
# VYSOKÉ UČENÍ TECHNICKÉ V BRNĚ
BRNO UNIVERSITY OF TECHNOLOGY

## FAKULTA ELEKTROTECHNIKY A KOMUNIKAČNÍCH TECHNOLOGIÍ
FACULTY OF ELECTRICAL ENGINEERING AND COMMUNICATION

## ÚSTAV BIOMEDICÍNSKÉHO INŽENÝRSTVÍ
DEPARTMENT OF BIOMEDICAL ENGINEERING


# METODY PRO VYLEPŠENÍ GENOMOVÉHO SESTAVENÍ ZALOŽENÉ NA SIGNÁLOVÉM ZPRACOVÁNÍ
SIGNAL PROCESSING BASED METHODS FOR GENOME ASSEMBLY REFINEMENT

**DIPLOMOVÁ PRÁCE**
MASTER'S THESIS


**AUTOR PRÁCE** — **Bc. Robin Jugas**
AUTHOR

**VEDOUCÍ PRÁCE** — **Mgr. Ing. Karel Sedlář**
SUPERVISOR



**BRNO 2016**

VYSOKÉ UČENÍ TECHNICKÉ V BRNĚ
FAKULTA ELEKTROTECHNIKY A KOMUNIKAČNÍCH TECHNOLOGIÍ

# Diplomová práce

magisterský navazující studijní obor **Biomedicínské inženýrství a bioinformatika**
Ústav biomedicínského inženýrství

***Student:*** Bc. Robin Jugas ***ID:*** 147488
***Ročník:*** 2 ***Akademický rok:*** 2015/16

**NÁZEV TÉMATU:**

## Metody pro vylepšení genomového sestavení založené na signálovém zpracování

**POKYNY PRO VYPRACOVÁNÍ:**

1) Zpracujte literární rešerši o základních přístupech využívaných pro sestavení genomu (assembly) ve vztahu k různým sekvenačním technologiím. 2) Prostudujte možnosti využití numerické reprezentace DNA (genomického signálu) pro vylepšení neúplných genomových sestavení. 3) Pro vhodně zvolenou/zvolené numerické reprezentace navrhněte metodu pro vyhledávání překryvů mezi sekvencemi. 4) Celou metodu, včetně nástroje pro převod sekvence do podoby signálu, implementujte ve vhodně zvoleném programovacím jazyce. 5) Funkčnost metody otestujte na vhodně vytvořeném souboru dat a statisticky vyhodnoťte její úspěšnost ve vztahu k různým délkám překryvů. 6) Metodu srovnejte se stávajícími algoritmy.

**DOPORUČENÁ LITERATURA:**

[1] UTTURKAR, S. M., et al. Evaluation and validation of de novo and hybrid assembly techniques to derive high-quality genome sequences. Bioinformatics. 2014, 30(19): 2709-2716.

[2] WALKER, B. J., et al. Pilon: An Integrated Tool for Comprehensive Microbial Variant Detection and Genome Assembly Improvement. PLoS ONE. 2014, 9(11): e112963.

***Termín zadání:*** 8.2.2016 ***Termín odevzdání:*** 20.5.2016

***Vedoucí práce:*** Mgr. Ing. Karel Sedlář
***Konzultant diplomové práce:***

**prof. Ing. Ivo Provazník, Ph.D.**, *předseda oborové rady*

## ABSTRAKT

Diplomová práce se zabývá sekvenačními metodami a metodami sestavení genomu s využitím numerických reprezentací. Teoretická část práce popisuje historii objevu DNA, jednotlivé generace sekvenačních metod, samotné metody sestavení genomu a definici numerických reprezentací. Numerické reprezentace slouží pro převod znakové podoby DNA do numerické podoby a umožňují tak využití metod digitálního zpracování signálu. V práci je navržen algoritmus pro sestavení genomu s využitím numerické reprezentace, který je dále otestován na datech sekvencí.

## KLÍČOVÁ SLOVA

bioinformatika, DNA, sestavení genomu, sekvenace, numerické reprezentace, neúplná genomová sestavení

## ABSTRACT

The diploma thesis deals with sequencing methods and genome assembly methods including usage of numerical representations. The theoretical part of thesis describes the history of DNA research, generations of sequencing methods, the assembly methods themselves and definiton of numerical representations. Numerical represenatations serve to convert character form of DNA to numerical form and so allow to use digital signal processing methods. There is an algorithm for genome assembly using numerical represenatation proposed in thesis, which is later tested at sequence data.

## KEYWORDS

bioinformatics, DNA, genome assembly, sequencing, numerical representations, draft assembly

JUGAS, Robin *Metody pro vylepšení genomového sestavení založené na signálovém zpracování*: diplomová práce. Brno: Vysoké učení technické v Brně, Fakulta elektrotechniky a komunikačních technologií, Ústav biomedicínského inženýrství, 2016. 65 s. Vedoucí práce byl Mgr. Ing. Karel Sedlář,

# PROHLÁŠENÍ

Prohlašuji, že svou diplomovou práci na téma „Metody pro vylepšení genomového sestavení založené na signálovém zpracování“ jsem vypracoval(a) samostatně pod vedením vedoucího diplomové práce a s použitím odborné literatury a dalších informačních zdrojů, které jsou všechny citovány v práci a uvedeny v seznamu literatury na konci práce.

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

Brno ..... .....

podpis autora(-ky)

# PODĚKOVÁNÍ

Rád bych poděkoval vedoucímu diplomové práce panu Mgr. Ing. Karlu Sedláři za odborné vedení, konzultace, trpělivost a podnětné návrhy k práci.


Brno ..... .....

podpis autora(-ky)

# OBSAH

# SEZNAM OBRÁZKŮ

# SEZNAM TABULEK

# ÚVOD

Růst v oblasti vědy a techniky se dotýká všech jejich oblastí a přesto, že jsme značně pokročili v našich možnostech zkoumání jsou před námi stále nepokořené překážky. Bioinformatika prodělala rychlý vývoj a stále se vyvíjí. Vývoj na poli zpracování DNA sekvencí napříč spektrem od medicínského využití, zkoumání funkce genů a expresi informací z DNA, vyvíjel tlak na „těžbu“ těchto dat - sekvenaci DNA a následné genomové sestavení. Současně se nacházíme v pomyslné třetí generaci sekvenátorů a jsme schopni osekvenovat velké množství dat za dostupnou cenu. Blížíme se k ceně 1000 amerických dolarů za osekvenování celého lidského genomu při 30× pokrytí. Přesto je v sekvenaci a následném sestavení genomu (assembly) mnoho výzev - chyby sekvenace, výpočetní náročnost sestavení genomu, opakující se oblasti DNA atd.

Nosným tématem práce je využití numerických reprezentací při sestavení genomu. Jsme zvyklí chápat DNA čistě jako sled znaků reprezentující nukleotidy, ale bylo již uvedeno několik prací zabývajících se převodem znakové sekvence DNA do numerické podoby. Na DNA se tak můžeme dívat jiným úhlem pohledu a aplikovat odlišné metody známé ze zpracování digitálních signálů. V této práci se zaměřím na využití numerických sekvenci pro neúplná sestavení genomu.

Text práce je rozdělen na několik kapitol zabývajících se hierachicky zadanou problematikou. První část práce bude teoretická rešerše oblastí týkajících se tématu práce. V první kapitole bude stručně popsána historie objevu DNA, která sahá až k pokusům Johanna Gregora Mendela s křížením zde v Brně. Na to naváže stručný popis informací o struktuře a funkci deoxyribonukleové kyseliny.

V druhé kapitole budou prostudovány sekvenační metody DNA jakožto nosný pilíř pro téma práce. Věnoval bych se historickému vývoji od prvních náročných metod až po dnešní strojově prováděné metody sekvenace třetí generace sekvenátorů.

V navazující kapitole by mělo být probráno téma samotného sestavení genomu, tzn. assemblery a jejich popis. Důležité bude zmínit vhodnost využití dané metody sestavování genomu pro konkrétní sekvenační metody. Stručně budou zmíněny okrajová témata, jež s procesem sekvenace a sestavení genomu souvisí.

Numerické reprezentace a princip jejich funkce bude probrán a demonstrován na ukázkách skutečných sekvencí. V současné době je numerických reprezentací k dispozici vícero, ale ne všechny budou vhodné pro téma práce.

Dále budou probrány teoretické aspekty fungování a využití numerických reprezentací pro konkrétní aplikaci v dokončování neúplných genomových sestaveních. Bude se diskutovat volba konkrétní numerické reprezentace a další možné parametry.

V další částí práce bych předvedl své řešení zadaného problému na základě předchozích poznatků. V praktické části bude algoritmus otestován na různých konkrétních sekvencích. Chtěl bych dosáhnout výsledků srovnatelných se znakovými metodami, ale ideálně při kratším čase zpracování.

# 1 DNA

## 1.1 Objev DNA

Historie výzkumu DNA (deoxyribonukleové kyseliny), její chemické vlastnosti a biologické funkce, začala dříve než byl učiněn slavný objev struktury DNA Francisem Crickem a Jamesem Watsonem v roce 1953. [50] Vzhledem k funkci DNA, jakožto nositele dědičnosti, je na počátku výzkumu práce Johanna Gregora Mendela a jeho zákony dědičnosti, které definoval po pokusech s hráchem setým v augustiniánském klášteře v Brně. [35] I přesto, že byla jeho práce na 30 let zapomenuta, vzbudila zájem o genetiku.

Molekula DNA byla objevena v 60. letech 19. století švýcarským chemikem Fridrichem Miescherem. Mladý student Miescher na začátku roku 1869 prováděl v laboratoři Felixe Hoppe-Seylera na univerzitě v německém Tübingenu expremienty s leukocyty, které vedly k objevu DNA. Tübingen byl tehdy centrem přírodních věd a sídlila zde první německá samostatná Fakulta přírodních věd. Mieschera k přírodním vědám vedla rodina - otec i strýc byly uznávanými fyziky a profesory anatomie a fyziologie a již od mládí byl v kontaktu s různými vědci. Sám vystudoval medicínu, ale věnoval se dále studiu chemie. Vedoucí laboratoře Felix Hoppe-Seyler, patřil mezi zakladatele biochemie a intenzivně se zabýval studiem krve a krevních buněk. Proto se i Miescher věnoval studiu leukocytů, které získával z nemocničních obvazů. Cílem bylo vyvinout efektivní metodu k izolaci protoplasmy z jádra, aby mohl dále studovat obsah jádra. Během těchto pokusů však objevoval precipitát neznámých vlastností. Šlo o sraženinu DNA. Pojmenoval ji jako nuclein, protože pocházela z jádra. Během další práce během které se snažil získat čistší sraženinu bez proteinů objevil další vlastnosti DNA, například vyšší obsah fosforu oproti proteinům, obsah dusíku a nerozložitelnost pepsinem. Byl si jist, že objevil látku odlišnou od proteinů. Jeho objev byl důležitý pro další práci na prokázání dědičnosti nesené nukleovými kyselinami. Sám Miescher uvažoval nad teorií, že by dědičnost byla založená na chemické bázi. To se prokázalo až později. [16]

V roce 1928 zveřejnil svůj experiment Frederick Griffith. V něm definoval tzv. proces transformace. Griffith se zabýval výzkumem vakcíny proti španělské chřipce. Používal různé kmeny pneumokoka Streptococcus pneumoniae, III-S virulentní a druhý II-R nevirulentní. Virulentní bakterie byla zlikvidována teplem a přidána k druhému kmeni. Pokud byly aplikovány zvlášť, pokusná myš nebyla infikována, pokud však byly zkombinovány, propukla nemoc. Griffith byl schopen z krve pokusné myši izolovat zvlášť oba kmeny. Definoval tak princip transformace, kdy se jeden kmen transformoval v druhý. Tímto principem byla DNA virulentního kmenu, která přežila zahřátí a exprimovala svoje geny v druhém kmeni. [19]

Molekula DNA jakožto transformační princip byla prokázana během navazujícího Avery-MacLeod-McCarty experimentu. Použili škálu analytických metod k určení DNA jakožto nositele dědičnosti. Také provedli pokus, kdy ke kmenům bakterií přidali ribonuklázy a proteázy, štěpící proteiny a RNA, přesto však transformační efekt nebyl deaktivoaán. [4]

V roce 1919 biochemik Phoebus Levene uveřejnil poznatky o nukleových kyselinách, tvořené sledem nukleotidů z nichž každý je složen z jedné dusíkové báze, molekuly cukru a fosfátové skupiny. Leven byl první, kdo objevil strukturu nukleotidu tvořeného bází, fosfátem a cukrem, objevil ribózu a deoxyribózu a fosfodiesterovou vazbu mezi molekulami sacharidů v DNA. Nejznámější je pro svou již překonanou tetranukleotidovu teorii, která předpokládala DNA jako opakující se tetramer. Teorie vyplývala z poznatků, že se čtyři nukleotidové báze vyskytují v rovnoměrném množství. Nadále se však za molekulu dědičnosti pokládaly proteiny, neboť takováto DNA nebyla dostatečná k nesení komplexní genetické informace. [8, 16]

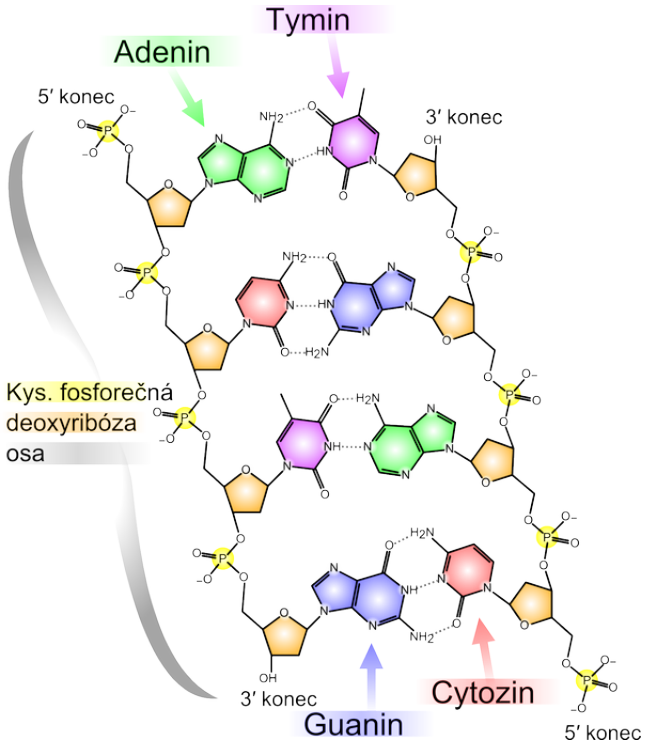
Studiu DNA se věnoval rakouský biochemik Erwin Chargaff, jež definoval tzv. Chargaffova pravidla. V 40. letech 20. století objevil vzorec mezi množstvím nukleotidových bazí. Zjistil, že množství thyminu se rovná množství adeninu, a že množství guaninu se rovná množství cytosinu. Dále zjistil, že počty nukleotidů se mezi živočišnými druhy liší. Objasnil tak párování bazí a naznačil, že molekulou dědičnosti je DNA. [10]

Nejdůležitější objev, který v podstatě shrnul předchozí poznatky učinili James Watson a Francis Crick, když v roce 1953 uveřejnili svůj model dvoušroubovice DNA. Navázali na pozorování DNA rentgenovými paprsky vědci Rosalind Franklinové a Maurice Wilkinse. Dva řetězce DNA mají opačné směrování, jsou značeny jako 3' a 5' konce a směr syntézy DNA je od 5' k 3' konci. Hydroxylová skupina na 3' konci se pojí k fosfátu na 5' konci. Je uvolněn pyrofosfát a nová báze utvoří fosfodiesterovou vazbu. Tato reakce je katalyzována DNA polymerázou. Odpovídající si báze jsou mezi řetězci spojeny vodíkovou vazbou. Párování bazí odpovídá Chargaffovým pravidlům. [50]

Ačkoliv byla již známa struktura DNA, rozluštění posloupnosti bazí v DNA sekvenci a její funkce při vytváření proteinů přišlo až se zrodem sekvenace v 50. letech 20. století. V roce 1958 Francis Crick publikoval práci, kde se věnoval právě uspořádání DNA sekvence jež determinovala sekvenci aminokyselin v proteinech, což umožnilo dále studovat funkce proteinů. [51] Zrodu sekvenace a jejím metodám je věnována druhá kapitola.

## 1.2 Struktura DNA

DNA (deoxyribonukleová kyselina) je nositelem genetické informace. Jde o biopolymer, využívá stavebnicový princip, složitější struktury jsou složené z jednodušších monomerů. Monomerem DNA je nukleotid, který je složen ze sacharidu pentózy, kyseliny fosforečné a organické dusíkaté báze, jichž existuje 5 typů (adenin, tymin, uracil, cytozin, guanin). Nukleotidy jsou k sobě navázány do lineárního řetězce vazbou mezi kyselinou fosforečnou nukleotidu a pátým atomem pentózy druhého nukleotidu. Primární struktura molekuly DNA je tvořena dvěma antiparalelními polynukleotidovými řetězci. Osu polynukletidového řetězce tvoří střídající se pentóza a fosfát, báze nukleotidů tvoří pomyslné "žebřiny"a obě báze se k sobě váží vodíkovými můstky dle komplementarity bazí. Adenin se váže s tyminem a guanin s cytozinem. Pro výzkum má stěžejní význam sekvence bazí nukleotidů v molekula DNA. Sekundární struktura DNA je tvořena stočenou dvoušroubovicí, tato konformace je stabilizována vodíkovými bázemi a vzniká samovolně jako stav s nejmenší volnou energií (na obrázku 1.1). Pro získání sekvence DNA používáme techniky sekvenace. [37]




Obr. 1.1: Struktura DNA (Zdroj: Wikipedia)

# 2 SEKVENACE DNA

## 2.1 Sekvenační techniky - 1. generace

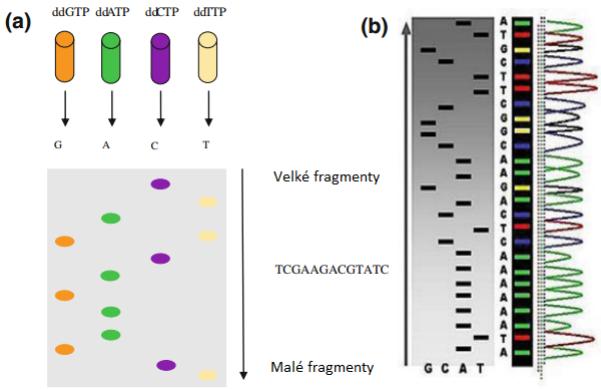
Objevem struktury a funkce deoxyribonukleové kyseliny začala nová éra biologie a medicíny. Rozšířilo se poznání biologických zákonitostí a jevů, rozvinuly se nové vědní obory a objasnění některých tajemství lidského života. To vše bylo umožněno i díky sekvenaci DNA. Díky ní bylo možné objevovat obsah DNA organismů a objasňovat genovou funkci DNA. První generace sekvenačních technik DNA tvořily dvě mětody vyvinuté nezávisle ve stejné době - metoda Fredericka Sangera (1977) a Allana Maxama a Waltera Gilberta (1977). [33, 34]

### 2.1.1 Sangerova sekvenace

Sangerova sekvenace, známá též jako „dideoxy metoda“, přinesla svému vynálezci již druhou Nobelovu cenu za chemii v roce 1980 a stala se všeobecně populární metodou sekvenace. Postupně byla inovována a automatizována. Pomocí Sangerovy metody byl proveden projekt sekvenace kompletního lidského genomu v roce 2001.

