

UNIVERZITA KOMENSKÉHO V BRATISLAVE
FAKULTA MATEMATIKY, FYZIKY A INFORMATIKY

VERIFIKÁCIA A OPTIMALIZÁCIA METÓDY
NA IDENTIFIKÁCIU CENTROMERICKÝCH
LOKUSOV V GENÓMOVÝCH SEKVENCIÁCH
KVASINIEK

BAKALÁRSKA PRÁCA

2023

OLÍVIA KREUTZOVÁ

UNIVERZITA KOMENSKÉHO V BRATISLAVE
FAKULTA MATEMATIKY, FYZIKY A INFORMATIKY

VERIFIKÁCIA A OPTIMALIZÁCIA METÓDY
NA IDENTIFIKÁCIU CENTROMERICKÝCH
LOKUSOV V GENÓMOVÝCH SEKVENCIÁCH
KVASINIEK
BAKALÁRSKA PRÁCA

Študijný program: Bioinformatika
Študijný odbor: Informatika a Biológia
Školiace pracovisko: Katedra informatiky
Školiteľ: prof. RNDr. Jozef Nosek, DrSc.
Konzultant: doc. Mgr. Bronislava Brejová, PhD.

Bratislava, 2023
Olívia Kreutzová



ZADANIE ZÁVEREČNEJ PRÁCE

Meno a priezvisko študenta: Olívia Kreutzová
Študijný program: bioinformatika (Medziodborové štúdium, bakalársky I. st., denná forma)
Študijné odbory: informatika
biológia
Typ záverečnej práce: bakalárska
Jazyk záverečnej práce: slovenský
Sekundárny jazyk: anglický

Názov: Verifikácia a optimalizácia metódy na identifikáciu centromerických lokusov v genómových sekvenciách kvasiniek
Verification and optimization of a method for the identification of centromeric loci in yeast genomic sequences

Anotácia: Centroméry sú funkčnými elementami eukaryotických chromozómov, ktoré majú kľúčovú úlohu v ich distribúcii počas bunkového delenia. Cieľom bakalárskej práce je využiť a prípadne aj optimalizovať bioinformatický nástroj pre identifikáciu centromerických lokusov v genómových sekvenciách rôznych druhov kvasiniek a uskutočniť komparatívnu analýzu identifikovaných centromér.

Literatúra: 1. Guin a kol. (2020) Implications of the evolutionary trajectory of centromeres in the fungal kingdom. *Annual Review of Microbiology* 74: 835–853.
2. McKinley a Cheeseman (2016) The molecular basis for centromere identity and function. *Nature Reviews Molecular Cell Biology* 17(1): 16–29.
3. Lynch a kol. (2010) Chromosomal G+C content evolution in yeasts: systematic interspecies differences, and GC-poor troughs at centromeres. *Genome Biology and Evolution* 2: 572–583.
4. Shen a kol. (2020) Genome-scale phylogeny and contrasting modes of genome evolution in the fungal phylum Ascomycota. *Science Advances* 6(45): eabd0079.

Vedúci: prof. RNDr. Jozef Nosek, DrSc.
Konzultant: doc. Mgr. Bronislava Brejová, PhD.
Katedra: FMFI.KI - Katedra informatiky
Vedúci katedry: prof. RNDr. Martin Škoviera, PhD.
Dátum zadania: 20.10.2022

Dátum schválenia: 20.10.2022
doc. Mgr. Bronislava Brejová, PhD.
garant študijného programu

.....
študent

.....
vedúci práce

PodĎakovanie: PodĎakovanie patrí môjmu školiťovi prof. RNDr. Jozefovi Nosekovi, DrSc. za spoluprácu, pomoc a ochotu, cenné rady a informácie. Tiež by som chcela poďakovať mojej konzultantke doc. Mgr. Bronislave Brejovej, PhD. za usmernenie pri písaní práce.

Abstrakt

Centroméry hrajú kľúčovú úlohu v bunkovom delení. Miera rekombinácie v tejto oblasti chromozómov býva potlačená, čo ovplyvňuje zastúpenie nukleotidov v sekvencii DNA. Obsah guanínu a cytozínu na tretích pozíciách kodónov (GC3), ktorý vo väčšine organizmov koreluje s mierou rekombinácie, varíruje pozdĺž chromozómu. GC3 metóda na základe nukleotidového zastúpenia pozdĺž chromozómov predikuje centromerické lokusy pre jednotlivé chromozómy a mapuje ho na oblasť s výrazným poklesom zastúpenia GC3. Na identifikáciu centromerickej oblasti sme v tejto práci implementovali metódu GC3 v Python skripte a overili jeho funkčnosť na genómových sekvenciách kvasiniek, pre ktoré je už pozícia centroméry známa a experimentálne overená. Následne sme optimalizovali parametre, pričom sme zistili, že veľkosť zvoleného okna ovplyvňuje predikciu centromerického lokusu, no vplyv vylúčenia niektorých kodónov naň je malý. Metódu sme aplikovali aj na genómové sekvencie kvasiniek, pre ktoré doposiaľ nebola centromerická oblasť predikovaná, napríklad *Candida auris* alebo *Candida parapsilosis*, kde sa nám podarilo centromerický lokus nájsť. Zároveň sme overili, že GC3 metóda nie je univerzálne použiteľná pre genómové sekvencie všetkých druhov kvasiniek, napríklad *Schizosaccharomyces pombe*, kde centroméra nebola v profiloch GC3 viditeľná. Táto bioinformatická metóda dokáže urýchliť a uľahčiť identifikáciu a analýzu centromér v chromozómoch.

Kľúčové slová: centroméra, GC3 metóda, genómová sekvencia, chromozóm, kvasinka

Abstract

Centromeres play a key role in cell division. The rate of the recombination process in this area tends to be suppressed, thus having impact on the nucleotide content. Guanine and cytosine content at the third codon positions (GC3) varies along the chromosome and it correlates with high recombination rates in most organisms. The GC3 method predicts centromeric locus for individual chromosomes and determines its position in an area with a significant decrease in GC3 content. To identify these loci, we implemented the GC3 method in a Python script. We verified its functionality using yeast genome sequences, for which the centromere positions have already been detected. By optimization of parameters we discovered that the size of the selected window affected the prediction of the centromeric locus. On the other hand, the exclusion of ATG and CTG codons had a negligible effect on it. In addition, we applied GC3 method to yeast genome sequences for which the centromeric region had not been predicted yet. We succeeded in finding the centromeric locus for species such as *Candida auris* and *Candida parapsilosis*. At the same time, we proved that the GC3 method is not universally applicable to genome sequences of all yeast species, such as *Schizosaccharomyces pombe*, where the centromere was not observable in the GC3 profiles. GC3 method is an efficient bioinformatical tool which can facilitate finding and analysis of centromeres.

Keywords: centromere, GC3 method, genome sequence, chromosome, yeast

Obsah

Úvod	1
1 Základné genetické pojmy	3
1.1 Genetická informácia a jej interpretácia	3
1.2 Nukleotidové zloženie DNA	3
1.3 Realizácia genetickej informácie	5
1.4 Delenie bunky	6
1.5 Štruktúra a funkcia chromozómu	8
1.6 Centroméry	9
1.6.1 Centroméry v ľudských bunkách	11
1.6.2 Centroméry v kvasinkových bunkách	11
2 Ciele práce	15
3 Metodika	17
3.1 Východiskové publikácie	17
3.2 Dáta a databázy	18
3.3 Použité programové balíčky	20
3.4 Princíp GC3 metódy a naša implementácia	20
3.5 Vstupy a výstupy	23
3.6 Ilustratívny príklad	23
4 Výsledky	27
4.1 Verifikácia GC3 metódy	27
4.1.1 <i>Clavispora lusitaniae</i>	27
4.1.2 <i>Scheffersomyces stipitis</i>	27
4.1.3 <i>Yarrowia lipolytica</i>	29
4.1.4 Porovnanie výsledkov	29
4.2 Optimalizácia parametrov	32
4.2.1 Veľkosť okna	32
4.2.2 Vylúčenie kodónov ATG a CTG	34

4.3	Aplikácia GC3 metódy na ďalšie genómy kvasiniek	36
4.3.1	<i>Candida auris</i>	36
4.3.2	<i>Candida metapsilosis</i>	36
4.3.3	<i>Candida orthopsilosis</i>	36
4.3.4	<i>Candida parapsilosis</i>	40
4.3.5	<i>Candida verbasci</i>	40
4.3.6	<i>Saprochaete ingens</i>	40
4.3.7	<i>Schizosaccharomyces pombe</i>	44
5	Diskusia	45
	Záver	49
	Príloha B	55

Úvod

Centroméry sú funkčnými elementami eukaryotických chromozómov, ktoré hrajú významnú úlohu v bunkovom delení. Zlyhanie ich funkcie môže byť pre bunku fatálne a preto je skúmanie centromerickej štruktúry a funkcie dôležité. V tejto práci sa budeme venovať identifikácii centromerického lokusu pomocou bioinformatickej analýzy genómových sekvencií kvasiniek.

Kvasinky je označenie pre jednobunkové huby predstavujúce ideálne modelové organizmy pre modernú experimentálnu biológiu. Gény a proteíny kvasiniek sú natoľko konzervované, že je možné mnohé poznatky aplikovať aj na ostatné eukaryotické organizmy, vrátane človeka. Vďaka kvasinkám je známa veľká časť základných biologických procesov, molekulárnych mechanizmov a biomakromolekulových interakcií v bunkách.

Význam týchto húb podčiarkuje aj počet Nobelových cien získaných za výskum kvasiniek. Nobelovu cenu za fyziológiu alebo medicínu získal v roku 2016 Yoshinori Ohsumi za objav mechanizmu autofágie v kvasinkách *Saccharomyces cerevisiae*. Nobelova cena z roku 2013, ktorú získal Randy W. Schekman, je za objavenie mechanizmu regulujúceho transport a sekréciu proteínov v kvasinkových bunkách. Ďalšiu nobelovu cenu získali vďaka výskumu *S. cerevisiae* v roku 2001 Leland Hartwell, Tim Hunt a Paul M. Nurse. Boli ocenení za prínosy poznatkov v oblasti bunkového cyklu.

Jednou z hlavných výhod využívania kvasiniek vo výskume sú nízke náklady na kultiváciu, jednoduchá genetická manipulácia a ich relatívne malé genómy, čo je aj hlavným dôvodom, prečo boli v tejto práci genómové sekvencie kvasiniek zvolené na bioinformatickú identifikáciu centromér.

Genómy kvasiniek sú variabilné a hoci v niektorých (napríklad *S. cerevisiae*) je centroméra podrobne charakterizovaná, v iných ešte preskúmaná nie je. Centroméry je možné identifikovať rôznymi experimentálnymi, ale aj bioinformatickými metódami, napríklad na základe nukleotidového zloženia genómovej sekvencie.

Cieľom tejto práce bolo implementovať program v Python skripte, ktorý predikuje centromerické lokusy v sekvenciách kvasinkových genómov na základe zastúpenia guanínu a cytozínu na tretích pozíciách kodónov (GC3 metóda) pomocou bioinformatickej analýzy. Na kvasinkách, pre ktoré boli predikované a experimentálne potvrdené centromerické lokusy, bola overená funkčnosť našej implementácie. Nasledovala optimalizácia parametrov a nakoniec bola naša implementácia GC3 metódy aplikovaná na genómové

sekvencie kvasiniek, pre ktoré doposiaľ centromerické lokusy nájdené neboli.

Prvá časť tejto práce obsahuje predstavenie základných genetických pojmov, funkcie a štruktúry DNA, spôsobu realizácie genetickej informácie, procesu bunkového delenia a funkcie centromér v ňom. Taktiež sú v nej uvedené bližšie informácie o centromérach ľudských a kvasinkových chromozómov.

Nasledujúca časť zahŕňa vysvetlenie GC3 metódy, informácie o databázach a dátach, Python balíčkoch použitých na analýzu a tiež interpretáciu získaných výsledkov a porovnanie výsledkov s podobnými výskumami. Nakoniec je zhodnotená aplikovateľnosť tejto metódy a perspektívy nadväzujúceho výskumu.

Kapitola 1

Základné genetické pojmy

1.1 Genetická informácia a jej interpretácia

Genetická informácia je uložená v molekulách DNA, ktoré majú štruktúru dvojzávitnicovitých špirál a v eukaryotických organizmoch sa nachádzajú v jadre. V eukaryotickej bunke je možné nájsť aj DNA nachádzajúce sa v semiautonómnych organelách, napríklad v mitochondriách a plastidoch.

Genetická informácia je reprezentovaná poradím nukleotidov v nukleových kyselinách. Nukleotid sa skladá z troch častí: fosfátová zložka, sacharidová zložka a dusíkatá báza. Sacharidovou zložkou môže byť ribóza alebo deoxyribóza, ktorá je súčasťou RNA, resp. DNA. Dusíkaté bázy patria do dvoch skupín: puríny, medzi ktoré sa zaraďuje adenín (A) a guanín (G), a pyrimidíny - cytozín (C), tymín (T). Puríny a pyrimidíny sú si navzájom komplementárne G:C a A:T. V RNA je tymín nahradený pyrimidínom uracil (U) [33].

Jednou z funkcií DNA je uchovávanie informácie o poradí aminokyselín v proteínoch. Proteínogénnych aminokyselín je 20 druhov. Každá aminokyselina je kódovaná trojicou báz - kodónom. Zo štyroch druhov báz (A, C, G, T) je teda možné zostaviť 64 rôznych kodónov. Jednu aminokyselinu môže kódovať viacero tripletov, pričom ich frekvencia v genóme nie je rovnomerná. Princíp priradenia aminokyselín ku kodónom sa nazýva genetický kód (Obr. 1.1). V štandardnom genetickom kóde sa nachádza niekoľko špeciálnych typov kodónov, ktorými sú štart-kodón: ATG, ktorý kóduje metionín a stop-kodóny: TAA, TAG a TGA, ktoré nekódujú žiadnu aminokyselinu [33].

1.2 Nukleotidové zloženie DNA

Zastúpenie párov GC a AT v sekvenciách nukleových kyselín nie je rovnomerné a môžu ho ovplyvniť replikácia, rekombinácia a ďalšie rôzne genetické procesy, ktorých dôsledkami sú mutácie.

Štandardný genetický kód					
1. poloha (5'-koniec)	2. poloha				3. poloha (3'-koniec)
	U	C	A	G	
U	Phe	Ser	Tyr	Cys	U
	Phe	Ser	Tyr	Cys	C
	Leu	Ser	STOP	STOP	A
	Leu	Ser	STOP	Trp	G
C	Leu	Pro	His	Arg	U
	Leu	Pro	His	Arg	C
	Leu	Pro	Gln	Arg	A
	Leu	Pro	Gln	Arg	G
A	Ile	Thr	Asn	Ser	U
	Ile	Thr	Asn	Ser	C
	Ile	Thr	Lys	Arg	A
	Met	Thr	Lys	Arg	G
G	Val	Ala	Asp	Gly	U
	Val	Ala	Asp	Gly	C
	Val	Ala	Glu	Gly	A
	Val	Ala	Glu	Gly	G

Obr. 1.1: Štandardný genetický kód

Všeobecne je známe, že v neexprimovaných sekvenciách a intrónoch dochádza k mutáciám častejšie, pretože ich vplyv na fungovanie organizmu nie je taký výrazný oproti kódujúcim častiam genómu a teda mutácia má väčšiu pravdepodobnosť prenosu do ďalšej generácie. Spomedzi zmien v génoch kódujúcich proteíny majú v evolúcii najvyššie zastúpenie zmeny na tretích pozíciách kodónov. Vyplýva to z genetického kódu, kde je možné vidieť, že zmena v niektorej z prvých dvoch báz väčšinou vedie k zmene druhu výslednej aminokyseliny, kým zmena tretej bázy je väčšinou synonymná [33].

V zmenách zastúpení nukleotidov hrá významnú úlohu podmienená génová konverzia (ďalej len gBGC - angl. *GC-biased gene conversion*) [15]. Ide o jav, kedy sa zvyšuje zastúpenie guanínu a cytozínu v evolučnom čase. Génové konverzie môžu prebiehať počas opravy dvojvláknových zlomov alebo počas rekombinácie. U mnohých druhov je správna segregácia chromozómov počas meiózy podmienená výskytom aspoň jedného crossing-over na chromozóme. Frekvencia rekombinácie je so zvyšujúcou sa dĺžkou chromozómu vyššia a zastúpenie GC priamo koreluje s rýchlosťou rekombinácie. To má za následok lokálne regionálne variácie v obsahu GC v genóme, pretože miera rekombinácie sa pozdĺž genómu mení. V oblastiach v blízkosti centromér býva rýchlosť rekombinácie nižšia, preto aj zastúpenie guanínu a cytozínu na tretích pozíciách kodónov (ďalej len GC3) by malo byť menšie [25]. Príkladom gBGC môže byť situácia, v ktorej bázy nie sú komplementárne - G:T. Podľa gBGC je pravdepodobnejšie, že sa T zmení na C a

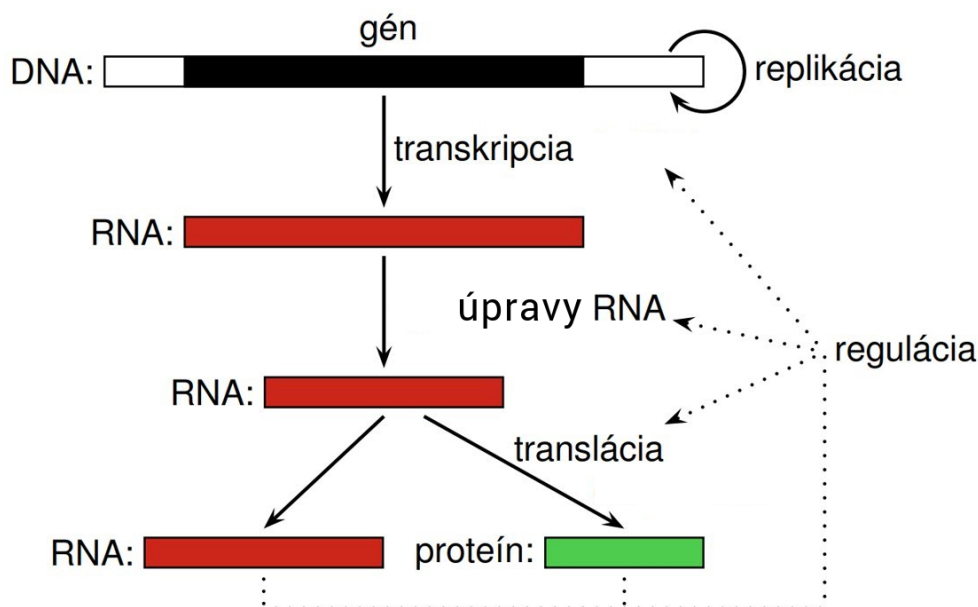
teda vznikne pár G:C. Vznik páru A:T bude menej pravdepodobný. Pravdepodobnosť fixácie alel GC je teda vyššia ako AT.

1.3 Realizácia genetickej informácie

Realizácia genetickej informácie prebieha v niekoľkých základných krokoch popísaných v centrálnej dogme molekulárnej biológie (Obr. 1.2), ktorú sformuloval Francis Crick v roku 1953 [37]. V eukaryotických bunkách prebieha realizácia genetickej informácie nasledovne [33]:

- Replikácia DNA je proces, pri ktorom z jednej molekuly DNA vzniknú dve identické molekuly DNA. V každej z novovzniknutých molekúl bude jedno vlákno z pôvodnej DNA a druhé novosyntetizované. V tomto procese je zapojených viacero enzýmov, pričom najdôležitejšiu funkciu plní DNA polymeráza.
- Transkripcia DNA je prvým krokom proteosyntézy. V tomto procese sa prepisom DNA tvoria viaceré typy RNA pomocou enzýmu RNA polymeráza. V eukaryotických bunkách sa nachádzajú napríklad mRNA, tRNA, rRNA, snRNA, snoRNA, miRNA, siRNA, piRNA, lncRNA, ncRNA.
- Posttranskripčné modifikácie sú úpravy RNA. V prípade eukaryotickej pre-mRNA sa počas úpravy koncov na 5'-konci sa vytvorí tzv. *čiapka* a 3'-koniec je polyadenylovaný. Prebehne zostrih, angl. *splicing*, kedy sa nekódujúce časti génov (intróny) vystrihnú a kódujúce úseky (exóny) sú enzymaticky pospájané. Výsledkom posttranskripčných modifikácií je zrelá mRNA kódujúca proteíny, ktorá cez jadrový pór vychádza z jadra do cytoplazmy. Súčasťou posttranskripčných modifikácií sú aj chemické zmeny RNA, napríklad metylácia alebo výmena uridínu za pseudouridín či dihydrouridín.
- Translácia je dej, pri ktorom je informácia z mRNA prekladaná pomocou genetického kódu do poradia aminokyselín v polypeptidových reťazcoch. Na tomto procese sa podieľajú aj tRNA a rRNA. tRNA (transferová RNA) prenáša aminokyseliny na miesto syntézy proteínov. rRNA (ribosomálna RNA) v komplexe s proteínmi vytvára ribozómy, na ktorých prebieha translácia mRNA. Druh aminokyseliny určuje kodón v mRNA, ku ktorému je komplementárny antikodón v tRNA nesúcej príslušnú aminokyselinu. Translácia prebieha na ribozómoch, kde sa po naviazaní tRNA na mRNA vytvorí peptidová väzba medzi aminokyselinou na tRNA a vznikajúcim polypeptidom, pričom preruší väzbu s tRNA. Po vytvorení celého syntetizovaného reťazca aminokyselín sa polypeptid odpojí od ribozómu. Už počas syntézy prebieha skladanie bielkovín do ich trojrozmernej štruktúry nazývané *foldng*.

