MINING FUNGAL EFFECTOR CANDIDATES IN BIOTROPHIC PLANT PATHOGENS; RUSTS AND MILDEWS

A THESIS SUBMITTED TO
THE GRADUATE SCHOOL OF INFORMATICS
OF
MIDDLE EAST TECHNICAL UNIVERSITY

BY

SİNAN UĞUR UMU

IN PARTIAL FULFILLMENT OF THE REQUIREMENTS FOR THE DEGREE
OF MASTER OF SCIENCE
IN
BIOINFORMATICS

JULY 2012

MINING FUNGAL EFFECTOR CANDIDATES IN BIOTROPHIC PLANT PATHOGENS; RUSTS AND MILDEWS

Submitted by **Sinan Uğur Umu** in partial fulfillment of the requirements for the degree of **Master of Science in Bioinformatics, Middle East Technical University** by,

Prof. Dr. Nazife Baykal ________________________
Director, **Informatics Institute**

Assist. Prof. Dr. Yeşim Aydın Son ________________________
Head of Department, **Health Informatics**

Prof. Dr. Mahinur S. Akkaya ________________________
Supervisor, **Chemistry, METU**

Assoc. Prof. Dr. Tolga Can ________________________
Co-Supervisor, **Computer Eng., METU**

Examining Committee Members:

Assoc. Prof. Dr. Hasan Oğul ________________________
Computer Eng., Başkent University

Prof. Dr. Mahinur S. Akkaya ________________________
Chemistry, METU

Assist. Prof. Dr. Yeşim Aydın Son ________________________
Information Institute, METU

Assist. Prof. Dr. Bala Gür Dedeoğlu ________________________
Biotechnology, Ankara University

Assist. Prof. Dr. Aslıhan Günel ________________________
Chemistry, Ahi Evran University

Date: 30 July 2012

I hereby declare that all information in this document has been obtained and presented in accordance with academic rules and ethical conduct. I also declare that, as required by these rules and conduct, I have fully cited and referenced all material and results that are not original to this work.

Name, Last Name : Sinan Uğur Umu

Signature :

ABSTRACT

MINING FUNGAL EFFECTOR CANDIDATES IN BIOTROPHIC PLANT PATHOGENS; RUSTS AND MILDEWS

Umu, Sinan Uğur

MSc., Bioinformatics Program

Supervisor: Prof. Dr. Mahinur S. Akkaya

Co-Supervisor: Assoc. Prof. Dr. Tolga Can

July 2012, 77 Pages

Biotrophic plant pathogens lead to huge crop losses and they have great economical drawbacks on wheat and barley production. These pathogens rely on formation of haustoria and transfer of effector proteins into the host cells for generating disease. The main role of effector proteins is to disable plant defense mechanisms. Effector proteins contain N-terminal signal peptides and they have little sequence similarity between each other. It is vital to detect as many effector proteins as possible to understand infection and disease formation processes of biotrophic plant pathogens. To this end, sequencing of pathogen genomes are being emerged, the data will be invaluable for identifying the candidate effectors in terms of biological and biochemical roles in infection and more. There are some bioinformatics based methods available that can be utilized to classify and distinguish effectors from other pathogenic genes. It is important to understand how candidate effectors can be searched from Expressed Sequence Tags or transcriptome sequences. Hereby, our attempt is to present a pipeline in establishing a methodology. As a consequence, here we propose new candidate effectors.

In plant-pathogen interactions also miRNAs are too proving to be an important factor which cannot be neglected. During disease infection, expression levels of miRNAs are varying which in turn may be a proof of miRNA regulation of pathogen genes. Therefore, cross-kingdom RNA interference may take place between plant and pathogen. We have tested plant pathogens for possible plant miRNA availability and found their most likely targets with in the pathogen genes.

Keywords: Effectors, microRNA, plant pathogen, cross-kingdom regulation

ÖZ

BİYOTROFİK BİTKİ PATOJENLERİNDE (PAS VE KÜF) ADAY EFEKTÖR TESPİTİ

Umu, Sinan Uğur

Yüksek Lisans, Biyoenformatik Programı

Tez Yöneticisi: Prof. Dr. Mahinur S. Akkaya

Ortak Tez Yöneticisi: Doç. Dr. Tolga Can

Temmuz 2012, 77 Sayfa

Biyotrofik bitki patojenleri buğday ve arpa gibi ekonomik değeri yüksek olan bitkilerde, büyük verim kaybına yol açan canlılardır. Bu patojenlerin bulaşma sistemleri, haustoria denen organların oluşumu ve efektör adlı proteinlerin hedef hücreye iletilmesine dayanır. Efektörlerin ana amacı, konak hücrenin savunma mekanizmasının kapatılmasıdır. Bu yüzden mümkün olduğunca çok efektör protein keşfedilmesi, hastalıkların anlaşılması ve bu patojenlerle mücadele edilmesi için gerekmektedir. Sayısı hızla artan genom sekansları da efektörlerin tanınması açısından çok değerli bir rol oynamaktadır. Dizilimleri arasında benzerliği düşük olan efektörlerin tespiti için pek çok farklı biyoenformatik yöntem kullanılmaktadır. Buradaki çalışmamızda efektör tespiti için bir dizi yöntemi kullanarak bir yol geliştirdik. Ayrıca tanıttığımız bu yolu kullanarak, yeni efektör adayları tespit ettik.

Bunların dışında son yıllarda varlığı tartışılmaya başlanan, alemler arası miRNA regülasyonu konusunda da bazı çalışmalar yaptık. Bilindiği üzere hastalık esnasında bitki hücrelerinde miRNA seviyelerinin değiştiği gözlenmektedir. Bu bağlamda bitki ve patojen arasında alemler arası miRNA-mRNA etkileşimi de mümkün olabilir. Yaptığımız çalışmada hem patojende miRNA bölgesi, hem de olası miRNA regülasyonu kanıtlarını da test etmiş, hem de olası hedef genler bulduk.

Anahtar Kelimeler: Efektör proteinler, bitki patojeni, microRNA, alemler arası regülasyon

To my mother, father, brother and my wife…

ACKNOWLEDGEMENTS

First of all, I thank to my advisor Prof. Dr. Mahinur S. Akkaya for her support in my studies, encouragement for my future career, happy home parties and also for her maternal aid outside school.

I am also thankful to Assoc. Prof. Dr. Tolga Can for co-supervise and his invaluable courses in Bioinformatics.

I am grateful to Assist. Prof. Dr. Yeşim Aydın Son. Without her, Bioinformatics program certainly has no meaning. Her gentle heart sharing with her husband, Assist. Prof. Dr. Çağdaş D. Son, gives me hope for humanity.

I am especially grateful to Assoc. Prof. Dr. Hasan Oğul. He has infinite patience and I hope I can return the favor of him one day. He has great ideas with gentle mind. I am also thankful to him for allowing me as a researcher in his TÜBİTAK project.

I am thankful to my lab friends and former lab partners; Yağmur Aksoy, Kutay Öztürk, Barış Boylu, Nesibe Tekiner, Dilay Kızışar, Ulaş Maraşlı, Bayantez Dagvadorj, Çağlar Özketen, Ayşe Andaç and also Assist. Prof. Dr. Aslıhan Günel.

I also acknowledge TÜBİTAK grant under Project 110E160 for the financial support and Project 110T445.

Lastly, I am grateful to my wife, Özgün Candan Onarman Umu. She deserves the best of all and sky is the limit for her. She bared all the stress while she was also completing her own thesis.

TABLE OF CONTENTS

LIST OF TABLES

TABLES

LIST OF FIGURES

FIGURES

LIST OF APPENDIX

APPENDIX

LIST OF ABBREVIATIONS

FAO: Food and Agricultural Organization

Pst: *Puccinia striiformis* f. sp. t*ritici*

Pgt: *Puccinia graminis* f.sp. *tritici*

Pt: *Puccinia Triticina*

Bgh: *Blumeria graminis* f.sp. *hordei*

BLAST: Basic Local Alignment Search Tool

SP: Signal Peptide

MSA: Multiple Sequence Alignment

EST: Expressed Sequence Tag

ORF: Open Reading Frame

WGS: Whole Genome Sequence

NCBI: National Center for Biotechnology Information

MUSCLE: Multiple Sequence Comparison by Log- Expectation

MEGA: Molecular Evolutionary Genetics Analysis

RNAi: RNA Interference

UTR: Untranslated Region

B2G: Blast2GO

CHAPTER I

INTRODUCTION

1.1 Wheat (*Triticum* spp.)

Cereals including wheat (*Triticum* spp.), barley (*Hordeum vulgare)*, rice (*Oryza sativa* L.) and maize (*Zae mays* L.) are very essential for human nutrition and they have major impacts on culture. Agricultural revolution (or the Neolithic revolution), around 10000 BC, was an influential event that changed the course of history. Surplus food resources led to large settlements and emerge modern civilization (Lev-Yadun, 2000). According to extensive molecular markers based genetic studies, very first domestication event had taken place in Karacadağ, Diyarbakır of Turkey (Heun, 1997).

Today much more areas of land are used for wheat production than any other crops including rice, maize and potatoes. Wheat growth generally located between the latitudes of 30° and 60°N and 27° and 40°S but it is also possible outside these limits. The optimum growth temperature is 25 °C, and it needs moisture during growth cycle; however, too much water leads to formation of diseases and rot (Curtis et al., 2002).

Wheat is a unique crop in different aspects. It is grown on more than 240 million hectares and world trade is more than all other crops combined. The wheat kernel contains gluten, a form of protein, initiating raise of dough. This is the best compared to all other cereals and offers better nourishment than any other food source. Wheat is a leading dietary component due to its agricultural adaptability, easiness of grain storage and flour production and capacity of being main ingredient for variety of foods. It has various vitamins and minerals as well as carbohydrate, protein and fiber (Curtis et al., 2002).

Not only in the world economy, but especially also for Turkey, wheat is a major component of economy and human consumption. Turkey is one of the top ten wheat producers in the world (Table 1).

Table 1. Top ten wheat producers.

Country	**Tonnes**
China	115180303
India	80710000
United States of America	60102600
Russian Federation	41507600
France	38207000
Germany	24106700
Pakistan	23310800
Canada	23166800
Australia	22138000
Turkey	*19660000*

Source**:** Statistics from (FAO, 2010)

1.2 Barley (*Hordeum vulgare*)

Barley (*Hordeum vulgare*) is an important cereal grain that has a substantial role in livestock feeding and beer production. Cultivated barley (*Hordeum vulgare* L.) originated from *Hordeum spontaneum* C. Koch, which is one of the first agricultural crops and seen in historical records 8th and 7th millennia BC. It was also one of the major crops convey the beginning of agriculture in Europe during 6th and 5th millennia BC (Jones et al., 2011).

Barley has a great economic value and Turkey is one of the top ten barley producers in the world too (Table 2).

Table 2. Top ten barley producers.

Country	**Tonnes**
Germany	10412100
France	10102000
Ukraine	8484900
Russian Federation	8350020
Spain	8156500
Canada	7605300
Australia	7294000
Turkey	*7240000*
United Kingdom	5252000

Source**:** Statistics from (FAO, 2010)

1.3 Biotrophic Plant Pathogens of Cereals and Diseases

Obligate parasitical plant pathogens cause most dangerous infectious diseases yet it is very hard to investigate them because they cannot be cultured outside of the host cells. The three important groups of biotrophic parasites are the powdery mildew, rust fungi and the downy mildews. They cause huge cereal production losses and economical drawbacks (Ridout et al., 2006).

Biotrophic plant pathogens live in close contact with their targets. They rely on transferring effector proteins into host cells and successful formation of haustorium, thereby generating diseases (Godfrey et al., 2010).

1.3.1 The Wheat Rusts

Wheat rust pathogens are members of genus Puccinia, family Pucciniaceae, order Uredinales and class Basidiomycetes. These fungi are highly specialized with very narrow target hosts. The causal organism of wheat stem rust (also called black rust or summer rust) is *Puccinia graminis* f. sp. *tritici* (Pgt). It is the first sequenced representative of the rust fungi. Pgt forms in the uredinium at the end of season or hostile conditions. It flourishes in humid conditions and warmer temperatures between 15 °C and 35 °C. Pgt can devastate 50 percent of yield and 100 percent of damage can occur in susceptible cultivars (Curtis et al., 2002).

Puccinia triticina (Pt) is the causative pathogen of leaf rust (also called brown rust). It develops swiftly between 10 °C and 30 °C. Leaf rust is present at some extent where wheat is grown. Losses due to disease are generally less than 10 percent but may be as high as 30 percent. It affects both durum and bread wheat (Curtis et al., 2002).

Stripe or yellow rust is caused by *Puccinia striiformis* f. sp. *tritici* (Pst) which favors cold climates. Due to its early occurrence, weakened and stunted plants often follow disease. Losses can be 50 percent and in some extreme conditions 100 percent losses can occur (Curtis et al., 2002).

Table 3. The rust diseases of wheat, their primary and alternate hosts and symptoms.

Disease	Pathogen	Primary hosts	Alternate hosts	Symptoms
Leaf rust	*Puccinia triticina*	Bread and durum wheats, triticale	*Thalictrum, Anchusa, Isopyrum, Clematis*	Isolated uredinia on upper leaf surface and rarely on leaf sheaths
Stem rust	*Puccinia graminis* f.sp. *tritici*	Bread and durum wheats, barley, triticale	*Berberis vulgaris*	Isolated uredinia on upper and lower leaf surfaces, stem and spikes
Stripe rust	*Puccinia striiformis* f.sp. *tritici*	Bread and durum wheats, triticale, a few barley cultivars	Unknown	Systemic uredinia on leaves and spikes and rarely on leaf sheaths

Source: Directly from Curtis et al., 2002.

1.3.2 Barley Powdery Mildew

Blumeria graminis f. sp. *hordei* (Bgh), the pathogen that causes barley powdery mildew, relies on formation of haustorium inside the host cell (Ridout et al., 2006). Successful haustorium formation is essential to take up nutrients from target cell. Due to barley's economic value; Bgh is the most extensively investigated powdery mildew fungi (Zhang et al., 2005).

Conidiospores are haploid, asexual form of fungus and distributed by wind during growth season (Ridout et al., 2006). The life cycle of Bgh is in Figure 1.

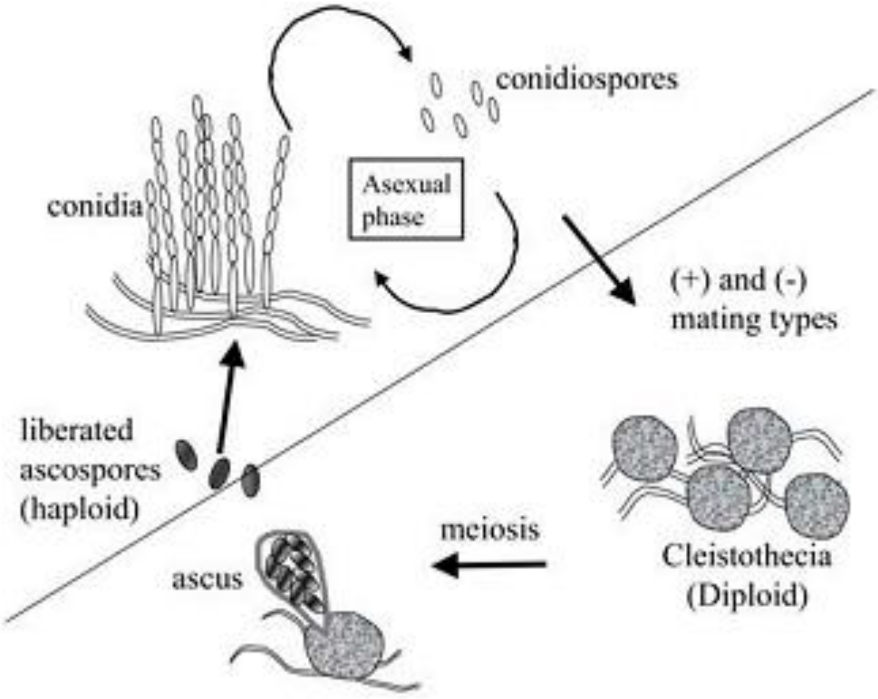


Figure 1. The life cycle of *Blumeria graminis* f. sp. *hordei*.

Source**:** Directly from Ridout et al., 2006.

In Figure 1, the line separates haploid and diploid stages of pathogen. In haploid stage, by forming of haustorium, pathogen feeds on epidermal cells of host plant and distributes conidiospores (Ridout et al., 2006).

1.4 Pathogen Effector Proteins

As mentioned before, all biotrophic plant pathogens depend on formation of haustorium (plural haustoria), which is a pocket like specialized feeding organ to take up nutrients from host cell. Successful haustoria growth is in parallel with transferring effector proteins into target cells. The main role of effector proteins is to disable plant defense mechanism. Powdery mildew, Rust fungi and Oomycete pathogens all develop haustoria inside host cells and together with this event, host cells create a membrane of unknown origin that surrounds haustoria. “Type three secretion system” (T3SS) is used to inject effectors by bacterial pathogens, but it is not clear how haustoria forming fungal pathogens achieve this transfer. It is a known fact that Oomycete effectors contain N-terminal signal peptides for secretion and use the default secretory pathway. In addition to that, Oomycete effectors contain amino acid double motif (RxLR-dEER), located a few amino acids downstream of signal peptide cleavage site. Both bacterial and Oomycete effector candidates are commonly small in their matured condition and they infrequently have homologues proteins in other species (Godfrey et al., 2010).

Unfortunately, there are not too many identified effectors and effector candidates available from haustoria-forming fungal pathogens due to difficulty of isolation. However, Bgh is very suitable for sequencing since it only attacks epidermal cell layer of plant. Therefore, it is possible to construct a library with enriched pathogen genes since epidermal cells can easily be separated. A set of 107 effector candidates had been identified in Bgh by EST sequencing. In addition to that 178 wheat stem rust (Pgt) and 57 wheat leaf rust (Pt) effector candidates were found. The analysis of these sequence sets show that, all contain N-terminal Y/F/WxC motif downstream of signal peptide cleavage site. Thus, they are also called Y/F/WxC-effector candidates (Godfrey et al., 2010).

1.5 Basic Local Alignment Search Tool

Basic Local Alignment Search Tool (BLAST) algorithm is a way to search DNA and protein sequence databases; it is faster than FASTA but equally sensitive. It is a heuristic algorithm like FASTA but it does not guarantee to find an optimal solution like dynamic programming algorithms. BLAST algorithm first looks for common words (k-tuples) in the query and target database sequences that increases speed of sequence alignment. BLAST restricts the query to the words that are the most significant while FASTA looks for all possible words. Significance is determined by

BLOSUM62 substitution matrix for protein sequences. The word lengths are 3 for protein and 11 for nucleotide sequences and these lengths are enough to find both significant and relatively short patterns (Mount, 2007). Table 4 shows specified BLAST algorithms and their properties.

Table 4. BLAST programs provided by NCBI.

Program	**Query sequence**	**Database**	**Type of alignment**
BLASTP	Protein	Protein	Gapped
BLASTN	Nucleic acid	Nucleic acid	Gapped
BLASTX	Translated nucleic acid	Protein	Each frame gapped
TBLASTN	Protein	Translated nucleic acid	Each frame gapped
TBLASTXc	Translated nucleic acid	Translated nucleic acid	Ungapped

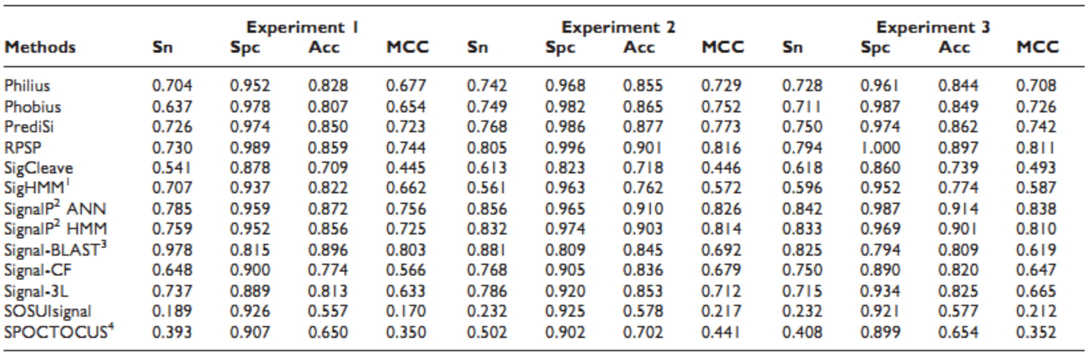
Source**:** Directly from (Mount, 2007)

1.6 SignalP

SignalP is one of the most popular secretory protein detection tools. It uses machine learning approach to predict possible signal peptide (SP) site. SignalP predicts both cleavage site and classification of secretory or non-secretory proteins (Bendtsen et al., 2004). Signal peptides are located N-terminus of non-mature protein sequences. Both prokaryotic and eukaryotic cells use this short peptide segments to achieve targeting and translocation. SPs are cut off from their passenger protein after getting into target location. In protein databases identification of SPs are very important annotation step. However, vast number of unprocessed sequences easily overcomes experimental methods to verify those sequences. Signal peptide prediction tools were developed due to requirement of faster SP annotation requirements (Choo et al., 2009).

Now there is SignalP version 4.0 available. It has purely neural network based method (Petersen et al., 2011). Benchmarking on SP detection tools show that, SignalP is more consistent and superior than others and it is able to more successfully distinguish cleavage sites of sequences as well (Choo et al., 2009)

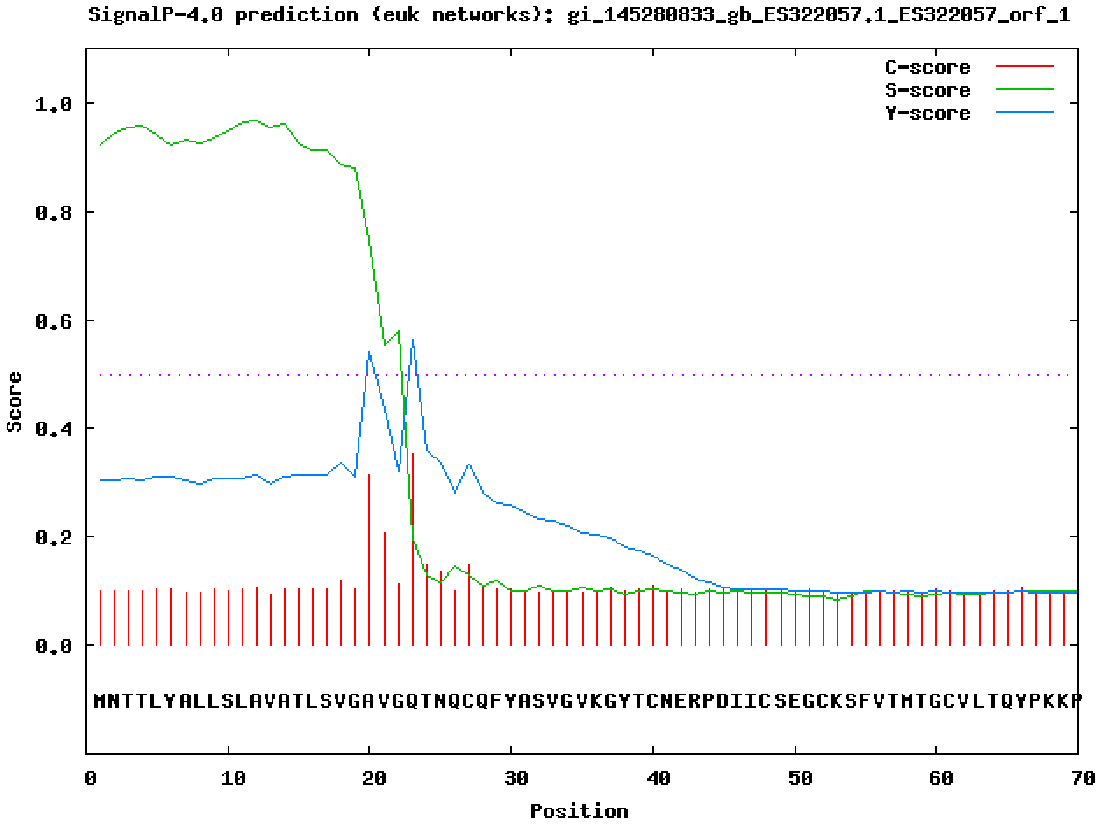
(Figure 2). SignalP ANN outperforms all other methods in all experiments referenced (Choo et al., 2009).



	Experiment 1				Experiment 2				Experiment 3			
Methods	Sn	Spc	Acc	MCC	Sn	Spc	Acc	MCC	Sn	Spc	Acc	MCC
Philius	0.704	0.952	0.828	0.677	0.742	0.968	0.855	0.729	0.728	0.961	0.844	0.708
Phobius	0.637	0.978	0.807	0.654	0.749	0.982	0.865	0.752	0.711	0.987	0.849	0.726
PrediSi	0.726	0.974	0.850	0.723	0.768	0.986	0.877	0.773	0.750	0.974	0.862	0.742
RPSP	0.730	0.989	0.859	0.744	0.805	0.996	0.901	0.816	0.794	1.000	0.897	0.811
SigCleave	0.541	0.878	0.709	0.445	0.613	0.823	0.718	0.446	0.618	0.860	0.739	0.493
SigHMM[1]	0.707	0.937	0.822	0.662	0.561	0.963	0.762	0.572	0.596	0.952	0.774	0.587
SignalP[2] ANN	0.785	0.959	0.872	0.756	0.856	0.965	0.910	0.826	0.842	0.987	0.914	0.838
SignalP[2] HMM	0.759	0.952	0.856	0.725	0.832	0.974	0.903	0.814	0.833	0.969	0.901	0.810
Signal-BLAST[3]	0.978	0.815	0.896	0.803	0.881	0.809	0.845	0.692	0.825	0.794	0.809	0.619
Signal-CF	0.648	0.900	0.774	0.566	0.768	0.905	0.836	0.679	0.750	0.890	0.820	0.647
Signal-3L	0.737	0.889	0.813	0.633	0.786	0.920	0.853	0.712	0.715	0.934	0.825	0.665
SOSUIsignal	0.189	0.926	0.557	0.170	0.232	0.925	0.578	0.217	0.232	0.921	0.577	0.212
SPOCTOCUS[4]	0.393	0.907	0.650	0.350	0.502	0.902	0.702	0.441	0.408	0.899	0.654	0.352

Figure 2. Comparison of signal peptide detection methods.

Source**:** Directly from Choo et al., 2009.



S-score: Amino acid score if it is a part of a SP or not; C-score: Cleavage score, it shows possible splitting point of SP; Y-max: Combination of S and C score, it is better cleavage site prediction; S-mean: Mean of S scores; D-score: Weighted average of S-mean and Y-max score, it is better to distinguish secretory and non-secretory proteins.

Figure 3. Example output of SignalP-4.0.

Ideally, all scores generated by SignalP, have to be high enough to consider one as a secretory protein. All of them are out of 1 and the graphic shows possible cut off position of SP. C score is at its highest at cut off position as expected in Figure 3.

The latest version of SignalP program is located at the online server (http://www.cbs.dtu.dk/services/SignalP-4.0/). It is also possible to run offline version in a Linux machine for batch jobs. Online version has 2000 sequence limitation. SignalP can also produce mature sequences in FASTA format.

1.7 Pairwise Alignment (Dynamic Programming)

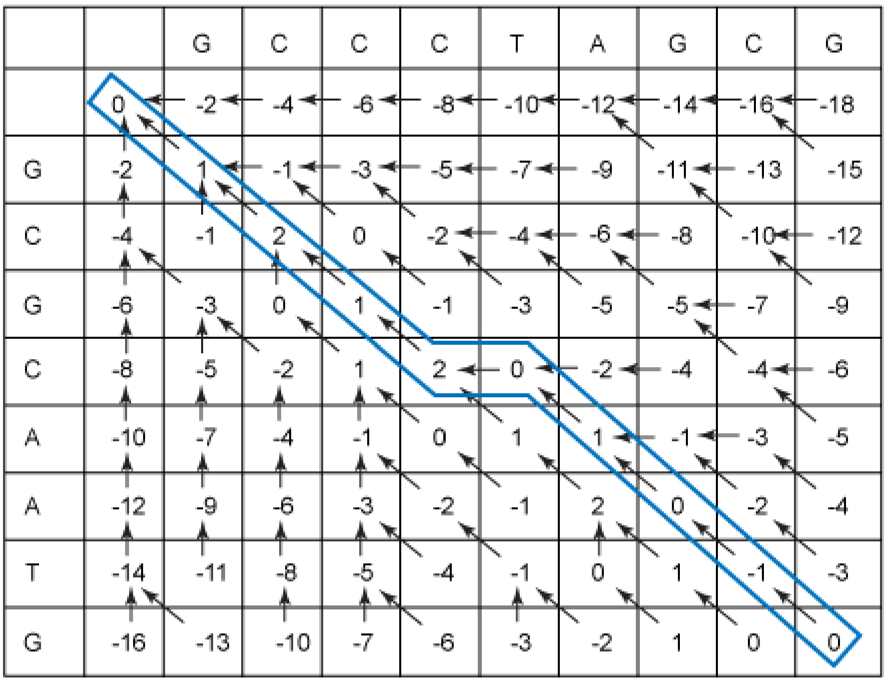
The basic sequence analysis method is to test the relation of two sequences. This is achieved by aligning two sequences or a part of them. The key issues of pairwise alignment are alignment sorts, scoring system, the algorithm and the statistical methods to evaluate significance (Durbin et al., 1998).

