinu v jednu diskrétní řadu dat. Je to zapříčiněno tím, že pro kvalitativní modely se využívají především microarray data, zatímco Western blot experimenty byly doposud využívány především pro kvantitativní modely, které diskretizaci nepotřebují. Nástroje pro zpracování microarray dat, které lze použít pro diskretizaci časových řad z Western blotů, pak umožňují převod pouze do dvou kvalitativních úrovní, neboť jsou uzpůsobeny pro práci s boolovskými modely. PLA modely totiž zatím nebyly tímto způsobem využity.

Úpravou stávajících technik však můžeme jednoduše implementovat nástroje pro diskretizaci na základě absolutní hodnoty. S využitím konsenzuálního výsledku je pak možné zpracovávat i více kvantitativních měření pro jeden kvalitativní výsledek. Obdobně pak můžeme upravit i techniky založené na shlukovacích algoritmech jako jsou k-means nebo UPGMA.

Bez úprav je také možné využít stávající nástroje pro zpracování microarray dat. Tyto nástroje jsou však limitované tím, že jsou vyhrazeny pro práci s boolovskými sítěmi a neumožňují tak rozhodnout o počtu

úrovní v diskrétních datech. Stejně tak je možné využít i standardní techniky pro diskretizace využívané v technikách strojového učení. Je ale potřeba počítat s tím, že tyto techniky nejsou uzpůsobeny na diskretizaci časových řad a výsledek tak nemusí dávat smysl.

5.2 Statistické srovnání

Protože od kolegů poskytujících laboratorní data víme, že všechna kromě kontrolních měření by měla být ve stavu sepnuto, můžeme diskretizace do dvou úrovní statisticky vyhodnotit. Můžeme vypočítat základní komponenty matice zmatení, tj. TP (true positives, skutečně pozitivní), TN (true negatives, skutečně negativní), FP (false positives, falešně pozitivní), FN (false negatives, falešně negativní) a z těch vypočítat přesnost (precision), pokrytí (recall) a celkovou správnost (accuracy):

$$PREC = \frac{TP}{TP+FP}, \tag{5.1}$$

$$REC = \frac{TP}{TP+FN}, \tag{5.2}$$

$$ACC = \frac{TP+TN}{TP+FP+TN+FN}. \tag{5.3}$$

Vyhodnocení vybraných algoritmů, které v přehledu vykazovaly použitelné výsledky, na všech naměřených datech ukazuje Tab. 5.1. Pro takové vyhodnocení je možné použít přiloženou funkci *stat.discretization*. Pro společné vyhodnocení FRS2 a ERK je vhodné data nejprve normalizovat.

Tab. 5.1: Vyhodnocení diskretizace

	parametry	TP	FP	TN	FN	PREC	REC	ACC
hodnota, průměr	-	119	0	66	79	1	0,601	0,701
hodnota, medián	-	132	0	66	66	1	0,667	0,75
hodnota, maximum	p=0,1	157	0	66	41	1	0,793	0,845
hodnota, percentil	p=0,2	186	11	55	12	0,944	0,939	0,913
BoolNet, 1D kmeans	100/250	110	0	66	88	1	0,556	0,667
BoolNet, edgeDetector	firstEdge	151	0	66	47	1	0,763	0,822
BoolNet, edgeDetector	maxEdge	107	0	66	91	1	0,54	0,655
BoolNet, scanStatistic	0,1/0,5	146	0	66	52	1	0,737	0,803
infotheo, equalfreq	-	88	33	33	110	0,727	0,444	0,458
infotheo, equalwidth	-	94	15	51	104	0,862	0,475	0,549

Nejlepší celkovou správnost vykazují algoritmy, u kterých je diskretizaci možné ovlivnit nějakým parametrem. Vůbec nejlépe tak dopadla diskretizace s volbou percentilu. Většina algoritmů má velmi dobrou přesnost, tedy schopnost správně predikovat nenulový stav, ovšem při velmi malém pokrytí, tj. schopnost predikovat všechny nenulové stavy. I v těchto parametrech dopadla nejlépe diskretizace percentilem. Dobré výsledky poskytují i funkce balíčku BoolNet, který je vyhrazen pro použití v systémové biologii. Obecné funkce z balíčku infotheo dopadly v hodnocení nejhůře, zaostávají ve všech hodnocených parametrech. Další obecné funkce nejsou v hodnocení zastoupeny, neboť jsou pro diskretizaci nevhodné, jak je patrné z ukázek v předchozích kapitolách.

