5362200

Binuclear zinc transcription factors and the regulation of patulin biosynthesis in the filamentous Ascomycete *Penicillium expansum*

DIVISION OF BIOTECHNOLOGY | FACULTY OF ENGINEERING | LUND UNIVERSITY
ROBERT HANSEN JAGRELIUS | MASTER THESIS 2021

Printed by Media-Tryck, Lund 2021 NORDIC SWAN ECOLABEL 3041 0903

LUND UNIVERSITY

LUND UNIVERSITY
FACULTY OF ENGINEERING
DIVISION OF BIOTECHNOLOGY

Binuclear zinc transcription factors and the regulation of patulin biosynthesis in the filamentous Ascomycete *Penicillium expansum*

Robert Hansen Jagrelius
June 17th, 2021

Master's thesis work carried out at the
Department of Synthetic Biology at DTU.

Supervisors: Assoc. Professor Rasmus JN Frandsen, rasf@bio.dtu.dk
Dr. Nélida Leiva Eriksson, nelida.leiva_eriksson@biotek.lu
Prof. Rajni Hatti Kaul, rajni.hatti-kaul@biotek.lu.se

Examiner: Dr. Magnus Carlquist, magnus.carlquist@tmb.lth.se

Abstract

This thesis investigated the regulation of patulin production in *Penicillium expansum* in relation to fungal global regulators as well as the cluster-specific binuclear zinc transcription factor (BZTF) patL *in silico*. It also explored the state of BZTF research in filamentous ascomycetes fungi, and questioned if results in *Saccharomyces cerevisiae* can be freely translated to other taxonomic groups. Results indicate that a similar percentage of BZTFs bind as homodimers in filamentous ascomycetes as in *S. cerevisiae*, however, they also suggest that BZTFs are more prone to bind as monomers than previously thought, which could be beneficial in binding site prediction. However, the results are inconclusive if the presence of a coiled-coil domain can be used to predict the type of binding site. Global regulators are involved in a complex regulation system of patulin and 5'- CCBRAAGGAG -3' is identified as a putative binding motif for patL. A new generation of meta binding site prediction tools is proposed.

Acknowledgements

I would like to offer my utmost gratitude to Associate Professor Rasmus John Normand Frandsen for his guidance and continued support through thick and thin in this master thesis, which was troubled by the COVID-19 pandemic. I would also like to thank everyone who has made these formative years a fantastic experience at Lund University.

Table of Contents

Abstract 1

Acknowledgements 1

Table of Contents 2

Aim 4

Background 4

Metabolism 4

Secondary Metabolites and their uses 5

Fungi and their impact on human industries 7

Plant pathogenic fungi and agricultural production 8

Biosynthetic Gene Clusters (BGCs) 10

Evolution of Biosynthetic Gene Clusters 11

Regulation of Secondary Metabolism 13

Global Regulators 15

Local Regulators 18

Binuclear Zinc Transcription Factors 19

Patulin 21

Motif Discovery and Confirmation 24

Electrophoretic mobility shift assay 24

DNase I footprinting analysis 26

Chromatin Immunoprecipitation and related techniques 27

Technical Background 29

AntiSMASH 29

RSAT 29

MEME 29

Qiagen CLC Main Workbench 30

DeepCoil 30

EasyFig 30

AUGUSTUS 30

Method 30

Sequence Acquisition 31

Identification of putative patulin clusters in patulin-producing filamentous ascomycetes 31

Identification of Global regulator homologues in silico in P. expansum IBT34672 31

Identification of binding sites for global regulators in the patulin biosynthetic gene cluster in P. expansum IBT34672 32

Identification of putative binding sites for patL in the patulin biosynthetic gene cluster in P. expansum IBT34672 through alignment 32

Identification of putative zinc binuclear transcription factors in P. expansum, A. nidulans, F. graminium, N. crassa and S. cerevisiae and characterization of coiled-coil domains 32
Characterization of coiled-coil domain and relation to type of binding site of characterized zinc binuclear transcription factors 33

Results 33
Identification of Global regulator homologues in silico in P. expansum IBT34672 34
Identification of binding sites for global regulators in the patulin BGC in P. expansum IBT34672 35
Identification of putative binding sites for patL in the patulin biosynthetic gene cluster in P. expansum IBT34672 through alignment 37
Identification of putative zinc binuclear transcription factors in filamentous ascomycetes and S. cerevisiae and characterization of coiled-coil domain **Fel! Bokmärket är inte definierat.**
Characterization of coiled-coil domain and relation to type of binding site of characterized zinc binuclear transcription factors 42

Discussion 43
Identification of Global regulator homologues in silico in P. expansum IBT34672 43
Identification of binding sites for global regulators in the patulin BGC in P. expansum IBT34672 43
Identification of putative binding sites for patL in the patulin biosynthetic gene cluster in P. expansum IBT34672 for alignment 45
Identification of filamentous ascomycetes and S. cerevisiae and characterization of coiled-coil domain 46
Characterization of coiled-coil domain and relation to type of binding site of characterized zinc binuclear transcription factors 48

Conclusion 49

References 50

Aim

Research on fungal binuclear zinc transcription factors (BZTFs) has historically mostly been done with *S. cerevisiae* and results are often directly translated to filamentous ascomycetes. The presence of fungal BZTFs which bind to non-consensus binding sites indicates that the commonly accepted homodimerically binding model BZTF may be outdated. This thesis explores BZTFs in filamentous ascomycetes, using the biosynthetic gene cluster (BGC) for the production of the mycotoxin patulin and the related BZTF patL in *P. expansum* as a model organism due to economic importance and research interest. This study aims to investigate if the presence of coiled-coil dimerization domain as characterized by *in silico* tools could be helpful when predicting BZTF binding sites, and if the typical model BZTF is still representative for both filamentous ascomycetes and yeast. Additionally, global regulators are important fundamental controllers of the fungal secondary metabolism, and their presence in *P. expansum* and involvement regarding the mycotoxin are to be investigated. The use of binding site prediction tools and difficulty of experimentally confirming binding sites is to be explored and discussed.

Background

Metabolism

Metabolism is one of the commonly recognized core properties of life (Jagers op Akkerhuis, 2010). Metabolism fulfills crucial functions for all living organisms and is central to life's ability to convert food into action (Blanco and Blanco, 2017).

Metabolism is driven forward by enzymatic reactions - which catalyse and enable the conversion of one compound into another. Compounds which are produced as a result of metabolism and its reactions are called metabolites (Blanco and Blanco, 2017) and are often classified as either primary or secondary. The primary metabolites are constantly produced (catabolism) and consumed (anabolism) and are involved with primary functions such as the generation of energy, growth, and reproduction (Reece et al., 2011). Secondary metabolites, on the other hand, are a bit more diverse in their functionality. They have key roles in ecological functions - relating to how the organism reacts and interacts with its environment (Assaf et al., 2020). In fungi, for example, the diversity of secondary metabolites is great -serving functions from UVs protection to ROS scavenging to communication with other microbes (Yu and Keller, 2005). The secondary metabolism is thus crucial to the behaviour of the organism and research thereof is a key part in the study of the taxonomy groups that produce them. Furthermore, secondary metabolites are of great interest to mankind as many of them have bioactive properties (Bérdy, 2005).

The first cornerstones of genetic engineering were set when James Watson and Francis Crick discovered the double helix structure of DNA in 1953. It was a foundation for new technologies which enabled control and more detailed investigation of secondary metabolism. It also set the stage for the field of biotechnology, which relies on the fact that microorganisms can be changed into miniature cell factories (Davy, Kildegaard and Andersen, 2017).

A large share of the compounds produced in the biotechnology industry are derived from primary metabolism, such as simple alcohols and smaller acids, and are crucial for industrial applications (Davy, Kildegaard and Andersen, 2017). However, secondary metabolites, and especially the bioactive ones are also of great importance being used as e.g., drugs for the treatment of diseases (Bérdy, 2005).

Secondary Metabolites and their uses

Secondary metabolites (SecMet) are commonly produced by bacteria, plants or fungi, typically as a response to external stimuli, such as the presence of competitors, low nutrition levels in the environment or abiotic stress factors such as high pH (Demain and Fang, 2000).

The SecMets cover a set of structural chemical classes that display a high diversity with respect to bioactivity. The production of individual classes and types of secondary metabolites are often restricted to a single phylogenetic group with similar niche, as a specific SecMet can often be useful to the host in a very narrow habitat and setting (Braga, Dourado and Araújo, 2016). Examples include the complex interaction between the mycoparasitic fungi *Stachybotrys elegenas* and its host *Rhizoctonia solani,* where both sides produce secondary metabolites to combat each other (Chamoun, Aliferis and Jabaji, 2015)

Secondary metabolites are also sometimes known as "natural products" - a term which originates from a time where the specific identity and functionality of most, if not all, of these compounds were unknown (Williams, Stone, Hauck and Rahman, 1989). It also calls back to the long history of mankind's utilization of these compounds. Clay tablets in ancient Mesopotamia dating almost 5000 years old document oils from cypress and myrrh as a way to treat colds - practices still in use today.

SecMet's have also proven very successful as leads for the development of new drugs for pharmaceutical applications (Dias, Urban and Roessner, 2012). Indeed, many ancient treatments have been used to develop modern medicine. The most notable recent example would be the work of Tu Youyou to analyse old chinese medicinal manuscripts to identify the hugely impactful malarial medicine artemisinin, for which she received the Nobel Prize in Medicine in 2015 (Liao, 2009).

Natural products from macroscopic fungi (basidiomycetes) have seen the same widespread use historically, with notable examples from ancient China and psilocybin-producing mushrooms from South America (Wasser, 2010). The use of SecMet's from microscopic organisms did not start until much later, which may be due to the practical difficulties of handling and culturing organisms often invisible to the naked eye. Today, microorganisms are the main source of new SecMet's and one of the most famous secondary metabolites ever is the antibiotic compound penicillin. It was discovered by serendipity in 1929 by Alexander Fleming from the filamentous ascomycete *Penicillium rubens* (Houbraken, Frisvad and Samson, 2011).

Interestingly enough, even though humans have massive use of these secondary metabolites, we still lack knowledge of what functions many of these metabolites have in the producing organisms. One exception to this are the metabolites utilized in microbial warfare, such as antibiotics and antifungal compounds. A high percentage of filamentous ascomycetes produce antibiotics and antifungals naturally which is a central part in many microorganisms' strategy of killing or inhibiting the action of competing microorganisms (Demain and Fang, 2000).

Many SecMets are important for pathogenic microorganisms and are crucial for the organisms' ability to break down the defenses of a plant or animal's hosts. These are called "virulence factors". They are commonly mentioned when discussing *Penicillium* molds and other filamentous ascomycetes that infect crops. One example are the pigments known as melanins (Hamilton and Gómez, 2002). In humans these are associated with UV-protection of the skin, however in fungi they may also serve other functions. Albino mutants of the several ascomycete genera deficient in melanin production have been shown to have reduced virulence on both animal and plant hosts (Jacobson, 2000). One possible explanation for this is thought to be that they are important cross-linkers in the cell walls of the fungi. In plant-pathogens such as *Magnaporthe grisea* this may enable them to form pressures high enough to properly penetrate the epithelium of the plants and enable infection (Jacobson, 2000).

In the post-harvest apple pathogen *Penicillium expansum*, the mycotoxin patulin has further been described to be a virulence factor. Patulin is however not essential for the success of the primary infection process, but rather impacts the spread of the fungus in the apple tissue impacting the severity of the disease, and hence has been deemed to be an aggressiveness factor. Little research has been done on how this works on a molecular level (Snini et. al, 2016)

Nevertheless, the function/role of most known SecMet's are complete enigmas. For some we can formulate hypotheses about their role in nature based on their bioactivity, identified in drug screens. One example of this is andrastins that are also produced by *P. expansum*. Andrastins falls under a new group of experimental cancer treatments drugs, which are known to inhibit the farnesylation of the prominent Ras protein, an well-known oncogene (Alton, Cox, Gerard Toussaint and Westwick, 2001). The function of andrastins in fungi themselves are unknown, but observations suggest it may have roles in adjusting the lipid composition of the cell membrane, sporulation, and regulating the sterol profile (Kozlovsky, 2019)(Kim et al., 2016).

The lack of knowledge regarding the natural function of SecMet is the norm for a large majority of the several thousands of secondary metabolites that have been characterized until now (Romero, Traxler, López and Kolter, 2011). The production of SecMets and preservation of their extensive pathways is very costly in terms of energy. They must be of benefit to the cell, otherwise they would have been lost due to natural selection.

Fungi and their impact on human industries

The fungal kingdom is denoted as distinctly separate from the other eukaryotic kingdoms, like plants and animals, see plantae and metazoa in Figure 1. There is considerable diversity within the kingdom itself, ranging from the well-known edible mushrooms to dangerous pathogenic molds and aquatic flagellate zoospores (Money, 2016). The kingdom has seen heavy revisions within the last twenty years as a result of improved techniques and increased application of molecular genetics within phylogenetics.

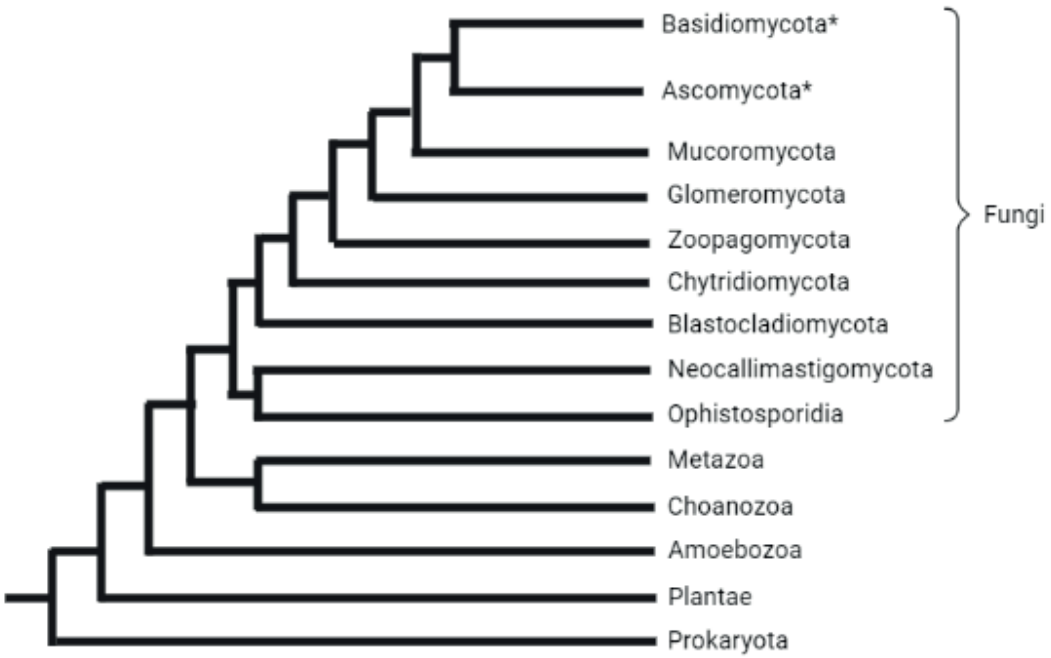


Figure 1: Phylogenetic tree of fungi with the inclusion of related taxonomic branches. *Basidiomycota and ascomycota are the two phylums which include the non-taxonomic groups filamentous fungi and yeast which are defined on their multicellular versus unicellular growth patterns. (Spatafora et al., 2016) (Naranjo-Ortiz and Gabaldón, 2019).

In 2007, as a result of a wide and sprawling collaboration, the kingdom had been divided into seven phyla, namely Basidiomycota, Ascomycota, Glomeromycota, Microsporidia, Chytridiomycota, Blastocladiomycota andNeocallimastigomycota (Hibbett et al., 2007). Revisions, elevations of subphylums and additions of two additional phylums; Ophistosporidia and Mucoromycota, being defined later on resulted in a total of 9 phyla and the phylogenetic tree seen in Figure 1 (Naranjo-Ortiz and Gabaldón, 2019). However, the composition of the tree is under constant revision and debate (Wijayawardene et al., 2020)

Two of phyla, the ascomycota and basidiomycota, contain most of all known fungal species and covers everything from mushrooms to *Saccharomyces cerevisiae*. A separate classification scheme is based on the

growth/morphology characteristics of the given species, as being either a filamentous fungus or yeasts. Neither are true phylogenetic groups, but are spread throughout the phylums Ascomycota and Basidiomycota. Filamentous fungi grow in multicellular structures called hyphae, which are elongated fungal cells with apical growth, resulting in thread-like chains of hyphae that forms a mycelium. Yeasts are unicellular and round fungi, which grow by budding. Some species are dimorphic and can switch between the two growth morphologies (Gauthier, 2017).

The most biotechnologically noteworthy member of the yeast group is the ascomycete *Saccharomyces cerevisiae* - colloquially known as baker's yeast or simply yeast. The species and its close relatives have been in use for e.g. bread baking and beer brewing for thousands of years, with records of yeasted bread existing all the way back to ancient Egypt in 1300 BCE, while the principle itself is probably much older than that (Samuel, 1996). *S. cerevisiae* is today widely used in the biotech industry as a production host. Relevant early large-scale examples include the mass productions of the yeast for beer brewing (Mattanovich, Sauer and Gasser, 2014). The rise of the biotech industry and genetic engineering tools have led to *S. cerevisiae* existing as thousands of commercial strains, optimized for producing everything from ethanol for biofuels, produced at industrial scale, to heterologous metabolites such as the antimalarial medicine artemisinin (Kung et al., 2018)(Mattanovich, Sauer and Gasser, 2014).

The first modern biotechnical industry however was not with yeast, but filamentous ascomycetes. In 1917, the chemist James Currie published a paper demonstrating that *Aspergillus niger* could produce high amounts of citric acid if grown in sugar solutions with low pH. This led to the first pilot plants for biochemical production - which were a huge success (Cairns, Nai and Meyer, 2018). Even today, 99% of global production of citric acid occurs via the fungus (Sweis and Cressey, 2018). The fungus and its diverse strains are also used within enzyme production, notably glucoamylase which is used to make glucose from starches (Silano et al., 2018).

As the field of biotechnology is working with living organisms, the process is always under-optimized with many byproducts being produced, reducing the yield of the desired product. This problem has spawned the emergence of new research fields focused on how to optimize metabolic pathways and perform metabolic engineering. There are many ways to optimize a process - from adjusting bioreactor parameters, to directed evolutions of enzymes, and the deactivation of pathways. This is why understanding of the secondary metabolism is crucial, as knowledge of its regulation is often applicable to the industry. However, there is more to fungi than industrial production.

Even though relationships to fungi are in parts very beneficial, with major applications for both the food and the biotechnology industry, many fungi are known pathogens. Not only are human fungal diseases often woefully underfunded for the impact they actually have compared to more recognizable diseases like malaria, pathogenic fungi also have drastic effects on agriculture (Almeida, Rodrigues and Coelho, 2019).

Plant pathogenic fungi and agricultural production

An estimated third of all crops worldwide are lost to fungal pathogens each year (Fisher et al., 2012). Most impactful are the fungi infecting staple crops e.g. *Magnaporthe oryzae* for rice and wheat and *Botrytis cinerea* for a wide variety of hosts including grapes, bulb fruits and strawberries (Almeida, Rodrigues and Coelho, 2019).

Food spoilage and waste is not the only dilemma, however, as many pathogenic fungi produce mycotoxins. Mycotoxins are secondary metabolites which pose a threat to animals and humans ingesting them through the consumption of infected foodstuff. Among the most prominent are the aflatoxins - which are produced by *Aspergillus* spp., some of which are strong carcinogens and regular ingestion of e.g. contaminated peanuts could lead to detrimental health effects (Mahato et al., 2019).

The genus *Aspergillus* was named and described by Italian priest and biologist Pier Antonio Micheli in 1729. The genus itself is diverse with various habitats. Alongside the food pathogens producing aflatoxin, the species can cause a lung infection in immunocompromised humans as well.

Another well-known genera within filamentous ascomycetes is *Penicillium*. The genus was first described in 1809 by the German naturalist Johann Heinrich Friedrich Link which gave them their name based on the brush- or pencil-like structures of their spores (Visagie et al., 2014). The genus is perhaps most famous for the ability of some of its members to produce the antibiotic and secondary metabolite penicillin. Species of the genus are decomposers but many are also prominent pathogens to the agricultural industry such as *P. digitatum* and *P. verrucosum*. A few of the species are crucial in the production of certain food products. Blue cheeses such as Roquefort for example, is made by inoculation of cheese with *P. roqueforti* (Visagie et al., 2014).

A species of filamentous ascomycetes which is responsible both for produce loss and mycotoxin production is *P. expansum*. The fungus causes the Blue mold disease, which is a very impactful post-harvest disease of apples and other produce. Estimates of the economic loss due to the fungi in the post-harvest stage lie between 10% and 30% depending on the industry (Errampalli, 2014). The fungi is present worldwide in natural environments and especially soil.

P. expansum is known to produce a myriad of interesting secondary metabolites, among which patulin and andrastins have been mentioned. Noteworthy compounds include citrinin - a mycotoxin and antibiotic, chaetoglobosins - which are cytotoxic compounds with potential cancer treatment capabilities and geosmin - a odorant compound with a characteristic earthy smell. (Tannous et al., 2017). Many of these compounds relate to how the fungi interacts with other organisms. The citrinin and chaetoglobosin are thought to be involved in competition, but even geosmin is likely relevant. It is hypothesized to be related to how the fungus can deter predators and also attract small soil animals to help with spore spreading (Stensmyr et al., 2012). The production of secondary metabolites is thus hypothesized to be essential for how it relates to its surroundings. Microbial communities that consist of a multitude of microorganisms, each with its own metabolic fingerprint, is likely that secondary metabolites are very impactful in how these develop. It is an interesting proposition that many of these compounds display a broader spectrum of functionalities than what we may be drawn to believe and are likely involved in microbial communication.