The algorithms that find optimal solutions *via* additive alignment score called dynamic programming. Dynamic programming algorithms find optimal solutions and optimal score of the alignment. In some cases, heuristic algorithms can perform the same with dynamic programming algorithms and they are faster than dynamic programming. Needleman-Wunsch global alignment algorithm (Equation 1) and Smith-Waterman local alignment algorithm (Equation 2) are two dynamic programming methods to align sequences (Durbin et al., 1998).

Equation 1. Needleman-Wunsch algorithm.

$$F(i,j) = max\begin{cases} F(i-1,j-1)+s(x_i,y_j), \\ F(i-1,j)-d, \\ F(i,j-1)-d. \end{cases}$$

Source**:** (Durbin et al., 1998).



		G	C	C	C	T	A	G	C	G
	0	-2	-4	-6	-8	-10	-12	-14	-16	-18
G	-2	1	-1	-3	-5	-7	-9	-11	-13	-15
C	-4	-1	2	0	-2	-4	-6	-8	-10	-12
G	-6	-3	0	1	-1	-3	-5	-5	-7	-9
C	-8	-5	-2	1	2	0	-2	-4	-4	-6
A	-10	-7	-4	-1	0	1	1	-1	-3	-5
A	-12	-9	-6	-3	-2	-1	2	0	-2	-4
T	-14	-11	-8	-5	-4	-1	0	1	-1	-3
G	-16	-13	-10	-7	-6	-3	-2	1	0	0

Figure 4. Needleman-Wunsch filled matrix and trace-back.

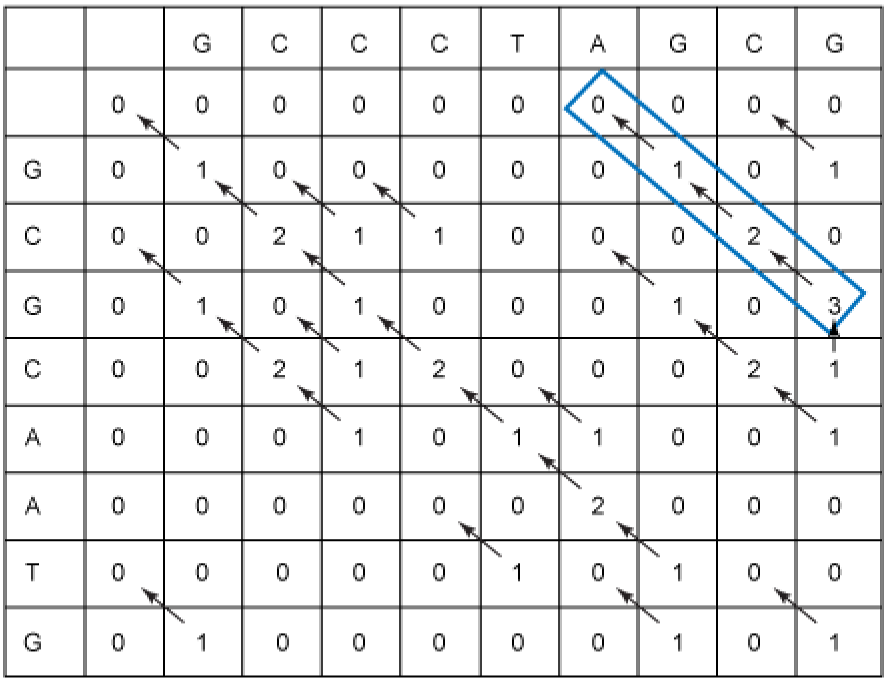
Needleman-Wunsch global alignment matrix, number at right bottom corner is alignment score. To find optimal alignment, a trace-back to upper left corner is required. Cell score that generates current cell score is the previous correct position of the matrix path which is denoted by arrows. The matrix is filled according to Equation 1 with scores for gap -2, mismatch -1 and match +1. The first row and column of the matrix is filled gap score -2; so, it sums up -2 in consecutive cells. Guanine is the first residue for both sequences which means a match score +1; thus, at that position alignment score becomes 1 and according to algorithm, the maximum score is +1. In every cell of the matrix, the scores are calculated similar and it continues as such to fill matrix completely.

Source: The matrix from (http://www.ibm.com/developerworks/java/library/j-seqalign/index.html).

Equation 2. Smith-Waterman algorithm.

$$F(i,j) = max \begin{cases} 0, \\ F(i-1,j-1) + s(x_i, y_j), \\ F(i-1,j) - d, \\ F(i,j-1) - d. \end{cases}$$

Source**:** (Durbin et al., 1998).



		G	C	C	C	T	A	G	C	G
	0	0	0	0	0	0	0	0	0	0
G	0	1	0	0	0	0	0	1	0	1
C	0	0	2	1	1	0	0	0	2	0
G	0	1	0	1	0	0	0	1	0	3
C	0	0	2	1	2	0	0	0	2	1
A	0	0	0	1	0	1	1	0	0	1
A	0	0	0	0	0	0	2	0	0	0
T	0	0	0	0	0	1	0	1	0	0
G	0	1	0	0	0	0	0	1	0	1

Figure 5. Smith-Waterman filled matrix and trace-back.

Smith-Waterman local alignment trace-back is similar with global alignment. This time maximum score number is the first location to start trace-back. Cell score that generates current cell score is the previous correct position of the matrix path which is denoted by arrows. Equation 2 generates this matrix with scores gap -2, mismatch -1 and match score +1. Zero is the lowest score possible local alignment matrix. First row and column were filled with 0 due to mismatch. In this example, the number 3 is the maximum alignment score. Ergo, "GCG" is the optimal local alignment here. Source: The matrix from (http://www.ibm.com/developerworks/java/library/j-seqalign/index.html).

1.8 Multiple Sequence Alignment

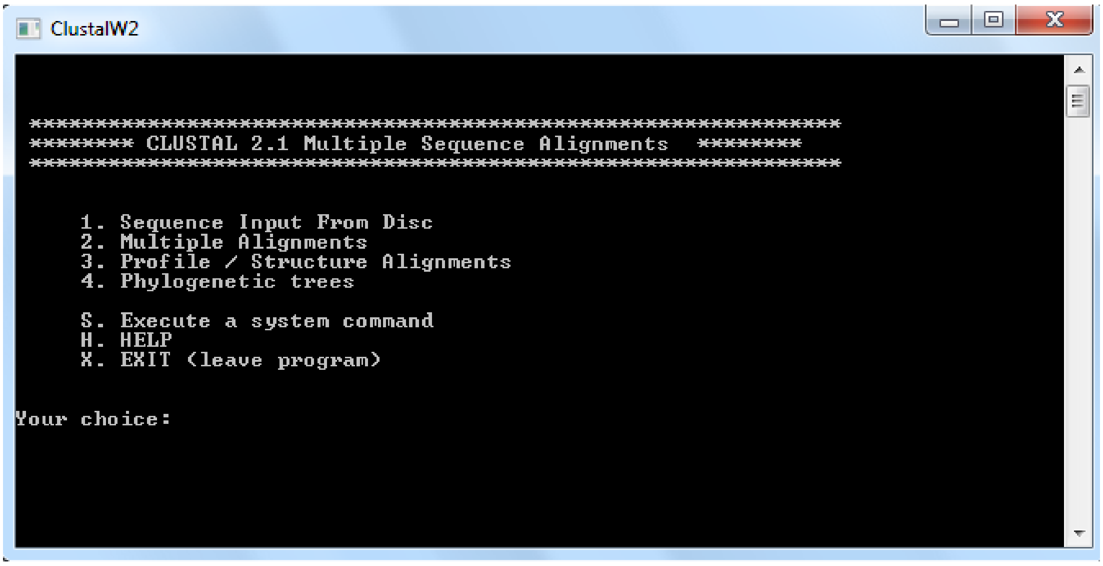
In a multiple sequence alignment (MSA), homologous residues (amino acids or nucleotides) in a set of sequences aligned together. Preferably, a column of aligned residues hold same structural positions and originate from same ancestor (Durbin et al., 1998). There are different algorithms and tools available. Most of them use progressive methods but also Hidden Markov Model (HMM) based algorithms are available too.

1.8.1 Clustal

The Clustal series of tools are extensively used for MSA of both protein and nucleic acid sequences in molecular biology. Their popularity is based on features like easy to use, robustness, multi-platform and online accessibility (Chenna et al., 2003). Clustal is the oldest of the currently used MSA tools and even it was distributed in a floppy disk at late 1980s. All Clustal derivations are based to ClustalW that uses a position-specific scoring scheme and a weighting scheme and it is a progressive method. Clustal also has a graphical user interface which developed at 1997. At the end of 90s, ClustalW (command line version) and ClustalX (visual version) were the most popular MSA programs (Larkin et al., 2007).

As mentioned, Clustal programs are easy to use and they can read FASTA, EMBL and SWISS-PROT database formats. Although Clustal programs are used widely, it does not mean it always produce best alignments. Clustal requires collinear sequences which means similar protein domains have to be in same order; otherwise, it may produce incorrect MSA (Jeanmougin et al., 1998).

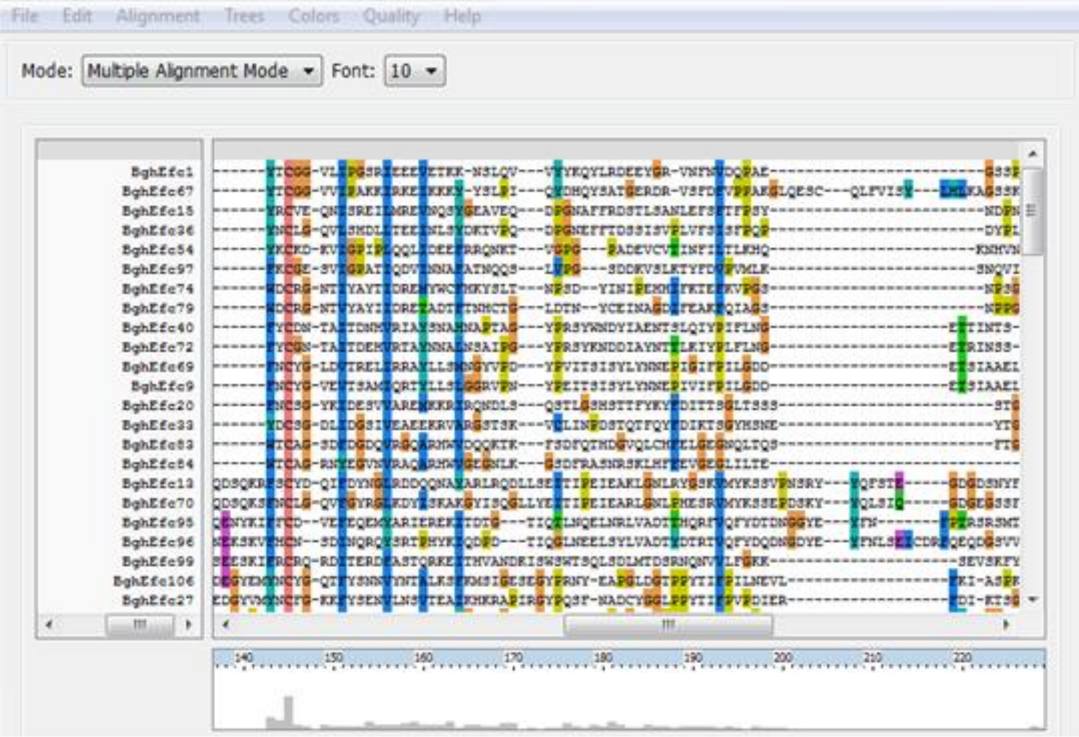
Figure 4 shows command line version of Clustal: ClustalW. It takes source file of sequences by pressing first command and the other sections to select alignment parameters, phylogenetic tree parameters. In Figure 5, graphical user interface (GUI) of ClustalX is presented, it has similar outputs and options but it could visualize the process.



```
 **************************************************************
 ******** CLUSTAL 2.1 Multiple Sequence Alignments  ********
 **************************************************************


     1. Sequence Input From Disc
     2. Multiple Alignments
     3. Profile / Structure Alignments
     4. Phylogenetic trees

     S. Execute a system command
     H. HELP
     X. EXIT (leave program)


Your choice:
```

Figure 6. A screen-shot of command line version of Clustal: ClustalW.

This image is obtained from Windows version of ClustalW, first option reads a FASTA file; second starts MSA and adjust alignment parameters; third option is for profile alignments and last option for phylogenetic trees where it is possible to change algorithm and bootstrap tree.



Mode: Multiple Alignment Mode

Font: 10

Figure 7. A screen-shot of ClustalX.

ClustalX has same functionality with command line version. However, GUI makes it easy to align and realign. Selecting and realigning of sequences are also possible. Conserved regions are indicated with corresponding colors.

There is an online version of Clustal available at EBI. It is possible to download guide tree and alignment file in different formats. From main menu, alignment parameters could be adjusted to tweak MSA. It has similar functionality with local versions.

1.8.2 MUSCLE

Multiple Sequence Comparison by Log- Expectation (MUSCLE) algorithm uses distance estimation using *k*mer counting, progressive aligment using the log-expectation score. MUSCLE claims to perform better than other MSA programs (Edgar, 2004). It is considered better on protein alignments.

A *k*mer distance for unaligned pair and the Kimura distance for an aligned pair are two distance measures used by MUSCLE. Related sequences generally have more mutual than estimated by chance. The *k*mer distance acquired from the element of kmers in common in a dense alphabet. This idea does not need an alignment and gives a speed advantage. Then for an aligned pair of sequences, the Kimura correction is applied for a single site. Distance matrices are clustered by UPGMA algorithm. MUSCLE uses a profile function called the log-expectation score; $\mathrm{LE}xy = (1 - fx_{\mathrm{G}})\ (1 - fy_{\mathrm{G}})\ \log\ \Sigma\, i\, \Sigma\, j\ fxi\ fyj\ pij/pi\ pj.$ The function for log-average is, $\mathrm{LA}xy = \log\ \Sigma\, i\, \Sigma\, j\ \alpha xi\ \alpha yj\ pij/pi\ pj$ and MUSCLE uses 240 PAM VTML matrix (Edgar, 2004).

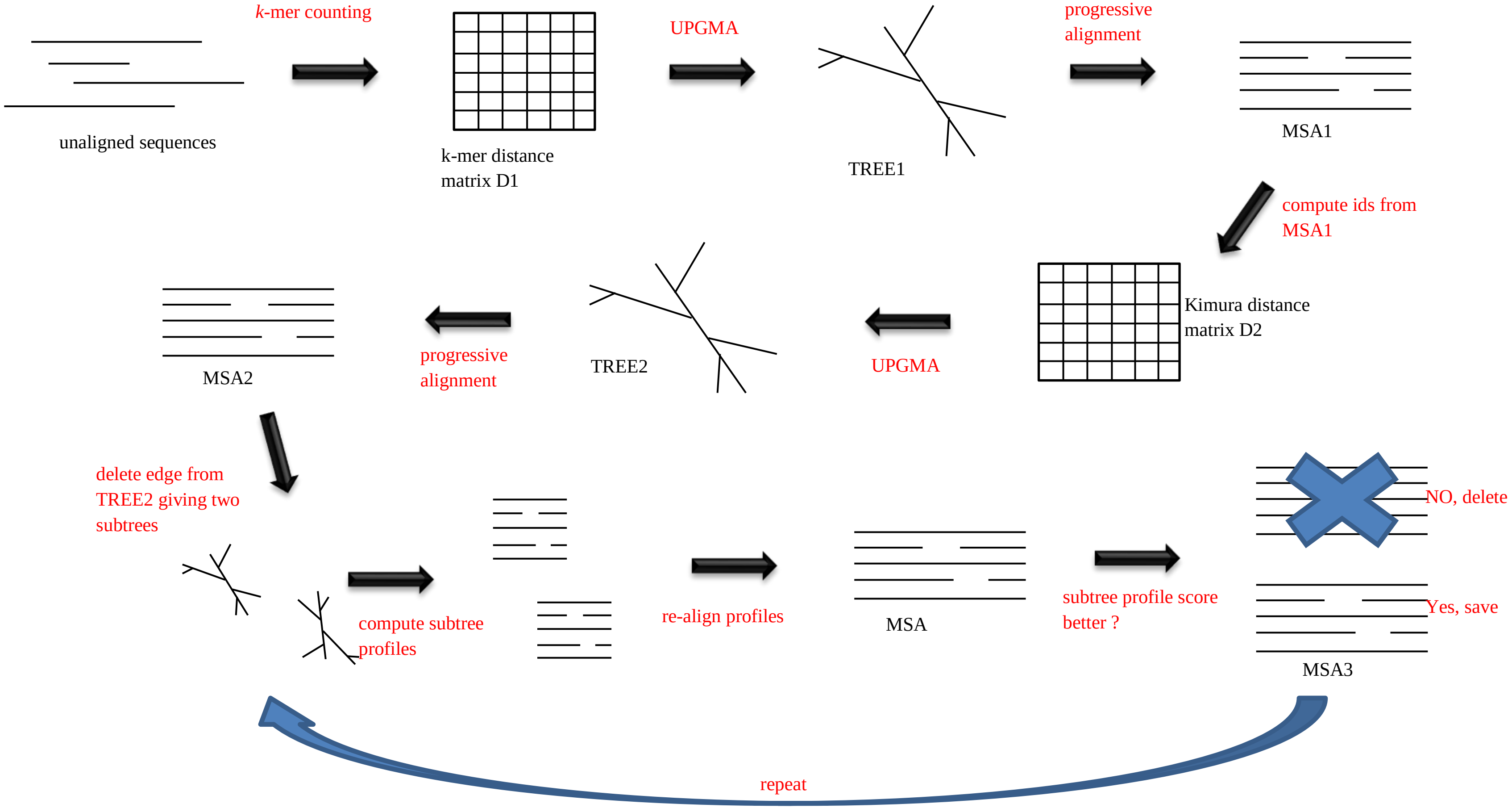


Figure 8. The flow chart of MUSCLE algorithm.

Source: Revised version of image from Edgar, 2004.

1.9 BioPython

Python is a high level programming language and it is well accepted in academic and bioinformatics world. It has object-oriented features, easy syntax and wide collection of modules (Cock et al., 2009). Without open source programming, it is hard to research in Bioinformatics. The open source projects make it easier to create custom pipelines or analysis. There are bio toolkits of common programming languages such as BioJava, BioPerl, BioPyhton (Mangalam, 2002).

BioPython project is a mature open source project that provides many different Python libraries to solve bioinformatics problems. It also uses BioSQL, which is a generic schema to store sequences, annotations and features, to retrieve and store data. BioPyhton could read many different common file formats to manipulate them. It can interact with popular databases like the NCBI Entrez Utilities, ExPASy, InterPro, KEGG and SCOP. It can call NCBI Blast and command line ClustalW via wrapper (Cock et al., 2009) .

Table 5. Selected formats manipulated by BioPython.

Format	Read/Write
fasta	R+W
genbank	R+W
embl	R
swiss	R
clustal	R+W
phylip	R+W
stockholm	R+W
nexus	R+W

Source**:** Revised version of original table from Cock et al., 2009.

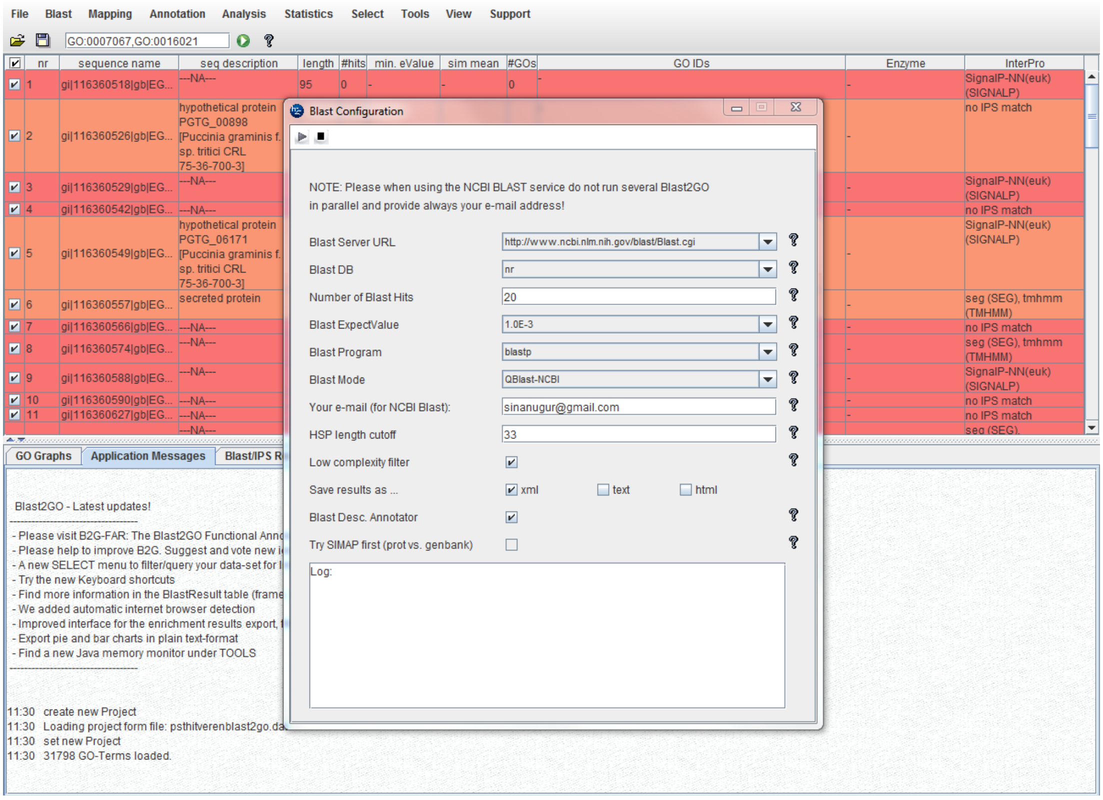
1.10 Blast2GO

Blast2GO (B2G) is a tool designed to enable Gene Ontology (GO) based data acquisition without any GO records. B2G combine GO annotation based similarity search with statistics and visualization. It is a Java based desktop application. B2G is freely accessible from “blast2go.org” (Conesa et al., 2005).

Functional annotation permits categorization of genes in functional classes that are suitable to understand physiological significance of vast amount of genes and to evaluate functional difference between sequences. Gene Ontology offers such a framework for that kind of analysis. B2G make high-throughput sequence annotation of non-model species with advanced functionalities, visualization and statistical framework. Therefore, B2G designed to allow automatic and high-throughput sequence annotation and incorporate functionality for annotation-based data mining (Conesa et al., 2005).

First step in B2G is blasting loaded sequences (Figure 9). Online NCBI or local BLAST databases can be used. To get GO terms for associated hits, mapping is made. Lastly, annotation step is performed (Conesa et al., 2005).

Figure 10 shows an application overview chart. In the middle of the chart all steps are numbered in an order to explain flow of Blast2GO.



☑	nr	sequence name	seq description	length	#hits	min. eValue	sim mean	#GOs	GO IDs	Enzyme	InterPro
☑	1	gi|116360518|gb|EG...	---NA---	95	0	-	-	0	-	-	SignalP-NN(euk) (SIGNALP)
☑	2	gi|116360526|gb|EG...	hypothetical protein PGTG_00898 [Puccinia graminis f. sp. tritici CRL 75-36-700-3]							-	no IPS match
☑	3	gi|116360529|gb|EG...	---NA---							-	SignalP-NN(euk) (SIGNALP)
☑	4	gi|116360542|gb|EG...	---NA---							-	no IPS match
☑	5	gi|116360549|gb|EG...	hypothetical protein PGTG_06171 [Puccinia graminis f. sp. tritici CRL 75-36-700-3]							-	SignalP-NN(euk) (SIGNALP)
☑	6	gi|116360557|gb|EG...	secreted protein							-	seg (SEG), tmhmm (TMHMM)
☑	7	gi|116360566|gb|EG...	---NA---							-	no IPS match
☑	8	gi|116360574|gb|EG...	---NA---							-	seg (SEG), tmhmm (TMHMM)
☑	9	gi|116360588|gb|EG...	---NA---							-	SignalP-NN(euk) (SIGNALP)
☑	10	gi|116360590|gb|EG...	---NA---							-	no IPS match
☑	11	gi|116360627|gb|EG...	---NA---							-	no IPS match
			---NA---								seg (SEG),

NOTE: Please when using the NCBI BLAST service do not run several Blast2GO in parallel and provide always your e-mail address!

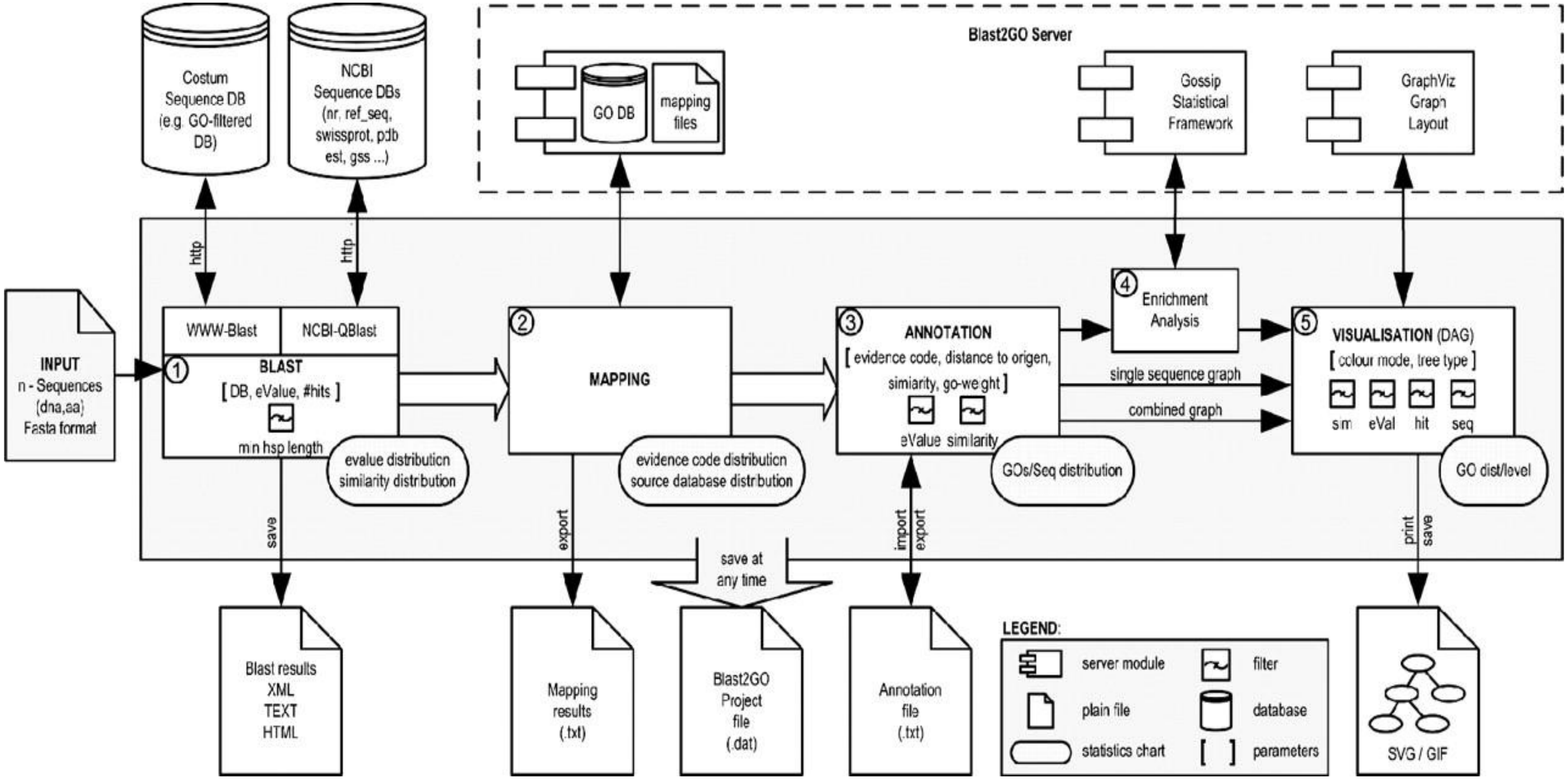
Blast Server URL: http://www.ncbi.nlm.nih.gov/blast/Blast.cgi
Blast DB: nr
Number of Blast Hits: 20
Blast ExpectValue: 1.0E-3
Blast Program: blastp
Blast Mode: QBlast-NCBI
Your e-mail (for NCBI Blast): sinanugur@gmail.com
HSP length cutoff: 33
Low complexity filter: ☑
Save results as ...: ☑ xml ☐ text ☐ html
Blast Desc. Annotator: ☑
Try SIMAP first (prot vs. genbank): ☐
Log:

```
Blast2GO - Latest updates!
--------------------------------
- Please visit B2G-FAR: The Blast2GO Functional Anno
- Please help to improve B2G. Suggest and vote new i
- A new SELECT menu to filter/query your data-set for l
- Try the new Keyboard shortcuts
- Find more information in the BlastResult table (frame
- We added automatic internet browser detection
- Improved interface for the enrichment results export, f
- Export pie and bar charts in plain text-format
- Find a new Java memory monitor under TOOLS
--------------------------------

11:30   create new Project
11:30   Loading project form file: psthitverenblast2go.da
11:30   set new Project
11:30   31798 GO-Terms loaded.
```

Figure 9. BLAST step of Blast2GO.