Původní sekvenace probíhá na jednovláknové DNA (templát) za přítomnosti DNA primerů, DNA polymerázy, čtyř běžných nukleotidů (deoxynucleosidtrifosfát) a jednoho ze čtyř dideoxynukleotidů (ddATP,ddCTP,ddGTP,ddTTP). Dideoxynukleotidy nemají funkční OH skupinu pro tvorbu fosfodiesterové vazby a jsou radioaktivně značeny. Sekvenace probíhá ve čtyřech oddělených reakcích, které se liší přítomností jednoho specifického dideoxynukleotidu. Daného dideoxynukleotidu je v reakci stokrát méně než daného nukleotidu. Využívá se replikace a amplifikace DNA pomocí PCR, DNA je degradována na jednovláknovou, DNA primer značí začátek sekvenovaného úseku od kterého probíhá syntéza DNA. Za přítomnosti DNA polymerázy se tvoří komplementární vlákna, která jsou na různých místech ukončena náhodně daným dideoxynukleotidem. Z každé reakce máme knihovnu fragmentů, které analyzujeme gelovou elektroforézou a autoradiografií. Znázornění procesu je na obrázku 2.1.

Metoda byla postupně modifikována, díky použití fluorescentního značení na různých vlnových délkách může reakce probíhat v jedné zkumavce a je analyzována kapilární elektroforézou. Jsou dostupná komerční řešení - sekvanátory, které zpracují výslednou analýzu a kompletní sety pro sekvenaci. Výsledná čtení jsou velmi dlouhá, mohou mít i více než 1000 bazí. [33, 40]




Obr. 2.1: Sangerova sekvenace (Zdroj: [33])

### 2.1.2 Maxam-Gilbertova sekvenace

Maxam-Gilbertova metoda vznikla v roce 1977 na Harvardské univerzitě společnou prací Waltera Gilberta a jeho doktoranda Allana Maxama. Spolu s Frederickem Sangerem dostal Gilbert v roce 1980 Nobelovu cenu za chemii.

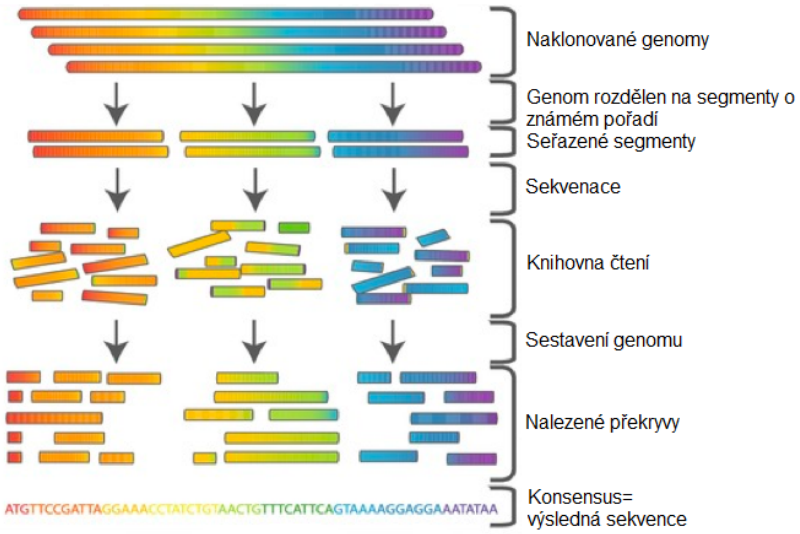
Metoda spočívá ve štěpení nukleotidů chemickými sloučeninami a je nejvhodnější pro menší nukleotidové řetězce. Vzorek sekvenované DNA je purifikován a na 5' konci je označen radioaktivním fosforem. Metoda pracuje s dvouvláknovou i jednovláknovou DNA. Dále metoda využívá přidávání chemikálií, štěpící sekvenci nukleotidů na úseky pomocí čtyř reakcí (G,A+G,C,C+T). Využívá se dvoustupňové katalytické reakce a dvou látek: dimethylsulfát štěpící puriny a hydrazin štěpící pyrimidiny. V prvním kroku buď puriny reagují s dimethylsulfátem nebo pyrimidiny s hydrazinem za rozložení glykosidové vazby mezi bazí a pentózou. V druhém kroku piperidin katalyzuje rozpad fosfodiesterové vazby na místě, kde byla odebrána báze. Vzniku fragmentů je dosaženo čtyřmi chemickými reakcemi: kyselina mravenčí depurinuje puriny (A+G), dimethylsulfát metyluje guanin (G), hydrazin hydrolyzuje pyrimidiny (C+T) a navázání soli k hydrazinové reakci (C). Koncetrace těchto látek je regulována, aby došlo průměrně k jedné modifikaci na DNA molekulu. Každá reakce probíhá v odděleném procesu. Úsek DNA je rozdělen reagenty na fragmenty. Knihovna fragmentů je analyzována gelovou elektroforézou na základě velikosti fragmentů a vizualizována pomocí autoradiografie. Výsledkem je znaková sekvence DNA. Metoda je oproti Sangerově metodě složitější a používané chemikálie jsou toxické, proto se prosadila v praxi více Sangerova metoda. [33, 34]

### 2.1.3 Příprava DNA pro sekvenaci

Téměř u všech sekvenačních metod je nutné připravit si DNA pro následnou sekvenaci. U DNA je nutné vytvořit si knihovnu fragmentů. Postup se liší v závislosti na

sekvenační metodě. Klasická Sangerova metoda neumožňuje sekvenaci DNA fragmentů delších jak 1000 bp z důvodu využití kapilární elektroforézy a nízké dělicí schopnosti mezi fragmenty. Větší sekvence, například celý chromozom, musí být před sekvenováním rozděleny na menší části - fragmenty a poté amplifikovány pro dosažení přijatelného množství kopií. Tento proces lze provést dvěmi postupy: technikou „whole-genome-shotgun“ (obrázek 2.3) a „map-based“ technikou (též hierarchické sekvenování) (obrázek 2.2). [33]

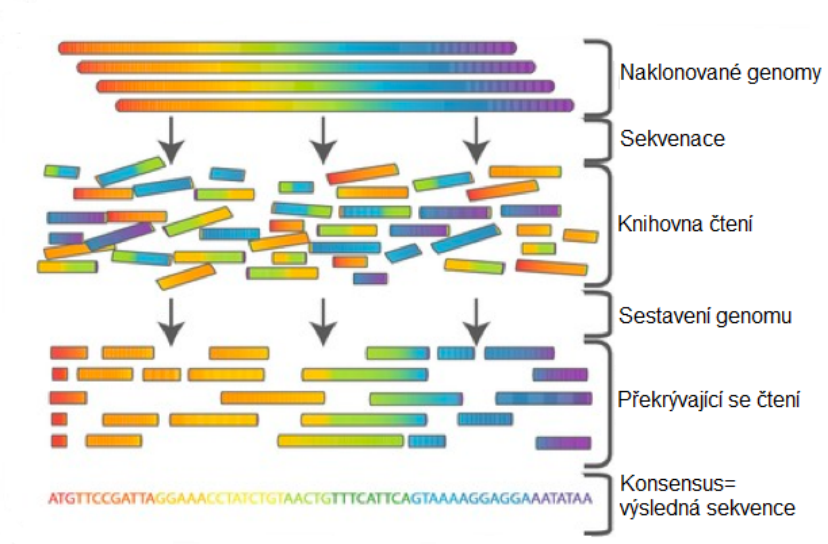
Map-based technika využívá bakteriálních plazmidů a chromozomů. Amplifikovaný úsek chromosomu je rozdělen náhodně na větší části (50 - 200 kb), jež jsou naklonovány do umělých bakteriálních chromosomů (BAC) o velikostech větších než 20 000 bází. Fragmenty v klonech se překrývají s určitým překryvem (coverage), aby bylo možné složit celý úsek. Poté, co je bakteriální chromozom amplifikován, je rozdělen na menší fragmenty velikosti 2-3 kb. Tyto jsou naklonovány do plazmidových vektorů a opět amplifikovány. Až teď je z nich extrahována DNA pro sekvenaci. Vysoké pokrytí nebo převzorkování každého fragmentu je vyžadováno k seskládání fragmentů do kontigů, vyšších souvislých celků z fragmentů, jež pak tvoří výslednou kompletní sekvenci. [33]




Obr. 2.2: Hierarchické sekvenování (Zdroj: Wikipedia)

U techniky whole-genome-shotgun je DNA sekvence náhodně rozstříhána do krátkých fragmentů a po přidání adaptorů k fragmentům je provedena sekvenace. Sestavení původní sekvence je u shotgun techniky složitější, protože nejsou známé žádné informace o relativní pozici fragmentů v chromosomu. [33]

K překování tohoto problému se používá metoda zpárovaných konců (paired-end nebo mate-pair), u které jsou oba konce DNA fragmentů osekvenovány. Celá geno-




Obr. 2.3: Whole-genome-shotgun sekvenování (Zdroj: Wikipedia)

mová DNA je rozstříhána na konkrétní velikost (např. 1500 bp) a fragmenty jsou pomocí dvou adaptorů převedeny na cirkulární DNA. V kruhovém fragmentu jsou oba konce původní lineární DNA vedle sebe. V dalším kroku se pomocí enzymu nukleázy rozdělí kruhový fragment na dvou místech, tak že vzniknou dva lineární fragmenty s rozdílnou délkou a jeden z nich obsahuje oba konce původního fragmentu a druhý fragment obsahuje báze mezi původními konci. Takto vzniklá knihovna fragmentů může být sekvenována a vzniklá data umožní „scaffolding“ - sestavení čtení do konečné sekvence obsahující i mezery známé délky. Obě metody jsou synonymní, ale liší se v délce čtení. Metoda mate-pair generuje čtení o délce 2-5 kb, zatímco paired-end má čtení o délce cca 500 bp. [33, 5]

## 2.2 Sekvenační techniky - NGS

Vzhledem k růstu odvětví bioinformatiky, genomiky a potřeby sekvenace celých genomů, se postupně prosadily nové přístupy v sekvenaci. Na rozdíl od dřívějších metod, již nejsou spojeny se jménem vynálezce, ale iniciativy ve vývoji sekvenačních technik se chopily soukromé firmy. Mezi nejvýznamnější firmy v odvětví patří Roche, Illumina a Life Technologies. Společným znakem nové generace sekvenačních technik je masivní paralelizace. Společně se označují pojmem „Next-generation sequencing“. [33, 44]

### 2.2.1 454 Pyrosekvenace

První technikou nové generace je metoda 454 pyrosekvenování vyvinutá firmou Life Science (USA) v roce 2005, později koupena firmou Roche. Firma nabízí dva typy sekvenátorů: GS Junior+ o délce čtení 700 bp a GS FLX+ o délce čtení 1000 bp [43]. Využívá alternativní sekvenační techniky pyrosekvenace využívající podobně jako Sangerova metoda koncept sekvenace syntézou. Liší se však v tom, že nezávisí na terminaci syntézy DNA dideoxynukleotidem, ale využívá detekci uvolnění pyrofosfátu uvolněného při začlenění nukleotidu DNA polymerázou do struktury DNA. Začlenění nukleotidu (deoxynukleotid trifosfátu - dNTP) , tedy vytvoření fosfodiesterové vazby a vodíkového můstku, má za následek vyloučení pyrofosfátu, který je za přítomnosti adenosin-5'-fosfátu přeměněn na ATP, reakce je katalyzována enzymem ATP sulfurylázou. Molekula ATP slouží jako substrát pro chemiluminiscentní enzym luciferázu, která přemění luciferin na oxyluciferin za vzniku viditelného záření. Toto záření je detekováno CCD čipem. Enzym apyráza odstraní nepoužité nukleotidy a ATP, aby se proces mohl opakovat. Jde tedy o využití čtyř enzymů - DNA polymerázy, ATP sulfurylázy, luciferázy a apirázy. [33, 44]

V procesu sekvenace je nejdříve genomická DNA nastříhána na kratší fragmenty a degradována na jednovláknovou, k tupým koncům ($3'$ a $5'$) DNA jsou ligovány krátké adaptory (A+B). Fragmenty s adaptory jsou smíchány s malé kuličky streptavidinu o velikosti $28\mu m$. Povrch kuliček je pokryt krátkými sekvencemi komplementárními s jedním adaptorem. Fragmenty jsou hybridizovány tak, že každá kuličky nese specifický fragment. Pro zesílení signálu z CCD čipu se provede amplifikace fragmentů s použitím emulzní PCR. Emulzní PCR obsahuje nutné reagencie v kapce vody, výsledkem je kulička pokrytá amplifikovanými fragmenty. Kuličky jsou umístěny na pikotitrační destičku, kde každá kulička zapadne do jedné jamky. Následně se přidají enzymy pro pyrosekvenování (sulfuryláza a luciferáza) a deoxynukleotidy. DNA polymeráza začne syntetizovat komplementární vlákna k fragmentům a CCD čip snímá světelný signál. Zesílení signálu je úměrné počtu nukleotidů. Výsledek je zakreslen do grafu (flowgram). [33, 44]

### 2.2.2 Illumina (Solexa)

Sekvenátor firmy Solexa byl uveden v roce 2006 a dále byl vyvíjen firmou Illumina. Knihovna fragmentů může být sestavena jakoukoliv metodou, která využívá ligaci adaptorů. Amplifikace templátových fragmentů je provedena metodou můstkové PCR, u níž jsou oba primery komplementární k adaptorům navázány k destičce. Pojem můstková PCR je odvozen od skutečnosti, že během dosednutí primerů (annealing) se oba primery spojí do pomyslného mostu odkud jsou syntetizována komplementární vlákna. Amplifikované fragmenty tvoří shluky, každý zhruba o milionu

kopií původního fragmentu. Každý shluk obsahuje dopředné i reverzní vlákna, ale s cílem mít homogenní shluk vláken se jeden typ odstraní. Metoda používá sekvenaci syntézou. K reakci jsou připojeny blokující nukleotidy, které zastaví syntézu DNA. Po navázání každého nukleotidu je pořízen obrazový snímek. Po akvizici snímků v každém cyklu, blokující nukleotidy jsou odstraněny a může následovat další cyklus. Firma Solexa nabízí několik platforem sekvenátorů Illumina: NextSeq,HiSeq, HiseqX (všechny produkují čtení $2 \times 150$ bp) a MiSeq (čtení $2 \times 300$ bp). [33, 44, 22]

### 2.2.3 SOLiD

Sekvenátory SOLiD byly komerčně dostupné roku 2007, nejdříve technologii zakoupila firma Applied Biosystems a posléze Life Technologies. Podobně jako u dalších metod nové generace sekvenátorů je zde vytvořena knihovna fragmentů s adaptory. Příprava DNA je podobná s metodou pyrosekvenování 454. Ligované fragmenty jsou denaturovány a navázány na magnetickou kuličku s emulzní PCR. Postup sekvenace je už odlišný. Kuličky jsou umístěny na destičku. Narozdíl od syntézové sekvenace, je zde sekvenace ligací skrze polymeázovou reakci. K fragmentům je hybridizován univerzální primer, komplementární s adaptorem. V každém cyklu probíhá hybridizace oligonukleotidu (oktameru), který obsahuje 2 specifické nukleotidy, 3 libovolné nukleotidy, 3 reverzibilní terminátory označené fluorescenční značkou. Pokud se shoduje dvojice nukleotidů na templátu a oktameru, proběhne hybridizace po které následuje ligace. Ostatní oktamery jsou vymyty. V tomto kroku je pořízen obraz (je detekováno fluorescenční záření specifické pro každý oktamer). Poslední 3 báze oktameru s fluorescenční značkou jsou odstraněny a zůstane jen 5 bazí. Tento krok se opakuje asi desetkrát. V dalším kroku je počáteční primer a všechny ligované části odstraněny a opakuje se celý proces s novým primerem o n-1 nukleotidech. Dekódování probíhá po dinukleotidech. Protože je zde 16 možných kombinací dinukleotidů, ale jen 4 fluorescenční vrstvy, identifikace nukleotidů není možná jen na základě takto získaných dat. Kompozice prvních dvou nukleotidů je snadno odvoditelná, protože známe předchozí bázi na primeru, který známe. Alternativně je možné použít oktamery s odlišnými pozicemi dinukleotidů. K dispozici je sekvenátor 5500W s maximální délkou čtení 75 bazí. [33, 44, 32]

### 2.2.4 Ion Torrent Sekvenace

Sekvenace metodou Ion Torrent je odlišná od všech metod nové generace sekvenátorů. Nepoužívají se žádné enzymatické reakce, žádní chemické značení, žádné optické systémy ani světlo. Využívá jednoduché chemické reakce - vyloučení vodíkového iontu po vzniku fosfodiesterové vazby. K dosažení paralelizace se používá matici

s jamkami, pod každou jamkou je čip ISFET (Ion-sensitive field effect transistor), který detekuje změny pH, které následují po vyloučení vodíkového iontu. Změna pH je zaznamenána jako změna napětí $\Delta V$. Příprava DNA probíhá stejně jako u pyrosekvenace 454. Proces zachycení probíhá takto: k DNA templátu se přidá jeden nukleotid, pokud je začleněn DNA polymerázou do DNA, uvolní se vodíkový iont a změní se pH, které je detekováno senzorovou vrstvou. V případě dvou stejných následujících nukleotidů bude změna napětí dvojnásobná. Pokud není detekována žádná změna napětí, nukleotid se nezačlenil. Metoda patří mezi nejjednodušší a nejlevnější, na platformě Ion S5 firmy ThermoFisher dosahuje délky čtení až 200 bp nebo až 400 bp v závislosti na použitém čipu. [33, 44, 46]

## 2.3 Třetí generace

Po druhé generaci, která se vyznačovala paralelizací a společným přístupem „wash-and-scan“ v podobě sledování syntézy DNA po nukleotidech. Pro třetí generaci je typický „single molecule“ přístup (sekvenace na úrovni molekul a vynechání PCR před čtením. Mezi dvě nejslibnější patří technologie sekvenace Nanopore a Single Molecule Real Time (SMRT) v provedení od firmy Pacific Biosciences (zkráceně PacBio sekvenátor). [33, 44, 41]

### 2.3.1 Single Molecule Real-Time Sequencing

Metoda byla vyvinuta firmou Pacific Biosciences v roce 2009. Stojí na pozorování DNA polymerázy během syntézy DNA. Jsou využity speciální čipy - „SMRT cell“, jež obsahují tisíce tzv. „zero-mode waveguides“ (ZMW), což jsou jamky o nano rozměrech, sloužící k zachycení světla o vlnové délce větší než je průměr jamky (desítky nm). Do každé jamky je umístěna jedna molekula DNA polymerázy. Přesnost a rychlost syntézy polymerázou je závislá na koncentracích nukleotidů a protože jsou nukleotidy fluorescentně značeny, povedou jejich vysoké koncentrace k šumu pozadí. Proto byl detekční objem SMRT čipu snížen, aby byl šum potlačen. Oproti ostatním metodám se liší způsobem značení fosforem, u jiných metod fluorescentní značka zůstává po syntéze navázána přímo na bázi nukleotidu, což opět vede k šumu, tak u SMRT metody je fluorescentní značka navázána na fosfátový řetězec nukleotidu a po tvorbě fosfodiesterové vazby bude odstraněna a vymyta z detekčního objemu. Přidaný nukleotid je pak epifluorescenčně detekován. [33, 44]