As *P. expansum* is an economically impactful and well-researched filamentous ascomycete, it can serve as a model organism for the investigation into the role of secondary metabolism for the fungus and the community it inhabits. It is further possible that an understanding of how metabolites shape communities can be used for preventing pathogens from infecting crops.

To properly analyse the role of SecMets in communities, we first need to understand more about how SecMets are biosynthesized and how their production is regulated in the fungus.

Biosynthetic Gene Clusters (BGCs)

Genes encoding the enzymes that are responsible for synthesis of the individual secondary metabolites such as patulin are in fungi generally grouped together in so-called biosynthetic gene clusters (BGCs). The BGCs generally include a core biosynthesis enzyme responsible for producing a complex carbon backbone from simple primary metabolites. The core synthases typically belong to one of three types - either a polyketide synthase (PKS), a terpene cyclase or a non-ribosomal peptide synthetase (NRPS) (Assaf et. al., 2020). The BGCs typically also include additional biosynthetic genes or "tailoring" enzymes - which modify the core synthase product (Fig. 2). Examples of these include methyltransferases and P450 monooxygenases. Several additional types of genes can also be present. These include genes encoding membrane transporters that enable proper localization of the end products, and resistance genes encoding proteins/enzymes that protect the cell itself from dangerous compounds such as gliotoxin (Schrettl et al., 2010) (Kwon et al., 2020).

Up to 50% of fungal BGCs also contain a gene coding for a cluster-specific transcription factor - which means they are self-regulating. These transcription factors are proteins which bind to fitting binding motifs found upstreams of the individual genes in the clusters resulting in upregulation or downregulation of gene expression, depending on the factor (Keller, 2018) (Li et al., 2015). Additionally, it has been shown that biosynthetic gene clusters are also under the control of so called 'global regulators' - which can adjust the expression of entire clusters based on more generic external stimuli such as iron levels or pH (Brakhage, 2012). In this thesis I have chosen to use the biosynthetic gene cluster for the mycotoxin patulin as a model cluster for my analysis. A choice that is based on the large interest in the compound,, alongside the presence of a local transcription factor (Fig. 2).

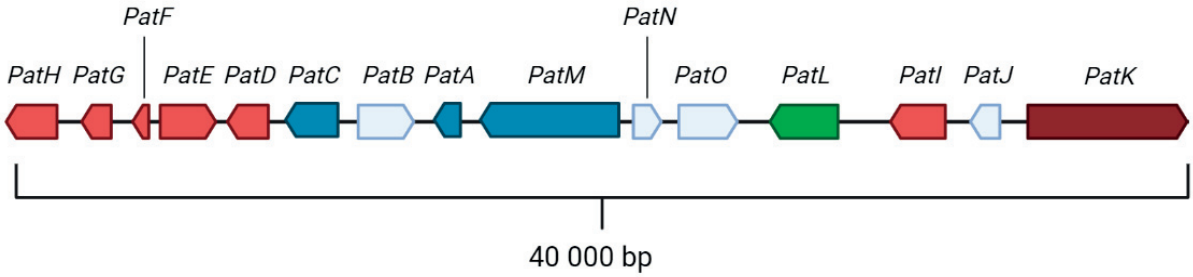


Figure 2: Biosynthetic gene cluster responsible for genes related to production of the mycotoxin patulin in the filamentous ascomycete *P. expansum*. The different colours correspond to different functionalities. Dark brown is the core enzyme, light red is additional biosynthetic enzymes, blue are transport-related and green is regulatory in function. Pale genes are undetermined (Li et al. 2019).

Evolution of Biosynthetic Gene Clusters

The production of a secondary metabolite with its associated BGC is most often relegated to a few closely related species, but there are exceptions with widely distributed BGCs in many different taxonomic groups (Rokas et al., 2020). There is also a consensus that BGCs are evolving at a rapid rate. An ongoing combination of gene duplication and horizontal gene transfer and how they interact with the BGCs is thought to be a cause of their versatility and their change (Wisecaver, Slot and Rokas, 2014). New secondary metabolites may then be favoured by natural selection - increasing fitness and evolution. Additionally, *de novo* assembly of BGCs is being researched as a possibly prominent way of new BGC formation.

Duplication or addition of genes within a BGC can result in a long-term change in the secondary metabolite being produced by the cluster. A noteworthy example is variation of mycotoxin production in the genus *Aspergillus*, where e.g. *A. flavus* produces aflatoxin and *A. nidulans* produces sterigmatocystin. The BGCs and toxins themselves are quite similar, but the addition of three new functional genes (Figure 3) to the aflatoxin BGC results in the different final product (Rokas et al., 2020).

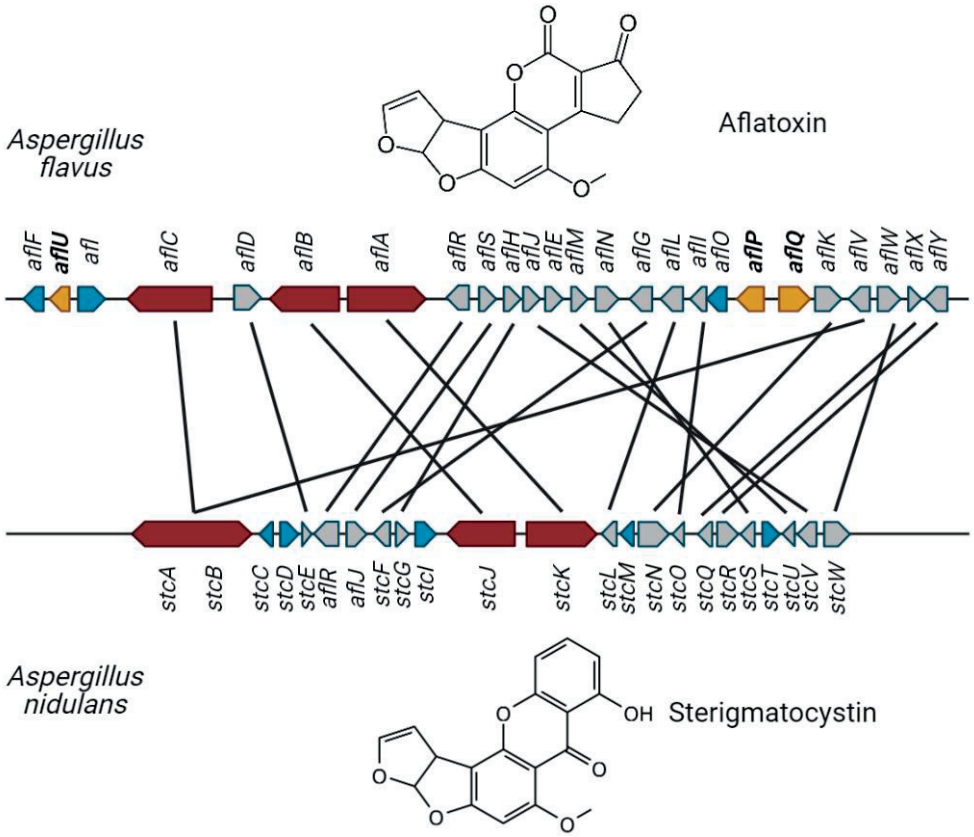


Figure 3: The two biosynthetic gene clusters responsible for genes related to production of the mycotoxins aflatoxin (above) and sterigmatocystin (below) in *A. flavus* respectively *A. nidulans*. Dark brown are the core enzymes.Grey are genes with orthologs in the other clusters, shown by lines. Blue are unique to each cluster. The genes *aflU, aflP* and *aflQ* are in yellow, and are the ones whose enzymes which have been proven to enable conversion from sterigmatocystin to aflatoxin in *A. nidulans*. Adapted from (Rokas et al., 2020).

Entire clusters can also be duplicated, but less research is available. An example of such a case is the origin of the patulin and yanuthone BGCs - with both being present in some members of the *Penicillium* genus. Several genes are shared, among them the 6-methylsalicylic acid core synthase (Nielsen et al., 2017). It is thus suggested that a duplication event occurred very long ago. Together with several other losses or recruitments of genes, this eventually resulted in the two clusters.

Horizontal gene transfer implies transfer of genetic material in a way that is not through reproduction - which can be seen as "vertical". It is a driving cause of evolution for many microbial species capable of it, and is very likely the answer to how some BGCs are spread in a disjointed manner. An interesting case of this occurring even in larger mushroom-forming fungi is the spread of the BGC responsible for producing the hallucinogenic compound psilocybin. A study by Reynolds et. al (2018) compares hallucinogenic and non-hallucinogenic fungi and concludes that the presence of the cluster was due to selection for similar niches, in this case dung and late-wood decay. Other documented examples include the transfer of the cluster for the pigment bikaverin between the distantly related filamentous ascomycete genera *Fusarium* and *Botrytis*, which resulted in a "silent" cluster - a BGC of which the regulation is missing or inactivated, meaning the compound is not produced (Campbell, Rokas and Slot, 2012). It should be noted that tracking horizontal gene transfers that go far back may be difficult, as the clusters are generally diverging as time goes on.

A method of BGC origination that should be mentioned in the context of BGC evolution is *De novo* BGC assembly. It is the process in which a biosynthetic gene cluster for a novel metabolite is constructed by itself in an organism. This grouping of genes likely originates from internal gene duplications from both primary and secondary metabolism combined with modules (independant smaller clusters within a BGC) and genes received through horizontal gene transfer. It is a possible explanation for clusters which are very hard to relate to existing clusters, either in the species itself or related ones (Rokas et al., 2020).

The acquisition of a new BGC or the disabling of one through loss of a crucial gene has some interesting implications. Recent research, spurred on by improving bioinformatic algorithms for finding BGCs has shown that there exists many so-called "orphan" or silent gene clusters. These are BGCs where the end product is either not known or not expressed under the conditions where the organism was cultivated (Hertweck, 2009). These are perplexing from a natural selection viewpoint when observed at first, but it seems likely that many of them have occurred through e.g., horizontal gene transfer events but where the associated transcription factor was lost or nonfunctional. However, many of them are also thought to function normally, but are likely only activated under conditions not found in a laboratory setting (Bok et al., 2009). This puts further emphasis on the value of studying these organisms and their metabolism from an ecological viewpoint, as many of these silent clusters are likely activated during interactions with their specific environmental niche and its inhabitants.

As can be understood from the rate of horizontal gene transfers and duplication events, a major advantage of BGCs seems to be that they are prone to constantly change and evolve, leading to better fitness for the host species. Among the other proposed advantages to clustering exists include the combating of accumulation of toxic intermediates in biosynthetic pathways (McGary, Slot and Rokas, 2013). A notable number of metabolic disorders occur due to the loss enzyme function that breaks down intermediates that may cause harm if a high concentration is reached. Therefore, it seems possible that the simultaneous expression and close positioning of the genes would result in lower accumulation and less likelihood of loss of just one gene (Pál and Hurst, 2003).

Another large notable hypothesized cause for the existence of BGCs is that they are enabling co-expression due to their spatial proximity. Prokaryotic organisms have functionally related genes condensed in operons, which are many genes under the control of a single promoter. It allows for very easy adjustment of expression

levels for the whole pathway simultaneously. Moreover, it is also discussed and hypothesized that the physical linkage of genes is very important even in eukaryotes - which would reasonably include BGCs (Sproul, Gilbert and Bickmore, 2005).

It is known that many BGCs activate and downregulate in tandem. Notable examples with primary metabolism include the *GAL* cluster in *S. cerevisiae,* which is enabled and disabled by galactose and glucose, respectively (Escalante-Chong et al., 2015). Together with transcription being adjusted by shared transcription factors, this is also thought to be heavily affected by chromatin modification. Chromatin structure is related to the folding and accessibility of DNA and this seems like it could be very useful in expressing silent clusters and controlling expression for BGCs as a whole (Bok et al., 2009). Biosynthetic gene clusters seem to be a tool for the organism to better control the production of specific compounds (Macheleidt et al., 2016).

Regulation of Secondary Metabolism

The central dogma is a showcase of how transfer of genetic information occurs in living organisms. It is quite simply stated as: DNA is used to make RNA, which is then used to make proteins which causes the desired effect on the organism - see figure 4 (Koonin, 2012).

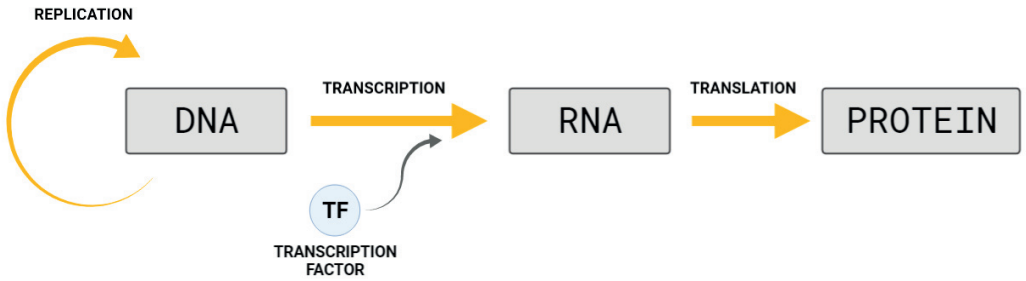


Figure 4: Simple figure of the central dogma. Note that transcription factors can both upregulate or downregulate the flow of information from left to right.

This explanation, while correct, fails to communicate on how in reality information can go the other way around as well. Cases such as reverse transcription, post-translational modification and epigenetics are not covered. However, the central dogma is a very useful starting point when discussing genetics and molecular biology. (Koonin, 2012)

Transcription is the reading and copying of DNA information into RNA using the language of the base pairs - often shortened as G,C,A,T for DNA with T substituted for U in RNA. The process starts with a type of protein called RNA polymerase, which binds to a region upstream of the actual gene called a promoter region. Other proteins bind simultaneously with RNA polymerase, some of which are transcription factors. The transcription then starts at a transcription start site, while the coding sequence is marked by start and stop codons. The coding sequence is what ends up being translated into RNA, and is seen in figure 5 below. The finished RNA is then further processed in the cell in the process known as translation to eventually result in a functional protein which can impact the cell and its surroundings.

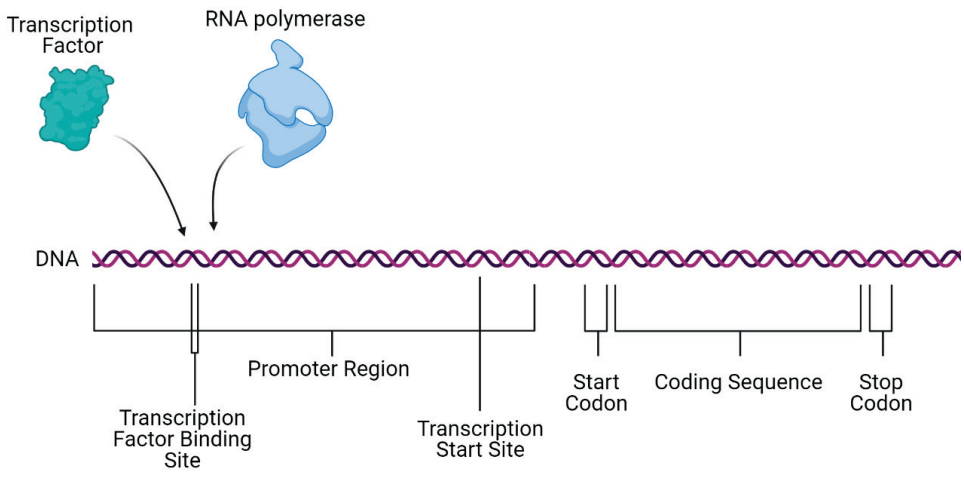


Figure 5: Schematic figure of transcription in fungi with illustrative sequences on the target DNA.

Transcriptional regulation is the process in which a cell controls the rate of transcription. This in turn influences the amount of synthesized proteins, which affects the behaviour of the cell and allows it to react to outside circumstances. This is a fundamental fact on which all life relies (Mitsis et al., 2020). The most prominent way this is done is by the interaction of the aforementioned transcription factors (TF). A transcription factor is a protein which binds to an upstream activation sequence of DNA during the transcription process and affects transcription of the gene by either inducing or repressing recruitment of RNA polymerase to the promoter and DNA. Transcription factors are ubiquitous to all life and thus there are a multitude of classes and groupings. Transcription factors are generally very narrow in their regulation due to the inherent specificity of protein folding and how they can interact with the nucleotides of DNA. (Mitsis et al., 2020).

However, very relevant to fungi due to the clustering of genes involved in secondary metabolism, there is also regulation through epigenetic regulation or chromatin modification (Macheleidt et al., 2016). Chromatin is a complex of folded genomic DNA wound around proteins, making up the chromosomal DNA of eukaryotes (Strauss and Reyes-Dominguez, 2011). This is due to size restraints, so the DNA has to take up as small a space as possible in the nucleus. Histones consist of nucleosomes, which is a unit defined as double-stranded DNA wrapped around an octamer of basic proteins called histones. This very high level of compartment means that physical limitations stop RNA polymerase from transcription per default and is known as heterochromatin. However, histones and the condensed structure are highly dynamic and adjustable. This is done by chromatin remodeling with ATP as well as post-translational modifications to the histones using specific enzymes. These enzymes are not categorized as TFs due to their inability to bind to DNA but can still regulate transcription due to affecting the state of the chromatin. The opened structure - called euchromatin - then allows access for e.g. RNA polymerase to transcribe (Strauss and Reyes-Dominguez, 2011). The resolution of chromatin modifications is of such a size that a change at one point in the structure has effects on the folding in the vicinity as well. This is likely even if a single gene is being expressed, in a cluster through e.g. a transcription factor binding, the euchromatin structure results in the increased transcription of nearby genes as well. The occurrence of BGCs in fungi is likely favoured by this phenomenon, as it likely reduces the odds of wasteful partial pathway expression and again, reduces the accumulation of toxic intermediates (Nützmann, Scazzocchio and Osbourn, 2018).

Notable enzyme examples which work through epigenetic means in fungi are the global regulators (also called global transcription factors) (Brakhage, 2012). Global regulators are mostly activated or downregulated through various environmental cues and modulate the expression of many pathways and metabolites simultaneously unlike local transcription factors which mostly only affect their own BGC.

The most important of these is the group of proteins called the Velvet Complex. The complex consists primarily of the proteins VeA, VelB and LaeA, where LaeA is a known methyltransferase. The complex controls the localization of the LaeA as a function of light exposure. In low-light situations, the complex is located in the nucleus and can methylate histones for epigenetic activity. If exposed to light, VeA is retained in the cytoplasm due to interactions with its nuclear localization sequence, which inactivates the complex (Bayram et al., 2008). It has been shown in several studies that LaeA inactivation and overexpression results in corresponding changes of expression in several BGCs (Bok and Keller, 2004)(Perrin et al., 2007).

Global Regulators

It is beneficial for metabolism to be controllable as a whole, as the occurrence of different stages of the cell cycle and other scenarios means that many times the expression of unrelated genes is a large waste of energy and resources. Secondary metabolism is often associated with a different stage of development or morphological change for an organism, occurring as the focus is switched from biomass production (Ruiz et al., 2010). These are logical and planned adjustments of metabolism. There are also other types of global regulators, which are expressed in reaction to a change in the environment.

These enzymes work independently or as a part of a complex and have an effect on a wide array of metabolic pathways - both primary and secondary. A handful have been categorized and investigated in filamentous ascomycetes and have been outlined below in table 1. The presence of global regulators have interesting implications on how and why secondary metabolites are synthesized.

Table 1: Overview of prominent global regulators known to be present in filamentous ascomycetes. Global regulators are enzymes which work alone or in a complex and can alter metabolism in a wide range within the organism as a response to stimuli.

Name	Stimuli	Type	Example:	Additional notes:	References
CreA	Carbon Source	2 Cys2His2 zinc fingers	Repression of penicillin production in *P. chrysogenum* based on glucose or sucrose as a carbon source.	Binding site: SYGGRG	(Cepeda-García et al., 2014) (Cubero and Scazzocchio, 1994)
AreA and AreB	Nitrogen level	Cys2Cys2-type zinc finger transcriptio n factor	De-represses genes involved in utilization of secondary nitrogen sources in absence of glutamine and ammonium.	GATA-type TF, AreA generally upregulator while AreB downregulates	(Tudzynski, 2014)
PacC	pH	3 Cys2His2 zinc fingers	Upregulated in alkaline situations. Deletion leads to lower patulin, especially when in alkaline situations.	5'-GCCARG-3' binding site,	(Chen et al., 2018)
LaeA (Velvet Complex)	Light exposure	Histone methyltrans ferase	In *P. expansum,* it regulates all 15 genes in the patulin cluster. LaeA together with VeA and VelB form the Velvet complex.	Epigenetic regulator	(Li et al., 2020), (El Hajj Assaf et al., 2020)
Skn7	Osmotic and Oxidative Stress	HSF/Stress TF	In *A. flavus* - deletion resulted in drastic decrease in aflatoxin.	Can work in tandem with Yap1. Both prominent in all eukaryotes but less documentation in fungi.	(Zhang et al., 2016 ,(He and Fassler, 2005)
Yap1	Oxidative Stress	bZIP	Ochratoxin A accumulation increases in *A. ochraceaus* by repressing yap1 ortholog.		(Reverberi et al., 2008) ,(He and Fassler, 2005)
CBC-complex and SreA	Iron levels	Heterotrim eric core complex + Cys2Cys2Z FTF	The complex in *Aspergillus* has an effect on penicillin production. Upregulates ipnA, aatA but downregulates acvA.	Binds to pentameric CCAAT box in promoter regions, SreA to GATA	(Furukawa et al., 2020) (Hortscansky et al, 2017)

Of all of the global regulators mentioned above, which are mainly outlined in a prominent review paper by Brakhage (2012), only the CBC-complex (consisting of subunits HapX, HapB, HapC and HapE) relating to iron has not been documented in penicillium, so publications based on the complex in *Aspergillus* are used (Furukawa et al., 2020). Also worth mentioning is that the exact function and specificity of these global regulators may vary depending on the exact organism. An example of this includes the functionality of Yap1p homologues in filamentous ascomycetes compared to each other and the model organism yeast. While the role the transcription factor has in antioxidant response is well conserved, the specifics are different. A yap1-homologue knock-out mutant of *A. fumigatus* is susceptible to H_2O_2, while in *A. parasiticus* the mutant was still resilient to H_2O_2 (Mendoza-Martínez, Cano-Domínguez and Aguirre, 2020)

Many of the global regulators are in play depending on what resources are available. CreA is related to carbon source regulation - when prime sources such as glucose are available, the metabolism of these may take precedence over other processes in the organism. This results in a reduction of transcription of genes related to carbon metabolism - for example penicillin in many filamentous ascomycetes. This is a notable problem when optimizing antibiotic production in the industry (Sánchez et al., 2010). Global regulators may also regulate transcription in cases where the synthesized protein would not function and be wasted (Peñalva, Tilburn, Bignell and Arst, 2008). PaCC, which is the fungal global regulator related to pH, works in a signal transmission network. It is an activator of genes expressed in alkaline conditions and a repressor of genes expressed in acidic conditions (Peñalva, Tilburn, Bignell and Arst, 2008).

