

UNIVERZITET U BEOGRADU
MATEMATIČKI FAKULTET

Milena Dukanac

**JEZIK ELIXIR SA PRIMENOM U
SEKVENCIJIRANJU GENOMA**

master rad

Beograd, 2019.

Mentor:

dr Milena VUJOŠEVIĆ JANIČIĆ, docent
Univerzitet u Beogradu, Matematički fakultet

Članovi komisije:

dr Jovana KOVAČEVIĆ, docent
Univerzitet u Beogradu, Matematički fakultet

dr Vesna MARINKOVIĆ, docent
Univerzitet u Beogradu, Matematički fakultet

Datum odbrane: _____

Mami i tati za nesebičnu ljubav, podršku i razumevanje

Naslov master rada: Jezik Elixir sa primenom u sekvencioniranju genoma

Rezime:

Ključne reči:

Sadržaj

1	Uvod	1
2	Elixir	3
2.1	Razvojno stablo	3
2.2	Uputstvo za instalaciju i pokretanje Elixir-a	7
2.3	Osnovne karakteristike	9
2.4	Osnovni tipovi podataka	9
2.5	Osnovni operatori	19
2.6	Poklapanje obrazaca	21
2.7	Nepromenljivost podataka	23
2.8	Odlučivanje	24
2.9	Moduli	25
2.10	Direktive	26
3	Sekvencioniranje genoma	28
3.1	Istorija sekvencioniranja genoma	31
3.2	Sekvencioniranje <i>shotgun</i> celokupnog genoma	32
3.3	Sekvencioniranje <i>de novo</i> genoma za kratka očitavanja	36
3.4	Korekcija grešaka	37
3.5	Konstrukcija kontiga	39
4	Implementirani algoritmi i rezultati	46
4.1	Algoritam <i>JellyFish</i>	46
4.2	Algoritam <i>DSK</i>	48
4.3	Algoritam <i>De_Brujin_graph</i>	50
4.4	Algoritam <i>All_Eulerian_cycles</i>	51
4.5	Uputstvo za pokretanje aplikacije	52

SADRŽAJ

4.6	Opis aplikacije	52
4.7	Rezultati	55
5	Zaključak	59
	Literatura	60

Glava 1

Uvod

U oblasti razvoja programskih jezika se stalno se dešavaju promene. Velika pažnja i značaj se pridaje razvoju novih programskih jezika koji omogućavaju brže i lakše programiranje, bržu obradu velike količine podataka i podršku za konkurentne procese. Upravo sa tim ciljem, u toku prethodnih godina nastao je veliki broj novih programskih jezika i razvojnih okruženja za njih. Jedan od novih jezika koji je nastao 2011. godine je *Elixir*. Za veoma kratko vreme postao je izuzetno popularan i danas ga veliki broj kompanija koristi za svoje projekte i softverska rešenja. Neki od razloga za popularnost Elixir-a su brza obrada velike količine podataka, skalabilnost, podrška za konkurentne procese i veoma visoka tolerancija na greške.

Rekonstrukcija genoma pomoću sekvencioniranja genoma je važan problem. Sekvencioniranje genoma omogućava istraživačima da proučavaju ne samo gene koji kodiraju važne proteine u našem organizmu, već i čitave oblasti DNK koji imaju druge važne uloge. Naši genomi ostaju relativno konstantni tokom čitavog života, tako da je genom dovoljno sekvencionirati jedanput kako bi se mogao proučavati iznova i iznova. U medicini, jedna od oblasti gde je sekvencioniranje genoma od izuzetnog značaja je dijagnostikovanje rizika od naslednih bolesti. Na primer, otkrivanje rizika od razvoja određenog karcinoma će uticati na obavljanje češćih skrininga. Češći skriningi će pomoći da se bolest na vreme otkrije, što je od presudnog značaja za proces lečenja.

Cilj ovog rada je prikazivanje osnovnih koncepata i osobina programskog jezika Elixir kroz primenu u implementaciji algoritama koji se koriste u oblasti sekvencioniranja genoma. Takođe, cilj je i upoznavanje sa osnovnim pojmovima i problemima ove oblasti, kao i prikazivanje značaja njenog proučavanja i unapređivanja. Poglavlje 2 je posvećeno Elixir-u. Prvo je prikazano razvojno stablo na kome se može

videti koji su jezici uticali na njegov nastanak i razvoj, kao i godine nastanka tih jezika. Zatim su predstavljeni osnovni tipovi podataka i osnovne osobine Elixir-a, zajedno sa primerima koda radi boljeg razumevanja. Poglavlje 3 je posvećeno oblasti sekvencioniranja genoma. Uvodni deo ovog poglavlja služi upoznavanju osnovnih pojmova iz ove oblasti i njenom istorijskom razvoju i napredovanju. Nakon toga su obrađeni osnovni koraci koji se sporovode u toku sekvencioniranja genoma, kao i algoritmi koji pomažu u ovom procesu. Poglavlje 4 je namenjeno opisu implementacije algoritama koji su obrađeni u okviru ovog rada. Dati su pseudokodovi zajedno sa razmatranjem njihove vremenske i prostorne složenosti. Potom sledi deo koji je posvećen interfejsu, njegovom izgledu i načinu pokretanja algoritama u okviru njega.

Glava 2

Elixir

Elixir je funkcionalan programski jezik nastao 2011. godine. Njegovim tvorcem se smatra José Valim (engl. *José Valim*). Elixir je dizajniran za izgradnju skalabilnih i lako održivih aplikacija. Posедуje jednostavnu i modernu sintaksu. Zbog svoje funkcionalne prirode, izuzetnog rada u distribuiranim sistemima i tolerancije na greške koja je na jako visokom nivou, u Elixir-u je rađeno mnogo zanimljivih projekata iz oblasti robotike. Takođe se uspešno koristi u razvoju veba i u domenu softvera za uređaje sa ugrađenim računarom (engl. *embedded software*).

2.1 Razvojno stablo

Na nastanak Elixir-a je uticao programski jezik **Erlang**. Pri njegovom kreiranju značajnu ulogu u smislu sintakse imao je programski jezik **Ruby**, a iz jezika kao što su **Python**, **Haskell** i **Clojure** je preuzeo mnoge koncepte. Razvojno stablo jezika Elixir može se videti na slici [2.1](#).

Erlang

Firma Erikson je 1981. godine oformila novu laboratoriju **Erikson CSLab** (engl. *The Ericsson CSLab*) sa ciljem da predlaže i stvara nove arhitekture, koncepte i strukture za buduće softverske sisteme [\[2\]](#). Jedan od zadataka novonastale laboratorije bio je dodavanje konkurentnih procesa u programski jezik **Prolog**¹ i

¹Prolog (engl. *PROgramming in LOGic*) je deklarativan programski jezik namenjen rešavanju zadataka simboličke prirode. Prolog se temelji na teorijskom modelu logike prvog reda. Početkom 1970-ih godina **Alen Kolmerur** (engl. *Alain Colmerauer*) i **Filip Rusel** (engl. *Philippe Rousel*) na Univerzitetu u Marselju, zajedno sa **Robertom Kovalskim** (engl. *Robert Kowalski*) sa



Slika 2.1: Razvojno stablo jezika Elixir

njegovo unapređivanje. Prolog predstavlja začetak novog programskog jezika koji je 1987. godine nazvan **Erlang**. Ime je nastalo zahvaljujući inicijativi zaposlenih koji su radili na telefonskim prekidačima, a za koje je jezik dizajniran. Naime, oni su predložili da jezik nosi ime Erlang u čast danskog matematičaru i inženjeru Agneru Krarupu Erlangu (engl. *Agner Krarup Erlang*), a što je ujedno odgovaralo i skraćenici od „**E**ricsson **L**anguage”. Erlang se smatrao dijalektom Prologa sve do 1990. godine, kada je postao potpuno samostalan programski jezik sa sopstvenom sintaksom. Međutim, neke delove sintakse i koncepte iz Prologa je zadržao (promenljive počinju velikim slovom, svaka funkcionalna celina se završava tačkom, poklapanje obrazaca (engl. *pattern matching*)).

Nakon mnogo godina rada nastajale su sve brže, bolje i stabilnije verzije jezika,

Departmana veštačke inteligencije na Univerzitetu u Edinburgu, razvili su osnovni dizajn jezika Prolog.

kao i **standardna biblioteka OTP** (engl. *The Open Telecom Platform*) [18]. Od decembra 1998. godine, kada su postali deo slobodnog softvera (engl. *open source software*), Erlang i OTP se mogu slobodno preuzeti sa zvaničnog sajta jezika Erlang [18]. Erlang dobija široko prihvatanje pojavom višejezgarnih procesora i njihovog novog skalabilnog pristupa konkurentnosti. Erlang je funkcionalan jezik idealan za svaku situaciju u kojoj su paralelnost, tolerancija na greške i brz odziv neophodni [4], te se koristi u velikom broju kompanija za razvoj njihovih glavnih softverskih rešenja (npr. Erikson (engl. *Ericsson*), Motorola, Votsap (engl. *WhatsApp*), Jahu (engl. *Yahoo!*), Fejsbuk (engl. *Facebook*)).

Elixir je preuzeo izmenjenu Erlangovu sintaksu i dopunjenu Erlangovu standardnu biblioteku. Pokreće se na vituelnoj mašini jezika Erlang, što znači da je nasledio i sve karakteristike Erlang platforme koja postoji već godinama i koja se pokazala pouzdanim rešenjem za skalabilne aplikacije.

Python

Python je interpretirani jezik opšte namene čiji tvorac je Guido van Rosum (engl. *Guido van Rossum*) [20]. Krajem 1980-ih je koncipiran kao naslednik jezika **ABC** [22], a prvi put je objavljen 1991. godine. Filozofija dizajna jezika Python naglašava čitljivost koda. Njegove jezičke konstrukcije i objektno-orijentisani pristup imaju za cilj da pomognu programerima da napišu jasan i logičan kod za male i velike projekte. Python je dinamički tipiziran jezik i poseduje sistem za prikupljanje smeća (engl. *garbage collector*). Podržava više paradigmi programiranja uključujući proceduralno, objektno-orijentisano i funkcionalno programiranje. Python interpreteri su dostupni za mnoge operativne sisteme. Globalna zajednica programera razvija i održava referentnu implementaciju otvorenog koda **CPython**. Neprofitna organizacija *The Python Software Foundation* upravlja i usmerava resursima za razvoj jezika Python i CPython. Jedna od osobina koje je Elixir nasledio od Python-a je podrška za dokumentaciju u vidu dokumentacionih stringova (engl. *docstrings*) koji omogućavaju povezivanje dokumentacije sa modulima, funkcijama, klasama, metodama.

Haskell

Haskell je čisto funkcionalni jezik koji je statički tipiziran [19]. Nazvan je po Haskellu Brooksu Kariju (engl. *Haskell Brooks Curry*), čiji rad u oblasti matematičke logike služi kao osnova za sve funkcionalne jezike. Haskell je zasnovan na lambda računu, pa se stoga lambda koristi kao logo jezika. Nudi kratak, jasan i održiv kod, mali procenat grešaka i veliku pouzdanost. Stoga je pogodan za pisanje velikih softverskih sistema, jer njihovo održavanje čini lakšim i jeftinijim. Jedna od karakteristika koje je Elixir preuzeo od ovog jezika je lenjo izračunavanje.

Ruby

Ruby je dinamički tipiziran programski jezik otvorenog koda nastao 1995. godine. Fokus kod ovog programskog jezika je na jednostavnosti i produktivnosti. Ruby ima elegantnu sintaksu koja je prirodna za čitanje i lako pisanje. Ruby je interpretirani programski jezik, što znači da se izvorni kôd prevodi u kôd razumljiv kompjuteru prilikom svakog izvršavanja programa. Interpretirani programski jezici su sporiji od kompajliranih, ali su fleksibilniji i potrebno je kraće vreme za izradu programa. Međutim, sve više iskusnih Ruby programera se okreće Elixir-u. Zapravo, Elixir je prvi jezik nakon Ruby-ja koji zaista brine o lepoti koda i korisničkom iskustvu vezanom za jezik, biblioteke i ekosistem.

Ruby je imao veliki uticaj na sintaksu programskog jezika Elixir. Na slici 2.2 se nalaze delovi koda napisani u jeziku Ruby i jeziku Elixir, koji imaju dosta sličnosti, a čiji je rezultat izvršavanja isti - dva stringa su nadovezana. U jeziku Ruby se definiše klasa *Concat* koja ima polje *value*, funkciju *initialize* koja se poziva pri kreiranju objekta klase radi inicijalizacije polja *value* i funkciju *join* koja vrši nadovezivanje dva stringa. U Elixir-u se umesto klase definiše modul *Concat* koji sadrži samo funkciju *join* koja vrši nadovezivanje dva stringa. Uočavaju se sličnosti u sintaksi koje su ilustrovane ovim primerom pri definisanju klase/modula, funkcija (ključne reči *def* i *end*), zatim pri nadovezivanju stringova (operatori *+* i *<>*) i pozivanju funkcija (ime klase/modula za kojim sledi tačka).

Clojure

Clojure je dinamički tipiziran programski jezik opšte namene nastao 2007. godine. Njegov tvorac je Rič Hiki (engl. *Rich Hickey*). Clojure kombinuje pristupačnost i interaktivni razvoj skriptnog jezika sa efikasnom i robusnom infrastrukturom

Ruby	Elixir
<pre>class Concat attr_reader :value def initialize(value) @value = value end def join(another_value) value + another_value end end irb> Concat.new("Yo").join("!!!") irb> => "Yo!!!"</pre>	<pre>defmodule Concat do def join(value, another_value) do value <> another_value end end iex> Concat.join("Yo", "!!!") iex> "Yo!!!"</pre>

Slika 2.2: Sintaksa jezika Ruby i Elixir

za višenitno programiranje. On je kompajlirani jezik, ali je i dalje potpuno dinamički tipiziran - svaka funkcija koju podržava Clojure je podržana u toku izvršavanja. Predstavlja dijalekt Lisp-a² i deli njegovu filozofiju *code-as-data* (program je funkcija koja se izvršava nad podacima) i moćan makro sistem. Clojure je pretežno funkcionalni programski jezik i sadrži bogat skup nepromenljivih i postojanih struktura podataka. Elixir je preuzeo neke od najboljih Clojure karakteristika - efikasne, nepromenljive strukture podataka, opcionalno lenjo izračunavanje i protokole.

2.2 Uputstvo za instalaciju i pokretanje Elixir-a

Za instaliranje Elixir-a je potrebno posetiti oficijalni sajt jezika Elixir [13] i pratiti uputstva za odgovarajući operativni sistem. Kada se instalacija završi, može se proveriti verzija Elixir-a pomoću komande `elixir -version` izvršene u komandnoj liniji.

Elixir programi imaju ekstenziju `.ex` ili `.exs` i mogu se pisati i izvršavati u okviru odgovarajućih integrisanih razvojnih okruženja. Neki od najpoznatijih su *Space-macs*, *Visual Studio Code*, *Emacs* i *Atom*, a čitav spisak, kao i link ka njihovim zvaničnim stranicama, može se videti na adresi [14].

²Lisp je programski jezik zasnovan na matematičkoj teoriji rekurzivnih funkcija (u kojoj se funkcija pojavljuje u sopstvenoj definiciji), a Lisp program je funkcija koja se primenjuje na podatke. Ime LISP je nastalo od „LIST Processor”, a povezane liste su jedan od glavnih tipova podataka. Osnova Lisp-a je funkcionalno programiranje, ali se Lisp zbog raznih drugih svojstava smatra multiparadigmatiskim programskim jezikom.

Pojedinačni izrazi, ali i Elixir programi, mogu se takođe izvršavati u komandnoj liniji nakon izvršavanja komande *iex* i pokretanja interaktivnog Elixir-a. Primer pokretanja interaktivnog Elixir-a i izvršavanja pojedinačnog izraza može se videti na listingu 2.1.

```
1 Microsoft Windows [Version 6.1.7601]
2 Copyright (c) 2009 Microsoft Corporation. All rights reserved.
3
4 C:\Windows\system32>iex
5 Interactive Elixir (1.8.0) - press Ctrl+C to exit (type h() for
  help)
6 iex(1)> 2 + 2
7 4
```

Listing 2.1: Pokretanje interaktivnog Elixir-a

U Elixir-u se sve funkcije definišu u okviru modula pomoću ključne reči *defmodule*. Za kompajliranje Elixir fajla u kome je definisan modul sa funkcijama i pokretanje interaktivnog Elixir-a nakon toga, najpre se poziva komanda *iex ime_fajla.ex*. Potom se može pokrenuti bilo koja funkcija unutar definisanog modula tako što se prvo navede ime modula, zatim ime funkcije i argumenti u okviru zagrada (*ime_modula.ime_programa(argument1, argument2)*). Primer definisanja modula i funkcije u okviru njega može se videti na listingu 2.2, a primer kompajliranja fajla i pokretanja funkcije dat je na listingu 2.3.

```
1 defmodule Example do
2   def print_args(arg1, arg2) do
3     IO.puts(arg1 <> " " <> arg2)
4   end
5 end
```

Listing 2.2: Primer definisanja modula

```
1 C:\Windows\system32>iex print_args.ex
2 Interactive Elixir (1.8.0) - press Ctrl+C to exit (type h() for
  help)
3 iex(1)>Example.print_args("Hello", "world")
4 Hello world
5 :ok
```

Listing 2.3: Primer kompajliranja fajla i pokretanja programa

Elixir programi koji imaju ekstenziju *.exs* se mogu kompajlirati i izvršavati iz komandne linije pomoću komande *elixir ime_programa.exs*. Nakon ove komande

neće biti pokrenut interaktivni Elixir, već će samo biti prikazani rezultati izvršavanja programa.