### 2.3.2 Nanopore

Sekvenace nanopórem byla představena v roce 1995. Společným znakem těchto metod je detekce nukleotidů v jednovláknové DNA procházejícím skrze tenkou membránu. Výhodou je možnost dosažení dlouhých čtení. Nejdále se ve vývoji dostala metoda firmy Oxford Nanopore Technologies. U této metody není třeba složité přípravy DNA podobně jako u Ion Torrentové sekvenace. Není třeba stříhat, amplifikovat a značit DNA fragmenty. Jako póry slouží bílkovinné póry, například $\alpha$-hemolyzin, který produkuje zlatý stafylokok. Na membránu je přiveden potenciál. Průběh vlákna DNA skrze pór má za následek změny proudu iontů. Přesun každé báze skrze nanopór způsobuje úbytek intenzity proudu, který je specifický pro každou bázi. V současné době je i vyřešeno řešení problému, kdy DNA vlákno prochází přes pór příliš rychle, což snižuje rozlišení detekce nukleotidových bazí. Metoda je stále ještě experimentální, ale mohla by dosahovat vysokých rychlostí. [33, 44]

Tab. 2.1: Srovnání sekvenátorů - zdroje: [43, 22, 46, 32, 33]

| **Sekvenátor** | **Délka čtení** | **Počet čtení** | **Doba** |
|---|---|---|---|
| Applied Biosystems 3500 (Sanger) | >1000 | | |
| Roche 454 GS Junior+ | 700 | $1 \times 10^5$ | 18 h. |
| Roche 454 GS FLX+ | 1000 | $1 \times 10^6$ | 23 h. |
| Illumina MiSeq | $2 \times 300$ | $15 \times 10^9$ | 56 h. |
| Illumina HiSeq2500 | $2 \times 250$ | $300 \times 10^9$ | 60 h. |
| SOLiD LifeTech 5500xl | 75 | $15 \times 10^9$ | 1 d. |
| Ion Torrent PGM 316 | >100 | $1 \times 10^6$ | 2 h. |
| Ion Torrent S5 | 200/400 | $0,6 - 10 \times 10^9$ | 2,5/4 h. |
| PacBio SMRT RS | >1000 | $1 \times 10^5$ | 1,5 h. |
| Oxford Nanopore | >100 000 | | |

## 2.4 Kvalita sekvenace - Phred skóre

Phred skóre (nebo Q skóre) se používá k posouzení kvality výstupu sekvenátoru. Indikuje pravděpodobnost s jakou by konkrétní báze mohla být nesprávně přiřazena sekvenátorem. U Sangerovy sekvence vyplývalo Phred skóre z parametrů Sangerovy sekvenace - rozlišení píku a tvaru křivky v chromatografu, kterým přiřazovalo již známé hodnoty přesnosti. Ačkoliv metody nové generace sekvenátorů pracují s

odlišnými principy, proces stanovení Phred skóre zůstal podobný. Parametry sekvenačního postupu jsou porovnány s empiricky získanými daty o známé přesnosti. Phred skóre je definováno rovnicí :

$$Q = -10 \log_{10} P \tag{2.1}$$

, kde $P$ je pravděpodobnost chybného přiřazení báze. Phred skóre jsou logaritmicky přiřazena k pravděpodobnostem chyby. Pokud je bázi přiřazeno skóre 30 (Q30), je to ekvivalentní pravděpodobnosti chyby 1:1000. To znamená, že přesnost přiřazení báze je $99,9\%$. Nižší skóre 20 (Q20) by značilo chybu 1:100, tedy čtení o 100 bazích obsahuje pravděpodobně 1 chybu. Pokud však skóre sekvenace dosahuje Q30, teoreticky by měly být všechny čtení bez chyb. Proto se Phred skore 30 používá jako měřítko (benchmark) pro posuzování kvality sekvenace. [20]

## 2.5 Formát souborů čtení - FASTQ

Stručně bude ještě zmíněna otázka struktur pro ukládání sekvencí získaných sekvenátory. Tyto data dále slouží pro sestavení genomu assemblery. Nejčastější formát je FASTQ, jež má několik variant (Sanger, Solexa, Illumina), jež se liší zápisem identifikátoru a užitým skóre kvality sekvenace. [11]

```
@SEQ_ID
GATTTGGGGTTCAAAGCAGTATCGATCAAATAGTAAATCCATTTGTTCAACTCACAGTTT
+
!''*((((***+))%%%++)(%%%%).1***-+*''))**55CCF>>>>>>CCCCCCC65
```

Obr. 2.4: FASTQ soubor

Standardní FASTQ (fastq-sanger) je textový soubor, kde každému čtení (sekvenci) odpovídá zápis na 4 řádky. První řádek začíná speciálním znakem '@', za kterým následuje identifikace sekvence a její popis. Na druhém řádku je samotná sekvence. Třetí řádek začíná znakem '+', za kterým může být volitelný popis sekvence. Čtvrtý řádek obsahuje informace o kvalitě a počet znaků se musí rovnat počtu znaků na druhém řádku. Ukázka zápisu souboru je na obrázku 2.4. Kvalita sekvenované báze je definována několika znaky: například '!' značí nejnižší kvalitu a naopak ' ' značí nejvyšší kvalitu (znaky ASCII 33-126 pro kódování kvality Phred skórem - rozsah 0-93). [11]

Varianta používaná softwarem sekvenátoru Illumina (fastq-illumina) používá na prvním řádku specifický systémový identifikátor. Řádek kvality kóduje Phred skórem

ASCII znaky 64-126 (rozsah 0-62), protože využívá odlišné mapování Phred skóre. [11]

# 3 SESTAVENÍ GENOMU

## 3.1 De novo assembly

Cílem sestavení genomu (ang. assembly) je ze čtení získaných sekvenací (ang. reads) složit původní sekvenci DNA, kterou jsme sekvenovali. Pokud tuto sekvenci vůbec neznáme, mluvíme o sestavení genomu „de novo“. Vychází z předpokladu, že mezi jednotlivými čteními jsou překryvy (ang. overlaps), které identifikujeme a jsme tak schopni složit výslednou sekvenci. Pokud už sekvenci známe, máme referenční sekvenci, jednotlivé čtení zpravidla jen mapujeme (či přiřazujeme) k již známé sekvenci a v podstatě jde o zarovnání sekvencí. Jde o výpočetně jednodušší problém než sestavení genomu de novo. I mapování však musí počítat s chybami sekvenace. [33, 36]

Většina sekvenátorů poskytuje čtení o délce desítek až tisíc bází. Pokud sekvenujeme celý genom, například lidský genom má přes 3 miliardy párů bazí (sekvenuje se i se 30× pokrytím) výsledný počet čtení je obrovský a sestavení genomu je náročnou výpočetní úlohou. Abychom sestavení genomu umožnili proveditelným, provádí se sekvenování s velkým pokrytím (ang. coverage), kdy číslo pokrytí označuje, kolikrát je daná báze zastoupena ve čteních. Jde o to, abychom získali dlouhé překryvy na různých místech sekvence. Kromě velkého množství čtení se sestavení genomu potýká s chybami v překryvech. S kratšími čteními je větší pravděpodobnost špatného zarovnání překryvů. Problémem jsou též repetice, opakující se částí DNA sekvence v genomu. Částečně se tento problém řeší vyšším pokrytím a metodou párových konců (paired-ends nebo mate-pair). [33, 36]

Je nutné vzít v potaz několik skutečností týkajících se procesu celého sekvenování. Za prvé, díky individuální genomické rozmanitosti není možné získat sekvenací jednu sekvenci, která by reprezentovala právě tento biologický druh. Dokonce buňky stejného organismu díky somatickým mutacím mohu mít rozdílný genomický obsah. Obecně sekvenujeme jen konkrétního jedince, lze však získat konsenzus vícero jedinců. Za druhé není možné osekvenovat a sestavit sekvence všech nukleotidů v genomu. Velké části DNA v heterochromatinové oblasti okolo centromer a telomer a další repetitivní oblasti nejsou dobře odlišitelné. Za třetí zde bude vždy jistá míra chybovosti v sekvenci, jak na úrovni sekvenačních chyb, tak chyb při sestavování genomu. Každá assembly by tak měla být vnímaná jako pracovní hypotéza a výsledek heuristického zpracování. [18]

## 3.2 Assemblery

### 3.2.1 Znakové metody

Prvotní úkol sestavení genomu ze čtení byl definován jako nalezení nejkratšího společného „superřettězce“ ze souboru všech čtení. Protože byl tento problém NP-úplný, využilo se při jeho řešení princip hladového algoritmu (greedy algoritmy), který opakuje sérii kroků a snaží se nalézt definované globální optimum. Ze souboru čtení je vypočítána párová vzdálenost mezi všemi dvojicemi čtení, v dalším kroku se vybere dvojice s nejvyšším překrytím a tyto dvě se spojí do celku. Kroky se postupně opakují dokud nezbudou žádná čtení. Hlavním problémem je uvíznutí v lokálním maximu, což se stane pokud se připojí čtení, které by však v dalších krocích vytvořilo ještě lepší variantu. Mezi algoritmy, jež využívají hladového principu patří SSAKE, SHARCGS a PE-Assembler, jež budou stručně popsány. [33, 28]

**SSAKE** (Short Sequence Assembly by progressive K-mer search and 3’ read Extension) je algoritmus pro sestavení genomu s využitím nepárových krátkých čtení stejné délky. Ukládá si soubor čtení do vyhledávací tabulky indexované dle prefixů čtení. Algoritmus cykluje sekvenčními daty v hashovací tabulce a hledá v prefixovém stromě nejdelší množný k-mer mezi dvěma sekvencemi. Hledá překryvy prefixu a sufixu, tak aby délka překryvu byla větší než daný práh. Tyto dvě sekvence spojí. Algoritmus se ukončí pokud narazí při slučování na větev stromu. Pokud zde již nejsou žádná čtení s překryvy vyhovující prahu, práh se sníží. Program se ukončí při vyčerpání všech čtení. [49, 33]

**SHARCGS** vychází z SSAKE, přidává však preprocessing a postprocessing, protože standardní SSAKE je citlivý na chyby čtení. Pracuje také s nepárovými čteními a stejné délky o vysokém pokrytí. V před-zpracovávání požaduje stanovenou minimální počet překryvů s ostatními čteními. Předpokládá se totiž, že v souboru o vysokém pokrytí jsou čtení s pouze pár překryvy pravděpodobněji zdrojem chyb. Dále se stanovuje minimální délka překryvu s možnými kandidáty na spojení, jež se stanovuje na alespoň polovinu délky čtení. Samotný algoritmus využívá strom prefixů k hledání překryvu s rostoucí výslednou sekvencí. Kontig je prodlužován dokud jsou zde čtení s prefixem minimální délky, které překryjí zcela konec kontigu. Když je kontig na 3’ konci dokončen, vypočítá se reverzní komplementární vlákno a prodlužuje vlákno stejným způsobem z druhé strany. SHARGCS filtruje soubor čtení třikrát o různé přísnosti kritérií, výsledkem jsou taktéž tři výsledné sekvence. Velmi přísné kritérium vede ke krátkým kontigům, protože mnoho čtení je vyřazeno, nejméně přísné kritérium generuje naopak nejdelší kontig. V posledním kroku jsou tyto tři sekvence kontigů spojeny do výsledného kontigu přes zarovnání. [17, 33]

**PE-Assembler** využívá paired-end čtení a je schopen pracovat s velkými ob-

jemy dat a generovat přesné a dlouhé kontigy. Celý algoritmus pracuje v pěti krocích: prověřování čtení, růst semínka, prodlužování kontigu, scaffolding a vyplňování mezer (read screening, seed-building, contig extension, scaffolding, gap-filling). První krok prověřování čtení identifikuje spolehlivá čtení za pomoci statistiky. Základní myšlenkou je, že pokud se daný k-mer vyskytuje pouze jednou, jde o chybu čtení, pokud se však daný k-mer vyskytuje často, jde o repetici. V dalším kroku se zvětšuje semínko. Semínko je navazující oblast, která je výsledkem prodlužování čtení a jež má stanovenou minimální délku. PE-Assembler prodlužuje čtení z obou konců sekvence, tak aby se překrývaly oba spárované konce čtení se sousedními čteními. Semínka jsou na konci kroku ověřena, zda se alespoň jedno čtení překrývá s 3' a 5' koncem. V dalším kroku prodlužování kontigu se ověřená semínka prodlužují, aby byl výsledný kontig co nejdelší. Při scaffoldingu je cílem najít správné řazení a orientaci kontigů. V dalším kroku se vyplní zbylé mezery, jež obvykle vyplňují repetitivní oblasti. [3, 33]

### 3.2.2 Grafové metody

Další popsané metody, OverLap Layout Consensus a metody De Bruijnových grafů, jsou založené na teorii grafů. Tu založil Leonhard Euler při řešení úlohy sedmi mostů v Královci. Grafy jsou tvořeny vrcholy vzájemně spojené hranami. Graf může být rozšířen o hodnocení vrcholů či hran. Dále se využívá tahů - sled vrcholů, ve kterém se neopakují hrany. Může být otevřený či uzavřený, pokud začíná a končí ve stejném bodě. Pokud vede přes všechny hrany, říká se mu eulerovský tah.[5, 12, 33]

#### Metoda překryvů - Overlap Layout Consensus

U algoritmů OLC je vrchol v grafu definován jako jedno čtení a dále orientované hrany definované jako překryvy s parametrem délky překryvu. Využívá se Hamiltonovské cesty - každým vrcholem grafu prochází právě jednou a končí v počátečním vrcholu. Najití Hamiltonovské cesty patří mezi NP-úplné problémy. Při velkém množství vrcholů je výpočetně náročný k vyřešení. V prvním kroku algoritmu se do grafu vynesou vrcholy (čtení) a mezi každou dvojicí vrcholů se hledá překryv. K tomu lze využít sufixové stromy. Dalším krokem je nalezení Hamiltonovské cesty grafem. Z výsledných kontigů se sestavena konečná konsensuální sekvence. [33, 30, 42]

**Celera** je algoritmus, který byl uveden pro sestavení celého osekvenovaného genomu druhu Drosophila. Algoritmus se zaměřuje na repetitivní úseky v genomu a jejich odstranění. K tomu využívá data z metody mate-pair a původně byl specifikován pro sekvenace Sangerovou metodou. Revidovaný algoritmus se nazývá CABOG (Celera Assembler with Best Overlap Graph). Algoritmus je rozdělen do několika po sobě jdoucích procesů. Prvním krokem je prověřování (screening) čtení, vstupní

čtení jsou porovnána s již existující databází známých repetitivních oblastí v genomu. Pokud se shodují je čtení vyřazeno. V dalším kroku se hledají překryvy za využití algoritmu na základě BLASTu. Následně se tvoří tzv. "unitigy", sestavené fragmenty jež se nepřekrývají s dalšími fragmenty. Unitigy obsahující unikátní DNA se nazývají U-nitigy a jsou prodlužovány. Posledním krokem je scaffolding, dvojice konců nebo konce BAC (bacterial artificial chromosome) se využijí ke spojení U-nitigů a vytvoření scaffoldů. Výstupem je finální konsenzus. [33, 12]

### Metoda De Bruijnových grafů

Novější metody jsou založené na De Bruijnových grafech. De Bruijnův graf je n-rozměrný orientovaný graf reprezentující překryvy mezi sekvencemi m symbolů. Je pojmenován po holandském matematikovi Nicolaasovi Govert de Bruijnovi. Namísto hledání hamiltonovské cesty je problém redukován na hledání eulerovské cesty v de Bruijnově grafu, což je jednodušší úloha. Grafické srovnání obou metod je znázorněno na obrázku 3.1. [33, 29]

De Bruijnův graf sestrojíme odlišně od OLC grafu. Čtení jsou zkrácena na k-mery o zvolené délce k. Vytvoříme vrcholy pro každý odlišný prefix nebo sufix k-merů tak, aby se každá sekvence o délce k-1 vyskytovala pouze jednou v celém grafu. Vrcholy jsou tedy tvořeny specifickými k-1 mery. Dále vytvoříme hrany, jež reprezentují překryvy mezi dvěma vrcholy a přiřadíme k hraně výsledný k-mer. V grafu hledáme eulerovskou cestu - každou hranu můžeme "navštívit"právě jednou, vrcholy vícekrát. V orientovaném grafu lze nalézt eulerovskou cestu, pokud má jeho každý vrchol vstupní stupeň roven výstupnímu. [33, 29, 38]

**AllPaths** je algoritmus vyvinutý pro krátká čtení. V předzpracování je každé čtení zkontrolováno a jen čtení vyskytující se často a o vysoké kvalitě postupují do dalšího zpracování. Jsou uloženy do datové struktury vhodné pro vyhledávání. Algoritmus generuje „unipath“ graf ze čtení a následně lokalizuje čtení před samotným sestavením genomu. Unipath je maximálně dlouhá nevětvící se sekvence, která je získána na základě zvoleného minimálního překryvu k v genomu. Lokalizace je způsob izolování malých regionů genomu, jež se sestavují samostatně. [7, 33]

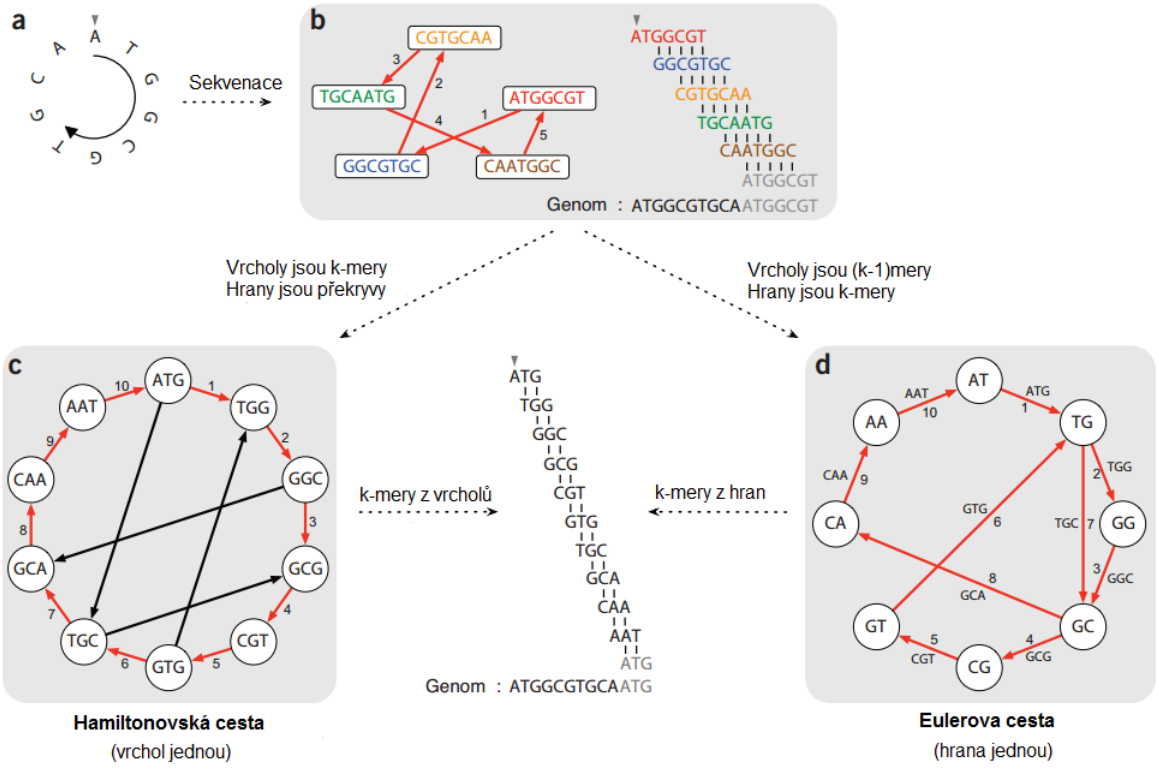
**Velvet** je algoritmus určený opět pro krátká čtení (25-50 bp) a skládá se z několika kroků. V prvním kroku je sestavena hashovací tabulka, což je datová struktura přiřazuje hodnotám (k-mer) jejich ID (číslo čtení obsahující daný k-mer). Pro zvolenou hodnotu k (pro čtení dlouhá 25 bp obvykle k=21) hashovací tabulka uloží hodnotu ID prvního čtení obsahující daný k-mer a také pozici k-meru v daném čtení. Další datová tabulka obsahuje informace o k-merech a jejich překryvech s dalšími čteními. Dále se sestaví de Bruijnův graf. Velvet pracuje namísto vrcholů s bloky, jež reprezentují překrývající se čtení (k-mery o k překrývajících se nukleotidech).

