Fungal Genetics and Biology 47 (2010) 736–741

Contents lists available at ScienceDirect

Fungal Genetics and Biology

journal homepage: www.elsevier.com/locate/yfgbi

SMURF: Genomic mapping of fungal secondary metabolite clusters

Nora Khaldi [a], Fayaz T. Seifuddin [b], Geoff Turner [c], Daniel Haft [b], William C. Nierman [b,d], Kenneth H. Wolfe [a], Natalie D. Fedorova [b,*]

[a] Smurfit Institute of Genetics, Trinity College, Dublin 2, Ireland
[b] Department of Infectious Disease, The J. Craig Venter Institute, Rockville, MD, USA
[c] Department of Molecular Biology and Biotechnology, University of Sheffield, Firth Court, Western Bank, Sheffield S10 2TN, UK
[d] Department of Biochemistry and Molecular Biology, The George Washington University School of Medicine, Washington, DC, USA

ARTICLE INFO

Article history:
Received 28 April 2009
Accepted 2 June 2010
Available online 8 June 2010

Keywords:
NRPS
PKS
Prenyltransferases
Polyketides
Antibiotics
Secondary metabolism
Filamentous fungi
Aspergillus
Genome annotation

ABSTRACT

Fungi produce an impressive array of secondary metabolites (SMs) including mycotoxins, antibiotics and pharmaceuticals. The genes responsible for their biosynthesis, export, and transcriptional regulation are often found in contiguous gene clusters. To facilitate annotation of these clusters in sequenced fungal genomes, we developed the web-based software SMURF (www.jcvi.org/smurf/) to systematically predict clustered SM genes based on their genomic context and domain content. We applied SMURF to catalog putative clusters in 27 publicly available fungal genomes. Comparison with genetically characterized clusters from six fungal species showed that SMURF accurately recovered all clusters and detected additional potential clusters. Subsequent comparative analysis revealed the striking biosynthetic capacity and variability of the fungal SM pathways and the correlation between unicellularity and the absence of SMs. Further genetics studies are needed to experimentally confirm these clusters.

1. Introduction

Secondary metabolites (SMs) are small bioactive molecules produced by many organisms including bacteria, plants and fungi. These compounds are particularly abundant in soil-dwelling filamentous fungi, which exist as multicellular communities competing with each other for nutrients, minerals and water (Keller et al., 2005). Unlike primary metabolites, most SMs – as their name suggests – are not essential for fungal growth, development, or reproduction under *in vitro* conditions. They can however provide protection against various environmental stresses and during antagonistic interactions with other soil inhabitants or a eukaryotic host. Scientific appreciation of the importance of fungal SMs grew in the 1940s as the massive impact of penicillin on human health began to be seen. Since then, many other beneficial SM compounds have been discovered including immunosuppressants, cholesterol-lowering drugs, antiviral drugs, and anti-tumor drugs (for a recent review see Hoffmeister and Keller, 2007). At the same time, fungi are also known to produce numerous mycotoxins such as aflatoxin, fumonisin, trichothecene, and zearalone.

The first committed step in biosynthesis of an SM is catalyzed by one of five proteins, which we refer to here as "backbone" enzymes. They include nonribosomal peptide synthases (NRPSs), polyketide synthases (PKSs), hybrid NRPS–PKS enzymes, prenyltransferases (DMATSs), and terpene cyclases (TCs). These multidomain enzymes are associated, respectively, with production of the five classes of SM: nonribosomal peptides, polyketides, NRPS–PKS hybrids, indole alkaloids, and terpenes (Hoffmeister and Keller, 2007). Terpenes, which are composed of isoprene units, are not considered further in our analysis, because terpene cyclases are highly variable in sequence and difficult to detect by bioinformatic methods (Keller et al., 2005; Townsend, 1997). Intermediate products formed by the backbone enzymes can undergo further modifications catalyzed by "decorating" enzymes. The final product is then often steered by a transporter outside the fungal cell wall or sometimes remains within the cell. All these genes tend to be found in contiguous gene clusters, which are coordinately regulated by a specific Zn_2Cys_6 transcription factor and/or by the global regulator of secondary metabolism, putative methyltransferase LaeA (Keller and Hohn, 1997; Keller et al., 2005).