2.3 Osnovne karakteristike

Hosé Valim je tokom 2010. godine bio zaposlen u kompaniji *Platformatec* [17] i radio je na poboljšanju performansi okruženja *Ruby on Rails* na višejezgarnim sistemima. Shvatio je da Ruby nije bio dovoljno dobro dizajniran da reši problem konkurentnosti, pa je započeo istraživanje drugih tehnologija koje bi bile prihvatljivije. Tako je otkrio Erlang i upravo ga je interesovanje prema virtuelnoj mašini jezika Erlang podstaklo da započne pisanje jezika Elixir. Uticaj projekta na kome je do tada radio odrazio se na to da Elixir ima sintaksu koja je nalik na sintaksu jezika Ruby. Ovaj jezik se pokazao veoma dobro pri upravljanju milionima simultanih konekcija: u 2015. je zabeleženo upravljanje nad 2 miliona *WebSocket* konekcija, dok je u 2017. za skalirani Elixir zabeležena obrada 5 miliona istovremenih korisnika. Elixir se danas koristi u velikim kompanijama, kao što su *Discord* i *Pinterest* [8].

Elixir je dinamički tipiziran, funkcionalni programski jezik koji se pokreće na virtuelnoj mašini jezika Erlang, pa samim tim i nasleđuje pogodna svojstva koje dolaze sa ovim okruženjem kao što su **konkurentnost** i **tolerisanje grešaka** [12]. Elixir je nadomestio mnoge koncepte koji su nedostajali jeziku Erlang. Neki od njih su **metaprogramiranje**³, **polimorfizam**, **makroi** i **podrška za alate**. Elixir poseduje podrazumevano okruženje, takozvani **Kernel**, koji obezbeđuje podršku sa osnovne tipove i funkcionalnosti jezika.

U ovom delu će biti opisani osnovni tipovi jezika Elixir, njegove osobine, osnove njegove sintakse, semantike, kao i podrška za osnovne koncepte funkcionalnih jezika poput poklapanja obrazaca i nepromenljivosti podataka.

2.4 Osnovni tipovi podataka

Elixir ima svoje ugrađene (primitivne) tipove. To su:

1. Atomi
2. Celi brojevi

³Tehnika koja omogućava da programi posmatraju druge programe kao svoje podatke i na taj način čitaju i modifikuju i njihov i svoj kôd u vreme izvršavanja.

3. Brojevi u pokretnom zarezu
4. Portovi
5. Ugrađene torke
6. Liste
7. Mape
8. Funkcije
9. Niske bitova
10. Reference

Svaki od ovih tipova, osim poslednja dva, ima odgovarajuće module koji sadrže funkcije koje se koriste za operacije nad tim tipom. Oni predstavljaju omotač oko primitivnog tipa koji nam omogućava korišćenje dodatnih funkcionalnosti nad njim. U nastavku će biti opisani neki od osnovnih tipova.

Atomi

Atomi su konstante ili simboli, pri čemu njihovo ime predstavlja njihovu vrednost. Počinju dvotačkom (`:`) i mogu sadržati slova, cifre, simbole `_`, `@`. Mogu se završavati sa `!` i `?`. Atomi se mogu naći svuda u Elixir-u. U *listama ključnih reči* koje će biti opisane u odeljku [2.4](#), predstavljaju prvu vrednost elementa liste i često se koriste da označe uspeh (`:ok`) ili grešku (`:error`).

Celi brojevi

Celi brojevi su slični kao i u većini programskih jezika i mogu biti dekadni, heksadekadni, oktalni i binarni. Karakter `_` se može koristiti za odvajanje blokova cifara. Veoma značajna stvar je da ne postoji fiksna veličina za čuvanje celih brojeva u memoriji, već interna reprezentacija raste kako bi broj mogao biti smešten u potpunosti.

Brojevi u pokretnom zarezu

Brojevi u pokretnom zarezu se zapisuju uz pomoć decimalne tačke po standardu *IEEE 754*. Pre i posle decimalne tačke mora biti najmanje jedna cifra (1.0, 0.2456). Može se koristiti i notacija koja obuhvata navođenje eksponenata (0.314159e1, 314159.0e - 5).

Liste

Liste se čuvaju u memoriji kao povezane liste, što znači da svaki element u listi čuva svoju vrednost i ukazuje na sledeći element sve dok se ne dostigne kraj liste. To znači da je pristup proizvoljnom elementu liste kao i određivanje dužine liste linearna operacija, jer je potrebno da prođemo celu listu da bismo odredili njenu dužinu. Slično, performanse spajanja dve liste zavise od dužine one koja se nalazi sa leve strane.

Elixir koristi uglaste zagrade (`[]`) da označi listu vrednosti. Vredosti mogu biti bilo kog tipa, a primer liste sa vrednostima različitih tipova prikazan je na listingu [2.4](#).

```
1 iex(1)>[1, 2, true, 3]
2 [1, 2, true, 3]
3 iex(2)>length([1, 2, true, 3])
4 4
```

Listing 2.4: Primer liste

Nadovezivanje ili oduzimanje 2 liste korišćenjem operatora `++` i `--` prikazano je na listingu 2.5.

```
1 iex(1)>[1, 2, 3] ++ [4, 5, 6]
2 [1, 2, 3, 4, 5, 6]
3 iex(2)>[1, false, 2, true, 3, false] -- [true, false]
4 [1, 2, 3, false]
```

Listing 2.5: Nadovezivanje i oduzimanje dve liste

Operatori liste nikada ne menjaju postojeću listu. Rezultat povezivanja listi ili uklanjanja elemenata iz liste je uvek nova lista, jer su strukture podataka u Elixir-u nepromenljive. Jedna od prednosti nepromenljivosti je jasniji kod. Omogućeno je slobodno prosleđivanje podatka sa garancijom da neće biti izmenjeni u memoriji.

Lista može biti prazna ili se može sastojati od **glave** i **repa**. Glava je prvi element liste, a rep je ostatak liste. Oni se mogu izdvojiti pomoću funkcija `hd/1` i `tl/1`. Dodeljivanje liste promenljivoj, dohvaćanje njene glave i repa prikazano je na listingu 2.6. Izdvajanje glave ili repa prazne liste rezultuje greškom.

```
1 iex(1)> lista = [1, 2, 3, 4]
2 [1, 2, 3, 4]
3 iex(2)>hd(lista)
4 1
5 iex(3)>tl(lista)
6 [2, 3, 4]
```

Listing 2.6: Izdvajanje glave i repa liste

Prilikom kreiranja liste, ukoliko Elixir vidi listu *ASCII* brojeva, ispisaće listu znakova. Liste znakova su uobičajene kada se povezuju sa postojećim Erlang kodom. Primer koda koji ilustruje ovo prikazan je na listingu 2.7.

```
1 iex(1)>[11, 12, 13]
2 '\v\f\r'
3 iex(2)>[104, 101, 108, 108, 111]
4 'hello'
```

Listing 2.7: Lista vrednosti pod jednostrukim navodnicima

Preuzimanje informacija o tipu neke vrednosti može se izvršiti pomoću funkcije `i/1` i može se videti na listingu 2.8.

```
1 iex(1)>i 'hello'
2 Term
3 'hello'
4 Data type
5 List
6 Description
7 This is a list of integers that is printed as sequence of
  characters delimited by single quotes because all the integers
  in it represent valid ASCII characters. Conventionally, such
  lists of integers are referred to as "charlists" (more precisely
  , a charlist is a list of Unicode codepoints, and ASCII is a
  subset of Unicode).
8 Raw representation
9 [104, 101, 108, 108, 111]
10 Reference modules
11 List
12 Implemented protocols
13 Collectable, Enumerable, IEx.Info, Inspect, List.Chars, String.
  Chars
```

Listing 2.8: Preuzimanje informacija o tipu vrednosti

Reprezentacije sa jednostrukim i dvostrukim navodnicima u Elixir-u nisu ekvivalentne i predstavljaju različite tipove. Primer se može videti na listingu 2.9.

```
1 iex(1)>'hello' == "hello"
2 false
```

Listing 2.9: Dva različita tipa

Torke

Torke se u Elixir-u definišu pomoću vitičastih zagrada `{}`. Kao i liste, mogu sadržati vrednosti bilo kog tipa. Primer torke sa vrednostima različitih tipova i određivanjem njene dužine prikazan je na listingu 2.10.

```
1 iex(1)>{:ok, "hello", 1}
2 {:ok, "hello", 1}
3 iex(2)>tuple_size({:ok, "hello", 1})
4 3
```

Listing 2.10: Primer torke i određivanje njene dužine

Torke su strukture fiksne dužine koje bi trebalo da sadrže svega nekoliko elemenata koji su zapisani u memoriji jedan za drugim. To znači da se pristup elementu torke ili određivanje dužine torke izvršava u konstantnom vremenu. Razlika u odnosu na liste je u semantici upotrebe. Liste se koriste kada se manipuliše kolekcijom, dok se torke, zbog brzine pristupa njihovim elementima, uglavnom koriste za smeštanje povratne vrednosti funkcije. Na primer, *File.read/1* je funkcija koja se može koristiti za čitanje sadržaja datoteke. Ako putanja do fajla postoji, povratna vrednost funkcije je torka sa prvim elementom koji je atom *:ok* i drugim elementom koji je sadržaj datog fajla. U suprotnom, povratna vrednost funkcije će biti torka gde je prvi element atom *:error*, a drugi element opis greške. Primer upotrebe ove funkcije može se videti na listingu 2.11.

```
1 iex(1)>File.read("C:\elixir\text_document.txt")
2 {:ok, "Hello, world!"}
```

Listing 2.11: Primer upotrebe funkcije *File.read/1*

Indeksi torke počinju od nule, a primer izdvajanje elementa sa indeksom 1 može se videti na listingu 2.12.

```
1 iex(1)>tuple = {:ok, "hello", 1}
2 {:ok, "hello", 1}
3 iex(2)>elem(tuple, 1)
4 "hello"
```

Listing 2.12: Izdvajanje elementa torke sa indeksom 1

Umetanje novog elementa na određeno mesto u torki vrši se pomoću funkcije *put_elem/3*. Ona vraća novu torku, dok originalna torka ostaje neizmenjena. Primer koda koji ilustruje upotrebu ove funkcije prikazan je na listingu 2.13.

```
1 iex(1)>tuple = {:ok, "hello", 1}
2 {:ok, "hello", 1}
3 iex(2)>put_elem(tuple, 1, "world")
4 {:ok, "world", 1}
5 iex(3)>tuple
6 {:ok, "hello", 1}
```

Listing 2.13: Umetanje novog elementa u torku

Kao i liste, torke su takođe nepromenljive. Svaka operacija nad torkom vraća novu torku i nikada ne menja postojeću. Ova operacija, kao i operacija ažuriranja torke je skupa, jer zahteva kreiranje nove torke u memoriji. Ovo se odnosi samo na samu torku, a ne na njen sadržaj. Na primer, prilikom ažuriranja torke, svi unosi se

dele između stare i nove torke, osim unosa koji je izmenjen. Drugim rečima, torke i liste u Elixir-u mogu da dele svoj sadržaj, što smanjuje količinu memorije koju jezik treba da zauzme. Ove karakteristike performansi diktiraju upotrebu struktura podataka.

Liste ključnih reči i mape

Elixir podržava asocijativne strukture podataka. Asocijativne strukture podataka su one koje su u stanju da pridruže određenu vrednost ili više vrednosti ključu. Dve glavne strukture među njima su **liste ključnih reči** i **mape**.

Liste ključnih reči

U mnogim funkcionalnim programskim jezicima, uobičajeno je da se koristi lista dvočlanih torki za predstavljanje strukture podataka ključ - vrednost. Lista torki gde je prvi element torke atom (tj. ključ) u Elixir-u se naziva **lista ključnih reči**. Elixir podržava posebnu sintaksu za definisanje takvih lista: `[key : value]`. Primer oba načina definisanja prikazan je na listingu 2.14.

```
1 iex(1)> lista = [{:a, 1}, {:b, 2}, {:c, 3}]
2 [a: 1, b: 2, c: 3]
3 iex(2)> lista == [a: 1, b: 2, c: 3]
4 true
```

Listing 2.14: Primer liste ključnih reči

Kako su liste ključnih reči liste, nad njima možemo primenjivati sve operacije dostupne nad listama. Na primer, korišćenjem operatora `++` može se izvršiti dodavanje nove vrednosti listi ključnih reči. Primer koda koji ilustruje ovo dodavanje dat je na listingu 2.15.

```
1 iex(3)> lista ++ [d: 4]
2 [a: 1, b: 2, c: 3, d: 4]
3 iex(4)> [a: 0] ++ lista
4 [a: 0, a: 1, b: 2, c: 3]
```

Listing 2.15: Dodavanje nove vrednosti listi ključnih reči

Elementima liste ključnih reči se pristupa na način prikazan na listingu 2.16.

```
1 iex(1)>lista = [a: 0, a: 1, b: 2, c: 3]
2 [a: 0, a: 1, b: 2, c: 3]
3 iex(2)>lista[:a]
4 0
```

Listing 2.16: Pristup elementu liste ključnih reči

Liste ključnih reči su važne, jer imaju tri posebne karakteristike:

1. Ključevi moraju biti atomi.
2. Ključevi su uredjeni, onako kako je navedeno od strane programera.
3. Ključevi se mogu ponavljati.

Elixir obezbeđuje modul koji omogućava manipulisanje listama ključnih reči. Liste ključnih reči su jednostavno liste, i kao takve pružaju iste karakteristike linearnih performansi kao i liste. Što je lista duža, više vremena će biti potrebno za pronalaženje ključa, prebrojavanje elemenata i tako dalje. Iz tog razloga, liste ključnih reči se u Elixir-u koriste uglavnom za prosleđivanje opcionih vrednosti. Za čuvanje mnogo elemenata ili garantovanje pojavljivanja jednog ključa sa maksimalno jednom vrednošću treba koristiti mape.

Mape

Mapa je kolekcija koja sadrži parove ključ : vrednost. Glavne razlike između liste parova ključ-vrednost i mape su u tome što mape ne dozvoljavaju ponavljanje ključeva (jer su to asocijativne strukture podataka) i što ključevi mogu biti bilo kog tipa. Mapa je veoma efikasna struktura podataka, naročito kada količina podataka raste. Ukoliko želimo da podaci u kolekciji ostanu baš u onom redosledu u kom smo ih naveli inicijalno, onda je bolje koristiti liste parova ključ : vrednost, jer mape ne prate nikakvo uređenje.

Mapa se definiše pomoću sintakse `%{}` na način prikazan na listingu 2.17.

```
1 iex(1)> mapa = %{a => 1, 2 => :b}
2 %{a => 1, 2 => :b}
3 iex(2)>mapa[:a]
4 1
5 iex(3)>mapa[2]
6 :b
7 iex(4)>mapa[:c]
```

```
8 nil
```

Listing 2.17: Primer mape i pristupa njenim elementima

Modul **Map** obezbeđuje razne funkcije za manipulaciju mapama, a neke od njih mogu se videti na listingu 2.18.

```
1 iex(1)>Map.get(%{:a => 1, 2 => :b}, :a)
2 1
3 iex(2)>Map.put(%{:a => 1, 2 => :b}, :c, 3)
4 %{2 => :b, :a => 1, :c => 3}
5 iex(3)>Map.to_list(%{:a => 1, 2 => :b})
6 [{2 => :b}, {:a => 1}]
```

Listing 2.18: Neke od funkcija modula Map

Mape imaju sintaksu za ažuriranje vrednosti ključa prikazanu na listingu 2.19

```
1 iex(1)>mapa = %{:a => 1, 2 => :b}
2 %{2 => :b, :a => 1}
3 iex(2)>%{mapa | 2 => "dva"}
4 %{2 => "dva", :a => 1}
5 iex(3)>%{mapa | :c => 3}
6 ** (KeyError) key :c not found in: %{2 => :b, :a => 1}
7 (stdlib) :maps.update(:c, 3, %{2 => :b, :a => 1})
8 (stdlib) erl_eval.erl:259: anonymous fn/2 in :erl_eval.expr/5
9 (stdlib) lists.erl:1263: :lists.foldl/3
```

Listing 2.19: Ažuriranje vrednosti ključa

Prethodno prikazana sintaksa zahteva da dati ključ postoji u mapi i ne može se koristiti za dodavanje novih ključeva. Na primer, korišćenje ove sintakse za ključ `:c` nije uspeo, jer ključ `:c` ne postoji u mapi.