V Tab. 5.2 je možné nalézt výsledky diskretizace pro konsenzuální zpracování FRS2 a ERK, stejně jako techniky, které zpracovávají naměřená data hromadně v jednu časovou řadu.

Tab. 5.2: Vyhodnocení diskretizace konsenzuálních technik

	parametry	TP	FP	TN	FN	PREC	REC	ACC
hodnota, průměr	-	10	0	6	8	1	0,556	0,667
hodnota, medián	-	12	0	6	6	1	0,667	0,75
hodnota, maximum	p=0,1	16	0	6	2	1	0,889	0,917
hodnota, percentil	p=0,2	18	0	6		1	1	1
BoolNet, 1D kmeans	100/250	9	0	6	9	1	0,5	0,625
BoolNet, edgeDetector	firstEdge	16	0	6	2	1	0,888	0,916
BoolNet, edgeDetector	maxEdge	9	0	6	9	1	0,5	0,625
BoolNet, scanStatistic	0,1/0,5	14	0	6	4	1	0,777	0,833
nD kmeans	100/250	100	0	6	8	1	0,555	0,667
hierarchické shluk.	-	8	0	6	10	1	0,444	0,583

Konsenzuální zpracování má velmi podobné statistické výsledky, jako zpracování po jednotlivých časových řadách. U technik s dobrými výsledky, pak může tyto ještě vylepšit. K nárůstu celkové správnosti došlo především u algoritmů z balíčku BoolNet, kdy například hranový detektor založený na detekci první hrany překročil hranici 0,9. Diskretizace percentilem dokonce dosáhla absolutní správnosti, dle výsledků očekávaných laboratorním specialistou. Techniky založené na shlukování, které zpracovává více časových řad v jeden výsledek, mají výsledky pouze průměrné. Musíme si však uvědomit, že toto hodnocení neříká nic o tom,

s jakou věrností jsou původní spojitá data reprezentována v nové diskrétní podobě. To můžeme zjistit korelací diskretizovaných časových řad s původními spojitými hodnotami.

5.3 Věrnost diskretizace

Protože srovnáváme diskrétní data a data spojitá, která mají naprosto rozdílný rozsah hodnot, není možné použít klasickou parametrickou korelaci. Je ale možné využít Spearmanovu pořadovou korelaci [41]. Přitom je potřeba korelovat výslednou diskrétní řadu se všemi naměřenými řadami spojitými. Zhodnocení věrnosti reprezentace časových řad FRS2 a ERK ukazují Tab. 5.3 a Tab. 5.4.

Tab. 5.3: Věrnost diskretizace FRS2 řad

	parametry	CCTL12 I	CCTL12 II	CCTL14 I	CCTL14 II	hESC I C	hESC II H	hESC III H	RCS I C	RCS II H	RCS III H
hodnota, průměr	-	0,86	0,86	0,86	0,86	0,86	0,86	0,86	0,81	0,81	0,86
hodnota, medián	-	0,87	0,87	0,82	0,87	0,72	0,87	0,87	0,72	0,72	0,87
hodnota, max	p=0,1	0,25	0,42	0,59	0,70	0,75	0,75	0,75	0,75	0,75	0,75
hodnota, percentil	p=0,2	0,25	0,42	0,59	0,70	0,75	0,75	0,75	0,75	0,75	0,75
BN, 1D kmeans	100/250	0,86	0,86	0,86	0,86	0,86	0,86	0,86	0,81	0,81	0,86
BN, edgeDetector	firstEdge	0,25	0,42	0,59	0,70	0,75	0,75	0,75	0,75	0,75	0,75
BN, edgeDetector	maxEdge	0,87	0,87	0,82	0,87	0,72	0,87	0,87	0,72	0,72	0,87
BN, scanStatistic	0,1/0,5	0,71	0,76	0,86	0,86	0,76	0,86	0,86	0,81	0,61	0,81
nD kmeans	100/250	0,86	0,86	0,86	0,86	0,86	0,86	0,86	0,81	0,81	0,86
hierarch. shluk.	-	0,82	0,82	0,82	0,77	0,67	0,82	0,82	0,82	0,82	0,82

