Comparative analysis of Gene Finding tools when applied to *Trichoderma* genomes

A thesis submitted to the
College of Graduate and Postdoctoral Studies
in partial fulfillment of the requirements
for the degree of Master of Science
in the Applied Computing of Computer Science
University of Saskatchewan
Saskatoon

By

Connor Burbridge, Dave Schneider, Tony Kusalik

Permission to Use

Disclaimer

Reference in this thesis to any specific commercial products, process, or service by trade name, trademark, manufacturer, or otherwise, does not constitute or imply its endorsement, recommendation, or favoring by the University of Saskatchewan. The views and opinions of the author expressed herein do not state or reflect those of the University of Saskatchewan, and shall not be used for advertising or product endorsement purposes.

Requests for permission to copy or to make other uses of materials in this thesis in whole or part should be addressed to:

Head of the Department of Computer Science
176 Thorvaldson Building, 110 Science Place
University of Saskatchewan
Saskatoon, Saskatchewan S7N 5C9 Canada

OR

Dean
College of Graduate and Postdoctoral Studies
University of Saskatchewan
116 Thorvaldson Building, 110 Science Place
Saskatoon, Saskatchewan S7N 5C9 Canada

Abstract

placeholder

Acknowledgements

I would like to acknowledge both Dr. Kusalik and Dave Schneider for their excellent support and mentorship during both COVID and my MSc. project. I would also like to thank Brendan Ashby and Dr. Leon Kochian for providing both data and additional financial support.

Contents

List of Tables

List of Figures

List of Abbreviations

DNA	Deoxyribonucleic acid
RNA	Ribonucleic acid
CDS	Coding sequence
Mb	Megabases
Kb	Kilobase
WGS	Whole Genome Shotgun (sequencing)
NGS	Next generation sequencing
PCR	Polymerase chain reaction
GMO	Genetically modified organism
CPU	Central processing unit
GC	Guanine cytosine
HMM	Hidden Markov model
UTR	Untranslated region
GFF	General feature format
BUSCO	Benchmarking Universal Single-Copy Orthologs
RSMI	Root, Soil and Microbial Interactions
MITE	Minature inverted-repeat transposable element
TIR	Terminal inverted repeat
TDR	Terminal direct repeat
maybe	Should I include tools like BLAST and resources like NCBI?

1 Background

Chapter one provides a brief backgound on *Trichoderma* and the context in which is being studied. Section 1.1 discusses the evolutionary mechanisms that may have led to the increased production of secondary metabolites in *Trichoderma* species, such as horizontal gene transfer, gene duplication, and transposable elements.

1.1 *Trichoderma* Species and their Evolution

Trichoderma are a genus of ascomycete fungi found in a wide variety of soils, and are well-known for their use in biomanufacturing of cellulases and hemicellulases as well as their roles as non-toxic, avirulent opportunistic plant symbionts[11][4]. Many of the features of *Trichoderma* can be attributed to the large number of secondary metabolites produced by these fungi, which are used to interact with plants and other organisms in the soil[**?**]. How some *Trichoderma* species evolved to produce more secondary metabolites than others, is a subject of frequent study.

There are many mechanisms organisms may leverage that result in gain or loss of function, and ultimately evolution of a species. One well-studied mechanism is that of horizontal gene transfer (HGT), which is the process of acquiring genetic material from another organism, rather than through inheritance from a parent organism[**?**]. HGT is a common mechanism in bacteria, but it has also been observed in fungi, including *Trichoderma* species[**?**]. Another mechanism of interest is the duplication of genes, which can result in the production of more than one copy of a gene, leading to increased expression of that gene[**?**]. This is particularly relevant in the case of secondary metabolite production, as many *Trichoderma* species have been shown to have multiple copies of genes involved in secondary metabolite production[**?**]. Lastly, transoposable elements (TEs) are another mechanism of interest, as they can insert themselves into the genome and disrupt or enhance the expression of genes[**?**]. TEs have been shown to play a role in the evolution of *Trichoderma* species, particularly in the case of secondary metabolite production[**?**].

