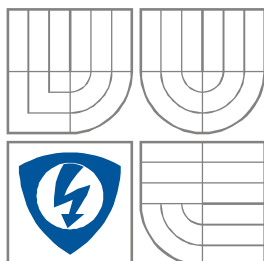


VYSOKÉ UČENÍ TECHNICKÉ V BRNĚ
BRNO UNIVERSITY OF TECHNOLOGY



FAKULTA ELEKTROTECHNIKY A KOMUNIKAČNÍCH
TECHNOLOGIÍ
ÚSTAV BIOMEDICÍNSKÉHO INŽENÝRSTVÍ

FACULTY OF ELECTRICAL ENGINEERING AND COMMUNICATION
DEPARTMENT OF BIOMEDICAL ENGINEERING

ODHAD ENTROPIE A KOMPRESSE BIOLOGICKÝCH SEKVENCÍ

ENTROPY RATE ESTIMATION AND COMPRESSION OF BIOLOGICAL SEQUENCES

DIPLOMOVÁ PRÁCE
MASTER'S THESIS

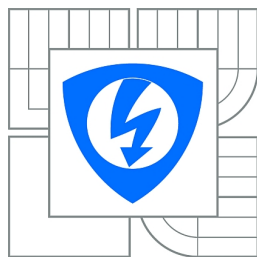
AUTOR PRÁCE
AUTHOR

Bc. Peter Miščík

VEDOUCÍ PRÁCE
SUPERVISOR

Ing. Helena Škutková

BRNO 2013



**VYSOKÉ UČENÍ
TECHNICKÉ V BRNĚ**

**Fakulta elektrotechniky
a komunikačních technologií**

Ústav biomedicínského inženýrství

Diplomová práce

magisterský navazující studijní obor
Biomedicínské a ekologické inženýrství

Student: Bc. Peter Miščík

ID: 106646

Ročník: 2

Akademický rok: 2012/2013

NÁZEV TÉMATU:

Odhad entropie a komprese biologických sekvencí

POKYNY PRO VYPRACOVÁNÍ:

1) Seznamte se s problematikou entropie a komprese dat v bioinformatice. 2) Proved'te srovnání odlišností v kódování DNA a proteinů, zohledněte obsaženou genetickou informaci a možnosti zjednodušení použité "abecedy" do jednodušších celků. 3) Navrhněte algoritmus pro vyhledání strukturálních podobností a opakovaných vzorů umožňující následnou kompresi biologických sekvencí. 4) Algoritmus otestujte na reálných sekvencích z veřejných databází pomocí programového prostředí Matlab. 5) Vytvořte program s uživatelským rozhraním umožňující odhad entropie, kompresi a zpětnou dekompresi biologické sekvence.

DOPORUČENÁ LITERATURA:

[1] HAYES, Brian. The Invention of the genetic code. American Scientist: the magazine of Sigma Xi, the Scientific Research Societ. 1998, roč. 86, č. 1, s. 8-14.

[2] RAJESWARI, Raja a Allam APPARAO. GENBIT COMPRESS – algorithm for repetitive and non-repetitive DNA sequences. Journal of Theoretical and Applied Information Technology. 2010, roč. 11, č. 1, s. 25-29.

Termín zadání: 11.2.2013

Termín odevzdání: 24.5.2013

Vedoucí práce: Ing. Helena Škutková

Konzultanti diplomové práce:

prof. Ing. Ivo Provazník, Ph.D.

Předseda oborové rady

UPOZORNĚNÍ:

Autor diplomové práce nesmí při vytváření diplomové práce porušit autorská práva třetích osob, zejména nesmí zasahovat nedovoleným způsobem do cizích autorských práv osobnostních a musí si být plně vědom následků porušení ustanovení § 11 a následujících autorského zákona č. 121/2000 Sb., včetně možných trestněprávních důsledků vyplývajících z ustanovení části druhé, hlavy VI. díl 4 Trestního zákoníku č.40/2009 Sb.

Abstrakt

Táto diplomová práca popisuje poznatky o biologických sekvenciách, princípy odhadu entropie a možnosti kompresie DNA sekvencií pomocou substitučných metód. Text obsahuje praktickú časť, kde sú využité kompresné algoritmy a praktický odhad entropie.

Kľúčové slová

entropia, kompresia, biologické sekvencie, DNA, matlab, fasta

Abstract

This master thesis describes theoretical knowledge of biological sequences, principles entropy rate estimates and possibilities of compression of DNA sequences using the substitution methods. Thesis includes practical application of the compression algorithm and practical estimation of entropy.

Keywords

entropy, compression, biological sequences, DNA, matlab, fasta

Bibliografická citácia

MIŠČÍK, P. *Odhad entropie a komprese biologických sekvencí*. Brno: Vysoké učení technické v Brně, Fakulta elektrotechniky a komunikačních technologií, 2013. 46 s. Vedoucí diplomové práce Ing. Helena Škútková.

Prohlášení

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

V Brně dne 24. května 2013

.....

podpis autora

Pod'akovanie

Ďakujem vedúcej diplomovej práce Ing. Helene Škútkovej za účinnú metodickú, pedagogickú a odbornú pomoc a ďalšie cenné rady pri spracovaní mojej diplomovej práce.

V Brně dne 24. května 2013

.....

podpis autora

Obsah

Úvod	1
1 Molekulárna biológia	2
1.1 Genetický kód	2
1.1.1 Štruktúra DNA	2
1.1.2 Párovanie dusíkatých báz	3
1.2 Centrálna dogma molekulárnej biológie	4
1.2.1 Replikácia DNA	5
1.2.2 Transkripcia do RNA	6
1.2.3 Preklad z RNA do proteínov	7
2 Biologické sekvencie	9
2.1 Sekvencie DNA, RNA	9
2.2 Organizácia nukleotidových sekvencií DNA	10
2.3 Sekvencie proteínov	11
2.4 Formáty biologických sekvencií	12
2.4.1 RAW	12
2.4.2 FASTA	12
2.4.3 PIR/NBRF	12
2.4.4 GenBank/DDBJ/EMBL <i>flatfile</i>	12
3 Odhad entropie, kompresné algoritmy	13
3.1 Odhad entropie	13
3.2 Kompresné algoritmy (Substitučná metóda)	15
3.2.1 Lempel-Ziv	16
3.2.2 Biocompress	17
3.2.3 Cfact	18
3.2.4 GenCompress	18
3.2.5 DNACompress	19

3.2.6	CTW+LZ.....	19
3.2.7	DNAC.....	19
3.2.8	GeNML	19
3.2.9	DNAPack	19
3.3	Experimentálne výsledky	20
4	Analýza entropie a kompresie DNA	22
4.1	Odhad entropie	22
4.2	Kompresia DNA	26
4.2.1	GenbitCompress	26
4.2.2	KompresDNA.....	29
4.3	Analýza kódujúcich a nekódujúcich úsekov DNA sekvencie	36
4.4	Porovnanie výsledkov.....	39
4.5	Popis programu.....	39
	Záver.....	42
	Zoznam literatúry	43
	Zoznam skratiek	45
	Prílohy	46

Zoznam obrázkov

<i>Obrázok 1.1 Dvoj závitnicová štruktúra [22].</i>	4
<i>Obrázok 1.2 Pohyb biologickej informácie v organizmoch [22].</i>	5
<i>Obrázok 1.3 Replikácia DNA [8].</i>	6
<i>Obrázok 1.4 Transkripcia do RNA [23].</i>	7
<i>Obrázok 2.1 Sekvencie DNA Sangerovou metódou [24].</i>	9
<i>Obrázok 2.2 Primárne sekvencie štruktúry jednoreťazovej molekuly tRNA [25].</i>	10
<i>Obrázok 2.3 Mechanizmus splicing.</i>	11
<i>Obrázok 2.4 Diagram príslušnosti aminokyselín do skupín [8].</i>	11
<i>Obrázok 2.5 Sekvencie aminokyselín vo formáte FASTA [7].</i>	12
<i>Obrázok 3.1: Logo pre E.coli [3].</i>	15
<i>Obrázok 3.2: Ukážka predošlých súvislosti medzi vstupným reťazcom, skomprimovanej časti a optimálnym prefixom [6].</i>	18
<i>Obrázok 4.1: Úsek DNA sekvencie vo formáte FASTA.</i>	22
<i>Obrázok 4.2: Shannonova entropia HUMGHCSA, HUMHDABCD, HUMHPRTB.</i>	23
<i>Obrázok 4.3: Shannonova entropia HUMDYSTROP.</i>	23
<i>Obrázok 4.4: Shannonova entropia MPOMTCG, VACCG.</i>	23
<i>Obrázok 4.5: Závislosť entropie na počtu znakov ($N = 100\,000$).</i>	24
<i>Obrázok 4.6: Maximálna entropia vírusu Copenhagen.</i>	25
<i>Obrázok 4.7: Príklad kompresie.</i>	27
<i>Obrázok 4.8: Príklad dekompresie.</i>	27
<i>Obrázok 4.9: Kompresný pomer chromozómu 7 a 10.</i>	29
<i>Obrázok 4.10: Myšlienka kompresnej metódy.</i>	30
<i>Obrázok 4.11: Kódovanie zhody.</i>	31
<i>Obrázok 4.12: Vývojový diagram kompresie.</i>	32
<i>Obrázok 4.13: Výstupné bitové reťazce a ich dĺžky.</i>	33
<i>Obrázok 4.14: Upravená metóda RLE.</i>	33

<i>Obrázok 4.15: Vývojový diagram dekompresie.</i>	<i>34</i>
<i>Obrázok 4.16: Vývoj kompresného pomeru na chromozómoch.</i>	<i>36</i>
<i>Obrázok 4.17: Grafické zobrazenie DNA sekvencie.</i>	<i>37</i>
<i>Obrázok 4.18: Shannova entropia kódujúceho a nekódujúceho úseku sekvencie.</i>	<i>37</i>
<i>Obrázok 4.19: Kompresia kódujúceho a nekódujúceho úseku metódou KompresDNA.</i>	<i>38</i>
<i>Obrázok 4.20: Kompresia kódujúceho a nekódujúceho úseku metódou GenbitCompress.</i>	<i>38</i>
<i>Obrázok 4.21: Priebeh kompresie.</i>	<i>40</i>
<i>Obrázok 4.22: Priebeh dekompresie.</i>	<i>40</i>
<i>Obrázok 4.23: Sequence Tool.</i>	<i>41</i>

Zoznam tabuliek

<i>Tabuľka 1.1: Prehľad aminokyselín.</i>	<i>8</i>
<i>Tabuľka 3.1: Príklad kompresie vstupnej sekvencie AABABBBABB pomocou LZ77.</i>	<i>16</i>
<i>Tabuľka 3.2: Príklad kompresie vstupnej sekvencie AABABBBABB pomocou LZ78.</i>	<i>17</i>
<i>Tabuľka 3.3: Informácie o štandardných DNA sekvenciách.</i>	<i>20</i>
<i>Tabuľka 3.4: Kompresný pomer dosiahnutý rôznymi kompresnými technikami [19].</i>	<i>20</i>
<i>Tabuľka 3.5: Kompresný zisk rôznych DNA kompresných metód [19].</i>	<i>21</i>
<i>Tabuľka 4.1: Prehľad výsledkov Shannonovej entropie.</i>	<i>24</i>
<i>Tabuľka 4.2: Shannonova entropia chromozómov, $N = 100\,000$.</i>	<i>25</i>
<i>Tabuľka 4.3: Výsledky kompresnej metódy GenbitCompress.</i>	<i>28</i>
<i>Tabuľka 4.4: GenbitCompress na chromozómoch človeka.</i>	<i>28</i>
<i>Tabuľka 4.5: Výsledky kompresnej metódy KompresDNA.</i>	<i>35</i>
<i>Tabuľka 4.6: KompresDNA na chromozómoch človeka.</i>	<i>35</i>
<i>Tabuľka 4.7: Porovnanie výsledkov.</i>	<i>39</i>

Úvod

Vzhľadom na narastajúce množstvo dát, uložených v genetických bankách, ktoré sa získavajú zdokonalenými technológiami sekvenovania, banky obsahujú genómy tisícich vírov, baktérii, mnohobunečných organizmov od rastlín až po ľudí. Tieto genómy sú uložené ako textové súbory, ktoré nepredstavujú najvhodnejšiu voľbu pri prenášaní týchto dát. Preto bolo za potreby vytvorenie kompresných algoritmov špecializovaných na biologické sekvencie. Ku kompresii sú veľmi vhodné DNA sekvencie so svojou vysokou redundanciou a dokonalou štruktúrou, pri poškodení časti sekvencie sa dokážu plne rekonštruovať na pôvodnú podobu, je to jeden zo samo opravných mechanizmov DNA.

Cieľom mojej diplomovej práce je poukázať na možnosti odhadu entropie a kompresie biologických sekvencií. Dôraz mal byť kladený na vyhľadávanie opakujúcich sa vzorov s možnosťou následnej kompresie DNA sekvencií.

Práca by sa dala rozdeliť na tri časti. V prvej časti sú popísané základy molekulárnej biológie, centrálna dogma molekulárnej biológie, štruktúry biologických sekvencií a ich vzájomných súvislostí, typy jednotlivých formátov biologických sekvencií použiteľných k následnej kompresii.

Druhá časť obsahuje základný popis informačnej teórie odhadu entropie, ktorá nám dáva matematický podklad k praktickému vypočítaniu entropie biologických sekvencií. Nasleduje prehľad kompresných algoritmov, bezstratovej kompresie substitučnou metódou, ktorá sa dá využiť k vyhľadávaniu opakujúcich sa vzorov. Ďalej sú rozobraté najznámejšie kompresné metódy, ktoré sú orientované ku kompresii DNA sekvencie.

Tretia časť je zameraná na praktické využitie vyššie popísaných znalostí. Realizovaný odhad entropie úsekov DNA sekvencií. Realizovanie kompresných algoritmov k vyhľadávaniu a kompresii opakujúcich sa vzorov, ich kompresia a výpočet kompresného pomeru. Práca ďalej obsahuje analýzu kódujúcich a nekódujúcich častí DNA sekvencie, taktiež popis programu s užívateľským rozhraním pre kompresiu a následnú dekompresiu DNA sekvencie.

1 Molekulárna biológia

Vedná disciplína zaoberajúca sa štúdiom bunecných biologických procesov na ich molekulárnej úrovni. Podstata niektorých biologických javov je odhaliteľná jedine štúdiom ich molekulárnej podstaty. Zvláštna pozornosť je predovšetkým venovaná funkcii makromolekúl podieľajúcich sa na dedičnosti organizmov, takže DNA, RNA a proteínom ich vzájomnej interakcii a regulácii ich funkcie. Molekulárne biologické znalosti sú široko využívané v medicíne a sú potrebné pre genetické inžinierstvo, ako aj pre akúkoľvek inú prácu s genetickou informáciou [9].

1.1 Genetický kód

Všetky živé tvory na našej planéte majú jednu spoločnú vlastnosť. Majú schopnosť uchovávať, spracovávať a predávať obrovské množstvo informácií. Je to základné kritérium, ktoré odlišuje všetko živé od veci neživých. Genetický kód predstavuje súbor pravidiel podľa ktorých sa genetická informácia uložená v DNA prevádza na primárnu štruktúru bielkovín.

1.1.1 Štruktúra DNA

DNA je nukleová kyselina, nositeľkou genetickej informácie všetkých organizmov, takže je pre život dôležitou látkou, ktorá vo svojej štruktúre kóduje a bunkám zadáva ich program a tým predurčuje vývoj a vlastnosti celého organizmu. DNA je uložená vo forme sekvencie nukleotidov. Každý nukleotid obsahuje jednu zo štyroch možných dusíkatých báz (purinových alebo pirimidonových báz):

Purinové bázy:

- **A** - Adenin
- **G** – Guanin

Pirimidonové bázy:

- **T** - Thymin
- **C** - Cytosin

Keďže kyselina fosforečná a deoxyribóza je spoločná zložka pre všetky nukleotidy, jednotlivé nukleotidy sa od seba odlišujú len bázou. Práve bázy sú zodpovedné za kľúčovú schopnosť DNA zaznamenávať a prenášať genetickú informáciu. Deoxyribózová a kyselinová zložka slúžia na to, aby držali bázy vo vhodných polohách a vzdialenostiach. Tieto dve zložky tvoria takzvanú pentózafosfátovú kostru DNA.

Väzba sa vytvára medzi zvyškom kyseliny trihydrogenfosforečnej na 5. uhlíkovom atóme pentózy jedného nukleotidu a hydroxilovou skupinou viazanou na 3. uhlíkovom atóme pentózy susedného nukleotidu. Sú teda viazané 3',5'-fosfodiesterovou väzbou, zapisovanou aj ako 5'-3'-fosfodiesterová väzba. Na vlákne DNA s nespojenými koncami (lineárnej molekule) rozlišujeme dva konce: na 5' konci je vlákno ukončené fosfátom a na 3' konci je vlákno ukončené hydroxilovou -OH skupinou. Primárna štruktúra obsahuje iba štyri nukleotidy, pričom pomer adenínu s tymínom (A : T) a guanínu s cytozínom (G : C) je rovnaký 1 : 1. Niektoré z báz (u človeka odhadom asi 1 %) môžu byť metylované. U zvierat je toto percento i niekoľkonásobne vyššie (napríklad u makakov až 6%).

Poradie nukleotidov – sekvencia DNA – sa udáva vždy pre jedno z dvoch vlákien vypísaním jednotlivých nukleotidov od 5' konca k 3' koncu. Každý nukleotid sa označuje jednopísmenovou skratkou dusíkatej bázy, ktorú obsahuje. Veľkosť genómu jednotlivých organizmov alebo organel sa udáva počtom báзовých párov, ktoré obsahuje ich DNA [9].

