Comenius University, Bratislava
Faculty of Mathematics, Physics and Informatics

Computational Design of Probes for the Hyb-Seq Protocol

Bachelor thesis

2017
Michaela Šandalová

Comenius University, Bratislava
Faculty of Mathematics, Physics and Informatics

Computational Design of Probes for the Hyb-Seq Protocol

Bachelor thesis

Study programme:	Bioinformatics
Study field:	2508 and 1536 Computer science and biology
Department:	Department of Computer Science
Supervisor:	Mgr. Matúš Kempa

Bratislava, 2017
Michaela Šandalová

Comenius University in Bratislava
Faculty of Mathematics, Physics and Informatics

THESIS ASSIGNMENT

Name and Surname:	Michaela Šandalová
Study programme:	Bioinformatics (Joint degree study, bachelor I. deg., full time form)
Field of Study:	Computer Science, Informatics Biology
Type of Thesis:	Bachelor´s thesis
Language of Thesis:	English
Secondary language:	Slovak

Title: Computational Design of Probes for the Hyb-Seq Protocol

Aim: The goal of the thesis is to design probes for the Hyb-Seq protocol for Alyssum genus based on raw sequencing data. The starting point is an existing software pipeline called Sondovač (https://github.com/V-Z/sondovac). The goal is to explore and describe methods used in this tool, propose and test apropriate parameter settings and potentially propose improvements of the pipeline.

Supervisor:	Mgr. Matúš Kempa
Department:	FMFI.KI - Department of Computer Science
Head of department:	prof. RNDr. Martin Škoviera, PhD.
Assigned:	26.10.2016

Approved: 26.10.2016 doc. Mgr. Bronislava Brejová, PhD.
Guarantor of Study Programme

.....
Student

.....
Supervisor

Univerzita Komenského v Bratislave
Fakulta matematiky, fyziky a informatiky

ZADANIE ZÁVEREČNEJ PRÁCE

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

Názov: Computational Design of Probes for the Hyb-Seq Protocol
Automatizovaný dizajn prób pre Hyb-Seq protokol

Cieľ: Cieľom práce je vytvoriť próby pre Hyb-Seq protokol pre rod Alyssum zo surových sekvenačných dát. Východiskom práce je existujúci nástroj Sondovač (https://github.com/V-Z/sondovac), pričom cieľom je naštudovať a popísať metódy použité v tomto nástroji, navrhnúť a otestovať vhodné nastavenia parametrov, prípadne navrhnúť vylepšenia nástroja.

Vedúci:	Mgr. Matúš Kempa
Katedra:	FMFI.KI - Katedra informatiky
Vedúci katedry:	prof. RNDr. Martin Škoviera, PhD.
Dátum zadania:	26.10.2016

Dátum schválenia: 26.10.2016 doc. Mgr. Bronislava Brejová, PhD.
garant študijného programu

.....
študent

.....
vedúci práce

Pod'akovanie: Thanks Obama.

Abstrakt

Slovenský abstrakt v rozsahu 100-500 slov, jeden odstavec. Abstrakt stručne sumarizuje výsledky práce. Mal by byť pochopiteľný pre bežného informatika. Nemal by teda využívať skratky, termíny alebo označenie zavedené v práci, okrem tých, ktoré sú všeobecne známe.

Kľúčové slová: jedno, druhé, tretie (prípadne štvrté, piate)

Abstract

Abstract in the English language (translation of the abstract in the Slovak language).

Keywords:

Contents

List of Figures

List of Tables

Introduction

Computational biology is currently a blooming discipline, its methods and tools having wide use among scientists, especially in the subject of genetics. One such use lies in taxonomy: a determination of phylogenetic relationships and evolutionary history among various species or families [5].

Research in the fields of phylogenesis and taxonomy allows us better understanding of biodiversity, evolution or ecology and aids in identification and classification of living organisms, effectively showing their differences and similarities. The analysis of evolutionary history is called phylogeny and is represented by a tree diagram called a phylogenetic tree [6]. We will provide a more detailed description of phylogenetic trees in chapter 1.

Creation of phylogenetic trees and comparing organisms in general requires a large amount of data, usually in the form of DNA or RNA sequences. In the last few yars, the price of sequencing has gone down rapidly. However, phylogenetic trees require data from several organisms, and can prove to be time and money consuming. Moreover, we are sequencing plants, which tend to have much bigger and more complex genomes, along with additional genetic information from chloroplasts and mitochondria [8].

A modern aproach to sequencing - the next generation sequencing offers reduced time and is affordable even with more data. The next generation sequencing, or NGS for short is a name for several methods that are more effective than the previously used Sanger sequencing. Specifically, Hyb-Seq protocol [10] is a method that utilizes target enrichment and genome skimming to forego sequencing all of the genome. Hyb-Seq uses specifically designed probes to find conserved places within the genome, and thus enabling comparison among the species.