In Figure 9, BLAST algorithm, expected value threshold, server, HSP length, hit number etc. could be adjusted. In addition to that, output file location and type can be selected for saving results. BLAST step takes time if there are a lot of sequences loaded to Blast2GO. There is also a possibility to run Blast2GO locally.

Figure 10. Application overview of Blast2GO.



Source: Directly from Conesa et al., 2005.

1.11 MEME Suite

The MEME Suite web server is a complete set of tools that is used to discover new motifs, search sequence databases with motifs, compare a motif with a database of motifs, annotate motif with Gene Ontology and analyze motif enrichment (Bailey et al., 2009).

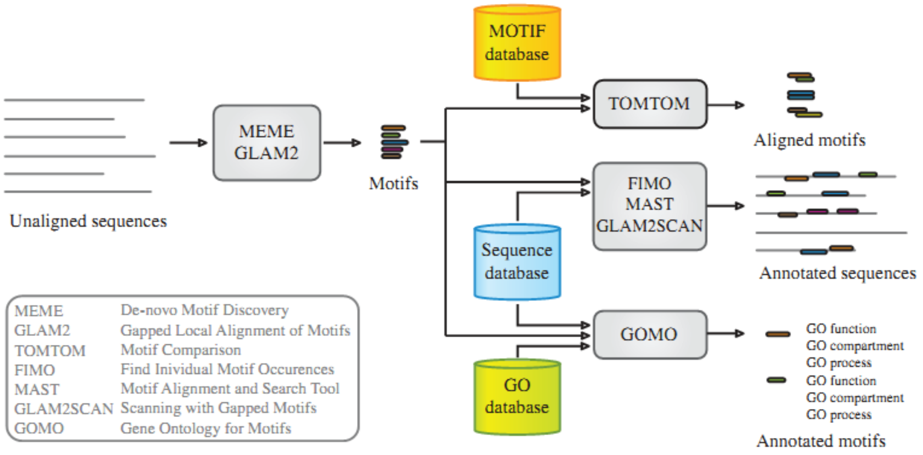


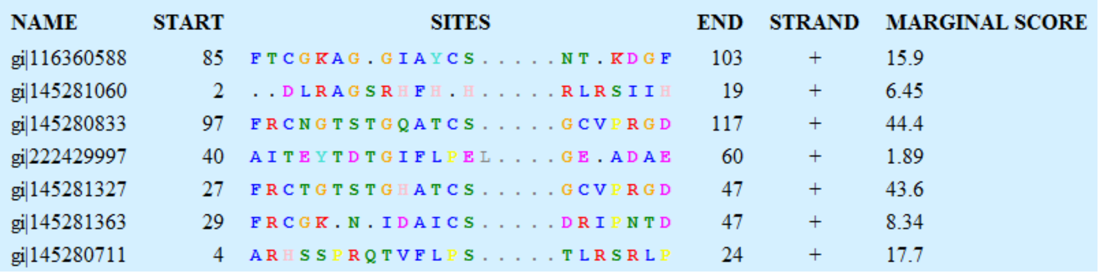
Figure 11. The MEME Suite overview.

Source**:** Directly from Bailey et al., 2009.

The MEME algorithm commonly used to find DNA and protein motifs. Basic MEME algorithm does not allow gapped motifs, thus a gapped version named GLAM2 is added to MEME Suite. GLAM2 returns with scores of each motif it finds. TOMTOM is used to check the similarity of motif with known motifs. If you want to find regulatory functions the motif, GOMO can search it from Gene Ontology annotations. FIMO, MAST and GLAM2SCAN are used to search sequence databases for discovered motif (Bailey et al., 2009).

The MEME Suite is hosted by National Biomedical Computation Resources (NBCR). There is also a toolkit named Opal adopted by NBCR to serve command line request of users. It is also possible to create own servers and redirect jobs to

NCBR MEME servers. Today NCBR servers try to handle more than 200 user requests per day (Bailey et al., 2009).



NAME	START	SITES	END	STRAND	MARGINAL SCORE
gi|116360588	85	FTCGKAG.GIAYCS.....NT.KDGF	103	+	15.9
gi|145281060	2	..DLRAGSRHFH.H.....RLRSIIH	19	+	6.45
gi|145280833	97	FRCNGTSTGQATCS.....GCVPRGD	117	+	44.4
gi|222429997	40	AITEYTDTGIFLPEL....GE.ADAE	60	+	1.89
gi|145281327	27	FRCTGTSTGHATCS.....GCVPRGD	47	+	43.6
gi|145281363	29	FRCGK.N.IDAICS.....DRIPNTD	47	+	8.34
gi|145280711	4	ARHSSPRQTVFLPS.....TLRSRLP	24	+	17.7

Figure 12. GLAM2 motif sample.

GLAM2 produces gapped motifs. In this figure, dots in sequences are gaps between two motifs. It is also possible to find motifs in two strands of sequences. In this example all of them are located in positive strand.

1.12 MEGA

Molecular Evolutionary Genetics Analysis (MEGA) software was developed to provide a suite of tools that make evolutionary analysis of DNA and protein sequences. It includes sequence alignment tools, phylogenetic tree reconstruction, visualization, evolutionary hypotheses testing, estimating sequence divergence and online sequence acquisition. In fifth version of MEGA the maximum likelihood (ML) methods are also added for molecular evolutionary analysis (Tamura et al., 2011). Now MEGA has its fifth version; MEGA5.

Table 6. Summary of analysis and substitution models in MEGA5.

Sequence alignments
CLUSTALW and MUSCLE* alignments DNA and protein
Major analyses (statistical approach in parentheses)
Models and parameters: Select Best-Fit Substitution Model* (ML); test pattern homogeneity; Estimate Substitution Pattern (MCL, ML*); Estimate Rate Variation Among Sites* (ML); Estimate Transition/Transversion Bias (MCL, ML*); Estimate Site-by-Site Rates* (ML). Infer phylogenies: Infer Phylogenetic Trees (NJ, ML*, ME, MP); Phylogeny Tests (Bootstrap and Branch-length tests); Branch-and-Bound Exact Search (MP); Heuristic Searches: Nearest-Neighbor-Interchange (NNI; ML*, ME, MP), Close-Neighbor-Interchange (CNI; ML*, ME, MP), and Max–Mini (MP) Compute distances: Pairwise and Diversity; Within- and Between-Group Distances; Bootstrap and Analytical Variances; separate distances by Site Degeneracy, Codon Sites; Separation of Distances in Transitions and Transversions; Separate Nonsynonymous and Synonymous Changes Tests of Selection: For Complete Sequences or Set of Codons; Sequence Pairs or Groups (Within and Between) Ancestral Sequences: Infer by ML with Relative Probabilities for bases or residues* or by MP (all parsimonious pathways) Molecular Clocks: Tajima's 3-Sequence Clock Test*; Likelihood Ratio Test (ML) for a Topology*; Estimate Branch Lengths under Clock
Substitution models (1F 5 with empirical frequencies; REV 5 reversible)
DNA: General Time Reversible (GTR)*, Tamura–Nei, Hasegawa–Kishino–Yano*, Tamura Three-Parameter, Kimura Two-Parameter, Tajima– Nei, Jukes–Cantor Codons: Nei–Gojobori (original and modified), Li–Wu–Lou (original and modified) Protein: Poisson, Equal-Input, Dayhoff (1F), Jones–Taylor–Thornton (1F), Whelan and Goldman (1F)*, Mitochondrial REV (1F)*, Chloroplast REV (1F)*, Reverse Transcriptase REV (1F)* Rate Variation and Base Compositions: Gamma rates (G) and Invariant sites (I)* models; Incorporate Compositional Heterogeneity

Source: Revised version of original table from Tamura et al., 2011.

1.13 MicroRNAs and RNA Interference

MicroRNAs (also called miRNAs) are 19-24 nucleotide long small RNAs, products of non-coding genes. They are abundant in many organisms and they have very important regulatory roles (Jiang et al., 2012). They are processed from RNAs which can form hairpin structures. MicroRNAs were first found and isolated from *Caenorhabditis elegans*. After their existence shown in animals, miRNAs were extensively demonstrated with their regulatory role in gene expression. In year 2001, miRNAs were identified in *Arabidopsis* which was the first evidence of plant miRNAs. Functional studies in Arabidopsis constructed a framework to understand miRNA function and biogenesis (Chen, 2008). Besides their regulatory endogenous gene expression function, microRNAs also provide intercellular communication (Jiang et al., 2012).

Discovery of plant miRNAs is a continuing procedure and lack of sequenced genome is a limiting factor. Cloning of small RNAs in *Arabidopsis* and rice show that only a small portion of cloned RNAs are miRNAs and the others are small interfering

RNAs (siRNAs) (Chen, 2008). In plants predicting miRNA targets is relatively direct that perfect or nearly perfect complementarity is essential, while in animals it is a little different. In animal miRNA target prediction, a region called seed, 5’ end of miRNA 2nd to 7th nucleotides, needs to be considered and it has to make perfect Watson-Crick pairing with targeted mRNA (Bartel, 2009). Figure 13 summarizes biogenesis of miRNAs with two different ways of function. In animals, as mentioned, microRNAs generally bind to 3’ untranslated region (UTR) of target mRNA, but also some studies reveal that 5’UTR and open reading frame (ORF) could be targeted though it is less frequent (Lytle et al., 2007). On the other hand, in plants, ORF targeting is very common (Millar and Waterhouse, 2005). Similarity between miRNAs and target mRNAs in plant suggests an evolutionary relationship among genes and miRNA genes. These miRNA genes are supposed to evolve from inverse duplication of their target genes (Zhang et al., 2011).

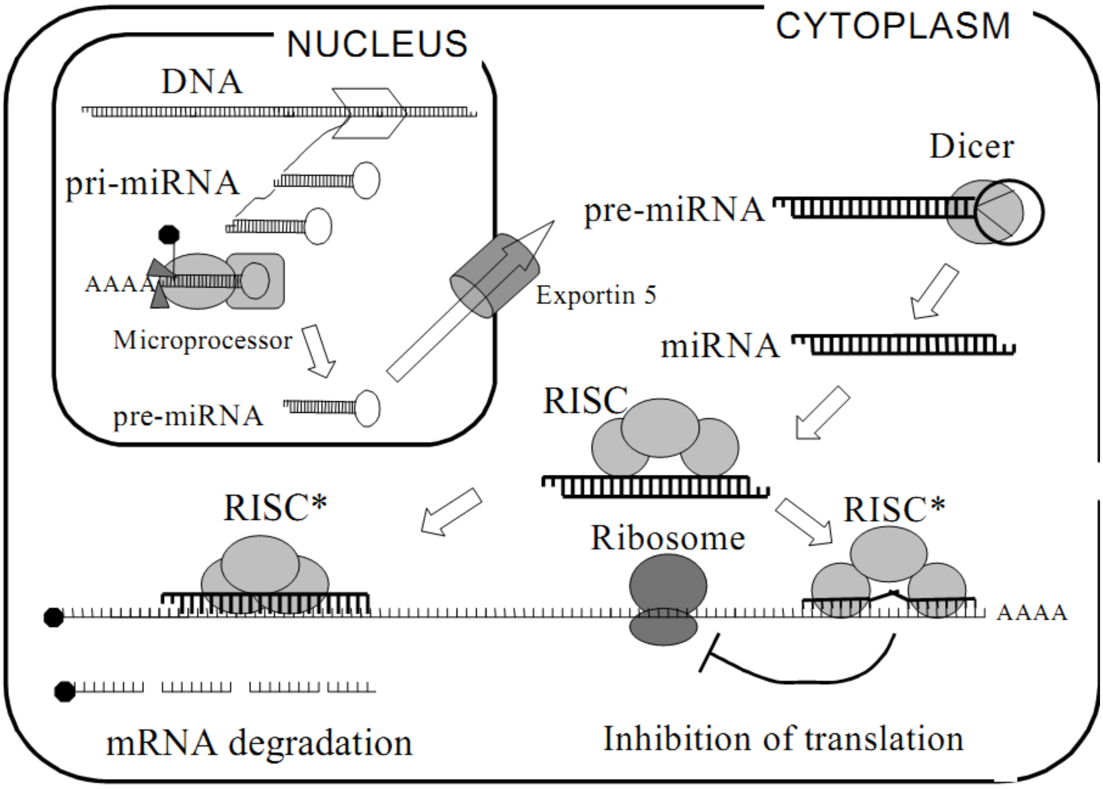


Figure 13. Biogenesis of miRNAs and two likely mechanisms of functions.

Source: Directly from Kusenda et al., 2006.

As it is shown in Figure 13, at the left bottom, RNA induced silencing complex (RISC) attached to miRNA and the target mRNA completely degraded. In the right part of the figure, this time miRNA and RISC inhibit translation and prevent ribosome movement on targeted mRNA.

1.13.1 Cross-Kingdom miRNA Regulation

The organisms in ecosystem are interconnected and they continuously communicate to each other. We know the cells communicate to others with hormones, growth factors etc. MicroRNAs are recently discovered to have similar inter cellular communication roles. In mammals miRNAs are found in body fluids like plasma, urine, saliva and serum. It was believed that extracellular RNA stability is quite low; on the other hand, biochemical experiments proved that microRNAs are very stable against pH, RNase activity and excess temperature as well. It is also interesting that those circulating microRNAs are related to diseases like cancer and diabetes, which are possible markers for disease detection. It is suggested that many microRNAs are wrapped into micro vesicle compartments and these membrane covered vesicles can be secreted by cells (Jiang et al., 2012).

A new phenomenon in miRNA regulation known as cross-kingdom regulation states that it is also possible for microRNAs to regulate genes of foreign cells belong to different kingdoms. Recent studies show that there are exogenous plant miRNAs available in serum and plasma of human and animals. Mir168a can pass through mouse gastrointestinal track and go into circulation; than regulates LDLRAP1 protein expression (Zhang et al., 2012). Therefore, now foods are not only supplier of nutrients, but also provide regulatory information for body (Jiang et al., 2012).

1.14 Aim of the Study

The first objective of the thesis is to discover novel effector candidates for *Puccinia striiformis* f. sp. *tritici* from the available EST sequence data and for which to propose an effector mining pipeline. The road map established in this thesis is going to facilitate candidate effector discovery computationally to lay ground rules for experimental testing of the functions and confirmations. The Pst candidate effectors are to be utilized in designing oligonucleotide microarray probe design.

Additionally, since microRNAs are being emerged as controlling many cellular processes, we aimed to conduct search for finding possible miRNAs in the pathogen and/or possible target genes in the Pst EST sequences and in the lists of candidate effectors of other rust and powdery mildew pathogens. It would be very interesting to find if plant miRNAs are also controlling any pathogen genes as another defense route, if so it will be a novel finding.

CHAPTER II

MATERIALS AND METHODS

2.1 Materials

In this thesis study, miscellaneous sequences (ESTs, mRNAs, nucleotide, protein) are downloaded from various databases and supplementary materials of reference articles. In the context, acquired and presented materials are either given as an appendix or a link to an online material which can be accessed.

The bioinformatics tools are either downloaded or used as an online tool. Self-developed BioPython scripts are used for batch jobs.

2.1.1 Bgh Candidate Effectors (Protein and mRNAs)

The candidate effectors of *Blumeria graminis* f. sp. *hordei* (Bgh) are available as both protein and nucleotide sequences (Godfrey et al., 2010).

The referenced article contains various PDF and EXCEL files as additional material. We converted them into relevant sequence formatted files. The number of effector candidates was reported to be 107.

2.1.2 Pt and Pgt Candidate Effectors (Protein and Nucleotide)

The effector candidates of *Puccinia graminis* f. sp. *tritici* (Pgt) and *Puccinia Triticina* (Pt) were obtained (Godfrey et al., 2010). Pt was reported to have 57 and Pgt was reported to have 178 effector candidates. Not all nucleotide sequences are available in Godfrey et al., 2010, thus they were gathered from NCBI database.

2.1.3 Pst ESTs

Puccinia striiformis f. sp. *tritici* (Pst) has various EST sequences in NCBI database. These sequences are gathered by using NCBI Pubmed and EST databases "file to FASTA" option. There are totally 2848 EST sequences obtained from (Ling et al., 2007; Zhang et al., 2008; Yin et al., 2009) (Supplementary Material 1). The other available Pst data were not included in this thesis.

2.1.4 Whole Genome Sequences (WGS)

WGS of Bgh, Pt and Pgt are available. Pst does not have full genome assembly yet.

Bgh genome sequence downloaded from BluGen (www.blugen.org). Bgh genome size is nearly 120 Mb. It shows losses on genes like enzymes of primary and secondary metabolism that result extremely parasitic life style (Spanu et al., 2010).

Pt and Pgt genome sequence downloaded from Broad Institute

(http://www.broadinstitute.org/annotation/genome/puccinia_group/MultiDownloads.html). Pgt genome size is nearly 80 Mb and Pt genome size is nearly 120 Mb.

2.1.5 Plant MicroRNAs

Populus euphratica, Populus trichocarpa, Zea mays, Hordeum vulgare, Oryza sativa, Triticum aestivum, Triticum turgidum and *Brachypodium distachyon* miRNAs were downloaded from miRBase (http://www.mirbase.org). All microRNAs were downloaded as miRNA precursors and mature miRNAs. 1435 mature miRNAs and 1201 miRNA precursors were found in total (Supplementary Material 2 & Supplementary Material 3). Unaligned FASTA format, stem-loop sequences and mature miRNA options were selected to fetch all sequences. All sequences in our dataset were updated on January 2012 from miRBase.

2.2 Methods

2.2.1 ORF Prediction of Pst ESTs

Open reading frames of Pst EST sequences are predicted by NCBI's ORF Finder (http://www.ncbi.nlm.nih.gov/projects/gorf/) tool.

However, ORF Finder does not support batch jobs, there are other ORF predictors available but it is best to use ORF Finder. To overcome batch job problem, two different BioPyhton scripts were written (Table 8). First script in Table 8 reads a FASTA file with BioPython extensions; it connects to ORF Finder server and finds all possible open reading frames then write all predicted ORFs to a text file. All ESTs in FASTA file creates a different text file. Second script in Table 8 parse the files written and combines all predicted ORFs as a single FASTA file.

Table 7. Python scripts to find Open Reading Frames.

First Script

```
import httplib
import urllib
list={}

from Bio import SeqIO
outputFasta=open("pstest.fasta","rU")
for seq_rec in SeqIO.parse(outputFasta,"fasta"):
    list[str(seq_rec.id)]=str(seq_rec.seq)

headers     =     {"Content-type":     "application/x-www-form-urlencoded",     "Accept":
"text/plain"}

for sequence in list:
    params=urllib.urlencode({'SEQUENCE':list[sequence]})
    while True:
        try:
            conn=httplib.HTTPConnection("www.ncbi.nlm.nih.gov:80")
            conn.request("POST","/projects/gorf/orfig.cgi",params,headers)
            response=conn.getresponse()
            readdata=response.read()
            break
        except:
            print 'Trying again this round'
    filei=open(sequence+".txt","w")
    filei.write(readdata)
    filei.close()
```

Second Script

```
import re
from Bio.Seq import Seq
from Bio import SeqIO

fastasupercontig=open("pstest.fasta","rU")
list={}
for seq_rec in SeqIO.parse(fastasupercontig,"fasta"):
    list[str(seq_rec.id)]=str(seq_rec.seq)

def fastaWrite(name,seq,file):
    fastaw='>' + name +'\n'+ seq +'\n'
    file.write(fastaw)

outPUT=open("orf_parsed.fasta","w")

for sequence in list:
    parseFile=open(sequence+".txt")
    parseFile=parseFile.read()
    parseFile=re.split('<tr><td
align=center>Frame</td><td></td><td>from</td><td></td><td>to</td><td>Length</td></tr>
',parseFile)
    parseFile=re.findall('([-+]?\d+)',parseFile[1])
    boy=len(parseFile)/6
    i=0
    while i<boy:
        posit=int(parseFile[i*6])
        seqord=int(parseFile[i*6+1])
        first_pos=int(parseFile[i*6+3])
        last_pos=int(parseFile[i*6+4])
        oku=list[sequence]
        oku=oku[(first_pos-1):(last_pos)]
        if posit<0:
            oku=Seq(oku)
            oku=oku.reverse_complement()
        fastaWrite(str(sequence)+"_orf_"+str(seqord),str(oku),outPUT)
        i=i+1
```

2.2.2 Signal Peptide Prediction

Predicted ORFs were sent to SignalP-4.0 program to detect signal peptide regions. The results written into an excel file and sorted the best ones with respect to their D-scores.

2.2.3 Local BLAST

A local BLAST database was created using Bgh, Pt and Pgt candidate effectors with command line "makeblastdb -in input.fasta -out outputdatabasename". This command is the same as both in Windows and Linux environment, if latest NCBI local BLAST is installed.

A blastp query, "blastp -query input.fas -db outputdatabasename -outfmt "6 qseqid;evalue;sseqid" -out hit.txt", shows the similarity between predicted ORFs and putative candidate effectors.

The similar ORFs, predicted by blastp, collected from FASTA file by BioPython script (Table 9). This script needs three file names to operate. It reads a source FASTA file and collects required sequences from that file and write them to another file.

Table 8. Python script to collect required sequences from a FASTA file.

```
list={}
from Bio import SeqIO

mainFile=raw_input("Fasta file name :")
requiredSeqName=raw_input("Required seq names text file :")
outputF=raw_input("Output Fasta file :")

outputFastaFile=open(mainFile,"rU")
readF=open(requiredSeqName,"r")
outputFasta=open(outputF,"w")

def fastaWriterFunc(name,seq,file):
    fastaWrite='>' + name +'\n' + seq +'\n'
    file.write(fastaWrite)

seqnames=readF.readlines()

for seq_rec in SeqIO.parse(outputFastaFile,"fasta"):
    for isim in seqnames:
        if str(seq_rec.id)==isim[:-1]:
            fastaWriterFunc(str(seq_rec.id),str(seq_rec.seq),outputFasta)
```

2.2.4 Multiple Sequence Alignment

There are two different alignment tools are used to align sequences because both algorithms have distinctive qualities and the quality of alignment could be verified by comparing them.

All candidate effectors gathered from NCBI database and reference articles were aligned with Clustal and MUSCLE.

Best predicted signal peptide containing ORFs and blasted ORFs were aligned in pairwise with candidate effectors of Bgh, Pgt and Pt to see difference and similarities between sequences. Moreover, phylogenetic trees were drawn to classify sequences.

2.2.5 Blast2GO Annotation

Annotation of sequences is a very important to find their function if available and define characteristics of sequences. Blast2Go annotation tool was used on all sequence sets to understand general properties by searching them on databases.

2.2.6 Local BLAST of Plant MicroRNAs

A local BLAST database was created using Bgh, Pt and Pgt WGS with command line “makeblastdb -in input.fasta -out outputdatabasename”.

A blastn query, “blastn -query input.fas -db outputdatabasename -outfmt "6 qseqid;evalue;sseqid" -out hit.txt”, shows the similarity between plant miRNAs and WGS of plant pathogens.

2.2.7 MicroRNA Target Prediction

Plant miRNAs generally prefer perfect or near perfect Watson-Crick complementarity in their target genes. To detect their possible target Smith-

Waterman algorithm was used. A BioPython script handled the batch process to run algorithm effectively.

Table 10 is the script. In the local alignment algorithm, match score is +2; mismatch and gaps scores are -1. Binding scores that are greater than 32 selected.

Table 9. BioPython script for local alignment.

```
import numpy
from Bio import SeqIO
from Bio.Seq import Seq
import string

mirnaFile=open("mirnas.fasta","rU")
effecFile=open("effectors.txt","rU")
outputFile=open("mirnares.txt","w")

idmirna=SeqIO.to_dict(SeqIO.parse(mirnaFile,"fasta"))
idefektor=SeqIO.to_dict(SeqIO.parse(effecFile,"fasta"))

for mirna in idmirna:
    idmirna[mirna].seq=idmirna[mirna].seq.lower()
    idmirna[mirna].seq=idmirna[mirna].seq.back_transcribe()
    idmirna[mirna].seq=idmirna[mirna].seq.reverse_complement()
    idmirna[mirna].seq='x'+idmirna[mirna].seq
for efektor in idefektor:
    idefektor[efektor].seq=idefektor[efektor].seq.lower()
    idefektor[efektor].seq=idefektor[efektor].seq.back_transcribe()
    if (string.find(idefektor[efektor].description,'3\''))!=-1:
        idefektor[efektor].seq=idefektor[efektor].seq.reverse_complement()
    idefektor[efektor].seq='x'+idefektor[efektor].seq

for efektor in idefektor:
    print idefektor[efektor].description
    print idefektor[efektor].seq


for mirna in idmirna:
    for efektor in idefektor:
        dptable=numpy.zeros((len(idmirna[mirna]),len(idefektor[efektor])))
        maxscore=0
        for i in range(1,len(idmirna[mirna])):
            for j in range(1,len(idefektor[efektor])):
                score_up=dptable[i-1][j]-1
                score_left=dptable[i][j-1]-1
                if idefektor[efektor].seq[j]==idmirna[mirna].seq[i]:
                    score_diagonal=dptable[i-1][j-1]+2
                else:
                    score_diagonal=dptable[i-1][j-1]-1
                dptable[i][j]=max(score_up,score_left,score_diagonal,0)
                if maxscore < dptable[i][j]:
                    maxscore=dptable[i][j]
                    maxi=i
                    maxj=j
        if (maxscore >=32):

outputFile.write(str(maxscore)+';'+str(idmirna[mirna].id)+';'+str(idmirna[mirna].seq)
+';'+str(idefektor[efektor].id)+';'+str(idefektor[efektor].seq)+'\n')
```

Plant miRNAs, from miRBase, are tested if they have a high score match (greater than 32) with Pst ESTs, Bgh, Pt, Pgt candidate effectors. The first assumption is plant miRNAs must have perfect or near perfect complementarity with pathogenic genes, mostly consist of effectors. The second assumption is mRNA can be targeted in all possible locations; 3’UTR, 5’UTR or coding region. There is no other study related to cross-kingdom regulation between plant and fungi. Thus, these assumptions based on plant RNAi.

CHAPTER III

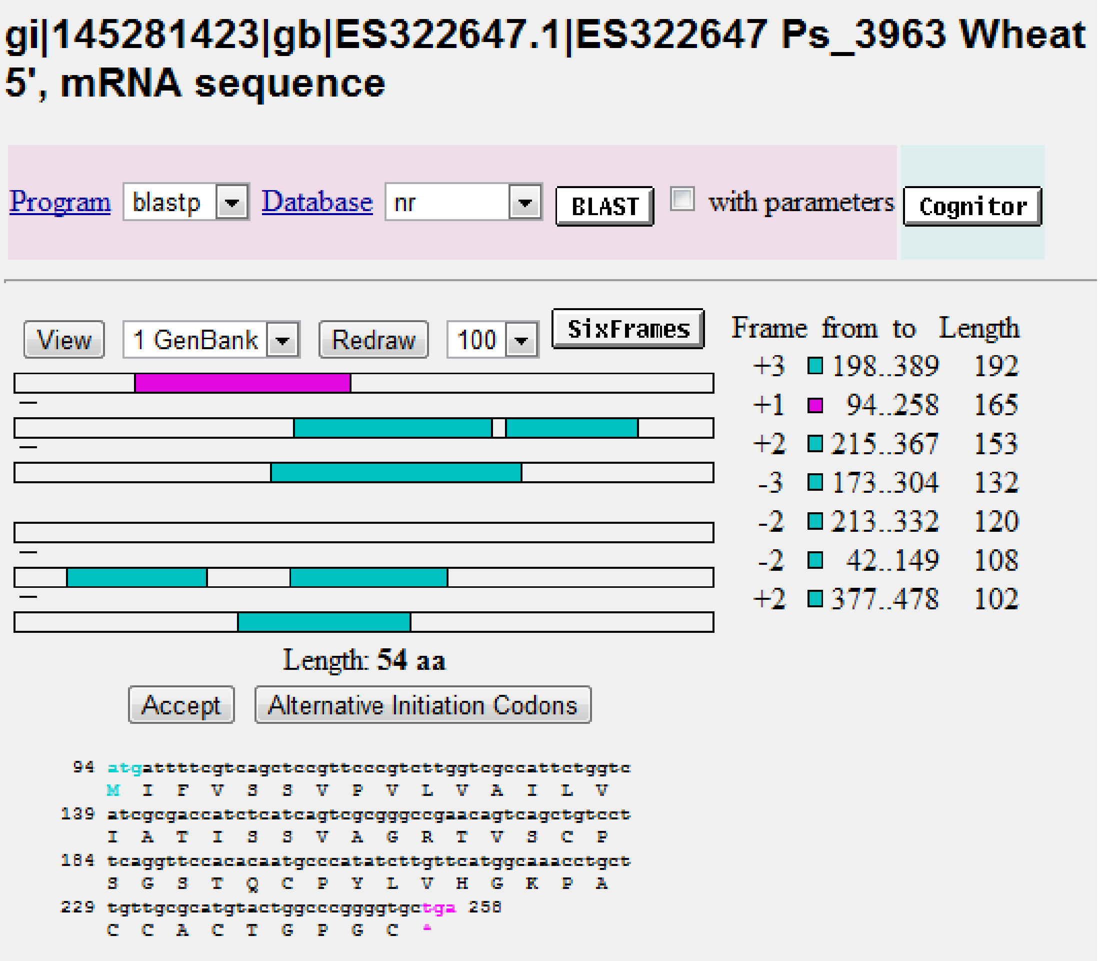
RESULTS

3.1 Results

All results are presented in logical and distinctive headings to make it easy understand the outputs. All figures without any reference source were created by Sinan Uğur Umu (SUU).