Tab. 5.4: Věrnost diskretizace ERK řad

	param.	CCTL12 I	CCTL12 II	CCTL14 I	CCTL14 II	hESC I C	hESC II H	RCS I C	RCS II C	RCS III C	RCS IV H	RCS V H	RCS VI H
průměr	-	0,86	0,86	0,86	0,86	0,86	0,47	0,86	0,86	0,86	0,86	0,86	0,86
medián	-	0,87	0,87	0,87	0,87	0,87	0,63	0,87	0,68	0,87	0,87	0,87	0,82
maximum	p=0,1	0,86	0,86	0,86	0,86	0,76	0,76	0,81	0,71	0,86	0,86	0,86	0,86
percentil	p=0,2	0,75	0,75	0,75	0,75	0,75	0,75	0,42	0,47	0,64	0,75	0,75	0,75
1D kmeans	100/250	0,77	0,77	0,72	0,82	0,82	0,41	0,82	0,82	0,82	0,82	0,82	0,82
edgeDetect	firstEd.	0,86	0,86	0,86	0,86	0,76	0,76	0,81	0,71	0,86	0,86	0,86	0,86
edgeDetect	maxEd.	0,64	0,64	0,70	0,75	0,70	0,42	0,75	0,75	0,75	0,70	0,70	0,75
scanStat	0,1/0,5	0,86	0,86	0,86	0,86	0,76	0,76	0,81	0,71	0,86	0,86	0,86	0,86
nD kmeans	100/250	0,71	0,71	0,81	0,76	0,71	0,61	0,76	0,56	0,76	0,71	0,71	0,71
hierar. sh.	-	0,72	0,72	0,82	0,77	0,72	0,46	0,77	0,77	0,77	0,72	0,72	0,77

Z výsledků je patrné, že dobrá statistika vůči očekávaným hodnotám ještě nemusí znamenat nejvěrnější reprezentaci naměřených dat. U všech technik se podařilo dosáhnout poměrně vysokých korelací, které potvrzují silnou závislost mezi naměřenými a diskretizovanými daty. Avšak techniky, které měly horší statistické výsledky, jako například multidimenzionální k-means diskretizace, mohou data reprezentovat věrněji než techniky se silnou statistikou, jako například percentilová diskretizace. Je to dáno rozdílnou povahou těchto technik. Percentilová diskretizace je založena na splnění podmínky, bez toho aniž by brala v potaz celkový průběh časové řady. Diskretizace založená na shlukování se snaží řadu rozdělit do diskrétních hladin, aby byla co nejlépe interpretovatelná. Z toho lze usuzovat, že pro sestavení FGFR3 signální dráhy bude vhodnější sestavit víceúrovňový model, protože takový bude mít k naměřeným datům silnější korelaci. Tento fakt můžeme ověřit korelací pro vyšší počet diskrétních hladin, jak ukazují Tab. 5.5 a Tab. 5.6.

Tab. 5.5: Věrnost diskretizace FRS2 řad pro více úrovní

	hladin	CCTL12 I	CCTL12 II	CCTL14 I	CCTL14 II	hESC I C	hESC II H	hESC III H	RCS I C	RCS II H	RCS III H
nD kmeans	3	0,67	0,76	0,85	0,88	0,84	0,94	0,94	0,94	0,94	0,94
hierarch. shluk.	3	0,67	0,76	0,85	0,88	0,84	0,94	0,94	0,94	0,94	0,94
nD kmeans	4	0,73	0,81	0,88	0,91	0,90	0,96	0,96	0,94	0,94	0,96
hierarch. shluk.	4	0,73	0,81	0,88	0,91	0,90	0,96	0,96	0,94	0,94	0,96

Tab. 5.6: Věrnost diskretizace ERK řad pro více úrovní

	hladin	CCTL12 I	CCTL12 II	CCTL14 I	CCTL14 II	hESC I C	hESC II H	RCS I C	RCS II C	RCS III C	RCS IV H	RCS V H	RCS VI H
nD kmeans	3	0,87	0,87	0,94	0,91	0,87	0,70	0,73	0,76	0,85	0,87	0,87	0,91
hierar. sh.	3	0,87	0,87	0,94	0,91	0,87	0,70	0,73	0,76	0,85	0,87	0,87	0,91
nD kmeans	4	0,87	0,87	0,94	0,90	0,87	0,74	0,75	0,70	0,85	0,87	0,87	0,88
hierar. sh.	4	0,87	0,87	0,94	0,90	0,87	0,74	0,75	0,70	0,85	0,87	0,87	0,88