These probes are usually determined from orthologous low-copy nuclear loci combined with other types of information, for example mitochondrial and plastid genomes. However, finding these loci for non-model organisms is a difficult task. Loci can be selected from transcriptomes (set of all messenger RNA molecules), genomes, gene expression studies, or the literature. There is a lack of automated bioinformatic pipelines for selection of low-copy nuclear loci.

Sondovac is a script that offers relatively easy and automated creation of othologous low-copy nuclear probes from transcriptome and genome skim data for target

enrichment [9]. Purpose of this thesis is to understand this tool, describe methods used in it and to create probes for plants from genus Alyssum using raw sequencing data. The resulting data are intended to be further used by the Hyb-Seq protocol.

In the first chapter, we will take a look at the biological motivation behind probe design and we will explain the most common terms used through the thesis and provide some information on phylogenetic trees. Chapter **??** offers insight into the biological data we are using and what we are trying to achieve with it. Chapter **??** will explain the Hyb-Seq protocol in more detail. Next, chapter **??** will take us through a detailed description of Sondovac and its methods. Chapter **??** will entail the results we got from creating probes for Alyssum. In conclusion, chapter **??** will go through possible improvements of the pipeline that is Sondovac, and their implementation.

Chapter 1

Biological motivation and background

In this chapter we will introduce the terms and concepts of biology, genetics and computational biology that are commonly used through this thesis. We will explain phylogenetic trees, their significance in evolutionary biology, their construction and their connection to the Hyb-Seq protocol and probes. In addition, we will specify the data we are using, provide a biological background for them and describe the way in which they are processed.

1.1 Terminology overview, basic terms and concepts

1.1.1 Nucleic acids

Nucleic acids carry the genetic information of all known living things. Two nucleic acids are called RNA (ribonucleic acid) and DNA (deoxyribonucleic acid). They both consist of sequence of nucleotides - monomers that are made of a pentose - a sugar with 5 carbons (ribose in RNA, deoxyribose in DNA), a phosphate group, and a nitrogenous-base. The five most common bases are cytosine (C), guanine (G), adenine (A), thymine (T) and uracil (U). RNA contains C, G, A, U and DNA contains C, G, A, T. These bases create hydrogen bonds between each other as follows: C-G, A-T in DNA, and C-G, A-U in RNA. RNA is usually single stranded and the bases pair with each other within the same strand, creating 3D-structures. Most DNAs is a double helix - it has two complementary strands that pair with each other.

1.1.2 Genes

A gene is a basic unit of heredity, a region in DNA that encodes some function, usually a protein. Several genes can encode a single trait or one gene can encode multiple traits. A position, or a place of a gene in DNA is called locus (plur. loci). Homologous genes are genes that share a common ancestor. More specificaly, homologous sequences are

called orthologs, if two copies of the same gene are in two different species and they are called paralogs, if the gene was duplicated within the same genome. During evolution, orthologs retain the same function while paralogs (or one of them) can can gain new functions. Sequences of DNA that are converted into mature mRNA are called exons. The sequences that are between exons are called introns. Introns do not code proteins and their sequences can change frequently over time. Exons, on the other hand, are much more conserved.

1.1.3 Transcription, translation

When making proteins, regions of DNA are transcribed into a shorter RNA that is complementary to the original DNA sequence. This RNA is called messenger RNA or mRNA. mRNA is then translated into proteins. The complete set of all mRNAs from a cell or a population of cells is called transcriptome. The transcriptome represents all genes that are being actively expressed.

1.2 Construction of phylogenetic trees

Here we will describe why is a construction of a phylogenetic tree important for the study of evolution. We will also look into how the tree can be made from the data we have.

1.3 Data and objectives

In this chapter we will take a closer look at the data we are using. We will cover the biological aspect of the data and compare the data we have to other kinds of data. We will also look at how such data are obtained and how they are further processed. We will define the problem we are facing from the biological point of view and why do we want to solve it.

1.3.1 The actual biological problem

In this section we will consider the actuall biological process and lab work behing the creation of probes and, by extension, the phylogenetic tree. We will answer why and how does the biological process work.