1.1.2 Párovanie dusíkatých báz

Objav dvoj závitnicovej štruktúry, čo predstavuje sekundárnu štruktúru DNA, sa spája s menami Jamesa Watsona a Francisa Cricka. Pred vypracovaním Watsonovho-Crickovho modelu už bolo známe, že DNA je tvorená nukleotidmi, ktoré sa líšia svojimi dusíkatými bázami. Dôležitou vlastnosťou je komplementárnosť jednotlivých nukleotidov, párovanie báz je základný spôsob párovania, na ktorom je postavený celý prenos genetickej informácie. Páruje sa vždy jedna purínová a pyrimidínová báza. Z chemického hľadiska sú adenín a guanín deriváty purínu, čo je heterocyklická zlúčenina tvorená dvoma kruhmi, thymin a cytosín sú deriváty pyrimidínu, ktorý patrí medzi heterocyklické zlúčeniny s jedným kruhom. Takže v prípade DNA sa teda adenín páruje s tymínom a guanín s cytozínom. Guanín s cytozínom sa viažu tromi vodíkovými väzbami, a adenín s tymínom zase dvomi vodíkovými väzbami. Keďže väzba adenín-tymín má menší počet vodíkových mostíkov, preto väčšina adenín-tymínových párov sa nachádza v miestach, kde je potrebné, aby sa dvoj závitnica rozdelila na dve jednotlivé vlákna. Spárovaním komplementárnych báz sa vytvorí dvojreťazová štruktúra DNA. Informácia je uchovaná rovnako v oboch vláknach. Výhodou je uchovanie informácie pri poškodení jedného z vlákien, kedy je možno danú informáciu obnoviť pomocou druhého vlákna.

Watsonovo-Crickovo párovanie báz je základné, pri tvorbe trojreťazových a štvorreťazových DNA je nutne uvažovať aj iné možnosti párovania báz. Obrátené Watsonovo-Crickovo párovanie báz znamená, že DNA môže byť zostavená z paralelných DNA reťazcov. Vyznačuje sa rovnakou orientáciou fosfodiesterových väzieb. Obrátený spôsob tvorí páry medzi: C – C, A – A, G – G, T – T. Ďalšími možnosťami párovania báz sú Hoogsteenovo alebo obrátené Hoogsteenovo párovanie báz. Vznikajú takzvané triády, trojice spárovaných báz, tvorba trojreťazových DNA. Vznik štvorreťazových DNA sa hypoteticky

vysvetľuje vznikom tetrády, párovanie medzi štyrmi bázami, ktoré sa uskutočňuje medzi molekulami guaninu a cytosinu a medzi molekulami adeninu a thyminu.

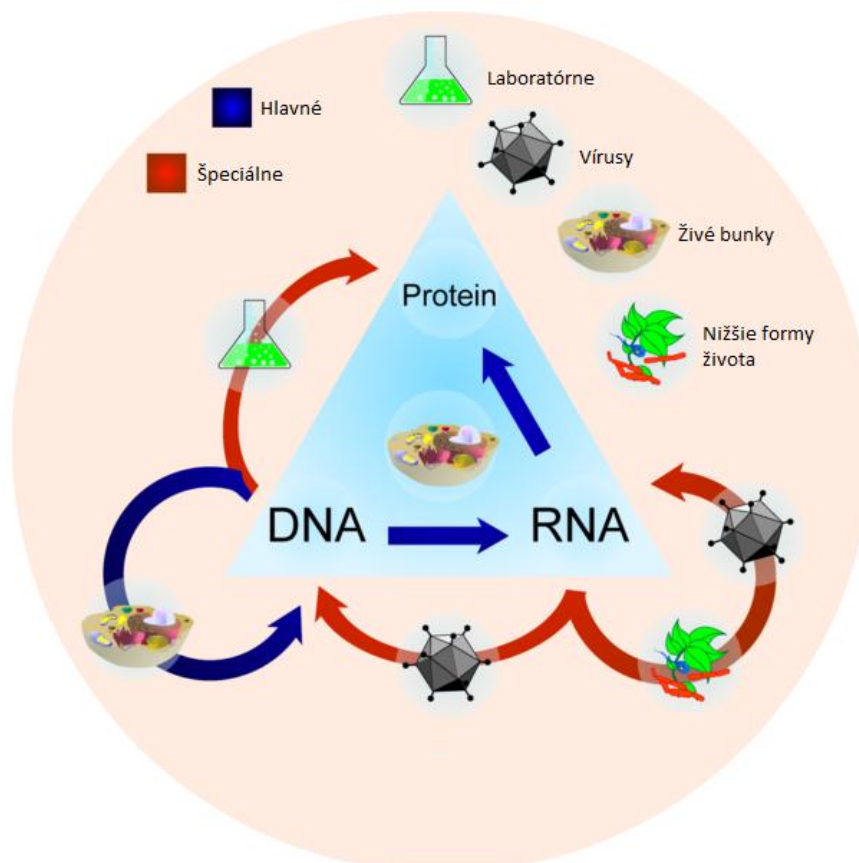
Reťazce sú vzhľadom na seba postavené v opačnom smere. Jeden reťazec je orientovaný v smere $5' \rightarrow 3'$, druhý bude orientovaný v smere $3' \rightarrow 5'$. Oddeliť od seba tieto vlákna možno zohriatím na vysokú teplotu, zmenou pH, zmenou iónovej sily roztoku a prítomnosťou niektorých organických látok, napríklad močoviny. Rozpadnutie dvoj vlákna na jednotlivé vlákna, sa nazýva denaturácia DNA. Pri teplotnej denaturácii je možné postupným ochladzovaním roztoku dosiahnuť opätovné spárovanie báz v oboch reťazcoch a obnovenie pôvodnej štruktúry [9].



Obrázok 1.1 Dvoj závitnicová štruktúra [22].

1.2 Centrálna dogma molekulárnej biológie

Popisuje cestu prenosu informácie medzi biopolymérmi, dovoľuje prepis medzi nukleovými kyselinami a preklad z RNA do proteínov [9]. To má za dôsledok nemožnosť toku informácií z bielkovín do nukleotidových kyselín a tak zanášanie zmien organizmu späť do genetickej informácie. Hlavnými zástupcami biopolymérov patrí DNA, RNA a proteíny. Medzi nimi existuje hypoteticky deväť ciest, obvyklá cesta je replikácia DNA, transkripcia do RNA a preklad z RNA do proteínov.

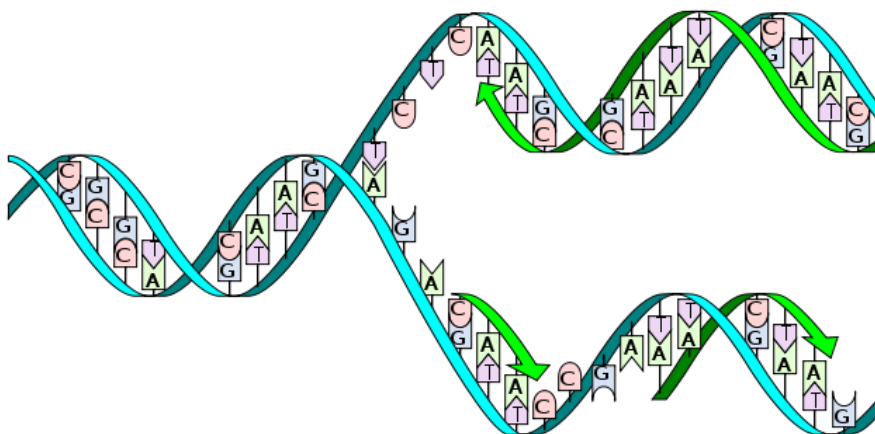


Obrázok 1.2 Pohyb biologickej informácie v organizmoch [22].

1.2.1 Replikácia DNA

Proces tvorby kópie molekuly DNA, čím sa genetická informácia prenáša z jednej molekuly na inú molekulu rovnakého typu. Celý proces je semikonzervatívny, každá nová molekula má jeden reťazec z pôvodnej molekuly a jeden nový, syntetizovaný. Replikácia nezačína na náhodnom mieste genómu, miesto je presne určené a označuje sa ako replikačný počiatočok. Každé vlákno je replikované odlišným spôsobom, pretože každé vlákno je orientované opačným smerom [9]. Replikáciu je možné rozdeliť do troch základných krokov:

- **Iniciácia** – určenie replikačného počiatočku, rozpletenie dvoj závitnice DNA, vznik replikačnej vidlice.
- **Elongácia** – pridávanie nukleotidov a postup replikačnej vidlice, pomáhajú k tomu svorkové proteíny.
- **Terminácia** – ukončenie replikácie, ak je zhotovená celá kópia DNA.



Obrázok 1.3 Replikácia DNA [8].

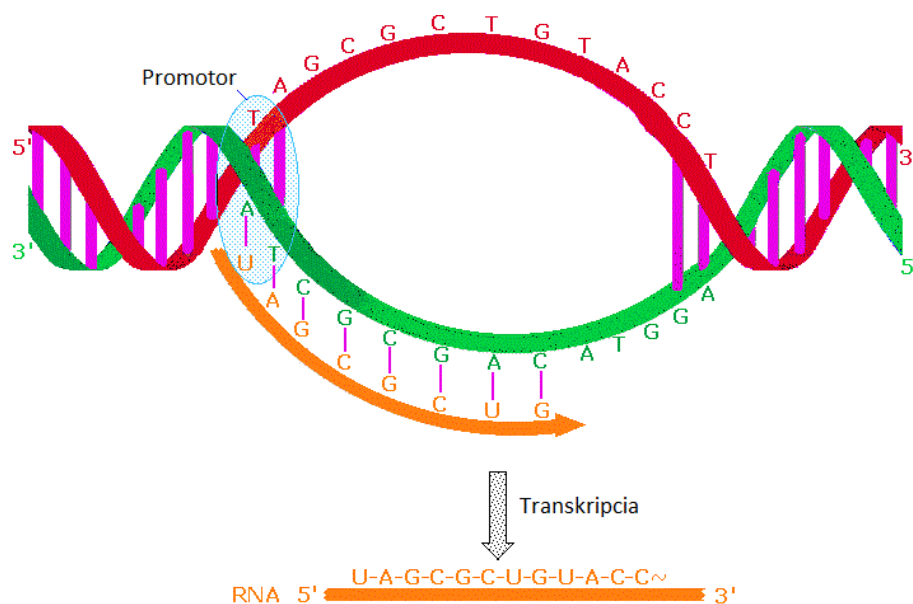
1.2.2 Transkripcia do RNA

Proces pri ktorom je podľa genetickej informácie zapísanej v reťazci DNA vyrábaný reťazec RNA. Prebieha u všetkých známych organizmov a je dôležitou súčasťou centrálnej dogmy molekulárnej biológie. Rýchlosť transkripcie v laboratóriu je asi 100 – 300 nukleotidov za minútu s chybovosťou približne jedna chyba za 10 000 bází.

Priebeh transkripcie je v niečom podobný s replikáciou, proces kedy na základe jedného reťazca DNA vytvorí vlákno iné. Proteíny v tele vznikajú na základe vzoru zapísaného v génoch DNA, tieto gény sú práve počas procesu transkripcie prepísané do RNA. Celý proces je kontrolovaný zložitou molekulárnou mašinériou, v centre stojí enzým RNA polymeráza. Táto polymeráza sa nadviaže na začiatok génu, ktorý má byť prepísaný. Nadviaže sa na oblasti DNA označované ako promótor a následná aktivácia RNA polymerázy v tomto bode sa prejaví komplikovaný regulačný systém ovládajúci transkripciu postupom troch bodov:

- **Iniciácia** – rozvinie sa dvoj závitnica DNA, začne sa vytvárať RNA a RNA polymeráza vystupuje z promotoru.
- **Elongácia** – predlžovanie reťazca.
- **Terminácia** – ukončenie transkripcie a uvoľnenie RNA molekuly, nasleduje niekoľko transkripčných úprav, ktoré nie sú súčasťou transkripcie.

K transkripcii dochádza v smere 5' → 3', podobne ako u replikácii. Sú nutné isté transkripčné faktory, tie sa viažu na začiatok alebo na koniec génu alebo priamo na samotnú RNA polymerázu a regulujú transkripciu. Molekula RNA je reprezentovaná nukleotidmi Uracil (U), Adenin (A), Cytosin (C), Guanin (G), Uracil nahrádza Thymin. Každé C je prepísané ako G, G prepísané ako C, T ako A, A je prepísané ako U [9].



Obrázok 1.4 Transkripcia do RNA [23].

1.2.3 Preklad z RNA do proteínov

Sekundárny proces syntézy bielkovín. Úlohou je podľa molekuly RNA ako vzoru vyrobiť odpovedajúcu bielkovinu. Informácia je zapísaná v RNA a podľa jasných pravidiel genetického kódu dekodovaná a podľa nej je zostavený reťazec aminokyselín. V RNA sa nachádzajú štyri rôzne bázy, keďže potrebujeme zakódovať dvadsať rôznych aminokyselín. Využívame k tomu triplet, kombináciu troch nukletidov, nazývaný ako kodón. Dostávame 64 možných kombinácií, 18 aminokyselín sú kódované viacerými kodónmi [8]. Každá aminokyselina patrí do jednej zo štyroch skupín: polárne, nepolárne, pozitívne nabité a negatívne nabité. Samotný preklad prebieha na orgánoch zvaných ribozomy a celý proces je rozdelený do troch fáz:

- **Iniciácia** – vznik iniciačného komplexu mRNA, skenovanie RNA a nájdenie štartovacieho kodónu AUG, ním začína každá vyrobená bielkovina.
- **Elongácia** – ribozom sa posúva o jeden kodón, pripojí sa príslušná aminokyselina, vytvorí sa peptidová väzba medzi ňou a predošlou aminokyselinou a znovu sa posúva.
- **Terminácia** – ak posúvanie ribozomu dostane kodón UAA, UAG a UGA proces končí. Jedná sa o terminačné kodóny, ktoré nesignalizujú žiadnu aminokyselinu.

Preklad je energeticky náročný. Odhaduje sa že baktéria *E.coli* spotrebuje 99% svojej celkovej spotreby energie práve na syntézu proteínov.

Tabuľka 1.1: Prehľad aminokyselín.

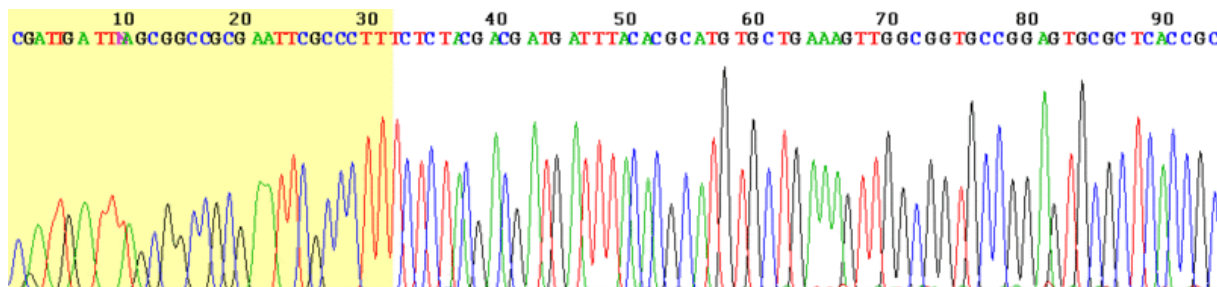
Označenie	Troj písmenkové označenie	Názov aminokyselín
A	Ala	Alanín
C	Cys	Cysteín
D	Asp	Kyselina asparágová
E	Glu	Kyselina glutámová
F	Phe	Fenylanín
G	Gly	Glicín
H	His	Histidín
I	Ile	Izoleucín
K	Lys	Lyzín
L	Leu	Leucín
M	Met	Metionín
N	Asn	Asparagín
P	Pro	Prolín
Q	Gln	Glutamín
R	Arg	Arginín
S	Ser	Serín
T	Thr	Treonín
V	Val	Valín
W	Trp	Tryptofán
Y	Tyr	Tyrozín

2 Biologické sekvencie

Pracujeme s dvomi základnými typmi biologických sekvencií: sekvencie nukleotidových kyselín alebo proteínov.

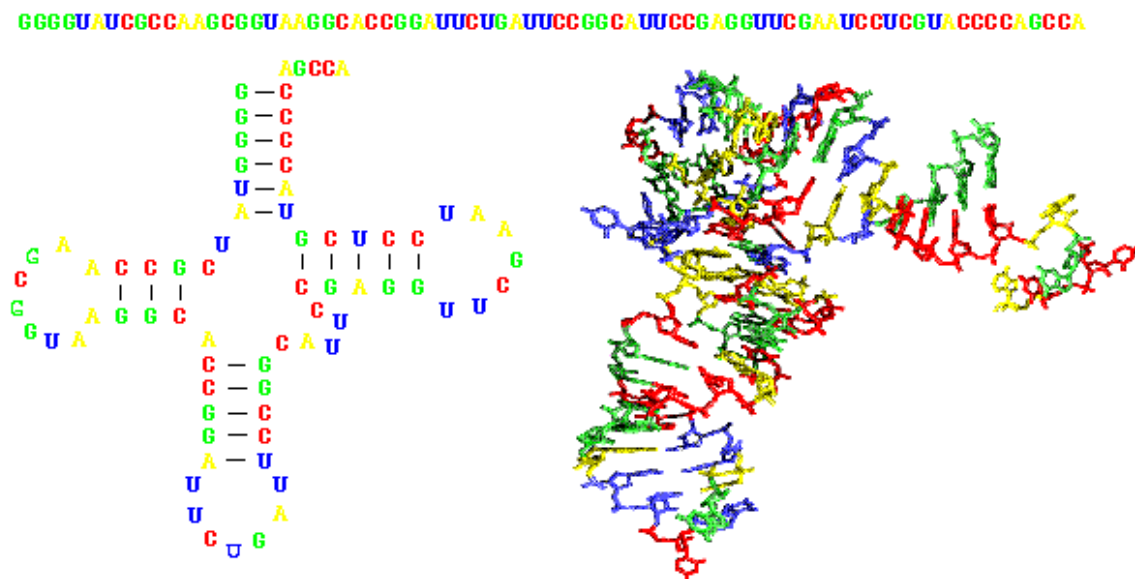
2.1 Sekvencie DNA, RNA

Medzi sekvencie nukleotidových kyselín patrí deoxyribonukleová kyselina (DNA), ktorá sa skladá zo štyroch nukleotidových báz (A, C, G, T) a kyselina ribonukleová (RNA), ktorá má taktiež štyri rozdielne nukleové bázy (U, A, C, G). V sekvencii zisťujeme ich poradie. Medzi prvé metódy sekvenovania patrí napríklad Maxam – Gilbertova metóda označovaná tiež ako chemické sekvenovanie, kde vzorka DNA je rozdelená na päť častí a každá je vystavená chemikálii, ktoré nám rozpoznajú nukleovú bázu. Samozrejme sa stále objavujú nové metódy, ktoré urýchľujú, zjednodušujú a spresňujú DNA sekvenovanie. Metóda SMRT (Single molecule real time) je metóda, ktorá v reálnom čase sleduje replikáciu DNA. Jednotlivé bázy sú fluorescenčne zafarbené, pri začlenení do reťazca sa farbivo uvoľní a vydá záblesk, ktorý je zachytený detektorom ako odpovedajúci nukleotid. Dnes sa sekvenovanie vyhodnocuje pomocou počítača, napríklad na základe Sangerovej metódy, ktorá využíva biologický proces replikácie DNA.