Additionally, global regulators commonly interact and are known to regulate each other in instances, which adds on another layer of complexity in certain cases. LaeA is a known positive global regulator of secondary metabolism and is suspected to be modulated itself by CreA. This is based on interesting study results regarding knockout mutants and the presence of CreA binding sites in the promoter region of LaeA (Li et al, 2020).

Knowledge and continued study of the global regulators pose an interesting opportunity to basal understanding of how and why secondary metabolism works the way it does. It is also important to have the global regulatory framework in mind when analysing individual biosynthetic gene clusters and their related secondary metabolites. It lends valuable insight in how external conditions may affect the microbes behaviour.

Global regulators are thus largely a tool for the fungi to optimize metabolism as a whole. The decreased production of penicillin related to carbon source is thus not "filamentous ascomycete can use carbon sources as signalling substances to produce or not produce antibiotics" but "filamentous ascomycete can shift metabolism to quickly utilize a limited resource in order to achieve better growth". Specific reactions are instead mostly driven by transcription factors that are smaller in scope.

Local Regulators

Local transcription factors are many times not separate to global regulators in structure or function, but in the scope of their influence. Up to 50% of all BGCs are estimated to have an encoded transcription factor which regulates itself and the cluster (Li et al., 2015). These transcription factors then bind to binding motifs which are present in all or just a few genes in the cluster (Kong et al., 2020). This then results in up- or downregulation of the cluster, possibly through assistance by epigenetic or pathway-specific means. A very prominent and well-characterized example is the BGCs for production of aflatoxin and its related regulatory genes in some *Aspergillus* species. The cluster is large and complex - reflecting the extensive synthetic pathway of Aflatoxins. In *A. flavus,* of the thirty genes it includes, seventeen are known to have the specific binding sequence the regulatory gene *AflR* is capable of binding to (Caceres et al., 2020). It is a positive regulator, as over-expression results in increased production of aflatoxin (Flaherty and Payne, 1997). *AflR,* which encodes for a binuclear zinc transcription factor, also has an interesting interaction with another gene in the BGC called *aflS*. Aflatoxin biosynthesis is additionally known to be heavily affected by global regulators, which serves as a useful guideline when investigating patulin.

Investigating the metabolic regulation of BGCs which lack a local transcription factor is somewhat more difficult, but these are currently thought to be controlled exclusively by global regulators and occasionally from cluster cross-talk, where local transcription factors may affect other clusters than it's own (Brakhage, 2012).

These local transcription factors, which are commonly denoted cluster-associated or pathway-specific can be divided into several protein families within fungi. Notable families include factors in fungi include basic leucine zippers or bZIPs, winged helix regulators and zinc-binding transcription factors.

Zinc-Binding Transcription Factors are the most prominent and abundant ones in fungi. These contain so-called zinc fingers motifs within their sequences which use cysteines and histidines to control zinc atoms. This results in unique peptide structures for which can be very specific. These can be further divided into three classes, in which the first two are shared between all eukaryotes and the third being unique to fungi. The three classes differ in which amino acids are coordinating the zinc atoms, which has effects on the sequence recognition (García-Estrada, Domínguez-Santos, Kosalková and Martín, 2018).

Class 1 has a signature sequence consisting of two cysteine and two histidines, and are therefore often called Cys_2His_2 or C_2H_2-type transcription factors. They are present throughout all eukaryotes and are often the largest class in many organisms - including humans. They are also prominent within fungi and the global regulators CreA and PaCC are of this type (García-Estrada, Domínguez-Santos, Kosalková and Martín, 2018). Recent examples of Cys_2His_2-TFs in local transcription include BcabaR1 which regulates abscisic acid production in the filamentous fungus *Botrytis cinerea* (Wang et al., 2018).

Class 2 are called Cys_4, C_4 or sometimes Cys_2Cys_2-transcription factors. This is due to their signature domain consisting of the four cysteines. GATA factors are a whole group of Cys_4 TFs known for their ability to bind to the nucleotide sequence GATA in promoters. The global regulator AreA and AreB are GATA-type transcription factors and are associated with nitrogen regulation, as outlined above in Table 1 (García-Estrada, Domínguez-Santos, Kosalková and Martín, 2018).

Class 3 are known as binuclear zinc transcription factors and are the largest family of fungal-specific TFs. Approximately 90% of the encoded local transcription factors of polyketide synthase BGCs are estimated to adhere to this group. Many of the interesting compounds synthesized by the *Penicillium* genus are polyketides, including the chosen model compound patulin.

Binuclear Zinc Transcription Factors

Binuclear zinc transcription factors (BZTFs) or Zn_2Cys_6 transcription factors, contain zinc finger domains consisting of six cysteines flanked with amino acids divided up into two sets of three that coordinate the zinc atoms. They are also occasionally known as Gal4-type transcription factors. Gal4 is a regulatory protein which is part of galactose metabolism in *S. cerevisiae a*nd is seen as a model protein for the class (García-Estrada, Domínguez-Santos, Kosalková and Martín, 2018). The DNA binding domain is most often located in the amino terminal end of the protein and the zinc finger has a distinct sequence structure. It is well conserved and usually written as Cys-X2-CysX6-Cys-X5-12-Cys-X2-Cys-X6-9-Cys (figure 6A), albeit certain prominent BZTFs such as AlcR has a middle spacer region which is notably longer (MacPherson, Larochelle and Turcotte, 2006). However, the two degenerate amino acids separating the adjacent cysteines are perfectly conserved in all BZTFs, and thus the entire sequence is quite easily identifiable using motif search tools (Schjerling and Holmberg, 1996). The DNA-binding domain as a whole is characterized as tripartite, being built up by the signal sequence Cys_6, a linker domain and a dimerization domain. The rest of the protein typically consists of a middle homology domain with regulatory properties and an acidic carboxy terminal domain, illustrated in figure 6B (MacPherson, Larochelle and Turcotte, 2006).

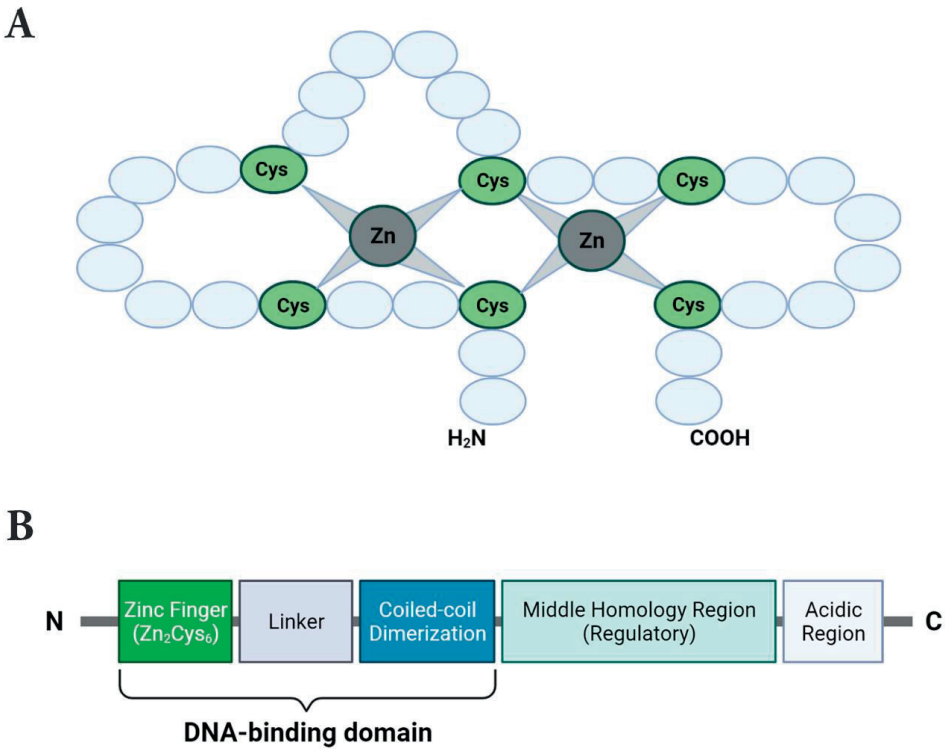


Figure 6: **A.** Primary structure of the zinc finger domain in a typical BZTF. The binuclear term stems from the two part division the zinc finger has from the two zinc atoms. Adapted from (García-Estrada, Domínguez-Santos, Kosalková and Martín, 2018). **B.** Schematic representation of the functional domains of a typical BZTF. The DNA-binding domain is divided into the three subregions and is the most researched. Adapted from (MacPherson, Larochelle and Turcotte, 2006)

As a more researched member of the class, GAL4 has had its crystal structure elucidated, unlike the vast majority of characterized BZTFs. Crystallisation is the process of stabilising a normal protein through a series of advanced and complex steps. Once stabilised, the protein crystal structure can be studied to predict how the protein interacts on a molecular level. Most of these notable crystallisation studies are from the late 1990s and were focused on BZTFs in yeast - either the aforementioned GAL4 or the protein PPR1, which is involved in pyrimidine regulation. These early publications reached several conclusions, among them the fact that BZTFs are homodimers, which means that two copies of the same transcription factor interact with the protein at the same time for regulation to occur, see Figure 7 (Marmorstein and Harrison, 1994). However, contemporary and later studies found that some BZTFs bind as monomers (Schjerling and Holmberg, 1996).

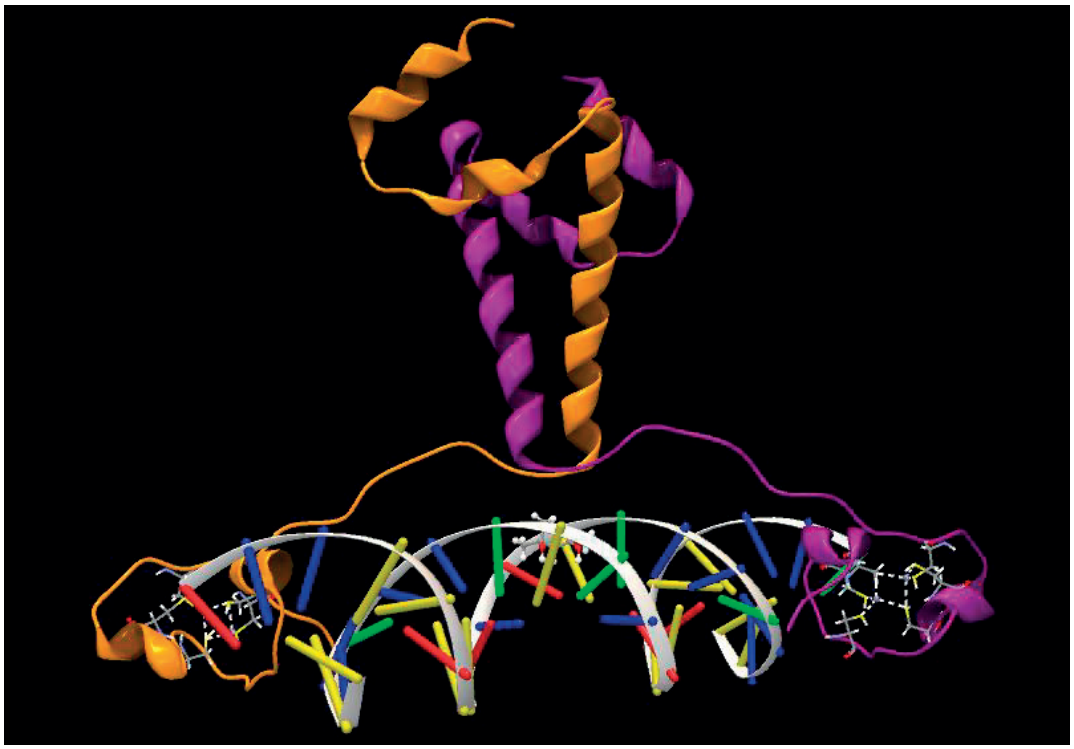


Figure 7: 3D Structure of dimeric binding of two GAL4 transcription factors to target DNA strand. GAL4 DNA-binding domains are coloured purple and orange. Coiled-coil structure at the top is the dimerization domain, the zinc finger domain with Zn_2Cys_6 is located at the bottom right and left. The linker domain is situated between these. GAL4 sequence from (Hong et al., 2008).

One early scientific consensus about BZTFs was that they recognize pairs of triplets consisting of the base pairs CGG, either in inverted, everted or direct repeat, e.g. CGG-n-GGC, CGG-n-CCG or CGG-n-CGG. 'These triplets have a TF-specific spacer domain between them. E.g, GAL4 is known to bind to the palindromic binding site 5'-CGG-n(11)-CCG and PPR1 to 5'-CGG-n(6)-CCG (Marmorstein and Harrison, 1994), where the n(x) indicates a number of base pairs whose type in less important. The spacing between the two triplets is known to vary widely depending on the exact BZTF, and seems to be dependent on the linker domains base pairs and length, which would impact conformation and electrostatic interactions with the backbone (Marmorstein and Harrison, 1994). Research has shown that while specific BZTFs may in cases bind to slightly longer or shorter spacer regions, this often comes at a heavy cost to binding affinity (Vashee et. al., 1993) It seems reasonable that this palindromic binding site stems from the homodimeric binding pattern, with each

copy of the protein binding one trinucleotide, especially if crystal structures are investigated such as from Marmorstein and Harrison (1994).

The dimerization domain itself consists of a protein structure interaction called a coiled coil. A coiled coil domain occurs when two or more α-helices wrap around each other in a supercoiled bundle (Ludwiczak et al., 2019). It is enabled by an extended α-helix structure in BZTFs, which can be seen in Gal4 in Figure 7. Since two copies of the same protein are binding, the coiled coil interaction is stabilizing the binding configuration to the target DNA. In Gal4, the dimerization is also partly stabilized by van der Waals interactions between amino acid side chains between the two copies of the protein (Hong et al., 2008). Hong (2008) also experimentally concludes that the dimerization is important for the ability of GAL4 to bind to DNA, albeit does not make any comment or comparison to monomerically binding BZTFs.

Schjerling and Holmberg (1996) discuss the dimerization domain of the BZTFs. They combine motif scanning with prediction of the coiled-coil domain, and conclude that a large majority (60 of 79) identified BZTFs display a coiled-coil domain within 150 bps after the Zn_2Cys_6-zinc finger and thus likely dimerize. 56 of these 79 BZTFs were from *S. cerevisiae*. Again, the consensus about the Zn_2Cys_6-zinc finger is that it is usually situated in the amino- or N-terminal of the protein, with exceptions. Ume6p - for example, is known to harbour the domain at the C-terminal instead (MacPherson, Larochelle and Turcotte, 2006). However, BZTFs have been documented where the domain is located centrally in the protein. The monomerically binding Pho7, has the domain start at amino acid 279 out 738, a stark contrast to e.g. GAL4, whose domain starts at 11 out of a total 881 amino acids (Garg et. al 2018).

Research on transcription factor binding sites indicates that many transcription factors in organisms may work bidirectionally. This means that if located between two genes located on the sense and antisense strands, the factor may modulate transcription of both of them. This seems to be true even in cases where the binding sequence is non-palindromic for BZTFs (Punt et al., 1995) (Li et al., 2020) and possibly for different classes of zinc finger transcription factors as well (Cepeda-García et al., 2014). While not true for all organisms - human transcription sites are known to be unidirectional - it could be assumed for fungi (Lis and Walther, 2016).

There somehow seems to be an predisposed consensus on the exact characteristics of the BZTFs, and there are multiple publications who refer to one of the early yeast research publications on BZTF binding sites and propose a putative binding sites *in silico* without properly gauging the state of the field (Guzmán-Chávez et al., 2017) (Fox, Gardiner, Keller and Howlett, 2008). A multitude of binding sites for BZTFs have been experimentally verified even in filamentous ascomycetes (Appendix A) and a notable portion do not follow the palindromic pattern (MacIsaac et al., 2006).

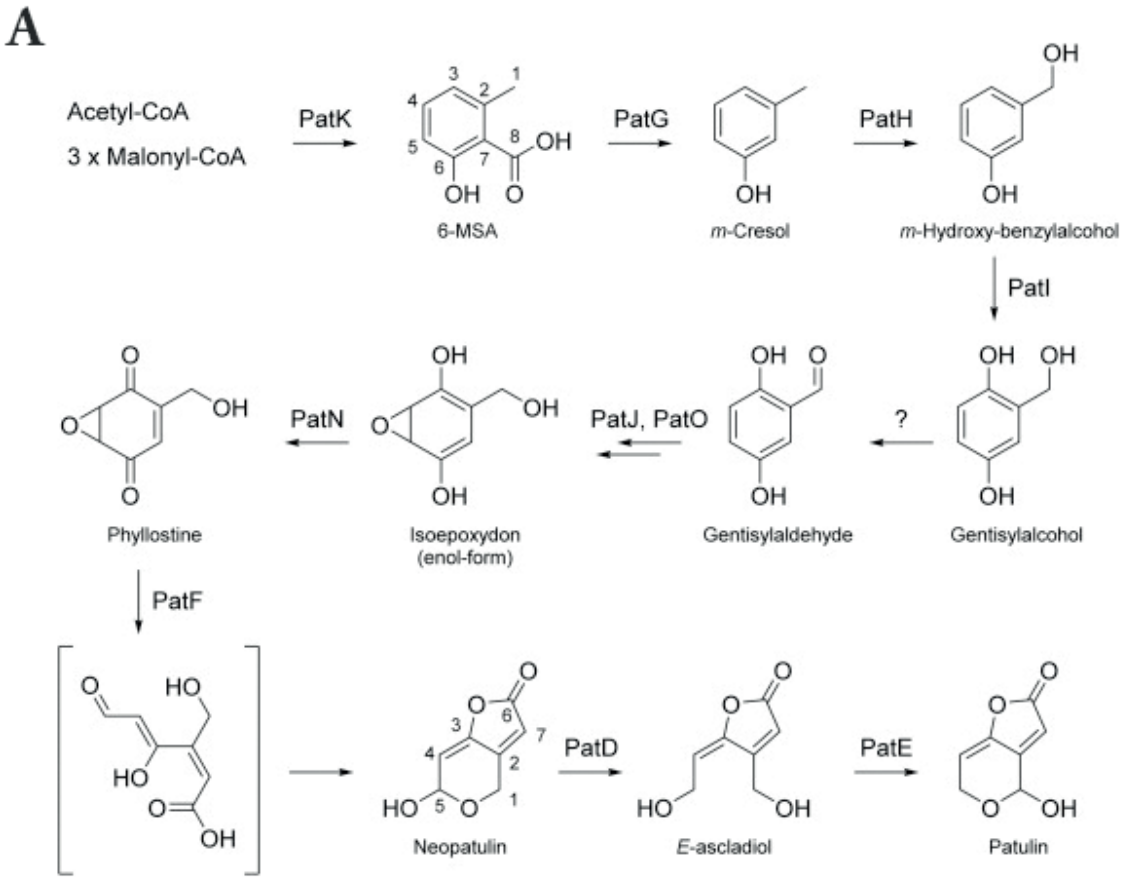
The preconception could partly be due to the higher focus on yeast in the field of secondary metabolism in fungi. There have been large reviews describing and compilating identified binding sites in yeast (MacIsaac et al., 2006), but none in filamentous ascomycetes. It is hard to know if results translate well to the different phylogenetic groups. BZTFs have been known in cases to vary in characteristics when comparing even closer taxonomic relatives. AflR has been characterized in *A. nidulans, A. flavus and A. parasiticus* and publications have reported three different binding sites, suggesting impactful structural differences (Kong et al., 2020).

Patulin

The mycotoxin patulin is a notorious secondary metabolite produced by a handful of filamentous ascomycete species (Frisvad, 2018). As mentioned previously, patulin poses a health risk to humans if ingested, and can be present in apples and other produce infected by pathogenic species such as *P. expansum*. This issue is exacerbated by the fact that the apples may contain notable concentrations of patulin even if the apple is visibly unharmed (Errampalli, 2014). There is currently extensive ongoing research about how to detect the infection in early stages to prevent the mycotoxin, and as such it is of great interest to find out how it functions and is regulated.

Patulin is not a virulence factor for the mold, but may have other related effects when infecting apples (Snini et al., 2016). Contrary to its characterization as a toxin, it originally saw use as an antibiotic and as treatment for the common cold, but this was discontinued after adverse effects (Clinical trial of patulin in the common cold, 2004). Therefore the role of patulin in the mold itself may be one of microbial warfare, as suggested by a few publications (Moake, Padilla-Zakour and Worobo, 2005).

The patulin biosynthetic pathway consists of approximately 10 steps, albeit the exact characteristic of each step has not been elucidated. The pathway as described by Li et al. (2019) is pictured below in Figure 8 A. The 15-gene cluster in *P. expansum* is shown below in Figure 8 B (Li et al., 2015).



