Mold allergomics: comparative and machine learning approaches

Ha X. Dang

Dissertation submitted to the faculty of the Virginia Polytechnic Institute and State University in partial fulfillment of the requirements for the degree of

Doctor of Philosophy
In
Genetics, Bioinformatics and Computational Biology

Christopher B. Lawrence
Lenwood S. Heath
T. M. Murali
Brett M. Tyler

May 12, 2014
Blacksburg, Virginia

Keywords: Fungal genomics, comparative genomics, allergy, allergen prediction

Copyright © 2014, Ha X. Dang

Mold allergomics: comparative and machine learning approaches

Ha X. Dang

ABSTRACT

Fungi are one of the major organisms that cause allergic disease in human. A number of proteins from fungi have been found to be allergenic or possess immunostimulatory properties. Identifying and characterizing allergens from fungal genomes will help facilitate our understanding of the mechanism underlying host-pathogen interactions in allergic diseases. Currently, there is a lack of tools that allow us to rapidly and accurately predict allergens from whole genomes. In the context of whole genome annotation, allergens are rare compared to non-allergens and thus the data is considered highly skewed. In order to achieve a confident set of predicted allergens from a genome, false positive rates must be lowered. Current allergen prediction tools often produce many false positives when applied to large-scale data set such as whole genomes, and thus lower the precision. Moreover, the most accurate tools are relatively slow because they use sequence alignment to construct feature vectors for allergen classifiers. This dissertation presents computational approaches in characterizing the allergen repertoire in fungal genomes as part of the whole genome studies of *Alternaria*, an important allergenic/opportunistic human pathogenic fungus and necrotrophic plant parasite. In these studies, the genomes of multiple *Alternaria* species were characterized for the first time. Functional elements (e.g. genes, proteins) were first identified and annotated from these genomes using computational tools. Protein annotation and comparative genomics approaches revealed the link between *Alternaria* genotypes and its prolific saprophytic lifestyle that provides at least a partial explanation for the development of pathological relationships between *Alternaria* and humans. A machine learning based tool (Allerdictor) was developed to address the neglected problem of allergen prediction in highly skewed large-scale data sets. Allerdictor exhibited high precision over high recall at fast speed and thus it is a more practical tool for large-scale allergen annotation compared with existing tools. Allerdictor was then used together with a comparative genomics approach to survey the allergen repertoire of known allergenic fungi. We predicted a number of mold allergens that have not been experimentally characterized. These predicted allergens are potential candidates for further experimental and clinical validation. Our approaches will not only facilitate the study of allergens in the increasing number of sequenced fungal genomes but also will be useful for allergen annotation in other species and rapid prescreening of synthesized sequences for potential allergens.

Dedication

To my parents

Acknowledgements

I would like to express my sincere thanks to my Ph.D. committee members: Drs. Christopher Lawrence, Lenwood Heath, T. M. Murali and Brett Tyler for their constant mentorship, patience and encouragement. This dissertation would not have been possible without their guidance and support.

I would like to thank Dr. Thanh Nguyen, for introducing me to the field of bioinformatics and computational biology.

I would like to thank Dr. David Bevan and Ms. Dennie Munson for their tremendous help and support during my time in the Genetics, Bioinformatics, and Computational Biology (GBCB) program at Virginia Tech.

I would like to thank my collaborators, Dr. Bernard Henrissat, Dr. Ronald de Vries, Dr. Jason Stajich, Dr. Braham Dhillon, Dr. Tobin Peever, and Dr. Barry Pryor. The success of the *Alternaria* genomes project would not have been possible without their collaboration.

I would like to thank the past and current members of the Lawrence lab (Dr. Sang-Wook Park, Dr. Kwang-Huang Kim, Dr. Mihaela Babiceanu, Dr. Mauricio La Rota, and Brad Howard) who have been my excellent teachers of biology. They made my life as a computational scientist in a biology lab very enjoyable. I thank the Virginia Bioinformatics Institute Core Computational Facility and IT support team for helping with high performance computing systems.

Lastly, I thank my wife, Trang Vu, for her love and sacrifice. She has quit her dream job at a bank in Vietnam to join and support me during my last years at Virginia Tech.

I gratefully acknowledge the financial support provided by the Vietnam Education Foundation (two-year Ph.D. fellowship), the Graduate School at Virginia Tech, the Genetics, Bioinformatics and Computational Biology (GBCB) Graduate Program, the National Science Foundation (Grant # NSF DEB-0918298), the National Institute of Health (Grant # 1R21AI094071-01), and the National Institute of Food and Agriculture (Grant # NIFA 2004-35600-15030).

Contents

List of Figures

List of Tables

Chapter 1

Introduction

1.1 Asthma, allergy and allergen

Allergy or type I hypersensitivity is the disorder of the immune system when it overreacts against innocuous foreign substances that enter the body. It is estimated that 20-30% of the world's population suffers from IgE-mediated allergic diseases including allergic asthma, allergic rhinitis, conjunctivitis, chronic rhinosinusitis (CRS) and atopic dermatitis. In allergic or atopic asthmatics, life threatening asthma exacerbations can be triggered by allergic reactions. Asthma alone is estimated to affect 250-300 million people worldwide according to the Global Initiative for Asthma (GINA) and the World Health Organization (WHO) [1, 2] and causes an annual death of approximately 250,000 people [1].

Type I hypersensitivity is often triggered by allergens, a type of protein defined by and that can interact with allergen-specific immunoglobulin E (IgE) antibodies. Adaptive IgE-mediated immunity consists of two main phases: allergic sensitization and re-exposure (Figure 1-1). In the sensitization phase, the immune system is exposed to allergens for the first time and antigen presenting cells such as dendritic cells take up allergens and process them displaying specific peptides to naïve T cells. The subsequent production of allergen specific IgE antibodies by B cells may occur in an environment favoring the skewing of naïve T cells towards a T_H2 phenotype. Allergen specific IgE molecules are attached on the surface of the mast cells and/or basophils. When re-exposure occurs, allergens can then bind to specific IgE on mast cells and basophils and trigger the degranulation process including massive releases of small inflammatory mediators such as histamine, prostaglandins, and leukotrienes. As a side effect, these

mediators cause a variety of symptoms from mild ones such as red eyes, sneezing, wheezing, coughing, and itching to severe asthma exacerbations and life threatening anaphylaxis [3, 4].

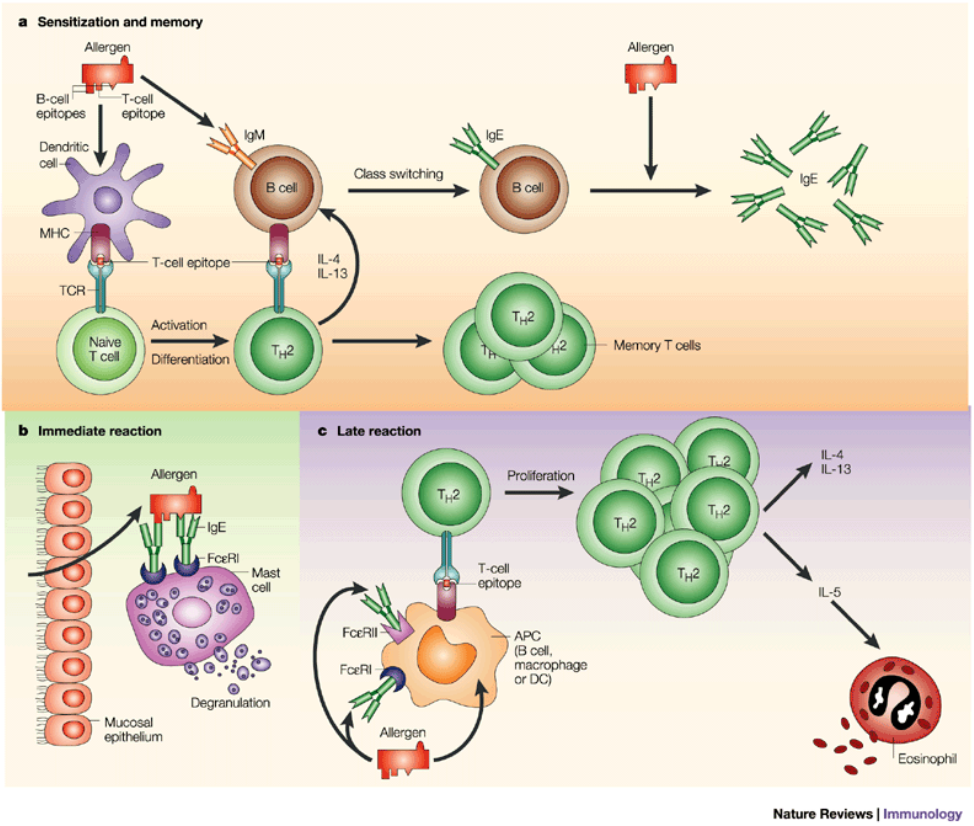


Figure 1-1. Molecular biology of allergy. **(a)** Sensitization and memory: Allergens entering human body are taken up and processed by antigen-presenting cells (APC, for example, dendritic cells). APCs present parts of allergens (epitopes) to naïve T cells via the major histocompatibility complexes (MHC), which in the presence of specific T_H2 driving cytokines and chemokines instructs naïve T cells to differentiate to T_H2 cells. T_H2 cells then produce cytokines such as interleukin-4 (IL-4) and IL-13 that signal the immunoglobulin-class switching of allergen specific B cells to IgE (sensitization). Many IgE B cells are established (memory) and subsequent allergen contact will boost IgE memory B cells to produce increased levels of allergen-specific IgE antibodies. These IgE antibodies are then attached onto mast cells, basophils, monocytes, dendritic cells and B cells. **(b)** Immediate reaction: The similar allergens entering sensitized individuals will bind to and crosslink IgEs on the surface of mast cells that triggers the releases of inflammatory mediators (histamine, leukotrienes) that causes multiple immediate symptoms of allergy. **(c)** Late reaction: T_H2 cells proliferate and produce cytokines (IL-4, IL-5, IL-13) when being presented with allergens by APCs. IL-5 induces tissue eosinophilia that release additional inflammatory mediators.

Reprinted by permission from Macmillan Publishers Ltd: Nature Reviews Immunology. 2(6): 446-53, copyright 2002 [5]

It is important to further our overall understanding of the molecular and biochemical nature of allergenic proteins as well as their environmental distribution. The

known allergens belong to many common functional categories such as proteolytic enzymes, binding proteins, lipid transfer proteins, seed storage proteins, and pathogenesis-related proteins [6, 7].

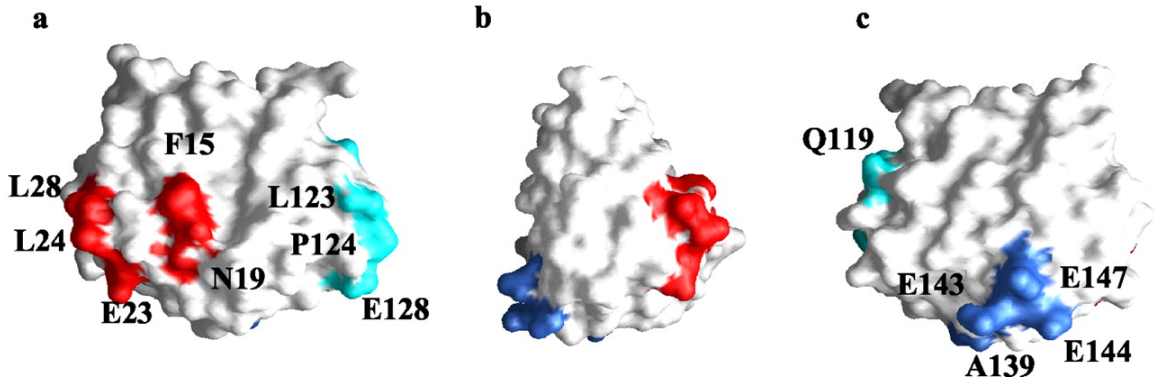


Figure 1-2. The exposed residues of three IgE epitopes (colored differently) on the molecular surface of a major cat allergen Fel d 1 with different angle views (b is 90^{o} angle of a, c is 180^{o} of a).

Reproduced with permission from ASBMB Journals: Kaiser L et al. J. Biol. Chem. 2003;278:37730-37735 [8]

One key component of an allergenic protein is the IgE epitope(s), an essential region of allergens that can be recognized by allergen-specific IgE antibodies (Figure 1-2). IgE epitopes exist in both linear form (continuous amino acids) and conformational form (discontinuous amino acids brought together via protein folding) [9, 10]. The length of IgE epitope sequences vary from as short as a few amino acids to as long as 30-70 amino acids according to the structural database of allergenic proteins (SDAP) (Figure 1-3). A major focus of allergen studies has been identifying and mapping IgE epitopes. However, allergens have diverse structures that are often difficult to crystalize and model. Currently there are only 92 allergens with a defined PDB structure and 29 allergens with experimentally characterized epitopes according to the Structural Database of Allergenic Proteins or SDAP [11].

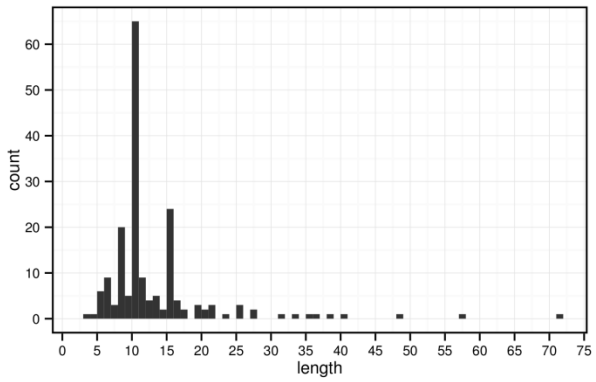


Figure 1-3. Distribution of the length of 183 known IgE epitopes collected from the SDAP database.

Besides having IgE epitopes, other aspects of protein structure relevant to allergenicity include solubility, stability, size, and protein compactness [12]. Our understanding of the features that contribute to the allergenicity of specific proteins is incomplete. Cross-reactivity of IgE may occur if a known allergen is structurally similar to another protein or shares sequence similarity, especially across IgE epitope containing regions. For example, allergic sensitization to specific birch pollen proteins can also induce allergy to similar proteins found in apple pollen [13].

1.2 Fungi as a human allergenic pathogen

Fungi or often referred to as molds are ubiquitous eukaryotic organisms that are found to grow in diverse habitats over a wide range of temperature although the optimal growth temperature varies between species and is often in the temperate range from 18^{o} to 32^{o}C [14]. They are primarily heterotrophic organisms and obtain nutrition from their environment. Besides being saprophytes, many fungi have developed relationships with other organisms to aid in nutrient acquisition including symbiotic and parasitic interactions with plants and animals. Fungi are found in both indoor and outdoor environments and often compose the largest portion of airborne particles, with the most common fungal genera being *Cladosporium*, *Alternaria*, *Penicillum*, and *Aspergillus*.

Human respiratory exposure to fungal spores and hyphae is almost constant throughout the year. Indoor and outdoor exposure to fungal components is a recognized triggering factor for respiratory allergy and asthma. More than 80 mold genera have been

shown to induce type I hypersensitivity in susceptible people whereas allergenic proteins have been identified in 25 genera, including 22 molds and 3 mushrooms (http://allergen.org). Allergic sensitization to fungi has been associated with moderate to severe asthma (sometimes life-threatening) in various countries [3, 4]. Interestingly, the most important allergenic fungal genera are *Alternaria, Aspergillus, Penicillium,* and *Cladosporium*. This is most likely due to the fact that these are the most abundant fungi in the environment. However, there may be other factors from these organisms that cause humans to develop allergies to them besides their ubiquitous nature such as the nature of their secretomes and growth rates at higher temperatures.

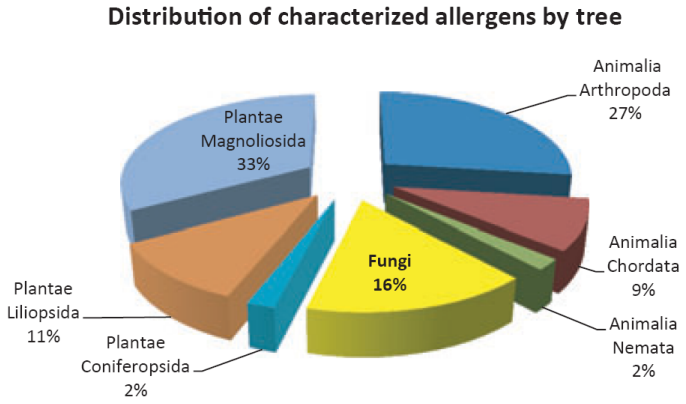


Figure 1-4. Taxonomic distribution of 753 allergens recognized by WHO/IUIS Allergen Nomenclature Subcommittee.

Reprinted by permission from John Wiley & Sons Ltd: Crameri R, Garbani M, Rhyner C, Huitema C. Fungi: the neglected allergenic sources. Allergy 2014; 69: 176–185 [15]

Fungi contribute to a large portion of causative agents of allergy. Sixteen percent of the currently recognized allergens by the World Heath Organization and the International Union of Immunology Society (WHO/IUIS) Allergen Nomenclature Sub-committee come from fungi (Figure 1-4). Studies have shown that fungi can produce a very diverse repertoire of allergenic proteins. Recent comparative studies showed that fungal genomes harbor many mold allergen homologs, including species specific and cross-reactive allergens [16, 17]. Other substances such as house dust mite, cat dander and tree pollen are also common environmental sources of allergens, but fungi alone have

the ability to actively germinate and secrete additional molecules in the host in an attempt to infect or colonize the skin and the respiratory tract. These non-allergenic toxins, enzymes, and other pro-inflammatory factors may play an accessory role in triggering and exacerbating chronic inflammatory disorders by stimulating innate immune responses and influencing the development of adaptive T_H2 immunity. Thus, these additiounal factors may contribute to the overall allergic inflammatory potential of fungi.

1.3 The fungal genus *Alternaria* and its pathogenicity

Alternaria is considered one of the most common saprophytic fungal genera on the planet. It is comprised of many species that exhibit a necrotrophic phytopathogenic lifestyle. Several species are clinically associated with allergic respiratory disorders although rarely found to cause invasive infections in humans. *Alternaria* spp. are also among the most well known producers of diverse fungal secondary metabolites, especially toxins.

Sensitivity to the fungus *A. alternata* is a common cause of allergic rhinitis and believed to be a common cause of atopic or allergic asthma. Epidemiological studies from a variety of locations worldwide indicate that *Alternaria* sensitivity is closely linked with the development of asthma and up to 70% of mold-allergic patients have skin test reactivity to *Alternaria* [26–30]. Additionally, *Alternaria* sensitization has been determined to be one of the most important factors in the onset of childhood asthma in the southwest deserts of the U.S. and other arid regions [27, 31]. *Alternaria* spores are routinely found in atmospheric surveys in the U.S. and in other countries and are the most frequently encountered fungal spore type [32]. Fungal exposure differs from pollen exposure in quantity (airborne spore counts are often 1,000-fold greater than pollen counts) and duration (e.g. *Alternaria* exposure occurs for months, whereas ragweed exposure for example occurs for weeks). This prolonged intense exposure is similar to that of other asthma-associated allergens such as cat dander and dust mites. Indeed it has long been speculated that this type of exposure may be partially responsible for both the chronic nature and severity of asthma in *Alternaria* sensitive individuals [33].

Although some research has been performed on the physiological and molecular identification of *Alternaria* allergens only three major and several minor allergenic proteins have been described [34]. The biological role of these allergens and other fungal products in the development of allergy and asthma is very poorly understood. Other than a few studies demonstrating binding of these allergens to IgE-specific antibodies in human sera from patients diagnosed as being *Alternaria* sensitive, virtually nothing is known about how these highly immunoreactive proteins interact with the host. It has been demonstrated in several recently published studies that protease activity in *A. alternata* culture filtrates has marked proinflammatory properties in treated lung cells [35, 36]. Thus there is clearly a need to identify and elucidate the role of *Alternaria* allergens, immunostimulatory proteins as well as perhaps small molecules (secondary metabolites) in the development of allergic airway diseases from both diagnostic and immunotherapeutic perspectives. Indeed in a recent survey by the National Institute of Environmental Health & Safety (NIEHS) of 831 homes containing 2,456 individuals it was found that the prevalence of current symptomatic asthma increased with increasing *Alternaria* concentrations. Higher levels of *Alternaria* antigens in the environment significantly increased odds of having had asthma symptoms during the preceding year, more so than other examined antigens [28].

In addition, *Alternaria* species are now gaining attention as emerging human invasive pathogens, particularly in immunocompromised patients [37, 38]. Several *Alternaria* species have been found associated with infections of the eye, oral and sinus cavities, respiratory tract, skin, and nails [39, 40].

1.4 Aims of this dissertation

This dissertation investigates computational approaches in allergen identification, develops a practical tool for large-scale allergen prediction, and uses this tool together with comparative approaches to study fungal allergen genomics in the context of fungal genome annotation. We centered our allergen-related informatics studies to complement the *Alternaria* genomes project where we annotated and compared *Alternaria* genomes sequenced from 25 species and isolates.

The specific aims are: (1) Analyzing *Alternaria* genome sequences in a search for genetic explanations of lifestyles and pathogenicity of this particular fungal genus, including human allergic pathogenicity; (2) Investigating computational approaches in allergen identification and developing a suitable tool for large-scale allergen prediction such as a whole genome scale; and (3) Analyzing allergen repertoire of a number of widely recognized allergenic fungal genera using computational approaches.

The dissertation is organized as follows. Chapter 1 provides the background and introduction on allergy and fungal pathogenicity. Chapter 2 and Chapter 3 focus on the first aim. In Chapter 2, we develop a robust computational infrastructure for fungal genomics including a pipeline for genome annotation and comparison and a genome database. In Chapter 3, we use the computational infrastructure to perform genome annotation and comparative studies of three *Alternaria* species, including two human allergenic isolates of *Alternaria alternata* and a necrotrophic plant pathogen *Alternaria brassicicola*. Chapters 4 and 5 focus on the second aim. In Chapter 4, we evaluate sequence similarity-based approaches in allergen identification with the goal of discerning optimal parameters for large-scale imbalance data sets. In Chapter 5, we develop a fast allergen prediction tool using supervised classification techniques in machine learning that is capable of producing high precision over high recall on imbalanced data sets and is suitable for large-scale allergen prediction in applications such as whole genome annotation. Chapter 6 focuses on the third aim. Using the newly developed method for allergen prediction, together with comparative approaches, we survey the allergen repertoires of the widely recognized allergenic molds. This provides an overall picture of the potential allergen distribution among these fungal species. Chapter 7 contains conclusions and discusses some future work.

Chapter 2

Identifying and annotating functional elements in *Alternaria* genomes

Parts of this chapter are included in:

Dang H and Lawrence C. (2014) Alternaria Comparative Genomics: The Secret Life of Rots. In Dean R, Lichens-Park A, and Kole C (Eds), Genomics of Plant-Associated Fungi and Oomycetes. Springer, 2014 (in press).

Dang H, Pryor B, Peever T, and Lawrence C. (2014) Alternaria genomes database: a resource for a ubiquitous fungal genus comprised of plant pathogens, saprophytes, and allergenic species (submitted for review).

2.1 Introduction

The number of fungal species is estimated to be in the millions representing one of the largest and most diverse branches of the tree of life [41, 42]. Fungal genome sequencing efforts have tremendously facilitated our understanding of the molecular basis of fungal pathogenicity as a whole system [43–46]. Approximately 600-700 fungal genomes have been sequenced by multiple sequencing centers worldwide (Figure 2-1) and many more species are being targeted for similar efforts. The majority of the sequenced fungal species belong to two phylums *Ascomycota* (molds) and *Basidiomycota* (mushrooms). We have recently sequenced 25 *Alternaria* species and isolates using multiple sequencing technologies, including Sanger, GS-FLX 454, and Illumina platforms. Genome annotation and comparison are the key and very first steps to uncover the genetic basis for *Alternaria's* plant pathogenicity, its ubiquitous saprophytic nature, its pathological relationship with mammals, and taxonomic aspects of this important fungal genus.

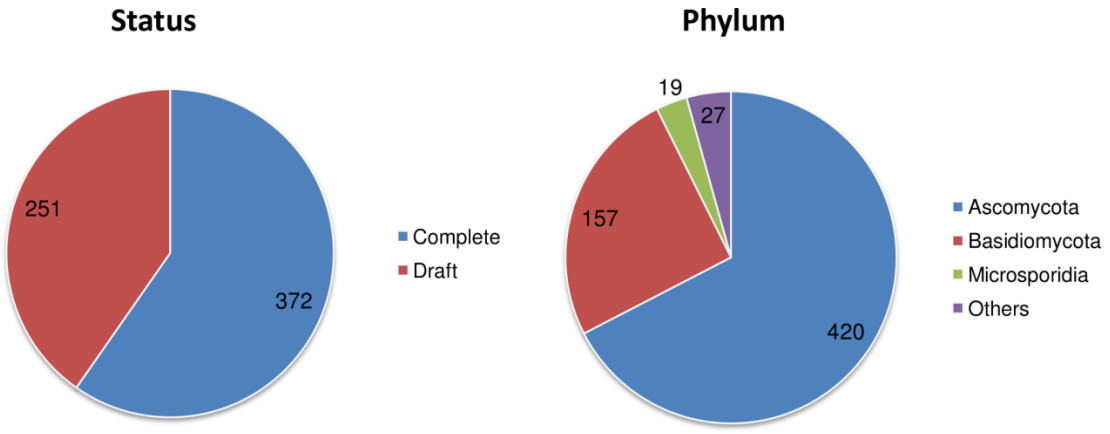


Figure 2-1. Number of draft or complete fungal genome projects. Data collected from the GOLD database (http://genomesonline.org) in March 2014.

In genome annotation, raw genomic DNA sequences are analyzed to identify protein coding genes and other important genomic components such as repetitive elements and non-coding genes. Protein coding genes/transcripts are then translated into protein sequences. Biological information is then attached to these protein sequences using various techniques from simple sequence similarity-based comparative genomics to sophisticated network-based and machine learning approaches. For example, hidden Markov model (HMM) profile searches can be used to identify conserved domains of proteins that allow one to predict the function of a gene and its protein products.

Comparative genomics is one of the most important computational tools in genome annotation in which biological functions are transferred from functionally known proteins to new proteins based on their sequence similarity, with the assumption that proteins with similar sequences are possible homologs. Comparative genomics can be performed at an individual gene level as well as at a whole genome scale. Comparison of different genomes also aids in the understanding of evolutionary events (e.g. genome rearrangement) as well as genetic similarities and differences between species/isolates by which we can identify important genetic components that may explain variable phenotypes and lifestyles. For example, a set of species-specific genes of a pathogenic fungus when compared with a closely related non-pathogenic fungus represents potential candidates for initial functional analysis of virulence genes. Indeed, comparative genomics has been used widely in many fungal genome projects to discover the

evolutionary and pathological relationship between closely and/or distantly related species [47, 48].

The Department of Energy Joint Genome Institute (DOE-JGI) and the Broad Institute have published a brief description of their in-house fungal genome annotation pipelines [49, 50] but these pipelines are not publicly available at this time. The Ensembl pipeline is publicly available but lacks documentation and is difficult to install locally [51]. To the best of our knowledge, it is not widely used outside of the European Bioinformatics Institute (EBI). Smaller laboratories often use simple annotation pipelines with very basic features (e.g. single gene prediction software, BLAST-based annotation). We developed a genome annotation and comparison pipeline for fungal genomes, optimized it for *Alternaria* species and used this pipeline to annotate and compare *Alternaria* genomes. A genome database was also developed based on the Ensembl genome browser platform [51] to house and present annotated *Alternaria* genomes to the public. The *Alternaria* genomes database is currently available at http://alternaria.vbi.vt.edu.

2.2 Materials and methods

To annotate *Alternaria* genomes, a custom pipeline for fungal genome annotation and comparative genomics was developed at the Virginia Bioinformatics Institute (VBI) (Figure 2-2). We herein call it the *Alternaria* pipeline. The pipeline receives inputs as assembled genomes (i.e. genomic sequences, often in the form of contigs or supercontigs). These genomic sequences were first scanned for repetitive sequences (both transposable elements and simple repeats). Repetitive sequences were then masked from the original genomic sequences. Masked genomic sequences were the starting point for various subsequent analyses, including gene prediction, functional annotation, and genome comparison.

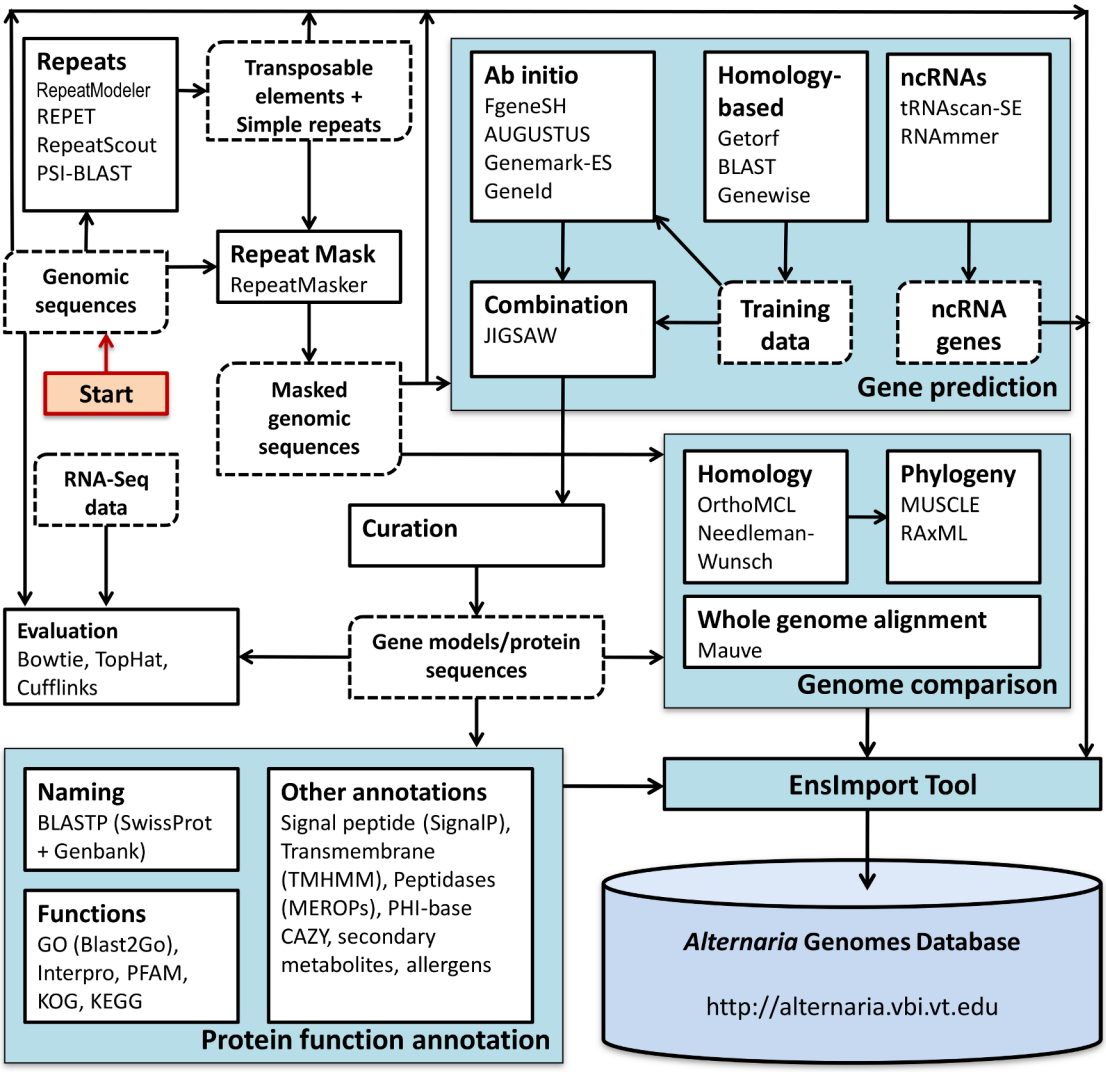


Figure 2-2. *Alternaria* genomes annotation and comparison pipeline (*Alternaria* pipeline).

2.2.1 Repetitive sequence annotation

For repetitive sequence discovery, multiple existing tools including RepeatModeler (http://www.repeatmasker.org), REPET [52], and RepeatScout [57] were used as part of the *Alternaria* pipeline. RepeatModeler combines repetitive sequences predicted by RECON, RepeatScout, and Tandam Repeat Finder [xxx Benson et al.]. For most *Altenaria* genomes, RepeatModeler was used to automatically annotate repetitive content.

For several *Alternaria* species including *A. alternata* ATCC 66981, ATCC 11680, and *A. brassicicola*, a more comprehensive repetitive sequence annotation was performed. REPET and RepeatScout was used for repetitive sequence prediction. REPET was developed specifically for fungi and uses BLAST [53], RECON [54] and PILER [55] for *de novo* repetitive element prediction. It was first used for predicting repeats in the genome of another closely related Dothideomycete, *L. maculans* [56]. In the annotation of *Alternaria* repetitive sequence, results from REPET were combined with repetitive sequences also discovered *de novo* by RepeatScout. All consensus repetitive elements were then compared using BLAST and clustered to remove redundancy. In the *Alternaria* pipeline, a consensus sequence was considered redundant and removed if ≥70% of it is covered by another sequence in a pairwise BLAST alignment with an e-value ≤ 10^{-10}. After repeat families were identified, they were classified by both an automatic classification pipeline and human curation. REPET was used again to classify repeat elements in which their repeat structures were identified and their sequences were compared with known elements from Repbase [58]. After the initial annotation by REPET, another round was performed where repeat families were searched against Genbank using PSI-BLAST and BLASTX to identify homologous transposable elements (e-value ≤ 10^{-4}). Repetitive sequences were then masked by RepeatMasker (http://repeatmasker.org) before further analyses were performed.

2.2.2 Gene prediction

The *Alternaria* pipeline used multiple gene predictors including *de novo* and evidence-based gene prediction tools. To date, evidence-based gene predictors such as homology-based methods are among the most accurate methods for eukaryotic gene structure annotation. Homology-based gene predictors build gene models by aligning DNA sequences with known proteins, genes, or mRNAs. It is often a powerful approach to recover orthologs of known genes in genomes of closely related species that can also be used as training gene models for gene prediction software. In order to maximize the accuracy of predicted gene models, multiple gene predictors for eukaryotic and fungal genomes were used in the *Alternaria* annotation pipeline, including homology-based tools such as Genewise [59], as well as machine learning based tools such as FgeneSH,

AUGUSTUS [60], Genemark-ES [61], and GeneID [62]. When annotating *Alternaria* gene structure using the *Alternaria* pipeline, supervised gene predictors were re-trained (when needed) using known *Alternaria* gene models surveyed from Genbank and high scoring genes built using Genewise. To speed up predictions with Genewise, Emboss getorf software [63] was used to identify all possible open reading frames (ORFs) that are longer than 500bp from genomic sequences which were then searched against Genbank using BLAST to identify possible homologous proteins. ORFs that were found to match with known proteins (e.g. e-value $\leq 10^{-20}$) were paired with these proteins and Genewise was then called to predict gene models for those ORFs from their paired proteins.

Gene models from individual predictors were then combined to produce the best models. The practice of combining results from different computational approaches is known as Ensemble learning and has proven useful in many fields. Combination approaches could potentially produce more accurate gene models than individual methods [64]. There are many tools to combine multiple gene models into one model. The JIGSAW combiner [65] was used for the *Alternaria* pipeline. JIGSAW uses decision trees to score different combinations of individual gene models derived from various predictors and selects the best scoring models. A subset of high quality gene models that was not used to train individual gene predictors was used to train JIGSAW for gene model combination. Beside protein coding genes, non-coding genes were predicted using rnammer [66] (rRNA genes) and tRNAScan-SE [67] (tRNA genes). Lastly, RNA-sequencing reads when available were mapped to the genome (and quantified) using the Tuxedo tool set [68] as an additional annotation feature and aids in evaluating gene prediction.

2.2.3 Protein functional annotation

Following gene prediction, gene models were transcribed and then translated to protein sequences. Various comparative computational tools were then used to infer the functions of the new proteins based on their sequence under the assumption that proteins with similar sequences are possible orthologs and share the same functions. This comparative approach is widely used in annotation of protein function and is relatively

accurate for proteins that possess a high level of sequence similarity with known proteins or protein families.

In the *Alternaria* pipeline, new protein sequences were searched (using BLAST) against publicly available sequence databases such as Swiss-Prot [69] and Genbank [70] to identify putative homologous proteins. Based upon results of these searches, new proteins were named according to the Broad Institute standard operating procedure [50] and predicted functions and other corresponding annotation were then assigned to the new proteins.

To gain more detailed information regarding putative functions of the proteins, domain annotation was performed in addition to whole sequence level BLAST-based approaches. In the *Alternaria* pipeline, InterproScan [71] was used to query predicted proteins against protein signature databases, including PFAM, Superfamily, ProDom, Panther, TIGR, PIR, SMART, and PROSITE, and domain architectures of the predicted proteins were derived and then used to infer protein functions in addition to BLAST-based annotation. The predicted proteins were also assigned to the eukaryote orthologous groups (KOGs) by searching the KOG profiles from the NCBI conserved domain database (CDD) [72] using RPS-BLAST [73]. Gene ontology (GO) terms were assigned to new proteins by a combination of results from InterproScan and Blast2GO [74].

2.2.4 Functional annotation relevant to saprophytic and human/plant pathogenic fungi

Specific annotation features of fungal proteins are important to predict especially those thought to play important roles in fungal pathogenicity and survival. These features may also include protein localization and secretion attributes, ability to acquire and digest different substrates, production of secondary metabolites (toxins, antibiotics, etc.), suppression of host defense, and features associated with the ability to detoxify antifungal chemicals and proteins from the host to name a few. For example, effector proteins are often small, cysteine-rich secreted virulence associated proteins [75]. PHI-base (http://www.phi-base.org) annotation also aids in identifying proteins that are similar to the known fungal plant pathogenicity-related proteins. BLAST searches against PHI-base database are used to link predicted proteins with experimentally verified pathogenicity,

virulence and effector genes. Carbohydrate active enzymes (CAZY) are often secreted and important for the fungal saprophytic lifestyle but also may play a role in pathogenesis via plant cell wall breakdown.

Protein localization

In the *Alternaria* pipeline, secreted proteins were predicted by SignalP [76] (using both HMM and neural network models), WoLF-psort [77], and Phobius [78]. A protein was predicted to be secreted when signal peptides were found by all three programs. Transmembrane proteins are often involved in sensing the environment and subsequent signaling. Transmembrane protein topology was predicted using TMHMM software [79].

Carbohydrate active enzymes

Many fungi including *Alternaria spp*. are ubiquitous saprophytes that have the capability of acquiring nutrients from dead or decaying organic sources in the environment. Nutrients are often acquired via the action of carbohydrate active enzymes, proteinases, and lipases. In addition to BLAST against Genbank and/or Swiss-Prot, carbohydrate active enzymes were annotated using the CAZY annotation tool [80] for several *Alternaria* species including *A. alternata* ATCC 66981, ATCC 11680, and *A. brassicicola* as described in more detail in Chapter 3.

Secondary metabolites

Another class of genes that are important for fungal pathogenicity is secondary metabolite biosynthetic genes. For several *Alternaria* species including *A. alternata* ATCC 66981, ATCC 11680, and *A. brassicicola*, in addition to BLAST, SMURF was used [81] to identify secondary metabolite gene clusters in addition to looking at domain architecture of the proteins to identify and characterize secondary metabolite production-related genes.

2.2.5 Genome comparison

Multiple genome comparison tasks were performed that utilized the genome sequences as well as the predicted genes/proteins from multiple species. Whole genome pairwise alignment was performed using Mauve progressive alignment software [82, 83].

Orthologs and paralogs were identified using bidirectional best BLAST hits and Markov clustering via OrthoMCL [84].

2.2.6 Housing, visualization and distribution of fungal genomic data

Annotation and comparison data of *Alternaria* genomes are presented via the popular Ensembl genome browser platform [51] that was customized and installed at the VBI. Outputs from the genome annotation as well as outputs from comparative genomics analyses were processed and converted to Ensembl compatible MySQL databases using EnsImport, a custom suite of scripts we developed in Perl/BioPerl (Figure 2-2). EnsImport supports multiple standard file formats such as FASTA, AGP, GFF3 and XMFA, and outputs from widely used tools such as BLAST, Interpro, RepeatMasker, OrthoMCL and Blast2GO that are parts of the *Alternaria* pipeline.

2.3 Results and discussion

2.3.1 Summary of annotated *Alternaria* genomes

Using the genome annotation pipeline, we annotated genomes from 25 *Alternaria* species, including saprophytes, necrotrophic plant pathogens and species associated with human diseases such as allergic airway disorders (Table 2-1). The genome sizes of the fungal genomes were in the range of ~30-40Mb. The total number of genes predicted for each genome ranged from ~10,000-15,000, which is comparable with the number of genes from other well annotated fungal genomes of similar overall size [43, 85, 86].

Transposable elements (and other types of repeats) in fungal genomes make up a portion from as small as ~1% to as large as ~65% [47, 56, 87, 88]. Although most fungal genomes have relatively small amounts of repetitive DNA, it is thought to play important roles in fungal evolution [89–91]. We found that repetitive sequences in *Alternaria* accounted for ~1-10% of a genome. For example, the percentage of repetitive sequences in *A. brassicicola* and *A. alternata* genomes were ~9% and ~1%, respectively. The larger repetitive content is a possible explanation for higher genome rearrangement rate in *A. brassicicola* (which will be discussed later in Chapter 4).

Table 2-1. Summary of genome sequence and gene prediction for *Alternaria* species

Species name	Strain codes	Additional Information	Sequencing technology	Genome sequence size (Mb)	Contigs/ super-contigs	N50 (kb)	Repetitive sequences (%)	Predicted genes (#)
A. alternata	ATCC 66891, EGS 34-016, BMP 0269	Allergic disease of human, leaf spot, rots of plants	454	33.2	499	300	0.82	11635
A. alternata	ATCC 11680, BMP0238, IHEM 4706	Allergic disease of human, leaf spot, rots of plants	454	33.8	797	450	0.81	12323
A. alternata	ATCC 66982, EGS 34-039, BMP 0270	Allergic disease of human, leaf spot, rots of plants	Illumina	33.5	393	757	1.26	12290
A. arborescens	ATCC 204491, EGS 39-128, BMP 0308	Stem canker disease of tomato	Illumina	34.0	1332	624	1.94	14741
A. brassicicola	ATCC 96836, EGS 42-002,BMP 1950	Blackspot disease of brassica	Sanger	29.6	4039/ 838	2400	8.6	10514
A. citriarbusti	EGS 46-140, BMP 2343, SH-MIL-8s	Brown/black spot disease of citrus	Illumina	34.1	2273	48	1.69	12606
A. carthami	BMP 1963, CBS 635.80	Leaf spot and blight ofsafflower	Illumina	34.5	9340	72	5.9	12071
A. capsici	ATCC MYA-998, EGS 45-075, BMP 0180	Leaf spot of solanaceae (pepper)	Illumina	34.0	13743	31	6.68	11487
A. crassa	BMP 0172, ACR1	Leaf spot of solanaceae	Illumina	35.0	12126	54	7.39	11663
A. dauci	ATCC 36613, BMP 0167	Leaf blight ofcarrots	Illumina	32.1	12030	13	4.17	11981
A. destruens	ATCC 204363, EGS 46-069, BMP 0317	Infecting and suppressing dodder (weed)	Illumina	41.8	31070	3	1.53	14814
A. fragariae	BMP 3062, NAF-8	Black spot disease of strawberry	Illumina	33.2	1027	78	0.98	12272
A. gaisen	EGS 90-0512, BMP 2338	Black spot, ring spot disease of pear	Illumina	34.6	7485	10	1.17	13902
A. limoniasperae	EGS 45-080, BMP 2327, BC2-RLR-1s	Leaf spot disease of citrus	Illumina	34.0	2459	37	1.4	12639
A. longipes	EGS 30-033, BMP 0313	Black/brown leaf spot of tobacco	Illumina	36.3	3412	137	2.29	13219
A. mali	BMP 3064, IFO8984	Leaf ring spot of apple	Illumina	34.7	2682	35	1.97	12715
A. mali	BMP 3063, M-71	Leaf ring spot of apple	Illumina	34.1	4439	21	2.22	12727
A. macrospora	BMP 1949, CH3	Leaf spot ofcotton	Illumina	31.7	3153	37	0.76	11961
A. porri	BMP 0178, Z6B	Purple blotch, leaf blight and bulb rot of Allium (onion)	Illumina	31.2	16767	9	2.94	12232
A. solani	BMP 0185	Early blight of potatoes and tomatoes	Illumina	32.9	5613	144	3.2	11726
A. tagetica	EGS 44-044, BMP 0179	Leaf spot of marigold	Illumina	35.1	16372	72	6.26	11999
A. tangelonis	BMP 3436, SH-MIL-20s	Leaf spot of tangelo, citrus	Illumina	34.0	2347	33	1.28	12739
A. tomatophila	BMP 2032, CBS 109156	Leaf spot oftomato	Illumina	34.1	10185	22	3.9	12601
A. tenuissima	ATCC 96828, EGS 34-015, BMP 0304	Leaf spot of plants	Illumina	33.5	676	662	1.51	12276
A. turkisafria	EGS 44-159, BMP 2335	Leaf spot of tangelo, citrus	Illumina	35.1	2796	50	1.97	12966

2.3.2 Overview of *Alternaria* gene functions

According to KOG annotation, the functional distribution of *Alternaria* genomes was overall very similar. The largest group was general function (category R), followed by signal transduction mechanisms (category T), posttranslational modification/protein

turnover/chaperones (category O), lipid transport and metabolism (category I), intracellular trafficking/secretion/versicular transport (category U), transcription (category T), carbohydrate transport and metabolism (category G), and secondary metabolites biosynthesis, transport, and catabolism (Figure 2-3). However, minor to moderate differences were observed for some individual species/isolates in various categories (Figure 2-4).

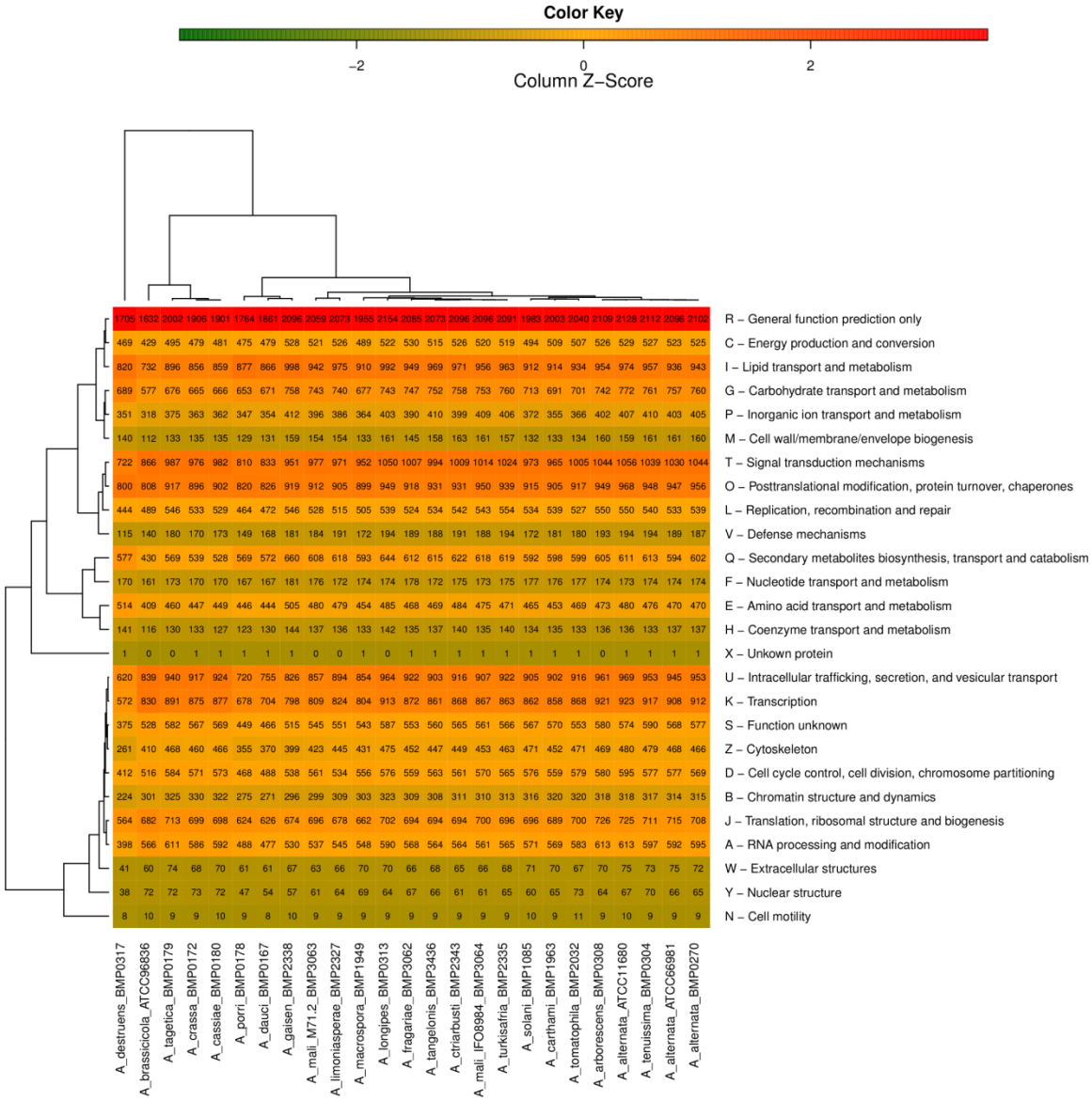


Figure 2-3. Number of genes classified into different KOG categories (colored by scaled values within columns).