Obrázok 2.1 Sekvencie DNA Sangerovou metódou [24].

RNA sa obvykle vyskytuje v jednovláknovej podobe, ktorá však môže tvoriť dvojreťazové úseky v rámci jednej molekuly. V bunkách sa nachádzajú tri základné typy: ribozomálna (rRNA) podieľajúca sa na stavbe ribozomu a katalýze syntézy proteínov, transferová (tRNA) prenášajúca jednotlivé kyseliny k ribozomu a mediatorová (mRNA), slúži ako vzor pre syntézu proteínov.



Obrázok 2.2 Primárne sekvencie štruktúry jednoreťazovej molekuly tRNA [25].

2.2 Organizácia nukleotidových sekvencií DNA

Sekvencie DNA sú buď jedinečné alebo repetitívne [9]. Jedinečná DNA sekvencia je taká, ktorá sa v génomu vyskytuje iba jedenkrát. Repetitívne sekvencie označované ako aj repetície, sú typmi DNA sekvencií, ktoré sa mnohonásobne opakujú v genóme, napríklad opakovanie krátkej sekvencie „ATAAT“ v chromozóme 2 *D.melanogaster*. Opakovaná sekvencia je jednotkou repetície, jej dĺžka je vyjadrená počtom, ktorý ju tvorí. Potom počet jednotiek repetície udáva koľkokrát sa daná repetitívna sekvencia vyskytuje v danom genóme.

Repetície, v ktorých sa určitá jednotka mnohonásobne opakuje bezprostredne za sebou, opakuje v tandeme, nazýva tandemová repetícia. Dĺžka je väčšinou 5 až 10 bp (bázových párov) ale u stavovcov a rastlín okolo 20 až 200 bp.

V DNA reťazci sa taktiež môžu objavovať obrátené repetície, ktoré vedú k tvorbe vlásieniek a krížových štruktúr. Repetície vo svojej komplementárnej podobe. Príklad

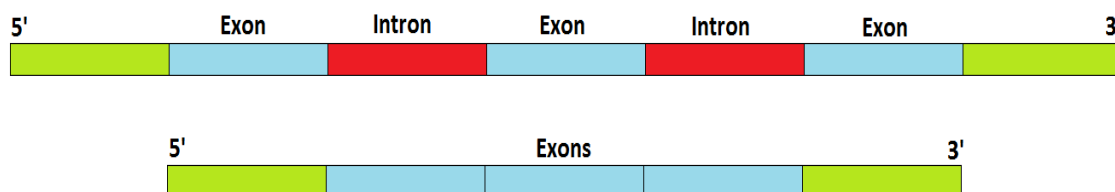
5'...ATGCTTGA...3'
3'...AGTTCGTA...5'

Priama repetícia, sekvencia opakovaná v rovnakom smere v DNA reťazci, môže sa vyskytovať bez prerušenia alebo môže byť prerušená inou repetíciou alebo sledom sekvencií. Napríklad:

5'...CGAC...AAT...CGAC...3'
3'...CAGCCAGC...5'

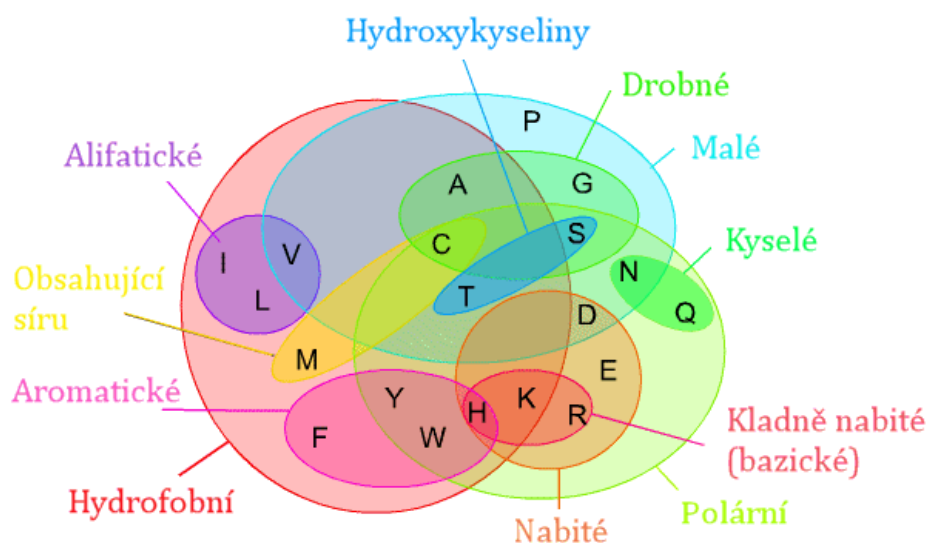
2.3 Sekvencie proteínov

DNA sekvencia pozostáva z dvoch typov oblastí. Jedna oblasť, ktorá nám kóduje proteíny a oblasť, ktorá naopak nekóduje proteíny tzv. „odpadová DNA“. Kódujúca oblasť je tak rozdelená do niekoľkých prerušených úsekov, ktoré sa nazývajú exony a nekodujúce oblasti medzi exonmi sa nazývajú introny. Pred prekladom do proteínov, introny sú vystrihnuté mechanizmom zvaným splicing a exony tak vytvoria jeden kódujúci neprerušný gén.



Obrázok 2.3 Mechanizmus splicing.

Proteínové sekvencie sú zložené z 20 rôznych aminokyselín, ktoré sú prekladané z kombinácii troch nukleotidov, sekvenčným motívom označujeme oblasť aminokyselinových sekvencií zdieľané nejakou skupinou proteínov, to znamená, že každá aminokyselina na základe jej chemických vlastností patrí do skupiny. Za skupinu sa považuje evolučná rodina homológnych proteínov, ktoré majú spoločný pôvod. Všetky vlastnosti proteínu sú dané jeho primárnou štruktúrou t.j. sekvenciami aminokyselín, kde jednopísmenkový kód nám udáva príslušnú aminokyselinu [9].



Obrázok 2.4 Diagram príslušnosti aminokyselín do skupín [8].

2.4 Formáty biologických sekvencií

Formáty biologických sekvencií [7] sa delia na niekoľko skupín, na jednoduchšie, ktoré väčšinou slúžia ako zdroj dát a zložitejšie, ktoré obsahujú informácie o danej sekvencii.

2.4.1 RAW

Pre ukladanie sekvencií, nukleotidov alebo aminokyselín je najjednoduchším formátom. Označujeme ho ako surové dáta, pretože neobsahuje informácie o sekvencii, ktoré neje možné ani prípadne doplniť. Formát sa preto využíva ako zdroj dát pre ostatné formáty.

2.4.2 FASTA

Formát FASTA sa používa pre sekvencie nukleotidov alebo aminokyselín. Prvým znakom je > (väčší než), za ním nasleduje názov sekvencie, po prvú medzeru, zvyšok sa využíva ako voliteľný komentár. Na ďalších riadkoch nám nasleduje sekvencia. Do jedného súboru je možné uložiť viacero sekvencií, ktoré sú rovnakého typu. Výhodou je jednoduchá editovateľnosť, jednoduchý syntax. V praxi najpoužívanejším formátom, možno ho napísať priamo z klávesnice.

```
>gi|230826|pdb|3CNA|  Concanavalin A
ADTIVAVELDTYPNTDIGDPSYPHIGIDIKSVRSKKTAKWNMQDGKVGTAHI IYNSVDKRLSAVVSYPNA
DATSVSYDVDLNDVLP EWVRVGLSASTGLYKETNTILSWSFTSKLKSNGTHQTDALHFMFNQFSKDQKDL
ILQGDATTGTDGNLELTRVSSNGSPEGSSVGRALFYAPVHIWESSAATVSFEATFAFLIKSPD SHPADGI
AFFISNIDSSIPSGSTGRLLGLFPDAN
```

Obrázok 2.5 Sekvencie aminokyselín vo formáte FASTA [7].

2.4.3 PIR/NBRF

Podobný formát ako je FASTA, ktorý navyše rieši problém určenia typu sekvencie. Prvý riadok začína taktiež znakom >, nasleduje dvojznakový kód určujúci typ sekvencie (P1 – proteín, DL – lineárna DNA), druhý riadok plný názov a anotácia a od tretieho riadku začína sekvencia ukončená hviezdikou.

2.4.4 GenBank/DDBJ/EMBL flatfile

Komplikované formáty, ktoré uchovávajú spolu so sekvenciami aj pridružené informácie, ktoré sa delia na kritéria a anotácie. Načítanie takýchto dát môže byť komplikované ak užívateľ neje vybavený potrebným softwarom, ktorý dokáže za pomoci vstupných filtrov načítať len čistú sekvenciu.

3 Odhad entropie, kompresné algoritmy

Život je silno spojený s organizáciou a štruktúrou. Živé organizmy možno považovať za zástupcov komunikácie s ich prostredím a ložiskom informácií potrebných k prispôsobeniu. Uložená informácia, jej obsah a forma, sa predáva z jednej generácie na inú prostredníctvom DNA molekuly. Biologický dôležitá molekula, ktorá je reprezentovaná ako biologická sekvencia sa stáva informáciou a správou. Teória a koncepcia kompresie sa stávajú prirodzeným nástrojom pre pochopenie štruktúry v týchto správach.

3.1 Odhad entropie

Základným pojmom teórie pravdepodobnosti je práve entropia. Stretávame sa s ňou všade kde hovoríme o pravdepodobnosti možných stavov systému alebo sústavy. Základy informačnej entropie položil C.E. Shannon [1] „otec teórie informácie“ vo svojej publikácii. Predpokladajme, že máme množinu možných udalostí, ktorých pravdepodobnosť je p_1, p_2, \dots, p_n . Následná entropia pre stochastické systémy je definovaná

$$H = -\sum p_i \log p_i \quad (3.1.)$$

ako suma všetkých pravdepodobností výskytu (p_i).

Definované H , má niekoľko vlastností vhodných pri výbere informácie [21]:

1. H je rovno nule v prípade, že existuje jediné i také, že $p_i > 0$, správa bude mať všetky znaky rovnaké, preto ju je možné zakódovať do iného znaku a entropia je idúca k nule.
2. Pre daný počet N , H je maximálna a rovná sa $\log n$ ak všetky p_i sa rovnajú $1/n$, to značí neistú situáciu a ide o náhodný proces.

Na základe tohto definoval Shannon entropiu pre sekvencie znakov

$$F_N = -\sum_{i,j} p(B_i, S_j) \log p_{B_i}(S_j) \quad (3.2.)$$

kde $p(B_i, S_j)$ je pravdepodobnosť sekvencie B_i nasledovanej symbolom S_j a $p_{B_i}(S_j) = p(B_i, S_j) / p(B_i)$ je podmienená pravdepodobnosť S_j po B_i . Ak nie sú k dispozícii žiadne štatistické vplyvy, ktorých rozsah presahuje N symbolov, máme podmienenú pravdepodobnosť ďalšieho symbolu, ktorého predchádzajúca ($N-1$) je známa.

Na základe týchto znalostí americká biofyzička Lila Gatlin [2], navrhla definíciu na informačný obsah DNA. Ak máme abecedu veľkosti N (4 – DNA, 20 – aminokyseliny), sú definované dve veličiny $D1$ a $D2$, ktoré merajú odlišnosti rovnako podobných entropií

$$D_1 = \log N - H_1(X) \quad (3.3.)$$

$$D_2 = H_1(X) - H(X|Y) \quad (3.4.)$$

kde $H_1(X)$ je prvá entropia sekvencie a $H(X|Y)$ je podmienená pravdepodobnosť, kde X a Y sú susedné znaky v sekvencii. Informačný obsah DNA sekvencie je definovaný ako súčet týchto dvoch odlišností, ktoré môžu byť definované ako rozdiel maximálnej entropie $\log N$ a podmienenej entropie $H(X|Y)$. Gatlin spája tieto informácie k definovaniu redundancie ako

$$R = 1 - \frac{H(X|Y)}{\log N} \quad (3.5.)$$

To nám vedie k

$$R \log N = D_1 + D_2 \quad (3.6.)$$

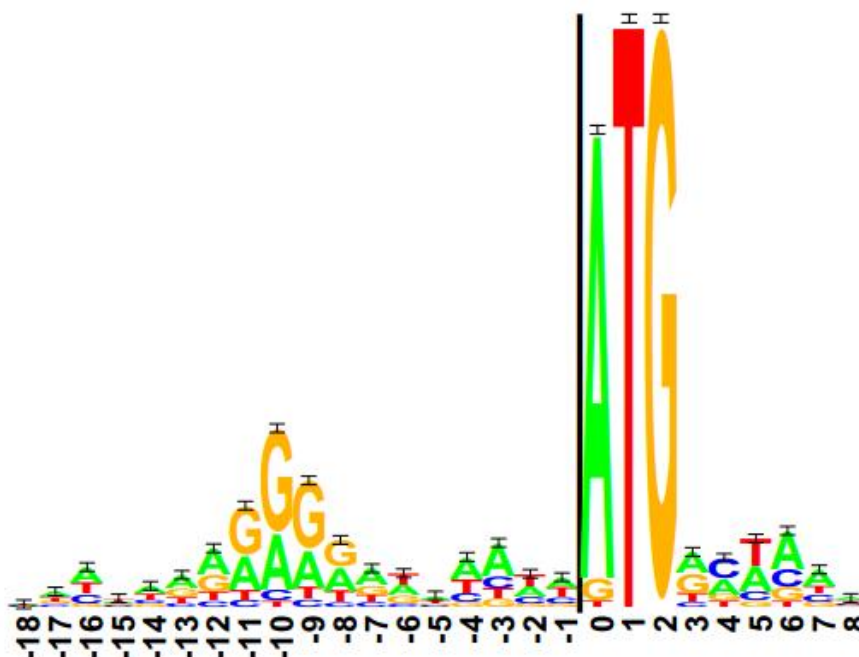
Gatlin poukázala na to, že DNA stavovcov, baktérií a vírusov možno rozlíšiť na základe ich informačného obsahu a že zvyšovaním redundancie v DNA, možno prejsť z nižšie zložitých organizmov ako je baktéria ku zložitejším, ako sú stavovce. Táto práca mala väčší vplyv vo fyziologickej oblasti, kde Gatlin predpokladala, že DNA je realizáciou ergodického procesu a odhad pravdepodobnosti sa realizuje pozdĺž sekvencie. Informačný obsah, ktorý má priamejší vplyv v bioinformatike, definovali Schneider a Stephens [3]. DNA je realizáciou náhodného procesu a informačný obsah sa počíta po bázach. Definícia sa mierne líši od Gatlinovej parametru DI , potom definovaná informácia je v lokalizácii l

$$R_{SEQUENCE}(l) = \log_2 N - (H(l) + e(n)) \quad (3.7.)$$

kde $H(l)$ je odhad entropie prvého radu daná ako

$$H(l) = -\sum f(x,l) \log_2 f(x,l) \quad (3.8.)$$

$e(n)$ je termín opravy pre malý počet sekvencií použitých k výpočtu entropie a $f(x,l)$ je frekvencia výskytu bázy x v mieste l . Na základe tejto hodnoty vytvorili logo sekvencie, kde každá poloha prvku zosúladenej sekvencie, ktoré sa objavujú na tomto mieste sú zastúpené v liste, ktorého výška je úmerná početnosti výskytu v danom mieste vynásobená informáciou na tomto mieste $R_{SEQUENCE}(l)$.



Obrázok 3.1: Logo pre *E.coli* [3].

Základným využitím kompresie biologických sekvencií je odhad entropie. Výpočet sa uskutočňuje skomprimovaním sekvencie a následným výpočtom, kde veľkosť skomprimovaného súboru v bitoch podelíme počtom znakov sekvencie, získavame horný odhad entropie sekvencie, ktorý nám slúži k ohodnotenie kvality komprimačného prístroja.

3.2 Kompresné algoritmy (Substitučná metóda)

Kompresia dát, zahŕňa poznatky o štruktúre informácie, jej mechanizmus a samozrejme k čomu je daná informácia určená. Najlepšie kompresné algoritmy, aj na základe odhadu entropie, sa snažia získať ako je informácia organizovaná adaptívnym spôsobom, určenie štruktúry, ktorá umožňuje kompresiu. Konceptne vyvinuté nástroje v oblasti zdrojového kódovania, ktoré riadia vývoj kompresných algoritmov sú tiež používaným nástrojom, pre analýzu štruktúry informácie, obzvlášť v biologických sekvenciách.

Základné kompresné algoritmy nie sú efektívne pri spracovaní biologických sekvencií, vedú skôr k expanzii ako ku kompresii. Preto bolo potrebné vytvoriť algoritmy špeciálne pre biologické sekvencie. Pravdepodobne jedným z najznámejších kompresných algoritmov DNA sekvencií je Gencompress, ktorý využíva skutočnosť, že DNA sekvencie obsahujú tandemové repetície, viac kópií génov a palindromických sekvencií. Jedná sa o upravený Lempel-Ziv, ktorý vyhľadáva a dopĺňa opakujúce sa časti sekvencie. Opakované sekvencie sa nazývajú pod reťazce (subsequences). Gencompress dosahuje lepšie výsledky ako predchádzajúce

známe programy Biocompress a Biocompress-2. Gencompress bol neskôr vylepšený na DNACompress, ktorý má lepšie vyhľadávanie modulov, no dnes je už aj vysoko sofistikovaný algoritmus DNAPack. Spomenuté algoritmy sú substitučnou kompresiou, ktorá pracuje tak, že opakujúce sa slová sú nahradzované odkazom na predchádzajúci výskyt alebo odkazom na slovník.

