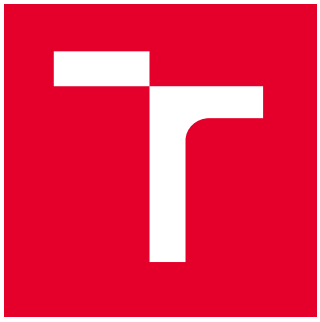


VYSOKÉ UČENÍ TECHNICKÉ V BRNĚ

BRNO UNIVERSITY OF TECHNOLOGY

FAKULTA ELEKTROTECHNIKY A KOMUNIKAČNÍCH TECHNOLOGIÍ

FACULTY OF ELECTRICAL ENGINEERING AND COMMUNICATION

ÚSTAV BIOMEDICÍNSKÉHO INŽENÝRSTVÍ

DEPARTMENT OF BIOMEDICAL ENGINEERING

VYHLEDÁVÁNÍ TRANSKRIPČNÍCH MOTIVŮ U NEMODELOVÝCH ORGANISMŮ

TRANSCRIPTION MOTIF FINDING IN NON-MODEL ORGANISMS

BAKALÁŘSKÁ PRÁCE

BACHELOR'S THESIS

AUTOR PRÁCE — **Klára Helešicová**

AUTHOR

VEDOUCÍ PRÁCE — **Ing. Jana Musilová**

SUPERVISOR

BRNO 2022

VYSOKÉ UČENÍ TECHNICKÉ V BRNĚ
FAKULTA ELEKTROTECHNIKY A KOMUNIKAČNÍCH TECHNOLOGIÍ

Bakalářská práce

bakalářský studijní program **Biomedicínská technika a bioinformatika**

Ústav biomedicínského inženýrství

Studentka: Klára Helešicová ***ID:*** 221508

Ročník: 3 ***Akademický rok:*** 2021/22

NÁZEV TÉMATU:

Vyhledávání transkripčních motivů u nemodelových organismů

POKYNY PRO VYPRACOVÁNÍ:

1) Vypracujte literární rešerši metod pro vyhledávání transkripčních motivů. Zaměřte se zejména na metody vyhledávání pomocí vzájemné informace. 2) Seznamte se s databázemi transkripčních motivů a vytvořte dataset sekvencí obsahujících transkripční motivy. 3) Navrhněte algoritmus pro vyhledávání transkripčních motivů a jeho dílčí části otestujte. 4) Implementujte navržený algoritmus v libovolném programovacím jazyce. 5) Algoritmus otestujte na vytvořeném datasetu. 6) Proveďte vyhodnocení a diskutujte výsledky.

DOPORUČENÁ LITERATURA:

[1] HASHIM, Fatma A, Mai S MABROUK a Walid AL-ATABANY. Review of Different Sequence Motif Finding Algorithms. Avicenna journal of medical biotechnology. 2019, 11(2), 130–148. ISSN 2008-2835.

[2] ELEMENTO, Olivier, Noam SLONIM a Saeed TAVAZOIE. A Universal Framework for Regulatory Element Discovery across All Genomes and Data Types. Molecular Cell. 2007, 28(2), 337–350. ISSN 10972765.

[3] INUKAI, Sachi, Kian Hong KOCK a Martha L BULYK. Transcription factor–DNA binding: beyond binding site motifs. Current Opinion in Genetics & Development. 2017, 43, 110–119. ISSN 0959437X.

Termín zadání: 7.2.2022 ***Termín odevzdání:*** 27.5.2022

Vedoucí práce: Ing. Jana Musilová

doc. Ing. Jana Kolářová, Ph.D.
předseda rady studijního programu

ABSTRAKT

Tato bakalářská práce se zabývá vyhledáváním transkripčních motivů u nemodelových organismů. V první části je vysvětlen proces transkripce, pojem vzájemná informace a algoritmus využívající vzájemnou informaci. V druhé části je popsáno rozdělení metod vyhledávající motivy a příklady algoritmů. Třetí část obsahuje přehled databází transkripčních motivů. Praktická část obsahuje popis vytvoření datasetu pro nemodelový organismus, popis navrženého algoritmu a jeho otestování na datasetu. Následně byly porovnány výsledky navrženého algoritmu s výsledky algoritmů FIRE a MEME.

KLÍČOVÁ SLOVA

DNA, transkripční motiv, nemodelový organismus, vzájemná informace, algoritmus

ABSTRACT

This bachelor thesis deals with the search for DNA motifs in non-model organisms. The first part explains the process of transcription, the concept of mutual information and an algorithm using mutual information. The second part describes the distribution of motif searching methods and examples of algorithms. The third part contains an overview of transcription motif databases. The practical part contains a description of the creation of a dataset for a non-model organism, a description of the proposed algorithm and its testing on the dataset. Then, the results of the proposed algorithm were compared with the results of FIRE and MEME algorithms.

KEYWORDS

DNA, DNA motif, non-model organism, mutual information, algorithm

HELEŠICOVÁ, Klára. *Vyhledávání transkripčních motivů u nemodelových organismů*. Brno: Vysoké učení technické v Brně, Fakulta elektrotechniky a komunikačních technologií, Ústav biomedicínského inženýrství, 2022, 43 s. Semestrální práce. Vedoucí práce: Ing. Jana Musilová

Prohlášení autora o původnosti díla

Jméno a příjmení autora:	Klára Helešicová
VUT ID autora:	221508
Typ práce:	Semestrální práce
Akademický rok:	2021/22
Téma závěrečné práce:	Vyhledávání transkripčních motivů u nemodelových organismů

Prohlašuji, že svou závěrečnou práci jsem vypracoval samostatně pod vedením vedoucí/ho závěrečné práce a s použitím odborné literatury a dalších informačních zdrojů, které jsou všechny citovány v práci a uvedeny v seznamu literatury na konci práce.

Jako autor uvedené závěrečné práce dále prohlašuji, že v souvislosti s vytvořením této závěrečné práce jsem neporušil autorská práva třetích osob, zejména jsem nezasáhl nedovoleným způsobem do cizích autorských práv osobnostních a/nebo majetkových a jsem si plně vědom následků porušení ustanovení § 11 a následujících autorského zákona č. 121/2000 Sb., o právu autorském, o právech souvisejících s právem autorským a o změně některých zákonů (autorský zákon), ve znění pozdějších předpisů, včetně možných trestněprávních důsledků vyplývajících z ustanovení části druhé, hlavy VI. díl 4 Trestního zákoníku č. 40/2009 Sb.

Brno ....

podpis autora*

*Autor podepisuje pouze v tištěné verzi.

PODĚKOVÁNÍ

Ráda bych poděkovala vedoucí bakalářské práce paní Ing. Janě Musilové za odborné vedení, konzultace, trpělivost a podnětné návrhy k práci. Taktéž bych ráda poděkovala celé své rodině za podporu nejenom během psaní bakalářské práce.

Obsah

Seznam obrázků

Úvod

Vědci již několik desetiletí zkoumají DNA, dědičnost, geny. V rámci poznávání genů a genové exprese se hledalo, na jakých úrovních v buňce se exprese genů ovlivňuje. Bylo zjištěno, že důležitá část regulace exprese se děje na úrovni transkripce, kde se na ovlivňování podílí transkripční faktory vážící se na transkripční motivy. Ty ovlivňují zásadním způsobem míru exprese daného genu. Právě objevování motivů DNA je základním krokem v mnoha systémech pro studium funkce genů.

V minulosti se vynalezlo několik algortimů na vyhledávání transkripčních motivů. Algoritmy byly stavěny na různých přístupech, ať už pravděpodobnostních, enumerativních či inspirovaných přírodou. Stále se vyvíjejí nové algoritmy a mnohé již existující se vylepšují. Pro spoustu hlavně modelových organismů je znalost transkripčních motivů velice dobrá, avšak stále mnohé organismy, obzvlášť nemodelové, nemají tak kvalitně zmapované motivy.

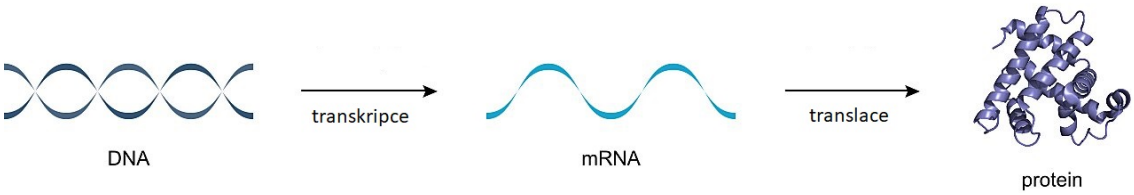
Cíl teoretické části je seznámení čtenáře s DNA, transkripcí, regulací genové exprese a vysvětlení vzájemné informace a relativní entropie. V další části jsou popsány přístupy vyhledávání transkripčních motivů a příklady algoritmů. V poslední části je uvedeno a popsáno několik databází transkripčních motivů.

Praktická část obsahuje popis vytoření testovacího datasetu pro nemodelový organismus. V následující části je popsány části kódu tvořící navržený algoritmus v programovém prostředí Python. Poslední část se věnuje porovnání výsledků navrženého algoritmu, algoritmů FIRE a MEME pro vytvořený dataset. Výsledky jsou porovnávány pomocí z-skóre a kontingenční tabulky.

1 Teoretický úvod

1.1 Od DNA k bílkovinám

Genetický kód je uchováván v jádře ve formě DNA. Ta se skládá ze 4 typů nukleotidů - adenin (A), cytosin (C), guanin (G) a thymin (T). Přes vzájemné navázání vodíkovými můstky a cukernými fosfáty se dohromady spojují do pravotočivé dvojšroubovice DNA. Části genomu, které obsahují informaci o tvorbě proteinu, se nazývají geny. Ty jsou aktivované transkripčními faktory, díky kterým různé buňky produkují rozdílné proteiny. Centrální dogma molekulární biologie definuje pořadí procesů v buňce v tomto pořadí - DNA se přepisuje do RNA a to následně na bílkovinu. V následujících odstavcích jsou detailněji popsané kroky přepisu DNA do proteinu. Při přenosu genetické informace - replikaci, se zajišťuje přenos z DNA do DNA (případně z RNA do RNA). Tento případ zde však nebude blíže popisován. [1] [2]



Obr. 1.1: Znázornění centrálního dogma molekulární biologie [3]

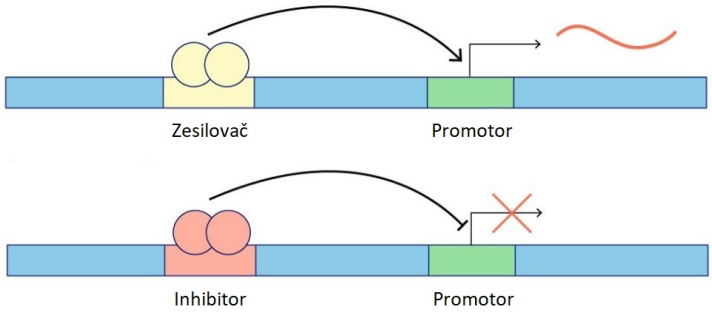
1.1.1 Transkripce

Transkripce je krok, kdy se přenáší informace z DNA do mRNA (mediátorová RNA). Transkripce začíná částečným rozpletením dvojšroubovice pomocí RNA polymeráz. Přepis se uskutečňuje na jednom vlákně DNA, kde dochází k párování bází podle Watson-Cricka (adenin se nepáruje s thyminem, ale s uracilem). RNA polymeráza spojuje nukleotidy do pre-mRNA do doby, než narazí na terminální sekvenci. Následně se z pre-mRNA, ve které se nachází kódující introny i nekódující exony, vystřihnou exony, proběhne metylace a polyadenylace.