Ukoliko su svi ključevi u mapi atomi, onda se radi pogodnosti može koristiti sintaksa ključnih reči data listingom 2.20.

```
1 iex(1)> mapa = %{a: 1, b: 2}
2 %{a: 1, b: 2}
```

Listing 2.20: Sintaksa ključnih reči

Još jedno zanimljivo svojstvo mapa je to što obezbeđuju sopstvenu sintaksu za pristup atomskim ključevima. Primer ove sintakse možemo videti na listingu 2.21.

```
1 iex(1)>mapa = %{:a => 1, 2 => :b}
2 %{2 => :b, :a => 1}
3 iex(2)>mapa.a
4 1
5 iex(3)>mapa.c
6 ** (KeyError) key :c not found in: %{2 => :b, :a => 1}
```

Listing 2.21: Sintaksa za pristup atomskim ključevima

Programeri koji programiraju u Elixir-u pri radu sa mapama češće koriste *map.field* sintaksu i poklapanje obrazaca nego funkcije iz modula Map, jer dovode do asertivnog stila programiranja.

Često se koriste mape unutar mapa ili čak liste ključnih reči unutar mapa. Elixir obezbeđuje pogodnosti za manipulisanje ugnježđenim strukturama podataka poput *put_in/2*, *update_in/2* i drugih naredbi koje daju iste pogodnosti koje se mogu pronaći u imperativnim jezicima, a da pritom zadrže svojstvo nepromenljivosti podataka (svojstvo koje će detaljnije biti objašnjeno u poglavlju 2.7).

Na listingu 2.22 je prikazana lista ključnih reči korisnika, gde je svaka vrednost mapa koja sadrži ime, starost i listu programskih jezika koje svaki korisnik voli.

```
1 iex(1)>users = [john: %{name: "John", age: 27, languages:
2 ["Erlang", "Ruby", "Elixir"]}, mary: %{name: "Mary", age: 29,
3 languages: ["Elixir", "F#", "Clojure"]}]
4 [
5   john:%{age:27,languages:["Erlang","Ruby","Elixir"],name:"John"},
6   mary:%{age:29,languages:["Elixir","F#","Clojure"],name:"Mary"}
7 ]
```

Listing 2.22: Struktura koja predstavlja listu korisnika

Pristup Džonovim godinama mogao bi se izvršiti na način prikazan na listingu 2.23.

```
1 iex(2)>users[:john].age
2 27
```

Listing 2.23: Pristup godinama od Džona

Ista sintaksa se može koristiti i za ažuriranje vrednosti i prikazana je na listingu 2.24.

```
1 iex(3)> users = put_in users[:john].age, 31
2 [
3   john:%{age:31,languages:["Erlang","Ruby","Elixir"],name:"John"},
4   mary:%{age:29,languages:["Elixir","F#","Clojure"],name:"Mary"}
5 ]
```


5]

Listing 2.24: Ažuriranje vrednosti

Makro `update_in/2` je sličan, ali daje mogućnost prosleđivanja funkcije koja kontroliše kako se vrednost menja. Na primer, uklanjanje programskog jezika Clojure sa Marijinog spiska jezika može se uraditi na način prikazan listingom 2.25.

```
1 iex(4)> users = update_in users[:mary].languages ,
2 fn languages -> List.delete(languages, "Clojure") end
3 [
4   john:%{age:31, languages:["Erlang","Ruby","Elixir"],name:"John"},
5   mary:%{age:29, languages:["Elixir","F#"],name:"Mary"}
6 ]
```

Listing 2.25: Brisanje jezika iz liste

Funkcija `get_and_update_in` omogućava izvlačenje vrednosti i ažuriranje strukture podataka odjednom, a funkcije `put_in/3`, `update_in/3` i `get_and_update_in/3` omogućavaju dinamički pristup strukturama podataka.

2.5 Osnovni operatori

Pored osnovnih aritmetičkih operatora `+`, `-`, `*`, `/`, kao i funkcija `div/2` i `rem/2` za celobrojno deljenje i ostatak pri celobrojnem deljenju, Elixir podržava i već pomenute operatore `++` i `--` za nadovezivanje i oduzimanje listi, kao i operator `<>` koji se koristi za nadovezivanje stringova.

Elixir obezbeđuje tri logička operatora: **and**, **or** i **not**. Oni su striktni u smislu da očekuju nesto što ima vrednost *true* ili *false* kao svoj prvi operand. Primer koda koji ilustruje ovu osobinu prikazan je na listingu 2.26.

```
1 iex(1) true and true
2 true
3 iex(2)> false or is_atom(:example)
4 true
```

Listing 2.26: Primer upotrebe logičkih operatora

Ukoliko kao prvi operand prosledimo nesto čija vrednost nije logička, rezultat je greška kao na listingu 2.27.

```
1 iex(1)> 1 and true
2 ** (BadBooleanError) expected a boolean on left-side of "and",
```

```
3 got: 1
```

Listing 2.27: Greška pri upotrebi logičkog operatora

And i *or* su lenji operatori, jer desni operand izračunavaju samo u slučaju da levi nije dovoljan za određivanje rezultata.

Pored ovih logičkih operatora, Elixir takođe obezbeđuje operatore `||`, `&&` i `!` koji prihvataju argumente bilo kog tipa. Sve vrednosti osim **false** i **nil** će biti procenjene na *true*, što se može videti na primeru prikazanom listingom 2.28.

```
1 iex(1)>1 || true
2 1
3 iex(2)>false || 11
4 11
5 iex(3)>nil && 13
6 nil
7 iex(4)true && 17
8 17
9 iex(5)>!true
10 false
11 iex(6)!1
12 false
13 iex(7)>!nil
14 true
```

Listing 2.28: Operatori koji prihvataju argumente bilo kog tipa

Pravilo je da kada se očekuju logičke vrednosti, treba koristiti operatore *and* i *or*, a ako bilo koji od operandada ima vrednost koja nije logička, onda treba koristi `||`, `&&` i `!`.

Elixir takođe obezbeđuje `==`, `!=`, `===`, `!==`, `<=`, `>=`, `<` i `>` kao operatore poređenja, pri čemu se operator `===` od operatora `==` razlikuje po tome što pored vrednosti poredi i tip.

Moguće je i poređenje tipova među sobom. Razlog zbog kojeg se mogu uporediti različiti tipovi podataka je pragmatizam. Algoritmi sortiranja ne moraju da brinu o različitim tipovima podataka da bi sortirali. Ukupan redosled sortiranja je definisan na način prikazan na listingu 2.29.

```
1 number < atom < reference < function < port < pid < tuple
2 < map < list < bitstring
```

Listing 2.29: Poređenje tipova

2.6 Poklapanje obrazaca

Poklapanje obrazaca je proveravanje da li se u datoj sekvenci tokena može prepoznati neki obrazac. Ovaj koncept će biti jasniji na praktičnom primeru operatora `=`. U većini programskih jezika, operator `=` je operator dodele koji levoj strani dodeljuje vrednost izraza na desnoj. U Elixir-u se ovaj operator naziva **operator uparivanja** (engl. *matching*). On se uspešno izvršava, ako pronade način da izjednači levu stranu (svoj prvi operand) sa desnom (drugi operand).

Na primer, izraz `2 + 2 = 5` bi rezultirao greškom datom na listingu [2.30](#).

```
1 iex(1) 5 = 2 + 2
2 ** (MatchError) no match of right hand side value: 4
```

Listing 2.30: Operator uparivanja

Na osnovu greške se može zaključiti da `2 + 2` zaista nije 5. U Elixir-u leva strana mora da ima istu vrednost kao i desna strana. Vrednost izraza dat listingom [2.31](#) nije greška, već uspešno poklapanje obrazaca:

```
1 iex(1) 4 = 2 + 2
2 4
```

Listing 2.31: Uspešno poklapanje obrazaca

Slično, dva identična stringa sa obe strane znaka jednakosti će dati rezultat prikazan listingom [2.32](#).

```
1 iex(1) > "pas" = = "pas"
2 "pas"
```

Listing 2.32: Uspešno poklapanje obrazaca sa stringovima

Poklapanje obrazaca se može prikazati i na primeru sa listama. Neka je data lista osoba koja je prikazana listingom [2.33](#).

```
1 iex(1) > lista = ["Milan Stamenkovic", "Petar Jovanovic",
2 "Milica Lazarevic", "Lena Markovic"]
3 ["Milan Stamenkovic", "Petar Jovanovic", "Milica Lazarevic",
4 "Lena Markovic"]
```

Listing 2.33: Lista osoba

Neka prve tri osobe treba da budu zapamćene. U te svrhe se može iskoristiti poklapanje obrazaca dato na listingu [2.34](#).

```
1 iex(2)>[prvi, drugi, treci | ostali] = lista
2 ["Milan Stamenkovic", "Petar Jovanovic", "Milica Lazarevic",
3  "Lena Markovic"]
```

Listing 2.34: Poklapanje obrazaca sa listama

Izvršeno je dodeljivanje prve, druge i treće stavke iz liste promenljivama *prvi*, *drugi* i *treci*. Ostatak liste je dodeljen promenljivoj *ostali* pomoću **pipe operatora** (`|`). Vrednost svake od ovih promenljivih može se iščitati na način prikazan na listingu 2.35.

```
1 iex(3)>prvi
2 "Milan Stamenkovic"
3 iex(2)>drugi
4 "Petar Jovanovic"
5 iex(3)>treci
6 "Milica Lazarevic"
7 iex(4)>ostali
8 ["Lena Markovic"]
```

Listing 2.35: Iščitavanje sadržaja promenljivih

Mape su vrlo korisne kod poklapanja obrazaca. Kada se koristi u poklapanju obrazaca, mapa će se uvek podudarati sa poskupom date vrednosti kao što se može videti na listingu 2.36.

```
1 iex(1)> %{ } = %{ :a => 1, 2 => :b }
2 %{ :a => 1, 2 => :b }
3 iex(2)>%{ :a => a } = %{ :a => 1, 2 => :b }
4 %{ 2 => :b, :a => 1 }
5 iex(3)>a
6 1
7 iex(4)>%{ :c => c } = %{ :a => 1, 2 => :b }
8 ** (MatchError) no match of right hand side value: %{ 1 => :b, :a =>
9    1 }
10 (stdlib) erl_eval.erl:453: :erl_eval.expr/5
11 (iex) lib/iex/evaluator.ex:257: IEx.Evaluator.handle_eval/5
12 (iex) lib/iex/evaluator.ex:237: IEx.Evaluator.do_eval/3
13 (iex) lib/iex/evaluator.ex:215: IEx.Evaluator.eval/3
14 (iex) lib/iex/evaluator.ex:103: IEx.Evaluator.loop/1
15 (iex) lib/iex/evaluator.ex:27: IEx.Evaluator.init/4
```

Listing 2.36: Mape pri poklapanju obrazaca

Mapa se podudara sve dok ključevi u obrascu postoje u datoj mapi. Tako, prazna mapa odgovara svim mapama.

Promenljive se mogu koristiti prilikom pristupa, podudaranja i dodavanja ključeva mape, što je dato listingom 2.37.

```
1 iex(1)>mapa = %{a => :jedan}
2 %{a => :jedan}
3 iex(2)>mapa[a]
4 :jedan
5 iex(3)>%{^a => :jedan} = %{1 => :jedan, 2 => :dva, 3 => :tri}
6 %{1 => :jedan, 2 => :dva, 3 => :tri}
```

Listing 2.37: Upotreba promenljivih u mapama

2.7 Nepromenljivost podataka

U mnogim programskim jezicima je dozvoljeno dodeljivanje vrednosti promenljivoj, a zatim njeno menjanje tokom izvršavanja programa. Mogućnost da zamene vrednost na određenoj memorijskoj lokaciji drugom vrednošću čini se legitimna i čini se da povećava čitljivost našeg programa. Tokom izvršavanja programa obično se ne zna tačno vreme izvršavanje ove promene i obično se i ne vodi računa o tome pri pisanju programa. Ali šta se dešava kada se vrednost u memoriji, ili čak i tip vrednosti, promeni u trenutku kada je koristi više instanci programa? Ovakvo ponašanje je poznato kao **promenljivost (mutabilnost)**. U konkurentnim okruženjima je izvor grešaka koje je veoma teško pratiti i reprodukovati. Promenljivost takođe vodi veoma komplikovanom kodu, koji se često piše ad-hoc kako bi se rešili problemi sinhronizacije.

Umesto toga, drugi jezici, kao što je Erlang, a samim tim i Elixir imaju osobinu **nepromenljivosti (imutabilnosti)**. Oni jednostavno ne dozvoljavaju promenu vrednosti na određenoj memorijskoj lokaciji. Na ovaj način, ako se promenljiva *a* poklopila sa vrednošću 1, onda se njena vrednost sigurno neće menjati tokom izvršavanja programa i ne mora se voditi računa o problemima sinhronizacije u konkurentnom okruženju.

2.8 Odlučivanje

Strukture odlučivanja zahtevaju da programer odredi jedan ili više uslova koje će program proceniti ili testirati zajedno sa naredbom ili naredbama koje treba izvršiti, ako je uslov određen ili tačan, i opciono, druge naredbe koje treba izvršiti, ako je utvrđeno da je uslov netačan.

Elixir obezbeđuje **if/else** uslovne konstrukte kao i mnogi drugi programski jezici. On takođe poseduje naredbu **cond** koja poziva prvu tačnu vrednost koju pronađe. **Case** je još jedan kontrolni tok koji koristi poklapanje obrazaca za kontrolu toka programa.

Elixir ima sledeće vrste naredbi za odlučivanje:

1. *if naredba* - If naredba se sastoji od logičkog izraza praćenog ključnom reči *do*, jedne ili više izvršnih naredbi i na kraju ključne reči *end*
2. *if..else naredba* - If naredba može biti praćena naredbom *else* (unutar *do..end* bloka), koja se izvršava, ako je loički izraz netačan.
3. *unless naredba* - Naredba *unless* ima isto telo kao i *if* naredba. Kod unutar *unless* naredbe se izvršava samo kada je navedeni uslov netačan.
4. *unless..else* - Naredba *unless...else* ima isto telo kao i naredba *if..else*. Kod unutar *unless..else* naredbe se izvršava samo kada je navedeni uslov netačan.
5. *cond* - Naredba *cond* se koristi ukoliko treba izvršiti neki kod na osnovu nekoliko uslova. Radi kao *if..else if..else* kod drugih programskih jezika.
6. *case* - Naredba *case* se može smatrati zamenom za **switch** naredbu u imperativnim programskim jezicima. Naredba *case* uzima promenljivu ili literal i primenjuje odgovarajući obrazac poklapanja u različitim slučajevima. Ako se bilo koji slučaj poklapa, Elixir izvršava kod povezan sa tim slučajem i izlazi iz *case* naredbe.

2.9 Moduli

U Elixir-u se može vršiti grupisanje nekoliko funkcija u module. Već su pomenuti različiti moduli u prethodnim odeljcima (*Map*, *Enum*, *List*, *String*,...). Za kreiranje sopstvenih modula u Elixir-u, koristi se makro **defmodule**, a za definisanje svojih funkcija, koristimo makro **def**. Primer koda koji ilustruje kreiranje modula i funkcija dat je listingom 2.38.

```
1 defmodule Math do
2   def sum(a, b) do
3     a + b
4   end
5 end
```

Listing 2.38: Kreiranje modula i funkcija

Moduli mogu biti ugnježdjeni u Elixir-u. Ova osobina jezika omogućava bolje organizovanje koda. Na listingu 2.39 se može videti definisanje dva modula: **Math** i **Math.Adding**, pri čemu je drugi ugnježđen unutar prvog. Drugom se može pristupati samo pomoću *Adding* unutar *Math* modula sve dok su u istom leksičkom opsegu. Ako se kasnije *Adding* modul premesti izvan definicije *Math* modula, onda se mora referencirati njegovim punim imenom *Math.Adding* ili pseudonim mora biti postavljen pomoću direktive aliasa.

```
1 defmodule Math do
2   defmodule Adding do
3     def sum(a, b) do
4       a + b
5     end
6   end
7 end
```

Listing 2.39: Ugnježdavanje modula

U Elixir-u nema potrebe za definisanjem modula *Math*, kako bi se definisao modul *Math.Adding*, pošto jezik prevodi sva imena modula u atome. Mogu se definisati i proizvoljno ugnježdjeni moduli bez definisanja bilo kog modula u lancu. Na primer, može se definisati modul *Math.Adding.Sum*, iako prethodno nije definisan modul *Math* i *Math.Adding*.

2.10 Direktive

Kako bi se olakšala ponovna upotreba koda, Elixir obezbeđuje tri direktive - **alias**, **require** i **import**, kao i makro **use**. Primer njihove upotrebe može se videti na listingu 2.40.

```
1 #Alias modula tako da se može pozvati sa Adding umesto sa
2 #Math.Adding
3 alias Math.Adding, as: Adding
4
5 #Obezbeđuje da je modul kompajliran i dostupan (obično za makroe)
6 require Math
7
8 #Uključuje prilagodjen kod definisan u Math kao prosirenje
9 use Math
```

Listing 2.40: Primer upotrebe direktiva

Direktiva *alias*

Direktiva *alias* nam služi za podešavanje pseudonima za bilo koje ime modula. Aliasi moraju uvek počinjati velikim slovom. Validni su samo unutar leksičkog opsega u kome su pozvani.

