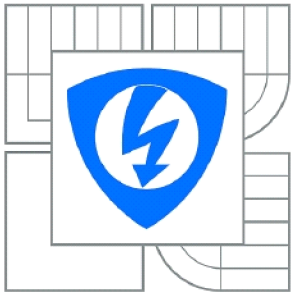
VYSOKÉ UČENÍ TECHNICKÉ V BRNĚ

BRNO UNIVERSITY OF TECHNOLOGY



FAKULTA ELEKTROTECHNIKY A KOMUNIKAČNÍCH TECHNOLOGIÍ

ÚSTAV BIOMEDICÍNSKÉHO INŽENÝRSTVÍ

FACULTY OF ELECTRICAL ENGINEERING AND COMMUNICATION

DEPARTMENT OF BIOMEDICAL ENGINEERING

VYHLEDÁVÁNÍ CPG OSTRŮVKŮ Z DNA SEKVENCÍ

CPG ISLANDS SEARCH IN DNA SEQUENCES

DIPLOMOVÁ PRÁCE

MASTER'S THESIS

AUTOR PRÁCE — Bc. VÁCLAV NERUŠIL

AUTHOR

VEDOUCÍ PRÁCE — Ing. VLADIMÍRA KUBICOVÁ

SUPERVISOR

BRNO 2014

VYSOKÉ UČENÍ TECHNICKÉ V BRNĚ

Fakulta elektrotechniky a komunikačních technologií

Ústav biomedicínského inženýrství

Diplomová práce

magisterský navazující studijní obor

Biomedicínské inženýrství a bioinformatika

Student: Bc. Václav Nerušil ***ID:*** 125064
Ročník: 2 ***Akademický rok:*** 2013/2014

NÁZEV TÉMATU:

Vyhledávání CpG ostrůvků z DNA sekvencí

POKYNY PRO VYPRACOVÁNÍ:

1) Proveďte literární rešerši o metodách vyhledávání CpG ostrůvku a významu vyhledávání CpG ostrůvků z DNA sekvencí. Pojednejte o výhodách a nevýhodách metod pro vyhledávání CpG ostrůvků z DNA sekvencí. 2) Zvolte metodu konstrukce barevného DNA spektrogramu a navrhněte metodu vyhledávání CpG ostrůvků založenou na jeho analýze. 3) Porovnejte výsledky z detekce CpG ostrůvku získaných z navrhnuté metody s výsledky vyhledávačů dostupných na internetu. 4) Zhodnoťte účinnost a využitelnost řešení.

DOPORUČENÁ LITERATURA:

[1] TAKAI, Daiya and Peter A. JONES. Comprehensive analysis of CpG islands in human chromosomes 21 and 22. Proceedings of the National Academy of Sciences, pp. 3740-3745, 2002.
[2] DIMITROVA, Nevenka, Yee H. CHEUNG and Michael ZHANG. Analysis and Visualization of DNA Spectrograms: Open Possibilities for the Genome Research. Proceedings of the 14th annual ACM international conference on Multimedia, pp. 1017-1024, 2006.

Termín zadání: 10.2.2014 ***Termín odevzdání:*** 23.5.2014

Vedoucí práce: Ing. Vladimíra Kubicová
Konzultanti diplomové práce:

prof. Ing. Ivo Provazník, Ph.D.
Předseda oborové rady

UPOZORNĚNÍ:

Abstrakt

Diplomová práce ze věnuje vyhledáváním CpG ostrůvků z DNA sekvencí na základě analýzy DNA spektrogramu. První část práce, teoretická, se zabývá významem CpG ostrůvku a popisem algoritmů, které se využívají nebo byly navrženy k jejich vyhledávání. Na teoretických podkladech byly realizovány dva algoritmy založené na analýze DNA spektrogramu. Jeden založen na předpokladu, že oblast CpG ostrůvku má vyšší obsah cytosinu a guaninu než oblast mimo CpG ostrůvek, a druhý na předpokladu vyššího frekvenčního výskytu CG dinukleotidu v oblasti CpG ostrůvku. Algoritmy jsou realizovány prostřednictvím programovacího rozhraní MATLAB. Za účelem zhodnocení využitelnosti a účinnosti řešení, jsou dosažené výsledky na zvolených DNA sekvencí realizovaných algoritmů porovnány s výsledky dosažených vyhledávači CpG ostrůvku, které jsou volně dostupné na internetu.

Klíčová slova

CpG, CpG ostrůvek, DNA spektrogram.

Abstract

This thesis focuses on searching for CpG islands of DNA sequences based on analysis of DNA spectrograms. The first part is theoretical and deals with the significance CpG island, and a description of the algorithms that are used or have been proposed for their search. The theoretical basis were implemented two algorithms based on the analysis of DNA spectrogram. One is based on the assumption that the region CpG islands has a higher content of guanine and cytosine than the region outside the CpG island and the other on the assumption of a higher frequency of occurrence of CG dinucleotides in the CpG island. The algorithms are implemented through MATLAB programming interface. For evaluation usefulness and effectiveness of solutions, results achieved on the selected DNA sequences implemented algorithms are compared with the results achieved by search engines CpG islands, which are freely available on the internet.

Key words

CpG, CpG island, DNA spectrogram.

Bibliografická citace

NERUŠIL, V. *Vyhledávání CpG ostrůvků z DNA sekvencí*. Brno: Vysoké učení technické v Brně, Fakulta elektrotechniky a komunikačních technologií, 2014. 55 s. Vedoucí diplomové práce Ing. Vladimíra Kubicová.

Prohlášení

Prohlašuji, že jsem svoji diplomovou práci na téma Vyhledávání CpG ostrůvků z DNA sekvencí, vypracoval samostatně pod vedením vedoucí semestrální 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 semestrální práce dále prohlašuji, že v souvislosti s vytvořením této práce jsem neporušil autorská práva třetích osob, zejména jsem nezasáhl nedovoleným způsobem do cizích autorských práv osobnostních a jsem si plně vědom následků porušení ustanovení § 11 a následujících autorského zákona č. 121/2000 Sb., včetně možných trestněprávních důsledků vyplývajících z ustanovení § 152 trestního zákona č. 140/1961 Sb.

V Brně dne 23.5.2014

.....

podpis autora

Poděkování

Děkuji vedoucí semestrální práce Ing. Vladimíře Kubicové za účinnou metodickou, pedagogickou a odbornou pomoc a další cenné rady při zpracování této semestrální práce.

V Brně dne 23.5.2014

.....

podpis autora

Obsah

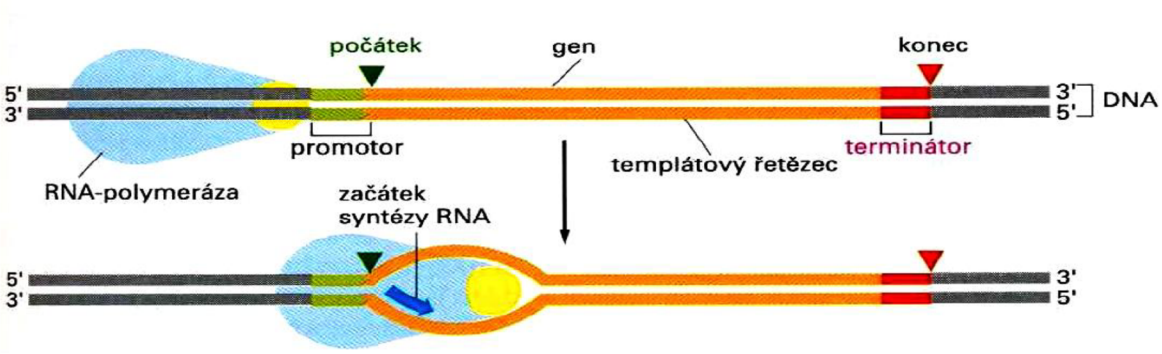
Úvod

DNA je nositelkou genetické informace, v bioinformatických datech je reprezentována čtyřmi písmeny A, C, G a T o různé délce a různé kombinaci, v závislosti na daném organismu. Písmena představují jednotlivé nukleotidy. V sekvenci DNA jsou určité úseky mající specifický účel či vlastnosti, např. exony, introny ale i CpG ostrůvky, což jsou oblasti s vyšším výskytem CG dinukleotidů, tedy CpG. CpG ostrůvek je význačný tím, že je součástí některých promotorů genů a jeho schopností omezit expresi genu vlivem řízené methylace. Ale pokud je ostrůvek methylován navzdory řízené methylaci (např. vlivem karcinogeneze), může dojít k útlumu exprese důležitého genu a k rozvoji rakoviny [2]. Analýzou DNA sekvence se věnuje vědní disciplína bioinformatika a vzhledem k současnému trendu poklesu ceny za přečtení genomu, se tak stává velmi důležitou oblastí, která má stále vysoký potenciál i do budoucnosti. V této práci je analýza DNA sekvence využita za účelem lokalizace pozic CpG ostrůvku.

Tato práce je rozdělena na tři hlavní kapitoly. První kapitola se zabývá CpG ostrůvky z pohledu významu, účelu vyhledávání a představením algoritmů metod, které byly pro jejich vyhledávání realizovány či představeny. Přestože bylo navrženo mnoho metod s různým přístupem pro vyhledávání CpG ostrůvků, je stále zapotřebí vyvíjet nové přístupy hledání CpG ostrůvků či hledat inovativní řešení, aby bylo dosaženo získání co nejpřesnějších výsledků. Ve druhé kapitole je popis algoritmů CpGI a CpGI2, které jsou součástí praktické části práce, realizace metody vyhledávání CpG ostrůvků. Jsou tedy vytvořeny dva přístupy určené k vyhledávání CpG ostrůvků. Realizace metod je založená na analýze DNA spektrogramu. DNA spektrogram má podobu dvojrozměrného barevného obrazu a je užitečným nástrojem ke zviditelnění repetičních vzorů v DNA sekvenci [4]. Algoritmus CpGI analyzuje DNA spektrogram na předpokladu, že CpG ostrůvek má vyšší zastoupení C a G oproti oblastem mimo ostrůvek. Druhý algoritmus CpGI2 realizuje a analyzuje DNA spektrogram na základě faktu, že CpG ostrůvek je tvořen CpG s vyšším frekvenčním zastoupením než oblasti mimo ostrůvek. Dosažené výsledky navržených metod na testovacích DNA sekvencích jsou ve třetí kapitole porovnány s výsledky osmi vyhledávačů, které jsou volně dostupné na internetu.

1. Metody a význam vyhledávání CpG ostrůvku v DNA sekvenci

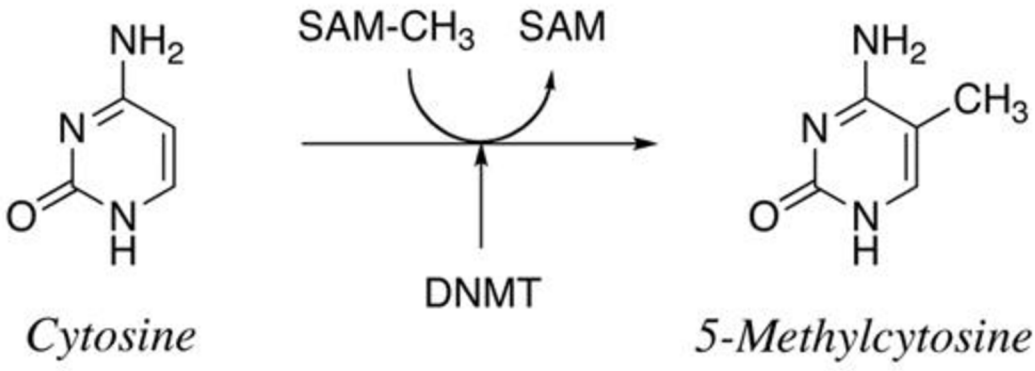
CpG je v podstatě dinukleotid CG a zařazením písmena „p“odlišen od párovaného CG a je zkratkou C – fosfát – G. Oblasti v genomu, které mají vyšší zastoupení těchto CpG se označuje jako CpG ostrůvek. Ty bývají často přidružené v místech výskytu (až u 70%) promotorů lidských genů. Promotor je úsek DNA sekvence obvykle umístěn před konkrétním genem a hraje důležitou roli v transkripčním aparátu (viz. Obr. 1). Je to tedy významné místo v DNA obsahující informaci o začátku transkripce, váže se zde molekula RNA polymeráza, která slouží k přepisu DNA sekvence do řetězce RNA [8]. Vyhledání CpG ostrůvků v DNA sekvenci za účelem predikce výskytu promotoru a tedy i počátku genu v DNA může být vhodnou alternativou k algoritmům využívající rozpoznávání funkční sekvence např. TATA box. [9]



Obr. 1: Struktura genu v molekule DNA během transkripce. CpG ostrůvek přidružený k promotoru může sloužit k predikci lokality genu. Převzato z [8].

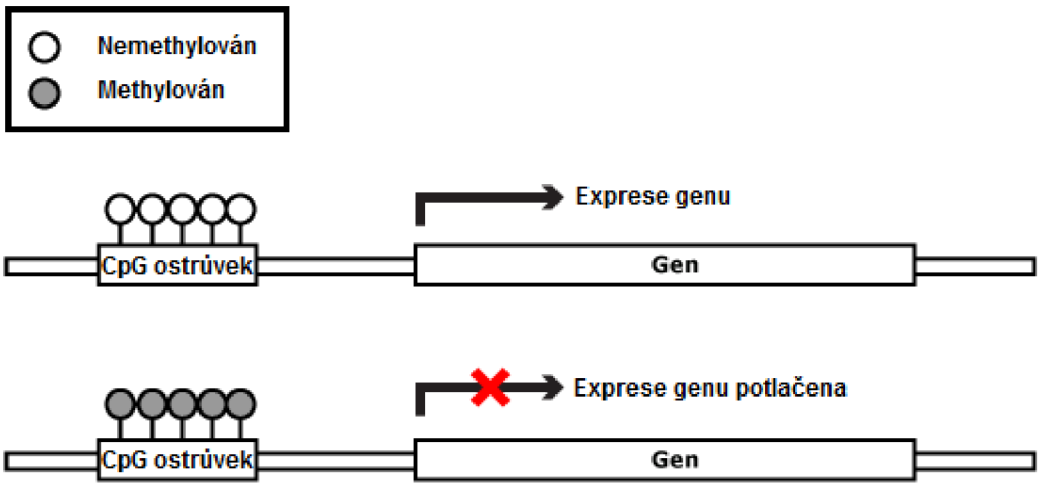
Pro CpG ostrůvky je významná modifikace cytosinu procesem methylace. Methylací dochází ke změně chemické struktury cytosinu na 5 – methylcytosin a nejčastěji se nachází na cytosinu v CpG (podrobněji viz Obr. 2). Potom CpG ostrůvek jenž je bohatý na výskyt CpG je lehce podléhající methylaci, avšak zde dochází k řízené methylaci. Bylo prokázáno, že methylace na cytosinu u CpG se vyskytují v inaktivních genech a tlumí tak expresi konkrétního genu (viz. Obr. 3). Methylace na CpG ostrůvku neumožní transkripčním proteinům nasednout na specifické místo v DNA k zahájení transkripce nebo transkripce není umožněna proteiny MBP (methyl-CpG binding proteins), které se

na methylované místo DNA sekvence navážou. CpG ostrůvek také v místech promotoru genu zastává důležitou roli v procesech jako je inaktivace na X-chromozomu. [3] [10], [11]



Obr. 2: Proces přeměny cytosinu na 5-methylcytosin pomocí enzymu methyltransferázy (DNMT), která zprostředkovává přenos methylové skupiny CH3 z S-adenosylmethionin (SAM) do polohy pátého uhlíku cytosinu. Převzato z [10].

Avšak pokud dojde k neřízené methylaci CpG ostrůvku vlivem karcinogeneze, může dojít např. k zamezení exprese tumor supresorového genu a následně k rozvoji rakoviny. Také bylo prokázáno, že mohou mít vliv i methylace CpG ostrůvku lokalizovaných mimo promotor genu, vzdálených až 2 kbp tzv. „CpG island shores“. [13] Na jiných místech lokalizované CpG ostrůvky, např. v místech exonu je biologická funkce neznámá. [2]



Obr. 3: Methylace na CpG ostrůvku a její vliv na expresi genu.

Dosavadní znalosti o CpG ostrůvku jsou dostačující motivací vyvíjet různé přístupy vyhledávání. V současné době existuje několik různých algoritmů pro vyhledávání CpG ostrůvků, které lze shrnout do několika skupin. Skupina kriteriálních metod hodnotí parametry CpG ostrůvku nejčastěji pomocí plovoucího okna. Parametry jsou délka CpG ostrůvku, procentuální obsah cytosinu a guaninu poměr získané a očekávané hodnoty výskytu CpG (Obs_{CpG}/Exp_{CpG}), pro který platí následující vztah:

$$Obs_{CpG}/Exp_{CpG} = \frac{Počet\ CpG}{Počet\ C \cdot Počet\ G} \cdot N, \tag{2.1}$$

kde *N* vyjadřuje celkovou délku zkoumaného úseku v sekvenci DNA. [1]

Další metody vyhledávání CpG ostrůvků jsou algoritmy, které se zabývají statistickými vlastnostmi, využívající statistického stavového modelu či metody hledání nejpravděpodobnější cesty. Další přístupy vyhledávání CpG ostrůvku je metoda založená na analýze DNA spektrogramu nebo metoda založená na optimalizačních algoritmech.

Cílem algoritmů vyhledávající CpG ostrůvky je lokalizace všech ostrůvků v analyzované sekvenci, určení pozic začátku a konce a pokročilé metody kategorizují nalezené CpG ostrůvky do specifických skupin.