Bloky tedy obsahují podobná si čtení a hrany mezi bloky existují, pokud poslední čtení prvního bloku má překryv o délce k-1 s prvním čtením druhého bloku. Poté, co je graf sestrojen dojde k jeho zjednodušení. Pokud má vrchol (blok) pouze jednu výstupní hranu směřující do dalšího bloku, který má jen jednu vstupní hranu, tyto dva bloky jsou spojeny. Také jsou aplikovány některé korekce chyb. Následně se odstraňují bubliny, což jsou redundantní cesty, které začínají a končí ve stejném vrcholu a obsahují podobné sekvence. Tyto bubliny mohli vzniknout jako chyby čtení či biologické varianty před samotným sekvenováním. Až teď se přistupuje k hledání eulerovské cesty v grafu. Finální cesta, jež je nalezena je považována za výsledek sestavení genomu. [52, 33]

**Euler-Usr** je postaven na představě „opakovaného“ grafu (repeat graph). Jde o zjednodušenou variantu de Bruijnova grafu, kde jsou odstraněny chybné cesty. Klíčovým bodem algoritmu je, že opakovaný graf celého genomu lze získat aproximací podobného grafu získaného ze čtení. Cílem tedy je sestavit opakovaný graf z přesných čtení a pak aproximovat výsledný graf. Algoritmus se skládá ze tří kroků: detekce přesných čtení, konstrukce opakovaného grafu, zjednodušení opakovaného grafu. Za přesné čtení je považováno čtení, kde jsou všechny k-mery důležité, tedy k-mer s minimální hodnotou výskytu v setu k-merů ze čtení. Následně je aplikován greedy algoritmus, který hledá minimální počet mutací k tomu, aby všechny k-mery byly považovány za spolehlivé. Je vytvořen de Bruinův graf, který ale i přesto obsahuje některé chyby - chyby čtení či jedno-nukleotidové polymorfismy, které vytváří v grafu smyčky. Transformace de Bruijnova grafu do opakovaného grafu znamená odstranění všech těchto chyb. Následně se transformují již v opakovaném grafu párové konce na párová čtení. Pokud existuje více cest mezi začátkem a koncem čtení, vybere se cesta s největší podporou a vyhovující zvolenému prahu. Posledním krokem je samotné sestavení genomu. Opakovaný graf je tvořen prefixy, každé čtení může být korigováno zpracováním všech podcest podél čtení. Toto se nazývá „threading“ a probíhá v pěti krocích: detekce přesných čtení; konstrukce repeat grafu a tvorba k-merových kontigů; threading (navlékání) celých čtení přes repeat graf a tvorba nových čtení takto vzniklých; konstrukce nového repeat grafu z takto vzniklých čtení a generování l-merových kontigů (l>k); zjednodušení repeat grafu. V tomto finálním sestavení, opakování o délce l a kratší jsou vyřešena. [9, 33]

## 3.3 Využití assemblerů pro konkrétní sekvenátory

Sekvenátory poskytují čtení o rozdílných parametrech čtení - délka, množství, obsah repetic a chybovost. Pro různé sekvenátory se využívají metody sestavení genomu, které poskytují lepší výsledky než jiné.




Obr. 3.1: Srovnání grafových metod (Zdroj: [12])

V roce 2003 byl ukončen projekt první sekvenace lidského genomu (Human Genome Project), který byl proveden s využitím Sangerovy sekvenace a následného sestavení genomu algoritmem Celera, využívající OLC metodu sestavení genomu. Pro Sangerovu sekvenace jsou typická dlouhá čtení (i více než 1000 bazí), jež tedy obsahují repetice a další artefakty. Pro dlouhá čtení se hodí právě metody založené na OLC, poskytující menší množství čtení o větší délce. [21, 31, 33, 12]

Algoritmy využívající OLC jsou však neefektivní pro větší množství čtení, které poskytují sekvenátory nové generace. U nich se ujaly algoritmy založené na De Bruijnových grafech s využitím k-merů a vypuštění opakujících se čtení. Volba délky k-meru závisí na délce čtení, pokrytí a chybovosti, obecně se ale volí větší než polovina délky čtení. Pro všechny metody nové generace sekvenátorů se tak ujaly algoritmy využívající De Bruijnovy grafy. Výjimkou může být pyrosekvenace 454 s délkou čtení 700 až 1000 v závislosti na typu sekvenátoru. Zde se spíše využijí metody OLC (Celera například). Data ze sekvenátoru SOLiD o délce čtení 75 bazí se dají sestavit i s využitím greedy assemblerů (SSAKE,VCAKE) a samozřejmě metodou deBruijnových grafů (ALLPATHS). Pro sekvenátory Illumina a Ion Torrent se běžně používají algoritmy Velvet a Abyss, stavějící na De Bruijnových grafech s volbou k-meru 119 pro 400 bazí dlouhá čtení. [21, 31, 33, 12]

Výpočetní náročnost sestavení genomu je závislá na počtu a délce čtení sekvenátoru. Nižší výpočetní náročnosti se dá dosáhnout dlouhými čteními, jakých jsou schopny dosáhnout sekvenátory třetí generace SMRT a Nanopore. Zatímco Oxford Nanopore je stále ještě experimentální metodou, pro čtení z PacBio SMRT se hodí OLC algoritmy, například Celera. [21, 31, 33, 12]

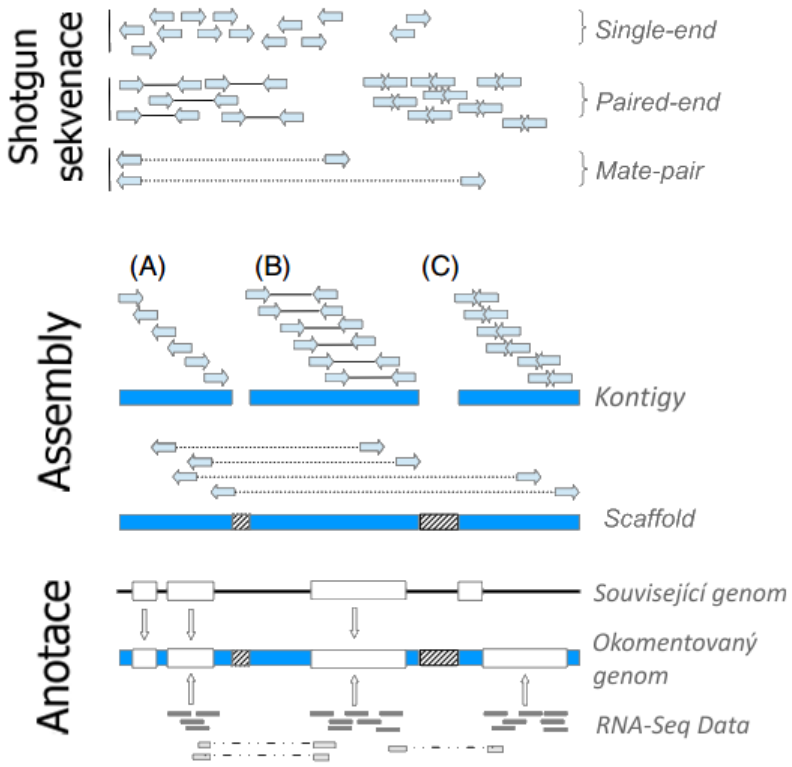
## 3.4 Neúplná genomová sestavení

V současné době většina sekvenačních projektů využívá shotgun sekvenaci. V prvním kroku je genomická DNA nastříhána na krátké náhodné fragmenty a v závislosti na sekvenátoru jsou tyto fragmenty sekvenovány nezávisle. Následuje výpočetní zpracování - sestavení genomu, které ma za cíl tyto fragmenty pomocí hledání překryvů složit zpět do delších souvislých úseků původní sekvence - kontigů. Takovému stavu sestavené genomu se říká neúplné genomové sestavení (ang. draft assembly). Obvykle jsou delší fragmenty sekvenovány z obou stran (paired-end sekvenace) za účelem poskytnutí dodatečné informace o správném umístění čtení v sekvenci. [18, 36]

Po počátečním sestavení genomu do kontigů, jsou kontigy spojeny do delších úseků zvaných scaffoldy (ang. scaffolds). K tomu se využívájí několik kilobází dlouhé fragmenty DNA, jejichž konce jsou osekvenovány. Pokud koncové sekvence několika těchto fragmentů leží na dvou odlišných kontizích, jsou tyto kontigy spojeny do jednoho scaffoldu. Očekávaná délka fragmentu poskytuje informaci o skutečné vzdálenosti mezi dvěma kontigy a vzniklá mezera je vyplněna znakem „N“. K vyplňování těchto mezer (ang. gap closing methods) pomáhají dlouhá čtení napříč těmito repetitivními úseky. V posledním kroku jsou výsledné scaffoldy spojeny do genetických map nebo umístěny na chromozom. Zjednodušený proces sekvenace a sestavení genomu je na obrázku 3.2. [18, 36]

Podobně jako existuje několik různých algoritmů pro hledání překryvů mezi čteními, existují také algoritmy, jež si kladou za cíl vylepšit či znova sestavit neúplné genomové sestavení a opravit chyby v genomovém sestavení. Pár z nich bude stručně popsáno.

Algoritmus **BIGMAC** se zaměřuje na sestavení genomu de-novo pro metagenomiku s využitím dlouhých čtení. Využívá post-processingový přístup. Vstupem algoritmu jsou jak sestavené kontigy, tak originální dlouhá čtení. Algoritmus nejdříve přeruší kontigy na potenciálně špatně sestavených pozicích a následně skládá kontigy do scaffoldů. Algoritmus je dvoukrokový - nejdříve rozbije kontigy s využitím signálové detekce. Detekuje tři různé signály spojené se špatně sestavenými genomy - palindromy, repetice a chyby pokrytí. Následně stanoví pozice, kde bude kontig rozdělen. V druhém kroku jsou tyto rozdělené kontigy spojovány a rozšiřovány. Al-




Obr. 3.2: Proces sestavení genomu (Zdroj: [18])

goritmus ověřitelně zvyšuje kvalitu kontigů snížením počtu chyb a současně zvyšuje hodnotu N50. [27]

**IMAGE** je algoritmus, jež se zaměřuje sekvenační platformu Illumina GA. V první fázi algoritmus zarovná čtení s počátečním sestavením, aby identifikoval čtení, která mohou být použita pro překlenutí mezer. Ve druhé fázi je provedeno jen lokální sestavení zvolených čtení a aktualizace počátečního sestavení vložením nově sestavených kontigů do mezer. Obě fáze mohou být prováděny iterativně. Tento přístup využívá toho, že Illumina GA produkuje čtení metodou paired-end z obou konců DNA fragmentu. Každé čtení tak má párové čtení a vzdálenost mezi těmito dvěma konci jednoho fragmentu může být odhadnuta na základě délky fragmentu v knihovně čtení. Pokud je čtení zarovnáno s kontigem, ale jeho druhý pár nikoliv, je pravděpodobné že tento pár je umístěn v mezeře, kde zatím chybí informace. Tyto čtení jsou spojeny a využity pro rozšíření kontigu nebo uzavření mezery. [47]

Algoritmus **Pilon** se zaměřuje na korekci chyb v sestavení - chybné báze, indely, mezery a nesouvislosti zarovnání čtení. V prvním kroku Pilon skenuje zarovnaná čtení a detekuje možné špatné báze a odlišnosti od vstupního genomu. V dalším kroku hledá nesrovnalosti v pokrytí a zarovnání, výsledkem je identifikace možných

špatných lokálních sestavení. V posledním kroku využije algoritmus čtení a párová čtení (mate-pairs) k znovusestavení dané oblasti a vyplnění mezer. Takto vylepšená sestavení poskytují kvalitnější výsledky pro genomovou anotaci. [48]

## 3.5 Anotace genomu

K využití osekvenovaného genomu pro další výzkum a studium DNA je nutné celý genom popsat. Mluvíme o genomové anotaci (ang. genome annotation). Jde o proces přiřazování biologických informací daným úsekům genomu. Je nutné identifikovat nekódující oblasti genomu, provést genovou predikci a správně tyto biologické informace přiřadit. Zahrnuje jak biologické informace o struktuře - čtecí rámce ORF, strukturu genu, kódující úseky a umístění regulačních motivů, tak informace o funkci - biochemické, biologické, regulaci a expresi. Anotace je obvykle vztažená k úsekům kódujícím proteiny - CDS (zkr. protein-coding sequence). [18]

Úspěch anotace závisí na kvalitě sekvenace a sestavení genomu. Pouze souvislé téměř kompletní genomy přerušené pouze malými mezerami poskytnou dostatečné výsledky [18]. Proces anotace lze rozdělit na dva kroky: výpočetní krok s predikcí genů na základě dalších již známých genomů a přiřazovací krok, kdy jsou získané informace shromážděny a na základě rozhodovacích pravidel přiřazeny k úsekům. V současné době již databáze jako NCBI umožňují uživatelům nahrát genomové sekvence spolu s jejich anotací, neboť z důvodu rostoucího množství osekvenovaných dat již proces anotace nezvládají databáze sami. [18]

## 3.6 Hodnocení kvality sestavení

Po dokončeném genomovém sestavení se provádí vyhodnocení kvality daného assembly. Za tímto účelem bylo vyvinuto několik měřítek kvality assembly. Lze je rozdělit na ty, které pracují s dalšími externími daty a na ty, které pracují čistě s daty z daného sestavení. Dalšími kategoriemi jsou měřítka, která pracují se skórem správnosti - jak správně je genomové sestavení přiřazeno k původnímu genomu, k tomu je zapotřebí referenční genom. A měřítka, která pracují se statistikou velikosti. [18, 33]

Do této kategorie patří nejpoužívanější měřítko kvality N50, jež slouží k porovnání různých metod sestavení genomu. Jde o numerickou hodnotu, která odpovídá počtu těch nejdelších kontigů, jejichž délka přesahuje 50 % celkové délky všech kontigů. Tedy pokud setřídíme kontigy dle velikosti a vybereme prvních N kontigů, jejichž délka překračuje polovinu celkové délky všech kontigů, tak číslo N odpovídá měřítku N50. Častěji užívanou variantou je měřítko „N50 contig size“, které odpovídá délce scaffoldu či kontigu, tak že 50 % sestavených sekvencí leží ve scaffoldu

této či větší délky. Hodnota prahu je volitelná, takže se lze setkat s měřítky N25 a N75. N50 ukazuje na schopnost assembleru vytvářet kontigy velké délky, ale nepostihuje další aspekty kvality assembly. Posuzování kvality assembly s pomocí N50 může být zavádějící pokud většina scaffoldů jsou kratší délky. V extrémním případě by assembly s nejvyšší hodnotou N50 a největší celkovou délkou mohla sestávat z pouze jednoho velmi dlouhého kontigu a tisíců velmi krátkých kontigů. Z hlediska genové predikce lze za použitelné sestavení považovat sestavení, které má nejvyšší počet scaffoldů, jež jsou delší než délka obvyklého genu. [6, 18, 33]

# 4 NUMERICKÉ REPREZENTACE

## 4.1 Využití numerických reprezentací

Standardní algoritmy užívané při analýzách DNA sekvencí pracují přímo s řetězcem znaků zastupujících nukleotidy. Znakovou sekvenci DNA lze však za využití různých číselných reprezentací převést na sekvenci čísel a pracovat dále s ní. Toto najde uplatnění ve škále činnosti od zobrazení průběhu sekvence a zvýraznění některých vlastností DNA sekvence, hledání periodických struktur a dalších oblastí v sekvenci, až po dosud neprozkoumané možnosti, jež se naskýtají. Převod lze znázornit schématem níže 4.1. Cílem je za užití definovaného algoritmu převodu (danou metodu nazýváme reprezentací) převést znakovou sekvenci na číselnou sekvenci a tu dále zpracovat.


Obr. 4.1: Schéma převodu na numerickou DNA sekvenci

Bylo by možné zde popsat množství numerických reprezentací a jejich rozdělení, což by však bylo na rámec a téma této diplomové práce. Proto zde bude popsáno jen několik numerických reprezentací, které jsou nejvýznamnější a nejčastěji užívané. Převod do numerické reprezentace spočívá v přiřazení různých numerických vyjádření jednotlivým nukleotidům v DNA sekvenci.