Direktiva *require*

Elixir obezbeđuje makroe kao mehanizam za meta-programiranje (pisanje koda koji generiše kod). Makroi su delovi koda koji se izvršavaju i proširuju tokom kompilacije. To znači da bi se mogao koristiti makro, mora se garantovati da su njegovi moduli i implementacija dostupni tokom kompilacije. Ovo se čini pomoću **require** direktive. Uopšteno, moduli nisu potrebni pre upotrebe, osim ako želimo da koristimo makroe koji su dostupni u njemu. **Require** direktiva je takođe leksički određena.

Direktiva *import*

Direktiva **import** se koristi kako bi se lakše pristupalo funkcijama i makroima iz drugih modula bez upotrebe potpuno kvalifikovanog imena. **Import** direktiva je takođe leksički određena.

Makro *use*

Iako nije direktiva, **use** je makro koji je usko povezan sa zahtevom koji omogućava korišćenje modula u trenutnom kontekstu. Makro *use* se često koristi za unos spoljne funkcionalnosti u trenutni leksički opseg, često modula.

Glava 3

Sekvencioniranje genoma

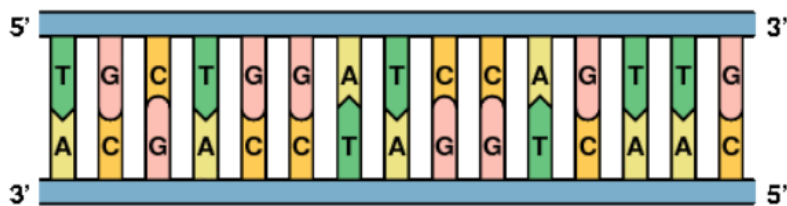
DNK (dezoksiribonukleinska kiselina) je nukleinska kiselina koja sadrži uputstva za razvoj i pravilno funkcionisanje svih živih organizama. Informacije u DNK se čuvaju kao kôd koji čine **četiri hemijske baze**: **adenin** (A), **guanin** (G), **citozin** (C) i **timin** (T). Ljudski DNK se sastoji od oko tri milijarde baza, a više od 99 procenata tih baza je isto kod svih ljudi. Redosled ili sekvenca ovih baza određuje informacije dostupne za izgradnju i održavanje organizma, slično načinu na koji se slova abecede pojavljuju određenim redosledom kako bi se formirale reči i rečenice. DNK baze se spajaju jedna sa drugom, adenin sa timinom i citozin sa guaninom, da bi formirale jedinice koje se nazivaju **bazni parovi**. Svaka baza je takođe vezana za molekul šećera dezoksiriboze i molekul fosfata. Zajedno, baza, šećer i fosfat nazivaju se **nukleotidima**. Nukleotidi su raspoređeni u dva dugačka lanca koji imaju antiparalelnu¹ orijentaciju i koji čine spiralu koja se naziva **dvostruka spirala** [15]. Struktura dvostruke spirale pomalo je nalik na merdevine, pri čemu bazni parovi formiraju lestvičaste trake, a molekuli šećera i fosfata formiraju vertikalne bočne delove merdevina (slika 3.1). Molekul DNK se može zamisliti odmotan i rotiran, tako da su trake merdevina orijentisane vertikalno, a bazni parovi se mogu čitati sa leva na desno (slika 3.2).

Krajevi šećer-fosfatnih lanaca se međusobno razlikuju po prirodi nevezanog atoma ugljenika - jedan kraj ima nevezani 5' (pet prim) atom ugljenika, dok drugi kraj ima nevezani 3' (tri prim) atom ugljenika. 5' ugljenik ima fosfatnu grupu koja je vezana za njega, a 3' ugljenik-hidroksilnu (-OH) grupu. Ova asimetrija daje DNK pravac 5' → 3'. Kada predstavljamo molekul DNK pomoću dijagrama, kakav je prikazan

¹U biohemiji, dva molekula su antiparalelna ako su locirani jedan pored drugog i usmereni u suprotnim pravcima, ili ako su dva lanca međusobno komplementarna.



Slika 3.1: Struktura DNK [15]



Slika 3.2: Odmotan DNK [3]

na slici 3.1, podrazumeva se da gornji lanac kreće od 5' kraja sa leve strane do 3' kraja sa desne strane. Donji lanac je obrnuto orijentisan, od 3' kraja sa leve strane do 5' kraja sa desne strane [3].

Sva živa bića svoj genetički materijal nose u obliku DNK, sa izuzetkom nekih virusa koji imaju ribonukleinsku kiselinu (RNK). DNK ima veoma važnu ulogu ne samo u prenosu genetskih informacija sa jedne na drugu generaciju, već sadrži i uputstva za građenje neophodnih ćelijskih organela, proteina i RNK molekula.

Geni su delovi DNK sekvence koja je sa računarske strane niska nad azbukom $\{A, C, G, T\}$. Svaki gen predstavlja pravilo za sintezu jednog proteina u ćeliji, koji je neophodan za njeno pravilno funkcionisanje. **Genom** je skup gena jednog organizma i sastoji od niza uparenih baza.

Sekvencioniranje genoma podrazumeva otkrivanje sastava genoma. U pitanju je eksperimentalan proces – da bi se saznalo šta se nalazi u sastavu jednog genoma, potreban je uzorak tkiva odgovarajuće vrste [10]. **Sekvencioniranje DNK** je proces određivanja preciznog redosleda nukleotida unutar molekula DNK, a mašine koje određuju redosled nukleotida nazivaju se **sekvenceri**. Sekvencioniranje DNK uključuje bilo koji metod ili tehnologiju koja se koristi za određivanje redosleda četiri baze iz ciljnog genoma (A, C, G, T) [7]. Savremene laboratorijske metode za dati uzorak krvi pacijenta mogu da očitaju podsekvence DNK koje se nazivaju **očitanja** (engl. *reads*), a koje je nakon toga neophodno sastaviti u polaznu DNK sekvencu pomoću posebnih alata za sklapanje, takozvanih **asemblera**. Podniske očitavanja dužine k nazivaju se **k-meri**. Genomi i očitavanja DNK se mere u **baznim parovima (bp)**. Očitavanja mogu biti **kratka** i **duga**. Kratka očitavanja imaju dužinu od 50 bp do 400 bp, a duga očitavanja dužinu veću od 400 bp.

Rekonstrukcija genoma kroz sekvencioniranje DNK je veoma važan problem. Ona se može uporediti sa kompletiranjem slagalice, gde su očitavanja delovi slagalice. Što su delovi veći, slagalicu je lakše sastaviti.

3.1 Istorija sekvencioniranja genoma

Sekvencioniranje Sanger predstavlja prvu tehniku sekvencioniranja. Nastala je 1977. godine, a njeni tvorci su Frederik Sanger (engl. *Frederick Sanger*) i njegove kolege. Razvijena su dva assemblera za assembliranje očitavanja sekvencioniranja Sanger: *OLC*² assembler **Celera** i assembler **Ojler** zasnovan na De Bruijnovim grafovima. **Humani referentni genom** sastavljen je korišćenjem ova dva pristupa. Humani referentni genom je digitalna baza podataka o nukleinskim kiselinama, koju su naučnici prikupili kao reprezentativni primer skupa gena čoveka. Kako su često sastavljeni sekvencioniranjem DNK većeg broja davalaca, referentni genomi ne predstavljaju skup gena nijedne pojedinačne osobe. Humani referentni genom **GRCh37** (*The Genome Reference Consortium human genome (build 37)*) je izveden 2009. godine iz DNK trinaest anonimnih dobrovoljaca iz Bafala (engl. *Buffalo*). Referentni genomi se obično koriste kao uputstvo na osnovu koga se grade novi genomi, što im omogućava da se sastave mnogo brže i ekonomičnije. Međutim, kako je sekvencioniranje Sanger niskopropusno (omogućava sekvencioniranje malog broja očitavanja odjednom) i neekonomično, samo nekoliko genoma je sastavljeno pomoću njega.

Razvoj tehnika za sekvencioniranje **druge generacije** značajno je doprineo efikasnijem i ekonomičnijem sekvencioniranju stotine miliona očitavanja. Međutim, očitavanja druge generacije sekvencioniranja su kratka. Njihova pojava je dovela do većeg broja uspešnih *de novo* assemblera³ projekata, uključujući rekonstrukciju genoma **Džejmsa Votsona** (engl. *James Watson*) i **panda** genoma. Iako je ovaj pristup ekonomičan, rezultat su bili fragmentisani genomi, jer su očitavanja kratka i ponavljajući regioni, takozvani **ponovci**, dugi. Ponorci su fenomen kada se šablon od nekoliko nukleotida pojavljuje više puta u nizu.

Od nedavno su na raspolaganju tehnike za sekvencioniranje **treće generacije**, koje proizvode duga očitavanja (dužine od oko 10000 bp). Duga očitavanja mogu obuhvatiti složene genomske karakteristike, omogućavajući njihovo ispravnije postavljanje u rekonstruisanom genomu, tj. mogu rešiti problem ponavljajućih regiona. Međutim, duga očitavanja imaju visoku stopu greške (15% – 18%). U cilju rešavanja ovog problema razvijen je veliki broj računskih metoda za korekciju grešaka u očitavanjima treće generacije sekvencioniranja.

²Skraćenica od *overlap layout consensus* (*overlap* - izgradnja grafa preklapanja, *layout* - spajanje putanja u grafu u kontige, *consensus* - određivanje najverovatnije sekvence nukleotida za svaku kontigu).

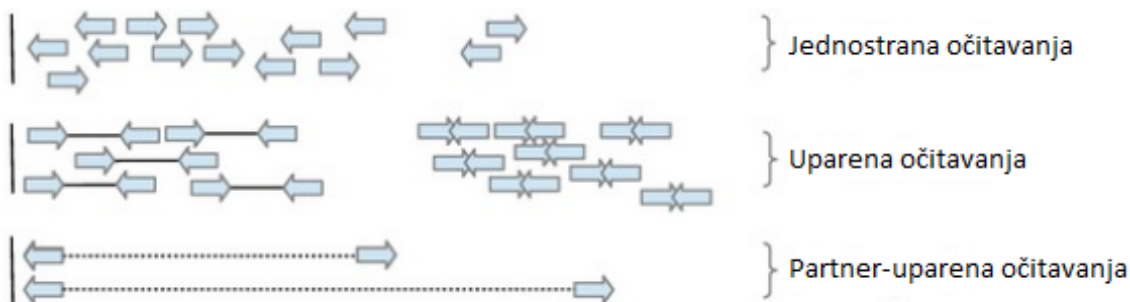
³*De novo* assembleri su programi koji vrše sklapanje tako što proširuju kratka očitavanja spajanjem susednih očitavanja u dužu sekvencu, bez korišćenja referentne sekvence.

3.2 Sekvencioniranje *shotgun* celokupnog genoma

Jedna od tehnika druge generacije sekvencioniranja je sekvencioniranje *shotgun* celokupnog genoma. Prvi korak u ovom procesu je razbijanje genoma na skup očitavanja, a na osnovu njegovog uzorka. Postoje tri vrste očitavanja:

- jednostrana očitavanja (engl. *single-end reads*)
- uparena očitavanja (engl. *paired-end reads*)
- partner-uparena očitavanja (engl. *mate-pair reads*)

Jednostrana očitavanja su očitavanja koja se generišu tako što sekvencer čita DNK fragment samo u jednom smeru, ili sa 5', ili sa 3' kraja. U slučaju uparenih očitavanja, sekvencer čita DNK fragment sa oba kraja. Partner-uparena očitavanja su 5' ili 3' očitavanja DNK fragmenta oba lanca. Prikaz ovih vrsta očitavanja može se videti na slici 3.3. Strelice označavaju smer u kome je sekvencer vršio očitavanje DNK fragmenta.



Slika 3.3: Vrste očitavanja [5]

Sekvenceri čitaju deo DNK fragmenta neke unapred zadate dužine koja se naziva **dužina očitavanja**. **Dubina pokrivanja** neke pozicije u DNK sekvenci je velika, ako je nukleotid na toj poziciji pročitao veliki broj puta u jedinstvenim očitavanjima. Tačnost sekvencioniranja za svaki pojedinačni nukleotid je veoma visoka, ali ukoliko se pojedinačni genom sekvencira samo jednom, zbog veoma velikog broja nukleotida u genomu, doći će do značajnog broja grešaka u sekvencioniranju. Pored toga, mnoge pozicije u genomu sadrže retke **jednonukleotidne polimorfizme**⁴ (engl. *single-nucleotide polymorphisms* - *SNPs*). Stoga, kako bi se napravila razlika između

⁴Pojava zamene mesta jednog nukleotida nekim drugim nukleotidom.

grešaka u sekvencioniranju i pravih *SNP*-ova, potrebno je još više povećati tačnost sekvencioniranjem pojedinačnih genoma veći broj puta [9].

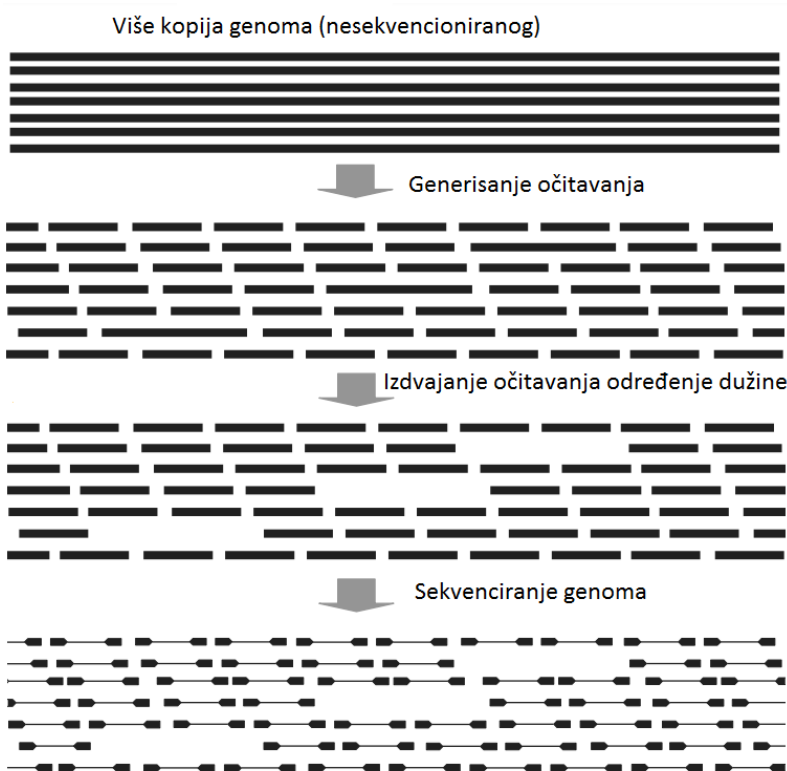
Za potrebe sekvencioniranja *shotgun* celokupnog genoma postoje dva protokola – **sekvencioniranje celokupnog genoma** i **sekvencioniranje partner-uparenih očitavanja**.

Sekvencioniranje celokupnog genoma

Sekvencioniranje celokupnog genoma uključuje tri koraka:

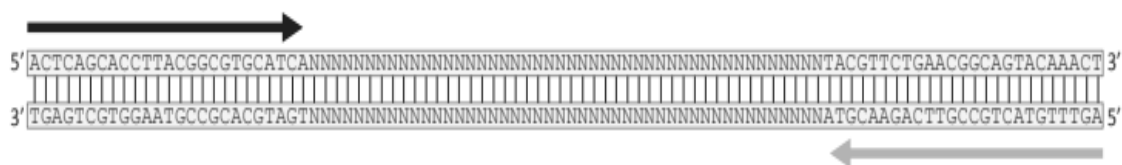
- razbijanje genoma
- izbor dužine očitavanja i izdvajanja fragmenata te dužine
- sekvencioniranje očitavanja

Prvo se uzorak genoma na slučajan način razbija na DNK fragmente. Zatim sledi korak odabira dužine očitavanja u kom se vrši ekstrahovanje DNK fragmenata te dužine. Na kraju se vrši sekvencioniranje jednostranih očitavanja ili sekvencioniranje uparenih očitavanja. Pri sekvencioniranju jednostranih očitavanja, sekvencer čita DNK fragment u jednom smeru, dok u slučaju uparenih očitavanja, fragment biva pročitao u oba smera. Pomenuti koraci su prikazani na slici 3.4.



Slika 3.4: Sekvencioniranje *shotgun* celokupnog genoma [11]

Na slici 3.5 prikazan je odmotan i rotiran DNK fragment ekstrahovan iz uzorka genoma. Sa N su označeni delovi koji se nalaze na poziciji koja premašuje dužinu očitavanja, pa ne ulaze u sastav očitavanja.