3.1.1 ORF Prediction Results

Pst ESTs (2848 sequences) were loaded into ORF Finder *via* Python scripts. ORF Finder resulted in 9854 ORFs (Supplementary Material 4). Figure 14 shows an example result of ORF Finder tool. It produces seven different open reading frames from EST with accession gi|145281423|gb|ES322647.1|ES322647. Python script can gather all of these frames successfully.



gi|145281423|gb|ES322647.1|ES322647 Ps_3963 Wheat 5', mRNA sequence

Frame	from to	Length
+3	🟦 198..389	192
+1	🟪 94..258	165
+2	🟦 215..367	153
-3	🟦 173..304	132
-2	🟦 213..332	120
-2	🟦 42..149	108
+2	🟦 377..478	102

Length: **54 aa**

```
 94 atgattttcgtcagctccgttcccgtcttggtcgccattctggtc
    M  I  F  V  S  S  V  P  V  L  V  A  I  L  V
139 atcgcgaccatctcatcagtcgcgggccgaacagtcagctgtcct
    I  A  T  I  S  S  V  A  G  R  T  V  S  C  P
184 tcaggttccacacaatgcccatatcttgttcatggcaaacctgct
    S  G  S  T  Q  C  P  Y  L  V  H  G  K  P  A
229 tgttgcgcatgtactggcccggggtgctga 258
    C  C  A  C  T  G  P  G  C  *
```

Figure 14. ORF Finder result of a single Expressed Sequence Tag.

In Figure 14, the query name is at the top; output of seven predicted ORFs are presented. When clicked to green sections, which represent frames, they turn into pink and show predicted sequence at the bottom.

3.1.2 Signal Peptide Prediction of Y/F/WxC Effector Candidates

N-terminal signal peptides are one of the most important features of effector proteins. It is important to detect possible signal peptide regions on candidate effectors and ORFs for a preliminary analysis (Supplementary Table 1).

SignalP-4.0 predicted that 275 of 342 reported effector candidates have SP regions. Though they are classified as effector candidates, SignalP did not predict that all of them are secretory.

3.1.3 Local BLAST Results

Although sequence similarity is low among effector candidates, there are some similar regions. Convergent evolution may also favor conservation of similar effectors in different pathogens. For example, Y/F/WxC motif containing candidate effectors also present level of similarity in their SP region according to MSA results.

Blastp is used in order to further classify and refine predicted ORFs (9854 sequences) and identify their similarity with putative effector candidates. It showed that 95 predicted ORF sequences are similar with Y/F/WxC candidate effectors (e-value threshold 0.05) (Appendix 1).

3.1.4 Signal Peptide Prediction of BLAST Validated Similar ORFs

Local BLAST analysis resulted in a new set of 95 sequences which have a significant sequence similarity with putative candidate effectors. This new set was tested with SignalP-4.0 (Appendix 2). The results showed that 32 of these sequences were predicted to have SP region; this group has a very high *D-score* average, 0.76. *D-score* is out of 1 and 0.76 score average is proficiently high.

Due to importance of SP region in candidate effectors, all predicted ORFs (9854 sequences) were tested in SignalP-4.0. We found that 880 out of 9854 sequences are predicted to have SP regions (Supplementary Table 1).

3.1.5 Multiple Sequence Alignments of Y/F/WxC Effector Candidates

Bgh, Pt and Pgt candidate effectors share common Y/F/WxC motif site (Godfrey et al., 2010). Multiple sequence alignment of all of these protein sequences demonstrates the similarity in motif region (Figure 15 pink lined area) and SP region (Figure 16 blue lined area).

In Figure 17, guide tree of Clustal is seen; Figure 18 is a cladogram after 1000 bootstrap sampling.

We made MSA analysis in order to verify our designated pipeline and conserved regions on reported effector candidates.

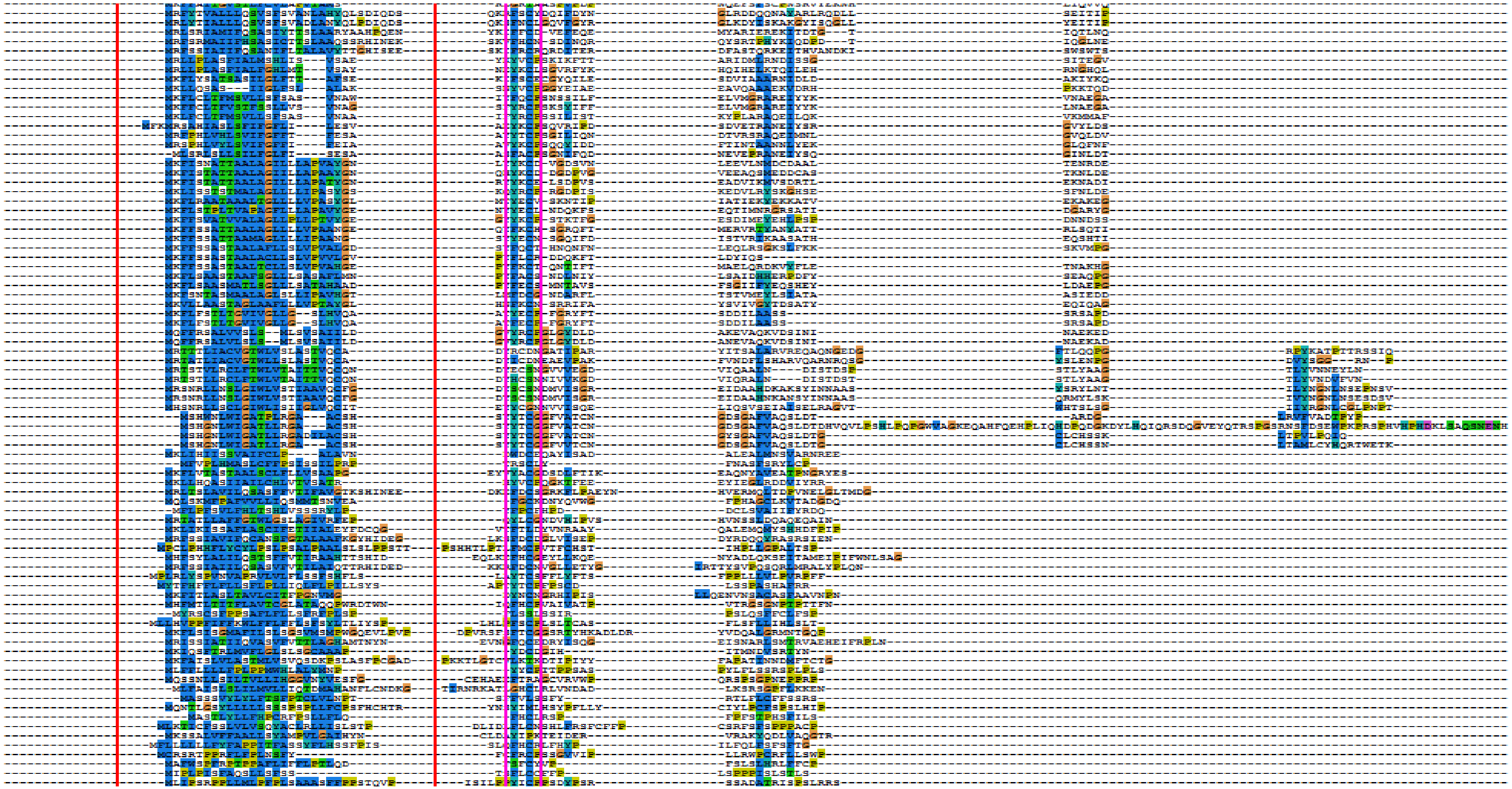


Figure 15. Multiple sequence alignment of Bgh, Pt and Pgt candidate effectors with Clustal.

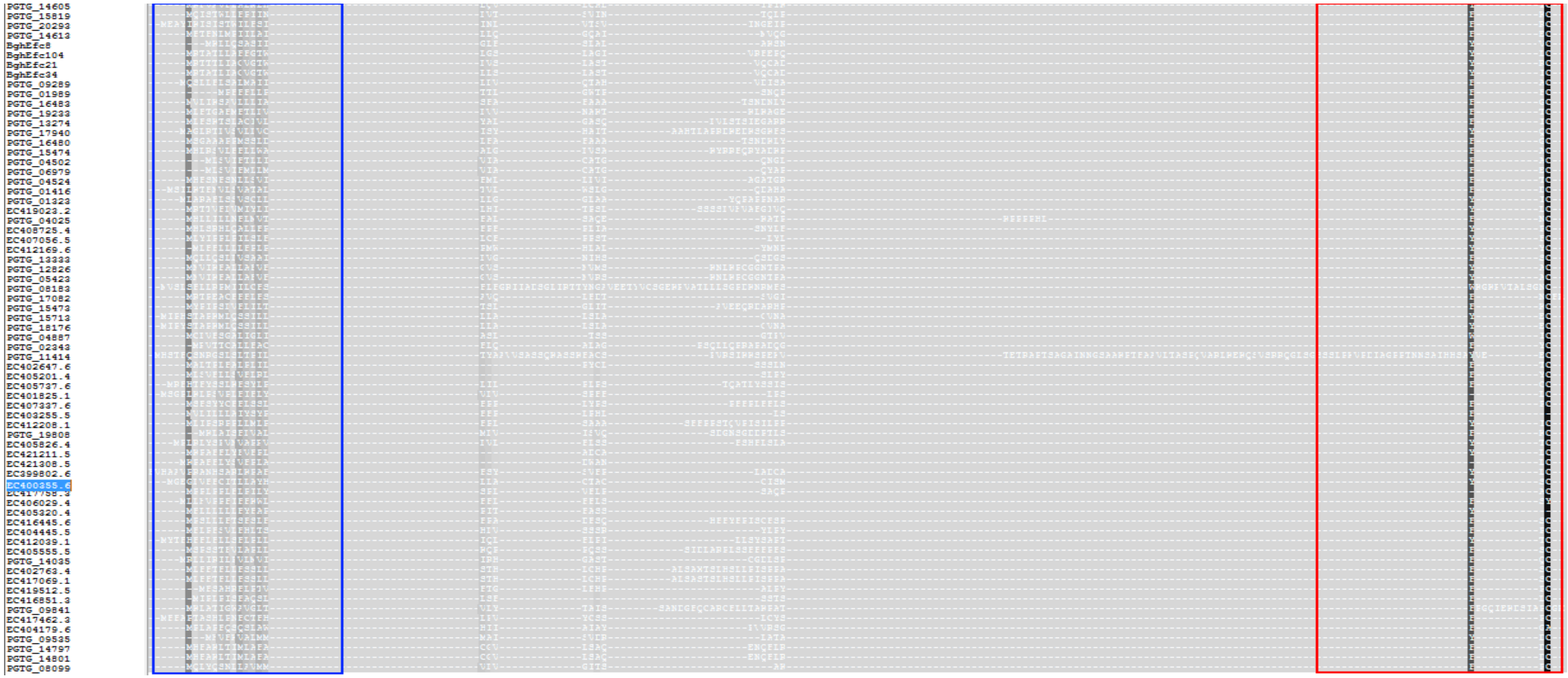


Figure 16. MUSCLE MSA of Bgh, Pt and Pgt candidate effectors (342 in total).

MUSCLE gives a little different MSA but it also points same similar regions of SP and conserved motif obtained as Clustal analysis. Red lined area shows conserved motif, and conserved residues are blacker than others. Blue lined area shows SP region and grey-black conserved regions are also visible in this figure. Visualized by BioEdit (Hall, 1999).

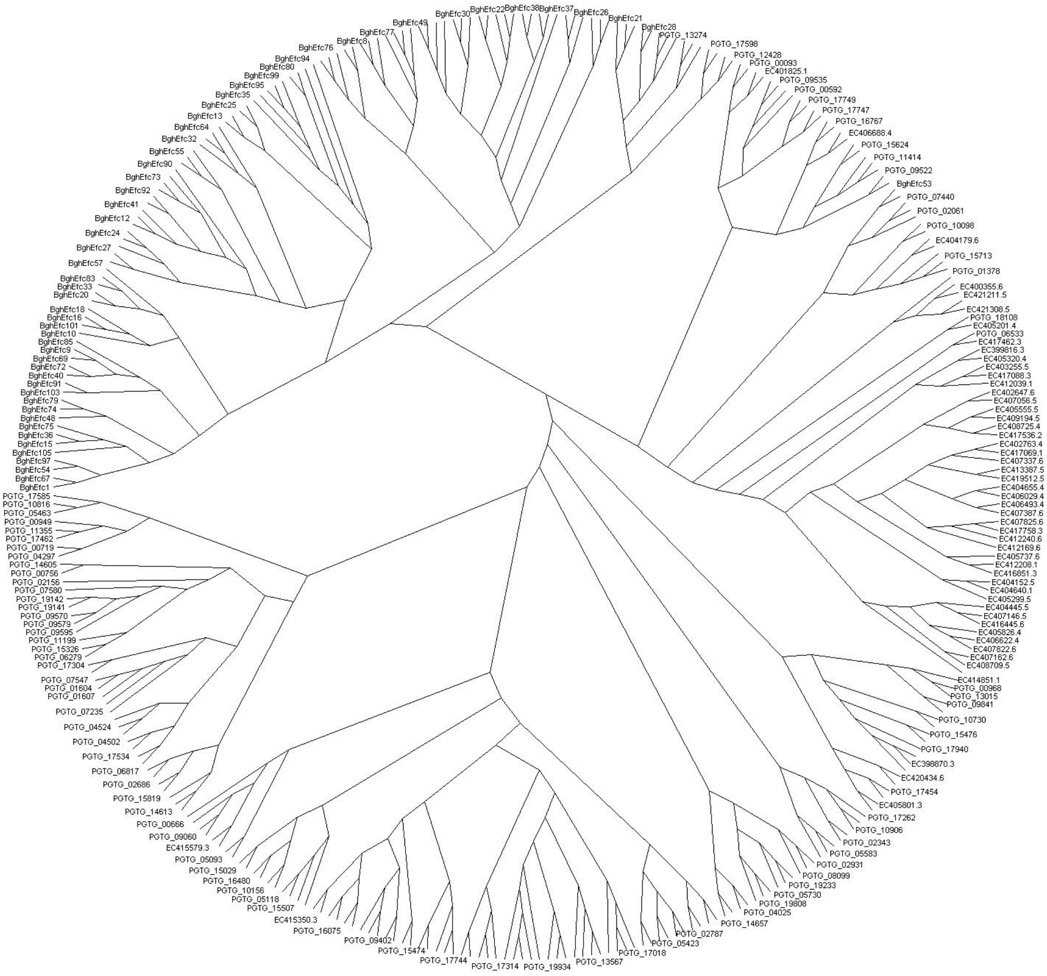


Figure 17. The guide tree of Bgh, Pt and Pgt candidate effectors (342 in total).

In Figure 17 all sequences were taken from Godfrey et al., 2010. This guide tree was created by ClustalX as a single dnd file. It is an un-rooted NJ tree and visualized by Dendroscope (Huson et al., 2007). PGT labeled sequences are Pgt effector candidates. EC labeled sequences are Pt effector candidates and so Bgh labeled is Bgh effector candidates. All sequences are grouped into their relative sequences but also outliers are observed.

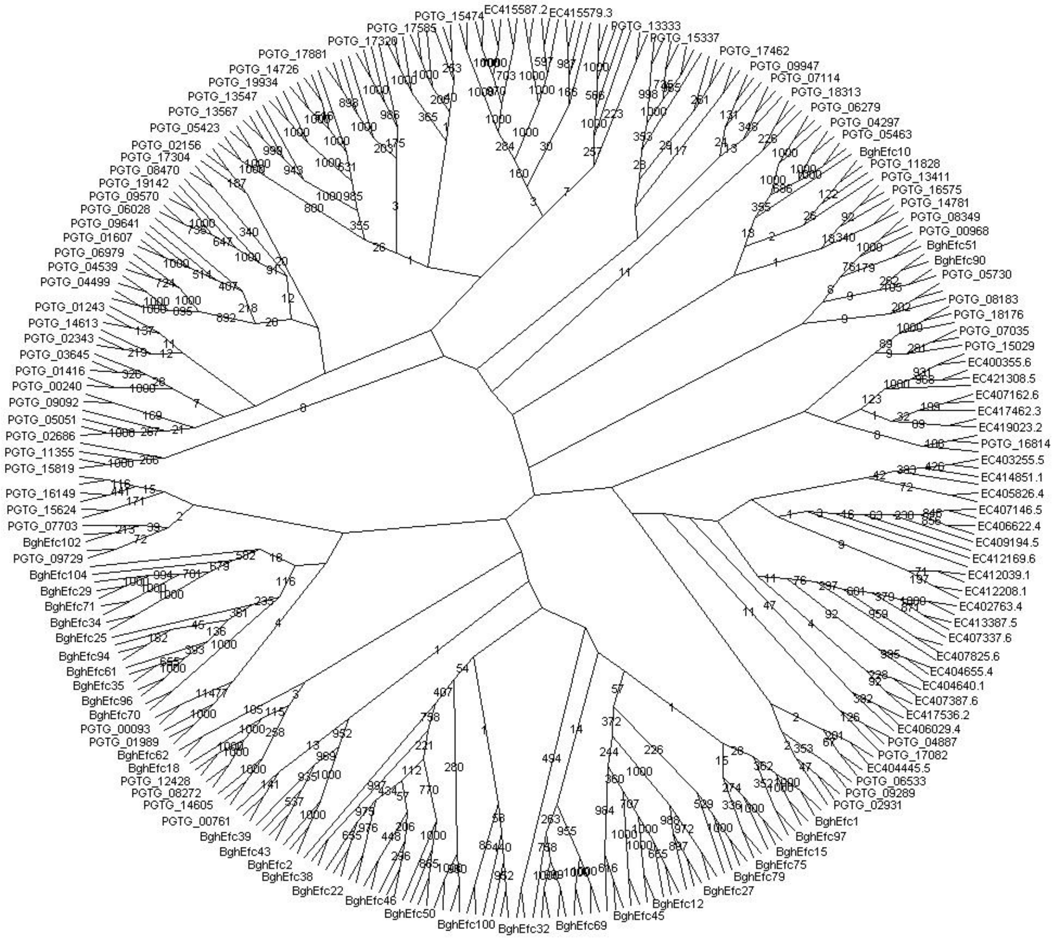


Figure 18. Cladogram of Bgh, Pt and Pgt candidate effectors (after 1000 bootstrap sampling).

In Figure 18 all sequences were taken from Godfrey et al., 2010. As expected, Bgh, Pt and Pgt effectors were clustered into different groups, Bgh effectors are at bottom, Pgt effectors are at top and Pt effectors are at left but also few of them are seen at different locations. It is an un-rooted NJ tree and visualized by Dendroscope (Huson et al., 2007). The numbers on bootstrap tree are confidence levels. It is similar with guide tree except bootstrapping and confidence values.

3.1.6 Multiple Sequence Alignment of Predicted ORFs

A new tree was drawn to obtain relationship between predicted and reported effectors (Godfrey et al., 2010) (Figure 19). Figure 19 cladogram reveals that predicted ORFs are generally much closer to Pgt candidate effectors.

The predicted 95 ORFs were also aligned alone using Clustal. As first glance, it is seen that they were not aligned like reported effector candidates. However, if they were aligned all together with reported effector candidates, the similar regions would have been visible. This shows the MSA prediction was insufficient to detect the best ORFs as candidate effector among the 95. They may have different properties or they are not complete ESTs, prematurely ended for experimental reasons etc. Figure 20 is the phylogram of these ORFs after 1000 bootstrap sampling.

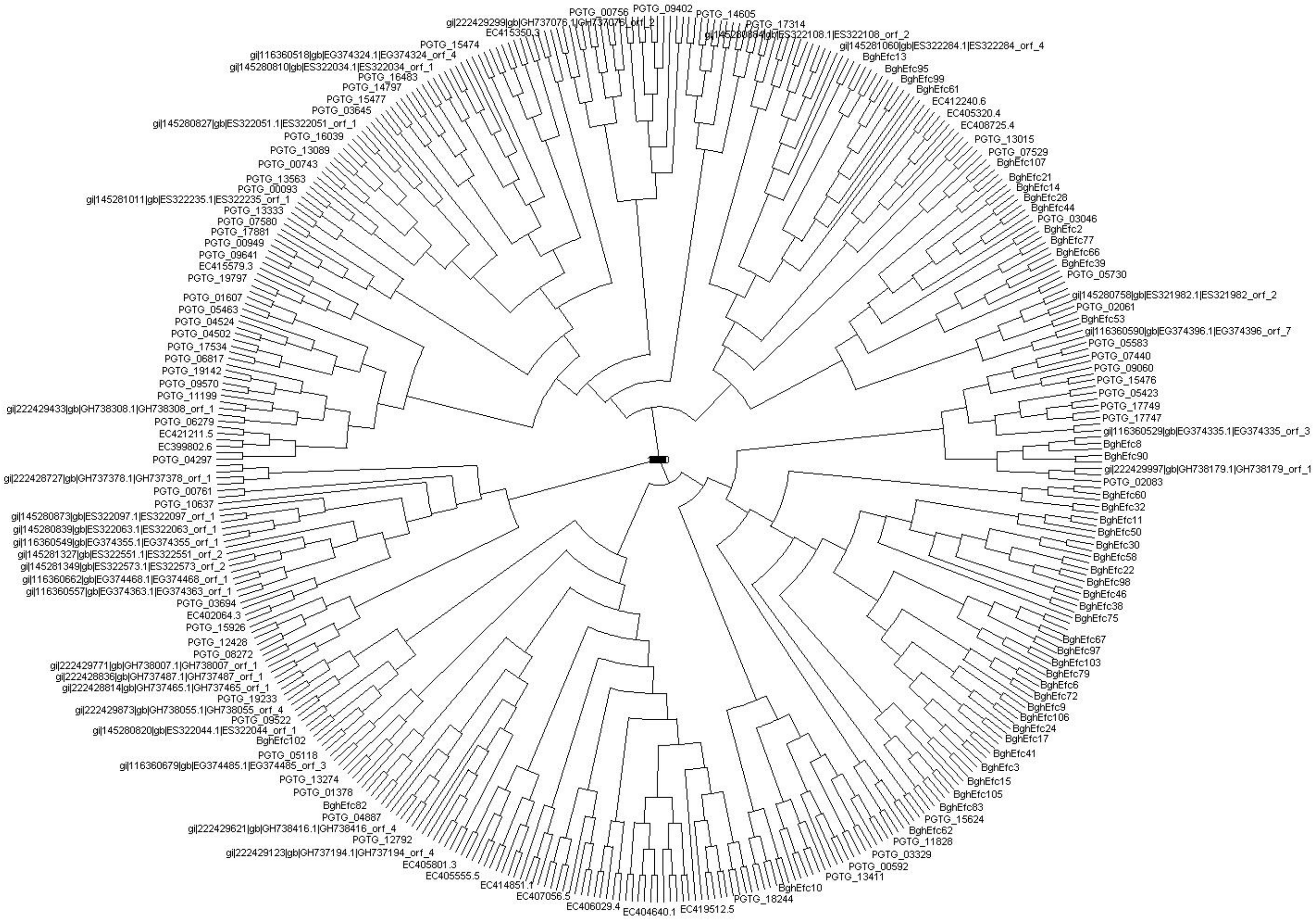

PGTG_00756
PGTG_09402
PGTG_14605
PGTG_17314
gi|145280884|gb|ES322108.1|ES322108_orf_2
gi|145281060|gb|ES322284.1|ES322284_orf_4
BghEfc13
BghEfc95
BghEfc99
BghEfc61
EC412240.6
EC405320.4
EC408725.4
PGTG_13015
PGTG_07529
BghEfc107
BghEfc21
BghEfc14
BghEfc28
BghEfc44
PGTG_03046
BghEfc2
BghEfc77
BghEfc66
BghEfc39
PGTG_05730
gi|145280758|gb|ES321982.1|ES321982_orf_2
PGTG_02061
BghEfc53
gi|116360590|gb|EG374396.1|EG374396_orf_7
PGTG_05583
PGTG_07440
PGTG_09060
PGTG_15476
PGTG_05423
PGTG_17749
PGTG_17747
gi|116360529|gb|EG374335.1|EG374335_orf_3
BghEfc8
BghEfc90
gi|222429997|gb|GH738179.1|GH738179_orf_1
PGTG_02083
BghEfc60
BghEfc32
BghEfc11
BghEfc50
BghEfc30
BghEfc58
BghEfc22
BghEfc98
BghEfc46
BghEfc38
BghEfc75
BghEfc67
BghEfc97
BghEfc103
BghEfc79
BghEfc6
BghEfc72
BghEfc9
BghEfc106
BghEfc24
BghEfc17
BghEfc41
BghEfc3
BghEfc15
BghEfc105
BghEfc83
PGTG_15624
BghEfc62
PGTG_11828
PGTG_03329
PGTG_00592
PGTG_13411
BghEfc10
PGTG_18244
EC419512.5
EC404640.1
EC406029.4
EC407056.5
EC414851.1
EC405555.5
EC405801.3
gi|222429123|gb|GH737194.1|GH737194_orf_4
PGTG_12792
gi|222429621|gb|GH738416.1|GH738416_orf_4
PGTG_04887
BghEfc82
PGTG_01378
PGTG_13274
gi|116360679|gb|EG374485.1|EG374485_orf_3
PGTG_05118
BghEfc102
gi|145280820|gb|ES322044.1|ES322044_orf_1
PGTG_09522
gi|222429873|gb|GH738055.1|GH738055_orf_4
PGTG_19233
gi|222428814|gb|GH737465.1|GH737465_orf_1
gi|222428836|gb|GH737487.1|GH737487_orf_1
gi|222429771|gb|GH738007.1|GH738007_orf_1
PGTG_08272
PGTG_12428
PGTG_15926
EC402064.3
PGTG_03694
gi|116360557|gb|EG374363.1|EG374363_orf_1
gi|116360662|gb|EG374468.1|EG374468_orf_1
gi|145281349|gb|ES322573.1|ES322573_orf_2
gi|145281327|gb|ES322551.1|ES322551_orf_2
gi|116360549|gb|EG374355.1|EG374355_orf_1
gi|145280839|gb|ES322063.1|ES322063_orf_1
gi|145280873|gb|ES322097.1|ES322097_orf_1
PGTG_10637
PGTG_00761
gi|222428727|gb|GH737378.1|GH737378_orf_1
PGTG_04297
EC399802.6
EC421211.5
PGTG_06279
gi|222429433|gb|GH738308.1|GH738308_orf_1
PGTG_11199
PGTG_09570
PGTG_19142
PGTG_06817
PGTG_17534
PGTG_04502
PGTG_04524
PGTG_05463
PGTG_01607
PGTG_19797
EC415579.3
PGTG_09641
PGTG_00949
PGTG_17881
PGTG_07580
PGTG_13333
gi|145281011|gb|ES322235.1|ES322235_orf_1
PGTG_00093
PGTG_13563
PGTG_00743
PGTG_13089
PGTG_16039
gi|145280827|gb|ES322051.1|ES322051_orf_1
PGTG_03645
PGTG_15477
PGTG_14797
PGTG_16483
gi|145280810|gb|ES322034.1|ES322034_orf_1
gi|116360518|gb|EG374324.1|EG374324_orf_4
PGTG_15474
EC415350.3
gi|222429299|gb|GH737076.1|GH737076_orf_2

Figure 19. Cladogram of all reported candidate effectors of Pgt, Pt, Bgh and 95 predicted ORFs.