3.2.1 Lempel-Ziv

Prvý a najznámejší kompresný algoritmus. Patrí sem LZ77 [4] a LZ78 [5], bezstratové kompresné algoritmy, ktoré publikovali Abraham Lempel a Jacob Ziv v rokoch 1977 a 1978. Tvoria základ pre ostatné substitučné algoritmy.

LZ77

Algoritmus kompresie dosiahne tým, že opakované výskyty dát nahradí odkazom na pôvodný výskyt. Ukazovateľ je reprezentovaný dvojicou čísel určujúcich dĺžku opakovanej sekvencie a vzdialenosť tohto opakovania v pôvodnom výskyte. To znamená vráti sa o určitý počet znakov a skopíruje potrebnú dĺžku. Nazýva sa tiež kompresia s posuvným oknom, ktorým prechádza sekvenciou a v každom kroku nájde najdlhšiu možnú zhodu okna s predchádzajúcim výskytom. Vypíše ukazovateľ a doplní ho o nasledujúci znak, ak ide o nový znak ukazovateľ je nulový. Príklad demonštruje využitie kompresie LZ77 na konkrétnej sekvencii (Tabuľka 4.1).

Algoritmy LZ77 pracujú na rovnakom základnom princípe, líšia sa v tom ako zakódujú svoje komprimovaná dáta. Môžu meniť a dopĺňať svoje číselne rozsahy ukazovateľov, meniť počet bitov spotrebovaných na dĺžku ukazovateľa, rozlišovať ukazovatele.

Tabuľka 3.1: Príklad kompresie vstupnej sekvencie AABABBBABB pomocou LZ77.

Pozícia	1.	2.	3.	4.	5.	6.	7.	8.	9.	10.
Znak	A	A	B	A	B	B	B	A	B	B
Krok algoritmu	Aktuálny znak		Zhoda okna so sekvenciou		Nasledujúci znak		Výstup			
1.	1		-		A		(0,0)A			
2.	2		A		B		(1,1)B			
3.	4		AB		B		(2,2)B			
4.	7		BAB		B		(4,3)B			

LZ78

Algoritmus kompresie nahradí opakovaný výskyt dát odkazom na slovník, ktorý je zostrojený na základe vstupného dátového toku. Odkaz po nájdení zhody je v tvare index a znak, kde

index je indexom predchádzajúceho miesta v slovníku a k nemu nasledujúci znak. V Tabuľke 4.2 je zobrazený priebeh funkcie LZ78 na konkrétnej sekvencii, ktorá bola použitá pri LZ77.

Tabuľka 3.2: Príklad kompresie vstupnej sekvencie AABABBBABB pomocou LZ78.

Pozícia	1.	2.	3.	4.	5.	6.	7.	8.	9.	10.
Znak	A	A	B	A	B	B	B	A	B	B

Krok algoritmu (index)	Obsah bunky	Výstup
1.	A	(0,A)
2.	AB	(1,B)
3.	ABB	(2,B)
4.	B	(0,B)
5.	ABB	(3,-)

LZW (Lempel-Ziv-Welch)

Rýchly, jednoduchý, bezstratový kompresný algoritmus, ktorého dáta ďalej nie sú kompresovateľné. Použitý algoritmus si najprv vytvára tabuľku zo znakov použitých v sekvencii. Potom algoritmus sériovo prehľadáva sekvenciu a ukladá všetky unikátne dvojznakové slova. Algoritmus pokračuje v kódovaní, ak na vstupe nájde známe slovo v tabuľke, na výstup pošle kódový znak plus pred ním prvý znak kódovaného slova.

3.2.2 Biocompress

Všetky kompresné algoritmy, ktoré sú špeciálne určené ku kompresii DNA sekvencií vychádzajú z toho, že každú bázu (A, C, G, T) je možno zakódovať dvoma bitmi (00, 01, 10, 11). Biocompress [11] bol prvý algoritmus, ktorý bol špeciálne určený na kompresiu DNA sekvencií navrhol ho v roku 1993 Grunbach a spol., neskôr bola vytvorená aj jeho druhá verzia Biocompress-2 [12]. Oba algoritmy sú založené na plávajúcom okne nazývanom tiež ako náhľadové okno, známe z kompresného algoritmu LZ77. Subsekvencia v náhľadovom okne je kódovaná podľa rovnakej postupnosti znakov, ktoré sa vyskytli v minulom priebehu sekvencie. Do výstupu sa zapíše iba miesto kde sa daná subsekvencia vyskytla a jej dĺžka, to znamená, že vyhľadáva opakujúce sa časti sekvencie, repetície. Biocompress-2 bol doplnený o aritmetické kódovanie druhého rádu, ak nebola nájdená opakujúca sa časť, repetícia sekvencie. Pre oba kompresné algoritmy, Biocompress a Biocompress-2, platí, že ak sekvencia obsahuje podobné repetície väčšej dĺžky tak je možno dosiahnuť dobrý kompresný pomer.

3.2.3 Cfact

Kompresný algoritmus založený taktiež na vyhľadávaní opakujúcich sa časti sekvencie, je to dvojprechodový algoritmus [13]. Prvým prechodom sekvencie, je celá sekvencia analyzovaná pomocou indexového stromu a je vytvorený list kde sú zoradené subsekvencie podľa dĺžky. Pri druhom prechode sekvenciou kóduje subsekvencie na základe predchádzajúceho výskytu. Oblasti sekvencie, ktoré neboli zakódované sa zakódujú pomocou dvoch bitov (00, 01, 10, 11) na bázu (A, C, G, T).

3.2.4 GenCompress

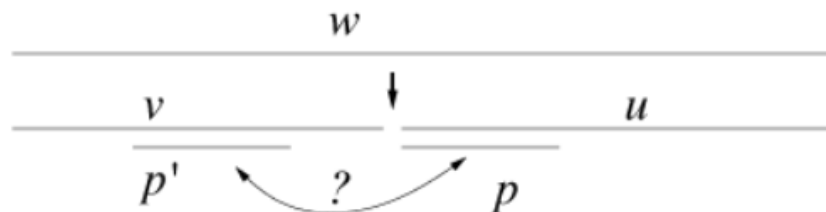
Algoritmus využíva základnú myšlienku LZ77, ale rozdiel je v tom, že vyhľadáva neexaktné opakovania, ktoré následne upravuje editačnými operáciami. GenCompress [6] prechádza sekvenciou a hľadá čo najdlhšiu zhodu s aktuálnym prefixom. Ukazovateľ je podobný ako u LZ77, ale je doplnený o editačnú operáciu oproti vzorovému prefixu. Editračné operácie sú tri:

1. Replace – záměna znaku na pozícii i , $(R, i, znak)$.
2. Insert – vloženie znaku na pozíciu i , $(I, i, znak)$.
3. Delete – zmazanie znaku na pozícii i , $(D, i, znak)$.

Príklad 4.1: Máme sekvenciu q , ktorú chceme editovať na sekvenciu p . $\lambda(gaccgtcatt, gaccttcatt)$ pomocou editačných operácií:

- $\lambda(gaccgtcatt, gaccttcatt) = (R, 4, t)$, pomocou záměny znaku.
- $\lambda(gaccgtcatt, gaccttcatt) = (D, 4), (I, 4, g)$, pomocou zmazania a následného vloženia znaku.

Algoritmus je založený na približnom vyhľadávaní. Pre vstupný reťazec w , predpokladá sa, že časť už bola skomprimovaná v a zostávajúca časť je u , $w = vu$. Algoritmus nájde optimálny prefix p z u a vyhľadá pod reťazec, ktorý môže byť zakódovaný ekonomickejšie. Prefix sa vymaže z u a pripojí k v .



Obrázok 3.2: Ukážka predošlých súvislostí medzi vstupným reťazcom, skomprimovanej časti a optimálnym prefixom [6].

3.2.5 DNACompress

DNACompress [14] využíva Ziv-Lempel kompresnú schému ako Biocompress-2 a GenCompress. Pozostáva z dvoch fáz. Počas prvej fázy sú vyhľadane opakujúce sa repetície, v druhej fáze dochádza k zakódovaniu sa týchto opakujúcich sa repetícií (subsekvencií) na základe predošlého výskytu a neopakujúce sa časti sekvencie sú kódované aritmetickým kódovaním druhého rádu. K identifikovaniu všetkých podobných subsekvencií je využívaný software PatternHunter [15], ktorý predstavuje rýchly a senzitívny vyhľadávač. To znamená, že DNACompress je rýchlejší ako GenCompress.

3.2.6 CTW+LZ

Kompresná technika založená na metóde CTW a LZ. Jednoducho dlhé opakujúce sa subsekvencie a repetície sú kódované LZ metódou a krátke subsekvencie sú komprimované pomocou CTW. Kompresný algoritmus CTW+LZ [16] dosahuje dobrý kompresný pomer ale čas komprimácie je dlhý najmä pri dlhých sekvenciách.

3.2.7 DNAC

Kompresná metóda DNAC [17] pozostáva zo štyroch fáz. Prvá fáza obsahuje vytvorenie indexového stromu pre opakujúce sa výsledky, v druhej sú všetky opakujúce sa repetície aproximované pomocou dynamického programovania. Tretia fáza sa všetky repetície, ktoré majú vysoké skóre s predošlými subsekvenciami vytiahnu z neprekývajúcich častí. Štvrtá fáza obsahuje kódovanie všetkých repetícií.

3.2.8 GeNML

DNA sekvencia je rozdelená na pevné bloky a tieto bloky GeNML [18] podľa predchádzajúcich subsekvencií s minimom editačných a substitučných operácií. V porovnaní s kompresným výkonom a rýchlosťou kompresie dosahuje lepšie výsledky ako Biocompress-2, GenCompress, CTW+LZ, DNACompress. K identifikovaniu repetícií využíva dynamické programovanie

3.2.9 DNAPack

DNAPack [19] využíva substitučné metódy pre opakujúce sa repetície a na kódovanie neopakujúcich častí využíva skoršie metódy ako CTW a aritmetické kódovanie druhého rádu. DNAPack ma väčší kompresný zisk ako predchádzajúce metódy.

3.3 Experimentálne výsledky

Kompresné algoritmy orientované na DNA sekvencie, ktoré sú popísané vyššie, sa testujú na štandardných DNA sekvenciách dostupných na stránkach verejnej databáze NCBI. Tabuľka 3.3 popisuje všetky sekvencie ich dĺžkou, zdrojom a veľkosťou v kB. Veľkosť v kilo bytoch je po použití 2 bitov na každú bázu. To je vlastne východisková veľkosť bez komprimácie.

Tabuľka 3.3: Informácie o štandardných DNA sekvenciách.

Názov Sekvencia	Dĺžka	Zdroj		Veľkosť (kB)
HUMGHCSA	66 495	rastový hormón	ľudské sekvencie	16,23
HUMDYSTROP	38 770	homo sapiens dystrophin		9,47
HUMHDABCD	58 864	sekvencia 3 kozmidov		14,37
HUMHPRTB	56 737	hypoxantín phosphoribosyltransferase		13,85
MPOMTCG	186 608	genóm mitochondrie		45,56
VACCG	191 737	vírus Copenhagen		46,81

Výsledky kompresie týchto sekvencií rôznymi kompresnými metódami sú zaznamenané v Tabuľke 3.4. Kompresný pomer je uvádzaný ako počet bitov na bázu (bpb – bits per base). Bez kompresie je tento kompresný pomer rovný 2 bpb. Kompresný pomer je definovaný ako

$$CompressionRatio = \frac{|O|}{|I|} \quad (3.9.)$$

kde $|O|$ je počet bitov na výstupe po kompresii a $|I|$ je dĺžka sekvencie na vstupe. Dobrý kompresný pomer dosiahnu všetky DNA orientované kompresné algoritmy približne od 1,69 do 1,78 bpb.

Tabuľka 3.4: Kompresný pomer dosiahnutý rôznymi kompresnými technikami [19].

Názov Sekvencie	Kompresná metóda						
	BioComp	GenComp	DNACopm	CTW+LZ	DNAC	GeNML	DNAPack
HUMGHCSA	1,3074	1,0969	1,0272	1,0972	1,0272	1,0089	1,0390
HUMDYSTROP	1,9262	1,9231	1,9116	1,9175	1,9116	1,9085	1,9088
HUMHDABCD	1,8770	1,8192	1,7951	1,8218	1,7951	1,7059	1,7394
HUMHPRTB	1,9066	1,8446	1,8165	1,8433	1,8165	1,7639	1,7886
MPOMTCG	1,9378	1,9058	1,8920	1,9000	1,8920	1,8822	1,8932
VACCG	1,7614	1,7614	1,7616	1,7616	1,7580	1,7644	1,7583
Priemer	1,7861	1,7252	1,7007	1,7236	1,7001	1,6723	1,6879

Z výsledkov kompresných pomerov plynie, že vývoj kompresných algoritmov orientovaných na DNA sa za desať rokov posunul iba o necelých 0,1 bpb, čo značí zložitosť komprimácie DNA sekvencií.

Ďalším faktorom hodnotenia kompresného algoritmu je kompresný zisk (Tabuľka 3.5), ktorý je definovaný ako

$$\left(1 - \frac{|O|}{2|I|}\right) \times 100\% \quad (3.10.)$$

kde $|O|$ je počet bitov na výstupe po kompresii sekvencie a $|I|$ je dĺžka alebo počet báz DNA sekvencie na vstupe. Kompresný zisk sa pohybuje medzi 11% až 16%, to znamená, že o toľko percent redukujú DNA sekvencie.

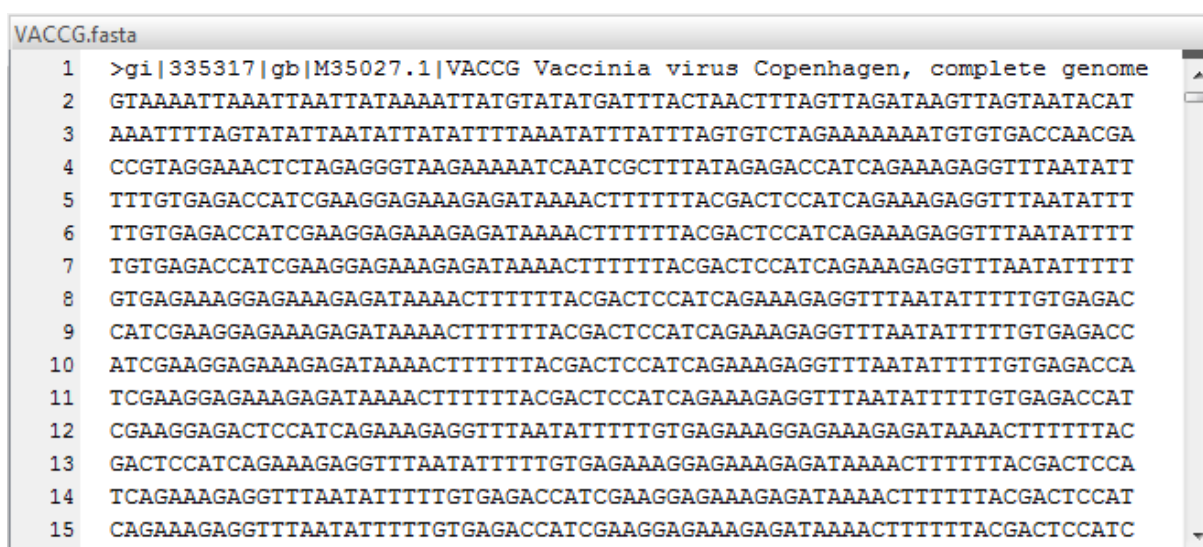
Tabuľka 3.5: Kompresný zisk rôznych DNA kompresných metód [19].

Názov Sekvencie	Kompresná metóda						
	BioComp	GenComp	DNACopm	CTW+LZ	DNAC	GeNML	DNAPack
HUMGHCSA	34,63%	45,16%	48,64%	45,14%	48,64%	49,56%	48,05%
HUMDYSTROP	3,69%	3,85%	4,42%	4,13%	4,42%	4,58%	4,56%
HUMHDABCD	6,15%	9,04%	10,25%	8,91%	10,25%	14,71%	13,03%
HUMHPRTB	4,67%	7,67%	9,18%	7,84%	9,18%	11,81%	10,57%
MPOMTCG	3,11%	4,71%	5,40%	5,00%	5,40%	5,89%	5,34%
VACCG	11,93%	11,93%	12,10%	11,92%	12,10%	11,78%	12,09%
Priemer	10,70%	13,73%	15,00%	13,82%	15,00%	16,39%	15,61%

Týmto je definovaný približný cieľ diplomovej práce, rozobraté kompresné algoritmy slúžia ako inšpirácia vo vytvorení kompresného algoritmu, ktorý sa zaoberá najmä vyhľadávaním opakujúcich sa častí. Funkčnosť a kvalita vytvoreného kompresného algoritmu je porovnaná s výsledkami kompresného pomeru a zisku algoritmov, ktoré sú orientované na DNA sekvencie.