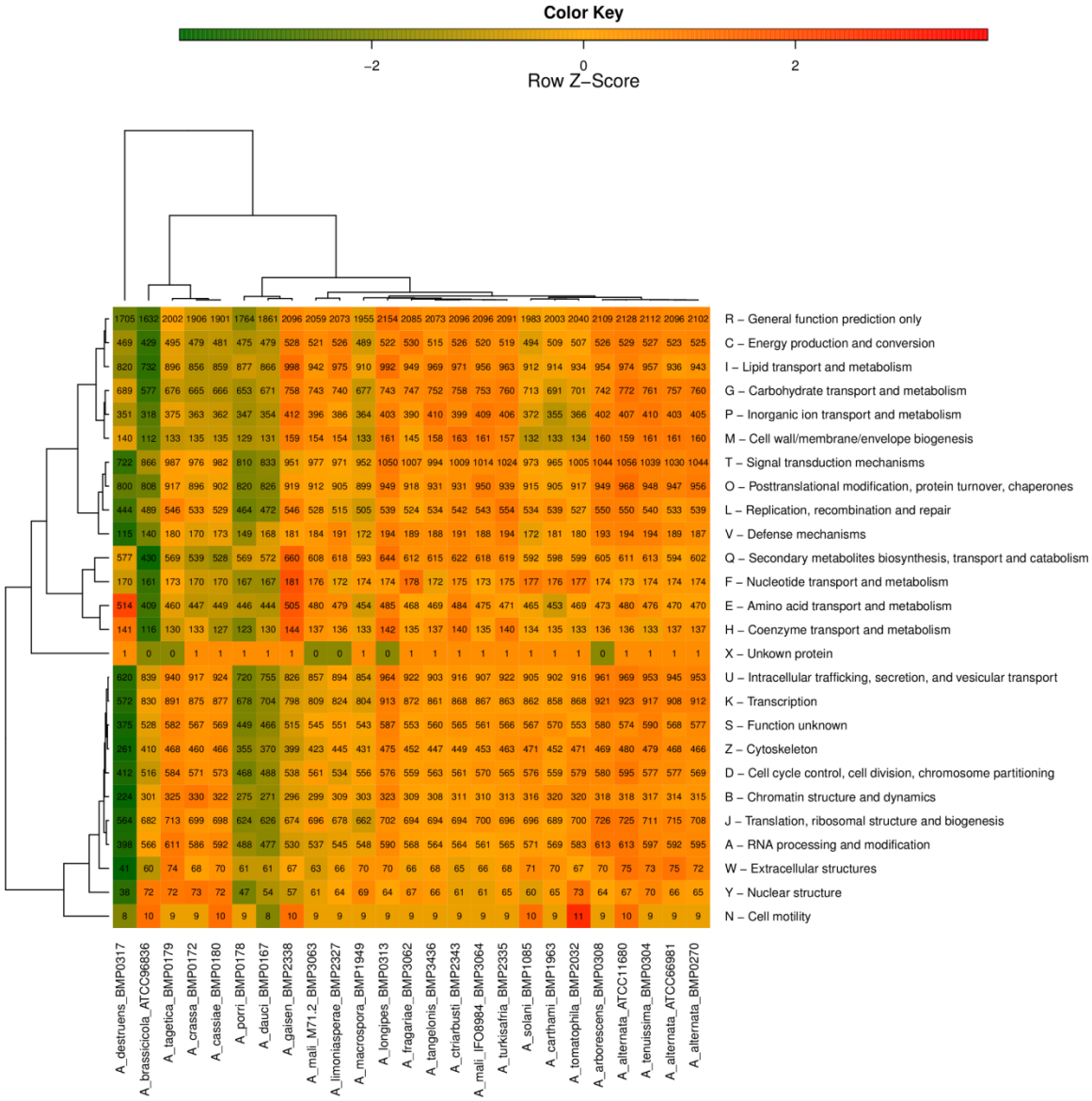


Figure 2-4. Number of genes classified into different KOG categories (colored by scaled values within rows).

2.3.3 *Alternaria* genomes database

We have annotated and compared 25 *Alternaria* genomes and used EnsImport to port genome annotation and comparison data to a local installation of the Ensembl genome browser platform at VBI (http://alternaria.vbi.vt.edu). Using the Ensembl genome browser platform, the *Alternaria* genomes database provides a rich set of user-friendly tools to browse and visualize sequences, annotation, and comparison data. Data export and search features are also available. Detailed instructions on how to use the

Ensembl browser are available on the 'Help & Documentation' section of the database. Here we only describe the most relevant features in the context of the *Alternaria* genomes project.

Genome region view

For each species, users can access and visualize a genomic region along with annotated functional and non-functional elements such as repetitive elements, predicted protein-coding gene models, and RNA coding gene models (Figure 2-5). A genomic region can be a whole (or part of) a contig or supercontig. Zooming functionality allows for intuitively scaling region views based on location. Each type of element (functional and non-functional) is displayed in a separate track using a unique color. Users can click on an individual element (e.g. repeats, genes, transcripts) to open a pop-up menu to access available annotation. The tracks can be displayed or hidden using the display configuration tool.

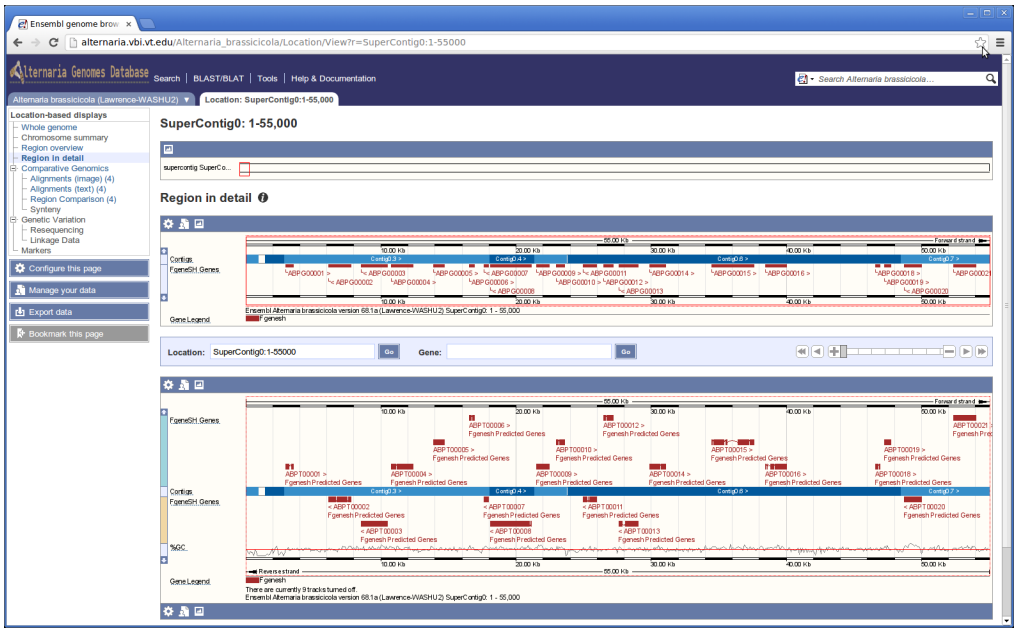


Figure 2-5. A screenshot of the *Alternaria* genomes database that shows a region of an *A. brassicicola* supercontig along with the predicted genes and transcripts.

Annotation view

The majority of functional annotation data in the database corresponds to predicted protein coding genes. For each gene/protein, extensive annotations include gene

structure and sequence, gene description, genomic location, protein domain architectures (e.g. Interpro, PFAM), gene ontology assignments, signal peptides, transmembrane structures and other annotation data (Figure 2-6). These annotation data are available and presented in multiple tightly linked web interfaces in the browser.

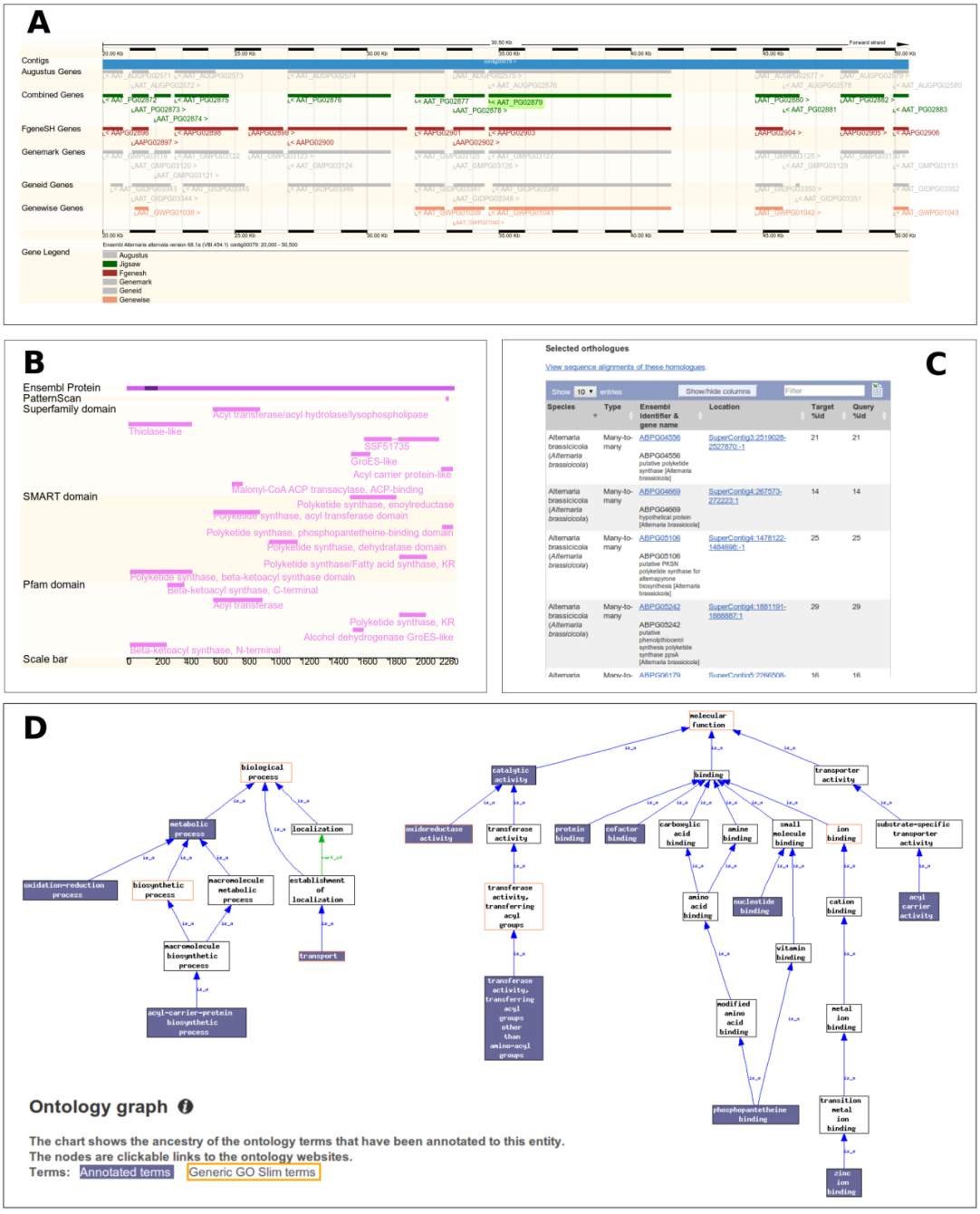


Figure 2-6. Examples of annotation and comparison views for an *A. alternata* polyketide synthase gene (AAT_PG02879).(A) Contig view of the gene, (B) Domain annotation, (C) Orthologs from other species, (D) Gene ontology annotation.

Comparative genomics view

The comparative browsing feature of Ensembl platform allows for conveniently viewing and visualizing comparative genomics data side-by-side with annotation data. Aligned regions between two genomes identified via whole genome pairwise alignments

are displayed together with functional and non-functional elements such as repetitive elements and gene models (Figure 2-7). This feature allows for easy investigation of the conserved genomic regions between multiple genomes. Whole genome alignments can be visualized using graphical representation as well as displayed in text formats such as FASTA and ClustalW. Orthologs and paralogs of a gene can be easily retrieved in a table that contains links to access protein alignments and related annotation data (Figure 2-6C).

Figure 2-7. An example of a syntenic region between *A. brassicicola* and *A. alternata*. The aligned blocks (in pink) between genomic sequences are connected by green bands.

Database search

Users may query the database using sequence alignment search (e.g. BLAST) and text search. The built-in search feature of the Ensembl platform allows for BLAST searches against genomic sequences, predicted transcript and protein sequences (Figure 2-8). Full text search for gene names is also available as a built-in feature in Ensembl platform. However, for newly sequenced species, a large portion of the predicted genes are not named or annotated with highly reliable descriptions. In such cases, information on the hits with known proteins or protein families and domains can be used to explore the functions of the genes. Therefore, we implemented a more comprehensive search module that allows for full text search within annotation from multiple sources including BLAST and Interpro hits and incorporated this module in the *Alternaria* genomes database (Figure 2-8).

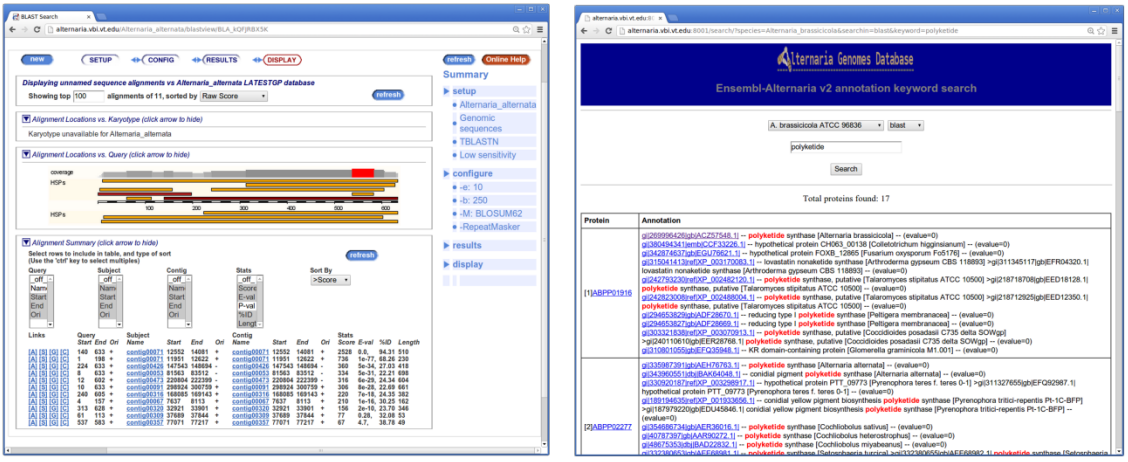


Figure 2-8. Search features of the *Alternaria* genomes database that allows for sequence alignment search using BLAST (left) and Interpro and BLAST hit description search (right).

Data export

Ensembl built-in functionality allows for exporting multiple types of data to various formats. Raw sequence and annotation data can be easily exported in multiple formats such as FASTA and GFF via available tools in Ensembl. A button to access the data export feature is located on the left pane in the interface of the database. It is also possible to export the graphical visualization of multiple types of annotation and

comparison data to multiple image formats that are suitable for publication or further editing.

2.4 Conclusion

We have annotated and compared 25 *Alternaria* genomes and made these data available to various fungal research laboratories. The *Alternaria* genomes database provides a comprehensive resource of genomics and comparative genomics data of an important plant and human pathogenic fungal genus. In addition, the database may prove useful for discovery of genes encoding industrial enzymes, antibiotics, and other molecules with utility in medicine and agriculture.

These genome annotation and comparison data have recently facilitated several large-scale functional genomics studies that resulted in the discovery of many new genes that contribute to virulence especially secondary metabolite genes, mitogen-activated protein (MAP) kinases, and transcription factors in *A. brassicicola* [92–102]. *Alternaria* genome annotation and comparison data have also enabled comprehensive comparative studies of *Alternaria* genomes in the context of plant and human pathogenicity [103].

The use of the familiar Ensembl browser platform makes browsing and visualizing *Alternaria* genome annotation and comparison data convenient. As we continue our efforts in *Alternaria* genome sequencing and analysis, we will update this database as new genomes and relevant annotation data become available.

We have also developed EnsImport, a tool that allows porting of genome data to Ensembl database schema. Although this tool was only used for the *Alternaria* genomes project, it can also be useful for other genome projects. EnsImport provides a "point-and-click" solution for ones who prefer to use only the browser features of Ensembl to present genomes and comparative data analyzed outside of Ensembl. For those who mainly rely on Ensembl pipelines, this tool can assist with integrating additional analyses not available in Ensembl into Ensembl databases. EnsImport has been extensively used to port *Alternaria* genome annotation and comparison to the customized installation of Ensembl at the VBI. EnsImport is available upon request.

Chapter 3

Alternaria comparative genomics and pathogenicity

3.1 Introduction

Alternaria species are a major cause of necrotrophic diseases of plants. Moreover, they are very often clinically associated with allergic respiratory disorders and are found, albeit rarely, to cause invasive infections in humans. They can be classified as belonging to kingdom Fungi, subkingdom Eumycotera, phylum *Fungi*, form class Hyphomycetes, form order Moniliales, form family Dematiaceae, and genus *Alternaria*. Some species of *Alternaria* are the asexual anamorph of the ascomycete *Pleospora,* others are speculated to be anamorphs of *Leptosphaeria* [18–20]. They are also broadly classified as members of the Dothideomycete group of fungi.

Alternaria species are some of the most common fungi encountered by humans. The number of species is estimated to range in the hundreds [18]. Many species are common saprophytes and have been recovered from very diverse substrates: plant material, sewage, leather, wood pulp, paper, textiles, building supplies, stone monuments, optical instruments, cosmetics, computer disks, and even jet fuel. This suggests that their genomes most likely contain many genes that encode proteins/enzymes that allow for the acquisition and utilization of quite diverse substrates as carbon and/or nitrogen sources for growth and reproduction.

Alternaria species are some of the most well known producers of diverse secondary metabolites especially toxins [21]. Over 70 small molecule compounds have been reported from *Alternaria* to date [21]. Some of these metabolites are potent mycotoxins (e.g. alternariol, alternariol methyl ether, tenuazonic acid, etc.) with mutagenic and teratogenic

properties, and have been linked to certain forms of cancer, such as esophogeal cancer, because of their genotoxic properties [22]. The occurrence of potentially harmful *Alternaria* metabolites in food and food products such as grains, peanuts, tomato products, and fresh fruits and vegetables is becoming an increasing environmental concern [23]. In addition, an emerging and critical area of research utilizes *A. alternata* f. sp. *lycopersici*, which produces the sphingolipid-like AAL toxin, as a model organism in the investigation of toxin-mediated programmed cell death (apoptosis) in animals and plants [24, 25]. Collectively these data strongly suggest that *Alternaria* is an ideal genus for studying the genome evolution and organization of genes involved in secondary metabolite biosynthesis and secretion.

In this chapter, we compared the genomes of a *Brassica* pathogen *A. brassicicola* and two saprophytic/allergenic *A. alternata* isolates. Although closely related taxonomically striking differences were revealed in regards to genomic content, the number of allergen homologues, the number and diversity of secondary metabolite biosynthetic gene clusters, and carbohydrate active enzymes, as well as proteases. All three fungi were found to be able to grow readily and more efficiently than 14 other fungi on diverse single carbon source substrates including xylan, cellulose, and lignin. The relationship of these findings in the context of plant pathogenicity, substrate utilization and biotechnological applications, and human health are discussed.

3.2 Materials and methods

3.2.1 Genome sequencing and assembly

DNA for both *A. brassicicola* and *A. alternata* genome sequencing projects were extracted using Qiagen Plant DNA extraction kit (Qiagen, Carlsbad, CA) and subject to RNAase treatment. The *A. brassicicola* isolate selected for sequencing, *A. brassicicola* ATCC 96866, was available from the American Type Culture Collection (Manassas, VA). This was originally isolated from infected *Brassica oleracea* (cabbage) seed by E.G. Simmons, designated as EGS 42-002, and deposited at ATCC. The genome was sequenced to a total of 6.4x whole genome coverage. A combination of whole genome shotgun plasmid, fosmid and BAC end sequences were used in the initial unfiltered

assembly. The combined sequence reads were assembled using the PCAP software [104]. A BAC fingerprint map were constructed and compared to the final assembly for further refinement of both the sequencing assembly and the fingerprint map. The physical map of *A. brassicicola* was constructed by generating fingerprints from the CSU-K35 *A. brassicicola* BAC library from Colorado State. The map consists of 4,056 BAC clones. These clones were fingerprinted, analyzed with Image (http://www.sanger.ac.uk/Software/Image) and assembled in an FPC database (http://www.agcol.arizona.edu/software/fpc/). The assembled contigs were manually edited to achieve proper order based upon fingerprint banding patterns and were then merged with one another to generate the largest contigs possible. The physical map covers 172 contigs. This map was built and edited with FPC Version 5.0.

The *A. alternata* genomes (ATCC 66891 and ATCC 11680) were sequenced using GS-FLX 454 technology with 20x whole genome coverage. The genomes were assembled using the Newbler software provided by 454 Life Sciences.

3.2.2 Genome annotation

Genomes were annotated using the *Alternaria* pipeline described in Chapter 2. Gene functions were summarized using gene ontology and KOG annotation. *Alternaria* GO slim set were created by expanding *Apsergillus* GO slim set [105] to cover terms found for *Alternaria genes* that were not mapped to any high level terms in the *Aspergillus* GO slim set. The Go-perl map2slim tool (http://search.cpan.org/~cmungall/go-perl) was used to map GO terms assigned to genes to the higher level GO slim terms used for general function categorization and comparison. The predicted proteins were searched against CDD KOG database using RPSBLAST [107] with e-value $\leq 10^{-5}$ to assign each protein to KOG groups.

3.2.3 Whole genome alignment

Whole genome pairwise alignments were performed using progressive Mauve software [108] with default parameters. The input to Mauve alignment software was contig and supercontig sequences of *Alternaria* genomes. When aligning large contig/supercontig sequences, local alignments were initiated and expanded that resulted

in multiple aligned blocks. Based on the direction (strand) of the alignment, the aligned blocks were classified to one of the two groups: non-inverted blocks (+/+ alignment, plus strand) and inverted blocks (+/- or -/+ alignment, reverse strand). Genomic inversion rates were estimated with the assumption that when aligning two contig/supercontig sequences, the total length of the inverted blocks was always smaller than the total length of the non-inverted blocks. Whole genome identity was estimated as the ratio of the identical bases in all alignment blocks and the total length of the genomic sequences.

3.2.4 Homology analysis

Homologous protein coding genes between genomes were identified by OrthoMCL software [84] using the predicted protein sequences. Two proteins that were reciprocal best BLAST hit of each other (e-value < 10^{-10} and alignment identity > 50%) were connected in a network. This network was then clustered using Markov clustering to identify homologous and paralogous groups of genes/proteins.

3.2.5 Phylogenetic tree construction

Single copy protein families between fungi were identified using orthoMCL, and multiple sequence alignment for all proteins in each family were performed using MUSCLE [109]. Phylogenetic tree was constructed using maximum likelihood method via RAxML 7.3.2 [110] with rapid bootstrapping algorithm and 100 bootstrap replicates. ProteinModelTest was also used to determine that Whelan and Goldman with empirical frequencies [111] was the best model of evolution of the sequences.

3.2.6 Synteny analysis

Syntenic blocks were identified using the OrthoCluster software [112] using orthologous genes identified via OrthoMCL. The parameters chosen for OrthoCluster were minimum of 5 genes per syntenic block and maximum of 20% number of non-homologous genes inside the syntenic blocks.

3.2.7 Carbohydrate active enzyme analysis

Sequence libraries were built with the full length as well as with the constitutive modules (glycoside hydrolase, GH; polysaccharide lyase, PL; carbohydrate esterase, CE; glycosyltransferase, GT; auxiliary activity, AA; and carbohydrate-binding module, CBM) isolated from the collection of carbohydrate-active enzymes of the CAZy database (www.cazy.org) [113]. A series of profile hidden Markov models (HMMs) were built from the module families (and subfamilies in a number of cases) described by the CAZy database. Assignment of the *Alternaria* protein models to a CAZy family (or to several in the case of multimodular proteins) was performed by a two-step procedure that involved first a BLAST [53] search against the full length CAZy proteins, keeping for further analysis all proteins that gave a e-value better (i.e. smaller) than 0.1. The selected proteins were then subjected to a combination of BLAST and HMMER3 [114] searches against the libraries of sequence made with the CAZy modules and the collection of HMM profiles, respectively. All positive results giving significant scores with both BLAST and HMMs were manually inspected, checked for the presence of catalytic residues or sequence motifs characteristic of the family, annotated for possible problems (for instance if the model was a fragment), and assigned to one or several CAZy families depending on the modularity. A heat map of CAZy profile for 17 fungi was analyzed and clustered using R [115]. The heat map included only the GH, PL, CE, AA, and CBM families, because they report on the environment of the fungi while GTs are more housekeeping.

3.2.8 Strains and growth conditions

All strains of 17 selected fungi were grown on malt extract agar (MEA) prior to the growth experiment. Growth profiling was performed on solid media. Media and inoculation time for the different fungi was as follows; *Ceriporiopsis subvermispora* was grown for 2 days on Serpula medium, *Laccaria bicolor* for 17 days on *Rhizoctonia solani* medium, *Schizophyllum commune* for 3 days on *Schizophyllum commune* medium, *Trichoderma reesei* for 1 day, *Myceliophthora thermophila* and *Thielavia terrestris* for 2 days, *Aspergillus fumigatus* for 3 days, *Aa1*, *Aa2*, and *Ab* were grown for 5 days, *Aspergillus niger* (ATCC1015) for 4 days and *Phaeosphaeria nodorum* for 7 days on

Aspergillus niger medium, *Neurospora crassa* was grown for 3 days on *Neurospora crassa* medium, *Podospora anserina* was grown for 2 days on *Podospora* medium and *Magnaporthe oryzae* was grown for 5 days on *Magnaporthe grisea* medium. A list with all media can be found on http://www.fung-growth.org. All strains were grown on 5 ml petri dishes (except that *Neurospora crassa* was grown in tubes) supplemented with different C-sources (25 mM D-glucose, 25 mM D-xylose, 25 mM cellobiose, 1% cellulose, 1% beechwood xylan, 1% guar gum, 1% apple pectin, 1% starch, 1% inulin and 1% lignin) and were grown in the dark at 25°C, except *Myceliophthora thermophila* and *Thielavia terrestris* were grown at 45°C. The growth test was executed in duplicate for each strain. Growth was stopped either if the fungus showed clear visible differences between the C-sources, or when the fungus reached the edges of the plate on one of the C-sources.

3.2.9 Allergen homologs analysis

To quickly survey allergen homologs, a traditional sequence homology-based approach was used to identify allergens in *Alternaria* genomes. Predicted proteins were aligned with known allergens collected from The Food Allergy Research and Resource Program (FARRP) AllergenOnline database version 12 (http://allergenonline.org) using BLAST. This database contained manually curated known allergens from multiple sources including fungi and non-fungi such as plant, bacteria, dustmites, and insects. An e-value of 10^{-10} was used as a cutoff score to identify allergen homologs. Allergen homologs were then classified as homologs of the group of the matched known allergens.

3.2.10 Statistical significance testing

Fisher's exact test was used to calculate the p-value of the statistical significance tests used in this study including GO and KOG enrichment tests. QVALUE software [116] was used to estimate False Discovery Rate (FDR) in multiple hypothesis testing, yielding q-values (adjusted p-values) for the tests, whose values ≤ 0.05 were considered significant.

3.3 Results

3.3.1 Genome sequences and characteristics

Two *A. alternata* allergenic strains, ATCC 66981 (*Aa1*) and ATCC 11680 (*Aa2*), and a *Brassica* pathogen *A. brassicicola* ATCC 96836 (*Ab*) strain were chosen for genome sequencing and subsequent analyses. Although initial results of *A. brassicicola* genome sequencing and annotation was reported previously [117], in depth genome analysis and comparison to other *Alternaria* genomes has not been reported until this time. *Aa1* is the type strain of *A. alternata* distributed by an international *Alternaria* taxonomy expert, E. G. Simmons and was found previously to be the most robust and consistent producer of allergens in vitro among the strains tested previously [118]. *Aa2* is an *A. alternata* commercial allergen production strain (GREER Laboratories, Lenoir, NC). It is noteworthy to mention that antigen/allergen extracts and most recently spore preparations from this strain are widely used in the allergy and immunology research community [119] (references too numerous to list herein).

The *A. alternata* genomes (*Aa1* and *Aa2*) were sequenced using the GS-FLX 454 Titanium technology (Roche-454). Roughly 1.43 million sequencing reads (~20x coverage) were obtained for *Aa1* of which ~1.3 million (~91%) reads were assembled using the Newbler software package (Roche) into 499 contigs (33.2 Mb, N50 ~0.3Mb, 278 contigs longer than 1kb). The longest contig was ~1Mb. The number of reads obtained for *Aa2* was ~1.54 million (~20x coverage) of which ~1.5 million (~98%) reads were assembled into 797 contigs (33.8Mb, N50 ~ 0.45Mb, 230 contigs longer than 1kb). Although physical maps were not obtained for the strains in this study, molecular karyotype analysis of 27 individual isolates previously revealed that *A. alternata* typically harbors 9-11 chromosomes with genome sizes ranging from 29.5-33.1 Mb [120]. Thus the genome sizes obtained from this present study are in basic agreement with karyotype analysis (genome size estimate) data reported previously. The *A. brassicicola* genome sequence was obtained previously using a Sanger-based whole genome shotgun approach resulting in ~6.4x coverage. Following assembly with PCAP software, the partially assembled genome of *Ab* consisted of 4,039 contigs with 3,402 contigs longer than 1kb. In order to provide a more robust assembly, a physical map of *Ab* was constructed by

generating fingerprints from the CSU-K35 *Ab* BAC library, and the contigs were then assembled into 838 supercontigs with N50 of ~2.4Mb. This resulted in an ~30Mb sequence assembly, consistent with the previous estimation of 29.6 Mb genome size of a different *A. brassicicola* isolate (0-263) determined using Pulse-Field Gel Electrophoresis (PFGE) [120]. Approximately 84% of the assembled *Ab* genome (~27 Mb) was distributed on the 11 largest supercontigs (each > 1Mb), which partially agrees with a karyotype estimate of 9 chromosomes [120]. Due to the variable nature of the number of chromosomes present in an individual *Alternaria* isolate regardless of species, it is difficult to make definitive conclusions regarding the number of chromosomes for this particular *A. brassicicola* isolate without additional research in the future. It is possible for example that several of these 11 supercontigs are portions of larger chromosomes. The sequenced genomes of *Aa1* and *Aa2* were larger than the *Ab* sequenced genome (~3-4Mb larger). The *Aa2* genome was slightly larger than the *Aa1* genome (~500kb larger).

Using *Alternaria* pipeline, we predicted 11,635 genes for *Aa1*, 12,323 genes for *Aa2*, and 10,514 genes for *Ab.* These numbers of predicted genes were comparable with those of other filamentous fungi with similar genome sizes such as *S. nodorum* [86], *N. crassa* [85], and *A. fumigatus* [43] (10,762, 10,082 and 9,457 genes, respectively). Both *A. alternata* strains had a slightly higher number of genes than *A. brassicicola* most likely because they had larger genomes and smaller repetitive content as describe hereafter. When searched against the NCBI Genbank non-redundant protein database using BLASTp [53], 11,088 (~95%) predicted proteins of *Aa1*, 11,606 (~94%) of *Aa2*, and 9,733 (~93%) of *Ab* exhibited homology to at least one protein (e-value $\leq 10^{-10}$). When scanned against the Interpro database [71], the numbers of predicted proteins that hit some protein families and domains were 8,728 (~75%), 9,031 (~73%), and 7,484 (~71%) for *Aa1, Aa2,* and *Ab*, respectively (Interpro default criteria). Non-coding RNA genes were also computationally annotated. Interestingly, *Ab* had more tRNA (156 vs. 127 and 155) and rRNA (12 vs. 9 and 11) than *Aa1* and *Aa2*. Table 3-1 summarizes the genome statistics for the three fungi.

Table 3-1. Genome statistics for *Alternaria* species

Statistics	***A. alternata* ATCC 66981 *(Aa1)***	***A. alternata* ATCC 11680 *(Aa2)***	***A. brassicicola* ATCC 96836 *(Ab)***
Total supercontig length	N/A	N/A	31,974,449
Total contig length	33,236,566	33,752,310	30,291,099
GC content (%)	51.11%	51.10%	50.54%
# of supercontigs	N/A	N/A	837
# of contigs	499	797	4,039
Super contig N50 (Mb)	N/A	N/A	2.48
Contig N50 (Mb)	0.304	0.451	0.018
# of predicted genes	11,635	12,323	10,514
Avg. gene length	1,626	1,594	1,542
Avg. transcript length	1,494	1,470	1,354
Longest gene length	29,525	29,513	33,207
Longest transcript length	29,361	29,349	31,557
# of singleton genes	2,765	3,069	1,510
# of genes with intron	8,870	9,254	9,004
# of exons	30,761	31,992	28,273
Avg. exon length	565	566	503
Exon coverage (over genome)	52.31%	53.69%	46.99%
Gene coverage (genic region/genome)	56.91%	58.18%	53.53%
Avg. exons per gene	2.6	2.6	2.7
# of introns	19,126	19,669	17,759
Avg. intron length	80	77	112
Avg. introns per gene	1.6	1.6	1.7
# of rRNA genes	9	11	12
# of tRNA genes	127	155	156

3.3.2 Expansion of repetitive sequences in A. *brassicicola*

Transposable elements (and other types of repeats) in fungal genomes make up a portion from as small as ~1% to as large as ~65% [47, 56, 87, 88]. Although most fungal genomes have relatively small amounts of repetitive DNA, it is thought to play important roles in fungal evolution [90, 91, 121] . Several tools for de novo repeat annotation including REPET [52] and RepeatScout [57] were used to annotate repeats in *Aa1, Aa2, and Ab* genomes. *A. alternata* and *A. brassicicola* both had small but significantly different repetitive content. Only 0.82% of the *Aa1* genome and 0.81% of the *Aa2* genome were predicted transposable elements compared to a much higher rate of 8.60% for the *Ab* genome (Table 3-2). The overrepresentation of transposable elements in *A. brassicicola* compared to *A. alternata* can be possibly explained by the higher evolutionary stress on *A. brassicicola* due to its necrotrophic pathogenic lifestyle. These

particular isolates of *A. alternata* are thought to be purely saprophytic capable of efficiently utilizing multiple sources of environmental nutrition and therefore possibly encountering less environmental stress. The larger transposable element content is also a possible explanation for the higher genome rearrangement rate in *A. brassicicola* observed in this study (discussed in more detail later).

Table 3-2. Transposable elements predicted for *Alternaria* genomes

		Aa1			***Aa2***			***Ab***		
Class		Total bases covered	Percent of repetitive DNA	Percent of genome	Total bases covered	Percent of repetitive DNA	Percent of genome	Total bases covered	Percent of repetitive DNA	Percent of genome
Class I	Ty1-Copia	3,039	1.11	0.01	-	-	-	30,382	1.17	0.10
	Ty3-Gypsy	82,411	30.22	0.25	40,598	12.15	0.12	1,455,489	55.84	4.81
	Non-LTR retrotransposon	-	-	-	-	-	-	2,914	0.11	0.01
	Unclassified	-	-	-	-	-	-	35,112	1.35	0.12
Class II	DDE_1	74,562	27.34	0.22	-	-	-	842,694	32.33	2.78
	hAT	6,214	2.28	0.02	14,321	4.28	0.04	72,328	2.78	0.24
	Tc1/Mariner	4,470	1.64	0.01	84,127	25.17	0.25	12,437	0.48	0.04
	MITEs	2,150	0.79	0.01	7,063	2.11	0.02	-	-	-
	Helicase containing	-	-	-	11,044	3.30	0.03	-	-	-
Unclassified		99,893	36.63	0.30	114,018	34.11	0.34	154,991	5.95	0.51
Total		**272,739**		**0.82**	**334,265**		**0.81**	**2,606,347**		**8.60**

* Repetitive sequence classification was performed by Braham Dhillon at the University of British Columbia.

The majority of the transposable elements from the three *Alternaria* genomes were class I Ty3-Gypsy (~30%, ~12%, and ~56% for *Aa1, Aa2,* and *Ab,* respectively) and class II DDE_1 (~27% in *Aa1* and ~32% in *Ab*). Some MITE elements were detected in *A. alternata* strains but not in *A. brassicicola* and some non-LTR retrotransposons were detected in *A. brassicicola* but not in *A. alternata*. Some minor differences to note between the two *A. alternata* strains were that the DDE_1 element was missing in *Aa2* while helicase-containing repeat was missing in *Aa1*. Many more transposable elements from *A. alternata* were unclassified (~37% and ~34% for *Aa1, Aa2*) when compared to that of *Ab* (~6%). Most of the *A. brassicicola* repeats were clustered together in regions of dense repeat contents (Figure B-1). These are possible centromeric and telomeric regions of *Ab* chromosomes. Telomeric regions in eukaryotic genomes in general usually possess simple repeats while centromeres often contain more transposable elements

[122]. For example, transposable elements clustering at centromeres have been observed in other fungal genomes [123].

3.3.3 Overview of gene function

In order to provide a general overview of predicted gene function in the *Alternaria* genomes, Eukaryote Ortholog Group (KOG) annotation was assigned. KOG annotation for the predicted proteins were performed using RPSBLAST search against the KOG database [124]. Approximately 6,179 (53%), 6,265 (51%), and 5,406 (51%) predicted proteins of *Aa1, Aa2,* and *Ab*, respectively, were able to be assigned to at least one KOG group. High level KOG classification for both *A. alternata* genomes showed very similar compositions in most functional categories when compared with *A. brassicicola*, except for the significant overrepresentation of genes associated with secondary metabolites/lipid/carbohydrate transport and metabolism (categories Q, G, I) (Figure 3-4). The overrepresentation of metabolism-related genes is a possible explanation for the successful saprophytic lifestyle of *A. alternata*.

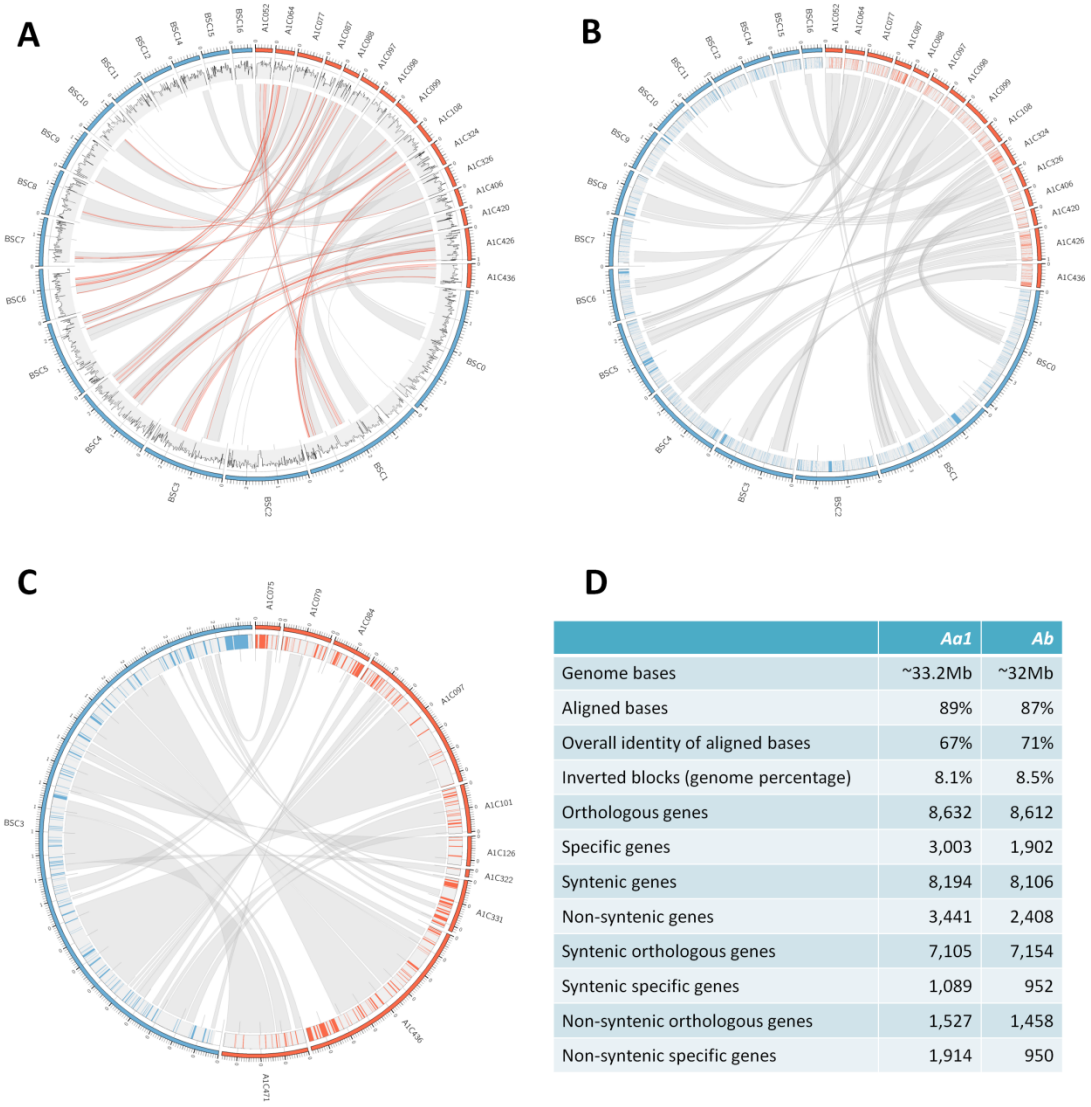
Using Blast2GO [74] and Interpro [71], 24,529 gene-GO term unique assignments were made for *Ab* that included 2,103 GO terms and 6,447 genes (~61%). For *Aa1*, 29,680 gene-GO term pairs were assigned for 7,093 genes (~61%) and 2,170 GO terms; and for *Aa2*, 30,523 assignments were made for 7,319 (~59%) genes and 2,232 GO terms. The ratios of total genes that were found to match with at least one GO term (and also previously with KOG) were roughly equal among the three fungi. This indicates the amounts of annotation information that we obtained via public databases for the three fungi are equal which made our comparative genomics approaches reliable despite the fact that different technologies were used to sequence *A. alternata* and *A. brassicicola* genomes to different coverage levels.

GO annotation, when mapped to the high level GO slim terms, also highlighted the general functional characteristics of the predicted proteins in *Alternaria* genomes. *Alternaria* GO slim subset were created by expanding the GO slim subset developed previously for Aspergillus [125] to include ten more high level terms. These terms were

ancestors of most of the GO terms that we were unable to map to the higher level *Aspergillus* GO slim terms other than the three top level terms (biological process, molecular function, and cellular component). The top GO slim categories (assigned to more than 15% the genes) for all species (Figure B-2, Figure B-3, Figure B-4) were metabolic process (GO:0008152), membrane (GO:0016020), nucleotide binding (GO:0000166), ion binding (GO:0043167), oxidoreductase activity (GO:0016491), and hydrolase activity (GO:0016787). The functional classifications of the two *A. alternata* and *A. brassicicola* genomes were very similar at high level GO terms in each category of biological process, molecular function, and cellular component. The *A. brassicicola* genome had more genes mapped to DNA, RNA, nucleotide, and nucleic acid bindings, while both *A. alternata* genomes had slightly higher numbers of genes mapped to GO terms related to metabolism. These data support the KOG annotation.

A more concrete functional view of these genomes were next obtained by searching the predicted proteins against the PFAM database (release 26) [126] using HMMER3 [114]. The total number of proteins that matched each PFAM domain was counted for each of the three fungi. One striking result of this analysis was that the *A. alternata* strains have dramatically more predicted proteins that possessed Heterokaryon domains (PF06985) than *A. brassicicola* (Figure B-8). Heterokaryon incompatibility genes are known to prevent hyphal fusion of genetically different filamentous fungi [127]. Interestingly, PFAM domain analysis revealed dipeptidyl peptidase family (X-Pro dipeptidyl-peptidase) was missing in *A. brassicicola* while the two *A. alternata* genomes each had 6 predicted copies of this gene.

3.3.4 Genome rearrangement



	Aa1	*Ab*
Genome bases	~33.2Mb	~32Mb
Aligned bases	89%	87%
Overall identity of aligned bases	67%	71%
Inverted blocks (genome percentage)	8.1%	8.5%
Orthologous genes	8,632	8,612
Specific genes	3,003	1,902
Syntenic genes	8,194	8,106
Non-syntenic genes	3,441	2,408
Syntenic orthologous genes	7,105	7,154
Syntenic specific genes	1,089	952
Non-syntenic orthologous genes	1,527	1,458
Non-syntenic specific genes	1,914	950

Figure 3-1. Whole genome alignment and synteny between *Aa1* and *Ab*. In the circos plots, the outer ring displays genomic sequences with *Aa1* contigs in red and *Ab* supercontigs in blue. (A) Whole genome alignment. The ribbons link aligned blocks (grey indicates putative non-inverted aligned block and red indicates putative inverted aligned block). The grey histogram shows percent similarity between aligned blocks. Only contigs/supercontigs longer than 500kb and alignment blocks longer than 1kb are displayed. (B) Syntenic regions between *Aa1* and *Ab* sequences that are at least 500kb. The inner ring shows orthologous genes (grey), *Aa1* specific genes (red), and *Ab* specific genes (blue). Syntenic regions are connected by the grey ribbons. (C) Syntenic regions originated from Ab supercontig BSC3. (D) Summary statistics for whole genome alignment and synteny analysis.

Whole genome pairwise alignment demonstrated that *A. alternata* and *A. brassicicola* genome had significant levels of variation. Alignment between *Aa1* contigs

and *Ab* supercontigs showed many genomic insertion/deletion blocks (Figure 3-1A and Table A-2). Total gap length was ~4.2Mb (12.5% of total alignment length) for *Aa1* and ~ 5.9Mb (17.4% of total alignment length) for *Ab*. This suggests that significant genome insertions and deletions have occurred in both genomes. Genome inversion was also significant and accounted for ~2.8Mb (8.5%) of *Aa1* and ~2.6Mb (8.1%) of *Ab* aligned bases, as demonstrated by the putative inverted alignment blocks. Approximately 59% of *Aa1* genomic DNA was identical to 61% of *Ab* genomic DNA (Table A-1). A very similar result was obtained from *Aa2* and *Ab* pairwise whole genome alignment. *Aa1* and *Aa2* genomes were very similar at the DNA level according to whole genome pairwise alignment with 93% bases of *Aa1* were identical to 92% bases of *Aa2*.

Gene order was generally conserved between *A. alternata* and *A. brassicicola* as demonstrated by microsynteny analysis between *Aa1* contigs and *Ab* supercontigs. We identified 389 syntenic regions between *Aa1* and *Ab* genomes. The majority of orthologous genes between the two species were found in these syntenic regions (~94% for each species). In comparison with specific genes identified from the homology analysis between *Aa1* and *Ab* (Figure 3-1B, C, D and Table A-2), *Ab* had a larger percentage of specific genes that were located in syntenic regions (50% vs. 36% in *Aa1*), suggesting either *Ab* genome had rearranged at a higher rate or the majority of gene acquisition in *Aa1* had occurred outside of the syntenic regions. A comparison of KOG annotation between non-syntenic genes from the two genomes showed an overrepresentation of several functional categories. For example, genes annotated with carbohydrate and lipid transport and metabolism (categories G and I), and secondary metabolites synthesis/transport/catabolism (category M) were enriched in *Aa1* non-syntenic regions. PFAM comparison showed an overrepresentation of heterokaryon domain (PF06985), NmrA-like family (PF05368), and NACHT domain (PF05729) in *Aa1* non-syntenic regions. The complete lists of PFAM domains and KOG groups that are significantly enriched/depleted are presented in Table A-3 and Table A-4.

3.3.5 Expansion of carbohydrate active enzymes (CAZY) especially in *A. alternata*

We examined the evolutionary relationship and CAZY contents of 17 selected filamentous fungi of different lifestyles (saprotroph, nectrotroph, biotroph, and hemibiotroph), including the three *Alternaria* species and other human and plant pathogenic and non-pathogenic molds and mushrooms from dothidiomycetes, sordariomycetes, and other classes. Carbohydrate active enzymes were annotated according to the CAZY database [80]. The phylogenetic tree was constructed using 100 protein families randomly selected from 1,486 single copy protein families computationally identified as orthologs across 17 fungi.

With over 280 GHs, *Aa1* and *Aa2* had the largest sets of glycoside hydrolase (GH) genes among the 17 examined fungi (Figure 3-2, Table A-5). *Ab* on the other hand had only ~240 GH genes. A similar trend was also observed for carbohydrate esterases (CEs), with *Aa1* and *Aa2* having more CE genes (57 genes) than the other fungi while with 43 genes *Ab* had fewer CE genes than *Magnaporthe oryzae* (52 genes) and *Phaeosphaeria nodorum* (49 genes) (Figure 3-2). For polysaccharide lyases (PLs) the three *Alternaria* fungi had the largest number of PL genes than any other examined fungi (each *Alternaria* species had 24 genes). Finally the two *A. alternata* genomes had more than 80 predicted enzyme genes of the recently described auxiliary activity (AA) category that groups together oxidoreductase enzymes such as ligninases and lytic polysaccharide monooxygenases that assist the breakdown of lignocellulosic substrates. Only *M. oryzea* had an equally rich oxidoreductase complement whereas *A. brassicicola* had a significantly lower number of AA (54 genes). Overall the total number of carbohydrate active enzymes (i.e. GHs, PLs, CEs, and AAs) for *Aa1* and *Aa2* approached 450 genes, addressing an exceptional range of possible substrates and providing a possible explanation that *A. alternata* is one of the most successful saprophytes in the fungal kingdom.