Interestingly, both HGT and TEs tend to present themselves in regions where GC content differs from the rest of the genome, [**?**]. Abnormal GC content is also associated with other genomic features, such as centromeres[8] and repetitive regions[10], both of which may have an effect on the production of secondary metabolites in *Trichoderma* species. As a result, it is important to consider the GC content of a genome when studying *Trichoderma* species, which forms the basis of the Research Question 2.6.

Given the proximity of *Trichoderma* species to other fungi, bacteria, and plants in soils, it is possible that

Trichoderma species acquired genes which are involved in response to antagonistic and sympathetic intercellular interactions from other organisms. It is also important to note that some *Trichoderma* species have higher numbers of genes involved in secondary metabolite production than others, indicating an evolutionary divergence, making comparative analysis of *Trichoderma* species and strains a useful tool for understanding the mechanisms behind secondary metabolite production[**?**].

1.2 Secondary Metabolites

While cellular products essential to an organism's viability are of great interest, there are a vast number of cellular products that while not essential, still provide great benefit to the organism and may be necessary for survival in some situations[**?**][**?**]. These other cellular products are known as secondary metabolites, and they are involved in several roles ranging from cellular signalling to antibiotic activity, making them a frequent subject of study in pharmaceutical research. Enzymes that produce secondary metabolites are comprised of non-ribosomal peptide synthetases (NRPS) and polyketide synthetases (PKS), which synthesize amino acids and other basic enzymatic building blocks[**?**] into proteins. Genes encoding the NRPSs and PKSs responsible for production of secondary metabolites are often found in clusters within a genome, but their products remain unknown[**?**]. The study of these products is difficult, as the genes encoding the modules that make up NRPSs and PKSs are not expressed under normal laboratory conditions[**?**]. The NRPSs and PKSs that have been studied have been shown to be large enzymatic structures containing several functional modules, each responsible for a specific step in the synthesis of a protein[**?**].

Given the modular nature of NRPSs and PKSs, one can assume that many genes are responsible for the production of secondary metabolites. Thus, is is important that methods used for the identification of genes involved in secondary metabolite production are able to identify as many of those genes as possible. Failure to identify these genes may result in an incomplete understanding of the secondary metabolite production process, and may also result in the loss of potential targets for further research. With this in mind, it is important to consider the methods used for identification of genes in a genome, as the methods used can have a significant impact on the results, forming the basis of Research Questions 2.4 and 2.5.

1.3 Novel *Trichoderma* Genomes

Recently, two strains of *Trichoderma* have been identified in the prairie regions of Alberta and Saskatchewan. These two strains, named Tsth20 and DC1, have been found to have beneficial properties when used as an inoculant for plants in the soils mentioned before. In addition to these beneficial properties, the two strains mentioned previously provide even further protection for plants in dry, salty soils and one strain also has potential for use as a bioremediation tool in soils contaminated with hydrocarbon content. Bioremediation and resistance to drought tolerance has also been investigated in other strains of *Trichoderma* as well[9]. However, little is known about the mechanisms at work in these strains, so DC1 and Tsth20 were sequenced by the Global Institute for Food Security (no publication yet) in an initial attempt to better understand the details of these genomes. While this research does not eplicitly identify genomic elements related the beneficial properties of these genomes, it may serve as a foundation for future research of *Trichoderma*. The assembly of these genomes come as a result of Research Question 2.2.

1.4 Gene Predictions and Complementing Similarity Searches

Gene prediction methods are generally based on the idea that genes can be identified by searching for patterns in a sequence that match an expected gene structure or model. These gene structures begin from a 5' start codon, continue through a series of exons and introns, and end with a 3' stop codon[5]. While the basic structure of a gene may be present in a sequence, that does not mean that the gene codes for a functional protein, or in the case of secondary metabolites, a protein that produces a secondary metabolite. As a result, gene prediction methods are often used in conjunction with similarity searches, which compare the predicted genes to known genes in other organisms to identify potential functions, and serve as a form of validation for the predicted genes[5].