Výsledky tuto hypotézu potvrzují, neboť korelace pro 3 a 4 úrovně jsou silnější, než pro diskretizaci pouze do 2 úrovní. Celá signální dráha FGFR3

obsahuje více komponent, než doposud naměřené 2 proteiny. Lze tedy předpokládat, že diskrétních úrovní pro věrný PLA model bude potřeba více než 2. Je však sympatické, že tento fakt potvrzují už prozatím 2 naměřené komponenty této signální dráhy.

Pro diskretizaci Western blot dat tedy není možné najít všeobecně nejvhodnější nástroj. Je nutné rozhodnutí, zda je vhodnější co nejlépe splnit nastavenou podmínku nebo co nejvěrněji reprezentovat data. Diskretizace tak sama o sobě může poskytnout důležitou informaci o tom, s kolika diskrétními hladinami by měl model pracovat. Je však nutné vybrat takovou techniku diskretizace, která se snaží data co nejvěrněji reprezentovat. Takovými technikami mohou být diskretizace s využitím shlukovacích algoritmů, jak prokázaly výše uvedené výsledky.

5.4 Implementace v jazyce R

Jazyk R je v komunitě bioinformatiků a systémových biologů velmi oblíbený o čemž vypovídá i fakt, že většina technik z přehledu je napsána právě v tomto jazyce. Proto i techniky navržené v rámci této práce byly implementovány jako funkce tohoto jazyka. Z funkcí byl sestaven balíček v R (3.1.2) pod názvem WBdiscretization. Balíček lze naistalovat příkazem *install.packages* s parametrem *repos=NULL*. Funkčnost byla ověřena i pro starší verze R (2.1.x) pro Windows (7 64 bit) a Linux (Ubuntu 12.04 LTS). Instalace balíčku tedy nevyžaduje specifickou verzi R, v případě problémů lze ale doporučit využít právě verzi 3.1.2, kde lze garantovat bezproblémovost.

Balíček obsahuje 7 funkcí, jejichž popis a návod na použití je přiložen na konci této práce. Jednotlivé funkce již také byly zmíněny v příslušných kapitolách o technikách, které v sobě implementují. Funkce jsou postaveny s využitím základních nástrojů R a není tudíž vyžadována instalace žádných dodatečných balíčků.

6 Závěr

Cílem diplomové práce bylo vypracování přehledu existujících technik a nástrojů pro zpracování dat z Western blot experimentů ve smyslu diskretizace těchto dat a porovnání těchto technik při zpracování poskytnutých experimentálních dat FGFR signální dráhy. Dílčími úkoly tak bylo nastudování základů FGFR signální dráhy, popis experimentální techniky a zhodnocení kvality dat, vypracování rešerše použitelných technik diskretizace a jejich zhodnocení.

Ve druhé kapitole jsem se zaměřil na popis nejdůležitějších komponent FGFR signální dráhy, jejichž naměřená data jsem měl k dispozici. Krátce jsem také popsal rozdíly mezi jednotlivými technikami používanými pro modelování.

Hlavním zaměřením třetí kapitoly byl popis experimentálních technik používaných v systémové biologii se zaměřením na techniku Western blot. Rozborem laboratorního postupu a zhodnocením dostupných naměřených dat se mi podařilo zjistit, že tato technika není dostatečně přesná pro kvantitativní modelování, byť tak bývá hojně využívána.

Jádrem práce je kapitola čtvrtá pojednávající o jednotlivých technikách diskretizace. Protože Western blot data byla doposud využívána pro kvalitativní modelování pouze okrajově, techniky vyhrazené přímo pro jejich zpracování neexistují. Zaměřil jsem se tedy na obecné techniky diskretizace využívané v systémové biologii i mimo ni. Všechny dostupné nástroje jsem otestoval na experimentálních datech. Zaměřil jsem se i na techniky, pro které nebyly dostupné diskretizační nástroje. Takové jsem implementoval v jazyce R, který je v systémové biologii hojně využívaným. V závěru kapitoly jsem pak navrhl dvě vlastní techniky založené na shlukovacích algoritmech.