1.1. Definice CpG ostrůvku Gardiner-Gardena a Frommera

Metoda určení CpG ostrůvku byla navržena Gardiner-Gardenem a Frommerem, v podstatě se také jedná o všeobecně přijímanou definici CpG ostrůvku. Aby byla určitá oblast prohlášena za CpG ostrůvek musí splňovat několik kritérií. CpG ostrůvek je definován o minimální délce 200 bp, musí obsahovat C+G více jak 50 % a poměr získané a očekávané hodnoty výskytu CpG (Obs_{CpG}/Exp_{CpG}) musí být větší nebo roven hodnotě 0,6. [1]

Nedostatkem této definice CpG ostrůvku je, že může dojít k problému rozlišení mezi CpG ostrůvkem a Alu elementem. Na tento problém poukazují autoři v článku [2], kde jako příklad uvádějí Alu repetice v lidském genomu, dlouhé přibližně 280 bp. Přičemž některé z nich mohou splňovat kritéria pro určení CpG ostrůvku. Tedy procentuelně vysoký obsah C+G a poměr Obs_{CpG}/Exp_{CpG}.

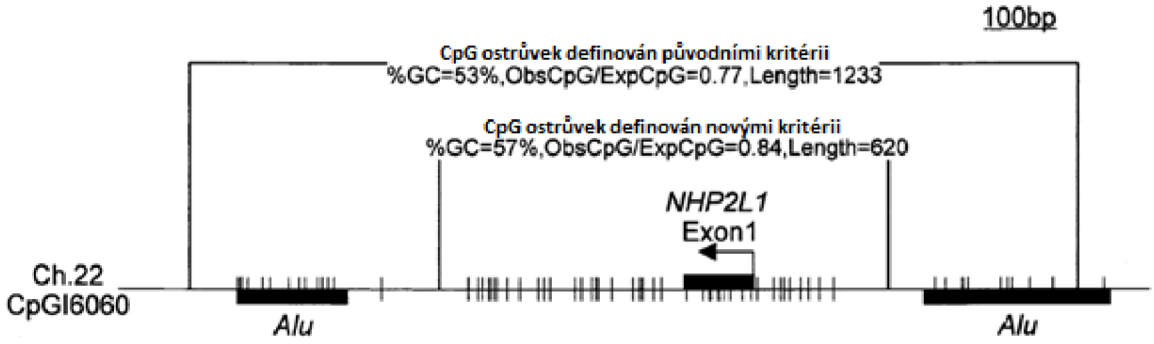
Na základě těchto kriterií určení CpG ostrůvku bylo navrženo spousta algoritmů, obvykle využívají plovoucího okna podél analyzované sekvence, například na internetu dostupná aplikace Cpgplot [19].

1.2. Modifikované kritéria Gardiner-Gardena a Frommera

Autoři Takai a Jones v článku [2] popisují metodu hledání a popisu CpG ostrůvků, přičemž poukazují na nevýhody určování CpG ostrůvků metody autorů v článku [1] a metodu optimalizují k dosažení lepších výsledků, především k eliminaci falešně pozitivní detekce CpG ostrůvků. Účelem modifikace Gardiner-Gardena a Frommerova kritéria jejich metodou má vyloučit Alu repetice a mnoho CpG ostrůvků, které nejsou umístěny v promotorech genů. CpG ostrůvek musí splňovat délku rovnu nebo větší než 500 bp, a dále musí obsahovat C+G více jak 55 % a hodnota Obs_{CpG}/Exp_{CpG} musí být větší nebo rovna hodnotě 0,65. K analýze použili kompletní genomové sekvence chromozomů 21 a 22 člověka. Realizovaný algoritmus vyhledávače je volně dostupný na internetu s názvem CpGIS.

Algoritmus je realizován v několika krocích, plovoucí okno pro délku 200 bp se posouvá po sekvenci vždy při kroku 1 bp dokud oblast nesplňuje dané podmínky definice CpG ostrůvku (%CG a Obs_{CpG}/Exp_{CpG}). Když jsou tyto podmínky splněny, hodnotí se následujících 200 bp, když i tato oblast splňuje dané podmínky (přiřadí se k původnímu oknu), krok se opakuje, dokud podmínky splněny nejsou. V tomto případě se plovoucí okno posouvá směrem k 5' konci dokud okno nesplní daná kritéria. Současným výsledkem je tedy oblast sekvence začínající na pozici prvního okna a končící na pozici konce posledního okna. Pro tento celý úsek jsou posouzeny podmínky (%CG a Obs_{CpG}/Exp_{CpG}), a pokud nejsou splněny, je úsek upraven postupným krácením začátku a konce vždy o 1 bp, až jsou splněny podmínky pro CpG ostrůvek. Za předpokladu, že jsou dva individuální CpG ostrůvky vzájemně vzdáleny 100 bp nebo méně, potom jsou tyto ostrůvky spojeny a hodnoty %CG a Obs_{CpG}/Exp_{CpG} přepočítány. [2]

V článku [2] autoři poukazují na výhodu jejich modifikace kritérií, která vede ke zlepšení kvality lokalizace CpG ostrůvku. Jak je vidět na Obr. 4, po úpravě kritérií byl CpG ostrůvek lokalizován bez Alu repetic. Po celkové analýze chromozomu 21 a 22 člověka zredukovali celkový počet CpG ostrůvků z 14062 na 1001, což je reálnější množství při předpokládaných cca 750 genů na oněch chromozomech.



Obr. 4: Zlepšení lokalizace CpG ostrůvku modifikací kriterií. Převzato z [2].

1.3. CpGProD

CpGProD je algoritmus určený pro vyhledávání CpG ostrůvků, a k ohodnocení ostrůvků za účelem stanovení pravděpodobnosti, že jsou lokalizovány právě v promotorech genu. Metoda je autory popsána v článku [9] a sestává ze dvou hlavních kroků. První krok spočívá ve vyhledání všech CpG ostrůvku založenou na Gardiner-Gardena a Frommerovy definice (viz kapitola 1.1), tedy úsek musí obsahovat C+G více jak 50 % a Obs_{CpG}/Exp_{CpG} musí být větší nebo roven hodnotě 0,6. Avšak autoři upravili kritérium délky CpG ostrůvků místo 200 bp na 500 bp a ostrůvky vzájemně vzdálených méně než 200 bp jsou sloučeny.

Ve druhém kroku algoritmu jsou nalezené CpG ostrůvky hodnoceny výpočtem pravděpodobnosti, že ostrůvek je součástí promotoru. Tato pravděpodobnost je vypočtena z délky ostrůvku, frekvence výskytu C+G a z Obs_{CpG}/Exp_{CpG}, pomocí vztahu, který byl stanoven na základě modelu složeného z dat se známými začátky ostrůvků. [17]

Přestože je metoda určená především k lokalizaci oblasti promotoru v DNA sekvencích, tak nalezené CpG ostrůvky mohou spolehlivě sloužit k porovnání výsledků s jinými metodami, např. autoři metody [16] nebo [18] využívají metodu CpGProD pro porovnání výsledků získané jejich metodou.

1.4. Metoda založená na Markovském řetězci

Markovský řetězec je označován za nejjednodušší typ z Markovských procesů. Definice tohoto stochastického procesu předpokládá diskrétní množinu hodnot, které pak odpovídá výsledná diskrétní množina výsledků neboli stavů. Pro pravděpodobnosti přechodu p_{ij}, platí, že ze stavu *i* do stavu *j* dojde v době od *n* do *n+1* a tím také přechod z proměnné X_n k proměnné X_{n+1}. Dále pravděpodobnosti přechodu musí splňovat podmínky

$$0 \leq p_{ij} \geq 1, \quad i,j = 0,1,2,\ldots \tag{2.2}$$

$$\sum_{i=0}^{+\infty} p_{ij} = 1, \quad i,j = 0,1,2,\ldots \tag{2.3}$$

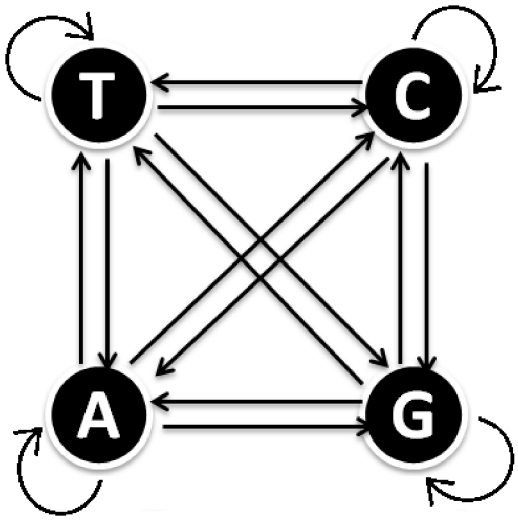
Jednotlivé pravděpodobnostní přechody jsou reprezentovány maticí označované jako pravděpodobnostní matice přechodu, v obecném tvaru:

$$\mathbb{P} = \begin{pmatrix} p_{01} & p_{02} & p_{03} & \cdots \\ p_{10} & p_{11} & p_{12} & \cdots \\ p_{20} & p_{21} & p_{23} & \cdots \\ \cdots & \cdots & \cdots & \cdots \end{pmatrix},$$

kde pro jednotlivé pravděpodobnostní přechody platí dle rovnic (2.2) a (2.3), že jsou vždy nezáporné a součet v řádku je roven jedné. [6]

Markovský řetězec lze uplatnit při hledání CpG ostrůvku, popis metody je uveden v článku [5]. Jednotlivé stavy jsou reprezentovány pozicí báze A, C, G nebo T a dále musí být stanoveny pravděpodobnostní přechody ze stavu i do stavu j.

Obr. *5* graficky reprezentuje strukturu onoho Markovského řetězce a také znázorňuje jednotlivé pravděpodobnostní přechody pomocí šipek.



Obr. 5: Grafická struktura Markovského řetězce pro všechny možné přechody.

Nejprve je zapotřebí určit parametry Markovského řetězce, tedy pravděpodobnostní přechody. Mohou nastat dva různé pravděpodobnostní stavy při přechodu od původní báze k aktuální. Jedním je pravděpodobnost, že CpG ostrůvek je a pravděpodobnost, že CpG ostrůvek není, přitom je očekávána vysoká pravděpodobnost výskytu ostrůvku mezi stavy C a G. Tedy, je zapotřebí mít dvě pravděpodobnostní matice přechodů. Na Obr. 6 je uveden příklad matice v konkrétních hodnotách.

$$P^{+} = \begin{bmatrix} & A & C & G & T \\ A & 0.18 & 0.27 & 0.43 & 0.12 \\ C & 0.17 & 0.37 & 0.27 & 0.19 \\ G & 0.16 & 0.34 & 0.37 & 0.13 \\ T & 0.08 & 0.36 & 0.38 & 0.18 \end{bmatrix} \quad P^{-} = \begin{bmatrix} & A & C & G & T \\ A & 0.30 & 0.20 & 0.29 & 0.21 \\ C & 0.32 & 0.30 & 0.08 & 0.30 \\ G & 0.25 & 0.25 & 0.29 & 0.21 \\ T & 0.18 & 0.24 & 0.29 & 0.29 \end{bmatrix}$$

Obr. 6: Možné hodnoty pravděpodobnostních matic přechodů, převzato z [5].

Máme-li libovolný úsek sekvence *x* o délce *n*, je spočítána *p(x)* pro každý řetězec, tedy $p(x|+)$ a $p(x|-)$, pak pomocí výpočtu poměru šancí vztahem

$$\log\frac{p(x|+)}{p(x|-)} = \sum_{i=1}^{n-1} \log\frac{p^{+}_{x_i x_{i+1}}}{p^{-}_{x_i x_{i+1}}} \tag{2.4}$$

je určena hodnota, která když splní podmínku $\log\frac{p(x|+)}{p(x|-)} > 0$, je úsek x prohlášen za CpG ostrůvek v opačném případě daný úsek není CpG ostrůvkem.

Vyhledáním CpG ostrůvků na základě Markovských modelů se zabývají autoři článku [12], využívají modely druhého a třetího řádu a také variabilního řádu, které jsou založeny na pravděpodobnostním „suffix tree“. Přičemž při různém řádu Markovského řetězce dosáhli různé přesnosti identifikace CpG ostrůvků. Obecně výše popsaná metoda určuje, zda zkoumaný úsek je anebo není CpG ostrůvek, je tedy zapotřebí (obdobně jako u kriteriálních metod) zvolit okno podél sekvence, které definuje zkoumanou oblast a pro tuto oblast je vždy nutné výpočtem určit zda úsek je CpG ostrůvek. Okno může být délky 100 [5] a následně jsou pospojována místa sekvencí splňující podmínky kladné hodnoty poměru šancí. Tuto nevýhodu lze vyřešit propojením obou řetězců do jednoho modelu, využitím skrytých Markovských modelů a hledání nejpravděpodobnější cesty Viterbiho algoritmem.

1.5. Skrytý Markovský model

Skrytý Markovský model označovaný zkratkou HMM (Hidden Markov Model) je pravděpodobnostní stavový model, který lze využít k hledání CpG ostrůvků. Pro něž je definován jako

$$HMM = (\Sigma, Q, P, e), \tag{2.5}$$

kde Σ= *{A,G,C,T}*, Q vyjadřuje množinu stavů, tedy Q=*{A+,G+,C+,T+,A-,G-,C-,T-}*, symbol „+“ reprezentuje stavy CpG ostrůvku a symbol „-“ stavy mimo oblast CpG ostrůvku, potom pro každý nukleotid jsou dva stavy. Dále P je pravděpodobnostní matice přechodů mezi jednotlivými stavy a $e_k(b)$ je emisní pravděpodobnost pro každé $k \in Q$ a $b \in \Sigma$. Tento HMM je vlastně sloučen z Markovských řetězců uvedených v předchozí kapitole do jednoho modelu s nízkou pravděpodobností přechodu mezi řetězci. [5],[14]

Pro rozlišení mezi posloupnosti stavů a posloupnosti písmenné abecedy x je posloupnost stavů nazvána cestou, označena π a je skryta. Je tedy zkoumána posloupnost x, kde jsou jednotlivé symboly emitovány ze stavu s určitou emisní pravděpodobností. Pak platí vztah

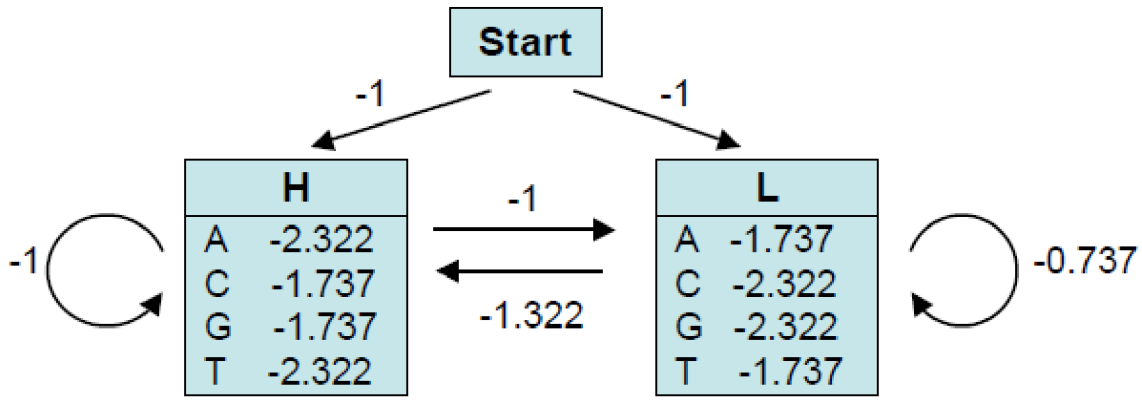
$$e_k(b) = \mathbb{P}(x_i = b|\pi_i = k), \tag{2.6}$$

kde $i \in (1,2 \ldots L)$ a vyjadřuje, že právě symbol b z posloupnosti x je ve stavu k. Snahou je získat z posloupnosti symbolů posloupnost stavů, tedy cesty, ve které jsou části sekvence klasifikovány jako CpG ostrůvek a nazývá se dekódování, to je pak řešeno s cílem najít nejvíce pravděpodobnější cestu, kterou lze řešit např. algoritmem dynamického programování. [14]

Praktická realizace metody vyhledávání CpG ostrůvků na základě HMM je popsána v článku [15]. Autoři poukazují na výhody metody oproti klasickým kriteriálním metodám, které mají nižší přesnost detekce způsobenou opakujícími elementy. Metodu HMM spojili s Baum-Welch algoritmem, který je používán k odhadu neznámých parametrů, a pro hledání nejpravděpodobnější cesty použili Viterbiho algoritmu.

1.6. Vyhledávání Viterbiho algoritmem

Viterbi algoritmus (navrhl jej Andrew Viterbi) je řazen mezi dynamické programování, které řeší problémy rozkladem na podproblémy. Algoritmus využívá skrytých Markovských modelů. V případě algoritmu pro vyhledávání CpG ostrůvku pak využívá dva stavy, stav **H** označující vysoký obsah páru CG (CpG) a naopak stav **L** označující nízký obsah CpG. Na Obr. 7 je pravděpodobnostní model skrytého Markovského modelu jako praktickou ukázkou. Hodnoty pravděpodobnosti jsou přepočteny pomocí logaritmu $log_2(p)$ za účelem snadnějšího výpočtu, malé hodnoty pravděpodobností jsou nepraktické, počítá se s nimi obtížněji a také násobení přechází na sčítání. [5], [6]



Obr. 7: Pravděpodobnostní model skrytého Markovského modelu, přičemž hodnoty pravděpodobností jsou přepočteny pomocí logaritmu $log_2(p)$.