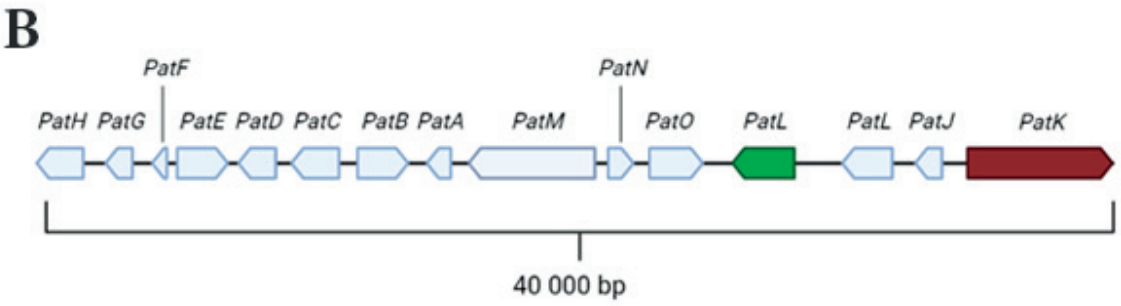


Figure 8: **A.** Biosynthetic pathway for patulin. Note that only eight of the genes from the patulin cluster have been attributed to steps in the pathway. From (Frisvad, Isbrandt and Larsen, 2020), which was in turn adapted from (Li et al., 2019). **B.** Simple schematic figure of the patulin biosynthetic gene cluster in *P. expansum*. Adapted from (Li et al., 2015). *patL* in green codes for a BZTF and *patK* for a polyketide synthase.

Of the 15 genes in the cluster, three are transport-related, 11 are biosynthetic in nature (including the core enzyme - the polyketide synthase *patK*) and one is denoted as a transcription factor (Li et al., 2019). The latter is marked in green in Figure 8 B as the putative BZTF patL. patL contains the characteristic Zn_2Cys_6 zinc finger associated with BZTFs and is localized in the nucleus. It also has been shown experimentally to be involved in the regulatory process of patulin due to absence of patulin production in Δ*patL* strains (Li et al., 2015). *patL* is also shown in Zong, Li and Tian (2015) to exhibit higher expression during environmental factors which favour the production of patulin such as acidic conditions, reinforcing the role of patL as a positive regulator. The presence of the typical transcription factor and the economic interest in researching the secondary metabolite makes it a useful model BGC for studies.

Several of the global regulators mentioned in Table 1 have been shown to be involved in the transcription of the cluster - LaeA, PaCC and CreA, influenced by light exposure, pH and carbon source, respectively. Other conditions have also been reported to affect the production such as nitrogen source and ROS exposure, meaning more of the global regulators seem likely to play a part somehow as well (Li et al, 2020). Zong, Li and Tian (2015) did research on the effect of a wide range of environmental factors on patulin production as well as individual gene expression in *P. expansum*. It would be interesting if the findings could be partly confirmed and validated through the identification of global regulator binding sites. Perhaps it could work the other way around as well - if identification of global regulator binding sites could serve as a predictive factor for environmental impact on the production of a secondary metabolite.

The binding site for patL has not been identified (Li et al, 2020). Its essential role in terms of the patulin production and the expression of the cluster means it should have binding sites in the promoter regions of most of the genes, if not all of the genes. Since it is a BZTF, it is a possibility that the binding site is of the characteristic palindromic type and should thus be identifiable using different motif discovery tools. If a putative binding site is identified, it is of high importance to verify it experimentally. Both of these steps, however, are not simple.

Motif Discovery and Confirmation

Motifs are defined as short patterns of nucleotides that have biological meaning (Zambelli, Pesole and Pavesi, 2012). Often this refers to transcription factor binding sites. But identification and confirmation of these sites is quite difficult, of which there are several reasons, including the fact that they are usually very short as well as degenerate, which means that the exact composition of the motif may vary due to the fact that the nucleotides therein may not be equally important (Zambelli, Pesole and Pavesi, 2012). Even though the promoter regions are generally quite short (ca. 1000 bp) in relation to a whole genome, the exact position of the binding site is also quite unspecific.

In order to confirm a binding site using various experimental means, it generally has to be identified first. This is called motif discovery. By using silico methods and programs such as RSAT or MEME (Bailey et al., 2009), statistics can be used to determine putative binding sites. RSAT uses a statistical model, which searches for motifs that are overrepresented compared to chance in the genome. It is a valid technique based on the size of the dataset and that overrepresentation implies meaning - it has been selected for again and again by natural selection. To further improve the quality of the candidates, an alignment with other genomes can be done to construct a consensus sequence (Zambelli, Pesole and Pavesi, 2012). This is again based on the fact that if the motif has a high level of conservation across even distant taxonomic relatives, it is likely due to the fact that point mutations or other changes there would result in fitness loss or death - meaning it essential.

After a putative binding site has been identified it has to be experimentally confirmed, which is nontrivial and tedious. There are several methods for confirming that the transcription factor does bind to the binding site. Some of the more notable ones include Electrophoretic mobility shift assays, DNase I footprinting analysis, Chromatin Immunoprecipitation ChiP-sequencing.

Electrophoretic mobility shift assay

Electrophoretic mobility shift assays (EMSA) is a technique for identifying DNA-protein bindings that relies on electrophoretic separation. In simple terms, if an electric current is placed over a specifically prepared agarose gel, molecules will move through the gel based on their size, shape and charge (Hellman and Fried, 2007). These characteristics are influenced when a DNA-protein complex forms.

The target DNA is generally amplified into desired fragments by PCR which are suspected to contain an interaction site. In the case of transcription factor binding sites, these would often be specific parts of the promoter regions of genes which the TF is known to regulate. By labeling the DNA with measurable markers (often fluorescent), the movement of the molecules through the gel can be tracked.

As can be seen below in Figure 9, results can be achieved by comparing to a control that is made by having one sample well consisting of only the free DNA fragment and no transcription factor. The samples that have achieved less movement are the ones containing DNA fragments that are very likely to have a binding site contained within them.

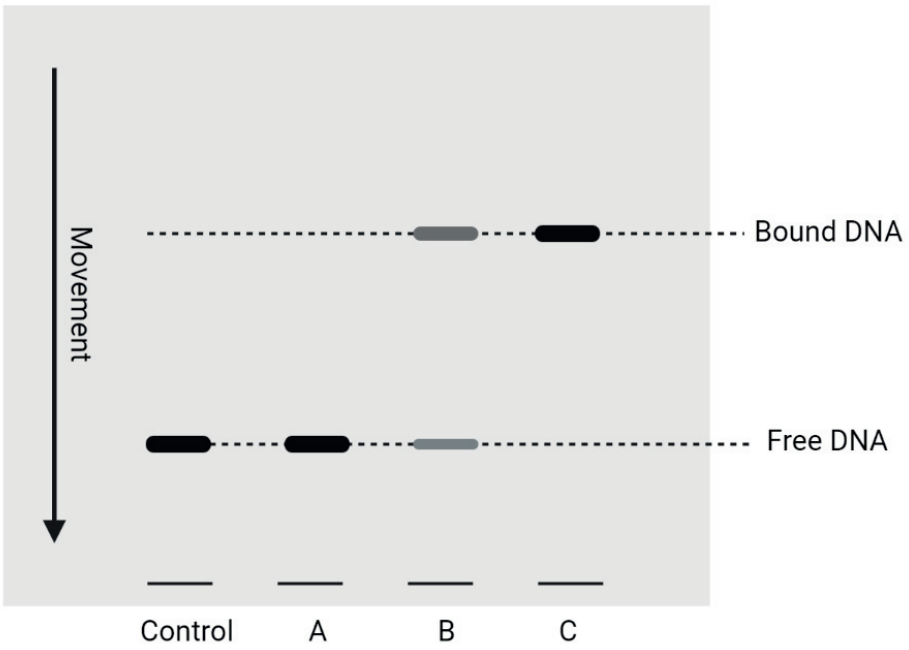


Figure 9: Example of electrophoretic mobility shift assay result. Control is a sample consisting only of the target DNA sample with no TF present. Sample A consists of DNA fragment A plus the TF, as the results are the same as the control sample, no interaction has occurred. Sample B consists of DNA fragment B plus the TF. The band at bound DNA indicates that there is interaction between the TF and the DNA fragment B. However, the still noticeable band at "free DNA" indicates that not all fragments interacted. This can be due to weaker interactions or stoichiometry. Sample C consists of DNA fragment C plus the TF. The prominent and sole band at "bound DNA" indicates that there are powerful and balanced interactions between the DNA and the TF. These results would indicate that a binding site is present in both fragment B and C. Note that all three fragments here are of the same size and charge for illustrative purposes.

EMSA is quite popular due to several reasons. Marking techniques can be varied to achieve greater sensitivity, allowing results even in poor test situations. It is also reasonably simple to perform and has a wide range in terms of what nucleic acid sizes and concentrations can be used. Stoichiometric relations can also be used to help identify proteins that bind as multimers (Hellman and Fried, 2007).

EMSA has seen much use in terms of binding site identifications for BZTFs in filamentous ascomycetes (Baba et. al, 2009)(Liu et al., 2018) (Punt et al., 1995). Multiple adaptations are also in use to specialise EMSA for unique cases, many of which are focused on removing the need for radioactive markers and expensive equipment (Song, Zhang and Huang, 2015).

DNase I footprinting analysis

DNase footprinting analysis or assay is a method that can be used to identify a specific protein binding site in a DNA sequence. It utilizes the principle that if a DNA is interacting with a protein, the DNA is protected from sequence-cleaving enzymes, most notably DNase I. This protection can be utilized together with gel electrophoresis to identify the exact binding site (Galas and Schmitz, 1978). DNase I is the most popular enzyme for DNase footprinting assays. It is a large endonuclease which can target double stranded sequences. The large size means that it is very unlikely to cut where the transcription factor binds due to steric hindrance (Ricci and El-Deiry, 2003).

DNase I footprinting uses 100 to 400 base pair regions which are suspected to contain a binding site for the transcription factor or protein. The fragments are then preferably made by PCR amplification, after which they are labeled, traditionally with ^{32}P. This enables analysis of the fragments during electrophoresis. The procedure is then outlined below in Figure 10.

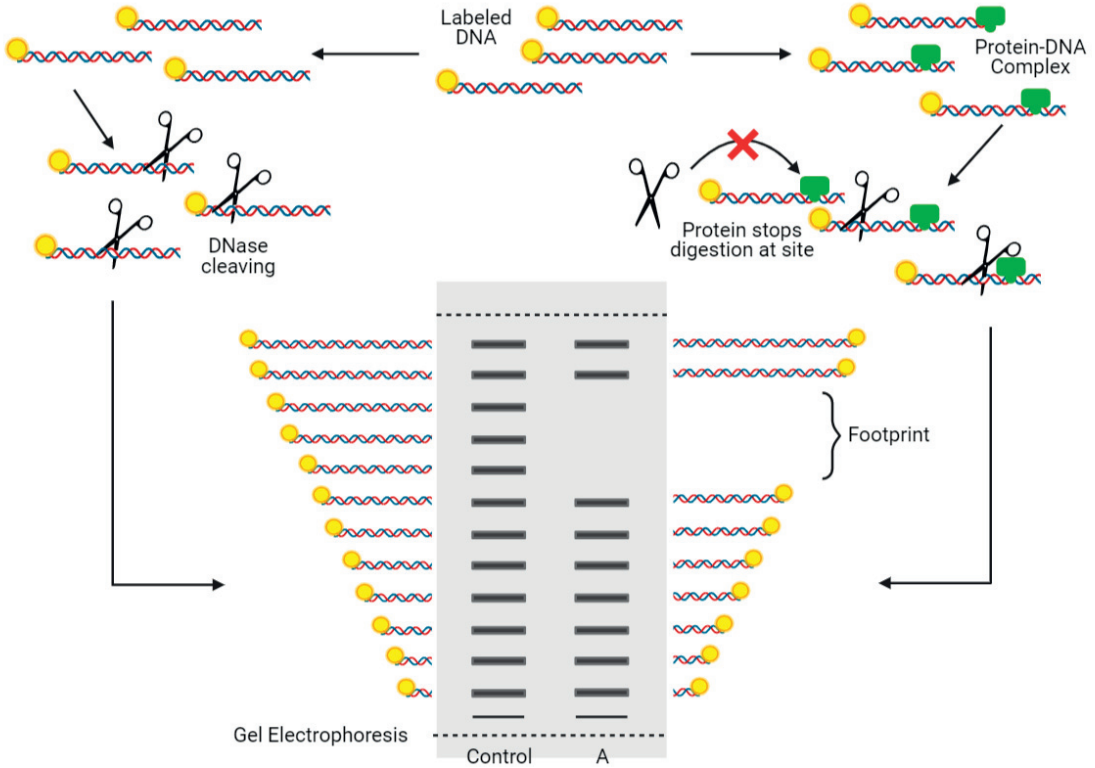


Figure 10: Simplified example of DNase 1 footprinting assay. End-labeled DNA is mixed with the transcription factor which binds to the specific binding site. DNase 1 is then inhibited from cleaving at that specific site. When compared to the control with no added proteins, a clear gap can be seen. Similar to EMSA, this occurs due to the DNA-protein complexes moving more slowly than their unencumbered counterparts through the gel. The control is used as a baseline and the exact region where the protein bound can be elucidated. Adapted from Song, Zhang and Huang, 2015.

DNase 1 footprinting has several drawbacks, mostly related to the higher protein concentrations required compared to EMSA (Song, Zhang and Huang, 2015). However, it is still widely in practice, and many times in combination with EMSA. An example includes (Liu et. al, 2018) where EMSAs were first used to deduce the presence of a binding site in promoter regions of genes related to cellulose metabolism in *Neurospora crassa*. DNase I footprinting was then used as a next step in the process to determine the exact binding site of the transcription factor. As an additional control, they performed EMSAs with site-directed mutation to see if binding was impaired.

EMSA and DNase 1 footprinting are both in-vitro techniques and thus enable greater control and surveillance of the process. However, they are also less true to how it actually occurs within the studied organism. The specificity and characteristics of the binding can be markedly different (Vashee et al, 1993).

Chromatin Immunoprecipitation and related techniques

One of the more advanced methods in use today is ChIP, short for Chromatin Immunoprecipitation. It relies on a combination of binding sites being protected by digestion if they have a bound protein, and that protein-specific antibodies when bound can be selected for through their precipitation. What makes ChIP very useful is that it is performed *in vivo*, meaning a more accurate result may be obtained compared to earlier methods such as EMSA and DNase 1 footprinting (Das, Ramachandran, vanWert and Singal, 2004).

A ChIP assay usually begins with cross-linkage of the chromatin-bounds using formaldehyde, which stabilises the DNA-protein complexes. The cells are then lysed and the DNA is cleaved through nucleases or sonication to approximately 500 bp segments. Antibodies specific for the desired protein are then added, which immunoprecipitate out of the cellular lysate. After heating to denature the complexes, the DNA can then be identified using PCR. For example, one's desired promoter region could be selected for this way - if a positive PCR result is obtained, it means a protein complex formed at the promoter region.

ChIP differs from the other techniques in another notable way, which is the throughput. ChIP allows for immunoprecipitation of *all* DNA sequences in a genome that the desired protein interacts with, not just for a specific gene or promoter sequence. To properly utilize the results from ChIP, it is preferable to combine the results with some sort of high-throughput technique. The most popular way is to utilize modern new-generation sequencing methods such as Illumina. All the precipitated DNA segments are sequenced simultaneously. This enables the construction of a database for that specific protein, which can be used to identify binding sites if compared to the whole genome. A typical workflow for ChIP-Seq, as the combination is often denoted, is outlined below in Figure 11.

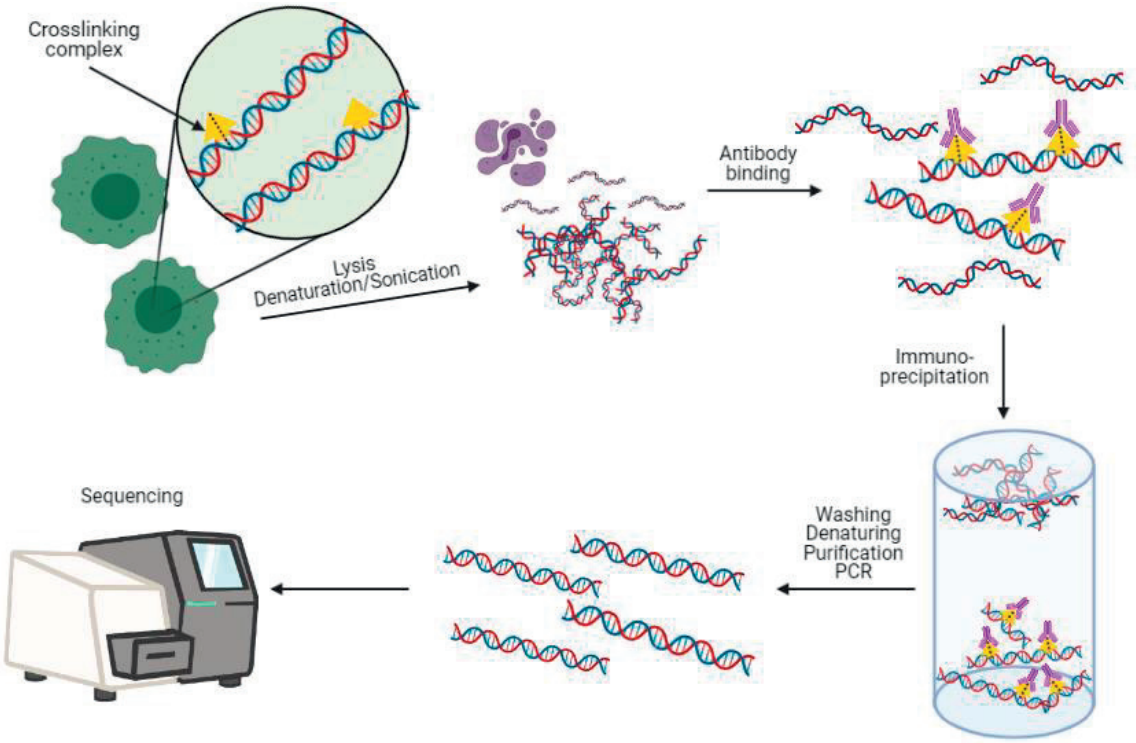


Figure 11: Workflow of an *in-vitro* ChIP-Seq experiment. First the desired protein and chromatin DNA bond in a living cell, which are stabilized through chemical crosslinking (often formaldehyde) The cells are then lysed and the genome is cleaved through nucleases or sonication. Specific antibodies bind to the desired proteins, which are then immunoprecipitated. Proteins, antibodies and crosslinks are removed and purification occurs before sequencing using e.g. Illumina. Adapted from Song, Zhang and Huang, 2015.

While immensely useful, the technique requires extensive preparatory work, notably the production of the specific antibody which can reliably bind to the correct protein. An additional difficulty is that while it can recognize a binding site for a given protein, it does not actually determine if that site serves any function at all. The technique is perhaps not suited as well as the two former to the recognition of a specific site for one BGC, but is very applicable in cases of where long-distance co-expression occurs or for the global regulators (Table 1). Some research groups have gathered hundreds of CHIP-seq results and combined them with motif discovery tools to construct regulatory maps for species such as *S. cerevisiae* , but it is impractical to implement on a species per species basis (MacIsaac et al., 2006). There are also difficulties in developing similar databases for filamentous fungi (Huang et al., 2019).

Overall, while identification of putative binding sites for transcription factors is quite accessible, actual confirmation is expensive and time consuming even with the more advanced techniques mentioned. This is a problem for the field of metabolic regulation and as a direct result many publications which outline or describe a newly identified transcription factor which is involved in interesting processes either do not touch upon the subject or propose a generic binding site. It is of high interest to the field with new advancements in bioinformatic confirmation such as determining binding sites on more parameters than just statistics, e.g. protein characteristics.

Technical Background

This thesis project utilized a combination of license software and freely available tools and databases. This is an outline of those tools.

AntiSMASH

Biosynthetic gene clusters follow certain patterns in terms of structure and can therefore be predicted in terms of localization in a genome. antiSMASH - short for Antibiotics and Secondary Metabolite Analysis SHell - is a freely available service and tool which utilizes this (Blin et al., 2019). It identifies, annotates and compares biosynthetic gene clusters in fungi and in bacteria as well. antiSMASH utilizes profile hidden Markov models of genes. It starts by identifying highly conserved core enzymes from database comparisons and then searches up- and downstream based on internal rules from the type of core enzyme. It then checks for overlaps and proposes a couple of candidate clusters. The program is capable of recognizing and predicting a wide range of enzyme types and other structures based on an internal rule set.

The user submits an annotated or unannotated sequence file such as FASTA or GenBank. The server then runs the job based on user settings and produces an output where individual putative gene clusters are available for analysis. This tool was used to identify putative patulin clusters in species where the species was known to produce the mycotoxin but gene sequences were unannotated.

RSAT

RSAT (van Helden, André and Collado-Vides, 1998) is short for Regulatory Sequence Analysis Tools and has freely available tools that can be used for motif discovery. Motif discovery is crucial when finding unique patterns and binding sites for transcription factors. A quite accessible way to do this *in silico* is by utilizing algorithms. The two general classes of algorithms are the enumerative approach and the probabilistic approach (Hashim, Mabrouk and Al-Atabany, 2019).

Probabilistic involves comparing oligonucleotide frequencies based on a model set. It can then deduce if motifs are overrepresented and are thus more likely to not be random by nature but instead have a biological meaning. A conserved motif is one that is selected for during natural selection. This is the model RSAT utilizes. RSAT can also be used for dyad-mer discovery, which tells the program to look for overrepresented motifs consisting of two trinucleotides with a spacer between them. This aligns with the consensus for the zinc binuclear transcription factor binding site. RSAT and MEME both produce overrepresented motifs based on statistics, and the factor of overrepresentation can be measured in an E-value (Ma, Noble and Bailey, 2014). An E-value is a quick way to appraise the significance of a given motif, and is the expected number of times that a given motif would occur if a sequence was generated randomly, using the submitted background model as a base.