V poslední kapitole jsem se zaměřil na statistické vyhodnocení jednotlivých technik. Srovnáním výsledků diskretizace s očekávaným výsledkem autorů experimentálních dat jsem vyhodnotil rozdíly v úspěšnosti diskretizace jednotlivých technik. Navíc jsem diskrétní data

vyhodnotil z hlediska věrnosti reprezentace původně spojitých dat pomocí neparametrických korelací.

V rámci praktické části práce tak vznikl balíček funkcí WBdisretization pro jazyk R. Tyto funkce poskytují diskretizace spojitých dat pomocí několika technik. Jedná se o techniky nově navržené nebo takové, které nebyly pro Western blot data dostupné. Jedna z funkcí je pak určena pro statistické hodnocení kvality diskretizačních algoritmů.

Reference

[1] ALDRIDGE, B. B., J. M. BURKE, D. A. LAUFFENBURGER a P. K. SORGER. Physicochemical modelling of cell signalling pathways. *Nature Cell Biology*. 2006, vol. 8, issue 11, pp. 1195-1203. DOI: 10.1038/ncb1497.

[2] ANTONY, P. M. A., C. TREFOIS, A. STOJANOVIC, A. S., BAUMURATOV a K. KOZAK. Light microscopy applications in systems biology: opportunities and challenges. *Cell Communication and Signaling*. 2013, vol. 11, issue 1, pp. 24-. DOI: 10.1186/1478-811X-11-24.

[3] BATT, G., M. PAGE, I. CANTONE, G. GOESSLER, P. MONTEIRO a H. DE JONG. Efficient parameter search for qualitative models of regulatory networks using symbolic model checking. *Bioinformatics*. 2010-09-07, vol. 26, issue 18, i603-i610. DOI: 10.1093/bioinformatics/btq387.

[4] BAUJAT, G., L. LEGEAI-MALLET, G. FINIDORI, V. CORMIER-DAIRE a M. LE MERRER. Achondroplasia. *Best Practice*. 2008, vol. 22, issue 1, pp. 3-18. DOI: 10.1016/j.berh.2007.12.008.

[5] BELOV, A. A. a M. MOHAMMADI. Molecular Mechanisms of Fibroblast Growth Factor Signaling in Physiology and Pathology. *Cold Spring Harbor Perspectives in Biology*. 2013-06-03, vol. 5, issue 6, a015958-a015958. DOI: 10.1101/cshperspect.a015958.

[6] BERESTOVSKY, N., L. NAKHLEH a P.V. BENOS. An Evaluation of Methods for Inferring Boolean Networks from Time-Series Data. *PLoS ONE*. 2013-6-21, vol. 8, issue 6, e66031-. DOI: 10.1371/journal.pone.0066031.

[7] BOULTON, T. G. a M. H. COBB. Identification of multiple extracellular signal-regulated kinases (ERKs) with antipeptide antibodies. *Molecular Biology of the Cell*. 1991-05-01, vol. 2, issue 5, pp. 357-371. DOI: 10.1091/mbc.2.5.357

[8] BOYER, F., B. BESSON, G. BAPTIST, J. IZARD, C. PINEL, D. ROPERS, J. GEISELMANN a H. DE JONG. WellReader: a MATLAB program for the analysis of fluorescence and luminescence reporter gene data. *Bioinformatics*. 2010-04-23, vol. 26, issue 9, pp. 1262-1263. DOI: 10.1093/bioinformatics/btq016.

[9] CIOS K.J. a L. KURGAN. Hybrid Inductive Machine Learning: An Overview of CLIP Algorithms. *New Learning Paradigms in Soft Computing*, 2001, pp. 276-322

[10] COUTTS, J. C. a J. T. GALLAGHER. Receptors for fibroblast growth factors. *Immunology and Cell Biology*. 1995, vol. 73, issue 6, pp. 584-589. DOI: 10.1038/icb.1995.92

[11] CRICK, F. What mad pursuit: a personal view of scientific discover. New York : Basic Books, 1988. ISBN 0-465-09137-7.

[12] DIMITROVA, E.S., M.P.V. LICONA, J. MCGEE, R. LAUBENBACHER. Discretization of Time Series Data. *Journal of Computational Biology*. 2010, vol. 17, issue 6, pp. 853-868. DOI: 10.1089/cmb.2008.0023.