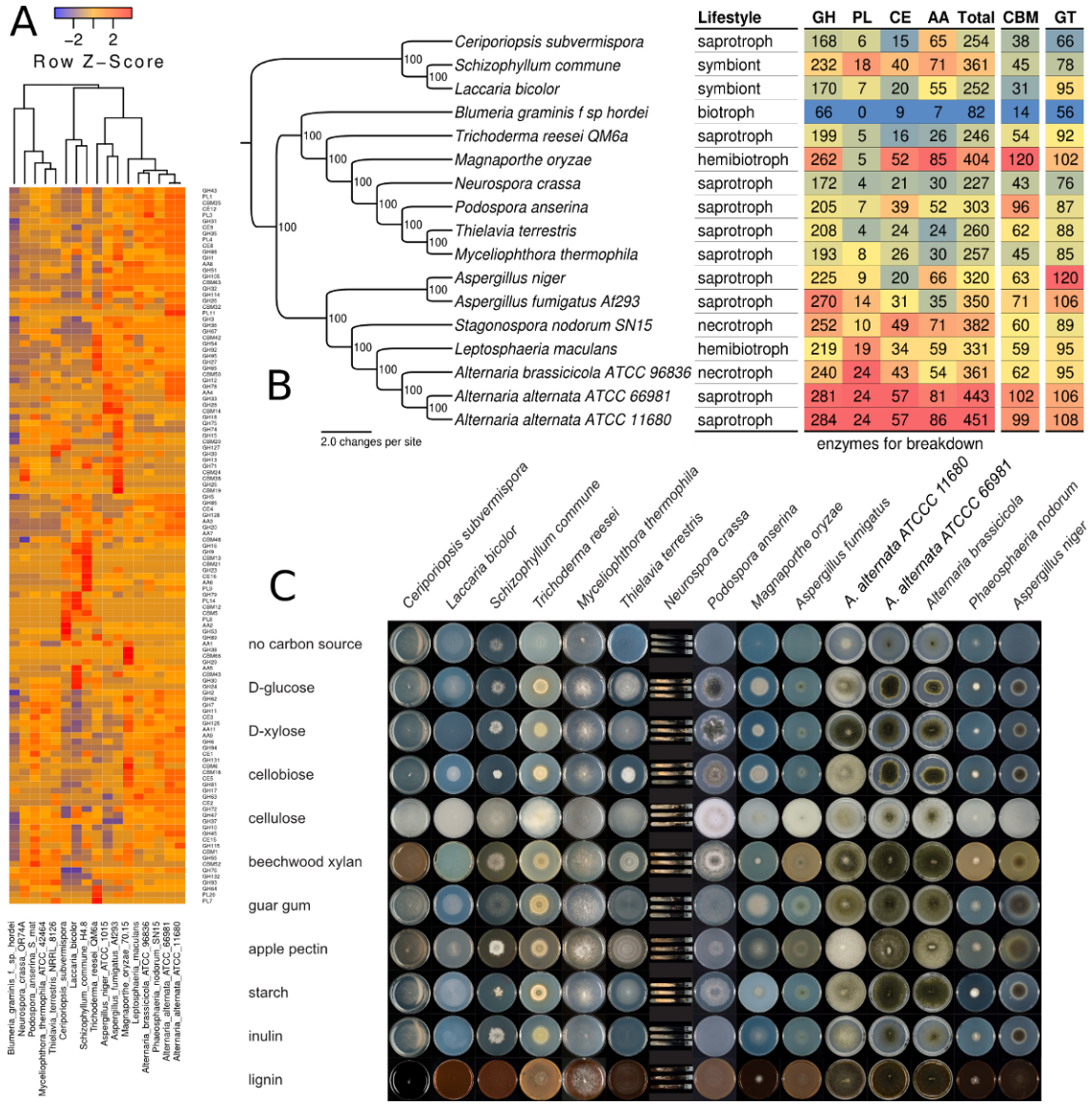


Figure 3-2. Phylogeny and carbohydrate active enzyme (CAZY) contents of 17 selected fungi of different lifestyles. (A) Clustering of CAZY profiles; (B) Phylogeny and the breakdown of CAZY; (C) Growth of fungi on different carbohydrate sources. The tree was built from the alignment of 100 random single-copy protein families using maximum likelihood method with RAxML. The number of carbohydrate active enzymes includes Polysaccharide Lyases (PL), Carbohydrate Esterases (CE), and Glycoside Hydrolases (GH). CAZY annotation was performed by Bernard Henrissat at French National Centre for Scientific Research, France. Phylogeny was constructed by Jason Stajich at the Univerisity of California – Riverside. Growth profiling was performed by Ronald de Vries and Eline Majoor at CBS-KNAW Fungal Biodiversity Centre, The Netherlands.

Genome based phylogeny agreed with the known evolutionary relationships among the examined fungi (Figure 3-2). The dothideomycete group (three *Alternaria* fungi, *L. maculans*, and *S. nodorum*) was clustered in a clade next to another clade that

included *A. fumigatus* and *A. niger*. Among other fungi, *L. maculans* was the closest species to *Alternaria*. A heat map built with the gene count in each family of catabolic carbohydrate-active enzymes (i.e. individual families of GHs, of PLs, of CEs, and of AAs), supplemented by the encoded carbohydrate-binding modules (CBMs) families revealed a clustering of the fungi according to their broad taxonomical divisions. Interestingly, *A. brassicicola* was placed in the same clade with *L. maculans*, while the two *A. alternata* strains were grouped in a different clade together with the wheat pathogen *S. nodorum* (Figure 3-2A). The families of CAZY enzymes group together enzymes of different specificity, and therefore the CAZY family profile of the fungi is shaped by two main (and perhaps opposite) evolutionary forces: taxonomy and specialization to environment (i.e. lifestyle and host). The former drives the global numbers of CAZY and the latter affected the composition of CAZY families that were represented in the heat map. Interestingly, there are a few CAZY families found in *A. brassicicola* but absent in both *A. alternata* strains, including GH29 (alpha L-fucosidase, *Ab/Lm specific*, 1 copy*)* and CBM24 (alpha-1,3-glucan-binding, 2 copies). In addition, there are several families found in the two *A. alternata* strains but absent in *A. brassicicola:* GH33 (sialidase or neuraminidase, 1 copy), GH71 (alpha-1,3-glucanase), PL11 (rhamnogalacturonan lyase, 1 copy), CBM32 (Binding to galactose and lactose, polygalacturonic acid, LacNAc, 2 copies), GH 71 (α-1,3-glucanase, 1 copy), and AA4 (vanillyl-alcohol oxidase, 1 copy). The GH 78 (α-L-rhamnosidase) had 7 copies in both *A. alternata* strains but only 1 copy in *A. brassicicola*. Several other noteworthy differences lie in other carbohydrate binding module family proteins especially chitin binding. For example, both *A. alternata* strains had over 40 copies of CBM 18 while *A. brassicicola* only had 25 copies.

3.3.6 *A. alternata* grows well on various substrates including cellulose and lignin

While it is not possible to directly compare two fungi on the same carbon source due to differences in species (strain)-specific growth rate and morphology, it is possible to compare relative growth to an internal reference. As D-glucose results in fastest growth of nearly all fungi compared to other monosaccharides, this substrate was used as

an internal reference. In the comparisons described below the relative growth on any substrate compared to growth on glucose was compared between the species/strains.

All three *Alternaria* strains grew well on most monomeric, oligomeric and polymeric carbon sources, but growth was reduced for all three strains on D-galacturonic acid and for *Aa1* and *Aa2* on D-galactose (Figure B-5). All three strains appeared to have a broad ability to degrade plant polysaccharides, suggesting that they are able to produce a wide range of enzymes acting on these substrates. Also in comparison to the other tested fungi, their growth on polysaccharides is equal or better (Figure 3-2C). Their growth profile was similar to that of *Aspergillus niger*, well-studied fungus that produces a broad range of enzymes acting on a variety of polysaccharides. The main differences with *A. niger* were weaker growth on inulin and substantially stronger growth on cellulose. The capability of *A. alternata* and *A. brassicicola* to grow well on cellulose is worth mentioning as most fungi grow poorly on cellulose as the sole carbon source. This is supported by a relatively high number of cellulase-encoding genes in their genomes (12 in families GH6, 7, 12 and 45 together compared to 6 in *T. reesei*). Also growth on the other polysaccharides can be explained by their genome content with 16-21 xylan-specific genes (GH10, 11, 62, 67, 115; CE15), 7-10 galactomannan-specific genes (GH26, 27, 36) and 47-56 pectin-specific genes (GH28, 53, 78, 88, 105; PL 1, 3, 4, 9, 11; CE8, 12), which are among the highest numbers of tested fungi in this study. This is complemented with high numbers of genes in GH5 (includes cellulose- and galactomannan-related genes) and GH43 (includes xylan- and pectin-related genes). Previous studies also demonstrated strong correlations between genome content and the ability to degrade plant polysaccharides [128–134]. The results obtained here indicate that the three *Alternaria* strains, and in particular *Aa1* and *Aa2*, have a high potential as sources of industrial enzymes and enzyme cocktails as their range of enzymes is similarly broad as the industrial workhorse *A. niger*. Growth on lignin was second only to *M. thermophila* and correlated with strong growth on cottonseed hulls (Figure B-5), which are rich in pectin. No clear explanation can be derived from the genomes for this phenomenon besides the expansion of AA families in *Alternaria* genomes. It cannot also be excluded that this reflects tolerance to lignin while growth is supported by carbohydrate and other impurities in the lignin.

3.3.7 Proteolytic enzyme content

Peptidases (also called proteinases or proteases) play important roles in fungal metabolism as well as pathogenicity. An examination of the peptidase-encoding gene content of the three *Alternaria* genomes showed that *A. alternata* genomes had many more predicted peptidase encoding genes than *A. brassicicola* (Figure B-6). The numbers of peptidases that were identified for *Aa1, Aa2,* and *Ab* were 456, 461, and 372, respectively. Serine proteases contributed the most to the overrepresentation of peptidases in *A. alternata*. *A. alternata* possesed 60-70 species specific serine proteases when compared with *A. brassicicola*. Many serine proteases have been known to be allergens in *Aspergillus* and *Penicillium* species [135]. A recent study showed that *A. alternata* serine protease activities induced IL-33 that contributed to the development of T_H2 inflammation in mice [136]. Therefore, this result suggests that serine protease overrepresentation may contribute to the allergenicity of *A. alternata.* On the other hand, *A. brassicicola* only harbored a small number of serine proteases. Among those serine proteases, Xaa-Pro dipeptidyl-peptidase (S15) was found to have 7 copies in both *A. alternata* (6 of them were also identified by PFAM analysis in the previous section) while none was found in *A. brassicicola*. These Xaa-Pro dipeptidyl-peptidases are possible allergen candidates for more in depth studies. Dipeptidyl peptidase has also been found to play an important role in *Aspergillus* infection [137]. The over-abundance of X-Pro dipetidyl-petidases in *A. alternata* when compared with *A. brassicicola* suggests that *A. alternata* has acquired important genes that may contribute to its allergenicity or its ability to be an opportunistic human pathogen in rare instances.

A few other peptidase families that were found to have at least five or more copies in *Aa1* when compared with *Ab* were endopeptidase subtilisin and its homologs (S08), serine-dependent peptidases (S09), D-Ala-D-Ala carboxypeptidases (S12), exopeptidases that act at the N-terminus of peptides (S33), and endopeptidases that lyse bacterial cell wall peptidoglycans (M23). A similar result was obtained when comparing the species-specific putative peptidases between *Aa1* and *Ab*, with *Aa1* has many more peptidases as well as serine proteases (Figure B-7).

3.3.8 Expansion of allergen homologs in *A. alternata*

Allergen homologs for the three *Alternaria* fungi were identified by comparing the predicted proteins with known allergens primarily from the *AllergenOnline* database version 11 [138] using BLAST. *AllergenOnline* included proteins that have peer-reviewed published clinical evidence of IgE binding using human sera from clinically defined subjects that were known to be allergic to the source of the proteins. Those proteins were classified in two groups: allergen (evidence for IgE binding and biological activities such as basophil activation or histamine release, skin test reactivity or challenge test reactivity using subjects allergic to the source); and putative allergen (missing evidence of biological activities). *AllergenOnline* version 11 contained approximately 1,491 allergens and putative allergens (18 source types, 165 species/organisms including fungi, insect, plant, food, etc., semi-non redundant type set of 553 groups of similar sequences). The two *A. alternata* genomes were found to have more predicted allergen homolog protein coding genes than *A. brassicicola* and the known allergenic fungus, *A. fumigatus* (*Af*). *Aa1* and *Aa2* had a total of 288 and 292 predicted allergen homologs, while *Ab* and *Af* only had 219 and 233 predicted allergen homologs, respectively (Table 3-3). *Af* has been known as one of the most successful allergic fungi being both primary and opportunistic human pathogens. This result suggests that the genus *Alternaria* in general has high allergenic potential particularly during the spore germination process. Many more predicted allergen homologs found in *A. alternata* genomes were also predicted to be secreted proteins when compared with *A. brassicicola* and *A. fumigatus* (Table 3-3).

Moreover, all surveyed fungi have high numbers of proteins homologous to known allergens in various groups including fungal allergens and other allergens from plant, bacteria, dust mite, animal, and insect. This homology-based survey result suggests that fungi in general possess the capability to produce many types of potential allergens originally found in other organisms and bio-sources. We speculate that long term exposure to high levels of fungal spores, especially *Alternaria*, may result in sensitization of humans to allergens from other organisms or sources such as pollens, insects, or dust mites.

Table 3-3. Allergen homologs in three *Alternaria* genomes in comparison with *Aspergillus fumigatus (Af)*

Types	**Aa1**	**Aa2**	**Ab**	**Af**
Total	288 (99)	292 (112)	219 (69)	233 (79)
Aero Fungi	143 (30)	145 (32)	125 (22)	136 (32)
Aero Plant	73 (24)	76 (30)	54 (23)	63 (22)
Contact	72 (30)	72 (34)	51 (15)	40 (8)
Food Plant	21 (13)	23 (12)	16 (8)	27 (17)
Aero Mite	16 (7)	16 (8)	9 (4)	10 (4)
Bacteria airway	12 (8)	11 (8)	5 (2)	3 (2)
Venom or Salivary	12 (3)	12 (4)	12 (4)	10 (1)
Aero Animal	8 (5)	8 (5)	6 (4)	7 (6)
Aero Insect	4 (2)	4 (2)	3 (2)	2 (1)
Worm (parasite)	1 (0)	1 (0)	1 (0)	1 (0)
Food Animal	0 (0)	0 (0)	0 (0)	1 (0)
Unassigned	1 (0)	1 (0)	1 (0)	1 (0)

* number of secreted proteins are in brackets

3.3.9 Homology analysis: Specific genes explain the pathogenicity and saprophyte of *A. brassicicola* and *A. alternata*

Homology analysis suggested that species specific protein coding genes may play important roles in pathogenicity of *A. brassicicola* and and the ubiquitous saprophytic lifestyle of *A. alternata*. Four dothideomycetes fungi were chosen for homology analysis, including *Aa1, Ab, L. maculans* (*Lm,* another brassica pathogen)*,* and *S. nodorum* (*Sn,* a wheat pathogen). The analysis identified 6,402 orthologous groups among four fungi (Figure 3-3). *Aa1* had 1,690 specific gene families and *Ab* had 1,589 specific gene families.

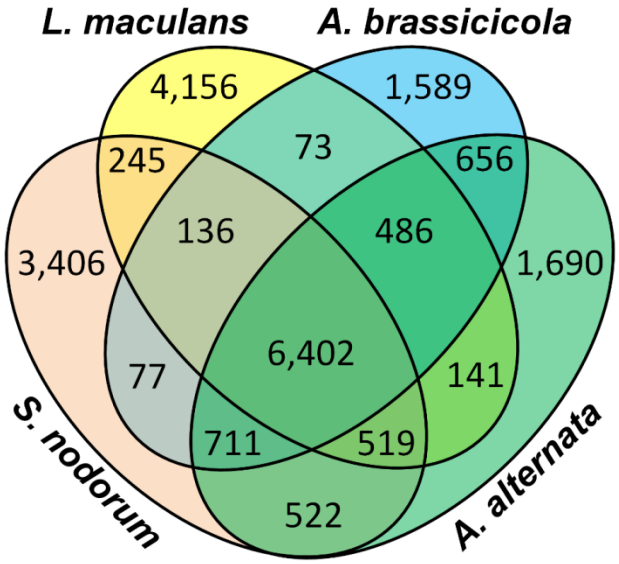


Figure 3-3. Ortholog groups between *Aa1, Ab, S. nodorum*, and *L. maculans (Lm)*. Ortholog groups were inferred by OrthoMCL with sequence identity ≥ 50% and e-value ≤ 10^{-10}.

We found 73 gene families that were specific to two brassica pathogens *A. brassicicola* and *L. maculans* (*Ab/Lm* specific proteins, Figure 3-3, Table A-6). Most of these gene families had unknown function. Two families (OG_109797, OG_109841) were annotated as major facilitator superfamily (MFS) transporters. One family (OG_109785) was annotated with glutathione-dependent formaldehyde-activating enzyme (GFA). Three CAZY enzymes were also among the *Ab/Lm* specific proteins, which included a predicted secreted alpha-L-fucosidase (glycoside hydolase family 29, OG_109875), a glycosyl transferase family 90 (OG_109875), and a predicted sercreted glycoside hydrolase family 105 - GH105 (OG_109769). Two peptidase families (OG_100044 and OG_109832) were also identified among *Ab*/*Lm* specific proteins. Several proteins in the OG_100044 family were found to hit an aspartic peptidase domain, while the OG_109832 family was annotated with G1 unassigned peptidases and was predicted to be secreted. Most of the *Ab/Lm* specific genes mentioned above were found to be highly expressed in *Ab* (top 10% expressed genes, FPKM > 170), according to results from Srivastava et al. [139]). GH105 was among the top 100 highly expressed genes in *Ab* with FPKM > 1,000 for both wild type and a virulence factor *Abvf19* mutant.

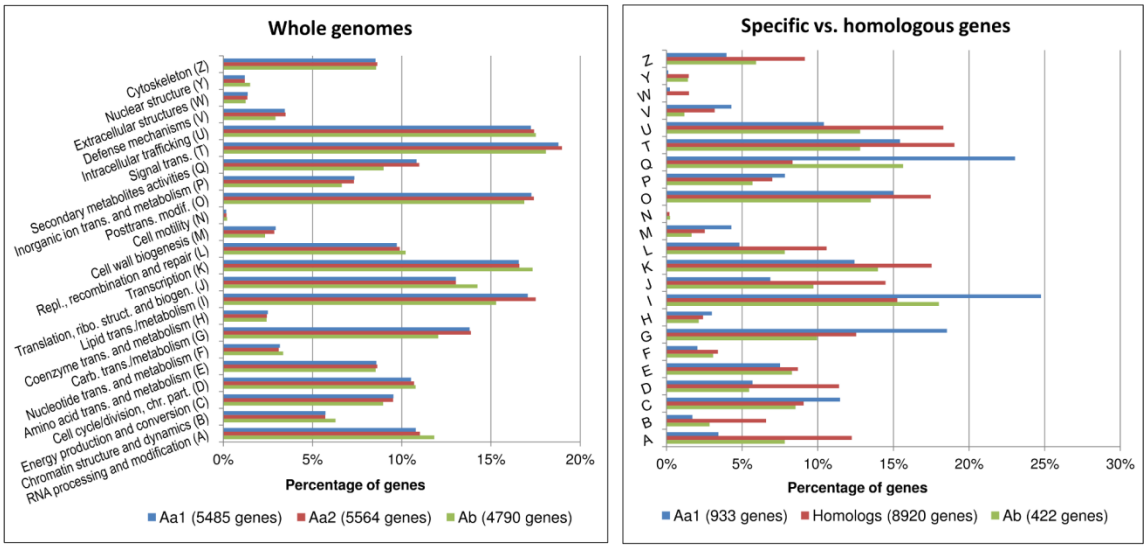


Figure 3-4. KOG comparison between *Alternaria* whole genomes, and between species specific and homologous genes. A protein is annotated with a KOG group if it hits the KOG group profile (RPSBLAST with e-value ≤ 1e-5). The bars indicate the percentages of KOG annotated proteins that belong to high level KOG groups. Hits to poorly characterized group (general function and unknown groups) are not included.

Results of four-species ortholog analysis were also used to infer orthologous and specific genes between *Aa1* and *Ab*. The two species shared 8,255 ortholog groups. *Aa1* had 2,872 specific gene families (3,003 genes) and *Ab* had 1,875 specific gene families (1,902 genes). Gene ontology analysis between the unique genes of the two species showed that *Ab* possessed many more specific genes related to nucleotide binding while *Aa1* had more specific genes that facilitate metabolism (Table A-8, Table A-9). For example, *Ab* specific genes were enriched in GO:0003676 (nucleic acid binding), GO:0003723 (RNA binding) while *Aa1* specific genes were enriched in GO:0008152 (metabolic process), GO:0016491 (oxidoreductase activity), GO:0016787 (hydrolase activity).

KOG comparison between *Aa1* specific genes, *Ab* specific genes, and *Aa1/Ab* orthologous genes revealed that *Aa1* had greater gene expansion in many categories, including secondary metabolites, carbohydrate and lipid metabolism that explained for the capability of *Aa1* in digesting nutrition from various sources in the environment (Figure 3-4). For example, 11% (1,119 genes) of the KOG characterized *Aa1/Ab* homologous genes were annotated with carbohydrate transport and metabolism (category G), while this percentage of *Aa1* specific genes was 16% (173 genes, increased) and that

of *Ab* specific genes was 9% (42 genes, reduced). This result agrees with CAZY analysis. For lipid transport and metabolism (category L), the ratios of KOG characterized genes for the homologous gene set and *Aa1, Ab* specific gene sets were 14% (1,361 genes), 22% (231 genes), and 16% (76 genes), respectively. This trend was also confirmed by gene ontology, PFAM, and CAZY analyses. This result again suggests that *Aa1* has acquired many more genes related to metabolism. Interestingly in several KOG categories, *Aa1* showed gene expansion, while, in *Ab,* gene reduction or unchanged was observed. These categories included carbohydrate transport and metabolism, defense mechanism, cell wall biogenesis - inorganic ion transport and metabolism, and energy production and conversion (Table A-11).

3.3.10 Secretome reveals saprophytic and pathological differences between *A. alternata* and *A. brassicicola*

Secreted proteins are very important in basic aspects of fungal metabolism and pathogenicity. Using SignalP [76], WoLF-psort [77] and Phobius [78], we predicted 1,085 secreted proteins for *Aa1*, 1,134 for *Aa2*, and 748 for *Ab*. Both *Aa1* and *Aa2* had significantly higher (~1.5 times) number of proteins predicted to be secreted than *Ab*. Most of the secreted proteins in three fungi were associated with GO terms related to metabolic functions/processes (Figure B-9). KOG annotation also showed very similar classification for *Aa1, Aa2,* and *Ab* with the only exception that *Ab* predicted secretome was overrepresented in cytoskeleton (Figure B-11).

We assigned the predicted secreted proteins to different groups of potential pathogenicity based on annotation data obtained from carbohydrate active enzymes (CAZY), peptidases (MEROPs), BLAST search against Swissprot and Genbank, and Interpro. Many secreted proteins from *A. alternata* species were chitin-related proteins, LysM containing proteins, secondary metabolites, peptidases, and glucose-methanol-choline oxidoreductases (Table 3-4). *A. brassicicola* had an equal number of secreted pectase lyases/pectin esterases and cytochrome P450 regardless of a much smaller number of predicted secreted proteins.

Table 3-4. Selected protein families identified by Interpro scan on three *Alternaria* genomes

Group	Whole genome			Secretome			Species specific genes (Aa1 vs. Ab)	
	Aa1	***Aa2***	***Ab***	***Aa1***	***Aa2***	***Ab***	***Aa1***	***Ab***
G-alpha	4	4	3	0	0	0	1	0
Hydrophobin	4	3	2	3	2	1	2	0
CFEM	13	12	7	6	6	4	4	0
Chitin binding/chitinase/chitin synthesis	48	47	35	24	25	13	15	5
Cytochrome P450	124	131	93	6	4	7	53	25
Cutinase	14	14	9	11	11	8	5	1
Glucose-methanol-choline oxidoreductase	43	43	26	25	28	13	21	6
Lipase	73	79	57	16	15	11	14	5
LysM	13	15	9	8	11	2	6	1
Pectate Lyase	37	37	36	32	32	31	7	5
Rhodopsin	2	2	3	0	0	0	0	1
Secondary metabolites	185	190	125	24	27	13	62	18
Subtilisin	21	21	13	12	12	5	10	3
Tannase/Feruroyl Esterase	9	8	3	4	4	2	5	1
Transcription factor	154	179	118	0	0	0	39	8
Toxin	13	15	8	4	4	2	5	1
Peptidase (MEROPs)	456	462	372	105	105	76	55	25
Total proteins annotated with Interpro	8,728	9,031	7,484	763	789	528	1,842	769

3.3.11 Survey of pathologically important gene families

Genes related to pathogenicity are important targets for further experimental studies. Using knowledge about plant and human fungal pathogens, we summarized important pathogenic genes and categories for the two *A. alternata* genomes and the *A. brassicicola* genome using PFAM annotation of the predicted proteins (Table 3-4). Overall, *A. alternata* had more genes in almost all of the categories. Several noticeable gene families and protein domains that were enriched in both *A. alternata* strains were CFEM, CP450, feruloyl esterase, chitin-related, glucose-methanol-choline oxidoreductase, LysM, and secondary metabolites. CFEM domain is an eight-cysteine-containing domain. Several CFEM-containing proteins were proposed to have important roles in fungal pathogenesis [140].

3.3.12 Polyketide synthases

Polyketides are one of the most important secondary metabolite classes produced by fungi and bacteria [141]. Fungal polyketides are most often synthesized by type I polyketide synthases (PKS), which are multidomain enzymes closely related to eukaryotic fatty acid synthases. Using SMURF [81] and BLAST search against known PKSs, we identified 11 PKS proteins for *Ab*, 13 PKS proteins for *Aa1* (10 of them were also previously identified using our older predicted gene models [102]), and 15 PKS proteins for *Aa2* (Table 3-5). The domain architecture of these PKS proteins were annotated (Figure B-12, Figure B-13, Figure B-14). For historical reasons, we named PKS genes using latin letters from A to O (PksA to PksO). The basic essential domains required for PKS were ketoacyl CoA synthase (KS), acyltransferase (AT) and acyl carrier (ACP). Those domains were found in most of the identified PKS proteins, except for PksE (similar to plant type III polyketide synthase [102]), PksK (a PKS-NRPS hybrid, missing AT domain), PksL (missing KS domain), and PksM (missing AT domain), *Ab* PksD (missing ACP domain), *Ab* PksP (missing both KS and AT domains), *Ab* PksR (missing ACP domain), *Ab* PksT (missing KS and ACP domains). Other domains not required but often found in PKS proteins included ketoreductase (KR), dehydratase (DH) and enoyl reductase (ER). Many of the identified *Alternaria* PKSs also had these domains. PksB (found in both *Aa1* and *Aa2* but not in *Ab*) was a large predicted protein (~4,000 amino acids) that possessed domains typically found in both PKS and NRPS (KS, AT, AA, CD), and therefore it was classified as PKS-NRPS hybrid protein.

Since the PKS proteins are generally conserved because of domain structures, we used a stringent criterion (80% sequence identity in a BLAST alignment) to define PKS orthologs between *Alternaria* species. PksR and PksT were identified as specific PKSs to *Ab*, and PksR was found previously to contribute to *Ab* pathogenicity. For example, the *Ab* PksR knocked-out mutant showed 50% reduction in virulence on green cabbage [142]. Other PKSs that were specific to *A. alternata* (*Aa1* and *Aa2*) were PksK, PksL, and PksM. These PKSs may be among the potential candidate genes for further pathogenicity studies. *AbPksA* deletion mutants were albino mutants that did not produce melanin but was not a pathogenicity factor (Kim et al., unpublished).

Table 3-5. PKS and NRPS genes in *Alternaria* species and their orthologs in *C. heterostrophus (Ch)*

Ab	***Aa1***	***Aa2***	***Ch***	**Notes**
PksA	PksA	PksA	PKS18	named AbPKS7 in Kim et al.2009 [142]
x	PksB	PksB		PKS-NRPS hybrid
PksC	PksC	PksC	PKS6	Ab mutant showed 20% reduced virulence, named AbPKS1 in Kim et al. 2009 [142]
PksD	PksD	PksD		Ab PksD missing ACP domain, named AbPKS2 in Kim et al. [142]
PksE	PksE	PksE		Similar to plant type III PKS
PksF	PksF	PksF	PKS16	named AbPKS6 in Kim et al. 2009 [142]
PksG	PksG	PksG	PKS8	named AbPKS3 in Kim et al. 2009 [142]
x	PksH	PksH		
x	PksI	PksI		
x	PksJ	PksJ	PKS9	
x	PksK	PksK		PKS-NRPS hybrid
x	PksL	PksL		missing KS, ACP domains
x	PksM	PksM		missing AT, ACP domains
x	x	PksN		
x	x	PksO (*)		has extra peptidase domain
PksP	x	x	PKS12	named AbPKS4 in Kim et al. 2009 [142]
PksQ	x	x	PKS15	named AbPKS5 in Kim et al. 2009 [142]
PksR	x	x	PKS25	named AbPKS8 in Kim et al. 2009 [142], Ab mutant has 50% reduced virulence
PksS	x	x		depudecin biosynthesis, Ab mutant showed 10% reduced virulence, named AbPKS9 in Kim et al. 2009 [142]
PksT	x	x		named AbPKS10 in Kim et al. 2009 [142]
NPS1	x	x		
NPS2	NPS2	NPS2		age-dependent reduction
NPS3	NPS3	NPS3		NPS-like
NPS4	x	x		NPS-like, Ab mutant showed 60% reduced virulence
NPS5	NPS5	NPS5		NPS-like
NPS6	NPS6	NPS6	NPS2	*Ch*NPS2 has ~62% ident with others
NPS7	NPS7	NPS7	NPS6?	Ab mutant showed 80% reduced virulence
NPS8	NPS8	NPS8		NPS-like
NPS9	NPS9	NPS9		NPS-like
x	NPS10	NPS10		NPS-like
x	NPS11	NPS11		PKS-NRPS hybrid (also Aa1/Aa2 PksK)
x	NPS12	NPS12	NPS3	
x	x	NPS13		
NPS14	NPS14	NPS14	NPS12	NPS-like, tmpL protein, named AbNPS8 in Kim et al. 2009 [142], Ab mutant showed 80% reduced virulence

x - missing , (*) species specific, (+) Aa1/Aa2 specific

3.3.13 Nonribosomal peptide synthetases

Another important class of microbial secondary metabolite synthesis enzymes is nonribosomal peptide synthetase (NRPS). NRPSs are large multi-domain enzymes that synthesize nonribosomal peptides which have a wide range of biological activity from being involved in plant pathogenesis to the production of beneficial antibiotics. Using an approach similar to that used in identifying PKSs, we also identified and annotated domain architecture of 10 NRPS and NRPS-like proteins for *Ab*, 11 for *Aa1*, and 12 for *Aa2* (Table 3-5, Figure B-15, Figure B-16, Figure B-17). Proteins that missed one of the essential domains required for NRPS (A/AA – adenylation/AMP-binding, C/CD – condensation, and T/ACP – thiolation) were classified as "NRPS-like" proteins. The NRPS-like proteins included *Ab NPS3,4,5,8,9,14, Aa1 NPS3,5,8,9,10,14,* and *Aa2 NPS3,5,8,9,10,14*. Previously we identified 7 putative NRPS encoding genes for *Ab*, and annotated and extensively studied functions of one of the *Ab* NRPS proteins (AbNPS2) [142, 143] (and thus we named the NRPSs following this study). The study showed that *AbNPS2* played an important role in development and virulence of *Ab* (age-dependent reduction of virulence). Mutants disrupted in *AbNPS2, AbNPS4,* and *AbNPS7* showed significant reductions in virulence (20% to 80%) indicating that the products of those genes were necessary for *Ab* infection on brassicas (Table 3-5). Several other mutants (*AbNPS6, AbNPS7* and *AbNPS8*) were hypersensitive to oxidative stresses. Among the identified NRPSs, *NPS4* was found only in *Ab* and showed 60% reduced virulence in *Ab* mutant on cabbage, *NPS13* was found only in *Aa2*, and *NPS11* was found only in *A. alternata*. *NPS11* had a *KS* domain and therefore it was classified as a hybrid PKS-NRPS. Among the NRPS-like proteins, *NPS14* was the *tmpL* protein that was found to be required for intracellular redox homeostasis and virulence in *Ab* and the human pathogenic fungus, *A. fumigatus* [144]. This protein was also found in both *A. alternata* strains with ~79% sequence identity (~85% similarity) to *Ab tmpL*.

3.4 Discussion

In this chapter, we have compared the genomes of three *Alternaria* isolates including two *A. alternata* species isolates and a *Brassica* pathogen *A. brassicicola*. We

found marked differences in genome rearrangements, repetitive DNA content, and the number and complexity of various genes and gene families at the species level in particular. We have also attempted to place these differences in the context of saprophytic plasticity, plant pathogenesis, and the ability to drive allergic inflammation in humans.

We found the two *A. alternata* genomes were only slightly larger in overall size than the *A. brassiciola* genome. The *A. alternata* genomes contained approximately 1/10 the amount of repetitive DNA content compared to the *A. brassicicola* genome. Not surprisingly, the *A. alternata* genomes had an approximately 10-15% higher number of total predicted genes than the *A. brassicicola* genome. Genome rearrangement appeared to have occurred more often in *A. brassicicola* perhaps due to its more specialized lifestyle and co-evolution as a brassica pathogen.

As mentioned previously, *A. alternata* is one of the most ubiquitous saprophytic fungi worldwide, whereas *A. brassicicola* is more specialized as a plant pathogen. *A. brassicicola* can also grow saprophytically but quite unlike *A. alternata*, is not generally reported as being found in general environmental surveys (atmospheric, dust samples, etc.). In our various genomic analyses, we found that *A. alternata* had a much higher number of genes dedicated to substrate utilization as depicted in our analyses of carbohydrate utilization enzymes (CAZY) than *A. brassicicola*. Moreover, *A. alternata* appeared to grow more robustly on diverse single carbon containing substrates including lignin compared to *A. brassicicola*. In general both of these species possessed a larger total number of CAZY encoding genes than most fungi that they were compared to including the known biotechnology saprophytic species, *Trichoderma reesei* and *Aspergillus niger*. Collectively, these data suggest that *Alternaria* fungi, particularly *A. alternata*, may be excellent sources of genes encoding enzymes with potential biotechnological applications. Some of these enzymes include cellulases, xylanases, feruoyl esterases, and members of the accessory enzyme family (AA) potentially involved in lignin breakdown. *A. brassicicola* appeared to possess several unique CAZY type genes including an alpha-fucosidase, which was also shared with another dothideomycete brassica pathogen, *L. maculans*. It will be interesting to determine in the future if this enzyme is important for brassica pathogenesis.

In regards to allergenicity and inflammatory potential, *A. alternata* genomes possessed a higher overall content of putative homologs to known allergens regardless of source organism (fungi, dustmites, insects, cats, dogs, pollens, etc.) in comparison to any other fungal genome investigated in this study including the well known allergenic fungus, *A. fumigatus*. Moreover, the number of predicted proteins that have protease or peptidase type activity was significantly higher in *A. alternata* compared to *A. brassicicola*. Serine type proteases are the most expanded class in *A. alternata*. In several recent studies, it has been reported that protease activity in *A. alternata* extracts, especially serine protease, have potent T_H2 driving inflammatory activity [136]. Thus, we might speculate that the expanded number and diversity of proteases, particularly the serine proteases, may contribute to the allergenic or inflammatory potential of *A. alternata* compared to *A. brassicicola*, which has never been reported to be the cause of human allergic disorders.

We also surveyed the number and complexity of secondary metabolite biosynthetic gene clusters (PKS and NRPS). In general we found an overall higher number of PKS and NRPS genes in *A. alternata* compared to *A. brassicicola*, suggesting an overall role for these types of genes in successful saprophytism. However, each species appeared to not only share several gene clusters (e.g. melanin biosynthetic PKS cluster and the siderophore iron chelating NRPS cluster) but have quite unique representatives as well. Interestingly, several of these unique NRPS and PKS genes in *A. brassicicola* have already been demonstrated to be important plant virulence factors [142]. More research will need to be done in the future to determine the actual products (small molecules) associated with these gene clusters and their potential roles in saprophytism and/or plant pathogenicity.

3.5 Conclusion

This is the first report of comparative genomics between multiple isolates within the *Alternaria* group of fungi. We have discovered some potential reasons in the context of genomics why these particular fungi have been so successful as both saprophytes and plant pathogens. Moreover, we have discovered that *A. alternata*, a fungus with historical and strong clinical ties to allergic airway diseases in humans, has tremendous

inflammatory potential in regards to the number and diversity of carbohydrate active enzymes, allergen-like proteins, and proteases.

Chapter 4

Evaluation of sequence comparison criteria for allergen prediction

4.1 Introduction

It is important to identify and eliminate potential allergens from biotechnology-derived products such as genetically modified crops, vaccines, and therapeutics, as well as identifying allergens from sequenced genomes. However, IgE mediated allergenicity is costly and difficult to assess without human data, because no single factor has been recognized as a primary identifier for allergenicity [4, 145]. High throughput allergen detection using proteomics has been used but with very limited success at identifying new allergens [146, 147]. Therefore, bioinformatics approaches have been widely used to pre-screen novel sequences [4, 145, 148–150].

Allergens have diverse structures. The presence of IgE epitopes is an important factor that makes a protein an allergen and identifying them from protein sequences is important for allergen prediction. However, our understanding of the molecular structure and biochemical nature of IgE epitopes is incomplete. IgE epitope regions are usually short sequences but can be conformational and therefore difficult to predict. The widely used guideline for allergenicity assessment of genetically modified (GM) crops was set forth by the International Food Biotechnology Council (IFBC) and the International Life Sciences Institute (ILSI) in 1996 and revised by FAO/WHO in 2001 (Figure 4-1). This guideline uses very relaxed sequence similarity criteria. A protein is identified as a potential allergen if it harbors ≥35% identity with a known allergen over a window of 80 amino acids or has 6 contiguous amino acids that are also found in a known allergen [148, 151]. These criteria are implemented in most of the allergen databases and tools

[150]. However, the FAO/WHO guideline focuses on sensitivity to prevent potential allergens/cross-reactive allergens entering the food market rather than accurate prediction. Therefore, these criteria yield very high false positive rates such that their application is limited [145, 152]. The 6 contiguous amino acid match criterion was designed to capture sequences that possibly share IgE epitopes with known allergens, but it is highly likely to find a match of 6 amino acids between unrelated sequences. The use of longer exact matches (7-8 amino acids) is sometimes recommended. The current evidence weight (based on both sequence similarity and experimental evidence) Codex guideline [149] does not recommend the use of the 6 contiguous amino acid match criterion.

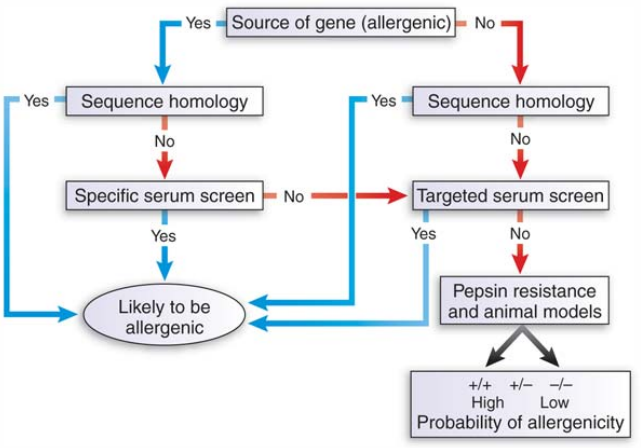


Figure 4-1. FAO/WHO 2001 "decision tree" for allergenicity assessment of GM crops

Reprinted by permission from Macmillan Publishers Ltd: Nature Immunology 6, 857 - 860, copyright 2005

Other sequence similarity based approaches were also used to survey the whole fungal genome pan-allergen repertoire in which allergen homologs were identified using sequence similarity search (BLAST) against a selected set of known fungal allergens [16, 17]. In this approach, a protein was considered an allergen homolog if it was found to have significant alignment with a known allergen sequence (i.e. BLAST e-value $\leq 10^{-10}$). It is very unlikely that an exact match of 6 contiguous amino acids produces a significant e-value score. We have adopted this approach to survey allergen homologs in Chapter 3. However, this approach also suffers from the same high false positive rate as the

FAO/WHO approach and the allergen homologs identified are not necessarily true allergens.

Recently, Mirsky et al. investigated different combinations of sequence similarity scores and the variations of FAO/WHO criteria in allergen prediction [153]. Using the AllergenOnline database (http://allergenoline.org) and a set of green plant proteins collected from SwissProt, they found that the best sensitivity and specificity scores can be obtained when e-value $\leq 10^{-4}$ was used to filter insignificant matches against known allergen sequences (using FASTA sequence alignment software [154]) and then a more stringent FAO/WHO criteria ($\geq$ 35% identity over a window of 80 amino acids or $\geq$13 contiguous amino acid matches) was used to predict allergens. However, the allergens and non-allergens in the test set were identified using the training set (AllergenOnline database) using sequence similarity search with stringent cutoff criteria of 100%, 95%, and 90% sequence identity. This made the test set dependent on the training set and therefore the results may be biased. Moreover, sensitivity and specificity analysis used in this study might not reveal low false positive rate for imbalanced data in large-scale allergen prediction.

In this chapter, we extend the idea of Mirsky et al. and evaluate different comparative genomics criteria using precision/recall evaluation scheme on several data sets that mimic the naturally skewed distribution of allergens. The purpose of this analysis is to obtain a comparative based allergen classifier that is suitable for large-scale allergen prediction in the context of whole genome annotation. Similarly to Mirsky et al.'s approach, we include sequence similarity scores (e-value and bit score using BLAST) and the variations of the FAO/WHO criteria in the evaluation. Compared with the Mirsky et al.'s approach, we further evaluate the effect of varying sequence identity over a window of 80 amino acids.

4.2 Materials and methods

4.2.1 Data sets

An initial set of allergens was built by combining sequences collected from the WHO/IUIS Allergen Nomenclature Subcommittee database (http://allergen.org),

Allergome [150], SDAP [11], AllergenOnline (http://allergenonline.org), and AllerMatch [155] databases. Duplicated sequences and sequences without experimental evidence, containing non-standard amino acids, or shorter than 100 amino acids were removed, and resulted in a set of 3,907 high quality allergen sequences. A portion of this set contains isoforms of the same allergens or allergens with very similar sequences. A putative non-allergen set was created from the Swiss-Prot database [156] by removing sequences tagged with 'predicted' or 'uncertain', and sequences annotated with allergen-related keywords (i.e. 'allerg*', 'antigen', or 'atopy'), similar to other approaches [157–160]. Because many allergens have yet to be identified, this putative non-allergen set may indeed contain some true allergens. Noise was reduced by further removing sequences that were highly similar to any of the sequences collected from the allergen databases (≥90% identity and ≥90% coverage on both query and subject sequences when aligned using BLAST [53]). Similar to the allergen set, sequences shorter than 100 amino acids or having non-standard amino acids were also removed. This resulted in a set of 464,101 putative non-allergens. From the sets of 3,907 allergens and 464,101 putative non-allergens, three data sets were derived and used in this study by the following procedures: All three data sets described below were designed to contain 10 times as many non-allergens as allergens, which represented the natural imbalanced distribution of allergens and non-allergens to some degree. Sequence-based allergen prediction methods often yield low performance on data sets that include many non-allergens that share sequence similarity with allergens. Our data sets exhibited low to high levels of sequence similarity between allergens and non-allergens (Figure 4-2), and thus allowed a more comprehensive evaluation of allergen prediction methods.

- **Data set A (3,907 allergens, 39,070 non-allergens):** All allergen sequences (including isoforms) were selected and 10 times that of putative non-allergen sequences were randomly selected from the putative non-allergen set. This data set exhibited a low level of overall sequence similarity between allergens and non-allergens. When clustered using BLASTClust [53], only 1,108 (~3%) non-allergens, together with 1,293 (~30%) allergens, were grouped in 131 clusters that contained both allergens and non-allergens (allergen/non-allergen clusters)

(Figure 4-2). The non-allergen sequences that were clustered with allergen sequences were designated "allergen-like non-allergens".

- **Data set B (1,990 allergens, 19,900 non-allergens):** All allergens were clustered using BLASTClust with ≥95% identity and ≥95% coverage on both query and subject sequences into 1,990 clusters. To remove sequence redundancy, only one sequence was selected randomly from each cluster to form a set of 1,990 allergens. Ten times as many non-allergen sequences were randomly selected from the putative non-allergen set. This data set also exhibited a low level of sequence similarity between allergens and non-allergens, with only 534 (~3%) non-allergens clustered with 473 (~24%) allergens in 91 allergen/non-allergen clusters (Figure 4-2).
- **Data set C (1,662 allergens, 16,620 non-allergens):** All allergen and putative non-allergen sequences were together clustered using BLASTClust with ≥50% identity and ≥50% coverage on both query and subject and resulted in 291 allergen only clusters, 233 allergen/non-allergen clusters, and 9,529 non-allergen only clusters. From each allergen-only and allergen/non-allergen cluster, at most three allergen sequences were randomly selected. Ten times the number of non-allergen sequences were selected in a similar fashion from the non-allergen only and allergen/non-allergen clusters. The final data set contained 1,662 allergens and 16,620 non-allergens, in which a significant number of non-allergens share sequence similarity with allergens. When being re-clustered using BLASTClust, 6,855 (~41%) non-allergens were grouped together with 725 (~44%) allergens in 232 allergen/non-allergen clusters (Figure 4-2).

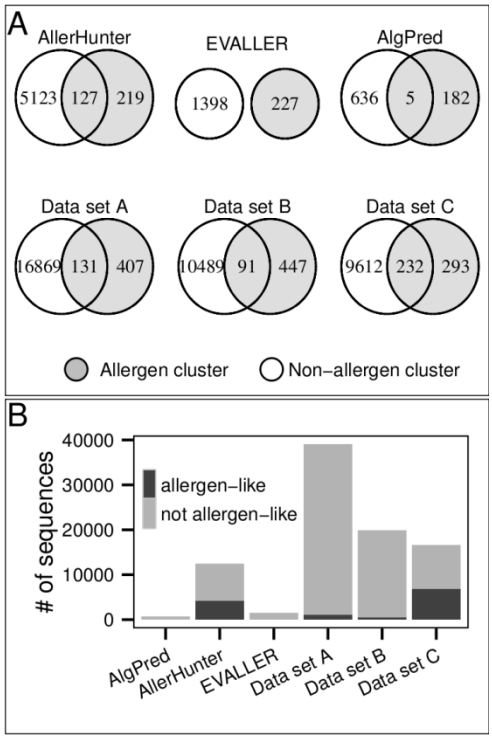


Figure 4-2. Sequence similarity between allergens and non-allergens in Allerdictor data sets and other data sets (BLASTClust cutoff ≥50% sequence identity over ≥50% query or subject coverage). The shared regions in Venn diagrams (A) are clusters that contain both allergen and non-allergen sequences. The total number of non-allergens that are allergen-like and not allergen-like are detailed in the column plot (B).

4.2.2 Identifying allergens using combination of sequence similarity criteria

Multiple sequence similarity criteria were evaluated for allergen prediction (Table 4-1). To classify a protein as an allergen or a non-allergen, the protein sequence was first searched against a database consisting of known allergen sequences using BLAST. Multiple alignments with different known allergen sequences could be found for a single protein. The minimum e-value of these alignments was recorded as the e-value score of the protein sequence against the allergen database. The maximum bit score of these alignments was recorded as the bit score of the protein sequence against the allergen database. Sequence identity over a window of 80 amino acids was calculated for each BLAST alignment with a known allergen. The maximum identity (80 aa identity) was recorded as the identity score against the allergen database. Finally, MEM was identified for the protein as the maximum of the maximum exact matches obtained when searching the protein sequence against known allergen database using SparseMEM software [161]. A protein was classified as an allergen if it passed a test of combination of sequence

similarity criteria (Figure 4-3). Otherwise, it was classified as a non-allergen. We herein call this approach the modified FAO/WHO approach or comparative approach (abbreviated as IDMEM).

Table 4-1. Sequence similarity criteria used in allergen prediction

Criterion name	**Description**	**Identified by**
E-value	Lowest e-value when aligning the sequence against the database of known allergen sequences	BLAST
Bit score	Highest bit score when aligning the sequence against the database of known allergen sequences	BLAST
80 aa ID	Highest sequence identity over an alignment that stretches 80 amino acids when aligning with the database of known allergen sequences	BLAST
MEM	Maximal exact matches or maximal contiguous amino acid matches against the database of known allergen sequences	SparseMEM

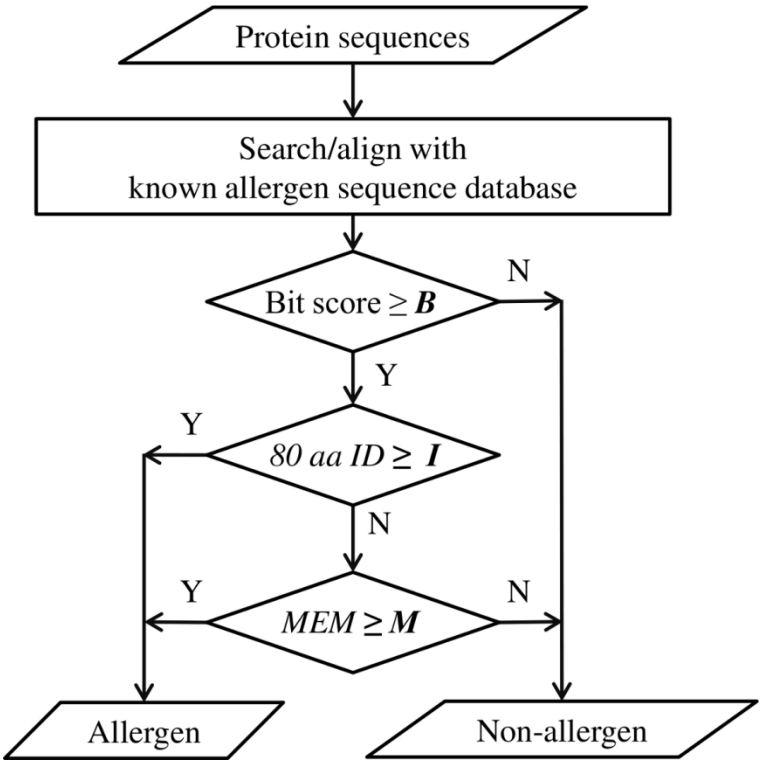


Figure 4-3. IDMEM approach to identify allergen homologs from fungal genomes. Parameters B, I, and M are optimized to maximize average F1 score of 10-fold cross validation run on data sets A, B, and C.

4.2.3 Performance evaluation criteria

The goal of this evaluation is to identify the best combinations of different sequence similarity criteria in the modified FAO/WHO approach that can be used to predict allergens at large-scale in applications such as whole genome annotation. In such cases, the number of allergens in the data is extremely low compared to the number of non-allergens. We used precision/recall (PR) performance evaluation because it can reveal the high number of false positives and therefore PR is a better evaluation method than sensitivity/specificity evaluation in ROC curves [162]. Given the number of true positives (TP), true negatives (TP), false positives (FP), and false negatives (FN), precision and recall are defined as follows.

$$\begin{aligned} Precision &= \frac{TP}{(TP+FP)} \\ Recall &= \frac{TP}{(TP+FN)} \end{aligned} \tag{4-1}$$

The harmonic mean of the precision and recall (F1 score) was used to rank the performance of different combined criteria. The F1 score is a widely used criterion in information retrieval that gives balanced weights on precision and recall [163]. The F1 score is defined as follows.

$$F1 = 2\times\frac{Precision\times Recall}{Precision + Recall} \tag{4-2}$$

4.2.4 Cross-validation

Cross-validation was used to estimate performance of comparative criteria in allergen prediction. A data set was randomly partitioned into 10 subsets of roughly equal size. Each subset (containing both allergens and non-allergens) was held out as a test set, and the remaining allergen sequences from 9 other subsets were used as a known allergen database for comparative based allergen prediction. Performance scores on the 10 subsets were then averaged and reported.