1.3.2 What kind of data we are using

This just a citation so I know what I need to address later: Script sondovac_part_a.sh requires as input files: 1) Transcriptome input file in FASTA format. Note: For techni-

cal reasons, the labels of FASTA sequences must be unique numbers (no other characters). Sondovač will check the labels, and if they are not in an appropriate form, a copy of this input file with correct labels will be created. 2) Plastome reference sequence input file in FASTA format. 3) Paired-end genome skim input file in FASTQ format (two files - forward and reverse reads). 4) OPTIONAL: Mitochondriome reference sequence input file in FASTA format. This file is not required. [9]

1.3.3 Where the data comes from

Chapter 2

Hyb-Seq Protocol

In this chapter we will describe in detail the Hyb-Seq Protocol and next generation sequencing, their pros and cons. We will address their position in the process of creating a phylogenetic tree and their connection to the Sondovac.

2.1 Orthologous genes

Orthologous genes are fundamental for phylogeny, since they can be used as markers. Selecting effective orthologous genes is difficult as gene duplication and deletion is making it hard to tell orthologs from paralogs. Single-copy paralogs that has undergone lineage-specific changes can be mistaken for orhologs. Distinguishing orthologs from paralogs is especially difficult in angiosperms, where polyploidization is a common occurence.

2.2 Low-copy nuclear genes

Low-copy nuclear genes are genes from nucleus that can be found in the genome in few or single copies. Highly conserved orthologous low-copy nuclear genes have found their use as a source of phylogenetic information. They proved to be effective markers to track organismal evolution. Most of the protein-coding cell genes can be found in nucleus. Compared to genes from other organelles, nuclear genes from eukaryotic organisms consist from several chromosomes. Because of this, genes of eucaryotic organisms are evolutionary unlinked; either on different chromosomes or sufficiently far apart.

Finding low-copy nuclear genes has been constrained by bechnical limitations. High sequencing throughput of current platforms, such as Illumina, combined with target enrichment enables us to sequence large amounts of low-copy nuclear loci effectively.

2.3 Target enrichment

Hyb-seq protocol is the combination of target enrichment and genome skimming. It enables data collection for low-copy nuclear genes and high-copy genomic targets for evolution studies and plant systematics. In hyb-seq, suitable probes are created to serve in target enrichment. The principle of target enrichment is selectively finding regions of interest in a genome before sequencing, thus making the sequencing process more effective.

[7]

2.4 Next generation sequencing

2.5 Hyb-Seq Protocol

Chapter 3

Existing work - Sondovac

In this chapter, we will take a closer look at the pipeline and the workflow of Sondovac. We will also specify the input and output data format and consider them from the informatic and more formal point of view. Finally, detailed description of the tools and software that are used by Sondovac will be listed.

3.1 Sondovac origins

Sondovac is a Czech neologism standing for "Probe-maker". It was created by Roswitha Smickl. Sondovac is an interactive and automated script to create orthologous low-copy nuclear probes for further use by the Hyb-Seq protocol. It uses transcriptome and genome skim data. Sondovac does not require strong bioinformatic skills nor high-performance computer. It is intended for either Linux or Mac OS X.

3.2 The pipeline

Sondovac is written in BASH, an Unix shell and a command language [4]. It has three parts: sondovac_part_a.sh, the geneious [2] intermediate part and sondovac_part_b.sh. Between the parts a and b of Sondovac it is necessary to manualy input the output from the part a into another software - Geneious, for data processing and then run the part b on the output from the software.

3.2.1 Sondovac part a

The part a of the script covers 6 steps:

1. Removing the transcripts that share $\geq 90\%$ sequence similarity
2. Removing the reads of plastid origin

3. Removing the reads of mitochondrial origin
4. Combining the paired-end reads
5. Matching the the unique transcripts and the filtered, combined genome skim reads sharing $\geq 85\%$ sequence similarity
6. Filtering the BLAT output
   (a) Choosing the transcript or genome skim sequences for further processing
   (b) Removing the transcripts with more than 1000 BLAT hits
   (c) Removing the transcript or genome skim BLAT hits containing masked nucleotides

3.2.2 Geneious

3.2.3 Sondovac part b

Sondovac part b covers 4 steps:

1. Retention of those contigs that comprise exons greater or equal than bait length and have a certain total locus length
2. Removal of probe sequences sharing greater or equal than 90% sequence similarity
3. Retention of those contigs that comprise exons greater or equal than bait length and have a certain total locus length
4. Removal of probe sequences sharing greater or equal than 90% sequence similarity with the plastome reference

3.3 Input and output data

3.3.1 Input data for part a

Input data for part a of Sondovac consist of 5 file, 1 of them being optional.

1. Transcriptome input file in FASTA format The file consists of several blocks with same format: On the first line of a block there is a '>' character followed with a unique description of the sequence, in this case a number. On the next few lines, there is the actual sequence composed of 'A', 'C', 'G' or 'T' characters.

2. Plastome reference sequence input file in FASTA format This file is used in both parts. It consists of a single long sequence and the starting line with '>' and the sequence's unique description.