[13] DUNGL, P. *Ortopedie*. 1. vyd. Praha: Grada Publishing, 2005, 1273 s. ISBN 80-247-0550-8.

[14] FAYYAD, U. M. a K. B. IRANI. Multi-interval discretization of continuous-valued attributes for classification learning. *Artificial intelligence*, 1993 , 13 , 1022–1027

[15] FOLDYNOVA-TRANTIRKOVA, S., W. R. WILCOX a P. KREJCI. Sixteen years and counting: The current understanding of fibroblast growth factor receptor 3 (FGFR3) signaling in skeletal dysplasias. *Human Mutation*. 2012, vol. 33, issue 1, pp. 29-41. DOI: 10.1002/humu.21636.

[16] GARINI, Y., B. J., VERMOLEN a I. T. YOUNG. From micro to nano: recent advances in high-resolution microscopy. *Current Opinion in Biotechnology*. 2005, vol. 16, issue 1, pp. 3-12. DOI: 10.1016/j.copbio.2005.01.003.

[17] GASSMANN, M., B. GRENACHER, B. ROHDE a J. VOGEL. Quantifying Western blots: Pitfalls of densitometry. *ELECTROPHORESIS*. 2009, vol. 30, issue 11, pp. 1845-1855. DOI: 10.1002/elps.200800720.

[18] GLAZ J., J. NAUS a S. WALLENSTEIN. *Scan Statistics*. New York: Springer, 2001. ISBN 978-1-4419-3167-2.

[19] GONZALEZ-ABRIL, L., F.J. CUBEROS, F. VELASCO a J.A. ORTEGA. Ameva: An autonomous discretization algorithm. *Expert Systems with Applications*. 2009, vol. 36, issue 3, pp. 5327-5332. DOI: 10.1016/j.eswa.2008.06.063.

[20] HAEFNER, James W. *Modeling biological systems: principles and applications*. 2nd ed. New York: Springer, c2005, xvi, 475 p. ISBN 03-872-5012-3.

[21] HARTEMINK, A.J., D.K. GIFFORD, T.S. JAAKKOLA a R.A. YOUNG. Bayesian methods for elucidating genetic regulatory networks. *IEEE Intelligent Systems*. vol. 17, issue 2, pp. 37-43. DOI: 10.1109/5254.999218.

[22] HEINER, M., D. GILBERT a R. DONALDSON. Petri Nets for Systems and Synthetic Biology. *Formal Methods for Computational Systems Biology*. Berlin, Heidelberg: Springer Berlin Heidelberg, 2008, pp. 215. DOI: 10.1007/978-3-540-68894-5_7

[23] KERBER, R.. ChiMerge : Discretization of numeric attributes. *Proceedings of the Tenth National Conference on Artificial Intelligence*, 1992, 123–128

[24] KREJCI, P., B. MASRI, L. SALAZAR, C. FARRINGTON-ROCK, H. PRATS, L. M. THOMPSON a W. R. WILCOX. Bisindolylmaleimide I Suppresses Fibroblast Growth Factor-mediated Activation of Erk MAP Kinase in Chondrocytes by Preventing Shp2 Association with the Frs2 and Gab1 Adaptor Proteins. *Journal of Biological Chemistry*. 2007-01-26, vol. 282, issue 5, pp. 2929-2936. DOI: 10.1074/jbc.M606144200.

[25] KURGAN, L.A. a K.J. CIOS. CAIM discretization algorithm. *IEEE Transactions on Knowledge and Data Engineering*. 2004, vol. 16, issue 2, pp. 145-153. DOI: 10.1109/TKDE.2004.1269594.

[26] LIU, H. a R. SETIONO. Chi2: Feature selection and discretization of numeric attributes. *Tools with Artificial Intelligence*, 1995, 388–391

[27] MACKAY, David J. *Information theory, inference, and learning algorithms*. Cambridge: Cambridge University Press, 2003, xii, 628 s. ISBN 978-0-521-64298-9.

[28] MARTIN, S., Z. ZHANG, A. MARTINO a J.-L. FAULON. Boolean dynamics of genetic regulatory networks inferred from microarray time series data. *Bioinformatics*. 2007-04-24, vol. 23, issue 7, pp. 866-874. DOI: 10.1093/bioinformatics/btm021.