4 Analýza entropie a kompresie DNA

Vyššie popísané teoretické znalosti, odhad entropie a kompresné algoritmy určené na DNA sekvencie sú spracované pomocou programového rozhrania MATLAB, sú realizované na DNA sekvenciách, ktoré boli použité pri testovaní kompresných algoritmov (Tabuľka 3.3) a na prvých desiatich chromozómoch človeka, dĺžky 10^5 báзовých párov. Biologické sekvencie sú vo formáte FASTA (Obrázok 4.1), ktorých obsah je vyhovujúci, prvý riadok obsahuje názov a kód sekvencie, zvyšok je samotná DNA sekvencia zložená zo znakov A,C,G a T. Miestami sa môže vyskytovať znak N, ktorý nedefinuje presne danú bázu, tento symbol je ignorovaný. Sekvencie sú k dispozícii na stránkach verejnej databázy NCBI (National Center for Biotechnology Information) [20], kde je možnosť si stiahnuť sekvencie aj v iných formátoch, ktoré obsahujú doplňujúce informácie.



```
VACCG.fasta
1 >gi|335317|gb|M35027.1|VACCG Vaccinia virus Copenhagen, complete genome
2 GTAA AATTAAATTATAAAATTATGTATATGATTACTTAAGTTAGTAGATAAGTTAGTAATACAT
3 AAATTTTAGTATATTAATATTATATTTTAAATATTTATTTAGTGTCTAGAAAAAATGTGTGACCAACGA
4 CCGTAGGAACTCTAGAGGGTAAGAAAAATCAATCGCTTTATAGAGACCATCAGAAAGAGGTTTAATATT
5 TTTGTGAGACCATCGAAGGAGAAAGAGATAAACTTTTTACGACTCCATCAGAAAGAGGTTTAATATTT
6 TTGTGAGACCATCGAAGGAGAAAGAGATAAACTTTTTACGACTCCATCAGAAAGAGGTTTAATATTTT
7 TGTGAGACCATCGAAGGAGAAAGAGATAAACTTTTTACGACTCCATCAGAAAGAGGTTTAATATTTT
8 GTGAGAAAGGAGAAAGAGATAAACTTTTTACGACTCCATCAGAAAGAGGTTTAATATTTTGTGAGAC
9 CATCGAAGGAGAAAGAGATAAACTTTTTACGACTCCATCAGAAAGAGGTTTAATATTTTGTGAGACC
10 ATCGAAGGAGAAAGAGATAAACTTTTTACGACTCCATCAGAAAGAGGTTTAATATTTTGTGAGACCA
11 TCGAAGGAGAAAGAGATAAACTTTTTACGACTCCATCAGAAAGAGGTTTAATATTTTGTGAGACCAT
12 CGAAGGAGACTCCATCAGAAAGAGGTTTAATATTTTGTGAGAAAGGAGAAAGAGATAAACTTTTTTAC
13 GACTCCATCAGAAAGAGGTTTAATATTTTGTGAGAAAGGAGAAAGAGATAAACTTTTTTACGACTCCA
14 TCAGAAAGAGGTTTAATATTTTGTGAGACCATCGAAGGAGAAAGAGATAAACTTTTTTACGACTCCAT
15 CAGAAAGAGGTTTAATATTTTGTGAGACCATCGAAGGAGAAAGAGATAAACTTTTTTACGACTCCATC
```

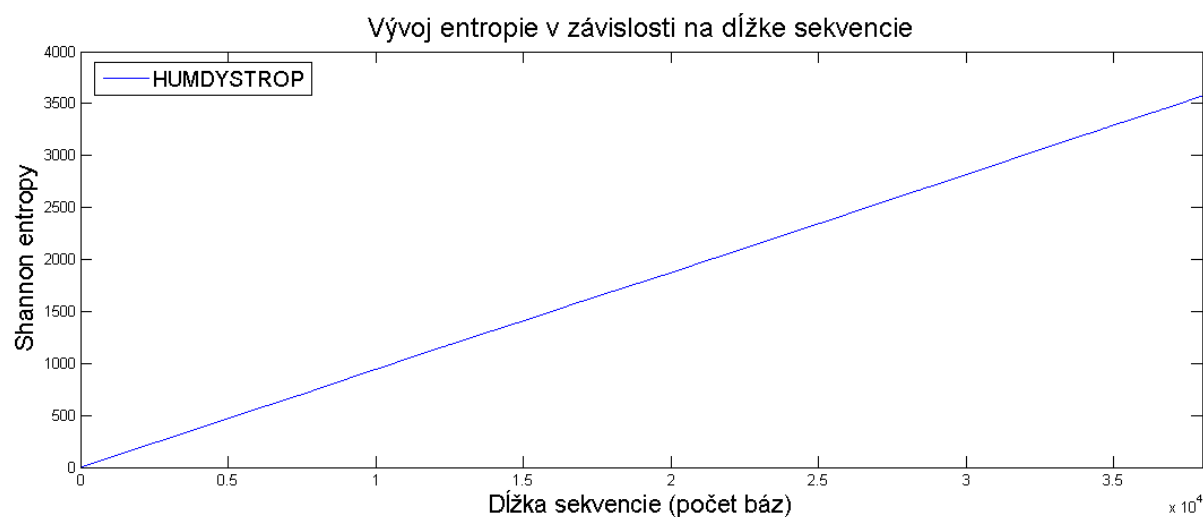
Obrázok 4.1: Úsek DNA sekvencie vo formáte FASTA.

4.1 Odhad entropie

Výpočet entropie sekvencie znakov je realizovaný pomocou definície Shannona (3.1.), ktorá je pomerne jednoducho aplikovateľná na DNA sekvenciách (Tabuľka 3.3). Entropia je zobrazená v závislosti na dĺžke sekvencie (Obrázok 4.2, 4.3, 4.4), lineárne stúpa s počtom báзовých párov.



Obrázok 4.2: Shannonova entropia *HUMGHCSA*, *HUMHDABCD*, *HUMHPRTB*.



Obrázok 4.3: Shannonova entropia *HUMDYSTROP*.

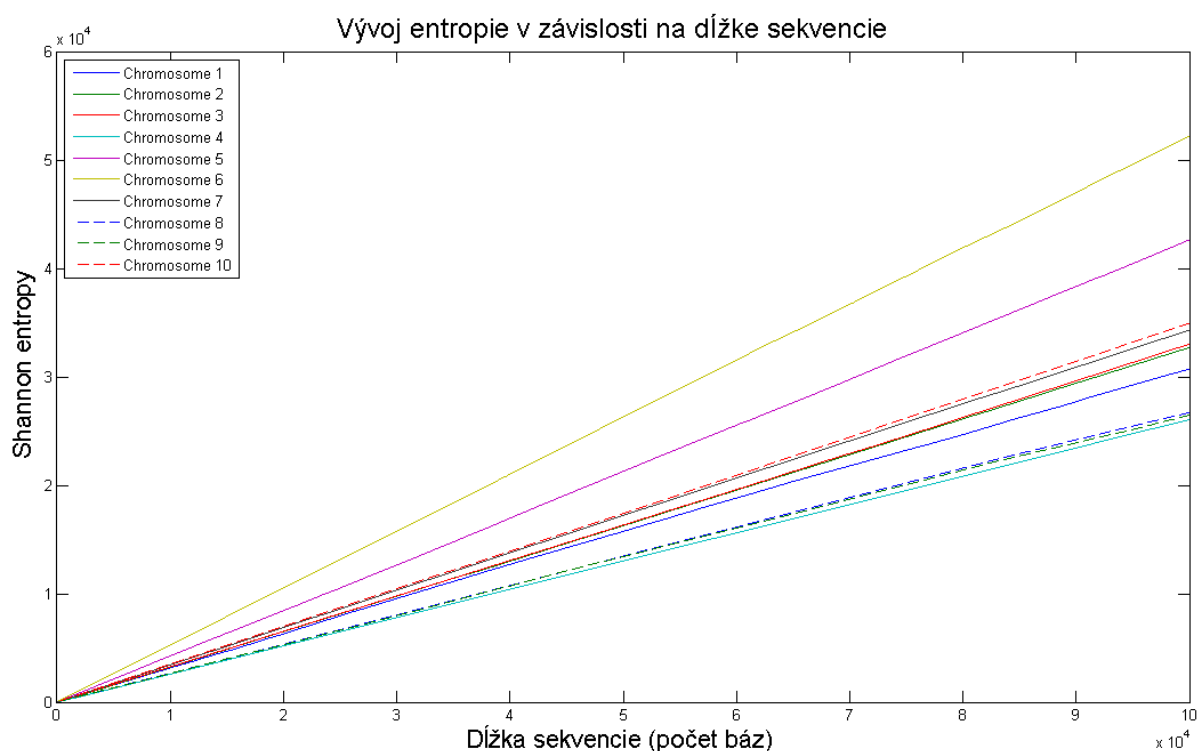


Obrázok 4.4: Shannonova entropia *MPOMTCG*, *VACCG*.

Tabuľka 4.1: Prehľad výsledkov Shannonovej entropie.

Sekvencia	Dĺžka	Shannon entropy
HUMGHCSA	66 495	5,505
HUMDYSTROP	38 770	3,569
HUMHDABCD	58 864	5,081
HUMHPRTB	56 737	5,087
MPOMTCG	186 608	16,280
VACCG	191 737	18,040
Priemer	99 869	8,927

Odhadnutá entropia prvých desiatich chromozómov človeka (Obrázok 4.5), kde vývoj entropie jednotlivých chromozómov je podobný, predovšetkým dvojice 2 a 3, 8 a 9, 7 a 10.



Obrázok 4.5: Závislosť entropie na počtu znakov ($N = 100\,000$).

Konečné výsledky Shannonovej entropie pri rovnakej dĺžke sekvencie sú rozdielne, to znamená, že jednotlivé chromozómy sú zložené z postupnosti iných znakov a tak aj podmienené pravdepodobnosti po sebe nasledujúcich znakov sú rozdielne, to je ovplyvnené aj celkovým počtom jednotlivých báz (A,C,G,T) sekvencie.

Tabuľka 4.2: Shannonova entropia chromozómov, $N = 100\,000$.

Chromozóm	Shannon entropy
1.	3,131
2.	3,517
3.	3,523
4.	2,612
5.	4,228
6.	5,201
7.	3,727
8.	2,637
9.	2,625
10.	3,879
Priemer	3,508

Keďže DNA sekvenciu môžeme považovať za sekvenciu náhodnú, to znamená, že s každým znakom prichádza nová informácia a nemôžeme nič predpovedať o tom čo bude nasledovať, v takomto prípade entropia dosahuje maximálnu hodnotu [1]. Entropia je rovná

$$H = \log N \quad (4.1.)$$

kde N je rovno dĺžke sekvencie, počtu znakov.



Obrázok 4.6: Maximálna entropia vírusu Copenhagen.

4.2 Kompresia DNA

Najjednoduchším spôsobom ako zakódovať respektíve skomprimovať DNA sekvencie je použitím kombinácii dvoch bitov (00, 01, 10, 11) na každú bázu (A, C, G, T). To znamená, že už len táto jednoduchá úprava zníži celkovú veľkosť sekvencie na 25% pôvodnej veľkosti. Táto skutočnosť nám definuje taktiež pojem kompresný pomer, pomocou ktorého sa posudzuje kvalita kompresného programu. Kompresný pomer, ktorý je totožný s hornou hranicou entropie a je definovaný celkovým počtom bitov vo výstupe na celkovú dĺžku sekvencie na vstupe. Za východiskový stav sa teda považujú 2 bity/bázu a cieľom komprimácie je znížiť tento pomer.

4.2.1 GenbitCompress

Jednoduchá bezstratová kompresná metóda, ktorá vstupnú sekvenciu rozdelí na segmenty pozostávajúce zo štyroch báz. Možných kombinácií 4 báz je 256 (4^4), každý DNA segment tak môže byť nahradený 8 bitmi a to tak, že každej báze sa priradia 2 bity (A = "00", C = "01", G = "10", T = "11"). V prípade, že po sebe idúce segmenty sú rovnaké, doplní sa 8 bitové slovo deviatym špecifickým bitom „1“, ak sa segmenty nezhodujú špecifický bit bude „0“.

Výsledok použitého algoritmu môže byť rozdielny a to, v najhoršom prípade kedy nebudú opakujúce sa segmenty pri dĺžke segmentu 4, v tomto prípade bude kompresný pomer veľmi vysoký (2,238). V lepšom prípade kde bude maximum opakujúcich sa segmentov a efektívnosť algoritmu plne preukázaná, kompresný pomer sa bude pohybovať okolo 1,125. Tento prípad bohužiaľ nie je možné dosiahnuť pre sekvencie z verejných databáz [10].

Celkový počet bitov na výstupe vypočítame ako

$$\mathfrak{R} = 9/4(N - \tau) + 2\tau - 9(\gamma) \quad (4.1.)$$

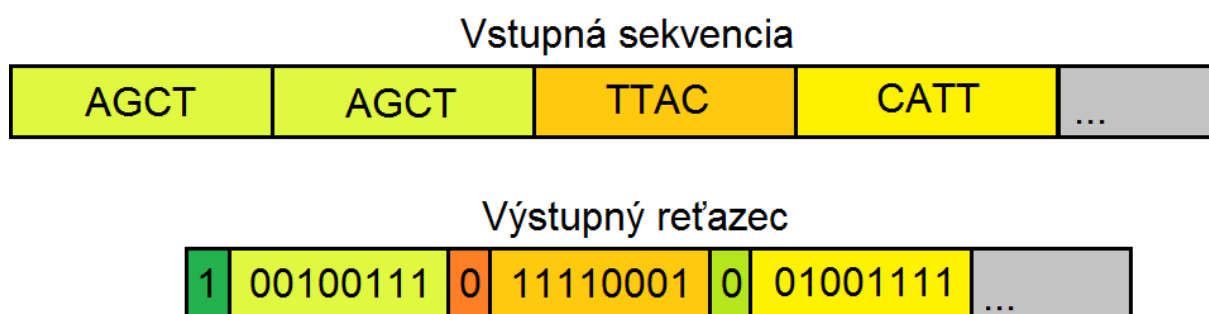
kde N je celková dĺžka sekvencie, τ je zvyšok po vydelení vstupnej sekvencie 4 a γ je počet opakujúcich sa segmentov nasledujúcich po sebe zložených zo štyroch báz.

Výpočet kompresného pomeru

$$CompressionRatio = \frac{\mathfrak{R}}{N}. \quad (4.2.)$$

Kompresia:

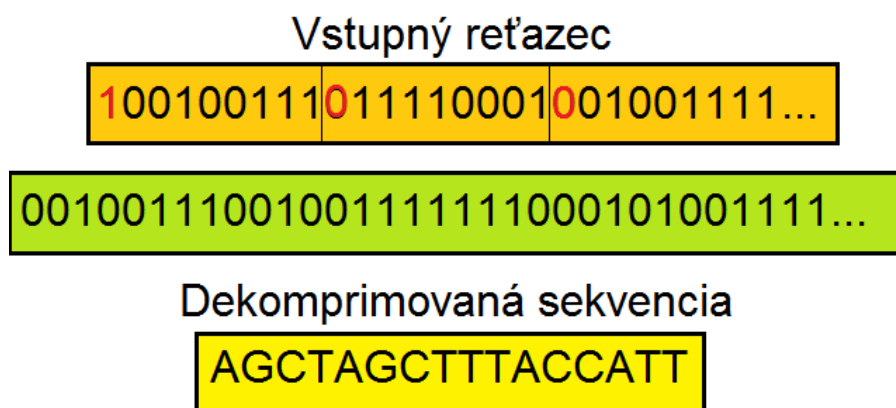
1. Vloženie DNA sekvencie pozostávajúcej zo znakov ACGT.
2. Rozdelenie sekvencie na segmenty pozostávajúce zo štyroch báz.
3. Zakódovanie každého segmentu na 8 bitové slovo.
4. Ak nasledujúce segmenty po sebe sa rovnajú, priradiť 1 ako 9. špecifický bit.
5. Ak po sebe nasledujúce segmenty nezhodujú, priradiť 0 ako 9. špecifický bit.
6. Opakuj 4. a 5. krok pokiaľ dĺžka sekvencie nie je $N - \tau$. (Kde N je dĺžka sekvencie a $\tau = N \bmod 4$).
7. Na výstupe reťazec 9 bitových čísel.



Obrázok 4.7: Príklad kompresie.

Dekompresia:

1. Načítanie vstupného reťazca.
2. Rozdelenie vstupu na 9 bitové segmenty.
3. Ak sa 9. špecifický bit rovná 1, nasledujúcu kombináciu ôsmich bitov vlož dvakrát, ak sa rovná 0 iba raz.
4. Opakuj 4. krok po koniec vstupného reťazca.
5. Dekóduj každú dvojicu bitov na príslušnú bázu.



Obrázok 4.8: Príklad dekompresie.

Výsledky kompresnej metódy na DNA sekvenciách (Tabuľka 3.3), ktoré boli použité aj pri experimentálnych výsledkoch predošlých kompresných metód (Tabuľke 4.3).

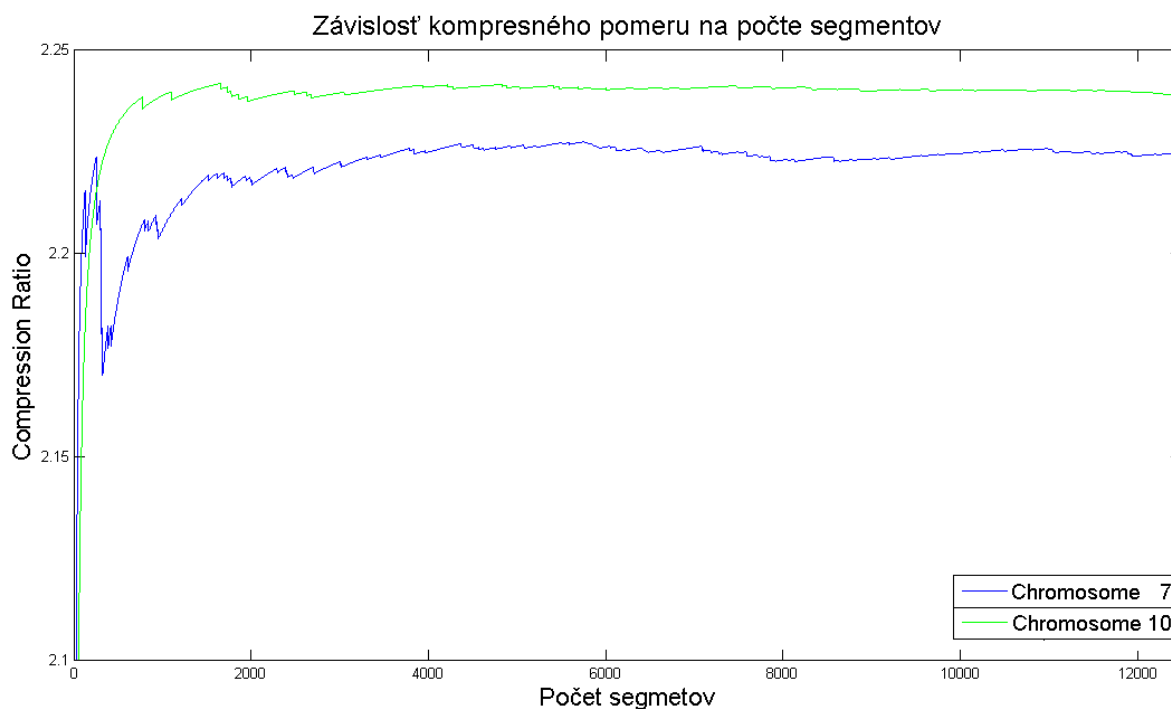
Tabuľka 4.3: Výsledky kompresnej metódy GenbitCompress.