Regulace transkripce je jednou z klíčových součástí buněčné aktivity. Díky transkripční aktivitě je buňka schopna ovládat genovou expresi. Z tohoto důvodu můžeme u rozdílně diferencovaných buněk, i přestože mají stejnou genetickou informaci, vidět rozdílnou aktivitu v exprimaci pouze části kódovaných genů.

Regulace exprese je řízena na mnoha úrovních, avšak jeden z nejdůležitějších kroků je regulace při transkripci. Kódující i nekódující geny obsahují promotorové a

enhancerové sekvence, na které se vážou zesilovače a inhibitory transkripce - transkripční faktory (TF). TF jsou nejčastěji proteiny, které se vážou na místa v DNA před nebo za exprimujícím se genem. Tato místa se nazývají transkripční motivy. Po navázání TF a obsazení aktivátorů transkripce se na promotoru rekrutuje transkripční iniciační aparát, který se skládá z RNA polymeráz a více než 50 dalších komponent. Následuje vyčištění promotoru, elongace a ukončení transkripce. Na zvyšování celkové hladiny genových transkriptů v buňce se podílejí i další složky posttranskripčního aparátu. Patří mezi ně regulovaný rozpad mRNA, stabilitou mRNA závislou na sekvenci a asociací mRNA s obecnými nebo specifickými RNA-vazebnými proteiny. [4]



Obr. 1.2: Transkripční faktory - zesilovače a inhibitory [5]

1.1.2 Translace

Translace je proces, při kterém se z mRNA vytváří bílkovina. mRNA vytvořená při transkripci je transportována z jádra do cytoplazmy. Translaci řídí ribozomy, které po vytvoření komplexu dokáží zahájit proces. Čtení RNA se děje po tripletech - kodonech, přičemž na začátku kaž́é transkripce je kodon methyonin a na konci stop kodon. Když mRNA prochází ribozomem, každý kodon interaguje s antikodonem specifické molekuly tRNA (transferová RNA), která nese aminokyselinu, jenž je začleněna do rostoucího proteinového řetězce. Po navázání aminokyseliny na řetězec se tRNA odpojuje z ribozomu.

1.2 Relativní entropie a vzájemná informace

Vzájemné informace (mutual information - MI) je míra informace, kterou jedna náhodná veličina obsahuje o druhé náhodné veličině. Její definování navazuje velice

těsně na relativní entropii, ze které je odvozena. [6]

Entropie náhodné veličiny je mírou neurčitosti náhodné veličiny. Je to míra množství informací, která je třeba k popisu náhodné veličiny. Na pojem entropie pak navazuje relativní entropie a vzájemná informace. Relativní entropie je mírou vzdálenosti mezi dvěma rozděleními. Ve statistice vzniká jako očekávaný logaritmus poměru pravděpodobnosti. Relativní entropie D(p||q) pravděpodobnostních hmotnostních funkcí p a q je mírou neefektivnosti předpokladu, že rozdělení je q, když skutečné rozdělení je p.

Relativní entropie je vždy nezáporná a je nulová pouze tehdy, když p = q, avšak není skutečnou vzdáleností mezi p a q, jak je tomu u entropie. Nyní zavedeme vzájemnou informaci, která je mírou vzájemné informace, kterou jedna náhodná veličina obsahuje o jiné náhodné veličině. Jedná se o snížení neurčitosti jedné náhodné veličiny v důsledku znalosti té druhé.

Uvažujme dvě náhodné veličiny X a Y se společnou pravděpodobnostní hmotnostní funkcí p(x, y) a okrajovými pravděpodobnostními hmotnostními funkcemi p(x) a p(y). Vzájemná informace I (X; Y) je relativní entropie mezi společným rozdělením a součinovým rozdělením p(x)p(y):

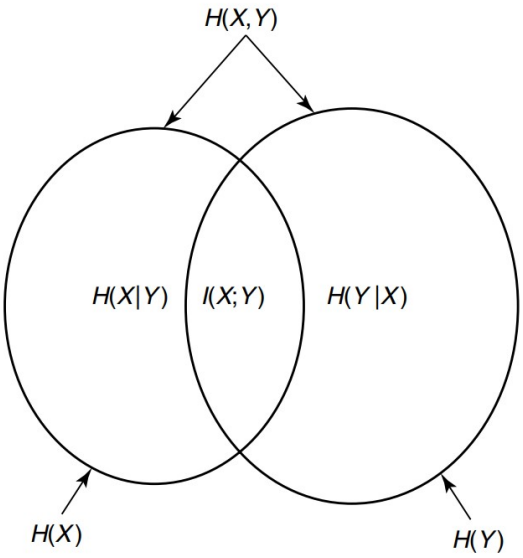
$$I(X;Y) = \sum_{x \in X} \sum_{y \in Y} p(x,y) log \frac{p(x,y)}{p(x)p(y)} = E_{p(x,y)} log \frac{p(X,Y)}{p(X)p(Y)} \tag{1.1}$$

Vzájemná informace náhodné veličiny se sebou samou je tedy entropie náhodné veličiny. To je důvod, proč se entropie někdy označuje jako informace o sobě samém.

Vztah mezi H (X), H (Y), H (X, Y), H (X|Y), H (Y|X), a I (X; Y)

$$I(X;Y) = H(X)H(X|Y) = H(Y)H(Y|X) \tag{1.2}$$

vzájemná informace I (X; Y) odpovídá průsečíku diagramů I (X; Y) a I (X; Y). informací v X s informacemi v Y. [7]



Obr. 1.3: Vennův diagram vztahu mezi vzájemnou informací a relativní entropií [7]

1.3 Z-skóre

Z-skóre udává počet směrodatných odchylek, o které je hodnota hrubého skóre vzdálena od střední hodnoty měřeného údaje či předpokladu. Nabývá kladných hodnot pro měřený údaj nad průměrem a záporných pod průměrem. Rozdílem populačního průměru od každého skóre a následným vydělením směrodatnou odchylkou získáme výpočet Z-skóre. Takovýto převod nazýváme standardizace a pojem standardní skóre označuje totéž co Z-skóre.

2 Metody vyhledávání transkripčních motivů

Vyhledávání transkripčních motivů je základním krokem v jejich určení. Jednotllivé metody reprezentují motivy buď jako konsenzuální řetezce nebo pozičně specifické váhové matice (PMW).V průběhu studia motivů se vyvinulo několik přístupů v samotném vyhledávání - proto dnes definujeme enumerativní přístup, pravděpodobnostní a přírodou inspirovaný přístup. Jednotlivé metody jsou využity pro návrh jednotlivých vyhledávacích algoritmů.

2.1 Enumerativní přístup

Tento přístup pracuje na bázi vyhledávání stejných sekvencí. Motivy se predikují na základě seznamu slov a podobností mezi slovy. Tyto algoritmy definují globální optimum - prohledávájí celý prostor všech možných kombinací (substitucí). Díky tomu jsou časově velice náročné, doba vyhledávání narůstá exponenciálně. Kvůli tomu nejsou vhodné pro dlouhé sekvence, nebo větší počet sekvencí. Tento nedostatek lze částečně eliminovat výběrem optimalizovaných datových struktur. Pro naše účely jsem vybrala několik nejznámějších algoritmů. [11]

2.1.1 Enumerace krátkých slov

Tato metoda je založena na výpočtu krátkých slov. Algoritmy, které můžeme zařadit do této kategorie, jsou DREME (Discriminative Regular Expression Motif Elicitation) a YMF (Yeast Motif Finder). [11]

DREME byl vytvořen především pro vyhledávání krátkých, neredundantních transkriptomických motivů u eukariot. DREME je součástí MEME Suite. [12] Jeho výhodou je rychlost a lineární časová náročnost, díky čemuž je vhodný pro využití na malých i velkých datasetech. Vstupními daty jsou dvě sady sekvencí DNA a práh významnosti. Algoritmus cyklicky vyhledává regulární výrazy motivů do doby, než některý z motivů nemá e-hodnotu nad práh významnosti. Pro nalezení statisticky významných motivů se v prvním kroku se vygenerují krátké k-mery a následně se ve dvou sadách sekvencí DNA vypočítá p-hodnota Fisherova exaktního testu motivu jako významnost relativního obohacení. Při překročení nastavené prahové hodnoty se daný motiv uloží a 100 nejvýznamějších motivů se označí jako počáteční regulární výrazy (regular expressions). Následně se použijí v paprskovém vyhledávání (beam search), které určuje jejich nejvýznamější zobecnění. U těchto výrazů se pak v každé iteraci zaměňuje jeden znak za znak zástupný (z IUPAC abecedy [13]). Díky těmto zástupným znakům se nevyhledávají stejná, ale heuristická slova. U takto modifikovaného slov se opět vypočte p-hodnota - pakliže přesáhne daný práh, slovo se uloží,

v opačném případě slovo smaže. Po vyčerpání všech původních slov se opět vybere 100 nejvýznamějších a cyklus se opakuje. [14]

YMF detekuje krátké, pozičně málo degenerované motivy u kvasinkového genomu pomocí konsenzuální reprezentace. Tento algoritmus enumeruje všechny motivy, u kterých vypočítává z-skóre. Z nich poté vybírá ty motivy s nejvyšším z-skórem.