Základní rovnice pro výpočet jednotlivých stavů je ve tvaru:

$$p_l(i,x) = e_l(i) \cdot \max[p_k(j, x-l) \cdot p_{kl}], \tag{2.7}$$

kde l je stav, který je počítán (**H** nebo **L**), $e_l(i)$ vyjadřuje pravděpodobnost elementu i ve stavu l a k nabývá jednotlivě všech hodnot stavů. Protože máme dva stavy, bude maximum vybíráno z dvou hodnot. Jak z uvedené rovnice vyplývá, algoritmus postupně počítá hodnoty pravděpodobností pro každý ze stavů a určují tak, zda konkrétní nukleotid na dané pozici je součástí anebo není součástí CpG ostrůvku. Tam, kde je nalezena vyšší míra pravděpodobnosti, je potom náležící stav součástí nejpravděpodobnější cesty. Ta je potom hledána tzv. zpětným trasováním.

1.7. CpGcluster

CpGcluster je název metody algoritmu vyhledávání CpG ostrůvků, založenou na fyzické vzdálenosti mezi sousedními CpG. Metodu navrhli autoři článku [16] a získané výsledky porovnali s jinými běžně používanými vyhledávači ostrůvků. Jako přední výhodu této metody autoři poukazují na nízký překryv mezi CpG ostrůvky a Alu retrotranspozomy, vysokou přesnost a specificitu a také nízkou mírou falešně pozitivních predikcí oproti jiným metodám. Jako další předností metody uvádí lepší odlišení CpG ostrůvků promotoru od zbylých ostrůvků, dále rychlost výpočtu díky operací s celými čísly a také, že každá lokalizace CpG ostrůvku začíná a končí CpG dinukleotidem.

Algoritmus metody CpGcluster sestává ze dvou základních kroků, vyhledávání shluků CpG v sekvenci a přiřazování *p-hodnoty* každému nalezenému shluku za účelem určení shluku jako CpG ostrůvek.

První krok je založen na statistických vlastnostech fyzické vzdálenosti mezi sousedními CpG dinukleotidy v sekvenci DNA. Za předpokladu, že by CpG byly podél sekvence rozmístěny zcela náhodně, pro vzdálenosti mezi sousedními CpG by platil vztah geometrického rozdělení:

$$P(d) = (1-p)^{d-1}p, \tag{2.7}$$

kde *P(d)* představuje pravděpodobnost nalezení vzdálenosti *d* mezi sousedními CpG a *p* je pravděpodobnost všech CpG v sekvenci, získaná jako poměr mezi všemi CpG a celkovým počtem dinukleotidů v sekvenci DNA. Avšak v CpG ostrůvku je vyšší míra výskytu CpG, tak tedy mají i kratší vzdálenosti mezi CpG dinukleotidy. To je také zásadní předpoklad algoritmu této metody.

Algoritmus na začátku zaznamenává pozici všech CpG v DNA sekvenci, vytvoří se tak množina $(x_1,x_2...x_N)$, kde *N* je počet všech CpG a vypočítá fyzickou vzdálenost mezi sousedními CpG podle vztahu

$$d_i = x_{i+1} - x_i - 1. \tag{2.8}$$

Dále jsou tyto vzdálenosti postupně hodnoceny, a pokud první vzdálenost spadá pod předem určený práh (odvozený z distribuce vzdáleností mezi CpG), stává se prvním shlukem. Postupně se tak přidávají další CpG, které splňují vzdálenost spadající pod danou mez. Pokud je vzdálenost následujícího CpG větší než stanovená mez, dojde k ukončení prvního shluku. Tímto způsobem jsou postupně hodnoceny všechny vzdálenosti a jsou nalezeny všechny shluky.

Pro nalezené shluky je v dalším kroku přiřazení *p-hodnoty,* tedy pravděpodobnosti, že se shluk objeví v náhodné sekvenci. Pouze ty shluky, které jsou dostatečně statisticky významné, jsou označeny za CpG ostrůvky, tj. *p-hodnota* splňuje hodnotu v daných mezích.

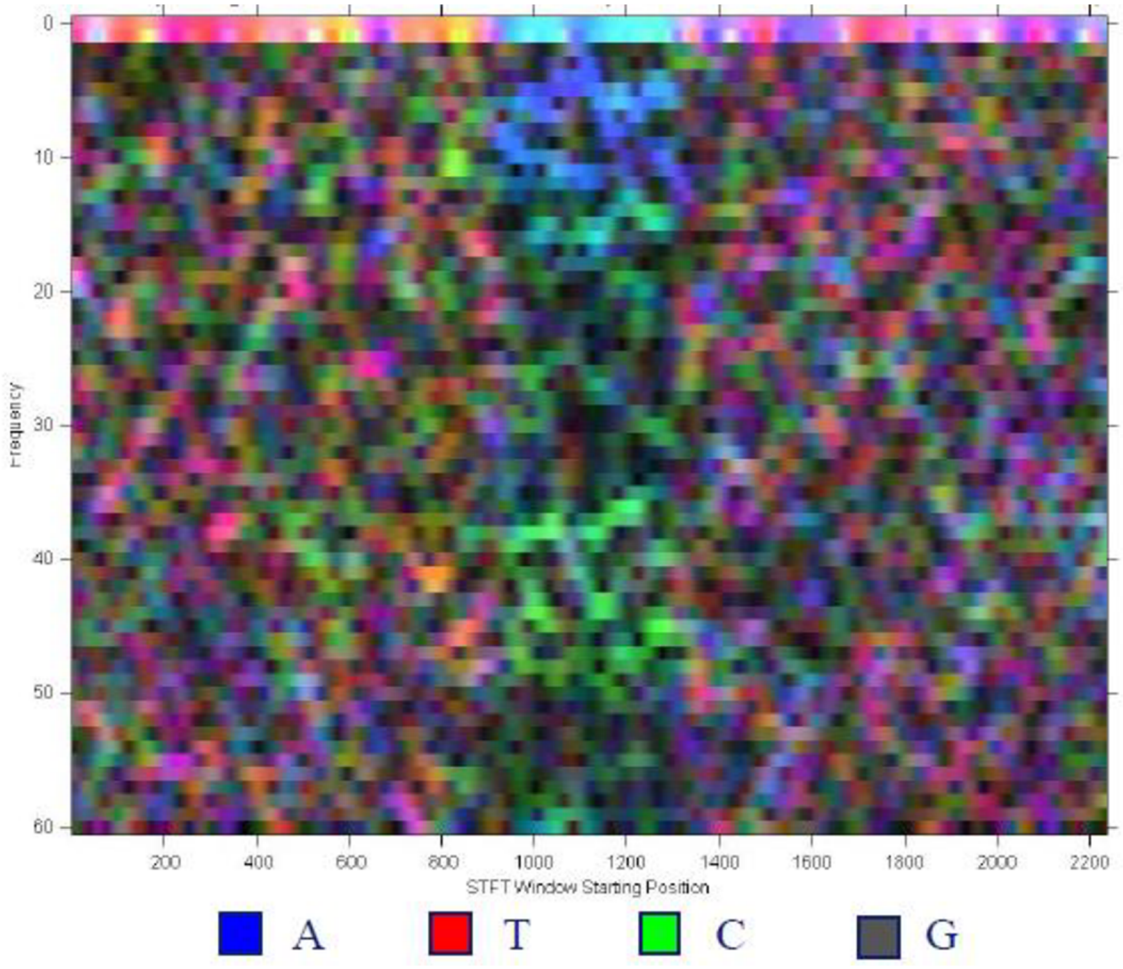
1.8. Analýza DNA spektrogramu

Autoři článku [4] popisují metodu hledání CpG ostrůvků na základě analýzy DNA spektrogramu, což je inovativní způsob, jak řešit některé problematiky v bioinformatice. Reprezentace DNA sekvence prostřednictvím spektrogramu jako barevného obrázku umožňuje lépe analyzovat specifické oblasti než při klasické reprezentaci znaky. [21]

Na začátku je samotná realizace DNA spektrogramu, která je provedena v několika krocích. Nejprve převod DNA sekvence 4D binární numerické reprezentace, jejímž výsledkem jsou čtyři vektory čtyř nukleotidových bází. Tedy pro každý vektor pro určitou bázi platí, že na pozici výskytu oné báze je přiřazena hodnota 1, v opačném

případě 0. Frekvenční spektrum pro DNA sekvenci je získáno pomocí výpočtu krátkodobé Fourierovy transformace (STFT), autoři popisovaného algoritmu použili velikost okna 120 při posunu o 1 pozici. Při skládání spektra jsou jednotlivé DFT hodnoty mapováním převedeny do RGB barev a normalizovány. Výsledný DNA spektrogram úseku DNA s výskytem CpG ostrůvku je zobrazen na

Obr. 8. Již na první pohled je z obrázku možné odhadem určit oblast, která představuje výskyt CpG ostrůvku.



Obr. 8: DNA spektrogram pro chr21-9905604-9907958 s výskytem CpG ostrůvku. Okno STFT: 120, interval posuvu okna: 1. Převzato z [4].

Samotnou analýzu DNA spektrogramu k hledání CpG ostrůvku autoři článku [4] založili na algoritmu založeného na porovnání dvou DNA spektrogramů. Jeden reprezentující pouze výskyt C a G vyjádřené zelenou barvou a druhý červený spektrogram reprezentující výskyt A a T v sekvenci. Hlavním předpokladem pro výskyt CpG ost-

růvku je, že intenzita zelené barvy je vyšší než červené v dané oblasti. Oba spektrogramy jsou zvlášť podrobeny hranovému detektoru, jehož výstupem je binární obraz. Hrany odpovídají pixelům, které jsou odlišeny různou intenzitou od sousedních. Pro výskyt CpG ostrůvku pak platí, že existuje více hran v zelené oblasti než v červené. Pixely detekovaných hran jsou v obou případech sečteny podél osy *x* (pozice STFT okna) a osy *y* (spektrální frekvence). V posledním kroku se provádí hodnocení histogramů a podle stanovených kritérií jsou určeny oblasti CpG. Za CpG ostrůvek je označena ta oblast, která splňuje, že rozdíl mezi zeleným a červeným histogramem (podél osy *x*) je větší než stanovený práh a současně je oblast delší než 200 bp. Jsou-li definované ostrůvky vzdáleny od sebe maximálně 100 bp, jsou sloučeny do jednoho.

Autoři této metody podle dosažených výsledků hodnotí svůj algoritmus jako striktnější oproti definici metody Gardiner-Gardena a Frommera a také i metody Takai a Jones. Jedním z nedostatků této metody může být, že definice CpG ostrůvku je založena na jeho délce a na výskytu nukleotidů C a G v sekvenci a přitom není brán ohled na podstatu dinukleotidu CpG.

1.9. CpGPAP

Za nejnovější metodu k vyhledávání CpG ostrůvků v DNA sekvencích je v této kapitole popisující metody vyhledávání aplikace CpGPAP (zkratka z anglického názvu CpG Island Predictor Analysis Platform), která byla představena v roce 2012 v článku [18]. Aplikace je volně přístupná na internetu a poskytuje přehledné rozhraní, které umožňuje analyzovat DNA sekvence za účelem predikce CpG ostrůvku metodami autorů, ale také poskytuje jiné vyhledávače jiných autorů (CpGIS, CpGProD a Cpgplot) za účelem umožnit uživateli porovnání dosažených výsledků různými algoritmy. Autoři navrhli dva samostatné algoritmy vyhledávání. Jeden CpGPSO založený na optimalizačním přístupu rojení částic a druhý CpGGA založený na genetickém algoritmu.

CpGPSO metoda využívá k predikci CpG ostrůvků optimalizačního algoritmu Rojení částic (PSO - Particle Swarm Optimalization). Který simuluje chování ptačího hejna prohledávajícího okolí se snahou najít nejvyšší vrchol. O vrcholu hejno neví, avšak po každém kroku vědí, který jedinec byl na dosud nejvyšším vrcholu a ostatní se jej snaží následovat. V prvním kroku je inicializace počátečních částic, s pozicemi začátku (*Fs*) a konce (*Fe)* v DNA sekvenci, tedy úseky DNA a pro každou částici je

určen vektor rychlosti, který pro následující iteraci udává směr pohybu (změnám pozic v DNA sekvenci). Ve druhém kroku je hodnocena kvalita *pbest* každé částice. Základem hodnocení je délka a kritéria, kdy CpG ostrůvek musí obsahovat C+G více jak 50 % a poměr získané a očekávané hodnoty výskytu CpG (Obs_{CpG}/Exp_{CpG}) musí být větší nebo roven hodnotě 0,6 a tedy dle kritéria viz [1]. Hodnota nejkvalitnější částice je označena jako *gbest* a označuje jako nejlepší hodnotu vůbec a mění se v dalších iteracích za hodnotu *pbest*, pokud je hodnota *gbest* horší než *pbest*. Pokud však hodnota *gbest* se nezlepší po pěti iteracích, pak náhodně vybraná polovina populace částic je nahrazena jejich komplementárními částicemi. V dalším kroku je upraven vektor rychlosti podle vztahu

$$v_{id}^{new} = w \times v_{id}^{old} + c_1 \times r_1 \times (pbest_{id} - x_{id}^{old}) + c_2 \times r_2 \times (gbest - x_{id}^{old}), \tag{2.9}$$

kde c_1 a c_2 jsou učící faktory, které jsou předem stanoveny a r_1 a r_2 jsou náhodně generované hodnoty v intervalu <0,1>. Potom jsou upraveny pozice *x* každé částice podle vztahu

$$x_{id}^{new} = x_{id}^{old} + v_{id}^{new}. \tag{2.10}$$

Ukončení algoritmu je předem stanoveno na počtu iterací. Výstupní pozice CpG ostrůvku jsou hodnoty pozic x_{id} každého jedince v poslední iteraci splňující podmínky definice CpG ostrůvků.

CpGGA využívá k predikci CpG ostrůvků Genetický algoritmus (GA), což je stochastický algoritmus, který se řadí mezi evoluční výpočetní techniky a je založen na přirozeném výběru, obdobně jako v přírodní evoluci. Algoritmus CpGGA je založen na dvojrozměrné funkci, na první pozici F_s CpG ostrůvku a jeho délky F_l. Na počátku dlouhou sekvenci rozdělí na několik menších úseků s náhodnými pozicemi začátků a délky v rozsahu 200-2000 bp. Základem hodnocení je opět délka a kritéria, kdy CpG ostrůvek musí obsahovat C+G více jak 50 % a poměr získané a očekávané hodnoty výskytu CpG (Obs_{CpG}/Exp_{CpG}) musí být větší nebo roven hodnotě 0,6. V algoritmu následují typické kroky genetického algoritmu jako je křížení mezi jedinci, mutace a turnajový výběr. Do další generace postupují pouze dva nejlepší jedinci, tedy z výběru dvou rodičů z minulé generace a dvou jejich potomků. Pokud se hodnocení kvality nezmění po pěti krocích, dojde ke změně hodnot F_s a F_l za jejich komplementární

hodnoty. Celý algoritmus je ukončen na základě předem stanoveného počtu iterací, potom výsledné hodnoty F_s a F_l jsou určeny poslední iterací a určují tak pozice všech CpG ostrůvku v analyzované DNA sekvenci, pokud splňují podmínky definice CpG ostrůvků.

Algoritmy CpGPSO a CpGGA jsou doplněny přístupem Reinforcement Learning (RL), jehož cílem je zlepšit rychlost učení, maximalizovat pozitivní výsledky a negativní minimalizovat.

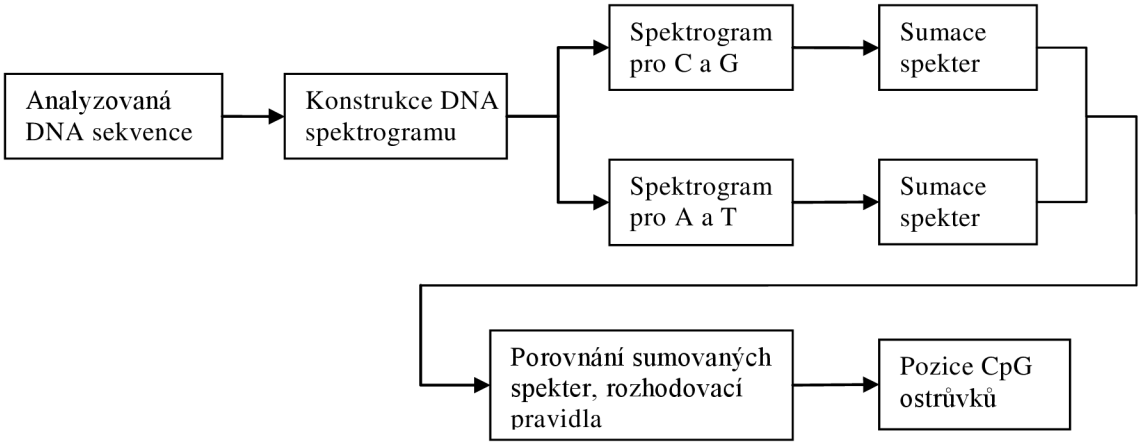
Autoři CpGPAP otestovali jejich algoritmy (CpGPSO, CpGGA) a výsledky porovnali s výsledky algoritmů jiných autorů (CpGcluster, CpGProd, Cpgplot a CpGIS) na DNA sekvenci chromozomů 21 a 22 člověka. Podle dosažených výsledků hodnotí, že algoritmy CpGPSO a CpGGA dosahují vyšší sensitivity než ostatní porovnávané algoritmy. Oba algoritmy hodnotí jako výpočetně méně náročné než ostatní porovnávané.

2. Realizace algoritmu vyhledávání CpG ostrůvku

Kapitola se zabývá realizací algoritmu vyhledávání CpG ostrůvku z DNA sekvencí, na základě analýzy DNA spektrogramu v programovacím prostředí MATLAB. Byly realizovány dva přístupy, jeden založen na předpokladu, že CpG ostrůvek má vyšší zastoupení obsahu C a G obdobně jak bylo popsáno v kapitole 1.8, soubor ke spuštění tohoto přístupu je součástí elektronické přílohy pod názvem *CpGI.m*. Druhý přístup byl založen na faktu, že CpG se v ostrůvcích vyskytují s vyšší frekvencí než na částech sekvence DNA bez výskytu CpG ostrůvku. Soubor pro spuštění je v elektronické příloze pod názvem *CpGI2.m*. Oba algoritmy lze rozdělit na dva základní kroky: konstrukce spektrogramu z DNA sekvence a jeho analýzu s rozhodovacími pravidly pro určení CpG ostrůvku. Vstupem obou programů je analyzovaná DNA sekvence ve formátu FASTA a výstupem jsou pozice začátků a konců detekovaných CpG ostrůvků.