## 4.2 Vossova reprezentace

Vossova reprezentace, též reprezentace binárními vektory četností je 4D reprezentace. Pracuje s čtyřmi indikačními vektory pro každý nukleotid zvlášť $u_A[n], u_T[n], u_G[n], u_C[n]$ dle vzorce 4.1, které indikují binární hodnotou 1 přítomnost nebo 0 nepřítomnost daného nukleotidu na dané pozici $n$ ve sekvenci DNA. [1, 2]

$$u_x(n) = \begin{cases} 1, s_n = X \\ 0, s_n \neq X \end{cases} \quad , X \in \{A, C, G, T\} \tag{4.1}$$

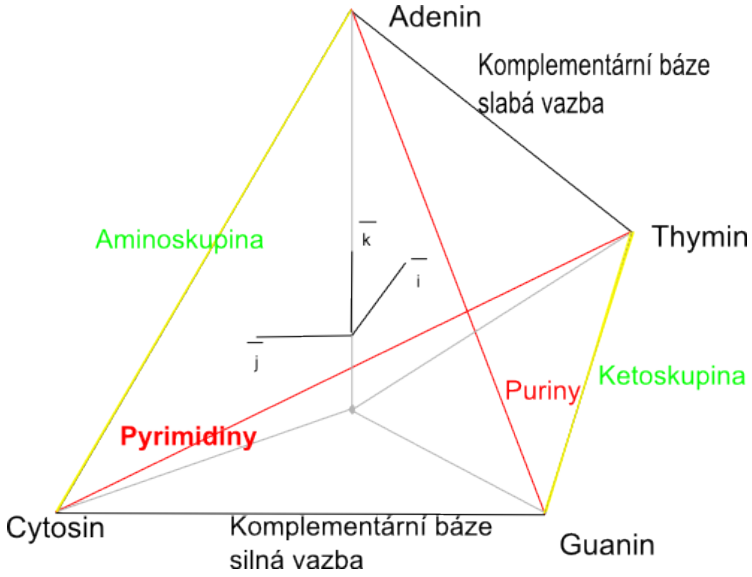
Reprezentace je vhodná pro spektrální analýzu DNA sekvencí a odlišení kódujících oblastí DNA, které jsou bohatší na nukleotidy guaninu a cytosinu oproti ostatním oblastem genomu. Výkonnové spektrum pak vykazuje "peak"na 1/3 pro kódující sekvence. [1, 2]

## 4.3 Tetrahedronová reprezentace

Redukci z 4D Vossovy reprezentace do 3D prostoru umožňuje reprezentace tetrahedronem (pravidelný čtyřstěn). Každý vrchol reprezentuje právě jeden nukleotid a je zadán samostatným vektorem dle vztahu 4.2. Binární sekvence Vossovy reprezentace $u_A[n], u_T[n], u_G[n], u_C[n]$ jsou redukovány do 3D odvozenými vektory os kartézského systému 4.2:

$$\begin{aligned} x_i[n] &= \frac{\sqrt{2}}{3}(2u_T[n] - u_C[n] - u_G[n]) \\ x_j[n] &= \frac{\sqrt{6}}{3}(u_C[n] - u_G[n]) \\ x_k[n] &= \frac{1}{3}(3u_A[n] - u_T[n] - u_C[n] - u_G[n]) \end{aligned} \tag{4.2}$$

Přiřazení nukleotidů k tetrahedronu je znázorněno na obrázku 4.2. Jsou též zaznačeny chemické vlastnosti, které jsou reprezentací zachovány. Přiřazením barev RGB systému každé ose trojrozměrné tetrahedronové reprezentace lze získat barevnou reprezentaci. [2, 1]




Obr. 4.2: Reprezentace tetrahedronem

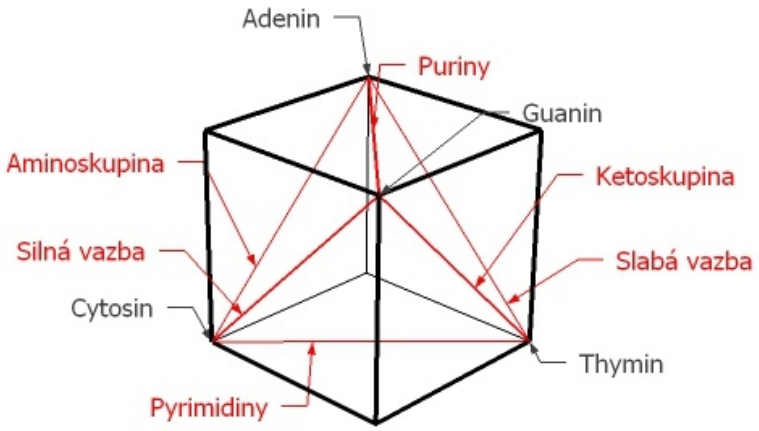
## 4.4 Reprezentace krychlí

Rotací tetrahedronu z předchozí reprezentace a jeho vepsáním do krychle získáme reprezentaci krychlí, kdy vrcholy tetrahedronu odpovídají vrcholům krychle. Zjednodušíme také zápis vektorů uvažováním délky hrany krychle rovné 1. Výsledné

vektory pro nukleotidy jsou popsány vztahy [2, 1]:

$$\begin{aligned} \vec{a} &= \vec{i} + \vec{j} + \vec{k} \\ \vec{c} &= -\vec{i} + \vec{j} - \vec{k} \\ \vec{g} &= -\vec{i} - \vec{j} + \vec{k} \\ \vec{t} &= \vec{i} - \vec{j} - \vec{k} \end{aligned} \tag{4.3}$$

Tato reprezentace zachovává chemické vlastnosti jednotlivých nukleotidů. Grafické znázornění reprezentace krychlí a přiřazení nukleotidů vrcholům spolu s chemickými vlastnosti je na obrázku 4.3.




Obr. 4.3: Reprezentace krychlí

## 4.5 Reprezentace komplexním číslem

Převedením do dvojrozměrného prostoru sklopením zvolené roviny krychle z krychlové reprezentace získáme reprezentaci, která bude nést požadovanou informaci. V komplexním prostoru přiřadíme vrcholům čtverce souřadnice podle vlastností, jež chceme zachovat. Pokud chceme zachovat komplementaritu bazí, přiřadíme jim souřadnice o opačném znaménku a umístíme je naproti sobě. Výsledné rovnice vypadají takto 4.4:

$$\begin{aligned} \vec{a} &= 1 + \jmath \\ \vec{c} &= -1 - \jmath \\ \vec{g} &= -1 + \jmath \\ \vec{t} &= 1 - \jmath \end{aligned} \tag{4.4}$$

Toto přiřazení respektuje komplementaritu bazí, komplementární dvojice jsou namapovány naproti sobě s opačnými znaménky. Dvojice A–T má kladnou reálnou část a naopak dvojice G–C má reálnou část zápornou.

## 4.6 Fázové reprezentace

Na reprezentaci bazí komplexními čísly stojí fázová reprezentace. Argument nebo fáze komplexního čísla je úhel s kladnou reálnou osou. Při násobení celými násobky $2\pi$ se fáze nezmění. Pro odstranění dvojznačnosti se obor hodnot stanovil na $(-\pi, pi]$ rad. Využíváme dvě fázové reprezentace - kumulovanou fázi a rozbalenou fázi. [13, 14]

Při použití mapování 4.4, které zachovává informace o síle vazby a typu báze, mají jednotlivé báze tyto argumenty:

$$\begin{aligned} \vec{a} = 1 + \jmath \rightarrow \quad & \varphi_a = \frac{\pi}{4} \\ \vec{c} = -1 - \jmath \rightarrow \quad & \varphi_c = -\frac{3\pi}{4} \\ \vec{g} = -1 + \jmath \rightarrow \quad & \varphi_g = \frac{3\pi}{4} \\ \vec{t} = 1 - \jmath \rightarrow \quad & \varphi_t = -\frac{\pi}{4} \end{aligned} \tag{4.5}$$

### 4.6.1 Kumulovaná fáze

Kumulovaná fáze je kumulativním součtem všech fází komplexních čísel od počátku sekvence až po aktuální pozici v sekvenci. Vektor kumulativního součtu je pak zobrazen v grafu a analyzován. Dá se získat též z četností výskytu jednotlivých nukleotidů podle vzorce:

$$s_c = \frac{\pi}{4}[3(G_n - C_n) + (A_n - Tn)], \tag{4.6}$$

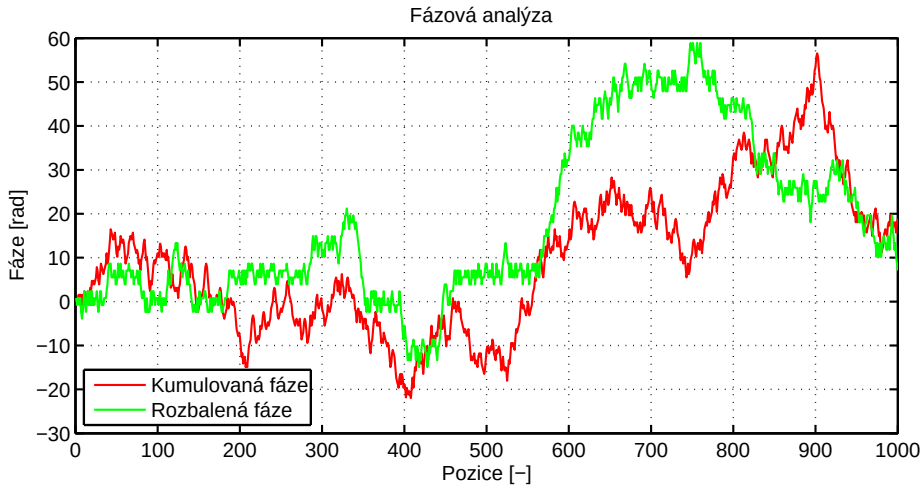
kde $G_n, C_n, A_n, Tn$ jsou kumulativní četnosti bazí až po aktuální pozici. Sklon křivky kumulované fáze ukazuje na poměr výskytu jednotlivých bazí, jež se odvíjí od DNA konkrétního původu. [15, 45]

### 4.6.2 Rozbalená fáze

Rozbalená fáze je korigovanou fází prvků v sekvenci. Absolutní hodnota rozdílu fází dvou po sobě jdoucích je záměrně snižována na hodnotu menší než $\pi$ přičítáním či odečítáním vhodného násobku čísla $2\pi$. Tím dojde k eliminaci velkých rozdílů ve fázové sekvenci. Na rozdíl od kumulované fáze rozbalená fáze ukazuje na výskyty tranzicí mezi nukleotidy. Při reprezentaci 4.5 jsou kladné tranzice $A \to G, G \to C, C \to T, T \to A$, jež určují kladný přírůstek křivky. Negativní tranzice $A \to T, T \to C, C \to G, G \to A$ určují záporný přírůstek křivky. Ostatní tranzice jsou neutrální, nedochází při nich ke změně fáze. Sklon křivky rozbalené fáze je

$$\frac{2}{\pi} s_u = f_+ - f_-, \tag{4.7}$$

kde $f_+$ je frekvence kladných tranzic nukleotidů a $f_-$ frekvence záporných tranzic.[13]




Obr. 4.4: Fázova analýza genomu Pragia fontium

### 4.6.3 Volba numerické reprezentace

Z množství reprezentací popsaných v odborné literatuře jich bylo několik popsáno. Pro další práci a návrh algoritmu je uvažoval jako ideální reprezentaci takovou, která má stejné vzorkování jako původní znaková sekvence, je jednorozměrná, vhodná pro korelační funkce a výpočetně nenáročná. Fázové reprezentace již byly ověřeny za dostatečné pro reprezentaci DNA sekvence s využitím pro klasifikace organismů. [45]. Z těchto důvodů jsem zvolil právě fázové reprezentace pro další návrh algoritmu.

# 5 VYUŽITÍ NUMERICKÉ REPREZENTACE PRO NEÚPLNÁ GENOMOVÁ SESTAVENÍ

## 5.1 Měřítka pro porovnání podobnosti mezi signály

V následující kapitole bude probrána problematika hledání podobností mezi číselnými signály. Pro hledání překryvů mezi znakovými sekvencemi se používají různé metody jako výše probrané De Bruijnovy grafové metody a OLC. Rovněž pro objektivní posouzení podobnosti mezi dvojicí numerických sekvencí - signálů, existují různé metody. Nejčastější jsou korelační a kovarianční funkce. Mezi použitelné metody patří Pearsonův korelační koeficient a dále vhodnější vzájemná korelace a vzájemná kovariance.

### 5.1.1 Korelační koeficient

Porovnat vztah mezi dvěma veličina lze nejčastěji korelačním koeficientem, jež sleduje vzájemný lineární vztah mezi dvěma veličinami. Definujme Pearsonův korelační koeficient [24] :

$$r_{xy} = \frac{\sum xy - \frac{1}{n}(\sum x)(\sum y)}{\sqrt{\left[\sum x^2 - \frac{1}{n}(\sum x)^2\right]\left[\sum y^2 - \frac{1}{n}(\sum y)^2\right]}} \tag{5.1}$$

Hodnota korelačního koeficientu nabývá hodnot $< -1, 1 >$. Hodnota koeficientu $-1$ značí zcela nepřímou závislost (antikorelaci). Hodnota 1 značí zcela přímou závislost. Při korelační koeficientu 0 není mezi znaky žádná lineární závislost (nelineární závislost být může). Hodnoty korelačního koeficientu nad $0,8$ se považují za dostačující k potvrzení korelace. [24]

Korelační koeficient je nevýhodný v tom, že nepočítá s různě posunutými (zpožděnými) signály, mezi kterými ale může existovat jistá korelace. Také při rozdílné délce hledaného překryvu mezi kontigy neposkytuje korelační koeficient korektní výsledek. Pro požadovaný účel hledání překryvů není pro využití při hledání překryvů dostatečný.

### 5.1.2 Vzájemná korelace

Na hledání překryvů mezi dvěma sekvencemi se také můžeme dívat jako na problematiku zarovnání vyslaného a přijatého signálu, protože překryv mezi dvěma sekvencemi může nabývat různých délek. Pro posouzení korelace mezi dvěma signály při

různém zpoždění slouží posloupnost vzájemné korelace, která vyjadřuje podobnost dvou signálů v různých pozicích vůči sobě [25].

Algoritmus korelační posloupnosti je založen na vzájemném násobení vzorků signálu, z nichž se následně provádí suma. Předpokládejme dvojici reálných signálů $x(n)$ a $y(n)$ s konečnou energií. Vzájemná korelační posloupnost signálů $x(n)$ a $y(n)$ je sekvence $r_{xy}(l)$, kterou lze definovat jako:

$$r_{xy}(l) = \sum_{n=-\infty}^{\infty} x(n)y(n-l) \quad l = 0, \pm 1, \pm 2, ... \tag{5.2}$$

Index $l$ je časový posun či zpoždění a indexy $xy$ u korelační posloupnosti $r_{xy}(l)$ značí, že sekvence jsou korelované [39]. Pořadí indexů, tedy že $x$ předchází $y$, ukazuje na směr, ve kterém je jeden ze signálů posunutý vzhledem k druhému. Pro rovnici 5.2 tak, že signál $x(n)$ je nalevo bez posunu a signál $y(n)$ je posunutý o $l$ napravo pro kladné $l$ a nalevo pro záporné $l$. Role signálů je možné prohodit podle vzorce 5.3 :

$$r_{yx}(l) = \sum_{n=-\infty}^{\infty} y(n)x(n-l) \quad l = 0, \pm 1, \pm 2, ... \tag{5.3}$$

Srovnáním těchto vzorců získáme jednu z vlastností korelační posloupnosti:

$$r_{xy}(l) = r_{yx}(-l) \tag{5.4}$$

Existuje rovněž vztah mezi konvolucí a korelační posloupností,

$$r_{xy}(l) = x(l) * y(-l) \tag{5.5}$$

kdy můžeme vypočítat korelační posloupnost pomocí konvoluce s obráceným signálem $y$. [39]

Další z vlastností korelační posloupnosti je, že škálování signálů ve výpočtu neovlivní tvar křivky korelační posloupnosti, ale pouze amplitudy, které jsou rovněž škálovány. Často je vhodné tedy normalizovat korelační posloupnost na rozsah $< -1, +1 >$, odtud definujeme normalizovanou korelační posloupnost jako [39]:

$$\rho_{xy}(l) = \frac{r_{xy}(l)}{\sqrt{r_{xx}(0)r_{xy}(0)}} \tag{5.6}$$

Proces výpočtu korelační posloupnosti mezi $x(n)$ a $y(n)$ zahrnuje posouvání jednoho ze signálů k získání sekvence $x(n-l)$ a násobení posouvané sekvence signálem $y(n)$ k získání součinu $y(n)x(n-l)$ a následné sumě všech hodnot k získání hodnoty $r_{yx}(l)$. Tento proces se opakuje pro různé hodnoty zpoždění $l$. V praxi se výpočet korelační posloupnosti dvou konečných signálů $x(n)$, $0 \leq n \leq N-1$ a $y(n)$, $0 \leq n \leq M-1$ pro $M \leq N$, provádí dle vztahu [39]:

$$r_{xy}(l) = \begin{cases} \sum_{n=l}^{M-1+l} x(n)y(n-l) & 0 \leq l \leq N-M \\ \sum_{n=l}^{N-1} x(n)y(n-l) & N-M < 1 \leq N-1 \end{cases} \tag{5.7}$$

Pro $M > N$ pak platí vztah pro korelační posloupnost [39]:

$$r_{xy}(l) = \sum_{n=l}^{N-1} x(n)y(n-l) \quad 0 \leq l \leq N - M \tag{5.8}$$

Korelační posloupnost nám umožňuje porovnávat signály koncových částí kontigů a vypočítávat i jejich vzájemný posun. Mezi další vlastnosti patří teoretická možnost detekce překryvu s reverzně komplementárním kontigem, pro takový signál poskytuje v případě velké míry korelace negativní hodnoty korelační posloupnosti. Na obrázku 5.1 nahoře je graf průběhu dvou identických signálů kumulované fáze a průběh jejich vzájemné korelační posloupnosti. Na obrázku 5.1 níže je graf průběhu signálu a jeho reverzně komplementárního protějšku a jejich vzájemná korelační posloupnost. Vidíme, že pro ideální případ identických signálů kumulované fáze dostáváme vyhovující výstup.