2.1.2 Shlukovací metoda

Mezi algoritmy, které využívají shlukovací metody, řadíme algoritmus CisFinder. Byl navržen pro detekování krátkých motivů v krátkém časovém úseku. Jeho výstupem je seznam motivů se sekvencemi DNA popsaných pomocí pozičně frekvenčních matic. Matice jsou počítány přímo z n-mer slov, která můžou být s i bez mezer. Matice se následně doplní o mezery a spojí se dohromady (do shluku), aby vytvořili neredundantní sadu motivů. [11] [15]

2.1.3 Stromová metoda

Stromová metoda vznikla pro zrychlení enumerativního vyhledávání hlavně u velkých datasetů. Toho bylo dosaženo díky nezapočítávání podstatného počtu cyklů, při nichž se znovu prohledávají sekvencí ve stromu přípon (suffix tree). Jedním z algoritmů využívajících tuto metodu je Weeder. [11]

Výpočet algoritmu je založený na vzorcích počítání se specifickými a největšími neshodami. Motivy jsou na začátku reprezentovány jednak konsenzuálními sekvencemi, které jsou vytvořeny pomocí rozdílu mezi nimi a k-mery vstupních sekvencí, a taktéž daným počtem substitucí. Výsledné seřazené k-mery jsou seskupeny a každá skupina je hodnocena se specifickou mírou významnosti. Weeder byl navržen na vyhledávání dlouhých sekvencí. Dalšími algoritmy založenými na této metodě jsou FMotif a MCES, které využívají jak stromové metody, tak paralelního zpracování. [16] [17]

2.1.4 Metoda založená na teorii grafů

Grafově teoretická metoda představuje případ motivu jako podgrafu. Graf je vytvořen tak, že každý l-mer ve vstupních sekvencích reprezentuje vrchol. Hrana mezi vrcholy tvoří dvojice l-merů, které mají vzdálenosti mezi podřetězci menší nebo rovnu dvojnásobku maximálního počtu neshod. Poté se v tomto grafu hledají podgrafy o velikosti N. Algoritmy vážící se k této metodě jsou WINNOWER, Pruner a cWINNOWER. [11]

2.1.5 Metoda založená na hashovací tabulce

Tato metoda se zakládá na promítání l-merů ve vstupních datech do menšího prostoru pomocí hashování. Prvním krokem je vytvoření projekce l-rozměrného prostoru na k-rozměrný podprostor pro všechny podsekvence. Následuje vytvoření náhodná projekce pomocí výběru náhodných k pozic z l pozice. Pomocí této projekce je každý l-mer přiřazen do příslušného segmentu. Po promítnutí každý klíč obsahuje l-mer vyšší než prahovou hodnotu, což nazýváme kvalifikovaný klíč. Náhodné hashování se opakuje nkrát, aby se zajistilo, že se kvalifikovaný segmentnalezne alespoň jednou. Nakonec by měl být vypočítán profil pro každou z nich, aby se získal nejpravděpodobnější l-mer v sekvenci, která byla reprezentována jako konsenzuální sekvence. [11]

2.2 Pravděpodobnostní přístup

Algoritmy řadící se k pravděpodobnostnímu přístupu využívá při výpočtech pravděpodobnost. Do této skupiny můžeme zařadit deterministický, stochastický, pokročilý a další přístup, pod který spadají i algoritmy využívající výpočet vzájemné informace a entropie.

2.2.1 Deterministický přístup

Jednou z nejznámějsích metod, která spadá do deterministických přístupů, je metoda očekávání-maximalizace (EM). Výpočet EM se skládá ze dvou hlavních kroků - prvním je krok očekávání, který odhaduje hodnoty neznámé množiny parametrů, druhým krokem je krok maximalizace, který využívá hodnoty z prvního kroku k upřesnění parametrů v následných iteracích. EM se používá k identifikaci konzervovaných oblastí s předpokladem, že každá sekvence obsahuje alespoň jeden motiv. [11] [18]

Na této metodě bylo založeno několik algoritmů. Jedním z nejznámějších je MEME (Multiple EM for Motif Elicitation), ze kterého se postupně vyvinulo několik jeho verzí. Dalším je algoritmus STEME (Suffix Tree EM for Motif Elicitation), který vznikl na základě zrychlení MEME algoritmu (podobně jako Weeder pro zrychlení enumerativního přístupu). Posledním je EXTREME (Online EM algorithm for motif discovery), webový server, na kterém je implementován MEME algoritmus a který dokáže zpracovávat i velké datasety bez zapomínání sekvencí. [18] [19]

2.2.2 Stochastický přístup

Stochastický přístup stojí na Gibbsovu vzorkování, které je velice podobné EM metodě. Vylepšení výpočtu motivů pomocí Gibbsova vzorkování můžeme nalézt u algoritmů Align ACE a BioProspector. Align ACE uvažuje obě vlákna vstupní sekvence a nepovoluje překrývaní motivů. BioProspector používá Markovův model. Ten zlepšuje specifičnost motivu díky reprezentaci nemotivového pozadí. [11]

2.2.3 Pokročilý přístup

Tento přístup využívá Bayesovský přístup. Jedním z algoritmů založených na tomto přístupu je LOGOS (Integrated LOcal and GlObal motif sequence model), jenž je kombinací HMDM zarovnávající lokálně každý jednotlivý motiv a HMM zarovnávající globálně distribuci vícero motivů. [11]

2.2.4 Další přístup

Algoritmus FIRE (Finding Informative Regulatory Elements) je založený na výpočtu vzájemné informace. Díky ní lze zachytit závislosti (přítomnosti či nepřítomnosti motivů při přítomnosti jiného motivu atd.) bez jejich předchozí specifikace. Tento přístup se tím pádem může použít jak pro diskrétní, tak pro spojitá data. Existuje i odvozený algoritmus FIRE-pro, jenž je vytvořen pro vyhledávání motivů u proteinů. Oba algoritmy lze nalézt a použít po přihlášení na oficiálních stránkách. [8] [9]

Do výpočtu motivů vstupuje několik proměnných. Nutné jsou fasta soubor obsahující sekvence a expresní profil, který je definován vektor s N prvky. Každý prvek odpovídá genu a označuje přidruženou hodnotu exprese pro tento gen. Profil exprese může být diskrétní nebo spojitý. Profil motivu definujeme jako binární vektor s N prvky, kde pro každý gen „1“ znamená, že motiv je přítomen v odpovídajícím promotoru a „0“ znamená, že chybí. FIRE začíná zvažováním všech možných k-merů. Pro každý k-mer se spočítá vzájemná informace mezi profilem motivu (přítomností/nepřítomností) a expresním profilem. Následně se k-mery seřadí podle hodnoty MI, přičemž ty nejinformativnější jsou ponechány jako semena. Tato semena jsou následně optimalizována do obecnějších reprezentací motivů pomocí degenerovaného kódu. Optimalizace je změna množiny možných nukleotidů na určitých pozicí motivu, kdy se ponechávají pouze změny vedoucí k informativnějším motivům. Tento postup se opakuje tak dlouho, dokud nelze provést další zlepšení a motiv je tím pádem maximálně informativní. Nakonec každý motiv prochází rozsáhlými neparametrickými randomizačními testy, díky kterým jsou na výstupu pouze motivy se statisticky významnou informací. [10]

2.3 Metody inspirované přírodou

Tyto metody se zakládají na buď na chování organismů v přírodě nebo fyzikálních a chemických interakcích. Avšak společné mají inspirování se přírodou. Podle typu inspirace dělíme tyto metody na fyzikálně-chemické, inspirované včelím rojem a neinspirované včelím rojem. [11] [20]

2.3.1 Fyzikálně-chemické

Algoritmy spadající do této kategorie se inspirovaly určitými fyzikálními či chemickými jevy, jejichž činnost byla nastudována a na tomto základě bylo navrhnotu oním inspirované vyhledávání. Inspirace přicházela z mnoha oblastí, kupříkladu z teorie elektrického náboje, přenosu genetické informace nebo difúze.

Jedním z nejznámějších algoritmů spadajících do této metody je GA algoritmus. GA se inspiroval biologickými procesy při přenosu genetické informace, jako jsou mutace, selekce nebo crossing-over. Záměr u vytváření algoritmu bylo zredukování počtu vyhledávání při velkém počtu sekvencí. Základem GA je populace kandidátů, ze kterých se během několika generací hledá nejlepší možné řešení (jedno či více).

2.3.2 Inspirované rojem

Tyto algoritmy, stejně jako roj (včelí, světluškovský, vosí), pracují na principu rozdělení úkolů mezi jednotlivé členy kolektivu, kteří se vzájemně ovlivňují a řídí se jednoduchými pravidly. Základem nejsou pouze roje, ale obecně kolektivní hmyz (mravenci, termiti) a hejna (ryb, ptáků, netopýrů). Stejně jako síla roje stojí na kolektivním chování a jeho důsledcích, takže zde nalézáme kolektivní inteligenci.

Příklady algoritmů jsou ABC (Artificial bee colony), PSO (Particle swarm optimalization), FA (Firefly algorithm), CS (Cuckoo search), ACO (Ant colony optimalization). [11] [21]

2.3.3 Neinspirované rojem

Do této kategorie spadají všechny algoritmy, které si sice inspirovaly přírodou, avšak ne organismy, které se shlukují do rojů, hejn apod.

Příkladem je květinový algoritmus a algoritmus opylování květů, které napodobují chování květin při opylování, kupříkladu jakým způsobem lákají opylující hmyz. [11]

3 Databáze transkripčních motivů

Mnoho organismů již bylo osekvenováno a na jejich genomy byly aplikovány algoritmy pro vyhledání transkriptomických motivů. Pro uložení jednotlivých motivů se založilo mnoho databází, které se od sebe liší jak velikostí, tak i taxonomickými oblastmi, které obsahují. Určité databáze jsou zaměřeny pouze na jeden, většinou modelový, organismus. Níže se nalézá popis databází, které obsahují více taxonomických skupin.