- Posttranslačné modifikácie zahŕňajú rozličné úpravy proteínov s cieľom regulovať ich biochemické funkcie. Ide o procesy proteolytického štiepenia, adície funkčných skupín alebo celkovú degradáciu proteínu. Príkladom môžu byť fosforylácia, glykozylácia alebo metylácia.



Obr. 1.2: Centrálna dogma molekulárnej biológie. Zdroj: <http://compbio.fmph.uniba.sk/vyuka/mbi/index.php/%C3%9Avod>

1.4 Delenie bunky

Bunkový cyklus reprezentuje život bunky od vzniku až po jej rozdelenie na dve dcérske bunky. Skladá sa z dvoch hlavných častí: interfáza a mitóza. Interfáza sa delí na [33]:

1. G1 fáza nasleduje po G0 alebo po cytokinéze a po nej nasleduje S fáza. Preto je občas označovaná aj ako postmitotická alebo presyntetická. G1 je zväčša časovo najdlhšou fázou bunkového cyklu. Počas nej dochádza k syntéze enzýmov potrebných pre replikáciu DNA a ribozómov. Hlavný kontrolný uzol medzi G1 a S fázou overuje, v akom stave je DNA a či je bunka pripravená na delenie. Zabráni sa tým replikácii poškodenej genetickej informácie. V tomto bode teda môže prísť k prerušeniu bunkového cyklu.
2. Ďalej nasleduje S fáza, v ktorej prebieha replikácia DNA, preto sa nazýva syntetická fáza. Na konci S fázy každý chromozóm diploidnej somatickej bunky pozostáva z dvoch identických chromatíd a bunka bude tetraploidná. V prechode medzi S a G2 fázou sa opäť skontroluje DNA - či replikácia prebehla v poriadku a vykonajú sa prípadné opravy.

3. G2 fáza, inak nazývaná aj postsyntetická alebo premitotická. Je zo všetkých fáz najkratšia. V ribozónoch sa počas G2 fázy syntetizuje bielkovina tubulín, ktorá hrá kľúčovú úlohu pri naväzovaní chromozómu na deliace vretienko.

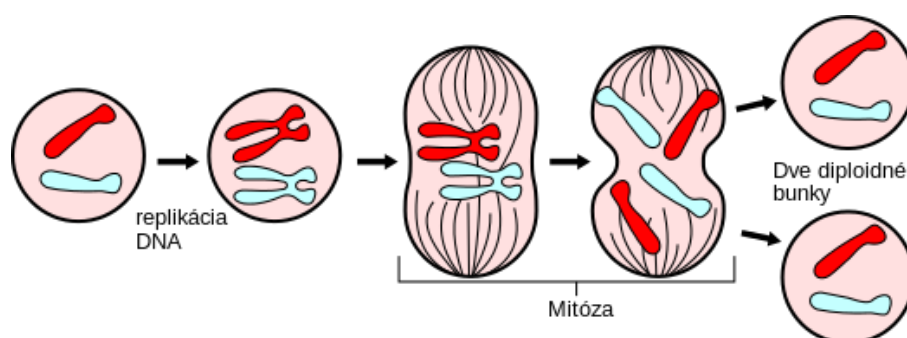
Po G2 fáze prebieha mitóza (M fáza, Obr. 1.3), kedy sa delí jadro (karyokinéza), pričom počet chromozómov sa vzhľadom na pôvodnú diploidnú bunku zachováva. M fáza pozostáva zo štyroch krokov [33]:

1. Profáza, v ktorej sa z centrozómov (koordinačné centrum pre mikrotubuly) začínú rozchádzať vlákna mikrotubulov k protiľahlým pólom bunky, čím sa vytvorí deliace vretienko. Ďalej sa začne narušovať jadrová membrána a mikrotubuly deliaceho vretienka sa začínú pripájať na chromozómy. Tam každú chromatídu naviažu v oblasti centroméry na kinetochór. Následne sa chromozómy začínú priestorovo preusporiadávať a špiralizovať.
2. V metafáze sú chromozómy už plne špiralizované a usporiadané do ekvatoriálnej roviny v strede bunky. V tomto momente bunka prejde tzv. mitotickým kontrolným uzlom, kde sa overí, či je každý chromozóm naviazaný na mikrotubuly deliaceho vretienka.
3. V anafáze po naviazaní všetkých chromatíd príde k odbúraniu proteínu, ktorý spája chromatídy. Zároveň sa kinetochórové mikrotubuly skracujú, čo ďalej spôsobí, že chromatídy sú od seba odťahované k opačným pólom bunky.
4. V telofáze sa bunka predlžuje a chromozómy sú už takmer pri pólach deliaceho vretienka. Tam sa okolo nich začne vytvárať jadrová membrána a chromozómy začínú dekonenzovať.

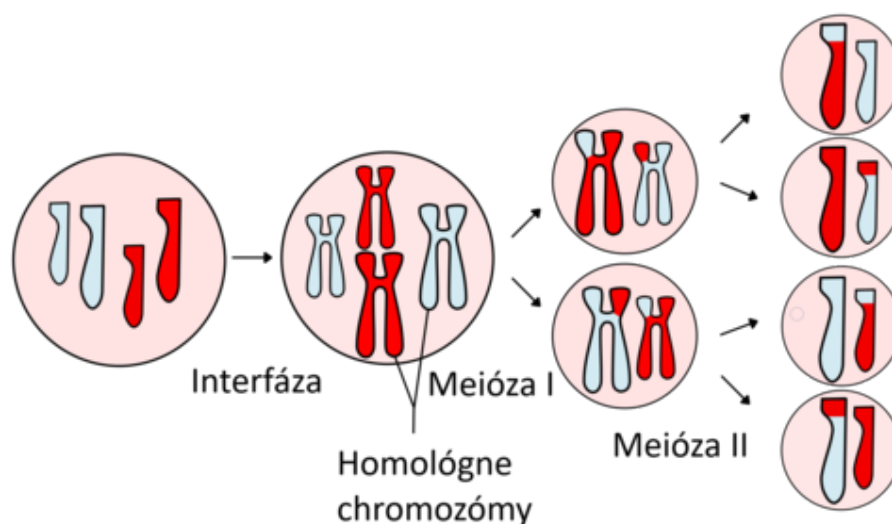
Bezprostredne po telofáze mitózy prebehne cytokinéza, v ktorej sa delí materská bunka na dve dcérske bunky. Následne sa bunka môže dostať opäť do G1 fázy alebo do G0, ktorá zodpovedá pokojovému štádiu bunky. V G0 sa nachádza väčšina somatických buniek a počas tohto obdobia vykonávajú určitú úlohu závislú od ich diferenciácie. Táto fáza je reverzibilná a teda bunka môže opäť vstúpiť do bunkového cyklu.

Ďalším typom bunkového delenia je meióza (Obr. 1.4) [33]. Možno ju charakterizovať aj ako dve po sebe nasledujúce mitotické delenia - heterotypické (meióza I) a homeotypické (meióza II). Výsledkom meiózy sú štyri dcérske bunky s redukovaným počtom chromozómov na polovicu.

Pri meióze I dochádza k redukcii počtu chromozómov. Podobne ako mitóza je rozdelená na štyri fázy. V profáze prebieha crossing-over, teda výmena homologických častí chromozómov a následne sa chromatídy oddeľujú. Metafáza a anafáza prebiehajú veľmi podobne ako pri mitóze. Telofáza mitózy a meiózy I sa odlišuje v tom, že pri meióze I sú v bunke chromozómy tvorené dvoma chromatídami. Následne začína profáza



Obr. 1.3: Zdroj: <https://sk.wikipedia.org/wiki/Mit%C3%B3za>



Obr. 1.4: Zdroj: <https://sk.wikipedia.org/wiki/Mei%C3%B3za>

meiózy II a teda sú potrebné dve deliace vretienka. Počas metafázy sa chromozómy usporadúvajú do ekvatoriálnej roviny a viažu sa na ne mikrotubuly. Anafáza a telofáza homeotypického delenia sú analogické k anafáze a telofáze mitózy.

1.5 Štruktúra a funkcia chromozómu

Jadrová DNA v komplexe s proteínmi a RNA vytvára chromatín. Základnou stavebnou jednotkou chromatínu sú nukleozómy. Jadro nukleozómu tvorí oktamér histónov H2A, H2B, H3 a H4, okolo ktorých je obtočená časť vlákna DNA. Nukleozómové vlákno s priemerom 11 nm a histón H1 spájajúci susedné nukleozómy spôsobuje ďalšiu špiralizáciu chromazínu. Ďalšou špiralizáciou vznikajú chromozómy.

Chromozóm je tvorený dvoma chromatídami, kde proteín kohezín spája dve sesterské chromatídy a kondenzín zodpovedá za správne zbaľovanie chromatínu do vyšších foriem. Základnými funkčnými elementami chromozómov eukaryotických organizmov sú: počiatky replikácie, centroméry a teloméry [33].

Počiatky replikácie sú miesta v molekulách DNA, ktoré sú rozoznávané komplexami

proteínov replikačného aparátu, kde sa po ich naviazaní spustí replikácia.

Centroméra je oblasť chromozómu, ktorá je v bunkách najlepšie viditeľná metafáze bunkového delenia ako miesto, v ktorom sa spájajú dve sesterské chromatídy. Plní dôležitú funkciu pri rovnomernej segregácii chromozómov do dcérskych buniek.

Na koncoch ramien chromozómov sa nachádzajú ich koncové časti - teloméry. Teloméry zabráňujú fúzii koncov chromozómov, čím chromozóm chránia. Vo väčšine somatických buniek sa veľkosť telomér každým bunkovým delením počas replikácie skracuje, čo súvisí so starnutím bunky. V zárodočných bunkových líniách je aktívny enzým telomeráza, ktorý pomáha zachovávať stabilnú dĺžku telomér. Tento enzým je aktívny aj v kmeňových bunkách kostnej drene a aj v bunkách väčšiny nádorov [33].

1.6 Centroméry

Centroméra je tiež označovaná ako miesto primárnej konstriktie na chromozóme. Sesterské chromatídy sú spojené v oblasti centroméry proteínmi - kohezínmi. Je to špecializovaná štruktúra na chromozómoch eukaryotických organizmov, ktorá hrá podstatnú úlohu v mitotickom aj meiotickom delení.

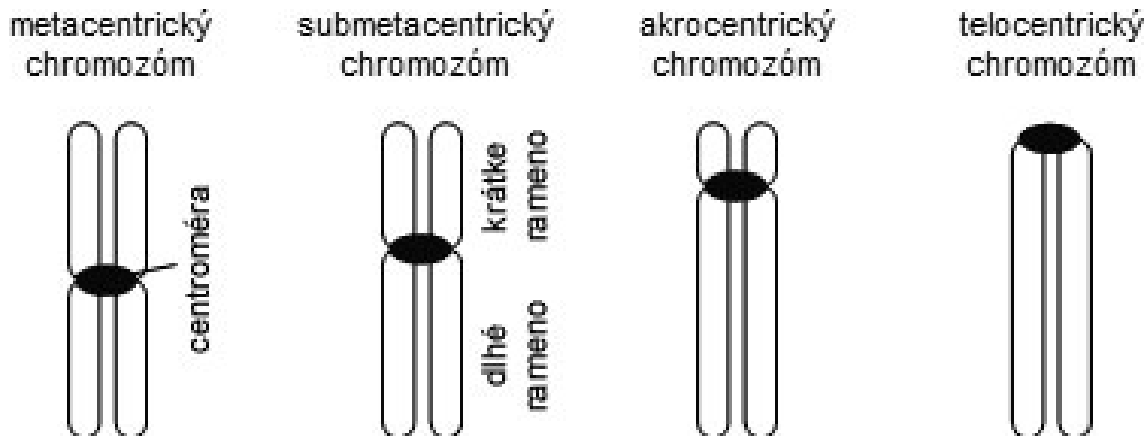
Na centroméru je pripojený kinetochór. Je to proteínový komplex, ktorý pomáha naviazaniu chromozómu na mikrotubuly deliaceho vretienka a zodpovedá za pohyb chromozómov k pólom deliacej sa bunky. Centroméra tiež zabezpečuje, aby sa zduplikovaný chromozóm rozdelil rovnomerne. Je to dôležitý predpoklad k tomu, aby vznikli zdravé dcérske bunky [33].

Zlyhanie funkcie centroméry alebo deliaceho vretienka počas bunkového delenia je pomerne signifikantné a môže viesť ku genetickým ochoreniam. Typickým príkladom je *Downov syndróm*, čo je neliečiteľná genetická porucha, ktorá sa prejavuje oneskoreným mentálnym vývojom, oblou tvárou so šikmými očami, menšou ústnou dutinou a vysokým podnebí, čo súvisí s poruchou reči. U ľudí s týmto ochorením sa chromozóm číslo 21 v meióze rozdelil nerovnomerne a po oplodnení prišlo k vytvoreniu bunky s tromi kópiami chromozómu 21 (trizómia 21) [12]. Podobným príkladom je *Edwardsov syndróm*, pri ktorom sa prejavujú defekty obličiek a srdca, ťažkosti s dýchaním, spomalený rast, či abnormality hlavy. Ide o trizómiu chromozómu 18 [4]. Poznanie funkcie centroméry je teda kľúčové pre hlbšie porozumenie mnohým genetickým ochoreniam.

Centroméra sa nemusí nachádzať v strede chromozómu. Podľa pozície centroméry rozdelíme niekoľko typov chromozómov (Obr. 1.5) [3]:

- metacentrický chromozóm - centroméra sa nachádza takmer v strede a ramená sú približne rovnako dlhé
- submetacentrický chromozóm - dlhé rameno je dlhšie o nanajvýš polovicu krátkeho ramena

- akrocentrický chromozóm - krátke rameno je výrazne (viac ako o polovicu) kratšie ako dlhé rameno
- telocentrický chromozóm - centroméra sa nachádza na konci chromozómu



Obr. 1.5: zdroj: <https://biopedia.sk/genetika/chromozomove-mutacie>

Centroméra pozostáva z jadra centroméry a pericentromérnych oblastí. Jadro centroméry je definované ako miesto pripojenia pre kinetochór a pericentroméry zabezpečujú pevnosť a súdržnosť dvoch chromatíd.

Pre jadrá centromér je typický centromerický chromatín. Je špecifický tým, že je obohatený o variant histónu H3, konkrétne CENP-A. Ide o epigenetickú značku, ktorá je nevyhnutná pre životaschopnosť buniek. Bolo preukázané, že poloha centroméry je špecificky udržiavaná a identifikovaná centromerickým chromatínom viazaným CENP-A. Tento histón sa zároveň podieľa na vytváraní kinetochórového komplexu počas mitózy. Vďaka špecifickej trojrozmernej štruktúre je chromatín obsahujúci CENP-A na povrchu kondenzovaného chromozómu a tým je umožnená interakcia s kinetochórom a mikrotubulami deliaceho vretienka. Zároveň tu počas segregácie vzniká napätie, vďaka ktorému sú dcérske chromatídy ťahané k opačným pólom bunky [2].

Centroméry sú ohraničené transkripčne neaktívnym chromatínom. Pericentroméra je tvorená pericentrickým chromatínom spolu s molekulami kohezín a kondenzín. Táto štruktúra zabezpečuje mechanickú odolnosť centromér. Pri prechode do anafázy mitózy, je kohezín proteolyticky štiepený enzýmom separáza, chromatídy sa uvoľnia a migrujú na protiľahlé póly bunky.

U eukaryotických organizmov sa rozlišuje niekoľko typov centromér: Bodové centroméry sa skladajú z jedného nukleozómu a vyskytujú sa napríklad u púciacich kvasiniek *S. cerevisiae*. Kvasinka *Schizosaccharomyces pombe* má regionálne centroméry, ktoré sú tvorené viacerými nukleozómami. Satelitné centroméry typické pre človeka sú založené na tandemových opakovaniach. Holocentroméry sú roztrúsené po celej dĺžke chromozómu a nachádzajú sa napríklad v bunkách červa *Caenorhabditis elegans* [27].

Chromozómy môžu obsahovať viacero oblastí s charakteristickou centromerickou DNA, avšak vždy len jedna je miestom tvorby kinetochóru. Za istých okolností, keď pôvodná centroméra stratí funkciu, môžu vzniknúť nové centroméry - neocentroméry. Narozdiel od štandardných centromér, neocentroméry neobsahujú repetitívne sekvenencie. Viacero centromér na jednom chromozóme sa v evolúcii mohlo objaviť pri spájaní chromozómov, no následne bola funkcia jednej z nich potlačená [9].

1.6.1 Centroméry v ľudských bunkách

V ľudských bunkách sa nachádza 23 párov chromozómov. V každom páre sa nachádza jeden chromozóm od matky a druhý od otca. 22 párov je označovaných ako autozómy a jeden pár ako pohlavné chromozómy (X a Y). Veľkosť ľudského jadrového genómu je 3,05 Gbp, pričom obsahuje 63 494 génov, z toho 19 969 kódujú proteíny. Veľkosť mitochondriálneho genómu je približne 16,6 kbp. Celkovo 2% DNA tvoria regióny kódujúce proteíny, 24% intróny a iné nekódujúce časti génov, 15% unikátna nekódujúca DNA a až 59% tvorí repetitívna DNA. V ľudských bunkách je počas replikácie aktivovaných viac ako 30 000 počiatkov replikácie, rozmiestnených približne každých 100 kbp. U ľudí je celkovo transkribovaných 80% genómu [21].

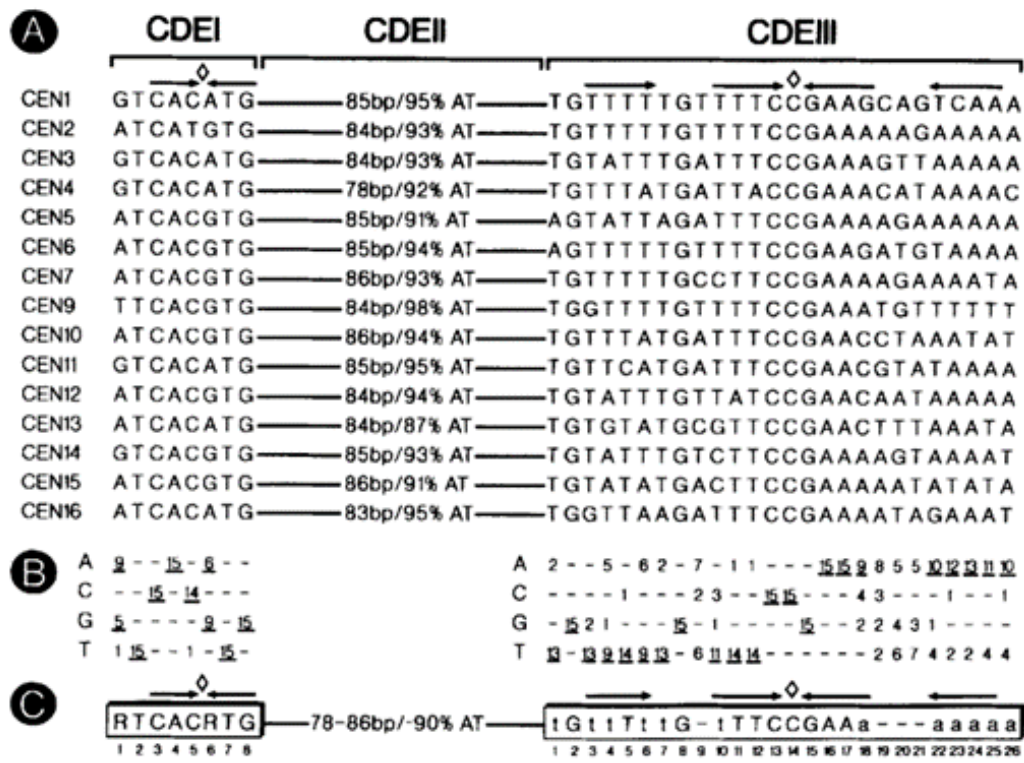
Centroméry v ľudských chromozómoch sú lokalizované v regiónoch bohatých na A-T páry s α -satelitnou DNA, ktorá je definovaná ako repetitívna oblasť chromozómu pozostávajúca z monomérov s dĺžkou 171 bp. Jedna centroméra môže obsahovať viacero odlišných α -satelitných DNA. Celkovo tvoria približne 5% genómu. Centroméry jednotlivých ľudských chromozómov sa líšia počtom a usporiadaním týchto monomérov. Jednotlivé monoméry sú sekvenčne identické na 50-70%. Na základe podobnosti rôznych typov monomérov ich klasifikujeme do suprachromozomálnych rodín. Rodiny 1 a 2 sa vyznačujú dimérnou periodicitou a rodina 3 pentamérnou.