Similarity searches can come in several different forms, such as blast searches, which compare the predicted genes to a database of known proteins, or InterProScan, which compares the predicted genes to a database of known protein domains and binding motifs[5]. These similarity searches can provide additional information about the predicted genes, such as potential functions, and can also help to validate the predicted genes by comparing them to known genes in other organisms. talk about BUSCO tomorrow

1.5 Genome Assembly

Sequence assembly has been a long-standing problem in the field of bioinformatics[7]. Determining the correct order and combination of smaller subsequences into an accurate complete sequence assembly is computationally difficult in terms of compute resources such as memory, CPU cycles and storage required for input sequences[7]. In addition to these difficulties, there can be other issues encountered during asssembly due to the nature of the data or genomes themselves, such as low quality base calls for long read data, which is not necessarily the case today, or the inherent content of genomes themselves using repetitive regions as an example. Insufficient data may result in short, fragmented assemblies, depending on the size of the genomes, while sequence data that is not long enough can fail to fully capture repetitive regions in an assembly. A wide range of assembly tools have been developed with their own unique approaches to the genome assembly problem, so it is important to use an appropriate assembler for the task at hand, and also important to evaluate the assembly thoroughly.

Genome assembly tools generally approach the assembly problem using a graph-based approach. The most common graph-based approach is the de Bruijn graph assembly [2]. A graph in this context, is set of nodes (k-mers from sequences) connected by edges (overlaps between k-mers). Traversing through this graph results in longer subsequences that ultimately result in a set of sequences referred to as an assembly. In the early years of long read sequence data, sequencing platforms encountered difficulties producing consistently high scores for base calls when sequencing. To combat this, some assembly workflows may also include a polishing or correction step once the initial assembly is completed in which high quality short read sequences are used as supplemental information to correct low quality regions in the assembly. These low quality base calls are typically not present in modern long read sequencing approaches as the methodology and quality of calls have improved drastically. While the polishing step is arguably unnecessary in modern assemblies, the polishing programs remain available should researchers be interested in applying additional reads for polishing.

One approach to aid in the previously mentioned issue of assembly correctness is to use a combination of long and short reads in what is known as a hybrid assembly. Combining both highly accurate short reads with deep coverage along with less accurate but much longer reads can produce high quality genome assemblies that capture long repetitive regions. Hybrid assembly approaches have been shown to produce high quality assemblies in a wide variety of organisms as the combine long read data with short data to produce assemblies that properly represent long repetetive regions with additionaly high quality Illumina sequences for correction. Once assembled, the sequences must also be evaluated with measures such as N50, L50, coverage, average contig length and total assembled length to ensure that the genomes are well assembled, at least based on these metrics[7]. Following appropriate assembly protocols is essential to the further success of a project as downstream processing such as annotation depends on a high-quality assembly.

1.6 Identification of Anomolous Genomic regions

One important aspect of interest when assembling any form of sequence is GC content or percent GC of the assembled sequence. Large regions of anomolous GC content may be of interest to researchers as they may contain repetitive regions and unique features responsible for traits specific to the organism in question.

1.7 Gene Finding Methods

Gene finding (or gene annotation) has been a long standing computational problem in bioinformatics, which concerns itself with identifying potential genes within assemblies based on patterns or pre-existing experimental evidence considered by the gene finding program. This process is critical for unraveling and understanding the complex processes occurring in all forms of life. In a general sense, gene finding programs operate by searching for patters or indicators showing that a gene of feature may be present. The most basic indicators being start and stop codons, with splice sites in between should the sequence match the applied model. The results produced by gene finding tools can vary considerably for a number of reasons, including quality of the assembly, the intrinsic model used by the gene finder, filtering criteria, and even the nature of the organism and assembly itself. Given the broad applications, choice of gene finding tools, and the variability of assemblies being considered, it is important that we gain a deeper understanding of these tools prior to putting them to use.

There are two common methods for gene finding, those methods being *ab initio* methods, where programs search for patterns and gene structures, and similarity or evidence-based searches, which use prior information such as RNAseq data, expressed sequence tags and expressd protein sequences to identify genes within a new genome[3]. Complicating the process more is the introduction of introns and alternative splicing in eukaryotes, making it possible for one gene to have several possible transcripts at the same locus. An example of an *ab initio* method would be GeneMark-ES[6], while an evidence based tool would be Braker2[1]. *Ab initio* gene finders typically predict genes using a Hidden Markov Model (HMM)[3]. These predictions are based on 'signals' or features associated with a gene, such as the usual start, stop, exon and intron portions of a gene as well as upstream promoter sequences and more. In this case, these signals would be considered states in the terminology associated with HMMs. Gene finders wish to predict these states based on observations, or sequences presented to the model. HMMs in gene finding tools are trained beforehand and then applied to a sequence. This means that a gene finding program may not be trained in the context of any assembly provided to it, and thus may miss genes that are unique to the assembly in question. On the other hand, while still relying on HMMs for a 'base' set of predictions, evidence-based gene finding tools leverage new evidence that may be outside the scope of the pre-existing model[**?**]. As an example, an evidence-based model would be useful in a situation where you are interested in annotating a new assembly for a non-model organism. The addition of experimental data provides context specific to your assembly of interest while still retaining the predictions from existing HMM models.

