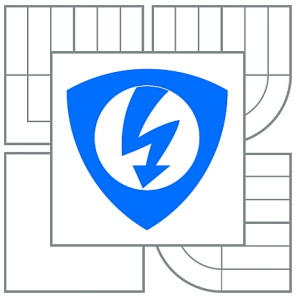
VYSOKÉ UČENÍ TECHNICKÉ V BRNĚ

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 POMOCÍ NUKLEOTIDOVÝCH DENZITNÍCH VEKTORŮ

NUCLEOTIDE DENSITY FOR FINDING CPG ISLANDS

BAKALÁŘSKÁ PRÁCE

BACHELOR'S THESIS

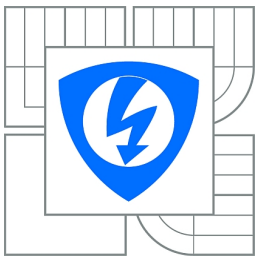
AUTOR PRÁCE — EVA SIKOROVÁ

AUTHOR

VEDOUCÍ PRÁCE — Ing. DENISA MADĚRÁNKOVÁ

SUPERVISOR

BRNO 2013



VYSOKÉ UČENÍ TECHNICKÉ V BRNĚ

Fakulta elektrotechniky a komunikačních technologií

Ústav biomedicínského inženýrství

Bakalářská práce

bakalářský studijní obor
Biomedicínská technika a bioinformatika

Studentka: Eva Sikorová ***ID:*** 136486
Ročník: 3 ***Akademický rok:*** 2012/2013

NÁZEV TÉMATU:

Vyhledávání CpG ostrůvků pomocí nukleotidových denzitních vektorů

POKYNY PRO VYPRACOVÁNÍ:

1) Proveďte literární rešerši metod pro predikci CpG ostrůvků v nukleotidových sekvencích. 2) Vytvořte v libovolném programovém prostředí funkci pro výpočet nukleotidových denzitních vektorů s možností volby velikosti výpočetního okna a s grafickým výstupem dle biochemických vlastností nukleotidů. 3) Vytvořenou funkci otestujte na uměle vytvořených sekvencích s vloženými CpG ostrůvky. 4) Vytvořte pseudokód a vývojový diagram pro postup analýzy sekvencí na výskyt CpG ostrůvků pomocí nukleotidových denzit. 5) Sestavte soubor alespoň 20 sekvencí, u kterých se předpokládá výskyt CpG ostrůvků, a soubor analyzujte vytvořenou funkcí. 6) Vyberte libovolné dva volně přístupné nástroje pro predikci CpG ostrůvků a proveďte analýzu souboru sekvencí. Výsledky přehledně porovnejte s výsledky analýzy pomocí nukleotidových denzit.

DOPORUČENÁ LITERATURA:

[1] ANTEQUERA, F. Structure, function and evolution of CpG island promoters. Cell. Mol. Life Sci. 2003, roč. 60, s. 1647-1658.
[2] GARDINER-GARDEN, M., FROMMER, M. CpG islands in vertebrate genomes. J. Mol. Biol. 1987, roč. 196., s. 261-282.

Termín zadání: 11.2.2013 ***Termín odevzdání:*** 31.5.2013

Vedoucí práce: Ing. Denisa Maděránková
Konzultanti bakalářské práce:

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

UPOZORNĚNÍ:

ABSTRAKT

Tato bakalářská práce se zabývá vyhledáváním CpG ostrůvků v sekvencích DNA pomocí nukleotidových denzitních vektorů. První část práce obsahuje náhled na strukturu DNA, expresi genetické informace, podrobnější rozbor CpG ostrůvku a především metody jejich detekce. V praktické části byl v prostředí MATLAB realizován algoritmus pro grafické zobrazení nukleotidů na základě jejich denzit a detekci CpG ostrůvků z uměle vytvořených i reálných sekvencí DNA. Součástí práce je analýza dvaceti sekvencí s očekávaným obsahem CpGIs a srovnání výsledků vytvořeného programu s dvěma internetovými vyhledávači.

KLÍČOVÁ SLOVA

DNA, exprese, transkripce, DNA metylace, CpG ostrůvky, nukleotidová denzita

ABSTRACT

This bachelor thesis deals with searching for CpG islands in the DNA sequences using the nucleotide density vectors. The first part includes view of the DNA structure, the genetic information expression, more detailed analysis of CpG islands and primarily their detection methods. In the practical part the algorithm for graphical representation of nucleotides on the basis of their densities and detections of CpG islands of artificially created and real DNA sequences was realized in the MATLAB program. The thesis includes the analysis of twenty sequences with the expected content of CpGIs and the comparison of results between the created program and two search engines.

KEYWORDS

DNA, expression, transcription, CpG islands, DNA methylation, nucleotide density

BIBLIOGRAFICKÁ CITACE

SIKOROVÁ, E. *Vyhledávání CpG ostrůvků pomocí nukleotidových denzitních vektorů.* Brno: Vysoké učení technické v Brně, Fakulta elektrotechniky a komunikačních technologii, 2013. 57 s. Vedoucí bakalářské práce Ing. Denisa Maděránková.

Prohlášení

Prohlašuji, že svou bakalářskou práci na téma „Vyhledávání CpG ostrůvků pomocí nukleotidových denzitních vektorů“ jsem vypracovala samostatně pod vedením vedoucího bakalářské 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 autorka uvedené bakalářské práce dále prohlašuji, že v souvislosti s vytvořením této práce jsem neporušila autorská práva třetích osob, zejména jsem nezasáhla nedovoleným způsobem do cizích autorských práv osobnostních a jsem si plně vědoma následků porušení ustanovení § 11 a následujících autorského zákona č. 121/2000 Sb., včetně možných trestněprávních důsledků vyplývajících z ustanovení části druhé, hlavy VI. díl 4 Trestního zákoníku č. 40/2009Sb.

V Brně dne .....

podpis autorky

Poděkování

Děkuji vedoucí bakalářské práce Ing. Denise Maděránkové za účinnou metodickou, pedagogickou a odbornou pomoc a další cenné rady pří zpracování mé bakalářské práce.

V Brně dne .....

podpis autorky

OBSAH

SEZNAM OBRÁZKŮ

SEZNAM TABULEK

ÚVOD

Účelem této bakalářské práce je seznámení se současnými metodami predikce CpG ostrůvků v nukleotidových sekvencích a jejich vyhledání pomocí nukleotidových denzitních vektorů jak v umělé sekvenci DNA s definovanými CpG ostrůvky, tak v sekvencích reálných. Vzhledem k jejich předpokládané funkci v transkripční regulaci a jejich významu jako genomických markérů v promotorové predikci, došlo v posledních letech ke značnému úsilí jejich předpovědi. Pro detekci CpG ostrůvků je k dispozici několik algoritmů založených buď na třech sekvenčních parametrech, kterými jsou délka CpG ostrůvku, obsah guaninu a cytosinu a poměr obdržené a očekávané hodnoty výskytu CpG (Obs_{CpG}/Exp_{CpG}), nebo na statistických vlastnostech v sekvencích. V této práci byla pro realizaci detekce CpG ostrůvků zvolena metoda, pracující na základě faktu, zda průměrný výskyt C a G nukleotidů v dané části sekvence překročí předem stanovený práh. [1], [2]

Pro snadnější orientaci a pochopení řešené problematiky je první teoretická část věnována základním poznatkům souvisejících se zvolenou tématikou. Jsou zde obsaženy informace o vlastnostech a struktuře DNA s bližším zaměřením na biochemické vlastnosti dusíkatých bází. Následující kapitola zmiňuje expresi genetické informace, především transkripci a regulaci genové exprese pomocí metylace DNA, která úzce souvisí s CpG ostrůvky. Ty jsou popsány z hlediska jejich charakteristických vlastností, lokalizace a výskytu v určitých genech. Jak již bylo naznačeno výše, poslední úsek teoretické části tvoří nástin metod predikce CpG ostrůvků od nejstarších algoritmů po algoritmy používající se dnes.

Zvolený postup a popis realizace nukleotidových denzitních vektorů použitých pro predikci CpG ostrůvku a výstupy jejich detekce jsou obsahem praktické části. Je zde okomentován princip převedení symbolů nukleotidových sekvencí do numerické reprezentace, která slouží jako vstup pro výpočet nukleotidových denzitních vektorů. Ty jsou rozebrány z hlediska jejich vlastností a volby parametrů, na nichž závisí dosažený výsledek. Vlastní část tvoří pojednání o volbě okna pro výpočet nukleotidových denzitních vektorů, jehož hodnota byla zvolena na základě provedené analýzy. Následující podkapitola je věnována grafické reprezentaci jak samostatných nukleotidů, tak sum denzitních vektorů jednotlivých nukleotidů na základě jejich chemických vlastností.

Samotná detekce CpG ostrůvku je prováděna na výsledné sumě nukleotidových denzitních vektorů cytosinu a guaninu na principu překročení zvoleného prahu. Jsou zde také krátce okomentovány umělé vytvořené sekvence s definovanými CpG ostrůvky. K dispozici jsou opět grafické výstupy příslušných denzit s vyznačeným prahem, kdy je sledován jeho vliv na samotnou detekci. Konkrétní stanovení začátků a konců CpG ostrůvku pro jednotlivé umělé zhotovené sekvence v závislosti na zvoleném prahu jsou přehledně uvedeny v tabulce a

okomentovány v dané kapitole. Na základě tohoto odzkoušení a chování programu při predikci CpGIs u zmíněných sekvencí byly vymyšleny a realizovány další vylepšení algoritmu pro přesnější stanovení požadovaných výstupů, jejichž popis je taktéž součástí práce.

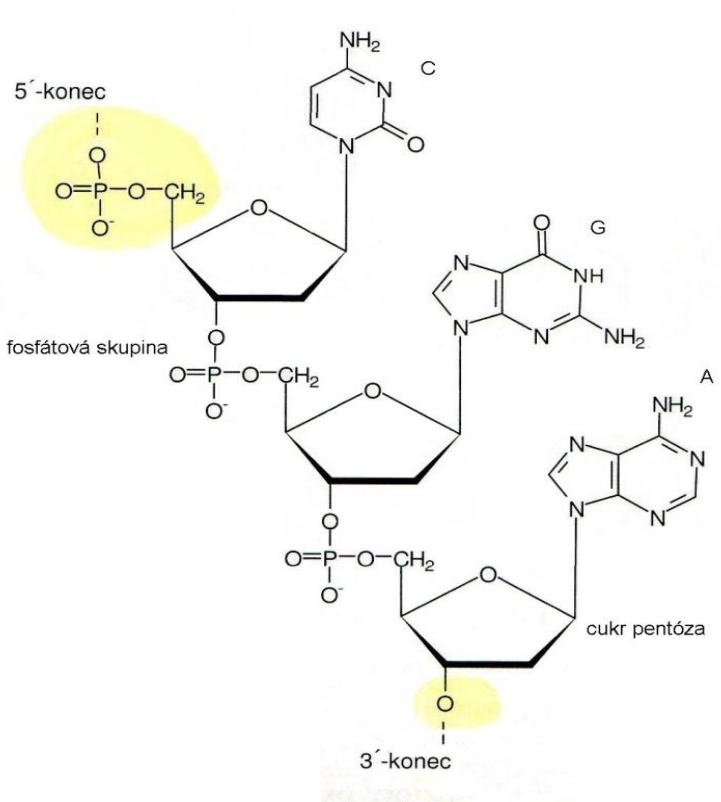
Poslední úsek praktické části je věnován testování konečné podoby programu na dvaceti vybraných reálných DNA sekvencích a jeho srovnání s dvěma volně dostupnými internetovými vyhledávači. Porovnávány byly jednotlivými programy jak ve stanovení počtu CpGIs, tak v určení jejich začátků, konců, délek a GC obsahu. Celá srovnávací analýza je přehledně uspořádána do podoby několika tabulek a patřičně okomentována. Navíc pro snadnější představu o podobnosti nebo rozdílnosti detekce ostrůvků u daných sekvencí bylo vytvořeno grafické zobrazení nalezených pozic CpGIs jednotlivými programy.

1 TEORETICKÁ ČÁST

1.1 DNA

Život všech živých organismů závisí na schopnosti buněk uchovávat, členit, překládat a předávat genetickou informaci, která je nezbytná pro vznik a udržení živého organismu. Nositeli genetické informace jsou molekuly deoxyribonukleové kyseliny, DNA. Informace uložené v DNA poskytují údaje pro syntézu všech proteinů organismu. [3], [4]

Základním stavebním prvkem DNA jsou nukleotidy, které se navzájem spojují fosfodiesterovými vazbami v polynukleotidový řetězec. Tyto vazby vznikají mezi 3‘-OH skupinou jednoho nukleotidu a 5‘-fosfátovou skupinou nukleotidu následujícího. Jednotlivá vlákna nukleových kyselin jsou tvořena repetitivními sekvencemi zbytky kyseliny fosforečné a sacharidové složky deoxyribózy. Na cukr-fosfátovou kostru je pak glykozidovou vazbou navázána konkrétní dusíkatá báze. [4], [5]



Obr. 1: Část řetězce DNA s dusíkatými bázemi cytosin (C), guanin (G), adenin (A). [4]

1.1.1 Dusíkaté báze

Existují čtyři druhy dusíkatých bází nacházejících se v DNA, kterými jsou adenin (A), cytosin (C), guanin (G) a tymin (T). Tyto nukleotidy lze podle biochemických vlastností roztřídit do tří skupin. [6]

V první skupině jsou dusíkaté báze rozděleny podle molekulární struktury na purinové A, G a pyrimidinové T, C. Puriny mají spojený šestičlenný a pětičlenný kruh, oba složené z atomů uhlíku a dusíku. Naopak pyrimidiny jsou menší, tvořeny pouze jedním šestičlenným kruhem. Do skupiny pyrimidinů patří i uracil, který se místo tyminu nachází v molekule RNA. [5], [6]

Druhá skupina podle síly vazby mezi komplementárními bázemi je rozčleněna na A a T, které jsou vzájemně vázány dvěma vodíkovými můstky. Na druhé straně je zde přítomen C a G umožňující spojení třemi vodíkovými můstky. Obsažený radikál je kritériem poslední skupiny, kde se nachází G a T zahrnující ve své stavbě keto skupinu C=O na šestém nebo čtvrtém uhlíku. Obdobným způsobem je umístěna aminoskupina NH_3 u A a C. [6]

1.1.2 Primární struktura DNA

Jako primární struktura DNA je označováno zastoupení a seřazení jednotlivých nukleotidů v polynukleotidovém řetězci. Nejčastěji se vyskytují molekuly DNA tvořeny dvěma řetězci s opačnou polaritou, takzvané antiparalelní uspořádání. První vlákno tvoří báze uspořádané ve směru 5‘→3‘, druhé má opačný směr 3‘→5‘. Jeden konec řetězce, který obsahuje OH skupinu pentózy, se označuje jako 3‘ konec, naopak druhý nazýván 5‘ je zakončen fosfátovou skupinou. [3], [4], [7]

Oba řetězce jsou vzájemně spojeny přes naproti sobě umístěné dusíkaté báze vodíkovými vazbami. Mezi sebou slučitelné jsou však ve vláknech pouze určité báze. Párováním bází je dáno, že jeden řetězec je předvídatelným doplňkem druhého. Adenin se vždy páruje s tyminem skrz dvě vodíkové vazby, na druhé straně pak cytosin s guaninem třemi vodíkovými vazbami. Tímto pravidlem je podmíněno, že naproti bázi purinové leží vždy pyrimidinová. Komplementární párování bází poskytuje možnost zaujmout z hlediska energie nejvýhodnější uspořádání. V této konformaci je zajištěna podobná vzdálenost mezi oběma páry bází a z toho vyplývající udržení stabilní vzdálenosti fosfát-pentózových sekvencí obou řetězců podél celé molekuly DNA. [3], [5]

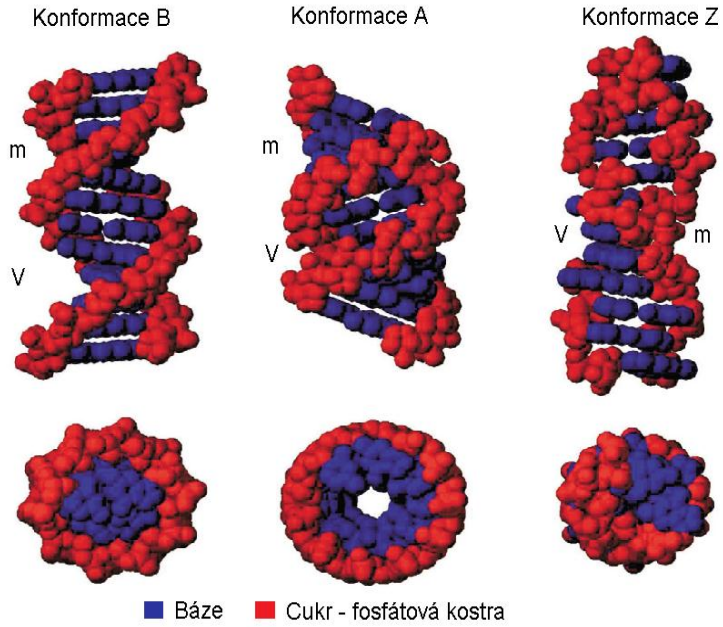
1.1.3 Sekundární struktura DNA

Molekuly DNA, které mají dva polynukleotidová vlákna, můžou vytvářet vyšší, sekundární strukturu. Řetězce molekuly se otáčí kolem imaginární podélné osy a tvoří dvojitou šroubovici, takzvaný helix. Právě tato pravotočivá struktura je nejčastěji se vyskytující formou. [5], [7]

Skelet dvoušroubovice tvoří diesterově vázané molekuly deoxyribózy a fosfátu, dusíkaté báze jsou orientovány do jejího jádra. Oba řetězce jsou navzájem připoutány vodíkovými vazbami mezi kompatibilními nukleotidy ležícími vodorovně nad sebou, což přispívá ke stabilitě dvoušroubovice. Mezi vnitřními bázemi působí van der Waalsovy síly. Vzájemné

obtáčení komplementárních řetězců je příčinou vzniku dvoušroubovicového vinutí, které se vyskytuje v pravotočivé nebo levotočivé formě. Povrch obtočených řetězců vytváří reliéf, na kterém se rozlišují dva žlábky, menší a větší. To je důsledkem skutečnosti, že jednotlivé páry bázi jsou posunuty od osy dvoušroubovice o určitou vzdálenost. Některé skupiny atomů nacházející se ve žlábcích šroubovice se podílí na interakcích s dalšími látkami, nejčastěji s proteiny. [5], [7]

Při určitém chování prostředí se může DNA transformovat do energeticky výhodnější konformace odpovídající daným okolnostem. Nejtypičtějším prostorovým uspořádáním pro živé buňky je konformace B, která odpovídá výše popsané pravotočivé dvoušroubovici. Druhou možností je taktéž pravotočivá konformace A, která se vyskytuje v buňkách za nižší vlhkosti. Obě konformace mohou přecházet jedna v druhou. Poslední ve výčtu základních konformací DNA je na rozdíl od předešlých levotočivá konformace Z. Má své uplatnění v procesu rekombinace, některých regulací genové exprese a jiných specifických procesech. [7]



Obr. 2: Trojrozměrné modely B-konformace, A-konformace a Z-konformace dvoušroubovice DNA. Malé (m) a velké (V) žlábky jsou vyznačeny u každé dvoušroubovice. [9]

1.2 Exprese genetické informace

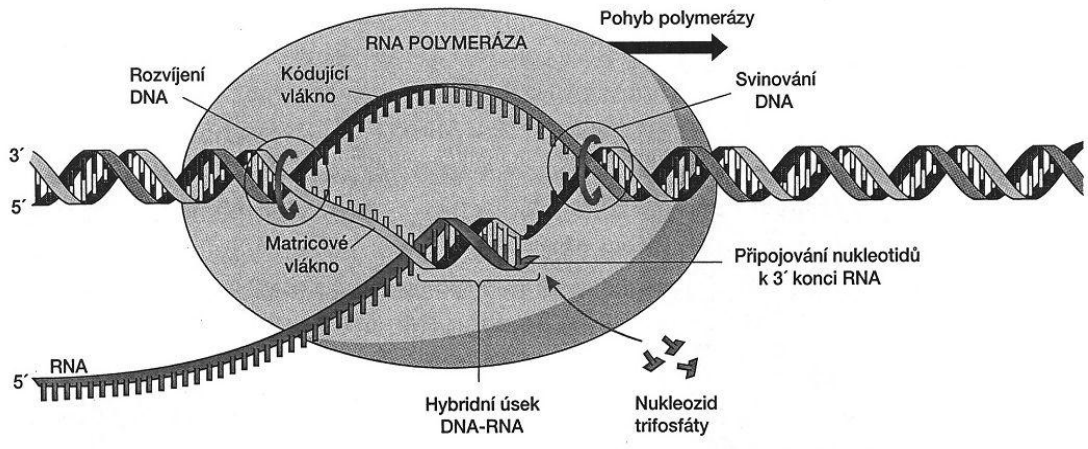
Vlastnosti organismu jsou souhrnem jednotlivých, konkrétních vlastností buněk, které jsou získány přečtením a převedením genetické informace uložené v genech jako lineární posloupnost nukleotidů. Prostřednictvím RNA kopie získané z DNA vznikají genové produkty, s jejichž aktivitou je obvykle spojeno fenotypové vyjádření. Těmito genovými produkty jsou nejčastěji specifické proteiny. Průběh jednosměrné cesty genetické informace z molekuly DNA přes RNA k proteinu je popsán centrálním dogmatem molekulární biologie. Dogma lze zjednodušeně znázornit zápisem DNA→ RNA→ protein. [7]