3.1 TRANSFAC

Transfac je databáze obsahující motivy eukariotických organismů, jejich vazebné sítě, konsenzuálně vazebné sekvence (pozičně váhové matice) a regulované geny. Databáze je kurátorovaná a anotovaná. Je rozdělena na dvě části - veřejnou a profesionální část. Veřejná část je dostupná zaregistrovaným uživatelům, kteří využívají Transfac pro neziskovou činnost. Oproti profesionální části je veřejná část o mnoho menší, pakliže bychom se podívali na počet transkripčních faktorů, tak ve verzích 2021.3 je v profesionální části 48,098 faktorů, kdežto ve veřejné je 6,133 faktorů. Ve veřejné databázi lze vyhledávat pomocí 6 metrik - Cell, Class, Factor, Gene, Matrix a Site, kde v každé metrice jde hledat i podle navazujících kritérií. V profesionální části je možné vyhledávat pomocí více metrik, kupříkladu lze vyfiltrovat vyhledávání podle tříd organismů nebo nemoci. Vyhledané motivy jsou popsané několika údaji, ve vyhledávání se zobrazuje přístupové číslo (Accesion No.), druh organismu a zdroj faktoru. Nalezený motiv je na vlastní stránce definován pomocí několika metrik, které nejsou rozepsané slovně, ale jsou použité zkratky, které si uživatel může nastudovat v dokumentaci. Vyhledané motivy a matice nelze stahovat ve fasta formátu. Ve veřejné části se nachází z velké části modelové organismy, avšak nachází se zde i nemodelové organismy. [22]

```
AC   R01203
XX
ID   PARS$PAL_02
XX
DT   20.06.1990 (created); ewi.
DT   28.01.1991 (updated); rge.
CO   Copyright (C), Biobase GmbH.
XX
TY   D
XX
DE   PAL (phenylalanine ammonia-lyase); G000659.
XX
OS   parsley, Petroselinum crispum (P. hortense)
OC   eukaryota; viridiplantae; embryobionta; magnoliophyta; magnoliopsida; rosidae; apiales; apiaceae
XX
SQ   CTCC.
XX
SF   -249
ST   -246
XX
SO   0501 dipl. parsley.
XX
MM   methylation protection
XX
DR   EMBL: X15473; PCPAL1GN (242:245).
DR   EPD: EP35057; PC_PAL1.
XX
RN   [1]; RE0000392.
RX   PUBMED: 2767049.
RA   Lois R., Dietrich A., Hahlbrock K., Schulz W.
RT   A phenylalanine ammonia-lyase gene from parsley: structure, regulation and identification of elicitor
     and light responsive cis-acting elements
RL   EMBO J. 8:1641-1648 (1989).
XX
//
```

Obr. 3.1: Záznam transkripčního motivu PAL v databázi Transfac

3.2 JASPAR

Jaspar je veřejná databáze, kde lze najít motivy mnohobuněčných eukariotických organismů. V databázi můžeme najít pozičně frekvenční matice (PMF), ve kterých jsou uloženy vazby transkripčních motivů. Ty shrnují výskyty každého nukleotidu v každé poloze v souboru pozorovaných interakcí. PMF lze transformovat na pravděpodobnostní nebo energetické modely za účelem vytvoření matic polohových hmotností (MPH) nebo pozičně specifických skórovacích matic (PSSM).

Jaspar databáze se dělí na dvě subdatabáze, větší z nich se nazývá JASPAR-CORE, druhá JASPAR UNVALIDATED. JASPAR-CORE obsahuje upravené, validované, neredundantní sady transkripčních faktorů. Zde lze vyhledávat pomocí subdatabáze, taxonu, druhu organismu, třídy nebo rodiny transkripčního faktoru. JASPAR UNVALIDATED obsahuje nevalidované transkripční faktory. Ve vyhledaném záznamu transkripčního motivu lze nalézt unikátní ID motivu, MPH a PSSM pro daný motiv. Soubory pro daný motiv si lze stáhnout v html a fasta formátu. V subdatabázi JASPAR-CORE se nalézají výhradně modelové organismy. [23]

Obr. 3.2: Záznam transkripčního motivu emBP-1 v databázi JASPAR

3.3 PROSITE

Prosite je databáze, která obsahuje sekce s dokumentacemi popisující proteinové domény, rodiny, funkční místa a související vzory a profily pro jejich identifikaci. PROSITE doplňuje ProRule, soubor pravidel založených na profilech a vzorech, poskytuje další informace o funkčně anebo strukturně kritických aminokyselinách. PROSITE obsahuje vzory a profily pro více než tisíc proteinových rodin nebo domén. Ke každému z těchto popisů je přiložena dokumentace poskytující základní informace o struktuře a funkci těchto proteinů. [24]

3.4 CisBP

CisBP je volně dostupná online databáze vazebně specifických transkripčních faktorů. V současné době obsahuje údaje od více než 700 druhů organicmů. CisBP shromažďuje data z jiných databází, jako jsou Transfac, JASPAR, HOCOMOCO atd. Kvůli zařazení určitých motivů z profesionální části databáze Transfac, která vyžaduje licenci, se některé motivy nezobrazují a v informacích mají pouze informaci, že se motiv nachází v této části databáze. Kromě přímo určených vazebných motivů obsahuje také odvozené motivy (díky mapování napříč druhy se hledají podobnosti ve vazebných doménách rodin TF). [25]

4 Praktická část

Tato kapitola popisuje praktickou část práce sestávající z vytvoření datasetu, návrhu algoritmu s využitím výpočtu vzájemné informace, otestování na datasetu a porovnání s jinými dostupnými algoritmy vyhledávající transkripční motivy. V první části popisuji proces vytvoření testovacího datasetu z volně dostupných databází. V druhé části popisuji vytvořený algoritmus a jeho dílčí části. Třetí část se zaměřuje na porovnání s algoritmy FIRE a MEME.

4.1 Vytvoření datasetu

Všechny popisované databáze ve třetí kapitole obsahují transkripční motivy mnoha organismů. V žádné ze zmíněných databází nebylo možno vyfiltrovat motivy podle toho, zda se jedná o modelové či nemodelové organismy. Bylo proto potřeba manuálně projít všechny záznamy a vyhodnotit, zda se organismus řadí mezi modelové či nikoliv. Z tohoto vyhledávání vyšlo, že ač databáze JAPSPAR-CORE na svých stránkách neuvádí tuto informaci, obsahuje pouze modelové organismy. Veřejná část databáze Transfac obsahuje několik nemodelových organismů, avšak jejich záznamy transkripčních motivů obsahují pouze motiv bez související sekvence. V Transfac jsem nalezla motivy a informaci, že se nachází v promotorové části sekvence. Z toho důvodu jsem gen, ve kterém se daný trankripční motiv nachází, vyhledala v NCBI v nukleotidové databázi. Zde se nachází celé promotorové sekvence genu, ve kterých je obsažen i daný transkripční motiv.

Konkrétně jsem jako testovací nemodelový organismus vybrala *Petroselinum crispum* - petržel zahradní. Dataset pojmenovaný jako "parsley_PAL_gene.fasta"je ve fasta formátu nacházející se v přílohách, obsahuje 16 záznamů promotorových sekvencí fenylalanin amoniak-lyáza 1, 2, 3 a 4 (PAL-1, PAL-2, PAL-3, PAL-4) vytvořený ze 4 neupravovaných sekvencí a dodatečně vytvořených 12 umělých sekvencí (označené zkratkou art od slova artificial - umělý) pro rozšíření datasetu z důvodu jeho malé velikosti. Umělé sekvence mají zaměněná místa transkripčního motivu a přidané motivy, konkrétně motivy PAL-1 a PAL-4. Přirozené a umělé výskyty jsou popsané v Příloze A. Jednotlivé nukleotidy jsou kódované podle IUPAC abecedy. Následně jsem v databázi CisBP ověřila, že v profesionální části databáze Transfac se nachází více transkripčních motivů pro petržel zahradní, než kolik se zobrazuje ve veřejné části.

Tab. 4.1: Motivy vyhledané v databázi TRANSFAC

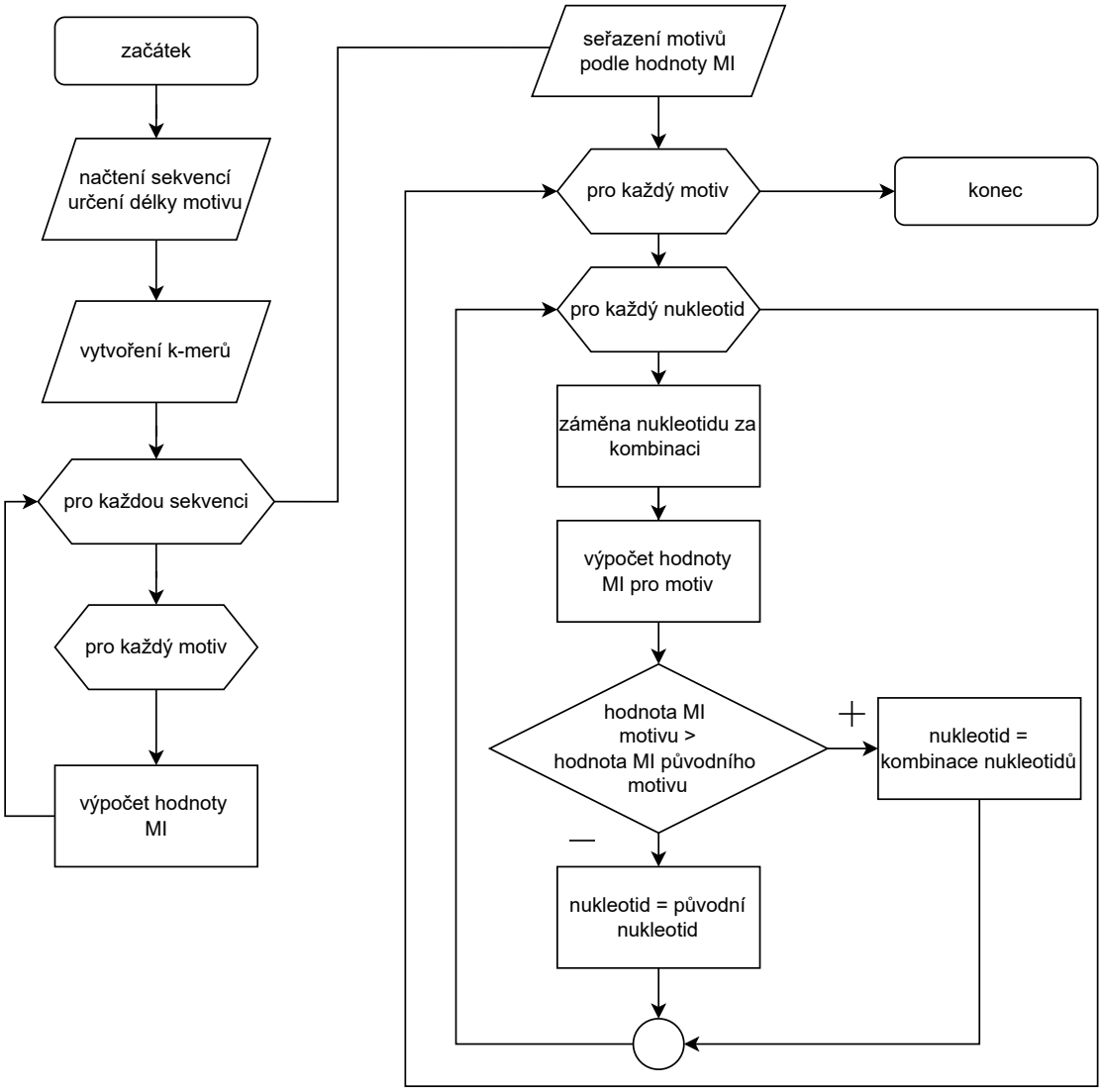
typ genu	**sekvence motivu**	**délka sekvence**
PAL-1	TCCACCT	7
PAL-2	CTCC	4
PAL-3	CTCCAACAAACCCCTTC	17
PAL-4	CCGTCC	6

4.2 MIne algoritmus

Vytvořený algoritmus obsahuje několik dílčích částí. Jednotlivé sekce lze pojmenovat jako načtení dat, vytvoření k-merů, výpočet vzájemné informace a optimalizace. Algoritmus je psán v programovacím jazyce Python, verze Python 3.8. Výpočet algoritmu se zakládá na výpočtu vzájemné informace, která je vždy počítána mezi dvojicí informací. Algoritmus FIRE využívá pro výpočet dvojici k-meru a hodnoty expresního motivu daného genu. Algoritmus MIA využívá ve výpočtu dvojice k-merů pro horizontální vzájemnou informaci, vertikální vzájemnou informaci a vertikální entropii. Ty pak dosazuje do výpočtu Jensen-Shannonovy divergence. Oba algoritmy tedy kombinují výpočet vzájemné informace s jinými metrikami.