2.1. CpGI

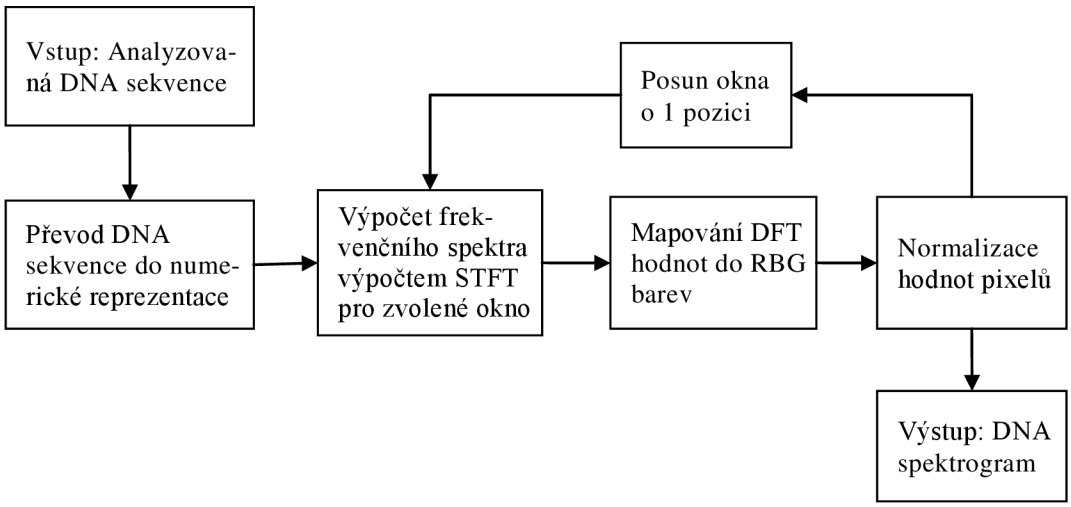
Je název metody vyhledávání CpG ostrůvků na základě předpokladu, CpG ostrůvek má vyšší zastoupení obsahu C a G než oblasti sekvence DNA neobsahující CpG ostrůvek. Blokové schéma algoritmu je uvedeno na Obr. 9.



Obr. 9: Blokové schéma pro navržený vyhledávač CpG ostrůvků CpGI.

2.1.1. Konstrukce DNA spektrogramu

DNA spektrogram je dvojrozměrným obrazem, ve kterém souřadnice na ose *y* odpovídá frekvenci a souřadnice na ose *x* odpovídá pozicím začátku plovoucího okna v sekvenci DNA. Samotný algoritmus konstrukce je realizován jako funkce pod názvem *spektrogram.m.* Realizace spektrogramu je rozdělena na několik postupných kroků. Blokové schéma algoritmu konstrukce DNA spektrogramu je uveden na Obr. 10. Výsledný spektrogram závisí na několika parametrech, na výběru délky STFT okna a jeho překryv následujícího okna a na způsobu mapování do RGB barev. Volba délky okna by měla být volena několikrát delší než je délka opakujícího se hledaného vzoru, ale současně by měla být menší než je délka hledané oblasti obsahující vzor. [4], [21]



Obr. 10: Blokové schéma konstrukce DNA spektrogramu.

Převod DNA sekvence do numerické reprezentace

Analyzovaná sekvence DNA ve FASTA formátu je v písmenném tvaru, pro výpočetní operace je nutné sekvenci převést do číselné formu. Pro konstrukci spektrogramu je DNA sekvence převedena do 4D numerické (binární) reprezentace, výsledkem jsou čtyři vektory $u_A[n]$, $u_C[n]$, $u_G[n]$ a $u_T[n]$. Reprezentují tak výskyt určité báze v sekvenci, tedy je-li daná báze na určité pozici v sekvenci, je na stejné pozici vektoru hodnota 1. V opačném případě 0. Například sekvence ACGTTAC má tvar v 4D numerické reprezentaci u_A=*100010*, u_C=*0100001*, u_G=*0010000* a u_T=*0001100*.

Výpočet frekvenčního spektra výpočtem DFT

Výpočet diskrétní Fourierovy transformace je pro každou bázi, tedy pro každý vektor 4D numerické reprezentace získán dle vzorce:

$$U_x[k] = \sum_{n=0}^{N-1} u_x[n] e^{-j \cdot \frac{2 \cdot \pi}{N} \cdot k \cdot n}, \tag{3.1}$$

kde $k = 0,1, \ldots, \lfloor N/2 \rfloor + 1 X = A, C, G \; nebo \; T$ [4]. Výpočet je postupně prováděn pro úsek zvoleného plovoucího okna o zvolené šířce a posuvu okna. V navrženém algoritmu je tento výpočet řešen pomocí funkce *fft.m* z knihovny funkcí MATLABu.

Mapování DFT hodnot do RBG barev

Čtyři vektory U_A, U_C, U_G a U_T jsou v tomto kroku redukovány do tří, tedy do RGB barev podle následujícího převodu:

$$X_r[k] = a_r|U_A[k]| + t_r|U_T[k]| + c_r|U_C[k]| + g_r|U_G[k]|$$

$$X_g[k] = a_g|U_A[k]| + t_g|U_T[k]| + c_g|U_C[k]| + g_g|U_G[k]|$$

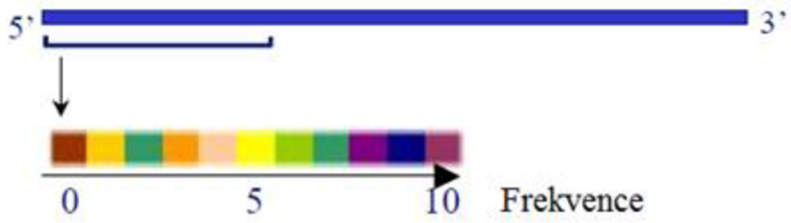
$$X_b[k] = a_b|U_A[k]| + t_b|U_T[k]| + c_b|U_C[k]| + g_b|U_G[k]|. \tag{3.2}$$

Rovnice (3.2) vyjadřují výslednou hodnotu barvy pixelu ($X_r[k]$, $X_g[k]$, $X_b[k]$), která je závislá na velikosti frekvence složky konkrétních nukleotidových bází. V navrženém algoritmu postačí mapovat DFT hodnoty pouze do dvou barev (viz blokové schéma na Obr. 9: Blokové schéma pro navržený vyhledávač CpG ostrůvků CpGI.), potom rovnice 3.2 má tvar:

$$X_r[k] = a_r|U_A[k]| + a_r|U_T[k]|$$

$$X_g[k] = a_g|U_C[k]| + a_g|U_G[k]|$$

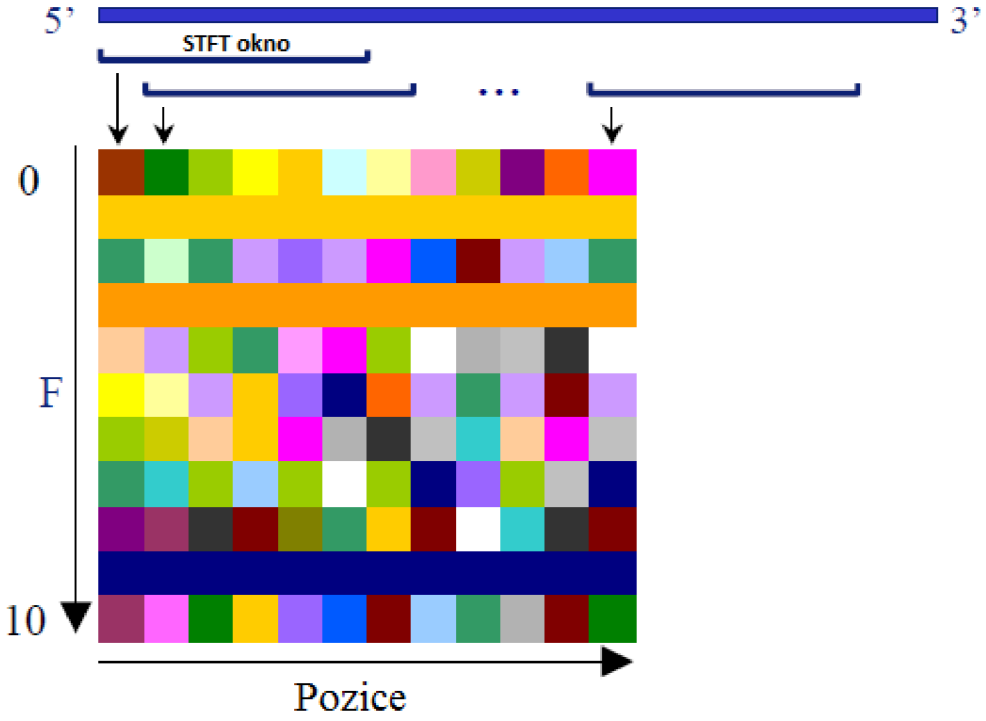
$$X_b[k] = 0$$



Obr. 11: Získané spektrum (po normalizaci hodnot pixelů) pro úsek DNA délky okna. Převzato z [4].

Normalizace hodnot pixelů

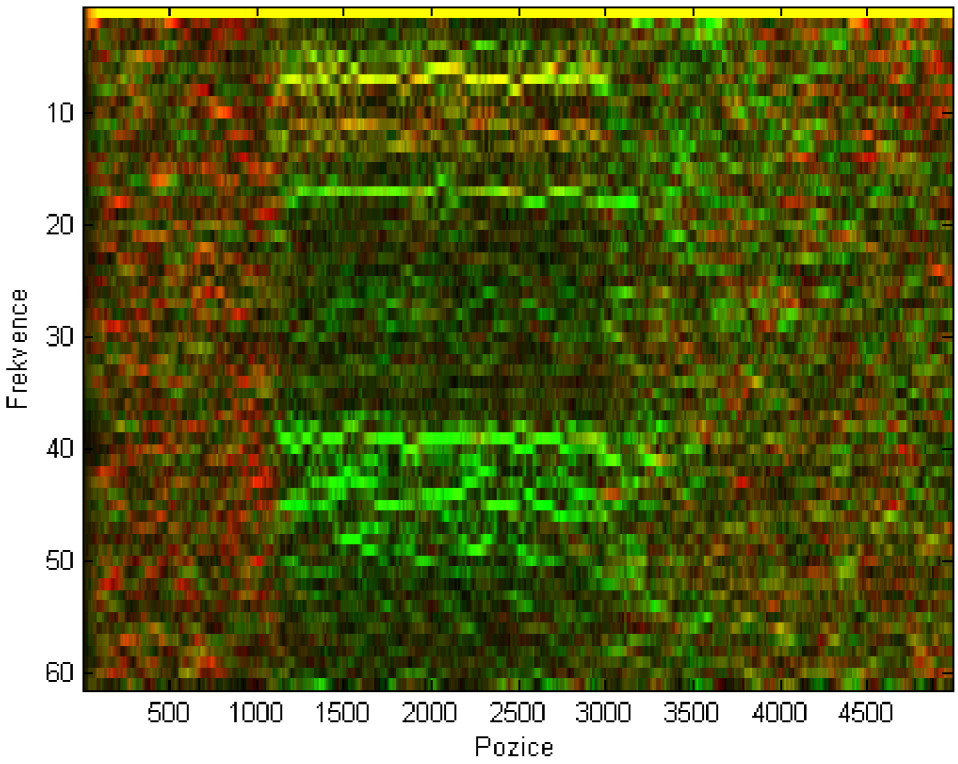
Pro vykreslení DNA spektrogramu je na závěr zapotřebí hodnoty pixelu normalizovat, tak aby hodnoty spadaly do rozsahu 0-1. Tímto krokem je získáno spektrum DNA úseku, viz Obr. 11. Posuvem okna vznikají další spektra, která pak tvoří sloupce v konečné matici spekter resp. spektrogramu jak je uvedeno na Obr. 12.



Obr. 12: Postup vzniku spektrogramu z DNA sekvence. Převzato z [4].

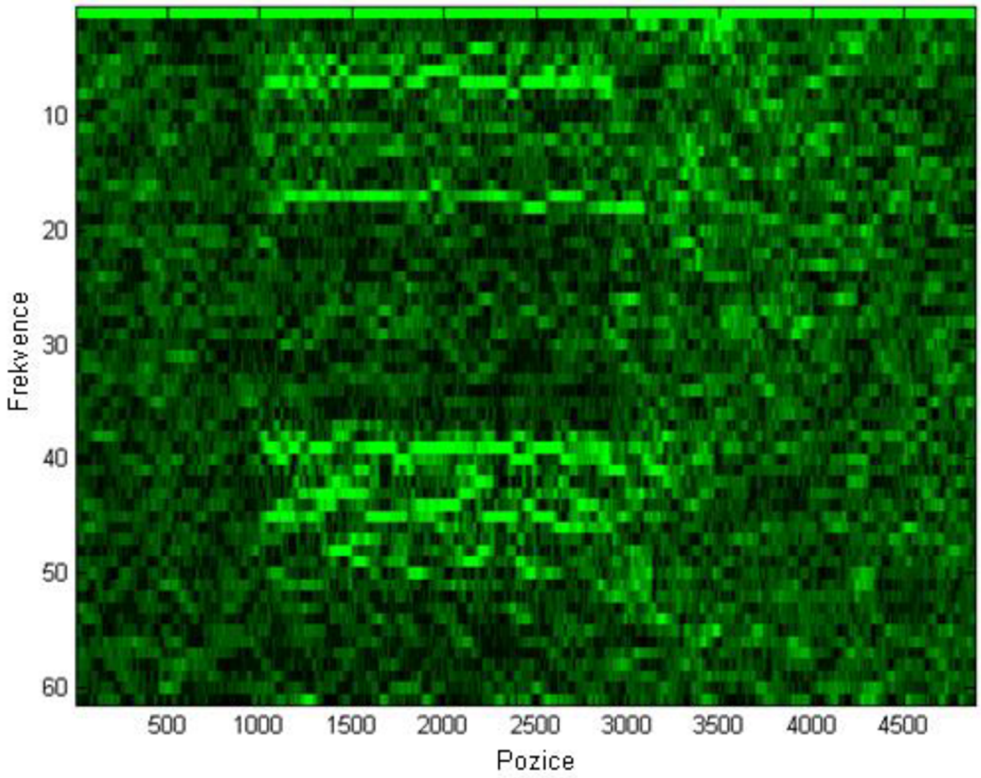
2.1.2. Analýza DNA spektrogramu a rozhodovací kritéria

Jak bylo uvedeno výše, mapování DFT hodnot do RGB je provedeno pouze do oblasti červené (A, T) a zelené (C, G), takový výsledný spektrogram je na Obr. 13, kde je na první pohled zřejmá odlišná oblast, která reprezentuje výskyt CpG ostrůvku.

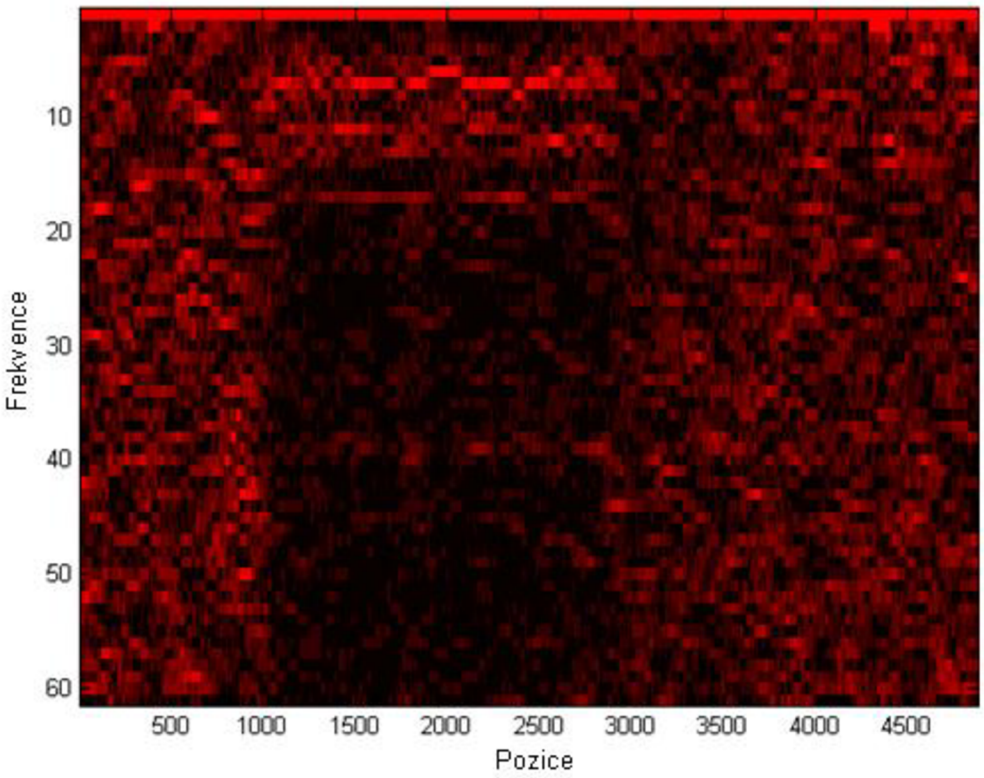


Obr. 13: Spektrogram v úseku DNA sekvence (chromozomu 21 člověka, hs_alt_CHM1_1.1_chr21:25000-30000) s výskytem CpG ostrůvku.