3. Mitochondriome reference sequence input file in FASTA format (Optional) The file contains a single sequence along with the first line describing it. It is optional, because the size of a plant mitochondrial genome can vary greatly and have high rearrangement rates.

4. Paired-end genome skim input file in FASTQ format (first file)

5. Paired-end genome skim input file in FASTQ format (second file)

3.3.2 Output data

3.3.3 Input data for part b

1. Input file in TSV format (output of Geneious assembly)

2. Input file in FASTA format (output of Geneious assembly)

3. Plastome reference sequence input file in FASTA format This file is used in both parts. It consists of a single long sequence and the starting line with '>' and the sequence's unique description.

3.3.4 Output data

3.4 Use of Sondovac

Here we will write about how to use Sondovac, what modes does it run and what flags or settings can be used.

3.5 Used software and tools

Sondovac uses a broad variety of tools and scientific software packages, both freeware and payware. It is mainly coded in BASH, but it also uses smaller python scripts. We will take a closer look at what each of the tools is and what does it do in the Sondovac script.

3.5.1 Programming languages

1. BASH

2. Python

3.5.2 Tools and software

This is just structure, there will be more about each of these tools.

1. Bam2fastq
2. BLAT
3. Bowtie2 Bowtie2 is a tool used for aligning sequencing reads to longer reference sequences. [1]
4. CD-HIT
5. FASTX toolkit
6. FLASH
7. Geneious
8. Htsjdk
9. Libgtextutils
10. Picard
11. SAMtools

Chapter 4

Results and conclusion

We will summarize the results of the thesis in this chapter and describe the final data.

4.1 Using the pipeline

4.2 Resulting data

Chapter 5

Possible Improvements

In this chapter we will suggest possible improvements to the pipeline and go through the various approaches to the problem. We will also describe main parts of the actual code and algorithms used.

5.1 Autometed selection of minimum total locus length

When making probes, we usually want to get exact number of bases, because the created probes are later passed to another company, such as MYcroarray (USA) [3] that will create the actual physical probes from them. These companies usually charge for processing of a precise number of bases. To save money, it is best to fill this number to the brim. Therefore, when we have several sequences available, we want to select such sequences, that the sum of thei lenghts makes the highest number possible, but does not exceed the given limit, let us say $1,000,000$ of base pairs.

Aside from this, we often have other requirements on the selected sequences, usually based on biological evidence. There might be sequences we do not want in our selection, or even sequences we might want to add.

This selection is usually done by hand or a series of programmes. It would be beneficial, if this part was, too, automated.

5.1.1 What we want to do, scheme of the program

We want to create a script, that would work with the second part of sondovac and would be able to select the best possible sequences to fill up the limit. This script would take into account any sequences we want to keep, or those we do not want in the result.

5.1.2 Additional requirements

We need specific genes - responsible for nickel and metal binding. If these genes are present among our sequences, they have to be in the result. We will take the largest output file from sondovac_part_b: the one with the lowest minimum total locus length. From this file, we will select the required genes. The rest will be chosen by a combination of greedy algorithm and a knapsack.

Nickel and metal binding genes

Input data

About aligning

We used bowtie2 for aligning the genes to our results.

5.1.3 Input and output

5.1.4 binary search

We will not use this.

5.2 Maximalization of space

5.2.1 Greedy vs Knapsack

5.2.2 Combined approach

5.3 results

Bibliography

[1] Bowtie2), note=http://bowtie-bio.sourceforge.net/bowtie2/index.shtml,.

[2] Geneious. http://www.geneious.com/.

[3] Mycroarray. http://www.mycroarray.com/.

[4] Free Software Foundation. Bash. https://www.gnu.org/software/bash/.

[5] Colective of authors. Biological dictionary. http://www.biology-online.org/dictionary/Phylogenetics.

[6] Colective of authors. Phylogenetic tree. http://www.biology-online.org/dictionary/Phylogeny.

[7] Tao Sang. Utility of low-copy nuclear gene sequences in plant phylogenetics. *Critical Reviews in Biochemistry and Molecular Biology*, 37(3):121–147, 2002.

[8] Michael C Schatz, Jan Witkowski, and W Richard McCombie. Current challenges in de novo plant genome sequencing and assembly. *Genome biology*, 13(4):243, 2012.

[9] Roswitha Smickl. Sondovac. https://github.com/V-Z/sondovac/.

[10] Kevin Weitemier, Shannon CK Straub, Richard C Cronn, Mark Fishbein, Roswitha Schmickl, Angela McDonnell, and Aaron Liston. Hyb-seq: Combining target enrichment and genome skimming for plant phylogenomics. *Applications in Plant Sciences*, 2(9):1400042, 2014.