[29] MESTDAGH, P., P. VAN VLIERBERGHE, A. DE WEER, D. MUTH, F. WESTERMANN, F. SPELEMAN a J. VANDESOMPELE. A novel and universal method for microRNA RT-qPCR data normalization. *Genome Biology*. 2009, vol. 10, issue 6, R64-. DOI: 10.1186/gb-2009-10-6-r64.

[30] MEYER, P. E. *Information-Theoretic Variable Selection and Network Inference from Microarray Data*. PhD thesis of the Universite Libre de Bruxelles, 2008.

[31] MISHRA, Kaushal. Basic western blotting procedure in reference to HIV test. *MEDICAL microbiology* [online]. 2009 [cit. 2014-12-27]. Dostupné z: <http://meromicrobiology.blogspot.cz/2011/07/western-blotting-for-hiv-test.html>

[32] MU, FI, Katedra informačních technologií. *Přednášky z předmětu PB050*: *Modelování a predikce v systémové biologii*. David Šafránek, Brno, 2012.

[33] MUSSEL, C., M. HOPFENSITZ a H. A. KESTLER. BoolNet--an R package for generation, reconstruction and analysis of Boolean networks. *Bioinformatics*. 2010-05-07, vol. 26, issue 10, pp. 1378-1380. DOI: 10.1093/bioinformatics/btq124.

[34] NOBLE, D. *The music of life: biology beyond the genome*. New York: Oxford University Press, 2006, xiii, 153 p. ISBN 978-019-9295-739.

[35] OZSOLAK, F. a P. M. MILOS. RNA sequencing: advances, challenges and opportunities. *Nature Reviews Genetics*. 2010-12-30, vol. 12, issue 2, pp. 87-98. DOI: 10.1038/nrg2934.

[36] PALSSON, B. *Systems biology: properties of reconstructed networks*. Cambridge: Cambridge University Press, 2006, xii, 322 s. ISBN 978-0-521-85903-5.

[37] SAUER, U., M. HEINEMANN a N. ZAMBONI. GENETICS: Getting Closer to the Whole Picture. *Science*. 2007-04-27, vol. 316, issue 5824, pp. 550-551. DOI: 10.1126/science.1142502.

[38] SHMULEVICH, I. a W. ZHANG. Binary analysis and optimization-based normalization of gene expression data. *Bioinformatics*. 2002-04-01, vol. 18, issue 4, pp. 555-565. DOI: 10.1093/bioinformatics/18.4.555.

[39] SOUTHERN, E. M. Detection of specific sequences among DNA fragments separated by gel electrophoresis. *Journal of Molecular Biology*. 1975, vol. 98, issue 3, pp. 503-517. DOI: 10.1016/S0022-2836(75)80083-0.

[40] SPARKMAN, O. *Mass spectrometry desk reference*. 1st ed. Pittsburgh, Pa.: Global View Pub., c2000. ISBN 09-660-8132-3.

[41] SPEARMAN, C. The Proof and Measurement of Association between Two Things. *The American Journal of Psychology*. 1904, vol. 15, issue 1, pp. 72-. DOI: 10.2307/1412159.

[42] STOUGHTON, R. B. APPLICATIONS OF DNA MICROARRAYS IN BIOLOGY. *Annual Review of Biochemistry*. 2005, vol. 74, issue 1, pp. 53-82. DOI: 10.1146/annurev.biochem.74.082803.133212.

[43] THE HUMAN GENOME MANAGEMENT INFORMATION SYSTEM (HGMIS). About the Human Genome Project [online]. 23.7.2013. [cit. 2014-08-28]. Dostupné z: <http://web.ornl.gov/sci/techresources/Human_Genome/project/index.shtml>

[44] TOWBIN, H., T. STAEHELIN a J. GORDON. Electrophoretic transfer of proteins from polyacrylamide gels to nitrocellulose sheets: procedure and some applications. *Proceedings of the National Academy of Sciences*. 1979-09-15, vol. 76, issue 9, pp. 4350-4354. DOI: 10.1073/pnas.76.9.4350.

[45] TSAI, C.-J., C. -I. LEE a W.-P. YANG. A discretization algorithm based on Class-Attribute Contingency Coefficient. *Information Sciences*. 2008, vol. 178, issue 3, pp. 714-731. DOI: 10.1016/j.ins.2007.09.004.