Pojem exprese genetické informace označuje sérii procesů, které se uskutečňují v průběhu realizace informace zakódované v pořadí nukleotidů. Prvním krokem vyjádření genetické informace je transkripce. Jedná se o proces přepisu genetické informace z DNA do RNA na základě komplementarity bází. Výsledný transkript je přesný přepis instrukcí genu pro vytvoření proteinu. Primární RNA kopie jsou zpravidla poněkud chemicky upraveny při posttranskripční úpravě. [5], [7]

Následující stupeň exprese se nazývá translace, která zaručuje překlad z jazyka nukleotidů do jazyka aminokyselin. Vlastní syntéza polypeptidu pod vedením mRNA je doprovázena posttranslačními modifikacemi, které zahrnují chemickou úpravu translačních výsledků, než se z nich vytvoří konečné bílkoviny. [5], [7]

1.2.1 Transkripce

Procesem transkripce vznikají všechny druhy RNA nacházející se v buňce. Předlohou pro budoucí vlákno RNA je pouze jedno nekódující vlákno duplexu DNA, templátové. Tím je zajištěno, že výsledný transkript je sekvenčně shodný s kódujícím vláknem, avšak tymin je nahrazen uracilem. Postup vzniku RNA probíhá obecně ve třech fázích, kterými jsou iniciace, elongace a terminace. Délka sekvence DNA, která je přepisována do RNA molekuly, se označuje transkripční jednotka. [3], [5], [7]



Obr. 3: Průběh transkripce. [7]

Iniciace je zahájení transkripce, kdy se na promotor naváže RNA polymeráza, která katalyzuje syntézu RNA prostřednictvím dřeňového enzymu a obsahuje sigma faktor, který umožní vazbu enzymu na promotor. Promotor je speciální sekvence dlouhá přibližně 40 bp sloužící jako vazebné místo pro RNA polymerázu, udává začátek transkripce a rozhoduje, který z řetězců DNA bude templátem. U eukaryotických buněk je charakteristický jak pro jednotlivé druhy RNA polymerázy, tak pro transkripční faktory. Tyto faktory jsou regulační proteiny vázající se na regulační místa promotorů nebo na zesilovače transkripce. Taktéž zprostředkovávají navázání RNA polymerázy na promotor, což je následováno částečným

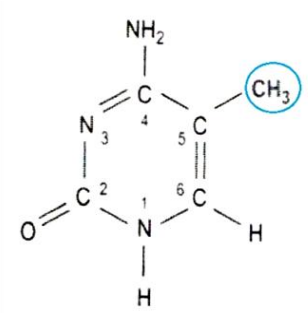
rozpletením obou řetězců dvoušroubovice DNA a zpřístupněním krátké sekvence templátového řetězce. Přiřazením prvního nukleotidu propuká RNA syntéza doprovázená odštěpením sigma faktoru od RNA polymerázy po získání délky kolem 9 nukleotidů. [5], [7]

Elongace řetězce RNA je časově definována v okamžiku odštěpení sigma faktoru s následným postupováním RNA polymerázy podél vlákna DNA. RNA polymeráza se posouvá krok za krokem a v místech styku rozvíjí dvoušroubovicovou strukturu DNA do délky 15-18 bp, tím jsou zpřístupňovány sekvence DNA potřebné k překladu. Vytvářející se RNA kopie je spolu s templátovým řetězcem DNA formována do hybridního úseku DNA-RNA o délce přibližně 12 bp, na jehož startu se RNA odpoutává od DNA molekuly. Nárůst vlákna RNA je uskutečňován ve směru 5‘ → 3‘. [3], [7]

Ukončení transkripce, zváno terminace, nastává v momentu dosažení RNA polymerázy ke specifické sekvenci DNA nazývané terminátor a jeho přepisu. Zastavení elongace RNA řetězce je realizováno buď změnou struktury RNA transkriptu do podoby vlásenky nebo navázáním specifického proteinu Rho faktoru na finální molekulu RNA. [7]

1.2.2 Regulace genové exprese pomocí metylace DNA

Pro regulaci genové exprese je pravděpodobně významná chemická modifikace nukleotidů, realizována napojením metylové CH_3 skupiny na cytosin. Většina metylovaných cytosinů je lokalizována v páru bází se strukturou mCpG, kde mC značí metylcytosin a p označuje fosfodiesterovou vazbu mezi oběma nukleotidy řetězce DNA. Reakci potřebnou pro navázání metylové skupiny na cytosin uskutečňuje enzym metyltransferáza, která rozpozná pouze nemetylovaný cytosin nacházející se vedle guaninu. [7], [10], [11]



Obr. 4: Struktura 5metylcytosinu. [10]

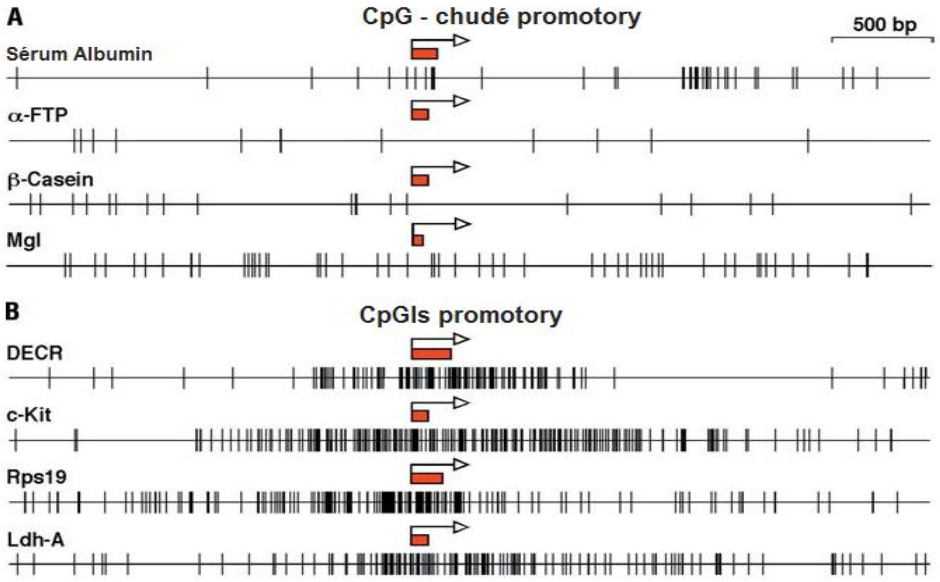
Frekvence dinukleotidů CpG (CpGs), většinou metylovaných, je v lidských genomech nízká, nejspíš proto, že během vývoje mutovaly na TpG dinukleotidy. CpG dinukleotidy nacházející se v DNA jsou navíc rozprostřeny nerovnoměrně, s hojnými krátkými částmi DNA, které mají mnohem větší výskyt CpGs než je jejich průměrná hodnota v genomu. Tyto oblasti s vysokou hustotou CpGs se nazývají CpG ostrůvky (CpGIs). Obsah metylovaného cytosinu v nich je buď velmi výjimečný, nebo žádný v případě aktivních genů a tento stav napomáhá transkripci. Přítomnost metylace v oblasti CpGIs je spojena s deaktivací genů.

Příčinou může být, že metylovaný cytosin zabrání navázání traknskripčních faktorů nebo podporuje vazbu inhibičních komplexů, které obsahují histon deacetylázy a jiné faktory podmiňující změny chromatinu do inaktivní formy. [10], [11]

U savců se vyskytují nezvyklé případy metylace DNA, kdy genetická exprese je podmíněna jeho parentálním původem. Tento stav je nazýván jako imprinting genu a vyjadřuje fakt, že gen byl nějak poznamenán, aby si pamatoval, u kterého rodiče má původ. Tímto označením, které řídí expresi genu, je právě metylace jistého množství dinukleotidů CpG v blízkém okolí genu. Původ tvorby této metylace je v rodičovské zárodečné linii. V průběhu embryogeneze se ne/metylované stavy zachovávají při všech replikacích, ale metylovaný gen, který byl získán od jednoho pohlaví, může být demetylován pří procházení přes potomka opačného pohlaví. V závislosti na pohlaví jsou charakteristické metylace nastaveny znovu v každém pokolení. Fakt, že některé geny jsou metylovány jen u jednoho z pohlaví, poukazuje na to, že metylační aparát je řízen pohlavně závislými faktory. [10]

1.3 CpG ostrůvky

Jak již bylo uvedeno výše, CpG ostrůvky jsou regiony nepodléhající nebo zcela vzácně podléhající metylaci, které mají přibližně desetkrát vyšší obsah cytosinu a guaninu než je jejich průměrný výskyt v genomu. Tyto nápadné sekvence dosahují délky asi 1 kb a jsou nejčastěji lokalizovány na 5‘ konci genu. Podle distribuce CpGs na 5‘ konci genu, mohou být RNA polymerázou II přepisované geny kategorizovány ještě do druhé skupiny. Do ní patří CpG chudé oblasti, většinou podléhají metylaci a jsou tvořeny CpGs jejichž denzita je stejná jako genomický průměr, který je přibližně jeden na každých 100 nukleotidů. Zde zapadají geny, jejichž exprese je omezena na limitovaný počet buněčných typů. [1]



Obr. 5: Dvě třídy lidských a myších promotoru asociovaných s CpG chudými oblastmi (A) a s CpGIs (B). Vertikální linky znázorňují rozložení CpGs, červené obdélníky s šipkami znázorňují první exon a transkripční iniciační místo. [1]

Asi 60 % všech promotorů vyskytujících se u lidí jsou asociovány s CpGIs. Obecně u obratlovců jsou CpGIs spojeny s 5‘ koncem všech provozních genů, které se projevují ve všech buněčných typech, a mnoha tkáňově specifických genů. Rovněž jsou v této skupině zahrnuty CpGIs asociovány s 3‘ koncem některých tkáňově specifických genů. Pár genů obsahuje jak 5‘, tak 3‘ CpGIs, které jsou odděleny několika tisíci bp CpG chudých oblastí. Na druhé straně 20 % lidských promotorů asociovaných s CpGIs jsou CpG nedostatečné, odpovídající myším ortologním genům. U ortologních genů obou organizmů byla pozorována přítomnost nebo absence CpGIs. Vysvětlením tohoto jevů může být, že některé lidské geny mohly CpGIs získat nebo odpovídající myší geny je mohly ztratit. Možnou příčinou je odchýlení obou druhů od společného předka před 65 milióny lety. [1], [12]

Nedávné srovnávací studie mezi oběma organismy umožnily z větší míry pochopit, jak jsou CpG ostrůvkové promotory organizované z hlediska protein-DNA interakcí a schématu exprese. Spouštění DNA replikace in vivo na CpGIs naznačuje, že jejich typické vlastnosti mohou být následkem této aktivity a otvírají možnost koordinované regulace transkripce a replikace. [1]

1.4 Metody pro predikci CpG ostrůvků v nukleotidových sekvencích

Pro identifikaci CpG ostrůvků byly vyvinuty řady algoritmů, které lze rozdělit do dvou skupin. Do první kategorie náleží tradiční algoritmy založené na třech sekvenčních parametrech, kterými jsou délka CpG ostrůvku, obsah guaninu a cytosinu a poměr obdržené a očekávané hodnoty výskytu CpG (Obs_{CpG}/Exp_{CpG}). Druhou skupinu tvoří algoritmy založené na statistických vlastnostech v sekvencích bez závislosti zmíněných tří kritérií uvedených výše. [2]

1.4.1 Gardiner-Gardenův a Frommerův algoritmus

V roce 1987 byl Gardiner-Gardenem a Frommerem navržen původní, na třech kritériích založený, algoritmus pro identifikaci CpG ostrůvku v sekvencích genu. Algoritmus pracuje na principu plovoucího okna o délce 200 bp, které se posouvá podél sekvence, GC obsah musí být vyšší než 50 % a Obs_{CpG}/Exp_{CpG} hodnota větší než 0,60. Poměr obdržené a očekávané hodnoty výskytu CpG je počítán podle následujícího vzorce:

$$\frac{Obs}{Exp} Cpg = \frac{počet\ CpG}{počet\ C \cdot počet\ G} \cdot N\ , \qquad (1)$$

kde N je celkový počet nukleotidů v analyzované sekvenci. [12], [13], [14]

Tento algoritmus, často s jistými úpravami, byl široce uplatňován v analýzách jednotlivých genů, nevelkých souborech genomických sekvencí nebo jednotlivých genomech. Problém nastává v genomech obratlovců s hojným počtem repetičních oblastí, jako jsou například Alu repetice, které také splňují uvedené tři kritéria. Alu repetice jsou krátké

roztroušené elementy o délce asi 280 bp s často vysokým obsahem guaninu a cytosinu opakujícího se v rámci genomu. Pro zamezení této falešné detekce byl algoritmus používán k vyhledávání CpG ostrůvku pouze u nerepetitívních částí genomu. [13], [14]

1.4.2 Takai a Jonesův algoritmus

Pomocí analýzy dat lidských chromosomů 21 a 22 hodnotili Takai a Jones tři parametry z Gardiner-Gardenova a Frommerova algoritmu a navrhli optimální set kritérií, kdy CG obsah ≥ 55 %, délka CpG ostrůvku ≥ 500 bp a Obs_{CpG}/Exp_{CpG} hodnota ≥ 0,65. Výsledkem je algoritmus, který může účinně vyloučit falešně pozitivní CpG ostrůvky v repeticích a spíše identifikovat CpG ostrůvky asociované s 5‘ koncem lidských genů. Opakující se oblasti, jako Alu repetice, nejsou ve výsledku zahrnuty tak často jako v předešlém případě. Použitím přísnějších kritérií je dosaženo lepších detekčních výsledků a zároveň se zdá, že tento postup je také vhodný pro ostatní geny. [13], [14]

1.4.3 CpGcluster

Kolektiv autorů v čele s Hackenbergem vyvinuli nový algoritmus, nazývaný CpGcluster, který zcela závisí na statistických vlastnostech CpG shluků v náhodných sekvencích lokalizovaných ve stejném chromosomu. CpGcluster detekuje shluky CpGs podle statistické významnosti na základě fyzických vzdáleností mezi sousedními CpGs na chromosomu na základě faktu, že vzdálenost mezi distribucí sousedních CpGs se liší na CpG ostrůvcích od zbylé části DNA sekvence. CpG shluky jsou definovány jako CG-husté fragmenty detekčně založeny na empirických druhově specifických měřítcích. [2], [14]

Celý algoritmus se skládá ze dvou hlavních kroků. Prvním z nich je nalezení CpG shluků v sekvenci chromozomu. Jak již bylo uvedeno výše tento krok je založený na statistických vlastnostech fyzické vzdálenosti mezi vedlejšími CpGs v DNA sekvenci, kdy bohaté CpGs v CpGIs mohou být odděleny kratšími vzdálenostmi za vzniku shluků. V zásadě, jestli je distribuce CpGs podél sekvence zcela náhodná, vzdálenosti mezi sousedními CpGs by měly sledovat geometrické rozdělení:

$$P(d) = (1-p)^{d-1}p \ , \tag{2}$$

kde *P(d)* je pravděpodobnost nalezení vzdálenosti d mezi sousedními CpGs a p reprezentuje pravděpodobnost CpGs v sekvenci, vypočtené jako poměr mezi CpGs a celkovým počtem dinukleotidů v DNA sekvenci. [15]

Samotná realizace začíná skenováním sekvence ve směru 5‘→3‘ a zaznamenáváním pozic obsazených cytozinem nebo guaninem. Poté následuje provedení výpočtu fyzické vzdálenosti oddělující dva sousední CpGs, kdy první vzdálenost pod předem stanovenou prahovou hodnotou určuje první CpG shluk. Růst prvního shluku po proudu pokračuje přidáváním dalších CpGs dokud je počítaná distance stále pod zvoleným prahem. Po

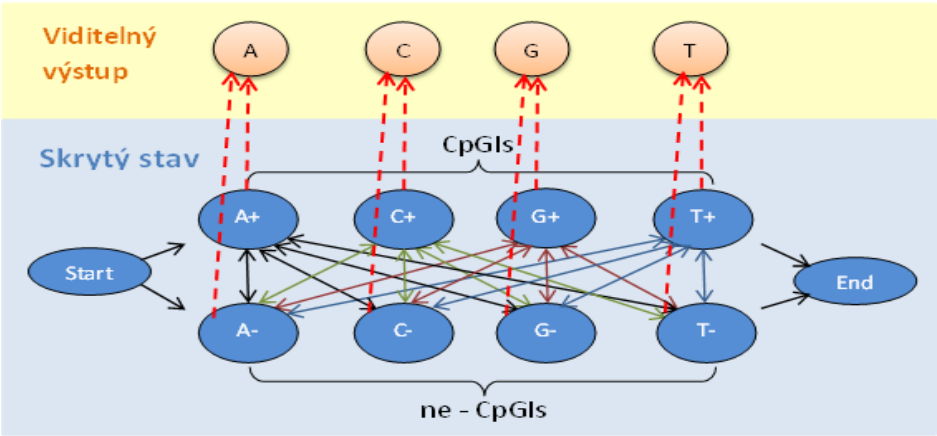
překročení hranice je shluk vyhodnocen jako kompletní a započíná hledání dalších, dokud nejsou identifikovány všechny CpG shluky v celé sekvenci. [15]

Druhým krok sestává z přiřazení p-hodnoty každému shluku nalezenému algoritmem uvedeným výše. P-hodnota pak předpovídá jako CpGIs pouze ty shluky s dostatečně velkou statistickou významností, tedy ty, jejichž p-hodnoty jsou pod nastavenou prahovou hodnotou. Její velikost lze odhadnout buď numericky pomocí náhodného testu na DNA sekvenci eventuálně prostřednictvím teoretických pravděpodobnostních funkcí. [15]

Protože CpGcluster nevyžaduje minimální délku, je pravděpodobné, že detekuje mnohem větší počet CpG ostrůvků než ostatní algoritmy. Především, může tento algoritmus přehánět počet CpG ostrůvků na chromosomech s nízkým obsahem GC, které mají často nízkou genovou denzitu, protože jejich CpG shluky byly identifikovány vzhledem k CpG vlastnostem pozadí. Na druhou stranu může CpGcluster najít krátké, ale plně funkční CpG ostrůvky obvykle nenalezené ostatními algoritmy, které navíc vždy začínají i končí CpG dinukleotidy. Jelikož algoritmus používá pouze celočíselné počty, vede k rychlému a výpočetně efektivnímu nálezu CpGIs. [14], [15]

1.4.4 Skrytý Markovův model

Nedávné studie ukázaly, že použití skrytých Markovových modelů (HMM) v detekčních algoritmech může vést k vylepšení dosažených výsledků. HMM je pravděpodobnostní stavový model, který na základě pravděpodobnosti přechází mezi jednotlivými stavy, kdy počty stavů jsou konečné. Tento model se skládá z Markova procesu, ve kterém je stav skrytý. To znamená, že zvenčí není stav, ve kterém se nachází přesně zjistitelný, ale je k dispozici informace o výstupu, který nadejde s určitou pravděpodobností. [13], [16]



Obr. 6: Schéma skrytého Markovova modelu pro CpGIs.