### 5.1.3 Vzájemná kovariance

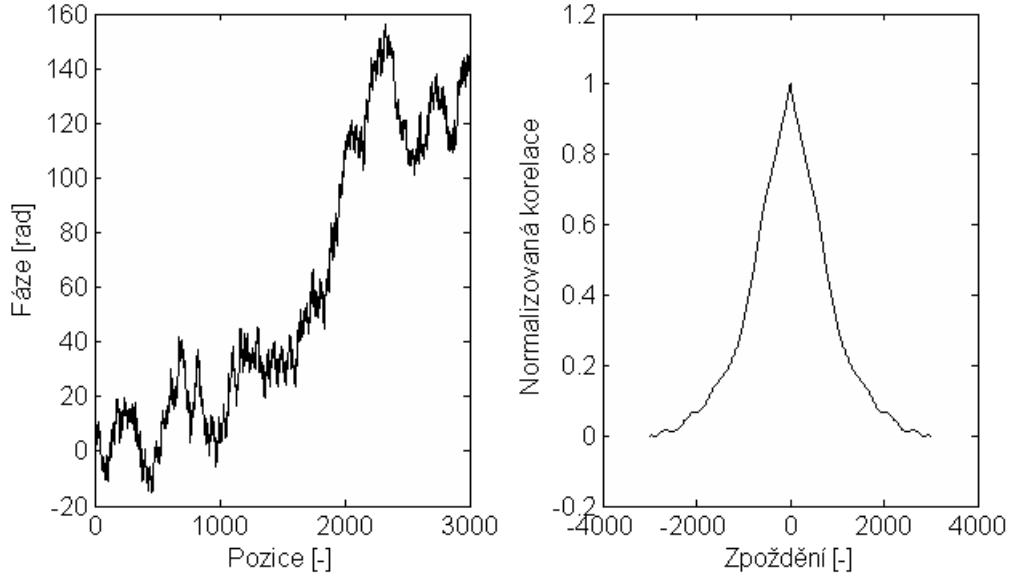
Zatímco korelace dvojice proměnných odpovídá střední hodnotě jejich součinu, tak kovariance odpovídá střední hodnotě součinu po odečtení příslušných středních hodnot. Stejně je tomu u vzájemné kovariance. Vzájemná kovariance je definována vzorcem [25]:

$$c_{yx}(l) = \sum_{n=0}^{N-1} [x(n) - \mu_x)][y(n+l) - \mu_y] \tag{5.9}$$

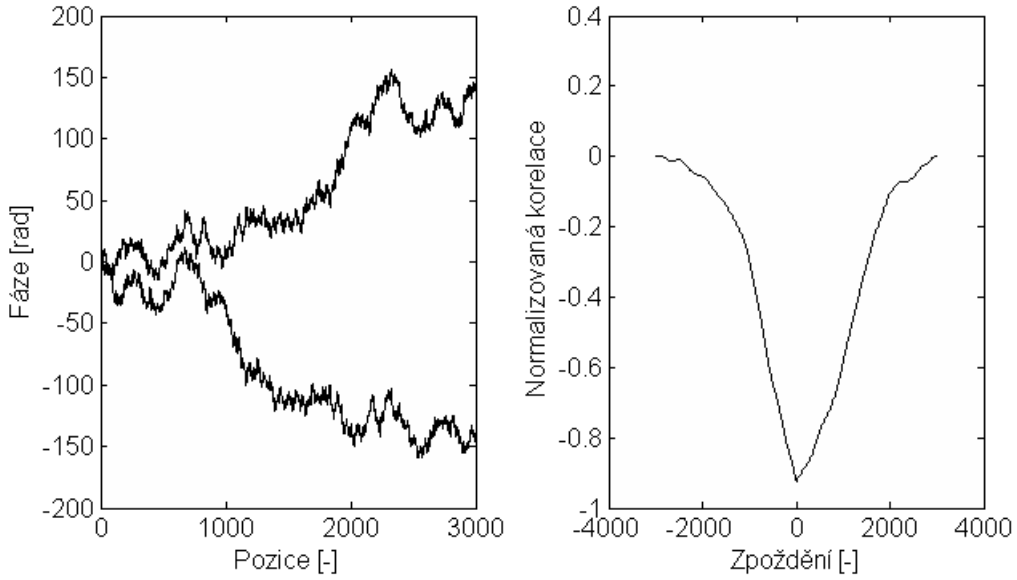
kde $\mu_x$ a $\mu_y$ jsou střední hodnoty. Jsou-li obě veličiny těsně vázány a vazba je pozitivní (odchylky obou veličin od středních hodnot mají stejné znaménko), pak většina součinů bude kladná a odhad kovariance bude také kladný. Při negativní vazbě bude mít většina součinů záporné znaménko a odhad bude záporný. Nulová hodnota kovariance se objeví pravděpodobně při nezávislosti obou veličin. Veličiny o nulové kovarianci nazýváme nekorelované. Absolutní hodnotu je nutno normalizovat na korelační koeficient[23] podle vzorce:

$$\rho_{xy} = \frac{c_{xy}}{\sqrt{\sigma^2(x)\sigma^2(y)}} \tag{5.10}$$

Jak vzájemná kovariance, tak vzájemná korelace jsou vhodné pro návrh algoritmu, jelikož umožňují hledat překryvy i při různých posunech kontigů. Z dané posloupnosti korelační či kovarianční funkce lze dále získat hodnotu maximální nebo minimální hodnoty a příslušný posun jednoho signálu vůči druhému, což je dále uplatnitelné při finálním skládání sekvencí ve znakové podobě. Grafy na obrázku 5.2 ukazují opět průběh dvou signálů kumulované fáze a posloupnost vzájemné kovariance, nahoře pro identické signály, dole pak pro signály reverzně komplementární.




(a)

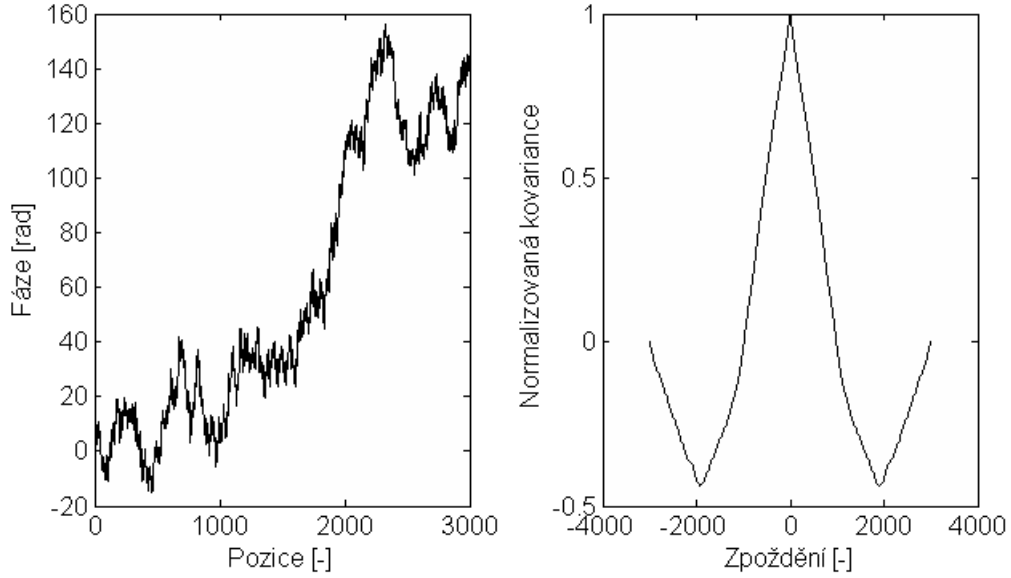



(b)

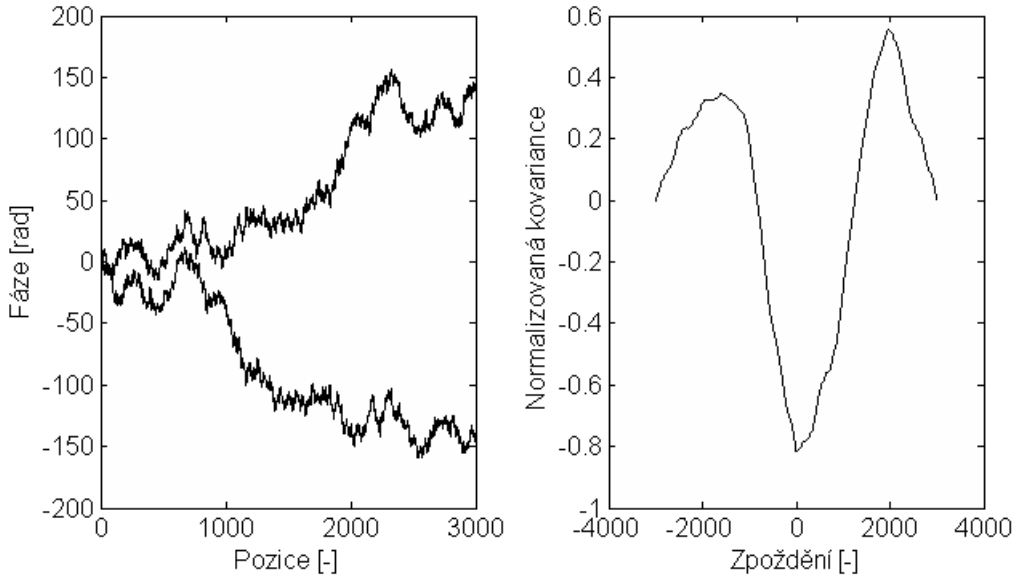
Obr. 5.1: Signály a jejich vzájemná korelace - nahoře identické, dole signály reverzně komplementárních sekvencí

## 5.2 Teoretické předpoklady pro hledání překryvů

V této části budou otestovány a prezentovány výsledky teoretických předpokladů. Jde o výběr správného funkce podobnosti, prahování korelačního koeficientu a schopnost detekovat správně reverzně komplementární kontig. Všechny výsledky pracují se signály kumulované fáze. Byla připravena data ze skutečné sekvence - dva kontigy




(a)




(b)

Obr. 5.2: Signály a jejich vzájemná kovariance - nahoře identické, dole signály reverzně komplementárních sekvencí

se skutečným překryvem o délce 1000 bazí a dva kontigy bez žádného překryvu. Pro různé délky oken, se kterými počítá algoritmus počítající korelační a kovarianční posloupnosti byly zjištěny maximální a minimální hodnoty dané posloupnosti. Cílem je vybrat vhodnější měřítko pro hledání překryvů, které bude schopno najít překryv ve stejném směru i překryv, jehož kontig je reverzně komplementární. Taktéž je nutné stanovit vhodný práh pro korelační koeficient, aby na sebe byly navázány kontigy

se skutečným překryvem a naopak, aby kontigy bez překryvu nebyly spojeny.

### 5.2.1 Volba funkce podobnosti

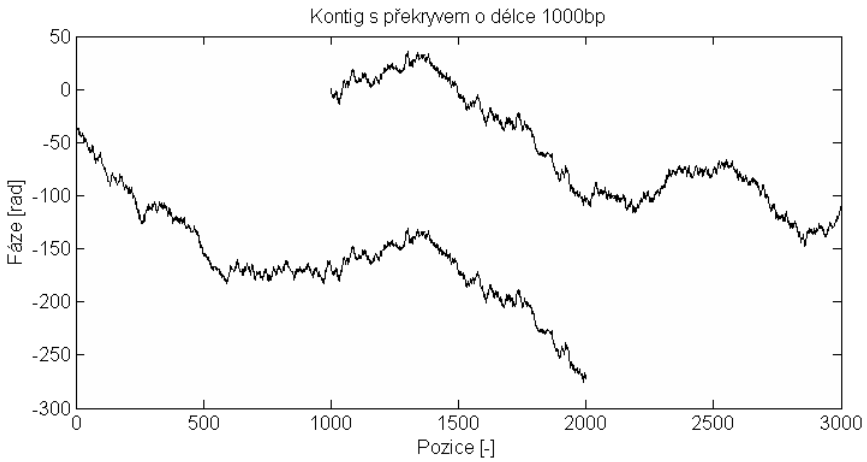
Nejdříve byla otestována vhodnost vzájemné korelace či kovariance pro hledání překryvů a reverzně komplementárních překryvů na skutečné sekvenci. Ideálním předpokládaným výstupem jsou vždy kladné hodnoty korelačního koeficientu pro překryv ve stejném směru a vždy záporné hodnoty pro překryv reverzně komplementárního kontigu. Koeficienty byly získány v závislosti na různé délce okna, ve kterém se počítá vzájemná korelace či kovariance a to v délkách 100, 1000, 2000 a 5000. Délka okna 100 byla pouze testovací, protože pro překryv kontigů o délce 1000 bází nemá okno o kratší délce smysl (velikost okna se odečítá od konce kontigu, tedy nepočítalo by se s odpovídajícími úseky signálu). Délka okna by měla být rovna či delší než skutečný překryv, který ale v reálném případě předem neznáme, ale právě hledáme.

Tab. 5.1: Kontig s překryvem 1000 bp

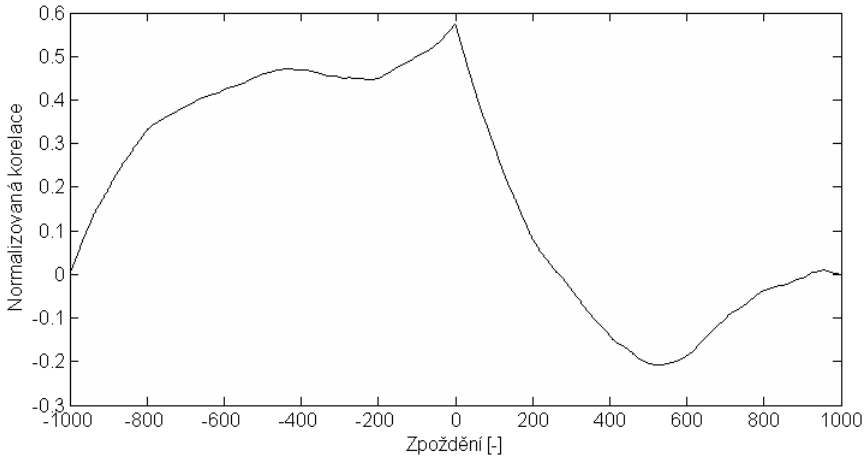
| **Měřítko** | **Hodnota** | **Překryv 100** | **1000** | **2000** | **5000** |
|---|---|---|---|---|---|
| Korelace | max | 0,3207 | 0,5743 | 0,8566 | 0,6099 |
| Korelace | min | -0,5393 | -0,2072 | -0,0616 | -0,3944 |
| Kovariance | max | 0,4689 | 1,0000 | 0,7471 | 0,8562 |
| Kovariance | min | -0,8508 | -0,4223 | -0,4244 | -0,3885 |

V tabulce 5.1 jsou výsledky prvního testování pro kontigy, které se překrývají podél úseku o délce 1000 bází. Jsou uvedeny maxima i minima korelačních a kovariančních posloupností. Pro korelaci byl nejlepší výsledek s oknem o délce 2000 vzorků, ostatní korelační koeficienty by nemohly být považovány za dostatečné. V tomto případě se vzájemná kovarianční posloupnost ukázala býti vhodnější, pro okno o stejné délce jako je délka překryvu 1000 byl dokonce korelační koeficient roven 1.

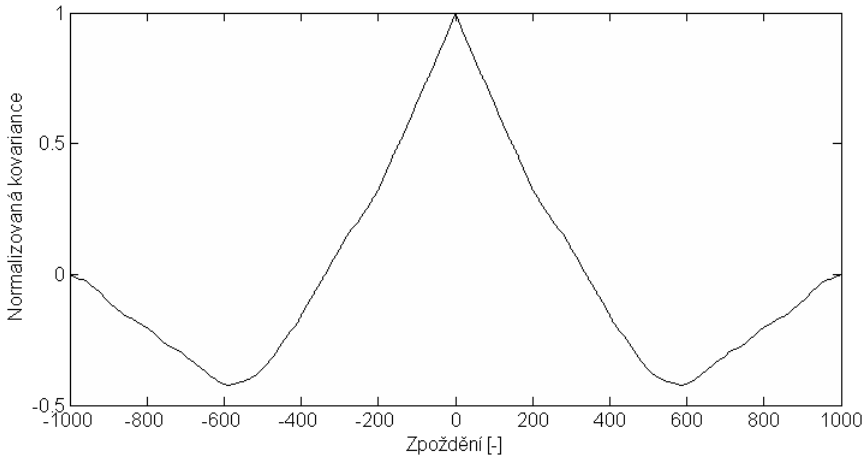
V tabulce 5.2 jsou výsledky pro reverzně komplementární kontig. Očekávaným výstupem by měly být hodnoty blížící se -1. Ideální hodnotě se však nejvíce blíží jen korelační koeficienty vzájemné kovariance, naopak vzájemná korelace ukazuje na opačné hodnoty blížící se +1. Z výsledku vzájemné korelace nebude schopné rozlišit, zda se jedná o kontig ve stejném či opačném směru. Na obrázku 5.3 je vykreslen průběh obou překrývajících se signálů a dále pak průběh korelační a kovarianční posloupnosti pro okno o délce 1000 vzorků. Na dalším obrázku 5.4 je pak průběh signálů z reverzně komplementárních sekvencích a níže opět průběh posloupností pro okno délky 1000 vzorků.




(a)

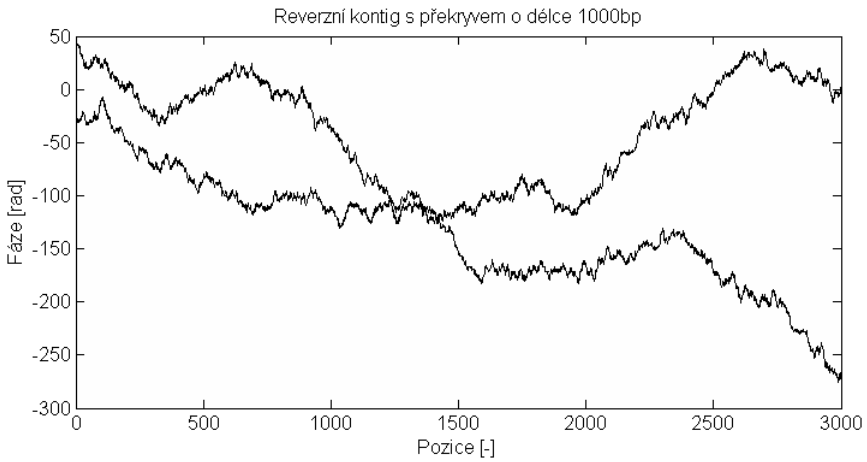



(b)

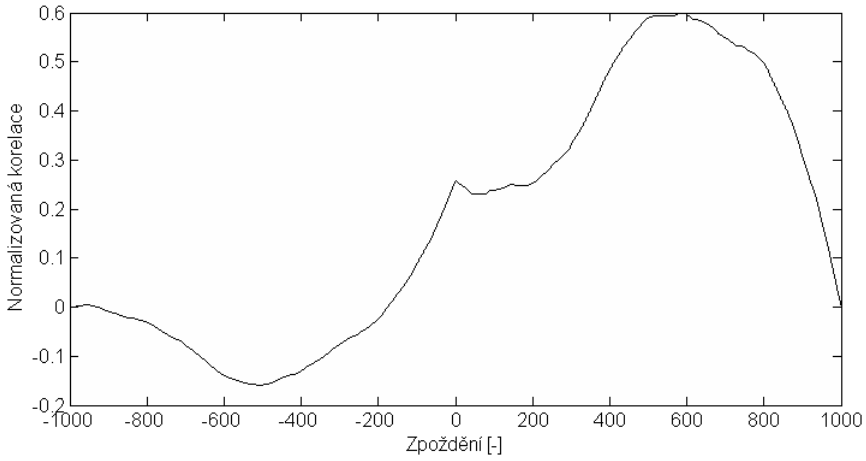



(c)

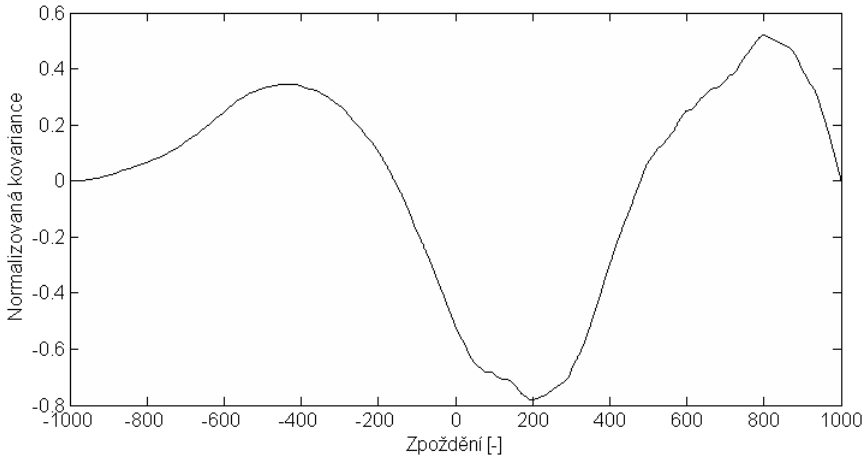
Obr. 5.3: Překryvy kontigů (a) a jejich vzájemná korelace (b) a kovariance (c) (okno 1000)




(a)




(b)




(c)

Obr. 5.4: Reverzně komplementární překryv (pozice 2000+) (a) a jejich vzájemná korelace (b) a kovariance (c) (okno 1000)

Tab. 5.2: Kontig s reverzně komplementárním překryvem 1000b

| **Měřítko** | **Hodnota** | **Překryv 100** | **1000** | **2000** | **5000** |
|---|---|---|---|---|---|
| Korelace | max | 0,2101 | 0,5999 | 0,8185 | 0,6208 |
| Korelace | min | -0,6231 | -0,1593 | -0,0308 | -0,2535 |
| Kovariance | max | 0,8117 | 0,5198 | 0,4731 | 0,4261 |
| Kovariance | min | -0,4675 | -0,7804 | -0,6360 | -0,6116 |

### 5.2.2 Volba prahu korelačního koeficientu

Další požadovanou vlastností je schopnost spolehlivě rozlišit, zda se kontigy skutečně překrývají či nikoliv a to volbou prahu pro korelační koeficient. Na základě hodnoty korelačního koeficientu jsou totiž kontigy spojovány. Při nepřekročení stanovené hodnoty prahu by už neměli být kontigy spojovány. Byly připraveny data - kontigy, které se překrývají na úseku o délce 1000 bazí a kontigy, mezi nimiž není žádný překryv.