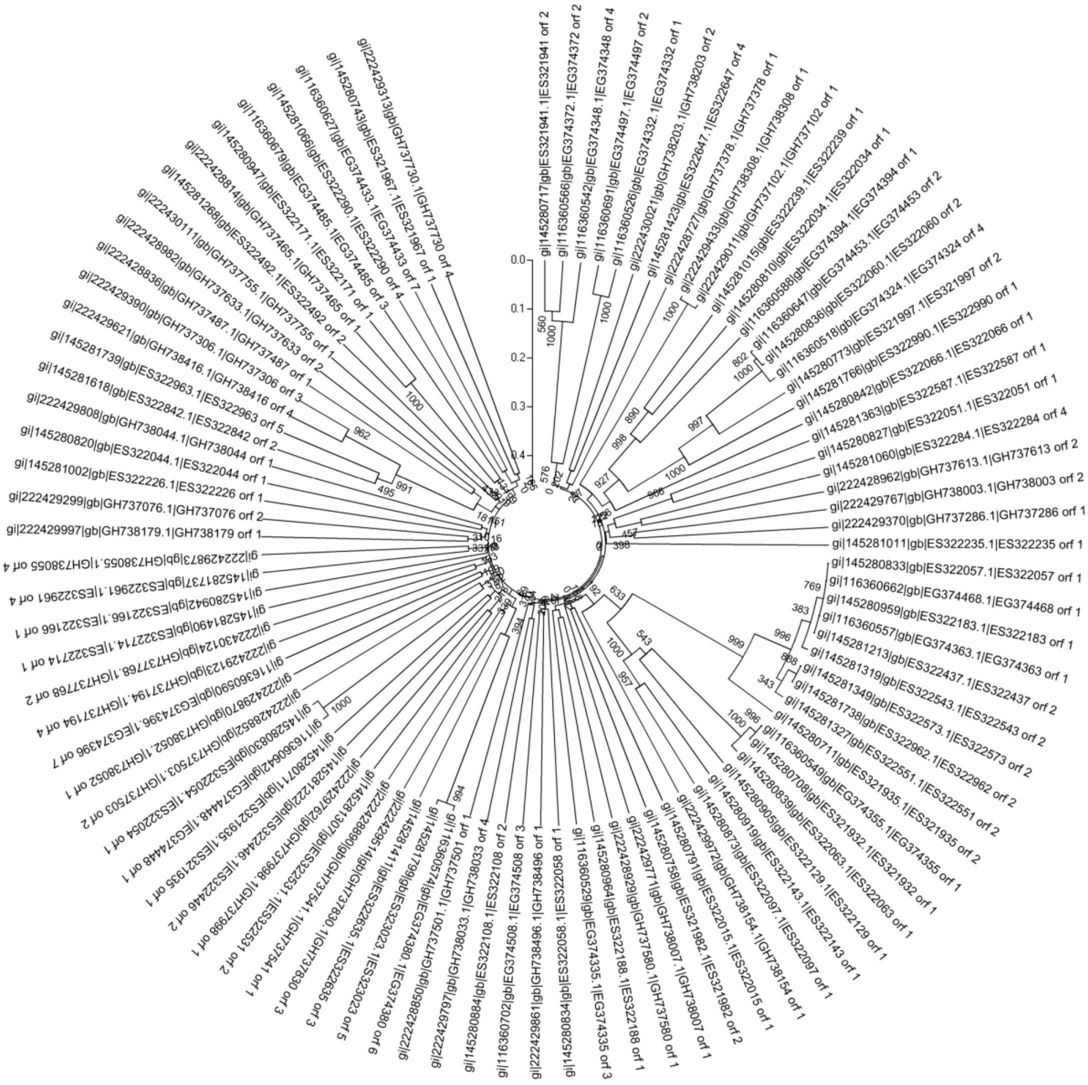

0.0
0.1
0.2
0.3
0.4
gi|145280717|gb|ES321941.1|ES321941 orf 2
gi|116360566|gb|EG374372.1|EG374372 orf 2
gi|116360542|gb|EG374348.1|EG374348 orf 4
gi|116360691|gb|EG374497.1|EG374497 orf 2
gi|116360526|gb|EG374332.1|EG374332 orf 1
gi|222430021|gb|GH738203.1|GH738203 orf 2
gi|145281423|gb|ES322647.1|ES322647 orf 4
gi|222428727|gb|GH737378.1|GH737378 orf 1
gi|222429433|gb|GH738308.1|GH738308 orf 1
gi|222429011|gb|GH737102.1|GH737102 orf 1
gi|145281015|gb|ES322239.1|ES322239 orf 1
gi|145280810|gb|ES322034.1|ES322034 orf 1
gi|116360588|gb|EG374394.1|EG374394 orf 1
gi|116360647|gb|EG374453.1|EG374453 orf 2
gi|145280836|gb|ES322060.1|ES322060 orf 2
gi|116360518|gb|EG374324.1|EG374324 orf 4
gi|145280773|gb|ES321997.1|ES321997 orf 2
gi|145281766|gb|ES322990.1|ES322990 orf 1
gi|145280842|gb|ES322066.1|ES322066 orf 1
gi|145281363|gb|ES322587.1|ES322587 orf 1
gi|145280827|gb|ES322051.1|ES322051 orf 1
gi|145281060|gb|ES322284.1|ES322284 orf 4
gi|222428962|gb|GH737613.1|GH737613 orf 2
gi|222429767|gb|GH738003.1|GH738003 orf 2
gi|222429370|gb|GH737286.1|GH737286 orf 1
gi|145281011|gb|ES322235.1|ES322235 orf 1
gi|145280833|gb|ES322057.1|ES322057 orf 1
gi|116360662|gb|EG374468.1|EG374468 orf 1
gi|145280959|gb|ES322183.1|ES322183 orf 1
gi|116360557|gb|EG374363.1|EG374363 orf 1
gi|145281213|gb|ES322437.1|ES322437 orf 2
gi|145281319|gb|ES322543.1|ES322543 orf 2
gi|145281349|gb|ES322573.1|ES322573 orf 2
gi|145281738|gb|ES322962.1|ES322962 orf 2
gi|145281327|gb|ES322551.1|ES322551 orf 2
gi|145280711|gb|ES321935.1|ES321935 orf 2
gi|116360549|gb|EG374355.1|EG374355 orf 1
gi|145280708|gb|ES321932.1|ES321932 orf 1
gi|145280839|gb|ES322063.1|ES322063 orf 1
gi|145280905|gb|ES322129.1|ES322129 orf 1
gi|145280919|gb|ES322143.1|ES322143 orf 1
gi|145280873|gb|ES322097.1|ES322097 orf 1
gi|222429972|gb|GH738154.1|GH738154 orf 1
gi|145280791|gb|ES322015.1|ES322015 orf 1
gi|145280758|gb|ES321982.1|ES321982 orf 2
gi|222429771|gb|GH738007.1|GH738007 orf 1
gi|222428929|gb|GH737580.1|GH737580 orf 1
gi|145280964|gb|ES322188.1|ES322188 orf 1
gi|116360529|gb|EG374335.1|EG374335 orf 3
gi|145280834|gb|ES322058.1|ES322058 orf 1
gi|222429861|gb|GH738496.1|GH738496 orf 1
gi|116360702|gb|EG374508.1|EG374508 orf 3
gi|145280884|gb|ES322108.1|ES322108 orf 2
gi|222429797|gb|GH738033.1|GH738033 orf 4
gi|222428850|gb|GH737501.1|GH737501 orf 1
gi|116360574|gb|EG374380.1|EG374380 orf 6
gi|145281799|gb|ES323023.1|ES323023 orf 5
gi|145281411|gb|ES322635.1|ES322635 orf 3
gi|222429514|gb|GH737830.1|GH737830 orf 3
gi|222428890|gb|GH737541.1|GH737541 orf 1
gi|145281307|gb|ES322531.1|ES322531 orf 2
gi|222429762|gb|GH737998.1|GH737998 orf 1
gi|145281222|gb|ES322446.1|ES322446 orf 2
gi|145280711|gb|ES321935.1|ES321935 orf 1
gi|116360642|gb|EG374448.1|EG374448 orf 1
gi|145280830|gb|ES322054.1|ES322054 orf 1
gi|222428852|gb|GH737503.1|GH737503 orf 2
gi|222429870|gb|GH738052.1|GH738052 orf 1
gi|116360590|gb|EG374396.1|EG374396 orf 7
gi|222429123|gb|GH737194.1|GH737194 orf 4
gi|222430124|gb|GH737768.1|GH737768 orf 2
gi|145281490|gb|ES322714.1|ES322714 orf 1
gi|145280942|gb|ES322166.1|ES322166 orf 1
gi|145281737|gb|ES322961.1|ES322961 orf 4
gi|222429873|gb|GH738055.1|GH738055 orf 4
gi|222429997|gb|GH738179.1|GH738179 orf 1
gi|222429299|gb|GH737076.1|GH737076 orf 2
gi|145281002|gb|ES322226.1|ES322226 orf 1
gi|145280820|gb|ES322044.1|ES322044 orf 1
gi|222429808|gb|GH738044.1|GH738044 orf 1
gi|145281618|gb|ES322842.1|ES322842 orf 2
gi|145281739|gb|ES322963.1|ES322963 orf 5
gi|222429621|gb|GH738416.1|GH738416 orf 4
gi|222429390|gb|GH737306.1|GH737306 orf 3
gi|222428836|gb|GH737487.1|GH737487 orf 1
gi|222428982|gb|GH737633.1|GH737633 orf 2
gi|222430111|gb|GH737755.1|GH737755 orf 1
gi|145281268|gb|ES322492.1|ES322492 orf 2
gi|222428814|gb|GH737465.1|GH737465 orf 1
gi|145280947|gb|ES322171.1|ES322171 orf 1
gi|116360679|gb|EG374485.1|EG374485 orf 3
gi|145281066|gb|ES322290.1|ES322290 orf 4
gi|116360627|gb|EG374433.1|EG374433 orf 7
gi|145280743|gb|ES321967.1|ES321967 orf 1
gi|222429313|gb|GH737730.1|GH737730 orf 4

Figure 20. Phylogram of 95 predicted PST ORFs.

In Figure 20, phylogram also demonstrates the difference between predicted PST ORFs. It is visualized by MEGA5 (Tamura et al., 2011) and created by ClustalX (Larkin et al., 2007).

3.1.7 Blast2GO Annotation

Blast2GO annotation tool was used on all sequence sets to determine their availability in databases.

The Pst EST sequences of 2848 fed into Blast2GO to annotate, which resulted with hits of 1453 and 1395 with no hits. Sequence annotation is very important step in sequence analysis. Figure 21 is a detailed chart of Blast2GO results.

In addition to that, all Bgh, Pt and Pgt reported effectors were also tested by Blast2GO. Pgt has more BLAST hits than others. Pgt sequences were added to databases as hypothetical proteins but annotations of Bgh and Pt were not present in databases. Many of them showed no significant hit in other organisms.

Blast2GO is also used to annotate 95 predicted effectors. According to annotation, 40 of them have no significant BLAST hits. The rest of them are generally related to Pgt and few of them resulted with different database information (Supplementary Table 2).

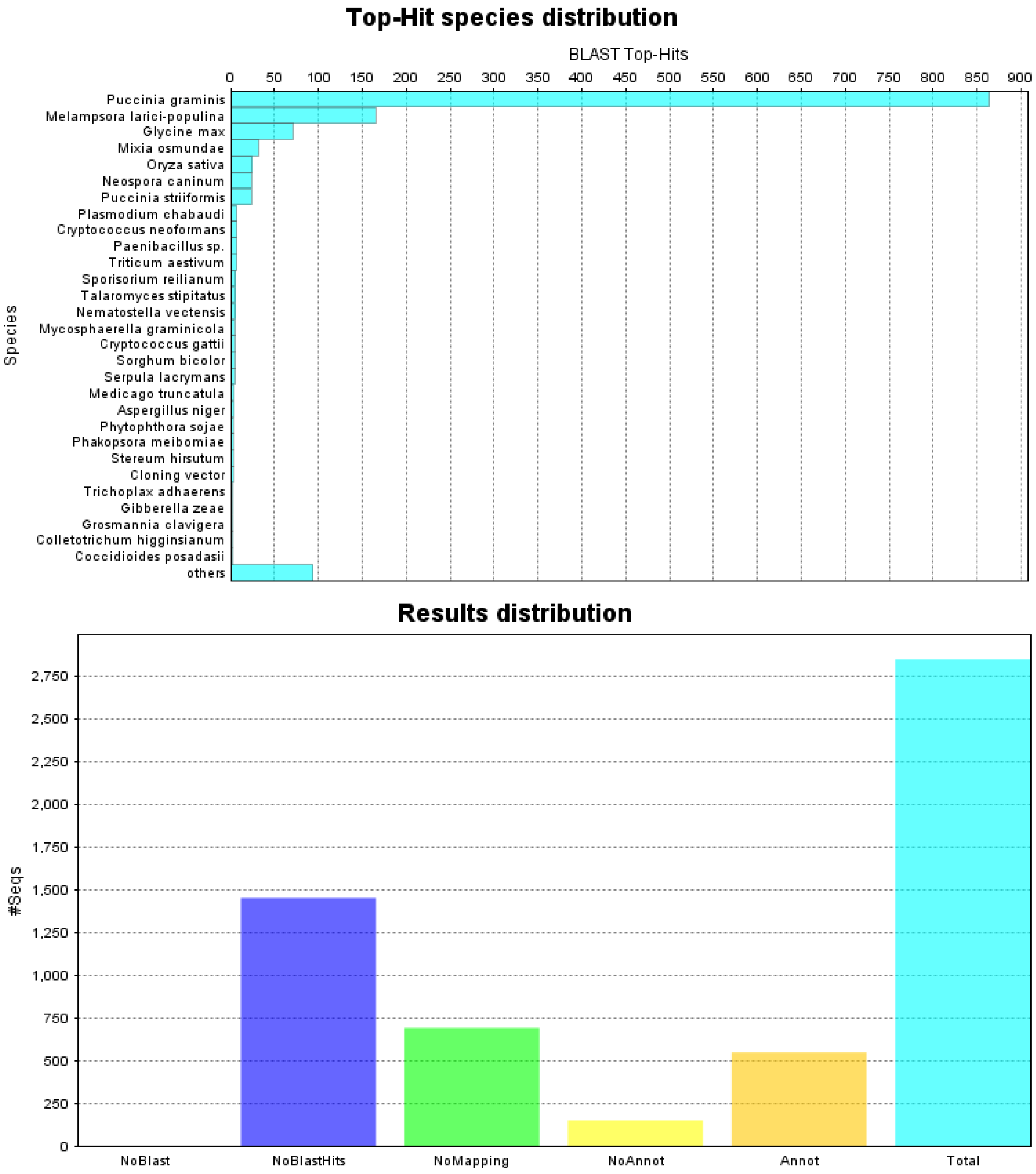


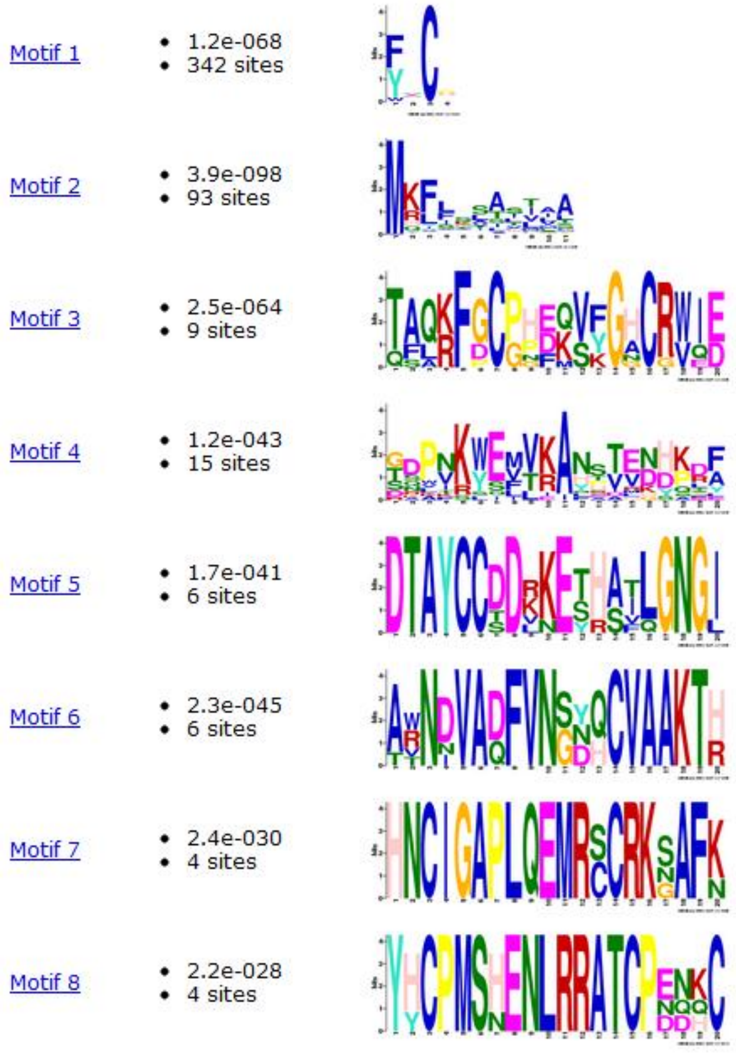
Figure 21. Species distribution and results distribution of Pst ESTs, from Blast2GO

Puccinia graminis has the most hits as expected because it is the closest relative of Pst. The other hits, other than plant pathogen family are not effectors and they are probably other genes. Second best hit organism, *Melampsora larici-populina,* is also a pathogen. Results distribution chart shows the annotation, blast and mapping results of ESTs.

3.1.8 MEME Suite Results

We wanted to show if we are obtaining similar MEME Suite results with reported effector candidates. Though Godfrey et al., 2010 found motif region, they did not perform any MEME Suite analysis.

All of the Bgh, Pgt and Pt reported effectors resulted in Y/F/WxC motif in MEME Suite analysis (Motif 1 in Figure 23). Figure 23 is graphical representation of all results and it shows the conserved regions in among reported effector candidates.



Motif 1	• 1.2e-068 • 342 sites
Motif 2	• 3.9e-098 • 93 sites
Motif 3	• 2.5e-064 • 9 sites
Motif 4	• 1.2e-043 • 15 sites
Motif 5	• 1.7e-041 • 6 sites
Motif 6	• 2.3e-045 • 6 sites
Motif 7	• 2.4e-030 • 4 sites
Motif 8	• 2.2e-028 • 4 sites

Figure 22. MEME results of Bgh, Pgt and Pt effectors.

In MEME Suite parameter menu, minimum motif size was adjusted to 2 and maximum motif size was adjusted to 20. Only first and second motifs are available in a certain number of effectors which is also another proof of low sequence similarity as it is detected in MSA. However, MEME analysis also emphasizes importance of these conserved regions.

GLAM2, which is gapped motif finder of MEME Suite, resulted in interesting results for predicted ORFs. We used GLAM2 to analyze predicted ORFs and candidate effectors. Figure 24 and Figure 25 are two GLAM2 prediction results.

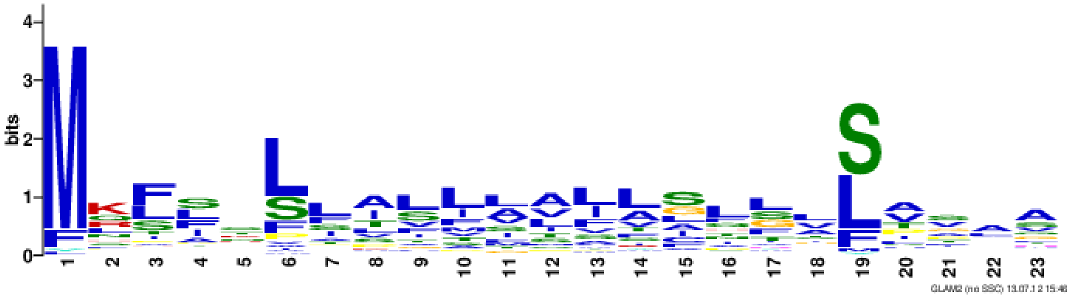


Figure 23. GLAM2 logo of SP region.

When all reported effectors and our predicted ORFs send to GLAM2, it produces this logo. It obviously shows a certain level of conservation in SP region. More interestingly, nearly all of the candidate effectors produce this logo in GLAM2 (333 out of 342) which is a proof of SP region similarity in putative effector candidates. Moreover, 61 of our predicted ORFs have this region.

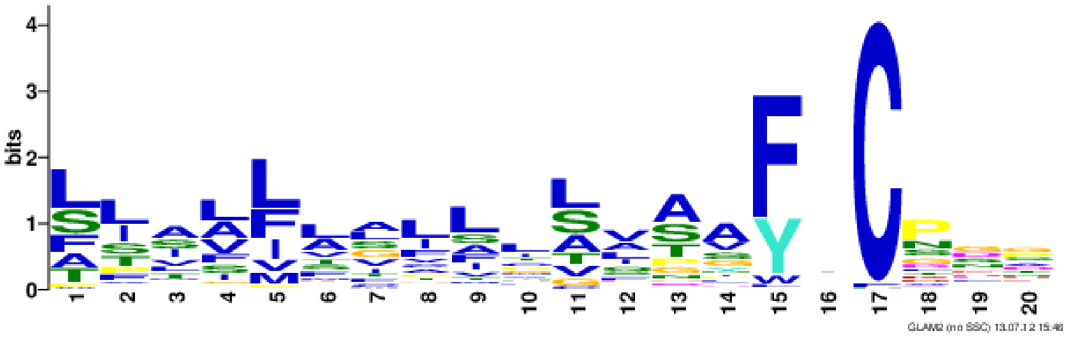


Figure 24. GLAM2 logo of motif region.

Figure 24 shows the reported candidate effectors and our predicted ORFs sharing a common Y/F/WxC motif and also there is a certain level of conservation detected according to Figure 23 logo. MEME Suite shows, all putative candidate effectors have this motif site. Furthermore, 50 of the predicted ORFs share this logo.

Though Y/F/WxC motif containing reported effectors are considered to have motif site a few amino acids downstream of SP cleavage site, our analysis with MEME Suite shows that this is not true in all cases (Figure 25).

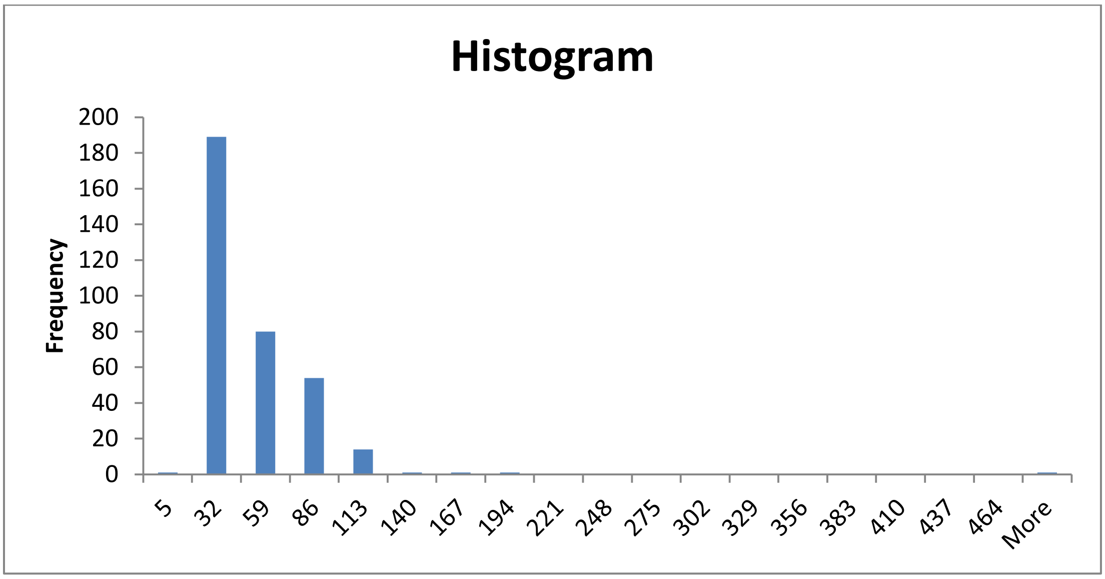


Figure 25. Histogram of motif starting locations in reported candidate effectors.

In Figure 25 shows a histogram of motif starting locations of Y/F/WxC motif containing reported effectors (Godfrey et al., 2010). According to MEME Suite, motif locations are varying. They are not a few downstream of SP cleavage site in all effector candidates as mentioned.

3.1.9 MicroRNA Mining and Target Prediction

Local blasts of plant miRNAs in WGS of Bgh, Pt and Pgt showed no significant results but target prediction has interesting results.

The prediction presented that *tae-miR1134* may regulate most of the Pst ESTs and those ESTs are derived from expressed genes of Pst. *Tae-miR1134* is a wheat (*Triticum aestivum*) microRNA. It has no known targets in wheat (Yao et al., 2007).

Tae-miR1134 has also some possible targets in Pgt too. O*sa-miR1877* and *osa-miR1881* have high score predictions against Pgt effector candidates. Both *osa-miR1877* and *osa-miR1881* are rice microRNAs and they have no predicted targets (Zhu et al., 2008).

Osa-miR2124 family generally has high score predictions for Bgh effector candidates. This family has been predicted to target f-box protein, hydrolase, leucine rich repeat domain-containing proteins and some other unknown proteins (Huang et al., 2009). However, this miRNA family was removed from database while this thesis was being written but we did not remove because it may show another important relationship.

Osa-miR2097-3p and *osa-miR1877* have high score prediction for Pt effector candidates. *osa-miR2097* family was predicted to target NBS-LRR disease resistance protein (Xue et al., 2009). *osa-miR1877* has no known targets (Zhu et al., 2008).

These are just a summary of our prediction. All predictions were also written into an excel file and can be accessed as a supplementary material (Supplementary Material 6).

CHAPTER IV

DISCUSSION

4.1 ORF Prediction

We have compiled 2848 *Puccinia striiformis* f. sp. *tritici* (Pst) ESTs. Initially we have predicted 9854 open reading frames (ORF). All research related to effector candidates based on this ORF set and needless to say, most of these predictions are meaningless but the real task is to find right ORF sequences.

There are some other open reading frame prediction methods but ORF Finder is a quite popular one. The possible mistakes and exceptions were manually tested randomly to ensure the predictions.

4.2 Multiple Sequence Alignment

Multiple sequence alignments were executed by both MUSCLE and Clustal algorithms. MUSCLE integrated to MEGA was used and ClustalX GUI is used to run Clustal.

To understand general picture of all putative candidate effectors (also known as Y/F/WxC effector candidates) they were aligned with each other and these alignments compared to predicted ORFs alignments. Some of these results had been added as figures and statistics in Results section.

MSA gives invaluable information to classify, refine and validate ORF predictions. Y/F/WxC motif is the most important feature of candidate effectors. On the other hand, MSA cannot show that motif region very successfully, both MUSCLE and Clustal cannot align correctly all sequences which contain that motif region.

Effector candidates are known to have very little sequence similarities among them. We have also confirmed this, but we detected that their SP region has a certain level of similarity (Figure 24). Furthermore, there are similarities among some of the sequences but not pronounced as Y/F/WxC motif.

4.3 Signal Peptide Prediction

Likewise to MSA step, SP prediction is very important for annotation and organization step. All the reported effector candidates were tested. However, not all of them contain SP region which is an interesting result. Although they are referred as candidates, effectors should have some kind of secretion machinery for translocation. May be they can be classified as false negatives of SignalP-4.0 or they have completely different secretion system.

For all the predicted ORFs 880 (out of 9854) of them have possible SP region. SignalP-4.0 showed that, 32 of 95 similarity predicted ORFs cotain probable SP region. 278 of 342 putative effector candidates are predicted to have SP region. Figure 26 and Figure 27 show the distribution of *D-value* of SignalP prediction results. Though, all reported effectors are thought to be secreted, SignalP prediction differs; but we can assume that SignalP positive results are likely to be secreted and they are possible effector candidates.

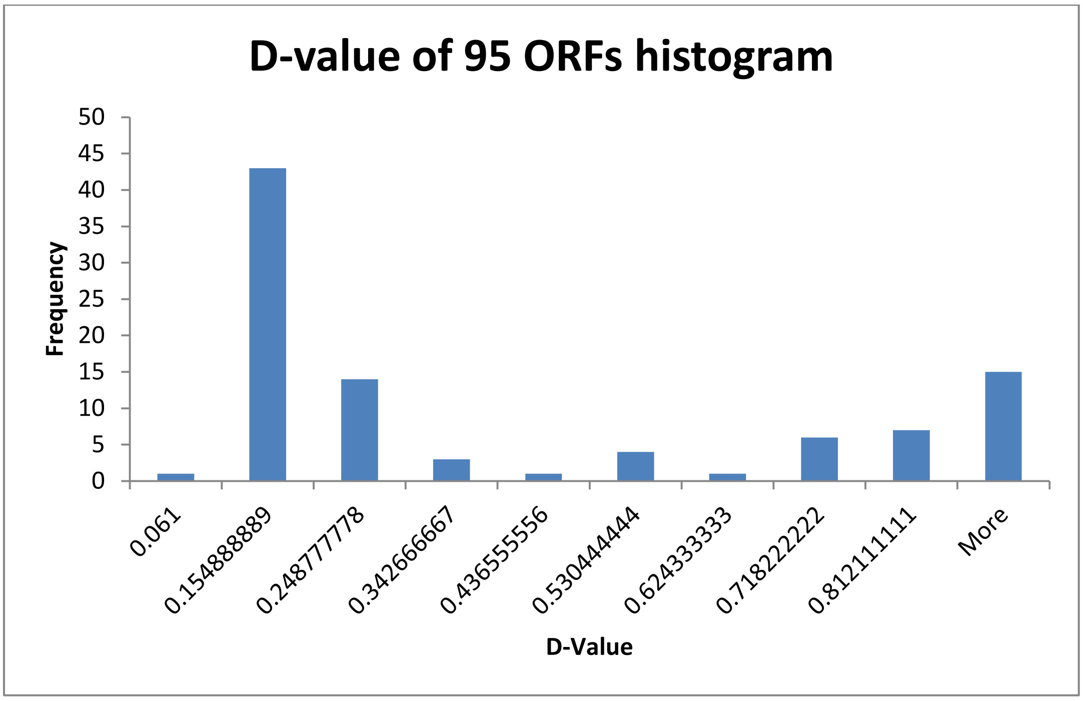


Figure 26. Histogram of 95 predicted ORFs' D-values.