MEME

MEME (Bailey et al., 2009) is short for Multiple EM for Motif Elicitation, which is a freely available tool used to discover ungapped motifs in a given sequence. MEME produces sequence logos and has the ability to predict motifs longer than 50 bps. It uses a background model to perform similar probabilistic calculations as RSAT.

It is useful as a complement to RSAT due to the ability to detect longer oligonucleotide motifs, as well as handling degenerate bases. (Ma, Noble and Bailey, 2014)

Qiagen CLC Main Workbench

CLC Main workbench is a license-based basic bioinformatics software for working with sequencing and omics datasets. It features tools such as sequence reading and trimming, primer design, alignment, BLAST database construction, molecular modeling, motif recognition, annotation and more. (QIAGEN CLC Genomics Workbench).

DeepCoil

DeepCoil is a freely available tool based on neural network-algorithms that can be used to predict the presence of coiled-coil domains in a given protein sequence. It is written in Python and utilizes the Keras machine learning library. It is useful in the scope of this thesis to estimate if BZTFs are prone to bind as dimers or not depending on the presence of a coiled-coil interaction close to the Zn_2Cys_6 zinc finger domain. DeepCoil has been shown to outperform other coiled-coil prediction tools such as Multicoil2, LOGICOIL and RFCoil (Ludwiczak et al., 2019). A newer version of the program called DeepCoil2 has been released, but a study proving its efficacy has not been published. This thesis utilizes the DeepCoil version seen in the paper by Ludwiczak et a., (2019).

EasyFig

EasyFig (Sullivan, Petty and Beatson, 2011) is a freely available tool based in Python which utilizes a BLAST algorithm to visualise an alignment for comparative genomics (Altschul et al., 1990). While it can plot regions of varying sizes, it is well suited for comparison of a limited length, such as a BGC. By using a gene sequence and annotation files, it can generate TBLASTx files, enabling a very user-friendly and clear comparison.

AUGUSTUS

AUGUSTUS (Hoff and Stanke, 2013) is a freely available tool for eukaryotic prediction of genes in a given genome sequence. As it is based on a Hidden Markov Model, It utilizes a training set for many organisms, but can also be trained by using an accompanying program if so desired.

Method

Sequence Acquisition

Most DNA sequences and related data were obtained from the National Center for Biotechnology Information or NCBI (NCBI, 1988). The genome for *P. expansum* IBT34672 was obtained from the DTU Bioengineering IBT fungal collection.

Identification of putative patulin clusters in patulin-producing filamentous ascomycetes

A study made by Frisvad (2018) was used as a basis for assessment of suitable candidates for alignment construction of patulin clusters. *Penicillium* species described in the study with quality genomes available in Genbank NCBI (1988) were chosen. The accession numbers from NCBI are listed in parentheses following the species. The chosen organisms were *P. expansum, P. vulpinum* (GCA_002072255.1)*, P. paneum* (GCA_000577715.1)*, P. antarcticum* (GCA_002072345.1)*, P. griseofulvum* (GCA_001561935.1), *P. dipodomyicola* (GCA_015585785.1) and *P. carneum* (GCA_000577495.1).

The species were then individually processed with AntiSMASH, using relaxed detection strictness. The candidate cluster with the highest percentage identify score with patulin was putatively chosen and the exact perimeters of the cluster was moved by using the alignment tool with standard settings in CLC Main Workbench with a characterized patulin cluster (Li et al., 2015) and by using EasyFig to visualise the movement of the individual parts of the cluster. In the case that AntiSMASH did not return any match for the patulin cluster, BLASTn was used within CLC Main Workbench instead to search within the *P. expansum* IBT34672 genome. All BLAST searches were performed in this thesis project using the complete genomes of the *Penicillium* spp. as bases for a BLAST database.

If the genome was originally unannotated, the putative patulin region was processed through AUGUSTUS. The AUGUSTUS prediction jobs were performed using *Fusarium graminearum* as a background training model and with no additional hints or external evidence. This due to supervisor recommendation and being the closest taxonomic relative to *Penicillium.*

Identification of Global regulator homologues *in silico* in *P. expansum* IBT34672

Publications from Brakhage (2012) and Assaf (2020) were used as guidelines for global regulation in *P. expansum*. Homologues were obtained from NCBI (1988) and the protein accession numbers are listed in the parentheses. The global regulators to be identified in *P. expansum* were LaeA (KGO39425.1), CreA (XP_014537353), AreA (KGO47814.1), AreB (NreB Homolog) (AAC09045), PaCC (AFS18474), Skn7 (KAF3030017), Yap1(KAF3015432) and the CBC-complex (HapX - Q4WER3, HapB - AAP92404, HapC - AAC49411, HapE - AAD12363) plus SreA (AAD25328).

The protein sequences were then imported into CLC Main Workbench. They were then run through TBLASTn against the *P. expansum* IBT34672 genome. The best matches were notated and compared to open reading frames and prior annotations in *P. expansum* IBT34672.

Identification of binding sites for global regulators in the patulin biosynthetic gene cluster in *P. expansum* IBT34672

Contemporary data for binding sites in *P. expansum* or close taxonomic relatives were identified. CLC Main Workbench was then used to map out the patulin BGC. An upstream region of a thousand bps was denoted as putative promoter regions and extracted for each individual gene, regardless if they included open reading frames or not . Studies were consulted for putative binding sites for the DNA-binding global regulators, see Table 1. CLC Main Workbench was then used to search for these motifs in the promoter regions of the *patL* genes. The number of hits per gene were annotated as well as if they were positioned within an ORF and if they were present on the sense or antisense strand.

Identification of putative binding sites for patL in the patulin biosynthetic gene cluster in *P. expansum* IBT34672 through alignment

RSAT was used to detect over-represented dyads (spaced pairs of k-mers) in the promoter regions of the genes making up the patulin BGC. First, an upstream region of a 1000 bps was denoted as putative promoter regions and extracted for each individual gene, regardless if they included open reading frames or not. These sequences were then added to RSAT, which was set to search for any repeating dyad (inverted, everted or direct) on both strands with a spacing region between 3 and 11 bps, based on a motif developed during this project. The background model for statistical purposes was set to *Penicillium digitatum* GCF 000315645.1 PdigPd1 v1. The experiment was then repeated with the "any dyad" setting, which includes non-repeating dyads. MEME was then used by submitting the upstream regions of the patulin cluster as 0th degree background and sequence to be analysed. MEME was set to detect any number of sites per sequence with motifs that were from 6 to 50 bps wide.

An alignment was then made of the biosynthetic gene clusters from seven *Penicillium* spp. known to produce patulin. It was first constructed using tBlastX and Easyfig (Sullivan, Petty and Beatson, 2011) to see how the placement of individual genes within the genome had moved. The individual genes were then adjusted accordingly in CLC Main Workbench to fit the patulin cluster in *P. expansum,* whereafter CLC Main Workbench was used to calculate the alignment using a BLAST-like algorithm. In combination with the motifs identified by RSAT, MEME and the CLC Main Workbench motif search tool, evaluation of the possible biological meaning of sites and sequences of logos were made.

Identification of putative zinc binuclear transcription factors in *P. expansum, A. nidulans, F. graminium, N. crassa* and *S. cerevisiae* and characterization of coiled-coil domains

The consensus sequence of BZTFs was based on MacPherson, Larochelle and Turcotte (2006), but the motif that was used is Cys-X2-Cys-X6-Cys-X5-16-Cys-X2-Cys-X6-9-Cys, which was developed during the project. The coding sequence (CDS) motifs of annotated Genbank file of *P. expansum* IBT34672 were then translated to amino acid sequences, which were then analysed for the consensus sequence using the tool "Motif Search". Hits were then extracted and the location of the start of the motif was noted. Additionally, a separate search was performed specifically to find pathway-specific BZTFs. This was done by applying the motif identification

procedure exclusively on translated CDS which resided in candidate cluster regions, as annotated by AntiSMASH.

Following this, the sequences were run through the Deepcoil program in Python. The result was then fed into MatLAB using a short script developed for the project (Appendix B). The script noted proteins which inhabited residues which had a coiled-coil probability of over 0.45 and the position of the first residue with that characteristic which was located after the last cysteine in the Zn_2Cys_6 motif.

The entire procedure was then repeated with *Aspergillus nidulans* (GCA_000149205.2), *Fusarium graminearum* (GCA_000240135.3), *Neurospora crassa* (GCA_000182925.2), as well as with *Saccharomyces cerevisiae* (GCA_000146045.2) as a control.

Characterization of coiled-coil domain and relation to type of binding site of characterized zinc binuclear transcription factors

Publications were identified based on the criteria that it experimentally verified a binding site for a BZTF in a filamentous ascomycete. The binding sites were then characterized as oligonucleotides or repeat. The protein sequences of the studied transcription factors were extracted (Appendix A), after which they were run through the Deepcoil program in Python. The result was then fed into MatLAB using a script developed for the project (Appendix B). The script noted proteins which inhabited residues which had a coiled-coil probability of over 0.45 and the position of the first residue with that characteristic which was located after the last cysteine in the Zn_2Cys_6 motif.

Results

Identification of Global regulator homologues *in silico* in *P. expansum* IBT34672

The global regulators mentioned in Table 1 were identified in the genome using TBLASTn. Results are presented below in Table 2.

Table 2: TBLASTn matches for global regulators in the genome sequence of *P. expansum IBT34672*. E-value and Score for top two hits are listed, if relevant (Altschul et al., 1990).

Global Regulator	Relevant Matches	E-value	Score
LaeA	Scaffold 23, Scaffold 23	0, 0	1632, 666
CreA	Scaffold 25	0	1820
AreA	Scaffold 15, Scaffold 15	0, 1.41E10-55	3154, 533
AreB	Scaffold 1, Scaffold 1	1.76E-128, 7.88E-46	1035, 418
PaCC	Scaffold 1	0	2517
Skn7	Scaffold 3, Scaffold 3	0, 0	1682, 1015
Yap1	Scaffold 25	0	2158
HapX	Scaffold 18	3.63E-137	1123
HapB	Scaffold 30	2.60E-111	916

HapC	Scaffold 4, Scaffold 4	2.09E-75, 1.83E-7	626, 118
HapE	Scaffold 13	7.46E-110	892
SreA	Scaffold 10, Scaffold 10	1.86E-85, 1.57E-21	741, 246

Identification of binding sites for global regulators in the patulin BGC in *P. expansum* IBT34672

Most of the global regulators are known to bind to specific DNA sequences and to then regulate the expression of the associated genes. Shown below in table 3 are identified potential binding sites and the occurrence thereof within the patulin BGC.

Table 3: Putative binding site identification 1kb upstream from the transcription start site of all the 15 genes in the patulin BGC. * indicates that one or more of the binding sites is located within the ORF for another gene.† indicates that one or more of the binding sites is on the non-coding (antisense) strand.

Global Regulator	**Stimuli**	**Binding site**	**Potential binding sites in the promoter regions of genes (incidence per gene)**	**Present in promoter regions of X out of 15 genes**
CreA	Carbon Source	SYGGRG	*patH* (5*†), *patG* (3*), *patF* (8*†), *patE*(6*†), *patD* (7*†), *patC* (2*†), *patB* (3*), *patM* (4*†),*patN* (4*†), *patO* (8*†), *patL* (2), *patI*(3*†), *patJ* (8*†), *patK* (6†*)	14
AreA and AreB	Nitrogen level	GATA.	*patH* (9*†), *patG* (5*†), *patF* (9*†), *patE* (5*†), *patD* (18*†), *patC* (7*†), *patB* (7*†), *patA* (13*†), *patM* (7*†), *patN* (11*†), *patO* (7†), *patL* (10†), *patI* (7*†), *patJ* (12*†), *patK* (13*†)	15
PacC	pH	GCCARG	*patG* (1*), *patE* (2*†), *patD* (1*), *patA* (2†), *patO* (2*†), *patL* (3†), *patI* (1*), *patJ* (1), *patK* (1†)	9

Skn7	Osmotic and Oxidative Stress	GGCCCAGA and GGCGAGATCT	*patM* (1), *patN (1†)*	2
Yap1	Oxidative Stress	TTACTAA, TGACAAA, TGAGTAA	*patH* (1†), *patC* (1), *patB* (1†), *patN* (1†*)	4
CBC-complex	Iron	CSAAT-n(12-)RWT (HapB,C,E bind to CSAAT)	*patH* (1), *patC* (1†), *patJ* (1), *patK* (2*†)	4
SreA	Iron	GATA.	*patH* (9*†), *patG* (5*†), *patF* (9*†), *patE* (5*†), *patD* (18*†), *patC* (7*†), *patB* (7*†), *patA* (13*†), *patM* (7*†), *patN* (11*†), *patO* (7†), *patL* (10†), *patI* (7*†), *patJ* (12*†), *patK* (13*†)	15

Identification of putative binding sites for patL in the patulin biosynthetic gene cluster in *P. expansum* IBT34672 through alignment

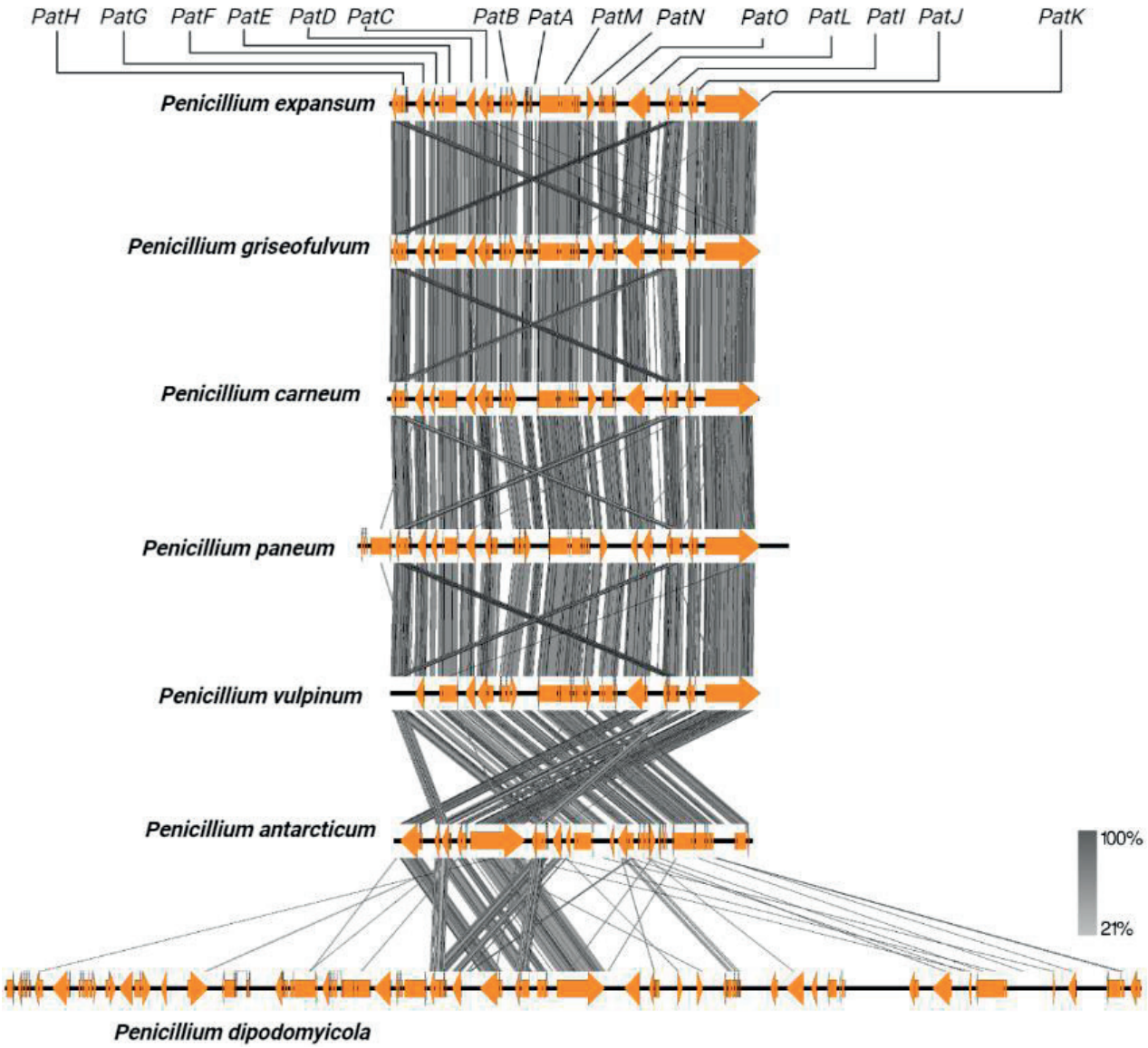


Figure 12, Alignment of the patulin biosynthetic gene cluster in seven *Penicillium* spp. which are known to produce the compound (Frisved, 2018). The orange arrows represent CDS for individual genes in the BGC. The lines between clusters indicate a stronger or weaker homology, measured using tBlastX. Constructed using Easyfig.

The results from RSAT for repeating dyads resulted in three overrepresented motifs. These were 5'-CCC-n(4)-GGG with a E-value of 7.0E-01, 5'-TAT-n(5)-ATA with a E-value of 8.3E-01 and 5'-CCG-n(9)-CCG with a E-value of 8.7E-01. The results from CLC Main workbench for the 5'-CCC-n(4)-GGG motif yielded four putative decently conserved palindromic motifs for *patH, patM, patN, patJ* and *patK*. The sequence logo for the putative sites is seen in Figure 13 A, made by the alignment of the seven *Penicillium* spp. shown in Figure 12. The consensus logo and putative results for the other motifs are presented in Appendix C.

The results from RSAT for any dyads resulted in four overrepresented motifs. These were 5'-ACA-n(7)-GGG with a E-value of 5.30E-03, 5'-CCG-n(3)-GAG with a E-value of 3.0E-01, 5'-CGG-n(6)-GGA with a E-value of 5.6E-01 and 5'-ACC-n(6)-GAG withya E-value of 6.1E-01. The results from CLC Main workbench for the 5'-CGG-n(6)-GGA motif yielded four putative decently conserved palindromic motifs for *patC, patB, patD patM, patJ* and *patK*. The sequence logo for the putative sites is seen in Figure 13 B, made by the alignment of the seven *Penicillium* spp. shown in Figure 12. The consensus logo and putative results for the other motifs are presented in Appendix C.

The results from MEME for oligonucleotides resulted in one overrepresented motif. This was 5' -CCBRAAGGAG with an E-value of 9.3E-016. Note that E-value for MEME has been calculated based on another background model than RSAT. The results from CLC Main workbench for the motif yielded eight putative decently conserved palindromic motifs for *patH, patG, patF, patE, patC, patB, patM, patN, patJ, and patK* The sequence logo for the putative sites is seen in Figure 13 C, made by the alignment of the seven *Penicillium* spp. shown in Figure 12.

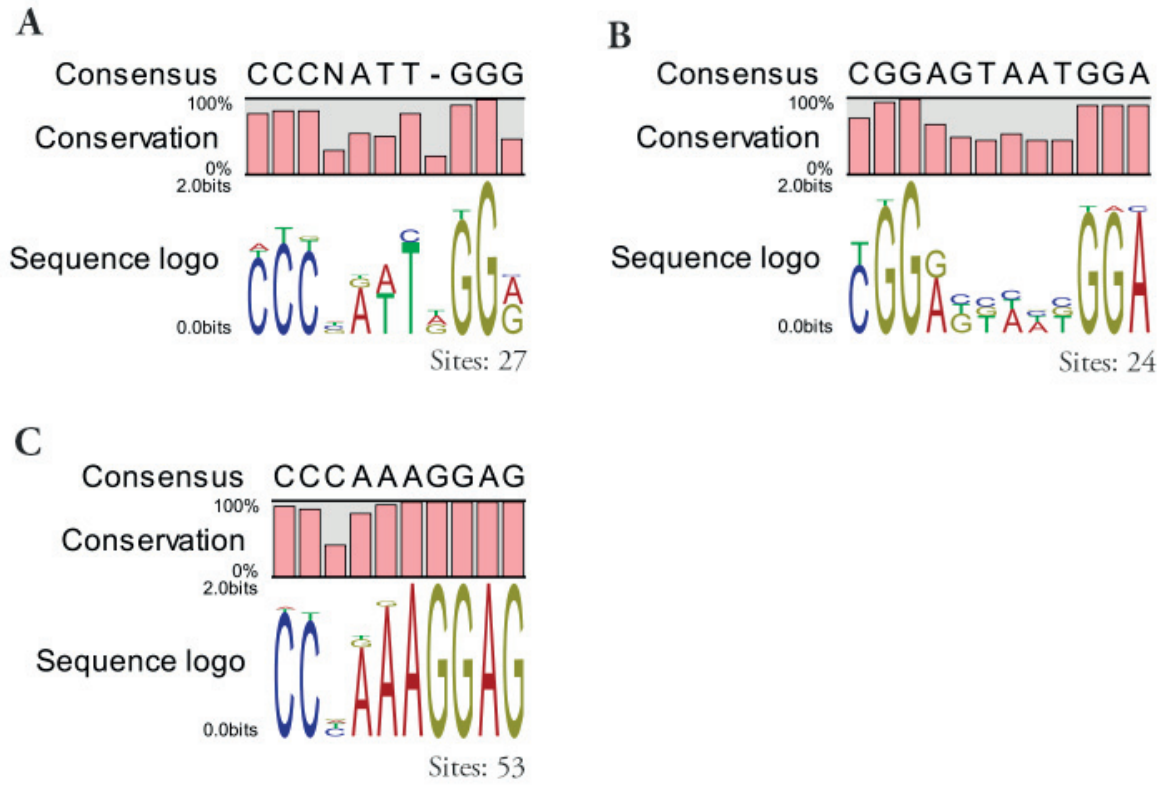


Figure 13: Sequence logos of probabilistically calculated putative binding sites in seven *Penicillium* spp. **A.** Sequence logo for the motif 5'-CCC-n(4)-GGG (E-value 7.0E-01) over 27 sites. Located in 1kb promoter regions of *patH, patM, patN, patJ, patK*. **B.** Sequence logo for the motif 5'-CGG-n(6)-GGA (E-value 5.6E-01) over 24 sites. Located in 1kb promoter regions of *patC, patB, patD, patM, patJ, patK*. **C.** Sequence logo for the motif 5'- CCBRAAGGAG (E-value 9.3E-016*) , over 53 sites. Located in 1kb promoter regions of *patH, patG, patF, patE, patC, patB, patM, patN, patJ, patK*. Note that E-value for MEME has been calculated based on another background model than RSAT.