[46] TSE-WEN, C. Binding of cells to matrixes of distinct antibodies coated on solid surface. *Journal of Immunological Methods*. 1983, vol. 65, 1-2, pp. 217-223. DOI: 10.1016/0022-1759(83)90318-6.

[47] WANEY, D. L., G. C. MCALISTER a J.J. COON. Decision tree–driven tandem mass spectrometry for shotgun proteomics. *Nature Methods*. 2008-10-19, vol. 5, issue 11, pp. 959-964. DOI: 10.1038/nmeth.1260.

[48] WANG, R-S., A. SAADATPOUR a R. ALBERT. Boolean modeling in systems biology: an overview of methodology and applications. *Physical Biology*. 2012-10-01, vol. 9, issue 5, pp. 055001-. DOI: 10.1088/1478-3975/9/5/055001.

[49] YANG, Y a G.I. WEBB. Discretization for naive-Bayes learning: managing discretization bias and variance. *Machine Learning*. 2009, vol. 74, issue 1, pp. 39-74. DOI: 10.1007/s10994-008-5083-5.

[50] ZI, Z. a E. KLIPP. SBML-PET: a Systems Biology Markup Language-based parameter estimation tool. *Bioinformatics*. 2006-10-24, vol. 22, issue 21, pp. 2704-2705. DOI: 10.1093/bioinformatics/btl443.

A. Obsah CD

K této práci je přiloženo CD s následujícím obsahem adresářů:

`src` – zdrojové kódy implementovaných technik

`doc` – elektronická verze práce v pdf souboru

`data` – zdrojová data FRS2 a ERK časových řad pro načtení do R (*read.table*)
(pro elektronickou verzi utajeno)

`R package` – instalační balíček pro jazyk R

B. Manuál k balíčku WBdiscretization

Instalaci balíčku je možné provést z lokálního umístění příkazem install.packages("WBdiscretization", repos=NULL, type="source"). Podrobnější nápověda k jednotlivým funkcím v angličtině je dostupná přímo v balíčku.

`data.norm(X)`
Provede normalizaci každého řádku vstupní matice (příp. data frame) X.

`abs.discretization(X, method, p)`
Provede diskretizaci každého řádku vstupní matice (příp. data frame) X dle absolutní hodnoty. Volba metody se provádí parametrem `method`:

- `mean` – (standardně) dle průměru časové řady
- `median` – dle mediánu časové řady
- `max` – dle maxima časové řady
- `percentile` – dle percentilu časové řady

Parametr `p` nastavuje percentil nebo procento maxima, standardně 0,5.

`cons.discretization(X, method, p)`
Provede diskretizaci všech řádků vstupní matice (příp. data frame) X do jedné konsenzuální diskrétní řady dle absolutní hodnoty. Volba metody se provádí parametrem `method`:

- `mean` – (standardně) dle průměru časové řady
- `median` – dle mediánu časové řady
- `max` – dle maxima časové řady
- `percentile` – dle percentilu časové řady

Parametr `p` nastavuje percentil nebo procento maxima, standardně 0,5.

`get.consensus(D)`
Ze vstupní matice diskrétních dat D vytvoří jedinou řadu konsenzuální.

`hier.discretization(X, l)`
Provede diskretizaci všech řádků vstupní matice (příp. data frame) X do jedné diskrétní řady dle principu hierarchického shlukování. Při výpočtu využívá euklidovské vzdálenosti a UPGMA metody shlukování. Parametr `l` nastavuje počet diskrétních úrovní, standardně 2.

`kmeans.discretization(X, l)`
Provede diskretizaci všech řádků vstupní matice (příp. data frame) X do jedné diskrétní řady dle k-means shlukování. Parametr `l` nastavuje počet diskrétních úrovní, standardně 2.

`stat.discretization(X, D, states)`
Vyhodnotí úspěšnost diskretizace diskrétní matice X vůči očekávaným výsledkům diskretizace v matici D. Na základě určení matice zmatení vypočte přesnost, pokrytí a celkovou správnost. Vyhodnocení je možné pouze pro binární diskretizace, obor diskrétních hodnot lze určit parametrem `states`, standardně 0 a 1.