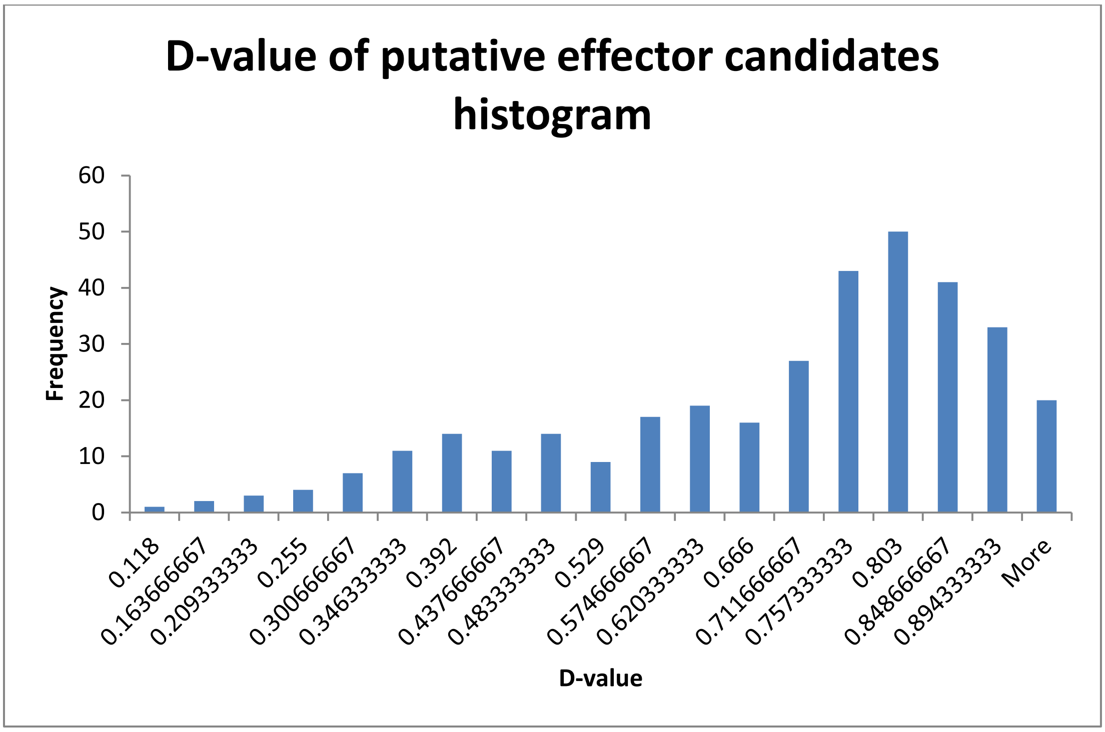


Figure 27. Histogram of reported effector candidates' D-values.

4.4 Predicted Effector Candidates

Various comparisons between predicted ORFs and candidate effectors make it clear that similarity between SP region and motif region is obvious in certain ORFs.

A table in Appendix 3 is presented with all 95 predicted ORFs. All of them are most likely effector candidates or they have a relation with putative effector candidates. In this table we marked our most likely prediction with bold. The sequences marked with a star (*) are also found in other studies. Co-finding of those sequences can be considered as a validation of our findings. Table 10 shows the best of our predictions refined from 95 predicted ORFs. Figure 28 is phylogram of these 30 refined ORFs.

More detailed version of Appendix 3 table can be accessed as an online excel table (Supplementary Table 1).

Table 10. Accession numbers of top candidates (30 in total).

Accession Numbers of best predicted ORFs.	
gi|116360518|gb|EG374324.1|EG374324_orf_4	gi|145281011|gb|ES322235.1|ES322235_orf_1
gi|116360529|gb|EG374335.1|EG374335_orf_3	gi|145281015|gb|ES322239.1|ES322239_orf_1
gi|116360642|gb|EG374448.1|EG374448_orf_1	gi|145281363|gb|ES322587.1|ES322587_orf_1
gi|116360647|gb|EG374453.1|EG374453_orf_2	gi|145281766|gb|ES322990.1|ES322990_orf_1
gi|145280708|gb|ES321932.1|ES321932_orf_1	gi|222428929|gb|GH737580.1|GH737580_orf_1
gi|145280711|gb|ES321935.1|ES321935_orf_2	gi|222429011|gb|GH737102.1|GH737102_orf_1
gi|145280743|gb|ES321967.1|ES321967_orf_1	gi|222429433|gb|GH738308.1|GH738308_orf_1
gi|145280758|gb|ES321982.1|ES321982_orf_2	gi|222429771|gb|GH738007.1|GH738007_orf_1
gi|145280773|gb|ES321997.1|ES321997_orf_2	gi|222430111|gb|GH737755.1|GH737755_orf_1
gi|145280791|gb|ES322015.1|ES322015_orf_1	gi|145280836|gb|ES322060.1|ES322060_orf_2
gi|145280810|gb|ES322034.1|ES322034_orf_1	gi|145280839|gb|ES322063.1|ES322063_orf_1
gi|145280827|gb|ES322051.1|ES322051_orf_1	gi|145280842|gb|ES322066.1|ES322066_orf_1
gi|145280830|gb|ES322054.1|ES322054_orf_1	gi|145280873|gb|ES322097.1|ES322097_orf_1
gi|145280833|gb|ES322057.1|ES322057_orf_1	gi|145280905|gb|ES322129.1|ES322129_orf_1
gi|145280959|gb|ES322183.1|ES322183_orf_1	gi|145280919|gb|ES322143.1|ES322143_orf_1

All of these sequences are predicted to have a similar motif region, SP region and predicted to be secretory by SignalP-4.0. Sequences are available in Appendix 1 and more extensive tables available as Appendix 3 and Supplementary Material 2.

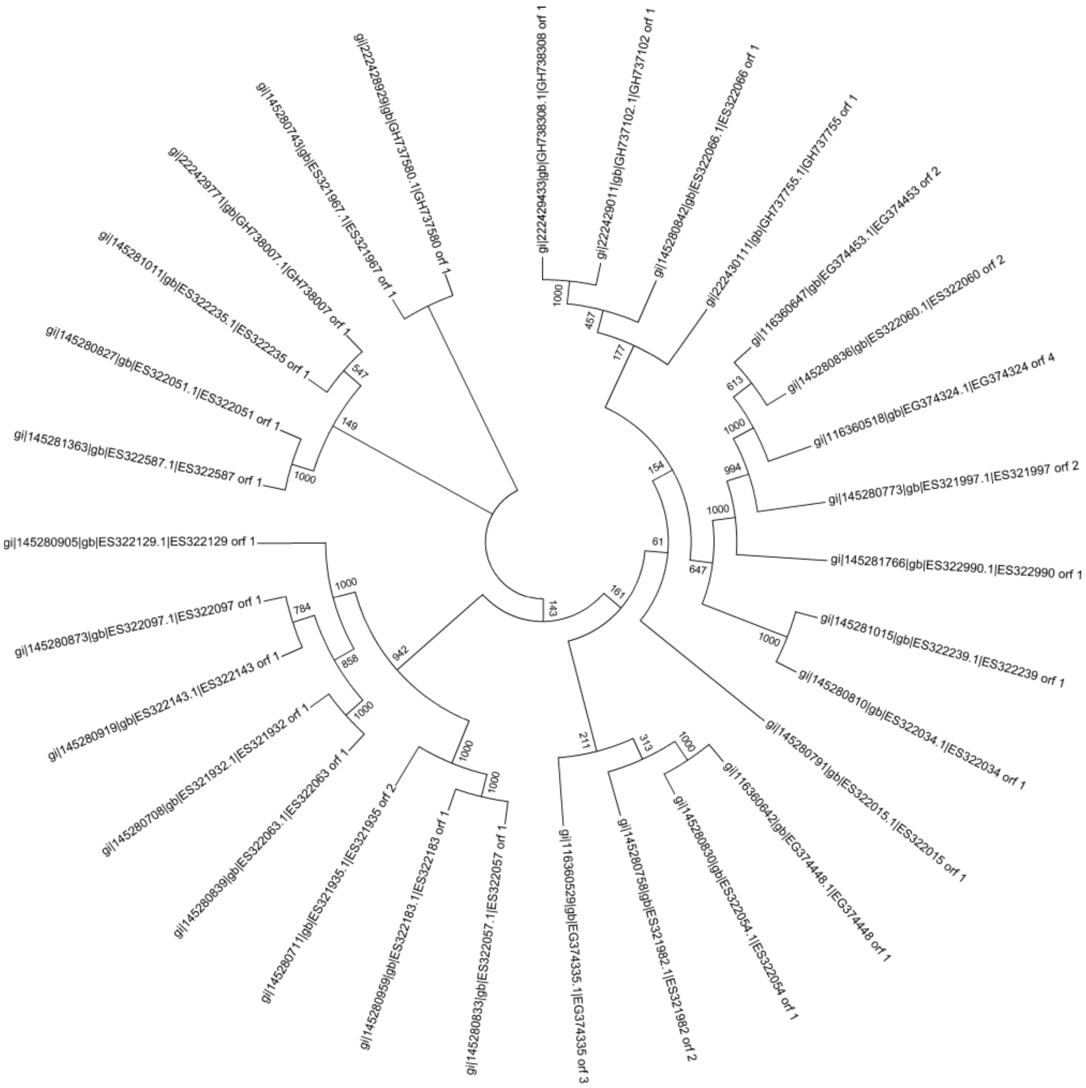

gi|222429433|gb|GH738308.1|GH738308 orf 1
gi|222429011|gb|GH737102.1|GH737102 orf 1
gi|145280842|gb|ES322066.1|ES322066 orf 1
gi|222430111|gb|GH737755.1|GH737755 orf 1
gi|116360647|gb|EG374453.1|EG374453 orf 2
gi|145280836|gb|ES322060.1|ES322060 orf 2
gi|116360518|gb|EG374324.1|EG374324 orf 4
gi|145280773|gb|ES321997.1|ES321997 orf 2
gi|145281766|gb|ES322990.1|ES322990 orf 1
gi|145281015|gb|ES322239.1|ES322239 orf 1
gi|145280810|gb|ES322034.1|ES322034 orf 1
gi|145280791|gb|ES322015.1|ES322015 orf 1
gi|116360642|gb|EG374448.1|EG374448 orf 1
gi|145280830|gb|ES322054.1|ES322054 orf 1
gi|145280758|gb|ES321982.1|ES321982 orf 2
gi|116360529|gb|EG374335.1|EG374335 orf 3
gi|145280833|gb|ES322057.1|ES322057 orf 1
gi|145280959|gb|ES322183.1|ES322183 orf 1
gi|145280711|gb|ES321935.1|ES321935 orf 2
gi|145280839|gb|ES322063.1|ES322063 orf 1
gi|145280708|gb|ES321932.1|ES321932 orf 1
gi|145280919|gb|ES322143.1|ES322143 orf 1
gi|145280873|gb|ES322097.1|ES322097 orf 1
gi|145280905|gb|ES322129.1|ES322129 orf 1
gi|145281363|gb|ES322587.1|ES322587 orf 1
gi|145280827|gb|ES322051.1|ES322051 orf 1
gi|145281011|gb|ES322235.1|ES322235 orf 1
gi|222429771|gb|GH738007.1|GH738007 orf 1
gi|145280743|gb|ES321967.1|ES321967 orf 1
gi|222428929|gb|GH737580.1|GH737580 orf 1
1000
457
177
613
1000
994
1000
154
61
647
1000
161
143
211
313
1000
1000
1000
942
1000
784
858
1000
547
149
1000

Figure 28. Phylogram of top 30 predicted ORFs.

4.5 Effector Mining Pipeline

After all data acquisition, similarity detection, motif prediction and alignment steps, we have created a chart to show effector mining pipeline. MEME Suite is the main indicator, if the motif availability is assumed. MSA and signal peptide prediction are also important to validate final data set. Therefore, we propose a pipeline to summarize all steps covered (Figure 27). In this figure, we write the number of sequences that we acquired and found in the brackets to ease the follow our procedure.

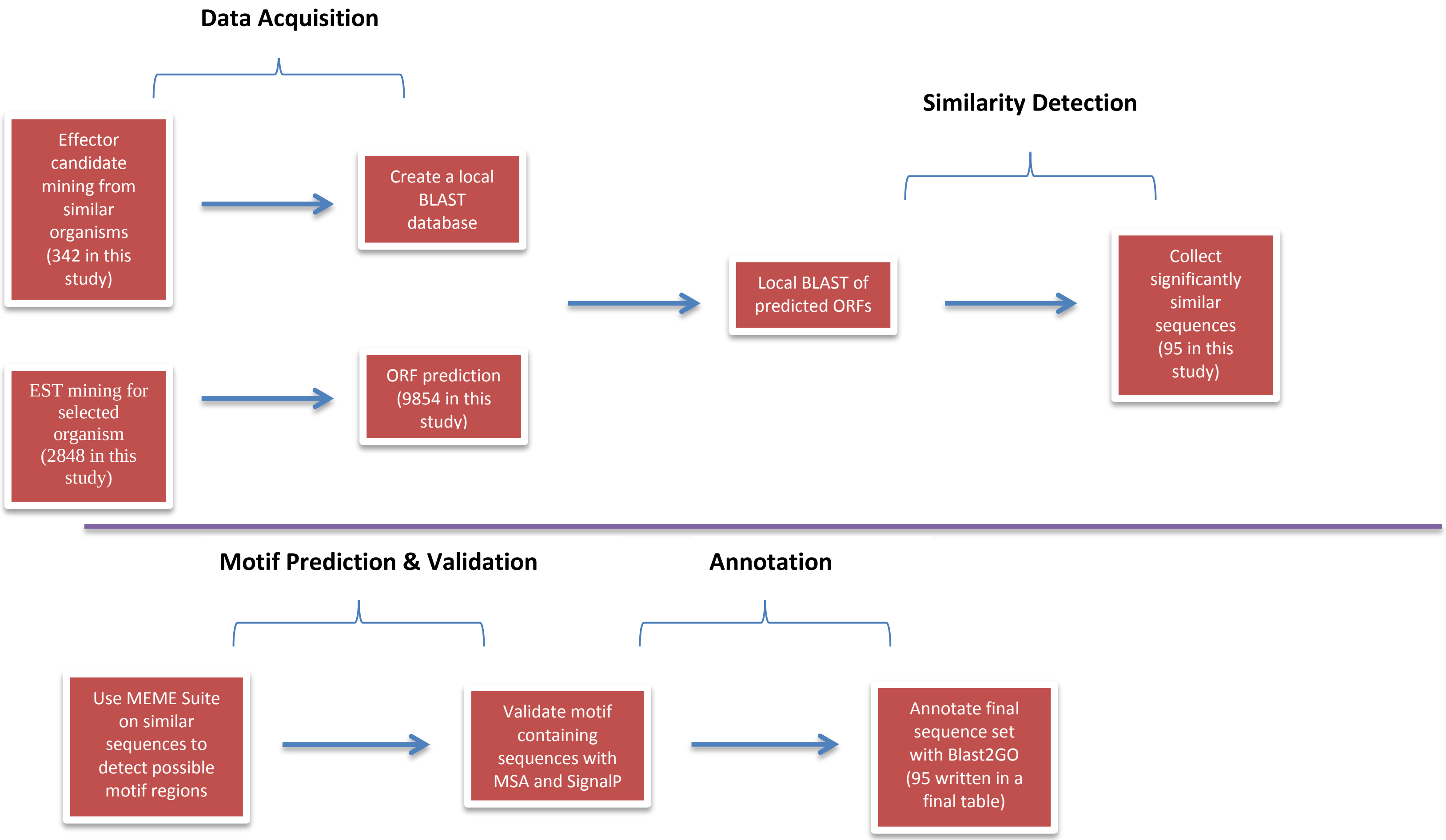

Data Acquisition
Effector candidate mining from similar organisms (342 in this study)
Create a local BLAST database
EST mining for selected organism (2848 in this study)
ORF prediction (9854 in this study)
Similarity Detection
Local BLAST of predicted ORFs
Collect significantly similar sequences (95 in this study)
Motif Prediction & Validation
Use MEME Suite on similar sequences to detect possible motif regions
Validate motif containing sequences with MSA and SignalP
Annotation
Annotate final sequence set with Blast2GO (95 written in a final table)

Figure 29. Effector mining pipeline, produced in this thesis.

4.6 MicroRNA Similarity and Target Prediction

BLAST analysis of plant microRNA precursors in WGS of Bgh, Pt and Pgt did not produce any significantly similar hits. This shows plant pathogens have no significantly similar miRNA genes to that of plants.

However, target prediction of plant miRNAs produced some likely target genes in pathogens. These miRNAs may be candidates for experimental analyses. Their possible targets in their originated plants are referenced in the Results section.

On the other hand, some probabilistic and energy minimization approaches might be better for target prediction between plant and plant pathogen. Moreover, the genes in pathogens except effectors could be tested for possible targeting. It is too immature to predict the presence of miRNA involved cross-kingdom regulation in plant-pathogen.

4.7 Microarray Design

We have created a microarray design as an additional study for Pst EST gene expression detection by using Agilent eArray software (https://earray.chem.agilent.com/earray/). Microarray analysis is an excellent method to detect gene expression levels of certain transcriptomes and metabolic pathways.

We have used default parameters with 4 oligonucleotide probes for a single EST and considered both sense and antisense strand of target ESTs because their orientation is not clear. All files and probe data can be accessed as an online material (Supplementary Material 5).

4.8 Future Studies

2848 ESTs were compiled and mined from which were presented three different articles (Ling et al., 2007; Zhang et al., 2008; Yin et al., 2009). The pipeline developed here can be applied to other EST sequences of Pst that were not part of this thesis and different pathogens to find more effector candidates. These results are also beneficial for experimental work; most likely candidates could be experimentally validated. This can decrease cost of experimental work. Microarray is also a suitable way to detect high-throughput expression detection.

The machine learning methods might be useful to detect possible effector candidates to separate effector candidates from other genes. We have already started to develop a new tool to classify sequences. It is located at http://www.baskent.edu.tr/~hogul/TRAINER/. This tool uses machine learning methods to determine and distinguish sequences. It might also be useful to detect other conserved assets of candidate sequences.

CHAPTER V

CONCLUSION

The objectives of this study are to analyze putative effector candidates of biotrophic plant rust and powdery mildew pathogens, to predict a set of novel effectors and to detect possible miRNA regions or miRNA targets between pathogen and their hosts.

Various bioinformatics tools and approaches were utilized to construct a logical analysis tool for effector candidate prediction using self-developed BioPython scripts (SUU). *Puccinia striiformis* f. sp. *tritici* (Pst), *Puccinia graminis* f. sp. *tritici* (Pgt), *Puccinia Triticina* (Pt) and *Blumeria graminis* f. sp. *hordei* (Bgh) are the four pathogens that candidate effectors, ESTs and various sequences were collected. Pst has no known effectors available, thus it was the main focus to find likely effectors. A new set of predicted sequences found which contain similar conserved regions. Y/F/WxC motif and signal peptide region are the main features of these group of candidate effectors in spite of low sequence similarity among them. It is thought that these features have some role in haustorium formation, protein-protein interaction and infection. Multiple sequence alignment (MSA) and MEME Suite results also confirmed this. It is possible to develop some other probabilistic approach, machine learning methods or gene detection algorithms to directly mine effector candidates from whole genome sequences (WGS) or better prediction using transcriptome sequences. In this study, 2848 Pst ESTs from NCBI database had been used and for further research other ESTs can be collected and analyzed. There is no WGS of Pst available yet. If it is completed, it will be useful to grasp full gene and intron-exon structure. Intron-exon structures have certain similarities in Bgh candidate effectors; therefore, this may also refine predictions of Pst. Furthermore, there are also other members available in biotrophic pathogen group and they may be mined for similar effectors. Blast2GO annotation tool is a good accessory to analyze sequences from an unknown origin. In every step of analysis, it was used to annotate sequences. Our pipeline in Figure 27 is a summarized chart of the thesis. There is no certain pipeline designated earlier, but by both literature search and our findings, we created a pipeline to guide further research.

Fungal microRNAs are disputed subject. There are certain similar miRNA families in plants, which is a starting point to find possible miRNA regions on plant pathogen genomes that belong to same families. On the other hand, BLASTs of microRNA genes of *Populus euphratica, Populus trichocarpa, Zea mays, Hordeum vulgare, Oryza sativa, Triticum aestivum, Triticum turgidum* and *Brachypodium distachyon* give no significant results on genomes of Pgt, Pt and Bgh. This means that pathogens have different kind of miRNA families than that of reference plants if any at all. Cross-kingdom regulation of miRNAs is a very new concept. Though, there is no proven miRNA existing in fungal pathogens, there may be an interaction between plant miRNAs and pathogenic genes. Varying expression level of plant miRNA during infection is an observed phenomenon. Smith-Waterman local alignment algorithm shows possible binding regions between pathogenic genes and plant miRNAs if they are assumed to have plant style interaction. Energy minimization and probabilistic approaches may refine this prediction, but in this study only nearly perfect Watson-Crick complementarity between miRNA and possible target were taken into account. As a result of these assumptions, we concluded that some plant miRNAs may regulate the genes in pathogens. The candidate targets used in this thesis are effector candidates of Bgh, Pt and Pgt, but it is also possible to test other type of genes to grasp a full picture.

We have added a future studies section to clarify our future route and possible implications of our findings. There is also a microarray design available as an online material which includes Pst ESTs.

SUPPLEMENTARY MATERIALS

We have supplied a list of supplementary materials throughout the sections of the thesis. In online version of thesis, they are clickable and they can be accessed as online materials. The list below also provides the web links of those materials.

Supplementary Material 1

http://curie.chem.metu.edu.tr/~akkayalab/suu/supp. material 1.fasta

Supplementary Material 2

http://curie.chem.metu.edu.tr/~akkayalab/suu/supp. material 2.fasta

Supplementary Material 3

http://curie.chem.metu.edu.tr/~akkayalab/suu/supp. material 3.fasta

Supplementary Material 4

http://curie.chem.metu.edu.tr/~akkayalab/suu/supp. material 4.fasta

Supplementary Material 5

http://curie.chem.metu.edu.tr/~akkayalab/suu/supp. material 5.zip

Supplementary Material 6

http://curie.chem.metu.edu.tr/~akkayalab/suu/supp. material 6.xlsx

Supplementary Table 1

http://curie.chem.metu.edu.tr/~akkayalab/suu/Supplementary Table 1.xlsx

Supplementary Table 2

http://curie.chem.metu.edu.tr/~akkayalab/suu/Supplementary Table 2.xlsx

REFERENCES

Bailey, T.L., Boden, M., Buske, F.A., Frith, M., Grant, C.E., Clementi, L., Ren, J., Li, W.W., Noble, W.S., 2009. MEME SUITE: tools for motif discovery and searching. Nucleic Acids Research 37, W202–8.

Bartel, D.P., 2009. MicroRNAs: target recognition and regulatory functions. Cell 136, 215–33.

Bendtsen, J.D., Nielsen, H., von Heijne, G., Brunak, S.D.A.-J.D.O.-S. [pii], 2004. Improved prediction of signal peptides: SignalP 3.0. J Mol Biol 340.

Chen, X., 2008. MicroRNA metabolism in plants. Current Topics in Microbiology and Immunology 320, 117–36.

Chenna, R., Sugawara, H., Koike, T., Lopez, R., Gibson, T.J., Higgins, D.G., Thompson, J.D., 2003. Multiple sequence alignment with the Clustal series of programs. Nucleic Acids Research 31, 3497–500.

Choo, K.H., Tan, T.W., Ranganathan, S.C.-P.D.O.-1471-2105-10-S.-S. [pii], 2009. A comprehensive assessment of N-terminal signal peptides prediction methods. BMC Bioinformatics 10 Suppl 1, S2 ST – A comprehensive assessment of N–terminal .

Cock, P.J.A., Antao, T., Chang, J.T., Chapman, B.A., Cox, C.J., Dalke, A., Friedberg, I., Hamelryck, T., Kauff, F., Wilczynski, B., de Hoon, M.J.L., 2009. Biopython: freely available Python tools for computational molecular biology and bioinformatics. Bioinformatics (Oxford, England) 25, 1422–3.

Conesa, A., Götz, S., García-Gómez, J.M., Terol, J., Talón, M., Robles, M., 2005. Blast2GO: a universal tool for annotation, visualization and analysis in functional genomics research. Bioinformatics (Oxford, England) 21, 3674–6.

Curtis, B.C., Rajaram, S., Macpherson, H.G., 2002. Bread Wheat [WWW Document]. FAO. URL http://www.fao.org/docrep/006/y4011e/y4011e00.htm#Contents

Durbin, R., Eddy, S., Krogh, A., Mitchison, G., 1998. Biological sequence analysis. Cambridge University Press.

Edgar, R.C., 2004. MUSCLE: multiple sequence alignment with high accuracy and high throughput. Nucleic Acids Research 32, 1792–7.

Godfrey, D., Böhlenius, H., Pedersen, C., Zhang, Z., Emmersen, J., Thordal-Christensen, H.C.-P.D.O.-1471-2164-11-317 [pii], 2010. Powdery mildew fungal effector candidates share N-terminal Y/F/WxC-motif. BMC Genomics 11.

Hall, T.A., 1999. BioEdit: a user-friendly biological sequence alignment editor and analysis program for Windows 95/98/NT. Nucleic Acids Symposium Series 41, 95–98.

Heun, M., 1997. Site of Einkorn Wheat Domestication Identified by DNA Fingerprinting. Science 278, 1312–1314.

Huang, S.Q., Peng, J., Qiu, C.X., Yang, Z.M., 2009. Heavy metal-regulated new microRNAs from rice. Journal of Inorganic Biochemistry 103, 282–7.

Huson, D.H., Richter, D.C., Rausch, C., Dezulian, T., Franz, M., Rupp, R., 2007. Dendroscope: An interactive viewer for large phylogenetic trees. BMC Bioinformatics 8, 460.

Jeanmougin, F., Thompson, J.D., Gouy, M., Higgins, D.G., Gibson, T.J., 1998. Multiple sequence alignment with Clustal X. Trends in Biochemical Sciences 23, 403–5.

Jiang, M., Sang, X., Hong, Z., 2012. Beyond nutrients: food-derived microRNAs provide cross-kingdom regulation. BioEssays : News and Reviews in Molecular, Cellular and Developmental Biology 34, 280–4.

Jones, H., Civáň, P., Cockram, J., Leigh, F.J., Smith, L.M., Jones, M.K., Charles, M.P., Molina-Cano, J.-L., Powell, W., Jones, G., Brown, T.A., 2011. Evolutionary history of barley cultivation in Europe revealed by genetic analysis of extant landraces. BMC Evolutionary Biology 11, 320.

Kusenda, B., Mraz, M., Mayer, J., Pospisilova, S., 2006. MicroRNA biogenesis, functionality and cancer relevance. Biomedical papers of the Medical Faculty of the University Palacký, Olomouc, Czechoslovakia 150, 205–15.

Larkin, M.A., Blackshields, G., Brown, N.P., Chenna, R., McGettigan, P.A., McWilliam, H., Valentin, F., Wallace, I.M., Wilm, A., Lopez, R., Thompson, J.D., Gibson, T.J., Higgins, D.G., 2007. Clustal W and Clustal X version 2.0. Bioinformatics (Oxford, England) 23, 2947–8.

Lev-Yadun, S., 2000. ARCHAEOLOGY:Enhanced: The Cradle of Agriculture. Science 288, 1602–1603.

Ling, P., Wang, M., Chen, X., Campbell, K.G., 2007. Construction and characterization of a full-length cDNA library for the wheat stripe rust pathogen (Puccinia striiformis f. sp. tritici). BMC Genomics 8, 145.

Lytle, J.R., Yario, T.A., Steitz, J.A., 2007. Target mRNAs are repressed as efficiently by microRNA-binding sites in the 5’ UTR as in the 3' UTR. Proceedings of the National Academy of Sciences of the United States of America 104, 9667–72.

Mangalam, H., 2002. The Bio* toolkits -- a brief overview. Briefings in Bioinformatics 3, 296–302.

Millar, A.A., Waterhouse, P.M., 2005. Plant and animal microRNAs: similarities and differences. Functional & Integrative Genomics 5, 129–35.

Mount, D.W., 2007. Using the Basic Local Alignment Search Tool (BLAST). CSH Protocols 2007, pdb.top17.

Petersen, T.N., Brunak, S., von Heijne, G., Nielsen, H., 2011. SignalP 4.0: discriminating signal peptides from transmembrane regions. Nature Methods 8, 785–6.

Ridout, C.J., Skamnioti, P., Porritt, O., Sacristan, S., Jones, J.D.G., Brown, J.K.M., 2006. Multiple avirulence paralogues in cereal powdery mildew fungi may contribute to parasite fitness and defeat of plant resistance. The Plant Cell 18, 2402–14.

Spanu, P.D., Abbott, J.C., Amselem, J., Burgis, T.A., Soanes, D.M., et al., 2010. Genome Expansion and Gene Loss in Powdery Mildew Fungi Reveal Tradeoffs in Extreme Parasitism. Science 330, 1543–1546.

Tamura, K., Peterson, D., Peterson, N., Stecher, G., Nei, M., Kumar, S., 2011. MEGA5: molecular evolutionary genetics analysis using maximum likelihood, evolutionary distance, and maximum parsimony methods. Molecular Biology and Evolution 28, 2731–9.

Xue, L.-J., Zhang, J.-J., Xue, H.-W., 2009. Characterization and expression profiles of miRNAs in rice seeds. Nucleic Acids Research 37, 916–30.

Yao, Y., Guo, G., Ni, Z., Sunkar, R., Du, J., Zhu, J.-K., Sun, Q., 2007. Cloning and characterization of microRNAs from wheat (Triticum aestivum L.). Genome Biology 8, R96.

Yin, C., Chen, X., Wang, X., Han, Q., Kang, Z., Hulbert, S.H., 2009. Generation and analysis of expression sequence tags from haustoria of the wheat stripe rust fungus Puccinia striiformis f. sp. Tritici. BMC Genomics 10, 626.

Zhang, Henderson, C., Perfect, E., Carver, T.L.W., Thomas, B.J., Skamnioti, P., Gurr, S.J., 2005. Of genes and genomes, needles and haystacks: Blumeria graminis and functionality. Molecular Plant Pathology 6, 561–75.