There are also other aspects of gene finding tools that are important to consider. These include features such as whether or not the gene finders find non-coding RNAs, annotation of 5' and 3' UTR regions, and in the case of ab-initio methods, the assumptions made by the underlying models used for gene finding. These features and others can influence a user's decision on which gene finding tool to consider and will complicate comparative analysis of multiple gene finding tools. (citation needed somewhere in here)

2 Research Questions

2.1 Research Questions

With an ever-increasing number of gene prediction tools available to users, it is important to assess and understand their behaviour and performance in the context, particularly in the context of new genome assemblies of lesser studied organisms, where a reference prediction set may not be available. The main purpose of this research is to evaluate and compare gene finding tools in the context of *Trichoderma* assemblies where a gold standard set of gene predictions does not exist. To assess behaviour and performance in these contexts, we have defined X problems to profile the selected gene finding tools. In addition to applying selected gene finding tools to novel *Trichoderma* isolates, Tsht20 and DC1, we also applied selected gene finding tools to existing *Trichoderma* assemblies from the National Center for Biotechnology Information (NCBI).

2.2 *Trichoderma* Assembly Results

Since gene finding tools operate on an assembled genomic sequence, it must follow that the results will be influenced by the supplied assembly. Before applying gene finders to the new assemblies, we should first investigate the new assemblies by generating general assembly metrics for the new assemblies to contrast and compare with existing assemblies. We ask: **how do assemblies of DC1 and Tsth20 compare to existing *Trichoderma* assemblies?** With these isolates being from the *Trichoderma* family, we expect assembly metrics to be similar in nature to existing assemblies from NCBI, but do they?

2.3 Profiling of Gene Finding Tools

Different gene finding tools may predict different types of features associated with gene structures. The question arises: **which (if any) gene finders predict additional features outside of the standard gene model?** Additional features in this case include promoter sequences, transcription binding sites, activating sequences and other upstream or downstream sequences. In addition, different gene finding tools employ differing programming languages and algorithms which raises several questions. **How are these gene finding tools implemented? Is the software straightforward to install? Are the tools user-friendly? What is the processing time and memory consumption of different gene finding**

tools in the context of *Trichoderma*?

2.4 Number of Features Predicted

Do gene finders predict similar numbers of features in the context of *Trichoderma* genomes? One common method for evaluating gene finding tools is by looking at the number of features predicted by each tool. These features make an obvious point of comparison for selected gene finding tools. The term 'feature' here is somewhat ambiguous, referring to many possible categories of genomic feature. For each gene finding tool, we compare the counts for predicted genes, transcripts, and coding sequences.

2.5 Lengths of Predicted Genes

Genome assemblies can contain a wide range of gene lengths. For some users, genes of a specific length may be a key point of interest, so the ability of a gene prediction tool to capture the broad range of possible gene lengths is another important metric for comparison. Thus we ask the question: **do different gene finders predict genes of similar lengths in *Trichoderma*?**

2.6 Performance in Regions of Anomalous Sequence Content

One of the inspirations for this research is the unique composition of genomic sequence in *Trichoderma*. Results from the assembly process show that GC content in *Trichoderma* strains is abnormal throughout most assemblies. These regions of assemblies present an interesting opportunity to assess gene finding performance in regions of anomalous GC content. The question follows: **do gene finders behave differently in regions of anomalous sequence content?**

2.7 BUSCO Completeness

The use of existing benchmarks for gene finding performance is useful when assessing performance of gene finding tools, particularly in the case of genes that should be evolutionarily conserved. **Do gene finders predict conserved single-copy orthologs expected in fungal genomes?**