Slika 3.5: DNK fragment ekstrahovan iz uzorka genoma [11]

Prilikom sekvencioniranja jednostranih očitavanja, dobija se očitavanje sa 5' kraja gornjeg lanca, tj. *ACTCAGCACCTTACGGCGTGCATCA*. Prilikom sekvencioniranja uparenih očitavanja, dobijaju se 5' očitavanja i gornjeg i donjeg lanca:

- *ACTCAGCACCTTACGGCGTGCATCA*
- *AGTTTGTACTGCCGTTCAAGAACGTA*

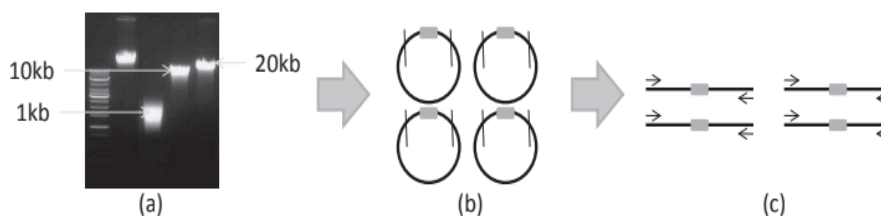
Drugo očitavanje *AGTTTGTACTGCCGTTTCAGAACGTA* je inverzni komplement očitavanja *TACGTTCTGAACGGCAGTACAAACT*, jer je drugo očitavanje dobijeno iz donjeg lanca koji je komplementaran sa gornjim.

Za asembliranje genoma, dužina očitavanja je veoma važan parametar. Različite tehnologije sekvencioniranja imaju različita ograničenja na dužinu očitavanja. Na primer, *Illumina Hi-seq*⁵ sekvencioniranje može izvršiti samo sekvencioniranje uparenih očitavanja dužine manje od 1000 bp. Za treću generaciju sekvencioniranja, ograničenje dužine očitavanja može biti veće od 10000 bp.

Sekvencioniranje partner-uparenih očitavanja

Sekvenceri druge generacije mogu izdvojiti partner-uparena očitavanja sa oba kraja kratkih fragmenata DNK (dužine očitavanja manje od 1000 bp). Za izdvajanje ovakvih očitavanja možemo koristiti **sekvencioniranje partner-uparenih očitavanja**.

Sekvenceri prvo izdvajaju duge DNK fragmente neke fiksirane dužine očitavanja (npr. 10000 bp). Potom dodaju takozvane **adapter sekvence**⁶ na krajeve svakog fragmenta (slika 3.6(a)). Zatim se vrši sečenje fragmenta levo i desno od adapter sekvence (slika 3.6(b)). Na kraju se uparena očitavanja ekstrahuju sekvencioniranjem uparenih očitavanja (slika 3.6(c)).



Slika 3.6: Sekvencioniranje partner-uparenih očitavanja [11]

Orijentacije uparenih očitavanja očitanih od strane sekvencera partner-uparenog i uparenog očitavanja se razlikuju. Sekvenceri partner-uparenih očitavanja daju dva očitavanja sa oba kraja svakog fragmenta DNK u spoljašnjoj orijentaciji umesto u unutrašnjoj. Npr. za DNK fragment sa slike 3.5 sekvencioniranje partner-uparenih očitavanja će dati:

⁵Illumina je američka kompanija koja je proizvodila visokopropusne sekvencere (omogućavaju sekvencioniranje velikog broja očitavanja odjednom (engl. *high-throughput sequencers*)).

⁶Adapter sekvence su kratke hemijski sintetizovane sekvence nukleotida koje se mogu vezati za krajeve nepoznatih DNK sekvenci i neophodne su u nekim koracima sekvencioniranja.

- TGATGCACGCCGTAAGGTGCTGAGT
- TACGTTCTGAACGGCAGTACAAACT

Iako protokol za sekvencioniranje partner-uparenih očitavanja može izdvojiti uparena očitavanja velike dužine očitavanja, on zahteva veći broj ulaznih kopija genoma za pripremu sekvencerskih biblioteka i sklon je greškama pri preklapanju očitavanja (engl. *ligation errors*).

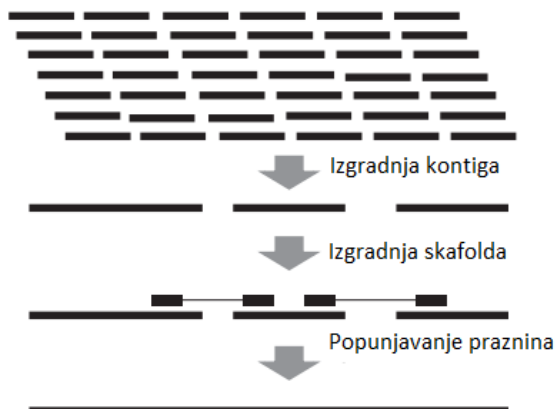
3.3 Sekvencioniranje *de novo* genoma za kratka očitavanja

Druga generacija tehnika sekvencioniranja omogućava dobijanje skupa jednostranih ili uparenih kratkih očitavanja celokupnog genoma. Sekvencioniranje *de novo* ima za cilj da izvrši preklapanje očitavanja u ispravnom redosledu i rekonstruiše genom.

Problem asembliranja genoma je računski težak. Čak i kada ne postoji greška sekvencioniranja, ovaj problem je ekvivalentan **problemu superstringa** za koji se zna da je NP-kompletno [6]. Problem superstringa predstavlja problem pronalaženja superstringa na osnovu skupa stringova S , gde je superstring najkraći string P takav da je svaki string s iz skupa S podstring stringa P . Na primer, ako je $S = \{ACATGC, ATGCGTGT, GTGTACGT\}$, onda je superstring $ACATGCGTGTACGT$.

Mnogi *de novo* asembleri predlažu asembliranje kratkih očitavanja. Opšte rešenje uključuje četiri koraka. U prvom koraku se koriguju greške sekvencioniranja u očitavanjima (opisano u poglavlju 3.4). Na osnovu korigovanih očitavanja, u drugom koraku se vrši spajanje očitavanja preklapanjem. U idealnom slučaju, teži se spajanju svih očitavanja tako da se formira kompletan genom, ali kako se zbog ponovaka javljaju dvosmislenosti, to nije moguće. Postojeće metode preklapanjem očitavanja grade kontinuiranu sekvencu, takozvanu **kontigu** (engl. *contig*). Kontige obično predstavljaju jednu **konsenzus nisku**⁷. Zatim, koristeći uparena očitavanja, vrši se rekonstrukcija redosleda kontiga tako da se formiraju **skafoldi** (engl. *scaffolds*). Svaki skafold je niz kontiga, a još se naziva i **superkontig** ili **metakontig**. Na kraju se vrši preuređivanje očitavanja u skafoldima kako bi se popunile praznine između susednih kontiga. Opisani koraci su prikazani na slici 3.7.

⁷Niska sastavljena od najfrekventnijih nukleotida na pozicijama poravnatih sekvenci.



Slika 3.7: Sekvencioniranje *de novo* genoma za kratka očitavanja [11]

3.4 Korekcija grešaka

Ukoliko se neki k -mer pojavljuje u ulaznim očitavanjima jednom ili veoma mali broj puta, velika je verovatnoća da on je nastao kao posledica grešaka prilikom sekvencioniranja očitavanja i da se neće naći u rekonstruisanom genomu. Sa druge strane, za k -mer koji se pojavljuje veliki broj puta u ulaznim očitavanjima sa velikom sigurnošću se može tvrditi da nije posledica grešaka pri očitavanju. Ove razlike se mogu iskoristiti za filtriranje grešaka u k -merima i selektivno uklanjanje iz skupa podataka očitavanja koja sadrže greške [11]. Greške u očitavanjima mogu zbuniti *de novo* asemblere. Da bi se to izbeglo, vrši se korigovanje tih grešaka pre početka asembliranja genoma.

Brojanje k -mera

Jedan konceptualno jednostavan, ali osnovni problem je **brojanje k -mera**. Brojanje k -mera se prvenstveno koristi u korekciji grešaka u očitavanjima, ali može biti korišćeno i u koraku asembliranja, detekciji ponovaka i kompresiji genomskih podataka. Ima značajnu ulogu u utvrđivanju da li je došlo do grešaka u očitavanju ili je u pitanju jednonukleotidni polimorfizam.

Problem brojanja k -mera se definiše nad ulaznim skupom očitavanja i parametrom k . Iz ovog skupa, potrebno je izdvojiti podskup Z svih mogućih k -mera i odrediti njihove frekvencije. Iako jednostavan problem, zbog velike količine podataka, on se ne može rešavati naivnim pristupima, već su potrebni napredni algoritmi i strukture podataka kako bi problem bio efikasno rešen. U nastavku će biti razma-

trana četiri rešenja:

- jednostavno heširanje
- JellyFish
- BFCOUNTER
- DSK

Jednostavno heširanje - Problem brojanja k -mera može biti rešen implementacijom asocijativnog niza koristeći **heširanje**. Heširanje je tehnika kojom se vrši preslikavanje skupa ključeva na tabelu značajno manjih dimenzija. Idealno bi bilo da funkcija za svaki ključ daje jedinstvenu poziciju. Ta funkcija naziva se **heš funkcija**, a tabela koja se koristi u tom postupku zove se **heš tabela**. Kada je k malo (npr. manje od 10), u procesu brojanja k -mera koristi se **savršeno heširanje**. Savršeno heširanje garantuje da neće doći do **kolizije**⁸. To je moguće kada se tačno zna koji skup ključeva će biti heširan prilikom dizajniranja heš funkcije. Ukoliko je k veliko, algoritam jednostavnog heširanja će zahtevati previše prostora.

Algoritam JellyFish - Moguće je smanjiti prostornu složenost algoritma jednostavnog heširanja koristeći mehanizam **otvorenog adresiranja**. Otvoreno adresiranje je način rešavanja kolizije u heš tabelama. Kada se desi kolizija, traži se sledeća slobodna lokacija u heš tabeli za smeštanje vrednosti. Postoje tri metode otvorenog adresiranja: **linearno popunjavanje**, **kvadratno popunjavanje** i **duplo heširanje**. Mehanizam otvorenog adresiranja se koristi u algoritmu *JellyFish* čija će implementacija detaljnije biti obajšnjena u poglavlju 4.1.

Algoritam BFCOUNTER - U mnogim aplikacijama, od značaja su samo k -meri koji se pojavljuju najmanje q puta. Kada bi moglo da se izbegne čuvanje k -mera koji se pojavljuju manje od q puta, sačuvalo bi se mnogo memorije. Pol Melsted (engl. *Páll Melsted*) je predložio algoritam **BFCOUNTER** koji broji samo k -mere koji se pojavljuju najmanje q puta. On koristi *counting Bloom filter*⁹ da odredi da li se k -mer pojavljuje najmanje q puta [11].

Algoritam DSK - Iako je *BFCOUNTER* prostorno efikasan, njegova prostorna složenost i dalje zavisi od broja k -mera u skupu očitavanja. Gijom Rizk (engl. *Guillaume Rizk*) predlaže metod koji se naziva **DSK**. Ideja ovog metoda je da se

⁸Izraz koji potiče od latinske reči *collisio* i znači sudar, sukob. U ovom kontekstu, moguće je da heš funkcija za dva različita k -mera da istu vrednost, te kažemo da su ta 2 k -mera u koliziji, tj. sukobu.

⁹Prostorno efikasna probabilistička struktura podataka koja dozvoljava dodavanje bilo kojih k -mera u nju i ispitivanje da li se k -mer pojavljuje najmanje q puta.

skup k-mera podeli u različite liste tako da svaka lista bude smeštena na disk koristeći D bitova. Zatim, za svaku listu, k-meri iz liste se dalje dele u podliste tako da svaka podlista može biti sačuvana u memoriji koristeći M bitova. Na kraju, frekvencije k-mera u svakoj podlisti se izračunavaju algoritmom *JellyFish*. Algoritam *DSK* će biti detaljno objašnjen u poglavlju 4.2.

3.5 Konstrukcija kontiga

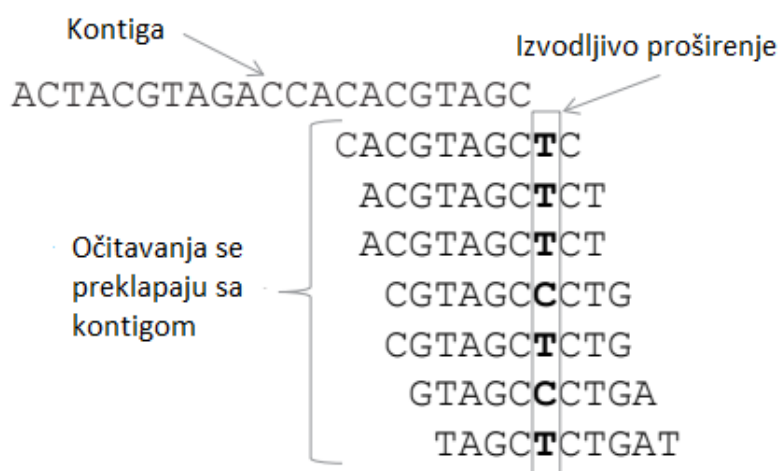
Nakon što su sva očitavanja korigovana, može se vršiti njihovo spajanje radi formiranja kontiga. Postoje dva pristupa za konstrukciju kontiga:

- pristup baznog proširenja (*base-by-base* pristup)
- De Bruijnov grafovski pristup

Pristup baznog proširenja

Pristup baznog proširenja rekonstruiše svaku kontigu tako što je proširuje bazu po bazu. Metod počinje tako što se nasumično bira očitavanje koje će služiti kao šablon. Zatim se očitavanja poravnavaju na oba kraja šablona (5' kraj i 3' kraj). Na osnovu poravnanja se dobija **konsenzusna baza**¹⁰ i šablon se njome proširuje.

Na slici 3.8 je prikazan jedan korak baznog proširenja. Na vrhu je data sekvenca koja predstavlja šablon, a ispod nje sedam očitavanja poravnatih na 3' kraju šablona.



Slika 3.8: Jedan korak baznog proširenja [11]

¹⁰Baza koja se pojavljuje najveći broj puta na određenoj poziciji u poravnanju.

Pravougaonik pokazuje da su baze C i T izvodljivo proširenje šablona. Kako je T konsenzusna baza, metod baznog proširenja će proširiti kontigu bazom T . Bazno proširenje se ponavlja sve dok ima konsenzusa. Zatim se prestaje sa proširenjem i dobija se kontiga. Proširenje se izvodi i na 3' kraju i na 5' kraju šablona.

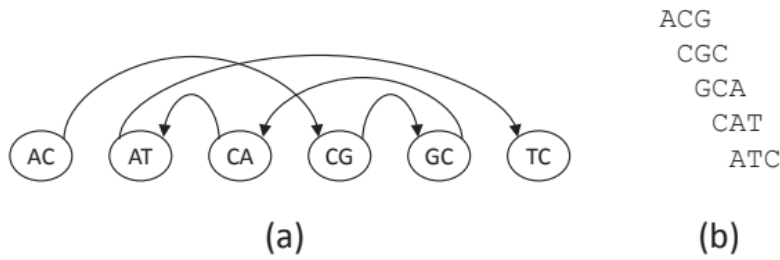
Iako je bazno proširenje jednostavno, ono često daje kratke kontige zbog dva problema. Prvo, početni šablon je nasumično izabrano očitavanje. Ukoliko ono sadrži greške sekvencioniranja ili se nalazi u ponovku, to će uticati na proširenje. Drugi problem je što se može desiti da se šablon proširi u neki od ponovaka. Ponovak stvara grane koje gore navedeni pristup ne može da razreši.

De Bruijinov grafovski pristup

De Bruijinov grafovski pristup je zasnovan na De Bruijinovim grafovima. Ovaj pristup su uveli Induri (engl. *Indury*) i Voterman (engl. *Waterman*) u njihovoj knjizi „*A new algorithm for DNK sequence assembly: Journal of Computational Biology*”. On je danas glavni pristup za asembliranje kratkih očitavanja.

Neka je dat skup očitavanja R i parametar k . De Bruijinov graf je graf $H_k = (V, E)$, gde je V skup svih k -mera skupa R , a E skup svih grana u grafu. Ako su u i v prefiks dužine k i sufiks dužine k neke podniske dužine $k + 1$ iz R , respektivno, k -meri u i v formiraju granu $(u, v) \in E$. Pod pretpostavkom da je N ukupna dužina svih očitavanja iz R , De Bruijinov graf može biti konstruisan u $O(N)$ vremenu.