Tab. 5.3: Korelační koeficienty vzájemné kovariance pro kontigy s překryvem (1000b) a bez překryvu

| **Varianta** | **Hodnota** | **Překryv 100** | **1000** | **2000** | **5000** |
|---|---|---|---|---|---|
| bez překryvu | max | 0,7004 | 0,8643 | 0,4569 | 0,5653 |
| bez překryvu | min | -0,4553 | -0,3793 | -0,4278 | -0,3986 |
| s překryvem | max | 0,4689 | 1,000 | 0,7471 | 0,8562 |
| s překryvem | min | -0,8508 | -0,4223 | -0,4244 | -0,3885 |

Z výsledků v tabulce 5.3 pro vzájemnou kovarianci vyplývá, že ne vždy se lze spolehnout na danou metodiku hledání překryvu. Pro délku okna 1000 vzorků by byl falešně pozitivní výsledek. Naopak u překrývajících se kontigů byla pro délku okna 1000 zjištěna správně pozitivní korelace.

Výsledky pro vzájemnou korelaci v tabulce 5.4 se potvrdila nevhodnost vzájemné korelace pro hledání překryvů. Došlo by k falešně pozitivním detekcím překryvů pro různé délky okna a navíc by byly považovány za reverzně komplementární.

Z výsledků předcházejících testů lze učinit několik závěrů. Jednak lze za vhodnou funkci pro měření podobnosti mezi signály pro účel hledání překryvů určit vzájemnou kovarianci, která je i vhodná pro detekci reverzně komplementárních kontigů. Vyloučit lze vzájemnou korelaci. Také by bylo vhodné stanovit dva rozdílné prahy

Tab. 5.4: Korelační koeficienty vzájemné korelace pro kontigy s překryvem (1000b) a bez překryvu

| **Varianta** | **Hodnota** | **Překryv 100** | **1000** | **2000** | **5000** |
|---|---|---|---|---|---|
| bez překryvu | max | 0,0877 | 0,0012 | 0,0004 | 0,0002 |
| bez překryvu | min | -0,7282 | -0,8355 | -0,9226 | -0,7196 |
| s překryvem | max | 0,3207 | 0,5743 | 0,8566 | 0,6099 |
| s překryvem | min | -0,5393 | -0,2072 | -0,0616 | -0,3944 |

pro maximální a minimální hodnoty kovariančních posloupností a to na 0,5 pro minima a 0,75 pro maxima, jako prahové hodnoty pro navázání dalšího kontigu.

# 6 NÁVRH PROVEDENÍ ALGORITMU

Cílem práce je prozkoumat možnosti využití numerických reprezentací pro vylepšení neúplných genomových sestavení. Tedy sestavení finální sekvence z kontigů, získaných některým z assemblerů. Jednotlivé sekvence kontigů, které jsou vstupem algoritmu jsou převedeny do signálové podoby s využitím fázové reprezentace kumulovanou fází nebo rozbalenou fází. Takto převedené sekvence mají výhodu v jednorozměrnosti a stejném vzorkování jako výchozí znakové sekvence, signál však není zpětně rekonstruovatelný.

Z důvodu menší výpočetní náročnosti u kontigů o délce až několik stotisíc bazí a také z důvodu potenciálních lepších výsledků algoritmu se v algoritmu nepočítá s celými kontigy, ale pouze s jejich konci o zvolené délce překryvu (tzv. okno). Vzhledem k tomu, že sekvenátor poskytuje čtení ve směru 5' k 3' označuji v algoritmu konce názvy prefix (počáteční část kontigu) a sufix (koncová část kontigu). Z každého kontigu jsou tedy vystřiženy dvě koncové části o stejné délce shodné u všech kontigů. Délka překryvu by neměla být větší než polovina délky nejkratšího kontigu. Takto získáme knihovnu koncových částí kontigů v signálové podobě.

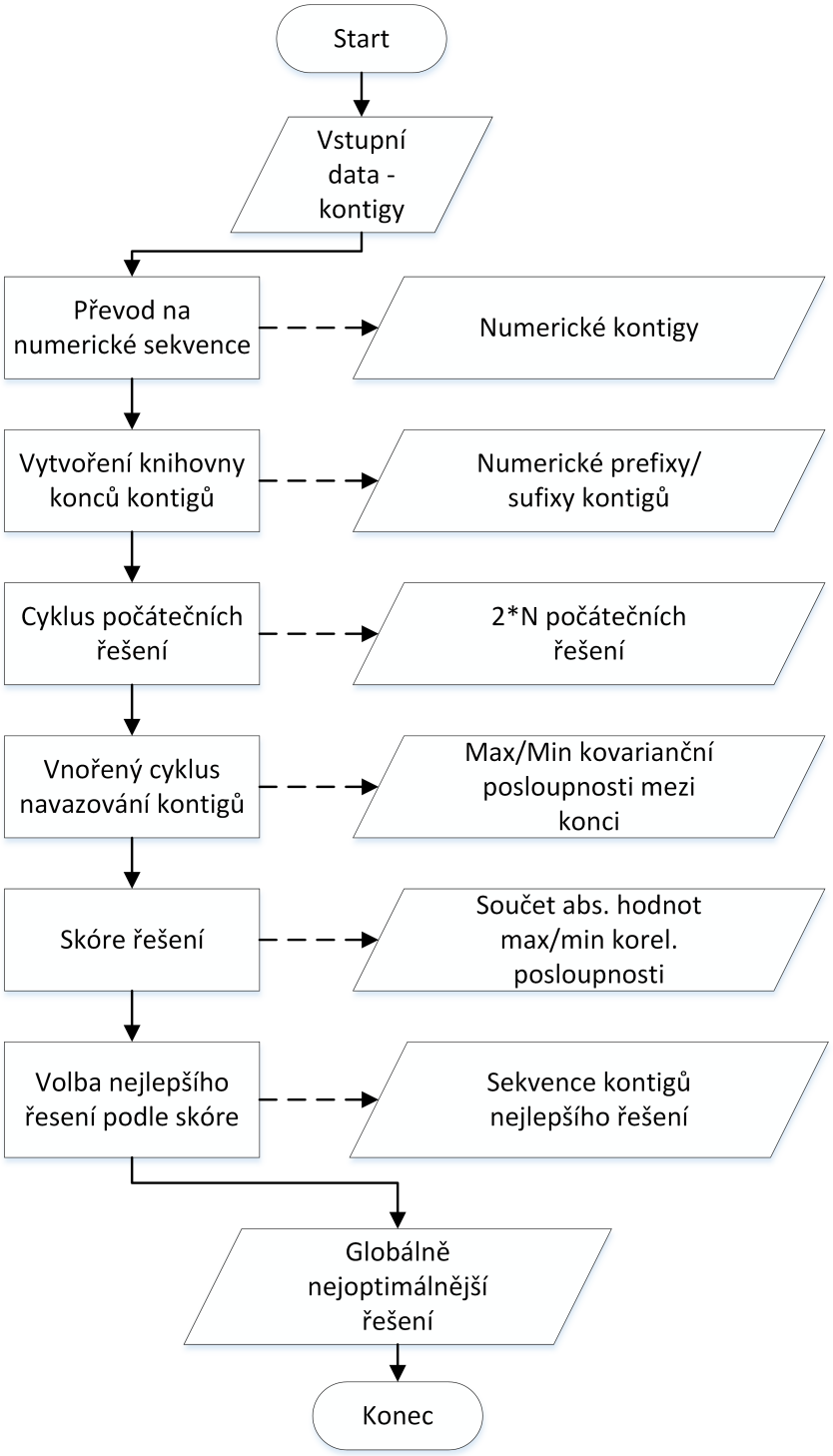
S využitím znalostí o vzájemné kovarianční posloupnosti byl sestaven algoritmus pro hledání překryvů mezi konci kontigů. Algoritmus využívá tzv. greedy přístup, v každém kroku hledá vždy maximální korelační koeficient pro kontig navazující ve směru 5' k 3' a minimální korelační koeficient pro kontigy, které jsou v opačném směru (bylo osekvenováno komplementární vlákno). Jako měřítko pro porovnání obou signálů slouží vzájemná kovariance, jež byla otestována a přijata za vhodný nástroj. Z greedy návrhu algoritmu vyplývá, že řešení algoritmus hledá v daném kroku lokálně nejoptimálnější řešení, ale není zaručeno, že nalezne globální optimum. Samotný návrh algoritmu vychází z řešení problému obchodního cestujího metodou greedy algoritmu. Greedy algoritmus se obecně skládá z několika složek - množiny počátečních řešení, výběrové funkce, funkce proveditelnosti, hodnotící funkce a funkce, která rozhodne o nalezení kompletního řešení. [26]

Vstupem do algoritmu pro hledání překryvů jsou konce kontigů. Generuje se $2N$ počátečních řešení ($N$ je počet kontigů) pro každý prefix a sufix. Ve vnořeném cyklu se dále provádí průběžný výpočet vzájemné kovarianční posloupnosti se všemi ostatními konci kontigů s výjimkou koncové části stejného kontigu. Hodnotícím kritériem v algoritmu je hodnota maxima či minima kovarianční posloupnosti. Na základě stanovených pravidel je vybrán následující kontig, který rozšíří aktuální řešení. Je zohledněno, zde jde o prefix či sufix, tak že se vybírá maximum kovarianční posloupnosti pro kontigy ve stejném směru, či minimum kovarianční posloupnosti pro kontigy, jež odpovídají reverzně komplementárním sekvencím. Dalším kritériem je porovnání s prahovou hodnotou korelačního koeficientu, která je definována zvlášť

pro maxima a minima kovarianční posloupnosti. Pro každé počáteční řešení se postupně sčítají absolutní hodnoty těchto hodnot a řešení s nejvyšším dosaženým skóre je vybráno jako globálně nejoptimálnější řešení.

Algoritmus byl naprogramován v programovém prostředí Matlab. Vstupem funkce je datový soubor s kontigy ve formátu FASTA, prahové hodnoty korelačního koeficientu pro maxima a minima, délka překryvu na kterém se počítá vzájemná kovarianční posloupnost a výběr numerické reprezentace pro převod do signálové podoby (kumulovaná či rozbalená fáze). Výstupem pak je pole celkového řešení pro jednotlivá počáteční řešení a globálně nejoptimálnější řešení.

Na diagramu 6.1 je znázorněno celkové blokové schéma navrženého algoritmu.




Obr. 6.1: Blokové schéma algoritmu

# 7 VÝSLEDKY ALGORITMU

## 7.1 Umělá data

Dalším krokem vyhodnocení kvality algoritmu je testování na datech. Jako data byly použity vytvořené kontigy z celogenomových sekvencí, jež byla nastříhána na náhodně dlouhé úseky s náhodně dlouhými překryvy. Rozsah délek kontigů i rozsah délek překryvů se dal předem určit. Byla tak získána data, u nichž byla apriorní znalost o správném uspořádání správně navazujících kontigů. Pořadí kontigy nebylo nutné dále nějak měnit, takže správné pořadí kontigů je vždy od 1 k poslednímu kontigu v souboru. Vzhledem k odlišnému chování fázových numerických reprezentací pro různé organismy [13] bude algoritmus otestován na datech organismů patřících do říší Archea, Prokaryota a Eukaryota získaných z databáze NCBI. V tabulce 7.1 jsou informace o použitých sekvencích.

Tab. 7.1: Přehled sekvencí DNA z NCBI

| **Latinský název** | **Říše** | **ID NCBI** | **Délka genomu bp** |
|---|---|---|---|
| Ignicoccus hospitalis | Archea | $NC_0$09776.1 | 1 297 538 bp |
| Ostreococcus tauri | Eukaryota | $NC_0$14426.1 | 1 076 297 bp |
| Pragia fontium | Prokayrota | CP010423.1 | 4 094 629 bp |

Tab. 7.2: Přehled parametrů kontigů

| **Sekvence** | **Počet kontigů** | **Délka kontigů** | **Délka překryvů** |
|---|---|---|---|
| Ignicoccus hospitalis | 7 | 160-220 tis. | 500 - 2000 |
| Ostreococcus tauri | 5 | 160-220 tis. | 1000 - 4000 |
| Pragia fontium | 6 | 160-450 tis. | 500-2000 |

## 7.2 Získané uspořádání kontigů a úspěšnost zařazení kontigů

Pro vygenerované soubory obsahující kontigy byly vypočteny vytvořeným algoritmem uspořádání kontigů. Pro každý soubor reprezentující daný organismus bylo

vypočteno toto uspořádání na základě kumulované a rozbalené fáze pro různé hodnoty překryvů. Hodnoty překryvů byly zvoleny v širokém rozmezí délek 500, 1000, 2000, 5000 a 10000. Na velikosti překryvu pro výpočet vzájemné kovariance a (ve skutečném případě) nám neznámé velikosti překryvu mezi kontigy závisí správnost výstupu funkce vzájemné korelace a tedy i algoritmu vůbec. Získaná uspořádání kontigů byla srovnána se skutečným apriorně známým uspořádáním a byla vyhodnocena mírou úspěšnosti. Míra úspěšnosti uspořádání zde se stanovuje jako poměr správně navázaných kontigů a celkovému počtu dvojic kontigů. Správně navázaný je ten kontig, který sousedí s dalším kontigem, který je o jednu hodnotu větší či menší. Mezi takovými kontigy existuje totiž sktečný překryv.

Tab. 7.3: Pragia fontium - uspořádání kontigů

| **Reprezentace** | **Překryv : 500** | **1000** | **2000** | **5000** | **10000** |
|---|---|---|---|---|---|
| Kumulovaná fáze | 1342– | 543612 | 163452 | 356214 | 326541 |
| Úspěšnost [%] | 20 | 60 | 40 | 40 | 60 |
| Rozbalená fáze | 654321 | 2143– | 246351 | 25346- | 243561 |
| Úspěšnost [%] | 100 | 40 | 0 | 20 | 40 |

Pro sekvenci Pragia fontium byly vytvořeny kontigy s překryvy o délce $500-2000$ vzorků a celkem 6 kontigů. Srovnáme-li nejdříve úspěšnost z hlediska volby numerické reprezentace, pro tuto sekvenci nejde zcela jasně rozhodnout, jaká reprezentace poskytuje lepší výsledky. Vyšší průměrnou úspěšnost vykazuje reprezentace kumulovanou fází, ale s reprezentací rozbalenou fází naopak získáme 100% korektní řešení. Pro tuto reprezentaci a výpočetní okno o délce 500 byly nejoptimálnější korelační koeficienty ve zpětném směru. Z hlediska délky výpočetního okna a známé informaci o rozsahu délek překryvů lze v tomto případě usoudit, že kumulovaná fáze lépe pokryvá rozsah délek překryvů, naopak rozbalená fáze dává lepší výsledek pro kratší výpočetní okno.

Z celogenomové sekvence druhu Ignicoccus hospitalis bylo vytvořeno 7 kontigů s rozsahem překryvů $500-2000$. Průměrná úspěšnost uspořádání vychází lépe opět pro kumulovanou fázi, ale nejvyšší výsledek dosahuje uspořádání získané rozbalenou fází - 66% úspěšnost. Reprezentace rozbalenou fází opět poskytuje vyšší úspěšnost při výpočtech s kratším výpočetním oknem, zatímco výslekdy kumulované reprezentace jsou vyrovnanější.

Z celogenomové sekvence druhu Ostreococcus tauri bylo vytvořeno 5 kontigů s rozsahem překryvů $1000-4000$. Již ze znalosti předchozích výsledků a této informace o rozsahu překryvů lze očekávat, že vyšší úspěšnost bude náležet výpočetním oknům

Tab. 7.4: Ignicoccus hospitalis - uspořádání kontigů

| **Reprezentace** | **Překryv : 500** | **1000** | **2000** | **5000** | **10000** |
|---|---|---|---|---|---|
| Kumulovaná fáze | 124367- | 7612435 | 5126734 | 1534762 | 4256713 |
| Úspěšnost [%] | 50 | 50 | 50 | 33 | 33 |
| Rozbalená fáze | 4563217 | 376125- | 4256137 | 1327456 | 315247- |
| Úspěšnost [%] | 66 | 33 | 16 | 50 | 0 |

Tab. 7.5: Ostreococcus tauri - uspořádání kontigů

| **Reprezentace** | **Překryv : 500** | **1000** | **2000** | **5000** | **10000** |
|---|---|---|---|---|---|
| Kumulovaná fáze | 43512 | 52134 | 42315 | 21543 | 12543 |
| Úspěšnost [%] | 50 | 50 | 25 | 75 | 75 |
| Rozbalená fáze | 24135 | 23145 | 23145 | 5134- | 53412 |
| Úspěšnost [%] | 0 | 25 | 25 | 25 | 50 |

o větší délce. Tento fakt se potvrzuje pro obě reprezentace. Pro kontigy této sekvence však lze konstatovat, že reprezentace kumulované fáze poskytuje lepší výsledky.

Tab. 7.6: Ostreococcus tauri + 2 cizí kontigy - uspořádání kontigů

| **Reprezentace** | **Překryv : 500** | **1000** | **2000** | **5000** | **10000** |
|---|---|---|---|---|---|
| Kumulovaná fáze | 45826— | 7634521- | 28516743 | 82654173 | 1653872- |
| Senzitivita | 0,25 | 1 | 1 | 1 | 0,5 |
| Specifita | 0 | 0,25 | 0 | 0 | 0 |
| Rozbalená fáze | 523187– | 621578– | 61234578 | 84126357 | 65213478 |
| Senzitivita | 0,33 | 0,33 | 1 | 1 | 1 |
| Specifita | 0 | 0 | 0 | 0 | 0 |

K předchozímu souboru kontigů Ostreococcus tauri byly přidány navíc dva cizí kontigy (celkem 8), které do souboru nepatří, aby se otestoval algoritmus v úspěšnosti (ne)detekce neodpovídajících kontigů. Při pohledu na výsledky a srovnáním s výsledky teoretických předpokladů, kde se ukázalo, že metodika vzájemné kovariance v některých případech může poskytnou falešně pozitivní výsledek. Pokud

se zde zaměříme na úspěšnost z hlediska nezařazení nepatřících kontigů a využijeme možnosti aplikovat výpočet senzivitiy a specifity na základě přiřazení, že FP - špatně zařazená dvojice, TP - správně zařazená dvojice, FN - chybějící správné zařazení, TN - chybějící, ale špatné zařazení, získáme výsledky senzitivity a specifity pro algoritmus a detekování kontigů bez překryvu. Specifita, označující schopnost detekovat jako negativní všechny skutečně negativní objekty, vychází velmi nedostatečně. Senzitivita, jež vyjadřuje schopnost správně detekovat všechny skutečně pozitivní objekty jako pozitivní, vychází na hodnoty v některých případech rovné 1.