Repetície v α -satelitnej DNA výrazne ovplyvňujú presnosť skladania súvislých sekvencií genómu, obzvlášť pri krátkych čítaniach, preto je pomerne ťažké analyzovať tieto oblasti. Z existujúcich sekvenčných zostáv okolo centroméry sú najlepšie charakterizované pericentromérne oblasti [21].

1.6.2 Centroméry v kvasinkových bunkách

Veľkosť genómu kvasinky *S. cerevisiae* je 12,07 Mb, ktorý je usporiadaný v 16 chromozómoch [23].

DNA v oblastiach centromér sa označuje ako CEN DNA. CEN DNA kvasiniek *S. cerevisiae* pozostáva z približne 120 báзовých párov. Pritom ju rozdelujeme na tri časti: CDEI, CDEII a CDEIII, ktoré sú v rámci rodu *Saccharomyces* vysoko konzervované (Obr. 1.6). CDEI pozostáva z 8 báзовých párov, z ktorých šesť je v tvare palindrómu. Viaz sa naň proteín s názvom Cbf1. CDEII sa nachádza medzi CDEI a CDEIII. Jej



Obr. 1.6: Rozdelenie centroméry *S. cerevisiae* na CDEI, CDEII a CDEIII [11].

veľkosť sa pohybuje okolo 83 básových párov, pričom viac ako 90% párov je A-T. Delécie v tejto oblasti viedli k úplnej strate funkcie centroméry. Celková dĺžka CDEIII je 26 bp, pričom je palindromatická medzi jedenástou a sedemnástou pozíciou. Viaz sa naň komplex Cbf3, ktorý je pre funkciu centroméry esenciálny. Zaujímavosťou je, že poloha, resp. orientácia CDEIII voči CDEI a CDEII môže potlačiť funkciu centroméry. Dokázalo sa, že efekt mutácii v CDEI a CDEII nie je až taký závažný ako pri CDEIII. Dodnes je však o presných funkciách CDE v CEN DNA kvasiniek známe len veľmi málo [11].

Centroméra bola prvýkrát funkčne charakterizovaná práve u kvasinky *S. cerevisiae*. Bola opísaná ako časť chromozómu schopná poskytnúť mitotickú a meiotickú stabilitu. Táto oblasť chromozómu je viazaná na práve jeden nukleozóm a teda tvorí bodový centroméru. Izolovaná bola v roku 1980 výskumníkmi Louiseom Clarkom a Johnom Carbonom pomocou tzv. *funkčného testu* [6].

Základom funkčného testu sú plazmidy. Sú to relatívne krátke cirkulárne dvojvláknové molekuly DNA. Prvou generáciou umelo-vytvorených plazmidov je YIp (angl. *Yeast Integrative plasmid*). Tento plazmid však nie je schopný replikácie a môže byť integrovaný do genómu. Experimentálnymi metódami boli doň vkladane rôzne náhodné sekvencie kvasinkového genómu, vďaka čomu vznikol *kvasinkový replikatívny plazmid* (YRp), ktorý sa v bunkách vyskytuje vo veľkom počte kópii. YRp obsahuje autonómne sa replikujúce sekvencie (ARS) z kvasinkového chromozómu. Plazmidy, ktoré obsahujú

ARS (počiatky replikácie), sa môžu replikovať nezávisle od chromozomálnej DNA v jadre. YRp má nízku mitotickú stabilitu, takže nie je schopný rovnomernej segregácie pri bunkovom delení. Následne bola podobným spôsobom experimentálne identifikovaná sekvencia centroméry, vďaka čomu vznikol *kvasinkový centromerický plazmid* (YCp). YCp sa vyskytuje v nižšom počte kópií, typicky jedna až dve na bunku. Obsahuje ARS spolu s centromerickou sekvenciou (CEN). Na rozdiel od YRp, YCp sú stabilné počas delenia [29].

Vo funkčnom teste bolo možné pozorovať, že kvasinky s YCp vykazovali najvyššiu schopnosť zachovať si plazmid pri delení (viac ako 90%). Z toho vyplýva, že na správny priebeh mitózy je potrebný ako ARS, tak aj centromerická časť chromozómu.

Spôsobov, ktorými je možné nájsť centroméry je omnoho viac a neustále sa vyvíjajú nové. Medzi experimentálne spôsoby hľadania centromér patrí metóda, ktorá využíva proteínový komplex Cbf3 (angl. *Centromere Binding Factor*) a fluorescenčné proteíny. Tento komplex bol izolovaný a purifikovaný z kvasiniek *S. cerevisiae*. Cbf3 sa špecificky viaže na CEN DNA, rozoznáva CDEIII v centromére, no neviaže sa na nefunkčnú CEN DNA. Tento komplex sa podieľa na tvorbe ďalších zložiek kinetochóru, ako napríklad nukleozómov. Po naviazaní fluorescenčne označeného Cbf3 na oblasť centroméry, vieme určiť jej pozíciu [38].

Metóda *Centurion* sa opiera o metódu Hi-C, ktorá zachytáva konformácie chromatinu a využíva masívne paralelné sekvenovanie. Autori ukázali, že *Centurion* funguje správne aj pri obmedzenej hĺbke sekvenovania Hi-C knižníc. Metóda *Centurion* predpovedá pozíciu s presnosťou 89% do 5 kb u kvasiniek. Je aplikovateľná aj na iné mikroorganizmy, pre ktoré štandardné metódy na identifikáciu centromér zlyhávajú [35].

Ďalšou metódou je bioinformatická GC3 metóda [18]. Vychádza z faktu, že zastúpenie guanínu a cytozínu na tretích pozíciách kodónov je v oblasti centromér nižšia. Touto metódou sa budeme zaoberať v tejto práci.

Kapitola 2

Ciele práce

Táto práca má tri hlavné ciele:

1. Implementácia a verifikácia metódy na identifikáciu centromerických lokusov na základe zastúpenia GC3 v genómoch kvasniiek *Clavispora lusitaniae*, *Scheffersomyces stipitis* a *Yarrowia lipolytica*. Pre tieto tri kvasinky bola predikovaná pozícia centroméry vo východiskovom článku [18] a zároveň bola experimentálne overená [5]. Kým prvé dve spomenuté používajú alternatívny genetický kód (CTG(Ser1) vetva), *Y. lipolytica* používa štandardný genetický kód.
2. Zistenie vplyvu parametrov v používanom algoritme a ich optimalizácia. Parametre zahŕňajú veľkosť okna a vylúčenie kodónov ATG a CTG.
3. Aplikácia GC3 metódy na nové kvasinkové genómové sekvencie, napríklad *Candida auris*, *Candida metapsilosis*, *Candida orthopsilosis*, *Candida parapsilosis*, *Candida verbasci*, *Saprochaete ingens* a *Schizosaccharomyces pombe*.

Kapitola 3

Metodika

3.1 Východiskové publikácie

Vo väčšine eukaryotov sa zastúpenie guaníu a cytozínu na tretích pozíciách kodónov mení pozdĺž celého chromozómu. Tento jav bol prvýkrát opísaný v roku 1993 britskými výskumníkmi Paulom M. Sharpom a Andrewom T. Lloydom [32]. Už je známe, že v kvasinkách *S. cerevisiae* miesta s vysokým zastúpením GC3 korelujú s miestami, kde často prebieha rekombinácia. Z toho logicky vyplýva, že miesta s nízkym GC3 korelujú s nízkou mierou rekombinácie a teda vzniká hypotéza, že miesta na chromozóme s nízkym GC3 súvisia s centromérmi, okolo ktorých býva rekombinácia potlačená.

Zastúpenie GC3 samozrejme súvisí s celkovým zastúpením guanínu a cytozínu v molekulách DNA. Celkové zastúpenie G+C v genómoch kvasiniek varíruje v rozmedzí od 22% až do 49%. Medzi kvasinky s najnižším zastúpením GC patrí *C. tropicalis* a s najvyšším *C. lusitaniae*. Kvasinky so zastúpením GC pohybujúcim sa okolo 35% vykazujú menšiu premenlivosť zastúpenia GC pozdĺž genómu ako kvasinky s GC pod 30%, resp. nad 40% [18].

Autori pionierskej štúdie Denise B. Lynchová a kol. (2010) [18] ukázali, že v genome kvasinky *Y. lipolytica* je na každom chromozóme jedna oblasť s výrazne nižším zastúpením GC3 ako inde, pričom sa táto pozícia zhoduje s experimentálne identifikovanými centromérmi. Zamerali sa aj na to, ako sa kvasinky z rodu *Saccharomyces* líšia od CTG(Ser1) vetvy. Z CTG(Ser1) vetvy im u *C. lusitaniae* a *S. stipitis* vyšli lokalizácie centromér v GC3 chudobných oblastiach pomerne jednoznačne. Hypotézou autorov bolo, že centromerická oblasť *C. lusitaniae* a *S. stipitis* podlieha mutačným zmenám a v evolúcii sa stala chudobnou na GC3.

Zaujímavosťou, ktorú tiež zistili je, že medzi kvasinkami *Candida albicans* a *Candida dubliniensis*, ktoré sú v CTG(Ser1) vetve, ostávajú zachované pozície minimálnych a maximálnych hodnôt GC3 napriek tomu, že sa v nich vyskytuje veľké množstvo nukleotidových substitúcií. Vedie to ďalej k záverom, že miesta s vysokou frekvenciou

rekombinácii ostávajú medzi druhmi zachované [18].

Na prácu Denise Lynchovej a kol. (2010) [18] naviazala Marie-Claude Marsolier-Kergoat, ktorá tiež objasňuje tému spojenú s centromérmi a zastúpením guanínu a cytozínu na tretích pozíciách kodónov [19]. V tomto článku vychádzajú najavo dôvody, kvôli ktorým zastúpenie G+C vyzerá inak v niektorých kvasinkách (napríklad *C. albicans*) a nie je na ne možné aplikovať metódu uvedenú vyššie. Autorka vychádza z predpokladov, že gBGC v meiotickej rekombinácii je najpravdepodobnejšia príčina vzniku nerovnováhy v zastúpení G a C pozdĺž chromozómu. Tieto variácie sa však nevzťahujú na organizmy, ako napríklad *C. albicans* a *C. dubliniensis*. Pravdepodobne ide teda o gBGC v mitotickej rekombinácii a pri oprave poškodení DNA alebo k vytvoreniu nerovnováhy dochádza počas replikácie DNA v S fáze, kedy je koncentrácia dGTP a dCTP vyššia ako dATP a dTTP. V závere naznačila, že jednou z príčin môže byť selekcia používaných kodónov.

Na predošlé dve štúdie naviazali autori Mingfeng Cao a kol. (2017) [5]. Experimentálne dokázali výsledky získané GC3 metódou použitou v roku 2010 [18] na genómových sekvenciách niektorých kvasiniek. Do kvasinkových kultúr *S. stipitis* transformovali YRp. Počas delenia buniek bola segregácia nerovnomerná a plazmidy ostávali v rodičovských bunkách a po 48 hodinách kultivácie viac ako 70% populácie plazmidy neobsahovalo. Následne do plazmidov integrovali CEN chromozómu 5 nájdeného pomocou GC3 metódy [18]. Tieto plazmidy boli po kultivácii detegované až v 93% populácie [5]. Zistili, že v oblasti centromér *S. stipitis* nájdených pomocou GC3 metódy sa nachádza nezvyčajne dlhá intergénová oblasť a to v rozmedzí 14,6-38 kb. Tieto hodnoty sú signifikantné, nakoľko priemerná dĺžka ostatných intergénových oblastí v tejto kvasinke je 0,7-2 kb.

3.2 Dáta a databázy

Pre aplikáciu GC3 metódy v tejto práci je pre každý druh kvasinky potrebná nukleotidová sekvencia genómu vo formáte MULTIFASTA, spolu s anotáciou v GFF3 formáte. FASTA formát (Obr. 3.1) pozostáva z identifikátora a prípadných doplňujúcich informácií na samostatnom riadku začínajúcim znakom > spolu s názvom chromozómu, organizmu a ďalšími doplňujúcimi informáciami. Potom nasledujú riadky nanajvýš 80 znakov dlhé, v ktorých je samotná nukleotidová sekvencia. Okrem písmen označujúcich štyri základné bázy, môžu .fasta súbory obsahovať aj iné písmená, napríklad N označujúce akúkoľvek bázu. MULTIFASTA je súbor obsahujúci viacero záznamov vo FASTA formáte.

```
>NW_003101578.1 Clavispora lusitaniae ATCC 42720 scaffold_1 genomic scaffold
GTTTTCATTTATAGAATGATCTATATGGAGTCTCTAAAATGTCATTTCTGACGGTTGCACCAAGATATGGGTGAAGGT
TCAATTCCAAGACGGCGATACGTTCTGGTCAGACTAACAACTCAACTATGGAGTAGATTTTTTTGGGAGTATCGATGTCT
TCCACTGCTTAAGAAGAAATGTAAATGCTGGATTAAGGCAGATCCAAATATGTTGCCATCCATAGTGTGGCGAGGGTAC
```

Obr. 3.1: Časť .fasta súboru genómovej sekvencie kvasinky *C. lusitaniae*.

Súbor .gff (Obr. 3.3) je viazaný na .fasta súbor a obsahuje k nemu doplňujúce informácie. Jednotlivé polia sú oddelené tabulátorom. .gff súbor obsahuje názov sekvencie, čiže názov chromozómu alebo tzv. scaffold-u. Nasleduje zdroj, čiže názov databázy alebo projektu, z ktorej sekvencia pochádza. Potom charakterizácia danej sekvencie - či sa jedná o gény, exóny, CDS, mRNA, prípadne iné. Ďalšie dve polia zodpovedajú štartovacej a končiacej nukleotidovej pozícii danej sekvencie v .fasta. Nasleduje skóre, vlákno (+ ako vedúce, - ako zaostávajúce) a rámcové číslo, ktoré označuje, či je prvá báza prvou bázou kodónu. V ďalšom poli sú atribúty, ktoré sú oddelené bodko-čiarkou a obsahujú dodatočné informácie, napríklad jedinečné ID, typ génu, názov génu, pri proteín kódujúcich génoch ID proteínu a funkciu proteínu [28].

```
NW_003101575.1      RefSeq CDS      399      1046      .      -      0
ID=cds-XP_002615956.1; Parent=rna-XM_002615910.1; Dbxref=Genbank:
XP_002615956.1, GeneID:8497195; Name=XP_002615956.1; gbkey=CDS;
locus_tag=CLUG_03197; product=hypothetical protein; protein_id=
XP_002615956.1; transl_table=12
```

Obr. 3.2: Časť .gff súboru genómovej sekvencie kvasinky *C. lusitaniae*.

Naša implementácia GC3 metódy pracuje aj s .fai súbormi, ktoré sú tiež viazané na .fasta súbor. Obsahuje šesť stĺpcov oddelených tabulátorom. V prvom sa nachádza názov sekvencie, druhý obsahuje dĺžku (počet nukleotidov) danej sekvencie. V treťom stĺpci sa nachádza tzv. *offset*, ktorý označuje index, kde sa kontigová sekvencia začína vo .fasta súbore. Štvrtý a piaty stĺpec hovorí o počte báz, resp. počte bytov v jednom riadku .fasta súboru [16].

```
NW_003101578.1      2423571 7143121 80      81
NW_003101577.1      2215196 1810468 80      81
```

Obr. 3.3: Časť .fai súboru genómovej sekvencie kvasinky *C. lusitaniae*.

Dáta používané v tejto práci pochádzajú z databázy *RefSeq* [23] a *GenBank* [1]. Databáza RefSeq je verejne prístupná anotovaná databáza referenčných sekvencií genómov a proteínových produktov, pričom poskytuje jeden záznam pre biologickú molekulu. Je vytvorená a spravovaná *Národným centrom pre biotechnologické informácie*

- NCBI [31]. RefSeq tvoria záznamy rôznych organizmov od vírusov, cez baktérie až po eukaryoty.

GenBank je databáza všetkých verejne dostupných sekvencií DNA. Spolu s *DataBank of Japan* [34] a *European Nucleotide Archive* [14] sú súčasťou *International Nucleotide Sequence Database Collaboration* [7] a vzájomne si vymieňajú dáta. Táto databáza je prístupná prostredníctvom NCBI. NCBI poskytuje aj údaje o DNA, taxonomii, proteínových sekvenciách, proteínovej štruktúre a aj biomedicínsku literatúru archivovanú v *PubMed* [36]. V súčasnosti (24.4.2023) sa v tejto databáze nachádza viac ako 1,6 bil. nukleotidových sekvencií takmer 450 000 známych druhov organizmov.

3.3 Použité programové balíčky

Samotný algoritmus implementujeme v programovacom jazyku Python a Bash. Na prácu s .fasta a .gff súbormi (napríklad rozdelenie po chromozómoch, zistenie nukleotidovej pozície, vybratie ORF), používame *samtools* [16] a *gffread* [24]. V Python scripte používame na spustenie Bash príkazov balíček *subprocess* [8]. Na iterovanie po sekvenciách jednotlivých ORF a ich názvov aplikujeme *SeqIO* [26] z balíčka *Bio*. Pre vizualizáciu GC3 profilov používame *matplotlib* [30], konkrétne *pyplot*. Na vyberanie údajov z anotácie používame *gffpandas*. Ďalšími použitými balíčkami na prácu s dátovými štruktúrami sú *pandas* a *numpy* [10]. Najvýznamnejšie lokálne minimum sme vypočítali pomocou dvoch funkcií zo *scipy.signal* [17].

Scipy.signal pomocou porovnávania susedných hodnôt v poli nájde lokálne maximum a ich prominencie. Pomocou funkcie na nájdenie dôležitosti jednotlivých maxím (ang. *prominences*) vieme určiť najvýraznejšie lokálne maximum. Používateľ vie nastaviť rôzne parametre, z ktorých v tejto práci využijeme dva. Prvým nastavíme minimálnu hodnotu prominencie, ktorá zabezpečí, aby bolo minimum výrazné. Druhým parametrom je nastavenie počtu susedných hodnôt, s ktorými chceme daný údaj porovnávať.

3.4 Princíp GC3 metódy a naša implementácia

V GC3 metóde v informatickej reprezentácii ide o nájdenie súvislej časti reťazca nad abecedou {A, C, G, T} v určitých častiach genómovej sekvencie, kde má zastúpenie prvkov {G, C} na každej tretej pozícii najvýraznejšie lokálne minimum.

V každom chromozóme je žiadúce nájsť práve jednu centroméru, preto je potrebné rozdeliť sekvencie vo formáte MULTIFASTA podľa chromozómov. V našej implementácii program po načítaní .fasta súboru použitím *samtools* vytvorí .fai súbor. Na základe neho bude program iterovať po FASTA sekvenciách (chromozómových sekvenciách) v MULTIFASTA formáte. Ďalej overí, či je chromozóm dlhší ako 100 000 báz. V tomto

kroku sa filtrujú najmä mitochondriálne DNA a príliš krátke kontigy. V prípade, že používateľ zvolil variabilnú veľkosť okna, v nasledujúcom kroku program vypočíta vhodnú veľkosť okna pre daný chromozóm.

Následne program vyberie z .fasta súborov sekvencie, ktoré kódujú proteíny - otvorené čítacie rámce (ORF), čo sú nukleotidové sekvencie, ktoré začínajú štart-kodónom, končia niektorým stop-kodónom a počet báz medzi nimi je deliteľný tromi. V GFF3 sú tieto časti označené ako *CoDing Sequence* (CDS). Niektoré ORF sa skladajú z viacerých CDS prerušenými nekódujúcimi sekvenciami. Na to slúži *gffread*, ktorý pospája tieto CDS aby vznikol kompletný ORF. Následne algoritmus postupne iteruje po sekvenciách týchto ORF, pričom do analýzy nebude zarátavať príliš krátke (menej ako 100 kodónov), ORF, v ktorých celkový počet báz nie je deliteľný tromi alebo také, ktoré začínajú iným ako štart-kodónom, resp. končia iným ako jedným zo stop-kodónov.