Navržený algoritmus však pro analýzu využívá dva spektrogramy, které jsou získány selekcí vektoru pixelu, tedy $(X_r[k], 0, 0)$ a $(0, X_g[k], 0)$. Jeden reprezentuje zvýraznění pouze výskyt nukleotidů C a G (zelený) a druhý spektrogram vyjadřuje výskyt nukleotidů A a T (červený) v analyzované sekvenci, prakticky dojde k zobrazení červené nebo zelené složky barevného DNA spektrogramu. Tyto spektrogramy jsou zobrazeny na Obr. 14 a Obr. 15. Již vizuálním pohledem na spektrogramy je zřejmá oblast v analyzované sekvenci, která je bohatší na frekvenční výskyt nukleotidů C a G, a současně tatáž oblast je chudší na výskyt A a T. Platí, že čím je intenzita dané barvy vyšší, tím je vyšší přítomnost daných nukleotidů na určitém frekvenčním pásmu. Na těchto poznatcích můžeme založit následující analýzu k lokalizaci CpG ostrůvku v sekvenci DNA.

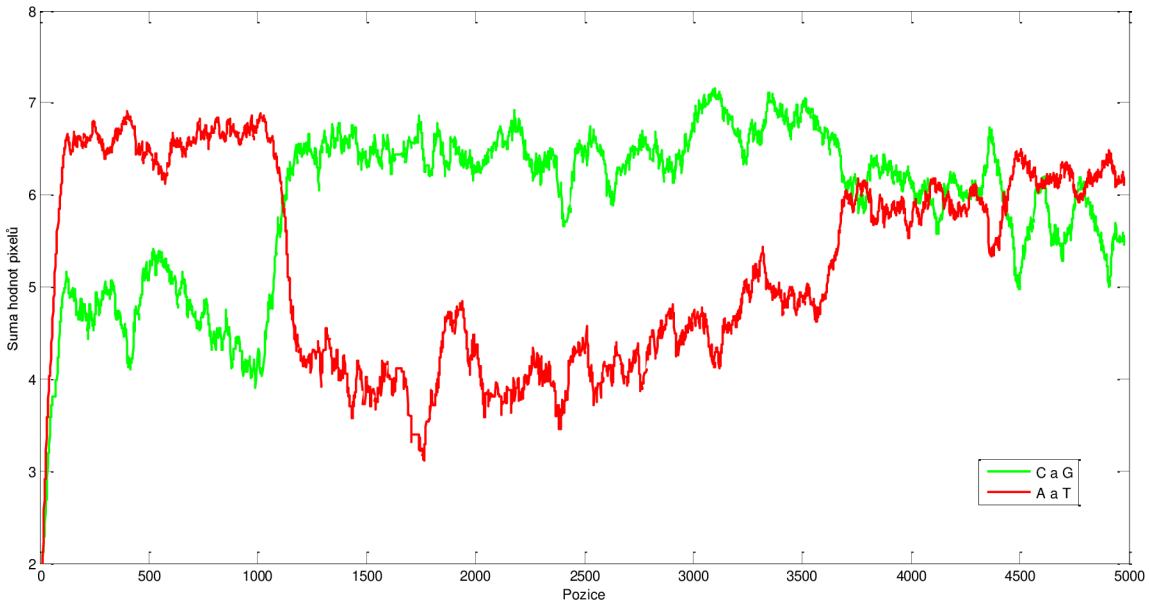


Obr. 14: Spektrogram pro nukleotidy C a G v úseku DNA sekvence (chromozomu 21 člověka, hs_alt_CHM1_1.1_chr21:25000-30000) s výskytem CpG ostrůvku.



Obr. 15: Spektrogram pro nukleotidy A a T v úseku DNA sekvence (chromozomu 21 člověka, hs_alt_CHM1_1.1_chr21:25000-30000) s výskytem CpG ostrůvku.

Porovnání obou spektrogramů je provedeno tak, že barevné spektrogramy jsou podle jednotlivých spekter (podél osy *x*) sumována, viz Obr. 16. Jak je z obrázku zřejmé, suma hodnot pixelů zeleného spektrogramu v oblasti výskytu CpG ostrůvku je vyšší než suma hodnot červeného spektrogramu.



Obr. 16: Sumovaná spektra, zeleného a červeného spektrogramu DNA sekvence (chromozomu 21 člověka, hs_alt_CHM1_1.1_chr21:25000-30000).

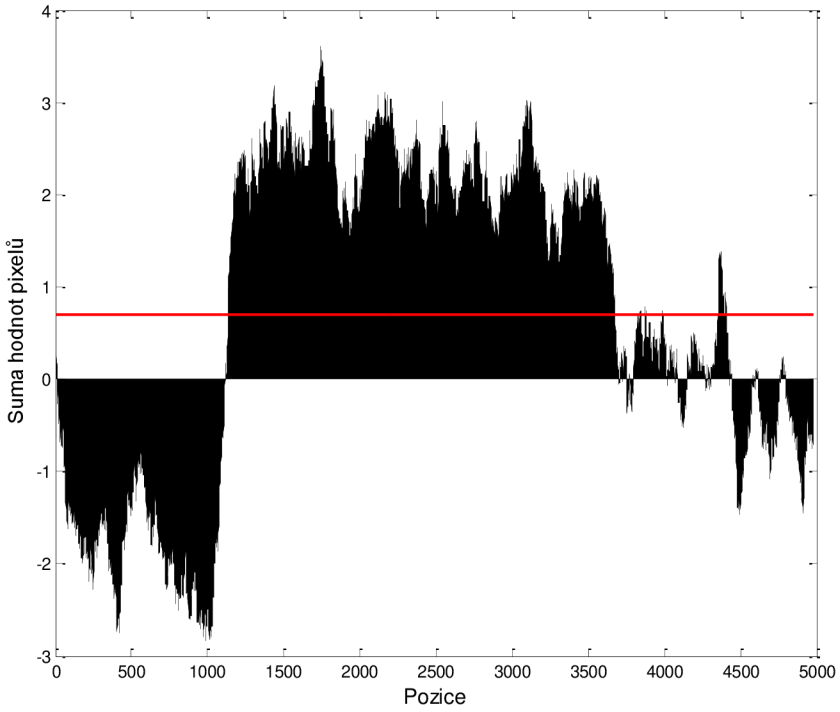
Rozhodovací kritéria a stanovení CpG ostrůvku

Jako vhodný způsob jak stanovit CpG ze sumovaných spekter je provést rozdíl sumovaných spekter a použít pevnou prahovou hodnotu k jeho určení. Tedy, CpG ostrůvek je stanoven podle následujících kritérií:

a) Rozdíl mezi hodnotami zeleného a červeného sumovaného spektra musí být větší než předem stanovený práh. Ukázková prahová hodnota 0,7 je červenou čarou zobrazen na Obr. 17.

b) Za CpG ostrůvky jsou potom označeny ty, které jsou delší než 200 bp a současně splňují kritéria CpG ostrůvku (viz 1.1), tedy obsah C+G vyšší než 50 % a poměr získané a očekávané hodnoty výskytu CpG (O/E) větší než 0,6.

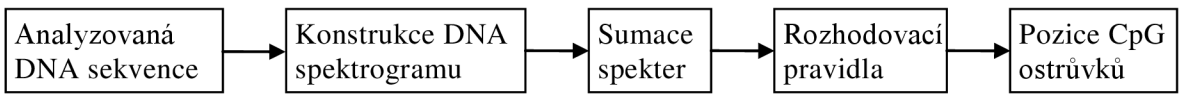
c) Lokalizované ostrůvky, které jsou od sebe vzájemně vzdáleny méně než 100 bp, jsou sloučeny do jednoho CpG ostrůvku.



Obr. 17: Rozdíl sumovaných spekter zeleného a červeného spektrogramu DNA sekvence (chromozomu 21 člověka, hs_alt_CHM1_1.1_chr21:25000-30000), stanovená prahová hodnota je označena červenou barvou. Začátek CpG ostrůvek byl v tomto případě algoritmem CpGI stanoven na pozici 1134, konec na pozici 3676, tedy o délce 2542 bp.

2.2. CpGI2

Pod názvem CpGI2 byl pojmenován druhý navržený algoritmus pro vyhledávání CpG ostrůvků. Tato verze programu je založena na předpokladu, že CpG ostrůvek má vyšší obsah CpG a tedy se v této oblasti vyskytuje s vyšší frekvencí. Oproti předchozí metodě je tato metoda zjednodušena především při výpočtu DNA spektrogramu, protože algoritmus CpGI2 realizuje pouze tzv. jednokanálový DNA spektrogram. Blokové schéma metody je uvedeno na Obr. 18.



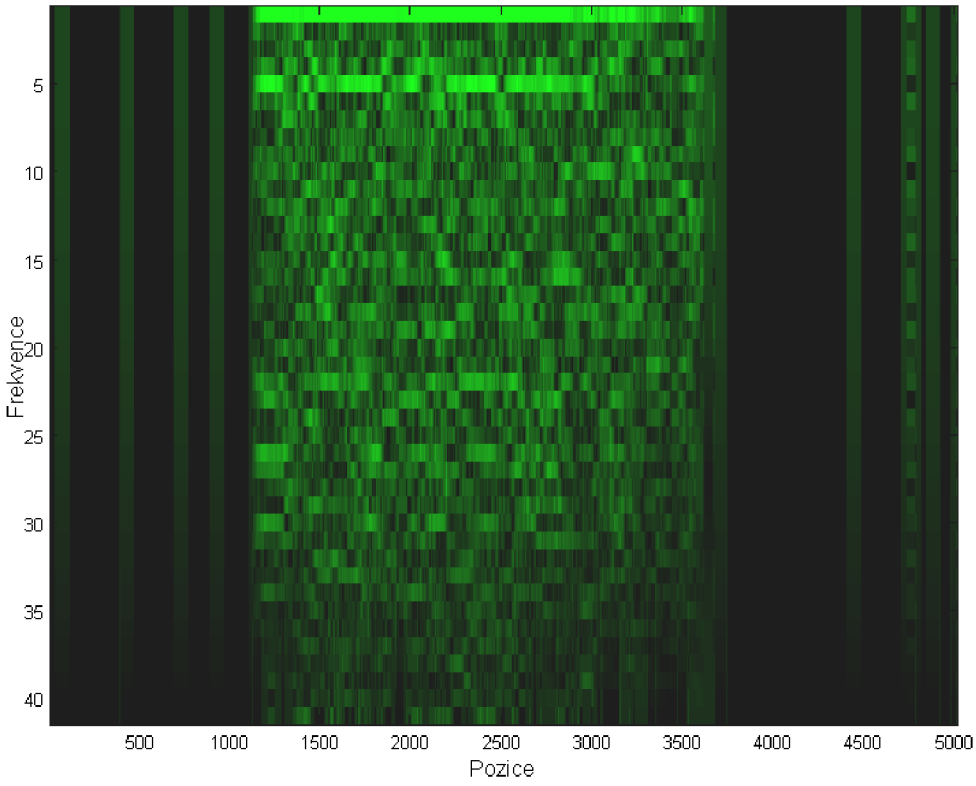
Obr. 18: Blokové schéma pro navržený vyhledávač CpG ostrůvků CpGI2.

2.2.1. Konstrukce DNA spektrogramu

V tomto případě je konstrukce DNA spektrogramu obdobná jak bylo uvedeno v kapitole 2.1.1, s několika rozdíly. Vstupní DNA sekvence je převedena do binární numerické reprezentace pouze do jednoho rozměru, tedy do binární posloupnosti čísel $u_{CG}[n]$. Převod je uskutečněn na základě výskytu CG dinukleotidu. Například pro sekvenci:

ACGTAATCGCTGAGTGCGACCGTTACGAATGCG,

je potom u_{CG}=*0110000110000000110011000110000011*. Následuje výpočet frekvenčního spektra výpočtem DFT viz rovnice (3.1). Tento výpočet je postupně prováděn pro úsek zvoleného plovoucího okna o zvolené šířce a posuvu plovoucího okna. Protože výsledkem je jednokanálový DNA spektrogram, krok mapovaní DFT hodnot do RGB barev je proveden za účelem lepšího vizuálního vyjádření jednotlivých spekter do zelené barvy. Potom rovnice (3.2) mají tvar: $X_r[k] = 0$; $X_g[k] = a_g|U_{CG}[k]|$; $X_b[k] = 0$, na závěr jsou hodnoty normalizovány, aby spadaly do rozsahu 0-1.

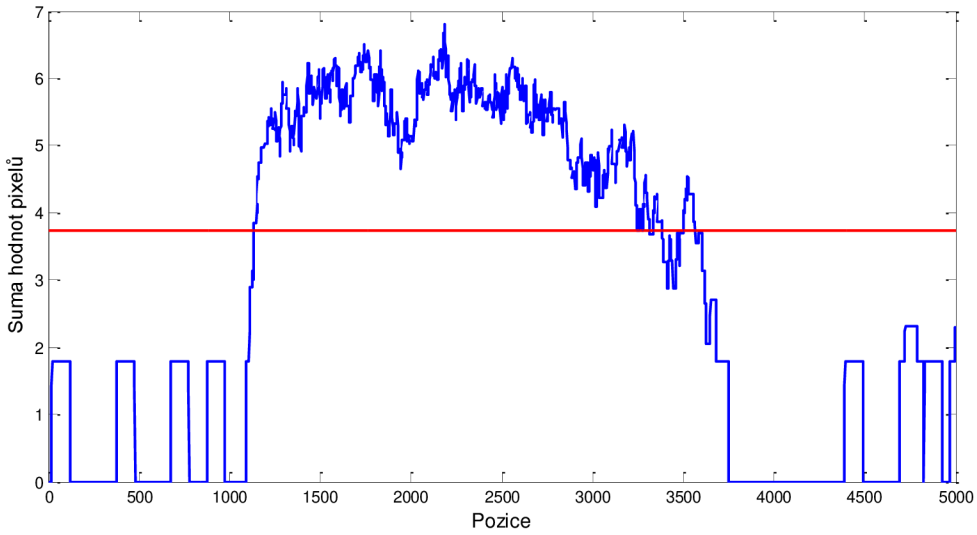


Obr. 19: Spektrogram metody CpGI2 v úseku DNA sekvence (chromozomu 21 člověka, hs_alt_CHM1_1.1_chr21:25000-30000) s výskytem CpG ostrůvku. Okno délky 80, posun okna o 1.

Samotný algoritmus konstrukce je realizován jako funkce pod názvem *spektrogram2.m.* Výsledný spektrogram realizovaný metodou CpGI2 je na Obr. 19. Oblast CpG ostrůvku je více ze spektrogramu odlišitelná než u metody CpGI, viz Obr. 13.

2.2.2. Analýza DNA spektrogramu a rozhodovací kritéria

Na Obr. 19 je uveden spektrogram, který byl vytvořen z ukázkové DNA sekvence s oblastí bohatou na CpG. Tato oblast je z obrázku zřejmá a cílem algoritmu je co nejpřesněji vymezit hranice CpG ostrůvku. To je provedeno na základě sumace hodnot pixelů jednotlivých spekter, tento způsob se jeví jako vhodný, neboť na aktuální pozici okna je zkoumán frekvenční výskyt CpG.



Obr. 20: Sumovaná spektra spektrogramu DNA sekvence (chromozomu 21 člověka, hs_alt_CHM1_1.1_chr21:25000-30000). Stanovená prahová hodnota je označena červenou barvou. Začátek CpG ostrůvek byl v tomto případě algoritmem CpGI2 stanoven na pozici 1207, konec na pozici 3339, tedy o délce 2132 bp.

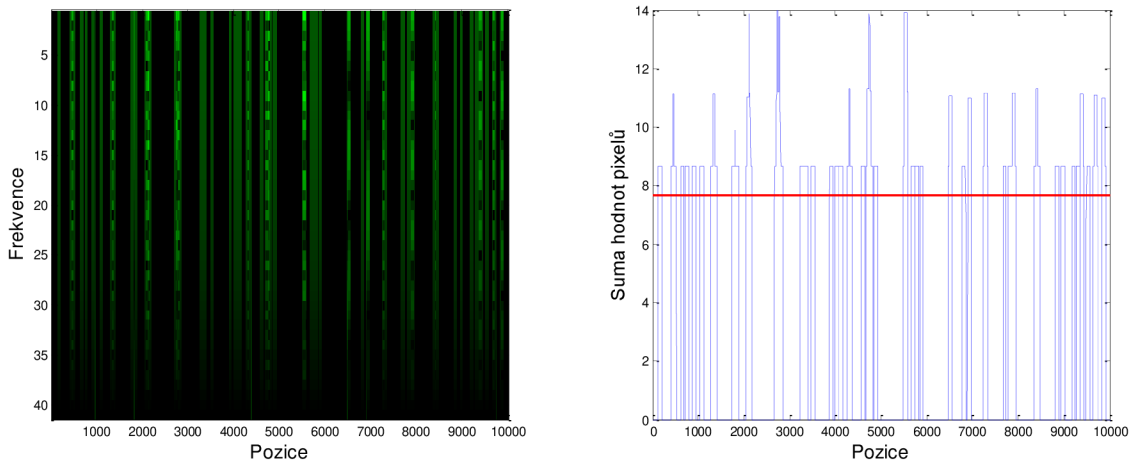
CpG ostrůvky jsou potom stanoveny podle následujících kritérií:

Sumované spektrum musí být větší než stanovený práh o hodnotě, která je větší než násobek předem určené hodnoty k a maximální hodnoty sumovaného spektrogramu. Aktuální hodnota tohoto prahu je červenou čarou vyznačena na Obr. 20, kde k = 0,7 . (Nevýhodou volby prahu tímto způsobem může být, že detekce může být falešně pozitivní v případě, kdy v sekvenci není CpG ostrůvek přítomen. Avšak, jak bylo uvedeno v kapitole 1, CpG v DNA sekvenci v oblastech mimo CpG ostrůvek se vyskytuje

velmi zřídka. Pro důkaz byl otestován úsek DNA sekvence bez výskytu CpG ostrůvku viz Obr. 21. Nad prahovou hodnotou nalezené CpG ostrůvky nejsou dostatečně dlouhé, aby byly označeny za ostrůvek).

a) Za CpG ostrůvky jsou potom označeny ty, které jsou delší než 200 bp a současně splňují kritéria CpG ostrůvku (viz 1.1), tedy obsah C+G vyšší než 50 % a poměr získané a očekávané hodnoty výskytu CpG (O/E) větší než 0,6. Tyto hodnoty je možné podle požadavků před analýzou nastavit.

b) Lokalizované ostrůvky, které jsou od sebe vzájemně vzdáleny méně než 100 bp, jsou sloučeny do jednoho CpG ostrůvku.