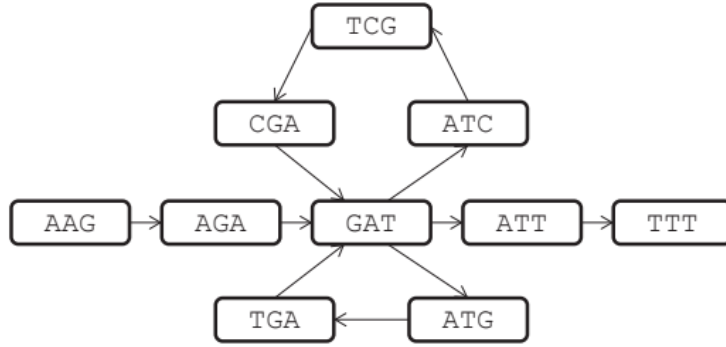
Na slici 3.9(a) je prikazan De Bruijinov graf H_3 izgrađen na osnovu skupa k -mera $R = \{ACGC, CATC, GCA\}$, pri čemu je $\{AC, AT, CA, CG, GC, TC\}$ skup čvorova. Preklapanje k -mera koji odgovaraju granama grafa H_3 je prikazano na slici 3.9(b).



Slika 3.9: Konstrukcija De Bruijinovog grafa [11]

Uzorak genoma se može predvideti identifikovanjem **Ojlerove putanje** grafa H_k . Ojlerova putanja je putanja koja obilazi svaku granu grafa H_k tačno jednom. Ojlerova putanja grafa H_k može biti izračunata u $O(n)$ vremenu, ako H_k ima n grana. Na primer, na slici 3.9(a) postoji jedinstvena putanja od čvora AC do čvora TC . Preklapanjem svih ivica 3-mera u redosledu putanje (slika 3.9(b)) dobija se sekvenca $ACGCATC$. Sekvenca $ACGCATC$ je zapravo superstring formiran preklapanjem očitavanja i dobijen na osnovu De Bruijinovog grafa za skup R . Međutim, Ojlerova putanja ne mora biti jedinstvena u H_k .

Neka je skup očitavanja $R = \{AAGATC, GATCGAT, CGATGA, ATGATT, GATTT\}$ i neka je $k = 3$. U De Bruijinovom grafu H_3 za dati skup R postoje dva ciklusa, te će postojati i dve Ojlerove putanje. Na slici 3.10 je prikazan graf H_3 . Ukoliko se prvo obiđe gornji ciklus, dobija se $AAGATCGATGATTT$, a ukoliko se prvo obiđe donji ciklus, dobija se $AAGATGATCGATTT$. Primer sa slike 3.10



Slika 3.10: De Bruijnov graf sa dva ciklusa [11]

pokazuje da Ojlerova putanja možda neće uvek dati ispravnu sekvencu. Čak još gore, Ojlerova putanja možda neće postojati u nekom grafu H_k .

U nastavku će se govoriti o De Bruijinovom grafovskom assembleru u slučaju kada:

- ne postoji greška sekvencioniranja
- postoji greška sekvencioniranja

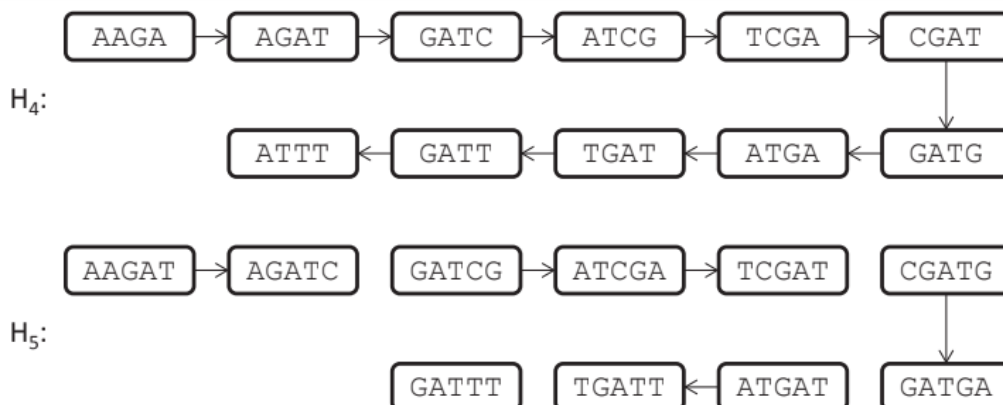
De Bruijnov assembler bez greške sekvencioniranja

Kako Ojlerova putanja nije jedinstvena i možda i ne postoji, ne teži se dobijanju kompletnog genoma. Umesto toga, teži se dobijanju skupa kontiga. Kontiga je maksimalna prosta putanja u H_k . Preciznije, svaka maksimalna prosta putanja je maksimalna putanja u H_k tako da svaki čvor (osim početnog i krajnjeg) ima unutrašnji i spoljašnji stepen 1.

Za De Bruijnov graf H_3 sa slike 3.10 mogu se konstruisati četiri kontige: $AAGAT$, $GATCGAT$, $GATGAT$, $GATTT$. Parametar k je veoma važan. Za različito k , mogu se dobiti različiti skupovi kontiga. Slika 3.11 ilustruje ovaj problem na De Bruijinovom grafu za skup $R = \{AAGATC, GATCGAT, CGATGA, ATGATT, GATTT\}$ u slučajevima kada je $k = 4$ i $k = 5$.

Kada je $k = 4$, H_4 je prosta putanja i može se konstruisati jedna kontiga $AAGATCGATGATTT$. U slučaju da je $k = 5$, H_5 sadrži pet povezanih komponenti, svaka je prosta putanja i može se konstruisati pet kontiga: $AAGATC$, $GATCGAT$, $CGATGA$, $ATGATT$, $GATTT$.

De Bruijnov graf je dobar, ukoliko se zna ispravno k . Kada je k malo (pogledati H_3 na slici 3.10), postoji veliki broj grana zbog ponovaka i rezultat je veliki broj kratkih kontiga. Kada je k veliko (pogledati H_5 na slici 3.11), neki k-meri nedostaju,



Slika 3.11: De Brujinovi grafovi za $k = 4$ i $k = 5$ [11]

što rezultuje nepovezanim komponentama, koje takođe generišu veliki broj kratkih kontiga. Tako da se mora identifikovati ispravno k kako bi se pronašla ravnoteža između ova dva problema.

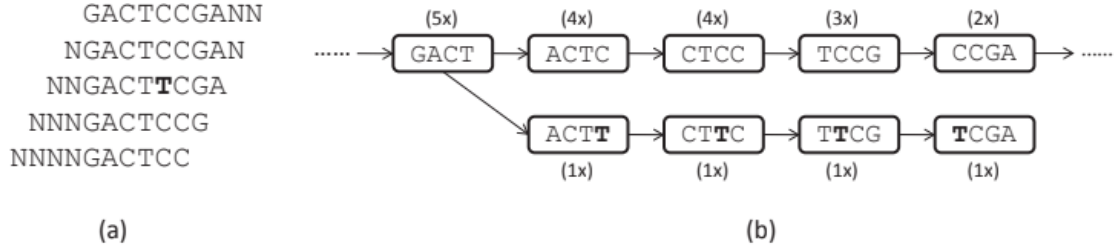
De Brujinov assembler sa greškama sekvencioniranja

Kod De Brujinovog assemblera bez grešaka sekvencioniranja pretpostavka je bila da ne postoje greške sekvencioniranja u očitavanjima, što je nerealno. Kada postoje greške, pokušava se njihovo uklanjanje redukovanjem šuma u De Brujinovom grafu. Rešenje koje će se razmatrati su predložili Zerbino (engl. *Zerbino*) i Birni (enlg. *Birney*). Kratka očitavanja imaju nisku stopu greške (sadrže jednu greška na svakih 100 baza), a većina k -mera sadrži najviše jednu grešku. Ovakvi k -meri sa greškom mogu kreirati dva moguća anomalna podgrafa u De Brujinovom grafu: **vrh (špic)** i **mehurić**.

Vrh je putanja dužine najviše k gde svi unutrašnji čvorovi imaju ulazni i izlazni stepen 1, dok jedan od njihovih krajeva ima ulazni ili izlazni stepen 0. To može proizvesti potencijalnu kontigu dužine najviše $2k$.

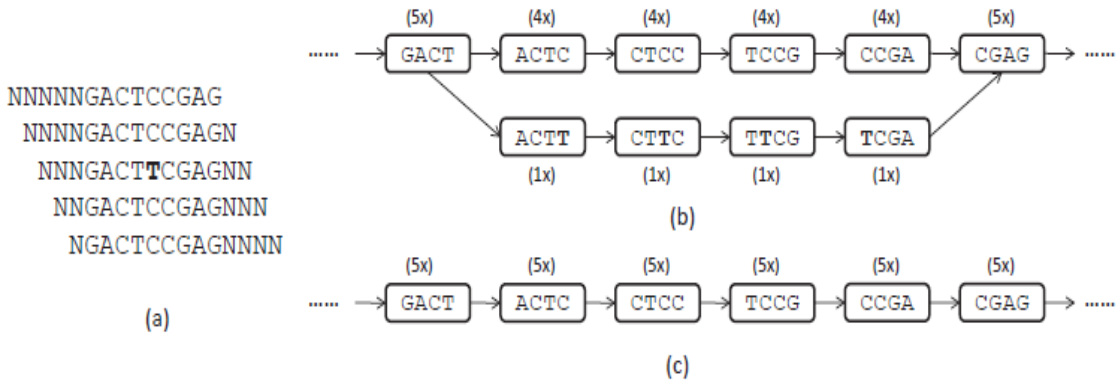
Slika 3.12(a) predstavlja višestruko poravnanje skupa od pet očitavanja, gde treće očitavanje ima grešku sekvencioniranja (prikazana podebljanim fontom). Slika 3.12(b) predstavlja De Brujinov graf koji odgovara skupu očitavanja sa prvog dela slike i vrh koji je formiran usled jednog nepoklapanja u jednom očitavanju, tj. usled pomenute greške. Brojevi u zagradama označavaju mnogostrukost 4-mera (broj pojavljivanja 4-mera u očitavanjima). Ovaj vrh se može ukloniti iz De Brujinovog grafa, ali uklanjanje jednog vrha može generisati nove vrhove. Tako da procedura

uklanjanja mora uklanjati vrhove rekurzivno.



Slika 3.12: Formiranje vrha [11]

Mehurić se sastoji od dve različite kratke putanje sa istim početnim i krajnim čvorovima u De Brujinovom grafu, pri čemu su te dve putanje zapravo kontige koje se razlikuju u samo jednom nukleotidu. Na primer, slika 3.13(a) sadrži skup očitavanja gde treće očitavanje ima jedno nepoklapanje. Slika 3.13(b) prikazuje odgovarajući De Brujinov graf i jedan mehurić koji se formira.



Slika 3.13: Formiranje mehurića [11]

Gornja putanja mehurića predstavlja *GACTCCGAG*, dok donja putanja predstavlja *GACTTCGAG*. Kada su dve putanje u mehuriću veoma slične, putanja sa nižom mnogostukošću će verovatno biti ona koja će biti odbačena. Može se pokušati sa spajanjem mehurića. U ovom primeru, dve putanje imaju samo jedno nepoklapanje. Budući da su čvorovi u donjoj putanji niže mnogostrukosti, vrši se spajanje mehurića i dobija se graf na slici 3.12(c). Preciznije, može se definisati **težina putanje** $w_1 \rightarrow w_2 \rightarrow \dots \rightarrow w_p$ kao $\sum_{i=1}^p f(w_i)$, gde je $f(w_i)$ mnogostrukost od w_i . Onda, kada se dve putanje spoje u mehuriću, zadržava se ona sa većom težinom.

Nakon što se uklone vrhovi i spoje mehurići u De Bruijnovom grafu, može se dalje redukovati šum u grafu uklanjanjem k -mera sa mnogostrukošću manjom ili jednakom nekom pragu.

Kombinovanjem ove tri tehnike: (1) otklanjanje vrhova, (2) spajanje mehurića i (3) filtriranje k -mera sa niskom mnogostrukošću dobija se algoritam *Velvet* [16] za manipulaciju De Bruijnovim grafovima i sastavljanje genomske sekvence.

Kako izabrati k ?

Odabir broja k može uticati na performanse De Bruijnovog algoritma. Jednostavno rešenje je pokretanje *Velvet* algoritma za različite k . Zatim se vrši grupisanje i spajanje kontiga. Jedan od problema sa ovim jednostavnim rešenjem je što kontige dobijene za različito k imaju različit kvalitet. Kontige dobijene na osnovu H_k , gde je k malo, veoma su tačne. Međutim, takve kontige su kratke, jer postoji veliki broj grana zbog ponovaka. Kontige dobijene na osnovu H_k gde je k veliko su duže, ali one mogu sadržati mnogo grešaka.

IDBA assembler¹¹ počiva na ideji da ne treba graditi De Bruijnov graf nezavisno za različite k . Umesto toga, *IDBA* gradi De Bruijnov graf H_k postepeno krećući od malih k i idući ka većim. Kada je k malo, mogu se dobiti visokokvalitetne kontige, iako su kratke. Zatim se ove kontige koriste za ispravljanje grešaka u očitavanjima. Postepeno se izgrađuje De Bruijnov graf H_k za sve veće k . Kako su očitavanja u R korigovana, to je šum u grafu H_k redukovan. Na ovaj način se za veliko k omogućava dobijanje dugih kontiga visokog kvaliteta.

¹¹IDBA - A Practical Iterative de Bruijn Graph De Novo Assembler.

Glava 4

Implementirani algoritmi i rezultati

U poglavlju 3 data je biološka osnova za algoritme koji se primenjuju u sekvencioniranju genoma. U ovom poglavlju će biti dati opisi implementacije tih algoritama u programskom jeziku Elixir sa osvrtom na njihovu prostornu i vremensku složenost, kao i opisi aplikacije za izvršavanje algoritama i rezultati njihovog izvršavanja.

4.1 Algoritam *JellyFish*

Algoritam *JellyFish* se koristi za brojanje k-mera, a kao potprogram programa za korigovanje grešaka u sekvencioniranju očitavanja. Pseudokod ovog algoritma dat je na slici 4.1.

Neka je $h()$ heš funkcija i $H[0..\frac{N}{\alpha} - 1]$ heš tabela koja čuva niz k-mera, gde je $N = |Z|$ i α faktor opterećenja ($0 < \alpha \leq 1$). Potrebno je izgraditi tabelu $Count[0..\frac{N}{\alpha} - 1]$, gde $Count[i]$ čuva broj pojavljivanja za k-mer $H[i]$. Za svaki k-mer iz Z vrši se njegovo heširanje u neku vrednost $H[i]$, gde je $i = h(z)$. Ako $H[i]$ nije prazan i $H[i] \neq z$, z se ne može čuvati u $H[i]$, odnosno došlo je do kolizije. Ona može biti razrešena pomoću mehanizma otvorenog adresiranja. Na primer, kolizija se može razrešiti linearnim popunjavanjem. Ovom metodom se indeks i uvećava za 1, kada se kolizija dogodi, sve dok je $H[i] = z$ ili je ulaz $H[i]$ prazan.

Funkcija $hashEntry()$ sa slike 4.2 ilustruje šemu linearnog popunjavanja za razrešavanje kolizije. Ako $hashEntry(z, h, \frac{N}{\alpha})$ vraća prazan ulaz $H[i]$, onda z ne postoji u heš tabeli i postavljamo $H[i] = z$ i $Count[i] = 1$. U suprotnom, ako $hashEntry(z, h, \frac{N}{\alpha})$ vraća ulaz $H[i] = z$, uvećava se $Count[i]$ za jedan. Nakon sto su svi k-meri iz Z obrađeni, prikazuje se $(H[i], Count[i])$ za sve ulaze $H[i]$ različite od 0.

Algoritam JellyFish(Z, α, h)

Ulaz: Z je skup od N k-mera, α je faktor opterećenja koji kontroliše veličinu heš tabele i h je heš funkcija.

Izlaz: Broj pojavljivanja svakog k-mera u skupu Z .

- 1: Kreirati praznu heš tabelu $H[1..\frac{N}{\alpha}]$ tako da svaki ulaz zahteva $2k$ bitova
- 2: Kreirati praznu tabelu $Count[1..\frac{N}{\alpha}]$ tako da svaki ulaz zahteva 32 bita
- 3: foreach k-mer $z \in Z$ do:
- 4: $i = hashEntry(z, h, \frac{N}{\alpha})$;
- 5: if $H[i]$ is empty then
- 6: $H[i] = z$ and $Count[i] = 1$;
- 7: else
- 8: $Count[i] = Count[i] + 1$;
- 9: end if
- 10: end for
- 11: Output $(H[i], Count[i])$ za sve ulaze $H[i]$ različite od 0.

Slika 4.1: *JellyFish* algoritam [11]**Algoritam hashEntry**($Z, \alpha, size$)

- 1: $i = h(z) \bmod size$;
- 2: while $H[i] \neq z$ do
- 3: $i = i + 1 \bmod size$; /* linearno popunjavanje */
- 4: end while
- 5: Return i ;

Slika 4.2: Funkcija *hashEntry* [11]

Algoritam *JellyFish* je efikasniji, ukoliko ne postoji kolizija. U praksi je očekivani broj kolizija manji, ukoliko za faktor opterećenja važi $\alpha \leq 0.7$. Zatim, očekivano vreme izvršavanja je $O(N)$. Što se tiče prostorne složenosti, tabele H i $Count$ zahtevaju $\frac{N}{\alpha}(2k + 32)$ bitova, pod pretpostavkom da broj zauzima 32 bita.