Markovův proces je náhodný jev, u kterého jsou budoucí pravděpodobnosti stanoveny na základě nejnovějších hodnot. Zmíněný model vyžaduje znalost několika pravděpodobností, které mohou být rozděleny do tří skupin. Do první skupiny patří počáteční pravděpodobnosti, které jsou často stejné pro různé stavy, a stanovují, ve kterém stavu bude systém na začátku

algoritmu. Přechodové pravděpodobnosti, patřící do druhé skupiny, určují s jakou pravděpodobností, se může model ze stavu *i* dostat do stavu *j*. Poslední skupinou jsou výstupní pravděpodobnosti určující, pravděpodobnost symbolu *b* na výstupu náležícího do stavu *k*. [13]

Skrytý Markovův model se skládá ze dvou Markovových řetězců, jeden popisuje stav pro CpGIs (+) a druhý řetězec označuje stav pro ne-CpGIs (-). Přepínání z jednoho řetězce na druhý na každém bodu přechodu probíhá pouze s malou pravděpodobností. S existencí těchto dvou stavů nastává problém, kdy není známo, z kterého stavu pochází aktuální nukleotid na výstupu. Cílem je právě určení sekvence stavů ze známé sekvence symbolů. Tato sekvence stavů je nazývána cesta a značí se symbolem π. [17]

Pravděpodobnost získání sekvence symbolů $x = (x_1, \ldots, x_n)$ pro konkrétní cestu $\pi = (\pi_1, \ldots, \pi_n)$ je obecně dána následujícím vzorcem:

$$p(x_1 \ldots x_n, \pi_1 \ldots \pi_n) = p(x_1 \ldots x_n, \pi_0 = 0, \pi_1 \ldots \pi_n, \pi_{n+1} = 0) =$$

$$= a_{0\pi_1} \prod_{i=1}^{n} e_{\pi_i}(x_i) a_{\pi_i \pi_{i+1}}, \qquad (3)$$

kde $\pi_{n+1} = 0$, $e_\pi(x)$ je výstupní pravděpodobnost, a_{ij} je přechodová pravděpodobnost a n je délka sekvence. Bohužel, obvykle není známa cesta skrz model. Proto následuje dekódování, tedy určení, které části sekvence budou klasifikovány jako CpGIs a které nikoli. Protože je více cest, které vedou ke generování stejné sekvence s rozdílnou pravděpodobností, cílem je nalezení nejpravděpodobnější cesty. To lze provést například pomoci Viterbiho algoritmu popsaného níže. [17], [18]

1.4.5 Viterbi algoritmus

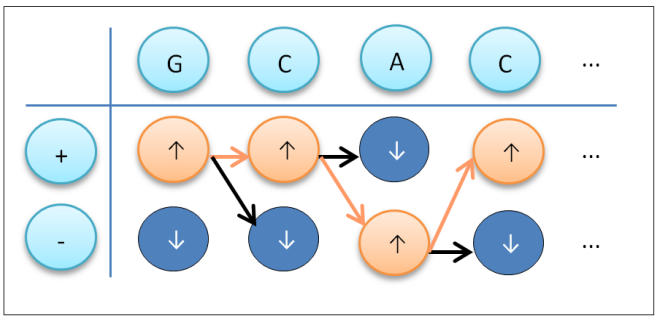
Viterbiho algoritmus je dynamický programovací algoritmus vytvořený Adnrewem Viterbim v roce 1967. Algoritmus umožňuje vypočítat nejpravděpodobnější cestu vedoucí k dané sekvenci, protože ve skutečnosti existuje několik cest vedoucích přes skryté stavy k dané sekvenci, které však nemají stejnou pravděpodobnost. Viterbiho cesta obvykle souvisí se skrytými Marokovovými modely, kdy vyžaduje znalost parametrů HMM a konkrétní výstupní sekvenci, poté pracuje na nalezení maxima přes všechny možné stavy sekvence (CpGIs a ne-CpGIs). [13], [19]

Pravděpodobnost nejpravděpodobnější cesty končící ve stavu *k* s pozorováním prvku *i* je dána následujícím vztahem:

$$p_l(i, x) = e_l(i) max_k(p_k(j, x-1) p_{kl}) , \qquad (4)$$

kde $e_l(i)$ je pravděpodobnost sledovaného prvku i ve stavu l, $p_k(j, x-1)$ je pravděpodobnost nejpravděpodobnější cesty končící na pozici $x-1$ ve stavu k s elementem j a p_{kl} je pravděpodobnost přechodu ze stavu l do stavu k. Pro větší efektivitu a přesnost je vhodné použití logaritmu pravděpodobnosti, který umožňuje počítat součty místo součinů. [19]

Z výše uvedené rovnice (4) Viterbiho algoritmus iteračně počítá pravděpodobnosti $p_+(i,x)$ a $p_-(i,x)$, které udávají, že nukleotid i na pozici x byl vytvořen stavem CpGIs nebo ne-CpGIs. Nejvyšší získaná pravděpodobnost pro nukleotid na aktuální pozici je pravděpodobnost nejpravděpodobnější cesty. Po proběhnutí celé sekvence výstupních symbolů pak lze zpětným sledováním získat nejpravděpodobnější posloupnost skrytých stavů. [19]

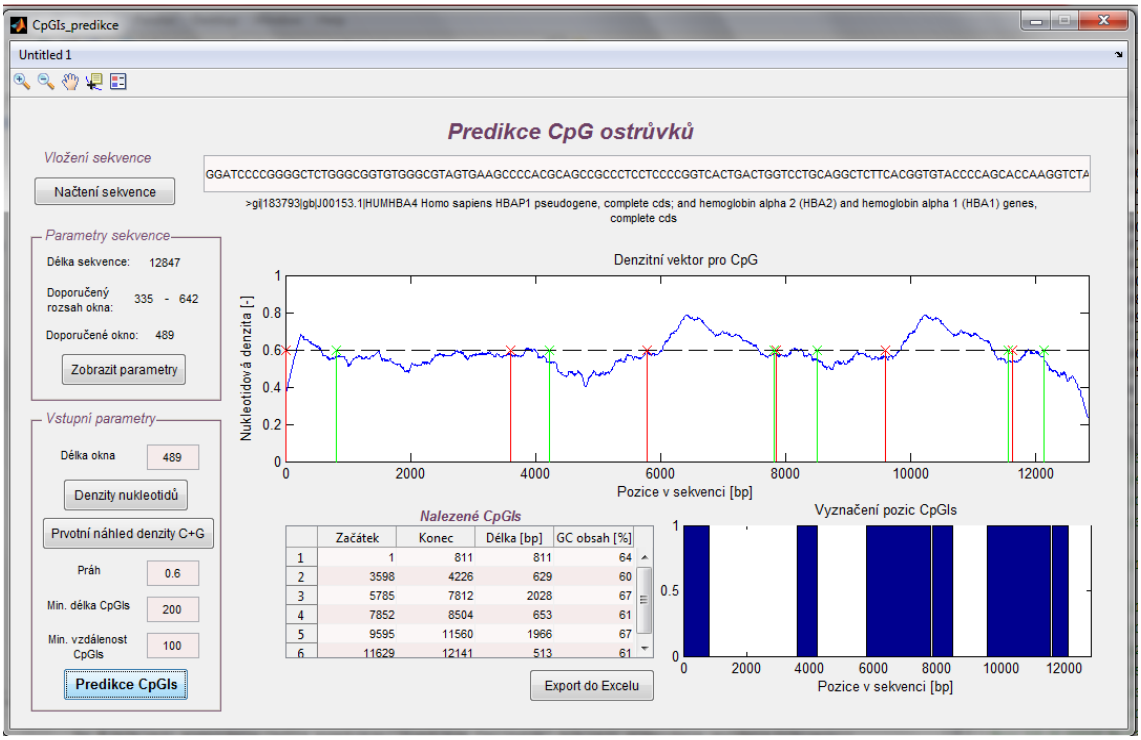


Obr. 7: Schéma Viterbiho algoritmu, kde znaménka +/- jsou stavy CpGIs/ ne-CpGIs, symbol ↑ pro vyšší hodnotu a ↓ nižší hodnotu pravděpodobnosti ze sloupce. Oranžově je vyznačena nejpravděpodobnější cesta vzniku dané sekvence.

2 PRAKTICKÁ ČÁST

2.1 Program CpGIs_predikce

Program CpGIs_predikce spolu s dalšími doprovodnými funkcemi byl navržen v programovacím prostředí MATLAB 7.11.0 R2010b. Pro snadnější práci a orientaci uživatele bylo v editoru GUIDE vytvořeno přehledné grafické rozhraní.



Obr. 8: Uživatelské rozhraní programu CpGIs_predikce.

CpGIs_predikce je algoritmus umožňující výpočet a grafickou reprezentaci denzit jak jednotlivých, tak dle biochemických vlastností uspořádaných nukleotidů. Hlavní funkcí programu je, na základě zvolených parametrů, především předpověď CpGIs, které jsou reprezentovány začátky, konci, délkami a GC obsahy. Navíc je pro lepší představu k dispozici jejich grafické znázornění. Testovaná data lze do grafického prostředí zadat ručně nebo je k dispozici načtení formátu FASTA, který obsahuje hlavičku a samotnou sekvenci. Nalezené CpGIs, popsány výše uvedenými parametry, je možné exportovat do tabulkového editoru MS Excel pro případné další zpracování. Základní principy programu jsou rozepsány v následujících kapitolách, grafické prostředí je okomentováno v manuálu umístěnému v příloze D .

2.2 Indikační a nukleotidové denzitní vektory

Pro detekci CpG ostrůvků z uměle vytvořené DNA sekvence, která je posloupností čtyř opakujících se symbolů, byla nejprve provedená vhodná transformace do numerické reprezentace. Zvolena byla 4D binární reprezentace. Základem metody je vytvoření čtyř indikačních vektorů, jeden pro každý nukleotid, $u_A(n)$, $u_T(n)$, $u_G(n)$, $u_C(n)$, které obsahují na pozici *n* hodnotu 1, pokud je daný nukleotid přítomný nebo 0 při jeho absenci a platí následující vztah:

$$u_x[k] = 1 \; jestliže \; s[k] = X \; , \tag{5}$$

kde *s[k]* pro *k = 0, 1,..., N-1* je symbolická sekvence o délce *N*, *X* je zvolený nukleotid. [20]

Zmíněné čtyři indikační vektory jsou vstupem pro výpočet vektorů nukleotidové denzity, které vyjadřují průměrný výskyt jednotlivých nukleotidů v dané části DNA sekvence. Vznik nukleotidového denzitního vektoru je realizován pomocí posuvného okna zvolené délky, které se postupně posouvá po indikačním vektoru a počítá průměr hodnot nacházejících se uvnitř okna. Z obsahu denzitních vektorů není možná zpětná rekonstrukce originální posloupnosti symbolů, následkem je ztráta informačního obsahu o přesné poloze nukleotidů v sekvenci. Denzitní vektory jsou vypočítány pro jednotlivé nukleotidy A, T, G, C podle níže uvedeného vzorce:

$$d_X[n] = \frac{\sum_{i=n-\frac{W}{2}}^{n+\frac{W}{2}} u_x[i]}{W}, \quad n = 1 \ldots N, \tag{6}$$

kde *n* je pozice v sekvenci, *X* je typ nukleotidu, *W* je velikost posuvného okna, u_x je indikační vektor zvoleného nukleotidu a *N* je délka sekvence. [21]

Realizace v programovém prostředí MATLAB je zřejmá z níže uvedeného pseudokódu funkce, využívající dříve uvedený vztah (6), pro výpočet čtyř řádkové nukleotidové denzitní matice, kdy jeden řádek odpovídá danému denzitnímu vektoru pro jeden ze čtyř nukleotidů. Vstupní proměnná *m1* je označení pro indikační matici a *w* opět symbolizuje velikost posuvného okna.

```
denzitni_matice(sekvence,m1,w)
1 for i ← 1 to 4
2    for j ← 1 to délka sekvence
3         denzitni_matice_{i,j} ← Σ_j^{j+w−1} m1_i /w;
4 return denzitni_matice
```

2.2.1 Posuvné okno

Velikost posuvného okna musí splňovat podmínku lichého čísla, důvodem je nutné umístění pozice *n* do středu posuvného okna. Nastavení velikosti okna se provádí v závislosti na požadovaném rozlišení, délce sekvence a dobrém vizuálním hodnocení. Zvolené hodnoty délky okna by měly ležet v intervalu od 5, což je nejmenší smysluplná velikost, po 1/20 délky sekvence. Aby byl potlačen vliv začátku a konce sekvence, je přidáno W/2 počet nul na obě strany indikačních vektorů. [21]

Jelikož uvedené rozmezí pro volbu velikosti posuvného okna je příliš velké a posuvné okno do značné míry ovlivňuje hledání CpGIs, byla provedena analýza vlivu velikosti okna na detekci CpGIs. Pro tento účel byla zhotovena umělá sekvence s vpravenými CpGIs na pozicích 51-100, 201-300 a 410-424. Protože reálné sekvence nemají v CpG ostrůvcích GC obsah 100 %, obsahují ostrůvky zhotovené sekvence 10 % ostatních dvou nukleotidů A a T, aby se reálným sekvencím přiblížily. Je však nutné brát v úvahu, že CpGIs nejsou přesně definovány, protože v jejich blízkosti se můžou nacházet další cytosiny a guaniny, které je mohou do jisté míry zkreslit. Z tohoto důvodu byly sledovány nejen odchylky začátků a konců od stanovených CpGIs, počet nalezených nebo nenalezených CpGIs, ale také jejich délka a GC obsah. Testováno bylo okno velikosti 5 a poté okna délky od 11-25 pro prahy 0,9/0,8 /0,7.

Tabulka 1: Skutečné pozice CpG v sekvencích, kde Z je začátek, K je konec, D je délka CpG ostrůvku.

Skutečné CpGIs		
Z	**K**	**D**
51	100	50
201	300	100
410	424	15

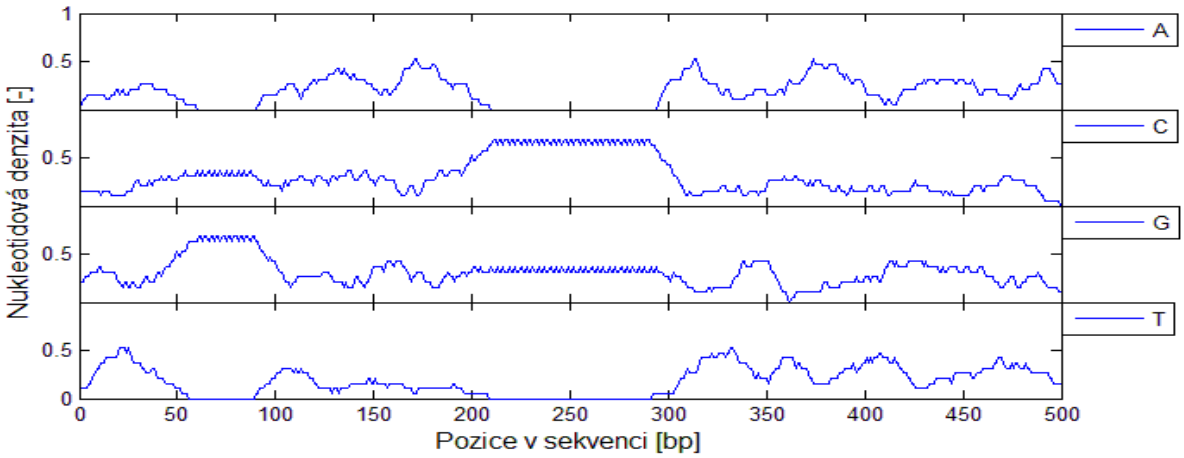
Po bližším přezkoumání níže uvedené Tabulka 2 a dalších dat, které jsou součástí přílohy B , byl odhadnut doporučený interval hodnot pro volbu délky okna od 13 do 25. Tento interval vychází z faktu, že u oken délky menších nebo rovno 11 bylo falešně detekováno větší množství CpGIs než u oken následujících. Naopak hodnoty okna větší než 25 by nesplňovaly podmínku, že maximální délka okna by neměla přesáhnout 1/20 délky sekvence. Příliš velké okno by nukleotidovou denzitu vyhladilo do takové míry, která by zamezila některé CpGIs detekovat. Konkrétní hodnota okna 19 pro výpočet nukleotidových denzit byla stanovena jako střední hodnota výše zmíněného intervalu. Pro obecnou platnost byla délka okna vztažena k délce sekvence. Doporučený rozsah je určen 2,6 %-5 % délky sekvence a z něj vyplývá střední hodnota okna 3,8 % délky sekvence.

Tabulka 2: Analýza vlivu velikosti posuvného okna na detekci CpGIs, kde *w* je velikost okna, *Z* je začátek, *K* je konec CpGIs, *GC%* je obsah G a C v CpGIs a *O* je odchylka od správných pozic. Červeně jsou znázorněny chybně detekované CpGIs.

Práh = 0,7															
w	**Z**	**O**	**K**	**O**	**D**	**O**	**GC%**	**w**	**Z**	**O**	**K**	**O**	**D**	**O**	**GC%**
5	4		12		9		67	**15**	40	11	101	-1	62	-12	80
	44		69		26		85		144		163		20		65
	69	-18	99	1	31	19	81		189	12	306	-6	118	-18	84
	104		109		6		67		341		356		16		69
	112		121		10		70		406	4	433	-9	28	-13	68
	128		133		6		67		*Počet chybně detekovaných CpGIs: 2*						
	147		162		16		69		*Počet nenalezených CpGIs: 0*						
	180		187		8		63	**17**	39	12	102	-2	64	-14	77
	190		199		10		70		144		166		23		66
	199	2	303	-3	105	-5	88		188	13	307	-7	120	-20	83
	325		332		8		63		404	6	435	-11	32	-17	66
	340		352		13		77		*Počet chybně detekovaných CpGIs: 1*						
	408	2	425	-1	18	-3	73		*Počet nenalezených CpGIs: 0*						
	427		435		9		67	**19**	39	12	102	-2	64	-14	77
	447		454		8		63		188	13	306	-6	119	-19	84
	476		484		9		67		404	6	435	-11	32	-17	66
	Počet chybně detekovaných CpGIs: 13								*Počet chybně detekovaných CpGIs: 0*						
	Počet nenalezených CpGIs: 0								*Počet nenalezených CpGIs: 0*						
11	42	9	101	-1	60	-10	80	**21**	35	16	103	-3	69	-19	76
	144		162		19		69		186	15	307	-7	122	-22	82
	190	11	305	-5	116	-16	86		403	7	438	-14	36	-21	64
	339		354		16		63		*Počet chybně detekovaných CpGIs: 0*						
	408	2	428	-4	21	-6	67		*Počet nenalezených CpGIs: 0*						
	427		440		14		65	**23**	35	16	103	-3	69	-19	76
	Počet chybně detekovaných CpGIs: 3								186	15	307	-7	122	-22	82
	Počet nenalezených CpGIs: 0								409	1	433	-9	25	-10	72
13	42	9	100	0	59	-9	82		*Počet chybně detekovaných CpGIs: 0*						
	148		162		15		74		*Počet nenalezených CpGIs: 0*						
	190	11	305	-5	116	-16	86	**25**	34	17	106	-6	73	-23	74
	339		354		16		63		182	19	309	-9	128	-28	82
	408	2	425	-1	18	-3	73		408	2	435	-11	28	-13	68
	Počet chybně detekovaných CpGIs: 2								*Počet chybně detekovaných CpGIs: 0*						
	Počet nenalezených CpGIs: 0								*Počet nenalezených CpGIs: 0*						

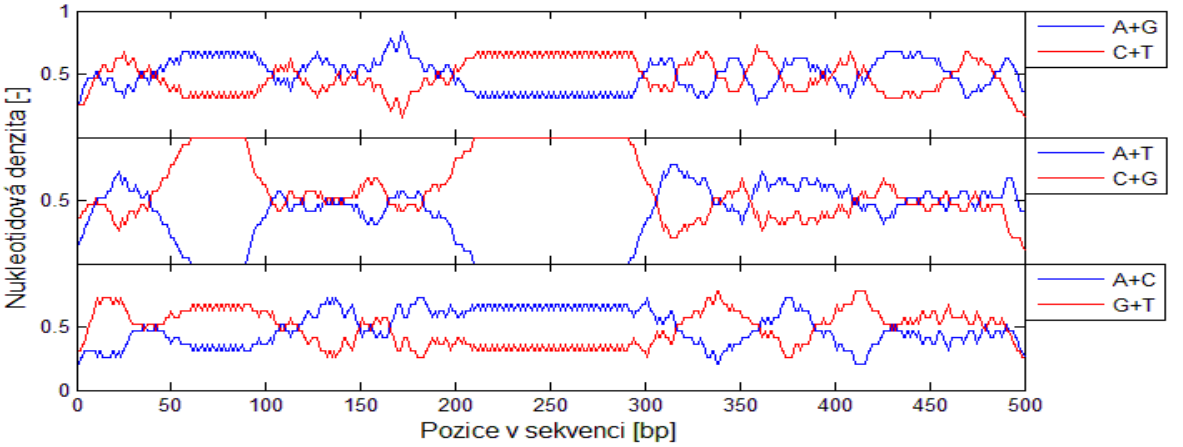
2.2.2 Grafická reprezentace nukleotidových denzit

Nejzákladnějším grafickým zobrazením nukleotidových denzit je jejich samostatné vykreslení pro jednotlivé nukleotidy. Na Obr. 9 jsou vyobrazeny denzity všech čtyř nukleotidů z uměle vytvořené sekvence dlouhé 500 bp s přesně definovanými CpG ostrůvky na pozicích 51-98 a 201-300, kdy délka okna byla analýzou zvolena na hodnotu 19, jako kompromis mezi rozlišením a dobrou vizualitou. Délka okna 19 byla použita pro všechny následující zobrazení. Ačkoliv je konstrukce jednoduchá, neposkytuje dostatečnou vizuální informaci.



Obr. 9: Denzitní vektory jednotlivých nukleotidů z umělé vytvořené sekvence, W = 19.

Druhou možností grafické reprezentace je zobrazení součtu nukleotidových denzitních vektorů na základě jejich biochemických vlastností, kterými jsou molekulární struktura, síla vazby mezi komplementárními bázemi a obsah radikálů. Výsledný obraz je rozdělen do tří bloků, kdy první vykresluje sumy denzit A, G oproti C, T, druhý A, T a C, G a poslední blok zobrazuje nukleotidy A, C a G, T. Biochemické vlastnosti nukleotidů jsou podrobněji rozepsány v kapitole 1.1.1. Je patrné, že vizuální informace je již hodnotnější než v předešlém případě. Zejména v druhém oddílu na Obr. 10 jsou již na součtu denzit C a G dobře rozpoznatelné uměle vytvořené CpG ostrůvky. Právě tato suma nukleotidových denzitních vektorů cytosinu a guaninu bude použita v další analýze.