Identification of putative zinc binuclear transcription factors in filamentous ascomycetes and *S. cerevisiae* and characterization of coiled-coil domain

For the analysis of the translated CDS for the entire genome of *P. expansum* IBT34672, 234 putative Zn_2Cys_6 zinc finger motifs were recognized out of 10 801 annotated CDS. The placement of the consensus motif within the CDS varied, and the distribution is pictured below in Figure 14, with most motifs starting between residue 0 and 50.

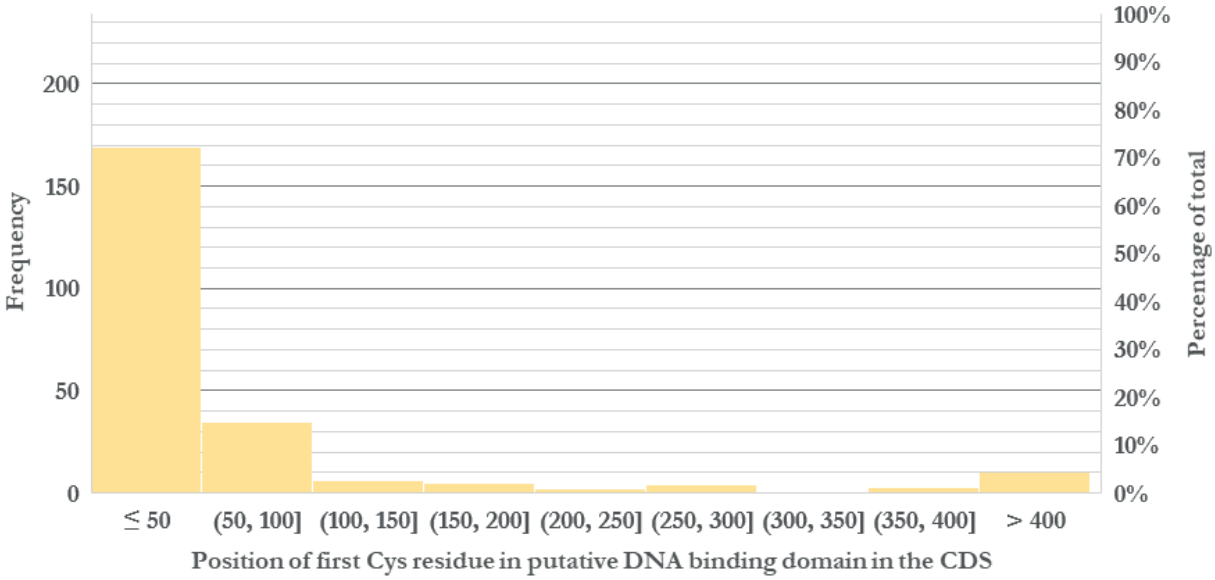


Figure 14: Histogram of the distribution of where the consensus sequence is placed within a given translated CDS. The identified motif is Cys-X2-Cys-X6-Cys-X5-16-Cys-X2-Cys-X6-9-Cys,. The diagram is for all putative BZTFs which were identified in the translated CDS for the whole *P. expansum* IBT34672 genome.

The DeepCoil tool yielded that out of the 234 putative BZTFs 95 displayed regions with a coiled coil probability of over 0.45, where the region was located within 150 bps after the Zn_2Cys_6 zinc finger motif. The distance between the last Cysteine in the motif and the start of the putative coiled-coil domain is shown below in Figure 15.

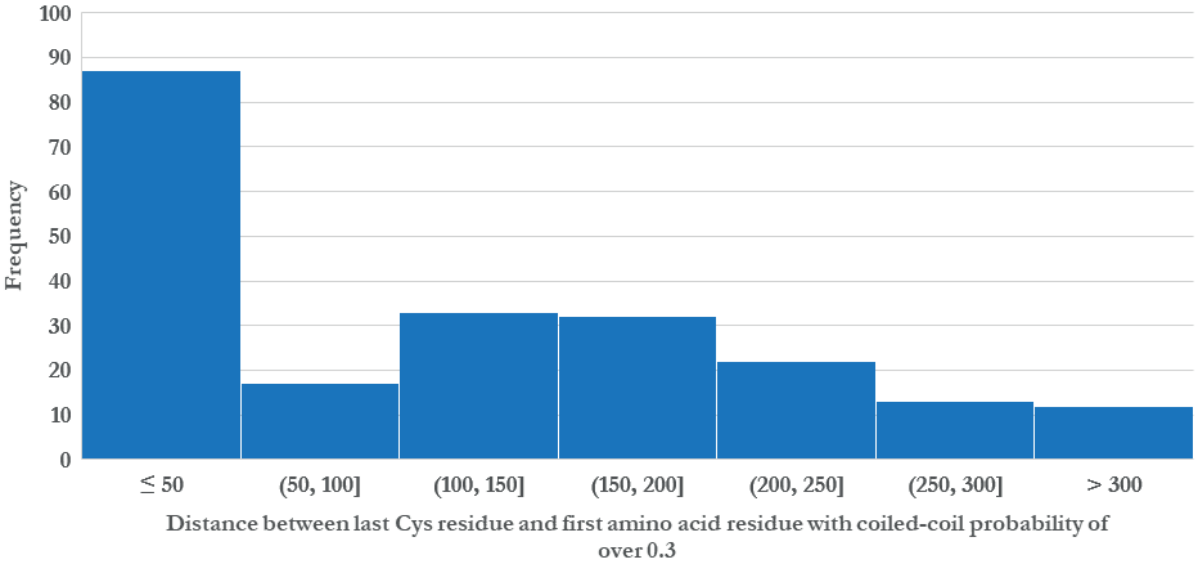


Figure 15: Histogram diagram of the distribution of the distance between the last Cys residue and the first amino acid residue that is suspected to be within a coiled-coil domain. All putative BZTFs which displayed a coiled-coil domain after the Zn_2Cys_6 zinc finger motif are included in the diagram. 150 bps were chosen to be the cut-off point, based on Holmberg and Schjerling (1996).

The results from the remaining investigated fungi can be seen compiled below in table 4. A total of 770 putative BZTFs were identified in filamentous ascomycetes, where of 313 (40.6%) were predicted to inhabit a coiled-coil dimerization domain within 150 bps of the Zn_2Cys_6 motif. A total of 187 candidate cluster regions were identified by ANTIsmash, of which 56 were predicted to contain a BZTF. 24 out of the 56 were predicted to inhabit the coiled-coil dimerization domain.

Table 4: Identified putative BZTFs for different fungal which contained a Zn_2Cys_6-motif (Cys-X2-Cys-X6-Cys-X5-16-Cys-X2-Cys-X6-9-Cys) in five fungal species. Also contains the number of identified putative BZTFs which were predicted to inhabit a coiled-coil domain within 150 bps of the last cysteine in the Zn_2Cys_6-motif by DeepCoil. BZTFs in a candidate cluster means how many of the identified BZTFs were included in a candidate cluster as described by ANTIsmash. BZTF located within clusters was not measured for *S. cerevisiae.*

	P. expansum	***A. nidulans***	***F. graminearum***	***N. crassa***	***S. cerevisiae***
Total BZTFs	234	145	247	144	59
Homo-dimer BZTFs and share of total	95 (40.6%)	44 (30.3%)	118 (47.8%)	56 (38.9%)	25 (42.4%)
BZTFs in candidate cluster	18	18	16	4	
Homo-dimer BZTFs in clusters	12	4	7	1	

Characterization of coiled-coil domain and relation to type of binding site of characterized zinc binuclear transcription factors

15 BZTFs were identified as following criteria, and DeepCoil produced the following results seen in table 5, characterizing 10 of 15 correctly.

Table 5: Binuclear zinc transcription factors in filamentous ascomycetes with confirmed binding sites. Rep indicates that the binding site is of a repeating type, typically with a spacer region. Oli indicates that it is an oligonucleotide. Yes and No indicates whether the protein in question was predicted to inhabit a coiled-coil domain within 150 bps of the last cysteine in the Zn_2Cys_6-motif by DeepCoil. Match means that the two other columns coincide in the proposed way - ergo, repeating binding sites have Coiled-coil domains and oligonucleotides do not.

TF	*Poxc-xra*	*AflR*	*AlcR*	*MlcR*	*CLR-4*	*NirA*	*FarA*	*VerZ*	*UaY*	*PrnA*	*XYR1*	*RhaR*	*CLR-2*	*LeuB*	*Amy*R
Binding site type	Oli	Rep	Oli	Oli	Rep	Oli	Oli	Oli	Rep	Rep	Oli	Rep	Rep	Rep	Rep
Id:ed CC-domain	No	No	No	No	Yes	Yes	No	No	Yes	Yes	Yes	No	Yes	Yes	No
Match	Yes	No	Yes	Yes	Yes	No	Yes	Yes	Yes	Yes	No	No	Yes	Yes	No

Discussion

Identification of Global regulator homologues *in silico* in *P. expansum* IBT34672

As can be seen in the results (Table 2), it was possible to identify most of the characterized global regulators in the specific strain of *P. expansum*. This suggests that continued study of the strain can be expected to follow results made in other closely related fungi However, some of the results were less conclusive - notably SreX and HapC. These results are not unexpected, as notably less research has been done on the CBC-complex in *Penicillium* than it has in e.g. *Aspergillus*. The complex and its structure is also known to vary notably between even closely related groups (Hortscansky et al., 2017). Therefore it is reasonable that *Penicillium* uses a different conformation of the CBC-complex, but the decent scores on the remaining proteins indicate that there is a sort of complex and it is likely involved in iron regulation. However, further analysis is recommended to analyse if all of the functional domains are conserved in the proteins.

As all global regulators are known to vary significantly in function and sequence, this *in silico* experiment is not enough to draw any conclusions, but is a useful tool to estimate what factors are capable of regulating metabolism in this strain. Assessment of conserved functional domains has not also not been done.

Identification of binding sites for global regulators in the patulin BGC in *P. expansum* IBT34672

There is an inherent difficulty in isolating the effect of a global regulator on a single pathway or gene. It requires in-depth research to determine if the e.g. lower patulin production depends on loss of a transcription factor, or due to the loss of fitness associated with the loss of a global regulator. Additionally, the presence of binding sites does not necessarily signify involvement of a given global regulator or transcription factor in the gene whose promoter sequence has the binding site - existence does not necessarily imply function.

LaeA, PaCC and CreA are known to be related to the production of patulin in *P. expansum* (Li et al., 2020). This has been shown in studies where ΔCreA- and ΔPaCC-strains have greatly inhibited patulin production (Chen et al., 2018) (Tannous et al., 2018). Nevertheless, the results of this *in silico* analysis confirms the results of the presence of the PaCC binding sites as stated by Chen et al. (2018), if antisense binding sites are deemed functional as well.

PacC is an important fungal regulator of metabolism as a response to pH changes in the environment. The regulator is known to be an repressor of genes expressed in acidic conditions, (Peñalva, Tilburn, Bignell and Arst, 2008). Since patulin production is higher at lower pHs (such as in an apple), it is reasonable to assume that it represses patulin production in alkaline conditions. If one investigates the chemical stability of patulin at higher pH levels (Collin et al., 2008), this could possibly be to prevent degradation of the compound in those conditions. ΔPaCC-strains have reduced patulin production in acidic conditions as well, but this is reasonably due to the aforementioned widespread metabolic problems from losing a global regulator. Exactly how much of patulin production is preserved at alkaline conditions still seems to be under research, as publications even from the same research group show different results (Zong, Li and Tian, 2015) (Chen et al. 2018) (Tannous et al., 2015).

The heavy prominence of CreA binding sites in almost all of the genes in the patulin BGC are not unexpected. If comparison is made to the well characterized regulation of aflatoxin in *Aspergillus* spp., we can see that carbon sources are correlated to different levels of mycotoxin. Simple sugar such as glucose leads to higher levels of mycotoxin, which may be due to the increased generation of polyketide precursors such as Acetyl-CoA in the TCA cycle (Caceres et al., 2020). However, this stands in contrast to cases such as in *P. chrysogenum* where CreA induces lower penicillin production if simple sugars are present - CreA is generally identified as a gene repressor, and activates in the presence of simple sugars (Assaf et al., 2020).

This also brings up the interesting results in Tannous et al., (2018), where the ΔCreA-mutant as prior stated led to decimated patulin production - but an increased expression of the genes in the patulin BGC. Could the loss of CreA result in lower utilization of simple sugars and therefore less Acetyl-CoA for patulin production, even though the genes are more active? Further research is perhaps required, but the data would point towards the fact that the presence of precursors is more important than the increased expression of the genes in the case of patulin biosynthesis.

The prominent study by Zong, Li and Tian (2015) is compelling to use as a basis for these results - can the environmental impact on the variance of individual gene expression can be attributed to the presence of a related global regulator binding site at that gene? However, the dataset is too small to properly make any decent conclusion and it seems likely that co-expression is in effect by the global regulators. *patA* is the only gene without any identified CreA binding site according to the motif discovery, but it's expression can still be seen to be affected by alternating carbon sources (Zong, Li and Tian, 2015). This could possibly be due to the change in gene expression of *patL*, which would in turn affect *patA* as well.

There are a multitude of studies identifying an effect of nitrogen source on patulin production (Stott and Bullerman, 1975) (Zong, Li and Tian, 2015). The involvement of AreA in aflatoxin biosynthesis has been confirmed as well (Caceres et al., 2020). Due to the abundance of putative binding sites, the likewise involvement of AreA in the regulation of patulin production is a plausible statement. However, nitrogen metabolism has prior been attributed mainly to AreA and AreB, but it would seem that both CBC and CreA may be related to regulation of nitrogen sources as well (Hortshansky et al., 2007)(Tannous et al., 2018).

The CBC-complexes extensive binding site was identified in four of the 15 genes in the genome, which is a notable number to have an effect on transcription. The complex is quite undocumented in fungi and much fewer experiments have been done in comparison to e.g. LaeA. Iron is an essential component of all life, serving vital roles in the TCA cycle, DNA replication, amino acid biosynthesis, and so on (Misslinger, Hortschansky, Brakhage and Haas, 2021). Interestingly enough, iron is also essential for the function of cytochrome P450 enzymes through the heme group. *patH* and *patI* have been characterized as cP450s in studies by Li et al. (2019). The presence of the putative binding site in the promoter region of *patH* would therefore seem reasonable.

Identification of putative binding sites for patL in the patulin biosynthetic gene cluster in *P. expansum* IBT34672 for alignment

The patulin biosynthetic gene cluster is well conserved in closely related *Penicillium* spp. as can be seen in Figure 12. *P. expansum, P. griseofulvum, P. carneum, P. paneum* and *P. vulpinum* have largely identical structuring of the cluster, with the possible exception of the absence of *patA*, but it is possible that it is only the result of a missed annotation. *P. antarctictum* has had a clear scrambling and inversion event, with *patK, patJ, patI* and *patL* being dislocated in comparison to the first five *Penicillium* spp. *P. dipodomyicola* is an outlier, however, with only *patK, patJ, patI, patL, patG* and *patH* being identifiable in the cluster region. More research is needed to properly identify how analogs or the scattering of genes still results in the fungi being capable of producing the mycotoxin.

The parameters used for the RSAT and MEME analysis were based on identified BZTF binding sites in filamentous fungi (Appendix A). The motifs proposed by RSAT were weakly overrepresented and did not have any ubiquitous candidates, with the possible exception of 5'-CCG-n(6)-GGA, which had decently conserved binding sites positioned in putative promoter regions in 6 out of 15 genes in the patulin cluster. However, the very low E-value can attribute the sites to chance, and two of the putative binding sites were located within CDS, which are generally more well conserved. While it is known that the presence of a BZTF binding site is not crucial in every promoter region for regulation to occur (Kong et al., 2020), a higher number likely indicates preservation and function.

MEME analysis is of lower quality due to the smaller background used for the calculations, but the E-values are of such an order of magnitudes different that the result is likely more significant. Additionally, the sequence is conserved to a greater extent across species. The proposed motif, 5'-CCBRAAGGAG, was identified in 10 of 15 putative promoter regions and none of the regions were located within a CDS. Due to equipment limitations caused by the Covid-19 pandemic, this binding site could not be experimentally verified in this thesis. If so was the case, I would likely have applied the same procedures used by e.g. Liu et al. (2018). However, an evaluation would have to be done on which promoter regions the EMSA should be performed on and if a DNase 1 footprinting assay is necessary.

While the putative binding site may not follow the conventional appearance of a BZTF binding site, this kind of asymmetric binding site is not unheard of. As an example, the *P. oxalicum* BZTF PoxCxra is known to bind to 5'-ATCAGATCCTCAAAGA-3' and 5'-GCTGAGTCCTT-3 (Liao et al., 2019). Baba et al. (2009) investigates MlcR, a BZTF involved in HMG-CoA reduction and is contained within a BGC with its associated genes. MlcR, like Pho7, was determined to bind monomerically to a (A/T)CGG site as a part of a larger asymmetric repeat motif. Baba then proposes that BZTFs in filamentous ascomycetes may have greater plasticity in their recognition of binding sites. However, the article has received little attention as of 2021. A way to explore this putative difference is to have a genome-wide perspective.

Identification of filamentous ascomycetes and *S. cerevisiae* and characterization of coiled-coil domain

The consensus sequence for the Zn_2Cys_6-motif was described by Schjerling and Holmberg (1996) as Cys-X2-Cys-X6-Cys-X5-12-Cys-X2-Cys-X6-9-Cys. The middle spacing region was here adjusted to be X5-16 instead, due to newer studies describing BZTFs such as AlcR which did not fit the original motif. It is also presumed that the internal placement of the Zn_2Cys_6 motif cannot be used as a screening method due to the existence of BZTFs such as Pho7 and Ume6p that fall outside the common notion of the amino terminal localization. The 150 bp range was adapted from Schjerling and Holmberg (1996), as anecdotal evidence suggests that most functional BZTFs are within the boundaries with some margin. GAL4 (CAA97969.1) for example, has a distance of 23 bps using DeepCoil.

Studies have proposed that the linker domain; which is located between the zinc finger domain and the coiled-coil domain, has important rigid properties that help retain specificity (MacPherson, Larochelle and Turcotte, 2006). An extended linker domain would likely lose this function However, this study does not consider BZTFs that might be able to dimerize in spite of no conventional dimerization domain or the counterpart - proteins possessing the domain but binding as monomers. A notable share of the identified putative BZTFs are likely false positives or non-functioning, but 234 out of 10801 putative CDS is in line with the estimate that 6-7% of an eukaryotic genome encodes DNA-binding proteins and the share of those that may be BZTFs in fungi (Walter et al., 2009) (Etxebeste, 2021). False positives include cysteine-heavy proteins, such as keratins. Even though Deepcoil is a sophisticated coiled-coil prediction tool, it is not guaranteed to be perfectly accurate. As an experiment, a characterized dimer BZTF such as GAL4 yielded a coiled-coil site with an probability of 0.45, and was thus used as a lower limit for the analysis. Additionally, since the exact characteristics of the training set for DeepCoil is unknown, it is possible that the shorter coiled-coil domains that are seen in BZTFs may be underdetected (Schjerling and Holmberg, 1996).

A large cause for this thesis objective of exploring BZTFs in filamentous ascomycetes was the scholarly bias towards yeast within the metabolic study of fungi, and if the taxonomic differences could putatively affect the study of the shared BZTFs. The results of this analysis - 95 out of 234 putative BZTFs displaying a putative coiled-coil dimerization domain within the first 150 bps after the Zn_2Cys_6-domain - stands in stark contrast to the result with 60 out of 79 BZTFs, the majority of which was from *S. cerevisiae,* as investigated by Schjerling and Holmberg (1996). However, *S. cerevisiae* was investigated using DeepCoil as well which resulted in a new estimate of 25 out of 59 putative BZTFs binding homodimerically in the yeast. This result would imply that the results from Schjerling and Holmberg (1996) are obsolete if DeepCoil is trusted to be a more accurate prediction procedure than Paircoil and the algorithm they used. These results also suggests that while palindromic binding sites are seen as the norm, BZTFs in *P. expansum, A. nidualns, F. graminearum, N. crassa* and *S. cerevisiae* are more likely to bind as monomers than dimers.

While statistically robust evaluations have not been done, there does appear to be slight variations in the share of BZTFs that bind dimerically in different fungi, but it does not appear to be divided between filamentous and non-filamentous fungi. This has some implications if BZTFs are to be characterized in e.g. *Aspergillus* or *Fusarium,* as the notably lower respectively higher share of dimerically binding TFs could affect binding site identification. While the sample size makes it unsatisfactory to make any conclusions, it is interesting to discuss

what could cause this discrepancy between e.g. *Aspergillus* and *Fusarium*. They are both Ascomycetes but in the different classes - Eurotiomycetes respectively Sordariomycetes.

BZTFs which resided in potential BGC regions were handled in a unique test as well to identify any difference in coiled-coil dimerization domain occurrence compared to those located outside clusters. A total of 187 candidate cluster regions were identified by AntiSMASH over the four filamentous ascomycete species, of which 56 had putative BZTFs. Out of those 56, 24 were predicted to have the coiled-coil dimerization domain, which is approximately the same share as what was seen in the rest of the genome.