4.3 Results and discussion

We first evaluated individual sequence similarity criterion as potential allergen prediction scores using 10-fold cross-validation on each of the three data sets described earlier (A, B, and C). The evaluated criteria were the BLAST e-value, bit score, 80 aa identity, and MEM (Table 4-1). The results demonstrated that sequence similarity scores were acceptable criteria for allergen prediction in data sets that exhibited low levels of sequence similarity between allergens and non-allergens. However, sequence similarity approaches performed poorly on data sets that exhibited high levels of sequence similarity between allergens and non-allergens. These approaches were also heavily affected by the prevalence of allergens in the data sets. For example, in skewed data sets with a low ratio of allergens, sequence similarity approaches yielded low precision and therefore were not suitable for allergen prediction in large-scale data such as whole genomes/proteomes.

4.3.1 BLAST sequence alignment scores in allergen prediction

BLAST e-value and bit scores were evaluated for allergen prediction on each of the three data sets (A, B, and C). As expected, the majority of non-allergens had high e-values and bit scores when aligned with known allergens while the reverse was true for allergens (Figure 4-4). However, many non-allergens had lower e-values (higher bit scores), and many allergens had higher e-values (lower bit scores).

For data sets A and B, using an e-value cutoff of 10^{-30} - 10^{-10} would classify the majority of non-allergens correctly. However, BLAST similarity scores were a poor criterion for allergen classification on data set C. This can be explained by the fact that data set C had many non-allergens that showed high levels of sequence similarity with allergens. This problem becomes more severe in skewed data sets because the number of false positives (non-allergens misclassified as allergens) is much higher than the number of true positives (correctly predicted allergens) and thus the precision is low.

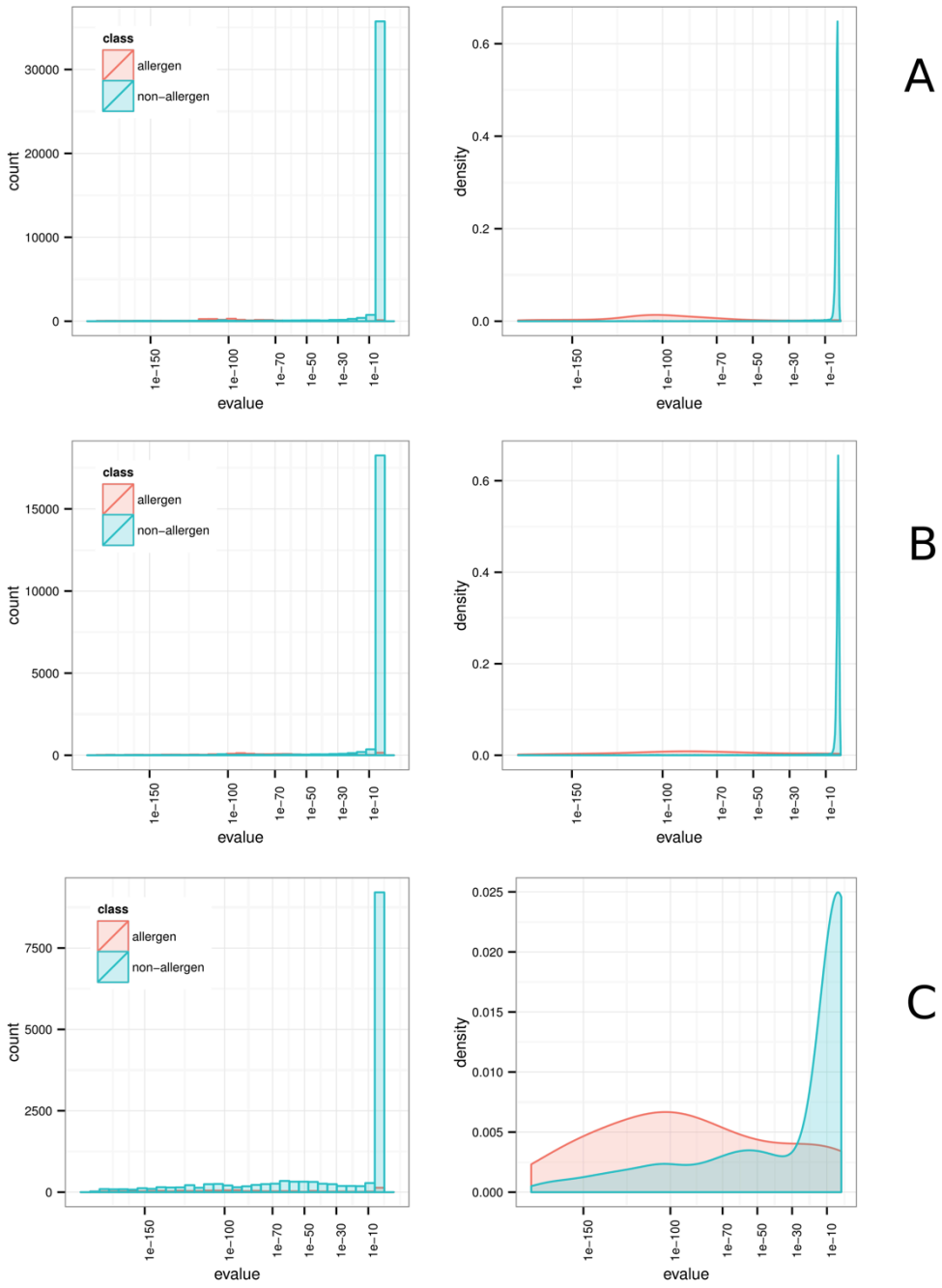


Figure 4-4. Distribution of BLAST e-value scores using data sets A, B, and C. Histograms of the e-values are shown on the left. Esmitated densities of the e-values are shown on the right.

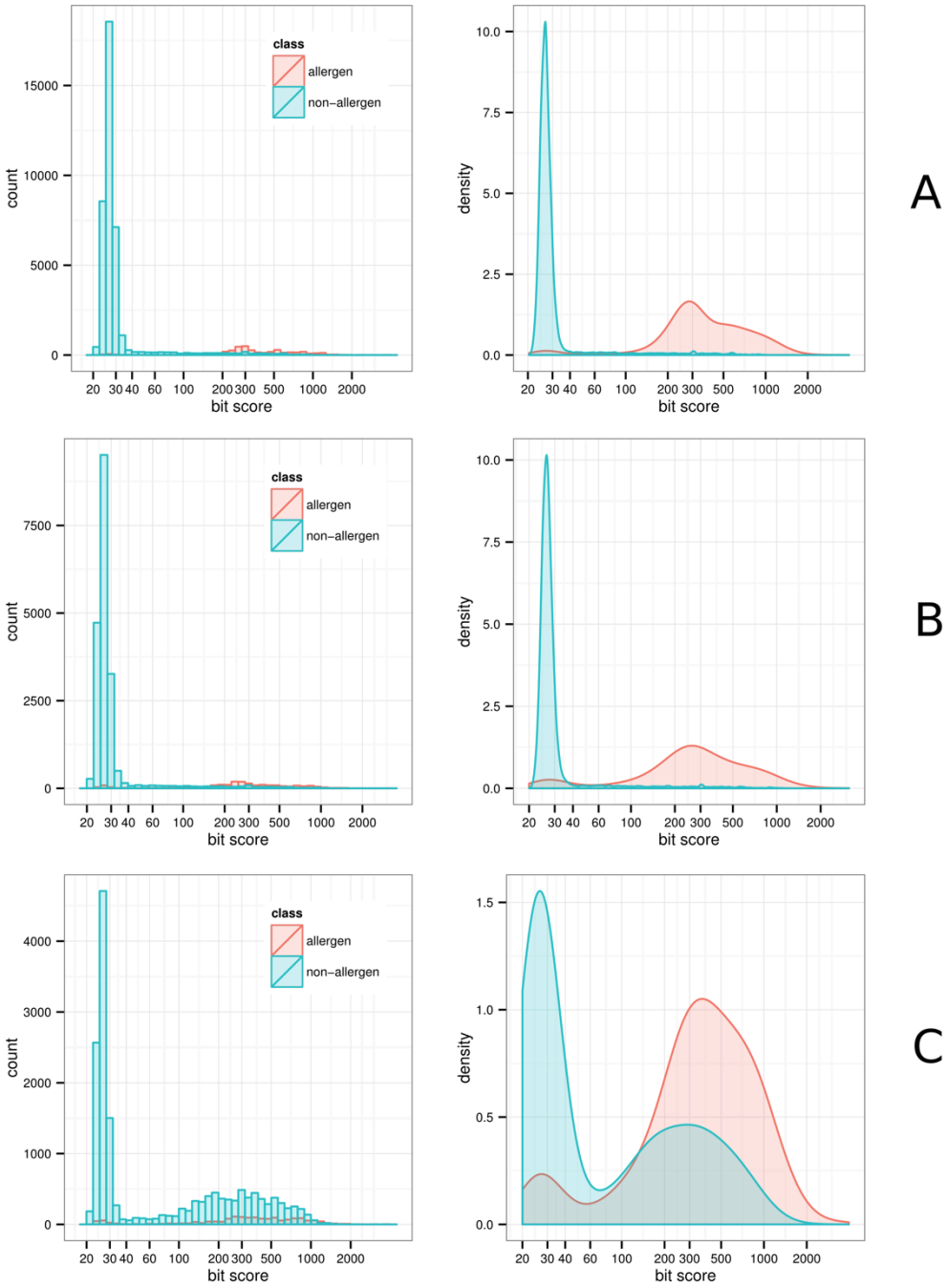


Figure 4-5. Distribution of BLAST bit scores using data sets A, B, and C. Histograms of the bit scores are shown on the left. Estimated densities of the bit scores are shown on the right.

4.3.2 Sequence identity in allergen prediction

Sequence identity has been used as a good criterion for estimating structural similarity between proteins. In a homology modeling approach to construct protein model

from a template, a sequence identity or similarity of ≥ 30-50 % is recommended [164]. Sequence identity over a window of 80 amino acids was investigated on data sets A, B, and C using BLAST search. Similarly to BLAST sequence similarity scores, sequence identity was found to be a good criterion for data sets that exhibited low levels of sequence similarity between allergens and non-allergens (data sets A and B). FAO/WHO recommended using sequence identity of 35% over 80 amino acids for allergen classification. However, our analysis showed that increasing the 80 aa identity cutoff would increase classification performance. A cutoff of 80 aa identity in the range of 40%-70% was effective in classifying allergens from non-allergens for data sets A and B (Figure 4-6A and B). When the level of sequence similarity between allergens and non-allergens increased (data set C) sequence identity was a poor criterion to discriminate non-allergens from allergens (Figure 4-6C).

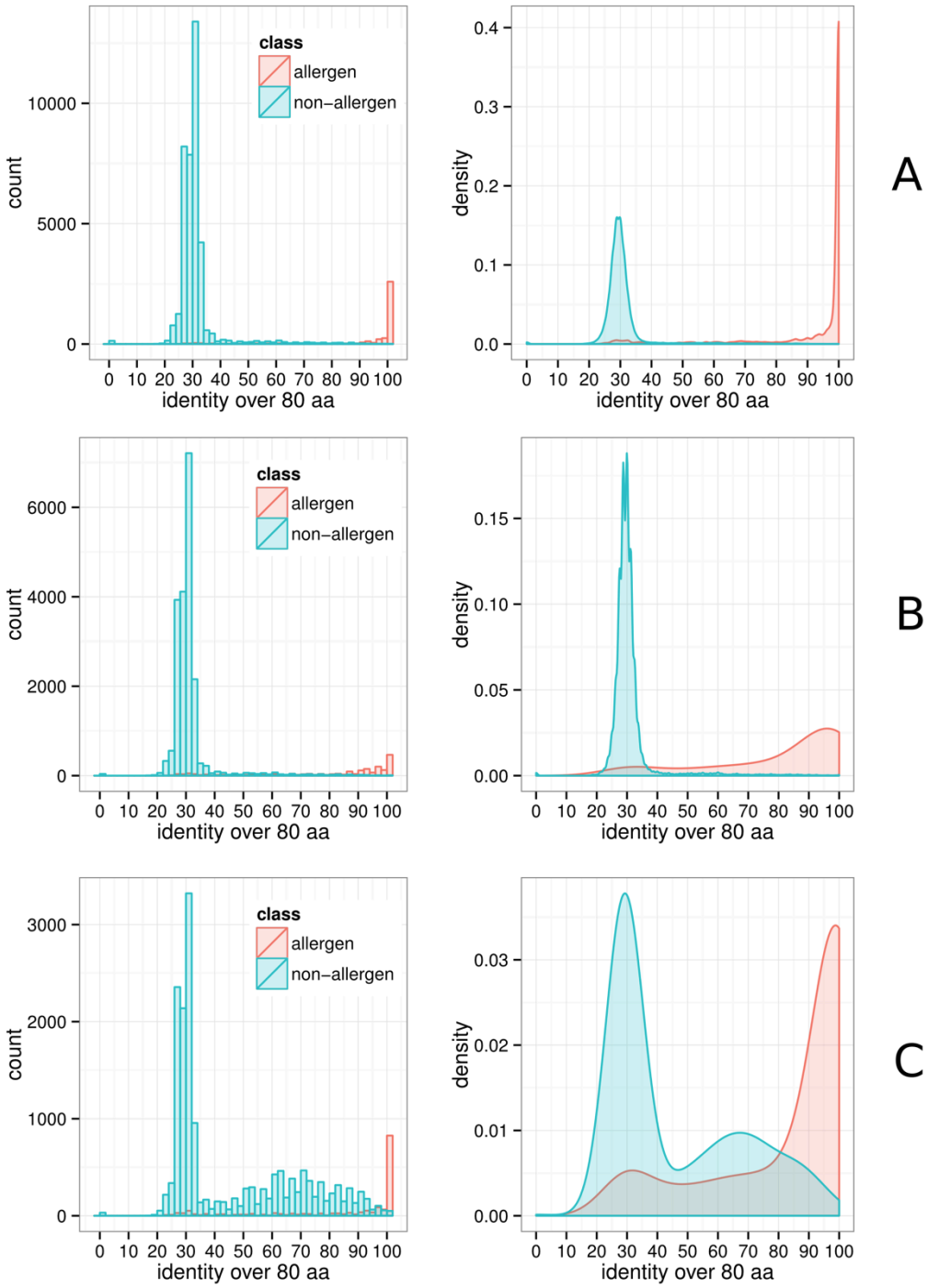


Figure 4-6. Distribution of sequence identity using data sets A, B, and C. Histograms of the identities are shown on the left. Estimated densities of the identities are shown on the right.

4.3.3 Maximal exact matches in allergen prediction

MEM showed some capability of distinguishing allergens from non-allergens in data sets with lower levels of sequence similarity between allergens and non-allergens

(data sets A and B). However, it suffered from the same high false positive issues as BLAST sequence similarity scores and 80 aa identity on data set C with high level of sequence similarity between allergens and non-allergens.

For all data sets, a MEM cutoff of 6 contiguous amino acid matches was a poor criterion for allergen prediction (Figure 4-7). A large number of non-allergens had MEM $\geq$ 6. Other studies also found that MEM cutoff of 6 was a poor choice for allergen prediction such that many non-allergens were misclassified as allergens [145, 152]. It was proposed to increase MEM in FAO/WHO and in practice, MEM is often chosen between 6 and 8 in evaluation of allergenicity of novel proteins introduced to genetically modified crops. Our analysis showed that increasing MEM cutoff significantly reduced the number of false positives. A large number of false positives were eliminated by increasing the MEM cutoff from 6 to 8 (Figure 4-7). Our analysis also showed that the best performing MEM cutoffs were in the range of 10-20 exact amino acid matches.

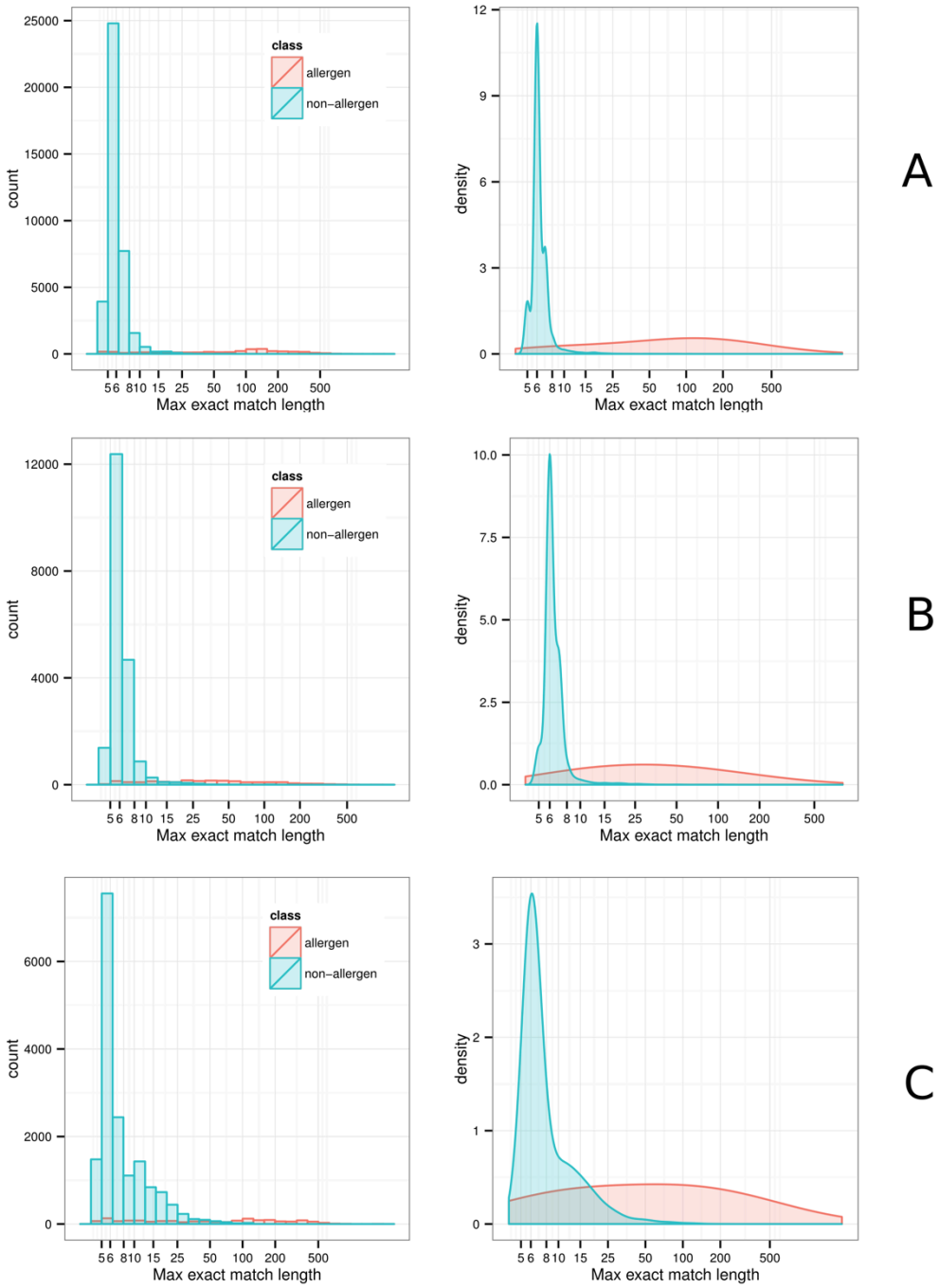


Figure 4-7. Distribution of maximal exact maches (MEM) using data sets A, B, and C. Histograms of the MEMs are shown on the left. Estimated densities of the MEMs are shown on the right.

4.3.4 Variations of FAO/WHO criteria in combination with BLAST similarity scores in allergen prediction

We varied the 80 aa sequence identity and maximal exact match cutoffs of the FAO/WHO guideline in combination with the BLAST sequence similarity bit score cutoff to identify the best combination of criteria for large-scale allergen prediction (IDMEM, Figure 4-3). The harmonic mean (F1 score) of precision and recall was chosen as the prediction performance score when comparing different criteria combinations. The results showed that for large-scale allergen prediction, 80 aa identity and MEM cutoffs in FAO/WHO guideline should be increased to reduce the false positive rate and increase the F1 score. When 80 aa identity or MEM cutoffs were small, additionally using BLAST similarity bit score cutoff helped increase F1 score. However, when 80 aa identity and MEM cutoffs were increased, the effect of bit score cutoff on overall performance was only subtle. This can be explained by the fact that using higher sequence identity and higher maximal exact match cutoffs already filtered out sequences with low BLAST similarity scores.

The optimal criteria for allergen prediction were also found to be dependent on the data sets used for evaluation. For data sets with significant levels of sequence similarity or duplication (data sets A and C) the optimal performances were obtained when 80 aa identity and MEM cutoffs were high. For example, when bit score cutoff = 60, the optimal criteria identified using data set A was 80 aa identity cutoff ~ 76-90 and MEM cutoff ≥ 17, and using data set C was 80 aa identity cutoff ~ 93-97 and MEM cutoff ≥ 40. Data set B had the least level of sequence duplication (i.e. duplicated allergen sequences were removed and non-allergens were randomly selected from a large set of putative non-allergens identified from Swiss-Prot), and therefore the optimal 80 aa identity and MEM cutoffs were smaller. Using bit score cutoff of 60, the optimal 80 aa identity cutoff was ~57-72 and the optimal MEM cutoff was ≥ 16.

Because sequence identity, maximal exact matches, and BLAST similarity scores are not independent of each other, performance might not change in response to the variation of a single criterion within a specific range. For example, after MEM cutoff reached some value, increasing it would not change performance. Similarly, low range bit

score cutoff does not affect performance when 80 aa identity and MEM cutoffs were already increased to the optimal values.

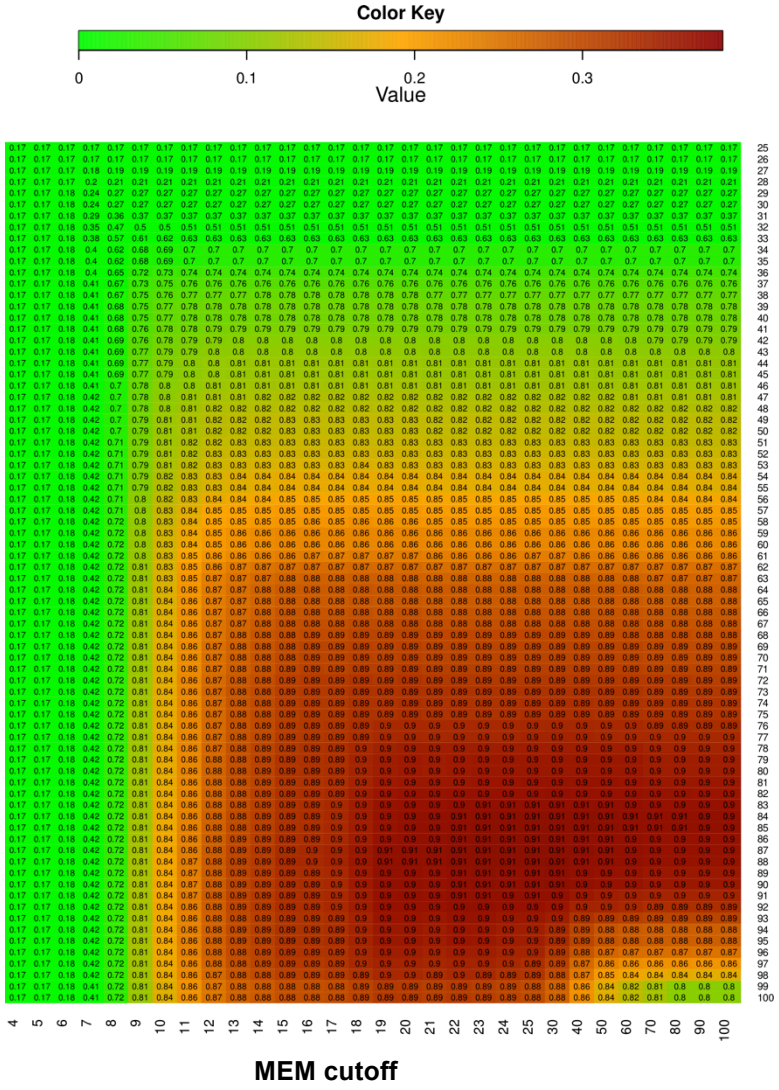


Figure 4-8. Heatmap of average F1 scores obtained from 10-fold cross-validation following the IDMEM approach on data set A with bit score cutoff = 0. Color value is scaled by a power of 15 for visualization purpose.

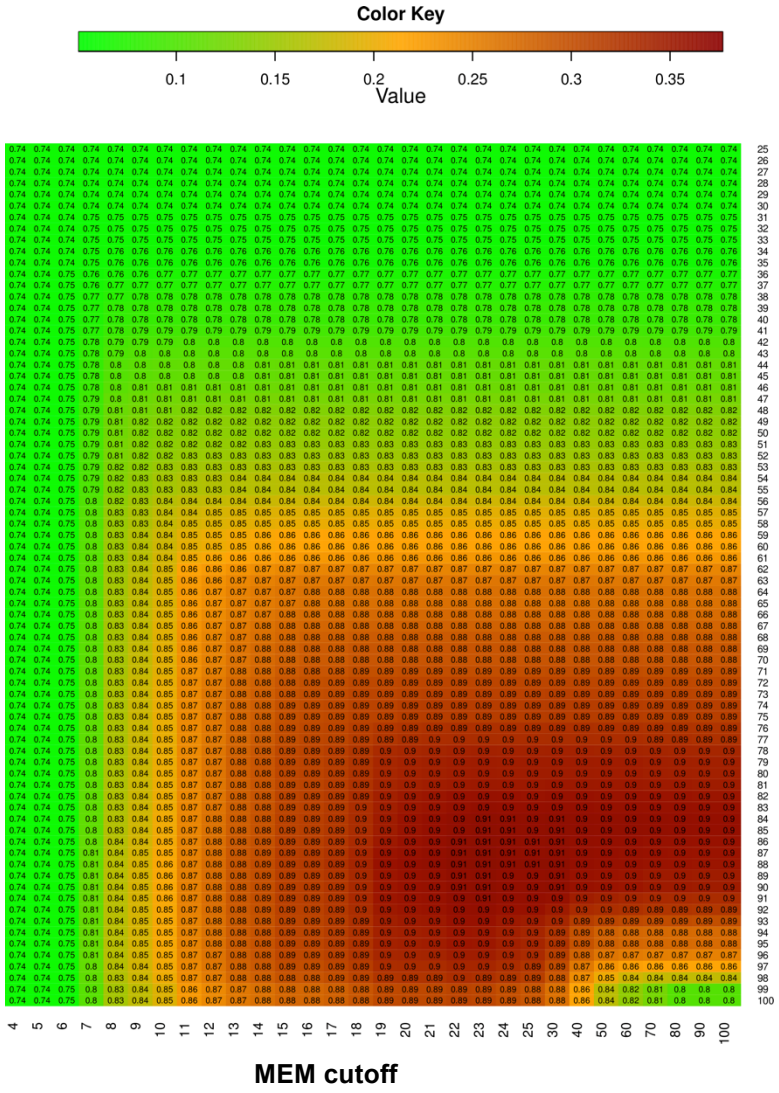


Figure 4-9. Heatmap of average F1 scores obtained from 10-fold cross-validation following the IDMEM approach on data set A with bit score cutoff = 60. Color value is scaled by a power of 15 for visualization purpose.

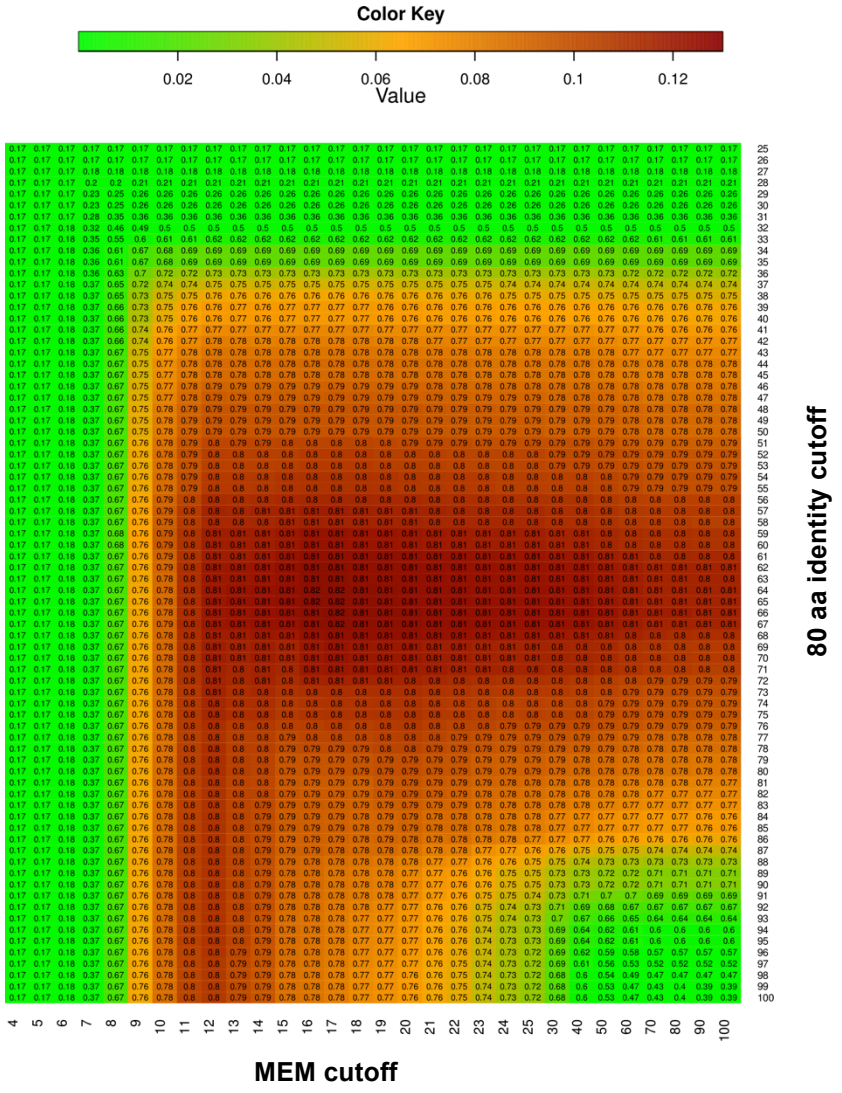


Figure 4-10. Heatmap of average F1 scores obtained from 10-fold cross-validation following the IDMEM approach on data set B with bit score cutoff = 0. Color value is scaled by a power of 15 for visualization purpose.

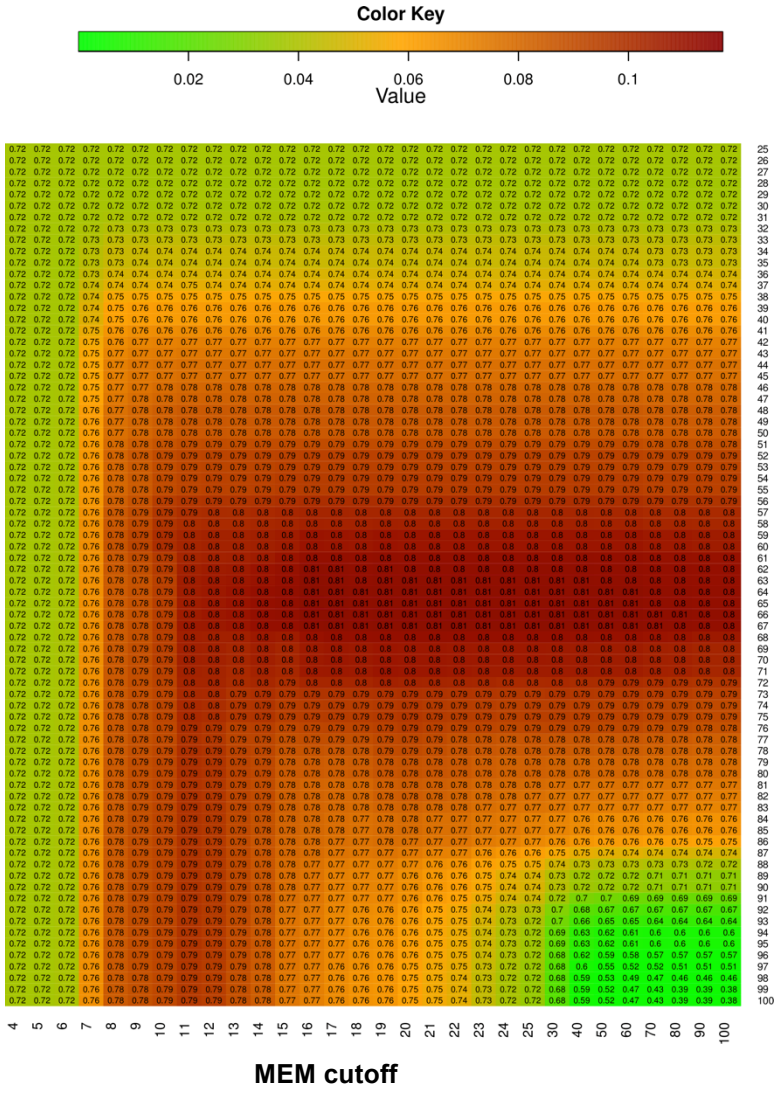


Figure 4-11. Heatmap of average F1 scores obtained from 10-fold cross-validation following the IDMEM approach on data set B with bit score cutoff = 60. Color value is scaled by a power of 15 for visualization purpose.

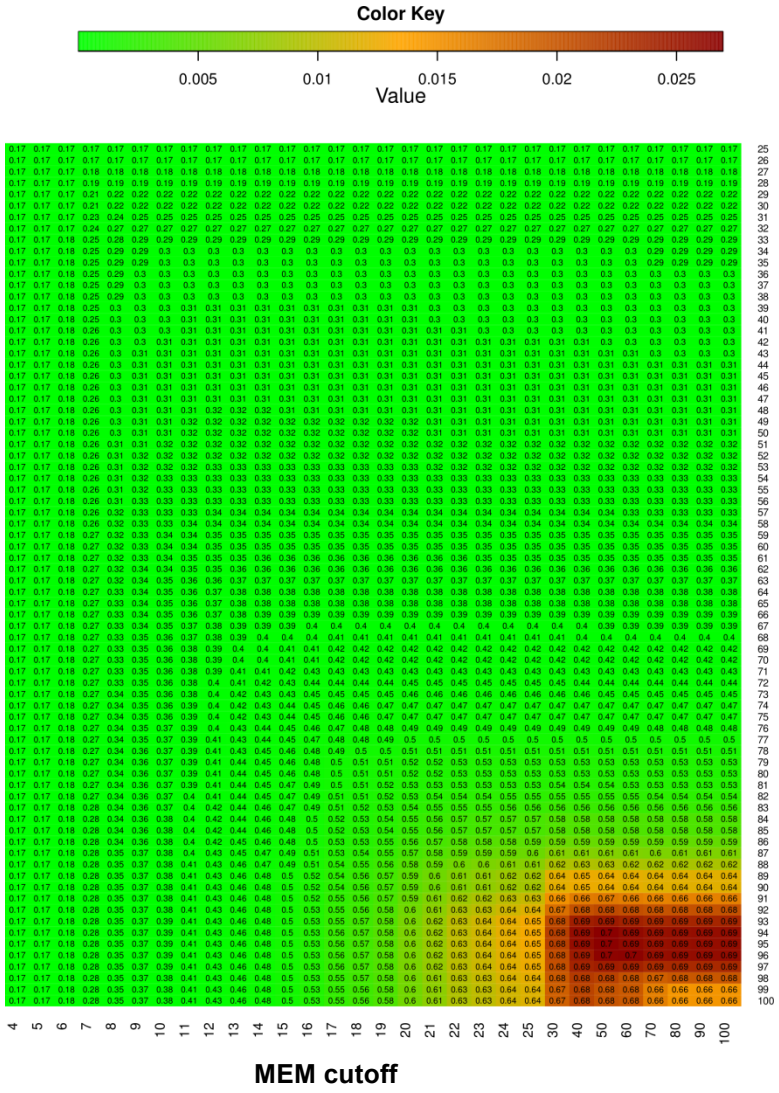


Figure 4-12. Heatmap of average F1 scores obtained from 10-fold cross-validation following the IDMEM approach on data set C with bit score cutoff = 0. Color value is scaled by a power of 15 for visualization purpose.

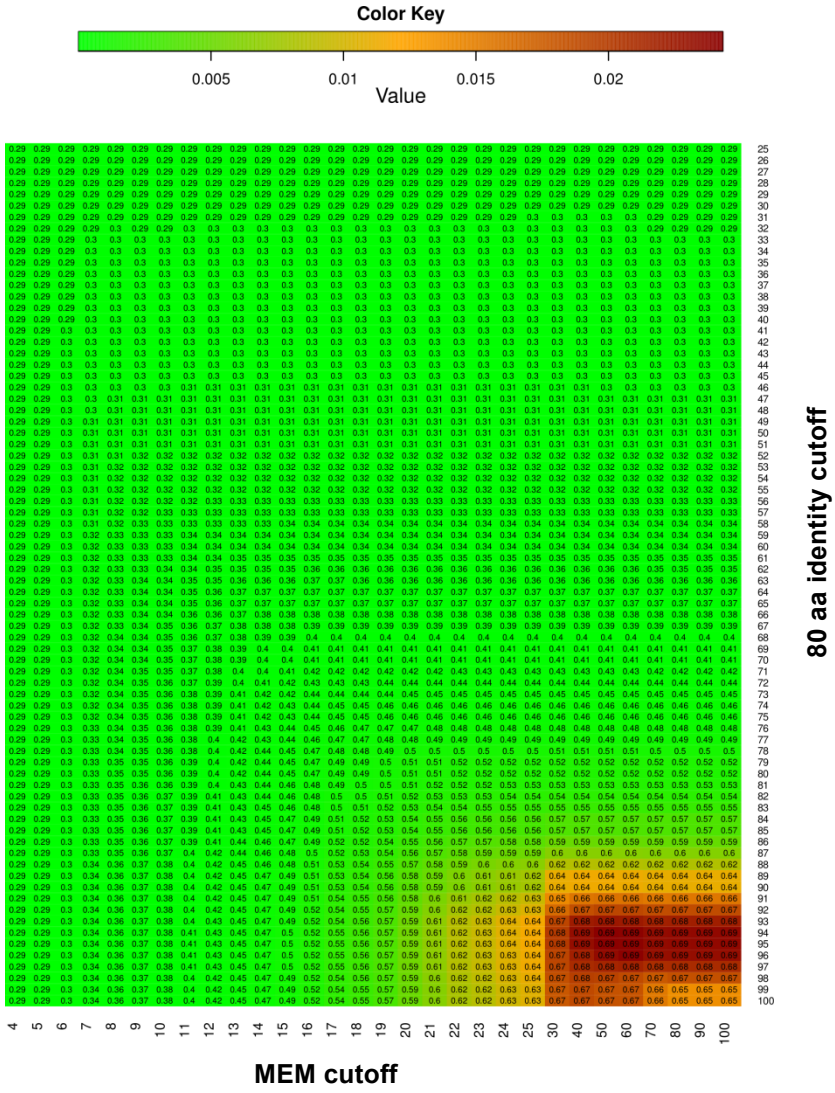


Figure 4-13. Heatmap of average F1 scores obtained from 10-fold cross-validation following modified FAO/WHO approach on data set C with bit score cutoff = 60. Color value is scaled by a power of 15 for visualization purpose.

4.4 Conclusion

We have evaluated different combinations of sequence similarity criteria surrounding the FAO/WHO guideline to assess allergenicity of proteins based on the sequences. In order to obtain a balanced precision and recall that are more practical for large-scale allergen prediction, we recommend increasing the 80 aa identity and MEM cutoffs. We found the optimal range for 80 aa identity cutoff was 57-72% and for MEM

cutoff was ≥16. BLAST similarity scores' cutoffs did not significantly affect performance and can be left out.

One remaining issue is that the sequence similarity approach is comprised of a sequence alignment step whose computation time is dependent on the length of the known allergen database and the length of the predicting protein sequence. This will become relatively slow when the number of known allergens in the database increases and the total number of proteins to analyze increases.

Chapter 5

Machine learning based large-scale allergen prediction

Parts of this chapter are included in:

Dang H and Lawrence C. Allerdictor: Fast allergen prediction using text classification techniques. Bioinformatics (2014) 30 (8): 1120-1128.

5.1 Introduction

In Chapter 4, we have evaluated comparative strategies to obtain an allergen classifier that can be applied to large-scale imbalanced data such as a whole genome. Although some improvement was obtained, comparative approaches are not capable of distinguishing sequence matches in regions that are unrelated to allergens from regions that are more likely to be related to allergenicity of a protein. Moreover, comparative approaches use sequence alignment and string matching to find the optimal similarity between sequences thus they are relatively slow. It often takes hours to analyze a whole genome depending on the size of the predicted corresponding proteome. Other approaches to allergen prediction employ more sophisticated machine learning techniques that attempt to learn some characteristics of the allergen proteins using a set of known allergens and putative non-allergens that have achieved some improvement over the traditional comparative approaches [157–160, 165–167].

Many methods for allergen prediction have been developed and are more accurate than the FAO/WHO pure sequence similarity-based approach. The majority of these methods are based on supervised machine learning and differ in ways to extract useful features from amino acid sequences. Most of them rely on sequence similarity with allergen specific peptides or motifs including the method by Stadler and Stadler [152],

the method by Li et al. [165] , WebAllergen [168], EVALLER [157, 158] and SORTALLER [160], or with known IgE epitopes such as AlgPred [166], or with known allergens and putative non-allergens such as AllerHunter [159]. Other methods use physicochemical representation of protein structure such as APPEL [167] and SDAP [169], or amino acid/dipeptide composition such as AlgPred [166].

Although current methods are significantly more accurate than the FAO/WHO approach, large-scale allergen prediction using these methods is still ineffective and inefficient. On large-scale data where non-allergens are naturally more abundant, the number of false positives often exceeds the number of true positives, which lowers the precision and thus the usefulness of the prediction. Moreover, the most accurate methods are relatively slow because they rely on homology and use sequence alignment to construct feature vectors. Additionally, current allergen prediction methods come pre-trained in the form of web servers without the capability of large batch submission, making large-scale allergen prediction even more difficult.

Here we propose a new sequence-based allergen prediction method (Allerdictor) that runs in time linear in sequence length and is capable of producing high precision over high recall, even on highly skewed data. Allerdictor models sequences as text documents in which words are represented as overlapping *k*-mers generated from the sequences. We found that the *k*-mer approach is particularly effective in allergen prediction. Feature construction is much faster than sequence alignment based methods and can be performed in time linear in sequence length. Allerdictor was implemented with both Naive Bayes (NB) and Support Vector Machine (SVM) classifiers. SVM outperformed NB on more difficult data sets where the level of sequence similarity between allergens and non-allergens is higher. Thus, we will mostly discuss results for the SVM-based version.

The advantages of Allerdictor make it very practical for large-scale allergen prediction in applications such as whole genome annotation, biotechnology-derived gene product screening, and allergen discovery from large public sequence databases. Allerdictor is implemented in Python and available as standalone and web server versions at http://allerdictor.vbi.vt.edu.

5.2 Methods

Allerdictor represents sequences as text documents and uses NB or SVM for allergen classification. We used the data sets described in Chapter 4 to evaluate Allerdictor.

5.2.1 Text representation of sequences

To represent an amino acid sequence of length n as a text document, Allerdictor uses a small sliding window of size k to break the sequence into $n - k + 1$ overlapping k-length peptides (k-mers). This collection of k-mers is used as a new sequence representation. If we consider a k-mer as a word, this representation is similar to the bag-of-words in document modeling [163]. The set of all unique k-mers generated from training data is similar to the dictionary used in text modeling and is herein called a k-mer dictionary. The feature vector for a sequence can be constructed by recording the appearance/absence of the k-mers (binary representation) or counting the frequencies of the k-mers (k-mer frequency representation). Given that N is the size of the k-mer dictionary built from training data, the k-mer frequency vector for a sequence is:

$$X = [x_1, x_2, \ldots, x_N] \tag{5-1}$$

where x_i is the frequency of emitting the ith k-mer of the dictionary from the sequence using the sliding window. When k is small (i.e. ≤ 2), the size of the k-mer feature vector is small. When k is larger (i.e. $k \geq 3$), because only a very small fraction of the k-mer dictionary can be generated from a limited length protein sequence, the k-mer feature vector is extremely sparse with the maximum of $n - k + 1$ non-zero elements.

When k is large, the k-mer representation of sequences also shares similar properties with the bag-of-words approach in text modeling [170] such as: (i) the feature space is very high dimensional (the number of possible unique k-mers is 20^k, given 20 amino acid alphabet size); (ii) feature vectors are sparse; (iii) the distribution of k-mer frequencies generally follows Zipf's law [171] in which the number of rare k-mers are much higher than the number of frequent k-mers (Figure 5-1).

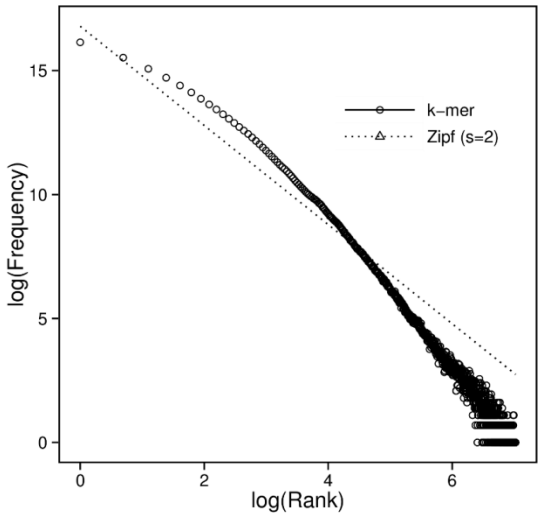


Figure 5-1. Distribution of the frequencies of k-mer (k=6) for the Swiss-Prot protein sequences compared with Zipf distribution (exponent parameter s = 2).

Many text classification methods can then be applied on the *k*-mer sequence representation. NB and SVM were chosen for Allerdictor because they were among the best methods for text classification and fast on high dimensional sparse vectors.

5.2.2 Naive Bayes

NB is a simple yet effective method for text classification, especially for spam filtering [163]. Using a multinomial NB model, Allerdictor-NB models the distributions of *k*-mer frequencies over allergen/non-allergen classes with a relaxed assumption that *k*-mer frequencies are independent of each other given the class. The probability of being an allergen for a sequence represented by a *k*-mer frequency vector X in (5-1) is given by:

$$\begin{aligned} p(alg|X) &= \frac{p(alg).p(X|alg)}{p(X)} \\ &= \frac{p(alg).\prod_{i=1}^{N} p(k_i|alg)^{x_i}}{\sum_{c\in\{alg,nlg\}} p(c).\prod_{i=1}^{N} p(k_i|c)^{x_i}} \end{aligned} \tag{5-2}$$

where *alg* and *nlg* are allergen and non-allergen classes, respectively, and k_i is the *ith k*-mer in the dictionary. The probability of seeing the *ith k*-mer in the allergen/non-allergen class $p(k_i|c)$ and the prior probability of the classes $p(c)$ can be estimated from training

data of known allergen and non-allergen sequences. The probability that the sequence is a non-allergen *p(nlg|X)* can be calculated by a similar formula or equal to *1 – p(alg|X)*.

5.2.3 Support vector machine

SVM has been successfully used in numerous applications across many fields, including text classification [170, 172–174]. Allerdictor-SVM uses a linear SVM model and *k*-mer frequencies are further normalized by the total number of *k*-mers generated from the sequence by the sliding window. The normalized vector *X'* of a *k*-mer frequency vector *X* given in (5-1) of a sequence of length *n* is:

$$X' = [x'_1, x'_2, \ldots, x'_N] \text{ with } x'_i = x_i/(n-k+1) \tag{5-3}$$

Each sequence represented by *X'* is now a point in *N*-dimensional space. Given a training data set of M sequences $\{X'_1, X'_2, \ldots, X'_M\}$ labeled with $\{y_1, y_2, \ldots, y_M\}$ (y_i = *1* if X'_i is allergen, y_i = *-1* otherwise), a soft margin linear binary SVM classifier finds the optimal hyperplane *h* that separates allergens from non-allergens with the maximum margin of classification, which is equivalent to solving:

$$\begin{aligned} \min_{w,\xi,b} \quad & \frac{1}{2}w^T w + C\sum_{i=1}^{M}\xi_i \\ \text{subject to} \quad & y_i(w^T X'_i + b) \geq 1 - \xi_i \text{ with } i = 1..M \\ & w \in \mathbf{R}^N, b \in \mathbf{R} \end{aligned} \tag{5-4}$$

where *N*-dimensional vector w is the normal vector of *h*, *b/*||w|| is the distance from the origin to *h*, slack variables ξ_i designate how far the points can pass the margin boundaries (misclassification) in cases of non-linear separable data, and *C* is the regularization constant to control how much of the training data can be misclassified.

A new sequence *X'* is then classified by what "side" of *h* it lies via an SVM score (with positive score being an allergen):

$$SVMscore = w^T X' + b \tag{5-5}$$

This score is then converted to a posterior probability of being an allergen by fitting a sigmoid function [175, 176]. Training and testing were conducted using the

SVMLight software [177] with an allergen misclassification penalty weight parameter $j = 10$ to address data imbalance. The regularization constant C was chosen by optimizing the performance via cross-validation described below.

5.2.4 Cross-validation and dimension reduction

Nested 10-fold cross-validation was used to evaluate Allerdictor performance on three data sets A, B and C. Each data set was randomly partitioned into 10 subsets containing roughly equal numbers of both allergen and non-allergen sequences. In each evaluation fold, one subset was held out (test set), and the remaining 9 subsets were combined and randomly partitioned into 10 other subsets for an inner 10-fold cross-validation to choose the best parameters. Mutual information [163] was used to generate feature selection scores. All *k*-mers were ranked by mutual information between the class variable (allergen/non-allergen) and *k*-mer frequency variables, and the top ranked *k*-mers were selected to build the prediction model. A feature abstraction technique was also used to group *k*-mers with the same frequency distribution in an allergen training set and in a non-allergen training set. This is a special case of distributional clustering that has been used successfully in text classification [178, 179]. The *k*-mers that were grouped together have the same frequency distribution over the allergen/non-allergen classes (observed from training data) and therefore they received the same weights in the classification model.