Následne program iteruje cez jednotlivé kodóny daného ORF, ktoré filtruje na základe zoznamu vylúčených kodónov. V prípade, že kodón sa v tomto zozname nenachádza, program si k premennej zodpovedajúcej celkovej dĺžke ORF pripočíta 1. Následne overí, či sa na tretej pozícii tohto kodónu nachádza G alebo C. Ak áno, k premennej označujúcej počet kodónov končiacich na G alebo C, pripočíta 1. Obe premenné si zapamätá v dvoch poliach na rovnakom indexe.

Hodnoty, ktoré budeme vizualizovať, sa vypočítajú pomocou techniky posuvného okna. V jednom okne sa nachádza určitý počet ORF, pre ktoré program spočíta celkový počet GC3, celkový počet kodónov a na základe podielu týchto dvoch hodnôt zistí percentuálne zastúpenie GC3. Okno sa posúva o jeden ORF, čiže, ak v okne s veľkosťou 3 boli ORF (1, 2, 3), tak v nasledujúcom okne budú ORF (2, 3, 4).

V našej implementácii to bude fungovať nasledujúcim spôsobom. Program postupne prechádza po poli, v ktorom má uložené počty G a C na tretích pozíciách pre jednotlivé ORF. V prvom kroku spočíta celkový počet GC3 prvých W ORF, pričom W zodpovedá zvolenej veľkosti okna. Vo všetkých ďalších krokoch odčíta hodnotu GC3 v prvom ORF daného okna a pripočíta hodnotu prvého ORF, ešte nezarátaného. To isté urobí pre pole, v ktorom sa nachádzajú dĺžky ORF (počty kodónov v nich). Do poľa *graf_gc* uloží zastúpenie GC3 v danom okne v percentách, čo získa podielom hodnoty GC3 a dĺžky pre dané okno. V grafoch budeme určovať x-ovú súradnicu do stredu daného okna. Do poľa *graf_indexy* teda vložíme hodnoty $W//2$, $W//2+1$, $W//2+2$, až po hodnotu rozdielu počtu ORF daného chromozómu a polovice W .

V nasledujúcich krokoch program počíta centromerickú oblasť. Autori východiskovej štúdie [18] mapovali centroméru na miesto v okolí lokálneho minima, ktoré má celkové zastúpenie $GC3 < 50\%$. Inak povedané, po nájdení najvýraznejšieho lokálneho minima hľadali prvé okno, v ktorom je zastúpenie $GC3 > 50\%$. Toto však nie je univerzálne aplikovateľný postup, nakoľko celkové zastúpenie GC3 je u rôznych kvasiniek rôzne a mnohokrát nevystúpi nad 50% pozdĺž celého chromozómu.

Algoritmus v tejto práci nájde okno, ktoré predstavuje najvýznamnejšie lokálne minimum, hodnotu zastúpenia GC3 v tomto okne (ďalej len *gc3min*). Následne zaráta všetky okná v blízkosti indexu *gc3min*, pre ktoré platí podmienka, že podiel GC3 v danom okne je menší ako $gc3min * 1,15$. Inak povedané: súvislá oblasť, kde je zastúpenie GC3 nanajvýš o 15% vyššie ako *gc3min*, sú považované za centromerickú oblasť. Z toho ďalej program určí nukleotidovú pozíciu predpovedaného začiatku a konca centroméry.

V samotnej implementácii je aplikovaný nasledujúci postup. Najprv program obráti všetky hodnoty poľa *graf_gc*. Na nich pomocou balíčka *Scipy* vypočíta všetky lokálne maximá. Pre tento výpočet použije parametre *wlen*, ktorú nastaví na dĺžku poľa *graf_gc*. Z nich nájde maximum s najväčšou prominenciou a zapamätá si jej index. Tento index zároveň zodpovedá indexu lokálneho minima (ďalej len *index_min*) pôvodných *graf_gc* hodnôt.

Potom postupne preiteruje cez okná v *graf_gc*, pričom začne od *index_min* a pohybuje sa do oboch smerov. Pre ilustráciu rozoberieme prípad, kedy sa posúva smerom k indexu 0. Pre druhý prípad to bude fungovať analogicky. Označme si index okna, ktoré chceme porovnať ako *i1*. Pokiaľ je hodnota *graf_gc[i1]* menšia ako *graf_gc[index_min]* vynásobená 115% a zároveň je *i1* väčšie ako 0, tak sa zmenší *i1* o 1 a porovná okno na tomto indexe. V prípade, že *graf_gc[i1]* prevyšuje hodnotu *graf_gc[index_min] * 115%*, prehľadávanie sa zastaví a program si zapamätá index okna, ktoré ako posledné zodpovedalo podmienkam.

Ďalej program pomocou *gffpd.read_gff3* načíta anotáciu v priečinku do tabuľky (angl. *DataFrame*). Z neho vyberie tzv. *locus_tag*, ktorý zodpovedá ORF uprostred okna na indexe *i1* a koordinátu. Tiež program vytvorí grafy. Každý graf zodpovedá GC3 profilu jedného chromozómu. Na x-ovej osi je poradové číslo ORF (*gene number*), ktoré sme zarátavali. Na y-ovej osi je zastúpenie GC3 v %. Červenou čiarou je vyznačený centromerický lokus predikovaný programom a červenými značkami v tvare "x" sú znázornené gény ohraničujúce túto oblasť.

Do výpočtov zastúpenia GC3 v práci Denise Lynchovej [18] nezarátavali viacero kodónov: štart a stop kodóny a kodón CTG. Kodón CTG bol vylúčený z analýzy pri kvasinkách, ktoré používajú alternatívny jadrový genetický kód, nakoľko je to jediný triplet z rodiny CUN kódujúci serín. To znamená, že pri ňom odpadá možnosť variability, na ktorej je táto metóda založená. Podobným prípadom je štart-kodón ATG, ktorý je tiež ako jediný prekladaný ako metionín a preto bol z analýzy vylúčený. Tak tiež nepočítali so sekvenciami, ktoré boli kratšie ako 100 kodónov. Používali metódu posuvného okna, kde v jednom okne bolo 15 génov (15 ORF) a veľkosť kroku bola jeden gén (ORF). Toto sú parametre, ktoré sú variabilné a v našej práci sme sa zaoberali aj ich vplyvom na celkové výsledky.

3.5 Vstupy a výstupy

Ako sme už spomenuli, program pre analýzu potrebuje dva súbory vo formátoch FASTA a GFF3 v jednom priečinku pomenované ako *data.fasta* a *data.gff*. Na vstupe potrebuje cestu do priečinka, v ktorom sú tieto dáta uložené. Nasleduje veľkosť okna, ktorú používateľ chce použiť. V prípade, že dané číslo bude -1, program zvolí vhodnú veľkosť okna pre každý chromozóm samostatne. Posledným údajom je pole kodónov, ktoré chce používateľ vylúčiť z analýzy. Používateľ má možnosť zadať *zaratat_ctg*, kedy sa do analýzy nebudú započítavať kodóny ATG, TAA, TAG a TGA. Druhou prednastavenou možnosťou je *vyLucit_ctg*, kde sa okrem štart a stop-kodónov do analýzy nezapočíta ani CTG. Pokiaľ chce používateľ nastaviť rôzne iné kombinácie kodónov, ktoré chce z analýzy vylúčiť, môže ich zadať ako súvislý reťazec oddelený čiarkami bez medzier. Pokiaľ chce používateľ spustiť program pre viacero genómových sekvencií naraz, môže tak urobiť pridaním daných troch argumentov pre ďalšie genómové sekvencie. Príkladom môže byť:

```
python3 gc3method.py /home/user/Desktop/kvasinka1/ -1 TAG,tgg,tga,TAA /
home/user/Desktop/kvasinka2/ 15 zaratat_ctg
```

Výstupom budú GC3 profily v priečinku *pictures*, dopĺňujúce .fasta súbory, napríklad všetky CDS sekvencie z jedného chromozómu. Tiež na konzolu program vypíše gén pred začiatkom a gén po konci centromerického lokusu.

Nakoľko bioinformatické dáta pochádzajú z rôznych zdrojov, formát dát sa môže vymykať štandardom. Naša implementácia je nastavená na štandardizovaný formát FASTA a GFF3. Preto je možné, že pre niektoré dáta tento program nebude fungovať správne, resp. bude potrebná malá úprava programu.

3.6 Ilustratívny príklad

Na zjednodušenom príklade pomocou pseudokódu vysvetlíme fungovanie programu. Nech sekvencia $S=ATGGTCGGGTGCTAG ATGCAAGTCCTGTGA ATGGAGC-TAGTATAA$ je genómová sekvencia kvasinky, ktorú chceme analyzovať GC3 metódou, v ktorej použijeme veľkosť okna 2 gény a z analýzy vylúčime kodón CTG. Sekvencia je uložená v súbore *data.fasta* a anotácia v *data.gff*.

Program (Obr. 3.4) pomocou *gffread* vyberie nukleotidové sekvencie ORF z .fasta súboru. V ďalšej podmienke program overí, či táto sekvencia naozaj zodpovedá ORF a overí jeho dĺžku.

V druhom for-cykle program iteruje cez jednotlivé kodóny a overí, či sa nenachádzajú medzi vylúčenými kodónmi. Ak sa nenachádza, overí, či tento kodón končí na G alebo C. Po preiterovaní všetkých kodónov v jednom ORF si program vo výslednom

```

orfs = gffread vyber ORF v tejto sekvencii
for orf in geny:
    if orf zodpoveda definicii ORF:
        for kodony in orf:
            if nie je vyluceny kodon:
                pocet_celkovy+=1
                if je na tretej pozicii G alebo C:
                    pocet_gc+=1
            dlzky.append(pocet_celkovy)
        gc.append(pocet_gc)

```

Obr. 3.4: Pseudokód časť 1.

poli, v ktorom budú ORF v pôvodnom poradí, zapamätá celkový počet kodónov (bez vylúčených kodónov) v ORF a počet kodónov končiacich na G alebo C.

V sekvencii S by teda v prvom ORF: ATGGTCGGGTGCTAG zistil, že zodpovedá požadovanej dĺžke a podmienkam. Nachádzajú sa tu dva kodóny, ktoré vylúčime a vznikne nám sekvencia GTCGGGTGC. Táto sekvencia je dlhá 3 kodóny a obsahuje 3 kodóny končiace na G alebo C. Po analýze všetkých ORF v sekvencii S by sme získali polia $dlzky=[3, 2, 3]$ a $gc=[3, 1, 1]$.

Pri technike posuvného okna (Obr. 3.5) v prvom kroku for-cyklu program spočíta celkový počet GC3 (premenná $okno_gc$) a celkový počet zarátaných kodónov (premenná $okno_dlzka$) prvých $okno$ ORF, pričom premenná $okno$ zodpovedá veľkosti zvoleného okna. V ďalších krokoch odčíta od $okno_gc$ hodnotu, ktorá prislúcha prvému ORF v danom okne. Následne pripočíta počet GC3 kodónov nového, ešte nezarátaného ORF. To isté sa zopakuje pre celkové dĺžky kodónov ($okno_dlzka$). Následne program vypočíta GC3 v %.

Z našej sekvencie S by teda vzniklo pole $graf_gc$, kde $okno_gc = gc[0] + gc[1]$, čiže $3 + 1$. Podobne by sme zistili, že $okno_dlzka = 5$ a teda $graf_gc[0] = 4/5 \cdot 100\% = 80\%$. Následne od $okno_gc$ odčítame $gc[0]$ a pripočítame $gc[2]$. Podobne aj pre $okno_dlzka$. $graf_gc[1]$ bude teda 40%. Toto sú dáta pripravené na vizualizáciu.

Program (Obr. 3.6) prvom kroku vynásobí všetky hodnoty v $graf_gc$ číslom -1, čím obráti tieto hodnoty a teda oblasti, ktoré predstavovali pokles teraz predstavujú nárast a opačne. Toto môžeme využiť, pretože pozícia lokálneho maxima pre obrátené hodnoty GC3 je rovnaká ako pozícia lokálneho minima pre pôvodné hodnoty GC3. *Scipy.signal* [17] nájde tieto pozície. Potom sa program pozrie na susedné hodnoty v poli $graf_gc$ a porovnáme ich s hodnotou $graf_gc[min_index]$. Pokiaľ túto hodnotu prevyšujú o viac ako 15%, prehľadávanie zastavíme a zapamätáme si indexy týchto hodnôt. V ďalších krokoch je na základe nich možné získať požadované informácie, napríklad koordináty génov.

```

for i in gc:
    if i == 0:
        okno_gc = sum(gc[:okno-1])
        okno_dlzka = sum(dlzky[:okno-1])
    else:
        okno_gc = okno_gc - gc[i] + gc[i+okno]
        okno_dlzka = okno_dlzka - dlzky[i] + dlzky[i+okno]
    indexy.append(okno//2+i)
    graf_gc3.append(okno_gc/okno_dlzka*100)

```

Obr. 3.5: Pseudokód časť 2.

```

gc_naopak = [i * (-1) for i in graf_gc]
min_index = index najvyznamnejšieho lokálneho maxima najdeného pomocou
               scipy
i = min_index
k = min_index

while i >= 0 and graf_gc[i] < graf_gc[min_index]/100*115:
    i -= 1

while k < len(graf_gc) and graf_gc[k] < graf_gc[min_index]/100*115:
    k += 1

i = i+swg//2+1
k = k+swg//2+1

```

Obr. 3.6: Pseudokód časť 3.

Kapitola 4

Výsledky

4.1 Verifikácia GC3 metódy

Na verifikáciu našej implementácie GC3 metódy sme použili genómové sekvencie kvasiniek *C. lusitaniae*, *S. stipitis* a *Y. lipolytica*. Cieľom bolo porovnať naše výsledky so závermi, ktoré vyšli v publikácii Denise Lynchovej a kol. [18]. Údaje o použitých sekvenciách genómov jednotlivých kvasiniek je možné nájsť v Prílohe A (5).

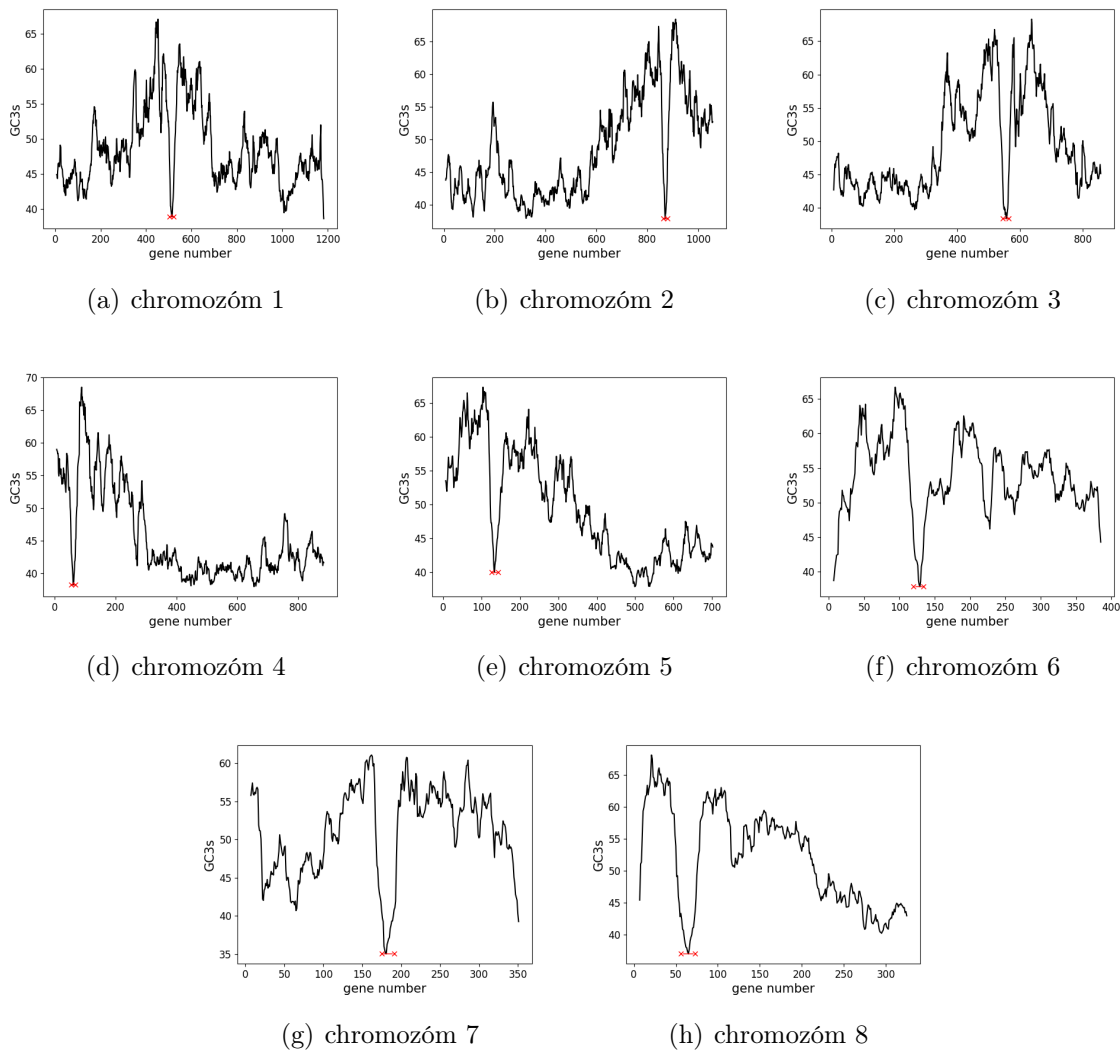
4.1.1 *Clavispora lusitaniae*

Kvasinka *C. lusitaniae*, ktorá bola v minulosti klasifikovaná ako *Candida lusitaniae*, je haploidná patogénna kvasinka, má osem jadrových chromozómov s celkovou dĺžkou genómu približne 12,11 Mb. Používa alternatívny genetický kód a preto sme spolu so štart- a stop-kodónmi vylúčili z analýz GC3 aj kodón CTG. V databáze RefSeq je genómová sekvencia kmeňa *C. lusitaniae* ATCC 42720, ktorý obsahuje 5 936 CDS. Celkové zastúpenie GC v tomto kmeni je 44,5% [20].

Pre konštrukciu GC3 profilov (Obr. 4.1) bola použitá veľkosť okna 15 génov. Na všetkých chromozómoch je v získaných grafoch centromerický lokus veľmi dobre viditeľný. V grafe je ho možné identifikovať ako oblasť s výrazným poklesom GC3, teda najvýraznejšie lokálne minimum. Najmä na chromozómoch 2, 3 a 4 je očividné, že centroméra nemusí obsahovať globálne minimum GC3, inak povedané: na chromozómoch sú oblasti, ktoré majú nižšie zastúpenie GC3 ako centroméra. Aj z týchto profilov je zrejmé, že centroméry sa nenachádzajú v strede chromozómov. Zastúpenie GC3 v centromerických oblastiach na jednotlivých chromozómoch siaha pod 41%.

4.1.2 *Scheffersomyces stipitis*

Kvasinka *S. stipitis*, ktorá bola v minulosti klasifikovaná ako *Pichia stipitis*, je nepatogénna kvasinka známa svojou schopnosťou fermentácie xylózy na etanol a iné produkty.



panel	chromozóm	gén pred centromérou	pozícia	gén za centromérou	pozícia
(a)	1	CLUG_00513	1040135	CLUG_00531	1075750
(b)	2	CLUG_02096	1791934	CLUG_02111	1818060
(c)	3	CLUG_02860	1139062	CLUG_02880	1176649
(d)	4	CLUG_03252	121683	CLUG_03268	157316
(e)	5	CLUG_04233	270454	CLUG_04249	309908
(f)	6	CLUG_04958	257515	CLUG_04972	291690
(g)	7	CLUG_05412	360609	CLUG_05429	399488
(h)	8	CLUG_05659	143232	CLUG_05676	176623

Obr. 4.1: Analýza distribúcie GC3 pozdĺž chromozómov 1 až 8 kvasinky *C. lusitaniae* ATCC 42720 a tabuľka génov pred a po centromére a ich vonkajšie nukleotidové koordináty pre jednotlivé chromozómy.

Je to haploidná kvasinka a veľkosť jadrového genómu je približne 15,44 Mb organizovaných v ôsmich chromozómoch a celkový obsah GC párov je 41,2%. Tiež používa alternatívny genetický kód. Pre analýzu sme použili genómovú sekvenciu kmeňa CBS 6054, ktorá obsahuje 7 094 CDS [20].

V grafoch zastúpenia GC3 pre túto kvasinku (Obr. 4.2) je tiež centromerický lokus dobre viditeľný pre každý chromozóm. V každej centromerickej oblasti kleslo zastúpenie GC3 pod 39%. Do analýzy sme nezapočítavali kodón CTG a bolo použité okno s veľkosťou 15 génov tak, ako vo východiskovej publikácii [18].

4.1.3 *Yarrowia lipolytica*

Y. lipolytica je haploidná kvasinka a patrí medzi nepatogénne kvasinky a používa štandardný genetický kód. Ako zdroje uhlíka môže využívať uhl'ovodíky a rôzne tuky. Jej jadrový genóm je veľký približne 20,5 Mb v šiestich chromozómoch s celkovým zastúpením GC 49%. Genóm z kmeňa CLIB89(W29) obsahuje 7 949 CDS [20].