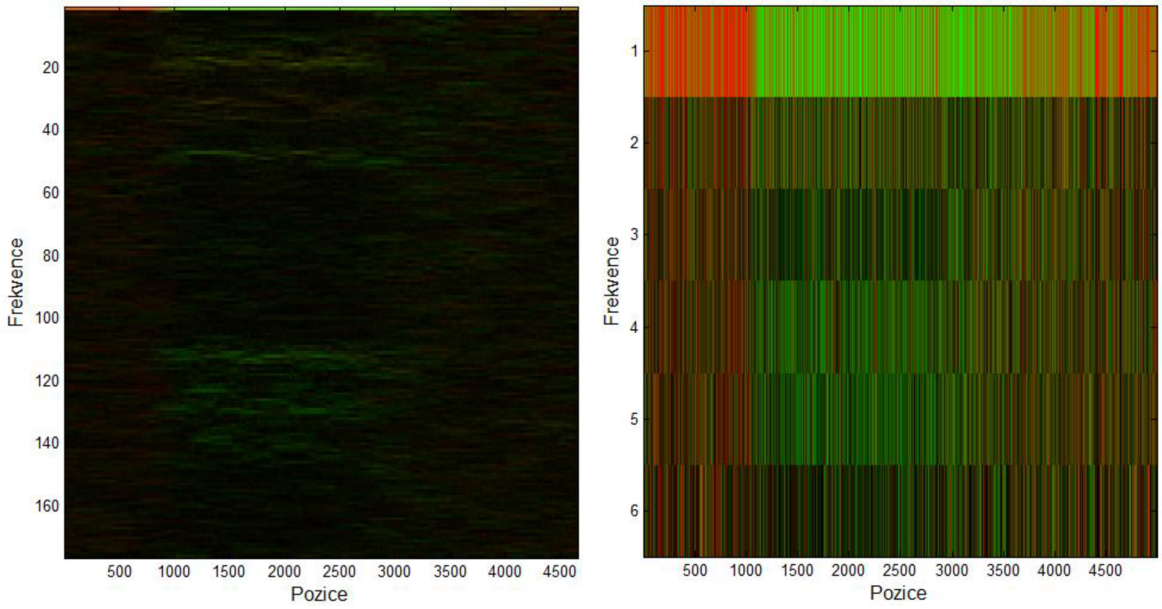
Obr. 21: Úsek DNA sekvence bez přítomnosti CpG ostrůvku (chromozomu 21 člověka, hs_alt_CHM1_1.1_chr21:40000-50000). Metodou CpGI2 správně ostrůvky nenalezeny

2.3. Vliv volby délky STFT okna na výsledný spektrogram

Jak bylo uvedeno již dříve, výsledný spektrogram závisí na několika parametrech, na výběru délky STFT okna a jeho posun a na způsobu převodu do RGB barev. Pro DNA spektrogram dále platí, že volba délky okna by měla být několikrát delší než je délka opakujícího se hledaného vzoru, kterým je např. CpG, přitom by okno mělo být menší než je délka hledané oblasti obsahující vzor, tedy CpG ostrůvek[4].

Obecně pro spektrogram platí, že je zapotřebí volit délku okna tak, aby byl zvolen kompromis. Protože pokud bude STFT okno příliš dlouhé, sice bude dosaženo vysoké-

ho frekvenčního rozlišení, ale časové rozlišení bude nízké. Pokud je však zvoleno okno příliš krátké, časové rozlišení spektrogramu je potom vysoké avšak frekvenční rozlišení bude nízké. Porovnání spektrogramů s extrémní volbou STFT okna je na Obr. 22.



Obr. 22: Porovnání DNA spektrogramů při extrémně zvolených STFT oken. Vpravo okno o délce 350, vlevo o délce10. Úsek DNA sekvence (chromozomu 21 člověka, hs_alt_CHM1_1.1_chr21:25000-30000) s výskytem CpG ostrůvku.

2.4. Optimalizace vstupních parametrů CpGI a CpGI2

Dosahování výsledků detekce CpG ostrůvku u navržených algoritmů záleží na zvolených stupních hodnotách. Tyto hodnoty byly na základě testování optimalizovány za účelem získání co nejlepších a současně spolehlivých výsledků, a tedy jsou již přednastaveny v programovém kódu, avšak uživatel jak programu CpGI a CpGI2 může tyto hodnoty v určitých mezích libovolně měnit.

Tab. 1: Optimalizované vstupní hodnoty algoritmů CpGI a CpGI2

	CpGI	CpGI2
Minimální obsah CG	0,5	0,5
Min. hodnota O/E	0,65	0,65
Velikost okna	120	60
Posun okna	20	20
Prahová hodnota	0,81	0,50

3. Hodnocení dosažených výsledků

Algoritmy popsané v předchozí kapitole byly doposud otestovány pouze na ukázkové sekvenci DNA k vysvětlení principu metod. Jedním z cílů této práce je navržené řešení metody vyhledávání CpG ostrůvku otestovat a na základě dosažených výsledků z volně dostupných vyhledávačů na internetu porovnat a posoudit využitelnost navrženého řešení.

Aby bylo možné dostatečně kvantitativně hodnotit dosažené výsledky, bylo by zapotřebí testovat DNA sekvenci alespoň o délce několik miliónů bp, například v článku [18] při testování použili databázi chromosomu 22 o délce téměř 50 Mbp. Takový objem dat je pro algoritmy CpGI i CpGI2 příliš výpočetně náročný a zdlouhavý, především pro samotnou konstrukci DNA spektrogramu. Hodnocení je tedy provedeno na kratších DNA sekvencích kvalitativně. Jsou zkoumány hodnoty pozic začátku a konců CpG ostrůvků, délky, obsah C+G v % a poměr získané a očekávané hodnoty výskytu CpG (Obs_{CpG}/Exp_{CpG}) zkr. O/E.

3.1. Vyhledávače dostupné na internetu

Dostupné vyhledávače CpG ostrůvků na internetu jsou nejčastěji založeny na definici CpG ostrůvku Gardiner-Gardena a Frommera ale podmínky definice je možné na vstupu vyhledávače upravit dle vlastní volby, pro získání výsledků budou vždy ponechány vstupní hodnoty, které jsou již přednastaveny.

Cpgplot

Je aplikace k vyhledávání CpG ostrůvků v DNA sekvencích volně dostupná na internetu. Metoda je založená na kriteriálních parametrech a plovoucím oknu podél analyzované sekvence uvedených v kap. 1.1. Výstupem je grafické vyjádření hodnot zkoumaných hodnot O/E, obsah CG a pozice lokalizovaných CpG ostrůvků (ostrůvky také numericky). [22]

cpgisland

Je algoritmus založený na obdobných parametrech jako aplikace Cpgplot. Avšak není dostupná on-line na internetu, ale je součástí knihovny funkcí programu MATLAB. Grafický výstup je také obdobný. [20]

CpGFinder (CpGF)

Volně přístupná aplikace pro určení CpG ostrůvku zkoumající parametry, viz výše. Aplikace neumožňuje grafické vyjádření, pouze seznam nalezených CpG ostrůvků s popisem jejich parametrů. Dostupnost aplikace je v odkaze [23].

CpG island Finder (CpGIF)

Má omezeně volitelné vstupní parametry, pouze nastavení minimální hodnot O/E a minimální délky. Nelze přednastavit minimální obsah CG. V průběhu testování bylo odhaleno několik nedostatků této aplikace. Při různé volbě minimální délky CpG ostrůvku byly dosažené výsledky pozic CpG ostrůvku odlišné. Přestože je volitelným parametrem hodnota O/E, nalezené CpG ostrůvky tuto hodnotu nerespektují. Výhodou je přehledná grafická úprava výstupu a jako jediný vyhledávač zobrazuje sekvenci lokalizovaných ostrůvků.

CpGIS

Je algoritmus popsaný v kap. 1.2 a realizována autory článku [2]. Vzhledem k omezené dostupnosti aplikace na internetu, bylo pro testování použitá verze programu obsluhovaná v příkazovém řádku. Běžně je dostupnost aplikace on-line na internetu viz [25].

CpGProd

Vyhledávač CpG ostrůvků s cílem lokalizovat promotory genů. Metoda byla popsána v kap. 1.3 a představena autory v článku [9]. Vstupní parametry jsou volba organismu analyzované sekvence mezi člověkem a myší. CpG ostrůvky na výstupu jsou pouze delší než 500 bp. Aplikace je dostupná on-line na internetu viz [17].

CpGGA a CpGPSO

Jsou algoritmy součástí webové aplikace CpGPAP popsané v kap1.9. Metoda CpGGA je založená na optimalizačním přístupu genetických algoritmů a metoda CpGPSO na optimalizačním přístupu rojení částic. Algoritmy byly představeny autory v článku [18]. Aplikace umožňuje grafickou reprezentaci výsledků i jako seznam výsledků. V průběhu testování se objevili některé nesrovnalosti, na výstupu obou algoritmů byla uvedena celková délka analyzované DNA sekvence odlišně než byla skutečná délka. Dále po přepočtení hodnot obsahu CG a O/E byly hodnoty rozdílné než uvedené na výstupu těchto algoritmů. Aplikace je on-line dostupná na internetu viz [26].

Další vyhledávače

V kap 1.7 je uvedený rozbor algoritmu CpGcluster, jehož princip je z uvedených metod ojedinělý. Jeho on-line dostupnost je omezena chybou webové stránky [29], ani verze přes programovací rozhraní PERL není funkční. Další nedostupnou metodou je algoritmus založený na skrytých Markovských modelech uvedený v kap 1.5. Přestože je aplikace na internetu volně dostupná [30], soubor po spuštění nepracuje správně.

3.2. Testované databáze sekvencí

Pro otestování vyhledávače CpG ostrůvku jsou nejčastěji používány sekvence DNA lidského chromozomu 21 a 22, tyto chromozomy k testování použili například v literatuře [2], [4], [16] nebo [18]. Avšak další sekvence, kde je výhodou zaručený výskyt CpG ostrůvků je z článku [1]. Vybrané DNA sekvence k testování jsou tak založeny na tomto výběru a byly získány z databáze NCBI.

Uměle vytvořená sekvence s CpG ostrůvkem (*seq_1.fasta*)

Sekvence byla vytvořena ze dvou částí, z náhodně vygenerované sekvence, která byla ochuzena o CpG s náhodným nahrazením jiným dinukleotidem a druhá část oblastí CpG ostrůvku s poměrně vysokým obsahem CpG. Sekvence je délky 5501 bp a skutečné pozice začátku a konce CpG ostrůvku jsou 2001 a 3500. Tato sekvence je součástí elektronické přílohy pod názvem *seq_1.fasta*.

Úsek lidského chromozomu 21 (*seq_2.fasta*)

Pokud je k testování použitá krátká sekvence s výrazným CpG ostrůvkem, jehož délka zaujímá poměrně značnou část zkoumané sekvence, tak obvykle je ostrůvek lokalizován všemi vyhledávači, tedy sekvence není k hodnocení příliš náročná. K testování potom byl použit úsek DNA sekvence lidského chromozomu 21 [27] o délce okolo 185 kbp. Cílem je pomocí této sekvence zhodnotit, zda navržené algoritmy CpGI a CpGI2 nedetekují falešně pozitivní detekce v místech, kde CpG ostrůvek neexistuje a také zda nedochází k falešně negativní detekci v místech, kde CpG ostrůvek existuje. Sekvence je součástí elektronické přílohy pod názvem *seq_2.fasta*.

DNA sekvence vybraných genů člověka

Na základě článku [1] byly vybrány tři DNA sekvence genů, ve kterých je výskyt CpG ostrůvku:

a) c-Ha-ras1 p21, DNA sekvence délky 6453 bp (*seq_3.fasta*),

b) beta-tubulin, klon m40, DNA sekvence délky 5117 bp (*seq_4.fasta*),

c) somatostatin I, DNA sekvence délky 2667 bp (*seq_5.fasta*).

3.3. Dosažené výsledky a porovnání

První testovací sekvence *seq_1.fasta*, tedy uměle vytvořená DNA sekvence s přesně vymezenou oblastí CpG ostrůvku bude sloužit k hodnocení, jak přesně daný algoritmus tyto pozice lokalizuje. Souhrn výsledků je uveden v Tab. 2.

Pro tuto sekvenci vyhledávač CpGI jako jediný lokalizoval pozici začátku CpG ostrůvku na přesné pozici, avšak kratší délky na úkor nižší hodnoty O/E. Naopak CpGI2 detekoval přesnou pozici konce ostrůvku, ale pozici začátku o 40 bp dříve. V porovnání mezi CpGI a CpGI2 bylo lepších výsledků dosaženo vyhledávačem CpGI2, ve srovnání s výsledky ostatních vyhledávačů jsou výsledky poměrně uspokojivé. V této sekvenci byl CpG ostrůvek lokalizován všemi použitými vyhledávači, nejpřesněji byl CpG ostrůvek detekován vyhledávačem CpGF, dále dobrý výsledek poskytují Cpgplot a cpgisland.

Tab. 2: Dosažené výsledky hodnot jednotlivých vyhledávačů CpG ostrůvku na umělé DNA sekvenci (*seq_1.fasta*) s přesně vymezeným CpG ostrůvkem na pozicích od 2001 do 3500, CG =0,8447, O/E = 1,0634.

	Cpgplot	cpgisland	CpGF	CpGIF	CpGIS	CpGProd	CpGGA	CpGPSO	CpGI	CpGI2
Začátek	1951	2004	2000	1807	1672	1596	1792	1935	2001	1961
Odch. [bp]	-50	3	-1	-194	-329	-405	-209	-66	0	-40
Konec	3530	3583	3496	3685	3902	3913	3750	3643	3420	3500
Odch. [bp]	30	83	-4	185	402	413	250	143	-80	0
Délka [bp]	1580	1580	1497	1879	2231	2318	1959	1709	1420	1540
Odch. [bp]	80	80	-3	379	731	818	459	209	-80	40
Obsah CG	0,8228	0,8234	0,8444	0,7632	0,7082	0,6967	0,7499	0,7934	0,8423	0,8318
Odch.	-0,0219	-0,0213	-0,0003	-0,0815	-0,1365	-0,148	-0,0948	-0,0513	-0,0024	-0,0129
[%]	2,59	2,52	0,04	9,65	16,16	17,52	11,22	6,07	0,28	1,53
Hod. O/E	1,0674	1,0606	1,0607	1,0516	1,0404	1,0344	1,0498	1,0644	1,0394	1,0679
Odch.	0,0040	-0,0028	-0,0027	-0,0118	-0,0230	-0,0290	-0,0136	0,0010	-0,0240	0,0045
[%]	0,38	0,26	0,25	1,11	2,16	2,73	1,28	0,09	2,26	0,42

Další testovanou DNA sekvencí je úsek DNA sekvence chromozomu 21 člověka, o celkové délce 184379 bp (*seq_2.fasta*). Dosažené výsledky všech pozic začátků a konců detekovaných CpG ostrůvku jednotlivými vyhledávači jsou uvedeny v příloze, v Tab. 8 a v grafickém zpracování na Obr. 27. Stručný souhrn těchto výsledků je uveden v Tab. 3. Během testování byly odhaleny některé nesrovnalosti vyhledávače jak CpG-GA tak CpGPSO, kde z výsledků analyzované sekvence po přepočtení CG obsahu a O/E hodnoty, nebyly totožné s výstupními hodnotami programu a také délka sekvence byla programem uvedená odlišně než délka skutečná. Další nesrovnalost byla vyhledávače CpGIF, kdy na výstupu byly sekvence s nižším hodnotou O/E než bylo přednastaveno.

Tab. 3: Dosažené výsledky hodnot jednotlivých vyhledávačů CpG ostrůvku na DNA sekvenci chromozomu 21 (*seq_2.fasta*). Podrobné výsledky jsou součásti přílohy viz Tab. 8.

	Cpgplot	cpgisland	CpGF	CpGIF	CpGIS	CpGProd	CpGGA	CpGPSO	CpGI	CpGI2
Počet	4	4	7	12	3	3	8	6	3	3
Délka [bp]										
Celková	4061	3995	6041	8463	4992	5442	7259	7023	5274	3654
Minimální	272	275	226	221	1000	1044	214	247	1338	398
Maximální	2250	2214	2433	2554	2473	2673	1996	2607	2558	2158
Průměrná	1015	999	863	705	1664	1814	907	1171	1758	1218
Pokrytí ostrůvků [%]	2,20	2,17	3,28	4,59	2,71	2,95	3,94	3,81	2,86	1,98
Obsah CG	0,67	0,64	0,62	0,58	0,64	0,63	0,39	0,42	0,63	0,70
Hod. O/E	1,07	1,14	0,87	0,71	1,05	1,02	0,48	0,57	1,04	1,18

Z výsledků je zřejmé, že algoritmy CpGI a CpGI2 dosažené výsledky jsou ve srovnání striktní a lokalizují výhradně významné ostrůvky, tedy pouze ty, které byly nalezeny všemi vyhledávači. Počet a lokalizace nalezených ostrůvků se shodují s výsledky CpGIS a CpGProd, které však definují ostrůvky pouze o délce delší než 500 bp. Z celkového procentuálního pokrytí analyzované DNA sekvence bylo nejnižší hodnoty dosaženo CpGI2 a současně nejvyšší hodnoty obsahu CG a hodnoty O/E, což znamená, že algoritmus vyhledává ostrůvky s nejvyšším zastoupením CpG.