Obr. 10: Součty denzitních vektorů jednotlivých nukleotidů umělé sekvence dle biochemických vlastností, W = 19.

2.3 Realizace detekce CpG ostrůvku

Samotná detekce CpG ostrůvku je prováděna na výsledné sumě nukleotidových denzitních vektorů cytosinu d_C a guaninu d_G, která je vypočtena na základě následujícího vzorce:

$$S_{CG}[n] = d_C[n] + d_G[n],\ n = 1 \dots N. \tag{7}$$

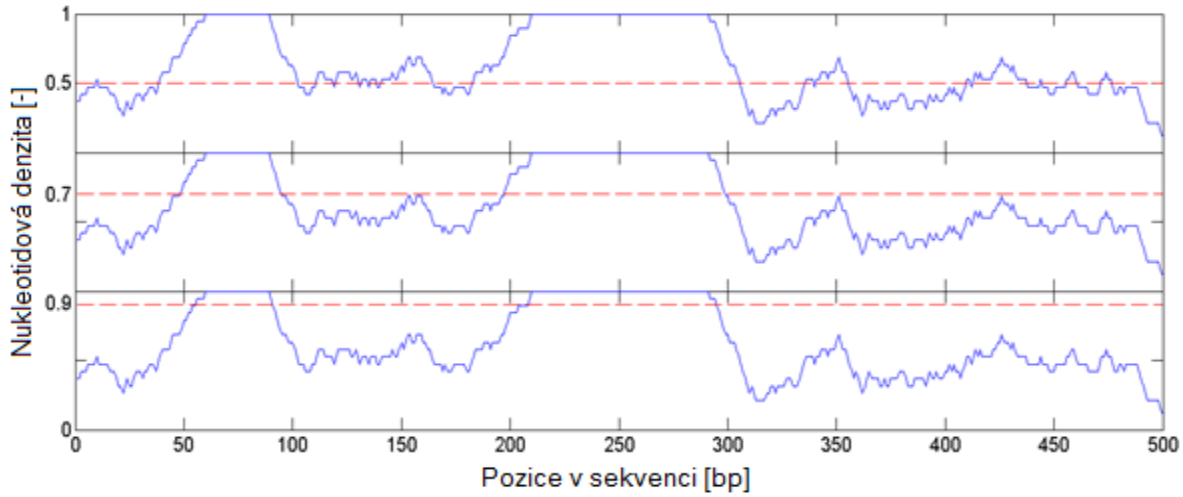
Určení zda se na daném místě vyskytuje významný shluk CpG dinukleotidů je realizováno na základě faktu, zda průměrný výskyt C a G v dané části sekvence překročí předem stanovený práh. Zvolena velikost prahu významně ovlivňuje dosažené výsledky. Pro přesnější stanovení začátku a konce CpG ostrůvku, se při jejich indexaci na začátku odečítá a na konci přičítá hodnota *W/2*.

Provedení v MATLABU je vysvětleno na níže uvedeném pseudokódu. Hodnota začátku *z* CpGIs je do matice vektoru uložena, pokud zkoumaná hodnota denzity je menší než stanovený práh a zároveň hodnota následující je větší než práh. Konec *k* CpGIs je získán obdobným způsobem s tím rozdílem, že první hodnota musí být větší a příští hodnota menší než práh. Vstup *cpg* je vektor získaný sumou nukleotidových denzitních vektorů cytosinu a guaninu.

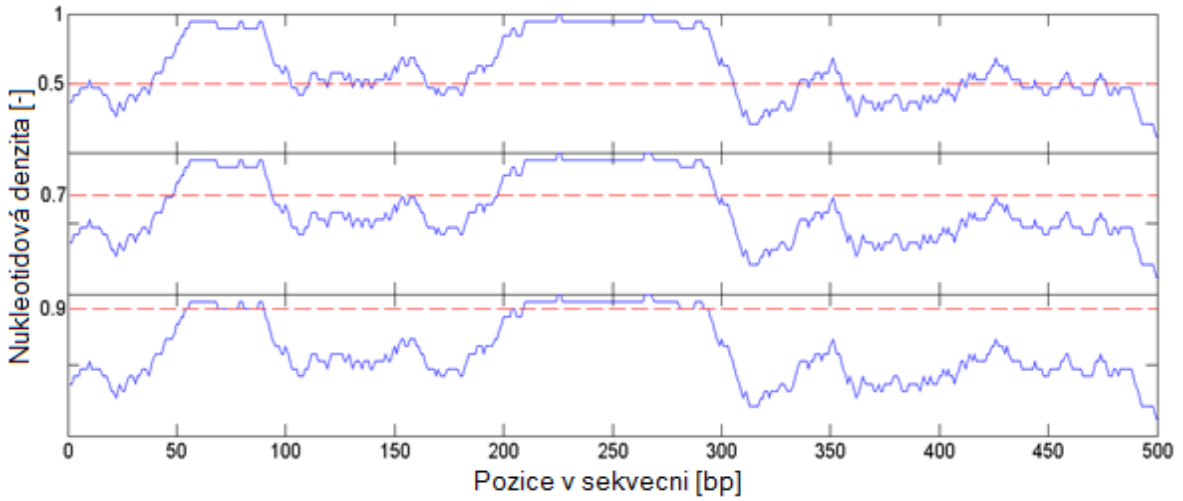
```
Detekce_z_k(sekvence,cpg,prah)
1 for i ← 1 to délka sekvence
2    if cpg_{i+1} ≥ práh and cpg_i<práh
3         z_i ←i
4         return z_i
5    elseif cpg_{i+1}<práh and cpg_i ≥ práh
6         k_i ←i
7         return k_i
8 return začátky a konce CpGIs
```

2.3.1 Testování funkce na umělých sekvencích

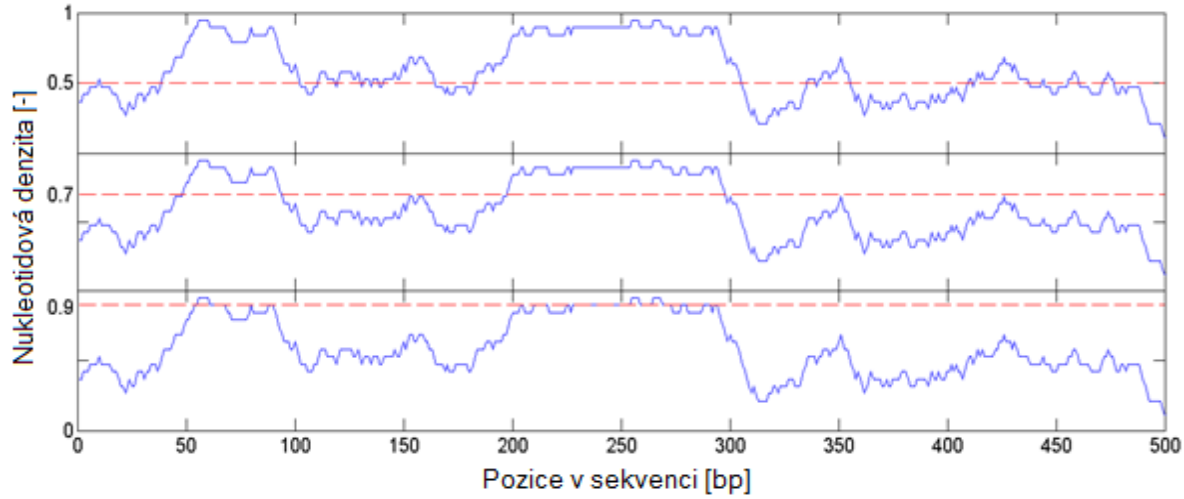
Výstupem celé detekce je jak grafické zobrazení denzitního vektoru S_{GC} s vyznačeným prahem, tak matice obsahující v prvním sloupci začátky CpG ostrůvku, v druhém sloupci jejich konce a v posledním, třetím, sloupci jejich délku. Pro vytvoření, vyzkoušení a vyladění programu, byly použity čtyři uměle vytvořené DNA sekvence o délce 500 bp se dvěma přesně definovanými CpG ostrůvky. K dispozici je informace o poloze, délce a obsahu ostrůvku. Všechny čtyři sekvence mají CpG ostrůvky délky 48 bp a 100 bp přítomny na pozicích 51 až 98 a 201 až 300. Liší se však obsahem ostatních dvou nukleotidů v ostrůvcích, kdy první sekvence obsahuje pouze C a G, avšak následující tří mají již zastoupení A, T 5 %, 10 % a 15 %.



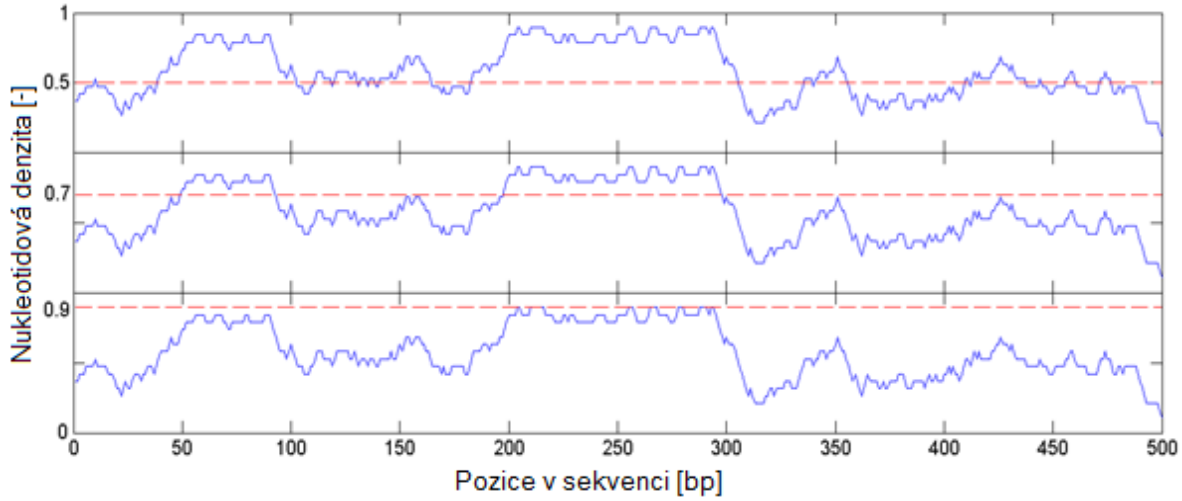
Obr. 11: Jednotlivé prahy pro sekvenci obsahující CpGIs s 0 % obsahu A, T. W = 19.



Obr. 12: Jednotlivé prahy pro sekvenci obsahující CpGIs s 5 % obsahu A, T. W = 19.



Obr. 13: Jednotlivé prahy pro sekvenci obsahující CpGIs s 10 % obsahu A, T. W = 19.



Obr. 14: Jednotlivé prahy pro sekvenci obsahující CpGIs s 15 % obsahu A, T. W = 19.

Vliv zvoleného prahu a délky posuvného okna na výslednou detekci je zobrazen na předešlých obrázcích (Obr. 11, Obr. 12, Obr. 13, Obr. 14). Při příliš nízké hodnotě prahu jsou jako CpG ostrůvky vyhodnoceny i oblasti, které nemají dostatečný GC obsah. Naopak při nadměrně velkém prahu nejsou přítomné CpG ostrůvky zaznamenány vůbec nebo jsou zobrazeny kratší, než jsou ve skutečnosti.

Tabulka 3: Skutečné pozice CpG v sekvencích, kde Z je začátek, K je konec, D je délka CpG ostrůvku.

Skutečné CpG		
Z	**K**	**D**
51	98	48
201	300	100

Výše zmíněný vliv volby prahu je potvrzen i v Tabulka 4, kdy pro práh 0,5 bylo jako CpG ostrůvek vyhodnoceno nadměrné množství dat u všech zkoumaných sekvencí. Z celkového pohledu, který zahrnuje všechny připravené sekvence, bylo nejlepší detekce dosaženo s prahem 0,7, kdy v sekvencích byly správně klasifikovány pouze dva CpG ostrůvky, i když hodnoty začátků a konců jsou poměrně zkresleny. Práh 0,9 dosáhl nejpřesnějšího stanovení všech parametrů, avšak pouze u sekvence, v které byl 0% výskyt A a T v CpG ostrůvcích. U následných dvou sekvencí bylo opět vyhodnoceno větší množství ostrůvku než ve skutečnosti. Naopak v poslední sekvenci detekce selhala a nebyl prokázán žádný ostrůvek, protože průměrný výskyt C a G v celé části sekvence nepřekročil takto stanovený práh. Volba prahu je ponechána na uživateli a měla by především vycházet z pohledu na grafické znázornění nukleotidové denzity GC.

Tabulka 4: Zobrazení obsahu jednotlivých výstupních matic pro zvolené sekvence a prahy, kde Z je začátek, K je konec, D je délka CpG ostrůvku.

	V CpG 0 % A, T			V CpG 5 % A, T			V CpG 10 % A, T			V CpG 15 % A, T		
Práh	**Z**	**K**	**D**	**Z**	**K**	**D**	**Z**	**K**	**D**	**Z**	**K**	**D**
0,5	0	19	19	0	19	19	0	19	19	0	19	19
	29	111	82	29	111	82	29	111	82	29	111	82
	101	127	26	101	127	26	101	127	26	101	127	26
	110	139	29	110	139	29	110	139	29	110	139	29
	122	143	21	122	143	21	122	143	21	122	143	21
	126	147	21	126	147	21	126	147	21	126	147	21
	131	173	42	131	173	42	131	173	42	131	173	42
	173	314	141	173	314	141	173	314	141	173	314	141
	326	348	22	326	348	22	326	348	22	326	348	22
	331	364	33	331	364	33	331	364	33	331	364	33
	400	420	20	400	420	20	400	420	20	400	420	20
	403	446	43	403	446	43	403	446	43	403	446	43
	434	453	19	434	453	19	434	453	19	434	453	19
	446	469	23	446	469	23	446	469	23	446	469	23
	462	485	23	462	485	23	462	485	23	462	485	23
0,7	39	103	64	39	102	63	39	102	63	40	101	61
	188	307	119	188	306	118	188	306	118	188	306	118
0,9	46	99	53	46	77	31	46	69	23	Nedetekováno		
	199	303	104	69	89	20	245	267	22			
				78	98	20	255	277	22			
				200	289	89						
				278	300	22						

2.3.2 Vylepšení detekce CpGIs

Po provedení testování programu na uměle vytvořených sekvencích byly sledovány určité nedostatky, které vedly k nepříliš přesnému odhadu CpGIs. Na základě tohoto zjištění byl program rozšířen o funkce, které realizují sčítání ostrůvku, omezují jejich zobrazení jen nad zvolené délky, vypočítávají GC obsah v nich a zprostředkovávají lepší vizualizaci získaných výsledků.

Na analýze uměle vytvořených sekvencích byla vypozorována detekce krátkých CpG ostrůvků s malou vzdáleností mezi sebou. Z tohoto důvodu byl program obohacen funkcí pro sčítání detekovaných CpGIs v takových případech. Základní princip spočívá v počítání vzdáleností v matici *CpG*, která obsahuje parametry původních ostrůvků, mezi koncem aktuálního ostrůvku a začátkem ostrůvku následujícího. Pokud je vzdálenost menší nebo rovna zadané minimální vzdálenosti *min_vzd* jsou ostrůvky sečteny. Konec prvního ostrůvku

je nahrazen koncem ostrůvku následujícího. Výsledkem je detekování menšího množství delších CpGIs.

```
ScitaniCpGIs(CpG, min_vzd)
1  CpGIs ← matice 0 o velikosti CpG     // pro ukládání
   upravených pozic CpGIs
2  a ←1         // řádek nové matice, uloží se upravený CpGIs
3  i ←0
4  b ←0         // určuje, o který ostrůvek se jedná
5  rozmer ← počet řádků CpG
6  while b<rozmer-1
7    c ←1+i          // určuje začátek sečteného ostrůvku
8    i ←i+1
9    if CpG_{i+1,1} − CpG_{i,2} ≤ min_vzd
10        while CpG_{i+1,1} − CpG_{i,2} ≤ min_vzd
11             CpGIs_{a,2} ← CpG_{i+1,2}
12             i ←i+1
13             b ←i+1
14        a ←a+1
15   else
16        CpGIs_{a,1} ←CpG_{c,1}
17        CpGIs_{a,2} ←CpG_{i,2}
18        b←b+1
19        a←a+1
20 return na základě sčítání upravené pozice CpGIs
```

I po sečtení relativních CpGIs se mohou vyskytnout osamocené krátké útvary, které jsou při překročení prahu označeny jako CpG ostrůvky. Protože není známa nejnižší délka, které mohou CpGIs dosahovat, byla uživateli ponechána možnost nastavení minimální délky predikovaných ostrůvků. V následujícím pseudokódu je patrné, že pokud bude ostrůvek kratší než předem zvolená minimální délka *min_delka* bude smazán.

```
Vymazani_kratkych_CpGIs(min_delka,CpGIs)
1 rozmer ← počet detekovaných CpGIs
2 do ← délka ostrůvku
3 for i ← 1 to rozmer
4    if do_i ≤ min_delka
5         CpGIs_i smazán
6 return pouze CpGIs s dostatečnou délkou
```

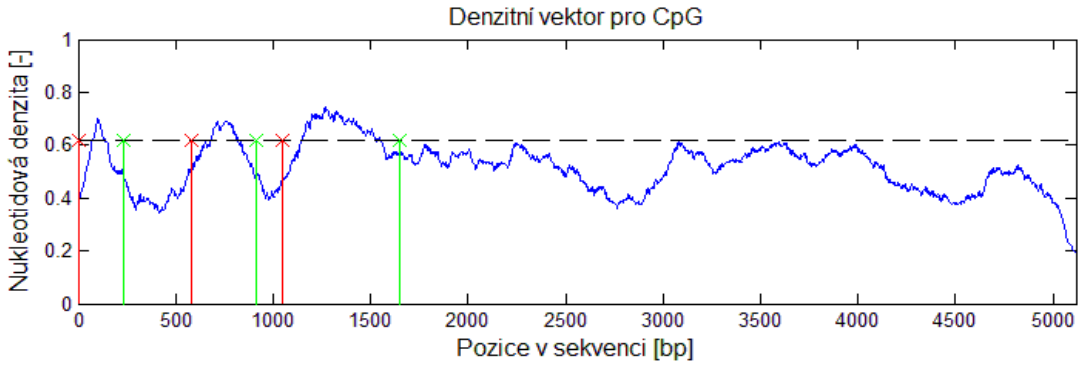
Existuje další parametr, který dokáže vystihnout významnou vlastnost nalezených CpGIs. Touto veličinou je GC obsah, který podává informaci o procentuálním zastoupení guaninu a cytosinu v ostrůvku a lze jej vypočítat podle následujícího vztahu:

$$GC\ \% = \frac{Počet\ G + Počet\ C}{Počet\ N} \cdot 100, \tag{8}$$

kde ve jmenovateli se objevuje součet počtu guaninu a cytosinu nacházejících se ve zvoleném ostrůvku, čitatel je zastoupen počtem všech nukleotidů v daném CpGIs. GC obsah lze spočítat i pro celou sekvenci, kdy jsou do vzorce dosazeny hodnoty vztažené k celé sekvenci. Níže uvedený pseudokód znázorňuje funkci pro potřebný výpočet. Z proměnné *matice* jsou vybrány řádky popisující indikační vektory C a G a jsou v nich sečteny hodnoty na pozicích od začátku do konce daného CpGIs. Potřebné začátky a konce ostrůvků jsou uloženy v matici *CpGIs*. Konečná úprava je provedena vynásobením stem a vydělením délkou ostrůvku *do*.

```
GC_obsah(CpGIs, matice)
7 rozmer ← počet detekovaných CpGIs
8 do ← délka ostrůvku
9 for i ← 1 to rozmer
10        GC_obsah_i ← (∑_{CpGIs_{i,1}}^{CpGIs_{i,2}} matice_G + ∑_{CpGIs_{i,1}}^{CpGIs_{i,2}} matice_C) · 100/do
11 return GC_obsah
```

Pouze pro informativní účel a lepší vizuální sdělení, byly do grafu vykreslení denzity nukleotidů C a G zakomponovány vertikální čáry, kdy červené označující začátek a zelené konec stanovených CpGIs. Ze stejného důvodu byl ještě zhotoven graf plošně znázorňující veškeré nalezené CpGIs v zadané DNA sekvenci. Obě možnosti jsou na následujících obrázcích.



Obr. 15: Grafické zobrazení nukleotidové denzity C a G u genu HUMTBBM40.

Obr. 16: Zobrazení nalezených CpGIs u genu HUMTBBM40.

2.4 Detekce CpGIs v reálných sekvencích

Pro ověření funkčnosti programu bylo z veřejně dostupné databáze NCBI vybráno dvacet odlišně dlouhých sekvencí, pocházejících od různých organismů. Jelikož je nutné, aby zvolené sekvence obsahovaly CpG ostrůvky, byl jejich výběr inspirován článkem CpG Island in Vertebrae Genomes [12]. Veškeré testované sekvence, včetně základních informací, jsou přehledně vylíčeny v následující Tabulka 5, a zároveň jsou ve FASTA formátu součástí přiloženého CD.