Kompresná metóda GenbitCompress		
Sekvencia	Kompresný pomer	Kompresný zisk
HUMGHCSA	2,221	-10,50%
HUMDYSTROP	2,228	-11,40%
HUMHDABCD	2,224	-11,20%
HUMHPRTB	2,219	-10,95%
MPOMTCG	2,231	-11,55%
VACCG	2,217	-10,85%
Priemer	2,223	-11,08%

Ďalej bola kompresná metóda testovaná na prvých desiatich chromozómoch človeka (Tabuľka 4.2) s dĺžkou sekvencie 100 000 znakov. Vývoj kompresného pomeru na počte znakov chromozómu 7 a 10 je vynesý do grafu (Obrázok 4.9). Kde pokles kompresného pomeru naznačuje výskyt opakujúceho sa segmentu.

Tabuľka 4.4: GenbitCompress na chromozómoch človeka.

Kompresná metóda GenbitCompress		
Chromozóm	Kompresný pomer	Kompresný zisk
1.	2,238	-11,90%
2.	2,231	-11,55%
3.	2,222	-11,10%
4.	2,235	-11,75%
5.	2,219	-10,95%
6.	2,231	-11,55%
7.	2,221	-11,05%
8.	2,223	-11,15%
9.	2,218	-10,90%
10.	2,223	-11,15%
Priemer	2,227	-11,37%



Obrázok 4.9: Kompresný pomer chromozómu 7 a 10.

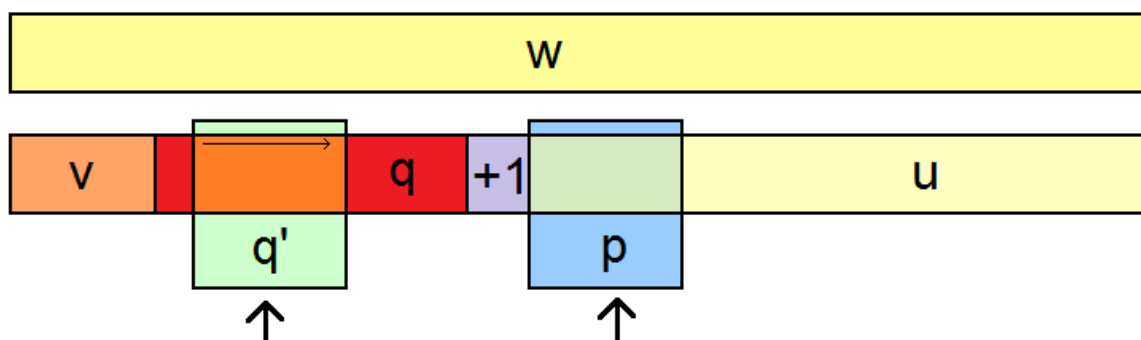
Z výsledkov je zrejme, že kompresná metóda nedosahuje prijateľné výsledky. Ako je popísané vyššie (kapitola 4.2.1), efektivita kompresnej metódy sa nedá plne preukázať na sekvenciách z verejných databáz. Ak by kompresný pomer mal byť pod 2 bpb (bits per base), tak podľa rovnice (4.1.), ktorá definuje celkový počet bitov na výstupe, opakujúci sa segment zložený zo štyroch báz musel by byť každý ôsmy v poradí, čo bohužiaľ je možné iba v krátkych úsekoch sekvencie nie v celej, to znamená, že algoritmus môže slúžiť práve na vyhľadávanie takýchto úsekov.

4.2.2 KompresDNA

Kompresné algoritmy rozobraté v kapitole 3.2 sú inšpiráciou na vytvorenie kompresného algoritmu orientovaného na DNA sekvencie, ktorý sa stáva hlavným predmetom diplomovej práce. Kompresia je založená na vyhľadávaní opakujúcich sa častí sekvencie, tzv. repetícií na základe predošlého výskytu, s možnosťou jednej editačnej operácie konkrétne zámene znaku (Replace) na dvoch pozíciách.

Kompresný algoritmus je jednoprechodovou bezstratovou kompresnou metódou. Pre danú vstupnú sekvenciu w , má predpoklad, že časť v je skomprimovaná a zostávajúcou časťou na komprimáciu je u (Obrázok 4.10). Algoritmus vyhľadá optimálnu subsekvenciu náhľadového okna p z u , ktorá približne zodpovedá určitej subsekvencii q' alebo jej prevrátenému tvaru v prehľadovom okne q z v , potom ju zakóduje ekonomicky aj pomocou editačných úprav, ak sú potrebné. Po zakódovaní, odstráni subsekvenciu p z u a pridá

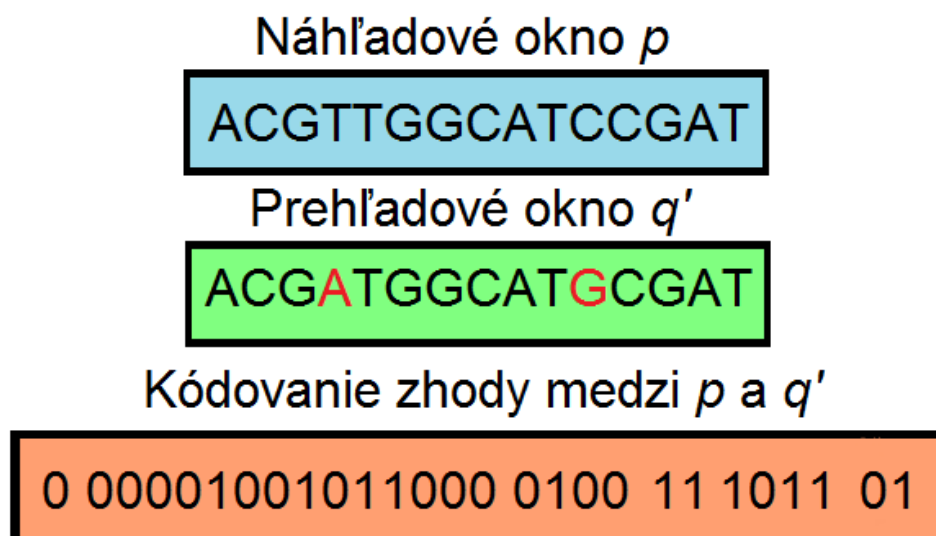
subsekvenciu p do v . Pokračuje kým sa $w = v$. Postup je inšpirovaný kompresnou metódou GenCompress [6].



Obrázok 4.10: Myšlienka kompresnej metódy.

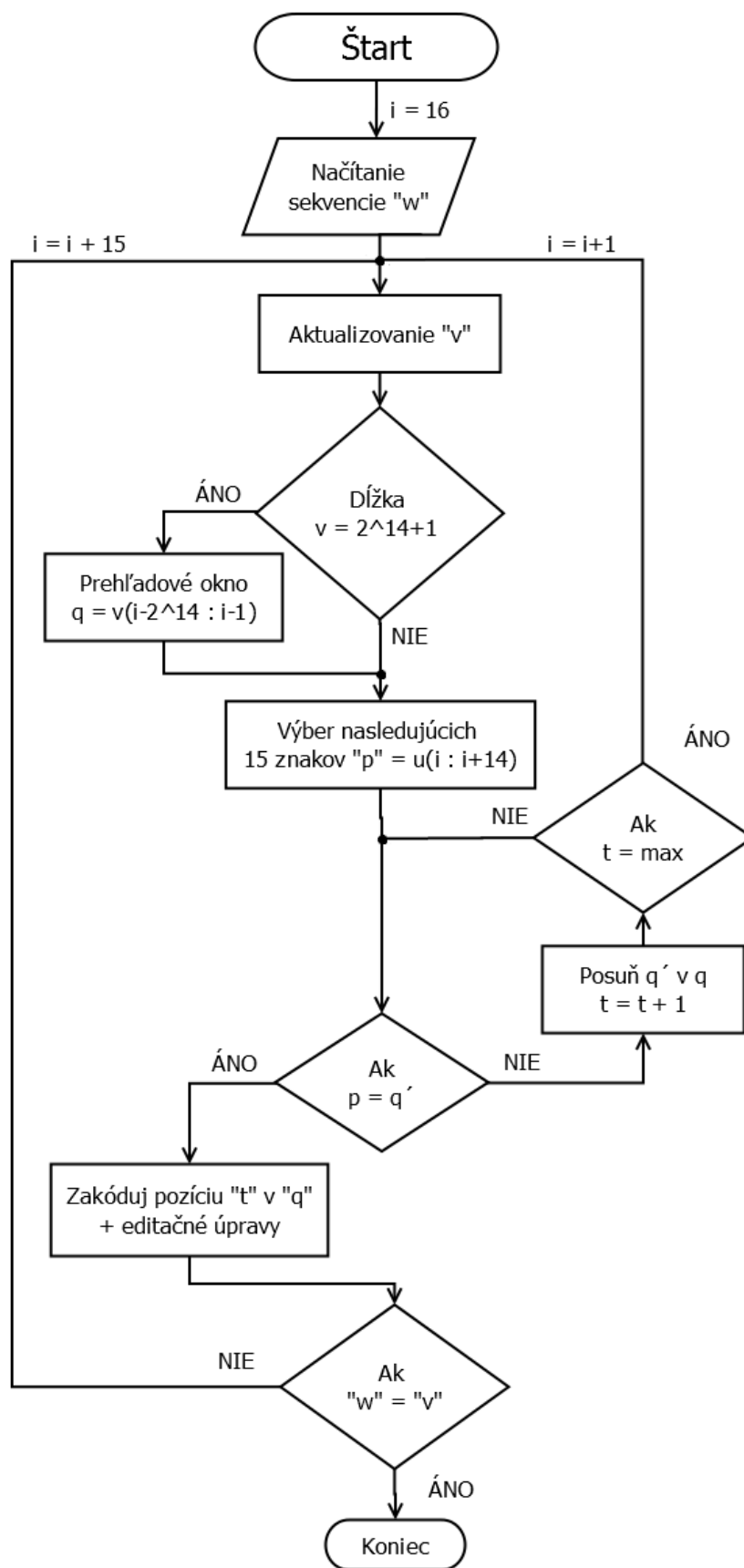
Kompresia:

1. Vloženie DNA sekvencie w vo formáte FASTA, s použitím podpríkazu IgnoreGaps, ktorý ignoruje iné znaky ako A,C,G,T.
2. Aktualizovanie skomprimovanej časti v , ktorá obsahuje prvých 15 znakov sekvencie, zakódovaných dvoma bitmi na bázu.
3. Do náhľadového okna sa vloží ďalších 15 znakov (Obrázok 4.8), subsekvencia p z u a začne sa porovnávať už so skomprimovanou časťou v , ktorá tvorí prehľadové okno q (max. dĺžka 2^{14}). Prehľadávanie je v poprednom a taktiež spätnom smere, ktoré zaručí väčšiu pravdepodobnosť nájdenia zhody medzi p a q' . Spätným smerom sa myslí výber q' z q a následným otočením subsekvencie (kapitola 2.2).
4. Pri nájdení zhody medzi p a q' zakóduje aktuálnu pozíciu v q (14 bitov), taktiež zakóduje či išlo o popredný alebo spätný smer (1 bit), spolu s editačnými operáciami, to je miesto zameny znaku (2×4 bity) a znakom k náhrade (2×2 bity).
5. Ak sa zhoda nenájde z náhľadového okna p sa prvý znak priradí ku skomprimovanej časti v , aktualizuje sa prehľadové q aj náhľadové okno p a opakuje sa krok 3. a 4. Znak je opäť kódovaný dvoma bitmi na bázu.
6. Kompresia je ukončená ak $w = v$ alebo ak náhľadové okno obsahuje menej ako 15 znakov, ktoré sa pred ukončením kompresie priradí k v ako zvyšok.



Obrázok 4.11: Kódovanie zhody.

Hlavnou podstatou kompresného algoritmu je 30 bitov (zakódovaných 15 znakov dvoma bitmi na bázu) nahradiť 27 bitmi a tým ušetriť miesto vo výstupnom reťazci. Prvý bit kóduje smer (popredný – 0, spätný - 1), ďalších 14 bitov kóduje pozíciu v q a nasledujúce dvojice 4 a 2 bitov kódujú editačné operácie. Ako je zrejmé z obrázka (Obrázok 4.11) prvá editačná operácia na štvrtej pozícii záměna znaku T za A, druhá editačná operácia na jedenástej pozícii záměna znaku C za G.



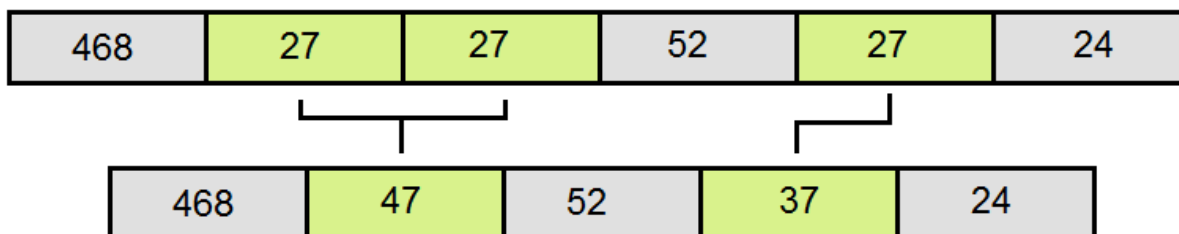
Obrázok 4.12: Vývojový diagram kompresie.

Výstup potom tvoria bitové reťazce rôznych dĺžok (Obrázok 4.13), ktoré sú usporiadané za sebou tak ako sa nachádzali počas kompresie. Dĺžke 27 bitov pripadá stav zhody medzi q' a p , ostatné dĺžky, rôzne od 27 bitov a hodnoty párnych čísel, pripadajú miestam kedy nebola nájdená zhoda medzi q' a p , v čase aktualizovania prehľadového okna, tzv. nezakódované miesta. Spojením reťazcov vznikne na výstupe jeden bitový reťazec z ktorého je počítaný kompresný pomer delením dĺžkou sekvencie na vstupe.

'000000110010100010000101001'	27
'01'	2
'000000011111111000100110010'	27
'000000100001110000000000000'	27
'000001'	6
'000000100011110010110011001'	27
'011010110010100111'	18
'000000100110110011110101110'	27
'000000101000101010000101001'	27
'000000101010100101110111000'	27
'1000000101010010101010101001'	28
'000000101110000010001100010'	27
'0101000010001101'	16
'000000110000111001001011101'	27

Obrázok 4.13: Výstupné bitové reťazce a ich dĺžky.

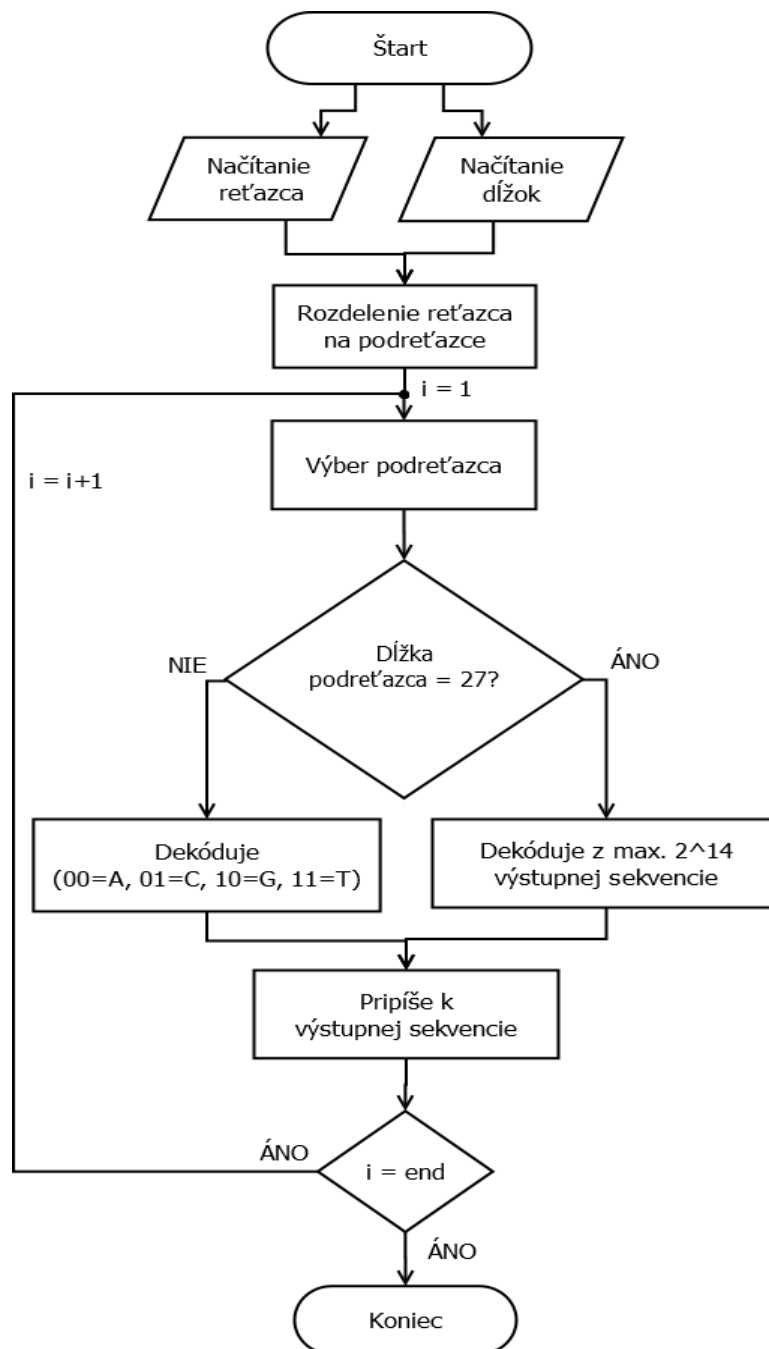
Kompresný pomer tak dosiahne hodnotu pod 2 bpb. Problém vznikne pri dekompresii, kde vstupom je dlhý bitový reťazec (0,1) a nie je jednoznačné aký úsek reťazca ma byť práve dekódovaný. Preto pri kompresii sa musí uložiť súbor, ktorý bude obsahovať jednotlivé dĺžky pod reťazcov. Keďže pod reťazce majú často krát dĺžku 27 bitov a vyskytujú sa aj niekoľkokrát za sebou, tak na úpravu je použitá myšlienka metódy RLE (Run length encoding), ktorá viacnásobný výskyt hodnoty zakóduje na jedno miesto. Pri každom výskyte dĺžky 27 pričíta k číslu 27 hodnotu 10 (Obrázok 4.14), potom pri dekompresii je jasné koľko 27 bitových úsekov nasleduje za sebou.