Dalším hodnocením je posuzováno, jak programy CpGI a CpGI2 nalezené CpG ostrůvky lokalizovaly z pohledu pozic začátků a konců, tedy i délky ve srovnání s dosaženými pozicemi ostatních vyhledávačů. Souhrn těchto hodnot pro tři CpG ostrůvky odhalené současně všemi programy je uveden v Tab. 4.

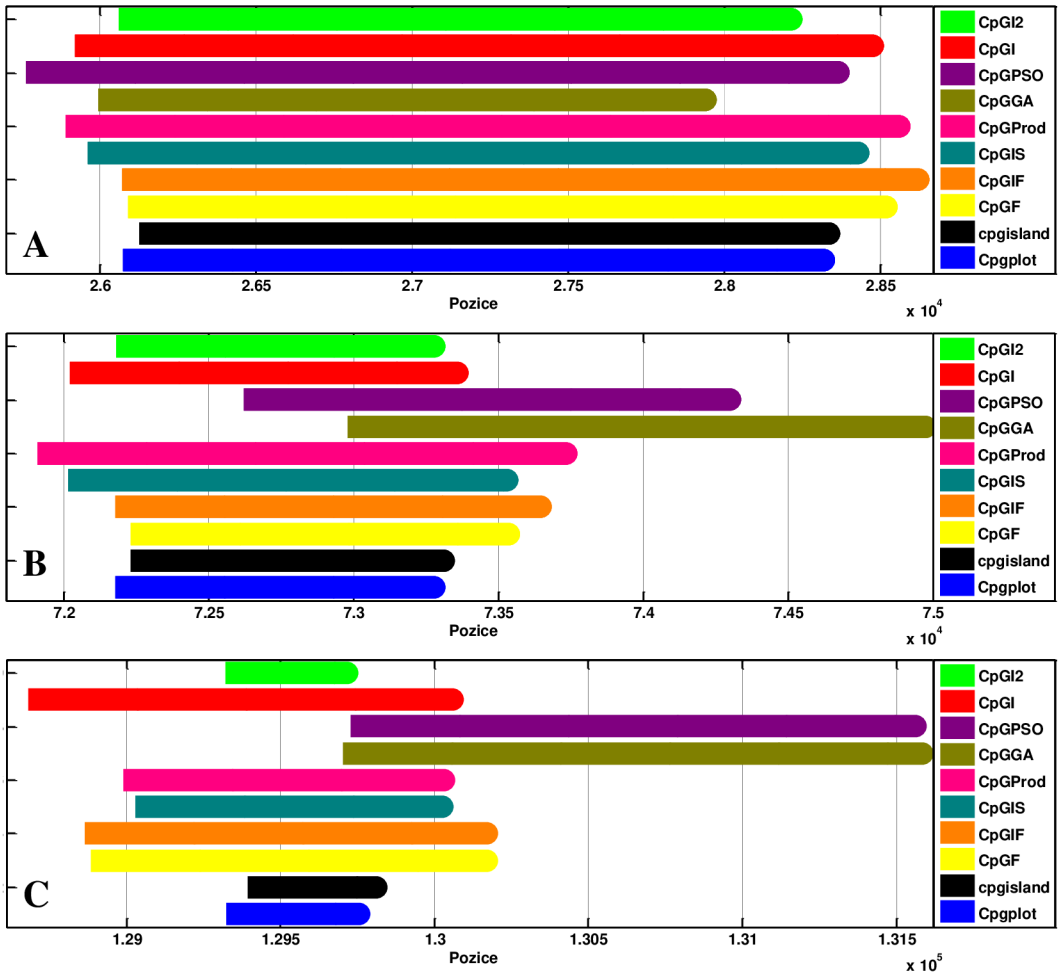
Tab. 4: Porovnání nalezených CpG ostrůvků vyhledávačem CpGI a CpGI2 s výsledky shodných ostrůvků nalezených ostatními vyhledávači v DNA sekvenci chromozomu 21 (*seq_2.fasta*).

	Cpgplot	cpgisland	CpGF	CpGIF	CpGIS	CpGProd	CpGGA	CpGPSO	CpGI	CpGI2
Začátek										
1	26074	26127	26091	26071	25961	25891	25995	25764	25921	26060
2	72178	72231	72229	72177	72013	71908	72980	72622	72021	72180
3	129323	129394	128884	128864	129030	128991	129702	129727	128682	129320
Konec										
1	28323	28340	28523	28624	28433	28563	27942	28370	28480	28219
2	73281	73314	73537	73647	73531	73735	74975	74300	73360	73279
3	129757	129815	130174	130172	130029	130034	131586	131564	130061	129719
Délka [bp]										
1	2250	2214	2433	2554	2473	2673	1948	2607	2558	2158
2	1104	1084	1309	1471	1519	1725	1996	1679	1338	1098
3	435	422	1291	1309	1000	1044	1885	1838	1378	398

Pro lepší vizuální porovnání nalezených oblastí jednotlivých CpG ostrůvků jsou na Obr. 23 tyto CpG ostrůvky zobrazeny v posloupnosti DNA sekvence pro každý algoritmus. Nejodlišnější detekce mají vyhledávače CpGGA a CpGPSO. Nejvíce různých detekcí, odlišné délky bylo dosaženo u třetího CpG ostrůvku (na Obr. 23 oblast C). CpGI byl tento ostrůvek detekován začátkem na pozici 128682, což je v porovnání s ostatními výsledky poměrně předčasně. Tento ostrůvek byl při detekci metodou CpGI2 detekován jako nejkratší, avšak délka i pozice začátků i konců ostrůvku jsou velmi blízké jako výsledky algoritmů CpGplot i cpgisland.

Z dosažených výsledků testované DNA sekvence úseku chromozomu 21 člověka hs_alt_CHM1_1.1_chr21:1-184379(*seq_2.fasta)* nelze posoudit, který výsledek je nejlepší, ale pro hodnocení účinnosti algoritmů CpGI a CpGI2 výsledky ukazují, že

v delší DNA sekvenci nelokalizují falešně pozitivní detekce ale také nelokalizují krátké CpG ostrůvky.



Obr. 23: Porovnání CpG ostrůvků lokalizovaných všemi vyhledávači v DNA sekvenci chromozomu 21 člověka, hs_alt_CHM1_1.1_chr21:1-184379, (seq_2.fasta).

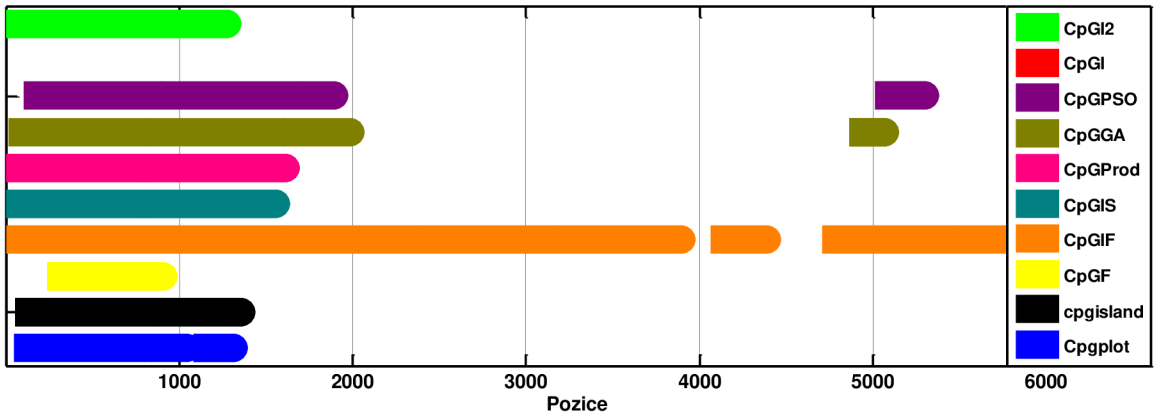
Další skupinou testovaných DNA sekvencí, jsou DNA sekvence vybraných tří konkrétních genů, kde je právě výskyt CpG ostrůvků. Byly vybrány takové sekvence, aby byly odhaleny některé nedostatky algoritmů CpGI či CpGI2.

První zkoumanou DNA sekvencí genu c-Ha-ras1 p21 (*seq_3.fasta*) byla velmi pravděpodobně odhalena falešně negativní detekce algoritmem CpGI. V této sekvenci byl CpG ostrůvek lokalizován alespoň jeden, a začínající nejčastěji na první pozici sekvence. Algoritmus CpGI2 lokalizoval shodný začátek ostrůvku jako algoritmy CpGIS a CpGProd, avšak kratší délky.

Souhrn výsledků nalezených pozic začátků, konců a délky CpG ostrůvků v DNA sekvenci genu c-Ha-ras1 p21 jednotlivými algoritmy je uveden v Tab. 5. Grafické porovnání je pak uvedeno na Obr. 24.

Tab. 5: Porovnání nalezených pozic začátků, konců a délky CpG ostrůvků lokalizovaných jednotlivými vyhledávači v DNA sekvenci genu c-Ha-ras1 p21 (*seq_3.fasta*).

	Cpgplot	**cpgisland**	**CpGF**	**CpGIF**	**CpGIS**	**CpGProd**	**CpGGA**	**CpGPSO**	**CpGI**	**CpGI2**
Začátek										
1	48	57	241	3	1	1	19	107	-	1
2	1082			4065			4862	5010		
3				4705						
Konec										
1	1049	1366	916	3903	1565	1621	1993	1900	-	1280
2	1323			4394			5073	5304		
3				5760						
Délka [bp]										
1	1002	1310	676	3903	1565	1621	1975	1794	-	1280
2	242			330			212	295		
3				1056						

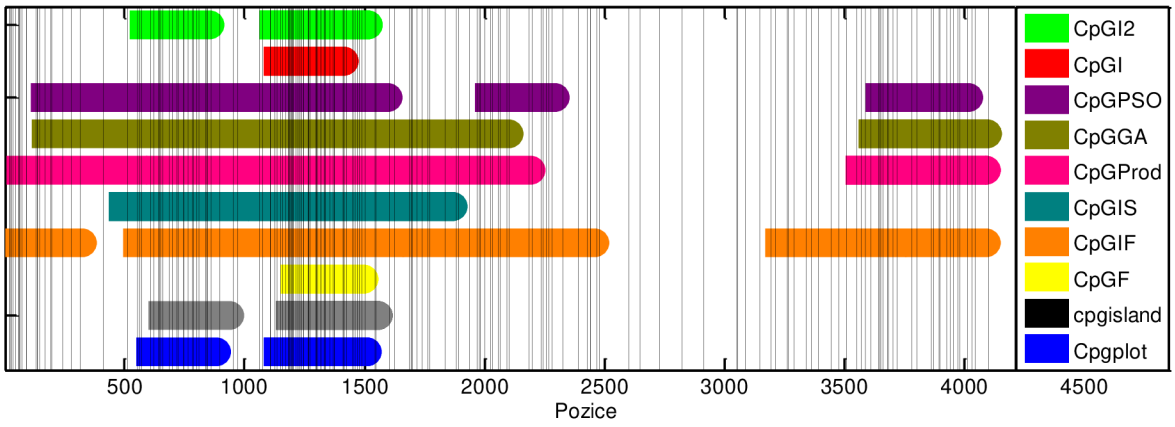


Obr. 24: Porovnání nalezených CpG ostrůvků lokalizovaných jednotlivými vyhledávači v DNA sekvenci genu c-Ha-ras1 p21 (seq_3.fasta).

Další zkoumaná DNA sekvence je zvláštní tím, že dosažené výsledky se velmi různí podle zvoleného vyhledávače. Avšak podobnosti jsou zřejmé, kdy algoritmus CpGI lokalizoval CpG ostrůvky obdobně jako algoritmus CpGplot a cpgisland, zatímco celkové výsledky CpGI2 jsou podobné s výsledky algoritmu CpGF a které jsou podobné s druhým (a nejvíce zastoupeným CpG viz Obr. 25) nalezeným ostrůvkem algoritmy CpGI2, CpGplot a cpgisland. Souhrn výsledků nalezených pozic začátků, konců a délky CpG ostrůvků v DNA sekvenci genu beta-tubulinu jednotlivými algoritmy je uveden v Tab. 6. Grafické porovnání je pak uvedeno na Obr. 25.

Tab. 6: Porovnání nalezených pozic začátků, konců a délky CpG ostrůvků lokalizovaných jednotlivými vyhledávači v DNA sekvenci genu beta-tubulinu, (*seq_4.fasta*).

	Cpgplot	cpgisland	CpGF	CpGIF	CpGIS	CpGProd	CpGGA	CpGPSO	CpGI	CpGI2
Začátek										
1	548	600	1150	1	432	7	115	107	1081	521
2	1079	1132		492	432	3506	3559	1962		1061
3				3171				3589		
Konec										
1	887	945	1502	330	1875	2200	2107	1603	1420	860
2	1517	1562		2465	1875	4098	4100	2300		1520
3				4096				4023		
Délka [bp]										
1	340	346	398	330	1444	2194	1993	1497	340	340
2	439	431		1974		593	542	339		460
3				926				435		



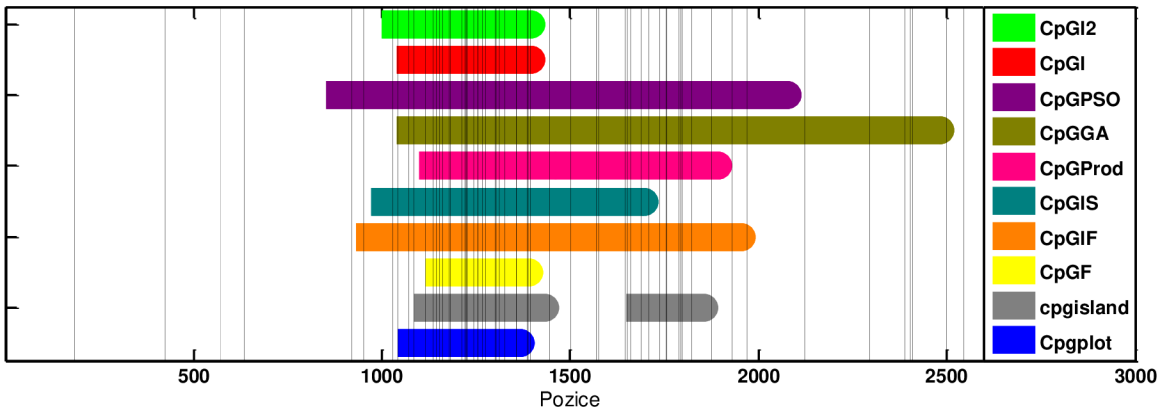
Obr. 25: Porovnání nalezených CpG ostrůvků lokalizovaných jednotlivými vyhledávači v DNA sekvenci genu beta-tubulinu, (seq_4.fasta). Svislé čáry zobrazují výskyt CpG na konkrétní pozici v sekvenci.

Ve třetí zkoumané sekvenci bylo dosaženo algoritmy CpGI a CpGI2 nejpodobnějších výsledků alespoň jednoho ostrůvku jako algoritmy CpGpolt, cpgisland a CpGF.

Souhrn výsledků nalezených pozic začátků, konců a délky CpG ostrůvků v DNA sekvenci genu somatostatin I jednotlivými algoritmy je uveden v Tab. 7. Grafické porovnání je pak uvedeno na Obr. 26.

Tab. 7: Porovnání nalezených pozic začátků, konců a délky CpG ostrůvků lokalizovaných jednotlivými vyhledávači v DNA sekvenci genu somatostatin I (*seq_5.fasta*).

	Cpgplot	cpgisland	CpGF	CpGIF	CpGIS	CpGProd	CpGGA	CpGPSO	CpGI	CpGI2
Začátek										
1	1042	1084	1115	931	972	1099	1040	853	1041	1001
2		1650								
Konec										
1	1372	1436	1395	1957	1700	1895	2486	2081	1400	1400
2		1858								
Délka [bp]										
1	331	353	281	1027	729	797	1447	1229	360	400
2		208								



Obr. 26: Porovnání nalezených CpG ostrůvků lokalizovaných jednotlivými vyhledávači v DNA sekvenci genu somatostatin I (seq_5.fasta). Svislé čáry zobrazují výskyt CpG na konkrétní pozici v sekvenci.

Na Obr. 25 a Obr. 26 je pro hlubší analýzu pomocí svislých čar vyjádřen výskyt CpG v sekvenci. Konkrétně na Obr. 25 na pozicích cca 500 až 1600 je viditelný rozdíl algoritmů v detekci. CpGI2, Cpgplot a cpgisland detekují dva samostatné CpG ostrůvky, CpGI a CpGF detekuje pouze jeden ale totožný alespoň s jedním z obou. Ostatní algoritmy tyto dva CpG ostrůvku lokalizovali jako součást jednoho delšího CpG ostrůvku.

Vzhledem k tomu, že nejsou k dispozici DNA sekvence s přesně vymezenými CpG ostrůvky, není ani možné přesně určit, jak je navržené řešení účinné. Přesto, na základě výsledků dosažených jinými přístupu, jejich porovnáním a především podle shody nalezených CpG ostrůvků lze určit, zda řešení poskytuje dobré výsledky. Na základě porovnání dosažených výsledků, lze obecně usoudit, že v porovnání mezi CpGI a CpGI2 lepších výsledků je dosahováno algoritmem CpGI2. V porovnání dosažených výsledků metody CpGI2 a ostatních vyhledávačů, jsou výsledky touto metodou horší porovnání při lokalizaci krátkých CpG ostrůvků ale na druhou stranu je metoda striktnější a lokalizuje spíše významnější ostrůvky při vyšším obsahu CG a vyšší hodnoty O/E.