5.3 Results and discussion

Evaluation results of Allerdictor on data sets A, B, C, and a data set previously built and used for AllerHunter (http://tiger.dbs.nus.edu.sg/AllerHunter) demonstrated that Allerdictor was capable of obtaining high precision over high recall rates. We evaluated Allerdictor using precision/recall (PR) measures that are widely used in information retrieval [163] instead of sensitivity/specificity measures and ROC analysis. A PR curve plots precision against recall obtained by varying the prediction score cutoff. Precision, recall, and F1 scores were defined previously in (4-1) and (4-2).

A ROC curve plots true positive rate (also called sensitivity or recall) against false positive rate (1 - specificity). Because non-allergens are much more abundant than

allergens in nature, a method that predicts many more false positives than true positives (low precision) can still produce a good ROC curve as long as its sensitivity is high and thus is often misleading [162]. PR curves can reveal high false positive rates, and thus provide a more meaningful evaluation on naturally skewed allergen/non-allergen distributions that are also represented in data sets A, B, and C.

In addition to the F1 score, we also reported Matthews correlation coefficient (MCC), a similar performance evaluation score of binary classifiers that put balanced weights on precision and recall. MCC is defined as follows.

$$MCC = \frac{TP \times TN - FP \times FN}{\sqrt{(\mathrm{TP} + \mathrm{FP})(\mathrm{TP} + \mathrm{FN})(\mathrm{TN} + \mathrm{FP})(\mathrm{TN} + \mathrm{FN})}} \quad (5\text{-}6)$$

5.3.1 Length of *k*-mer peptides

The length of *k*-mer peptides is the most important parameter for Allerdictor prediction models. Allerdictor performed very differently with different *k* values. Performance peaked at *k* = 5 or 6 and decreased as *k* moved away from the peaks (Figure 5-2). This interesting result agrees with the debatable criterion of 6 contiguous amino acid matches with a known allergen used by FAO/WHO guideline. The classification power of Allerdictor comes with its ability to distinguish and assign higher weight to *k*-mers that are more likely associated with allergens (see 5.3.4). With *k* = 5, Allerdictor produced near perfect false positive rates while *k* = 6 allowed for better sensitivity and still maintained very low false positive rates. We chose *k* = 6 for the analyses and results reported in the following sections.

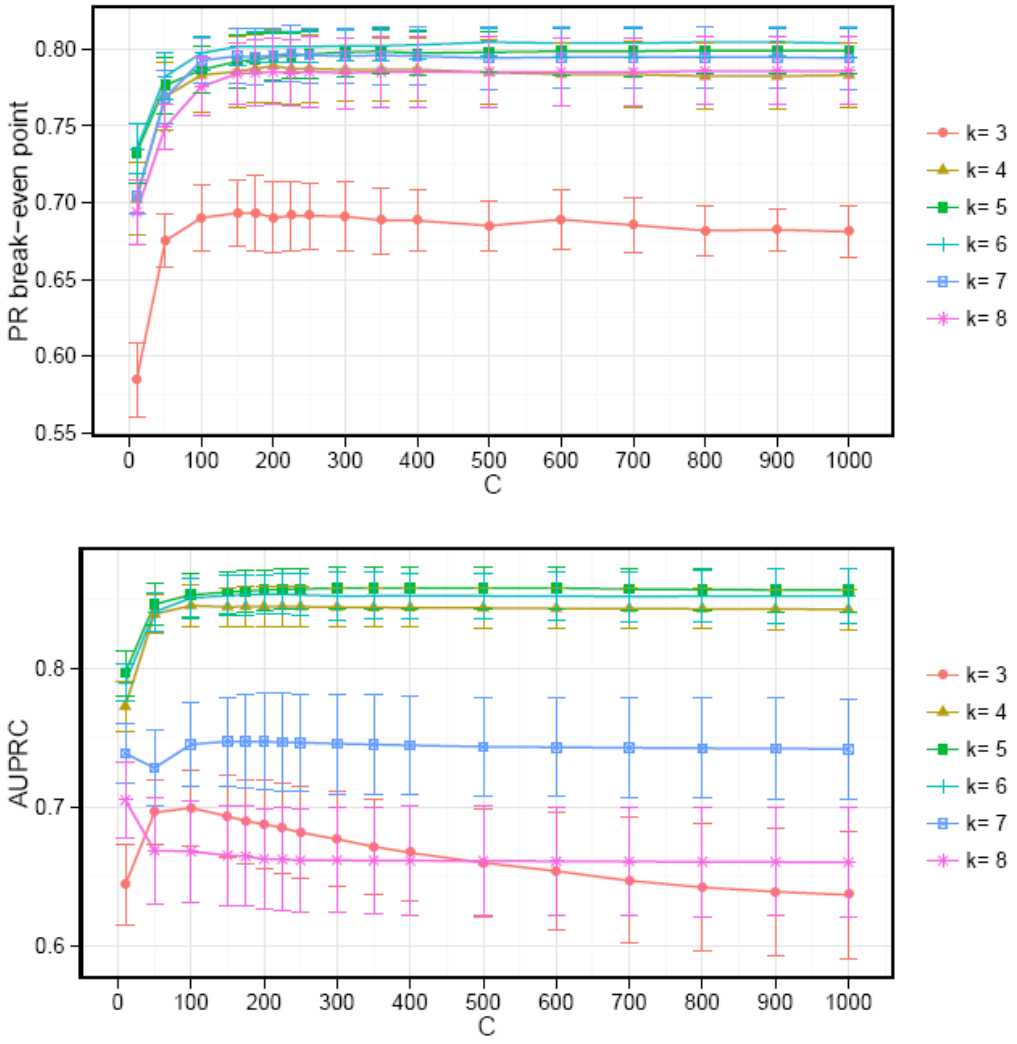


Figure 5-2. AUPRCs and PR break-even points for Allerdictor 10-fold cross-validation on data set C, with the different *k-mer* length (k) and regularization parameter (C) of the SVM model. The error bars show standard deviations of performance scores of 10 fold evaluation.

5.3.2 Allerdictor produces high precision over high recall

We performed nested 10-fold cross-validation to evaluate Allerdictor performance in comparison with the baseline classifiers we derived from FAO/WHO guidelines on our three data sets. In the BLAST method, a protein was classified based on the best BLAST similarity score (e-value) against a database of known allergens from the training set. In MEM (maximal exact matches), the longest subsequence of contiguous amino acid matches against the allergens in the training set was chosen as the classification score. MEM was implemented using SparseMEM software [161].

For all three data sets, both Allerdictor-NB and Allerdictor-SVM performed better than BLAST and MEM with higher precision over the same recall rate as well as larger area under the PR curve (AUPRC) (Figure 5-3). NB and SVM performed equally on data sets A and B, while, with data set C, SVM exhibited better performance. The AUPRCs for Allerdictor-SVM averaged at 0.97, 0.91, and 0.85 for data sets A, B, and C, respectively. BLAST and MEM's performance was acceptable on data sets containing proteins exhibiting low levels of sequence similarity between allergens and non-allergens (AUPRC ≈ 0.7-0.8 for data sets A and B). However, their performance dropped dramatically when the level of sequence similarity between allergens and non-allergens increased (data set C). BLAST appeared to be more vulnerable to a drop in performance with AUPRC ≈ 0.25 and precision rarely reaching 0.5 on data set C. MEM was less vulnerable (AUPRC ≈ 0.63) compared to BLAST yet failed to produce >0.6 precision over >0.6 recall. On the other hand, Allerdictor still yielded high performance on data set C. The AUPRC was ~0.81 for Allerdictor-NB and ~0.85 for Allerdictor-SVM, and both still produced ~0.8 precision over 0.8 recall. Since SVM performed more robustly than NB, studies were concentrated on SVM.

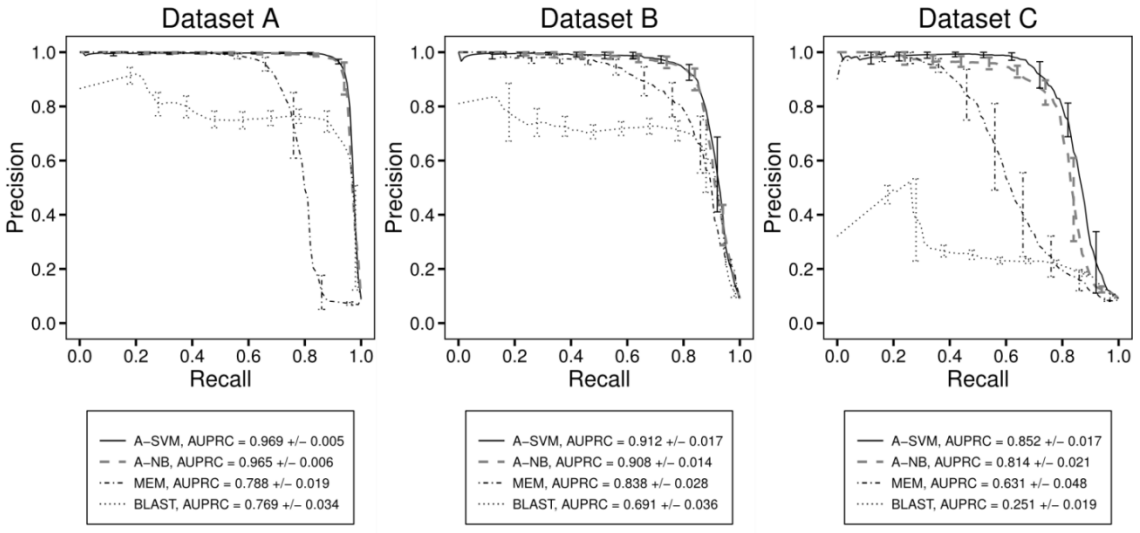


Figure 5-3. PR curves for Allerdictor-SVM (A-SVM) and Allerdictor-NB (A-NB), MEM and BLAST on three data sets of increasing level of sequence similarity between allergens and non-allergens (A, B, and C). The curves were averaged on nested 10-fold cross-validation with standard deviations as error bars.

The capability of Allerdictor to produce high precision over high recall rates is due to its extremely high specificity (low false positive rate). In order to assess Allerdictor specificity, we trained Allerdictor-SVM with each of the three data sets and predicted allergens for the whole Swiss-Prot database. The results confirmed Allerdictor had a high level of specificity with <1% of proteins in the Swiss-Prot database predicted as allergens (Table 5-1). Regardless of the level of similarity between allergens and non-allergens in the training data sets, Allerdictor still predicted <1% of Swiss-Prot as allergens. Homology-based methods often produce many false positives when trained with data sets exhibiting low levels of sequence similarity between allergens and non-allergens. Allerdictor specificity, on the other hand, is consistent.

5.3.3 Allerdictor prediction time is linear

Sequence alignment based approaches, which are also the most accurate current allergen prediction methods, construct features from sequence alignment. Most of the prediction time for a sequence is spent on aligning the sequence against a database of full-length allergen/non-allergen sequences and/or allergen specific peptides. This depends on the length of both the sequence and the database. Moreover, aligning sequences requires non-linear time of the sequences' length, which makes large-scale allergen prediction a relatively time-consuming task.

Allerdictor feature construction and prediction times are both linear in the length of the sequence. Counting frequency of *k*-mers from a sequence can be achieved in time linear in the length of the sequence and does not depend on training data. Prediction time for both NB and SVM has two components. The first one is the time required to look up model parameters (e.g. SVM weights of the linear model) for the *k*-mers generated from the sequence. With proper hashing techniques, the total look up time is also linear of the number of *k*-mers on average. The second component is the time to compute the score of the model (NB or linear SVM), which is also linear in the number of *k*-mers, because it involves only non-zero elements of the sparse *k*-mer frequency vector. Overall, Allerdictor prediction time is linear in the length of the sequence.

Table 5-1. Whole Swiss-Prot (539,616 sequences) scan results for Allerdictor trained with different data sets.

Training data	Predicted allergens	Percent Swiss-Prot	Allergen-related[a]
Data set A	3025	0.56	1069
Data set B	4160	0.77	1109
Data set C	2150	0.40	976

[a]Predicted allergens that are true allergens or annotated with allergen-related keywords in Swiss-Prot

Table 5-2. Running time for 100 random test sequences (T) and whole Swiss- Prot (SP) of 539,616 sequences.

Method	T(s)	SP(h)	Implementation note
Allerdictor	32[a]	0.1	Standalone, implemented in Python
Allerdictor	24	34.5	Web server, implemented in Python, submission via Perl script
AlgPred-d	114	174[b]	Web server, submission via Perl script
AllerHunter	863	1318[b]	Web server, submission via Perl script
AllerHunter	15[c]	24[c]	BLAST, using AllerHunter data
APPEL	3731	5700[b]	Web server, submission via Perl script
EVALLER	4094	6255[b]	Web server, submission via Perl script
SORTALLER	158	241[b]	Web server, submission via Perl script

[a]Including time to read k-mer dictionary from disk. [b]Estimated time based on running time of 100 test sequences. [c]Lower bound estimate (time required to run BLAST against the training sequences).

The running time of Allerdictor (both web server and standalone versions) was estimated in comparison with the other methods (including EVALLER, AlgPred, AllerHunter, APPEL, and SORTALLER) on a random test set of 100 protein sequences (average length of 326 amino acids) and the whole Swiss-Prot database. Since only web server versions of the other methods were available, we wrote scripts to submit sequences one by one to the web servers, and measured the time needed to run 100 sequences, including time for data transmission over the web. For these methods, the estimated time required to run the whole Swiss-Prot database was derived from the time used for 100 sequences. For Allerdictor web server, true running time to scan the whole Swiss-Prot database was measured using a similar submission script. The lower bound for AllerHunter feature construction was also estimated by time required to align sequences against the database of training sequences using BLAST. Allerdictor was extremely fast compared with other methods (Table 5-2). Allerdictor standalone version only took ~6

minutes (on a single core PC) and Allerdictor web server submission took ~34.5 hours to scan the whole Swiss-Prot of 539,616 protein sequences. Web server performance depends on many factors such as web server configuration and internet connection speed, and therefore the rough estimates obtained in Table 5-2 were not necessarily the true performance. However, these estimates should correlate with true running time and were appropriate for comparison. The linear running time in addition to high precision over high recall makes Allerdictor more practical for large-scale allergen discovery compared to existing methods.

5.3.4 Allerdictor distinguishes allergen-related peptides

Table 5-3. Statistics on k-mers learned from three data sets A, B, and C in relation with 183 known IgE epitopes coming from 29 allergens.

Training data	**Total *k*-mers learned**	**IgE epitope-matched *k*-mers**	**Matched IgE epitopes***	**Allergens of matched IgE epitopes**
Data set A	9,082,690	1,238	174	25
Data set B	5,654,846	1,222	174	25
Data set C	4,561,099	1,068	170	25

*Not all epitopes are matched with *k*-mers due to some variations in the collected sequences as well as the removal of some allergen sequences in data preparation.

An IgE epitope is a region of an allergen that can be recognized by and interact with allergen-specific IgE antibodies. It is perhaps the most important allergenicity identification feature. However IgE epitopes exist in both linear form (continuous amino acids) and conformational form (discontinuous amino acids brought together via protein folding), and thus are difficult to model. Sequence similarity approaches in allergen prediction such as those corresponding to the FAO/WHO guideline are centered upon knowledge of the IgE epitope length, which ranges from 3-71 amino acids according to the known IgE epitopes from the structural database of allergen proteins (SDAP) [11]. These approaches, however, cannot distinguish between sequence similarity matches in regions that are related to allergenicity such as the IgE epitopes and those in regions that are commonly found in both allergens and non-allergens, and thus yield low performance.

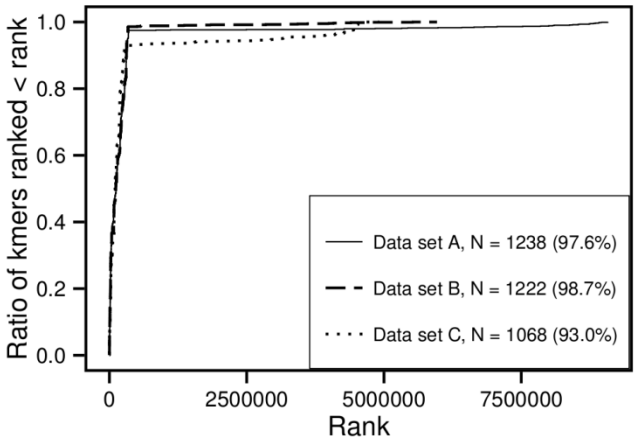


Figure 5-4. Empirical cumulative distribution of ranks of the k-mers (k=6) that are sub-sequences of at least one of 183 known IgE epitopes from SDAP database. The percentage in the brackets is the ratio of k-mers that are ranked in the top 10% of all k-mers obtained from each training set.

Allerdictor is effective in allergen prediction because it is capable of distinguishing allergen-related short peptides (not necessarily IgE epitopes per se). For example, Allerdictor, although it does not directly model IgE epitope structures, can learn and assign higher weight to *k*-mers that are subsequences of known IgE epitopes using a machine learning approach. We investigated this using a set of 183 known IgE epitopes (from 29 allergens) collected from the SDAP database (one of the most updated lists of known IgE epitopes). As given in (5), *k*-mers with higher weight in the linear SVM model represent more allergen predictive features (more commonly found in allergens). We ranked *k*-mers by their weight and investigated the distribution of the ranks of the *k*-mers that were subsequences of at least one known IgE epitope (IgE epitope-matched *k*-mers). We found >1,000 IgE epitope-matched *k*-mers learned from each of the data sets A, B, and C (Table S1). The majority of the IgE epitope-matched *k*-mers (>93%) were ranked in the top 10% among ~4.5-9.1 million *k*-mers obtained from the training data (or 90% of them were ranked in the top 3.4%, 5.5% and 5.9% for data sets A, B, and C, respectively) (Figure 5-4). This result suggests that Allerdictor is capable of assigning higher weight to *k*-mers that are more important for allergenicity such as those found in IgE epitopes than those that are often found in both allergens and non-allergens. For data sets A and B, almost all IgE epitope-matched *k*-mers were ranked among the top while a

small number of these *k*-mers for data set C had low ranks. This can be explained by the fact that data set C only contained a fraction of the known allergens (and IgE epitopes) via the relaxed sequence clustering criteria described earlier.

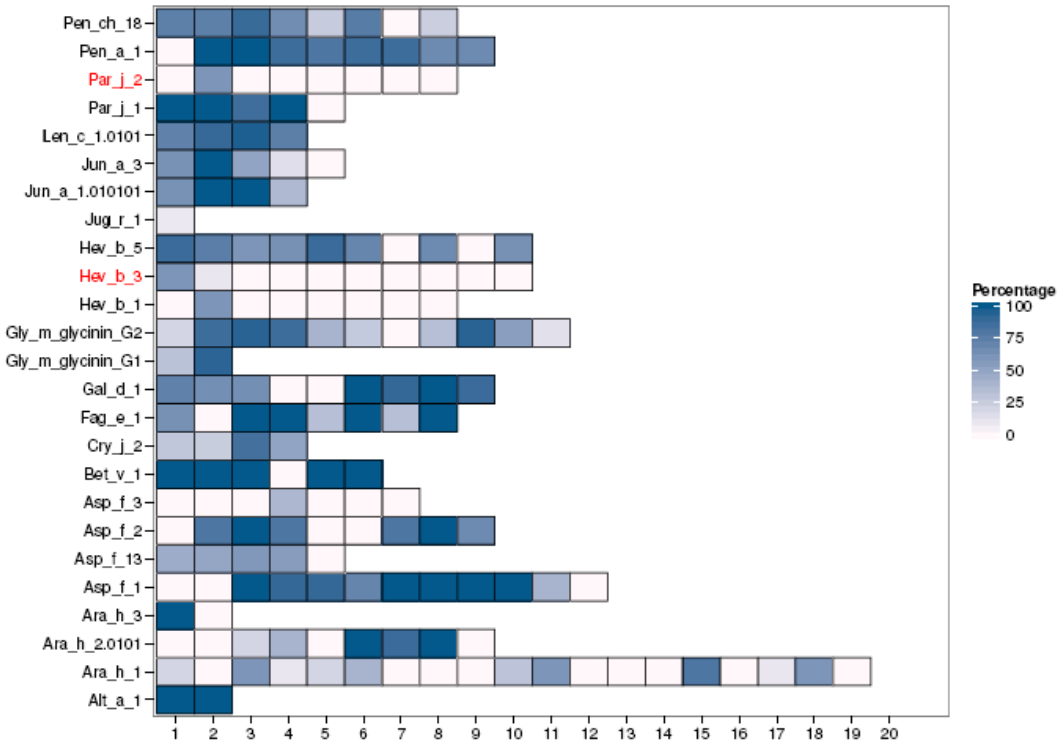


Figure 5-5. Epitope coverage by high scoring regions for allergens with known IgE epitopes. High scoring regions were formed by merging consecutive *k*-mers that ranked in the top 25,000. Each box represents one epitope sequence (color density indicates percentage of an epitope covered by high scored regions). Sequences predicted to be allergen are labeled in black /non-allergen are in red. Training was performed on data set A (known-epitope allergens and sequences that had a BLAST HSP $\geq$ 99% identity with these allergens were removed from training data).

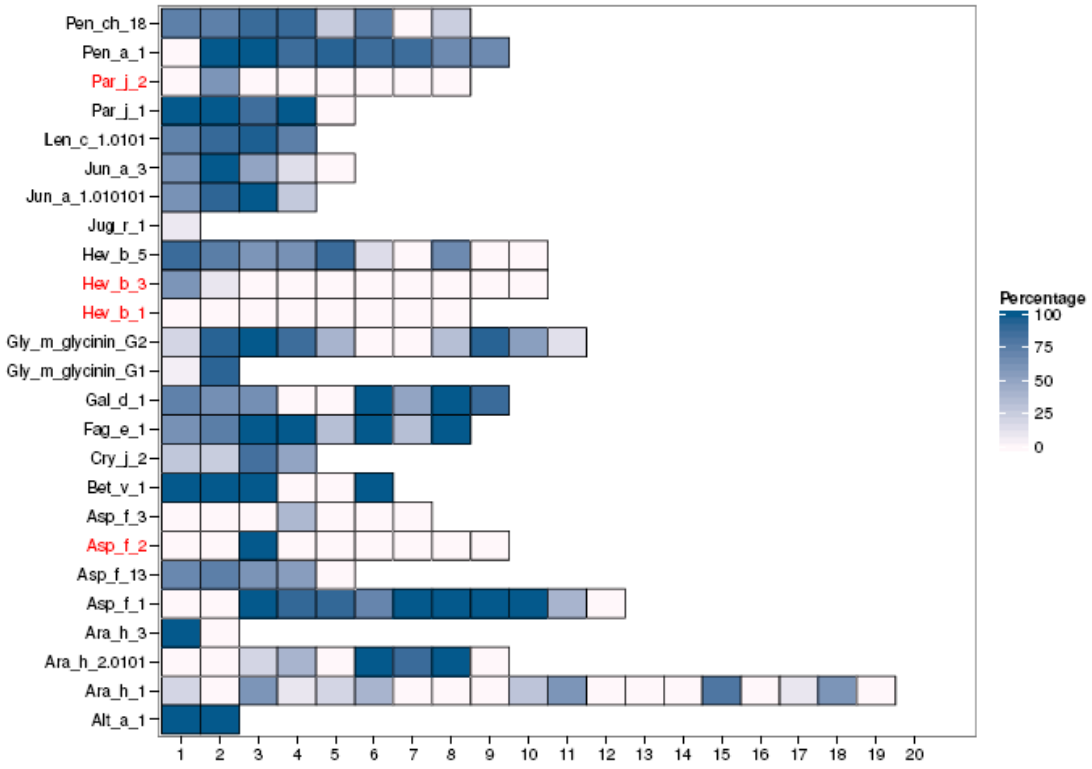


Figure 5-6. Epitope coverage by high scoring regions for allergens with known IgE epitopes. High scoring regions were formed by merging consecutive *k*-mers that ranked in the top 25,000. Each box represents one epitope sequence (color density indicates percentage of an epitope covered by high scored regions). Sequences predicted to be allergen are labeled in black /non-allergen are in red. Training was performed on data set B (known-epitope allergens and sequences that had a BLAST HSP >= 99% identity with these allergens were removed from training data).

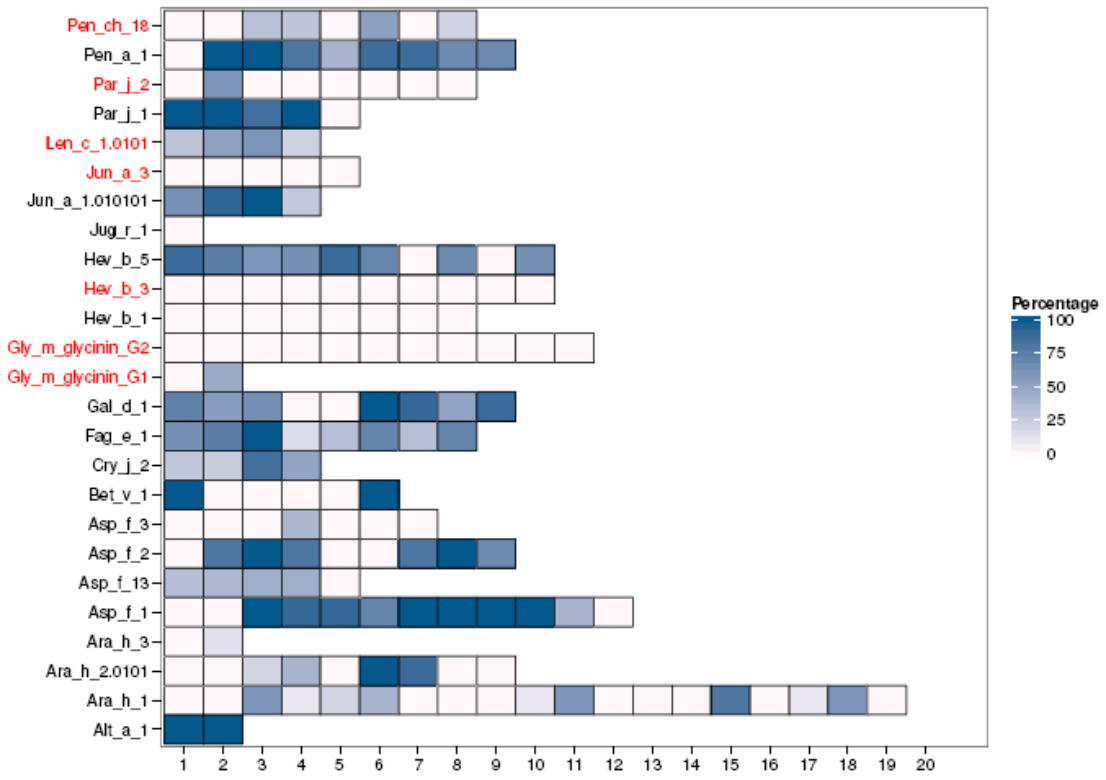


Figure 5-7. Epitope coverage by high scoring regions for allergens with known IgE epitopes. High scoring regions were formed by merging consecutive *k*-mers that ranked in the top 25,000. Each box represents one epitope sequence (color density indicates percentage of an epitope covered by high scored regions). Sequences predicted to be allergen are labeled in black /non-allergen are in red. Training was performed on data set C (known-epitope allergens and sequences that had a BLAST HSP ≥ 99% identity with these allergens were removed from training data).

In fact, many of the highly ranked *k*-mers formed continuous peptides overlapping with known IgE epitopes. We ran Allerdictor on 25 allergen proteins with IgE epitopes previously mapped (prepared from the set of 29 allergens with known IgE epitopes collected from the SDAP database). The majority of the known IgE epitopes overlapped with regions formed by highly ranked *k*-mers and many of them were fully covered by these regions (Figure 5-5, Figure 5-6 and Figure 5-7). This result suggests that the regions of a protein sequence that contain highly ranked *k*-mers have higher probability of being part of IgE epitopes or other immunologically relevant features and thus they are highlighted in the prediction output of Allerdictor server for further computational and/or experimental investigation by the end-users.

5.3.5 Comparison with other methods

Current allergen prediction tools were first evaluated on a set of randomly drawn ~10% of data set C (167 allergens and 1,663 non-allergens, test set X). For methods that produced monotonous prediction scores (AlgPred, AllerHunter, SORTALLER), the score cutoff was varied to obtain PR curves. For other methods (EVALLER, APPEL), fixed default performance measures were calculated from the number of correct and incorrect predictions. The results showed that all methods evaluated yielded low precision on the chosen test set (Table 5-4, Figure 5-8). None of the methods yielded precision >0.4 over recall >0.6 for PR curves. For default performance, only EVALLER and AllerHunter yielded MCC >0.5 with both precision and recall >0.5. Sequence similarity based methods (AllerHunter and EVALLER) appeared to perform better in this test. As expected, performance was correlated with the time the methods were released, where later methods performing better (with the exception that SORTALLER performed poorly although it was the latest method in this test). AllerHunter performed better than other methods, partly because it was trained on a data set that contained many allergen-like non-allergens, a characteristic that was also exhibited by the test data.

Table 5-4. Default performance measures of current methods on test set X of 167 allergens and 1,663 non-allergens randomly drawn from data set C.

Method	TP	FP	TN	FN	Recall	Precision	Accuracy	F1	MCC
AllerHunter*	90	43	1620	70	0.56	0.68	0.94	0.61	0.58
AlgPred-c	133	860	803	34	0.80	0.13	0.51	0.23	0.16
AlgPred-d	125	766	897	42	0.75	0.14	0.56	0.24	0.17
APPEL*	67	49	1614	93	0.42	0.58	0.92	0.49	0.45
EVALLER*	106	100	1560	61	0.63	0.51	0.91	0.57	0.52
SORTALLER	144	340	1323	23	0.86	0.30	0.80	0.44	0.43

TP -true positive, FP -false positive, TN -true negative, FN -false negative, MCC -Mathews correlation coefficient; *server returned error on a few (<10) sequences

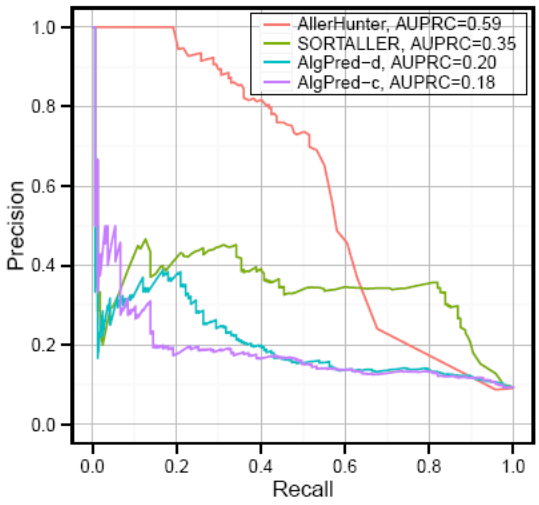


Figure 5-8. PR curves for AllerHunter, AlgPred composition (AlgPred-c), AlgPred dipeptide (AlgPred-d), and SORTALLER on a test set of 167 allergens and 1,663 non-alergens randomly drawn from data set C.

Because performance of supervised machine learning methods depends heavily on training and testing data, we avoided comparison of Allerdictor with other methods trained with different data sets. Current allergen prediction methods were pre-trained with specific data sets and only available in the form of web servers, and thus prevented re-training them for comprehensive comparison with Allerdictor. Therefore, we investigated these data sets on whether they are appropriate to train and compare Allerdictor with the pre-trained web servers of these methods. Among three publicly available data sets, AllerHunter was the only data set that possessed a significant level of sequence similarity between allergens and non-allergens and had many more non-allergens than allergens (Figure 4-2). The AlgPred data set was small and sequence names were masked while the EVALLER non-allergen sequences that were used to derive allergen specific peptides were not available. The level of sequence similarity between allergens and non-allergens for AlgPred and EVALLER was very low as determined by BLASTClust (Figure 4-2).

The AllerHunter data set was considered the only complete data set for the purpose of the comparison. However, this data set was redundant and contained noise (obsolete sequences and allergens mislabeled as non-allergens). The AllerHunter data set was reviewed and found to contain 48 obsolete sequences (deleted from Genbank and Swiss-Prot), 233 duplicated sequences and 328 short sequences (3-50 amino acids). Also,

among the non-allergens, 135 were found to be true allergens, 176 were antigens and 165 contained allergen-related ambiguous annotation (from Swiss-Prot and Allergome databases). The noise possibly resulted from new annotation added to the public databases after AllerHunter data were collected. We accepted the noise for training data and trained Allerdictor on the same training set of sequences (1,266 allergens and 11,229 non-allergens) that were used to train AllerHunter server (personal communication with Martti Tammi), and compared Allerdictor with AllerHunter server on two test sets: the original AllerHunter test set (139 allergens and 1,245 non- allergens) that contained noise, and the revised version of the test set with reduced noise (149 allergens and 1,141 non-allergens). The review process moved the newly discovered allergens from the non-allergen set to the allergen set, and removed duplicated, obsolete, or ambiguous sequences and non-allergen sequences that had ≥ 90% identity over ≥ 90% coverage with a known allergen (similar to the procedure used to reduce noise from Allerdictor data sets).

Comparison results on AllerHunter data set showed that Allerdictor slightly outperformed AllerHunter with slightly larger AUPRCs (Figure 5-9). An interesting trend was that Allerdictor produced higher high-range precision (>0.8) at lower recall (<0.8). At recall >0.85, both Allerdictor and AllerHunter produced many false positives and thus the precision for both methods dropped below 0.6. AllerHunter performed slightly better in lower-range precision (<0.75) at a very narrow recall range from ~0.85-0.9. High-range precision is particularly useful in large-scale prediction. For example, one often chooses the top scoring candidates from computational predictions for further experimental validation, which is equivalent to lowering recall to obtain higher precision. Along with higher high-range precision, Allerdictor also runs much faster than AllerHunter (see 3.3) which makes it the better choice for large-scale allergen prediction.

The amino acid composition and dipeptide approaches in AlgPred are special cases of the *k*-mer approach in Allerdictor with $k = 1$ and 2. We found that such small values of *k* yielded very low performance on multiple data sets, including data sets A, B and C. Also pointed out by AlgPred authors, when tested with Swiss-Prot non-allergens, AlgPred falsely predicted ~40% of them to be allergens [166].

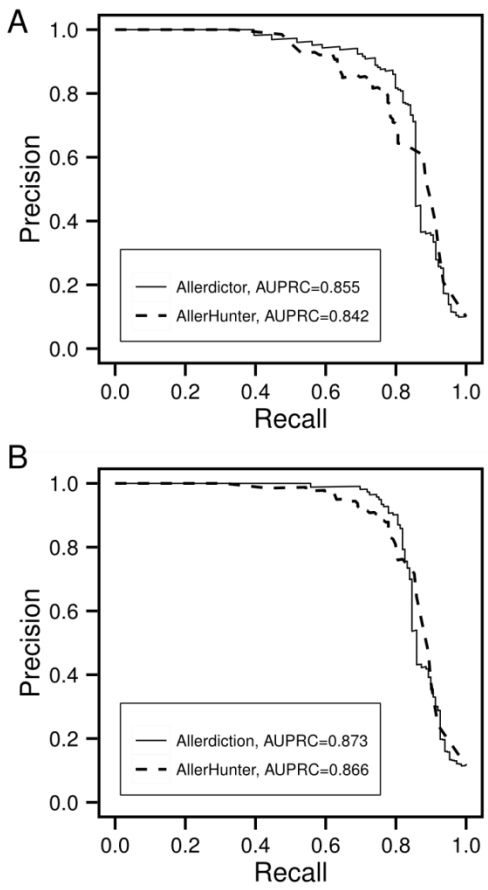


Figure 5-9. PR curves for Allerdictor and AllerHunter, both trained on the original AllerHunter training set and tested with the original AllerHunter test set (A) and the reviewed AllerHunter test set (B).

5.3.6 Allerdictor prefers larger number of *k*-mers

One of the drawbacks that prevents many machine learning approaches from utilizing the *k*-mer sequence representation is the size of the *k*-mer dictionary (also the feature vector size) is exponential in *k* (20^k) and can become very large as *k* increases. In reality, the size of the *k*-mer dictionary depends on training data sets and is often smaller than the number of possible *k*-mers. Allerdictor when trained with $k = 6$ on data sets A, B, and C had feature space dimension of approximately 4.6, 5.7, and 9.1 million, respectively. To test if we can reduce the number of *k*-mers without lowering

performance, feature selection using mutual information and feature abstraction (as described in 5.2.4) were performed with Allerdictor-SVM using $k = 6$.

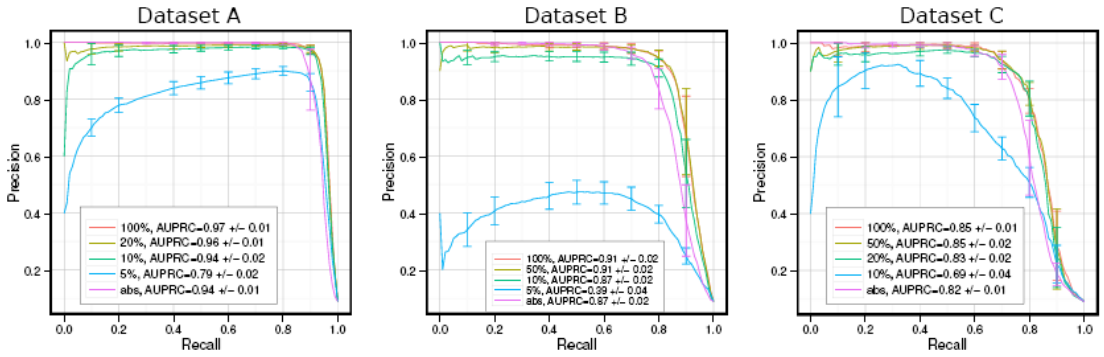


Figure 5-10. PR curves for mutual information based feature selection (at 5%-100% top *k*-mers selected) and feature abstraction (abs) on data sets A, B, and C. The curves are average of 10-fold cross-validation with standard deviations as error bars.

The results on data sets A, B, and C showed that no performance gain was achieved with both feature selection and feature abstraction (Figure 5-10). Using ≥20-50% *k*-mers, performance was very similar to that obtained with all *k*-mers. Allerdictor performance slightly dropped when the number of selected *k*-mers ≤10-20% and dramatically dropped when ≤5-10% *k*-mers were selected. This result suggests that Allerdictor generally performs better with more *k*-mers. Feature abstraction reduced ~4.6-9.1 million *k*-mers down to ~1,000 abstract features when trained using $k = 6$. Surprisingly, performance for feature abstraction was very close to the performance using all *k*-mers. This interesting result opens doors for using other classification methods that can only handle a small number of features.

5.3.7 Effects of allergen prevalence

Supervised machine learning based allergen prediction methods are often available to end-users as tools pre-trained on some specific data set. The predictive values including positive predictive value (PPV, also called precision) and negative predictive value (NPV) of such tools are subject to the prevalence of allergens in data. We have shown that Allerdictor produced high precision over high recall when training and testing using data that exhibited low ratios of allergen sequences.

To provide a complete picture of allergen prediction performance, we also investigated the effects of allergen prevalence (in testing data) on PPV and NPV of Allerdictor and the current allergen prediction tools. As expected, predictive values of all methods were affected by the prevalence of allergens in testing data (Figure 5-11, Figure 5-12, and Figure 5-13). When the prevalence of allergens was low, AllerHunter, APPEL, and EVALLER exhibited higher predictive values than AlgPred and SORTALLER on a random set of sequences drawn from data set C (Figure 5-12). Allerdictor exhibited stable PPV on data sets A, B, C (Figure 5-11) as well as on AllerHunter data set (Figure 5-13). When allergen ratio ≤ 0.5, Allerdictor achieved both PPV and NPV ≥ 0.8 in all data sets. In comparison with AllerHunter on AllerHunter data set, Allerdictor PPV and NPV were better when allergen prevalence was low but AllerHunter exhibited more balanced PPV and NPV when the ratio of allergens was higher. The NPV of all methods including Allerdictor decayed rapidly as the ratio of allergens in the test sets increased. However, this behavior does not significantly limit the application of machine learning based allergen prediction methods because allergen prevalence is low in nature as well as in many applications. For example, there exist ~20 known allergens among >9,000 proteins coded by the genome of the allergenic fungus *Aspergillus fumigatus* (Fedorova et al., 2008). Low allergen ratio is a characteristic of large sequence sets often seen in large-scale sequence annotation that is also Allerdictor main application.

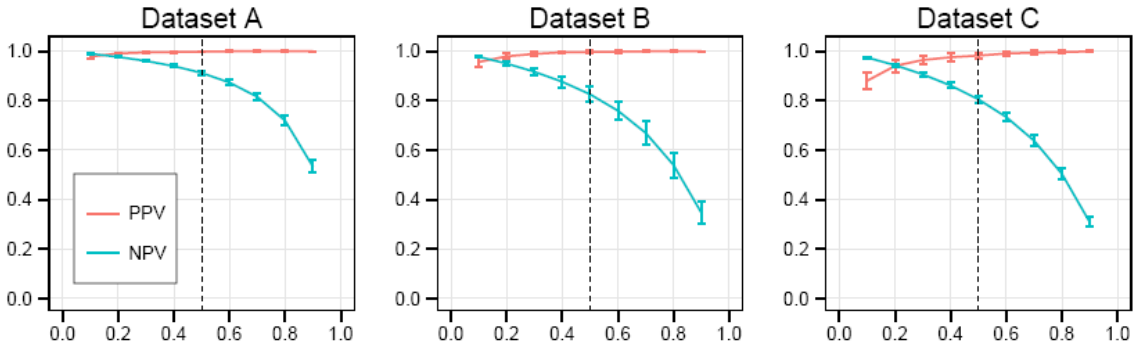


Figure 5-11. Allerdictor positive predictive value (PPV) and negative predictive value (NPV) in relation to the ratio (prevalence) of allergens in test sets when trained and tested on data sets A, B, and C using nested 10-fold cross-validation. Error bars represent standard deviation of 10-fold nested cross-validation.

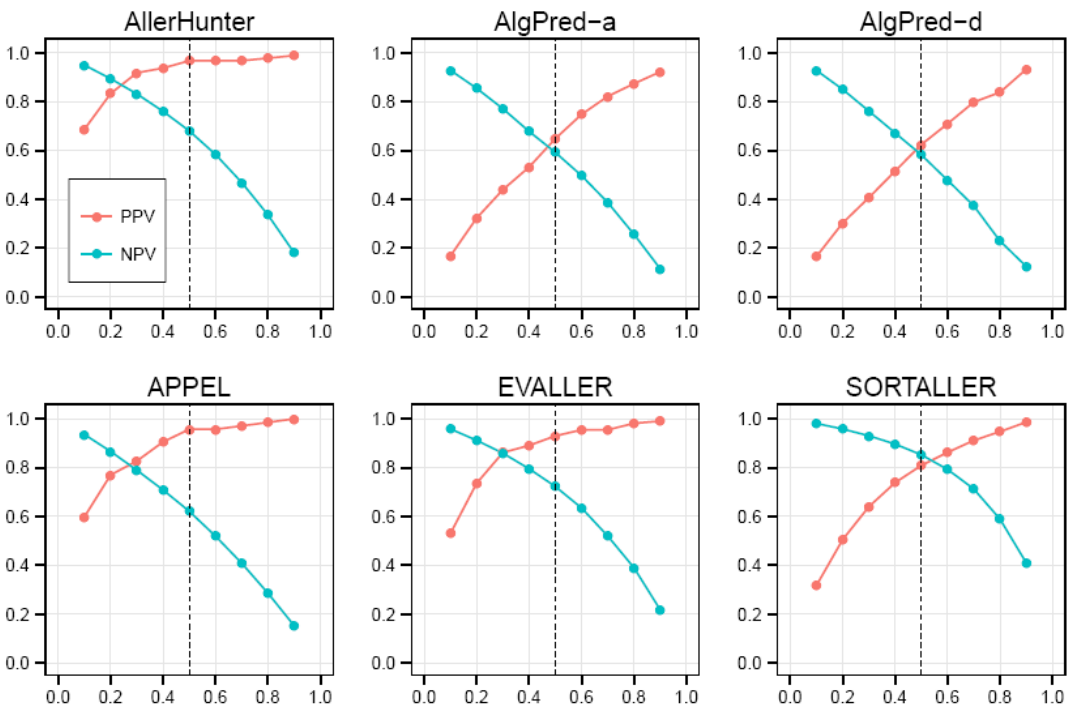


Figure 5-12. Positive predictive value (PPV) and negative predictive value (NPV) of current allergen prediction tools in relation to the ratio (prevalence) of allergens in test sets. The tools were evaluated using a test set of 167 allergens and 1,663 non-allergens randomly drawn from data set C (test set X used in section 5.3.5).

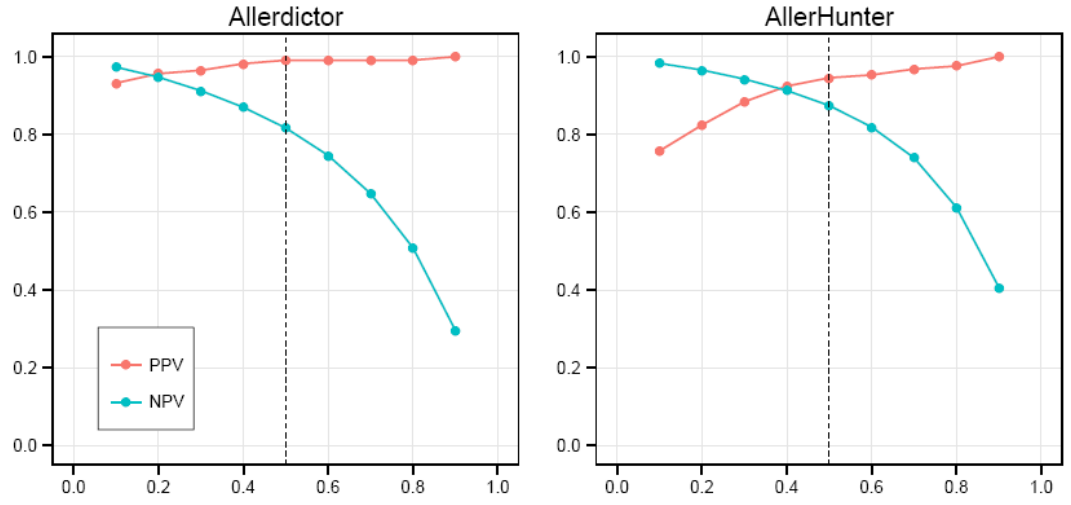


Figure 5-13. Positive predictive value (PPV) and negative predictive value (NPV) of Allerdictor and AllerHunter in relation to the ratio (prevalence) of allergens in test sets. Both methods were trained using AllerHunter training set and evaluated using the revised AllerHunter test set.

5.4 Conclusion

This chapter presented an accurate sequence-based allergen protein prediction method (Allerdictor) that is much faster than the current most accurate methods while still maintaining comparable or better predictive performance (when compared with AllerHunter). The main idea is the use of the *k*-mer feature representation of sequences and thus linear prediction time is achieved for both feature construction and prediction using a linear SVM model. Moreover, the *k*-mer approach is particularly effective for allergen prediction because supervised machine learning methods such as SVM can learn the *k*-mers shared by many allergens such as the ones found in IgE epitopes and assign higher weights to these *k*-mers.

The prevalence of asthma has been an increasing human health issue. Approximately 235-300 million people worldwide were diagnosed with asthma with annual deaths of ~250,000 [1, 2]. The majority of asthmatic patients have allergic asthma in which allergic reactions (caused by allergens) exacerbate asthmatic symptoms. To facilitate our understanding and prevention of this disease, it is important to identify potential allergens from massive amounts of protein sequences produced every day via both genome sequencing and sequence synthesis. While experimental allergenicity assessment is still expensive and difficult (especially at large-scale), computational allergen identification is an alternative first step.

Allerdictor addresses the shortcomings of the current allergen prediction tools. With high precision over high recall and fast speed, Allerdictor is not only useful for general sequence allergenicity assessment in applications such as screening of novel proteins introduced to genetically modified crops but also particularly suitable for allergen discovery on a large-scale in applications such as whole genome annotation and quick screening of synthesized sequences.

Chapter 6

Analysis of mold allergens

6.1 Introduction

Fungi are potent sources of allergenic molecules including both allergens and cross-reactive proteins. Indoor and outdoor exposure to fungal spores is a triggering factor for respiratory allergy and asthma. The ability of fungi to germinate and grow at human body temperature allows them to have longer interaction with humans than other allergens like pollen or dustmites. Most of the allergenic fungi belong to the *Ascomycete* phylum (mold). It was reported that approximately 80 mold genera can induce type I hypersensitivity in atopic individuals. A number of allergens have been characterized in these fungi (Table 6-1). *Aspergillus fumigatus* is one of the most extensively studied allergenic fungi. More than 20 clinically validated allergens have been characterized in this species. *Alternaria alternata* is another important allergenic fungus, and more than 10 allergens have been clinically characterized in this species.

Although many allergens have been experimentally characterized and clinically validated in various molds, it is possible that more potential allergens and cross-reactive proteins in these species remain undiscovered. A recent survey of allergen homologs revealed that fungi harbor many allergen-like proteins or homologs. Experimental methods to detect allergens at a whole genome scale was also used but with limited success [146, 147]. Using a cDNA phage display, a recent study detected 81 cDNA clones from *A. fumigatus* that produce IgE-binding proteins [180]. However, it was not reported how many of these cDNA clones corresponded to different genes; a unisequence set was not established in the study. Many fungal genomes have been sequenced over the past few years, including a number of allergenic fungi described above, that now allows for systematic studies of their allergen repertoire.

In this chapter, we performed a computational survey of the allergomes of the WHO/IUIS recognized allergenic mold genera that have genome sequence available, using comparative and machine learning approaches that we developed and evaluated in Chapter 4 and Chapter 5.

6.2 Methods

6.2.1 Fungal protein sequences

The protein sequences of three *Alternaria* fungi are the output of the *Alternaria* genome annotation pipeline that we used to annotate *Alternaria* genomes. Protein sequences for *Aspergillus* species were downloaded from Ensembl fungi (http://fungi.ensembl.org) and protein sequences for other fungi were downloaded from the DOE-JGI Genome Portal for the JGI fungal genomics program (http://genome.jgi.doe.gov/programs/fungi).