V této práci jsem se zaměřila na otestování využití pouze výpočtu vzájemné informace mezi k-mery ze sekvencí. V první verzi bylo vyzkoušeno, zda je možné najít motivy v rámci jedné sekvence - pro vytvořené k-mery z právě jedné sekvence byla počítána hodnota MI proti ostatním k-merům v rámci oné sekvence. Druhá verze ověřovala, zda je možné najít motiv z jedné sekvence pomocí zbylých sekvencí z datasetu. MI se tedy počítala pro dvojici k-merů - první byl postupně každý k-mer z jedné vybrané sekvence, druhý k-mer byl vybírán ze všech zbylých sekvencí. Algoritmus se nachází v přílohách pojmenovaný "MIne_alg.ipynb",

Na vývojovém diagramu níže můžeme vidět jednotlivé kroky navrženého algoritmu. První sloupec odpovídá částem načtení dat, vytvoření k-merů a výpočet vzájemné informace. Druhý sloupec znázorňuje proces optimalizace. Výstupem algoritmu je seznam optimalizovaných motivů s hodnotou vzájemné informace a počtem výskytu.



Obr. 4.1: Vývojový diagram algoritmu, levý sloupec načtení dat a výpočet vzájemné informace, prostřední sloupec optimalizace

4.2.1 Načtení dat a vytvoření k-merů

V rámci funkce načtení dat, pojmenované v kódu jako "load_data", se dataset "parsley_PAL_gene.fasta"ve fasta formátu načítá pomocí balíčku Biopython a knihovny Bio, konkrétně jejích balíčků Seq a SeqIO. V rámci tohoto kroku uživatel definoval, v jaké délce požaduje hledané motivy a kolik motivů chce vyhledat. Výstupem funkce načtení dat byl seznam sekvencí ve formě řetězců. Pro každou zadanou délku motivu následovaly níže popsané úkony. Každá sekvence v datasetu byla rozdělěna na k-mery délky předem zadané uživatelem ve funkci "make_kmers". Pro snížení výpočetní náročnosti se v tomto kroku odstranily duplicitní k-mery. Výstupem této

funkce byl seznam vytvořených k-merů.

4.2.2 Výpočet vzájemné informace

Tento krok zahrnuje načtení vytvořených k-merů a spočítání hodnoty vzájemné informace každého s každým k-merem, funkce je nazvaná "count_mi". Vstupem je výstup z funkce, která vytvářela k-mery. Pro výpočet byla použita knihovna Scikit Learn [26], konkrétně její funkce mutual_info_score. Tato funkce je zespodu omezena 0, vrchní hranici avšak nemá přesně danou. Z dat vyšlo, že nejvyšší hodnota dosažené vzájemné informace byla 1.36. Pro vyfiltrování pouze těch nejrelevantnějších k-merů, které lze vyhodnotit jako motivy, byla nastavena spodní hranice vzájemné informace na hodnotu 0.6. Ta byla vybrána jako polovina hodnoty vzájemné informace - nízké hodnoty vzájemné informace naznačují méně signifikantní vztah mezi k-mery. Ty s vyšší hodnotou poukazují na vyšší pravděpodobnost, že se jedná o hledané tanskripční motivy. V případě, že hodnota vzájemné informace překročila tuto hranici, byl k-mer zařazen do slovníku spolu se svou vypočítanou hodnotou MI. Dle uživatelského vstupu byl vybrán počet k-merů, které se nejčastěji objevovaly ve slovníku. Tyto k-mery označíme jako primární motivy. V této části se právě ony výše zmíněné verze lišily. První verze, která počítala vzájemnou informaci v rámci jedné sekvence, v této části ukládala oba k-mery, které přesáhly hranici 0.6. Druhá verze ukládala z vyhledané dvojice pouze k-mery ze sekvence, která byla počítána proti k-merům z ostatních sekvencí. Výsledky byly předávány dál v podobě slovníku, ve kterém se nacházel primární motiv, četnost jeho výskytu a hodnota vzájemné informace.

4.2.3 Optimalizace

V této části kódu byly motivy z předchozí funkce optimalizovány do informativnější podoby. Proteiny vázající DNA a RNA se obvykle mohou vázat k více lehce se lišícím místům, proto se v tomto kroku motivy zobecní. Díky obecnějšímu charakteru se stanou informativnějšími a reprezentují přesněji vazebná místa, na která se vážou transkripční faktory. Informativnější motivy mívají na určitých místech kombinaci nukleotidů, nejsou přesně definované. Z toho důvodu bylo vyzkoušeno i v navrženém algoritmu pro nukleotidy v rámci motivu kombinace s jiným nukleotidem. Funkce s názvem "better_mi"v cyklu pro každý nukleotid vyzkoušela, zda v případě záměny za jiný nukleotid bude hodnota MI podobná. V případě, že možnost se zaměněným nukleotidem přesáhla hranici 95 % hodnoty MI pro původní motiv, byl onen nukleotid přidán jako vhodná kombinace. V případě, že by kupříkladu v motivu na pozici nukleotidu G vycházel jako vhodná možnost i nukleotid A, vznikla by kombinace [GA]. V optimalizovaných verzích motivů je v kombinacích nukleotidů vždy

na prvním místě původní nukleotid z ještě neoptimalizovaného motivu. Výstupem této funkce je seznam optimalizovaných motivů.

4.3 Vyhodnocení a porovnání s dalšími algoritmy

Výše popsaný algoritmus byl otestován na vytvořeném datasetu. Následující kapitola se věnuje jeho zhodnocení pomocí z-skóre a matice záměn a z ní vypočítaných hodnoty správnosti (accuracy) vůči výsledkům algoritmů FIRE a MEME.

4.3.1 Vyhodnocení MIne algoritmu

MIne algoritmus jsem otestovala ve dvou nastaveních. První nastavení počítalo vzájemnou informaci v rámci jedné sekvence - hledala se podobnost sama se sebou. Druhé nastavení vyhledávalo možné motivy v ostatních sekvencích. Obě verze vypisovaly prvních 20 výsledných motivů. První verze trvala podle délky k-merů (od délky 4 po délku 17) od 18 minut po 6 hodin 34 minut (CPU 2.20 GHz, 4 jádra) pro vytvořený dataset při spuštění na Google Colaboratory. Jednalo se o téměř 22násobné navýšení času pro výpočet motivu nejkratšího a nejdelšího k-meru. Motiv TCCACCT byl nalezen primárně jako CACCTGT a následně optimalizován do podoby uvedené v tabulce níže. Ta obsahuje vyhledané optimalizované motivy. Motiv je posunut o dva nukleotidy. Motiv CTCC byl nalezen primárně jako GTCC, se záměnou jednoho nukleotidu. Motiv CCGTCC byl primárně nalezen jako ACCGTC, byl tedy posunut o jeden nukleotid. Motiv CTCCAACAAACCCCTTC nebyl nalezen z důvodu nevhodné optimalizace výsledků.

Tab. 4.2: Motivy vyhledané algoritmem MIne verze 1

sekvence motivu	**motiv nalezený MIne verze 1**	**hodnota MI**
TCCACCT	[CT]A[CGT][CT][TA]G[TA]	1.2770
CTCC	[GA][TA][CAGT][CAGT]	1.0397
CTCCAACAAACCCCTTC	nenalezen	-
CCGTCC	A[CT][CAGT]GT[CAT]	1.2425

Pro zhodnocení přesnosti algoritmu byl použit výpočet přesnosti z matice záměn. Matice záměn byla vytvořena ze všech k-merů, které vstupovaly do výpočtu primárních motivů. V tabulkách můžeme pozorovat nárůst počtu k-merů na délce požadovaného motivu, ač by bylo očekávatelné, že jejich počet bude klesat. Tato skutečnost je zapříčiněna odstraňováním duplicitních motivů - u kratších motivů je pravděpodobnější, že se určitý k-mer vytvoří vícekrát než pouze jednou, narozdíl

od delších motivů, u kterých je pravděpodobnost duplicit menší. Z důvodu, že ani jeden vyhledaný motiv, primární či optimalizovaný, se přesně neshodoval s vloženým motivem, byl určen jako správně nalezený motiv nejpodobnější primární motiv (tyto motivy jsou zmiňované v předešlém odstavci).

Rovnice pro výpočet správnosti:

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN}$$

délka 4 verze 1	**vyhledání motiv**	**vyhledání k-mer**
skutečnost motiv	4	52
skutečnost k-mer	21	3 110

$\text{Accuracy} = 0,977$

délka 6 verze 1	**vyhledání motiv**	**vyhledání k-mer**
skutečnost motiv	4	14
skutečnost k-mer	5	9 951

$\text{Accuracy} = 0,998$

délka 7 verze 1	**vyhledání motiv**	**vyhledání k-mer**
skutečnost motiv	4	14
skutečnost k-mer	4	11 809

$\text{Accuracy} = 0,998$

Druhá verze kvůli násobně větší náročnosti trvala v závislosti na délce motivu od 1 minuty pro délku 4 po 2 hodiny 20 minut pro délku 17, celkově výpočet pro všech 16 sekvencí trval od 29 minut po 28 hodin 20 minut (CPU 2.20 GHz). Jednalo se o téměř 57násobné navýšení času pro výpočet motivu nejkratšího a nejdelšího k-meru. Časy pro výpočet se lišily i v rámci jednotlivých sekvencí kvůli jejich rozdílné délce, nejdelší sekvence byla PAL-3 artificial 2 s délkou 1365, nejkratší pak sekvence PAL-2 s délkou 411. Motiv TCCACCT byl nalezen primárně jako GTTCCAC, posunutý o dva nukleotidy. Motiv CTCC byl nalezen primárně jako CTCG, se záměnou jednoho nukleotidu. Motiv CCGTCC byl primárně nalezen jako TCCGTA, byl tedy posunut o jeden nukleotid a obsahoval jednu záměnu. Motiv CTCCAACAAACCCCTTC nebyl nalezen z důvodu nevhodné optimalizace výsledků (algoritmus vyhodnotil vhodné zobecnění jako kombinaci všech nukleotidů). Obecně lze říci, že MIne algoritmus verze 1 i 2 více zobecňoval motivy než algoritmy FIRE a MEME.