Závěr

V této diplomové práci bylo teoreticky pojednáno o významu vyhledávání CpG ostrůvků a o metodách, které tyto ostrůvky vyhledávají v DNA sekvencích. Byl tak vytvořen souhrn snad veškerých používaných či jen navržených metod, který tak může poskytovat přehled o těchto metodách na jednom místě.

Stěžejní částí práce byla samotná realizace vyhledávače CpG ostrůvků, která je založená na analýze DNA spektrogramu. Vzhledem k možnostem řešení dané problematiky byly realizovány dva přístupy analýzy DNA spektrogramu. Jeden založen na předpokladu, že CpG ostrůvek má v DNA vyšší zastoupení obsahu cytosinu a guaninu než je oblast bez výskytu CpG ostrůvku. Na těchto poznatcích byly realizovány dva DNA spektrogramy pro nukleotidy C, G a pro A, T. Tyto spektrogramy pomocí sumací jednotlivých spekter byly porovnány a tím tak lokalizovány (predikovány) CpG ostrůvky. Protože tento přístup má jistá omezení a za účelem dosáhnout lepších výsledků, byl realizován druhý přístup. Druhý přístup je založen na faktu, že CpG se v ostrůvcích vyskytují s vyšší frekvencí než na částech sekvence DNA bez výskytu CpG ostrůvku. Tím bylo dosaženo značného zjednodušení a lepších výsledků.

Zásadní vliv na dosažené výsledky má samotná konstrukce DNA spektrogramu kdy změnou délky plovoucího okna či změnou posunu okna podél pozic v DNA sekvenci dochází k jiným výsledkům. Proto bylo zapotřebí prozkoumat vliv změny těchto parametrů na dosažené výsledky, a na základě těchto poznatků vstupní hodnoty optimalizovat. Byl tak získán kompromis mezi hodnotami těchto parametrů. Přesto uživatel programu může tyto hodnoty na vstupu algoritmu upravit.

V poslední části práce byly dosažené výsledky porovnány s jinými osmi přístupy vyhledávání CpG ostrůvků, které jsou volně dostupné na internetu. Bylo zde také uvedeno stručné představení těchto vyhledávačů s odkazem na teoretickou část práce a autorů metod, díky tomu tato práce může sloužit zájemcům o studium metodách vyhledávání CpG ostrůvků z DNA sekvencí.

Protože nejsou k dispozici sekvence s přesně známými pozicemi CpG ostrůvků, k hodnocení navržených algoritmů postačily výsledky ostatních metod. Na základě porovnání výsledků algoritmů CpGI a CpGI2 bylo potvrzeno, že vyhledávání CpG

ostrůvků na základě analýzy DNA spektrogramu je vhodným nástrojem k lokalizaci CpG ostrůvků a vede v porovnání k dobrým výsledkům. Při porovnání algoritmu CpGI a CpGI2 jsou dosažené výsledky algoritmu CpGI2 lepší a je tedy vhodnější volbou pro vyhledávání CpG ostrůvků.

Seznam použité literatury

[1] GARDINER-GARDEN, M a M FROMMER. *CpG islands in vertebrate genomes.* Journal of Molecular Biology. 1987, vol. 196, s. 261-282.

[2] TAKAI, Daiya and Peter A. JONES. *Comprehensive analysis of CpG islands in human chromosomes 21 and 22.* Proceedings of the National Academy of Sciences, pp. 3740-3745, 2002.

[3] *DNA Methylation and CpG Island.* Web - books [online]. [cit. 2014-01-02]. Dostupné z: http://www.web-books.com/MoBio/Free/Ch7F2.htm

[4] DIMITROVA, Nevenka, Yee H. CHEUNG and Michael ZHANG. *Analysis and Visualization of DNA Spectrograms: Open Possibilities for the Genome Research.* Proceedings of the 14th annual ACM international conference on Multimedia, pp. 1017-1024, 2006.

[5] MNEIMNEH, Saad. *Computational Biology. Lecture 9: CpG islands, Markov Chains, Hidden Markov Models HMMs*. [online]. [cit. 2014-01-02]. Dostupné z: http://www.vutbr.cz/elearning

[6] BAŠTINEC, J.; SVOBODA, Z. Náhodné procesy. Brno: 2011. s. 1-182.

[7] *HMM: Viterbi algorithm – a toy expample* [online prezentace]. [cit. 2014-01-04]. Dostupné z: http://www.vutbr.cz/elearning

[8] ALBERTS, Bruce. *Základy buněčné biologie: úvod do molekulární biologie buňky.* 2. vyd. Překlad Arnošt Kotyk, Bohumil Bouzek, Pavel Hozák. Ústí nad Labem: Espero, c1998, 1 sv. (různé stránkování). ISBN 80-902-9062-0.

[9] PONGER, L. a D. MOUCHIROUD. CpGProD: identifying CpG islands associated with transcription start sites in large genomic mammalian sequences. Bioinformatics. 2002, vol. 18, issue 4, s. 631-633.

[10] MURRELL, Adele. DNA Methylation and Genome Stability. University of Cambridge [online]. [cit. 2014-04-21]. Dostupné z: http://www-medchem.ch.cam.ac.uk/lab_rotations/murrell.php

[11] BUCK-KOEHNTOP, Bethany A. a DEFOSSEZ. *On how mammalian transcription factors recognize methylated DNA*. Epigenetics. 2013, vol. 8, issue 2, s. 131-137.

[12] LIU, Zhenqiu, Dechang CHEN a Xue-wen CHEN. *CpG Island Identification with Higher Order and Variable Order Markov Models*. Data Mining in Biomedicine. Boston, MA: Springer US, 2007, s. 47. DOI: 10.1007/978-0-387-69319-4_4.

[13] IRIZARRY, Rafael A, Christine LADD-ACOSTA, Bo WEN, Zhijin WU, Carolina MONTANO, Patrick ONYANGO, Hengmi CUI, Kevin GABO, Michael RONGIONE, Maree WEBSTER, Hong JI, James B POTASH, Sarven SABUNCIYAN a Andrew P FEINBERG. *The human colon cancer methylome shows similar hypo- and hypermethylation at conserved tissue-specific CpG island shores.* Nature Genetics. 2009-1-18, vol. 41, issue 2, s. 178-186.

[14] DIETERICH, Christoph. *Algorithms in Bioinformatics I: Markov chains and Hidden Markov Models.* In: University of Tübingen: Faculty of Computer Science [online]. 6.2.2007. 2007 [cit. 2014-04-07]. Dostupné z: http://ab.inf.uni-tuebigen.de/teaching/ws06/albi1/script/MarkovChainsAndHMMs_complete.pdf

[15] SPONTANEO, Leah a Nick CERCONE. *Correlating CpG islands, motifs, and sequence variants in human chromosome 21.* BMC Genomics 2011, vol. 12, Suppl 2, S10-.

[16] HACKENBERG, M., PREVITI, C., LUQUE-ESCAMILLA, P. L., CARPENA, P., MARTÍNEZ-AROZA, J., & OLIVER, J. L. (2006). *CpGcluster: a distance-based algorithm for CpG-island detection.* BMC Bioinformatics, vol. *7*, s. 446-458.

[17] SOFTWARE AND PACKAGES: CpGProD (CpG Island Promoter Detection). PRABI Rhone-Alpes Bioinformatics Center [online]. 2001 [cit. 2014-05-01]. Dostupné z:http://doua.prabi.fr/software/cpgprod_query

[18] CHUANG, Li-Yeh, Cheng-Huei YANG, Ming-Cheng LIN a Cheng-Hong YANG. *CpGPAP: CpG island predictor analysis platform. BMC Genetics*. 2012, vol. 13, issue 1.

[19] RICE, Peter, Ian LONGDEN a Alan BLEASBY. EMBOSS: The European Molecular Biology Open Software Suite. *Trends in Genetics*. 2000, vol. 16, issue 6, s. 276-277.

[20] *Cpgisland: Locate CpG islands in DNA sequence.* MathWorks: Documentation Center [online]. 2014 [cit. 2014-05-17]. Dostupné z: http://www.mathworks.com/help/bioinfo/ref/cpgisland.html

[21] SUSSILLO, David, Anshul KUNDAJE, Dimitris ANASTASSIOU a Sean A. FULOP. *Spectrogram Analysis of Genomes. EURASIP* Journal on Advances in Signal Processing. 2004, vol. 2004, issue 1, s. 127-165.

[22] *EMBOSS Cpgplot: Identify and plot CpG islands in nucleotide sequence(s).* The European Bioinformatics Institute: Part of the European Molecular Biology Laboratory. [online]. 2014 [cit. 2014-05-19]. Dostupné z: https://www.ebi.ac.uk/Tools/seqstats/emboss_Cpgplot/

[23] *Services Test Online: CpGFinder.* Softberry [online]. 2014 [cit. 2014-05-19]. Dostupné z: http://www.softberry.com/berry.phtml?topic=cpgfinder&group=programs&subgroup=promoter

[24] *CpG island Finder.* DataBase of CpG islands and Analytical Tool (DBCAT) [online]. [cit. 2014-05-19]. Dostupné z: http://dbcat.cgm.ntu.edu.tw/

[25] *CpG Island Searcher* [online]. 2004 [cit. 2014-05-19]. Dostupné z: http://cpgislands.usc.edu/

[26] *CpG island Prediction*. CpGPAP: CpG island predictor analysis platform [online]. 2011 [cit. 2014-05-19]. Dostupné z: http://bio.kuas.edu.tw/CpGPAP/CpG%20islands%20prediction.jsp

[27] *Genomová databáze. Index/genomes/H_sapiens/*. NCBI: National Center for Biotechnology Information [online]. 2014 [cit. 2014-05-19]. Dostupné z: ftp://ftp.ncbi.nlm.nih.gov/genomes/H_sapiens/

[28] *National Center for Biotechnology Information* [online]. 2009 [cit. 2014-05-22]. Dostupné z: http://www.ncbi.nlm.nih.gov/

[29] CpGcluster2. *Computational Genomics and Bioinformatics* [online]. 2006 [cit. 2014-05-23]. Dostupné z: http://bioinfo2.ugr.es/CpGcluster/

[30] Hidden Markov Models, the Viterbi Algorithm, and CpG Islands (in VB6). *Tanner Helland* [online]. 2009 [cit. 2014-05-23]. Dostupné z: http://www.tannerhelland.com/1187/hidden-markov-models-viterbi-algorithm-cpg-islands-in-vb6/

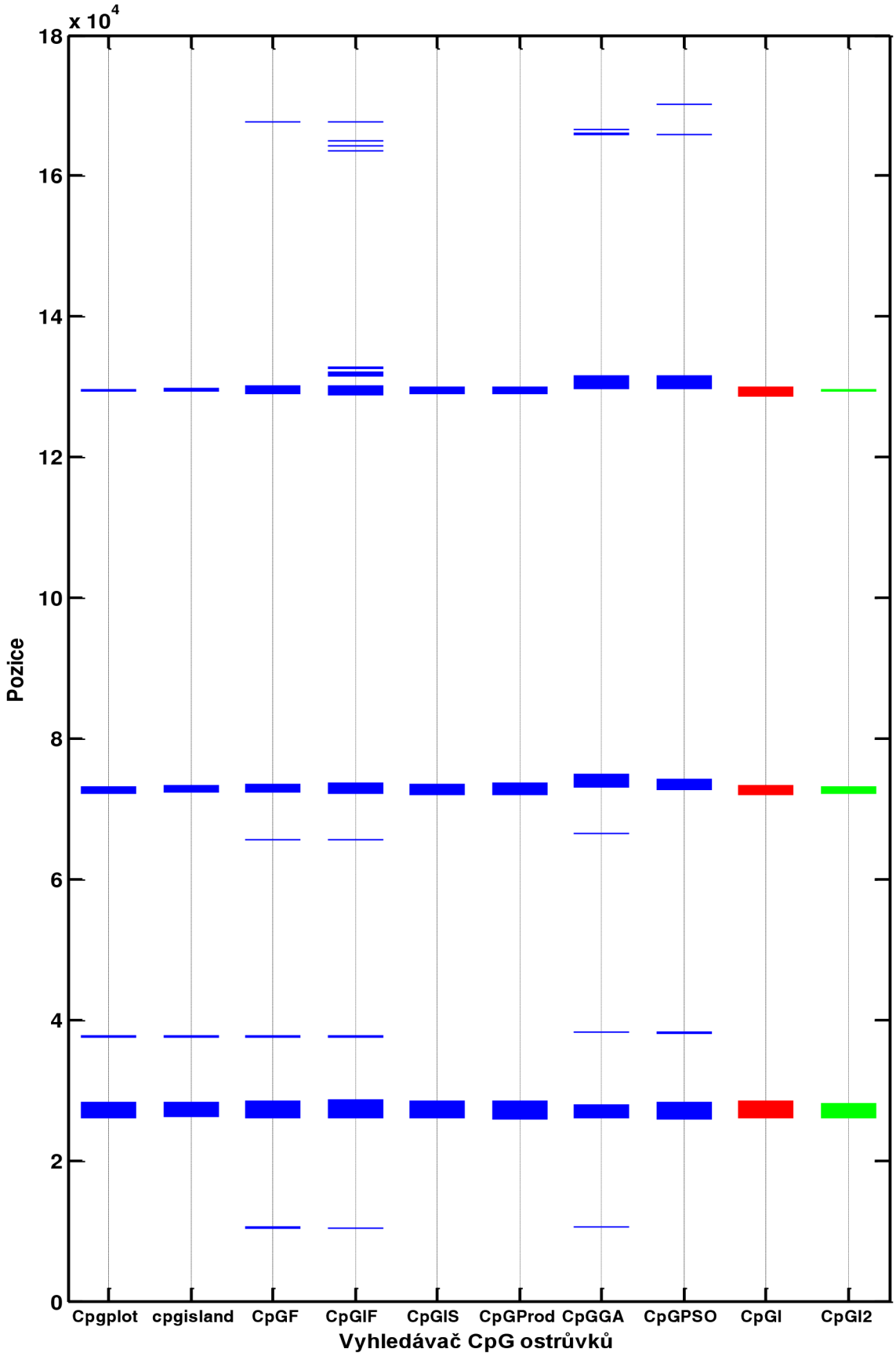
Seznam příloh

A Přehled výsledků analýzy DNA sekvence *seq_2.fasta*

Tab. 8: Souhrn všech výsledků dosažených jednotlivými vyhledávači CpG ostrůvku. Červeně jsou zvýrazněny hodnoty, které se neshodují s nastavením vstupních hodnot vyhledávače.

Číslo	Začátek	Konec	Délka [bp]	CG	E/O
Cpgplot					
1	26074	28323	2250	0,6671	1,1981
2	37472	37743	272	0,6140	0,7884
3	72178	73281	1104	0,6821	1,2869
4	129323	129757	435	0,7126	1,0060
cpgisland					
1	26127	28340	2214	0,6707	1,1935
2	37525	37799	275	0,5345	0,9774
3	72231	73314	1084	0,6670	1,3510
4	129394	129815	422	0,6872	1,0397
CpGF					
1	10275	10554	270	0,5714	0,6222
2	26091	28523	2433	0,6695	1,1405
3	37482	37762	281	0,6192	0,8653
4	65457	65687	231	0,5541	0,7555
5	72229	73537	1309	0,6341	1,3892
6	128884	130174	1291	0,6770	0,6293
7	167577	167842	226	0,6015	0,7093
CpGIF					
1	10255	10531	277	0,5271	0,5787
2	26071	28624	2554	0,6660	1,1001
3	37462	37891	430	0,5000	0,8626
4	65437	65697	261	0,5172	0,7711
5	72177	73647	1471	0,6105	1,3411
6	128864	130172	1309	0,6730	0,6212
7	131511	132158	648	0,6188	0,3891
8	132491	132887	397	0,6322	0,3808
9	163399	163697	299	0,5418	0,6428
10	164094	164408	315	0,5238	0,6508

11	164844	165064	221	0,5520	0,5464
12	167557	167837	281	0,5836	0,6702
CpGIS					
1	25961	28433	2473	0,6494	1,1658
2	72013	73531	1519	0,6011	1,3354
3	129030	130029	1000	0,6840	0,6379
CpGProd					
1	25891	28563	2673	0,6382	1,1367
2	71908	73735	1725	0,5574	1,2987
3	128991	130034	1044	0,6810	0,6251
CpGGA					
1	10368	10715	348	0,4052	0,6309
2	25995	27942	1948	0,6504	1,3459
3	38127	38340	214	0,2430	0,3208
4	66372	66656	285	0,3018	0,1555
5	72980	74975	1996	0,4474	0,7080
6	129702	131586	1885	0,4255	0,3903
7	165742	166100	359	0,3510	0,2716
8	166468	166691	224	0,2946	0
CpGPSO					
1	25764	28370	2607	0,6168	1,2170
2	38011	38322	312	0,2724	0,1818
3	72622	74300	1679	0,5015	1,1205
4	129727	131564	1838	0,4217	0,3450
5	165746	165992	247	0,3320	0,2940
6	169996	170335	340	0,3765	0,2500
CpGI					
1	25921	28480	2558	0,6418	1,1616
2	72021	73360	1338	0,6037	1,3506
3	128682	130061	1378	0,6435	0,6133
CpGI2					
1	26060	28219	2158	0,6690	1,2216
2	72180	73279	1098	0,6836	1,2848
3	129320	129719	398	0,7325	1,0432



Obr. 27: Grafické porovnání detekovaných CpG ostrůvků dle metod.