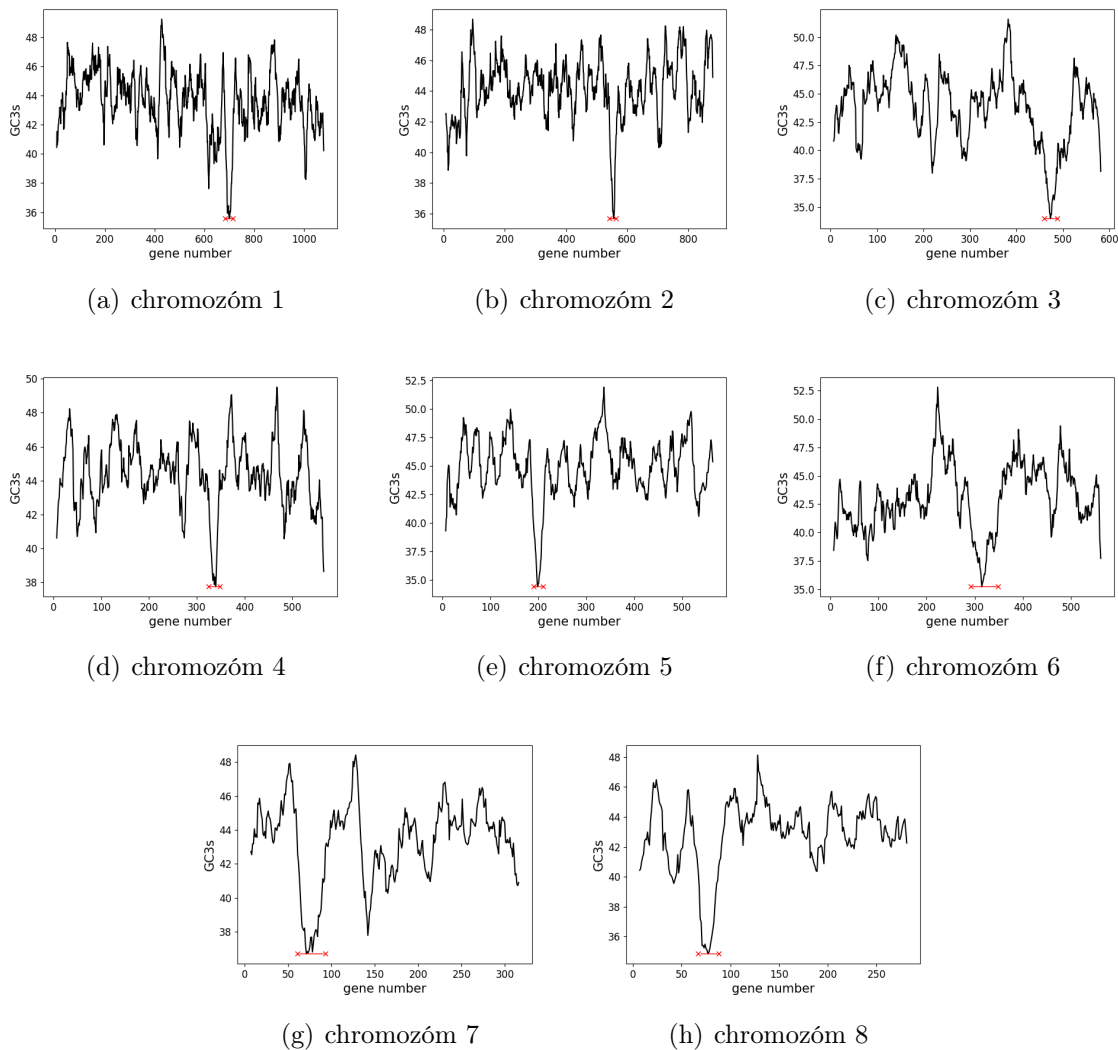
Pri analýze zastúpenia GC3 v chromozómoch *Y. lipolytica* (Obr. 4.3) bolo použité okno 15 génov a z analýzy boli vylúčené len štart- a stop-kodóny. Centromerická oblasť je na všetkých chromozómoch veľmi dobre viditeľná.

4.1.4 Porovnanie výsledkov

Porovnávaním získaných výsledkov s východiskovou publikáciou Denise Lynchovej a kol. (2010) [18] bola overená správnosť našej implementácie. Ako prvá prebehla kontrola grafickej reprezentácie GC3 profilov (Obr. 4.3 a 4.4). Pre všetky tri kvasinky boli grafy veľmi podobné. Rozdiely nastali na x-ovej osi v poradovom čísle génu ("gene number") pri *C. lusitaniae* a *Y. lipolytica*. V našich profiloch je x-ová os približne o 200 génov dlhšia. Napríklad pre *Y. lipolytica* chromozóm 1 náš GC3 profil obsahuje približne 810 génov zarátaných do analýzy, kým v práci z 2010 [18] ich bolo približne 640, avšak tvar krivky GC3 profilu je porovnateľný. Preto sme namiesto porovnávania poradového čísla génu zobrali do úvahy ID génov pred a po centromerickom lokuse a ich koordináty.

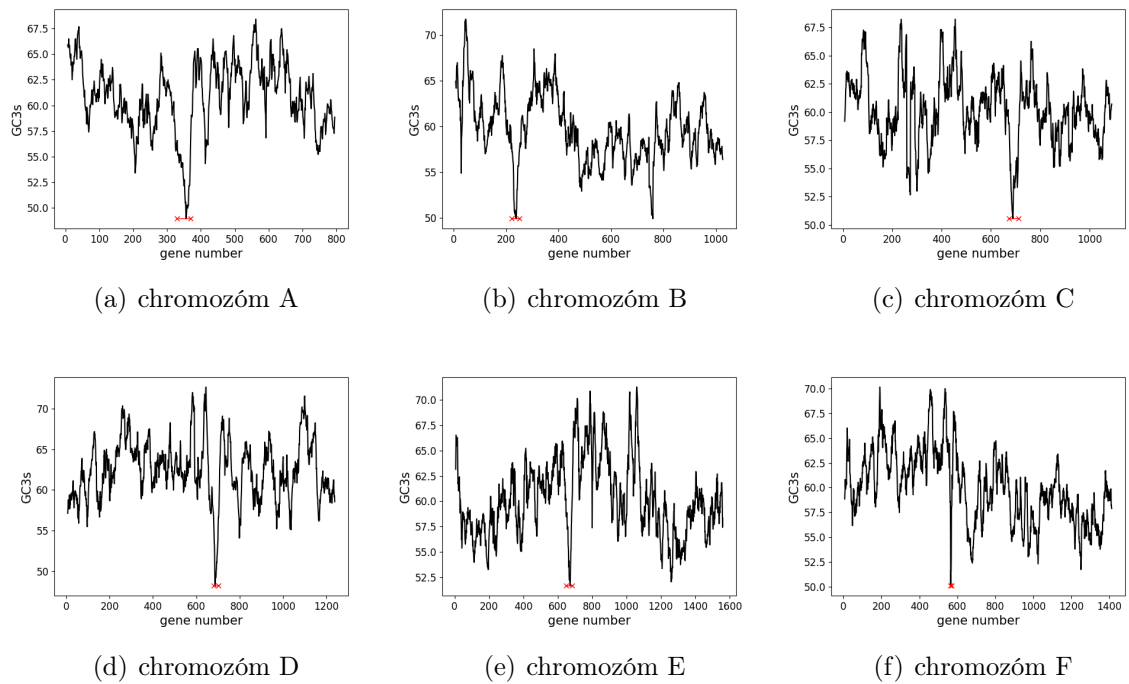
Naša predikcia génu pred centromérou a po nej na jednotlivých chromozómoch je porovnateľná s výsledkami Denise Lynchovej a kol. (2010) [18]. Napríklad na chromozóme 1 *C. lusitaniae* nami predikovaný gény pred centromerickou oblasťou a po nej sú CLUG_00513 a CLUG_00531. Vo východiskovej práci sú to CLUG_00511 a CLUG_00537. Pre chromozóm 7 sú naša predikcie CLUG_05412 a CLUG_05429. U Lynchovej CLUG_05411 a CLUG_05430. Pri ostatných chromozómoch ako aj sekvenciách iných kvainiek je to obdobne.

Koordináty génov ohraničujúcich centromerický lokus predikovaný našou implementáciou GC3 pri chromozóme 1 *S. stipitis* vyšiel 2 221 833 a 2 375 948. Vo východiskovej práci [18] predikovali túto oblasť medzi 2 230 753 a 2 332 231. Pre chromozóm 3 nám



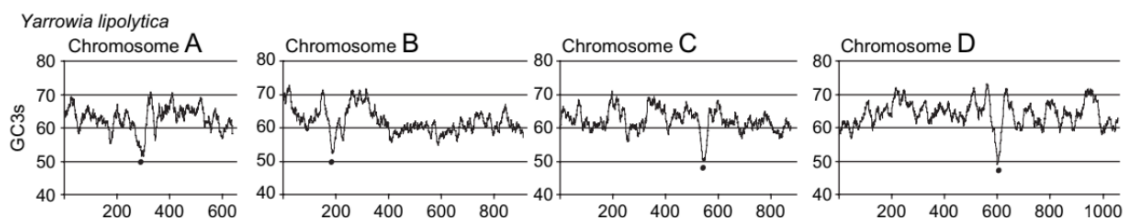
panel	chromozóm	gén pred centromérou	pozícia	gén za centromérou	pozícia
(a)	1	PICST_81158	2221833	PICST_53664	2375948
(b)	2	PICST_29980	1633544	PICST_35133	1731945
(c)	3	PICST_30962	1378361	PICST_82710	1486160
(d)	4	PICST_31523	1006692	PICST_58749	1098481
(e)	5	PICST_67821	617142	PICST_60854	689729
(f)	6	PICST_32861	806332	PICST_47586	1015645
(g)	7	PICST_85406	219647	PICST_85181	342402
(h)	8	PICST_64420	248922	PICST_73961	351425

Obr. 4.2: Analýza distribúcie GC3 pozdĺž chromozómov 1 až 8 kvasinky *S. stipitis* CBS 6054 a tabuľka génov pred a po centromére a ich vonkajšie nukleotidové koordináty pre jednotlivé chromozómy.



panel	chromozóm	gén pred centromérou	pozícia	gén za centromérou	pozícia
(a)	A	YALI1_A09165g	915946	YALI1_A10254g	1025387
(b)	B	YALI1_B06735g	673492	YALI1_B07675g	767522
(c)	C	YALI1_C20901g	2088817	YALI1_C22556g	2256103
(d)	D	YALI1_D18845g	1884552	YALI1_D19365g	1939961
(e)	E	YALI1_E17786g	1778583	YALI1_E18854g	1887168
(f)	F	YALI1_F16218g	1621421	YALI1_F16366g	1636624

Obr. 4.3: Analýza distribúcie GC3 pozdĺž chromozómov A až F kvasinky *Y. lipolytica* CLIB89(W29) a tabuľka génov pred a po centromére a ich vonkajšie nukleotidové koordináty pre jednotlivé chromozómy.



Obr. 4.4: Analýza distribúcie GC3 pozdĺž chromozómov 1 až 3 kvasinky *Y. lipolytica* v práci D. Lynchovej a kol. (2010) [18].

vyšla predikcia centromerickej oblasti medzi 1 378 361 a 1 486 160. Vo východiskovej práci medzi 1 399 674 a 1 477 779, čo sú porovnateľné výsledky. Podobne to platí aj pre ostatné chromozómy.

Na základe týchto údajov sa dá potvrdiť správnosť našej implementácie, napriek drobným nezrovnalostiam. Je možné, že určité výkyvy sú spôsobené aj zmenou v dátach, teda v samotnej sekvencii alebo v anotácii, ktoré boli v databázach od roku 2010 aktualizované.

4.2 Optimalizácia parametrov

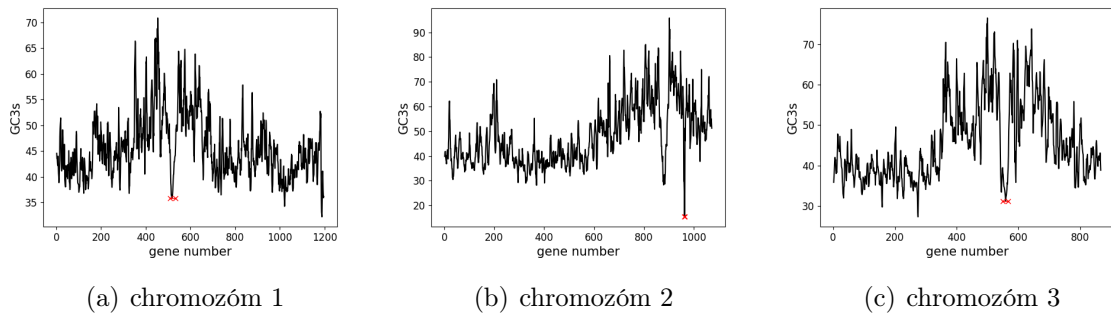
N výsledky výpočtov vplyva viacero parametrov. Na príklade *C.lusitaniae* sme skúmali vplyv veľkosti zvoleného okna a vylúčenia kodónov ATG a CTG. Kvasinku *C.lusitaniae* sme zvolili pre najjednoduchšie viditeľné centromerické oblasti, využívanie alternatívneho genetického kódu a vysoké výkyvy zastúpenia GC3 pozdĺž celých chromozómov, ktoré rôzne vplyvy parametrov výborne podčiarkujú.

4.2.1 Veľkosť okna

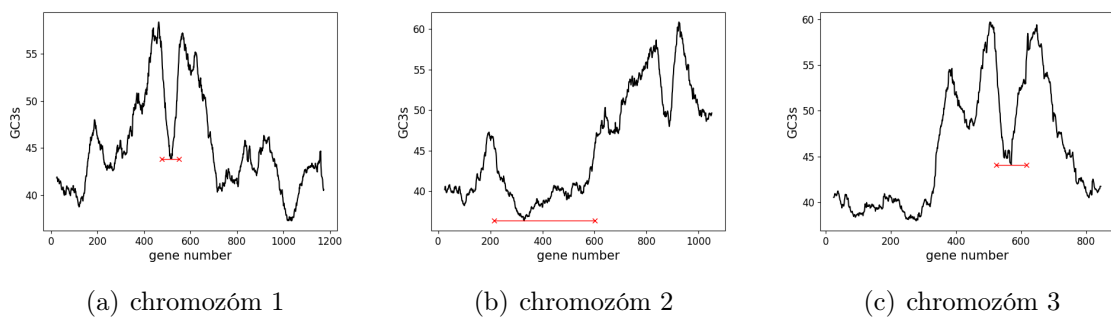
Tento parameter má vplyv ako na grafický výsledok, tak aj na výpočet pozície centroméry. Pri aplikácii kódu z tejto práce má používateľ možnosť nastaviť veľkosť okna manuálne uvedením počtu génov v jednom okne alebo automaticky zadáním čísla -1. Následne program na základe veľkosti chromozómu vypočíta primeranú veľkosť okna. Pre veľké chromozómy sú prirodzene vhodnejšie väčšie okná ako pre malé chromozómy.

Na grafoch s veľkosťou okna 5 génov (Obr. 4.5) je oblasť centroméry viditeľná, no nie je tak zreteľná, pretože grafy obsahujú veľa šumu, ktorý je spôsobený veľkým množstvom údajov. Vtedy je vhodné zmenšiť citlivosť a teda zväčšiť veľkosť okna. V porovnaní s primeranou veľkosťou okna 15 génov (Obr. 4.1) sa objavilo viacero výraznejších lokálnych GC3 poklesov (napríklad v okolí génu 870 na chromozóme 2), ktoré sa však nedajú považovať za centromerickú oblasť. Vypočítaný centromerický lokus s použitím veľkosti okna 5 génov sa prekrýva s predikciou tejto oblasti s veľkosťou okna 15 génov, no táto oblasť je oveľa menšia.

Naopak v prípade GC3 profilov s veľkosťou okna 50 génov (Obr. 4.6) je šum oveľa menší, no tentokrát je to na úkor presnosti. Oblasť centroméry je stále viditeľná, no vizuálne zasahuje väčšiu oblasť. Tieto oblasti sa síce v chromozóme 1 a 3 prekrývajú s oblasťami predikovanými pomocou okna s veľkosťou 15 génov, no na chromozóme 2 algoritmus predikoval centroméru do miesta, ktorá sa s pôvodnou predikciou nestretáva. V porovnaní s GC3 profilmi s veľkosťou okna 15 génov (Obr. 4.1) nie je možné vidieť detaily zmien zastúpení GC3 pozdĺž chromozómu. Veľmi výrazné sú zmeny vrámci y-ovej súradnice. Napríklad na chromozóme 1 na Obr. 4.1 je zastúpenie GC3 v oblasti



Obr. 4.5: Grafická reprezentácia GC3 profilu pre chromozómy 1, 2 a 3 *C. lusitaniae* pri použití okna o veľkosti 5 génov s krokom o jeden gén.



Obr. 4.6: Grafická reprezentácia GC3 profilu pre chromozómy 1, 2 a 3 *C. lusitaniae* pri použití okna o veľkosti 50 génov s krokom o jeden gén.

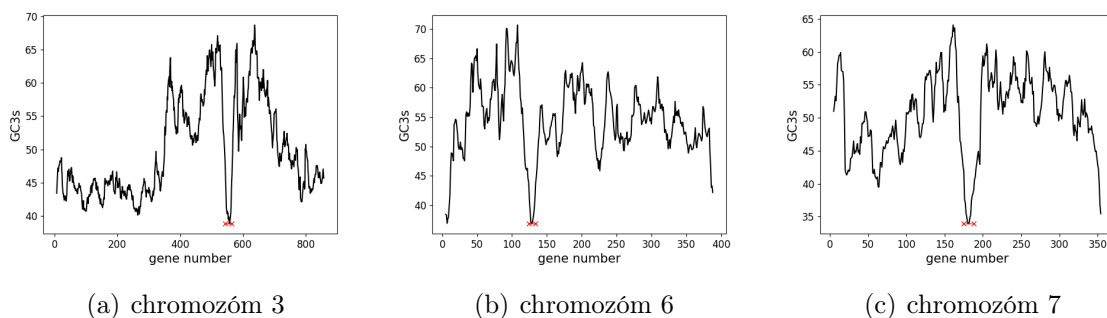
centroméry približne 37,5%, pri okne s veľkosťou 50 génov je to približne 44%.

Preto prikladáme náš návrh (Obr. 4.7), ktorý na základe veľkosti chromozómu určí vhodnú veľkosť okna. Pre chromozómy kratšie ako 1 mil. nukleotidov bude veľkosť okna 10 génov, pre chromozómy väčšie ako 1 mil. a menšie ako 2,5 mil. zvolí veľkosť 15 nukleotidov. Ak veľkosť leží v intervale (2,5 mil., 4 mil.) určí veľkosť okna 25 génov a pre chromozómy väčšie ako 3 mil. génov bude veľkosť okna 35 génov.

Vidíme, že aj pre chromozómy väčšie (chromozóm 1), aj pre stredné (chromozóm 5) a aj pre menšie (chromozóm 7) je veľkosť okna primeraná. Program pri tomto nastavení vytvorí grafy, ktoré sú dobre čitateľné a zároveň presnosť je dostatočne vysoká, čo sme overili aj výpočtom.

Program pre chromozóm 3 program zvolil okno 15 génov, tak ako v pôvodnej analýze (Obr. 4.1), takže aj výpočet koordinát vyšiel rovnaký. Pre chromozómy 6 a 7 bolo zvolené okno s veľkosťou 10 génov, pretože veľkosť týchto chromozómov je menšia. Presnosť výpočtu to ovplyvnilo len minimálne. Podobné výsledky vyšli aj pri analýzach genómových sekvencií iných druhov kvasiniek.

Veľkosť okna je teda jeden z faktorov, ktorý má vplyv na vyobrazenie zastúpenia GC3 aj na predikciu centromerickej oblasti.



Obr. 4.7: Grafická reprezentácia GC3 profilu pre chromozómy 3 a 7 *C. lusitaniae* pri použití variabilnej veľkosti okna.