6.2.2 Categorizing fungal proteins

The list of widely accepted mold allergens was compiled from the WHO/IUIS Allergen Nomenclature Subcommittee website (http://allergen.org). Corresponding sequences of these allergens were downloaded from Swiss-Prot and NCBI. These sequences were searched against the database compiled from all mold proteins to identify sequences with very high similarity (≥ 90% identity and coverage on the query) coming from the same fungus and designated as known allergens.

Based on sequence similarity with the set of known allergens from data set A described in Chapter 4, a predicted allergen was also classified into three groups: (i) *highly similar* sequence had ≥90% sequence identity and ≥90% sequence coverage in an alignment with a known allergen; (ii) *similar* sequence had ≥70% (but <90%) sequence identity and coverage with a known allergen; and (iii) *different* sequence had <70% sequence identity or coverage with a known allergen.

6.2.3 Training data and prediction criteria

In Chapter 4, we have evaluated the effects of different sequence similarity criteria for allergen prediction in the modified FAO/WHO approach (comparative approach). The optimal combinations of comparative criteria were dependent on data sets used. Among the three data sets, data set B had the lowest level of sequence similarity/duplication, and therefore the parameters chosen using this data set were the most appropriate for predicting allergens from newly sequenced genomes. Based on evaluation on data set B, we chose bit scores ≥ 0, sequence identity over 80 amino acids ≥ 65 and maximal exact match ≥ 17 as the combined criteria to identify a protein as a potential allergen.

Data set A (described in Chapter 4) was used as the training data for Allerdictor and the allergen sequences from data set A were used as the allergen database in the comparative allergen classifier. The reason for choosing data set A was because this data set contained the largest number of known allergens (although redundant) when compared with other data sets.

In the comparative approach, BLAST was used to find the maximum 80 aa identity and bit score and SparseMEM was used to find maximal exact matches for a fungal protein. A protein was classified as an allergen if it was found to have either 80 aa identity ≥ 65 or maximal exact matches ≥ 17.

6.3 Results and discussion

We were able to map most of the known mold allergens to the proteins in the collected proteomes of the corresponding mold species (Table 6-1). One allergen in *A. alternata* (Alt a 14) did not show high level of sequence similarity (≥ 70%) with the proteins we predicted from the genome.

Table 6-1. Known allergenic mold genera and species that have been sequenced

Species	**Number of proteins**	**Number of WHO/IUIS allergens**	**Number of WHO/IUIS allergens mapped in the proteome**
Alternaria alternata ATCC 66981	11635	11	10
Alternaria alternata ATCC 11680	12323	11	10
Alternaria brassicicola[*]	10514	0	0
Aspergillus flavus	13487	1	0
Aspergillus fumigatus a1163	9916	23	22
Aspergillus fumigatus af293	9630	23	22
Aspergillus niger	14068	3	2
Aspergillus oryzae	12074	2	3
Aspergillus versicolor	13228	1	1
Candida albicans SC5314	6221	2	3
Candida albicans WO-1	6160	2	3
Cladosporium fulvum[*]	14127	0	0
Cochliobolus lunatus	12131	4	3
Penicillium brevicompactum	11536	2	1
Penicillium chrysogenum	11396	6	6
Penicillium oxalicum	9979	1	1
Trichophyton rubrum	8707	2	2
Trichophyton tonsurans	8521	2	0

Some allergens are not found or multiple copies are found in the sequenced genomes
***** These fungi are not recognized as allergenic fungi, but belong to a genus that has IUIS recognized allergenic species.

6.3.1 Comparative analysis of fungal allergens

Using the modified FAO/WHO approach outlined in Chapter 4 and the set of known allergens from data set A as a database for sequence comparison, we surveyed the genomes of 18 fungi of the known allergenic genera in *Ascomycota* phylum (Table 6-1). Each predicted allergen was then tagged as *'characterized'* allergens (if it was classified as a known allergen by the WHO/IUIS allergen nomenclature sub-committee) or *'uncharacterized'* allergens (if it was not classified as a known allergen by the WHO/IUIS allergen nomenclature sub-committee). The predicted allergens were also tagged with a status of either *'highly-similar', 'similar', or 'different'* based on sequence similarity with known allergens in the database (see methods).

As expected, all known allergens were recovered using this comparative approach. The results also demonstrated that the surveyed fungal genomes harbored more potential allergens in addition to the known allergens that have been experimentally studied and classified (Figure 6-1, Table 6-2). These potential allergens included predicted proteins that had significant sequence similarity with known allergens (categories novel/highly similar and novel/similar) and predicted proteins that showed low levels of sequence similarity with known allergens (categories novel/different). This approach also predicted one allergen from *A. alternata* (found in both strains surveyed) that was recently classified in WHO/IUIS but not included in the knowledge set of known allergens.

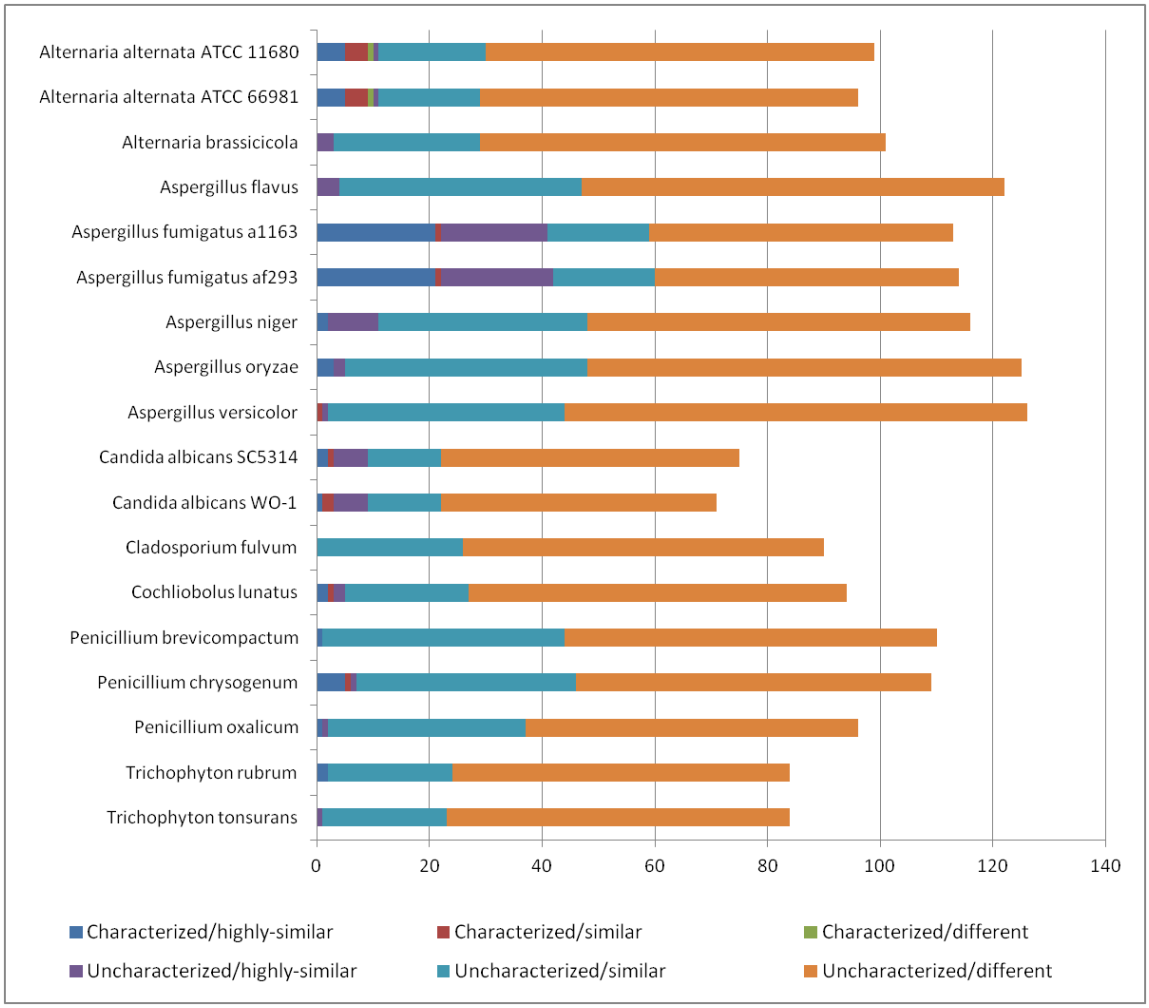


Figure 6-1. The number of allergens identified from known allergenic mold fungi.

Table 6-2. Summary of allergens predicted using the modified FAO/WHO approach

Species	**Characterized**			**Uncharacterized**			**Total**
	highly-similar	similar	different	highly-similar	similar	different	
Alternaria alternata ATCC 11680	5	4	1	1	19	69	99
Alternaria alternata ATCC 66981	5	4	1	1	18	67	96
Alternaria brassicicola	0	0	0	3	26	72	101
Aspergillus flavus	0	0	0	4	43	75	122
Aspergillus fumigatus a1163	21	1	0	19	18	54	113
Aspergillus fumigatus af293	21	1	0	20	18	54	114
Aspergillus niger	2	0	0	9	37	68	116
Aspergillus oryzae	3	0	0	2	43	77	125
Aspergillus versicolor	0	1	0	1	42	82	126
Candida albicans SC5314	2	1	0	6	13	53	75
Candida albicans WO-1	1	2	0	6	13	49	71
Cladosporium fulvum	0	0	0	0	26	64	90
Cochliobolus lunatus	2	1	0	2	22	67	94
Penicillium brevicompactum	1	0	0	0	43	66	110
Penicillium chrysogenum	5	1	0	1	39	63	109
Penicillium oxalicum	1	0	0	1	35	59	96
Trichophyton rubrum	2	0	0	0	22	60	84
Trichophyton tonsurans	0	0	0	1	22	61	84

6.3.2 Analysis of fungal allergens using machine learning approach with Allerdictor

Using Allerdictor trained with data set A, we predicted fewer number of allergens for each of the 18 mold species genomes compared to the comparative approach. The distribution of the predicted allergens among species was similar to the results obtained with the comparative approach, with *Aspergillus* exhibiting the most number of allergens, followed by *Penicillium* and *Alternaria*.

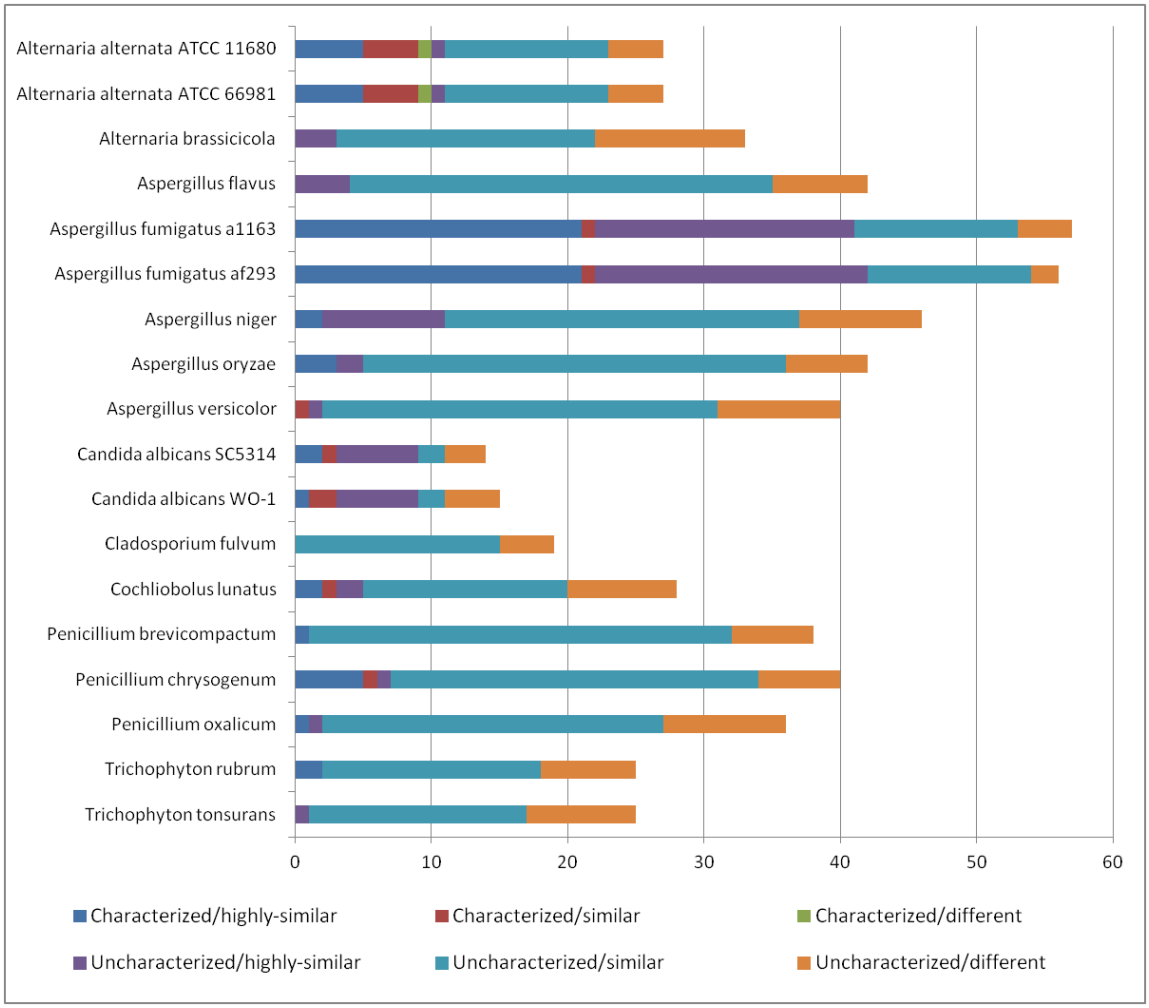
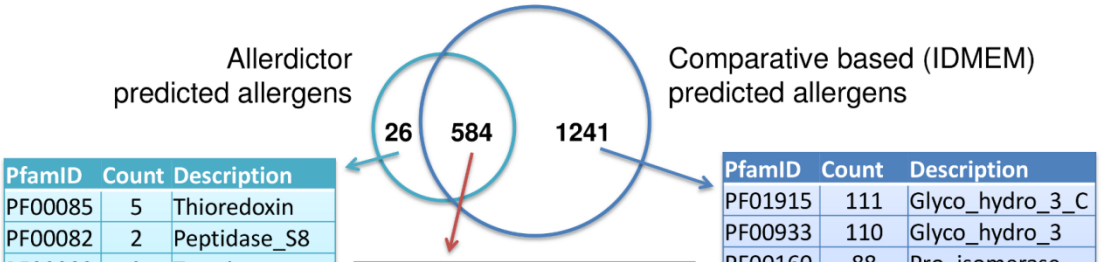


Figure 6-2. The number of allergens identified from known allergenic mold fungi using Allerdictor.

Table 6-3. Summary of allergens predicted using Allerdictor

Species	**Characterized**			**Uncharacterized**			**Total**
	highly-similar	similar	different	highly-similar	similar	different	
Alternaria alternata ATCC 11680	5	4	1	1	12	4	27
Alternaria alternata ATCC 66981	5	4	1	1	12	4	27
Alternaria brassicicola	0	0	0	3	19	11	33
Aspergillus flavus	0	0	0	4	31	7	42
Aspergillus fumigatus a1163	21	1	0	19	12	4	57
Aspergillus fumigatus af293	21	1	0	20	12	2	56
Aspergillus niger	2	0	0	9	26	9	46
Aspergillus oryzae	3	0	0	2	31	6	42
Aspergillus versicolor	0	1	0	1	29	9	40
Candida albicans SC5314	2	1	0	6	2	3	14
Candida albicans WO-1	1	2	0	6	2	4	15
Cladosporium fulvum	0	0	0	0	15	4	19
Cochliobolus lunatus	2	1	0	2	15	8	28
Penicillium brevicompactum	1	0	0	0	31	6	38
Penicillium chrysogenum	5	1	0	1	27	6	40
Penicillium oxalicum	1	0	0	1	25	9	36
Trichophyton rubrum	2	0	0	0	16	7	25
Trichophyton tonsurans	0	0	0	1	16	8	25

Compared with the comparative approach, Allerdictor predicted an average of ~3 times fewer allergens for each species. For example, Allerdictor predicted 27 allergens in *A. alternata* ATCC 66891 compared with 97 allergens predicted by the comparative approach (Table A-12). Most allergens identified using Allerdictor were also identified by the comparative approach (Figure 6-3). Only 26 proteins were identified as allergens by Allerdictor but not classified as allergens by the modified FAO/WHO comparative analysis. These predicted allergens together with other predicted allergens with low level of sequence similarity with known allergens represent possible novel allergens that deserve further experimental validation.



PfamID	Count	Description
PF00085	5	Thioredoxin
PF00082	2	Peptidase_S8
PF00089	2	Trypsin
PF00081	1	Sod_Fe_N
PF00428	1	Ribosomal_60s
PF00838	1	TCTP
PF02777	1	Sod_Fe_C
PF05922	1	Inhibitor_I9

PfamID	Count	Description
PF00082	38	Peptidase_S8
PF05922	36	Inhibitor_I9
PF00160	34	Pro_isomerase
PF00428	30	Ribosomal_60s
PF00171	22	Aldedh
PF00199	22	Catalase
PF06628	22	Catalase-rel
PF00389	20	2-Hacid_dh
PF00081	19	Sod_Fe_N
PF00297	19	Ribosomal_L3

PfamID	Count	Description
PF01915	111	Glyco_hydro_3_C
PF00933	110	Glyco_hydro_3
PF00160	88	Pro_isomerase
PF00012	76	HSP70
PF00171	65	Aldedh
PF14310	56	Fn3-like
PF00128	44	Alpha-amylase
PF00400	40	WD40
PF00722	37	Glyco_hydro_16
PF01565	32	FAD_binding_4

Figure 6-3. Predicted allergens shared/not shared by the comparative approach and Allerdictor and top 10 PFAM domains found in these predicted allergens.

6.4 Conclusion

Using computational analyses, we have identified a significantly higher number of allergens in mold genomes than the number of WHO/IUIS characterized allergens. Among uncharacterized allergens identified, many of them do not have significant sequence similarity with existing known allergens. The modified FAO/WHO comparative approach predicted more allergens than the machine learning based approach with Allerdictor. Allerdictor helped reduce the potential allergens in a species to manageable numbers for subsequence experimental functional analyses. A further step is to experimentally validate the predicted set of allergens using a targeted sera screen as well as animal models. This task is deemed feasible because for a fungus, Allerdictor shortened the list of candidates to 30-60 predicted allergens.

Chapter 7

Conclusions and Future Perspectives

The contributions of this dissertation in the field of bioinformatics and fungal genomics can be summarized in several aspects.

Firstly, we have used various computational approaches to analyze genomic data from a variety of fungi, including a number of important necrotrophic plant parasites and human opportunistic respiratory pathogens of the *Alternaria* genus. These approaches employ available bioinformatics tools in genome annotation and comparison that were incorporated into a comprehensive fungal genome annotation and comparison pipeline and have been extensively used to analyze *Alternaria* genomes. The *Alternaria* genomes database made annotated genome data analyzed in this research available to the public. This database has recently facilitated a number of large-scale functional genomics studies that resulted in the discovery of many new genes that contribute to pathogenicity of *Alternaria* species.

Secondly, we have discovered important genetic components of *Alternaria* species that may help explain this genus' prolific ubiquitous saprophytic lifestyle. Moreover, these components may be linked to in some respects to pathogenicity on plants and human allergic airway diseases. For example, *Alternaria* species in general possesses many more carbohydrate active enzymes (CAZY) than other fungi that correspond to an exceptional range of substrates that can be utilized by this fungus as nutrition sources. In particular, *A. alternata* harbors the largest and most extensive collection of carbohydrate active enzymes in a comparison of 17 fungi of different lifestyles. We also confirmed this finding experimentally by growing these fungi on various single carbohydrate sources in vitro. Thus, based on our collective data (both computational and experimental), we feel that *Alternaria* may be an excellent source of genes encoding industrial enzymes useful

in a wide variety of fields or applications including biomass deconstruction (cellulases, ligninases, xylanases, pectate lyases, feruoyl esterases etc.), detergents (lipases, proteases), allergen immunotherapy (recombinant allergens), and the pulp/paper industry (xylanaes, feruoyl esterases, peroxidases, etc.). We also found that *A. alternata*, when compared to a closely related species *A. brassicicola*, possesses a more extensive set of proteolytic enzymes. This may be one important factor in the clinical linkage of *A. alternata* but not *A. brassicicola* to moderate to severe allergic airway disorders like asthma in humans.

Thirdly, we have evaluated different approaches in computational prediction of allergenicity including comparative genomics and machine learning approaches, and developed a machine learning based tool (Allerdictor) to tackle the problem of identifying allergens from large-scale data (e.g., whole genome annotation). This work has addressed several of the shortcomings of the current allergen prediction tools including improvements in precision and prediction speed on large data. To the best of our knowledge, this is the first time the problem of rapid large-scale allergen identification is considered thoroughly.

Lastly, we have identified an extensive pan-allergen repertoire of the widely known allergenic mold fungi including *Aspergillus*, *Alternaria,* and *Candida*, using the approach that we developed. Each allergenic fungal species harbored an average of ~34 allergens with *Aspergillus* having the largest collection of allergens (e.g., >55 allergens were found in both *A. fumigatus* strains analyzed). Many of the uncharacterized predicted allergens might now be the basis for further experimental validation in the future. As the number of sequenced genomes increases rapidly, the approaches described here has the potential of becoming a useful tool to study the genomes of other human allergenic pathogens besides fungi.

Appendix A

Supplementary tables

Table A-1. Whole genome pairwise alignment statistics between *Alternaria* species: *A. alternata* ATCC 66981 (*Aa1*), *A. alternata* ATCC 11680 (*Aa2*), and *A. brassicicola* (*Ab*) genomes

	Ab vs. Aa1		***Ab vs. Aa2***		***Aa1 vs. Aa2***	
	Ab	***Aa1***	***Ab***	***Aa2***	***Aa1***	***Aa2***
Genome size (bases)	31,974,449	33,236,566	31,974,449	33,752,310	33,236,566	33,752,310
Total aligned bases	27,838,765	29,518,871	27,848,640	29,678,726	32,620,578	32,655,959
Total identical bases	19,629,341	19,629,341	19,659,726	19,659,726	31,131,903	31,131,903
Total gaps (bases)	5,881,433	4,201,327	6,008,698	4,178,612	907,141	871,760
Percent gaps/aligned bases	17.44%	12.46%	21.58%	14.08%	2.78%	2.67%
Longest aligned block (bases)	303,436	317,263	365,062	363,939	828,752	830,661
Aligned bases' percent identity	70.51%	66.50%	70.59%	66.24%	95.44%	95.33%
Genome percent aligned	87.07%	88.81%	87.10%	87.93%	98.15%	96.75%
Genome percent identical	61.39%	59.06%	61.49%	58.25%	93.67%	92.24%
Bases in non-inverted blocks	25,247,871	26,693,899	25,143,538	26,715,529	31,832,583	31,858,752
Bases in inverted blocks	2,590,894	2,824,972	2,705,102	2,963,197	787,995	797,207

Table A-2. Comparison of specific and non-syntenic genes between *A. alternata* ATCC 66981 (Aa1) and *A. brassicicola* (Ab)

	All		**Syntenic**		**Non-syntenic**		**Specific**	
	Aa1	***Ab***	***Aa1***	***Ab***	***Aa1***	***Ab***	***Aa1***	***Ab***
Number of genes	11,635	10,514	8,194	8,106	3,441	2,408	3,003	1,902
Secreted proteins	1,085	748	693	556	392	192	335	108
Small cysteine-rich secreted proteins	231	153	145	113	86	40	93	35
Carbohydrate active enzymes	512	436	353	334	159	102	105	39
Putative peptidases	456	372	324	296	132	76	93	26
Putative allergens	288	219	254	211	34	8	75	26
Secondary metabolites	185	125	121	90	64	35	62	18
Specific genes (% over non-synthenic genes)					1,914 (56%)	950 (40%)		
Homologous genes (% over non-synthenic genes)					1,527 (44%)	1,458 (60%)		
Syntenic genes (% over specific genes)							1,089 (36%)	952 (50%)
Non-syntenic genes (% over specific genes)							1,914 (64%)	950 (50%)

Table A-3. PFAM domains significantly different when comparing *A. alternata* ATCC 66981 (Aa1) and *A. brassicicola* (Ab) non-syntenic genes in their whole genome pairwise alignment

Term	**Description**	**Aa1**	**Ab**	**P-value**
PF06985.6	HET	75	18	0.0001
PF05368.8	NmrA	34	4	0.0003
PF05729.7	NACHT	21	1	0.0006
PF07690.11	MFS_1	82	31	0.0230
PF01370.16	Epimerase	16	2	0.0253
PF12796.2	Ank_2	31	8	0.0297
PF00106.20	adh_short	68	25	0.0299
PF11951.3	Fungal_trans_2	17	3	0.0373
PF12697.2	Abhydrolase_6	25	6	0.0395
PF03211.8	Pectate_lyase	2	6	0.0599
PF03184.14	DDE_1	4	8	0.0685
PF04082.13	Fungal_trans	40	14	0.0888
PF02129.13	Peptidase_S15	6	0	0.0892
PF08530.5	PepX_C	6	0	0.0892
PF13577.1	SnoaL_4	6	0	0.0892
PF01494.14	FAD_binding_3	28	9	0.0913
PF00023.25	Ank	12	2	0.0957

* Enriched numbers are underlined.

Table A-4. KOG group significantly different when comparing *A. alternata* ATCC 66981 (Aa1) and *A. brassicicola* (Ab) non-syntenic genes in their whole genome pairwise alignment

Group	**Description**	***Aa1***	***Ab***	**p-value**
R	General function prediction only	<u>602</u>	317	2E-06
G	Carbohydrate transport and metabolism	<u>244</u>	124	0.003
Q	Secondary metabolites biosynthesis, transport and catabolism	<u>231</u>	119	0.007
M	Cell wall/membrane/envelope biogenesis	<u>55</u>	20	0.012
I	Lipid transport and metabolism	<u>295</u>	162	0.014

* Enriched numbers are underlined.

Table A-5. List of selected 17 fungi used in phylogeny analysis

Name	**Class**	**Characteristics**	**Lifestyle**
Alternaria alternata ATCC 11680	Dothideomycetes	human allergy	saprotroph
Alternaria alternata ATCC 66981	Dothideomycetes	human allergy	saprotroph
Alternaria brassicicola ATCC 96836	Dothideomycetes	cabbage pathogen	saprotroph
Leptosphaeria maculans	Dothideomycetes	cabbage pathogen	necrotroph
Stagonospora nodorum SN15	Dothideomycetes	wheat pathogen	necrotroph
Aspergillus fumigatus Af293	Eurotiomycetes	human allergy/invasive infection	saprotroph
Aspergillus niger	Eurotiomycetes	black mold on plants	necrotroph
Myceliophthora thermophila	Eurotiomycetes	cellulose degrading, thermophilic	saprotroph
Blumeria graminis f sp hordei	Leotiomycetes	powdery mildew on grasses	biotroph
Trichoderma reesei QM6a	Sordariomycetes	cellulose degrading	saprotroph
Magnaporthe oryzae	Sordariomycetes	rice blast	hemibiotrophic
Neurospora crassa	Sordariomycetes	red bread mould	saprotroph
Podospora anserine	Sordariomycetes		saprotroph
Thielavia terrestris	Sordariomycetes	thermophilic, cellulose degrading	saprotroph
Laccaria bicolor	Agaricomycetes	mushroom	symbiont
Ceriporiopsis subvermispora	Agaricomycetes	ligninotic activity	saprotroph
Schizophyllum commune	Basidiomycetes	mushroom, wood decaying, recently identifed as cause of allergic infection	symbiont

Table A-6. *Brassica* pathogens specific genes (*A. brassicicola and L. maculans*)

Ortholog group	Protein	Annotation
OG_100044	ABPP00527	-
	ABPP00528	IPR001969[1]{Peptidase aspartic, active site}
	ABPP02813	-
	ABPP08270	-
	ABPP09934	IPR001969[1]{Peptidase aspartic, active site}
	ABPP10434	-
	ABPP10679	-
OG_106546	ABPP03486	IPR004045[1]{Glutathione S-transferase, N-terminal}; IPR010987[1]{Glutathione S-transferase, C-terminal-like}; IPR012336[1]{Thioredoxin-like fold}
	ABPP10578	IPR004045[1]{Glutathione S-transferase, N-terminal}; IPR010987[1]{Glutathione S-transferase, C-terminal-like}; IPR012336[1]{Thioredoxin-like fold}
OG_106552	ABPP09424	IPR004875[1]{DDE superfamily endonuclease, CENP-B-like}; IPR006600[1]{HTH CenpB-type DNA-binding domain}; IPR007889[1]{DNA binding HTH domain, Psq-type}; IPR009057[1]{Homeodomain-like}
OG_109738	ABPP00119	IPR003749[1]{ThiamineS/Molybdopterin converting factor subunit 1}; IPR016155[1]{Molybdopterin synthase/thiamin biosynthesis sulphur carrier, beta-grasp}
OG_109741	ABPP00271	-
OG_109742	ABPP00347	IPR001810[2]{F-box domain, cyclin-like}
OG_109743	ABPP00399	-
OG_109744	ABPP00523	IPR001453[2]{Molybdopterin binding}
OG_109747	ABPP00910	-
OG_109748	ABPP00935	IPR007736[1]{Caleosin}
OG_109753	ABPP01265	IPR018744[1]{Domain of unknown function DUF2293}
OG_109760	ABPP01662	-
OG_109765	ABPP02104	IPR009057[1]{Homeodomain-like}
OG_109767	ABPP02150	-
OG_109768	ABPP02285	IPR007918[1]{Mitochondrial distribution/morphology family 35/apoptosis}
OG_109769	ABPP02305	IPR008928[1]{Six-hairpin glycosidase-like}; IPR010905[1]{Glycosyl hydrolase, family 88}
OG_109773	ABPP02681	-
OG_109775	ABPP02877	-
OG_109780	ABPP03114	-
OG_109782	ABPP03195	-
OG_109785	ABPP03484	-
OG_109786	ABPP03485	IPR007219[1]{Transcription factor, fungi}
OG_109787	ABPP03518	IPR002198[1]{Short-chain dehydrogenase/reductase SDR}
OG_109789	ABPP03778	-
OG_109791	ABPP03786	IPR021858[1]{Protein of unknown function DUF3468}
OG_109793	ABPP03805	-
OG_109794	ABPP03891	IPR001214[1]{SET domain}; IPR009207[1]{Histone methyltransferase, PBCV type putative}

OG_109795	ABPP03973	-
OG_109796	ABPP04047	-
OG_109797	ABPP04050	IPR011701[1]{Major facilitator superfamily}; IPR016196[1]{Major facilitator superfamily domain, general substrate transporter}
OG_109801	ABPP04323	-
OG_109802	ABPP04326	-
OG_109804	ABPP04403	-
OG_109805	ABPP04415	-
OG_109806	ABPP04416	IPR010730[1]{Heterokaryon incompatibility}
OG_109809	ABPP04564	-
OG_109811	ABPP04610	-
OG_109814	ABPP04806	IPR001810[3]{F-box domain, cyclin-like}
OG_109816	ABPP04889	-
OG_109817	ABPP04914	-
OG_109821	ABPP05326	-
OG_109822	ABPP05349	-
OG_109824	ABPP05704	-
OG_109826	ABPP05807	-
OG_109830	ABPP06025	-
OG_109832	ABPP06147	IPR000250[1]{Peptidase G1}; IPR008985[1]{Concanavalin A-like lectin/glucanase}
OG_109834	ABPP06569	-
OG_109835	ABPP06626	-
OG_109836	ABPP06900	-
OG_109837	ABPP07010	-
OG_109838	ABPP07056	-
OG_109839	ABPP07083	-
OG_109841	ABPP07092	IPR011701[1]{Major facilitator superfamily}; IPR016196[1]{Major facilitator superfamily domain, general substrate transporter}
OG_109842	ABPP07250	-
OG_109845	ABPP07582	IPR007175[1]{RNAse P, Rpr2/Rpp21 subunit}
OG_109846	ABPP07596	IPR002198[1]{Short-chain dehydrogenase/reductase SDR}; IPR020904[1]{Short-chain dehydrogenase/reductase, conserved site}
OG_109848	ABPP07794	-
OG_109854	ABPP08708	-
OG_109856	ABPP08896	IPR009071[3]{High mobility group, superfamily}
OG_109858	ABPP09041	IPR011054[1]{Rudiment single hybrid motif}; IPR020559[1]{Phosphoribosylglycinamide synthetase, conserved site}; IPR020560[1]{Phosphoribosylglycinamide synthetase, C-domain}; IPR020561[1]{Phosphoribosylglycinamide synthetase, ATP-grasp (A) domain}
OG_109859	ABPP09128	-
OG_109860	ABPP09156	-
OG_109861	ABPP09215	-
OG_109862	ABPP09291	-
OG_109867	ABPP09613	IPR006598[2]{Lipopolysaccharide-modifying protein}

OG_109869	ABPP09855	-
OG_109871	ABPP10086	IPR024630[1]{Stc1 domain}
OG_109875	ABPP10154	IPR000933[3]{Glycoside hydrolase, family 29}; IPR017853[1]{Glycoside hydrolase, superfamily}
OG_109876	ABPP10179	-
OG_109877	ABPP10192	-
OG_109878	ABPP10266	-
OG_109879	ABPP10631	-
OG_109880	ABPP10632	-

Table A-7. Gene ontology comparison between *A. alternata* ATCC 66891 (Aa1) and *A. brassicicola* (Ab) specific genes in their pairwise homology analysis

GO Term	Description	Aa1	Ab	p-value	q-value
GO:0003676	nucleic acid binding	33	59	5.9E-08	5.8E-05
GO:0008152	metabolic process	220	68	3.9E-07	1.9E-04
GO:0005737	cytoplasm	16	37	8.8E-07	2.9E-04
GO:0016491	oxidoreductase activity	255	86	1.2E-06	3.0E-04
GO:0045449	regulation of transcription, DNA-dependent	16	34	8.5E-06	1.7E-03
GO:0005524	ATP binding	76	85	1.9E-05	3.1E-03
GO:0008026	ATP-dependent helicase activity	0	10	4.3E-05	6.2E-03
GO:0005667	transcription factor complex	15	30	5.6E-05	7.0E-03
GO:0006886	intracellular protein transport	2	12	2.3E-04	2.6E-02
GO:0042967	acyl-carrier-protein biosynthetic process	9	21	3.7E-04	3.6E-02
GO:0016787	hydrolase activity	84	23	6.4E-04	5.8E-02
GO:0004713	protein tyrosine kinase activity	16	0	9.8E-04	7.7E-02
GO:0003723	RNA binding	12	22	1.0E-03	7.7E-02
GO:0000036	acyl carrier activity	1	8	2.0E-03	1.3E-01
GO:0000287	magnesium ion binding	1	8	2.0E-03	1.3E-01
GO:0009069	serine family amino acid metabolic process	6	15	2.1E-03	1.3E-01
GO:0006260	DNA replication	0	6	2.4E-03	1.4E-01
GO:0003700	sequence-specific DNA binding transcription factor activity	21	29	2.6E-03	1.4E-01
GO:0006099	tricarboxylic acid cycle	2	9	3.0E-03	1.5E-01
GO:0004674	protein serine/threonine kinase activity	6	14	3.8E-03	1.9E-01
GO:0008565	protein transporter activity	1	7	4.8E-03	2.1E-01
GO:0046983	protein dimerization activity	1	7	4.8E-03	2.1E-01
GO:0055114	oxidation-reduction process	294	131	5.0E-03	2.1E-01
GO:0004672	protein kinase activity	38	8	5.1E-03	2.1E-01
GO:0016301	kinase activity	4	11	5.3E-03	2.1E-01
GO:0048037	cofactor binding	2	8	6.7E-03	2.6E-01
GO:0016310	phosphorylation	10	17	7.7E-03	2.8E-01
GO:0005622	intracellular	39	41	8.9E-03	3.1E-01
GO:0003824	catalytic activity	136	53	9.2E-03	3.1E-01
GO:0005515	protein binding	98	35	9.4E-03	3.1E-01
GO:0044237	cellular metabolic process	18	2	1.0E-02	3.3E-01
GO:0006200	ATP catabolic process	1	6	1.2E-02	3.5E-01
GO:0046487	glyoxylate metabolic process	1	6	1.2E-02	3.5E-01
GO:0016614	oxidoreductase activity, acting on CH-OH group of donors	44	12	1.7E-02	4.8E-01
GO:0005739	mitochondrion	0	4	1.8E-02	5.2E-01
GO:0006633	fatty acid biosynthetic process	3	8	2.3E-02	6.4E-01

GO:0006355	regulation of transcription, DNA-dependent	51	47	2.4E-02	6.4E-01
GO:0050662	coenzyme binding	16	2	2.6E-02	6.7E-01
GO:0003677	DNA binding	72	61	2.6E-02	6.7E-01
GO:0019643	reductive tricarboxylic acid cycle	1	5	2.8E-02	6.9E-01
GO:0050660	flavin adenine dinucleotide binding	51	16	2.8E-02	6.9E-01
GO:0008762	UDP-N-acetylmuramate dehydrogenase activity	21	4	3.6E-02	8.2E-01
GO:0006144	purine base metabolic process	4	8	3.8E-02	8.2E-01
GO:0005509	calcium ion binding	3	7	4.4E-02	8.2E-01
GO:0005083	small GTPase regulator activity	0	3	5.0E-02	8.2E-01
GO:0005741	mitochondrial outer membrane	0	3	5.0E-02	8.2E-01
GO:0005789	endoplasmic reticulum membrane	0	3	5.0E-02	8.2E-01
GO:0005819	spindle	0	3	5.0E-02	8.2E-01
GO:0006777	Mo-molybdopterin cofactor biosynthetic process	0	3	5.0E-02	8.2E-01
GO:0018106	peptidyl-histidine phosphorylation	0	3	5.0E-02	8.2E-01
GO:0042575	DNA polymerase complex	0	3	5.0E-02	8.2E-01

Table A-8. Gene ontology terms enriched in *A. alternata ATCC 66981 (Aa1)* specific genes, compared with *Aa1* genes that have homologs in *A. brassicicola (Ab)*

GO Term	**Description**	**# specific genes**	**# homologous genes**	**p-value**	**q-value**
GO:0016491	oxidoreductase activity	255	508	2.2E-22	4.8E-19
GO:0055114	oxidation-reduction process	294	777	3.6E-12	3.9E-09
GO:0008152	metabolic process	220	591	1.3E-08	5.7E-06
GO:0016705	oxidoreductase activity, acting on paired donors, with incorporation or reduction of molecular oxygen	55	84	2.7E-08	9.9E-06
GO:0005506	iron ion binding	67	117	4.3E-08	1.3E-05
GO:0016614	oxidoreductase activity, acting on CH-OH group of donors	44	69	1.6E-06	3.4E-04
GO:0020037	heme binding	61	126	1.3E-05	1.9E-03
GO:0009055	electron carrier activity	62	128	1.5E-05	2.0E-03
GO:0055085	transmembrane transport	151	421	2.1E-05	2.6E-03
GO:0006066	alcohol metabolic process	19	21	5.3E-05	6.0E-03
GO:0008080	N-acetyltransferase activity	21	27	1.3E-04	1.2E-02
GO:0008812	choline dehydrogenase activity	17	19	1.4E-04	1.2E-02
GO:0016702	oxidoreductase activity, acting on single donors with incorporation of molecular oxygen, incorporation of two atoms of oxygen	19	27	5.6E-04	3.6E-02
GO:0022857	transmembrane transporter activity	35	69	6.2E-04	3.9E-02

Table A-9. Gene ontology terms depleted in *A. alternata* ATCC 66981 (Aa1) specific genes, compared with Aa1 genes that have homologs in *A. brassicicola* (Ab)

GO Term	Description	Specific genes	Homologous genes	p-value	q-value
GO:0005737	cytoplasm	16	239	2.2E-09	1.6E-06
GO:0005524	ATP binding	76	595	5.6E-09	3.0E-06
GO:0003676	nucleic acid binding	33	323	1.3E-07	3.6E-05
GO:0005622	intracellular	39	352	2.4E-07	5.8E-05
GO:0006886	intracellular protein transport	2	79	7.2E-06	1.4E-03
GO:0008026	ATP-dependent helicase activity	0	55	9.2E-06	1.6E-03
GO:0006260	DNA replication	0	57	1.0E-05	1.6E-03
GO:0045449	regulation of transcription, DNA-dependent	16	182	1.0E-05	1.6E-03
GO:0005739	mitochondrion	0	51	2.3E-05	2.7E-03
GO:0005515	protein binding	98	608	5.8E-05	6.3E-03
GO:0004386	helicase activity	4	85	6.9E-05	7.1E-03
GO:0005840	ribosome	23	208	9.3E-05	9.0E-03
GO:0005667	transcription factor complex	15	161	9.6E-05	9.0E-03
GO:0006200	ATP catabolic process	1	53	1.9E-04	1.6E-02
GO:0009069	serine family amino acid metabolic process	6	96	2.0E-04	1.6E-02
GO:0005525	GTP binding	7	102	2.0E-04	1.6E-02
GO:0005789	endoplasmic reticulum membrane	0	41	2.2E-04	1.6E-02
GO:0006511	ubiquitin-dependent protein catabolic process	0	41	2.2E-04	1.6E-02
GO:0003723	RNA binding	12	133	2.7E-04	1.9E-02
GO:0016887	ATPase activity	1	52	3.1E-04	2.1E-02
GO:0004674	protein serine/threonine kinase activity	6	89	5.6E-04	3.6E-02
GO:0042254	ribosome biogenesis	12	127	7.0E-04	4.2E-02

Table A-10. CAZY modules for plant cell-wall degradation in different fungal genomes assigned by substrate categories

Species	**Cellulose**	**Xyloglucan**	**beta-1,3-1,4-glucan**	**Xylan**	**Galactomannan**	**Starch**	**Inulin**	**Pectin**	**Chitosan**	**Chitin**	**alpha-1,3-glucan**	**beta-1,3-glucan**	**beta-1,6-glucan**
Ceriporiopsis subvermispora	42	9	50	19	26	15	0	27	24	19	7	54	19
Laccaria bicolor	39	9	61	2	26	17	0	16	34	14	10	69	32
Schizophyllum commune	58	10	57	52	22	20	1	71	28	21	15	70	21
Blumeria graminis f. sp. hordei	5	1	19	2	3	5	0	3	7	9	2	25	3
Trichoderma reesei QM6a	30	11	33	24	25	11	0	20	16	24	9	58	13
Myceliophthora thermophila	54	9	35	47	20	17	0	48	22	12	10	53	12
Thielavia terrestris	54	12	38	39	23	16	1	44	25	17	11	57	16
Neurospora crassa	40	7	31	30	14	17	1	33	17	14	12	48	9
Podospora anserina S mat+	69	8	36	54	23	17	0	46	24	22	9	54	15
Magnaporthe oryzae	65	16	47	64	27	18	5	60	30	21	9	73	14
Aspergillus fumigatus Af293	51	15	41	53	28	28	5	87	26	22	16	67	15
Aspergillus niger	42	15	29	38	23	22	4	76	17	20	13	48	10
Alternaria alternata ATCC 11680	84	17	57	64	36	20	3	102	37	20	13	80	24
Alternaria alternata ATCC 66981	82	17	57	61	36	19	3	102	37	21	13	80	24
Alternaria brassicicola ATCC 96836	67	12	52	49	35	18	2	86	30	16	10	70	20
Phaeosphaeria nodorum SN15	77	15	51	58	36	20	4	75	38	20	13	70	23
Leptosphaeria maculans	60	13	46	36	32	18	2	69	26	15	12	65	18

Table A-11. Comparison of KOG annotation between specific genes and homologous genes of *A. alternata* ATCC 66981 *(Aa1)* and *A. brassicicola (Ab)*.

Category	Descriptipon	Aa1 specific genes (count)	Aa1/Ab homologous genes (count)	Ab specific genes (count)	Aa1 specific genes (%)	Aa1/Ab homologous genes (%)	Ab specific genes (%)
R	General function prediction only	493	3107	128	46.33%	30.96%	26.39%
T	Signal transduction mechanisms	144	1698	54	13.53%	16.92%	11.13%
U	Intracellular trafficking, secretion, and vesicular transport	97	1633	54	9.12%	16.27%	11.13%
K	Transcription	116	1563	59	10.90%	15.57%	12.16%
O	Posttranslational modification, protein turnover, chaperones	140	1558	57	13.16%	15.52%	11.75%
I	Lipid transport and metabolism	231	1361	76	21.71%	13.56%	15.67%
J	Translation, ribosomal structure and biogenesis	64	1292	41	6.02%	12.87%	8.45%
G	Carbohydrate transport and metabolism	173	1119	42	16.26%	11.15%	8.66%
A	RNA processing and modification	32	1093	33	3.01%	10.89%	6.80%
D	Cell cycle control, cell division, chromosome partitioning	53	1017	23	4.98%	10.13%	4.74%
L	Replication, recombination and repair	45	944	33	4.23%	9.41%	6.80%
Z	Cytoskeleton	37	816	25	3.48%	8.13%	5.15%
C	Energy production and conversion	107	809	36	10.06%	8.06%	7.42%
E	Amino acid transport and metabolism	70	774	35	6.58%	7.71%	7.22%
Q	Secondary metabolites biosynthesis, transport and catabolism	215	743	66	20.21%	7.40%	13.61%
P	Inorganic ion transport and metabolism	73	624	24	6.86%	6.22%	4.95%
B	Chromatin structure and dynamics	16	587	12	1.50%	5.85%	2.47%
F	Nucleotide transport and metabolism	19	303	13	1.79%	3.02%	2.68%
V	Defense mechanisms	40	284	5	3.76%	2.83%	1.03%
M	Cell wall/membrane/envelope biogenesis	40	226	7	3.76%	2.25%	1.44%
H	Coenzyme transport and metabolism	28	216	9	2.63%	2.15%	1.86%
W	Extracellular structures	2	133	0	0.19%	1.33%	0.00%
Y	Nuclear structure	1	131	6	0.09%	1.31%	1.24%
N	Cell motility	0	18	1	0.00%	0.18%	0.21%
S+X	Unkown function	48	1018	31	4.51%	10.14%	6.39%
Total		1064	10036	485			

Table A-12. Predicted allergens in *A. alternata* ATCC 66891 using Allerdictor and the comparative approach.