Tab. 4.3: Motivy vyhledané algoritmem MIne verze 2

sekvence motivu	**motiv nalezený MIne verze 2**	**hodnota MI**
TCCACCT	G[TGA][TGA][CTA][CAG]A[CAGT]	1.2770
CTCC	[CATG][TA][CAGT][GA]	1.0397
CTCCAACAAACCCCTTC	nenalezen	-
CCGTCC	[TAG][CGA][CAG]G[TGA]A	1.3297

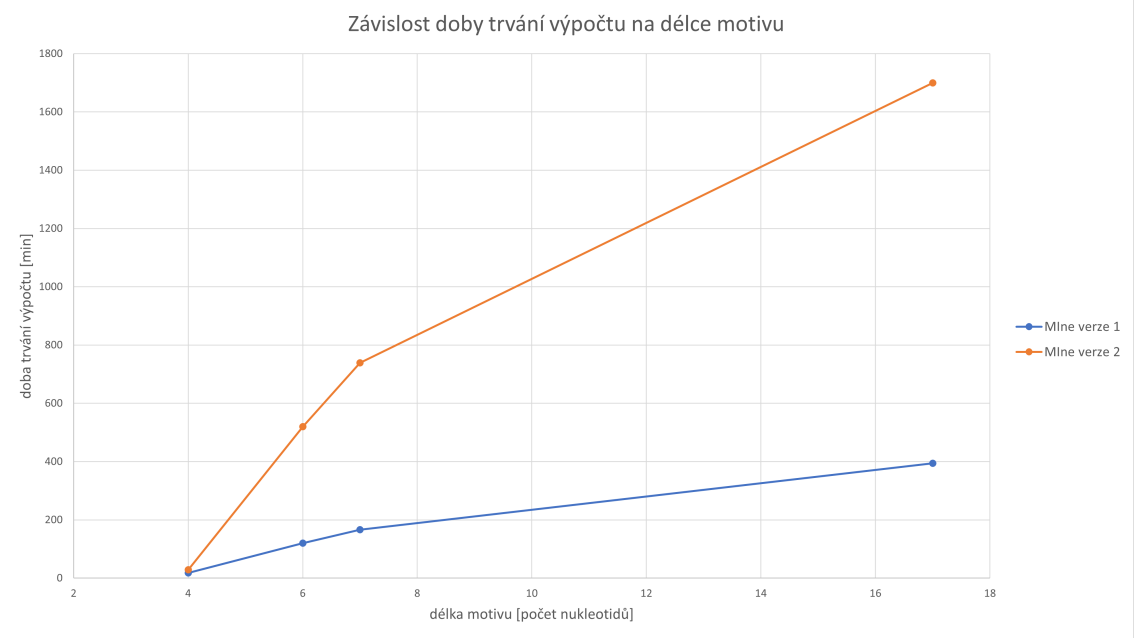
délka 4 **verze 2**	**vyhledání motiv**	**vyhledání k-mer**
skutečnost motiv	4	52
skutečnost k-mer	8	3 123

$$\text{Accuracy} = 0,981$$
$$\text{Accuracy} = 0,998$$
$$\text{Accuracy} = 0,998$$

délka 6 **verze 2**	**vyhledání motiv**	**vyhledání k-mer**
skutečnost motiv	4	14
skutečnost k-mer	5	9 951

délka 7 **verze 2**	**vyhledání motiv**	**vyhledání k-mer**
skutečnost motiv	1	17
skutečnost k-mer	7	11 806

U obou verzí algoritmu MIne lze pozorovat lineární asymptotickou složitost O(n) při nárůstu času s přibývající délkou požadovaného motivu, jak naznačuje graf níže.



Obr. 4.2: Graf znázorňující závislost doby trvání výpočtu na délce vyhledávaného motivu

4.3.2 Porovnání s dalšími algoritmy

Tato podkapitola se zaměřuje na porovnání výsledků MIne algoritmu s výsledky algoritmů FIRE a MEME. Algoritmus FIRE byl vybrán z důvodu využívání výpočtu vzájemné informace, stejně jako v MIne algoritmu. MEME algoritmus byl vybrán jako jeden ze standardně využívaných algoritmů.

Pro spuštění **algoritmu FIRE** je potřeba fasta soubor se sekvencemi a taktéž soubor s expresními profily daných genů. Pro absenci profilu pro petržel zahradní byl použit profil *Solanum lycopersicum* - rajče jedlé. Pro zjištění jeho podobnosti byl použit nástroj BLAST [27]. Do zarovnání se pro nízkou podobnost vytvořeného datasetu s jiným organismem, než je právě petržel, použil přímo gen PAL-1 pro petržel a ne pouze jeho promotorová sekvence. Pro tuto sekvenci vyšel jako nejbližší organismus, který má známý expresní profil pro gen PAL, právě rajče jedlé - konkrétně jeho varianta PAL-3, hodnota pokrytí byla 33 %. Expresní profil byl stáhnut z Expression Atlas [28], kde se nachází expresní profily 65 organismů - živočichů, rostlin a hub. Algoritmus FIRE je psaný pro Linux, je tedy buď potřeba nainstalovat rozšíření pro Windows či spouštět na Linuxu. FIRE vyhledal po vložení sekvence ve formátu fasta a expresního profilu 2 ze 4 motivů.

Tab. 4.4: Motivy vyhledané algoritmem FIRE

sekvence motivu	**motiv nalezený FIRE**
TCCACCT	ATCCACC
CTCC	nenalezen
CTCCAACAAACCCCTTC	nenalezen
CCGTCC	GTATCC

Motiv TCCACCT byl nalezen posunutý o jeden nukleotid. Motiv CTCC sice byl nalezen, avšak neprošel všemi randommizačními testy pro optimalizaci motivů, stejně jako jiný motiv o délce 4. Motiv CTCCAACAAACCCCTTC nebyl nalezen z důvodu přílišné délky, algoritmus FIRE je schopný vyhledávat motivy pouze do délky 9 včetně. Motiv CCGTCC nebyl přesně nalezen, nejvíc se mu podobal vyhledaný motiv GTATCC, který obsahuje inzerce i delece nukleotidů. Kvůli malé velikosti datasetu trvalo spuštění jednoho běhu necelou sekundu. Algoritmus FIRE je vytvořený na násobně větší datasety, než vytvořený pro tuto práci.

Algoritmus MEME vyhledává transkripční motivy na principu podobnosti mezi sekvencemi. Pro vyhledávání je vyžadován pouze fasta soubor. Uživatel si může navolit metriky jako rozsah délky motivu, kolik by jich měl MEME vyhledat či vyhledávací mód. Algortimus MEME má standardně zadaný rozsah délky motivu od 6 do 50 nukleotidů. Při nenavolení délky se pro dataset vyhledávají motivy nejčastěji s délkou 50, ne méně však než 48. Po definování délky na rozsah 4 až 7 (pro vyhledání motivu PAL-1 stejné délky, trvání 27 vteřin) a 4 až 6 (pro motiv PAL-4, trvání 24 vteřin) byly nalezeny oba motivy.

Tab. 4.5: Motivy vyhledané algoritmem MEME

sekvence motivu	**motiv nalezený MEME**	**e-hodnota**
TCCACCT	CCACC[CG]T	2.9e-011
CTCC	nenalezen	-
CTCCAACAAACCCCTTC	TC[CT][AC][AC]C[AC]A[AC]CCCCT[TGC][CG][TA]	1.1e-055
CCGTCC	[CAG]CGTCC	9.3e-003

Pro motiv TCCACCT délky 7 se nejvíce podobá motiv 2 (viz Příloha C), ač je zde posunut o jeden nukleotid. Jeho e-hodnota je 2.9e-011 a byl nalezen ve všech 16 sekvencích v datasetu. Motiv CTCC nebyl nalezen - ani jeden výsledek se s motivem neshodoval a neměl dostatečně nízkou e-hodnotu natolik, aby ho algoritmus vyhodnotil jako motiv. Motiv CTCCAACAAACCCCTTC délky 17 byl nalezen jako motiv 1 s e-hodnotou 1.1e-055 a nalezením ve všech 16 sekvencích (viz Příloha D).

Motiv CCGTCC délky 6 byl nalezen přesně v motivu číslo 5 (viz Příloha B) s e-hodnotou 9.3e-003 a nalezením ve všech sekvencích.

Pro vyhodnocení výsledků algoritmu bylo vybráno z-skóre jako všeobecný ukazatel rozptylu dat a matice záměn. Z-skóre bylo počítáno pro všechny správně nalezené motivy. V rámci vyhledaného motivu TCCACCT se jako nejlepší ukazuje algoritmus MIne verze 2 se z-skórem -5.5876. MIne verze 1 i MEME jsou co se týče výsledku z-skóre velice podobné MIne algoritmu. Znatelně větší rozptyl byl vypočítán pro výsledek algoritmu FIRE, jehož hodnota z-skóre dosáhla hodnoty -7.2114. Pro motiv CTTC byly hodnoceny pouze výsledky z algoritmů MIne verze 1 a 2, ze kterých vyšla jako přesnější verze 1 s hodnotou z-skóre -0.0025. Motiv CTCCAACAAACCCCTTC byl hodnocen pouze pro algoritmus MEME s hodnotou z-skóre -35.1608. Pro motiv CCGTCC se hodnoty z-skóre pro jednotlivé algoritmy příliš nelišily. Nejmenší rozptyl pro tento motiv lze vidět u algoritmu MIne verze 2 s hodnotou -3.4618, o jednu setinu horšího výsledku dosáhl FIRE s hodnotou z-skóre -3.4798. O setinu z-skóre se lišily taktéž hodnoty algoritmů MEME a MIne verze 1.