Skup očitavanja Z je prilikom implementacije algoritma predstavljen listom, dok su za predstavljanje heš tabele H i $Count$ tabele korišćene mape. Pri početnoj inicijalizaciji, vrednosti H tabele se postavljaju na 0, a vrednosti $Count$ tabele na prazan string, dok im je skup ključeva isti. Početna inicijalizacija se vrši rekursivno za svaki ključ. Izgled funkcija koje inicijalizuju H i $Count$ tabelu, može se videti na listingu 4.1. Argumenti pri pozivanju obe funkcije su prazna mapa ($\%{\}$) i veličina mape koju treba kreirati. Pri svakom rekursivnom pozivu, veličina mape se uma-

njuje za 1, a u mapu se ubacuje novi element funkcijom *Map.put*/3. Kada veličina mape postane jednaka nuli, tada se izlazi iz rekurzije i funkcija vraća inicijalizovanu mapu.

```

1 def build_h_table(h_table_empty, 0), do: h_table_empty
2 def build_h_table(h_table_empty, size_of_H) do
3   size_of_H = size_of_H - 1
4   h_table_empty = Map.put(h_table_empty, size_of_H, "")
5   build_h_table(h_table_empty, size_of_H)
6 end
7
8 def build_count_table(count_table_empty, 0), do: count_table_empty
9 def build_count_table(count_table_empty, size_of_count_table) do
10  size_of_count_table = size_of_count_table - 1
11  count_table_empty = Map.put(count_table_empty,
12    size_of_count_table, 0)
13  build_count_table(count_table_empty, size_of_count_table)
14 end

```

Listing 4.1: Funkcije za inicijalizaciju *H* i *Count* tabele

4.2 Algoritam *DSK*

DSK je još jedan algoritam za brojanje k-mera, ali vremenski i prostorno efikasniji od algoritma *JellyFish*. Njegov pseudokod dat je na slici 4.3.

Neka je dat skup k-mera Z , broj bitova D na disku za čuvanje svake liste u koju se smeštaju k-meri, broj bitova M u memoriji za čuvanje svake podliste k-mera i heš funkcija $h()$. Prvo se k-meri iz skupa Z dele u n_{list} lista približno slične dužine. Kako disk ima D bitova i svaki k-mer može biti reprezentovan u $2k$ bitova, svaka lista može čuvati $l_{list} = \frac{D}{2k}$ k-mera. Kako je N broj k-mera u Z , to je $n_{list} = \frac{N}{l_{list}} = \frac{2kN}{D}$. Ovo deljenje se obavlja heš funkcijom $h()$ koja ravnomerno mapira sve k-meru u n_{list} lista. Preciznije, za svaki k-mer z iz skupa Z , z se dodeljuje i-toj listi, ako je $h(z) \bmod n_{list} = i$. Zatim, svaka lista se dalje deli u podliste, pri čemu je svaka dužine $l_{sublist}$. Svaka podlista će biti obrađena u memoriji pomoću algoritma *JellyFish*, koji zahteva $\frac{l_{sublist}}{0.7}(2k + 32)$ bitova. Kako memorija ima M bitova, tako je $l_{sublist} = \frac{0.7M}{(2k+32)}$. Broj podlista je $n_{sublist} = \frac{n_{list}}{l_{sublist}} = \frac{D(2k+32)}{0.7(2k)M}$.

Svaka lista je podeljena u podliste heš funkcijom $h()$. Preciznije, za svaki k-mer s u i-toj listi, s je dodeljen j-toj podlisti, ako je $(\frac{h(s)}{n_{list}}) \bmod n_{sublist} = j$. Za svaku podlistu dužine $l_{sublist} = 0.7M2k + 32$, koristeći M bitova, vrši se brojanje

Algoritam DSK(Z, M, D, h)

Ulaz: Z je skup od N k-mera, M (bitova) je veličina memorije za čuvanje lista, D (bitova) je veličina diska za čuvanje podlista i h je heš funkcija.

Izlaz: Broj pojavljivanja svakog k-mera u skupu Z .

```

1:  $n_{list} = \frac{2kN}{D}$ ;
2:  $n_{sublist} = \frac{D(2k+32)}{0.7(2k)M}$ ;
3: for  $i = 0$  to  $n_{list} - 1$  do
4:     Inicijalizovati skup praznih podlista  $d_0, \dots, d_{n_{sublist}-1}$  na disku
5:     for each k-mer  $z \in Z$  do
6:         if  $h(z) \bmod n_{list} = i$  then
7:              $j = (h(z) \text{div } n_{list}) \bmod n_{sublist}$ ;
8:             Upisati  $z$  na disk u podlistu  $d_j$ 
9:         end if
10:    end for
11:    for  $j = 0$  to  $n_{sublist} - 1$  do
12:        Učitaj  $j$ -tu podlistu  $d_j$ ;
13:        Pokreni  $JellyFish(d_j, 0.7, h)$  za ispisivanje broja pojavljivanja
        svakog k-mera u podlisti  $d_j$ 
14:    end for
15: end for
    
```

 Slika 4.3: Funkcija *DSK* algoritam [11]

pojavljivanja svakog k-mera u podlisti koristeći algoritam $JellyFish(d_j, 0.7, h)$ sa slike 4.1.

DSK algoritam će zapisati samo jednom svaki k-mer iz Z , iako će svaki k-mer pročitati n_{list} puta. Stoga, algoritam neće generisati mnogo pristupa disku radi pisanja. Što se tiče vremenske složenosti, za i -tu iteraciju, algoritam numeriče sve k-mere u Z , što oduzima $O(n)$ vremena. Zatim, algoritam identifikuje $\frac{D}{2k}$ k-mera koji pripadaju i -toj listi i zapisuje ih na disk, što oduzima $O(\frac{D}{k})$ vremena. Nakon toga, algoritam čita $\frac{D}{2k}$ k-mera i izvodi brojanje, što oduzima $O(\frac{D}{k})$ vremena. Tako da svaka iteracija zahteva $O(N + \frac{D}{k}) = O(N)$ vremena, gde je $N > \frac{D}{2k}$. Kako je $n_{list} = \frac{2kN}{D}$ broj iteracija, algoritam se izvršava u $O(kN^2)$ očekivanom vremenu. Kad je $D = \theta(N)$, algoritam se izvršava u $O(kN)$ očekivanom vremenu.

Heš funkcija $h()$, koja se koristi i u algoritmu *DSK* i u algoritmu *JellyFish*, predstavljena je funkcijom *pattern_to_number* koja koristi pomoćnu funkciju *symbol_to_number*. Implementacija ovih funkcija u Elixir-u može se videti na listingu 4.2. Pomoćna funkcija *symbol_to_number* kodira nukleotide pomoću mape, gde

su A, T, C i G ključevi sa vrednostima 00, 01, 10, 11 respektivno. Funkcija *pattern_to_number* rekurzivno kodira svako slovo niske *pattern* koja predstavlja k-mer, počevši od nulte pozicije. Kodovi se nadovezuju na početnu praznu nisku *begin_string* pomoću operatora <>. Kad se obradi cela niska, izlazi se iz rekurzije i funkcija vraća nisku transformisanu u broj koji predstavlja kodiran k-mer.

```
1 def symbol_to_number(c) do
2   map = %{"A" => "00", "T" => "01", "C" => "10", "G" => "11"}
3   map[c]
4 end
5
6 def pattern_to_number(_pattern, begin_string, index,
7   length_of_pattern) when index == length_of_pattern do
8   String.to_integer(begin_string, 10)
9 end
10
11 def pattern_to_number(pattern, begin_string, index,
12   length_of_pattern) do
13   begin_string = begin_string <> symbol_to_number(String.at(pattern
14     , index))
15   pattern_to_number(pattern, begin_string, index + 1,
16     length_of_pattern)
17 end
```

Listing 4.2: Funkcija *pattern_to_number* i pomoćna funkcija *symbol_to_number*

4.3 Algoritam *De_Brujin_graph*

De Brujinov grafovski pristup je jedan od pristupa u izgradnji kontiga koji se danas najčešće koristi. Cilj ovog algoritma je da pronađe sve kontige određene veličine, a uz pomoć De Brujinovog grafa. Slika 4.4 daje pseudokod ovog jednostavnog metoda.

Neka je dat skup očitavanja R i parametar k . Prvo se izgrađuje De Brujinov graf H_k , a zatim se pronalaze sve maksimalne proste putanje i od njih se formiraju kontige.

U okviru ovog rada, implementiran je algoritam *De_Brujin_graph* koji se koristi u prvom koraku De Brujinovog grafovskog asemblera. Argumenti ovog algoritma su string koji predstavlja jedno očitavanje iz skupa R i parametar k . Prvi korak je kreiranje skupa k-mera na osnovu očitavanja i parametra k . Zatim sledi korak for-

Algoritam *De_Brujin_Assembler*(R, k)**Ulaz:** R je skup očitavanja i k je parametar De Brujinovog grafa.**Izlaz:** Skup kontiga A .1: Generisati De Brujinov graf H_k za skup R ;2: Izdvojiti sve maksimalne proste putanje u H_k kao kontige;

Slika 4.4: De Brujinov grafovski assembler [11]

miranja De Brujinovog grafa na osnovu kreiranog skupa k -mera. Povratna vrednost algoritma je mapa koja predstavlja De Brujinov graf. Svaki ključ mape predstavlja jedan čvor grafa, a njegova vrednost je lista stringova koji takodje predstavljaju čvorove grafa i sa kojima pomenuti čvor formira granu. Primer pokretanja ovog algoritma biće dat u poglavlju 4.7

4.4 Algoritam *All_Eulerian_cycles*

U poglavlju 3.5 je objašnjen pojam Ojlerove putanje i značaj u procesu rekonstrukcije genoma. U tom procesu, algoritam za pronalaženje svih Ojlerovih putanja je neophodan. Pseudokod pronalaženja Ojlerove putanje na osnovu De Brujinovog grafa može se videti na slici 4.5. Algoritam za pronalaženje Ojlerove putanje u grafu nije NP-kompletna i može se efikasno pronaći [10].

Problem Ojlerove putanje: Pronaći Ojlerovu putanju u grafu.**Ulaz:** Graf.**Izlaz:** Putanja koja posećuje svaku granu u grafu tačno jednom.

Slika 4.5: Problem Ojlerove putanje

U odeljku 4.3 je prikazano kako se na osnovu očitavanja i parametra k može konstruisati De Brujinov graf. Konstruisani graf se može iskoristiti kao primer grafa u kome se pronalaze sve Ojlerove putanje. Rezultati pokretanja ovog algoritma će detaljno biti prikazani u poglavlju 4.7.

4.5 Uputstvo za pokretanje aplikacije

Celokupna implementacija grafičkog korisničkog interfejsa i algoritama je dostupna i može se preuzeti sa adrese [1]. Nakon preuzimanja implementacije i izvršenih koraka iz poglavlja 2.2, potrebno je preuzeti i instalaciju za neko od integrisanih razvojnih okruženja za programski jezik Python. Za potrebe ovog rada izabran je *PyCharm*. Njegova instalacija se može preuzeti sa zvanične stranice [21].

Na osnovu preuzetog materijala sa [1], potrebno je napraviti projekat u *PyCharm*-u. Projekat se sastoji od glavnog *.py* fajla u kome se nalazi implementacija čitavog interfejsa. Paket *Tkinter* je iskorišćen za implementaciju grafičkog korisničkog interfejsa, a za njegovo korišćenje je neophodno na početku glavnog *.py* fajla navesti komandu *from tkinter import **. Zatim, u okviru projekta se nalaze *.exs* fajlovi sa implementacijama algoritama (*jellyfish.exs*, *dsk.exs*,...). Tekstualni fajlovi iz kojih se čitaju ulazni parametri za algoritme se mogu nalaziti na bilo kojoj lokaciji, jer se njihov izbor vrši preko posebnog dijaloga. Nakon izvršavanja svakog algoritma, kreira se novi tekstualni fajl unutar projekta, koji nosi naziv tog algoritma i u kome su smešteni rezultati njegovog izvršavanja. Izgled ulaznih i izlaznih fajlova za svaki pojedinačni algoritam biće opisan u poglavlju .

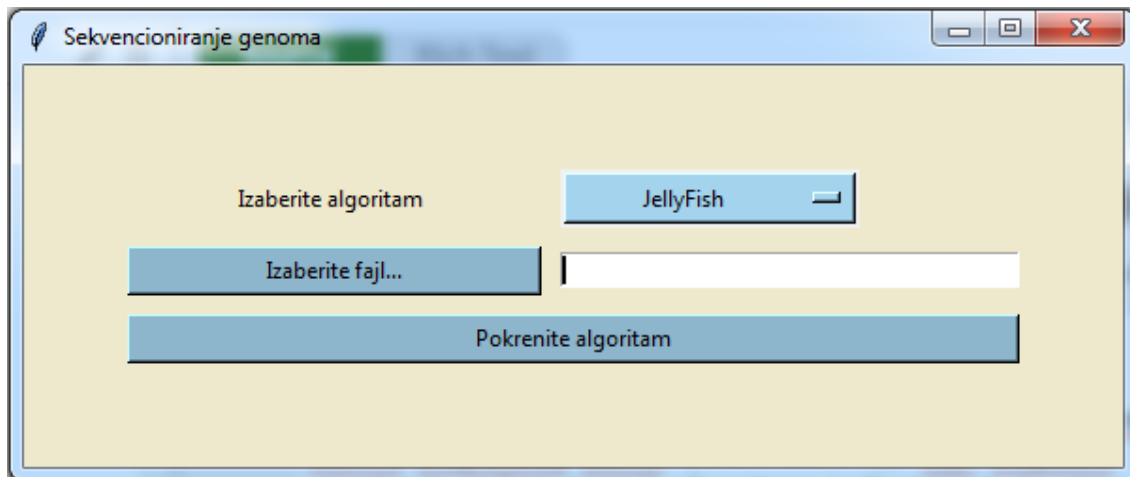
Implementirani algoritmi se mogu pokrenuti i iz komandne linije. Potrebno je izvršiti komandu *elixir ime_algoritma.exs* za kojom slede argumenti razdvojeni razmakom. Rezultati izvršavanja će biti prikazani unutar komandnog prozora.

4.6 Opis aplikacije

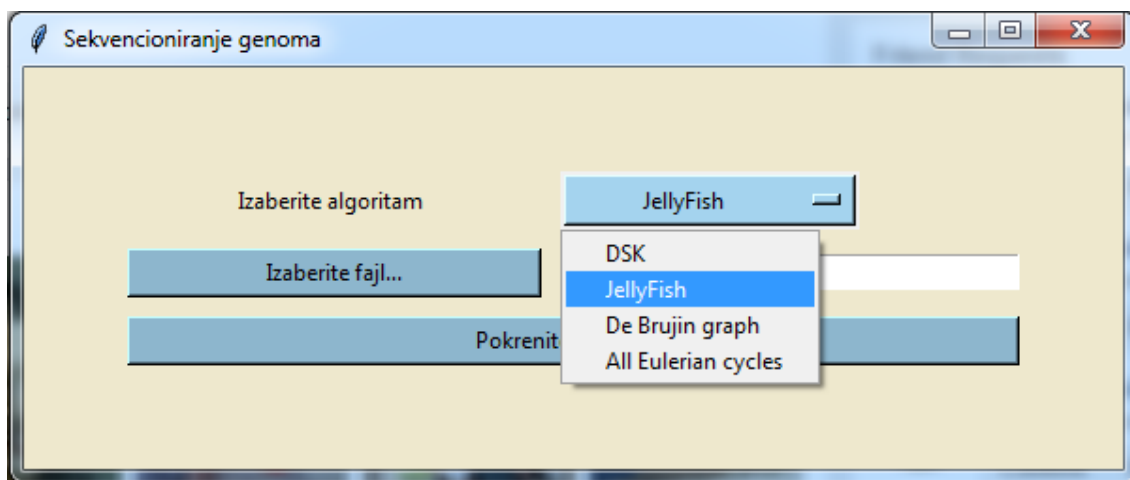
Svi algoritmi su implementirani u programskom jeziku Elixir. Prilikom sekvenciranja genoma potrebno je obraditi veliki broj ulaznih očitavanja. Elixir omogućava brzu obradu velike količine podataka odjednom i pritom je veoma tolerantan na greške, te ga to čini dobrim izborom. Međutim, Elixir nema naprednu podršku za pravljenje interfejsa i to nije njegova osnovna namena. Zbog toga je za potrebe ovog rada iskorišćena i podrška programskog jezika Python. Interfejs koji omogućava pokretanje algoritama opisanih u poglavlju 4 implementiran je upravo u ovom jeziku, a koristeći integrisano razvojno okruženje *PyCharm*. Kratak prikaz i opis izgleda i funkcionalnosti interfejsa biće dat u nastavku.

Pri pokretanju programa otvara se glavni prozor koji se može videti na slici 4.6. Klikom na padajuću listu mogu se videti dostupni algoritmi za izvršavanje, a potom

i izabrati neki od njih. Prikaz padajuće liste i izbor jednog od algoritama može se videti na slici 4.7.