Obrázok 4.14: Upravená metóda RLE.

Dekompresia:

1. Načítanie vstupného reťazca a súboru s dĺžkami, rozdelenie reťazca.
2. Ak sa dĺžka pod reťazca nerovná 27, tak dekoduj reťazec pomoc dvoch bitov (00, 01, 10, 11) na bázu.
3. Ak sa rovná, tak dekoduj podľa schémy. Prvý bit smer, ďalších 14 bitov pozícia z 2^{14} už dekomprimovanej sekvencie, ďalšie bity pripadajú editačným operáciám.
4. Opakuj krok 4 a 5 po koniec reťazca.



Obrázok 4.15: Vývojový diagram dekompresie.

Pred dekompresiou je vstupný reťazec rozdelený na úseky jednotlivých dĺžok. Postup dekompresie je potom opačný kompresii. Dekompresia je časovo nenáročná oproti kompresii, keďže sa len dekódujú jednotlivé úseky a neprehľadáva sa 2^{14} znakov sekvencie. Pri dĺžke pod reťazca sa tak určí iba pozícia z maximálne 2^{14} dĺžky už dekomprimovanej sekvencie a vyberie sa príslušná subsekvencia dĺžky 15 báz, na ktorej sa ak je to potrebné prejaví editačné úpravy a pripíše sa k výstupnej sekvencii.

Algoritmus bol testovaný na DNA sekvenciách predchádzajúcich experimentov z verejnej databázy (Tabuľka 4.5) a taktiež na prvých desiatich chromozómoch dĺžky 100000 znakov (Tabuľka 4.6).

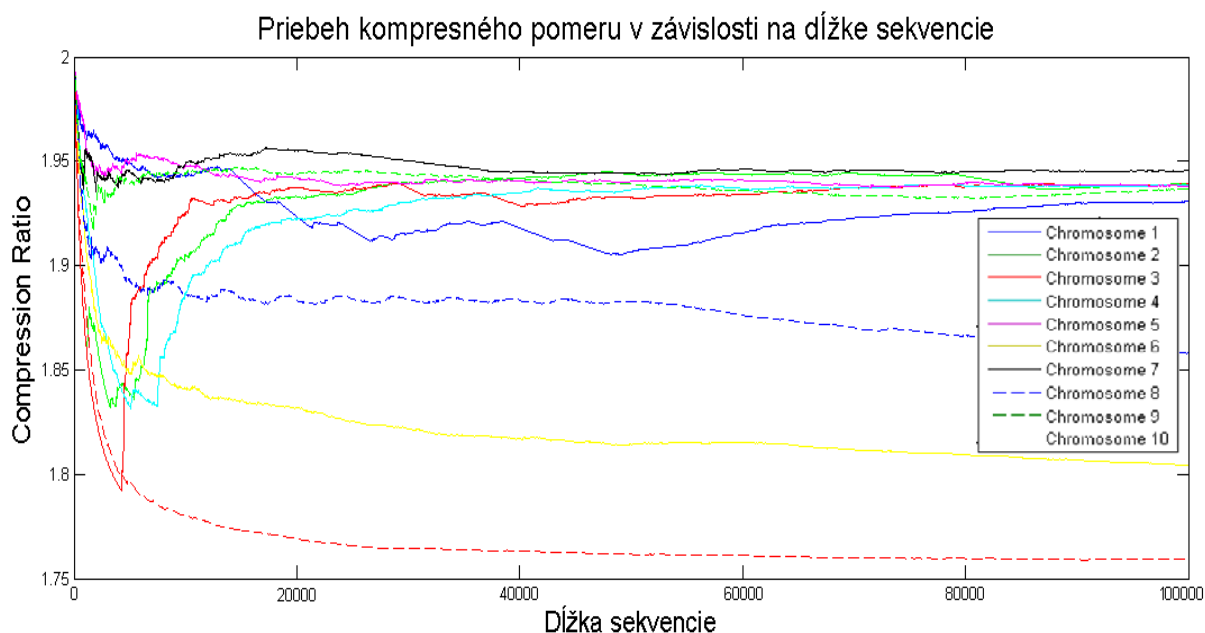
Tabuľka 4.5: Výsledky kompresnej metódy KompresDNA.

Kompresná metóda KompresDNA		
Sekvencia	Kompresný pomer	Kompresný zisk
HUMGHCSA	1,867	6,65%
HUMDYSTROP	1,912	4,40%
HUMHDABCD	1,905	4,75%
HUMHPRTB	1,903	4,85%
MPOMTCG	1,883	5,85%
VACCG	1,852	7,40%
Priemer	1,887	5,65%

Tabuľka 4.6: KompresDNA na chromozómoch človeka.

Kompresná metóda KompresDNA		
Chromozóm	Kompresný pomer	Kompresný zisk
1.	1,878	6,10%
2.	1,888	5,60%
3.	1,889	5,55%
4.	1,885	5,75%
5.	1,883	5,85%
6.	1,702	14,90%
7.	1,891	5,45%
8.	1,711	14,45%
9.	1,925	3,75%
10.	1,577	21,15%
Priemer	1,823	8,86%

Priebeh zmien kompresného pomeru v závislosti na počte znakov je vynesý do grafu (Obrázok 4.16). Miesta s veľkým poklesom kompresného pomeru, určujú tu časť sekvencie, kde sa nachádza najväčší počet repetícií, výskyt zhody medzi p a q .



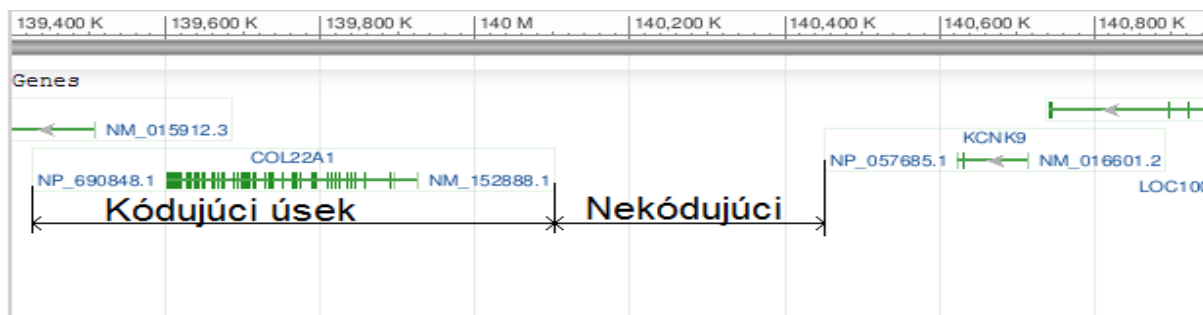
Obrázok 4.16: Vývoj kompresného pomeru na chromozómoch.

Z výsledkov je zrejme, že kompresná metóda dosahuje kompresný pomer pod hodnotu 2 bpb, čo je dôkazom toho, že sekvencie obsahujú opakujúce sa úseky, ktoré je možné zakódovať ekonomicky. Problém kompresnej metódy nastáva pri uložení ako bolo spomenuto vyššie a taktiež je veľmi časovo a pamäťovo náročná.

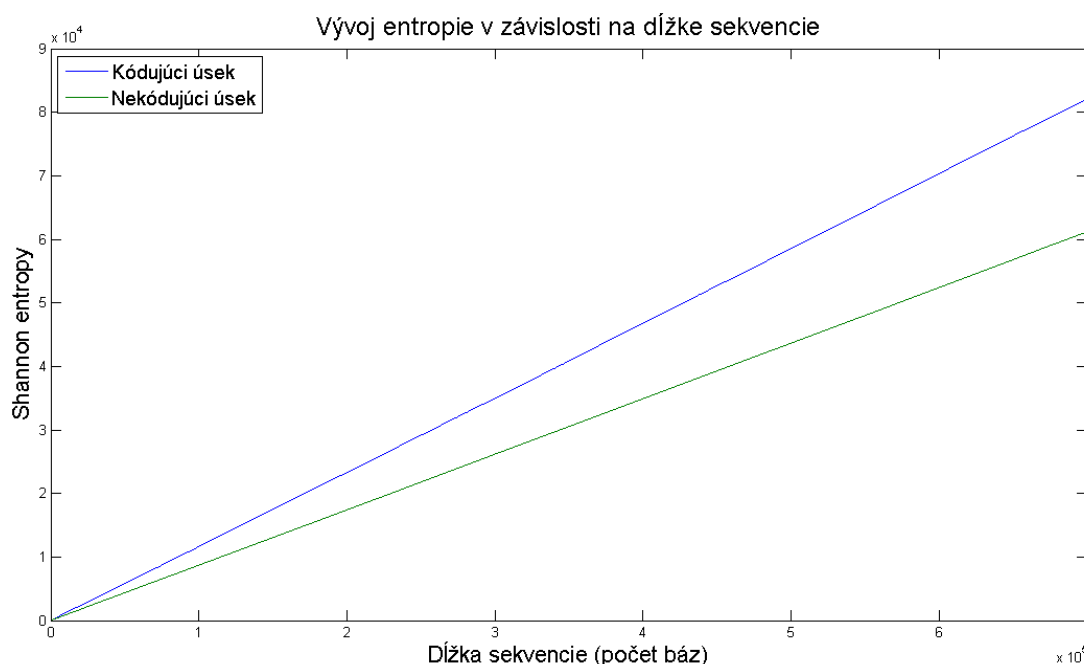
U chromozómoch 2,3,4 je vidieť rýchly pokles kompresného pomeru na začiatku kompresie to je spôsobené tým že sekvencie sú úsekom začiatku chromozómu a tvoria jeden z telomérických koncov, ktoré slúžia k udržiavaniu polohy chromozómov a taktiež chránia chromozómy aby spolu nerekombinovali. Tieto úseky sú nekódujúcou časťou DNA sekvencie, tvorí ju opakujúci segment dĺžky 6 – 10 bázových párov, ktorý má 100 až tisíc opakovaní.

4.3 Analýza kódujúcich a nekódujúcich úsekov DNA sekvencie

Ako bolo spomenuté vyššie DNA sekvencie obsahujú kódujúce (introny) a nekódujúce úseky (exony). Kódujúce úseky sú charakteristické tým, že obsahujú veľké množstvo opakujúcich sa častí sekvencie, naopak nekódujúce úseky neobsahujú opakujúce sa časti sekvencie, tzv. repetície a preto ich kompresia takto ladenými kompresnými metódami nemusí dosahovať pozitívnych výsledkov. V tejto časti sú kompresné algoritmy KompresDNA a GenbitCompress vyskúšané na kódujúcom a nekódujúcom úseku DNA sekvencie chromozómu 8 (Obrázok 4.17). Taktiež je odhadnutá Shannova entropia (Obrázok 4.18), kde je zrejme, že pri väčšom počte znakov sa entropia úsekov pomaly rozchádza.

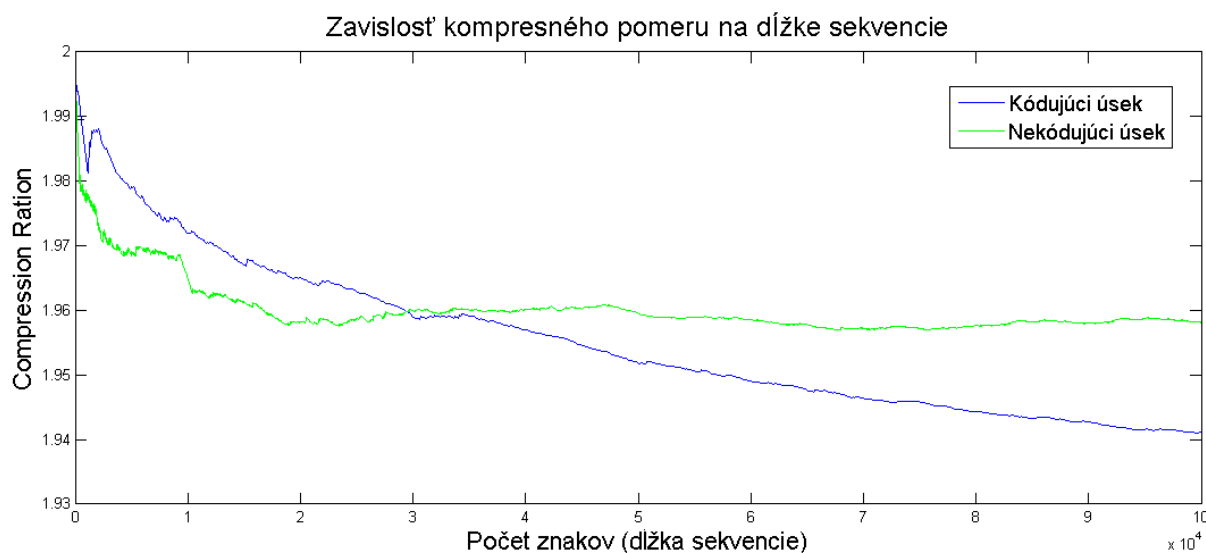


Obrázok 4.17: Grafické zobrazenie DNA sekvencie.



Obrázok 4.18: Shannova entropia kódujúceho a nekódujúceho úseku sekvencie.

Rozdiel pri kompresii kódujúcich a nekódujúcich úsekov by sa mal prejavíť na vývoji kompresného pomeru v závislosti na počte znakov sekvencie, kedy pri väčšom skomprimovanom úseku kódujúcej časti sekvencie malo dochádzať k častejším nachádzaním opakujúcich sa časti sekvencie, túto vlastnosť je možné pozorovať pri kompresnej metóde KompresDNA (Obrázok 4.19). Z obrázka je tak zrejmé, že pri kódujúcom úseku dochádza k pravidelnejšiemu nachádzaniu repetícií a tak k väčšiemu spádu priebehu.



Obrázok 4.19: Kompresia kódujúceho a nekódujúceho úseku metódou KompresDNA.

Pri použití kompresnej metódy GenbitCompress (Obrázok 4.20) je priebeh očakávaný. Metóda nerobí rozdiely medzi kódujúcimi a nekódujúcimi úsekmi DNA sekvencie, keďže vyhľadáva len zhodné segmenty zložené zo štyroch báz, ktoré sú susedné. Preto vývoj kompresného pomeru je totožný pre oba úseky.



Obrázok 4.20: Kompresia kódujúceho a nekódujúceho úseku metódou GenbitCompress.

4.4 Porovnanie výsledkov

Výsledky sú porovnané s priemernými hodnotami kompresných pomerov a ziskov DNA orientovaných kompresných metód (kapitola 3.2), ktoré boli testované na zhodných sekvenciách (Tabuľka 4.7).

Metóda GenbitCompress má veľmi dobrú myšlienku, je jednoduchá, no pri kompresii DNA sekvencii z verejných databáz je nepoužiteľná. Výhodou tohto algoritmu je, že je jednoduchý, rýchly a na jeho výsledkoch nebádať rozdiely v kompresii kódujúcich a nekódujúcich častí DNA sekvencie. Kompresná metóda je skôr použiteľná pri porovnávaní sekvencií (Multiple analysis), tiež môže byť kompromis medzi pamäťovou a časovou zložitou.

Kompresná metóda KompresDNA dosahuje obstojné výsledky v porovnaní s oveľa sofistikovanejšími metódami, ktoré na kódovanie neopakujúcich častí využívajú dynamické programovanie, jej hlavnou výhodou je, že bádať rozdiely v kompresii kódujúcich a nekódujúcich úsekov DNA sekvencie. Oproti metóde GenbitCompress má vysokú pamäťovú a časovú náročnosť.

Tabuľka 4.7: Porovnanie výsledkov.