### 7.2.1 Srovnání s jinými algoritmy

Algoritmy zmíněné v práci v kapitole o neúplných genomových sestaveních jsou užitečné nástroje, ale všechny ke svému běhu požadují jako vstup data navíc, ať už jde o surová čtení či zarovnané úseky sekvencí. Algoritmus navržený v této práci se zaměřuje pouze na práci se soubory kontigů. Využití numerických reprezentací je vhodné zařadit jako dílčí případně srovnávací mezikrok při práci na neúplných genomových sestaveních. Výpočetní náročnost a rychlost zpracování instrukcí při práci s jednorozměrnými genomickými signály je velmi nízká i na standardním počítači, zatímco srovnávané algoritmy vyžadovaly serverové výkonnější počítače. Algoritmus také nevyžaduje žádné zarovnávání sekvencí, což je jedna z nejnáročnějších operací prováděných v bioinformatice (zvláště k faktu, že pracujeme s kontig a čteními). Výpočetní nenáročnost tedy patří mezi nesporné výhody tohoto řešení spolu s nutností mít pouze datový soubor neúplného genomového sestavení.

# 8 ZÁVĚR

Cílem práce bylo navržení algoritmu pro využití numerických reprezentaci při sestavování genomu. Nosným tématem práce byly metody sestavování genomu. Aby byla práce seřazena chronologicky, byla nejdříve uvedena stručná historie objevu DNA a popis struktury DNA. Navazovala kapitola o sekvenačních metodách zahrnující sekvenátory od Sangerovy metody, která je i dnes považována za zlatý standard, až po experimentální metody poslední generace sekvenátorů jako je Nanopore. Stručně byly zmíněn nejpoužívanější formát pro ukládání čtení ze sekvenátorů FASTQ a Phred skóre pro posuzování kvality sekvenace. Pro pochopení sestavování genomu a náhledu do komplexní problematiky je nastudování sekvenačních metod důležité, neboť v další kapitole popsané metody sestavování genomu na práci sekvenátorů navazují. Je též důležité uvědomit si rozdíly mezi assemblery a jejich vhodnosti při sestavování genomu z dat různých sekvenátorů. Tomu se věnuje závěrečná část kapitoly o genomovém sestavování. Ve čtvrté kapitole byly představeny numerické reprezentace, z nichž fázové reprezentace byly vybrány pro využití v praktické části práce. Numerickými reprezentacemi jsem se zabýval už ve své bakalářské práci a fázové reprezentace se prokázaly jako vhodné z důvodu svojí jednorozměrnosti podél sekvence.

Na teoretickou část práce navazuje praktická, kde jsem navrhl algoritmus pro hledání překryvů mezi kontigy na základě numerických reprezentací. Využití numerických reprezentací poskytuje možnost, jak využít metody digitálního zpracování signálů při standardně znakově prováděné operaci, jakou sestavení genomu je. Zařazuje se tak mezi další možnosti, kdy byly numerické reprezentace využity. Rovněž byla diskutována volba ideální porovnávací funkce pro genomické signály kontigů, kde se uplatnila vzájemná kovarianční posloupnost. Diskutovány byly teoretické předpoklady pro správné fungování algoritmu jako správné detekování signálu reverzně komplementární sekvence, prahování korelačních koeficientů a případných nedostatků řešení.

V poslední části byl algoritmus otestován na souboru kontigů skutečných sekvencí s apriorní znalostí o jejich správném pořadí, které pak mohlo být dále vyhodnoceno a diskutováno. Pro využití se osvědčily obě fázové reprezentace, ale s odlišnými dílčími vlastnostmi. Pro úspěch algoritmu je důležité nastavení délky překryvu, na kterém se počítá vzájemná kovariance a následně optimální hodnoty této posloupnosti pro hledání překryvu. Ačkoliv navržená metoda hledání překryvů nemá absolutní úspěšnost, je výpočetně nenáročná a nepotřebuje další pomocná data. Lze ji využít jako dílčí krok či referenci při práci s běžně užívanými nástroji.

# LITERATURA


[1] Abo-Zahhad, M.; Ahmed, S. M.; Abd-Elrahman, S. a.: Genomic Analysis and Classification of Exon and Intron Sequences Using DNA Numerical Mapping Techniques. *International Journal of Information Technology and Computer Science*, ročník 4, č. 8, 2012: s. 22–36, ISSN 20749007, doi:10.5815/ijitcs.2012.08.03.

[2] Anastassiou, D.: Genomic signal processing. *IEEE Signal Processing Magazine*, ročník 18, č. 4, 2001: s. 8–20, ISSN 10535888, doi:10.1109/79.939833.

[3] Ariyaratne, P. N.; Sung, W. K.: PE-Assembler: De novo assembler using short paired-end reads. *Bioinformatics*, ročník 27, č. 2, 2011: s. 167–174, ISSN 13674803, doi:10.1093/bioinformatics/btq626.

[4] Avery, O. T.; Macleod, C. M.; McCarty, M.: Studies on the Chemical Nature of the Substance Inducing Transformation of Pneumococcal Types : Induction of Transformation By a Desoxyribonucleic Acid Fraction Isolated From Pneumococcus Type Iii. *The Journal of experimental medicine*, ročník 79, č. 2, 1944: s. 137–58, ISSN 0022-1007, doi:10.1084/jem.79.2.137.
URL <http://www.pubmedcentral.nih.gov/articlerender.fcgi&artid=2135445&tool=pmcentrez&rendertype=abstract>

[5] Baker, M.: De novo genome assembly: what every biologist should know. *Nature Methods*, ročník 9, č. 4, 2012: s. 333–337, ISSN 1548-7091, doi:10.1038/nmeth.1935.

[6] Bradnam, K. R.; Fass, J. N.; Alexandrov, A.; aj.: Assemblathon 2: evaluating de novo methods of genome assembly in three vertebrate species. *GigaScience*, ročník 2, č. 1, 2013: str. 10, ISSN 2047-217X, doi:10.1186/2047-217X-2-10, <1301.5406>.
URL <http://www.gigasciencejournal.com/content/2/1/10http://www.pubmedcentral.nih.gov/articlerender.fcgi&artid=3844414&tool=pmcentrez&rendertype=abstract>

[7] Butler, J.; MacCallum, I.; Kleber, M.; aj.: ALLPATHS : De novo assembly of whole-genome shotgun microreads. ročník 18, 2008: s. 810–820, doi:10.1101/gr.7337908.

[8] Carr, S. M.: Tetranucleotide Hypothesis.
URL <https://www.mun.ca/biology/scarr/Tetranucleotide_Hypothesis.html>

[9] Chaisson, M. J.; Brinza, D.; Pevzner, P. a.: De novo fragment assembly with short mate-paired reads: Does the read length matter? *Genome Research*, ročník 19, č. 2, 2009: s. 336–346, ISSN 10889051, doi:10.1101/gr.079053.108.

[10] Chargaff, E.; Zamenhof, S.; Green, C.: Composition of Human Desoxypentose Nucleic Acid. *Nature*, ročník 165, č. 4202, may 1950: s. 756–757, ISSN 0028-0836, doi:10.1038/165756b0.
URL <http://www.nature.com/nature/journal/v165/n4202/pdf/165756b0.pdf>

[11] Cock, P. J. A.; Fields, C. J.; Goto, N.; aj.: The Sanger FASTQ file format for sequences with quality scores, and the Solexa/Illumina FASTQ variants. *Nucleic Acids Research*, ročník 38, č. 6, 2010: s. 1767–1771, ISSN 0305-1048, doi:10.1093/nar/gkp1137.
URL <http://nar.oxfordjournals.org/lookup/doi/10.1093/nar/gkp1137>

[12] Compeau, P. E. C.; Pevzner, P. a.; Tesler, G.: How to apply de Bruijn graphs to genome assembly. *Nature Biotechnology*, ročník 29, č. 11, 2011: s. 987–991, ISSN 1087-0156, doi:10.1038/nbt.2023, <nbt.2023>.
URL <http://dx.doi.org/10.1038/nbt.2023>

[13] Cristea, P.: Phase analysis of DNA genomic signals. *Proceedings of the 2003 International Symposium on Circuits and Systems, 2003. ISCAS '03.*, ročník 5, 2003: s. V–25–V–28, doi:10.1109/ISCAS.2003.1206163.

[14] Cristea, P. D.: Conversion of nucleotides sequences into genomic signals. *Journal of cellular and molecular medicine*, ročník 6, č. 2, 2002: s. 279–303, ISSN 1582-1838.

[15] Cristea, P. D.: Large scale features in DNA genomic signals. *Signal Processing*, ročník 83, č. 4, apr 2003: s. 871–888, ISSN 01651684, doi:10.1016/S0165-1684(02)00477-2.

[16] Dahm, R.: Discovering DNA: Friedrich Miescher and the early years of nucleic acid research. *Human Genetics*, ročník 122, č. 6, 2008: s. 565–581, ISSN 03406717, doi:10.1007/s00439-007-0433-0.

[17] Dohm, J. C.; Lottaz, C.; Borodina, T.; aj.: SHARCGS, a fast and highly accurate short-read assembly algorithm for de novo genomic sequencing. *Genome Research*, ročník 17, č. 11, 2007: s. 1697–1706, ISSN 10889051, doi:10.1101/gr.6435207.

[18] Ekblom, R.; Wolf, J. B. W.: A field guide to whole-genome sequencing, assembly and annotation. *Evolutionary Applications*, ročník 7, č. 9, 2014: s. 1026–1042, ISSN 17524571, doi:10.1111/eva.12178.

[19] Griffith, F.: The Significance of Pneumococcal Types. *Journal of Hygiene*, ročník 27, č. 02, 1928: str. 113, ISSN 0022-1724, doi:10.1017/S0022172400031879. URL <http://www.journals.cambridge.org/abstract_S0022172400031879>

[20] Illumina: Quality Scores for Next-Generation Sequencing. 2000: s. 1–2. URL <http://www.illumina.com/documents/products/technotes/technote_Q-Scores.pdf>

[21] Illumina: A Sampling of Assemblers for Short Reads. 2010. URL <https://www.illumina.com/Documents/products/technotes/technote_denovo_assembly_ecoli.pdf>

[22] Illumina.com: Illumina Next-Generation Sequencing Platforms. 2015. URL <http://www.illumina.com/systems/sequencing-platform-comparison.html>

[23] Jan, J.: *Číslicová filtrace, analýza a restaurace signálů*. Brno: VUTIUM, druhé vydání, 2002, ISBN 80-214-1558-4, 427 s.

[24] Kozumplík, J.: Umělá inteligence v medicíně : Shluková analýza. 2013. URL <https://www.vutbr.cz/studium/ects-katalog/detail-predmetu&apid=160494>

[25] Kozumplík, J.: Pokročilá analýza biologických signálů. 2014. URL <https://www.vutbr.cz/studium/ects-katalog/detail-predmetu&apid=149086>

[26] Kozumplík, J.; Mézl, M.: Evoluční algoritmy : Problém TSP. 2015. URL <https://www.vutbr.cz/studium/ects-katalog/detail-predmetu&apid=160958>

[27] Lam, K.-k.; Hall, R.; Clum, A.: BIGMAC : Breaking Inaccurate Genomes and Merging Assembled Contigs for long read metagenomic assembly. 2016, doi: 10.1101/045690.

[28] Langmead, B.: Assembly  shortest common superstring. 2014. URL <http://www.cs.jhu.edu/$\sim$langmea/resources/lecture_notes/assembly_scs.pdf>

[29] Langmead, B.: De Bruijn Graph and Genome Assembly. 2014.
URL <http://www.cs.jhu.edu/$\sim$langmea/resources/lecture_notes/assembly_dbg.pdf>

[30] Langmead, B.: Overlap Layout Consensus assembly. 2014.
URL <http://www.cs.jhu.edu/$\sim$langmea/resources/lecture_notes/assembly_olc.pdf>

[31] LifeTechnologies: De novo assembly using Ion semiconductor sequencing. 2013.
URL <https://tools.thermofisher.com/content/sfs/brochures/de_novo_assembly_Ion_CO23721_App_Note_V9.pdf>

[32] Lifetechnologies: 5500 W Series Genetic Analyzers Sheet. 2015.
URL <https://tools.thermofisher.com/content/sfs/brochures/5500-w-series-spec-sheet.pdf>

[33] Masoudi-Nejad, A.; Narimani, Z.; Hosseinkhan, N.: *Next Generation Sequencing and Sequence Assembly*. SpringerBriefs in Systems Biology, New York, NY: Springer New York, 2013, ISBN 978-1-4614-7725-9, 86 s., doi:10.1007/978-1-4614-7726-6.
URL <http://link.springer.com/10.1007/978-1-4614-7726-6>

[34] Maxam, a. M.; Gilbert, W.: A new method for sequencing DNA. *Proceedings of the National Academy of Sciences of the United States of America*, ročník 74, č. 2, 1977: s. 560–564, ISSN 0027-8424, doi:10.1073/pnas.74.2.560.

[35] Mendel, G.: Versuche über Plflanzenhybriden. *Verhandlungen des naturforschenden Vereines in Brünn, Bd. IV für das Jahr 1865, Abhandlungen,*, 1866: s. 3–47.
URL <http://publikationen.ub.uni-frankfurt.de/home/index/language/language/de/rmodule/frontdoor/rcontroller/deliver/raction/index/docId/9193/file/mendel_screen.pdf>

[36] Miller, J. R.; Koren, S.; Sutton, G.: Assembly algorithms for next-generation sequencing data. *Genomics*, ročník 95, č. 6, 2010: s. 315–327, ISSN 08887543, doi:10.1016/j.ygeno.2010.03.001.
URL <http://dx.doi.org/10.1016/j.ygeno.2010.03.001>

[37] Nečas, O.; Svoboda, A.; Hejtmánek, M.; aj.: *Obecná biologie pro lékařské fakulty*. Nakladatelství H+H, Jinočany, třetí vydání, 2000, ISBN 80-86022-46-3, 555 s.

[38] Pevzner, P. a.; Tang, H.; Waterman, M. S.: An Eulerian path approach to DNA fragment assembly. *Proceedings of the National Academy of Sciences of the*

*United States of America*, ročník 98, č. 17, 2001: s. 9748–9753, ISSN 00278424, doi:10.1073/pnas.171285098.

[39] Proakis, J. G.; Manolakis, D. G.: *Digital signal processing : principles, algorithms, and applications*. New Jersey: Prentice Hall, třetí vydání, 1996, ISBN 0133737624, 968 s.

[40] Sanger, F.; Nicklen, S.; Coulson, A. R.: DNA sequencing with chain-terminating inhibitors. *Proceedings of the National Academy of Sciences of the United States of America*, ročník 74, č. 12, 1977: s. 5463–7, ISSN 0027-8424, doi:10.1073/pnas.74.12.5463, <0402594v3>.
URL <http://www.pubmedcentral.nih.gov/articlerender.fcgi&artid=431765&tool=pmcentrez&rendertype=abstract>

[41] Schadt, E. E.; Turner, S.; Kasarskis, A.: A window into third-generation sequencing. *Human Molecular Genetics*, ročník 19, č. R2, 2010: s. 227–240, ISSN 09646906, doi:10.1093/hmg/ddq416.

[42] Scheibye-Alsing, K.; Hoffmann, S.; Frankel, A.; aj.: Sequence assembly. *Computational Biology and Chemistry*, ročník 33, č. 2, 2009: s. 121–136, ISSN 14769271, doi:10.1016/j.compbiolchem.2008.11.003.
URL <http://linkinghub.elsevier.com/retrieve/pii/S1476927108001497>

[43] Science, R. L.: 454 Products. 2015.
URL <http://454.com/products/index.asp>

[44] Sedlář, K.: Laboratorní přístroje v genomice a proteomice - Sekvenace DNA. 2015.

[45] Skutkova, H.; Vitek, M.; Babula, P.; aj.: Classification of genomic signals using dynamic time warping. *BMC Bioinformatics*, ročník 14, č. Suppl 10, 2013: str. S1, ISSN 1471-2105, doi:10.1186/1471-2105-14-S10-S1.

[46] Thermofisher: Ion S5 and Ion S5 XL Systems. 2015.
URL <https://tools.thermofisher.com/content/sfs/brochures/Ion-S5-S5XL-Brochure.pdf>

[47] Tsai, I. J.; Otto, T. D.; Berriman, M.: Improving draft assemblies by iterative mapping and assembly of short reads to eliminate gaps. *Genome biology*, ročník 11, č. 4, 2010: str. R41, ISSN 1465-6906, doi:10.1186/gb-2010-11-4-r41.

[48] Walker, B. J.; Abeel, T.; Shea, T.; aj.: Pilon: An integrated tool for comprehensive microbial variant detection and genome assembly improvement. *PLoS ONE*, ročník 9, č. 11, 2014, ISSN 19326203, doi:10.1371/journal.pone.0112963.

[49] Warren, R. L.; Sutton, G. G.; Jones, S. J. M.; aj.: Assembling millions of short DNA sequences using SSAKE. *Bioinformatics*, ročník 23, č. 4, 2007: s. 500–501, ISSN 1367-4803, doi:10.1093/bioinformatics/btl629.
URL <http://bioinformatics.oxfordjournals.org/cgi/doi/10.1093/bioinformatics/btl629>

[50] Watson, J.; Crick, F.: A structure for deoxyribose nucleic acid. *Nature*, 1953.

[51] Whatisbiotechnology.org: The path to sequencing nucleic acids. 2015.
URL <http://www.whatisbiotechnology.org/exhibitions/sanger/path>

[52] Zerbino, D. R.; Birney, E.: Velvet: algorithms for de novo short read assembly using de Bruijn graphs. *Genome research*, ročník 18, č. 5, 2008: s. 821–9, ISSN 1088-9051, doi:10.1101/gr.074492.107, <0209100>.
URL <http://www.ncbi.nlm.nih.gov/pubmed/18349386>

# SEZNAM PŘÍLOH

# A PŘÍLOHA A

## A.1 Obsah přiloženého CD

- Složka *Funkce DraftAssembler+data* - Obsahuje volatelnou funkci a sekvence pro spuštění spolu s návodem
- Složka *Skript DraftAssembler+data* - Obsahuje skript a sekvence
- Složka *x data fasta ncbi* - Obsahuje původní sekvence pro tvorbu kontigů
- Složka *x skripty pro tabulky a grafy* - Obsahuje pomocné skripty