4.2.2 Vylúčenie kodónov ATG a CTG

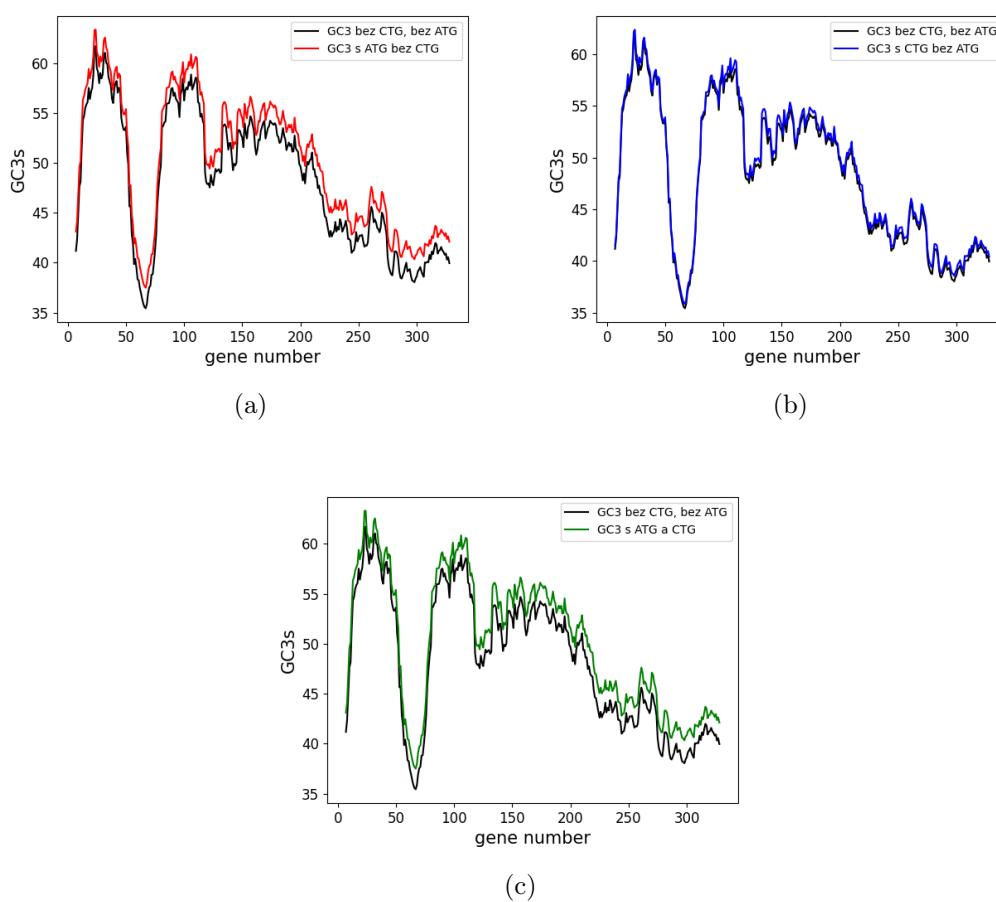
Ako sme už spomenuli, pri kodónoch ATG a CTG odpadá možnosť variability. Preto bol kodón ATG spolu so stop-kodónmi vylúčený pri analýze všetkých druhov kvasiniek. Kodón CTG bol z analýzy vylúčený pri kvasinkách, ktoré používajú alternatívny genetický kód.

Nasledujúce grafy prislúchajúce chromozómu 8 kvasinky *C. lusitaniae* (Obr. 4.8) ukazujú zmeny spôsobené kodónom ATG a CTG. Vo všetkých troch grafoch sú čiernou farbou profily GC3 pôvodnej analýzy (Obr. 4.1), pri ktorých sú vylúčené stop-kodóny, štart-kodón ATG aj kodón pre serín CTG. Na chromozóme 8 sú zmeny tohto parametra najlepšie viditeľné. Celkový počet kodónov zarátaných do analýzy na chromozóme 8 je 154 931.

Na prvom grafe je červenou farbou zobrazený profil GC3 s vylúčeným kodónom CTG, no zarátaným ATG. V tomto prípade je viditeľný výraznejší rozdiel opäť smerom k vyššiemu zastúpeniu GC3, kvôli guanínu na tretej pozícii kodónu ATG pre metionín. Výraznejšia zmena je spôsobená väčším počtom kodónu ATG ako kodónu CTG v génoch *C. lusitaniae* na chromozóme 8 takmer dvojnásobne. Počet kodónov ATG je 2 936.

Na druhom grafe je modrou farbou je naznačený GC3 bez vylúčenia CTG kodónu. Keďže na tretej pozícii tohto kodónu je guanín, hodnoty na y-osi sa vynechaním CTG z analýzy mierne znížili, avšak pozícia centroméry ostáva bezo zmeny, čo sme si overili aj numericky. Počet kodónov CTG je na tomto chromozóme 1 598.

Tretí graf je vyobrazením zastúpenia GC3 pôvodnej analýzy v porovnaní s GC3 profilom zelenou farbou, v ktorom sa do analýzy započítava aj ATG aj CTG. Prirodzene je tu rozdiel najvýraznejší, avšak na samotnú predikciu pozície centroméry nemá veľký vplyv. Mení sa len hodnota GC3.



Obr. 4.8: GC3 profily pre chromozóm 8 *C. lusitaniae* s viditeľným rozdielom pri vylúčení kodónov ATG alebo/a CTG z analýzy.

4.3 Aplikácia GC3 metódy na ďalšie genómy kvasiniek

4.3.1 *Candida auris*

C. auris patrí do čeľade *Metschnikowiaceae*. Veľkosť jadrového genómu *C. auris* z kmeňa B11220 je 12,25 Mb, je usporiadaný v 7 chromozómoch, obsahuje 5 327 CDS a celkové zastúpenie GC je 45,2%. Používa alternatívny genetický kód [20].

Na grafoch zastúpenia GC3 pre jednotlivé chromozómy *C. auris* (Obr. 4.9) je vizuálna predikcia centromerického lokusu pomerne jednoducho uskutočniteľná. Na chromozóme 6 vidíme dve oblasti s výrazným poklesom, kde môže ísť o centroméru a neocentroméru. Na všetkých chromozómoch siaha hodnota GC3 v oblasti s najvýraznejším lokálnym minimom pod 42%. Analýza prebehla s veľkosťou okna 15 génov.

4.3.2 *Candida metapsilosis*

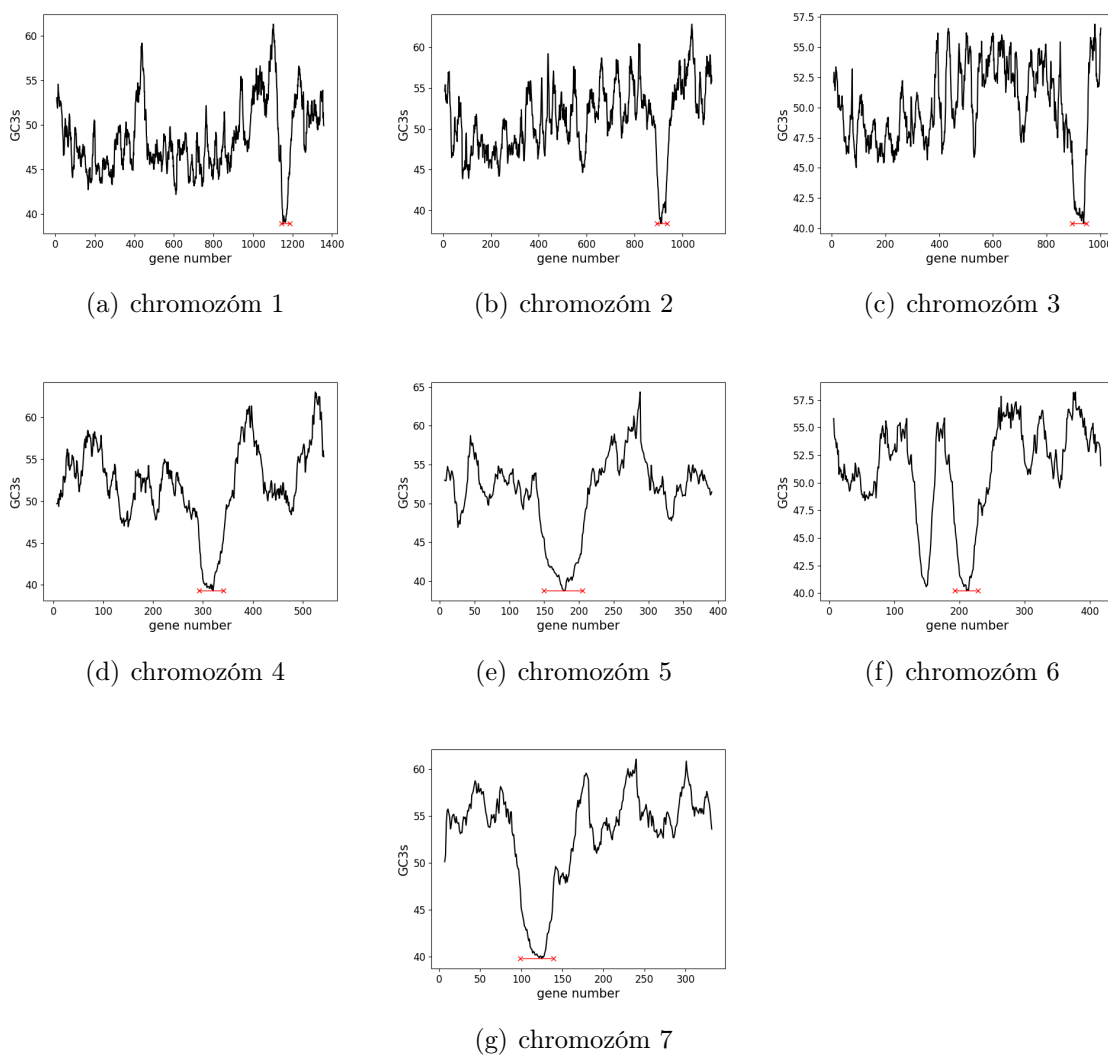
C. metapsilosis z čeľade *Debaryomycetaceae* má genóm veľký 13,09 Mb, obsahuje 5 726 proteín-kódujúcich génov a celkové zastúpenie GC je 38,1%. Používa alternatívny genetický kód. Pre túto kvasinku z kmeňa BP57 máme dostupné dáta, ktoré sú spracované na úroveň *scaffolds*, ktorých je 9 [20].

Na niektorých GC3 profiloch (Obr. 4.10) je pokles v zastúpení GC3 viditeľný. Obzvlášť pre scaffolds 4 a 7, pokles bol veľmi výrazný. Na scaffolds 1 a 2 sú frekventované vysoké výkyvy GC3, čo sťažuje identifikáciu centromerickej oblasti. Čiastočným riešením by mohlo byť použitie väčšieho okna. Dva scaffolds tu neuvádzame, pretože ich veľkosť bola oproti ostatným príliš malá a teda sa v nich centromerická oblasť nachádza s malou pravdepodobnosťou. Pre analýzu bolo použité okno s veľkosťou 15 génov.

4.3.3 *Candida orthopsilosis*

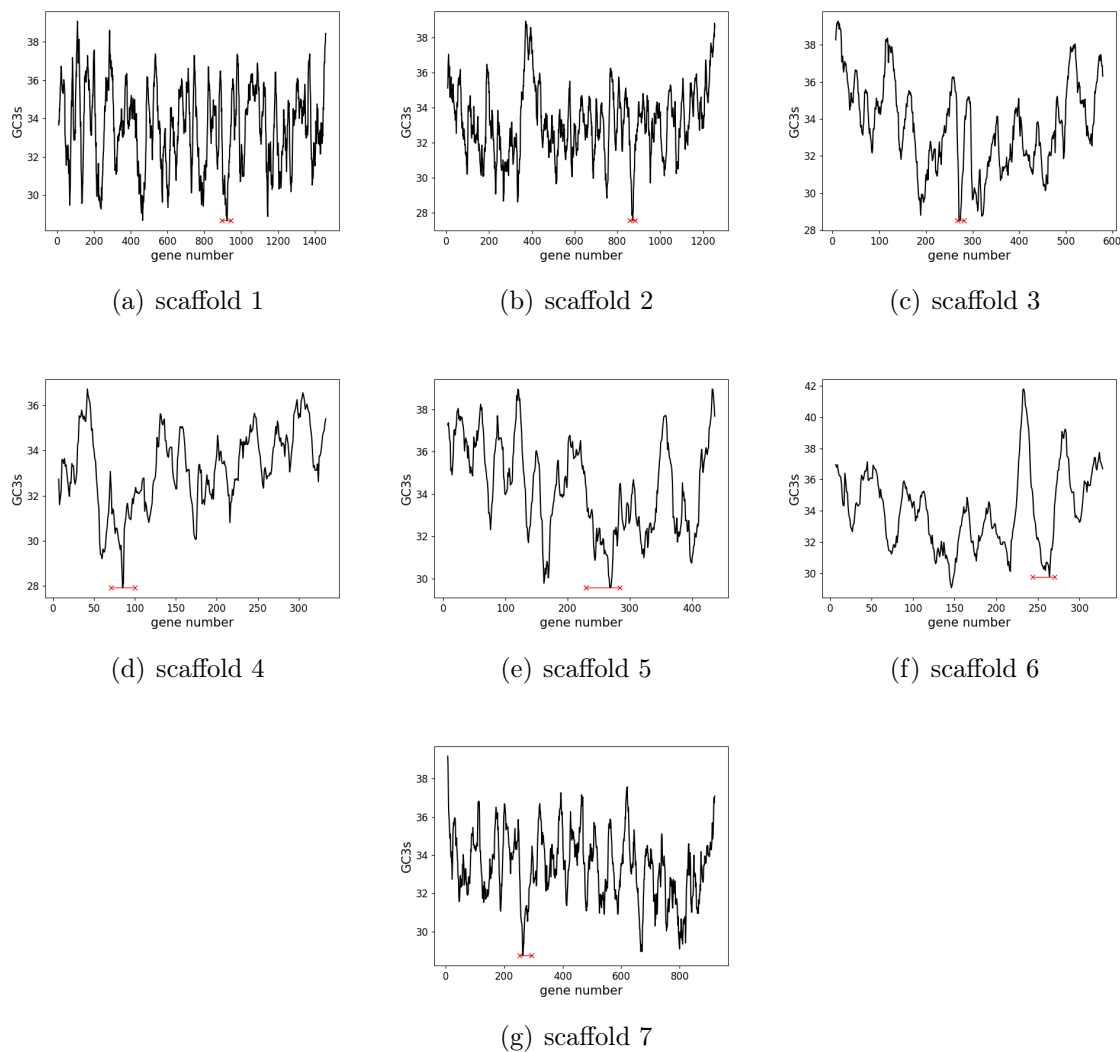
C. orthopsilosis z čeľade *Debaryomycetaceae* má hybridný genóm. Veľkosť genómu tejto kvasinky je 12,66 Mb organizovaných v ôsmich chromozómoch. Genóm *C. orthopsilosis* kmeňa Co 90-125 obsahuje 5 678 CDS, pričom používa alternatívny genetický kód. Celkové zastúpenie GC v jadrovom genóme je 37,5% [20].

GC3 profily chromozómov *C. orthopsilosis* (Obr. 4.11) sú veľmi odlišné. Na niektorých je oblasť centroméry jednoznačne viditeľná (napríklad chromozóm C, D, F alebo H) a na niektorých sú viacero oblastí s výrazným poklesom GC3 (napríklad chromozóm B alebo G). Na chromozóme G zastúpenie GC3 sa po celej dĺžke veľmi rýchlo a výrazne mení a preto je ťažké z grafu jednoznačne určiť centromerický lokus. Pre analýzu bolo použité okno s veľkosťou 15 génov.



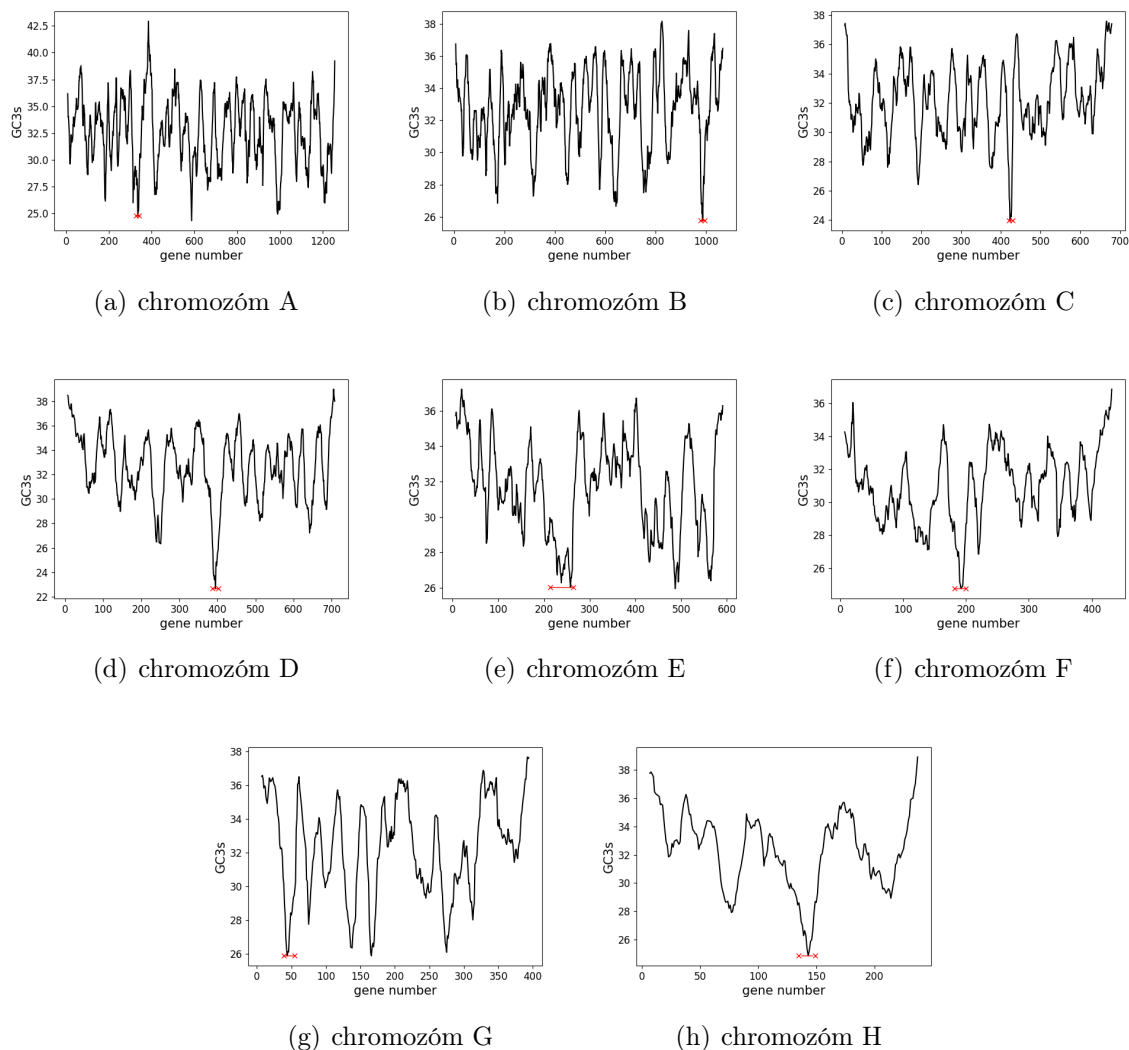
panel	chromozóm	gén pred centromérou	pozícia	gén za centromérou	pozícia
(a)	1	CJI96_0001202	2588345	CJI96_0001246	2692836
(b)	2	CJI96_0002380	2012283	CJI96_0002422	2117055
(c)	3	CJI96_0003557	2024020	CJI96_0003611	2168864
(d)	4	CJI96_0003993	658266	CJI96_0004044	787730
(e)	5	CJI96_0004422	345231	CJI96_0004479	471740
(f)	6	CJI96_0004893	464386	CJI96_0004928	554288
(g)	7	CJI96_0005241	268533	CJI96_0005285	375171

Obr. 4.9: Analýza distribúcie GC3 pozdĺž chromozómov 1 až 8 pre *C. auris* B11220 a tabuľka génov pred a po centromére a ich vonkajších koordinát pre jednotlivé chromozómy.



panel	scaffold	gén pred centromérou	pozícia	gén za centromérou	pozícia
(a)	1	I9W82_000917	2041927	I9W82_000965	2127107
(b)	2	I9W82_002374	2038826	I9W82_002399	2094082
(c)	3	I9W82_003064	617513	I9W82_003080	648449
(d)	4	I9W82_003464	170744	I9W82_003494	230220
(e)	5	I9W82_003982	563418	I9W82_004036	703907
(f)	6	I9W82_004453	608197	I9W82_004480	658272
(g)	7	I9W82_004810	575879	I9W82_004853	678975

Obr. 4.10: Analýza distribúcie GC3 pozdĺž chromozómov 1 až 8 pre *C. metapsilosis* BP57 a tabuľka génov pred a po centromére a ich vonkajších koordinát pre jednotlivé scaffolds.



panel	chromozóm	gén pred centromérou	pozícia	gén po centromére	pozícia
(a)	A	CORT_0A03600	777101	CORT_0A03740	804930
(b)	B	CORT_0B10570	2204968	CORT_0B10760	2252623
(c)	C	CORT_0C04510	1017904	CORT_0C04600	1040094
(d)	D	CORT_0D04230	816219	CORT_0D04390	855683
(e)	E	CORT_0E02450	519539	CORT_0E02970	643248
(f)	F	CORT_0F02080	437925	CORT_0F02260	473582
(g)	G	CORT_0G00550	90822	CORT_0G00730	135716
(h)	H	CORT_0H01570	322257	CORT_0H01720	355206

Obr. 4.11: Analýza distribúcie GC3 pozdĺž chromozómov A až F pre *C. orthopsilosis* Co 90-125 a tabuľka génov pred a po centromére a ich vonkajších koordinát pre jednotlivé chromozómy.