Tabulka 5: Testované sekvence obsahující CpGIs. Poslední dva sloupce zobrazují délku a GC obsah celé sekvence. [22]

Kódovaný protein	Gen Bank kód	Organismus	D [bp]	GC %
c-Ha-ras1	HUMRASH	Homo sapiens	6453	68
α1 Globin	GOTHBAI	Capra hircus	1894	61
α2 Globin	GOTHBAII	Capra hircus	1691	64
α2 Hemoglobin	HUMHBA4	Homo sapiens	12847	59
Metallothionein-II	HUMMET2	Homo sapiens	1703	56
Metallothionein-I	MUSMETI	Mus musculus	1561	54
Metallothionein-II	MUSMETII	Mus musculus	1400	57
Oxytocin/neurophysin	RATOXTNP	Rattus norvegicus	1053	61
Ribosomal protein L32-3A	MUSRUPL3A	Mus musculus	5380	50
Somatostatin I	HUMSOMI	Homo sapiens	2667	45
β Tubulin	HUMTBBM40	Homo sapiens	5117	52
Cytochrome c, allele CC10	CHKCYC10	Gallus gallus	1620	44
ζ Hemoglobin	HUMHBA1	Homo sapiens	2685	65
MHC, class II HLA-DC-β	HUMMHDCB	Homo sapiens	7272	47
MHC, class II HLA-DC-3β	HUMMHDC3B	Homo sapiens	8090	47
Histone H1	CHKH11A1	Gallus gallus	1098	61
Histone H2A/H2B	CHKH2A2B	Gallus gallus	1564	57
N-myc proto-oncogene protein	MYCN	Homo sapiens	6447	59
Ras-related protein Rab-42 isoform 1	RAB42	Homo sapiens	2387	60
Bromodomain adjacent to zinc finger domain protein 1A	BAZ1A	Mus musculus	6179	46

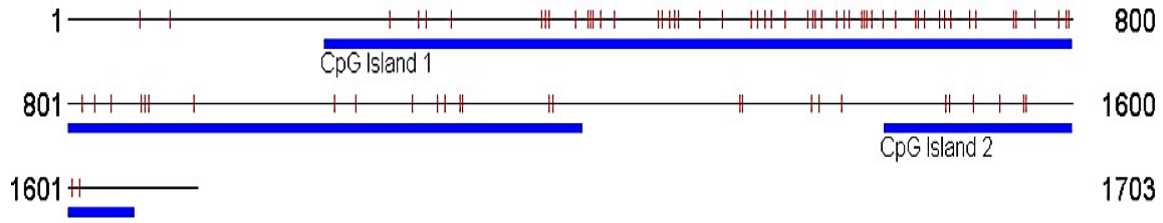
2.4.1 Nástroje pro predikci CpGIs

Jelikož nejsou k dispozici informace o přesných polohách CpGIs v sekvencích DNA, byly pro interpretaci a srovnání výsledků získaných z programu CpGIs_predikce (CpGIsP), vybrány dva volně dostupné internetové vyhledávače CpGIs. Oba vyhledávače CpG Island Searcher (CpGIsS) [23] a DBCAT [24], byly zvoleny, na základě uspořádaného uživatelského prostředí, snadné obsluhy a přehledných výstupů ve formě pozic, délek a grafického znázornění CpGIs v sekvenci. Výhodou je také možnost volby parametrů nutných pro realizaci detekce CpGIs, které byly pro následné porovnání nastaveny na stejné hodnoty.

Parametry byly zvoleny podle Gardiner-Gardenových a Frommerových kritérií na hodnoty CG obsah ≥ 50 %, délka CpG ostrůvku ≥ 200 bp a Obs_{CpG}/Exp_{CpG} hodnota > 0,65. U programu CpG Island Searcher byl navíc k dispozici parametr určující minimální vzdálenost mezi ostrůvky, který byl nastaven na minimální možnou hodnotu 100 bp. Naopak nevýhodou je pouze ruční vkládání sekvencí, protože není k dispozici přímé načítání FASTA formátu.

Tabulka 6: Vstupní parametry internetových vyhledávačů CpGIs.

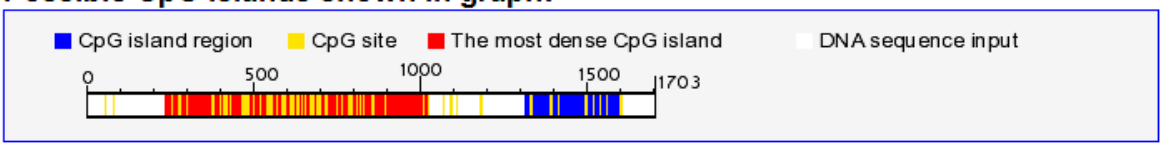
	DBCAT	CpGIsS
GC obsah [%]	≥ 50	≥ 50
Délka CpGIs [bp]	≥ 200	≥ 200
Obs_{CpG}/Exp_{CpG}	> 0,65	> 0,65
Min. vzdálenost [bp]		100



Select lower limits: %GC=50, ObsCpG/ExpCpG=0.60, Length=200, Distance=100
CpG island 1 start=204, end=1210, %GC=59.1, ObsCpG/ExpCpG=0.782, Length=1007
CpG island 2 start=1450, end=1653, %GC=50.5, ObsCpG/ExpCpG=0.615, Length=204

Obr. 17: Výstup vyhledávače CpG Island Searcher. [23]

Possible CpG islands shown in graph:



Sequence of each CpG island:

1. start site:236 end site:1021 length:786 GC content:61%

TTCCTACAGTCCCTGTTACACGCTAAAAGTACTCAACTAGCTTCGGATACGTCATCAGCAACCACCCCACGGGTTACTGTGATGCTGC
ACAATTATTAAGCCCTGGCTGCTACAGAGTTGTAACCTGTCTGCACTTCCAACCGGCGCCGCAAGCAGCATTCCCAGTCCCGCTTTCA
CCCGCGCGCTAACGGCTCAGGTTCGAGTACAGGACAGGAGGGAGGGGAGCTGTGCACACGGCGGAGGCGCACGGCGTGGGCACCCAGC
ACCCGGTACACTGTGTCCTCCCGCTGCACCCAGCCCCTTCAGCCCGAGGCGTCCCCGAGGCGCAAGTGGGCCGCCTTCAGGGAACTGA
CCGCCCGCGGCCCGTGTGCAGAGCCGGGTGCGCCCGGCCCAGTGCGCGCGGCCGGGTGTTTCGCTTGGAGCCGCAAGTGACTTCTAGC
GCGGGGCGTGTGCAGGCACGGCCGGGGCGGGGCTTTTGCACTCGTCCCGGCTCTTTCTAGCTATAAACACTGCTTGCCGCGCTGCACT
CCACCACGCCTCCTCCAAGTCCCAGCGAACCCGCGTGCAACCTGTCCCGACTCTAGCCGCCTCTTCAGCACGCCATGGATCCCAACTG
CTCCTGCGCCGCCGGTAAGAGGCTGGGGATGCCCAGTGTAGACTGTAGCGCTAGAGAAGCAATTTCTGACCCCTCTTTCTTTCTCTGG
TCACTCAATTTCAGGCACAGGAGTTGCTCTTCCCAAAGAGTTTTGGTATCTTTCTCTCCATTCTAGGTTATTCGGAGCCCCC

2. start site:1315 end site:1598 length:284 GC content:58%

AGTCTGGGGATGCCCCATTGCGCGGAAATTGTTGCCTCCTCAGTGATCTTATCAGGGAGAGCAGGAATCCTTATTCCCGGTGTCGCTA
GTACTCATCTCTGCGCCTCCTGTCTGCCCCCAGGCTGCTGCTCCTGCTGCCCTGTGGGCTGTGCCAAGTGTGCCCAGGGCTGCATCTG
CAAAGGGGCGTCGGACAAGTGCAGCTGCTGCGCCTGATGCTGGGACAGCCCCGCTCCCAGATGTAAAGAACGCGACTTCCACAAACCT
GGATTTTTTATGTACAACCC

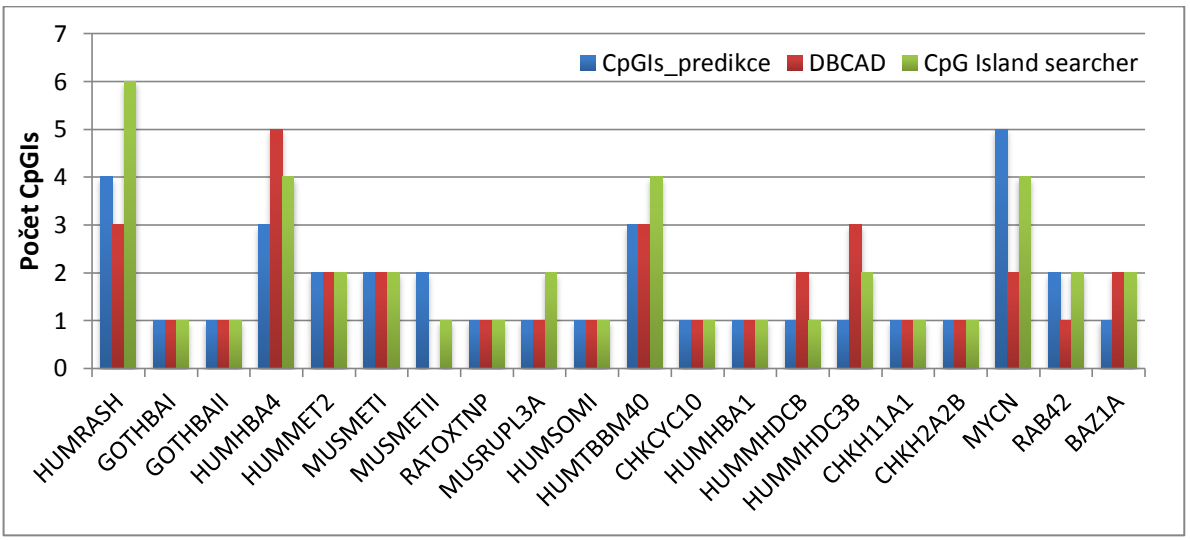
Obr. 18: Výstup vyhledávače DBCAT. [24]

2.4.2 Testování funkce na reálných sekvencích a srovnání s vyhledávači

Obsahem závěrečné analýzy bylo testování výše zmíněných dvaceti sekvencí programem CpGIsP a volně dostupnými nástroji CpGIsS a DBCAT. U CpGIsP byla používána automaticky počítaná délka okna, zjištěná na základě provedené analýzy, práh byl volen podle konkrétního vzhledu denzitního vektoru C a G, minimální délka ostrůvku byla nastavena na 200 bp a jejich minimální vzdálenost na 100 bp. Parametry vyhledávačů byly zvoleny podle Gardiner-Gardenových a Frommerových kritériích uvedených v kapitole 1.4.1. Zjišťovány byly jak pozice začátku a konců CpGIs, tak jejich délka a GC obsah. Sledovány byly rozdíly v samotném počtu nalezených ostrůvků, ale i odchylky od výsledných pozic určených jednotlivými programy. Následující Tabulka 7 spolu se sloupcovým Graf 1, ukazuje rozdíl mezi jednotlivými programy ve stanovení počtu CpGIs v testovaných sekvencích.

Tabulka 7: Srovnání jednotlivých programů z hlediska počtu nalezených CpGIs.

	CpGIsP	DBCAT	CpGIsS		CpGIsP	DBCAT	CpGIsS
Gen Bank kód	**Počet CpGIs**			**Gen Bank kód**	**Počet CpGIs**		
HUMRASH	4	3	6	HUMTBBM40	3	3	4
GOTHBAI	1	1	1	CHKCYC10	1	1	1
GOTHBAII	1	1	1	HUMHBA1	1	1	1
HUMHBA4	3	5	4	HUMMHDCB	1	2	1
HUMMET2	2	2	2	HUMMHDC3B	1	3	2
MUSMETI	2	2	2	CHKH11A1	1	1	1
MUSMETII	2	0	1	CHKH2A2B	1	1	1
RATOXTNP	1	1	1	MYCN	5	2	4
MUSRUPL3A	1	1	2	RAB42	2	1	2
HUMSOMI	1	1	1	BAZ1A	1	2	2



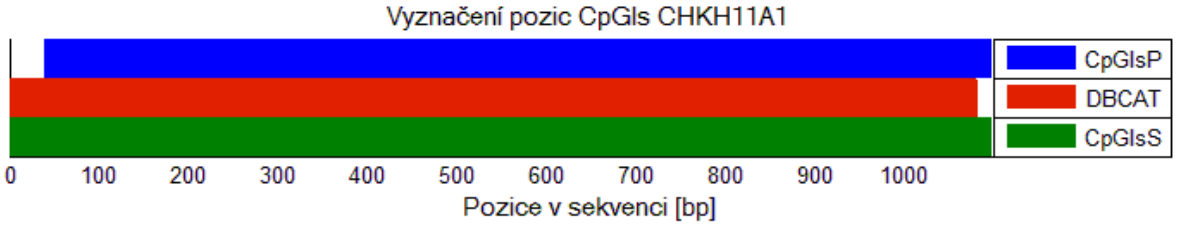
Graf 1: Srovnání jednotlivých programů z hlediska počtu nalezených CpGIs.

Jak je patrné z tabulky nebo grafu, byly rozlišeny dvě skupiny dat. První skupinu tvoří sekvence, u kterých byl počet CpGIs shodně stanoven všemi programy. Ve druhé jsou již vidět nepatrné i značné rozdíly v určeném množství. Právě na tomto základě byly vytvořeny a následně popsány dvě kategorie testovaných sekvencí.

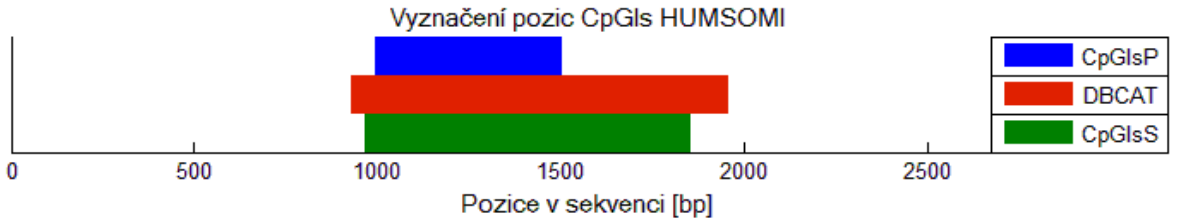
Tabulka 8: Sekvence, u kterých byl u všech tří programů stanoven stejný počet CpGIs.

CpGIs_predikce						DBCAT				CpG Island Searcher			
w	p	Z	K	D	GC%	Z	K	D	GC%	Z	K	D	GC%
GOTHBAI													
71	0,6	694	1813	1120	67	640	1733	1094	65	663	1610	948	67,4
GOTHBAII													
63	0,6	124	1668	1545	66	18	1573	1556	64	21	1553	1533	64,3
RATOXTNP													
39	0,65	241	950	710	68	184	980	797	65	197	1044	848	62,7
HUMSOMI													
101	0,55	992	1503	512	61	931	1957	1027	52	964	1855	892	50,1
CHKCYC10													
61	0,5	1	340	340	71	1	457	457	52	1	431	431	64,5
HUMHBA1													
101	0,7	1374	2581	1208	76	1355	2535	1181	75	1515	2672	1158	72,5
CHKH11A1													
41	0,5	39	1098	1060	63	1	1082	1082	68	1	1098	1098	61,1
CHKH2A2B													
59	0,5	127	1461	1335	62	4	1465	1462	58	99	1510	1412	59,1
HUMMET2													
65	0,65	361	887	527	69	236	1021	786	61	1	204	1007	59,1
		1388	1554	167	66	1315	1598	284	58	1450	1653	204	50,5
MUSMETI													
59	0,6	64	493	430	61	1	703	703	50	1	678	678	56
		762	1019	258	60	1017	1478	462	51	1136	1414	279	53,4

První skupina vytvořena na základě shodného stanovení počtu CpGIs je z většiny tvořena sekvencemi u kterých byl nalezen pouze jeden ostrůvek. Pouze dvě sekvence obsahují ostrůvky dva. Bylo provedeno srovnání detekce CpGIs mezi vytvořeným programem a internetovými vyhledávači. Nejshodnější detekce ve vypsaných pozicích bylo dosaženo u genu CHKH11A1, kdy programy CpGIsP a CpGIsS stanovily začátek s rozdílem 38 bp a konec určily zcela shodně. Naopak největší nepřesnosti dosáhl realizovaný program vzhledem k vyhledávači DBCAT u genu HUMSOMI. Zde se lišil začátek CpGIs o 61 bp a konec o 454 bp. Průměrně se pak odchylky detekovaných pozic pohybovaly kolem hodnoty 88,59 bp při srovnání s programem DBCAT a 90,85 bp u CpGIsS.

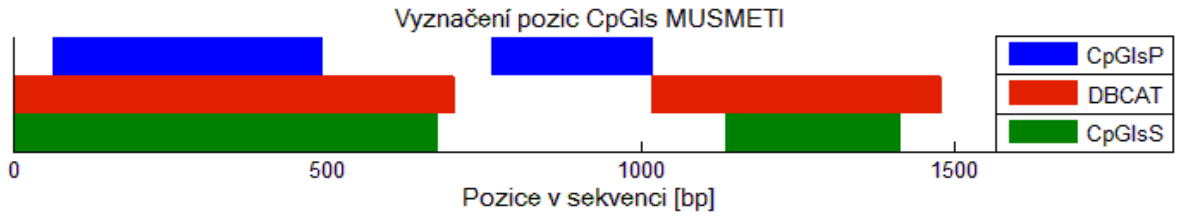


Obr. 19: Srovnání programů v detekci CpGIs u genu CHKH11A, nejshodnější detekce.

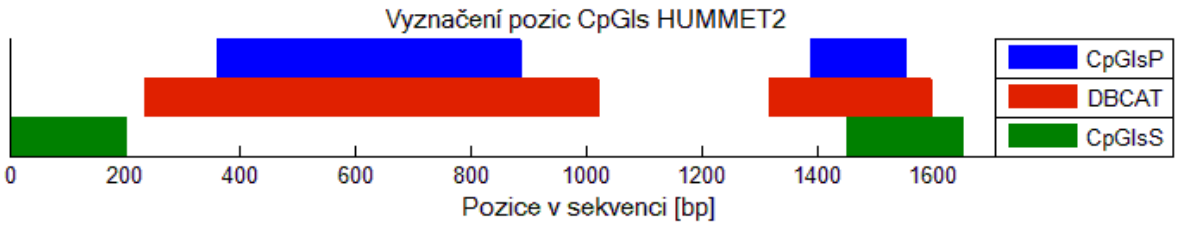


Obr. 20: Srovnání programů v detekci CpGIs u genu HUMSOMI, nejvíce odlišná detekce.

Do výše zmíněných průměrných odchylek však nejsou započítány dva případy, kdy došlo k detekci úplně odlišného ostrůvku, než který byl stanoven ostatními dvěma vyhledávači. Tento případ nastal u sekvence MUSMETI, kdy CpGIsP označil jako CpGIs úsek s 61 % GC, který nebyl stanoven ani jedním s ostatních programů, které naopak jako ostrůvek vyhodnotily jinou část DNA sekvence s GC obsahem pouze kolem 50 %. Naopak u genu HUMMET2 byl programy CpGIsP a DBCAT první ostrůvek stanoven přibližně stejně, ale CpGIsS se v predikci podstatně lišil. Souhrnně řečeno, byly zhotoveným programem ve většině případu nalezeny ostrůvky sice kratší, ale ve všech případech s vyšším obsahem GC než u zbylých dvou programů. Na vložených obrázcích lze sledovat odchylky detekce všech tří programů u výše zmíněných genů, zobrazení ostatních případů je součástí přílohy C .



Obr. 21: Srovnání programů v detekci CpGIs u genu MUSMETI. CpGIsP nalezl zcela odlišný CpGIs.

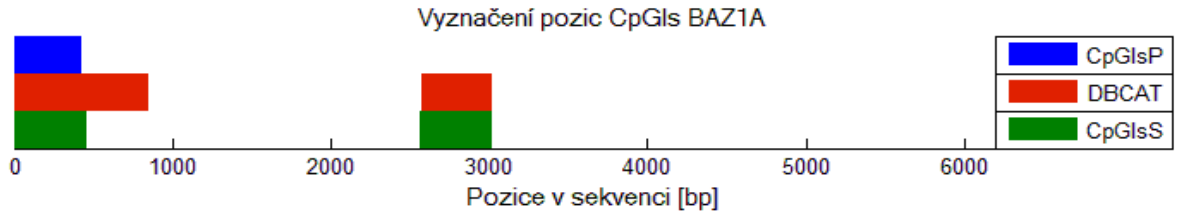


Obr. 22: Srovnání programů v detekci CpGIs u genu HUMMET2. CpGIs nalezený CpGIsS se zcela vymyká ostatním dvěma detekcím.