Tab. 4.6: Porovnání z-skóre pro jednotlivé motivy vyhledané algoritmy

sekvence motivu	**MIne verze 1**	**MIne verze 2**	**FIRE**	**MEME**
TCCACCT	-5.5703	-5.5876	-7.2114	-5.6145
CTCC	-0.0025	0.0086	-	-
CTCCAACAAACCCCTTC	-	-	-	-35.1608
CCGTCC	-3.5278	-3.4618	-3.4798	-3.5107

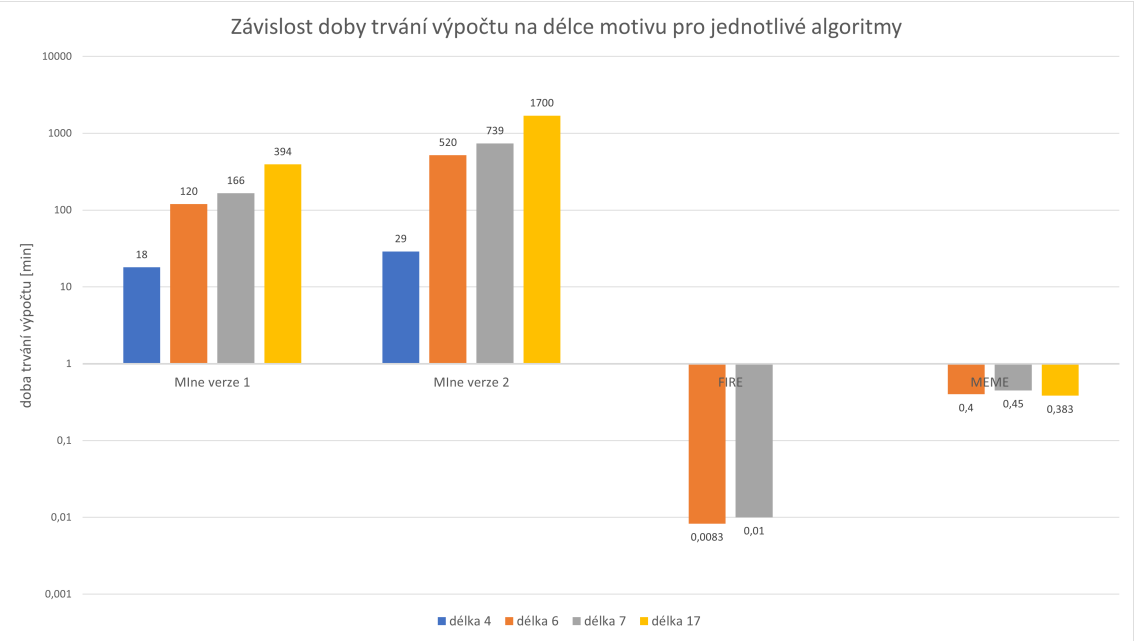
Časová náročnost se signifikantně liší pro jednotlivé algoritmy. Jako nejrychlejší se ukázal FIRE, výpočty pro délky 6 i 7 trvaly méně jak 1 vteřinu. Výpočty v rámci jedné minuty měl algoritmus MEME, z časových údajů nelze říct, že doba trvání výpočtu se prodlužuje v závislosti s rostoucí délkou sekvence. Tuto skutečnost lze pozorovat u času 23 vteřin pro motiv délky 17 a čas 27 vteřin pro motiv délky 7.

Tab. 4.7: Porovnání času pro výpočet motivů algoritmy

sekvence motivu	**MIne verze 1**	**MIne verze 2**	**FIRE**	**MEME**
TCCACCT	2 hodiny 46 minut	12 hodin 19 minut	0.6 vteřiny	27 vteřin
CTCC	18 minut	29 minut	-	-
CTCCAACAAACCCCTTC	6 hodin 34 minut	28 hodin 20 minut	-	23 vteřin
CCGTCC	2 hodiny	8 hodin 40 minut	0.5 vteřiny	24 vteřin

Na grafu níže je znázorněna časová náročnost jednotlivých algoritmů. Z důvodu velkého rozptylu údajů bylo použito pro lepší přehlednost logaritmické měřítko, se

základem 10, pro časovou osu. Můžeme zde pozorovat velice znatelné rozdíly mezi jednotlivými algoritmy.



Obr. 4.3: Graf znázorňující závislost doby trvání výpočtu na délce vyhledávaného motivu pro jednotlivé algoritmy

Závěr

Bakalářská práce se věnovala tématu transkripčních motivů a jejich vyhledáváním pomocí algoritmů s rozdílnými přístupy k vyhledávání. V teoretické části jsou vysvětleny základy k transkripčním motivům a vzájemné informaci. Díky nastudování databází s transkripčními motivy byl vytvořen dataset nemodelového organismu petržele zahradní, na kterém byl testován navržený algoritmus MIne ve dvou verzích. První verze počítající vzájemnou informaci v rámci sebe samotné, jako nejlepší motiv byl podle hodnoty vzájemné informace optimalizovaný [CT]A[CGT][CT][TA]G[TA] s délkou 7. Nejnižší hodnotu vzájemné informace měl motiv GA][TA][CAGT][CAGT] s délkou 4. Přesnost se pro jednotlivé motivy vypočítané v první verzi pohybovala mezi 0,977 a 0,998. Druhá verze počítající vzájemnou informaci v ostatních sekvencích vyhledala nejlépe motiv TAG][CGA][CAG]G[TGA]A s délkou 6. Jako nejhorší motiv s nejnižší hodnotou MI vyšel motiv [CATG][TA][CAGT][GA] s délkou 4. Z výsledků MIne algoritmu vyplynulo, že algoritmus spíše zobecňuje motivy narozdíl od zbylých dvou algoritmů. Přesnost se pro jednotlivé motivy vypočítané v druhé verzi pohybovala mezi 0,981 a 0,998. Z výsledků vzájemné informace i hodnot přesnosti vyplývá, že algoritmus nejlépe hledá motivy délky 6 a 7, délka 4 má nižší hodnoty v obou kategoriích hodnocení. Algoritmus v obou verzích má lineární asymptotickou složitost, která hlavně pro větší délky prodlužuje hodnotu výpočtu do řádu desítek hodin. Ke zrychlení by bylo možné přispět paralelizací výpočtů, především výpočtu vzájemné informace. Při paralelizaci výpočtů by bylo možné taktéž efektivněji pracovat s většími datasety.

Při porovnání s algoritmy FIRE a MEME byly nejvíce znatelné časové rozdíly. Algoritmu FIRE, tvořený pro mnohem větší datasety než vytvořený dataset, byl nejrychlejším algoritmem. Vysokou rychlost výpočtu však kompenzovalo nepřesné nalezení motivů a omezenost délek vyhledávaných motivů. Algoritmus MEME jako jediný dokázal najít všechny motivy, které nezobecňoval do takové míry jako MIne algoritmus obou verzí. Z pohledu z-skóre jako nejpřesnější výsledků dosáhly algoritmus MEME a MIne verze 2.

Literatura

[1] ALBERTS, Bruce. *Základy buněčné biologie: úvod do molekulární biologie buňky.* 2. vydání. Ústí nad Labem: Espero, 1998. ISBN 80-902906-0-4.

[2] ROSYPAL, Stanislav. *Nový přehled biologie.* Praha: Scientia, 2003. ISBN 978-80-86960-23-4.

[3] DNA, RNA and protein synthesis. *ATDBio* United Kingdom. Dostupné z: <https://atdbio.com/nucleic-acids-book/Transcription-Translation-and-Replication

[4] TAM, Wai-Leong a Bing LIM. Genome-wide transcription factor localization and function in stem cells. *StemBook*. 2008. doi:10.3824/stembook.1.19.1

[5] DNA binding proteins, transcription factors. *Jack Westin* Dostupné z: <https://jackwestin.com/resources/mcat-content/control-of-gene-expression-in-eukaryotes/dna-binding-proteins-transcription-factors

[6] STEUER, Ralf, Juergen KURTHS, Carsten Oliver DAUB, Joachim WEISE a Joachim SELBIG. The mutual information: Detecting and evaluating dependencies between variables *Bioinformatics*. 2002, 2002(18(2)), 231–240. *https* : $//doi.org/10.1093/bioinformatics/18.suppl_2.S231$ Dostupné z: <https://academic.oup.com/bioinformatics/issue/18/$suppl_2$

[7] COVER, T. M. a Joy A. THOMAS. *Elements of information theory.* 2nd ed. Hoboken: Wiley-Interscience, 2006. ISBN 978-0-471-24195-9.

[8] Tavazoie lab Dostupné z: <https://tavazoielab.c2b2.columbia.edu/FIRE

[9] LIEBER, Daniel S., Olivier ELEMENTO a Saeed TAVAZOIE. Large-Scale Discovery and Characterization of Protein Regulatory Motifs in Eukaryotes. *PLOS ONE*. 2010, 2010(5(12): e14444) Dostupné z: <https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0014444

[10] ELEMENTO, Olivier, Noam SLONIM a Saeed TAVAZOIE. A universal framework for regulatory element discovery across all genomes and data-types *Molecular cell*. 2007, 2007(28(2)), 337-350. doi: 10.1016/j.molcel.2007.09.027

[11] A. HASHIM, Fatma, Mai S. MABROUK a Walid AL-ATABANY. Review of Different Sequence Motif Finding Algorithms. *Avicenna journal of medical biotechnology*. 2019, 2019(11(2), 130-148. Dostupné z: <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6490410/

[12] BAILEY, Timothy L., James JOHNSON, Charles E. GRANT a William S. NOBLE. The MEME Suite. *Nucleic Acids Research*. 2015, 2015(43(W1), 39-49. doi: https://doi.org/10.1093/nar/gkv416

[13] IUPAC codes. Sequence Manipulation Suite Dostupné z: <https://www.bioinformatics.org/sms2/iupac.html

[14] BAILEY, Timothy L. DREME: motif discovery in transcription factor ChIP-seq data. *Bioinformatics*. 2011, 2011(27(12), 1653-1659. doi:10.1093/bioinformatics/btr261

[15] SHAROV, Alexei A. a Minoru S.H. KO. Exhaustive Search for Over-represented DNA Sequence Motifs with CisFinder. *Bioinformatics*. 2009, 2009(16(5), 261-273. doi:10.1093/dnares/dsp014

[16] ZHANG, Yipu, Ping WANG a Maode YAN. An Entropy-Based Position Projection Algorithm for Motif Discovery. *BioMed research international*. 2016, 2016((2016): 9127474). doi: 10.1155/2016/9127474

[17] PAVESI, Giulio, Federico ZAMBELLI a Graziano PESOLE. WeederH: an algorithm for finding conserved regulatory motifs and regions in homologous sequences. *BMC Bioinformatics*. 2007, 2007(8(46). doi: 10.1186/1471-2105-8-46.

[18] MACHANICK, Philip a Timothy L. BAILEY. MEME-ChIP: motif analysis of large DNA datasets. *Bioinformatics*. 2011, 2011(27(12), 1696-1697. doi:10.1093/bioinformatics/btr189

[19] MA, Wenxiu, William S. NOBLE a Timothy L. BAILEY. Motif-based analysis of large nucleotide data sets using MEME-ChIP. *Bioinformatics*. 2014, 2014(9(6), 1428-1450. doi:10.1038/nprot.2014.083

[20] FISTER JR, Iztok, Xin-She YANG, Iztok FISTER, Janez BREST a Dušan FISTER. A Brief Review of Nature-Inspired Algorithms for Optimization. *Elektrotehniški vestnik*. 2013, 2013(80(3), 1-7. Dostupné z: <https://arxiv.org/pdf/1307.4186.pdf

[21] KARABOGA, Dervis a Selcuk ASLAN. A discrete artificial bee colony algorithm for detecting transcription factor binding sites in DNA sequences. *Genetics Molecular Research*. 2016, 2016(15(2): gmr8645. https://doi.org/10.4238/gmr.15028645

[22] MATYS, Volker, Olga KEL-MARGOULIS, Ellen FRICKE, et al. TRANSFAC and its module TRANSCompel: transcriptional gene regulation in eukaryotes. *Nucleic Acids Research.* 2006, 2006(34, D1), 108-110. doi:10.1093/nar/gkj143.