ID	IDMEM	Allerdictor	Description
AAT_PP00328	allergen	allergen	Peptidyl-prolyl cis-trans isomerase
AAT_PP00463	allergen	non-allergen	Peptidyl-prolyl cis-trans isomerase D
AAT_PP00557	allergen	non-allergen	Pectate lyase A
AAT_PP00650	allergen	non-allergen	Pectate lyase B
AAT_PP01023	allergen	allergen	L-xylulose reductase
AAT_PP01058	allergen	non-allergen	putative beta-hexosaminidase precursor
AAT_PP01180	allergen	non-allergen	Endopolygalacturonase AN8327
AAT_PP01205	allergen	non-allergen	Carboxypeptidase Y
AAT_PP01273	allergen	non-allergen	putative thioredoxin TrxA
AAT_PP01510	allergen	allergen	Major allergen Alt a 1
AAT_PP01832	allergen	allergen	Heat shock protein 90
AAT_PP01902	allergen	non-allergen	Aldehyde dehydrogenase
AAT_PP01945	allergen	allergen	Protein disulfide-isomerase
AAT_PP01992	allergen	non-allergen	Pectate lyase plyB
AAT_PP02000	allergen	non-allergen	Probable L-asparaginase 1
AAT_PP02191	allergen	allergen	Alkaline protease 2
AAT_PP02462	allergen	non-allergen	Peptidyl-prolyl cis-trans isomerase ppi1
AAT_PP02569	allergen	allergen	Catalase B
AAT_PP02643	allergen	non-allergen	Probable beta-glucosidase M
AAT_PP02931	allergen	non-allergen	putative purine catabolism protein pucG
AAT_PP03395	allergen	allergen	Minor allergen Alt a 7
AAT_PP03408	allergen	non-allergen	Tubulin alpha-2 chain
AAT_PP03411	allergen	non-allergen	putative alcohol dehydrogenase
AAT_PP03470	allergen	non-allergen	Guanine nucleotide-binding protein subunit beta-like protein
AAT_PP03496	allergen	non-allergen	putative aldehyde dehydrogenase (NAD+)
AAT_PP03911	allergen	allergen	Peptidyl-prolyl cis-trans isomerase, mitochondrial
AAT_PP04110	allergen	non-allergen	putative alpha-glucosidase
AAT_PP04206	non-allergen	allergen	Proteinase T
AAT_PP04207	allergen	allergen	Cuticle-degrading protease
AAT_PP04228	allergen	non-allergen	Superoxide dismutase [Cu-Zn]
AAT_PP04229	allergen	allergen	Transaldolase
AAT_PP04235	allergen	non-allergen	putative extracellular cell wall glucanase Crf1/allergen Asp F9
AAT_PP04281	allergen	non-allergen	Casein kinase II subunit beta-2
AAT_PP04328	allergen	allergen	Aldehyde dehydrogenase
AAT_PP04412	allergen	non-allergen	Elongation factor 1-beta
AAT_PP04544	allergen	non-allergen	5-methyltetrahydropteroyltriglutamate--homocysteine methyltransferase

AAT_PP04753	allergen	allergen	60S acidic ribosomal protein P1
AAT_PP04839	allergen	non-allergen	putative trypsin
AAT_PP04915	allergen	allergen	Translationally controlled tumour protein
AAT_PP05047	allergen	non-allergen	Phosphoglycerate kinase
AAT_PP05326	allergen	allergen	Heat shock 70 kDa protein 2
AAT_PP05514	allergen	non-allergen	Peptidyl-prolyl cis-trans isomerase B
AAT_PP05533	allergen	allergen	Enolase
AAT_PP05751	allergen	non-allergen	Heat shock protein Hsp88
AAT_PP05809	allergen	allergen	Superoxide dismutase [Mn], mitochondrial
AAT_PP05896	allergen	non-allergen	Nascent polypeptide-associated complex subunit alpha
AAT_PP05996	allergen	non-allergen	putative FAD binding domain-containing protein
AAT_PP06039	allergen	allergen	Cytochrome c
AAT_PP06082	allergen	non-allergen	Endoglucanase 1
AAT_PP06148	allergen	allergen	Malate dehydrogenase, mitochondrial
AAT_PP06273	allergen	non-allergen	Probable exo-1,4-beta-xylosidase bxlB
AAT_PP06449	allergen	non-allergen	Heat shock protein SSB
AAT_PP06558	allergen	non-allergen	hypothetical protein
AAT_PP06651	allergen	non-allergen	Catalase-peroxidase
AAT_PP06706	allergen	allergen	Nuclear transport factor 2
AAT_PP07231	allergen	non-allergen	putative isoamyl alcohol oxidase
AAT_PP07252	allergen	non-allergen	Nucleoside diphosphate kinase
AAT_PP07548	allergen	non-allergen	putative glycosyl hydrolase family 16
AAT_PP07599	allergen	non-allergen	Fructose-bisphosphate aldolase
AAT_PP07685	allergen	non-allergen	Probable beta-glucosidase G
AAT_PP07687	allergen	non-allergen	3-isopropylmalate dehydratase
AAT_PP07795	allergen	non-allergen	putative isoamyl alcohol oxidase
AAT_PP07927	allergen	non-allergen	Aldehyde dehydrogenase
AAT_PP08036	allergen	non-allergen	putative lipase 3 precursor
AAT_PP08109	allergen	non-allergen	Alpha-glucosidase
AAT_PP08204	allergen	non-allergen	Triosephosphate isomerase
AAT_PP08292	allergen	non-allergen	Aldehyde dehydrogenase
AAT_PP08339	allergen	non-allergen	Catalase-1
AAT_PP08358	allergen	non-allergen	putative major allergen Asp f 2 precursor
AAT_PP08423	allergen	non-allergen	1,3-beta-glucanosyltransferase gel2
AAT_PP08435	allergen	non-allergen	Alkaline proteinase
AAT_PP08456	allergen	non-allergen	Eukaryotic translation initiation factor 6
AAT_PP08492	allergen	non-allergen	Peptidyl-prolyl cis-trans isomerase H
AAT_PP08540	non-allergen	non-allergen	putative extracellular cell wall glucanase Crf1/allergen Asp F9
AAT_PP08659	allergen	non-allergen	Extracellular metalloproteinase
AAT_PP08946	allergen	allergen	60S ribosomal protein L3
AAT_PP09056	allergen	non-allergen	putative dipeptidyl-peptidase 5 precursor

AAT_PP09150	allergen	allergen	Alcohol dehydrogenase 1
AAT_PP09630	allergen	non-allergen	Calnexin
AAT_PP09715	allergen	non-allergen	putative dipeptidyl-peptidase 5 precursor
AAT_PP09843	allergen	non-allergen	Probable beta-glucosidase E
AAT_PP09861	allergen	non-allergen	Eukaryotic translation initiation factor 3 subunit I
AAT_PP09950	allergen	allergen	Formate dehydrogenase
AAT_PP09981	allergen	non-allergen	Glyceraldehyde-3-phosphate dehydrogenase
AAT_PP10037	allergen	non-allergen	Tubulin alpha chain
AAT_PP10069	allergen	non-allergen	Aspartic protease PEP1
AAT_PP10196	allergen	non-allergen	putative alkaline protease
AAT_PP10285	allergen	non-allergen	Heat shock 70 kDa protein C
AAT_PP10517	allergen	non-allergen	Lysophospholipase 1
AAT_PP10781	allergen	non-allergen	Beta-glucosidase 1
AAT_PP10872	allergen	non-allergen	Peptidyl-prolyl cis-trans isomerase H
AAT_PP10979	allergen	non-allergen	Malate dehydrogenase, cytoplasmic
AAT_PP11204	allergen	allergen	Glucose-6-phosphate isomerase
AAT_PP11221	allergen	non-allergen	Heat shock 70 kDa protein
AAT_PP11349	allergen	non-allergen	Polygalacturonase
AAT_PP11362	allergen	allergen	40S ribosomal protein S3
AAT_PP11384	allergen	non-allergen	Mannosyl-oligosaccharide alpha-1,2-mannosidase 1B
AAT_PP11538	allergen	allergen	60S acidic ribosomal protein P2

Appendix B

Supplementary figures

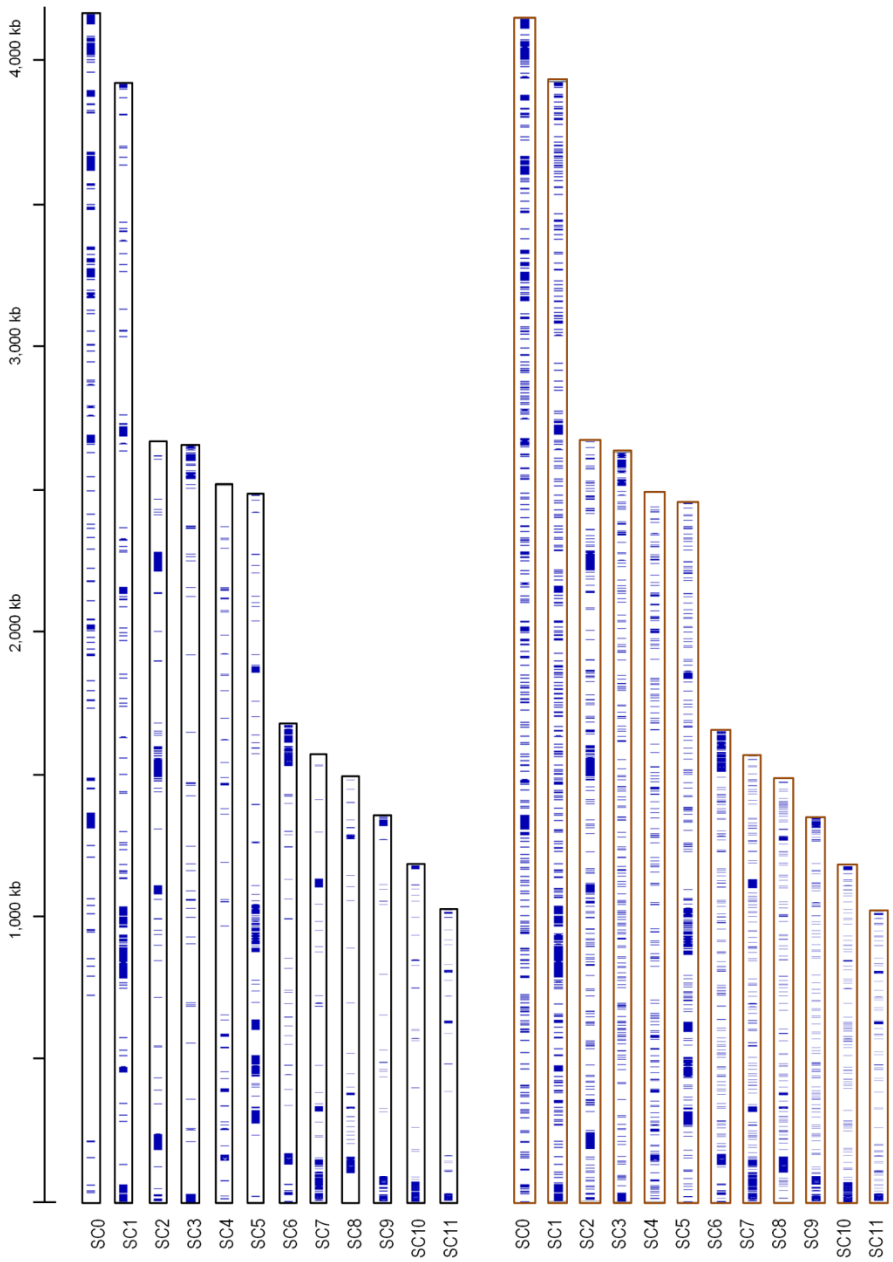


Figure B-1. *A. brassicicola* repetitive sequence distribution over the longest 11 supercontigs. Blue bands represent repetitive sequences, excluding simple repeats (left) and including simple repeats (right).

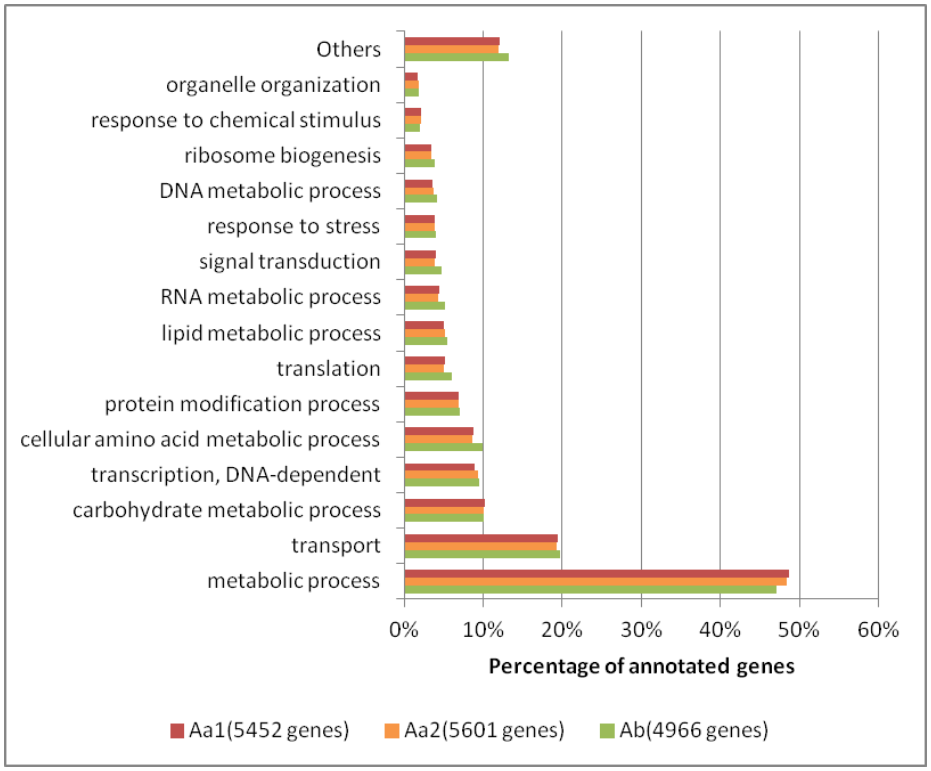


Figure B-2. Biological proccess GO slim classification for three *Alternaria* genomes. GO terms were associated with genes by BLAST2GO analysis and mapped to *Alternaria* high level GO slim terms using GO-Perl software. The bars indicate the percentage of genes that were mapped to each GO slim term. The total number of genes that has mappable GO slim terms for each species is in the brackets.

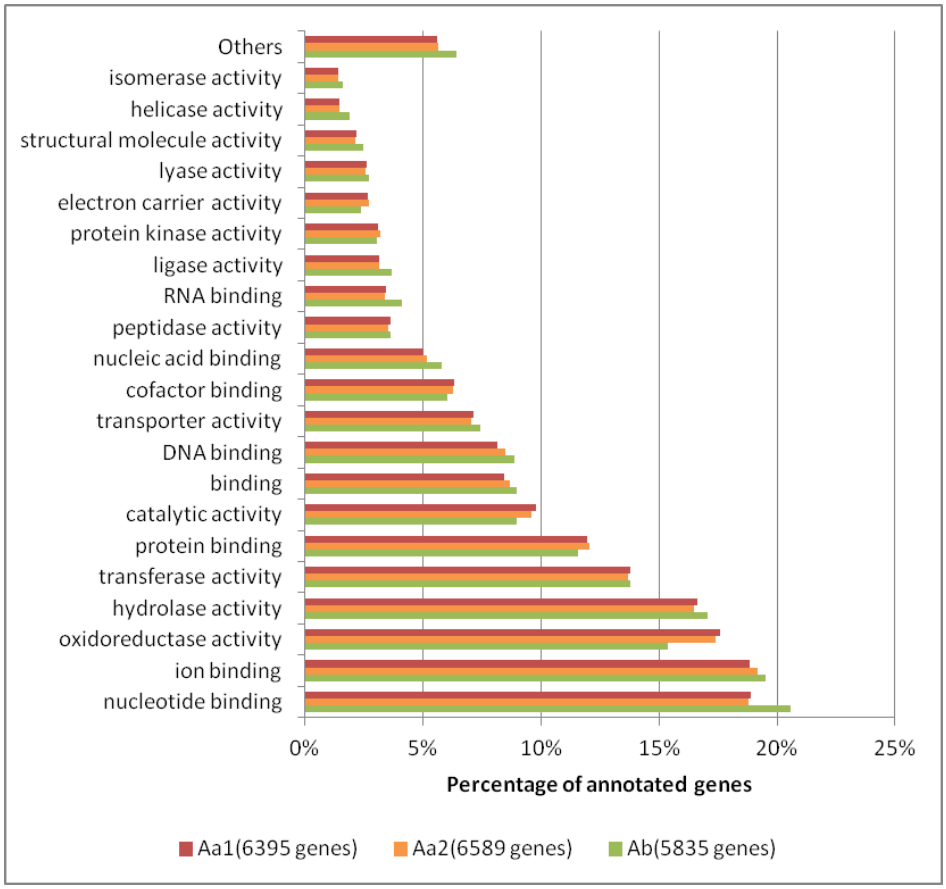


Figure B-3. Molecular function GO slim classification for three *Alternaria* genomes. GO terms were associated with genes by BLAST2GO analysis and mapped to *Alternaria* high level GO slim terms using GO-Perl software. The bars indicate the percentage of genes that were mapped to each GO slim term. The total number of genes that has mappable GO slim terms for each species is in the brackets.

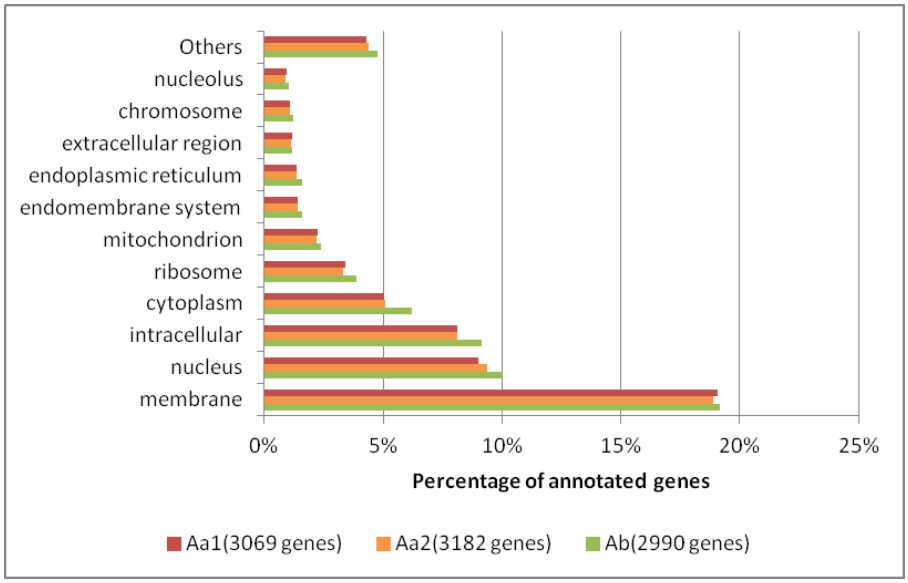


Figure B-4. Cellular component GO slim classification for three *Alternaria* genomes. GO terms were associated with genes by BLAST2GO analysis and mapped to *Alternaria* high level GO slim terms using GO-Perl software. The bars indicate the percentage of genes that were mapped to each GO slim term. The total number of genes that has mappable GO slim terms for each species is in the brackets.

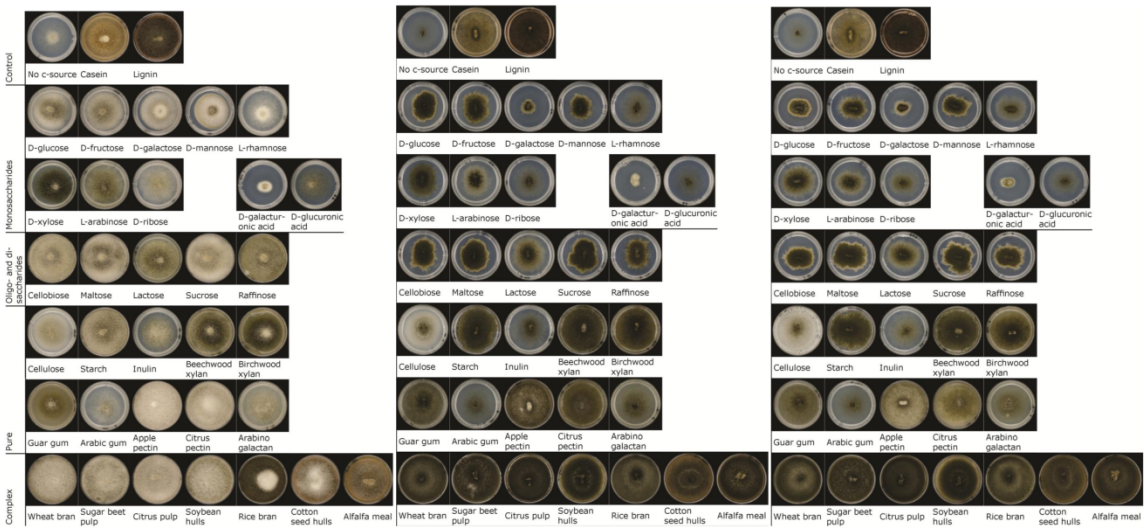


Figure B-5. Growth profiles of *A. alternata* ATCC 11680 (left), *A. alternata* ATCC 66981 (middle), and *A. brassicicola* ATCC 96836 (right) on monomeric, oligomeric and polymeric carbon sources. Growth profiling was performed by Ronald de Vries and Eline Majoor at CBS-KNAW Fungal Biodiversity Centre, The Netherlands.

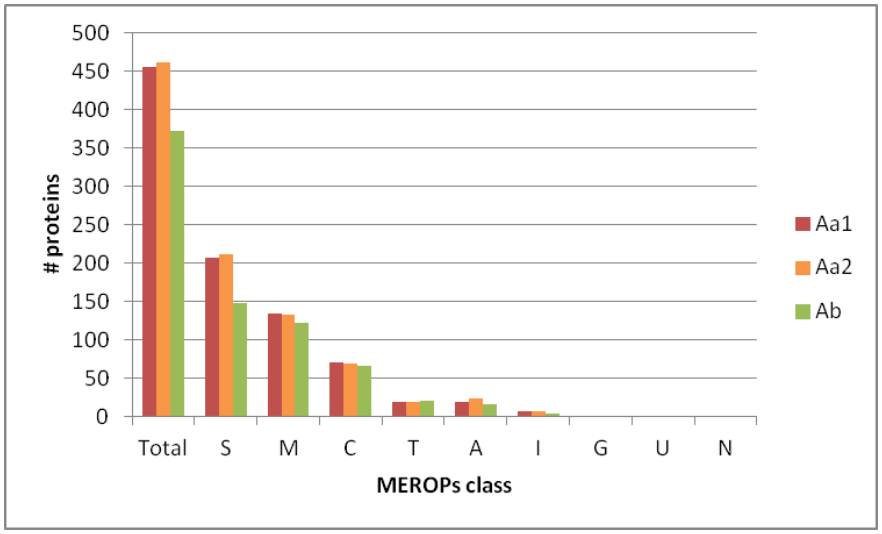


Figure B-6. Number of peptidases found in three *Alternaria* genomes. The peptidases were identified by MEROPs database batch BLAST search tool. A – Aspartic, C – Cysteine, G – Glutamic, M – Metallo, N – Asparagine, S – Serine, T – Threonine, I – Inhibitor, U – Unknown

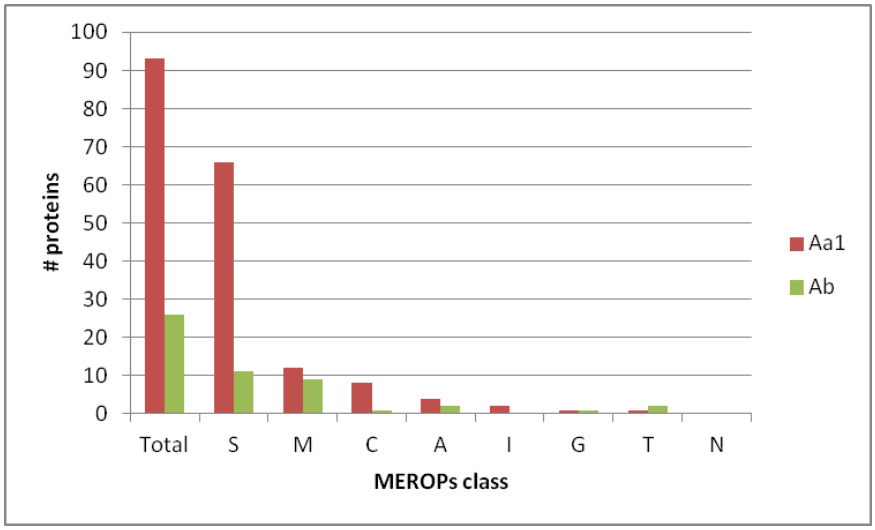
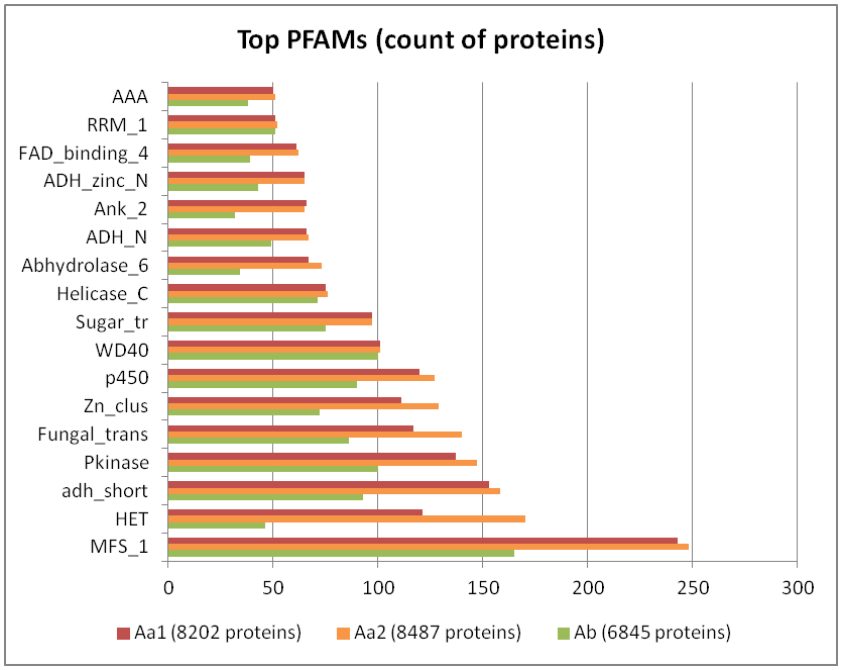


Figure B-7. Peptidase classes of *A. alternata* ATCC 66981 (Aa1) and *A. brassicicola* (Ab) specific proteins. Specific proteins were identified by OrthoMCL and peptidases were identified by MEROPs database batch BLAST search tool. A – Aspartic, C – Cysteine, G – Glutamic, M – Metallo, N – Asparagine, S – Serine, T – Threonine, I – Inhibitor, U – Unknonwn



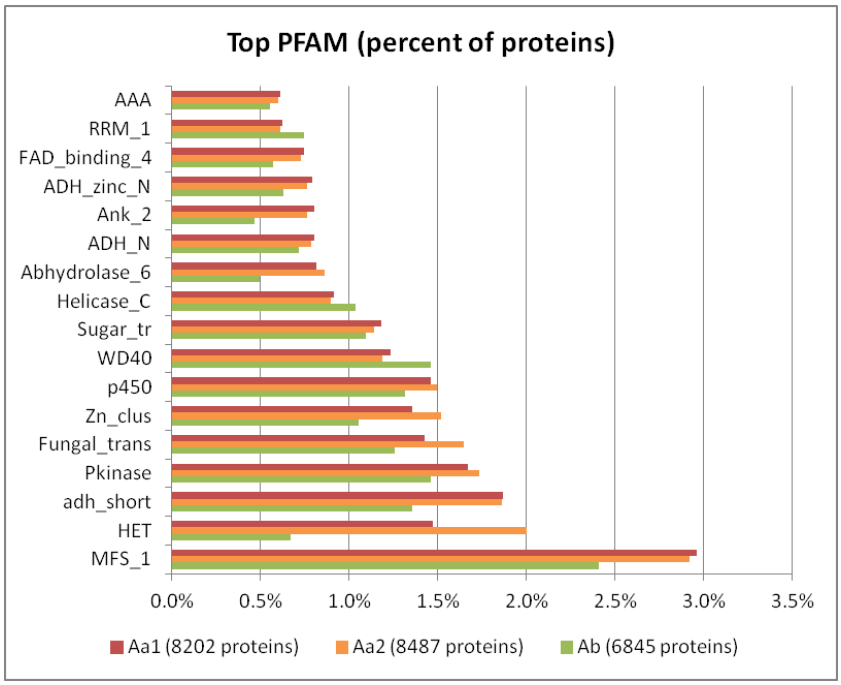
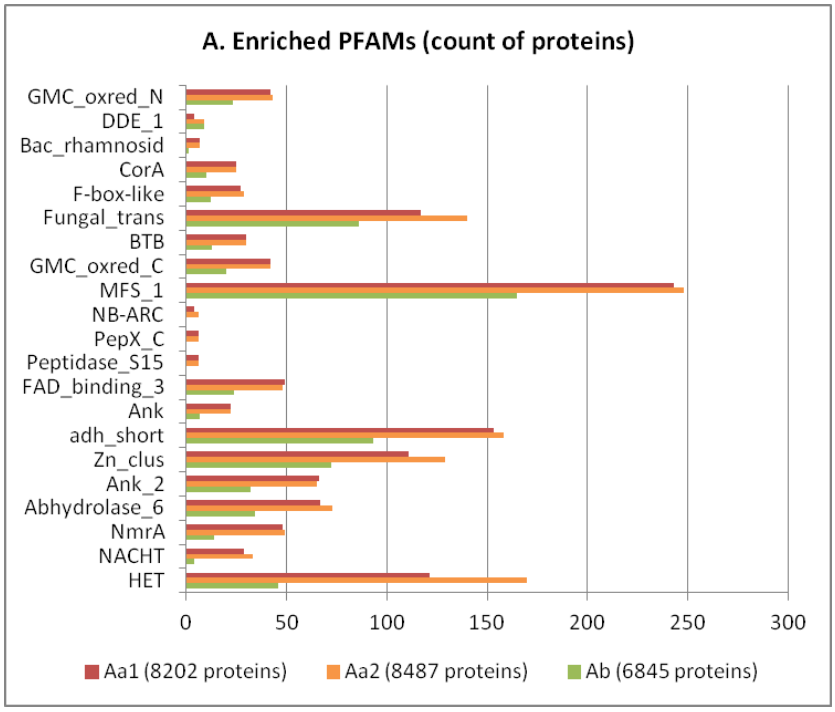


Figure B-8. Top PFAM (hit by ≥ 50 proteins) found in *Alternaria* proteins. PFAMs were identified using HMMER3 scan on Pfam-A database version 26.



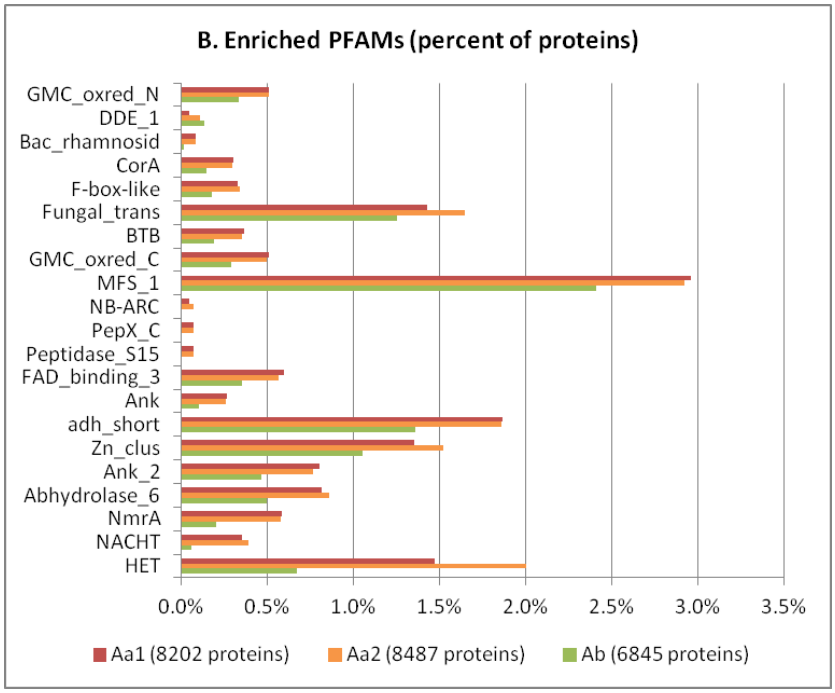


Figure B-9. PFAM enriched/depleted in Aa1 and/or Aa2 when compared with Ab genome.

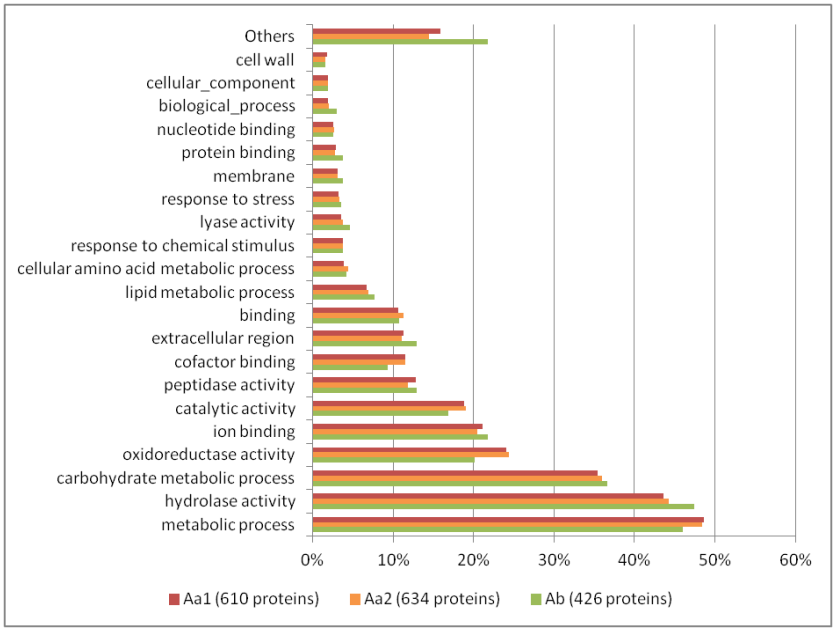


Figure B-10. Top 25 GOslim terms associated with predicted secreted proteins from *Alternaria* proteins

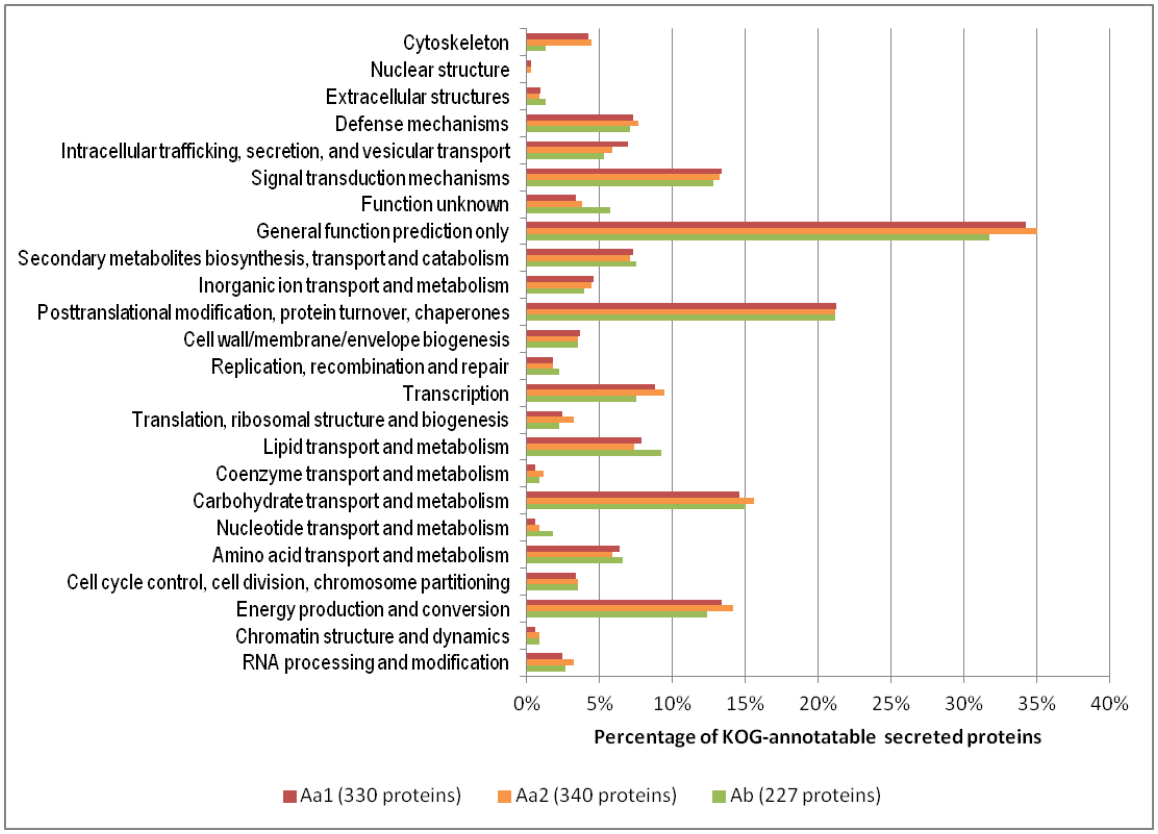


Figure B-11. KOG classification for *Alternaria* predicted secretomes

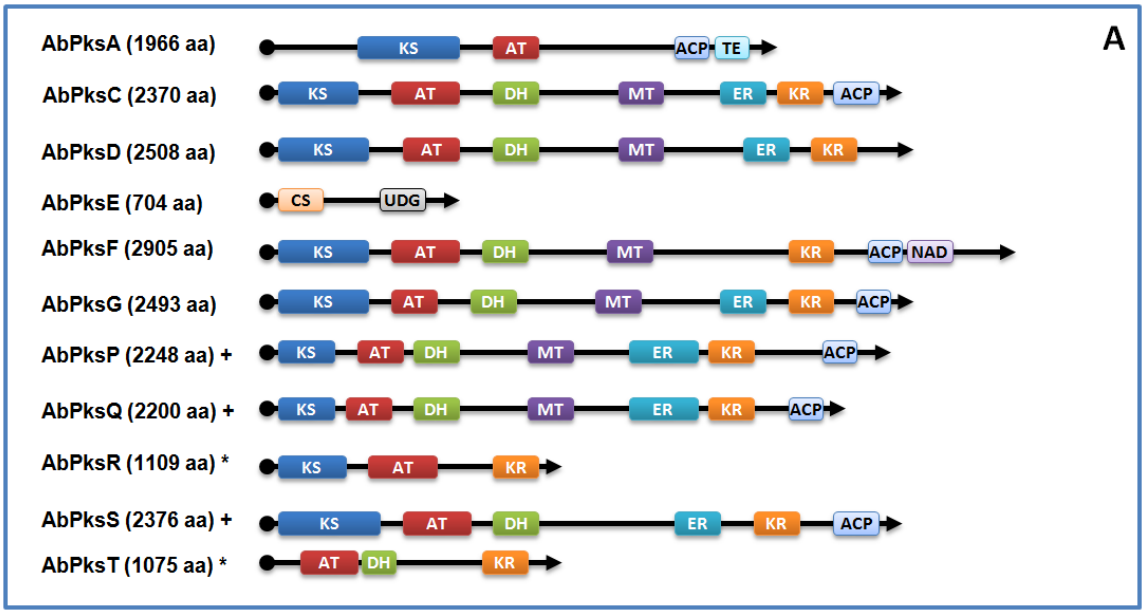


Figure B-12. *A. brassicicola* PKS proteins' domain architecture. Domain legends: KS – β–ketoacyl synthase; AT – acyltransferase; DH – dehydratase; MT – methyltransferase; ER – enoyl reductase; KR – ketoreductase; ACP – acyl carrier protein; CD – condensation domain; AA – Amino acid adenylation domain; TE – Thioesterase; NAD – NAD binding domain; CS – Chalcone and stilbene synthases; UDG – Uracil DNA Glycolase Superfamily; DUF – Domain of unknown function; PEP – Peptidase domain; Pvc – Pyoverdine/dityrosine biosynthesis protein; ABH – Abhydrolase; ETR – Esterase; Acps – 4'–phosphopantetheinyl transferase; Hxx – HxxPF–repeated domain; MFS – MFS transporter.

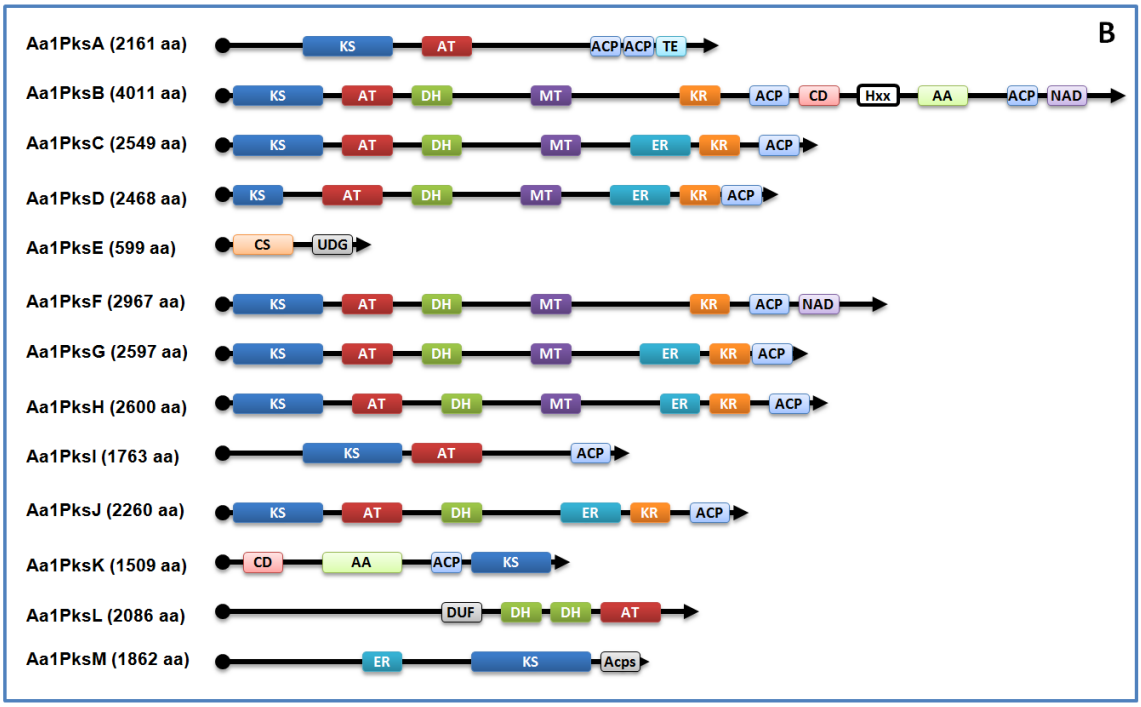


Figure B-13. *A. alternata* ATCC 66981 PKS proteins' domain architecture.Domain legends: KS – β–ketoacyl synthase; AT – acyltransferase; DH – dehydratase; MT – methyltransferase; ER – enoyl reductase; KR – ketoreductase; ACP – acyl carrier protein; CD – condensation domain; AA – Amino acid adenylation domain; TE – Thioesterase; NAD – NAD binding domain; CS – Chalcone and stilbene synthases; UDG – Uracil DNA Glycolase Superfamily; DUF – Domain of unknown function; PEP – Peptidase domain; Pvc – Pyoverdine/dityrosine biosynthesis protein; ABH – Abhydrolase; ETR – Esterase; Acps – 4'–phosphopantetheinyl transferase; Hxx – HxxPF–repeated domain; MFS – MFS transporter.

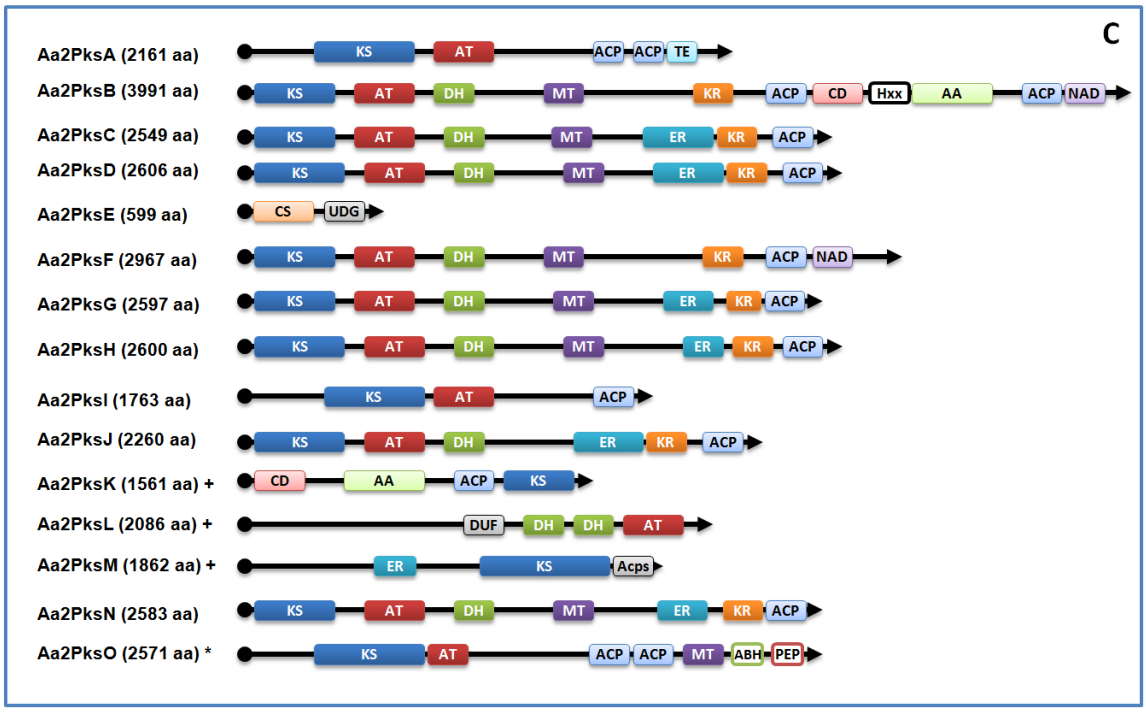


Figure B-14. *A. alternata* ATCC 11680 PKS proteins' domain architecture. Domain legends: KS – β–ketoacyl synthase; AT – acyltransferase; DH – dehydratase; MT – methyltransferase; ER – enoyl reductase; KR – ketoreductase; ACP – acyl carrier protein; CD – condensation domain; AA – Amino acid adenylation domain; TE – Thioesterase; NAD – NAD binding domain; CS – Chalcone and stilbene synthases; UDG – Uracil DNA Glycolase Superfamily; DUF – Domain of unknown function; PEP – Peptidase domain; Pvc – Pyoverdine/dityrosine biosynthesis protein; ABH – Abhydrolase; ETR – Esterase; Acps – 4'–phosphopantetheinyl transferase; Hxx – HxxPF–repeated domain; MFS – MFS transporter.

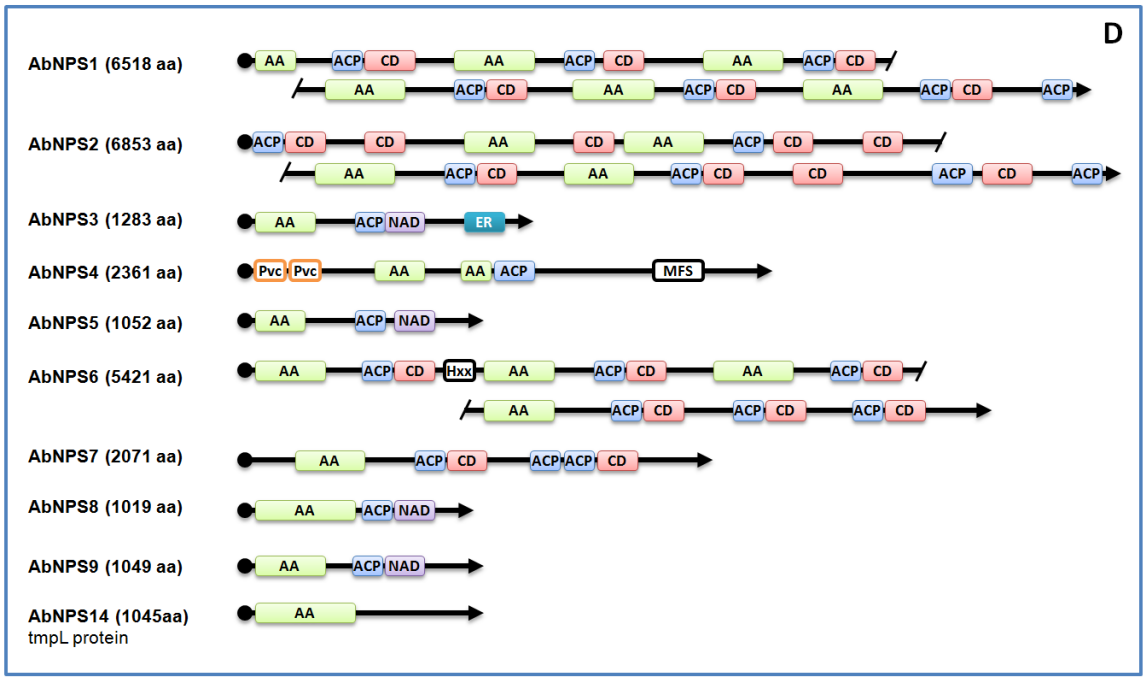


Figure B-15. *A. brassicicola* NRPS proteins' domain architecture. Domain legends: KS – β–ketoacyl synthase; AT – acyltransferase; DH – dehydratase; MT – methyltransferase; ER – enoyl reductase; KR – ketoreductase; ACP – acyl carrier protein; CD – condensation domain; AA – Amino acid adenylation domain; TE – Thioesterase; NAD – NAD binding domain; CS – Chalcone and stilbene synthases; UDG – Uracil DNA Glycolase Superfamily; DUF – Domain of unknown function; PEP – Peptidase domain; Pvc – Pyoverdine/dityrosine biosynthesis protein; ABH – Abhydrolase; ETR – Esterase; Acps – 4'–phosphopantetheinyl transferase; Hxx – HxxPF–repeated domain; MFS – MFS transporter.

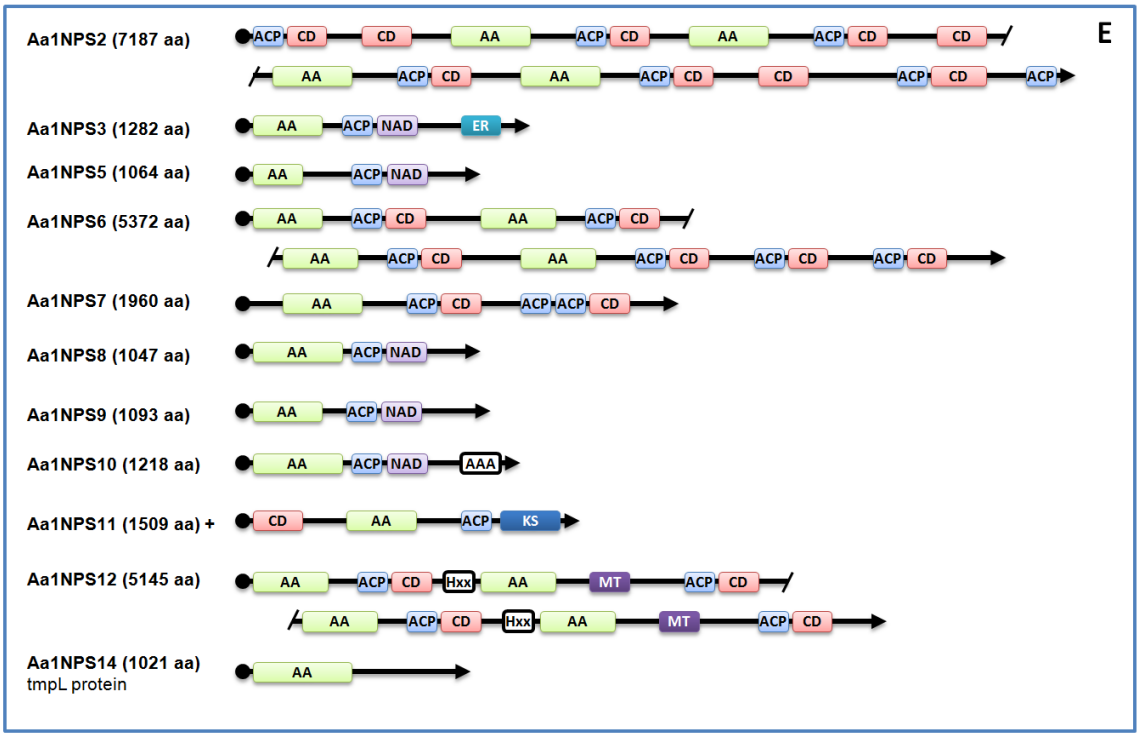


Figure B-16. *A. alternata* ATCC 66981 NRPS proteins' architecture. Domain legends: KS – β–ketoacyl synthase; AT – acyltransferase; DH – dehydratase; MT – methyltransferase; ER – enoyl reductase; KR – ketoreductase; ACP – acyl carrier protein; CD – condensation domain; AA – Amino acid adenylation domain; TE – Thioesterase; NAD – NAD binding domain; CS – Chalcone and stilbene synthases; UDG – Uracil DNA Glycolase Superfamily; DUF – Domain of unknown function; PEP – Peptidase domain; Pvc – Pyoverdine/dityrosine biosynthesis protein; ABH – Abhydrolase; ETR – Esterase; Acps – 4'–phosphopantetheinyl transferase; Hxx – HxxPF–repeated domain; MFS – MFS transporter.