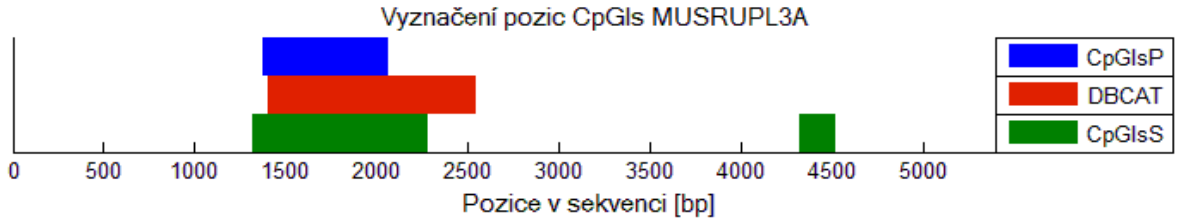
Tabulka 9: Sekvence, u kterých nebyl u všech tří programů stanoven stejný počet CpGIs.

CpGIs_predikce						DBCAT				CpG Island Searcher			
w	**p**	**Z**	**K**	**D**	**GC%**	**Z**	**K**	**D**	**GC%**	**Z**	**K**	**D**	**GC%**
HUMRASH													
245	0,7	1	1345	1345	78	3	3903	3901	69	1	1438	1438	76,3
		1401	1736	336	69	4065	4394	330	63	1608	1847	240	65
		2457	2971	515	70	4705	5760	1056	66	2242	2555	314	66,9
		3345	3596	252	70					3237	3443	207	67,1
										4768	5125	358	66,8
										5303	5504	202	68,3
HUMHBA4													
489	0,65	1	614	614	66	1	868	868	53	1	855	855	61,8
		5846	7618	1773	69	5021	5664	644	54	5014	5550	537	52
		9657	11452	1796	69	5847	7530	1684	69	5981	7505	1525	70
						8391	8675	285	51	9788	11200	1413	70,5
						9658	11487	1830	68				
MUSMETII													
53	0,6	1	524	524	65	0	0	0	0	1	1020	1020	59,7
		786	1060	275	63								
MUSRUPL3A													
205	0,6	1377	2063	687	64	1397	2545	1149	57	1312	2277	966	58,7
										4314	4515	202	50
HUMTBBM40													
195	0,65	1	217	217	65	1	330	330	57	1	328	328	57,9
		595	900	306	64	492	2465	1974	57	466	1862	1397	58,4
		1066	1607	542	67	3171	4096	926	55	1863	2200	338	54,1
										3497	4115	619	53,2
HUMMHDCB													
277	0,6	1575	2742	1168	65	394	614	221	58	1617	2795	1179	63,2
						1475	2803	1329	61				
HUMMHDC3B													
307	0,6	1861	2991	1131	63	586	895	310	55	1933	2913	981	63,4
						1900	3038	1139	61	3364	3770	407	50,4
						5444	5664	221	57				
MYCN													
245	0,6	1	459	459	65	958	2509	1552	68	1	1075	1075	57,5
		335	869	535	61	4953	5534	582	56	1231	2427	1197	70,8
		920	2585	1666	68					2758	2993	236	53,8
		2848	3136	289	58					4908	5549	642	55
		4967	5385	419	60								
RAB42													
91	0,65	127	760	634	74	8	1365	1358	64	17	761	745	70,3
		1289	1499	211	64					981	1393	413	56,9
BAZ1A													
235	0,55	1	426	426	67	1	854	854	50	1	457	457	65
						2572	3014	443	50	2564	3023	460	50

Tabulka 9 zobrazuje druhý souboru dat, zahrnující sekvence, u nichž se jednotlivé programy neshodly v počtu detekovaných CpGIs. Jelikož zde dochází k více případům, kdy byly jako CpGIs označeny zcela jiné úseky DNA sekvence je vzájemné porovnávání složitější. Pokud jsou v úvahu brány jen programy stanovené ostrůvky, které se z větší části překrývají, bylo v predikci dosaženo průměrné odchylky 97,3 při srovnání s programem DBCAT a 113,7 u CpGIsS. Nejshodnější detekce bylo dosaženo v případě BAZ1A, kdy CpGIsP a CpGIsS určily začátek prvního ostrůvku zcela stejně a ve stanovení konce se lišily o pouhých 31 bp. Zhotovený algoritmus však na rozdíl od porovnávaného programu nedokázal druhý ostrůvek o délce 460 bp s pouhým 50% GC obsahem nalézt vůbec. Na druhé straně nejvyšší odlišnosti bylo dosaženo u jediného ostrůvku sekvence MUSRUPL3A. Zde se CpGIsP od vyhledavače DBCAT ve stanovení začátku lišil jen o 20 bp, ale u konce o 482 bp.

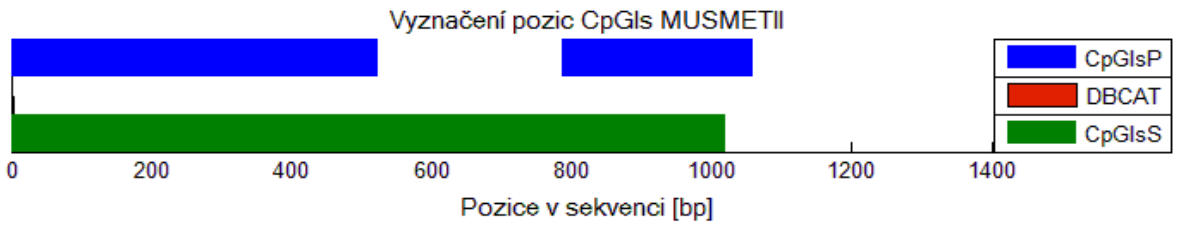


Obr. 23: Srovnání programů v detekci CpGIs u genu BAZ1A, nejshodnější detekce.



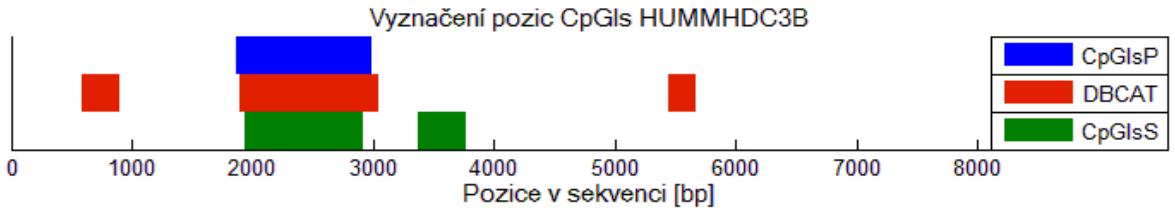
Obr. 24: Srovnání programů v detekci CpGIs u genu MUSRUPL3A, nejodlišnější detekce.

Jelikož u z hotoveného programu GC obsah hraje pouze informativní roli pro uživatele a neovlivňuje konečnou predikci CpGIs, mají ve většině případu stanovené ostrůvky vyšší GC obsah, ale kratší délku ve srovnání s vyhledávači. Na základě tohoto faktu byl u některých genů internetovým vyhledávačem určený ostrůvek, programem CpGIsP rozdělen na několik dílčích úseků. Tento případ nastal například u genů HUMRASH, HUMTBBM40 nebo MUSMETII. Poslední z nich je ukázán níže.

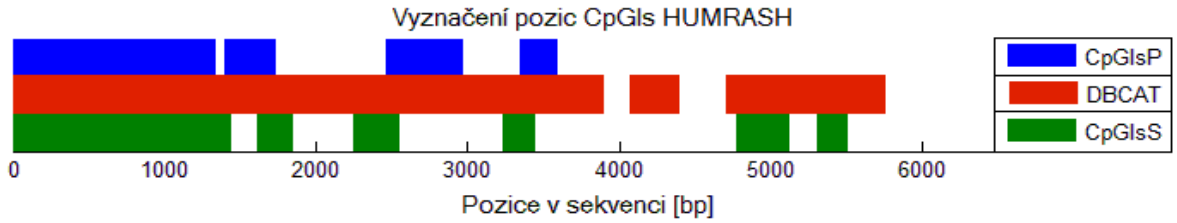


Obr. 25: Srovnání programů v detekci CpGIs u genu MUSMETII, ostrůvek rozdělen na více částí.

Tak jako v první skupině i zde byly jako CpGIs u některého ze tří programů občas stanoveny zcela odlišné části sekvence. Nebylo tomu tak pouze u srovnání CpGIsP s ostatními dvěma vyhledávači, ale lišily se mezi sebou i DBCAT a CpGIsS. Zmíněné chování bylo pozorováno například u HUMRASH, HUMMHDCB, HUMMHDC3B, a dalších. Výsledek označení CpGIs u posledního z uvedených genů HUMMHDC3B je možné vidět na Obr. 26. Co se týče stanovení počtu CpGIs programy u jednotlivých sekvencí velký rozpor byl u genu HUMRASH. V případě tohoto genu, jak je zřetelné z Obr. 27, DBCAT stanovil pouze tři ostrůvky, CpGIsP čtyři a CpGIsS šest.



Obr. 26: Srovnání programů v detekci CpGIs u genu HUMMHDC3B, stanovení zcela odlišných CpGIs.



Obr. 27: Srovnání programů v detekci CpGIs u genu HUMRASH, stanovení rozdílného počtu CpGIs.

Celkovou úspěšnost programu CpGIs_predikce bylo těžké hodnotit, protože jak bylo uvedeno výše, není dostupná informace o přesné poloze CpGIs v sekvencích DNA. Nastala taktéž komplikace při srovnávání výstupů mezi zhotoveným programem a internetovými vyhledávači, protože i mezi nimi nastávaly různé míry shody nebo neshody ve vymezení umístění CpGIs. Navíc použité programy nepracují na principu nukleotidových denzitních vektorů jako zhotovený program, nicméně bylo pro větší věrohodnost použito alespoň stejné nastavení společných parametrů.

Na základě provedeného testování bylo pozorováno, jak už bylo zmíněno výše, že realizovaný program vždy vyhodnotil ostrůvky s větším GC obsahem, ačkoli ve většině případů kratší než ostatní dva vyhledávače. Zároveň, měl problém s určením částí označených jako CpGIs jejichž GC obsah se pohyboval kolem hodnoty 50-55 %. Pro odhalení ostrůvku s takto nízkým GC obsahem by bylo nutné posunout práh k nižším hodnotám, což by na druhé straně mohlo v konkrétních případech vést k falešnému označení dalších částí sekvence jako CpGIs. Z celkových dvaceti sekvencí se u deseti z nich CpGIsP shodoval ve stanovení počtu ostrůvku s oběma programy a u čtyř se shodoval s jedním z nich. Ze zbylého počtu pak ve čtyřech případech bylo nalezeno CpGIsP méně ostrůvku, většinou v souvislosti s jejich malým GC obsahem. Jak již bylo popsáno u konkrétních příkladů, odchylky stanovených

pozic u korespondujících ostrůvků stanovených jednotlivými programy se pohybují v širokém intervalu od 0 až do 515 bp. Závěrem je však nutné podotknout, že ve velké většině případů se stanovené ostrůvky mezi CpGIsP a oběma nebo alespoň jedním vyhledavačem z větší či menší míry překrývaly a nedošlo k žádné nesmyslné predikci.

ZÁVĚR

V této práci bylo dosaženo vyvinutí algoritmu CpGIs_predikce pro grafickou reprezentaci denzit jednotlivých nukleotidů, nukleotidů dle biochemických vlastností a zejména pro predikci CpG ostrůvků. Parametry popisující nalezené CpGIs jsou začátky, konce, délky a GC obsahy. Zároveň je pro lepší představu k dispozici jejich grafické znázornění. Pro pohodlnější práci a orientaci uživatele byl vytvořen přehledný čelní panel. Testovaná data lze do grafického prostředí zadat ručně nebo je k dispozici načtení formátu FASTA.

Nejdříve byly veškeré realizace spuštěny na uměle vytvořených sekvencích, s ostrůvky na přibližně známých lokalizacích, s definovaným obsahem adeninu a tyminu 0 %, 5 %, 10 % a 15 % v nich. Délka posuvného okna byla na základě analýzy zvolena na hodnotu 19, jako kompromis mezi rozlišením a dobrou vizualitou. Výsledný program dokáže na základě zvolené délky okna, prahu a vypočtené sumy nukleotidových denzit cytosinu a guaninu zobrazit jak grafický výstup s vyznačeným prahem a viditelnými CpG ostrůvky, tak matici hodnot začátků, konců, délky a GC obsahu detekovaných CpG ostrůvků. Hodnoty výsledných matic pro jednotlivé sekvence a prahy jsou přehledně uvedeny v tabulce, na jejímž základě může být odvozeno výsledné hodnocení úspěšnosti predikce CpG ostrůvků. Z celkového pohledu, který zahrnuje všechny připravené sekvence, bylo nejlepší detekce dosaženo s prahem 0,7, kdy v sekvencích byly správně klasifikovány pouze dva CpG ostrůvky, i když hodnoty začátků, konců a délky jsou poměrně zkresleny. Nejpřesnějšího stanovení všech parametrů bylo u prahu 0,9, avšak pouze u sekvence, u které byl v CpG ostrůvcích 0% výskyt A a T.

Na základě testování programu na uměle vytvořených sekvencích byly vypozorovány nedostatky, které negativně ovlivňovaly konečnou detekci CpGIs. Proto bylo realizováno shluknutí kratších, nepříliš vzdálených, úseků bohatých na CpG dinukleotidy do jednoho CpG ostrůvku. Zároveň byla nastavena minimální délka ostrůvků, čímž došlo k omezení zobrazení příliš krátkých úseků. Pouze pro informativní účel byly do grafu vykreslení denzity nukleotidů C a G zakomponovány vertikální čáry označující začátky a konce stanovených CpGIs.

Na závěr bylo uskutečněno testování programu na dvaceti reálných sekvencích s očekávaným obsahem CpGIs, získaných z databáze NCBI. Jelikož nejsou k dispozici informace o přesných polohách CpGIs v sekvencích DNA, byly pro interpretaci a srovnání výsledků získaných z programu CpGIs_predikce, vybrány dva volně dostupné internetové vyhledávače CpGIs. Vstupní parametry internetových vyhledávačů byly zvoleny podle Gardiner-Gardenových a Frommerových kritérií na hodnoty CG obsah ≥ 50 %, délka CpG ostrůvku ≥ 200 bp a Obs_{CpG}/Exp_{CpG} hodnota > 0,65. U zhotoveného programu bylo potřeba

zvolit délku okna, která se automaticky počítala na základě provedené analýzy, práh, který se odvíjel od konkrétního vzhledu nukleotidového denzitního vektoru C a G, minimální délka ostrůvku byla nastavena na 200 bp a jejich minimální vzdálenost na 100 bp.

Na základě shrnutí celé analýzy, bylo zjištěno, že realizovaný program vyhodnotil ve většině případů kratší ostrůvky, avšak vždy s větším GC obsahem než zbylé dva vyhledávače. Zároveň, měl potíže s určením úseků stanovených jako CpGIs jejichž GC obsah se pohyboval kolem hodnoty 50-55 %. Z celkových dvaceti sekvencí se u deseti z nich CpGIsP souhlasil v určení počtu ostrůvku s oběma programy a u čtyř se shodoval s jedním z nich. Ze zbylého počtu pak ve čtyřech případech bylo nalezeno CpGIsP méně ostrůvku, většinou v souvislosti s jejich malým GC obsahem. Odchylky stanovených pozic u korespondujících ostrůvků stanovených jednotlivými programy se pohybují v širokém intervalu od 0 až do 515 bp. Nakonec je však nezbytné podotknout, že v převážné většině případů se stanovené ostrůvky mezi CpGIsP a oběma nebo alespoň jedním vyhledavačem z větší či menší části překrývaly a u žádného genu nedošlo ke zcela zcestné predikci.

ZDROJE

[1] ANTEQUERA, F. *Structure, function and evolution of CpG island promoters.* Cell. Mol. Life Sci. 2003, vol. 60, p. 1647-1658.

[2] ZHAO, Zhongming a Leng HAN. *CpG islands: Algorithms and applications in methylation studies. Biochemical and Biophysical Research Communications.* [online]. 2009, vol. 382, issue 4, p. 643-645 [cit. 2013-05-13]. DOI: 10.1016/j.bbrc.2009.03.076. Dostupné z: http://www.ncbi.nlm.nih.gov/pmc/articles/PMC2679166/

[3] ALBERTS, Bruce. *Základy buněčné biologie: úvod do molekulární biologie buňky.* 2. vyd. Ústí nad Labem: Espero, 2005, 740 s. ISBN 80-902906-2-0.

[4] DOSTÁL, Jiří, et al. *Biochemie: pro posluchače bakalářských oborů.* 1. vyd. Brno: Masarykova univerzita, 2009, 158 s. ISBN 978-80-210-5020-4.

[5] CAMPBELL, Neil A a Jane B REECE. *Biologie.* 1. vyd. Brno: Computer Press, 2006, xxxiv, 1332 s. ISBN 80-251-1178-4.

[6] *Bioinformatika – Návod do 1. počítačových cvičení.* Brno: VUT, FEKT, ÚBMI, 2012, 6 s.

[7] NEČAS, Oldřich et al. *Obecná biologie: pro lékařské fakulty*. Jinočany: H+H, 2000, 554 s. ISBN 80-86022-46-3.

[8] IPSER, Jan. *Genetika.* Ústí nad Labem: Univerzita Jana Evangelisty Purkyně, 2006, 197 s.

[9] BOWATER, Richard P. *DNA structure*. *Encyclopedia of Life Sciences* [online]. Chichester, UK: John Wiley, 2006, vol. 12, Suppl 2 [cit. 2013-05-13]. DOI: 10.1038/npg.els.0006002. Dostupné z: http://doi.wiley.com/10.1038/npg.els.0006002

[10] SNUSTAD, D a Michael J SIMMONS. *Genetika*. Brno: Masarykova univerzita, 2009, xxi, 871 s. ISBN 978-802-1048-522.

[11] POLÍVKOVÁ, Z. Imprinting genů a lidské patologie. *Časopis lékařů českých*. 2005, roč. 144, č. 4, s. 245–250.

[12] GARDINER-GARDEN, M a M FROMMER. *CpG islands in vertebrate genomes.* J. Mol. Biol. 1987, vol. 196, p. 261-282.

[13] SPONTANEO, Leah a Nick CERCONE. *Correlating CpG islands, motifs, and sequence variants in human chromosome 21. BMC Genomics* [online]. 2011, vol. 12, Suppl 2, S10- [cit. 2013-05-13]. DOI: 10.1186/1471-2164-12-S2-S10. Dostupné z: http://www.biomedcentral.com/1471-2164/12/S2/S10

[14] HAN, Leng, et al. *CpG island density and its correlations with genomic features in mammalian genomes. Genome Biology* [online]. 2008, vol. 9, issue 5, R79-

[cit. 2013-05-13]. DOI: 10.1186/gb-2008-9-5-r79. Dostupné z: http://genomebiology.com/2008/9/5/R79

[15] HACKENBERG, M, et al. *CpGcluster: a distance-based algorithm for CpG-island detection.* BMC Bioinformatics. 2006, 7:446.

[16] RŮŽEK, Václav. *Algoritmy pro rozpoznání ručně psaných znaků.* Zlín: Univerzita Tomáše Bati ve Zlíně. Fakulta aplikované informatiky. Ústav řízení procesů, 2010, 66 s. Vedoucí bakalářské práce Ing. Petr Chalupa, Ph.D.

[17] DIETERICH, C. *Algorithms in Bioinforamtics I* [online pdf]. [cit. 2013-05-13]. WS06, ZBIT, 2007, p. 193 – 216. Dostupné z: http://ab.inf.uni-tuebingen.de/teaching/ws06/albi1/script/MarkovChainsAndHMMs_complete.pdf

[18] MNEIMNEH, Saad. *Computational Biology Lecture 9: CpG islands, Markov Chains, Hidden Markov Models HMMs* [online pdf]. [cit. 2013-05-13]. Dostupné z: http://www.cs.hunter.cuny.edu/~saad/courses/compbio/lectures/lecture9.pdf

[19] *HMM: Viterbi algorithm – a toy expample* [online prezentace]. [cit. 2013-05-13]. Dostupné z: http://homepages.ulb.ac.be/~dgonze/TEACHING/viterbi.pdf

[20] *Bioinformatika – Návod do 9. počítačových cvičení.* Brno: VUT, FEKT, ÚBMI, 2012, 4 s.

[21] MADĚRÁNKOVÁ, Denisa a Ivo PROVAZNÍK. *Motive Representation in Nucleotide Densities of Bird's Mitochondrial Gene COX1*. Brno: Brno University of Technology. Faculty of Electrical Engineering and Communications. Department of Biomedical Engineering, 5 p.