2.8 Identifying Regions of Agreement and Disagreement

With predicted genes from several tools available, the question we would like to ask is whether or not the gene finders agree with one another for any given prediction. To answer this question, we will identify 'regions' of overlapping predictions. A region can be defined as a start and stop position of a set of individual or overlapping features from one or more gene finding tools and external sources. With regions identified, we can determine agreement, or more importantly, disagreement in predictions between gene finding tools from which we can ask: **do gene finders agree on their predictions? If no, to what extent do they disagree? Are there genomic regions where agreement or disagreement are more prevalent?**

2.9 Validation of Predicted Genes via InterProScan

In an effort to validate, or at least provide supporting evidence for any given gene prediction we will apply InterProScan to coding sequences predicted by each of the gene finders to identify features associated with protein function. Genes will be considered as 'valid' if the gene's protein sequence contains binding sites, motifs, or other fucntional characteristics of proteins. Using the results from InterProScan, we ask the question: **in the context of *Trichoderma*, do proteins predicted by gene finders contain functional signatures?**

2.10 Selection of a Gene Finding Tool

With all the results generated, we can provide insight to the question: **which gene finding tool should one choose?**

References

[1] Tomáš Brůna, Katharina J Hoff, Alexandre Lomsadze, Mario Stanke, and Mark Borodovsky. Braker2: automatic eukaryotic genome annotation with genemark-ep+ and augustus supported by a protein database. *NAR genomics and bioinformatics*, 3:lqaa108, 3 2021.

[2] Phillip E C Compeau, Pavel A Pevzner, and Glenn Tesler. How to apply de bruijn graphs to genome assembly. *Nature biotechnology*, 29:987–991, 11 2011.

[3] Girum Fitihamlak Ejigu and Jaehee Jung. Review on the computational genome annotation of sequences obtained by next-generation sequencing. *Biology*, 9, 9 2020.

[4] Christian P Kubicek, Andrei S Steindorff, Komal Chenthamara, Gelsomina Manganiello, Bernard Henrissat, Jian Zhang, Feng Cai, Alexey G Kopchinskiy, Eva M Kubicek, Alan Kuo, Riccardo Baroncelli, Sabrina Sarrocco, Eliane Ferreira Noronha, Giovanni Vannacci, Qirong Shen, Igor V Grigoriev, and Irina S Druzhinina. Evolution and comparative genomics of the most common trichoderma species. *BMC Genomics*, 20:485, 2019.

[5] Brendan Loftus. *4 - Genome Sequencing, Assembly and Gene Prediction in Fungi*, volume 3, pages 65–81. Elsevier, 2003.

[6] Alexandre Lomsadze, Vardges Ter-Hovhannisyan, Yury O Chernoff, and Mark Borodovsky. Gene identification in novel eukaryotic genomes by self-training algorithm. *Nucleic Acids Research*, 33:6494–6506, 2005.

[7] Niranjan Nagarajan and Mihai Pop. Sequence assembly demystified. *Nature Reviews Genetics*, 14:157–167, 2013.

[8] Miroslav Plohl, Nevenka Meštrović, and Brankica Mravinac. Centromere identity from the dna point of view. *Chromosoma*, 123:313–325, 8 2014.

[9] Biancamaria Senizza, Fabrizio Araniti, Simon Lewin, Sonja Wende, Steffen Kolb, and Luigi Lucini. Trichoderma spp.-mediated mitigation of heat, drought, and their combination on the arabidopsis thaliana holobiont: a metabolomics and metabarcoding approach. *Frontiers in Plant Science*, 14, 2023.

[10] David J Winter, Austen R D Ganley, Carolyn A Young, Ivan Liachko, Christopher L Schardl, Pierre-Yves Dupont, Daniel Berry, Arvina Ram, Barry Scott, and Murray P Cox. Repeat elements organise 3d genome structure and mediate transcription in the filamentous fungus epichloë festucae. *PLOS Genetics*, 14:1–29, 2018.

[11] Sheridan L Woo, Rosa Hermosa, Matteo Lorito, and Enrique Monte. Trichoderma: a multipurpose, plant-beneficial microorganism for eco-sustainable agriculture. *Nature Reviews Microbiology*, 21:312–326, 2023.