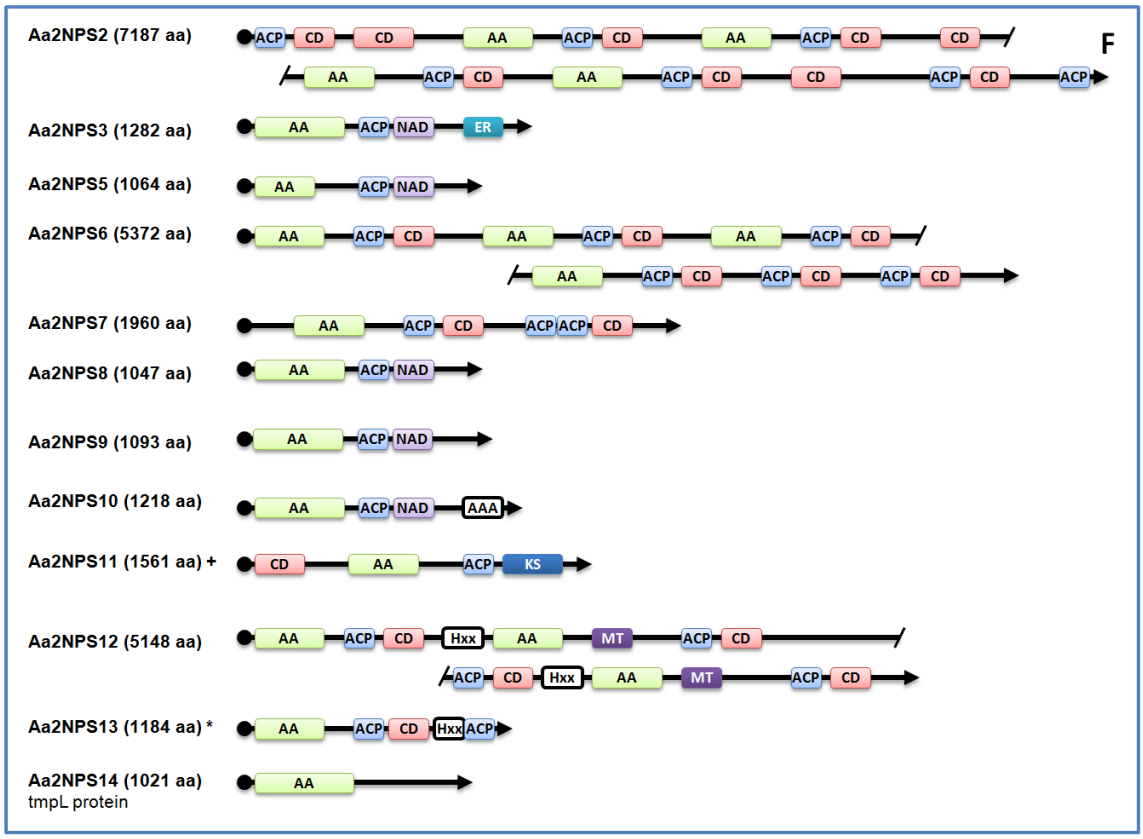


Figure B-17. *A. alternata* ATCC 11680 NRPS proteins' domain architecture. Domain legends: KS – β–ketoacyl synthase; AT – acyltransferase; DH – dehydratase; MT – methyltransferase; ER – enoyl reductase; KR – ketoreductase; ACP – acyl carrier protein; CD – condensation domain; AA – Amino acid adenylation domain; TE – Thioesterase; NAD – NAD binding domain; CS – Chalcone and stilbene synthases; UDG – Uracil DNA Glycolase Superfamily; DUF – Domain of unknown function; PEP – Peptidase domain; Pvc – Pyoverdine/dityrosine biosynthesis protein; ABH – Abhydrolase; ETR – Esterase; Acps – 4'–phosphopantetheinyl transferase; Hxx – HxxPF–repeated domain; MFS – MFS transporter.

References

1. GINA: *Global Strategy for Asthma Management and Prevention*. 2011.

2. WHO: **Asthma Fact sheet No. 307**. 2013.

3. Denning DW, O'Driscoll BR, Hogaboam CM, Bowyer P, Niven RM: **The link between fungi and severe asthma: a summary of the evidence**. *Eur Respir J* 2006, **27**:615–626.

4. Stagg NJ, Ghantous HN, Ladics GS, House RV, Gendel SM, Hastings KL: **Workshop Proceedings Challenges and Opportunities in Evaluating Protein Allergenicity Across Biotechnology Industries**. *Int J Toxicol* 2013, **32**:4–10.

5. Valenta R: **The future of antigen-specific immunotherapy of allergy**. *Nat Rev Immunol* 2002, **2**:446–453.

6. Chapman MD, Pomés A, Aalberse RC: **Molecular Biology of Allergens: Structure and Immune Recognition**. In *Allergy Front Epigenetics Allerg Risk Factors*. Edited by Pawankar R, Holgate ST, Rosenwasser LJ. Springer Japan; 2009:265–289. [*Allergy Frontiers*, vol. 1]

7. Bush RK: **Fungal Allergy as Yet Unsolved**. In *Allergy Front Clin Manif*. Edited by Pawankar R, Holgate ST, Rosenwasser LJ. Springer Japan; 2009:471–485. [*Allergy Frontiers*, vol. 3]

8. Kaiser L, Grönlund H, Sandalova T, Ljunggren H-G, van Hage-Hamsten M, Achour A, Schneider G: **The crystal structure of the major cat allergen Fel d 1, a member of the secretoglobin family**. *J Biol Chem* 2003, **278**:37730–37735.

9. Huby RDJ, Dearman RJ, Kimber I: **Why Are Some Proteins Allergens?** *Toxicol Sci* 2000, **55**:235–246.

10. Ivanciuc O, Schein CH, Garcia T, Oezguen N, Negi SS, Braun W: **Structural analysis of linear and conformational epitopes of allergens**. *Regul Toxicol Pharmacol* 2009, **54**(3, Supplement):S11–S19.

11. Ivanciuc O, Schein CH, Braun W: **SDAP: database and computational tools for allergenic proteins**. *Nucleic Acids Res* 2003, **31**:359–362.

12. Aalberse RC: **Structural biology of allergens**. *J Allergy Clin Immunol* 2000, **106**:228–238.

13. Fernández-Rivas M, Bolhaar S, González-Mancebo E, Asero R, van Leeuwen A, Bohle B, Ma Y, Ebner C, Rigby N, Sancho AI, Miles S, Zuidmeer L, Knulst A,

Breiteneder H, Mills C, Hoffmann-Sommergruber K, van Ree R: **Apple allergy across Europe: how allergen sensitization profiles determine the clinical expression of allergies to plant foods**. *J Allergy Clin Immunol* 2006, **118**:481–488.

14. Simon-Nobbe B, Denk U, Pöll V, Rid R, Breitenbach M: **The Spectrum of Fungal Allergy**. *Int Arch Allergy Immunol* 2008, **145**:58–86.

15. Crameri R, Garbani M, Rhyner C, Huitema C: **Fungi: the neglected allergenic sources**. *Allergy* 2014, **69**:176–185.

16. Bowyer P, Fraczek M, Denning DW: **Comparative genomics of fungal allergens and epitopes shows widespread distribution of closely related allergen and epitope orthologues**. *BMC Genomics* 2006, **7**:251.

17. Bowyer P, Denning DW: **Genomic analysis of allergen genes in Aspergillus spp.: the relevance of genomics to everyday research**. *Med Mycol* 2007, **45**:17–26.

18. Rotem J: *The Genus Alternaria: Biology, Epidemiology, and Pathogenicity*. St. Paul, Minn.: APS Press; 1994.

19. Morales VM, Jasalavich CA, Pelcher LE, Petrie GA, Taylor JL: **Phylogenetic relationship among several Leptosphaeria species based on their ribosomal DNA sequences**. *Mycol Res* 1995, **99**:593–603.

20. Dong J, Chen W, Crane JL: **Phylogenetic studies of the Leptosphaeriaceae, Pleosporaceae and some other Loculoascomycetes based on nuclear ribosomal DNA sequences**. *Mycol Res* 1998, **102**:151–156.

21. Montemurro N, Visconti A: **Alternaria metabolites-chemical and biological data**. *Alternaria Biol Plant Dis Metab* 1992:449–557.

22. Liu GT, Qian YZ, Zhang P, Dong ZM, Shi ZY, Zhen YZ, Miao J, Xu YM: **Relationships between Alternaria alternata and oesophageal cancer**. *IARC Sci Publ* 1991:258–262.

23. Bottalico A, Logrieco A: **Toxigenic Alternaria species of economic importance**. *Mycotoxins Agric Food Saf* 1998, **65**:108.

24. Abbas HK, Tanaka T, Shier WT: **Biological activities of synthetic analogues of Alternaria alternata toxin (AAL-toxin) and fumonisin in plant and mammalian cell cultures**. *Phytochemistry* 1995, **40**:1681–1689.

25. Wang H, Li J, Bostock RM, Gilchrist DG: **Apoptosis: a functional paradigm for programmed plant cell death induced by a host-selective phytotoxin and invoked during development.** *Plant Cell Online* 1996, **8**:375–391.

26. Gergen PJ, Turkeltaub PC: **The association of individual allergen reactivity with respiratory disease in a national sample: data from the second National Health and**

Nutrition Examination Survey, 1976–1980 (NHANES II). *J Allergy Clin Immunol* 1992, **90**:579–588.

27. Halonen M, Stern DA, Wright AL, Taussig LM, Martinez FD: **Alternaria as a major allergen for asthma in children raised in a desert environment.** *Am J Respir Crit Care Med* 1997, **155**:1356–1361.

28. Salo PM, Arbes Jr SJ, Sever M, Jaramillo R, Cohn RD, London SJ, Zeldin DC: **Exposure to Alternaria alternata in US homes is associated with asthma symptoms**. *J Allergy Clin Immunol* 2006, **118**:892–898.

29. O'Hollaren MT, Yunginger JW, Offord KP, Somers MJ, O'Connell EJ, Ballard DJ, Sachs MI: **Exposure to an aeroallergen as a possible precipitating factor in respiratory arrest in young patients with asthma**. *N Engl J Med* 1991, **324**:359–363.

30. Andersson M, Downs S, Mitakakis T, Leuppi J, Marks G: **Natural exposure to Alternaria spores induces allergic rhinitis symptoms in sensitized children**. *Pediatr Allergy Immunol* 2003, **14**:100–105.

31. Peat J, Tovey E, Mellis C, Leeder S, Woolcock A: **Importance of house dust mite and Alternaria allergens in childhood asthma: an epidemiological study in two climatic regions of Australia**. *Clin Exp Allergy* 1993, **23**:812–820.

32. Hoffman DR: **Mould allergens**. *Mould Allergy Phila PA Lea Febiger* 1984:104–116.

33. Van Leeuwen WS: **Bronchial asthma in relation to climate**. *Proc R Soc Med* 1924, **17**(Ther Pharmacol Sect):19.

34. Bush RK, Prochnau JJ: **Alternaria-induced asthma**. *J Allergy Clin Immunol* 2004, **113**:227–234.

35. Matsuwaki Y, Wada K, White TA, Benson LM, Charlesworth MC, Checkel JL, Inoue Y, Hotta K, Ponikau JU, Lawrence CB, others: **Recognition of fungal protease activities induces cellular activation and eosinophil-derived neurotoxin release in human eosinophils**. *J Immunol* 2009, **183**:6708–6716.

36. Kouzaki H, O'Grady SM, Lawrence CB, Kita H: **Proteases induce production of thymic stromal lymphopoietin by airway epithelial cells through protease-activated receptor-2**. *J Immunol* 2009, **183**:1427–1434.

37. Anaissie EJ, Bodey GP, Rinaldi MG: **Emerging fungal pathogens**. *Eur J Clin Microbiol Infect Dis* 1989, **8**:323–330.

38. Rossmann SN, Cernoch PL, Davis JR: **Dematiaceous fungi are an increasing cause of human disease**. *Clin Infect Dis* 1996, **22**:73–80.

39. De Bievre C de: **Alternaria spp. pathogenic to man: epidemiology**. *J Mycol Medicale* 1991:50–58.

40. De Hoog GS, Guarro J, Figueras M, Gené J: *Atlas of Clinical Fungi*. *Volume 1*. Centraalbureau voor Schimmelcultures Utrecht; 2000.

41. Bass D, Richards TA: **Three reasons to re-evaluate fungal diversity "on Earth and in the ocean."***Fungal Biol Rev* 2011, **25**:159–164.

42. Hibbett DS, Taylor JW: **Fungal systematics: is a new age of enlightenment at hand?** *Nat Rev Microbiol* 2013, **11**:129–133.

43. Nierman WC, Pain A, Anderson MJ, Wortman JR, Kim HS, Arroyo J, Berriman M, Abe K, Archer DB, Bermejo C, Bennett J, Bowyer P, Chen D, Collins M, Coulsen R, Davies R, Dyer PS, Farman M, Fedorova N, Fedorova N, Feldblyum TV, Fischer R, Fosker N, Fraser A, Garcia JL, Garcia MJ, Goble A, Goldman GH, Gomi K, Griffith-Jones S, et al.: **Genomic sequence of the pathogenic and allergenic filamentous fungus Aspergillus fumigatus**. *Nature* 2005, **438**:1151–1156.

44. Hane JK, Lowe RGT, Solomon PS, Tan K-C, Schoch CL, Spatafora JW, Crous PW, Kodira C, Birren BW, Galagan JE, Torriani SFF, McDonald BA, Oliver RP: **Dothideomycete–Plant Interactions Illuminated by Genome Sequencing and EST Analysis of the Wheat Pathogen Stagonospora nodorum**. *Plant Cell Online* 2007, **19**:3347–3368.

45. Mabey Gilsenan J, Cooley J, Bowyer P: **CADRE: the Central Aspergillus Data REpository 2012**. *Nucleic Acids Res* 2012, **40**(Database issue):D660–666.

46. Cerqueira GC, Arnaud MB, Inglis DO, Skrzypek MS, Binkley G, Simison M, Miyasato SR, Binkley J, Orvis J, Shah P, Wymore F, Sherlock G, Wortman JR: **The Aspergillus Genome Database: multispecies curation and incorporation of RNA-Seq data to improve structural gene annotations**. *Nucleic Acids Res* 2013, **42**:D705–D710.

47. Galagan JE, Calvo SE, Cuomo C, Ma L-J, Wortman JR, Batzoglou S, Lee S-I, Baştürkmen M, Spevak CC, Clutterbuck J, Kapitonov V, Jurka J, Scazzocchio C, Farman M, Butler J, Purcell S, Harris S, Braus GH, Draht O, Busch S, D'Enfert C, Bouchier C, Goldman GH, Bell-Pedersen D, Griffiths-Jones S, Doonan JH, Yu J, Vienken K, Pain A, Freitag M, et al.: **Sequencing of Aspergillus nidulans and comparative analysis with A. fumigatus and A. oryzae**. *Nature* 2005, **438**:1105–1115.

48. Ma L-J, van der Does HC, Borkovich KA, Coleman JJ, Daboussi M-J, Di Pietro A, Dufresne M, Freitag M, Grabherr M, Henrissat B, Houterman PM, Kang S, Shim W-B, Woloshuk C, Xie X, Xu J-R, Antoniw J, Baker SE, Bluhm BH, Breakspear A, Brown DW, Butchko RAE, Chapman S, Coulson R, Coutinho PM, Danchin EGJ, Diener A, Gale LR, Gardiner DM, Goff S, et al.: **Comparative genomics reveals mobile pathogenicity chromosomes in Fusarium**. *Nature* 2010, **464**:367–373.

49. Martinez D, Grigoriev I, Salamov AA: **Annotation of Fungal Genomes**. *Proc ANAS Biol Sci* 2010, **65**:177–183.

50. Haas BJ, Zeng Q, Pearson MD, Cuomo CA, Wortman JR: **Approaches to Fungal Genome Annotation**. *Mycology* 2011, **2**:118–141.

51. Flicek P, Amode MR, Barrell D, Beal K, Brent S, Carvalho-Silva D, Clapham P, Coates G, Fairley S, Fitzgerald S, Gil L, Gordon L, Hendrix M, Hourlier T, Johnson N, Kahari AK, Keefe D, Keenan S, Kinsella R, Komorowska M, Koscielny G, Kulesha E, Larsson P, Longden I, McLaren W, Muffato M, Overduin B, Pignatelli M, Pritchard B, Riat HS, et al.: **Ensembl 2012**. *Nucleic Acids Res* 2011, **40**:D84–D90.

52. Flutre T, Duprat E, Feuillet C, Quesneville H: **Considering Transposable Element Diversification in De Novo Annotation Approaches**. *PLoS ONE* 2011, **6**:e16526.

53. Altschul SF, Madden TL, Schäffer AA, Zhang J, Zhang Z, Miller W, Lipman DJ: **Gapped BLAST and PSI-BLAST: a new generation of protein database search programs**. *Nucleic Acids Res* 1997, **25**:3389–3402.

54. Bao Z, Eddy SR: **Automated De Novo Identification of Repeat Sequence Families in Sequenced Genomes**. *Genome Res* 2002, **12**:1269–1276.

55. Edgar RC, Myers EW: **PILER: identification and classification of genomic repeats**. *Bioinformatics* 2005, **21**(suppl 1):i152–i158.

56. Rouxel T, Grandaubert J, Hane JK, Hoede C, van de Wouw AP, Couloux A, Dominguez V, Anthouard V, Bally P, Bourras S, Cozijnsen AJ, Ciuffetti LM, Degrave A, Dilmaghani A, Duret L, Fudal I, Goodwin SB, Gout L, Glaser N, Linglin J, Kema GHJ, Lapalu N, Lawrence CB, May K, Meyer M, Ollivier B, Poulain J, Schoch CL, Simon A, Spatafora JW, et al.: **Effector diversification within compartments of the Leptosphaeria maculans genome affected by Repeat-Induced Point mutations**. *Nat Commun* 2011, **2**:202.

57. Price AL, Jones NC, Pevzner PA: **De novo identification of repeat families in large genomes**. *Bioinforma Oxf Engl* 2005, **21 Suppl 1**:i351–358.

58. Jurka J, Kapitonov VV, Pavlicek A, Klonowski P, Kohany O, Walichiewicz J: **Repbase Update, a database of eukaryotic repetitive elements**. *Cytogenet Genome Res* 2005, **110**:462–467.

59. Birney E, Clamp M, Durbin R: **GeneWise and Genomewise**. *Genome Res* 2004, **14**:988–995.

60. Stanke M, Waack S: **Gene prediction with a hidden Markov model and a new intron submodel**. *Bioinformatics* 2003, **19**(Suppl 2):ii215–ii225.

61. Ter-Hovhannisyan V, Lomsadze A, Chernoff YO, Borodovsky M: **Gene prediction in novel fungal genomes using an ab initio algorithm with unsupervised training**. *Genome Res* 2008, **18**:1979–1990.

62. Blanco E, Parra G, Guigó R: **Using geneid to identify genes**. *Curr Protoc Bioinforma Ed Board Andreas Baxevanis Al* 2007, **Chapter 4**:Unit 4.3.

63. Rice P, Longden I, Bleasby A: **EMBOSS: the European Molecular Biology Open Software Suite**. *Trends Genet TIG* 2000, **16**:276–277.

64. Bernal A, Crammer K, Pereira F: **Automated gene-model curation using global discriminative learning**. *Bioinformatics* 2012, **28**:1571–1578.

65. Allen JE, Salzberg SL: **JIGSAW: integration of multiple sources of evidence for gene prediction**. *Bioinformatics* 2005, **21**:3596–3603.

66. Lagesen K, Hallin P, Andreas Rodland E, Staerfeldt H-H, Rognes T, Ussery DW: **RNAmmer: consistent and rapid annotation of ribosomal RNA genes**. *Nucl Acids Res* 2007:gkm160.

67. Lowe T, Eddy S: **tRNAscan-SE: a program for improved detection of transfer RNA genes in genomic sequence**. *Nucl Acids Res* 1997, **25**:955–964.

68. Trapnell C, Roberts A, Goff L, Pertea G, Kim D, Kelley DR, Pimentel H, Salzberg SL, Rinn JL, Pachter L: **Differential gene and transcript expression analysis of RNA-seq experiments with TopHat and Cufflinks**. *Nat Protoc* 2012, **7**:562–578.

69. The UniProt Consortium: **The Universal Protein Resource (UniProt)**. *Nucl Acids Res* 2008, **36**:D190–195.

70. Benson DA, Karsch-Mizrachi I, Clark K, Lipman DJ, Ostell J, Sayers EW: **GenBank**. *Nucleic Acids Res* 2011, **40**:D48–D53.

71. Hunter S, Apweiler R, Attwood TK, Bairoch A, Bateman A, Binns D, Bork P, Das U, Daugherty L, Duquenne L, Finn RD, Gough J, Haft D, Hulo N, Kahn D, Kelly E, Laugraud A, Letunic I, Lonsdale D, Lopez R, Madera M, Maslen J, McAnulla C, McDowall J, Mistry J, Mitchell A, Mulder N, Natale D, Orengo C, Quinn AF, et al.: **InterPro: the integrative protein signature database**. *Nucleic Acids Res* 2009, **37**(Database):D211–D215.

72. Marchler-Bauer A, Lu S, Anderson JB, Chitsaz F, Derbyshire MK, DeWeese-Scott C, Fong JH, Geer LY, Geer RC, Gonzales NR, Gwadz M, Hurwitz DI, Jackson JD, Ke Z, Lanczycki CJ, Lu F, Marchler GH, Mullokandov M, Omelchenko MV, Robertson CL, Song JS, Thanki N, Yamashita RA, Zhang D, Zhang N, Zheng C, Bryant SH: **CDD: a Conserved Domain Database for the functional annotation of proteins**. *Nucleic Acids Res* 2011, **39**(suppl 1):D225–D229.

73. Marchler-Bauer A, Panchenko AR, Shoemaker BA, Thiessen PA, Geer LY, Bryant SH: **CDD: a database of conserved domain alignments with links to domain three-dimensional structure**. *Nucleic Acids Res* 2002, **30**:281–283.

74. Conesa A, Gotz S, Garcia-Gomez JM, Terol J, Talon M, Robles M: **Blast2GO: a universal tool for annotation, visualization and analysis in functional genomics research**. *Bioinformatics* 2005, **21**:3674–3676.

75. Stergiopoulos I, de Wit PJGM: **Fungal effector proteins**. *Annu Rev Phytopathol* 2009, **47**:233–263.

76. Bendtsen JD, Nielsen H, von Heijne G, Brunak S: **Improved prediction of signal peptides: SignalP 3.0**. *J Mol Biol* 2004, **340**:783–795.

77. Horton P, Park K-J, Obayashi T, Fujita N, Harada H, Adams-Collier CJ, Nakai K: **WoLF PSORT: protein localization predictor**. *Nucleic Acids Res* 2007, **35**(Web Server issue):W585–W587.

78. Käll L, Krogh A, Sonnhammer EL.: **A Combined Transmembrane Topology and Signal Peptide Prediction Method**. *J Mol Biol* 2004, **338**:1027–1036.

79. Krogh A, Larsson B, von Heijne G, Sonnhammer EL: **Predicting transmembrane protein topology with a hidden Markov model: application to complete genomes**. *J Mol Biol* 2001, **305**:567–580.

80. Cantarel BL, Coutinho PM, Rancurel C, Bernard T, Lombard V, Henrissat B: **The Carbohydrate-Active EnZymes database (CAZy): an expert resource for Glycogenomics**. *Nucleic Acids Res* 2009, **37**(Database):D233–D238.

81. Khaldi N, Seifuddin FT, Turner G, Haft D, Nierman WC, Wolfe KH, Fedorova ND: **SMURF: genomic mapping of fungal secondary metabolite clusters**. *Fungal Genet Biol FG B* 2010, **47**:736–741.

82. Darling ACE, Mau B, Blattner FR, Perna NT: **Mauve: Multiple Alignment of Conserved Genomic Sequence With Rearrangements**. *Genome Res* 2004, **14**:1394–1403.

83. Darling AE, Mau B, Perna NT: **progressiveMauve: Multiple Genome Alignment with Gene Gain, Loss and Rearrangement**. *PLoS ONE* 2010, **5**:e11147.

84. Li L, Stoeckert CJ, Roos DS: **OrthoMCL: Identification of Ortholog Groups for Eukaryotic Genomes**. *Genome Res* 2003, **13**:2178–2189.

85. Galagan JE, Calvo SE, Borkovich KA, Selker EU, Read ND, Jaffe D, FitzHugh W, Ma L-J, Smirnov S, Purcell S, Rehman B, Elkins T, Engels R, Wang S, Nielsen CB, Butler J, Endrizzi M, Qui D, Ianakiev P, Bell-Pedersen D, Nelson MA, Werner-Washburne M, Selitrennikoff CP, Kinsey JA, Braun EL, Zelter A, Schulte U, Kothe GO, Jedd G, Mewes W, et al.: **The genome sequence of the filamentous fungus Neurospora crassa**. *Nature* 2003, **422**:859–868.

86. Hane JK, Lowe RGT, Solomon PS, Tan K-C, Schoch CL, Spatafora JW, Crous PW, Kodira C, Birren BW, Galagan JE, Torriani SFF, McDonald BA, Oliver RP:

Dothideomycete Plant Interactions Illuminated by Genome Sequencing and EST Analysis of the Wheat Pathogen Stagonospora nodorum. *Plant Cell* 2007, **19**:3347–3368.

87. Wicker T, Sabot F, Hua-Van A, Bennetzen JL, Capy P, Chalhoub B, Flavell A, Leroy P, Morgante M, Panaud O, Paux E, SanMiguel P, Schulman AH: **A unified classification system for eukaryotic transposable elements**. *Nat Rev Genet* 2007, **8**:973–982.

88. Spanu PD, Abbott JC, Amselem J, Burgis TA, Soanes DM, Stüber K, Themaat EVL van, Brown JKM, Butcher SA, Gurr SJ, Lebrun M-H, Ridout CJ, Schulze-Lefert P, Talbot NJ, Ahmadinejad N, Ametz C, Barton GR, Benjdia M, Bidzinski P, Bindschedler LV, Both M, Brewer MT, Cadle-Davidson L, Cadle-Davidson MM, Collemare J, Cramer R, Frenkel O, Godfrey D, Harriman J, Hoede C, et al.: **Genome Expansion and Gene Loss in Powdery Mildew Fungi Reveal Tradeoffs in Extreme Parasitism**. *Science* 2010, **330**:1543–1546.

89. Daboussi M-J, Capy P: **Transposable Elements in Filamentous Fungi**. *Annu Rev Microbiol* 2003, **57**:275–299.

90. Thon MR, Pan H, Diener S, Papalas J, Taro A, Mitchell TK, Dean RA: **The role of transposable element clusters in genome evolution and loss of synteny in the rice blast fungus Magnaporthe oryzae**. *Genome Biol* 2006, **7**:R16.

91. Braumann I, van den Berg M, Kempken F: **Strain-specific retrotransposon-mediated recombination in commercially used <i>Aspergillus niger</i> strain**. *Mol Genet Genomics* 2008, **280**:319–325.

92. Cho Y, Davis JW, Kim K-H, Wang J, Sun Q-H, Cramer RA Jr, Lawrence CB: **A high throughput targeted gene disruption method for Alternaria brassicicola functional genomics using linear minimal element (LME) constructs**. *Mol Plant-Microbe Interact MPMI* 2006, **19**:7–15.

93. Kim K-H, Cho Y, LA Rota M, Cramer RA Jr, Lawrence CB: **Functional analysis of the Alternaria brassicicola non-ribosomal peptide synthetase gene AbNPS2 reveals a role in conidial cell wall construction**. *Mol Plant Pathol* 2007, **8**:23–39.

94. Cho Y, Cramer Jr. RA, Kim K-H, Davis J, Mitchell TK, Figuli P, Pryor BM, Lemasters E, Lawrence CB: **The Fus3/Kss1 MAP kinase homolog Amk1 regulates the expression of genes encoding hydrolytic enzymes in Alternaria brassicicola**. *Fungal Genet Biol* 2007, **44**:543–553.

95. Craven KD, Vélëz H, Cho Y, Lawrence CB, Mitchell TK: **Anastomosis is required for virulence of the fungal necrotroph Alternaria brassicicola**. *Eukaryot Cell* 2008, **7**:675–683.

96. Cho Y, Kim K-H, La Rota M, Scott D, Santopietro G, Callihan M, Mitchell TK, Lawrence CB: **Identification of novel virulence factors associated with signal transduction pathways in Alternaria brassicicola**. *Mol Microbiol* 2009, **72**:1316–1333.

97. Wight WD, Kim K-H, Lawrence CB, Walton JD: **Biosynthesis and role in virulence of the histone deacetylase inhibitor depudecin from Alternaria brassicicola**. *Mol Plant Microbe Interact* 2009, **22**:1258–1267.

98. Kim K-H, Willger SD, Park S-W, Puttikamonkul S, Grahl N, Cho Y, Mukhopadhyay B, Cramer RA Jr, Lawrence CB: **TmpL, a transmembrane protein required for intracellular redox homeostasis and virulence in a plant and an animal fungal pathogen**. *PLoS Pathog* 2009, **5**:e1000653.

99. Cho Y, Srivastava A, Ohm RA, Lawrence CB, Wang K-H, Grigoriev IV, Marahatta SP: **Transcription factor Amr1 induces melanin biosynthesis and suppresses virulence in Alternaria brassicicola**. *PLoS Pathog* 2012, **8**:e1002974.

100. Srivastava A, Ohm RA, Oxiles L, Brooks F, Lawrence CB, Grigoriev IV, Cho Y: **A zinc-finger-family transcription factor, AbVf19, is required for the induction of a gene subset important for virulence in Alternaria brassicicola**. *Mol Plant-Microbe Interact MPMI* 2012, **25**:443–452.

101. Srivastava A, Cho IK, Cho Y: **The Bdtf1 gene in Alternaria brassicicola is important in detoxifying brassinin and maintaining virulence on Brassica species**. *Mol Plant-Microbe Interact MPMI* 2013, **26**:1429–1440.

102. Saha D, Fetzner R, Burkhardt B, Podlech J, Metzler M, Dang H, Lawrence C, Fischer R: **Identification of a Polyketide Synthase Required for Alternariol (AOH) and Alternariol-9-Methyl Ether (AME) Formation in Alternaria alternata**. *PLoS ONE* 2012, **7**:e40564.

103. Hu J, Chen C, Peever T, Dang H, Lawrence C, Mitchell T: **Genomic characterization of the conditionally dispensable chromosome in Alternaria arborescens provides evidence for horizontal gene transfer**. *BMC Genomics* 2012, **13**:171.

104. Huang X, Wang J, Aluru S, Yang S-P, Hillier L: **PCAP: A Whole-Genome Assembly Program**. *Genome Res* 2003, **13**:2164–2170.

105. Arnaud MB, Chibucos MC, Costanzo MC, Crabtree J, Inglis DO, Lotia A, Orvis J, Shah P, Skrzypek MS, Binkley G, Miyasato SR, Wortman JR, Sherlock G: **The Aspergillus Genome Database, a curated comparative genomics resource for gene, protein and sequence information for the Aspergillus research community**. *Nucleic Acids Res* 2009, **38**(Database):D420–D427.

106. **Go-perl package** [http://search.cpan.org/~cmungall/go-perl/]

107. Marchler-Bauer A, Lu S, Anderson JB, Chitsaz F, Derbyshire MK, DeWeese-Scott C, Fong JH, Geer LY, Geer RC, Gonzales NR, Gwadz M, Hurwitz DI, Jackson JD, Ke Z, Lanczycki CJ, Lu F, Marchler GH, Mullokandov M, Omelchenko MV, Robertson CL, Song JS, Thanki N, Yamashita RA, Zhang D, Zhang N, Zheng C, Bryant SH: **CDD: a**

Conserved Domain Database for the functional annotation of proteins. *Nucleic Acids Res* 2010, **39**(Database):D225–D229.

108. Darling ACE, Mau B, Blattner FR, Perna NT: **Mauve: Multiple Alignment of Conserved Genomic Sequence With Rearrangements**. *Genome Res* 2004, **14**:1394–1403.

109. Edgar RC: **MUSCLE: multiple sequence alignment with high accuracy and high throughput**. *Nucleic Acids Res* 2004, **32**:1792–1797.

110. Stamatakis A: **RAxML-VI-HPC: maximum likelihood-based phylogenetic analyses with thousands of taxa and mixed models**. *Bioinformatics* 2006, **22**:2688–2690.

111. Whelan S, Goldman N: **A general empirical model of protein evolution derived from multiple protein families using a maximum-likelihood approach**. *Mol Biol Evol* 2001, **18**:691–699.

112. Zeng X, Nesbitt MJ, Pei J, Wang K, Vergara IA, Chen N: **OrthoCluster: A New Tool for Mining Synteny Blocks and Applications in Comparative Genomics**. In *Proc 11th Int Conf Extending Database Technol Adv Database Technol*. New York, NY, USA: ACM; 2008:656–667. [*EDBT '08*]

113. Lombard V, Golaconda Ramulu H, Drula E, Coutinho PM, Henrissat B: **The carbohydrate-active enzymes database (CAZy) in 2013**. *Nucleic Acids Res* 2014, **42**(Database issue):D490–495.

114. Eddy SR: **A new generation of homology search tools based on probabilistic inference**. *Genome Inform Int Conf Genome Inform* 2009, **23**:205–211.

115. R. Core Team: *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing; 2014.

116. Storey JD, Tibshirani R: **Statistical significance for genomewide studies**. *Proc Natl Acad Sci* 2003, **100**:9440–9445.

117. Ohm RA, Feau N, Henrissat B, Schoch CL, Horwitz BA, Barry KW, Condon BJ, Copeland AC, Dhillon B, Glaser F, Hesse CN, Kosti I, LaButti K, Lindquist EA, Lucas S, Salamov AA, Bradshaw RE, Ciuffetti L, Hamelin RC, Kema GHJ, Lawrence C, Scott JA, Spatafora JW, Turgeon BG, de Wit PJGM, Zhong S, Goodwin SB, Grigoriev IV: **Diverse Lifestyles and Strategies of Plant Pathogenesis Encoded in the Genomes of Eighteen Dothideomycetes Fungi**. *PLoS Pathog* 2012, **8**:e1003037.

118. Burge HA, Hoyer ME, Solomon WR, Simmons EG, Gallup J: **Quality control factors for Alternaria allergens**. *Mycotaxon* 1989, **34**:55–63.

119. Babiceanu MC, Howard BA, Rumore AC, Kita H, Lawrence CB: **Analysis of global gene expression changes in human bronchial epithelial cells exposed to spores of the allergenic fungus, Alternaria alternata**. *Front Microbiol* 2013, **4**.

120. Akamatsu H, Taga M, Kodama M, Johnson R, Otani H, Kohmoto K: **Molecular karyotypes for Alternaria plant pathogens known to produce host-specific toxins**. *Curr Genet* 1999, **35**:647–656.

121. Hatta R, Ito K, Hosaki Y, Tanaka T, Tanaka A, Yamamoto M, Akimitsu K, Tsuge T: **A Conditionally Dispensable Chromosome Controls Host-Specific Pathogenicity in the Fungal Plant Pathogen Alternaria alternata**. *Genetics* 2002, **161**:59–70.

122. Lamb JC, Theuri J, Birchler JA: **What's in a centromere?** *Genome Biol* 2004, **5**:239.

123. De Wit PJGM, van der Burgt A, Ökmen B, Stergiopoulos I, Abd-Elsalam KA, Aerts AL, Bahkali AH, Beenen HG, Chettri P, Cox MP, Datema E, de Vries RP, Dhillon B, Ganley AR, Griffiths SA, Guo Y, Hamelin RC, Henrissat B, Kabir MS, Jashni MK, Kema G, Klaubauf S, Lapidus A, Levasseur A, Lindquist E, Mehrabi R, Ohm RA, Owen TJ, Salamov A, Schwelm A, et al.: **The Genomes of the Fungal Plant Pathogens Cladosporium fulvum and Dothistroma septosporum Reveal Adaptation to Different Hosts and Lifestyles But Also Signatures of Common Ancestry**. *PLoS Genet* 2012, **8**:e1003088.

124. Tatusov R, Fedorova N, Jackson J, Jacobs A, Kiryutin B, Koonin E, Krylov D, Mazumder R, Mekhedov S, Nikolskaya A, Rao BS, Smirnov S, Sverdlov A, Vasudevan S, Wolf Y, Yin J, Natale D: **The COG database: an updated version includes eukaryotes**. *BMC Bioinformatics* 2003, **4**:41.

125. Arnaud MB, Chibucos MC, Costanzo MC, Crabtree J, Inglis DO, Lotia A, Orvis J, Shah P, Skrzypek MS, Binkley G, Miyasato SR, Wortman JR, Sherlock G: **The Aspergillus Genome Database, a curated comparative genomics resource for gene, protein and sequence information for the Aspergillus research community**. *Nucleic Acids Res* 2010, **38**(suppl 1):D420–D427.

126. Punta M, Coggill PC, Eberhardt RY, Mistry J, Tate J, Boursnell C, Pang N, Forslund K, Ceric G, Clements J, Heger A, Holm L, Sonnhammer ELL, Eddy SR, Bateman A, Finn RD: **The Pfam protein families database**. *Nucleic Acids Res* 2011, **40**:D290–D301.

127. Glass NL, Jacobson DJ, Shiu PK: **The genetics of hyphal fusion and vegetative incompatibility in filamentous ascomycete fungi**. *Annu Rev Genet* 2000, **34**:165–186.

128. Amselem J, Cuomo CA, van Kan JAL, Viaud M, Benito EP, Couloux A, Coutinho PM, de Vries RP, Dyer PS, Fillinger S, Fournier E, Gout L, Hahn M, Kohn L, Lapalu N, Plummer KM, Pradier J-M, Quévillon E, Sharon A, Simon A, ten Have A, Tudzynski B, Tudzynski P, Wincker P, Andrew M, Anthouard V, Beever RE, Beffa R, Benoit I,

Bouzid O, et al.: **Genomic analysis of the necrotrophic fungal pathogens Sclerotinia sclerotiorum and Botrytis cinerea**. *PLoS Genet* 2011, **7**:e1002230.

129. Battaglia E, Benoit I, Brink J van den, Wiebenga A, Coutinho PM, Henrissat B, Vries RP de: **Carbohydrate-active enzymes from the zygomycete fungus Rhizopus oryzae: a highly specialized approach to carbohydrate degradation depicted at genome level**. *BMC Genomics* 2011, **12**:38.

130. Berka RM, Grigoriev IV, Otillar R, Salamov A, Grimwood J, Reid I, Ishmael N, John T, Darmond C, Moisan M-C, Henrissat B, Coutinho PM, Lombard V, Natvig DO, Lindquist E, Schmutz J, Lucas S, Harris P, Powlowski J, Bellemare A, Taylor D, Butler G, de Vries RP, Allijn IE, van den Brink J, Ushinsky S, Storms R, Powell AJ, Paulsen IT, Elbourne LDH, et al.: **Comparative genomic analysis of the thermophilic biomass-degrading fungi Myceliophthora thermophila and Thielavia terrestris**. *Nat Biotechnol* 2011, **29**:922–927.

131. Coutinho PM, Andersen MR, Kolenova K, vanKuyk PA, Benoit I, Gruben BS, Trejo-Aguilar B, Visser H, van Solingen P, Pakula T, Seiboth B, Battaglia E, Aguilar-Osorio G, de Jong JF, Ohm RA, Aguilar M, Henrissat B, Nielsen J, Stålbrand H, de Vries RP: **Post-genomic insights into the plant polysaccharide degradation potential of Aspergillus nidulans and comparison to Aspergillus niger and Aspergillus oryzae**. *Fungal Genet Biol FG B* 2009, **46 Suppl 1**:S161–S169.

132. Espagne E, Lespinet O, Malagnac F, Da Silva C, Jaillon O, Porcel BM, Couloux A, Aury J-M, Ségurens B, Poulain J, Anthouard V, Grossetete S, Khalili H, Coppin E, Déquard-Chablat M, Picard M, Contamine V, Arnaise S, Bourdais A, Berteaux-Lecellier V, Gautheret D, de Vries RP, Battaglia E, Coutinho PM, Danchin EG, Henrissat B, Khoury RE, Sainsard-Chanet A, Boivin A, Pinan-Lucarré B, et al.: **The genome sequence of the model ascomycete fungus Podospora anserina**. *Genome Biol* 2008, **9**:R77.

133. Floudas D, Binder M, Riley R, Barry K, Blanchette RA, Henrissat B, Martínez AT, Otillar R, Spatafora JW, Yadav JS, Aerts A, Benoit I, Boyd A, Carlson A, Copeland A, Coutinho PM, Vries RP de, Ferreira P, Findley K, Foster B, Gaskell J, Glotzer D, Górecki P, Heitman J, Hesse C, Hori C, Igarashi K, Jurgens JA, Kallen N, Kersten P, et al.: **The Paleozoic Origin of Enzymatic Lignin Decomposition Reconstructed from 31 Fungal Genomes**. *Science* 2012, **336**:1715–1719.

134. Lévesque CA, Brouwer H, Cano L, Hamilton JP, Holt C, Huitema E, Raffaele S, Robideau GP, Thines M, Win J, Zerillo MM, Beakes GW, Boore JL, Busam D, Dumas B, Ferriera S, Fuerstenberg SI, Gachon CMM, Gaulin E, Govers F, Grenville-Briggs L, Horner N, Hostetler J, Jiang RHY, Johnson J, Krajaejun T, Lin H, Meijer HJG, Moore B, Morris P, et al.: **Genome sequence of the necrotrophic plant pathogen Pythium ultimum reveals original pathogenicity mechanisms and effector repertoire**. *Genome Biol* 2010, **11**:R73.

135. Kurup VP, Shen H-D, Banerjee B: **Respiratory fungal allergy**. *Microbes Infect* 2000, **2**:1101–1110.

136. Snelgrove RJ, Gregory LG, Peiró T, Akthar S, Campbell GA, Walker SA, Lloyd CM: **Alternaria-derived serine protease activity drives IL-33 mediated asthma exacerbations(☆)**. *J Allergy Clin Immunol* 2014.

137. Beauvais A, Monod M, Wyniger J, Debeaupuis JP, Grouzmann E, Brakch N, Svab J, Hovanessian AG, Latgé JP: **Dipeptidyl-peptidase IV secreted by Aspergillus fumigatus, a fungus pathogenic to humans.** *Infect Immun* 1997, **65**:3042–3047.

138. **Allergen Online Database** [http://www.allergenonline.org/]

139. Srivastava A, Ohm RA, Oxiles L, Brooks F, Lawrence CB, Grigoriev IV, Cho Y: **A Zinc-Finger-Family Transcription Factor, AbVf19, Is Required for the Induction of a Gene Subset Important for Virulence in Alternaria brassicicola**. *Mol Plant Microbe Interact* 2012, **25**:443–452.

140. Kulkarni RD, Kelkar HS, Dean RA: **An eight-cysteine-containing CFEM domain unique to a group of fungal membrane proteins**. *Trends Biochem Sci* 2003, **28**:118–121.

141. Cox RJ: **Polyketides, proteins and genes in fungi: programmed nano-machines begin to reveal their secrets**. *Org Biomol Chem* 2007, **5**:2010.

142. Kim K-H: **Functional Analysis of Secondary Metabolite Biosynthesis-Related Genes in Alternaria brassicicola**. *PhD Dissertation*. Virginia Tech; 2009.

143. Kim K-H, Cho Y, La Rota M, Cramer RA, Lawrence CB: **Functional analysis of the Alternaria brassicicola non-ribosomal peptide synthetase gene AbNPS2 reveals a role in conidial cell wall construction**. *Mol Plant Pathol* 2007, **8**:23–39.

144. Kim K-H, Willger SD, Park S-W, Puttikamonkul S, Grahl N, Cho Y, Mukhopadhyay B, Cramer RA, Lawrence CB: **TmpL, a Transmembrane Protein Required for Intracellular Redox Homeostasis and Virulence in a Plant and an Animal Fungal Pathogen**. *PLoS Pathog* 2009, **5**:e1000653.

145. Ladics GS, Cressman RF, Herouet-Guicheney C, Herman RA, Privalle L, Song P, Ward JM, McClain S: **Bioinformatics and the allergy assessment of agricultural biotechnology products: Industry practices and recommendations**. *Regul Toxicol Pharmacol* 2011, **60**:46–53.

146. Beyer K, Bardina L, Grishina G, Sampson HA: **Identification of sesame seed allergens by 2-dimensional proteomics and Edman sequencing: Seed storage proteins as common food allergens**. *J Allergy Clin Immunol* 2002, **110**:154–159.

147. Sancho AI, Mills ENC: **Proteomic approaches for qualitative and quantitative characterisation of food allergens**. *Regul Toxicol Pharmacol* 2010, **58**(3, Supplement):S42–S46.

148. FAO/WHO: *Evaluation of Allergenicity of Genetically Modified Foods. Report of a Joint FAO/WHO Expert Consultation on Allergenicity of Foods Derived from Biotechnology*. Rome, Italy; 2001.

149. Codex Alimentarius Commission: *Foods Derived from Modern Biotechnology*. Rome, Italy; 2009.

150. Mari A, Rasi C, Palazzo P, Scala E: **Allergen databases: Current status and perspectives**. *Curr Allergy Asthma Rep* 2009, **9**:376–383.

151. Metcalfe DD: **Genetically modified crops and allergenicity**. *Nat Immunol* 2005, **6**:857–860.

152. Stadler MB, Stadler BM: **Allergenicity prediction by protein sequence**. *FASEB J Off Publ Fed Am Soc Exp Biol* 2003, **17**:1141–1143.

153. Mirsky HP, Cressman Jr. RF, Ladics GS: **Comparative assessment of multiple criteria for the in silico prediction of cross-reactivity of proteins to known allergens**. *Regul Toxicol Pharmacol* 2013, **67**:232–239.

154. Pearson WR, Lipman DJ: **Improved tools for biological sequence comparison**. *Proc Natl Acad Sci U S A* 1988, **85**:2444–2448.

155. Fiers MW, Kleter GA, Nijland H, Peijnenburg AA, Nap JP, Ham RC van: **Allermatch™, a webtool for the prediction of potential allergenicity according to current FAO/WHO Codex alimentarius guidelines**. *BMC Bioinformatics* 2004, **5**:133.

156. Magrane M, Consortium U: **UniProt Knowledgebase: a hub of integrated protein data**. *Database* 2011, **2011**:bar009–bar009.

157. Soeria-Atmadja D, Lundell T, Gustafsson MG, Hammerling U: **Computational detection of allergenic proteins attains a new level of accuracy with in silico variable-length peptide extraction and machine learning**. *Nucleic Acids Res* 2006, **34**:3779–3793.

158. Barrio AM, Soeria-Atmadja D, Nistér A, Gustafsson MG, Hammerling U, Bongcam-Rudloff E: **EVALLER: a web server for in silico assessment of potential protein allergenicity**. *Nucleic Acids Res* 2007, **35**(suppl 2):W694–W700.

159. Muh HC, Tong JC, Tammi MT: **AllerHunter: A SVM-Pairwise System for Assessment of Allergenicity and Allergic Cross-Reactivity in Proteins**. *PLoS ONE* 2009, **4**:e5861.

160. Zhang L-D, Huang Y-Y, Zou Z-H, He Y, Chen X-M, Tao A-L: **SORTALLER: Predicting Allergens Using Substantially Optimized Algorithm on Allergen Family Featured Peptides**. *Bioinformatics* 2012.

161. Khan Z, Bloom JS, Kruglyak L, Singh M: **A practical algorithm for finding maximal exact matches in large sequence datasets using sparse suffix arrays**. *Bioinformatics* 2009, **25**:1609–1616.

162. Davis J, Goadrich M: **The relationship between Precision-Recall and ROC curves**. In *Proc 23rd Int Conf Mach Learn*. New York, NY, USA: ACM; 2006:233–240. [*ICML '06*]

163. Manning CD, Raghavan P, Schütze H: *Introduction to Information Retrieval*. New York: Cambridge University Press; 2008.

164. Baker D, Sali A: **Protein Structure Prediction and Structural Genomics**. *Science* 2001, **294**:93–96.

165. Li K-B, Issac P, Krishnan A: **Predicting allergenic proteins using wavelet transform**. *Bioinformatics* 2004, **20**:2572–2578.

166. Saha S, Raghava GPS: **AlgPred: prediction of allergenic proteins and mapping of IgE epitopes**. *Nucleic Acids Res* 2006, **34**(Web Server):W202–W209.

167. Cui J, Han LY, Li H, Ung CY, Tang ZQ, Zheng CJ, Cao ZW, Chen YZ: **Computer prediction of allergen proteins from sequence-derived protein structural and physicochemical properties**. *Mol Immunol* 2007, **44**:514–520.

168. Riaz T, Hor HL, Krishnan A, Tang F, Li K-B: **WebAllergen: a web server for predicting allergenic proteins**. *Bioinformatics* 2005, **21**:2570–2571.

169. Ivanciuc O, Midoro-Horiuti T, Schein CH, Xie L, Hillman GR, Goldblum RM, Braun W: **The property distance index PD predicts peptides that cross-react with IgE antibodies**. *Mol Immunol* 2009, **46**:873–883.

170. Joachims T: *Learning to Classify Text Using Support Vector Machines*. Boston: Kluwer Academic Publishers; 2002.

171. Zipf GK: *Human Behavior and the Principle of Least Effort: An Introduction to Human Ecology*. Mansfield Centre, CT: Martino Pub.; 1949.

172. Boser BE, Guyon IM, Vapnik VN: **A Training Algorithm for Optimal Margin Classifiers**. In *Proc 5th Annu ACM Workshop Comput Learn Theory*. ACM Press; 1992:144–152.

173. Cortes C, Vapnik V: **Support-vector networks**. *Mach Learn* 1995, **20**:273–297.

174. Burges CJC: **A Tutorial on Support Vector Machines for Pattern Recognition**. *Data Min Knowl Discov* 1998, **2**:121–167.

175. Platt JC: **Probabilistic Outputs for Support Vector Machines and Comparisons to Regularized Likelihood Methods**. In *Adv Large Margin Classif*. MIT Press; 1999:61–74.

176. Lin H-T, Lin C-J, Weng RC: **A note on Platt's probabilistic outputs for support vector machines**. *Mach Learn* 2007, **68**:267–276.

177. Joachims T: **Advances in kernel methods**. Edited by Schölkopf B, Burges CJC, Smola AJ. Cambridge, MA, USA: MIT Press; 1999:169–184.

178. Pereira F, Tishby N, Lee L: **Distributional clustering of English words**. In *Proc 31st Annu Meet Assoc Comput Linguist*. Stroudsburg, PA, USA: Association for Computational Linguistics; 1993:183–190. [*ACL '93*]

179. Baker LD, McCallum AK: **Distributional clustering of words for text classification**. In *Proc 21st Annu Int ACM SIGIR Conf Res Dev Inf Retr*. New York, NY, USA: ACM; 1998:96–103. [*SIGIR '98*]

180. Kodzius R, Rhyner C, Konthur Z, Buczek D, Lehrach H, Walter G, Crameri R: **Rapid identification of allergen-encoding cDNA clones by phage display and high-density arrays**. *Comb Chem High Throughput Screen* 2003, **6**:147–154.