[22] NCBI [online]. [cit. 2013-05-13]. Dostupné z: http://www.ncbi.nlm.nih.gov/

[23] CpG Island Searcher. [online]. [cit. 2013-05-13]. Dostupné z: http://cpgislands.usc.edu/

[24] DataBase of CpG islands and Analytical Tool [online]. 5. listopadu 2008 [cit. 2013-05-13]. Dostupné z: http://www.dbcat.cgm.ntu.edu.tw

PŘÍLOHY

A Seznam zkratek

DNA	Deoxyribonukleová kyselina
A	Adenin
C	Cytosin
G	Guanin
T	Tymin
OH	Hydroxylová skupina
RNA	Ribonukleová kyselina
mRNA	Mediátorová RNA
bp	Páry bazí
CpG	Cytosin a guanin s fosfodiesterovou vazbu
CpGs	CpG dinukleotidy
CpGIs	CpG ostrůvek/ky
HMM	Skrytý Markovův model (Hidden markov model)
CpGIsP	Program CpGIs_predikce
CpGIsS	Internetový vyhledávač CpG Island Searcher

B Přehled výsledků analýzy délky okna

Tabulka 10: Analýza vlivu velikosti posuvného okna na detekci CpGIs, kde *w* je velikost okna, *Z* je začátek a *K* je konec CpGIs, *GC%* je obsah G a C v CpGIs a *O* je odchylka od správných pozic. Červeně jsou znázorněny chybně detekované CpGIs.

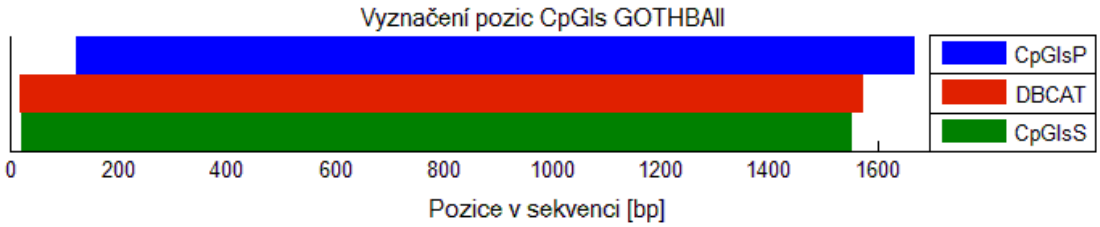
Práh = 0,8															
w	**Z**	**O**	**K**	**O**	**D**	**O**	**GC%**	**w**	**Z**	**O**	**K**	**O**	**D**	**O**	**GC%**
5	4		12		9		67	**15**	42	9	100	0	59	-9	82
	44		69		26		85		190	11	305	-5	116	-16	86
	69	-18	99	1	31	19	81		408	2	425	-1	18	-3	73
	104		109		6		67		*Počet chybně detekovaných CpGIs: 0*						
	112		121		10		70		*Počet nenalezených CpGIs: 0*						
	128		133		6		67	**17**	42	9	100	0	59	-9	82
	147		162		16		69		190	11	304	-4	115	-15	87
	180		187		8		63		*Počet chybně detekovaných CpGIs: 0*						
	190		199		10		70		*Počet nenalezených CpGIs: 1*						
	199	2	303	-3	105	-5	88	**19**	42	9	100	0	59	-9	82
	325		332		8		63		190	11	304	-4	115	-15	87
	340		352		13		77		*Počet chybně detekovaných CpGIs: 0*						
	408	2	425	-1	18	-3	73		*Počet nenalezených CpGIs: 1*						
	427		435		9		67	**21**	40	11	101	-1	62	-12	80
	447		454		8		63		189	12	305	-5	117	-17	85
	476		484		9		67		*Počet chybně detekovaných CpGIs: 0*						
	Počet chybně detekovaných CpGIs: 13								*Počet nenalezených CpGIs: 1*						
	Počet nenalezených CpGIs: 0							**23**	40	11	99	1	60	-10	82
11	44	7	99	1	56	-6	84		189	12	304	-4	116	-16	86
	148		159		12		75		*Počet chybně detekovaných CpGIs: 0*						
	194	7	304	-4	111	-11	87		*Počet nenalezených CpGIs: 1*						
	340		353		14		72	**25**	39	12	100	0	62	-12	80
	409	1	425	-1	17	-2	77		188	13	305	-5	118	-18	84
	Počet chybně detekovaných CpGIs: 2								*Počet chybně detekovaných CpGIs: 0*						
	Počet nenalezených CpGIs: 0								*Počet nenalezených CpGIs: 1*						
13	44	7	99	1	56	-6	84								
	194	7	304	-4	111	-11	87								
	409	1	424	0	16	-1	82								
	Počet chybně detekovaných CpGIs: 0														
	Počet nenalezených CpGIs: 0														

Tabulka 11: Analýza vlivu velikosti posuvného okna na detekci CpGIs, kde *w* je velikost okna, *Z* je začátek a *K* je konec CpGIs, *GC%* je obsah G a C v CpGIs a *O* je odchylka od správných pozic. Červeně jsou znázorněny chybně detekované CpGIs.

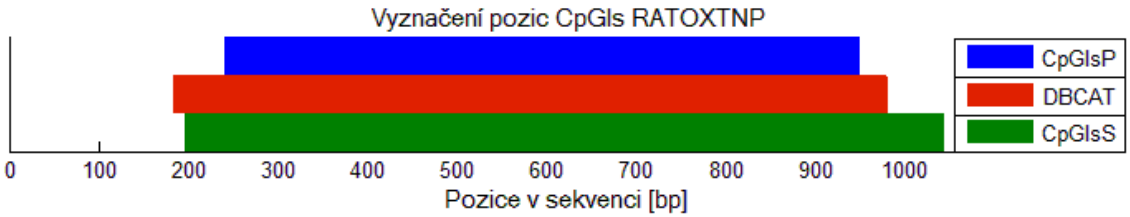
Práh = 0,9															
w	**Z**	**O**	**K**	**O**	**D**	**O**	**GC%**	**w**	**Z**	**O**	**K**	**O**	**D**	**O**	**GC%**
5	5		10		6		84	**15**	46	46	69	31	24	26	88
	50	1	68	32	19	31	95		80		98		19		90
	70		77		8		88		199		217		19		85
	80		89		10		90		218	-17	277	23	60	40	90
	90		98		9		89		283		300		18		89
	200	1	214	86	15	85	94		*Počet chybně detekovaných CpGIs: 3*						
	218		225		8		88		*Počet nenalezených CpGIs: 1*						
	226		235		10		90	**17**	46	5	69	31	24	26	88
	236		244		9		89		80		98		19		90
	245		254		10		90		200		217		18		89
	255		267		13		93		218	-17	277	23	60	40	90
	268		277		10		90		283		300		18		89
	283		289		7		86		*Počet chybně detekovaných CpGIs: 3*						
	290		300		11		91		*Počet nenalezených CpGIs: 1*						
	341		348		8		88	**19**	46	5	69	31	24	26	88
	416	-6	424	0	9	6	89		245	-44	277	23	33	67	91
	448		453		6		84		*Počet chybně detekovaných CpGIs: 0*						
	477		483		7		86		*Počet nenalezených CpGIs: 1*						
	Počet chybně detekovaných CpGIs: 15							**21**	44	7	77	23	34	16	86
	Počet nenalezených CpGIs: 0								200		225		26		89
11	46	5	69	31	24	26	88		218	-17	300	0	83	17	90
	78		98		21		86		*Počet chybně detekovaných CpGIs: 1*						
	199		217		19		85		*Počet nenalezených CpGIs: 1*						
	218	-17	303	-3	86	14	89	**23**	44	7	77	23	34	16	86
	341		352		12		84		200		225		26		89
	Počet chybně detekovaných CpGIs: 3								218	-17	282	18	65	35	90
	Počet nenalezených CpGIs: 1								*Počet chybně detekovaných CpGIs: 1*						
13	46	5	69	31	24	26	88		*Počet nenalezených CpGIs: 1*						
	80		98		19		90	**25**	50	1	77	23	28	22	90
	199		217		19		85		200		225		26		89
	218	-17	282	18	65	35	90		218	-17	282	18	65	35	90
	283		303		21		86		*Počet chybně detekovaných CpGIs: 1*						
	Počet chybně detekovaných CpGIs: 3								*Počet nenalezených CpGIs: 1*						
	Počet nenalezených CpGIs: 1														

C Srovnání detekce CpGIs programy

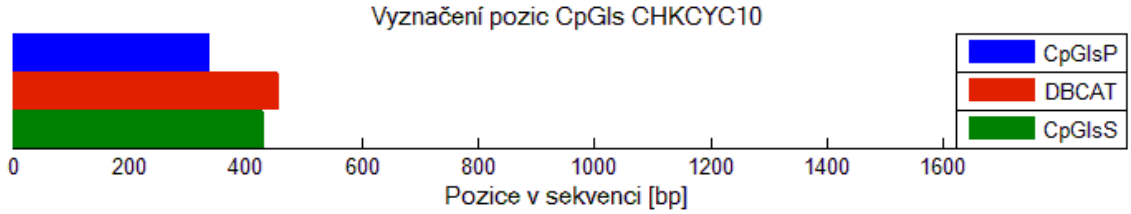
Obr. 28: Srovnání programů v detekci CpGIs u genu GOTHBAI.



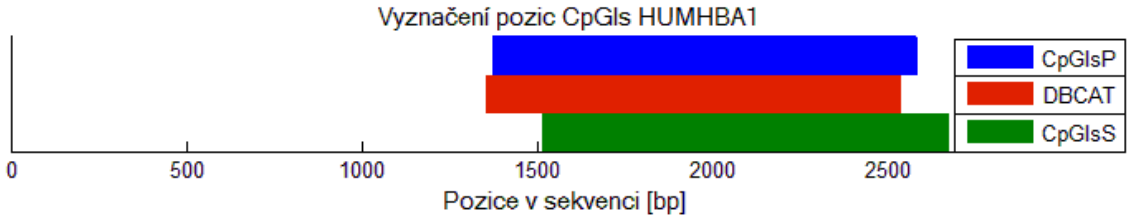
Obr. 29: Srovnání programů v detekci CpGIs u genu GOTHBAII.



Obr. 30: Srovnání programů v detekci CpGIs u genu RATOXTNP.

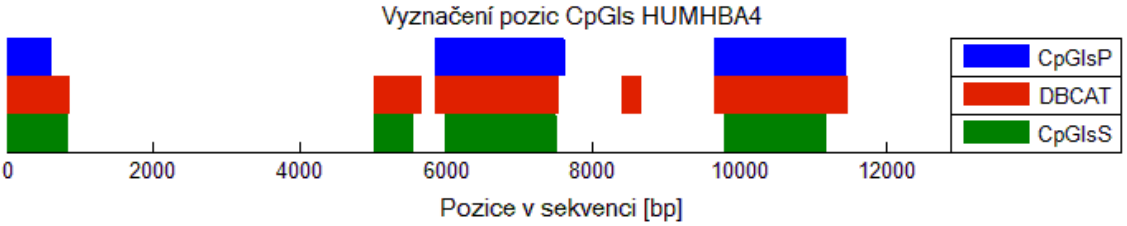


Obr. 31: Srovnání programů v detekci CpGIs u genu CHKCYC10.

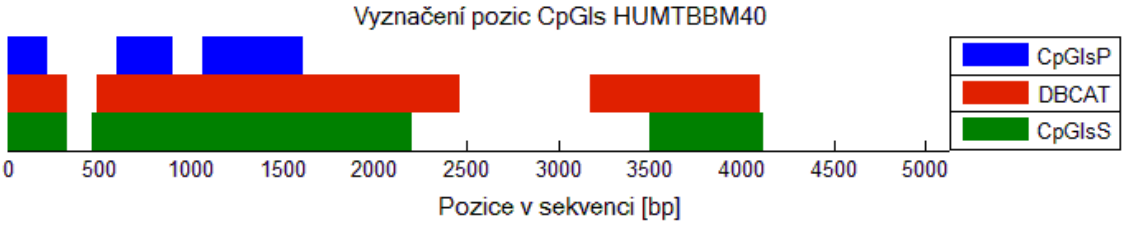


Obr. 32: Srovnání programů v detekci CpGIs u genu HUMHBA1.

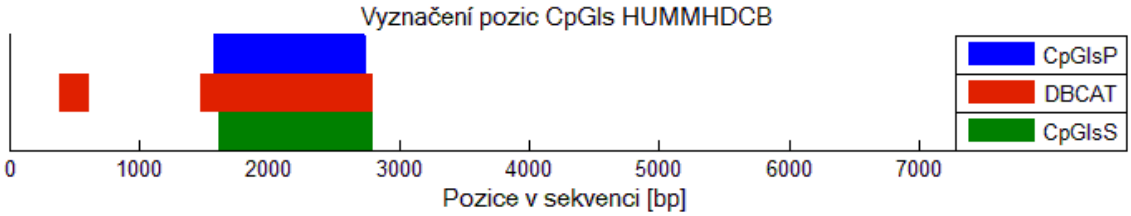
Obr. 33: Srovnání programů v detekci CpGIs u genu CHKH2A2B.



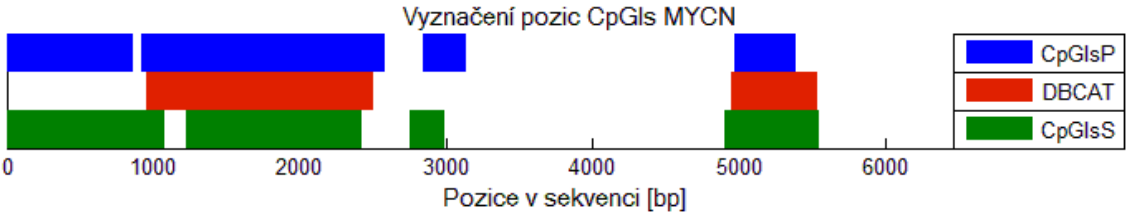
Obr. 34: Srovnání programů v detekci CpGIs u genu HUMHBA4.



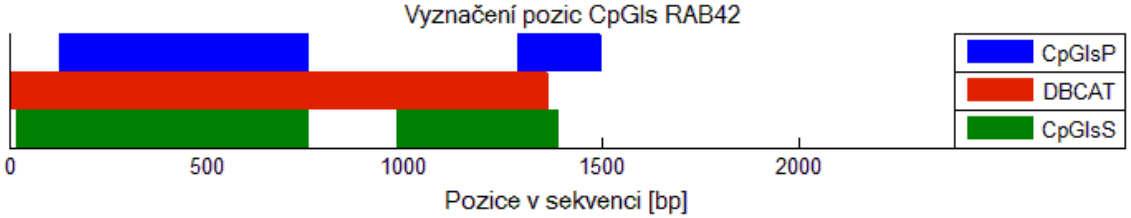
Obr. 35: Srovnání programů v detekci CpGIs u genu HUMTBBM40.



Obr. 36: Srovnání programů v detekci CpGIs u genu HUMMHDCB.



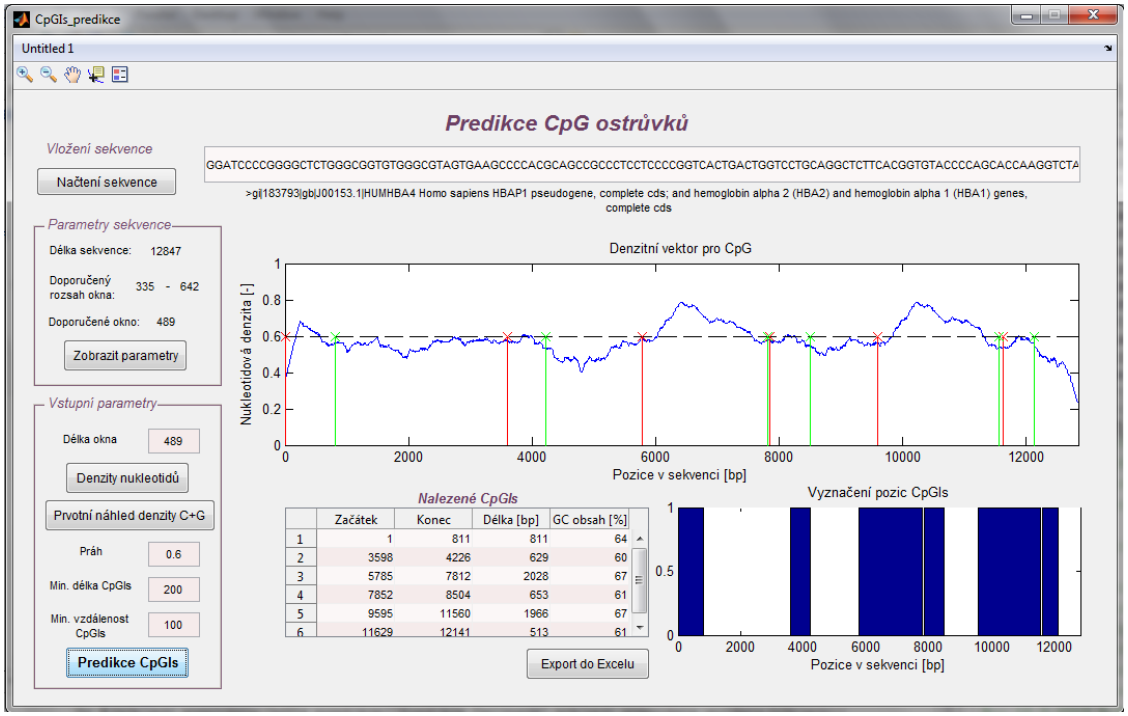
Obr. 37: Srovnání programů v detekci CpGIs u genu MYCN.



Obr. 38: Srovnání programů v detekci CpGIs u genu RAB42.

D Uživatelský manuál

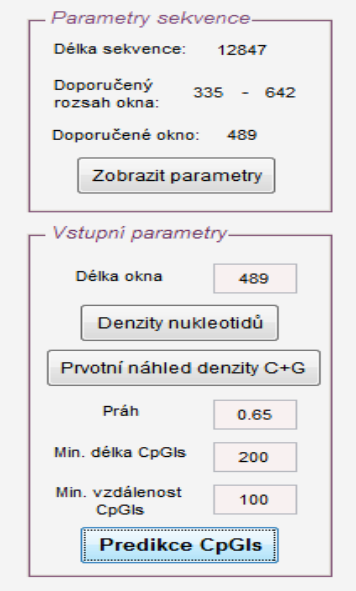
Program CpGIs_predikce umožňuje detekci CpGIs v DNA sekvencích zadaných ručně nebo načtením FASTA formátu. Uživatel získá informace o nalezených CpGIs v podobě tabulky jejich začátků, konců, délek a GC obsahů, navíc doprovázených grafickým znázorněním výsledné detekce.



Obr. 39: Uživatelské rozhraní programu CpGIs_predikce.

Doporučený postup:

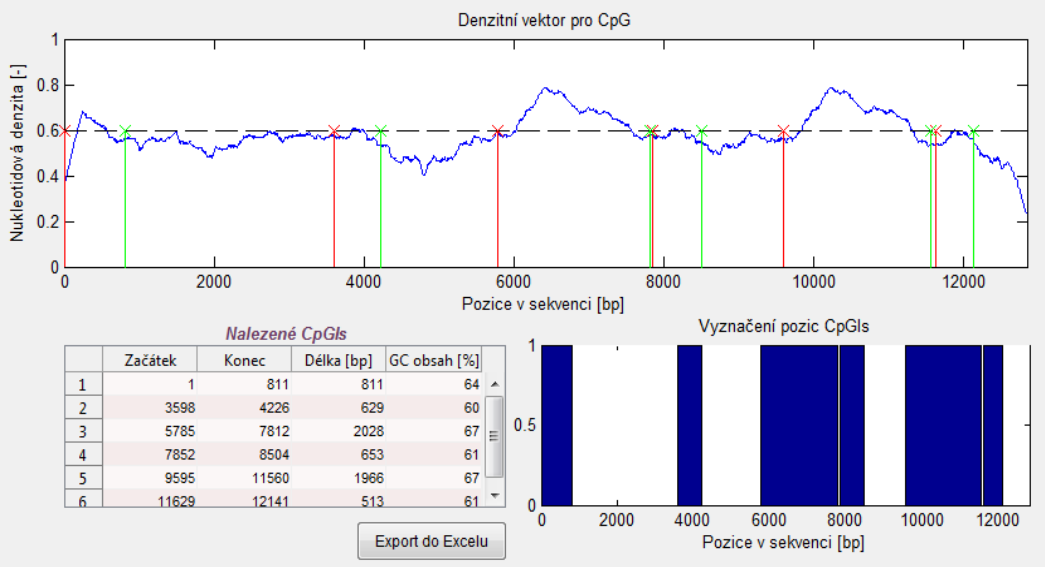
1) Program CpGIs_predikce zahajte v MATLABU stiskem ikony *run*, která je znázorněná zelenou šipkou a umístěna v horním panelu editoru přibližně ve středu.
2) Vložení sekvence:
   a) Zadání sekvence můžete provést ručně zapsáním do připraveného podélného, světlého pole, kdy není třeba brát ohled na malá či velká písmena.
   b) Druhou možností je načtení FASTA souboru tlačítkem *Načtení sekvence*, kterým je zobrazeno dialogové okno pro výběr sekvence. Při správném načtení je sekvence opět viditelná ve zmíněném světlém poli a zároveň je pod ním vypsána hlavička z FASTA souboru.



Obr. 40: Panely pro zobrazení parametrů sekvence a zadání požadovaných hodnot s tlačítky spouštěcími zvolené funkce.