Zhang, L., Hou, D., Chen, X., Li, D., Zhu, L., Zhang, Y., Li, J., Bian, Z., Liang, X., Cai, X., Yin, Y., Wang, C., Zhang, T., Zhu, D., Zhang, D., Xu, J., Chen, Q., Ba, Y., Liu, J., Wang, Q., Chen, J., Wang, J., Wang, M., Zhang, Q., Zhang, J., Zen, K., Zhang, C.-Y., 2012. Exogenous plant MIR168a specifically targets mammalian LDLRAP1: evidence of cross-kingdom regulation by microRNA. Cell Research 22, 107–26.

Zhang, Y., Jiang, W., Gao, L., 2011. Evolution of microRNA genes in Oryza sativa and Arabidopsis thaliana: an update of the inverted duplication model. PloS One 6, e28073.

Zhang, Y., Qu, Z., Zheng, W., Liu, B., Wang, X., Xue, X., Xu, L., Huang, L., Han, Q., Zhao, J., Kang, Z., 2008. Stage-specific gene expression during urediniospore germination in Puccinia striiformis f. sp tritici. BMC Genomics 9, 203.

Zhu, Q.-H., Spriggs, A., Matthew, L., Fan, L., Kennedy, G., Gubler, F., Helliwell, C., 2008. A diverse set of microRNAs and microRNA-like small RNAs in developing rice grains. Genome Research 18, 1456–65.

APPENDIXES

Appendix 1. Predicted ORFs which have significant BLAST hits (95 in total).

```
>gi|145281423|gb|ES322647.1|ES322647_orf_4
MYDEPMLNCWAFLGFISTPGQYMRNKQVCHEQDMGIVWNLKDS*
>gi|116360588|gb|EG374394.1|EG374394_orf_1
MPGNSAITAGDLQSFQSAKFYTTTIMQSVKPTIVLVALLAGAISFVSVEGSKVFPCLHDHPSGFCGVKKDDKFLLW
PAFPEGKGFTCGKAGGIAYCSNTKDGFEDSDPDRYDKWIGDHCKQA*
>gi|145281060|gb|ES322284.1|ES322284_orf_4
MDLRAGSRHFHHRLRSIIHPNSIDRIPFHHLYH*
>gi|145280833|gb|ES322057.1|ES322057_orf_1
MNTTLYALLSLAVATLSVGAVGQTNQCQFYASVGVKGYTCNERPDIICSEGCKSFVTMTGCVLTQYPKKPATTELC
TVGYGRDTAAFKACLTSQGAFRCNGTSTGQATCSGCVPRGDVTWAN*
>gi|222429997|gb|GH738179.1|GH738179_orf_1
LEFIKGWYLRLSLDQLATFCAEADQELEAVAPQIYHRAAAITEYTDTGIFLPELGEADAEVPPNWFRLDPEGGEDL
AEEIASSEEGEEEAKKKKKKKKKKKH
>gi|145281327|gb|ES322551.1|ES322551_orf_2
GTRDSAITAGGRDTAAFKACLTSQGAFRCTGTSTGHATCSGCVPRGDVTWAN*
>gi|145281363|gb|ES322587.1|ES322587_orf_1
MFMLRFLGLIASVLLIAPCKGEDVPYHYFRCGKNIDAICSDRIPNTDQQKLVWAVRLEKGKRRYKCPALLTSFCCW
QGKFDINGHHGELTVPRDATFDPCTQV*
>gi|145280711|gb|ES321935.1|ES321935_orf_1
LRDARHSSPRQTVFLPSTLRSRLPQRSAPLALDVTRPLSKLVSLAKVLSGAMAPRLEEPLAMVAFQTVASPGPTRD
IVNQLFWTLIQSLLYTHPLIRLLCVCSCQLTHRNAIEDTTSLSKQKKKKKKKKK
>gi|145280711|gb|ES321935.1|ES321935_orf_2
MNTALYALFSLAAATSSVGAVAETSQCQFYASVGVKGYTCNESPDYICSAGCSSFVTATNCVLTQYPKKPPTTEVC
TLGFGRDTAAFKACLTGQGSFRCNGTSVGRATCHGCVPNGGVTWAN*
>gi|222429299|gb|GH737076.1|GH737076_orf_2
MYLGADVWQSPNGHDILGIVIYRLVEKDGVKFELEAMPLDFVRRVKNHTGEYLAETMRVVVEKFGVQDKVR*
>gi|145281411|gb|ES322635.1|ES322635_orf_3
FFFFFFRLRSLSFSCSSLLLVADESKLRFLLLLWSVIGLISVAETRSVVSYPLKTTV*
>gi|145281737|gb|ES322961.1|ES322961_orf_4
MMMMMMMLWMRMERRVDKSIWIQWLISDTMTSLRRSPKSHLEGNDKTNRLTRLSNQKIRPPRKST*
>gi|145280964|gb|ES322188.1|ES322188_orf_1
MRQGGTILTVNGSQVAVLHTLLALGAFGTALVLGCYLHYQKIVKNEWYGYPQEWFPSVSATIGDYYPERPIFQILI
AFNSGTPTFFLYI*
>gi|145281066|gb|ES322290.1|ES322290_orf_4
MGQFRIEMVLRIQPKLKFGFQCFHLSIGFQTCVGGRPYPFRSSRTAS*
>gi|116360702|gb|EG374508.1|EG374508_orf_3
MNRRDRKNPKGSNPRKKPQKTPPRGIPKLKNHAWPKGLPETPPPARGPMAETPHANQIGPLHPSPKPPPFTVKPGR
NFPAAGIPKGTGPHDKTRGRPHPPPNYLGPGRAH
>gi|145280834|gb|ES322058.1|ES322058_orf_1
MSAPEKFVSQGMIPSIGSTLPASEGAAVPPKSSDVSNQAGSVTPGRPTTVSTIPVRSHGGSKPSFRLKRKTLSFER
LF*
>gi|145280905|gb|ES322129.1|ES322129_orf_1
MFHPVLPSLVVVCILGLLNVVRADDLDYAYRYYPSGNELRVDGTKDSYDCPANCQSFYHATGCTIDDGSSKEKTTQ
VCSNRYAPSGASGKACTNAALKRYICTGVEGSSKFKCSGCKIVPP*
>gi|222428836|gb|GH737487.1|GH737487_orf_1
LCMYSPGFDELNTPRSRSHDGYYITRGLFRLQSQEGRSSSWRTSDATPNIVESVQDLETENIDFSKRIMHASWLMA
TIFSLCSRQFGVNDSQKHKSQGWDWSDNSQKKKKKKKKTCRPPRP
>gi|145280791|gb|ES322015.1|ES322015_orf_1
MSLLRFLVVLACSATFGVSAANPKTTVEFECGQPRPIGWCAIKRPSKSIYMVADANVVVSGSGGRGYNCINKGESK
WCCPLTWVPDSRGNAIIIDFEMTCSRK*
>gi|145280717|gb|ES321941.1|ES321941_orf_2
MGSQRISEQEYKATKSTESTINDGQGGREGRLDSWLSAAACCELASGHAYVGAETRYGRKGIAWIERM*
>gi|145280743|gb|ES321967.1|ES321967_orf_1
```

```
MISSFNIIVKLALVGFFASLPEVLACSGQTSTLACWNAGWEQPDPSGPGGCPYGASVLCCSGDPLTQTNPRPSCIY
PNGAIFHVPSHPKRFKL*
>gi|222428962|gb|GH737613.1|GH737613_orf_2
FFFFFFFFGRAAQGRGEGGHRQQSAHPPHAWPLPSSGFLWCGRRPDEHRLIGRQTSPLPIANVADHAYVARLFPAQ
A*
>gi|222429313|gb|GH737730.1|GH737730_orf_4
MANISRLFHYPIDYHVQALPISLATTLGISVDFFSCSYLDVSVRRVRFAYLCIQHAMT*
>gi|145280919|gb|ES322143.1|ES322143_orf_1
MLFSVLAVFMMVQGRSVIGAGFQCLDPARAQALCSRPPTAPQDHTVTIVKPYRIGDDYFCPPRLDAEIPVCCKTDM
YMRYMASGWKTILPNDTYSAACFPPVHLPDPPKVDLTDALRYYPAGDGINLHVDTKTGGSFNCPVKTCKSSYGGIG
CTHDDIPGLGKANQTCSHLFGAKGATQISCGNLRNLEMIAFTCDRVDPASKFACSGCTFTDA*
>gi|116360590|gb|EG374396.1|EG374396_orf_7
MPFCPWKGQAFPVFCHGKSGLSSMPFFGNIMKRDEDTPFTEELPAMPSVPGGDSGSRSLVNGIRTFFHSDPWQHSQ
AR*
>gi|116360642|gb|EG374448.1|EG374448_orf_1
MFTSRLLLSSVLCVIAAVVTATTPAPLGPLNLCSDKDSVYKVTGVLQLNGTVSRVDGQGHPTGNPDGICECQPTGI
NCNTRPSAPIFDGPPVFCANPITNKCDAPAPQKLCSPAGSKYQVVGVIHPDGSVSSVDAKGTVSPRASGICTCTPN
GVPKCHLAPTNTVFSPLGLLC*
>gi|116360691|gb|EG374497.1|EG374497_orf_2
MSSWGFHPNGIYSQPEIQHPDYQAPVSNCRRTYNSCQSGCSQIWNWPSLPGFNPRLQPSQYAGYRGHGRDAPVQMC
PNVKETACPIFANATG*
>gi|116360526|gb|EG374332.1|EG374332_orf_1
MSSCGSHLNWIYSQPETQHPDYQAPVSNCRRTYNSCQFGCSQIWNWPSPPGFNPRPQPSQYAGYRRHRRDAPAQMC
PNVKETACPIFANATGYECLDVKTEIHLMWWM*
>gi|222428890|gb|GH737541.1|GH737541_orf_1
MWCHPTPPEHRIPQEVAVSPCKGHSIYFQTIFPIDCPEIYCILIPGHLLPVVPYPSSSHSNSPQIVPLPSTIYPAP
SAIQHLYFLIALPLPPIILEYFLTCANQFPSKFPAPSPLVILVIVQE*
>gi|116360647|gb|EG374453.1|EG374453_orf_2
MNFWGSAILLVASIGHLVAGQQVFHCPKSAPYAHCGTNNYAAVPPTWDITNAAKNGNTYDCPGGDQITLCCHIGGE
PSFSSKADYDKWVKDHCI*
>gi|145281618|gb|ES322842.1|ES322842_orf_2
LYCHFDRILWQMIFPSPSKKKKKKKKKKKKKKKKKKNLGGRL
>gi|222430111|gb|GH737755.1|GH737755_orf_1
MPCIKSSTILLIVLISSMFQVIMIKAFSANIICNTVPRTVGVCLLPSTNQGGAIANYWTLPASKKQGGVFTCDDQT
MGGSYAKTKTSCCDPLLQLSPQPPAAAKPQFKEIPVSDFNRFCNTPS*
>gi|116360662|gb|EG374468.1|EG374468_orf_1
MPREFGAITAGDPPHNSPAIRYHLPLLTSFFGLHLPAFNSTRSNISAKMNTTLYALLSLAVATLSVGAVGQTNQCQ
FYASVGVKGYTCNERPDIICSEGCKSFVTMTGCVLTQYPKKPATTELCTVGYGRDTAAFKACLTSQGAFRCNGTST
GQATCSGCVPRGDVTWAN*
>gi|222429370|gb|GH737286.1|GH737286_orf_1
MWIDFKARCRESAIDYESTRSRVGWCRFVYLDLALIFLLFRLSFLIRFLAHLALMLADC*
>gi|222429123|gb|GH737194.1|GH737194_orf_4
GPHSCHVLQISTNKHPSTSRSSGHPRFPASRSPQGRRTPRCRSRIGHRCRTIRCYCYRRWTWWIRCRHQGRSAGLQ
DCLC*
>gi|222429861|gb|GH738496.1|GH738496_orf_1
LASLIAPNRFPANEPIANPGKFPTMNPSVPPDAAPRYFHFVVFVSPVLSISGWSKSSTISSFFFFRFGFDQSR
>gi|145281011|gb|ES322235.1|ES322235_orf_1
MYIPNMSVMVFLTVSMVIGLATAEWREYTKVDLVCTGEKTQALCSTPITTGYSVILATPVDKTKGTNNCVNARTTH
KLCCEAETAPLNDVNQTPVNLSTETVGKKCTVWQSLE*
>gi|145281002|gb|ES322226.1|ES322226_orf_1
MENSTTASPLSPQSQELTTEQPSPADITVPSTNEAKVSKSPRQFKVSLAKRYFKSEPTQPAATNGLTDTLEGPFSS
VIGGIGYSPSAPRAKTLIKKRRTLSARVPTPVLKDLKLSGIISKILGRKTHMDEIVQGA*
>gi|116360627|gb|EG374433.1|EG374433_orf_7
MLLRRRVALFDCGDDPRCLMKDSSTAEMKGREKLFFPGRNGRIPGHMSGTVD*
>gi|116360557|gb|EG374363.1|EG374363_orf_1
MPGNSAITAGDPPHNSPAIRYHLPLLTSFFGLHLPAFNSTRSNISAKMNTTLYALLSLAVATLSVGAVGQTNQCQF
YASVGVKGYTCNERPDIICSEGCKSFVTMTGCVLTQYPKKPATTELCTVGYGRDTAAFKACLTSQGAFRCNGTSTG
QATCSGCVPRGDVTWAN*
>gi|116360679|gb|EG374485.1|EG374485_orf_3
MRIREPKTTALIFASGKMVVTGAKFEDDSRLAARKYARNCSETWASKQKFTEFKIQNIGWKLRRFAFPIWLGRFKP
TTKGHFFVV*
>gi|222428727|gb|GH737378.1|GH737378_orf_1
```

```
MVAETEYYDRLGVAPDVDEAALKKAYRKKALQLHPDKNPAGAEEFKAVSEAYDVLSTPEKRELYDQYGKKGLEGGG
GMGGMDPGDLFSQLFGGGGGMFGGRKPNGT
>gi|222429771|gb|GH738007.1|GH738007_orf_1
MNFLQLFVFMSGAITPVVMAAGTRAQAAGRGQQNQPKTTPGNFPCTGTLNAGWCVSAILPDSSLGARYFVKAKGAN
PNFSCNGAAQPNKACCKSNFAPDRNGVSLTLNVDFCKIL*
>gi|116360542|gb|EG374348.1|EG374348_orf_4
MGSQRISEQEYKATKSTESTINDGQGGREGRLNSWLKAAACCELASGHAYLRKLNPRYERKRELARDGKKGSSRPS
EGPISIRSFRVTE*
>gi|145280830|gb|ES322054.1|ES322054_orf_1
MFTSRLLLSSVLCVIAAVVTATTPAPLGPLNLCSDKDSVYKVTGVLQLNGTVPRVDGQGHPTGNPDGICECQPTGI
NCNTRPSAPIFDGPPVFCANPITNKCDAPAPQKLCSPAGSKYQVVGVIHPDGSVSSVDAKGTVSPRASGICTCTPN
GVPKCHLAPTNTVFSPLVFCANRQTNKC*
>gi|145280836|gb|ES322060.1|ES322060_orf_2
MNFLGSAILLVASIGHLVAGQQVFHCPKSAPYAHCGTNNYAAVPPTWDITNAAKNGNTYDCPGGDQITLCCHIGGE
PSFSSKADYDKWVKDHCI*
>gi|222428850|gb|GH737501.1|GH737501_orf_1
MGFFWFPLFCNLKLKHWWWSGRAREYERNENEKNGARSTMTSKEYRFNKSWRVYPGTWCPSKATRGFRKVPVRRWM
*
>gi|145280827|gb|ES322051.1|ES322051_orf_1
MLQSFRLIALVALIACREVISAEDKYFGCHRNVDAICATSAVRYTLKLTWAERLHRGKRDYVCRSDTNPICCNQGM
FDINATPNHFLMVVDTAINPCSIGGQ*
>gi|222428814|gb|GH737465.1|GH737465_orf_1
LDVLSCPATSVDVERAFSFGRDYVSSKRHRLSSESISRGMSVAFYSKNGLIKEGVLDRWKTGIQTGKKLNAKKKKK
KKKKKT
>gi|222429762|gb|GH737998.1|GH737998_orf_1
LETQKVDVLLLANKTAGWSCSVPKLTAQSCHSNKSLDWPVQAVLGHGCCGT
>gi|145281015|gb|ES322239.1|ES322239_orf_1
MQFVNSTVLLFVLLAGALNLVGVDAAGRVFPCRSPKPYALCGGRPDADYQLWFAPRVSGGHSCDSTNGIPYCCSIN
KRFSTTDPNAYDQYISAICANP*
>gi|145280942|gb|ES322166.1|ES322166_orf_1
MQWEKTPSSQRNQQDEIMSSFFGGLATGKQKTVIKPRKNLPEHTKQYQLKKYADATLGSGNLRSAVTLPEGEDLNE
WLAVNTLDFYNQINMLYGTVTEFCTPTECPVMSAGSRYEYHWHDGKEFKKATKVSAPEYVEYLMNWVQGFLDDEKI
FPSKIGQEFPKTFKSTIQSIVRRLFRVYAHLYNHHFAQICALGIEAHLNTSYRHFYFFIDEFELIKKDELIPLAEL
NTSIVNAELAAEDQKSHK*
>gi|222429873|gb|GH738055.1|GH738055_orf_4
MLTPSTMSISPPCGQFSPTVQKAGQVEQPKGIFITSRITRPYLKDFLEVNRTEFRFCRSARNV*
>gi|145280773|gb|ES321997.1|ES321997_orf_2
MNFFGSAILLVALTGPLVAGQIYFHCGKSAPYAHCGSNNSHAVPPTWDITYSYELGPGNIAHCPGGDQFKLCCHII
GEPGFQNKHDYDVYVKEHCS*
>gi|116360549|gb|EG374355.1|EG374355_orf_1
MPGNSAITAGDTHTILHSLQLVKTQRTHISKHILRLSTMFKAALPALVAVTLGMLSVVRAVDLTDAYRYYPDGDLL
HVDANAGSFKCPRNCPEFFRATRCTNNDVTGSKVTNETCSSTFGFNGAAHKTCGGFVNGKRHTYTCDHLDPASKFV
CSGCTATTS*
>gi|145280959|gb|ES322183.1|ES322183_orf_1
MNTTLYALLSLAVATLSVGAVGQTNQCQFYASVGVKGYTCNERPDIICSEGCKSFVTMTGCVLTQYPKKPATTELC
TVGYGRDTAAFKACLTSQGAFRCNGTSTGQATCSGCVPRGDVTWAN*
>gi|222428852|gb|GH737503.1|GH737503_orf_2
MNRQRHQTVLKLSCNHTTGCQTPNPPIKTNTPRIKSYYPYPRNWGQ
>gi|116360566|gb|EG374372.1|EG374372_orf_2
MGSQRISEQEYKATKSTEPTINDGQGGREGRLNSWLKAAACCELASGHAYVGTQQGMKEGGICPGLKER*
>gi|145281739|gb|ES322963.1|ES322963_orf_5
LLVICPDPRSKKKRRNNNISLQKKKKKKKKKNMSGRL
>gi|145281268|gb|ES322492.1|ES322492_orf_2
MLIEIETGIGTGMDETEIVHLFVTAAGTMSTETSIGTAMGIVIDPETPIEIAIDPETESEVKITRKATTETQIKTE
IGTAKGLVTETASAEI*
>gi|145280842|gb|ES322066.1|ES322066_orf_1
MHFSDFSHLLLVILLHLTVSCVTGEKKLFTCPHFGYCTNQELITIPVSYDIGPVYIPQPSITYTNFLTCKQSQLPI
GPAKNSCCDHAVPGVESRASNDPVNILYNDYTNKYKCHEVPYTN*
>gi|222428982|gb|GH737633.1|GH737633_orf_2
MHCHQSPKPFRKKYVMRYIYIGQSYSVTSNTIIFCYHLILYRARSDQKQKNLDKPVRACPPQKKKKKKKK
>gi|222429514|gb|GH737830.1|GH737830_orf_3
MSLCITPLPRFQFHTNISFVIFLFDQISRRKTRIIHQKSLMQ*
>gi|116360529|gb|EG374335.1|EG374335_orf_3
```

```
MLNALRQAQPKFWLLAMSITFPTTLDICCEVSCSGAELTGDAATSPLVASITCPLICTTKQPPQLPSVPARQRTSP
KSLVAPPVNLLPARKGVTTVILVFC*
>gi|145280884|gb|ES322108.1|ES322108_orf_2
KARTTTSNDDTPTTDQSHQPKHQEHTQDHPPSYLCHGLTSTDLSH*
>gi|145281349|gb|ES322573.1|ES322573_orf_2
VGTRDSAITAGGRDTAAFKACLTSQGAFRCNGTSTGQATCSGCVPRGDVTWAN*
>gi|222429621|gb|GH738416.1|GH738416_orf_4
LAGFQGFYFLLFKKKKKKKKKKKKKKKKRKKKKKKKKKKKHVGRLGPRE
>gi|222429972|gb|GH738154.1|GH738154_orf_1
MSQQDAKNGTQTDQKEKVDVSKLSQDKNMAICETATGGKTGLCDVSTCQVPQVTVCTQCTEVNKDSPDFAPAAGAT
PVEKMQCDSGYVLKTPENKQAGNICMTKTSAFTCAGECQGVLACNKCVTEDSAPKA*
>gi|222430021|gb|GH738203.1|GH738203_orf_2
MPPPCFFRDKPGDDVVWRSFNGAPWQSRGVPKGWMLG*
>gi|145280947|gb|ES322171.1|ES322171_orf_1
MALDVLSCPATSVDVDRAFSFGRDYVATKRHRLTACSLSRGMTVAFYSKNGLIKEGVLAKWKDGIQAEKKVMAKGQ
QNPRVIHLDDD*
>gi|145280839|gb|ES322063.1|ES322063_orf_1
MFKAALPALVAVTLGMLSVVRAVDLTDAYRYYPDGDLLHVDANAGSFKCPRNCAEFFRATGCTNNDVTGSKVTNET
CSSMFGANGAAHKTCGGFVNGKRHTYTCDHLDPASKFVCSGCTATTS*
>gi|145280758|gb|ES321982.1|ES321982_orf_2
MQFTTLMAVLATCSVVLTSPLSQQISAAAAVESAEAVESRWGGWAW*
>gi|222429797|gb|GH738033.1|GH738033_orf_4
MPACASVDSARRPIIFCAAHPYTSNADPQIYAIHACQSYVTNCSNSTV*
>gi|222428929|gb|GH737580.1|GH737580_orf_1
MDIVQLTLLVFLAGVCKSVISGRIQPPVMEAACCTTGDLDHADVYKPHNKNNDCYKTPDEVPYECPKGVLPPVLSI
KLREANARGCHKR*
>gi|145281307|gb|ES322531.1|ES322531_orf_2
VGTRDSAITAGTPCHPPSKETRQAQNATSFLLPRPRQTAPVHCGDGTTPPCHESSSQTDQ*
>gi|222430124|gb|GH737768.1|GH737768_orf_2
LRFHNTMLGSCSPSWHIRSCMRIQRRGSGRRRHSRTRSHFDPAVNRARSHIDPVSAARHPPDPPSYLIASAPSAAA
A
>gi|145280810|gb|ES322034.1|ES322034_orf_1
MRFANPTTLLVVLLAGALNLISVDGARIFPCPSAKPHGYCGEVKDNLYTLWHAFPEGGGNSCGMTSGIPYCCAMTN
AYSNPSPTYYDLIISEYCAQA*
>gi|222429390|gb|GH737306.1|GH737306_orf_3
MNQFFSPRKKKKKKKKKKKKKKRKKKKKKKKKKNLVGRLGPRKVF*
>gi|145280873|gb|ES322097.1|ES322097_orf_1
MLFSVLTVLLMIQGRSVIGNLFECPNPDRALALCSNPPDDDEDTTYVVKPYHIGAHYSCPPNLDAQTLNCCKTDYK
IGIRRPGTATRISTDTYSDVCSPGVDSPDPPTVDLTDAYRYYPDGNGYLYVDADAGSFHCPLTCESFKHAIGCTAS
DDVPGAEKTNETCSRMYGPDGATHKTCGNVVDGRTNSFSCDRTDHEINFACSGCISTTFK*
>gi|145281490|gb|ES322714.1|ES322714_orf_1
GNSAITAGDLNSSDYEPLSVIVKSINNFEGLVRRSIGLSVSSCFKSIDSDRLSKYLSFSASEHDQLIDWVKQFGWS
FSSDNRKVIIPNNSDNCPVTVVIRENTTIDDLQHFIAKSIVC*
>gi|222429870|gb|GH738052.1|GH738052_orf_1
MFTCDVARKKFQGGFHNTCCSKDLDLAKFTAPKAPFAKLDRATFDKFCTDTTPTP*
>gi|222429808|gb|GH738044.1|GH738044_orf_1
MFSYPICDKCNLLESVNHYLLTCKRYREQRQTLCNQLKALKLKGNDLTARYLLRNPKAAIPLANFIRNSRRFASYP
SYTTKFGPK*
>gi|145280820|gb|ES322044.1|ES322044_orf_1
LLIKTETHTSIRYSKLKMLCDLRLCIISLLSFSLLSLTEIPPERENSTTQSAPSSSKLSARQTSQEPPASPPHGSA
CKSIRVRKEWRTLSHDEQADYIRSVKSLARLPSKLLGSSYRRWDDFEYVHSQLRGRIHVRPLFLPWHRNLARIYEK
VLQDECNLKGTLPYWDWTLDYKNITQSPIWSSDTAIGFGSKGSFFGPGSDPANLDAGVVMDG
>gi|222429433|gb|GH738308.1|GH738308_orf_1
MQSLNFFMVFAVLLINTQFISVKSFKCPGLHGTPSQTHGYCTRSITDEERKAKKIGKEFTMWKEEIKTVDGKFSCD
KVDLNGSVATDSFCCDVAGRIGEVEKSKQAMWTNNCSKAS*
>gi|145281213|gb|ES322437.1|ES322437_orf_2
YPKKPATTELCTVGYGRDTAAFKACLTSQGAFRCNGTSTGQATCSGCVPRGDVTWAN*
>gi|116360574|gb|EG374380.1|EG374380_orf_6
MTSACLVGDLVCPVSMLIPSKFPLAKLELFHYHPSSLSLFLLFYLHIHFTSTCFFFYFLETPLF*
>gi|222429011|gb|GH737102.1|GH737102_orf_1
MQSFNFFIVFAVLLINTQFISVKSFKCPGLHGTPSQTHGYCTRSITDEERKAKKIGKEFTMWKEEIKTVDGKFSCD
KVDLNGSVATDSFCCDVAGRIGEVEKSKQAMWTNNCSKAS*
>gi|145281738|gb|ES322962.1|ES322962_orf_2
```

```
VGTRDSAITAGGCVLTHYPKKPATTELCTVGYGRDTAAFKACLTSQGAFRCNGTSTGQATCSGCVPRGDVTWAN*
>gi|145281799|gb|ES323023.1|ES323023_orf_5
MTSACLVGDLVCPVSMLIPSKFPLAKLELFHYHPSSLSLFLLSYLHIH
>gi|145281222|gb|ES322446.1|ES322446_orf_2
MLSREHDHDYYFSEVGMSSCRLMGLSPSVQLASFGTSSWLLLPSPQVRGVCVCVPGRNGRIPGTN
>gi|116360518|gb|EG374324.1|EG374324_orf_4
MNFLGSAILLVASIGHLVAGQQVFHCPKSAPYAHCGTNSYAAVPPTWDITNAAKNGNTYDCPGGDQITLGCHIGGE
PSFSSKADYDKWVKDHCI*
>gi|145280708|gb|ES321932.1|ES321932_orf_1
MFKAALPALVAVTLGMLSVVRAVDLTDAYRYYPDGDLLHVDANAGSFKCPRNCPEFFRATRCTNNDVTGSKVTNET
CSSTFGFNGAAHKTCGGFVNGKRHTYTCDHLDPASKFVCSGCTATTS*
>gi|222429767|gb|GH738003.1|GH738003_orf_2
MMSTIFLSCIDIFYTVRPFVFLQIHGFRIKYKFKKYIHLKYKYK*
>gi|145281766|gb|ES322990.1|ES322990_orf_1
MNFLGSAILFVALAGPLVAGQQYFRCGAAAPYGHCGSNNSHAVPPTWDITYIYHHTRSSPRFSPADHRGAATRRGG
IRPFSSTASSSHSQLWFPYHHVSRIFSQCI*
>gi|145281319|gb|ES322543.1|ES322543_orf_2
GRDTAAFKACLTSQGAFRCNGTSTGHATCSGCVPRGDVTWAN*
```

Appendix 2. SignalP-4.0 results of 95 ORFs.