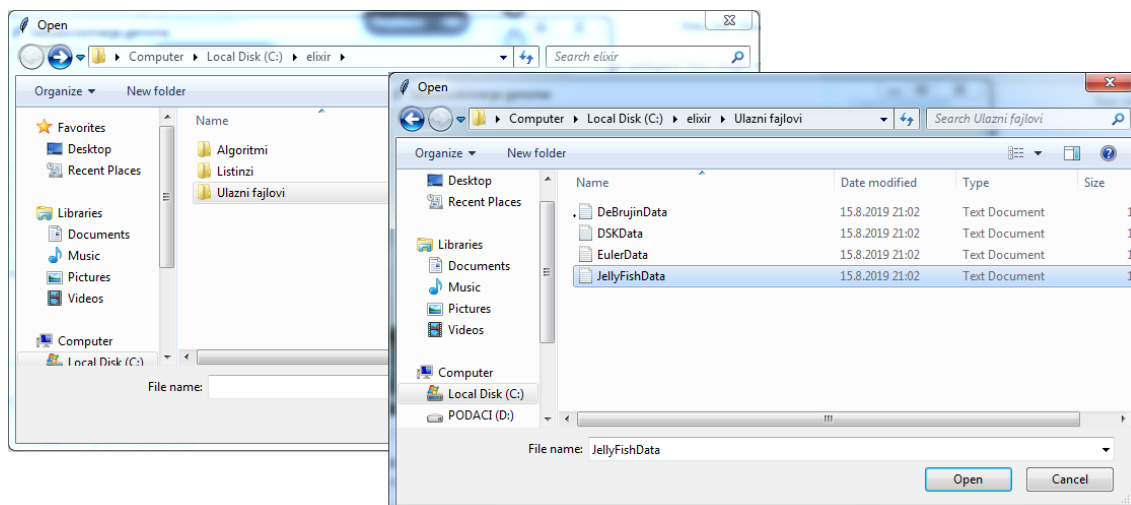
Slika 4.6: Glavni prozor programa



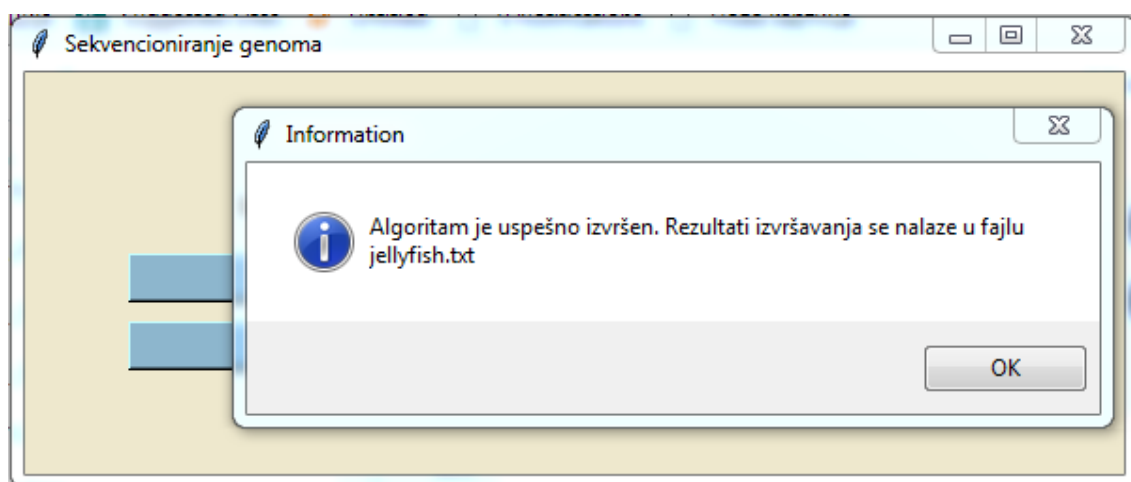
Slika 4.7: Izbor algoritma

Nakon toga je potrebno izabrati fajl iz kojeg će se čitati ulazni podaci. Klikom na dugme *Izaberite fajl...* otvara se prozor u okviru kojeg se može izabrati ulazni tekstualni fajl, što se može videti na slici 4.8.

Na kraju je potrebno pokrenuti algoritam klikom na dugme *Pokrenite algoritam*. Ukoliko je algoritam uspešno izvršen, rezultati pokretanja će biti upisani u odgovarajući tekstualni fajl i pojaviće se obaveštenje o uspešnom izvršavanju i nazivu rezultujućeg fajla. Prikaz ovog obaveštenja dat je na slici 4.9.

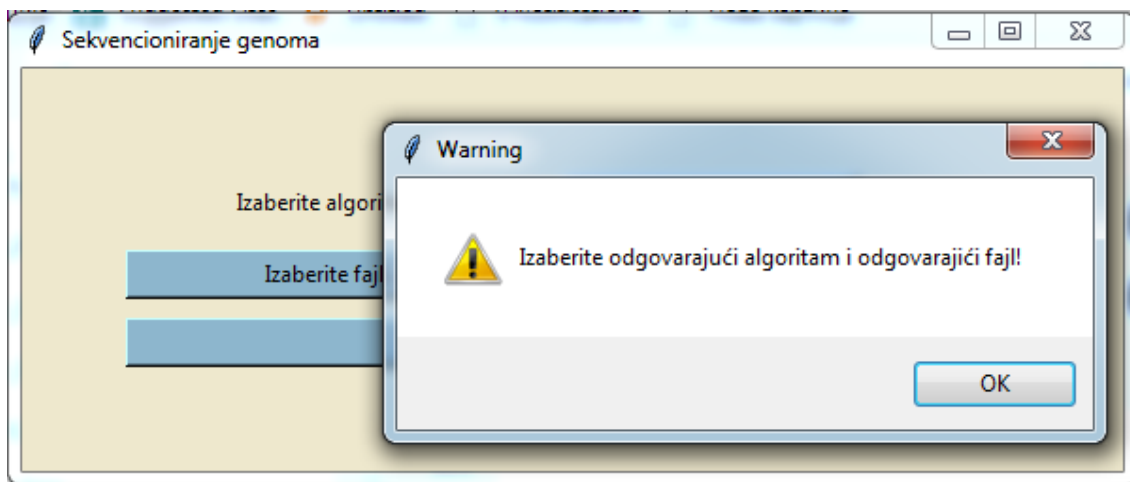


Slika 4.8: Izbor ulaznog fajla



Slika 4.9: Izbor algoritma

Ukoliko je prilikom pokretanja algoritma detektovano nepoklapanje u izabranom algoritmu i ulaznom fajlu, pojaviće se upozorenje koje je prikazano na slici 4.10. U tom slučaju je potrebno proveriti izabrane vrednosti i ponovo pokrenuti algoritam.



Slika 4.10: Izbor algoritma

4.7 Rezultati

U ovom poglavlju će biti prikazani rezultati pokretanja algoritama, koji su nakon izvršavanja smešteni u odgovarajućim tekstualnim fajlovima i mogu se upoređivati za različite ulazne vrednosti.

Algoritam *JellyFish*

Ulazni fajl *JellyFishData.txt* za algoritam *JellyFish* sadrži dve linije. U prvoj liniji se nalaze k-meri razdvojeni zarezom (*AC, CG, AC, GT, CA, GG, AC, GT*), dok se u drugoj liniji nalazi samo parametar alfa (0.7). Nakon izvršavanja algoritma, rezultat je smešten u tekstualnom fajlu *jellyfish.txt*, čiji prikaz je dat na listingu 4.3. Rezultati se očitavaju tako što k-meru, koji je dodeljen nekom ključu u prvoj mapi, odgovara broj pojavljivanja dodeljen istom ključu u drugoj mapi (npr. k-mer "CG" se pojavljuje jednom u ulaznom skupu k-mera, dok se k-mer "AC" pojavljuje tri puta).

```

1 {%{
2   0 => " ",
3   1 => " ",
4   2 => " ",
5   3 => "CG",
6   4 => "CA",
7   5 => " ",
8   6 => " ",

```

```

9      7 => "GG",
10     8 => "",
11     9 => "GT",
12    10 => "AC",
13    11 => ""
14 },
15 %{
16     0 => 0,
17     1 => 0,
18     2 => 0,
19     3 => 1,
20     4 => 1,
21     5 => 0,
22     6 => 0,
23     7 => 1,
24     8 => 0,
25     9 => 2,
26    10 => 3,
27    11 => 0
28 }%

```

Listing 4.3: Rezultati izvršavanja *JellyFish* algoritma

Broj kolizija u toku heširanja k-mera je manji, a algoritam je brži i efikasniji, ukoliko za parametar α važi $\alpha \leq 0.7$, tako da se brzina i efikasnost mogu uporediti pokretanjem algoritma za različito α . Na malom skupu podataka, unapređenje u pogledu brzine i efikasnosti je zanemarljivo, ali prilikom izvršavanja nad velikim skupom podataka je od velikog značaja.

Algoritam *DSK*

Ulazni fajl *DSKData.txt* za algoritam *DSK* sadrži tri linije. Prva linija ima isti sadržaj kao i prva linija ulaznog fajla za algoritam *JellyFish* 4.7. U drugoj liniji se nalazi parametar M (6) i parametar D (5). Rezultati izvršavanja algoritma se mogu videti na listingu 4.4. Ušteda prostora se ogleda u ograničavanju memorije koja će biti iskorišćena za smeštanje k-mera u liste i podliste, a sve to zahvaljujući parametrima M i D . Ukoliko je parametar D približno jednak broju N k-mera u datom skupu k-mera, algoritam će biti izuzetno brz i izvršiće se u $O(kN)$ očekivanom vremenu.

```

1  [[], ["AC", "AC", "AC"], [], [], [], [], ["CA"], [], [], []]

```

```
2 %{0 => "CA", 1 => ""}  
3 %{0 => 1, 1 => 0}  
4 %{0 => "AC", 1 => "", 2 => "", 3 => "", 4 => ""}  
5 %{0 => 3, 1 => 0, 2 => 0, 3 => 0, 4 => 0}  
6 [], [], [], ["GT", "GT"], [], [], [], [], ["CG"], []  
7 %{0 => "", 1 => "CG"}  
8 %{0 => 0, 1 => 1}  
9 %{0 => "GT", 1 => "", 2 => ""}  
10 %{0 => 2, 1 => 0, 2 => 0}  
11 [], [], [], [], [], ["GG"], [], [], [], []  
12 %{0 => "", 1 => "GG"}  
13 %{0 => 0, 1 => 1}
```

Listing 4.4: Rezultati izvršavanja *DSK* algoritma

Algoritam *De_Brujin_graph*

Ulazni fajl *DeBrujinData.txt* sadrži dve linije, pri čemu je u prvoj smešteno očitavanje (*TAATGCCATGGGATGTT*), a u drugoj parametar k (3). Nakon pokretanja *De_Brujin_graph* algoritma, rezultat se može videti u tekstualnom fajlu *debrujin_graph.txt*, čiji prikaz se nalazi na listingu 4.5.

```
1 %{  
2   "AA" => ["AT"],  
3   "AT" => ["TG", "TG", "TG"],  
4   "CA" => ["AT"],  
5   "CC" => ["CA"],  
6   "GA" => ["AT"],  
7   "GC" => ["CC"],  
8   "GG" => ["GG", "GA"],  
9   "GT" => ["TT"],  
10  "TA" => ["AA"],  
11  "TG" => ["GC", "GG", "GT"],  
12  "TT" => []  
13 }
```

Listing 4.5: Rezultati izvršavanja *De_Brujin_graph* algoritma

Na osnovu dobijene mape se može konstruisati graf sa čvorovima koje predstavljaju ključevi i granama ka čvorovima koji se nalaze unutar liste k -mera koje predstavljaju vrednosti mape.

***All_Eulerian_cycles* algoritam**

Na listingu 4.5 je prikazan graf koji je dobijen pokretanjem *De_Brujin_graph* algoritma i on će biti iskorišćen kao ulazni parametar za algoritam *All_Eulerian_cycles*. Dakle, ulazni fajl *EulerData.txt* sadrži pomenuti De Brujinov graf, koji je u formatu *AT : TC; TC : CG; CG : GA, GG; GA : AT, AC; AC : CG; GG : GA*. Elementi odvojeni sa ; predstavljaju jedan element mape, čiji je ključ k-mer sa leve strane znaka :, a vrednost lista k-mera koji se nalaze sa desne strane znaka : i koji su razdvojeni zarezom. Nakon izvršavanja algoritma, dobija se lista kontiga koja je prikazana na listingu 4.6.

```
1 ["ACGATCGGAC", "ATCGGACGAT", "CGGACGATCG", "CGATCGGACG", "  
   GACGATCGGA",  
2 "GATCGGACGA", "GGACGATCGG", "TCGGACGATC", "ACGGATCGAC", "  
   ATCGACGGAT",  
3 "CGACGGATCG", "CGGATCGACG", "GATCGACGGA", "GACGGATCGA", "  
   GGATCGACGG",  
4 "TCGACGGATC"]
```

Listing 4.6: Rezultati izvršavanja *All_Eulerian_cycles* algoritma

Glava 5

Zaključak

U ovom radu su prikazane osnove sintakse i semantike funkcionalnog programskog jezika Elixir. On je iskorišćen za implementaciju algoritama u oblasti sekvencioniranja genoma, pa je na taj način prikazana i njegova primena u rešavanju realnih problema. Sekvencioniranje genoma je proces otkrivanja sastava genoma i od izuzetnog je značaja za modernu dijagnostiku, a posebno za dijagnostiku rizika od naslednih bolesti. Za potrebe ovog rada, algoritmi su testirani na veoma malom skupu podataka, te pun potencijal Elixir-a nije prikazan. Radi uvida u sposobnost brze obrade velike količine podataka, potrebno je implementirati algoritme pokretati za realne skupove podataka koji se inače u ovoj oblasti obrađuju. U pitanju su milijarde nukleotida od kojih se može sastojati jedan genom.

Literatura

- [1] Adresa za preuzimanje materijala sa github-a. URL: <https://github.com/MilenaDukanac/Master-rad>.
- [2] Joe Armstrong. “Making reliable distributed systems in the presence of software errors”. PhD thesis. 2003.
- [3] David Austin. *What is DNA?* URL: <http://www.ams.org/publicoutreach/feature-column/fcarc-dna-puzzle>.
- [4] Francesco Cesarini and Simon Thompson. *Erlang Programming. A Concurrent Approach to Software Development*. O’Reilly Media, June 2009.
- [5] Robert Eklblom and Jochen B W Wolf. “A field guide to whole-genome sequencing, assembly and annotation”. In: *Evolutionary Applications* 7 (Nov. 2014), pp. 1026–1042. URL: <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4231593/>.
- [6] Martin Middendorf. “The shortest common nonsubsequence problem is NP-complete”. In: *Theoretical Computer Science* (1993), pp. 365–369. URL: <https://www.sciencedirect.com/science/article/pii/030439759390200D>.
- [7] Mahmoud Parsian. English. Ed. by O’Reilly Media. 2015, pp. 486–502.
- [8] Charles Petit. “A Brief History of Erlang and Elixir”. In: *Coder Stories* (Jan. 2019). URL: <https://www.welcometothejungle.co/en/articles/history-erlang-elixir>.
- [9] *Sequencing Coverage*. URL: <https://www.illumina.com/science/technology/next-generation-sequencing/plan-experiments/coverage.html>.
- [10] Faculty of Mathematics Students of University of Belgrade. *Uvod u bioinformatiku - beleške sa predavanja*. 2018, str. 47–68.
- [11] Wing-Kin Sung. English. Ed. by CRC Press. 2017, pp. 123–146.
- [12] Dave Thomas. *Programming Elixir 1.3. The Pragmatic Bookshelf*. 2016.

- [13] *Uputstvo za instaliranje Elixir-a.* URL: <https://elixir-lang.org/install.html>.
- [14] *What are the best Elixir IDEs/editors?* URL: <https://www.slant.co/topics/6921/~elixir-ides-editors>.
- [15] *What is DNA?* URL: <https://ghr.nlm.nih.gov/primer/basics/dna>.
- [16] Daniel R. Zerbino and Ewan Birney. In: *Genome Research* (May 2008). URL: <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2336801/>.
- [17] *Zvanična stranica kompanije Platformatec.* URL: <http://plataformatec.com.br/>.
- [18] *Zvanična stranica programskog jezika Erlang.* URL: <http://erlang.org>.
- [19] *Zvanična stranica programskog jezika Haskell.* URL: <https://www.haskell.org/>.
- [20] *Zvanična stranica programskog jezika Python.* URL: <https://www.python.org/>.
- [21] *Zvanična stranica PyCharm-a.* URL: https://www.jetbrains.com/pycharm/promo/?gclid=CjwKCAjw1_PqBRBIEiwA71rmtauYUPbqEGp8akacF41lZAV-AAJQsTXoTZGJOvnGYBwE.
- [22] Rolf Zwart. In: (Apr. 2018). URL: <https://reinout.vanrees.org/weblog/2018/04/25/origin-of-python-abc.html>.