3) Volbou tlačítka *Zobrazit parametry* se v panelu *Parametry sekvence* vypíše délka sekvence, doporučený rozsah okna i samotná hodnota okna doporučená pro další výpočet.
4) Zvažte velikost okna pro výpočet nukleotidových denzitních vektorů. Automaticky je zde vložena doporučená velikost okna, kterou si však podle uvážení můžete přepsáním změnit. Hodnota okna musí být liché číslo a měla by ležet v doporučeném intervalu a nesmí překročit jeho horní hranici, kterou je 1/20 délky sekvence. Okno musí být zadáno pro všechny další úkony.
5) Pokud se chcete podívat na grafickou reprezentaci denzit jednotlivých nukleotidů nebo nukleotidů uspořádaných podle biochemických vlastností, zvolte tlačítko *Denzity nukleotidů*. Po stisku tlačítka vyběhnou dva grafické okna s požadovanými výstupy.
6) Pro pohled na grafické zobrazení denzity GC je opět nutné mít zvolenou velikost okna. Poté vykreslení zahájíte stisknutím tlačítka *Prvotní náhled*. Tento krok je čistě na uživateli a může být vynechán, avšak umožňuje lépe rozvážit volbu prahu.
7) Zvolte ostatní požadované parametry nutné pro predikci CpGIs. Kromě prahu jsou všechny další parametry nastaveny na výchozí doporučené hodnoty, které opět můžete libovolně přepsat.
   a) Práh se může pohybovat v rozmezí 0-1 a je klíčový pro detekci CpGIs.
   b) Minimální délka CpGIs určuje, že detekované úseky kratší než zadaná hodnota budou z výsledku smazány.
   c) Minimální vzdálenost CpGIs udává, že pokud budou ostrůvky překročující práh od sebe vzdáleny méně než zvolená hodnota, budou sloučeny.
8) Pro detekci CpGIs stiskněte tlačítko *Predikce CpGIs*.

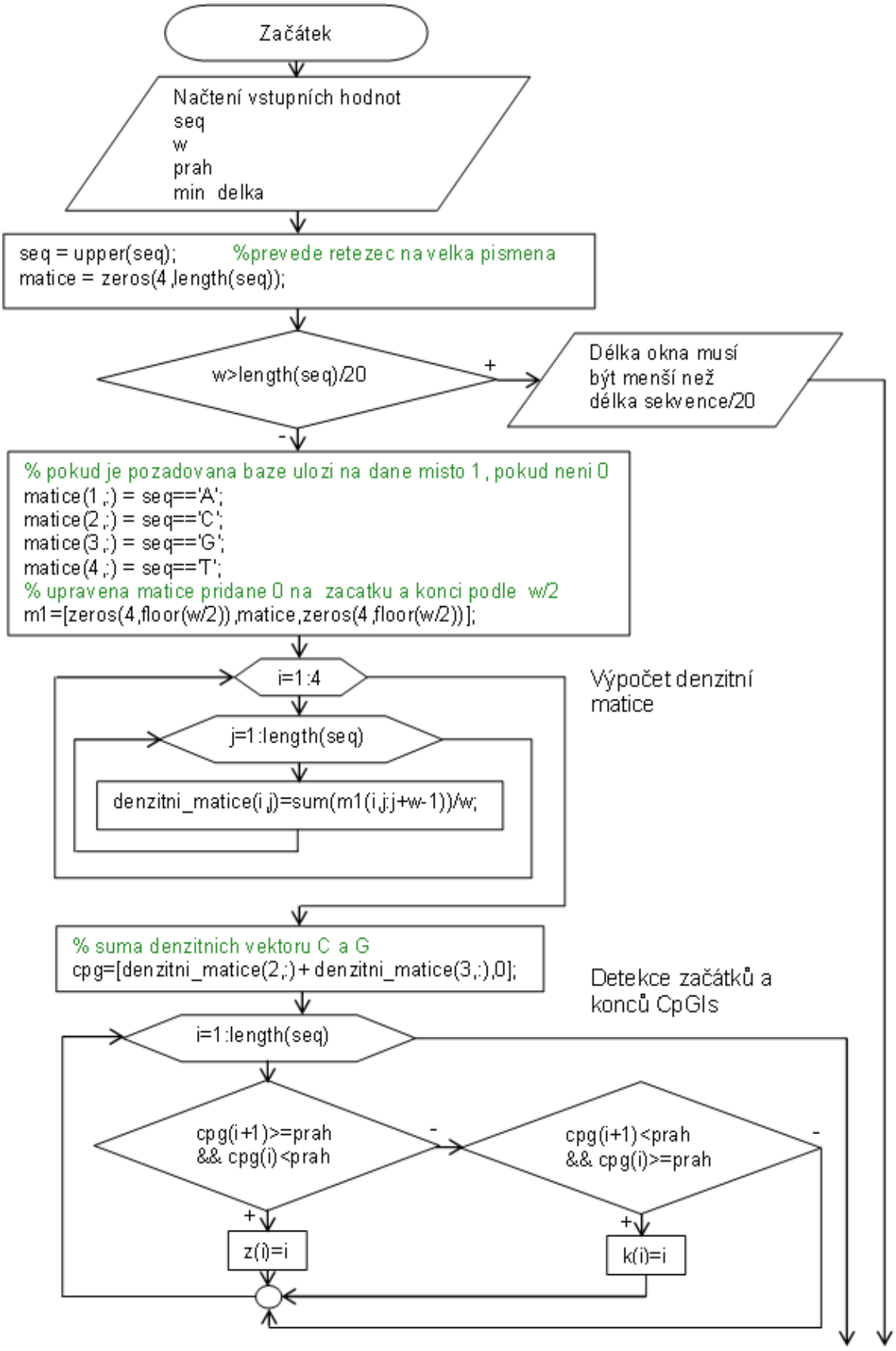
9) V případě, že v sekvenci nebyl nalezen žádný CpGIs, je zobrazeno okno s upozorněním.
10) Při úspěšné predikci ostrůvků, je graficky zobrazena denzita GC spolu z vyznačeným prahem a mezemi ostrůvku. Zároveň jsou do tabulky vypsány jejich parametry a ostrůvky jsou schematicky znázorněny.

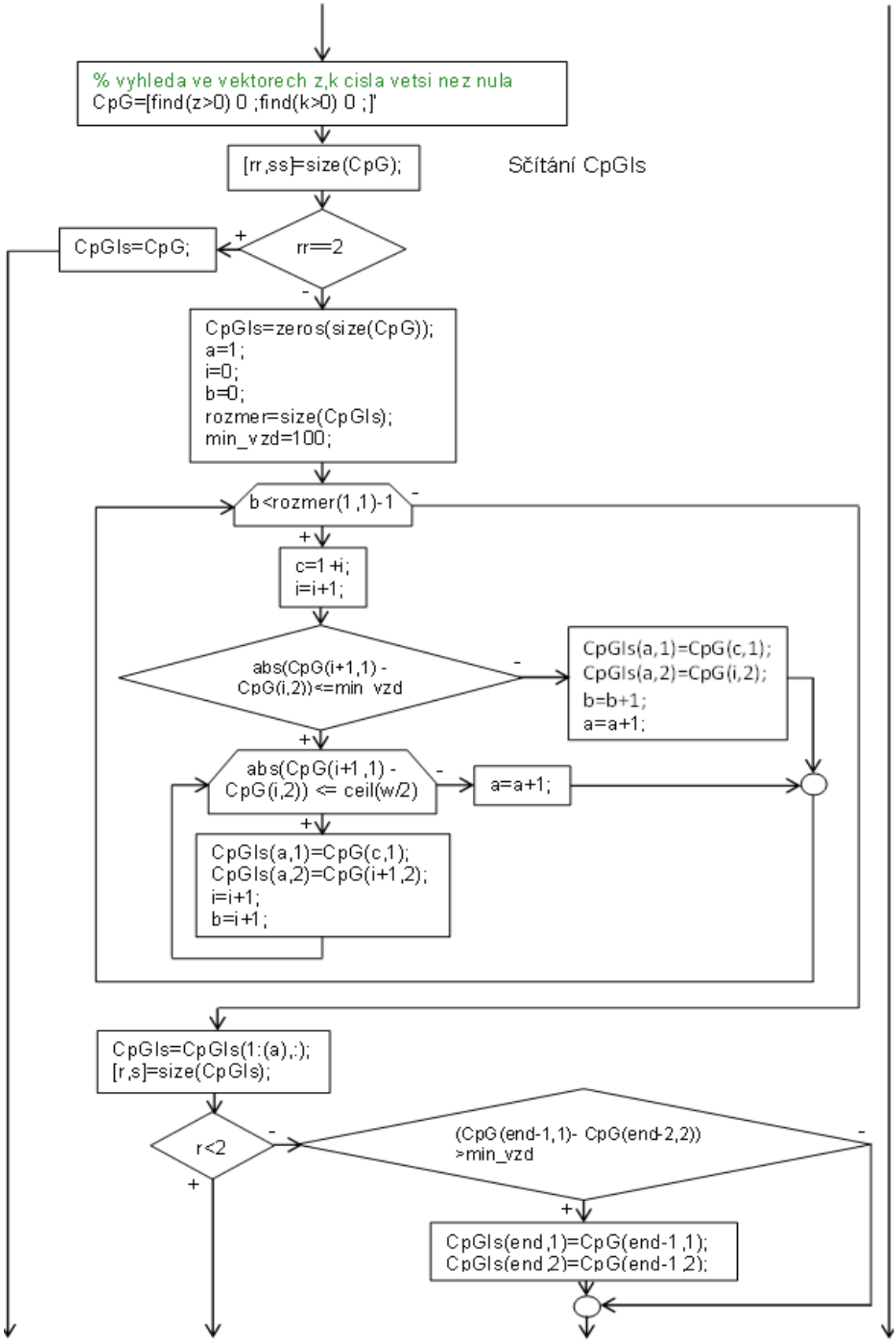


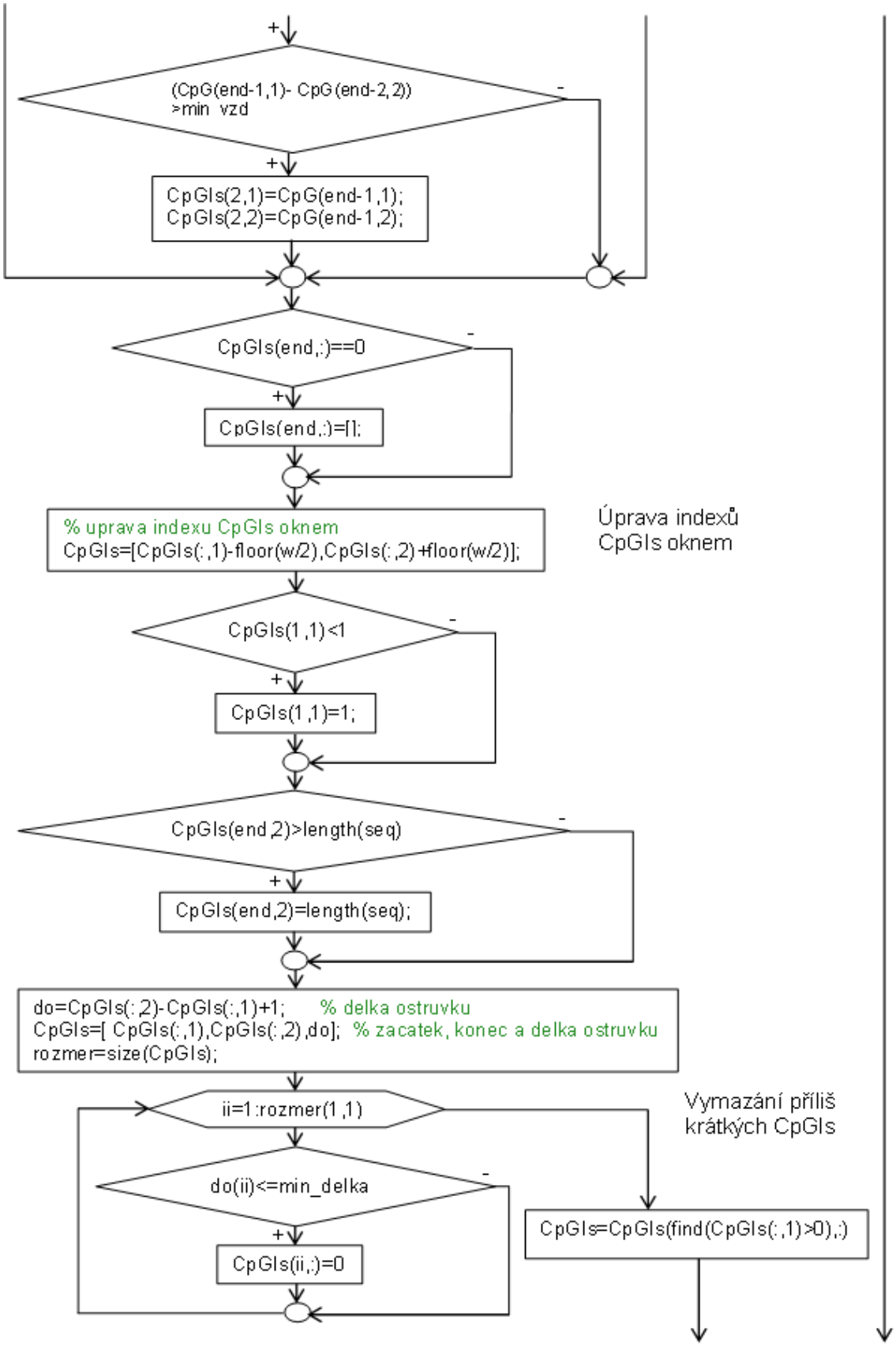
Obr. 41: Část čelního panelu zobrazující denzitní vektor pro CpG dinukleotidy, tabulku zjištěných parametrů CpGIs a schematicky znázorněné pozice ostrůvků.

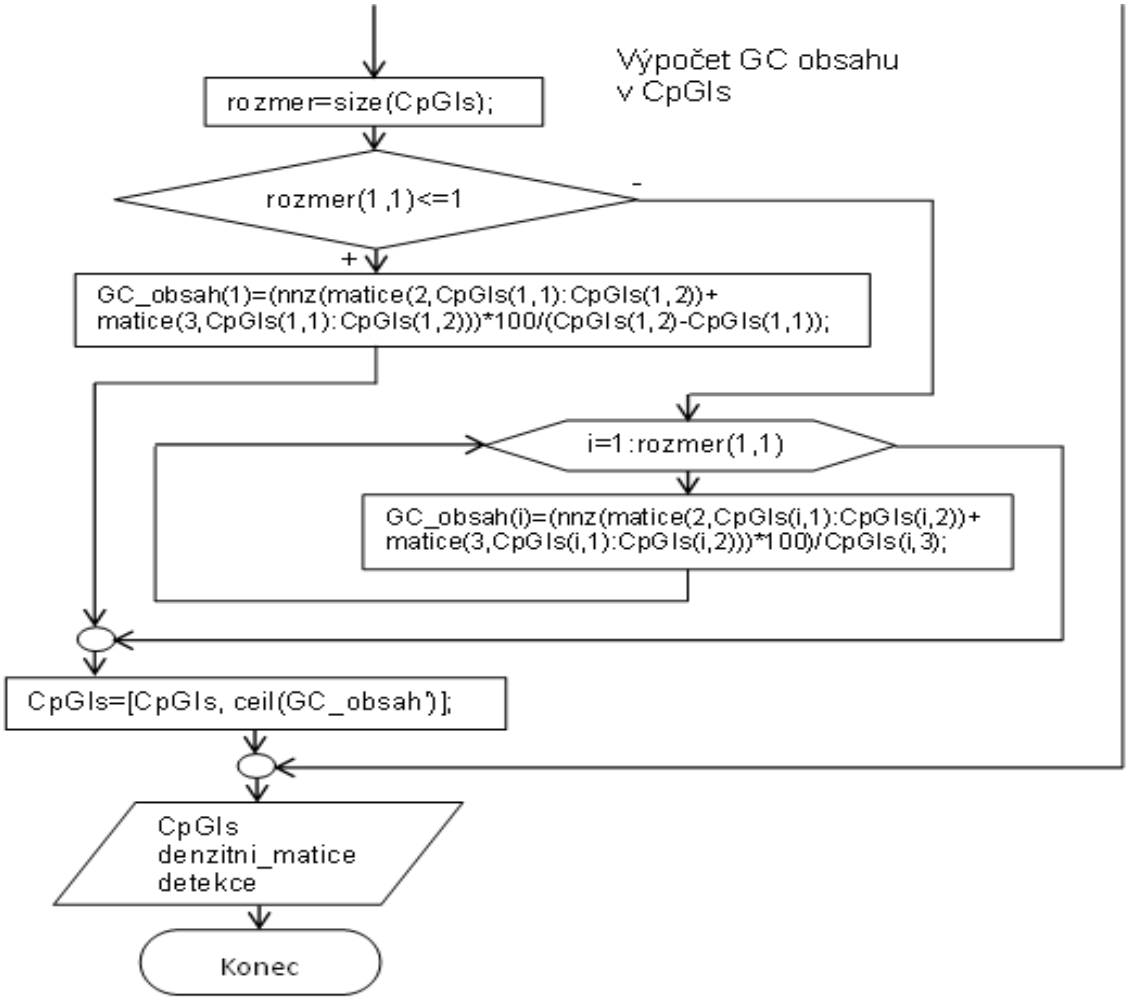
11) Na horní liště panelu jsou k dispozici ikony pro práci s grafy. Stiskněte zvolenou ikonu a přejeďte kursorem myši do oblasti grafu. Poté klikem nebo tahem můžete provádět zvolené změny.
    a) Přiblížení zvoleného úseku pomoci ikony lupy s plus.
    b) Vzdálení zvoleného úseku pomoci ikony lupy s mínus.
    c) Pohyb v oblasti grafu pomoci pačičky.
    d) Zobrazení souřadnic zvoleného bodu pomoci data kursoru.
    e) Vytvoření legendy grafu pomoci poslední ikony na liště.
12) Pokud máte zájem výsledky dále zpracovávat v tabulkovém editoru MS Excel, stiskněte tlačítko *Export do Excelu* umístěné pod tabulkou nalezených CpGIs a data se uloží do souboru s názvem data.
13) Program ukončete stisknutím křížku v pravém horním rohu.

E Vývojový diagram funkce denzita




% vyhleda ve vektorech z,k cisla vetsi nez nula
CpG=[find(z>0) 0 ;find(k>0) 0 ;]'
[rr,ss]=size(CpG);
Sčítání CpGIs
rr==2
+
CpGIs=CpG;
-
CpGIs=zeros(size(CpG));
a=1;
i=0;
b=0;
rozmer=size(CpGIs);
min_vzd=100;
b<rozmer(1,1)-1
-
+
c=1+i;
i=i+1;
abs(CpG(i+1,1) - CpG(i,2))<=min_vzd
-
CpGIs(a,1)=CpG(c,1);
CpGIs(a,2)=CpG(i,2);
b=b+1;
a=a+1;
+
abs(CpG(i+1,1) - CpG(i,2)) <= ceil(w/2)
-
a=a+1;
+
CpGIs(a,1)=CpG(c,1);
CpGIs(a,2)=CpG(i+1,2);
i=i+1;
b=i+1;
CpGIs=CpGIs(1:(a),:);
[r,s]=size(CpGIs);
r<2
-
(CpG(end-1,1)- CpG(end-2,2)) >min_vzd
-
+
+
CpGIs(end,1)=CpG(end-1,1);
CpGIs(end,2)=CpG(end-1,2);


(CpG(end-1,1)- CpG(end-2,2))
>min_vzd
+
-
CpGIs(2,1)=CpG(end-1,1);
CpGIs(2,2)=CpG(end-1,2);
CpGIs(end,:)==0
CpGIs(end,:)=[];
% uprava indexu CpGIs oknem
CpGIs=[CpGIs(:,1)-floor(w/2),CpGIs(:,2)+floor(w/2)];
Úprava indexů
CpGIs oknem
CpGIs(1,1)<1
CpGIs(1,1)=1;
CpGIs(end,2)>length(seq)
CpGIs(end,2)=length(seq);
do=CpGIs(:,2)-CpGIs(:,1)+1; % delka ostruvku
CpGIs=[CpGIs(:,1),CpGIs(:,2),do]; % zacatek, konec a delka ostruvku
rozmer=size(CpGIs);
ii=1:rozmer(1,1)
do(ii)<=min_delka
CpGIs(ii,:)=0
Vymazání příliš
krátkých CpGIs
CpGIs=CpGIs(find(CpGIs(:,1)>0),:)


Výpočet GC obsahu
v CpGIs
rozmer=size(CpGIs);
rozmer(1,1)<=1
-
+
GC_obsah(1)=(nnz(matice(2,CpGIs(1,1):CpGIs(1,2))+
matice(3,CpGIs(1,1):CpGIs(1,2)))*100/(CpGIs(1,2)-CpGIs(1,1));
i=1:rozmer(1,1)
GC_obsah(i)=(nnz(matice(2,CpGIs(i,1):CpGIs(i,2))+
matice(3,CpGIs(i,1):CpGIs(i,2)))*100)/CpGIs(i,3);
CpGIs=[CpGIs, ceil(GC_obsah')];
CpGIs
denzitni_matice
detekce
Konec