Quite surprisingly when discussing the motif discovery results, patL is denoted as possessing a coiled-coil dimerization domain (residues 75-105) in direct connection to the Zn_2Cys_6 motif. This would presumably mean that a classic palindromic binding site could be identified, as many times the binding site is overrepresented to a large degree and present in all of the promoter regions of the gene (Fox, Gardiner, Keller and Howlett, 2008). But any well-conserved palindromic or repeat binding sites were unable to be identified. The proposed binding site would appear to be more of an oligonucleotide. However, binding sites may behave unexpectedly in BZTFs.

While discussion in new publications pertaining BZTF is again mostly centered on the accepted palindromic binding site, there is research describing many BZTFs which bind sequences quite unlike either short oligonucleotide sites or the standard palindromic (MacPherson, Larochelle and Turcotte, 2006). Another well characterized example is Pho7, which is a crucial transcription factor in yeast related to phosphate starvation. Garg et. al (2018) noted that it is bound monomerically, but a short α-helix was determined to be present following the linker region. But, while it still targets a 5'-CGG binding site, the transcription factor has additional contact with the major groove of the target DNA through the short α-helix. The prefered final binding site is thus akin to 5'-TCG(G/C(A/T)NNTTNAA. Is it possible that 5'-CCBRAAGGAG, is the result of a similar α-helix interaction? However, analysing Pho7 (NP_595543) with DeepCoil did not result in any predicted coiled-coil domains.

Discovering these large asymmetrical motifs is quite difficult, due to the lack of documentation and expectations of a simple palindromic binding site. Other binding sites of similar complexity have been mostly discovered using large ChiP-Seq databases (MacIsaac et al., 2006). Therefore I would like to propose the need for a new generation of meta motif discovery tools. By combining the existing probability-based tools such as MEME or RSAT with DNA-protein interaction tools and homology-based comparison tools, perhaps the assessment can be strong enough to render an experimental confirmation redundant. There is research into the area, but many solutions rely on known protein structures or relation to binding site databases (Si, Zhao and Wu, 2015).

Characterization of coiled-coil domain and relation to type of binding site of characterized zinc binuclear transcription factors

There has been discussion whether the coiled-coil domain can be used as a predictor for proteins which do or do not bind monomerically and thus do not follow the commonly attributed binding motif of the triplet repeats (Schjerling and Holmberg, 1996)(Cahuzac, Cerdan, Felenbok and Guittet, 2001).

It would appear logical due to the BZTFs which have had their 3D structure validated through crystallization (e.g. Figure 7) that the palindromic binding site stems from the dimerization region, in which the repeat occurs due to the second protein binding slightly further down the strand. But the interest lies if coiled-coil dimerization prediction tools such as DeepCoil could be used when a new BZTF is being researched to assist in identifying the binding site. The results from the 15 TFs show some promise that this is the case, but due to the binary nature of the classification, chance is a heavy factor. Interestingly, the putative binding sites proposed for *patL* in this paper does not follow the expected pattern, due to the prediction of a coiled-coil domain in the protein. There is also some difficulty and discussion to be made regarding if the binding sites are oligonucleotides or not - the BZTF FarA for example, binds to the site 5'-CCT CGG (Hynes et al., 2006). Due to the two triplets it could definitely be seen as a repeat, but due to the short an asymmetric nature it could also be seen as an oligonucleotide.

This anecdotal experiment is not statistically robust enough to draw any conclusions, but is interesting as a proof of concept and in order to test the apparent efficacy of the DeepCoil tool. A full-scale test is outside of the scope and timeframe of this project. A good start would be to include ChiP-Seq databases and genome regulatory maps from *S. cerevisiae* e.g. from MacIsaac et al. (2006).

The DeepCoil used in this thesis was version 1, which is the one that was used to prove its efficacy (Ludwiczak et al., 2019). However, they have since released a new version with an updated algorithm called DeepCoil2, but the publication for it is not out yet. DeepCoil2 was considered for use in this project, but the predictions were drastically different for the first and second versions. E.g. while DeepCoil2 managed to characterize 11 out of 15 correctly in regards to the BZTFs with confirmed binding sites, it did so with a lower cut-off of 0.3. It also produced even fewer predicted coiled-coil domains in BZTFs in almost all cases. Due to the DeepCoil2 data and training set being unpublished, it decided to disregard it.

The difficulty of finding BZTFs with experimentally validated binding sites in filamentous ascomycetes highlights the fact that verifying is time-consuming and cumbersome, using protein-DNA interaction methods like EMSA, DNase 1 Footprinting or even ChiP-Seq. I would like to reinforce the importance of new accessible meta motif discovery tools.

Conclusion

The usage of CLC Main Workbench together with DeepCoil produced results that indicate that the binuclear zinc transcription factors (BZTFs) are as prone to bind homodimerically in filamentous ascomycetes as in *S. cerevisiae*. However, it also suggests that monomerically binding BZTFs may be more common than what was previously thought. The results are of potential use to the field due to a tendency to assume that BZTFs bind as dimers.

However, the analysis is rudimentary and should be improved upon to draw safer conclusions. It is of high importance to statistically determine if the presence of a coiled-coil dimerization region can be used to predict the type of binding site. It would appear logical due to the nature of the protein and crystal structures which have been elucidated so far, and here a simple anecdotal experiment comes to the same conclusion. *patL* putatively binds to an asymmetric oligonucleotide 5'- CCBRAAGGAG, which stands in contrast to the result that it may bind as a dimer.

The metabolic regulation of patulin is an interesting research field both as a model for the regulation of biosynthetic gene clusters in fungi, and due to the economic interest in the compound from the agricultural industry. Here prominent global regulators like CreA, PaCC and the CBC-complex are shown to be present in the *P. expansum* genome and that they likely have a role in controlling the transcription of patulin, reinforcing current findings in the field. The iron responsive regulative CBC-complex has putative binding sites in the several promoter regions and is likely to be involved, but the complex remains to be investigated thoroughly in *Penicillium*.

The experimental verification of transcription factor binding sites using techniques such as EMSA, DNase 1 Footprinting and ChIP-SEQ is slow and labour intensive. The development of a new generation of computer based tools which can combine homology-based, probabilistic and DNA-protein interaction approaches could lead to a bioinformatics breakthrough in regards to natural product research.

References

Almeida, F., Rodrigues, M. and Coelho, C., 2019. *The Still Underestimated Problem of Fungal Diseases Worldwide.* Frontiers in Microbiology, 10.

Altschul, S., Gish, W., Miller, W., Myers, E. and Lipman, D., 1990. *Basic local alignment search tool.* Journal of Molecular Biology, 215(3), pp.403-410.

Baba, S., Kinoshita, H., Hosobuchi, M. and Nihira, T., 2009. *MlcR, a zinc cluster activator protein, is able to bind to a single (A/T)CGG site of cognate asymmetric motifs in the ML-236B (compactin) biosynthetic gene cluster*. Molecular Genetics and Genomics, 281(6).

Bailey, T., Boden, M., Buske, F., Frith, M., Grant, C., Clementi, L., Ren, J., Li, W. and Noble, W., 2009. *MEME SUITE: tools for motif discovery and searching.* Nucleic Acids Research, 37(Web Server), pp.W202-W208.

Bayram, O., Krappmann, S., Ni, M., Bok, J., Helmstaedt, K., Valerius, O., Braus-Stromeyer, S., Kwon, N., Keller, N., Yu, J. and Braus, G., 2008. *VelB/VeA/LaeA Complex Coordinates Light Signal with Fungal Development and Secondary Metabolism.* Science, 320(5882), pp.1504-1506.

Bérdy, J., 2005. *Bioactive Microbial Metabolites*. The Journal of Antibiotics, 58(1), pp.1-26.

Blanco, A. and Blanco, G., 2017. *Metabolism. Medical Biochemistry,* pp.275-281.

Blin, K., Shaw, S., Steinke, K., Villebro, R., Ziemert, N., Lee, S., Medema, M. and Weber, T., 2019. *antiSMASH 5.0: updates to the secondary metabolite genome mining pipeline.* Nucleic Acids Research, 47(W1), pp.W81-W87.

Bok, J. and Keller, N., 2004. *LaeA, a Regulator of Secondary Metabolism in Aspergillus spp.* Eukaryotic Cell, 3(2), pp.527-535.

Bok, J., Chiang, Y., Szewczyk, E., Reyes-Dominguez, Y., Davidson, A., Sanchez, J., Lo, H., Watanabe, K., Strauss, J., Oakley, B., Wang, C. and Keller, N., 2009. *Chromatin-level regulation of biosynthetic gene clusters.* Nature Chemical Biology, 5(7), pp.462-464.

Bok, J., Chiang, Y., Szewczyk, E., Reyes-Dominguez, Y., Davidson, A., Sanchez, J., Lo, H., Watanabe, K., Strauss, J., Oakley, B., Wang, C. and Keller, N., 2009. *Chromatin-level regulation of biosynthetic gene clusters.* Nature Chemical Biology, 5(7), pp.462-464.

Braga, R., Dourado, M. and Araújo, W., 2016. *Microbial interactions: ecology in a molecular perspective.* Brazilian Journal of Microbiology, 47, pp.86-98.

Brakhage, A., 2012. *Regulation of fungal secondary metabolism.* Nature Reviews Microbiology, 11(1), pp.21-32.

Caceres, I., Al Khoury, A., El Khoury, R., Lorber, S., P. Oswald, I., El Khoury, A., Atoui, A., Puel, O. and Bailly, J., 2020. *Aflatoxin Biosynthesis and Genetic Regulation: A Review*. Toxins, 12(3), p.150.

Cahuzac, B., Cerdan, R., Felenbok, B. and Guittet, E., 2001. *The Solution Structure of an AlcR-DNA Complex Sheds Light onto the Unique Tight and Monomeric DNA Binding of a Zn2Cys6 Protein.* Structure, 9(9), pp.827-836.

Cairns, T., Nai, C. and Meyer, V., 2018. *How a fungus shapes biotechnology: 100 years of Aspergillus niger research.* Fungal Biology and Biotechnology, 5(1).

Campbell, M., Rokas, A. and Slot, J., 2012. *Horizontal Transfer and Death of a Fungal Secondary Metabolic Gene Cluster.* Genome Biology and Evolution, 4(3), pp.289-293.

Cepeda-García, C., Domínguez-Santos, R., García-Rico, R., García-Estrada, C., Cajiao, A., Fierro, F. and Martín, J., 2014. *Direct involvement of the CreA transcription factor in penicillin biosynthesis and expression of the pcbAB gene in Penicillium chrysogenum.* Applied Microbiology and Biotechnology, 98(16), pp.7113-7124.

Chamoun, R., Aliferis, K. and Jabaji, S., 2015. *Identification of signatory secondary metabolites during mycoparasitism of Rhizoctonia solani by Stachybotrys elegans*. Frontiers in Microbiology, 6.

Chen, Y., Li, B., Xu, X., Zhang, Z. and Tian, S., 2018. *The pH-responsive PacC transcription factor plays pivotal roles in virulence and patulin biosynthesis in Penicillium expansum.* Environmental Microbiology, 20(11), pp.4063-4078.

Collin, S., Bodart, E., Badot, C., Bouseta, A. and Nizet, S., 2008. Identification of the Main Degradation *Products of Patulin Generated Through Heat Detoxification Treatments.* Journal of the Institute of Brewing, 114(2), pp.167-171.

Cubero, B. and Scazzocchio, C., 1994. *Two different, adjacent and divergent zinc finger binding sites are necessary for CREA-mediated carbon catabolite repression in the proline gene cluster of Aspergillus nidulans.* The EMBO Journal, 13(2), pp.407-415.

Das, P., Ramachandran, K., vanWert, J. and Singal, R., 2004. *Chromatin immunoprecipitation assay*. BioTechniques, 37(6), pp.961-969.

Davy, A., Kildegaard, H. and Andersen, M., 2017. *Cell Factory Engineering. Cell Systems*, 4(3), pp.262-275.

Demain, A. and Fang, A., 2000. *The Natural Functions of Secondary Metabolites. History of Modern Biotechnology I*, pp.1-39.

Dias, D., Urban, S. and Roessner, U., 2012. *A Historical Overview of Natural Products in Drug Discovery.* Metabolites, 2(2), pp.303-336.

El Hajj Assaf, C., Zetina-Serrano, C., Tahtah, N., Khoury, A., Atoui, A., Oswald, I., Puel, O. and Lorber, S., 2020. *Regulation of Secondary Metabolism in the Penicillium Genus*. International Journal of Molecular Sciences, 21(24), p.9462.

Errampalli, D., 2014. *Penicillium Expansum (Blue Mold).* In: S. Bautista-Baños, ed., Postharvest Decay: Control Strategies, 1st ed. Academic Press, pp.189-231.

Escalante-Chong, R., Savir, Y., Carroll, S., Ingraham, J., Wang, J., Marx, C. and Springer, M., 2015. *Galactose metabolic genes in yeast respond to a ratio of galactose and glucose*. Proceedings of the National Academy of Sciences, 112(5), pp.1636-1641.

Etxebeste, O., 2021. *Fungal Transcription factors: Markers of genetic innovation, network rewiring and conflict between genomics and transcriptomics.* [Manuscript submitted for publication].

Fisher, M., Henk, D., Briggs, C., Brownstein, J., Madoff, L., McCraw, S. and Gurr, S., 2012. *Emerging fungal threats to animal, plant and ecosystem health.* Nature, 484(7393), pp.186-194.

Flaherty, J.E. and Payne, G.A., 1997. ***Overexpression of aflR Leads to Upregulation of Pathway Gene Transcription and increased Aflatoxin Production in Aspergillus flavus***. Appl. Environ. Microbiol. ,63, 3995–4000

Fox, E., Gardiner, D., Keller, N. and Howlett, B., 2008. *A Zn(II)2Cys6 DNA binding protein regulates the sirodesmin PL biosynthetic gene cluster in Leptosphaeria maculans*. Fungal Genetics and Biology, 45(5), pp.671-682.

Frisvad, J., 2018. *A critical review of producers of small lactone mycotoxins: patulin, penicillic acid and moniliformin*. World Mycotoxin Journal, 11(1), pp.73-100.

Frisvad, J., Isbrandt, T. and Larsen, T., 2020. *Fungal Partially Reducing Polyketides and Related Natural Products From Aspergillus, Penicillium, and Talaromyces. Comprehensive Natural Products III*, pp.313-332.

Furukawa, T., Scheven, M., Misslinger, M., Zhao, C., Hoefgen, S., Gsaller, F., Lau, J., Jöchl, C., Donaldson, I., Valiante, V., Brakhage, A., Bromley, M., Haas, H. and Hortschansky, P., 2020. *The fungal CCAAT-binding complex and HapX display highly variable but evolutionary conserved synergetic promoter-specific DNA recognition*. Nucleic Acids Research, 48(7), pp.3567-3590.

Galas, D. and Schmitz, A., 1978. *DNAase footprinting a simple method for the detection of protein-DNA binding specificity*. Nucleic Acids Research, 5(9), pp.3157-3170.

García-Estrada, C., Domínguez-Santos, R., Kosalková, K. and Martín, J., 2018. *Transcription Factors Controlling Primary and Secondary Metabolism in Filamentous Fungi: The β-Lactam Paradigm*. Fermentation, 4(2), p.47.

Garg, A., Goldgur, Y., Schwer, B. and Shuman, S., 2018. *Distinctive structural basis for DNA recognition by the fission yeast Zn2Cys6 transcription factor Pho7 and its role in phosphate homeostasis.* Nucleic Acids Research,.

Gauthier, G., 2017. *Fungal Dimorphism and Virulence: Molecular Mechanisms for Temperature Adaptation, Immune Evasion, and In Vivo Survival.* Mediators of Inflammation, 2017, pp.1-8.

Guzmán-Chávez, F., Salo, O., Nygård, Y., Lankhorst, P., Bovenberg, R. and Driessen, A., 2017. *Mechanism and regulation of sorbicillin biosynthesis by Penicillium chrysogenum. Microbial Biotechnology*, 10(4), pp.958-968.

Hamilton, A. and Gómez, B., 2002. *Melanins in fungal pathogens.* Journal of Medical Microbiology, 51(3), pp.189-191.

Hashim, F. A., Mabrouk, M. S., & Al-Atabany, W., 2019. *Review of Different Sequence Motif Finding Algorithms.* Avicenna journal of medical biotechnology, 11(2), 130–148.

He, X. and Fassler, J., 2005. *Identification of novel Yap1p and Skn7p binding sites involved in the oxidative stress response of Saccharomyces cerevisiae.* Molecular Microbiology, 58(5), pp.1454-1467.

Hellman, L. and Fried, M., 2007. *Electrophoretic mobility shift assay (EMSA) for detecting protein–nucleic acid interactions.* Nature Protocols, 2(8), pp.1849-1861.

Hertweck, C., 2009. *Hidden biosynthetic treasures brought to light.* Nature Chemical Biology, 5(7), pp.450-452.

Hibbett, D., Binder, M., Bischoff, J., Blackwell, M., Cannon, P., Eriksson, O., Huhndorf, S., James, T., Kirk, P., Lücking, R., Thorsten Lumbsch, H., Lutzoni, F., Matheny, P., McLaughlin, D., Powell, M., Redhead, S., Schoch, C., Spatafora, J., Stalpers, J., Vilgalys, R., Aime, M., Aptroot, A., Bauer, R., Begerow, D., Benny, G., Castlebury, L., Crous, P., Dai, Y., Gams, W., Geiser, D., Griffith, G., Gueidan, C., Hawksworth, D., Hestmark, G., Hosaka, K., Humber, R., Hyde, K., Ironside, J., Kõljalg, U., Kurtzman, C., Larsson, K., Lichtwardt, R., Longcore, J., Miądlikowska, J., Miller, A., Moncalvo, J., Mozley-Standridge, S., Oberwinkler, F., Parmasto, E., Reeb, V., Rogers, J., Roux, C., Ryvarden, L., Sampaio, J., Schüßler, A., Sugiyama, J., Thorn, R., Tibell, L., Untereiner, W., Walker, C., Wang, Z., Weir, A., Weiss, M., White, M., Winka, K., Yao, Y. and Zhang, N., 2007. *A higher-level phylogenetic classification of the Fungi*. Mycological Research, 111(5), pp.509-547.

Hoff, K. and Stanke, M., 2013. *WebAUGUSTUS--a web service for training AUGUSTUS and predicting genes in eukaryotes*. Nucleic Acids Research, 41(W1), pp.W123-W128.

Hong, M., Fitzgerald, M., Harper, S., Luo, C. and Speicher, D., 2008. *Structural Basis for Dimerization in DNA Recognition by Gal4*. Structure, 16(7), pp.1019-1026.

Hong, M., Fitzgerald, M., Harper, S., Luo, C., Speicher, D. and Marmorstein, R., 2008. *Structural Basis for Dimerization in DNA Recognition by Gal4*. Structure, 16(7), pp.1019-1026.

Hortschansky, P., Haas, H., Huber, E., Groll, M. and Brakhage, A., 2017. *The CCAAT-binding complex (CBC) in Aspergillus species*. Biochimica et Biophysica Acta (BBA) - Gene Regulatory Mechanisms, 1860(5), pp.560-570.

Houbraken, J., Frisvad, J. and Samson, R., 2011. *Fleming's penicillin producing strain is not Penicillium chrysogenum but P. rubens*. IMA Fungus, 2(1), pp.87-95.

Huang, L., Li, X., Dong, L., Wang, B. and Pan, L., 2019. *Profiling of chromatin accessibility across Aspergillus species and identification of transcription factor binding sites in the Aspergillus genome using filamentous fungi ATAC-seq*. [Manuscript submitted for publication].

Hynes, M., Murray, S., Duncan, A., Khew, G. and Davis, M., 2006. *Regulatory Genes Controlling Fatty Acid Catabolism and Peroxisomal Functions in the Filamentous Fungus Aspergillus nidulans*. Eukaryotic Cell, 5(5), pp.794-805.

International Journal of Epidemiology, 2004. *Clinical trial of patulin in the common cold.* 33(2), pp.243-246.

Jacobson, E., 2000. *Pathogenic Roles for Fungal Melanins*. Clinical Microbiology Reviews, 13(4), pp.708-717.

Jagers op Akkerhuis, G., 2010. *Towards a Hierarchical Definition of Life, the Organism, and Death*. *Foundations of Science*, 15(3), pp.245-262.

Keller, N., 2018. *Fungal secondary metabolism: regulation, function and drug discovery*. Nature Reviews Microbiology, 17(3), pp.167-180.

Kim, H., Heo, D., Park, H., Singh, D. and Lee, C., 2016. *Metabolomic and Transcriptomic Comparison of Solid-State and Submerged Fermentation of Penicillium expansum* KACC 40815. PLOS ONE,.

Kong, Q., Chang, P., Li, C., Hu, Z., Zheng, M., Sun, Q. and Shan, S., 2020. *Identification of AflR Binding Sites in the Genome of Aspergillus flavus by ChIP-Seq*. Journal of Fungi, 6(2), p.52.

Koonin, E., 2012. *Does the central dogma still stand?*. Biology Direct, 7(1), p.27.

Kozlovsky, A.G., Kochkina, G.A., Zhelifonova, V.P., Antipova T.V., Ivanushkina N.E and Ozerskaya S.M., 2019el. *Secondary metabolites of the genus Penicillium from undisturbed and anthropogenically altered Antarctic habitats*. Folia Microbiol 65, 95–102.

Kung, S., Lund, S., Murarka, A., McPhee, D. and Paddon, C., 2018. *Approaches and Recent Developments for the Commercial Production of Semi-synthetic Artemisinin.* Frontiers in Plant Science, 9.

Kwon, M., Steiniger, C., Cairns, T., Wisecaver, J., Lind, A., Pohl, C., Regner, C., Rokas, A. and Meyer, V., 2020. *Beyond the biosynthetic gene cluster paradigm: Genome-wide co-expression networks connect clustered and unclustered transcription factors to secondary metabolic pathways.*

Li, B, Chen, Y, Zhang, Z, Qin, G, Chen, T, Tian, S., 2020. M*olecular basis and regulation of pathogenicity and patulin biosynthesis in Penicillium expansum.* Compr Rev Food Sci Food Saf, 19: 3416– 3438.