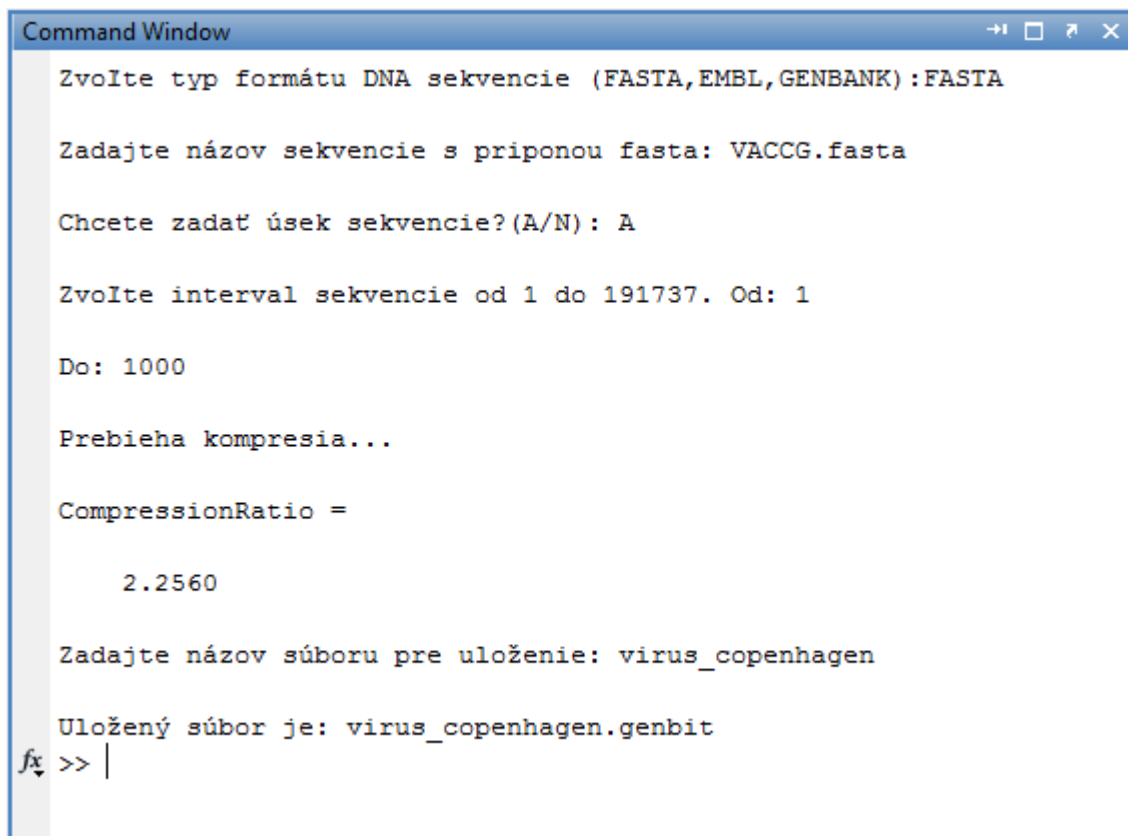
Kompresná metóda	Kompresný pomer	Kompresný zisk
BioComp	1,7861	10,70%
GenComp	1,7252	13,73%
DNACopm	1,7007	15,00%
CTW+LZ	1,7236	13,82%
DNAC	1,7001	15,00%
GeNML	1,6723	16,39%
DNAPack	1,6879	15,61%
GenbitCompress	2,2230	-11,08%
KompresDNA	1,8870	5,65%

4.5 Popis programu

Program bol vytvorený v programe MATLAB s jednoduchým užívateľským rozhraním pomocou dotazov v Command Window (Obrázok 4.21), tvorba m-filov pre GenbitCompress, KompresDNA a Shannon entropy (viď. Príloha).

Po spustení, program vyzve užívateľa k výberu formátu, všetky experimenty boli z formátov FASTA, no program dokáže čítať aj formáty EMBL a GENBANK. Po zvolení formátu vyzve k napísaniu názvu sekvencie spolu s príponou, ktorá sa musí nachádzať v spoločnom priečinku ako daný m-file. Dôjde k načítaniu sekvencie. Ďalším dotazom je, či má byť skomprimovaná celá sekvencia alebo len jej časť, úsek sekvencie. Potom začne

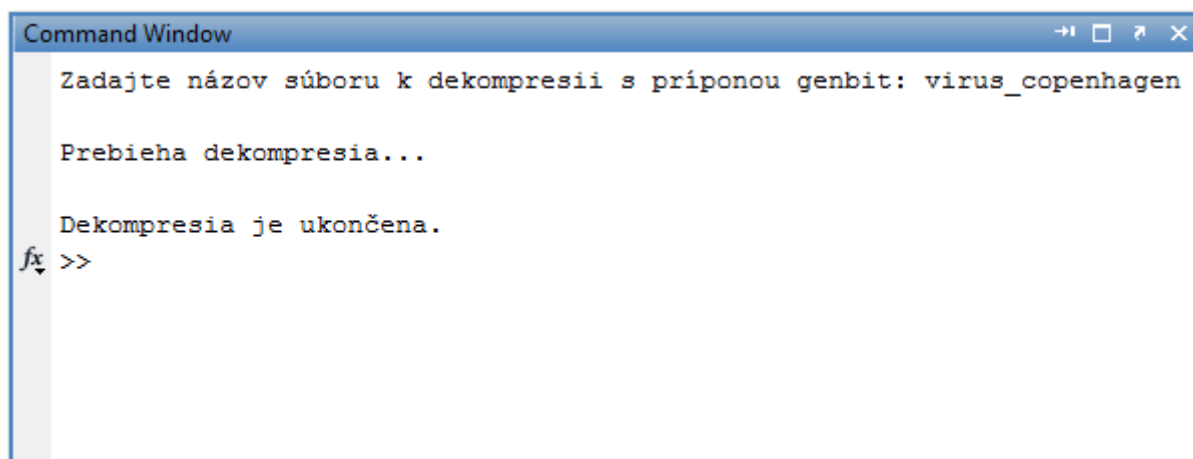
prebiehať kompresia, po jej ukončení je vypísaný kompresný pomer a vyzýva k napísaniu názvu skomprimovanej sekvencie. Podľa kompresnej metódy priradí príponu, pre GenbitCompress je to prípona *.genbit, pre metódu KompresDNA je to prípona *.komdna pre výstupný reťazec a pre jednotlivé dĺžky *.komdnal.



```
Command Window
Zvoľte typ formátu DNA sekvencie (FASTA,EMBL,GENBANK):FASTA
Zadajte názov sekvencie s príponou fasta: VACCG.fasta
Chcete zadať úsek sekvencie?(A/N): A
Zvoľte interval sekvencie od 1 do 191737. Od: 1
Do: 1000
Prebieha kompresia...
CompressionRatio =
2.2560
Zadajte názov súboru pre uloženie: virus_copenhagen
Uložený súbor je: virus_copenhagen.genbit
fx >> |
```

Obrázok 4.21: Priebeh kompresie.

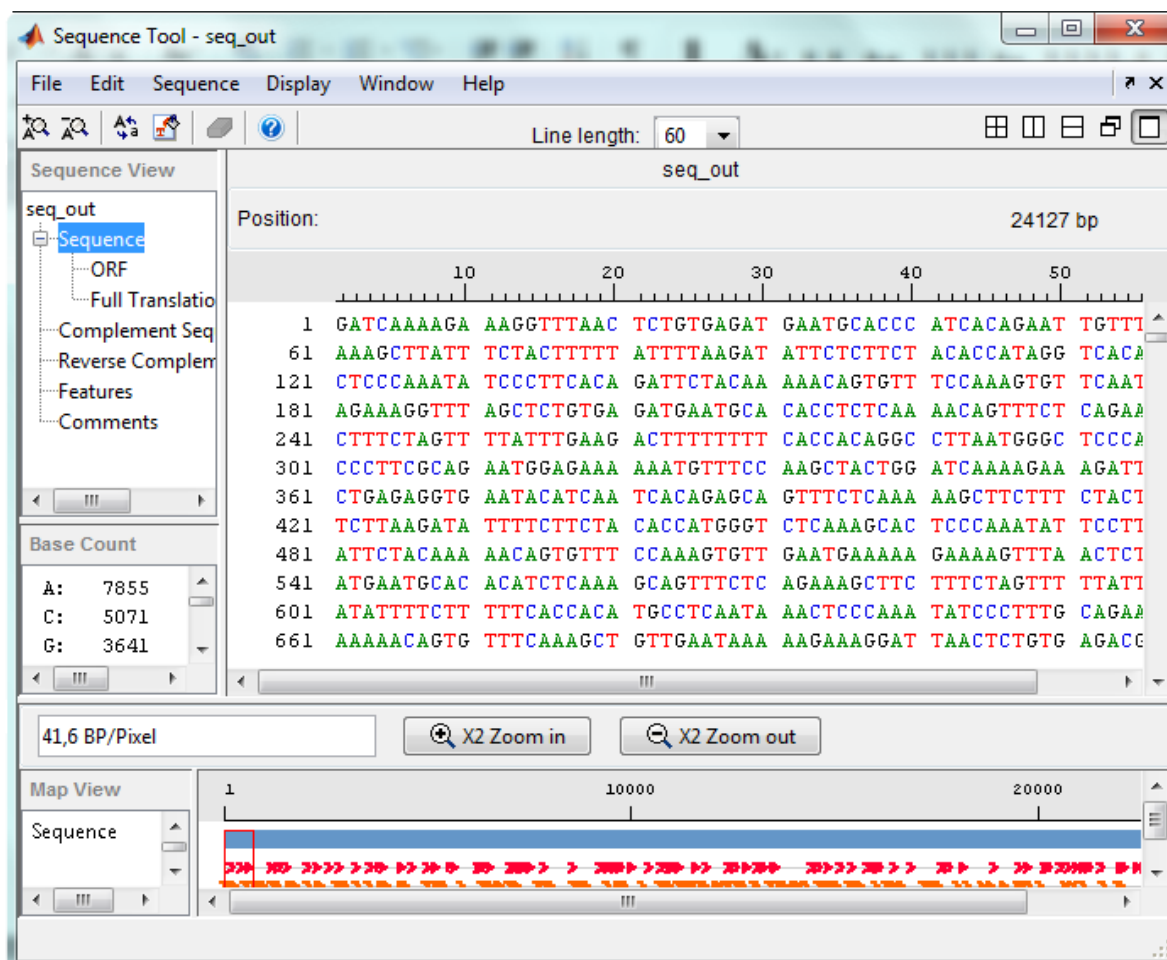
Dekompresia (Obrázok 4.22) je potom obdobná, po spustení programu sa len napíše názov skomprimovaného súboru bez prípony a začne prebiehať dekompresia.



```
Command Window
Zadajte názov súboru k dekompresii s príponou genbit: virus_copenhagen
Prebieha dekompresia...
Dekompresia je ukončena.
fx >>
```

Obrázok 4.22: Priebeh dekompresie.

Po ukončení dekompresie je sekvencia zobrazená pomocou nástroja Sequence Tool (Obrázok 4.23), ktorý obsahuje prehľadávanie sekvencie a mnoho ďalších funkcií. Dôležitá vec, ktorú tento nástroj ponúka je uloženie, **File>Save as FASTA...**, sekvencie vo formáte FASTA, s možnosťou vyplnenia hlavičky.



Obrázok 4.23: Sequence Tool.

Obdobne je vytvorený aj program pre odhad Shannonovej entropie, kde po výbere formátu sekvencie, načítaní sekvencie, zadania jej úseku dôjde k vypočítaniu entropie. Na konci je možnosť si prehliadnuť vývoj entropie v grafe.

Časová náročnosť jednotlivých kompresných metód je rozdielna, dôležitým faktorom, ktorý určuje čas kompresie je dĺžka sekvencie. Metódy boli testované na bežnom stolovom počítači. Pri metóde GenbitCompress sa pri dĺžke sekvencie 100 000 bázových párov, kompresia chromozómov, čas pohyboval v rozmedzí od 103 do 114 minút. Kompresná metóda KompresDNA je o čosi náročnejšia, pri kompresii chromozómov, tej istej dĺžky, trvala kompresia od 6,2 do 7,1 hodín.

Záver

Tvorbou diplomovej práce som sa oboznámil so základom molekulárnej biológie, špeciálne biologických sekvencií ich formátov a následným spracovávaním a to so zámerom odhadu entropie a kompresie.

Literárna rešerš slúži ako dobrý materiál k pochopeniu súvislostí v molekulárnej biológii s konkrétnym zámerom na biologické sekvencie a špeciálne na spracovanie DNA sekvencií. Pochopenie štruktúry sekvencie z akých častí sa skladá, ako je usporiadaná a čo sa deje s DNA sekvenciou, centrálna dogma molekulárnej biológie.

Teória informácie a jej odhad entropie je aplikovaný na DNA sekvencie. Zisťuje nové skutočnosti o DNA sekvenciách najmä ich podobnosť vo vývoji entropie pri danom počte báзовých párov, čo je preukázané odhadom Shannonovej entropie chromozómov človeka.

Realizáciou kompresnej metódy GenbitCompress je poukázané na to, že aj jednoduché algoritmy nájdu uplatnenie v bioinformatike. Výhodou tohto algoritmu je jednotná kompresia kódujúcich a nekódujúcich častí DNA sekvencie. Najväčšie uplatnenie algoritmu je pri porovnávaní jednotlivých sekvencií.

Hlavným predmetom práce je navrhnutý algoritmus KompresDNA pre vyhľadávanie štruktúrnych podobností a opakovaných vzorov s následnou bezstratovou kompresiou DNA sekvencií. Rozdiel tohto algoritmu oproti ostatným DNA orientovaným kompresným metódam je, že neprehľadáva skomprimovanú sekvenciu len v jednom smere ale aj v opačnom, čo má za následok dvakrát väčšiu možnosť nájdenia opakujúceho sa úseku. Algoritmus je jedným zo sofistikovanejších spôsobov vyhľadávania opakovaných vzorov, keďže pri kódovaní využíva aj editačné operácie. Algoritmus je prevedený dostatočným počtom testov, ktoré poukazujú na jeho dobrú kompresnú myšlienku a v porovnaní s inými zložitejšími kompresnými metódami dosahuje obstojný kompresný pomer. Taktiež sa preukázal ako dobrý kompresný nástroj pri kompresii kódujúcich častí DNA sekvencie.

Výstupom je taktiež program, ktorý umožňuje odhad entropie a následnú kompresiu DNA sekvencie pomocou kompresných metód GenbitCompress a KompresDNA, na ktorých výstupoch po kompresii je možno získať späťne DNA sekvenciu bezstratovou dekompresiou.

Zoznam literatúry

- [1] SHANNON, C. E. *A Mathematical Theory of Communication*. AT&T Tech. J. 1948, 27, 379–423, 623–656.
- [2] GATLIN, L. *Information Theory and the Living System*. Columbia University Press: New York, NY, USA, 1972.
- [3] SCHNEIDER, T., STEPHENS, R. *Sequence logos: A new way to display consensus sequences*. Nucleic Acids Res. 1990, 18, 6097–6100.
- [4] ZIV, J., LEMPEL, A. *A Universal Algorithm for Sequential data Compression*. IEEE Trans. Inform. Theory. Vol IT – 23, 337 – 343, 1977.
- [5] ZIV, J., LEMPEL, A. *Compression of Individual Sequences via Variable-Rate Coding*. IEEE Trans. Inform. Theory. Vol IT – 24, 530 – 536, 1978.
- [6] CHEN, Xi., KWONG, S., MING, Li. *A Compression Algorithm for DNA Sequences*. IEEE Engin. in Med. and Bio. 2001.
- [7] CVRČKOVÁ, F. *Úvod do praktické bioinformatiky*. Praha: Academia. 2006.
- [8] HAYES, B. *The invention of the genetic code*. Am sci 86(1):8-14, 1998.
- [9] ROSYPAL, S. *Úvod do molekulární biologie*. Brno: Rosypal, 1998.
- [10] RAJESWARI, R., APPARAO, A. *Genbit Compress – algorithm for repetitive and non-repetitive DNA sequences*. Journal of Theoretical and Applied Information Technology. 2010, roč. 11, č. 1, s. 25 – 29.
- [11] GRUMBACH, S., TAHI, F. *Compression of DNA sequences*. In Data compression conference, s. 340 – 350. IEEE Computer society press, 1993.
- [12] GRUMBACH, S., TAHI, F. *A new challenge for compression algorithms: Genetic sequences*. Journal of Information and Management, vol. 30, s. 857 – 866, 1994.
- [13] Rivls, E., Delahayem, P., Dauchet, M., Delgrange, O. *A Guaranteed compression scheme for repetitive DNA sequences*. Data Compression Conference, Snowbird, UT, s. 453, 1996.
- [14] CHEN, X., LI, M., MA, B., TROMP, J. *DNACompress: A fast and effective DNA sequence compression*. Bioinformatics, vol. 18, no. 12, s. 1696 – 1698, 2002.
- [15] LI, M., MA, B., TROMP, J. *PatternHunter - Faster and more sensitive homology search*. Bioinformatics, vol. 18, s. 440 – 445, 2002.
- [16] MATSUMOTO, T., SADAKANE, K., IMAI, H. *Biological sequence compression algorithm*. Genome Informatics Workshop, Universal Academy Press, vol. 11, s. 43 – 52, 2000.

- [17] CHANG, C. H. *DNAC: A compression algorithm for DNA sequences by non-overlapping approximate repeats*. Master thesis, 2004.
- [18] TABUS, I., KORODI, G., RISSANEN, J. *DNA sequence compression using the normalized maximum likelihood model for discrete regression*. Data Compression Conference, Snowbird, UT, s. 253 – 262, 2003.
- [19] BEHZADI, B., LE FESSANT, F. *DNA compression challenge revisited*. Symposium on Combinatorial Pattern Matching, s. 190 – 200, 2005.
- [20] National Center for Biotechnology Information. U.S. National library of Medicine. Bethesda MD, [cit. 4.5.2013]. Dostupné na: <http://www.ncbi.nlm.nih.gov/gquery/>
- [21] Bystrý, V. *Systém pro odhad entropie a detekci struktury v biologických sekvencích*. Diplomová práce. Masarykova Univerzita, Fakulta informatiky. Brno, 2008.
- [22] The Astrophysics & Astrochemistry Laboratory. Mountain View, 2009 [cit. 10.3.2013]. Dostupné na: http://www.astrochem.org/sci_img/dna.jpg
- [23] HyperPhysics. Georgia State University. Georgia, 2008 [cit. 10.3.2013]. Dostupné na: <http://hyperphysics.phy-astr.gsu.edu/hbase/organic/transcription.html>
- [24] Biogen, molekulární biologie a genetika. Biogen Praha. Praha, 2013 [cit. 10.3.2013]. Dostupné na: <http://eshop.biogen.cz/sekvenace-klasicka-sanger/sekvenace-klasicka-sangerovou-metodou>
- [25] Wiley, J. *Essential Biochemistry*. United Kingdom, 2004 [cit. 10.3.20013] http://www.wiley.com/college/pratt/0471393878/student/structure/trna_aars/trna.gif

Zoznam skratiek

DNA	deoxiribonucleic acid, deoxiribonukleová kyselina
RNA	ribonucleic acid, ribonukleová kyselina
EMBL	European molecular biology laboratory, Európske laboratórium molekulárnej biológie
DDBJ	DNA data bank of Japan, Japonská DNA databáza
NCBI	National center of biotechnology information, Národné centrum pre biotechnologické informácie

Prílohy

Obsah CD

1. GenbitCompress
 - 1.1 GenbitCompress.m
 - 1.2 GenbitDecompress.m
2. KompresDNA
 - 1.1 KompresDNAcompress.m
 - 1.2 KompresDNAdecompress.m
 - 1.3 decode.m – function
 - 1.4 encode.m – function
 - 1.5 findR.m - function
3. Shannon entropy
 - 1.1 ShannonEntropy.m
4. Elektronická verzia diplomovej práce