4.3.4 *Candida parapsilosis*

Kvasinka *C. parapsilosis* patrí medzi patogénne kvasinky z čeľade *Saccharomycetaceae*. Je haploidná, genóm má organizovaný v ôsmich chromozómoch a jeho veľkosť je približne 13 Mb. Zastúpenie GC je 38.5% a obsahuje 5 837 CDS. Dáta sú z kmeňa CAJEJE01. Táto kvasinka používa alternatívny genetický kód [20].

Na GC3 profiloch pre *C. parapsilosis* (Obr. 4.12) je na väčšine chromozómoch možné s ľahkosťou predikovať centromerický lokus. Na chromozómoch 4, 6 a 7 sú tieto oblasti menej výrazné. Pozdĺž celých chromozómov sú výkyvy v zastúpení GC3 veľmi veľké. Na chromozóme 5 vidíme v tesnej blízkosti dve oblasti s výrazným poklesom v GC3. Algoritmus predikoval teda centroméru do väčšej oblasti ako na ostatných chromozómoch a tým zahrnul obe výrazné lokálne minimá. Pre analýzu bolo použité okno s veľkosťou 15 génov.

4.3.5 *Candida verbasci*

C. verbasci je nepatogénny druh kvasinky, ktorý patrí do čeľade *Debaryomycetaceae*. Požíva alternatívny genetický kód. Veľkosť jej genómu je 12,17 Mb, je usporiadaný v siedmich chromozómoch. Pre analýzu bola použitá *C. verbasci* kmeňa CBS 12699, ktorá obsahuje 5 313 proteín-kódujúcich génov a celkové zastúpenie GC je 28% [20].

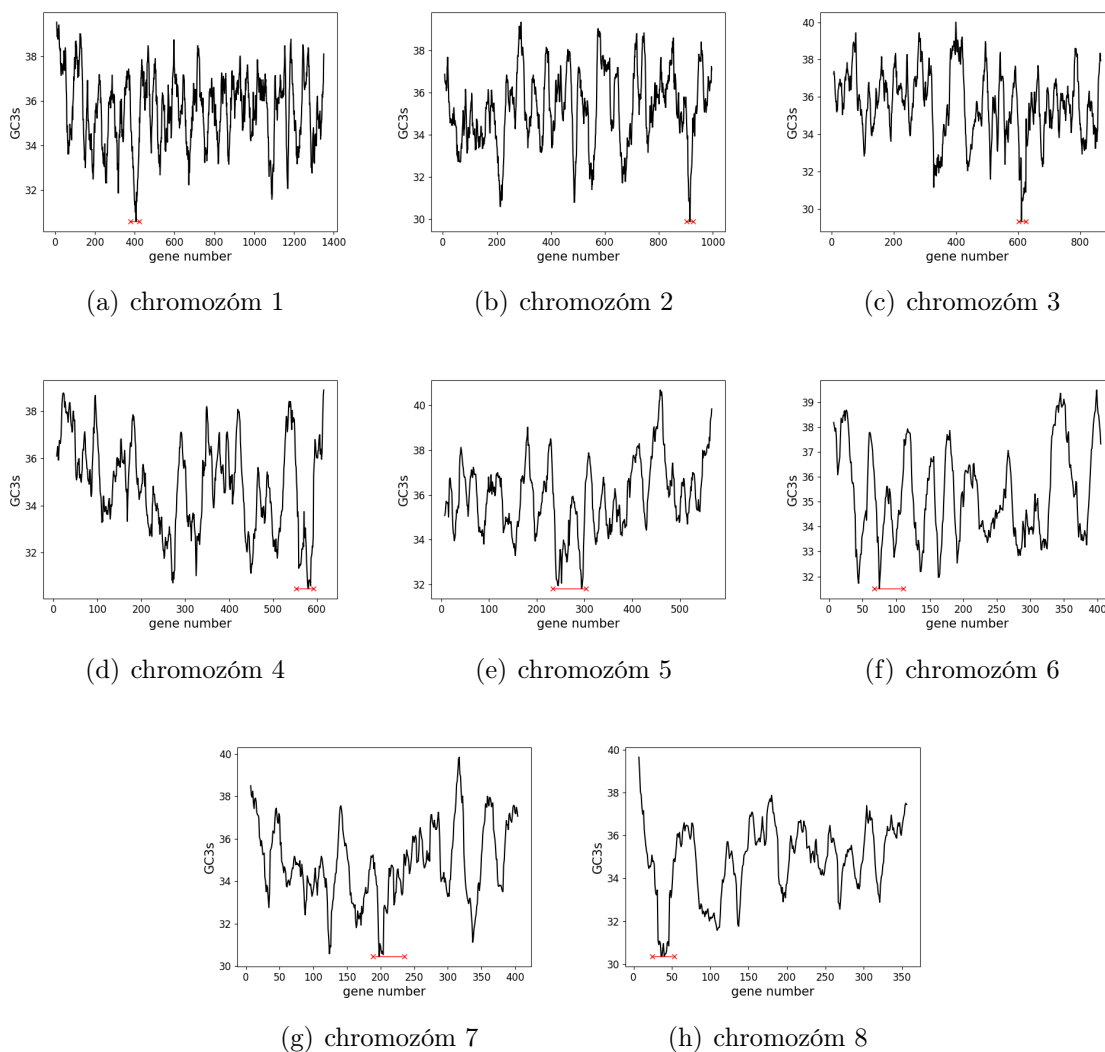
V analýze genómovej sekvencie tejto kvasinky (Obr. 4.13) je pomerne ťažké identifikovať centromerický lokus na všetkých chromozómoch, čo môže zodpovedať zisteniam Denise Lynchovej a kol. (2010) [18], že GC3 metóda nie je univerzálne aplikovateľná. Našou hypotézou je, že to môže byť spôsobené aj extrémne nízkym celkovým zastúpením GC.

4.3.6 *Saprochaete ingens*

S. ingens je nepatogénna kvasinka z čeľade *Dipodascaceae* z rodu *Magnusiomyces*. Používa štandardný genetický kód. Veľkosť jej genómu je 21,2 Mb a je organizovaný v piatich chromozómoch. Kvasinka *S. ingens* kmeňa CBS 517.90 obsahuje 6 475 CDS a zastúpenie GC je 36,5% [20].

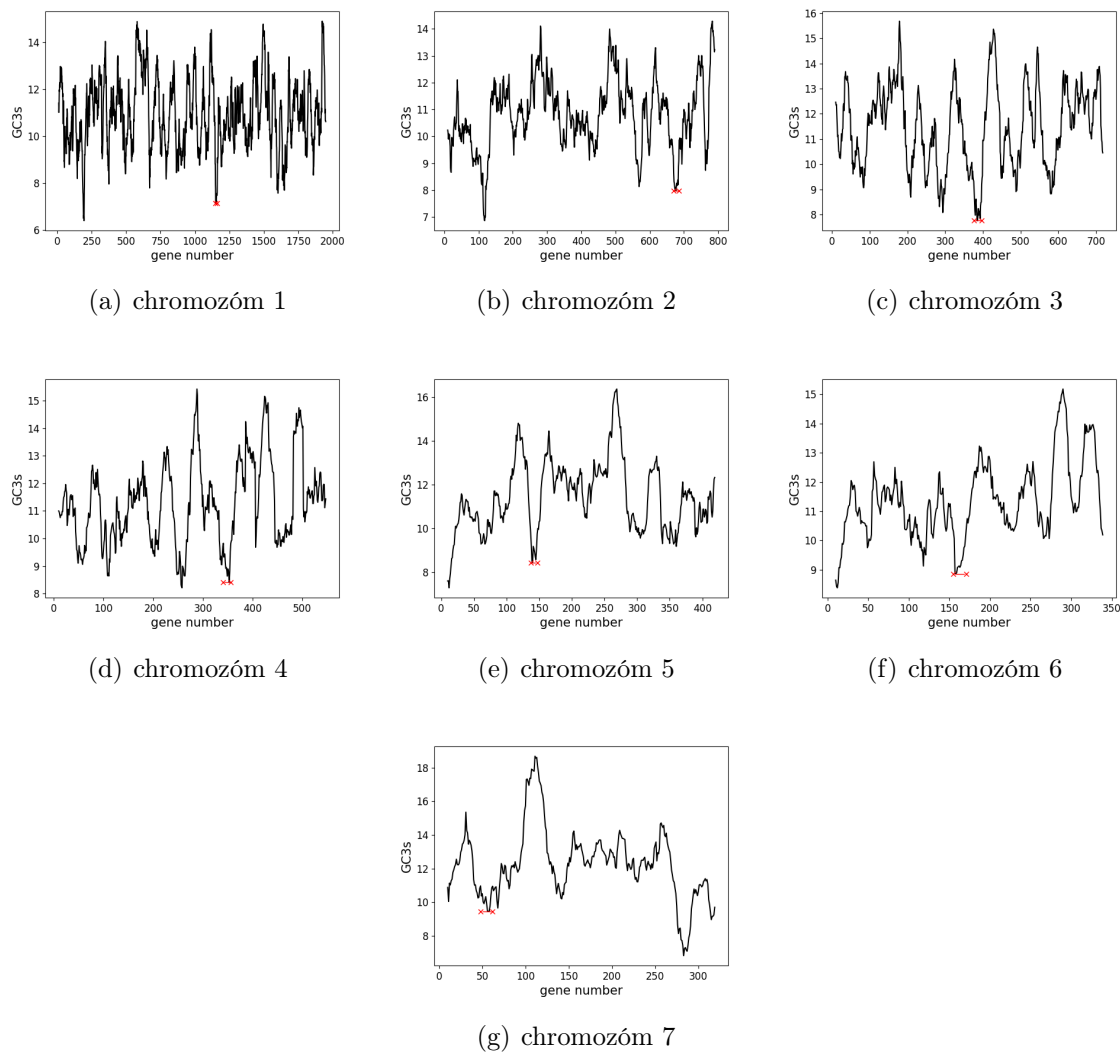
Na grafoch zastúpenia GC3 pre *S. ingens* (Obr. 4.14) je centromerická oblasť dobre viditeľná pre chromozóm 2. Na každom chromozóme kleslo najvýraznejšie lokálne minimum pod 29%, kým zastúpenie GC3 sa v ostatných oblastiach chromozómov pohybovalo okolo 35-40%. Na chromozóme 3 je viditeľných viacero oblastí s výrazným poklesom GC3. Na chromozómoch 4 a 5 nie je centromerická oblasť až taká výrazná. Pre výpočty bolo použité okno s veľkosťou 15 génov.

Na tejto kvasinke je možné podporiť argumenty o vplyve veľkosti zvoleného okna. Pri použití okna 30 génov sa grafy (Obr. 4.15) zmenili a algoritmus predikoval centro-



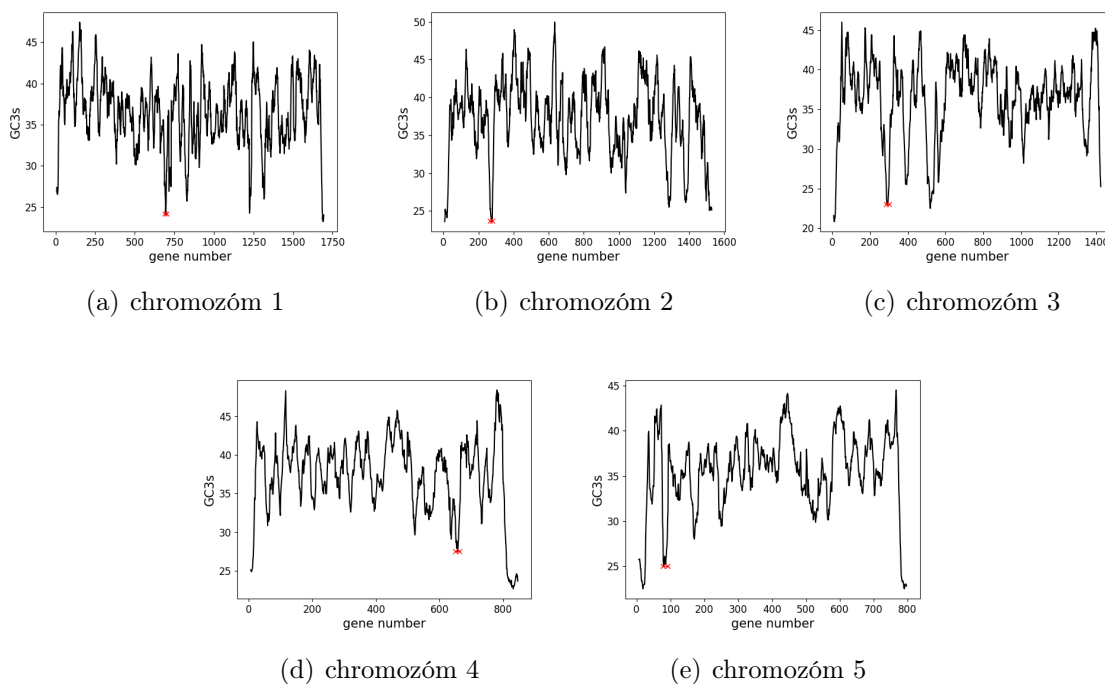
panel	chromozóm	gén pred centromérou	pozícia	gén za centromérou	pozícia
(a)	1	CANPARB_P03890	794500	CANPARB_P04360	899176
(b)	2	CANPARB_P23590	1994694	CANPARB_P23820	2057943
(c)	3	CANPARB_P30850	1415444	CANPARB_P31100	1486991
(d)	4	CANPARB_P39450	1302976	CANPARB_P39850	1418296
(e)	5	CANPARB_P42590	564213	CANPARB_P43310	733994
(f)	6	CANPARB_P46800	151964	CANPARB_P47260	254595
(g)	7	CANPARB_P52370	452598	CANPARB_P52830	555096
(h)	8	CANPARB_P54880	60274	CANPARB_P55190	146472

Obr. 4.12: Analýza distribúcie GC3 pozdĺž chromozómov 1 až 8 pre *C. parapsilosis* CA-JEJE01 a tabuľka génov pred a po centromére pre jednotlivé chromozómy



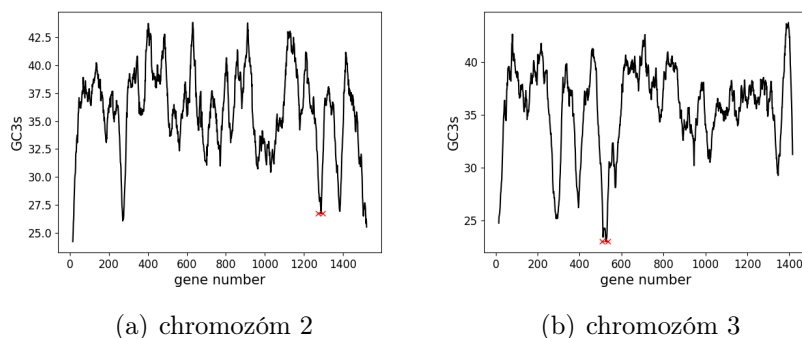
panel	chromozóm	gén pred centromérou	pozícia	gén za centromérou	pozícia
(a)	1	CANVERA_P1187	2688634	CANVERA_P1203	2722815
(b)	2	CANVERA_P2723	1483279	CANVERA_P2741	1516342
(c)	3	CANVERA_P3247	957220	CANVERA_P3268	998583
(d)	4	CANVERA_P3960	844205	CANVERA_P3977	884498
(e)	5	CANVERA_P4327	287246	CANVERA_P4337	311177
(f)	6	CANVERA_P4786	348705	CANVERA_P4802	381167
(g)	7	CANVERA_P5038	94627	CANVERA_P5053	135221

Obr. 4.13: Analýza distribúcie GC3 pozdĺž chromozómov 1 až 7 pre *C. verbasci* CBS 12699 a tabuľka génov pred a po centromére a ich vonkajších koordinát pre jednotlivé chromozómy.



panel	chromozóm	gén pred centromérou	pozícia	gén pred centromérou	pozícia
(a)	1	SAPINGB_P000711	2419952	SAPINGB_P000721	2451791
(b)	2	SAPINGB_P002014	911556	SAPINGB_P002030	959392
(c)	3	SAPINGB_P003607	988579	SAPINGB_P003625	1055323
(d)	4	SAPINGB_P005453	2321909	SAPINGB_P005466	2371433
(e)	5	SAPINGB_P005737	237560	SAPINGB_P005752	324302

Obr. 4.14: Analýza distribúcie GC3 pozdĺž chromozómov 1 až 5 pre *S. ingens* CBS 517.90 a tabuľka génov pred a po centromére a ich vonkajších koordinát pre jednotlivé chromozómy.



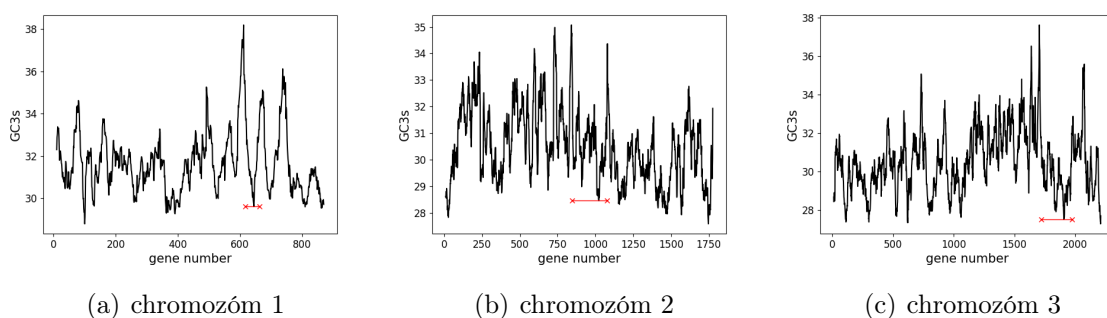
Obr. 4.15: Analýza distribúcie GC3 pozdĺž chromozómov 2 a 3 pre *S. ingens* CBS 517.90 s použitím veľkosti okna 30 génov.

méru na chromozómoch 2 a 3 do iných oblastí ako pri použití okna 15 génov.

4.3.7 *Schizosaccharomyces pombe*

S. pombe je haploidná kvasinka patriaca do čeľade *Schizosaccharomycetaceae* s troma chromozómami. Veľkosť genómu tejto kvasinky z kmeňa 972h- je 12,6 Mb, pričom obsahuje 5 131 proteín-kódujúceich génov a celkové zastúpenie GC je 36% [20].

Ako môžeme z GC3 profilov *S. pombe* (Obr. 4.16) vidieť, táto kvasinka má chromozómy pomerne veľké, preto pre analýzu bolo použité okno s veľkosťou 20 génov. Ani na jednom chromozóme nie je predikovaná oblasť veľmi výrazná, čo tiež podporuje hypotézu z východiskovej štúdie [18]. Tiež vidíme, že náš algoritmus predikoval centromerický lokus do väčšej oblasti ako pri ostatných kvasinkách.



panel	chromozóm	gén pred centromérou	pozícia	gén pred centromérou	pozícia
(a)	1	SPAC2C4.10c	4277153	SPAPB8E5.03	4913613
(b)	2	SPBC14C8.10	2220763	SPBC32F12.02	2787266
(c)	3	SPCC191.10	1721828	SPCC285.17	1832175

Obr. 4.16: Analýza distribúcie GC3 pozdĺž chromozómov 1 až 5 pre *S. pombe* 972h- a tabuľka génov pred a po centromére a ich vonkajších koordinát pre jednotlivé chromozómy.

Kapitola 5

Diskusia

Identifikácia centromerických lokusov na základe nukleotidového zastúpenia v genómových sekvenciách kvasiniek má mnoho výhod. Po porovnaní predikcii centromerického lokusu metódou GC3 s experimentálnymi zisteniami môžeme túto metódu považovať za perspektívnu. V prvom rade ide o čiste bioinformatickú analýzu, ktorá môže predchádzať experimentálnym analýzám a tým výrazne uľahčiť a urýchliť hľadanie centromér v chromozónoch.

Z nášho pohľadu je ďalšou veľkou výhodou tejto metódy možnosť grafickej reprezentácie. Na GC3 profiloch je možné predikovať centromerický lokus a rozpoznať rôzne iné charakteristiky chromozómov, ako napríklad extrémne nízke alebo vysoké zastúpenie GC3 pozdĺž celého chromozómu, variabilita, výkyvy a rýchlosť zmien zastúpení GC3 a tiež môžeme predikovať kryptické sekvencie, ktoré môžu potencionálne reprezentovať novo-vznikajúce centroméry.