[23] CASTRO-MONDRAGON, Jaime A., Rafael RIUDAVETS-PUIG, Ieva RAULUSEVICIUTE, et al. JASPAR 2022: the 9th release of the open-access database of transcription factor binding profiles. *Nucleic Acids Research.* 2021, 2021(gkab1113). doi:10.1093/nar/gkab1113.

[24] SIGRIST, Christian J. A., Edouard DE CASTRO, Lorenzo CERUTTI, Béatrice A. CUCHE, Nicolas HULO, Alan BRIDGE, Lydie BOUGUELERET a Ioannis XENARIOS. New and continuing developments at PROSITE. *Nucleic Acids Research.* 2013, 2013(vol. 41, D1), 344–347. doi:10.1093/nar/gks1067.

[25] WEIRAUCH, Matthew, Ally YANG, Mihai ALBU, et al. Determination and inference of eukaryotic transcription factor sequence specificity. *Cell.* 2014, 2014(158(6)), 1431-1443. doi: 10.1016/j.cell.2014.08.009.

[26] PEDREGOSA, Fabian, Gaël VAROQUAUX, Alexandre GRAMFORT, et al. Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research.* 2011, 2011(12(85)), 2825-2830. doi: 10.1016/j.cell.2014.08.009.

[27] ALTSCHUL, Stephen Frank, Warren GISH, Webb MILLER, Eugene MYERS a David LIPMAN. *Journal of molecular biology.* 1990, 1990(215(3)), 403–410. doi.org/10.1016/S0022-2836(05)80360-2

[28] PAPATHEODOROU, Irene, Pablo MORENO, Jonathan MANNING et al. Expression Atlas update: from tissues to single cells. *Nucleic Acids Research.* 2020, 2020(48(D1)), D77–D83. https://doi.org/10.1093/nar/gkz947

Seznam příloh

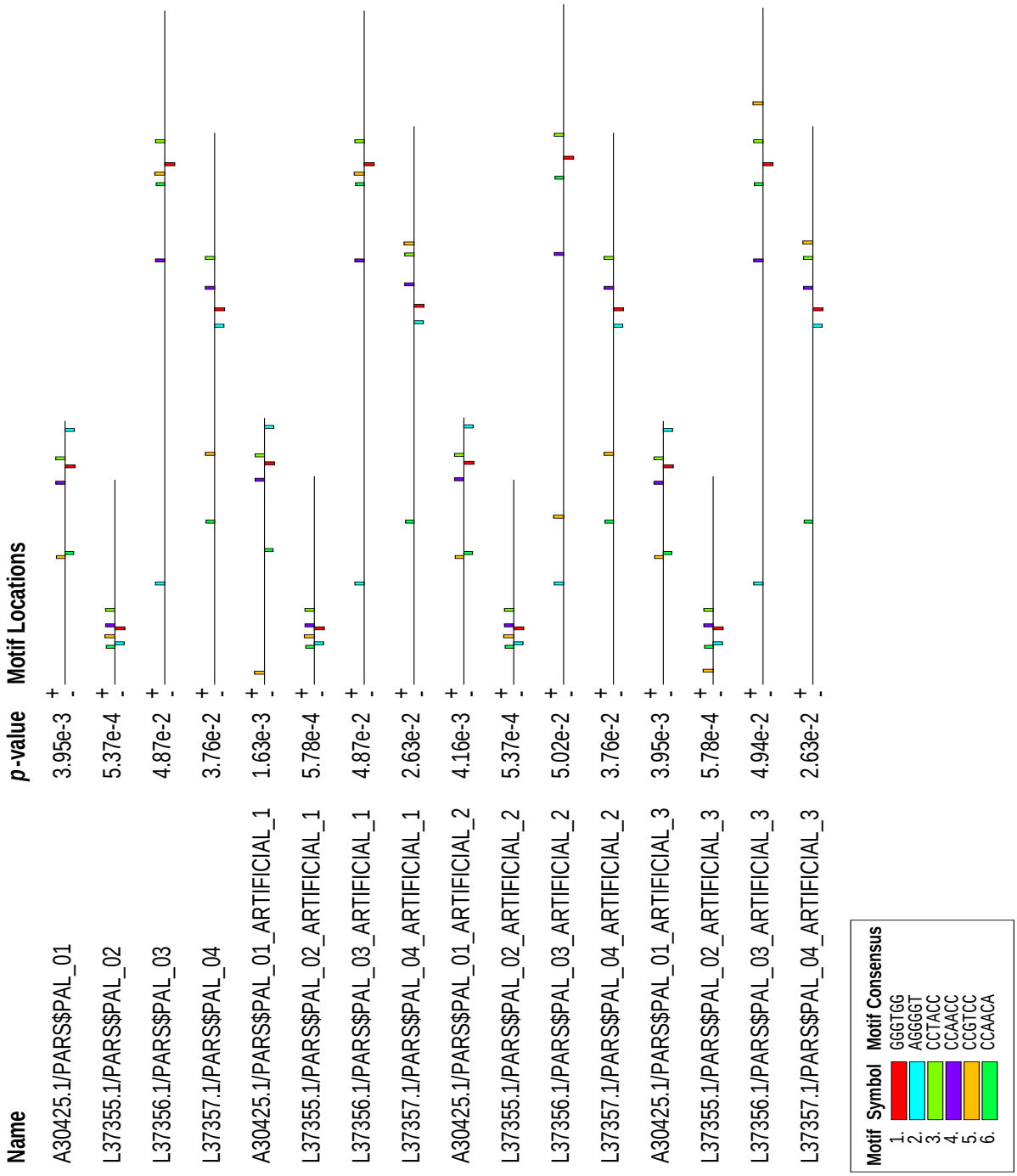
A Příloha: Popis výskytů přirozených a umělě vložených motivů sekvencí

TCCACCT - přirozeně jako CCACCT v PAL-3 2krát, v PAL-1 1krát
uměle dodáno do:
art 1 - PAL-2, PAL-4
art 2 - PAL-2, PAL-3
art 3 - PAL-2, PAL-4
celkem 18krát v datasetu

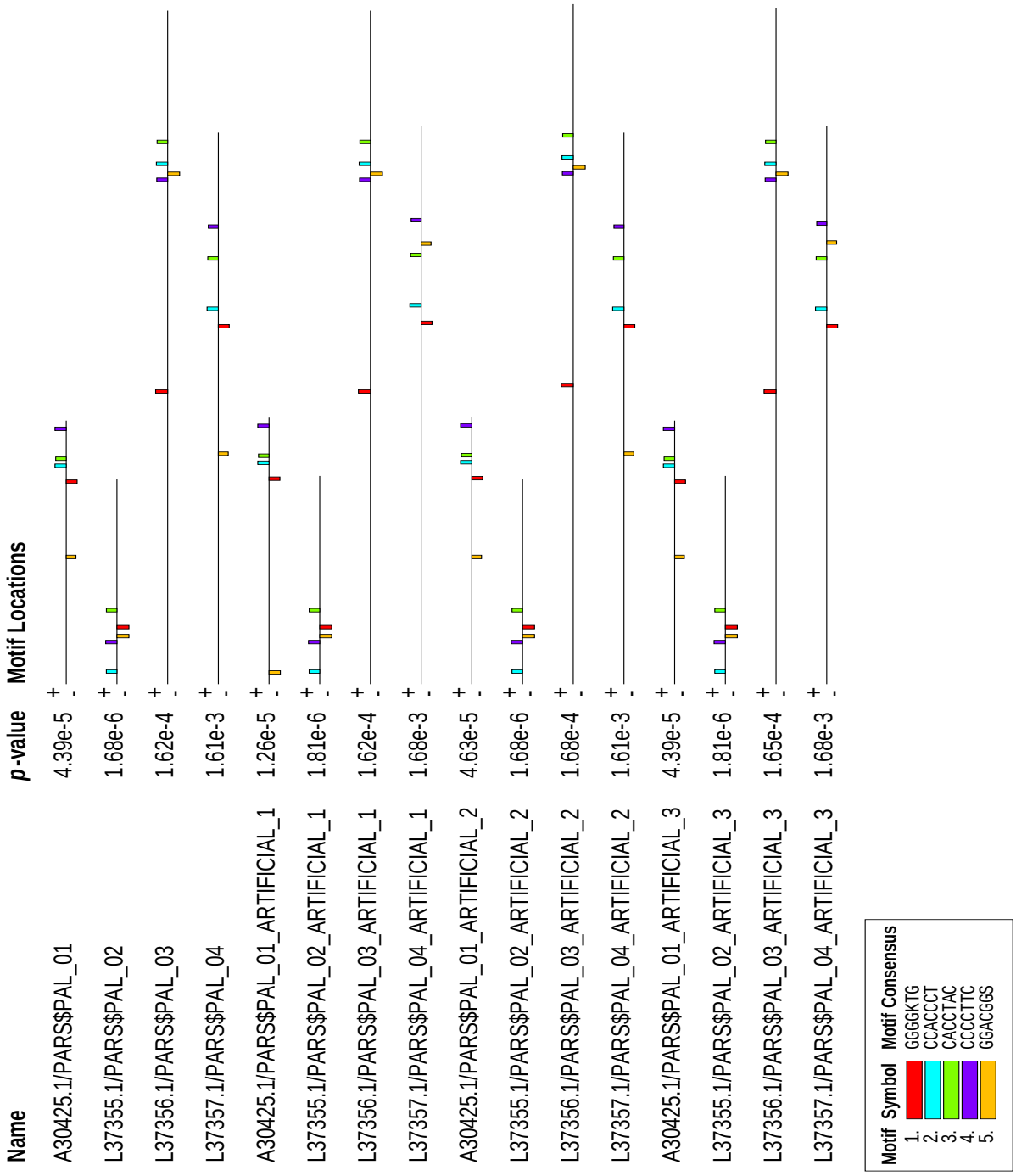
CTCC - přirozeně v PAL-1 3krát, v PAL-2 2krát, v PAL-3 4krát, v PAL-4 5krát
uměle nepřidáváno
celkem 56krát v datasetu

CCGTCC - přirozeně v PAL-2 2krát, v PAL-3 1krát
uměle dodáno do:
art 1 - PAL-1, PAL-4
art 2 - PAL-1, PAL-3
art 3 - PAL-3, PAL-4
celkem 18krát v datasetu

B Příloha: Výsledky algoritmu MEME pro zadanou délku motivu 4-6 nukleotidů



C Příloha: Výsledky algoritmu MEME pro zadanou délku motivu 4-7 nukleotidů



D Příloha: Výsledky algoritmu MEME pro zadanou délku motivu 10-17 nukleotidů