Li, B., Chen, Y., Zong, Y., Shang, Y., Zhang, Z., Xu, X., Wang, X., Long, M. and Tian, S., 2019. *Dissection of patulin biosynthesis, spatial control and regulation mechanism in Penicillium expansum.* Environmental Microbiology, 21(3), pp.1124-1139.

Li, B., Zong, Y., Du, Z., Chen, Y., Zhang, Z., Qin, G., Zhao, W. and Tian, S., 2015. *Genomic Characterization Reveals Insights Into Patulin Biosynthesis and Pathogenicity in Penicillium Species.* Molecular Plant-Microbe Interactions®, 28(6), pp.635-647.

Li, B., Zong, Y., Du, Z., Chen, Y., Zhang, Z., Qin, G., Zhao, W. and Tian, S., 2015. G*enomic Characterization Reveals Insights Into Patulin Biosynthesis and Pathogenicity in Penicillium Species.* Molecular Plant-Microbe Interactions®, 28(6), pp.635-647.

Liao, F., 2009. *Discovery of Artemisinin (Qinghaosu).* Molecules, 14(12), pp.5362-5366.

Liao, L., Li, C., Zhang, F., Yan, Y., Luo, X., Zhao, S. and Feng, J., 2019. *How an essential Zn2Cys6 transcription factor PoxCxrA regulates cellulase gene expression in ascomycete fungi?*. Biotechnology for Biofuels, 12(1).

Lis, M. and Walther, D., 2016. *The orientation of transcription factor binding site motifs in gene promoter regions: does it matter?*. BMC Genomics, 17(1).

Liu, Q., Li, J., Gao, R., Li, J., Ma, G. and Tian, C., 2018. *CLR-4, a novel conserved transcription factor for cellulase gene expression in ascomycete fungi.* Molecular Microbiology, 111(2), pp.373-394.

lton, G., Cox, A., Gerard Toussaint, L. and Westwick, J., 2001. *Functional proteomics analysis of GTPase signaling networks.* Methods in Enzymology, pp.300-316.

Ludwiczak, J., Winski, A., Szczepaniak, K., Alva, V. and Dunin-Horkawicz, S., 2019. *DeepCoil—a fast and accurate prediction of coiled-coil domains in protein sequences.* Bioinformatics, 35(16), pp.2790-2795.

Ma, W., Noble, W. and Bailey, T., 2014. *Motif-based analysis of large nucleotide data sets using MEME-ChIP*. Nature Protocols, 9(6), pp.1428-1450.

Macheleidt, J., Mattern, D., Fischer, J., Netzker, T., Weber, J., Schroeckh, V., Valiante, V. and Brakhage, A., 2016. *Regulation and Role of Fungal Secondary Metabolites.* Annual Review of Genetics, 50(1), pp.371-392.

MacIsaac, K., Wang, T., Gordon, D., Gifford, D., Stormo, G. and Fraenkel, E., 2006. BMC Bioinformatics, 7(1), p.113.

MacPherson, S., Larochelle, M. and Turcotte, B., 2006. *A Fungal Family of Transcriptional Regulators: the Zinc Cluster Proteins.* Microbiology and Molecular Biology Reviews, 70(3), pp.583-604.

Mahato, D., Lee, K., Kamle, M., Devi, S., Dewangan, K., Kumar, P. and Kang, S., 2019. *Aflatoxins in Food and Feed: An Overview on Prevalence, Detection and Control Strategies*. Frontiers in Microbiology, 10.

Marmorstein, R. and Harrison, S., 1994. *Crystal structure of a PPR1-DNA complex: DNA recognition by proteins containing a Zn2Cys6 binuclear cluster.* Genes & Development, 8(20), pp.2504-2512.

Mattanovich, D., Sauer, M. and Gasser, B., 2014. *Yeast biotechnology: teaching the old dog new tricks. Microbial Cell Factories,* 13(1), p.34.

McGary, K., Slot, J. and Rokas, A., 2013. *Physical linkage of metabolic genes in fungi is an adaptation against the accumulation of toxic intermediate compounds.* Proceedings of the National Academy of Sciences, 110(28), pp.11481-11486.

Mendoza-Martínez, A., Cano-Domínguez, N. and Aguirre, J., 2020. *Yap1 homologs mediate more than the redox regulation of the antioxidant response in filamentous fungi.* Fungal Biology, 124(5), pp.253-262.

Misslinger, M., Hortschansky, P., Brakhage, A. and Haas, H., 2021. *Fungal iron homeostasis with a focus on Aspergillus fumigatus.* Biochimica et Biophysica Acta (BBA) - Molecular Cell Research, 1868(1), p.118885.

Mitsis, T., Efthimiadou, A., Bacopoulou, F., Vlachakis, D., Chrousos, G. and Eliopoulos, E., 2020. *Transcription factors and evolution: An integral part of gene expression (Review).* World Academy of Sciences Journal,.

Moake, M., Padilla-Zakour, O. and Worobo, R., 2005. *Comprehensive Review of Patulin Control Methods in Foods.* Comprehensive Reviews in Food Science and Food Safety, 4(1), pp.8-21.

Money, N., 2016. *Fungal Diversity.* The Fungi, pp.1-36.

Naranjo-Ortiz, M. and Gabaldón, T., 2019. *Fungal evolution: diversity, taxonomy and phylogeny of the Fungi. Biological Reviews*, 94(6), pp.2101-2137.

National Center for Biotechnology Information (NCBI) [Internet]. Bethesda (MD): National Library of Medicine (US), National Center for Biotechnology Information; [1988] – [cited 2021 Apr 15]. Available from: https://www.ncbi.nlm.nih.gov/

Nielsen, J., Grijseels, S., Prigent, S., Ji, B., Dainat, J., Nielsen, K., Frisvad, J., Workman, M. and Nielsen, J., 2017. *Global analysis of biosynthetic gene clusters reveals vast potential of secondary metabolite production in Penicillium species.* Nature Microbiology, 2(6).

Nützmann, H., Scazzocchio, C. and Osbourn, A., 2018. *Metabolic Gene Clusters in Eukaryotes.* Annual Review of Genetics, 52(1), pp.159-183.

Ofran, Y., Mysore, V. and Rost, B., 2007. *Prediction of DNA-binding residues from sequence.* Bioinformatics, 23(13), pp.i347-i353.

Orejas, M., Espeso, E., Tilburn, J., Sarkar, S., Arst, H. and Penalva, M., 1995. A*ctivation of the Aspergillus PacC transcription factor in response to alkaline ambient pH requires proteolysis of the carboxy-terminal moiety*. Genes & Development, 9(13), pp.1622-1632.

Pál, C. and Hurst, L., 2003. *Evidence for co-evolution of gene order and recombination rate. Nature Genetics*, 33(3), pp.392-395.

Peñalva, M., Tilburn, J., Bignell, E. and Arst, H., 2008. *Ambient pH gene regulation in fungi: making connections*. Trends in Microbiology, 16(6), pp.291-300.

Perrin, R., Fedorova, N., Bok, J., Cramer, R., Wortman, J., Kim, H., Nierman, W. and Keller, N., 2007. *Transcriptional Regulation of Chemical Diversity in Aspergillus fumigatus by LaeA*. PLoS Pathogens, 3(4), p.e50.

Punt, P., Strauss, J., Smit, R., Kinghorn, J., van den Hondel, C. and Scazzocchio, C., 1995. *The intergenic region between the divergently transcribed niiA and niaD genes of Aspergillus nidulans contains multiple NirA binding sites which act bidirectionally*. Molecular and Cellular Biology, 15(10), pp.5688-5699.

QIAGEN CLC Genomics Workbench 20.0 (https://digitalinsights.qiagen.com/)

Reece, J. B., Urry, L. A., Cain, M. L., Wasserman, S. A., Minorsky, P. V., and Jackson, R. B., 2011. A*n introduction to metabolism*. In Campbell biology (10th ed., pp. 141-161). San Francisco, CA: Pearson.

Reverberi, M., Zjalic, S., Ricelli, A., Punelli, F., Camera, E., Fabbri, C., Picardo, M., Fanelli, C. and Fabbri, A., 2008. *Modulation of Antioxidant Defense in Aspergillus parasiticus Is Involved in Aflatoxin Biosynthesis: a Role for the ApyapA Gene*. Eukaryotic Cell, 7(6), pp.988-1000.

Reynolds, H., Vijayakumar, V., Gluck-Thaler, E., Korotkin, H., Matheny, P. and Slot, J., 2018. *Horizontal gene cluster transfer increased hallucinogenic mushroom diversity*. Evolution Letters, 2(2), pp.88-101.

Ricci MS, El-Deiry WS., 2003. *DNA Footprinting*. Methods in Molecular Biology. 223:117–27.

Rokas, A., Mead, M., Steenwyk, J., Raja, H. and Oberlies, N., 2020. *Biosynthetic gene clusters and the evolution of fungal chemodiversity*. Natural Product Reports, 37(7), pp.868-878.

Romero, D., Traxler, M., López, D. and Kolter, R., 2011. *Antibiotics as Signal Molecules*. Chemical Reviews, 111(9), pp.5492-5505.

Ruiz, B., Chávez, A., Forero, A., García-Huante, Y., Romero, A., Sánchez, M., Rocha, D., Sánchez, B., Rodríguez-Sanoja, R., Sánchez, S. and Langley, E., 2010. *Production of microbial secondary metabolites: Regulation by the carbon source*. Critical Reviews in Microbiology, 36(2), pp.146-167.

Ruiz, B., Chávez, A., Forero, A., García-Huante, Y., Romero, A., Sánchez, M., Rocha, D., Sánchez, B., Rodríguez-Sanoja, R., Sánchez, S. and Langley, E., 2010. *Production of microbial secondary metabolites: Regulation by the carbon source*. Critical Reviews in Microbiology, 36(2), pp.146-167.

Samuel, D., 1996. *Investigation of Ancient Egyptian Baking and Brewing Methods by Correlative Microscopy. Science,* 273(5274), pp.488-490.

Sánchez, S., Chávez, A., Forero, A., García-Huante, Y., Romero, A., Sánchez, M., Rocha, D., Sánchez, B., Ávalos, M., Guzmán-Trampe, S., Rodríguez-Sanoja, R., Langley, E. and Ruiz, B., 2010. *Carbon source regulation of antibiotic production.* The Journal of Antibiotics, 63(8), pp.442-459.

Schjerling, P. and Holmberg, S., 1996. *Comparative amino acid sequence analysis of the C6 zinc cluster family of transcriptional regulators*. Nucleic Acids Research, 24(23), pp.4599-4607.

Schrettl, M., Carberry, S., Kavanagh, K., Haas, H., Jones, G., O'Brien, J., Nolan, A., Stephens, J., Fenelon, O. and Doyle, S., 2010. S*elf-Protection against Gliotoxin—A Component of the Gliotoxin Biosynthetic Cluster, GliT, Completely Protects Aspergillus fumigatus Against Exogenous Gliotoxin*. PLoS Pathogens, 6(6), p.e1000952.

Si, J., Zhao, R. and Wu, R., 2015. *An Overview of the Prediction of Protein DNA-Binding Sites. International Journal of Molecular Sciences*, 16(12), pp.5194-5215.

Silano, V., Barat Baviera, J., Bolognesi, C., Brüschweiler, B., Cocconcelli, P., Crebelli, R., Gott, D., Grob, K., Lampi, E., Mortensen, A., Riviere, G., Steffensen, I., Tlustos, C., Van Loveren, H., Vernis, L., Zorn, H., Jany, K., Kärenlampi, S., Penninks, A., Želježic, D., Aguilera-Gómez, M., Andryszkiewicz, M., Arcella, D., Gomes, A., Kovalkovičová, N., Liu, Y., Rossi, A., Engel, K. and Chesson, A., 2018. *Safety of the food enzyme glucoamylase from a genetically modified Aspergillus niger (strain NZYM-BF*). EFSA Journal, 16(10).

Snini, S., Tannous, J., Heuillard, P., Bailly, S., Lippi, Y., Zehraoui, E., Barreau, C., Oswald, I. and Puel, O., 2015. *Patulin is a cultivar-dependent aggressiveness factor favouring the colonization of apples by Penicillium expansum*. Molecular Plant Pathology, 17(6), pp.920-930.

Song, C., Zhang, S. and Huang, H., 2015. *Choosing a suitable method for the identification of replication origins in microbial genomes.* Frontiers in MICROBIOLOGY, 6.

Spatafora, J., Chang, Y., Benny, G., Lazarus, K., Smith, M., Berbee, M., Bonito, G., Corradi, N., Grigoriev, I., Gryganskyi, A., James, T., O'Donnell, K., Roberson, R., Taylor, T., Uehling, J., Vilgalys, R., White, M. and Stajich, J., 2016. *A phylum-level phylogenetic classification of zygomycete fungi based on genome-scale data.* Mycologia, 108(5), pp.1028-1046.

Sproul, D., Gilbert, N. and Bickmore, W., 2005. *The role of chromatin structure in regulating the expression of clustered genes*. Nature Reviews Genetics, 6(10), pp.775-781.

Stensmyr, M., Dweck, H., Farhan, A., Ibba, I., Strutz, A., Mukunda, L., Linz, J., Grabe, V., Steck, K., Lavista-Llanos, S., Wicher, D., Sachse, S., Knaden, M., Becher, P., Seki, Y. and Hansson, B., 2012. A *Conserved Dedicated Olfactory Circuit for Detecting Harmful Microbes in Drosophila.* Cell, 151(6), pp.1345-1357.

Stott, W. and Bullerman, L., 1975. *Influence of Carbohydrate and Nitrogen Source on Patulin Production by Penicillium patulum*. Applied Microbiology, 30(5), pp.850-854.

Strauss, J. and Reyes-Dominguez, Y., 2011. *Regulation of secondary metabolism by chromatin structure and epigenetic codes.* Fungal Genetics and Biology, 48(1), pp.62-69.

Sullivan, M., Petty, N. and Beatson, S., 2011. *Easyfig: a genome comparison visualizer*. Bioinformatics, 27(7), pp.1009-1010.

Sweis, I. and Cressey, B., 2018. *Potential role of the common food additive manufactured citric acid in eliciting significant inflammatory reactions contributing to serious disease states: A series of four case reports.* Toxicology Reports, 5, pp.808-812.

Tannous, J., Atoui, A., El Khoury, A., Francis, Z., Oswald, I., Puel, O. and Lteif, R., 2015. *A study on the physicochemical parameters for Penicillium expansum growth and patulin production: effect of temperature, pH, and water activity.* Food Science & Nutrition, 4(4), pp.611-622.

Tannous, J., Keller, N., Atoui, A., El Khoury, A., Lteif, R., Oswald, I. and Puel, O., 2017. *Secondary metabolism in Penicillium expansum: Emphasis on recent advances in patulin research*. Critical Reviews in Food Science and Nutrition, 58(12), pp.2082-2098.

Tannous, J., Kumar, D., Sela, N., Sionov, E., Prusky, D. and Keller, N., 2018. *Fungal attack and host defence pathways unveiled in near-avirulent interactions of Penicillium expansum creA mutants on apples*. Molecular Plant Pathology, 19(12), pp.2635-2650.

Taylor J. W., J. Spatafora, K. O'Donnell, F. Lutzoni, T. James, D. S. Hibbett, D. Geiser, T.D. Bruns and M. Blackwell,.*The fungi.* In J. Cracraft and M. J. Donoghue [eds.], Assembling the Tree of Life, 171–194. Oxford University Press, Oxford, UK

Tudzynski, B., 2014. *Nitrogen regulation of fungal secondary metabolism in fungi.* Frontiers in Microbiology, 5.

van Helden, J., André, B. and Collado-Vides, J., 1998. *Extracting regulatory sites from the upstream region of yeast genes by computational analysis of oligonucleotide frequencies*. Edited by G. von Heijne. Journal of Molecular Biology, 281(5), pp.827-842.

Vashee, S., Xu, H., Johnston, S. and Kodadek, T., 1993. *How do "Zn2 cys6" proteins distinguish between similar upstream activation sites? Comparison of the DNA-binding specificity of the GAL4 protein in vitro and in vivo.* Journal of Biological Chemistry, 268(33), pp.24699-24706.

Visagie, C., Houbraken, J., Frisvad, J., Hong, S., Klaassen, C., Perrone, G., Seifert, K., Varga, J., Yaguchi, T. and Samson, R., 2014. *Identification and nomenclature of the genus Penicillium.* Studies in Mycology, 78, pp.343-371.

Walter, M., Rattei, T., Arnold, R., Guldener, U., Munsterkotter, M., Nenova, K., Kastenmuller, G., Tischler, P., Wolling, A., Volz, A., Pongratz, N., Jost, R., Mewes, H. and Frishman, D., 2009. P*EDANT covers all complete RefSeq genomes.* Nucleic Acids Research, 37(Database), pp.D408-D411.

Wang, Y., Zhou, J., Zhong, J., Luo, D., Li, Z., Yang, J., Shu, D. and Tan, H., 2018. *Cys2His2 Zinc Finger Transcription Factor BcabaR1 Positively Regulates Abscisic Acid Production in Botrytis cinerea*. Applied and Environmental Microbiology, 84(17).

Wasser, S., 2010. *Medicinal Mushroom Science: History, Current Status, Future Trends, and Unsolved Problems*. International Journal of Medicinal Mushrooms, 12(1), pp.1-16.

Wijayawardene, N.N., Hyde, K., Al-Ani, L., Tedersoo, L., Haelewaters, D., Rajeshkumar, K.C., Zhao, R., Aptroot, A., Leontyev, D., Saxena, R.K., Tokarev, Y., Dai, D., Letcher, P.M., Stephenson, S., Ertz, D., Lumbsch, H., Kukwa, M., Issi, I., Madrid, H., Phillips, A., Selbmann, L., Pfliegler, W.P., Horváth, E., Bensch, K., Kirk, P., Kolaříková, K., Raja, H., Radek, R., Papp, V., Dima, V., Ma, J., Malosso, E., Takamatsu, S., Rambold, G., Gannibal, P., Triebel, D., Gautam, A., Avasthi, S., Suetrong, S., Timdal, E., Fryar, S., Delgado, G., Réblová, M., Doilom, M., Dolatabadi, S., Pawłowska, J., Humber, R., Kodsueb, R., Sánchez-Castro, I.,

Goto, B., Silva, D.K., Souza, F.A., Oehl, F., Silva, G.A., Błaszkowski, J., Jobim, K., Maia, L.C., Barbosa, F.R., Fiuza, P.O., Divakar, P., Shenoy, B.D., Castañeda-Ruíz, R., Somrithipol, S., Lateef, A.A., Karunarathna, S., Tibpromma, S., Mortimer, P., Wanasinghe, D.N., Phookamsak, R., Xu, J., Wang, Y., Tian, F., Alvarado, P., Li, D.W., Kušan, I., Matočec, N., Mešić, A., Tkalčec, Z., Maharachchikumbura, S., Papizadeh, M., Heredia, G., Wartchow, F., Bakhshi, M., Boehm, E., Youssef, N., Hustad, V., Lawrey, J., Santiago, A.E., Bezerra, J., Souza-Motta, C., Firmino, A.L., Tian, Q., Houbraken, J., Hongsanan, S., Tanaka, K., Dissanayake, A., Monteiro, J.S., Grossart, H., Suija, A., Weerakoon, G., Etayo, J., Tsurykau, A., Vázquez, V., Mungai, P., Damm, U., Li, Q., Zhang, H., Boonmee, S., Lu, Y., Becerra, A., Kendrick, B., Brearley, F., Motiejūnaitė, J., Sharma, B., Khare, R., Gaikwad, S., Wijesundara, D., Tang, L.Z., He, M., Flakus, A., Rodriguez-Flakus, P., Zhurbenko, M., Mckenzie, E., Stadler, M., Bhat, D., Liu, J.K., Raza, M., Jeewon, R., Nassonova, E., Prieto, M., Jayalal, R., Erdoğdu, M., Yurkov, A.M., Schnittler, M., Shchepin, O., Novozhilov, Y., Silva-Filho, A.G., Gentekaki, E., Liu, P., Cavender, J.C., Kang, Y., Mohammad, S., Zhang, L.F., Xu, R., Li, Y.M., Dayarathne, M., Ekanayaka, A.H., Wen, T., Deng, C., Pereira, O.L., Navathe, S., Hawksworth, D., Fan, X., Dissanayake, L., Kuhnert, E., & Thines, M. 2020. *Outline of Fungi and fungus-like taxa.* Mycosphere, 11, 1060-1456.

Williams, D., Stone, M., Hauck, P. and Rahman, S., 1989. *Why Are Secondary Metabolites (Natural Products) Biosynthesized?*. Journal of Natural Products, 52(6), pp.1189-1208.

Wisecaver, J., Slot, J. and Rokas, A., 2014. *The Evolution of Fungal Metabolic Pathways.* PLoS Genetics, 10(12), p.e1004816.

Yu, J. and Keller, N., 2005. *Regulation of Secondary Metabolism in Filamentous Fungi.* Annual Review of Phytopathology, 43(1), pp.437-458.

Zambelli, F., Pesole, G. and Pavesi, G., 2012. *Motif discovery and transcription factor binding sites before and after the next-generation sequencing era.* Briefings in Bioinformatics, 14(2), pp.225-237.

Zhang, F., Xu, G., Geng, L., Lu, X., Yang, K., Yuan, J., Nie, X., Zhuang, Z. and Wang, S., 2016. *The Stress Response Regulator AflSkn7 Influences Morphological Development, Stress Response, and Pathogenicity in the Fungus Aspergillus flavus.* Toxins, 8(7), p.202.

Zong, Y., Li, B. and Tian, S., 2015. *Effects of carbon, nitrogen and ambient pH on patulin production and related gene expression in Penicillium expansum.* International Journal of Food Microbiology, 206, pp.102-108.