name	Cmax	pos	Ymax	pos	Smax	pos	Smean	D	?	Dmaxcut	Networks-used
gi_145280833_gb_ES322057.1_ES322057_orf_1	0.354	23	0.562	23	0.968	12	0.892	0.74	Y	0.45	SignalP-noTM
gi_145281363_gb_ES322587.1_ES322587_orf_1	0.865	22	0.874	22	0.974	10	0.88	0.877	Y	0.45	SignalP-noTM
gi_145280711_gb_ES321935.1_ES321935_orf_2	0.288	21	0.495	21	0.918	11	0.848	0.686	Y	0.45	SignalP-noTM
gi_145280905_gb_ES322129.1_ES322129_orf_1	0.891	24	0.896	24	0.947	4	0.901	0.899	Y	0.45	SignalP-noTM
gi_145280791_gb_ES322015.1_ES322015_orf_1	0.545	21	0.705	21	0.97	11	0.908	0.815	Y	0.45	SignalP-noTM
gi_145280743_gb_ES321967.1_ES321967_orf_1	0.561	26	0.697	26	0.956	12	0.864	0.787	Y	0.45	SignalP-noTM
gi_222428962_gb_GH737613.1_GH737613_orf_2	0.23	17	0.354	17	0.747	11	0.533	0.451	Y	0.45	SignalP-noTM
gi_145280919_gb_ES322143.1_ES322143_orf_1	0.243	31	0.413	20	0.91	14	0.745	0.592	Y	0.45	SignalP-noTM
gi_116360642_gb_EG374448.1_EG374448_orf_1	0.78	22	0.859	22	0.979	5	0.945	0.906	Y	0.45	SignalP-noTM
gi_116360647_gb_EG374453.1_EG374453_orf_2	0.748	21	0.825	21	0.939	3	0.909	0.87	Y	0.45	SignalP-noTM
gi_222430111_gb_GH737755.1_GH737755_orf_1	0.237	27	0.453	27	0.951	15	0.871	0.679	Y	0.45	SignalP-noTM
gi_145281011_gb_ES322235.1_ES322235_orf_1	0.727	24	0.796	24	0.93	15	0.872	0.837	Y	0.45	SignalP-noTM
gi_222429771_gb_GH738007.1_GH738007_orf_1	0.75	21	0.802	21	0.914	1	0.852	0.829	Y	0.45	SignalP-noTM
gi_145280830_gb_ES322054.1_ES322054_orf_1	0.78	22	0.859	22	0.979	5	0.945	0.906	Y	0.45	SignalP-noTM
gi_145280836_gb_ES322060.1_ES322060_orf_2	0.672	21	0.774	21	0.927	12	0.89	0.837	Y	0.45	SignalP-noTM
gi_222428850_gb_GH737501.1_GH737501_orf_1	0.374	24	0.423	24	0.669	17	0.487	0.458	Y	0.45	SignalP-noTM
gi_145280827_gb_ES322051.1_ES322051_orf_1	0.569	23	0.676	23	0.93	3	0.802	0.744	Y	0.45	SignalP-noTM
gi_145281015_gb_ES322239.1_ES322239_orf_1	0.596	26	0.747	26	0.988	14	0.936	0.849	Y	0.45	SignalP-noTM
gi_145280773_gb_ES321997.1_ES321997_orf_2	0.526	20	0.693	20	0.955	3	0.912	0.811	Y	0.45	SignalP-noTM
gi_145280959_gb_ES322183.1_ES322183_orf_1	0.354	23	0.562	23	0.968	12	0.892	0.74	Y	0.45	SignalP-noTM
gi_145280842_gb_ES322066.1_ES322066_orf_1	0.692	25	0.79	25	0.974	14	0.903	0.851	Y	0.45	SignalP-noTM
gi_116360529_gb_EG374335.1_EG374335_orf_3	0.16	27	0.307	27	0.724	13	0.576	0.453	Y	0.45	SignalP-noTM
gi_145280839_gb_ES322063.1_ES322063_orf_1	0.777	23	0.835	23	0.944	6	0.897	0.869	Y	0.45	SignalP-noTM
gi_145280758_gb_ES321982.1_ES321982_orf_2	0.33	19	0.534	19	0.918	1	0.854	0.707	Y	0.45	SignalP-noTM
gi_222428929_gb_GH737580.1_GH737580_orf_1	0.505	19	0.65	19	0.88	12	0.83	0.748	Y	0.45	SignalP-noTM
gi_145280810_gb_ES322034.1_ES322034_orf_1	0.703	26	0.794	26	0.966	15	0.897	0.85	Y	0.45	SignalP-noTM
gi_145280873_gb_ES322097.1_ES322097_orf_1	0.259	20	0.46	20	0.901	9	0.809	0.648	Y	0.45	SignalP-noTM
gi_222429433_gb_GH738308.1_GH738308_orf_1	0.445	25	0.595	25	0.911	4	0.801	0.706	Y	0.45	SignalP-noTM
gi_222429011_gb_GH737102.1_GH737102_orf_1	0.438	25	0.593	25	0.918	4	0.808	0.709	Y	0.45	SignalP-noTM
gi_116360518_gb_EG374324.1_EG374324_orf_4	0.671	21	0.773	21	0.925	12	0.888	0.835	Y	0.45	SignalP-noTM
gi_145280708_gb_ES321932.1_ES321932_orf_1	0.777	23	0.834	23	0.942	6	0.895	0.867	Y	0.45	SignalP-noTM
gi_145281766_gb_ES322990.1_ES322990_orf_1	0.468	21	0.648	21	0.949	2	0.892	0.78	Y	0.45	SignalP-noTM
gi_145281423_gb_ES322647.1_ES322647_orf_4	0.11	23	0.168	2	0.288	2	0.275	0.225	N	0.45	SignalP-noTM
gi_116360588_gb_EG374394.1_EG374394_orf_1	0.18	51	0.198	51	0.32	48	0.156	0.181	N	0.5	SignalP-TM
gi_145281060_gb_ES322284.1_ES322284_orf_4	0.108	5	0.118	5	0.139	3	0.122	0.12	N	0.45	SignalP-noTM
gi_222429997_gb_GH738179.1_GH738179_orf_1	0.184	24	0.23	24	0.411	13	0.291	0.263	N	0.45	SignalP-noTM
gi_145281327_gb_ES322551.1_ES322551_orf_2	0.108	32	0.149	4	0.215	1	0.195	0.174	N	0.45	SignalP-noTM
gi_145280711_gb_ES321935.1_ES321935_orf_1	0.117	54	0.136	3	0.186	49	0.157	0.147	N	0.45	SignalP-noTM
gi_222429299_gb_GH737076.1_GH737076_orf_2	0.111	28	0.12	4	0.148	5	0.12	0.12	N	0.45	SignalP-noTM
gi_145281411_gb_ES322635.1_ES322635_orf_3	0.177	46	0.264	8	0.833	5	0.745	0.456	N	0.5	SignalP-TM
gi_145281737_gb_ES322961.1_ES322961_orf_4	0.165	20	0.28	20	0.783	14	0.464	0.379	N	0.45	SignalP-noTM
gi_145280964_gb_ES322188.1_ES322188_orf_1	0.185	35	0.208	15	0.536	10	0.417	0.291	N	0.5	SignalP-TM
gi_145281066_gb_ES322290.1_ES322290_orf_4	0.172	35	0.15	35	0.189	3	0.136	0.142	N	0.45	SignalP-noTM
gi_116360702_gb_EG374508.1_EG374508_orf_3	0.11	49	0.108	5	0.117	3	0.113	0.111	N	0.45	SignalP-noTM
gi_145280834_gb_ES322058.1_ES322058_orf_1	0.115	28	0.12	2	0.15	17	0.137	0.129	N	0.45	SignalP-noTM
gi_222428836_gb_GH737487.1_GH737487_orf_1	0.11	19	0.105	29	0.134	11	0.099	0.102	N	0.45	SignalP-noTM
gi_145280717_gb_ES321941.1_ES321941_orf_2	0.119	54	0.115	54	0.149	47	0.104	0.109	N	0.45	SignalP-noTM
gi_222429313_gb_GH737730.1_GH737730_orf_4	0.117	19	0.146	25	0.239	21	0.161	0.152	N	0.5	SignalP-TM
gi_116360590_gb_EG374396.1_EG374396_orf_7	0.121	21	0.157	11	0.258	3	0.193	0.176	N	0.45	SignalP-noTM
gi_116360691_gb_EG374497.1_EG374497_orf_2	0.111	14	0.107	70	0.117	23	0.101	0.104	N	0.45	SignalP-noTM
gi_116360526_gb_EG374332.1_EG374332_orf_1	0.11	70	0.109	70	0.137	23	0.105	0.107	N	0.45	SignalP-noTM
gi_222428890_gb_GH737541.1_GH737541_orf_1	0.124	42	0.134	1	0.174	1	0	0.061	N	0.45	SignalP-noTM
gi_145281618_gb_ES322842.1_ES322842_orf_2	0.11	21	0.116	21	0.202	18	0.123	0.12	N	0.45	SignalP-noTM
gi_116360662_gb_EG374468.1_EG374468_orf_1	0.109	42	0.157	35	0.292	24	0.187	0.169	N	0.5	SignalP-TM
gi_222429370_gb_GH737286.1_GH737286_orf_1	0.108	58	0.111	24	0.19	56	0.104	0.108	N	0.5	SignalP-TM
gi_222429123_gb_GH737194.1_GH737194_orf_4	0.118	20	0.108	37	0.119	19	0.102	0.105	N	0.45	SignalP-noTM
gi_222429861_gb_GH738496.1_GH738496_orf_1	0.132	50	0.129	66	0.184	54	0.12	0.125	N	0.5	SignalP-TM
gi_145281002_gb_ES322226.1_ES322226_orf_1	0.108	67	0.104	37	0.112	35	0.097	0.1	N	0.45	SignalP-noTM
gi_116360627_gb_EG374433.1_EG374433_orf_7	0.111	37	0.125	5	0.15	1	0.137	0.131	N	0.45	SignalP-noTM
gi_116360557_gb_EG374363.1_EG374363_orf_1	0.113	70	0.161	34	0.307	23	0.195	0.174	N	0.5	SignalP-TM
gi_116360679_gb_EG374485.1_EG374485_orf_3	0.2	23	0.217	23	0.52	21	0.236	0.227	N	0.45	SignalP-noTM
gi_222428727_gb_GH737378.1_GH737378_orf_1	0.11	49	0.107	49	0.113	55	0.098	0.102	N	0.45	SignalP-noTM
gi_116360542_gb_EG374348.1_EG374348_orf_4	0.112	50	0.11	50	0.162	47	0.102	0.106	N	0.45	SignalP-noTM
gi_222428814_gb_GH737465.1_GH737465_orf_1	0.113	17	0.108	2	0.113	26	0.113	0.11	N	0.45	SignalP-noTM
gi_222429762_gb_GH737998.1_GH737998_orf_1	0.134	18	0.148	18	0.238	16	0.164	0.157	N	0.45	SignalP-noTM
gi_145280942_gb_ES322166.1_ES322166_orf_1	0.112	29	0.116	15	0.164	6	0.13	0.124	N	0.45	SignalP-noTM
gi_222429873_gb_GH738055.1_GH738055_orf_4	0.108	38	0.132	5	0.172	1	0.138	0.136	N	0.45	SignalP-noTM
gi_116360549_gb_EG374355.1_EG374355_orf_1	0.236	61	0.16	3	0.25	1	0.241	0.193	N	0.5	SignalP-TM
gi_222428852_gb_GH737503.1_GH737503_orf_2	0.163	21	0.148	21	0.17	17	0.132	0.139	N	0.45	SignalP-noTM
gi_116360566_gb_EG374372.1_EG374372_orf_2	0.12	50	0.121	50	0.209	47	0.11	0.115	N	0.45	SignalP-noTM
gi_145281739_gb_ES322963.1_ES322963_orf_5	0.108	20	0.11	3	0.121	19	0.111	0.111	N	0.45	SignalP-noTM
gi_145281268_gb_ES322492.1_ES322492_orf_2	0.143	31	0.12	31	0.125	26	0.103	0.111	N	0.45	SignalP-noTM
gi_222428982_gb_GH737633.1_GH737633_orf_2	0.131	46	0.116	57	0.197	43	0.097	0.108	N	0.5	SignalP-TM
gi_222429514_gb_GH737830.1_GH737830_orf_3	0.112	33	0.136	8	0.324	4	0.221	0.182	N	0.45	SignalP-noTM
gi_145280884_gb_ES322108.1_ES322108_orf_2	0.111	39	0.107	39	0.116	20	0.093	0.099	N	0.45	SignalP-noTM
gi_145281349_gb_ES322573.1_ES322573_orf_2	0.114	33	0.119	5	0.176	3	0.156	0.139	N	0.45	SignalP-noTM
gi_222429621_gb_GH738416.1_GH738416_orf_4	0.107	49	0.108	31	0.135	27	0.105	0.106	N	0.45	SignalP-noTM
gi_222429972_gb_GH738154.1_GH738154_orf_1	0.11	57	0.103	57	0.111	55	0.095	0.099	N	0.45	SignalP-noTM
gi_222430021_gb_GH738203.1_GH738203_orf_2	0.108	32	0.112	5	0.12	1	0.112	0.112	N	0.45	SignalP-noTM
gi_145280947_gb_ES322171.1_ES322171_orf_1	0.139	19	0.128	19	0.131	3	0.117	0.122	N	0.45	SignalP-noTM
gi_222429797_gb_GH738033.1_GH738033_orf_4	0.165	27	0.136	27	0.182	2	0.121	0.128	N	0.45	SignalP-noTM
gi_145281307_gb_ES322531.1_ES322531_orf_2	0.108	35	0.108	29	0.136	12	0.106	0.107	N	0.45	SignalP-noTM
gi_222430124_gb_GH737768.1_GH737768_orf_2	0.132	23	0.128	23	0.168	4	0.125	0.127	N	0.45	SignalP-noTM
gi_222429390_gb_GH737306.1_GH737306_orf_3	0.107	40	0.105	25	0.122	21	0.104	0.105	N	0.45	SignalP-noTM
gi_145281490_gb_ES322714.1_ES322714_orf_1	0.118	50	0.105	50	0.108	56	0.091	0.097	N	0.45	SignalP-noTM
gi_222429870_gb_GH738052.1_GH738052_orf_1	0.11	42	0.105	13	0.121	9	0.108	0.107	N	0.45	SignalP-noTM
gi_222429808_gb_GH738044.1_GH738044_orf_1	0.11	58	0.121	5	0.143	1	0.125	0.123	N	0.45	SignalP-noTM

gi_145280820_gb_ES322044.1_ES322044_orf_1	0.122	39	0.201	39	0.442	31	0.25	0.221	N	0.5	SignalP-TM
gi_145281213_gb_ES322437.1_ES322437_orf_2	0.121	22	0.116	22	0.137	6	0.112	0.114	N	0.45	SignalP-noTM
gi_116360574_gb_EG374380.1_EG374380_orf_6	0.136	53	0.177	3	0.306	1	0.264	0.212	N	0.5	SignalP-TM
gi_145281738_gb_ES322962.1_ES322962_orf_2	0.114	26	0.113	5	0.136	24	0.113	0.113	N	0.45	SignalP-noTM
gi_145281799_gb_ES323023.1_ES323023_orf_5	0.125	28	0.185	3	0.328	1	0.295	0.244	N	0.45	SignalP-noTM
gi_145281222_gb_ES322446.1_ES322446_orf_2	0.318	50	0.232	50	0.244	31	0.137	0.181	N	0.45	SignalP-noTM
gi_222429767_gb_GH738003.1_GH738003_orf_2	0.165	19	0.228	19	0.528	4	0.289	0.261	N	0.45	SignalP-noTM
gi_145281319_gb_ES322543.1_ES322543_orf_2	0.122	22	0.133	3	0.169	4	0.165	0.15	N	0.45	SignalP-noTM

Appendix 3. Table of 95 predicted ORFs of Pst with sequence annotation.

Seq. Name	SP Region	Motif	SignalP	Seq. Description	Seq. Length	Methionine
gi|116360518|gb|EG374324.1|EG374324_orf_4	YES	Y/F/WxC	YES	---NA---	95	YES
gi|116360526|gb|EG374332.1|EG374332_orf_1	NO	NO	NO	hypothetical protein PGTG_00898 [Puccinia graminis f. sp. tritici CRL 75-36-700-3]	109	YES
gi|116360529|gb|EG374335.1|EG374335_orf_3	YES	Y/F/WxC	YES	---NA---	102	YES
gi|116360542|gb|EG374348.1|EG374348_orf_4	NO	NO	NO	---NA---	90	YES
gi|116360549|gb|EG374355.1|EG374355_orf_1	YES	Y/F/WxC	NO	hypothetical protein PGTG_06171 [Puccinia graminis f. sp. tritici CRL 75-36-700-3]	162	YES
gi|116360557|gb|EG374363.1|EG374363_orf_1	YES	Y/F/WxC	NO	secreted protein	170	YES
gi|116360566|gb|EG374372.1|EG374372_orf_2	NO	NO	NO	---NA---	70	YES
gi|116360574|gb|EG374380.1|EG374380_orf_6	YES	Y/F/WxC	NO	---NA---	65	YES
gi|116360588|gb|EG374394.1|EG374394_orf_1	YES	Y/F/WxC	NO	---NA---	123	YES
gi|116360590|gb|EG374396.1|EG374396_orf_7	YES	NO	NO	---NA---	79	YES
gi|116360627|gb|EG374433.1|EG374433_orf_7	NO	NO	NO	---NA---	53	YES
gi|116360642|gb|EG374448.1|EG374448_orf_1	YES	Y/F/WxC	YES	---NA---	174	YES
gi|116360647|gb|EG374453.1|EG374453_orf_2	YES	Y/F/WxC	YES	---NA---	95	YES
gi|116360662|gb|EG374468.1|EG374468_orf_1	YES	Y/F/WxC	NO	secreted protein	171	YES
gi|116360679|gb|EG374485.1|EG374485_orf_3	YES	NO	NO	tata-box-binding protein	86	YES
gi|116360691|gb|EG374497.1|EG374497_orf_2	NO	NO	NO	hypothetical protein PGTG_00898 [Puccinia graminis f. sp. tritici CRL 75-36-700-3]	93	YES
gi|116360702|gb|EG374508.1|EG374508_orf_3	NO	NO	NO	---NA---	110	YES
gi|145280708|gb|ES321932.1|ES321932_orf_1	YES	Y/F/WxC	YES	hypothetical protein PGTG_06171 [Puccinia graminis f. sp. tritici CRL 75-36-700-3]	124	YES
gi|145280711|gb|ES321935.1|ES321935_orf_1	YES	Y/F/WxC	NO	---NA---	130	NO
gi|145280711|gb|ES321935.1|ES321935_orf_2	YES	Y/F/WxC	YES	secreted protein	123	YES
gi|145280717|gb|ES321941.1|ES321941_orf_2	NO	NO	NO	---NA---	69	YES
gi|145280743|gb|ES321967.1|ES321967_orf_1	YES	Y/F/WxC	YES	---NA---	94	YES
gi|145280758|gb|ES321982.1|ES321982_orf_2	YES	Y/F/WxC	YES	---NA---	47	YES
gi|145280773|gb|ES321997.1|ES321997_orf_2	YES	Y/F/WxC	YES	---NA---	97	YES
gi|145280791|gb|ES322015.1|ES322015_orf_1	YES	Y/F/WxC	YES	---NA---	104	YES
gi|145280810|gb|ES322034.1|ES322034_orf_1	YES	Y/F/WxC	YES	---NA---	98	YES
gi|145280820|gb|ES322044.1|ES322044_orf_1	YES	NO	NO	di-copper centre-containing protein	214	NO
gi|145280827|gb|ES322051.1|ES322051_orf_1	YES	Y/F/WxC	YES	hypothetical protein PGTG_17018 [Puccinia graminis f. sp. tritici CRL 75-36-700-3]	103	YES
gi|145280830|gb|ES322054.1|ES322054_orf_1	YES	Y/F/WxC	YES	---NA---	181	YES
gi|145280833|gb|ES322057.1|ES322057_orf_1	YES	Y/F/WxC	YES	secreted protein	123	YES
gi|145280834|gb|ES322058.1|ES322058_orf_1	YES	NO	NO	hypothetical protein PGTG_17073 [Puccinia graminis f. sp. tritici CRL 75-36-700-3]	79	YES
gi|145280836|gb|ES322060.1|ES322060_orf_2	YES	Y/F/WxC	YES	---NA---	95	YES
gi|145280839|gb|ES322063.1|ES322063_orf_1	YES	Y/F/WxC	YES	hypothetical protein PGTG_06171 [Puccinia graminis f. sp. tritici CRL 75-36-700-3]	124	YES
gi|145280842|gb|ES322066.1|ES322066_orf_1	YES	Y/F/WxC	YES	hypothetical protein PGTG_04524 [Puccinia graminis f. sp. tritici CRL 75-36-700-3]	121	YES
gi|145280873|gb|ES322097.1|ES322097_orf_1	YES	Y/F/WxC	YES	hypothetical protein PGTG_06171 [Puccinia graminis f. sp. tritici CRL 75-36-700-3]	213	YES
gi|145280884|gb|ES322108.1|ES322108_orf_2	NO	NO	NO	---NA---	46	NO
gi|145280905|gb|ES322129.1|ES322129_orf_1	YES	Y/F/WxC	YES	hypothetical protein PGTG_06171 [Puccinia graminis f. sp. tritici CRL 75-36-700-3]	122	YES
gi|145280919|gb|ES322143.1|ES322143_orf_1	YES	Y/F/WxC	YES	hypothetical protein PGTG_06171 [Puccinia graminis f. sp. tritici CRL 75-36-700-3]	215	YES
gi|145280942|gb|ES322166.1|ES322166_orf_1	YES	NO	NO	mps1 binder-like protein	247	YES
gi|145280947|gb|ES322171.1|ES322171_orf_1	NO	Y/F/WxC	NO	hypothetical protein PGTG_16568 [Puccinia graminis f. sp. tritici CRL 75-36-700-3]	88	YES
gi|145280959|gb|ES322183.1|ES322183_orf_1	YES	Y/F/WxC	YES	secreted protein	123	YES
gi|145280964|gb|ES322188.1|ES322188_orf_1	NO	NO	NO	calcofluor white hypersensitive protein	90	YES

gi|145281002|gb|ES322226.1|ES322226_orf_1	YES	NO	NO	hypothetical protein PGTG_07786 [Puccinia graminis f. sp. tritici CRL 75-36-700-3]	136	YES
gi|145281011|gb|ES322235.1|ES322235_orf_1	YES	Y/F/WxC	YES	hypothetical protein PGTG_01775 [Puccinia graminis f. sp. tritici CRL 75-36-700-3]	114	YES
gi|145281015|gb|ES322239.1|ES322239_orf_1	YES	Y/F/WxC	YES	---NA---	99	YES
gi|145281060|gb|ES322284.1|ES322284_orf_4	NO	NO	NO	---NA---	34	YES
gi|145281066|gb|ES322290.1|ES322290_orf_4	YES	Y/F/WxC	NO	---NA---	48	YES
gi|145281213|gb|ES322437.1|ES322437_orf_2	YES	Y/F/WxC	NO	secreted protein	58	NO
gi|145281222|gb|ES322446.1|ES322446_orf_2	YES	NO	NO	---NA---	65	YES
gi|145281268|gb|ES322492.1|ES322492_orf_2	NO	NO	NO	hypothetical protein PGTG_01800 [Puccinia graminis f. sp. tritici CRL 75-36-700-3]	93	YES
gi|145281307|gb|ES322531.1|ES322531_orf_2	NO	NO	NO	---NA---	61	NO
gi|145281319|gb|ES322543.1|ES322543_orf_2	YES	NO	NO	hypothetical protein PGTG_20495 [Puccinia graminis f. sp. tritici CRL 75-36-700-3]	43	NO
gi|145281327|gb|ES322551.1|ES322551_orf_2	YES	Y/F/WxC	NO	hypothetical protein PGTG_20495 [Puccinia graminis f. sp. tritici CRL 75-36-700-3]	53	NO
gi|145281349|gb|ES322573.1|ES322573_orf_2	YES	Y/F/WxC	NO	hypothetical protein PGTG_20495 [Puccinia graminis f. sp. tritici CRL 75-36-700-3]	54	NO
gi|145281363|gb|ES322587.1|ES322587_orf_1	YES	Y/F/WxC	YES	hypothetical protein PGTG_17018 [Puccinia graminis f. sp. tritici CRL 75-36-700-3]	104	YES
gi|145281411|gb|ES322635.1|ES322635_orf_3	YES	NO	NO	---NA---	58	NO
gi|145281423|gb|ES322647.1|ES322647_orf_4	NO	NO	NO	---NA---	44	YES
gi|145281490|gb|ES322714.1|ES322714_orf_1	NO	NO	NO	arm repeat-containing protein	119	NO
gi|145281618|gb|ES322842.1|ES322842_orf_2	NO	NO	NO	---NA---	42	NO
gi|145281737|gb|ES322961.1|ES322961_orf_4	NO	NO	NO	---NA---	67	YES
gi|145281738|gb|ES322962.1|ES322962_orf_2	YES	Y/F/WxC	NO	secreted protein	75	NO
gi|145281739|gb|ES322963.1|ES322963_orf_5	NO	NO	NO	---NA---	37	NO
gi|145281766|gb|ES322990.1|ES322990_orf_1	YES	Y/F/WxC	YES	---NA---	107	YES
gi|145281799|gb|ES323023.1|ES323023_orf_5	YES	NO	NO	---NA---	48	YES
gi|222428727|gb|GH737378.1|GH737378_orf_1	NO	NO	NO	mitochondrial protein import protein mas5	106	YES
gi|222428814|gb|GH737465.1|GH737465_orf_1	NO	Y/F/WxC	NO	hypothetical protein PGTG_05550 [Puccinia graminis f. sp. tritici CRL 75-36-700-3]	82	NO
gi|222428836|gb|GH737487.1|GH737487_orf_1	YES	Y/F/WxC	NO	---NA---	122	NO
gi|222428850|gb|GH737501.1|GH737501_orf_1	NO	NO	YES	---NA---	77	YES
gi|222428852|gb|GH737503.1|GH737503_orf_2	NO	NO	NO	---NA---	46	YES
gi|222428890|gb|GH737541.1|GH737541_orf_1	YES	Y/F/WxC	NO	---NA---	124	YES
gi|222428929|gb|GH737580.1|GH737580_orf_1	YES	Y/F/WxC	YES	hypothetical protein PGTG_16021 [Puccinia graminis f. sp. tritici CRL 75-36-700-3]	90	YES
gi|222428962|gb|GH737613.1|GH737613_orf_2	NO	Y/F/WxC	YES	---NA---	78	NO
gi|222428982|gb|GH737633.1|GH737633_orf_2	YES	NO	NO	---NA---	70	YES
gi|222429011|gb|GH737102.1|GH737102_orf_1*	YES	Y/F/WxC	YES	hypothetical protein PGTG_11199 [Puccinia graminis f. sp. tritici CRL 75-36-700-3]	117	YES
gi|222429123|gb|GH737194.1|GH737194_orf_4	NO	NO	NO	---NA---	81	NO
gi|222429299|gb|GH737076.1|GH737076_orf_2	NO	NO	NO	hypothetical protein PGTG_17994 [Puccinia graminis f. sp. tritici CRL 75-36-700-3]	72	YES
gi|222429313|gb|GH737730.1|GH737730_orf_4	YES	Y/F/WxC	NO	serine acetyltransferase	59	YES
gi|222429370|gb|GH737286.1|GH737286_orf_1	YES	NO	NO	---NA---	60	YES
gi|222429390|gb|GH737306.1|GH737306_orf_3	NO	NO	NO	---NA---	46	YES
gi|222429433|gb|GH738308.1|GH738308_orf_1	YES	Y/F/WxC	YES	hypothetical protein PGTG_11199 [Puccinia graminis f. sp. tritici CRL 75-36-700-3]	117	YES
gi|222429514|gb|GH737830.1|GH737830_orf_3	YES	NO	NO	---NA---	43	YES
gi|222429621|gb|GH738416.1|GH738416_orf_4	NO	NO	NO	---NA---	49	NO
gi|222429762|gb|GH737998.1|GH737998_orf_1	NO	NO	NO	---NA---	51	NO
gi|222429767|gb|GH738003.1|GH738003_orf_2	YES	Y/F/WxC	NO	---NA---	45	YES
gi|222429771|gb|GH738007.1|GH738007_orf_1*	YES	Y/F/WxC	YES	---NA---	116	YES
gi|222429797|gb|GH738033.1|GH738033_orf_4	NO	NO	NO	---NA---	49	YES
gi|222429808|gb|GH738044.1|GH738044_orf_1	NO	NO	NO	hypothetical protein PGTG_15276 [Puccinia graminis f. sp. tritici CRL 75-36-700-3]	86	YES
gi|222429861|gb|GH738496.1|GH738496_orf_1	YES	Y/F/WxC	NO	---NA---	73	NO
gi|222429870|gb|GH738052.1|GH738052_orf_1	NO	NO	NO	---NA---	56	YES

gi|222429873|gb|GH738055.1|GH738055_orf_4	YES	NO	NO	---NA---	64	YES
gi|222429972|gb|GH738154.1|GH738154_orf_1	YES	Y/F/WxC	NO	secreted protein	133	YES
gi|222429997|gb|GH738179.1|GH738179_orf_1	NO	NO	NO	---NA---	101	NO
gi|222430021|gb|GH738203.1|GH738203_orf_2	NO	NO	NO	---NA---	38	YES
gi|222430111|gb|GH737755.1|GH737755_orf_1	YES	Y/F/WxC	YES	---NA---	124	YES
gi|222430124|gb|GH737768.1|GH737768_orf_2	NO	NO	NO	---NA---	77	NO

`*` sequences are also predicted by (Yin et al., 2009) to be secreted.

Bold named sequences are our most likely predictions which contain motif region.

First column is the accession number of predicted ORF and EST.

Second column is SP region similarity according to GLAM2.

Third column is motif region availability according to GLAM2.

Fourth column is annotation of predicted ORF.

Fifth column is the ORF predicted amino acid sequence length.

The last column shows if predicted ORF starts with a Methionine amino acid.