Taktiež sú v analýze využívané rôzne parametre, ktoré sú variabilné a používateľ si ich môže upraviť podľa vlastných výskumných cieľov.

Pozoruhodnou hypotézou vyplývajúcou z výsledkov analýzy viacerých chromozómov *C. parapsilosis* a napríklad aj chromozómu 6 *C. auris* sú viaceré oblasti s výrazne nízkym zastúpením GC3 na profiloch. Našou hypotézou je, že niektoré z nich sú kryptické, resp. novo-vznikajúce centroméry, ktoré by mohli nahradiť pôvodnú centroméru, čo by mohlo byť predmetom ďalšieho výskumu.

Z našich výsledkov môžeme povedať, že najlepšie viditeľné centromerické oblasti sa nachádzajú v genómových sekvenciách kvasiniek *C. lusitaniae* a *C. auris*. V oboch je pokles GC3 veľmi výrazný. Z toho vyplýva hypotéza, že GC3 metóda by mohla fungovať aj pre ďalšie druhy kvasiniek z čeľade *Metschnikowiaceae*.

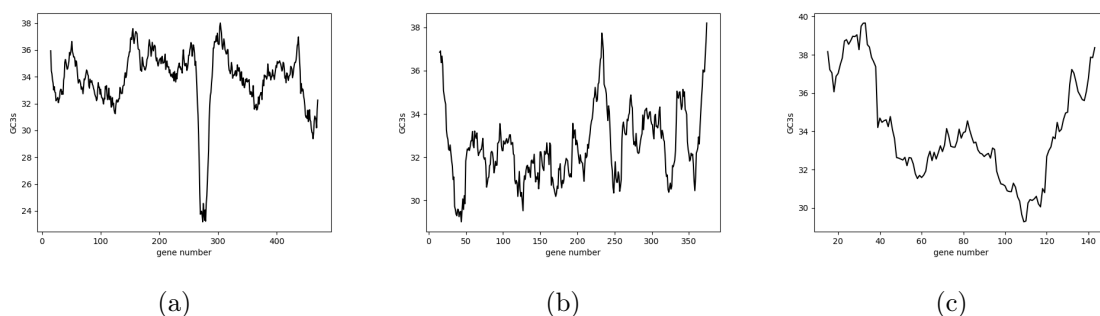
Nevýhodou tejto metódy je, že nie je aplikovateľná univerzálne. Presný dôvod zatiaľ nie je známy. Denise Lynchová a kol. (2010) [18] na *C. albicans* a *C. dubliniensis* tiež potvrdili, že GC3 metóda nie je použiteľná pre genómové sekvencie všetkých druhov kvasiniek. V tejto práci predpokladáme, že pre kvasinky s príliš nízkym zastúpením

GC (pod 30%) táto metóda nefunguje. Našou ďalšou hypotézou je, že aplikovateľnosť GC3 metódy sa netýka rôznych druhov kvasiniek, ale samotných chromozómov. Táto hypotéza vychádza z analýz GC3 profilov *C. metapsilosis* a *C. parapsilosis*, kde sme mali možnosť vidieť, že na niektorých grafoch bol centromerický lokus identifikovateľný, na iných nie.

Ďalším dôležitým predpokladom pre funkčnosť GC3 metódy sú dobre spracované dáta. Sekvencie vo FASTA formáte musia byť poskladané na celé chromozómy. V opačnom prípade je ťažké identifikovať centromerickú oblasť. Znamenalo by to, že v niektorých GC3 profiloch by nebolo možné označiť najvýraznejšie lokálne minimum ako centromerický lokus.

Celkovo vieme genómové sekvencie rozdeliť do troch skupín. V prvej (pre ktorú táto metóda môže fungovať) sú genómové sekvencie poskladané na celé chromozómy. V druhej sa nachádzajú vysoko súvislé sekvencie, no nie všetky sú poskladané na celé chromozómy. Tretia skupina pozostáva z vysoko fragmentovaných sekvencií, ktoré nie sú pre GC3 metódu vhodné. Príkladom druhej kategórie môže byť kvasinka *Lodderomyces elongisporus* z čeľade *Debaryomycetaceae*, ktorej genómová sekvencia je poskladaná na 11 kontigov.

Pre ilustráciu prikladáme tri GC3 profily kontigov *L. elongisporus* (Obr. 5.1). Prvý kontig prevdepodobne pokrýva časť chromozómu, v ktorej sa nachádza centromerická oblasť, ktorá je dobre viditeľná a pokles GC3 je veľmi výrazný. Vyplýva z toho hypotéza, že pre *L. elongisporus* by mohla GC3 metóda fungovať veľmi dobre. Z druhého kontigu nie je jasné, či sa tam centromerická oblasť nachádza. Tretí kontig je príliš krátky - cca 150 génov, čo pre analýzu nie sú použiteľné dáta.

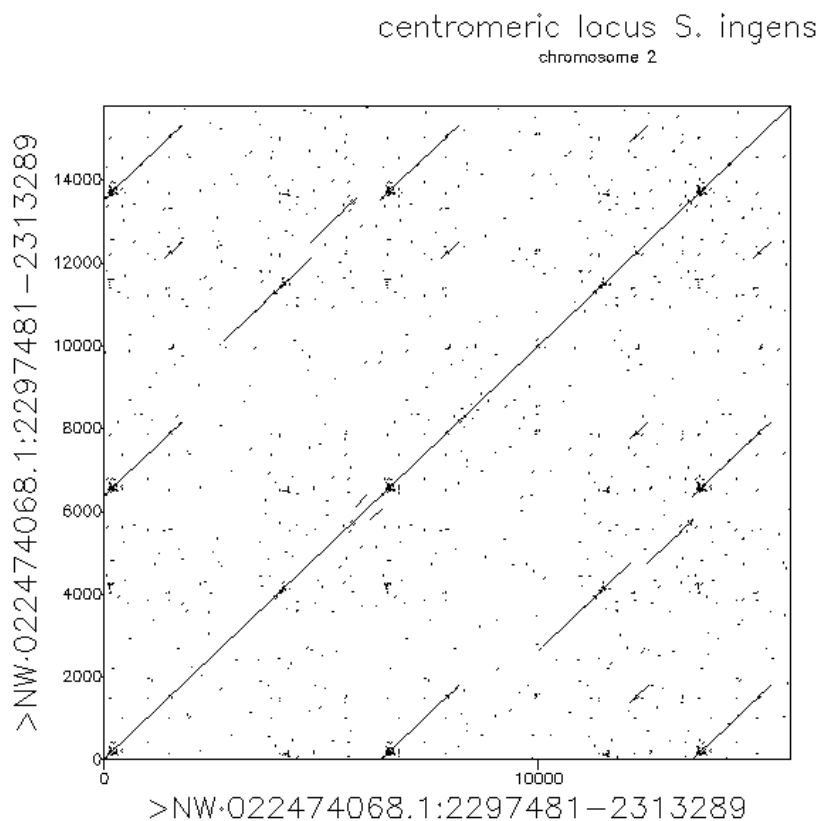


Obr. 5.1: Analýza distribúcie GC3 troch kontigov genómu kvasinky *L. elongisporus*.

V rámci detailnejšej analýzy centromerických lokusov je možné bioinformatickými nástrojmi podložiť predikcie centromér získané GC3 metódou. Jednou z možností je hľadanie repetícií, ktoré sú pre centromerické oblasti typické. Príkladom výstupu tejto analýzy by mohol byť tzv. *dotplot*.

Pre ilustráciu prikladáme dotplot (Obr. 5.2) centromerického lokusu predikovaného

pomocou GC3 metódy v genómovej sekvencii kvasinky *S. ingens*. Nukleotidovú sekvenciu sme z .fasta súboru vybrali pomocou *samtools* [16]. Dotplot bol zhotovený pomocou on-line nástroja *EMBOSS* [13]. Na x-ovej aj y-ovej je centromerický lokus toho istého chromozómu. Pomocou dotplotu je možné určiť dlhšie repetitívne oblasti, ktoré sú znázornené súvislými čiarami.



Obr. 5.2: Dotplot centromerickej sekvencie chromozómu 2 pre *S. ingens* [13].

Ďalšou možnosťou je hľadanie retrotranspozónov v sekvenciách v blízkosti centromér. Expresia v centromerických oblastiach býva potlačená a preto sa u niektorých kvasiniek retrotranspozóny vyskytujú v centromerických lokusoch [22].

Zaujímavým experimentom by bolo porovnanie sekvencie a pozície predikovaných centromerických oblastí medzi rôznymi príbuznými druhmi kvasiniek. Na základe týchto analýz by sme mohli zistiť, nakoľko sú tieto oblasti medzi druhmi konzervované a akým spôsobom sa centromerické lokusy v evolúcii vyvíjali.

Predpokladáme, že GC3 metóda by bola aplikovateľná aj na genómové sekvencie iných eukaryotických organizmov.

Záver

GC3 metóda predikuje centromerickú oblasť jednotlivých chromozómov na základe nukleotidového zastúpenia v rôznych druhoch kvasiniek. Prvým cieľom tejto práce bola implementácia GC3 metódy a overenie jej funkčnosti. Na kvasinkách *C. lusitanae*, *S. stipitis* a *Y. lipolytica*, pre ktoré už centromerická oblasť bola predikovaná a experimentálne overená [5], sme ukázali funkčnosť našej implementácie.

V druhej časti práce sme sa venovali optimalizácii jednotlivých parametrov. Ukázali sme, že veľkosť okna výrazne vplýva na presnosť GC3 metódy. Príliš malé okno spôsobí v GC3 profiloch šum, ktorý sťažuje identifikáciu centromerického lokusu, kým príliš veľké okno mapuje centromerický lokus na príliš veľkú alebo dokonca odlišnú oblasť chromozómu a tým znižuje presnosť výpočtu. Zároveň sme zistili, že vylúčenie niektorých kodónov má vplyv na zastúpenie GC3, no presnosť predikcie centromerického lokusu bola zasiahnutá len minimálne. V GC3 profiloch bol viditeľný posun krivky smerom nahor pozdĺž y-osi.

V tretej časti práce sme uviedli výsledky aplikácie našej implementácie GC3 metódy na genómové sekvencie kvasiniek, pre ktoré ešte centromerická oblasť nebola predikovaná. V GC3 profiloch chromozómov niektorých kvasiniek, napríklad *C. auris*, je centromerický lokus dobre viditeľný, no pre iné kvasinky, napríklad *S. pombe* je centromerická oblasť veľmi ťažko identifikovateľná. Tým sme ukázali, že GC3 metóda nie je univerzálne aplikovateľná na genómové sekvencie všetkých druhov kvasiniek. Môže to byť spôsobené extrémne nízkym zastúpením GC pozdĺž chromozómov alebo neštandardnou rekombináciou v danom druhu kvasiniek.

Limitáciou tejto metódy je tiež kvalita genómovej sekvencie a anotácia. Genómová sekvencia musí byť vysoko súvislá, pretože pri fragmentovaných genómových sekvenciách je veľmi ťažké určiť centromerickú oblasť.

Literatúra

- [1] Dennis A Benson, Mark Cavanaugh, Karen Clark, Ilene Karsch-Mizrachi, David J Lipman, James Ostell, and Eric W Sayers. Genbank. *Nucleic acids research*, 41(D1):D36–D42, 2012.
- [2] Sue Biggins. The composition, functions, and regulation of the budding yeast kinetochore. *Genetics*, 194(4):817–846, 2013.
- [3] Biopedia. Chromozómové mutácie. [Online; visited 5. 2. 2023].
- [4] Leslie Boghosian-Sell, Reema Mewar, Wilbur Harrison, Raymond M Shapiro, Elaine H Zackai, John Carey, Laura Davis-Keppen, Louanne Hudgins, and Joan Overhauser. Molecular mapping of the Edwards syndrome phenotype to two noncontiguous regions on chromosome 18. *American journal of human genetics*, 55(3):476, 1994.
- [5] Mingfeng Cao, Arun Somwarpet Seetharam, Andrew Josef Severin, and Zengyi Shao. Rapid isolation of centromeres from *Scheffersomyces stipitis*. *ACS Synthetic Biology*, 6(11):2028–2034, 2017.
- [6] Louise Clarke and John Carbon. Isolation of a yeast centromere and construction of functional small circular chromosomes. *Nature*, 287(5782):504–509, 1980.
- [7] Guy Cochrane, Ilene Karsch-Mizrachi, Toshihisa Takagi, and International Nucleotide Sequence Database Collaboration. The international nucleotide sequence database collaboration. *Nucleic acids research*, 44(D1):D48–D50, 2016.
- [8] Python Software Foundation.
- [9] Tatsuo Fukagawa and William C Earnshaw. Neocentromeres. *Current Biology*, 24(19):R946–R947, 2014.
- [10] Charles R Harris, K Jarrod Millman, Stéfan J Van Der Walt, Ralf Gommers, Pauli Virtanen, David Cournapeau, Eric Wieser, Julian Taylor, Sebastian Berg, Nathaniel J Smith, et al. Array programming with NumPy. *Nature*, 585(7825):357–362, 2020.

- [11] Johannes H Hegemann and Ursula N Fleig. The centromere of budding yeast. *Bioessays*, 15(7):451–460, 1993.
- [12] Diana Hernandez and Elizabeth MC Fisher. Down syndrome genetics: unravelling a multifactorial disorder. *Human molecular genetics*, 5(Supplement_1):1411–1416, 1996.
- [13] Hinxton Cambridge CB10 1SA UK. Ian Longden formerly at: Sanger Institute, Wellcome Trust Genome Campus. EMBOSS: dotmatcher.
- [14] Rasko Leinonen, Ruth Akhtar, Ewan Birney, Lawrence Bower, Ana Cerdeno-Tárraga, Ying Cheng, Iain Cleland, Nadeem Faruque, Neil Goodgame, Richard Gibson, et al. The European nucleotide archive. *Nucleic acids research*, 39(suppl_1):D28–D31, 2010.
- [15] Yann Lesecque, Dominique Mouchiroud, and Laurent Duret. GC-biased gene conversion in yeast is specifically associated with crossovers: molecular mechanisms and evolutionary significance. *Molecular biology and evolution*, 30(6):1409–1419, 2013.
- [16] Heng Li, Bob Handsaker, Alec Wysoker, Tim Fennell, Jue Ruan, Nils Homer, Gabor Marth, Goncalo Abecasis, Richard Durbin, and 1000 Genome Project Data Processing Subgroup. The sequence alignment/map format and SAMtools. *Bioinformatics*, 25(16):2078–2079, 2009.
- [17] Thomas Lindblad, Jason M Kinser, and JG Taylor. *Image processing using pulse-coupled neural networks*. Springer, 2005.
- [18] Denise B Lynch, Mary E Logue, Geraldine Butler, and Kenneth H Wolfe. Chromosomal G+C content evolution in yeasts: systematic interspecies differences, and GC-poor troughs at centromeres. *Genome biology and evolution*, 2:572–583, 2010.
- [19] Marie-Claude Marsolier-Kergoat. Models for the evolution of GC content in asexual fungi *candida albicans* and *c. dubliniensis*. *Genome Biology and Evolution*, 5(11):2205–2216, 2013.
- [20] NCBI. The National Center for Biotechnology Information.
- [21] Sergey Nurk, Sergey Koren, Arang Rhie, Mikko Rautiainen, Andrey V Bzikadze, Alla Mikheenko, Mitchell R Vollger, Nicolas Altemose, Lev Uralsky, Ariel Gershman, et al. The complete sequence of a human genome. *Science*, 376(6588):44–53, 2022.

- [22] Mihaela Ola, Caoimhe E O’Brien, Aisling Y Coughlan, Qinx Ma, Paul D Donovan, Kenneth H Wolfe, and Geraldine Butler. Polymorphic centromere locations in the pathogenic yeast *Candida parapsilosis*. *Genome Research*, 30(5):684–696, 2020.
- [23] Nuala A O’Leary, Mathew W Wright, J Rodney Brister, Stacy Ciufo, Diana Haddad, Rich McVeigh, Bhanu Rajput, Barbara Robbertse, Brian Smith-White, Danso Ako-Adjei, et al. Reference sequence (RefSeq) database at NCBI: current status, taxonomic expansion, and functional annotation. *Nucleic Acids Research*, 44(D1):D733–D745, 2016.
- [24] Geo Pertea and Mihaela Pertea. GFF Utilities: GffRead and GffCompare [version 1; peer review: 2. 2020.
- [25] Eugénie Pessia, Alexandra Popa, Sylvain Mousset, Clément Rezvoy, Laurent Duret, and Gabriel AB Marais. Evidence for widespread GC-biased gene conversion in eukaryotes. *Genome Biology and Evolution*, 4(7):675–682, 2012.
- [26] Simon L Peyton Jones and Philip Wadler. Imperative functional programming. In *Proceedings of the 20th ACM SIGPLAN-SIGACT symposium on Principles of programming languages*, pages 71–84, 1993.
- [27] Miroslav Plohl, Nevenka Meštrović, and Brankica Mravinac. Centromere identity from the DNA point of view. *Chromosoma*, 123:313–325, 2014.
- [28] Martin G Reese, Barry Moore, Colin Batchelor, Fidel Salas, Fiona Cunningham, Gabor T Marth, Lincoln Stein, Paul Flicek, Mark Yandell, and Karen Eilbeck. A standard variation file format for human genome sequences. *Genome biology*, 11:1–9, 2010.
- [29] Ann Reynolds, Victoria Lundblad, David Dorris, and Marie Keaveney. Yeast vectors and assays for expression of cloned genes. *Current protocols in molecular biology*, 39(1):13–6, 1997.
- [30] Nicolas P Rougier. *Matplotlib tutorial*. PhD thesis, INRIA, 2012.
- [31] Conrad L Schoch, Stacy Ciufo, Mikhail Domrachev, Carol L Hotton, Sivakumar Kannan, Rogneda Khovanskaya, Detlef Leipe, Richard Mcveigh, Kathleen O’Neill, Barbara Robbertse, et al. NCBI Taxonomy: a comprehensive update on curation, resources and tools. *Database*, 2020, 2020.
- [32] Paul M Sharp, Michele Stenico, John F Peden, and Andrew T Lloyd. Codon usage: mutational bias, translational selection, or both? *Biochemical Society Transactions*, 21(4):835–841, 1993.

- [33] Peter D Snustad, Michael J Simmons, Jiřina Relichová, Jiří Doškař, Jiří Fajkus, Petr Hořín, Aleš Knoll, Petr Kuglík, Jan Šmarda, Jana Šmardová, et al. *Genetika*. Masarykova univerzita, 2017.
- [34] Yoshio Tatenó, Tadashi Imanishi, Satoru Miyazaki, Kaoru Fukami-Kobayashi, Naruya Saitou, Hideaki Sugawara, and Takashi Gojobori. DNA Data Bank of Japan (DDBJ) for genome scale research in life science. *Nucleic acids research*, 30(1):27–30, 2002.
- [35] Nelle Varoquaux, Ivan Liachko, Ferhat Ay, Joshua N Burton, Jay Shendure, Maitreya J Dunham, Jean-Philippe Vert, and William S Noble. Accurate identification of centromere locations in yeast genomes using Hi-C. *Nucleic acids research*, 43(11):5331–5339, 2015.
- [36] Jacob White. PubMed 2.0. *Medical reference services quarterly*, 39(4):382–387, 2020.
- [37] Wikipedia. Centrální dogma molekulární biologie — Wikipedia, the free encyclopedia, 2004. [Online; accessed 22-October-2022].
- [38] Wenjuan Zhang, Natalya Lukoyanova, Shomon Miah, Jonathan Lucas, and Cara K Vaughan. Insights into centromere DNA bending revealed by the cryo-EM structure of the core centromere binding factor 3 with Ndc10. *Cell reports*, 24(3):744–754, 2018.

Príloha A: údaje o použitých dátach

druh kvasinky	číslo assembly	databáza	dátum
<i>Candida auris</i>	ASM301371v2	RefSeq	10/2019
<i>Candida lusitaniae</i>	ASM383v1	RefSeq	7/2009
<i>Candida metapsilosis</i>	BP57	RefSeq	4/2021
<i>Candida orthopsilosis</i>	ASM31587v1	RefSeq	4/2012
<i>Candida parapsilosis</i>	CAJEJE01	GenBank	10/2021
<i>Candida verbasci</i>	canVerA	GenBank	12/2022
<i>Saprochaete ingens</i>	sapIngB	RefSeq	9/2019
<i>Scheffersomyces stipitis</i>	ASM20916v1	RefSeq	3/2007
<i>Schizosaccharomyces pombe</i>	ASM294v2	RefSeq	7/2022
<i>Yarrowia lipolytica</i>	ASM176148v1	GenBank	10/2016