The availability of data from fungal genome sequencing projects has facilitated the discovery and characterization of new compounds and their biosynthetic pathways. Thus within months after completion of the first *A. fumigatus* genome (Nierman et al., 2005),

* Corresponding author.
E-mail address: natalief@jcvi.org (N.D. Fedorova).

1087-1845/$ - see front matter © 2010 Elsevier Inc. All rights reserved.
doi:10.1016/j.fgb.2010.06.003

several secondary metabolite clusters were characterized at the molecular level including the gliotoxin (Gardiner and Howlett, 2005), fumigaclavines (Coyle and Panaccione, 2005; Unsold and Li, 2005; Unsold and Li, 2006), fumitremorgin (Maiya et al., 2006), and siderophores (Reiber et al., 2005) biosynthesis clusters. Genome sequencing also revealed that the number of secondary metabolites characterized from a given species falls far behind the numbers of clusters that can be predicted based on its genomic sequence (Bok et al., 2006; Chiang et al., 2008). This has been attributed to the fact that not all clusters may be expressed under normal laboratory conditions.

Despite the medical and agricultural importance of fungal SMs, most putative SM clusters in fungal genomes have been predicted by *ad hoc* strategies based on manual reviews of BLAST searches generated for backbone genes and their neighbors (e.g. Nierman et al., 2005). Manual annotation of SM clusters, however, is time-consuming and may result in inconsistent annotation.

To facilitate systematic mapping of SM clusters in fungal genomes, we developed a web-based software tool, Secondary Metabolite Unknown Regions Finder (SMURF; www.jcvi.org/smurf/). It is based on three hallmarks of fungal SM biosynthetic pathways: (i) the presence of backbone genes, (ii) clustering, and (iii) characteristic protein domain content. Subsequent analyses of the predicted clusters present in 27 sequenced fungal genomes (Supplementary Table 1) shows SM gene enrichment in the genus *Aspergillus*, the absence of the clusters in unicellular fungi, and unexpected abundance and variability of the fungal clusters. Our results are also consistent with the view that SM profiles can be used as means of differentiating species and strains in filamentous fungi (Frisvad et al., 2008), and show that gene duplication plays an essential role in the creation and expansion of the SM repertoires of fungi.

2. Methods

2.1. Identification of putative backbone enzymes

SMURF relies on hidden Markov model (HMM) searches to detect backbone genes in sequenced fungal genomes. The HMMER program (http://hmmer.janelia.org) was used to search for conserved Pfam and TIGRFAM domains of backbone enzymes in the protein set of each sequenced species. Trusted threshold bit score cutoffs (predefined in HMMER) were used for each HMM search. NRPS enzymes were identified as enzymes with at least one module composed of an amino acid adenylation domain (A), a thiolation domain (PCP) and a condensation domain (C). PKS enzymes were identified as enzymes with at least one acyl transferase domain (AT), a beta-ketoacyl synthase C-terminal domain (BKS-C), and a beta-ketoacyl synthase N-terminal domain (BKS-N). Hybrid PKS–NRPS enzymes were identified as enzymes with at least one instance from each set of three domains listed above.

NRPS-like enzymes were identified with a combination of at least two domains from any of those in the NRPS enzyme module; or a combination of an A domain and a NAD_binding_4 domain; or a combination of an A domain and short chain dehydrogenase domain. PKS-like enzymes were identified with a combination of at least two domains from any of those in the PKS enzyme module. To eliminate false positives among PKS-like enzymes, they were defined as proteins with AT, BKS-C and BKS-N domains that scored below a trusted HMM cut-off. In contrast, to eliminate false positives such as alpha-aminoadipate reductase among NRPSs, we required the score of the C-terminal domain of L-aminoadipate-semialdehyde dehydrogenase alpha subunit to be above the cut-off.

Prenyltransferase enzymes were identified as enzymes with at least one DMATS-type prenyltransferase domain (DMATS). The corresponding *de novo* HMM model for this domain (TIGR03429) was created in this study from the seed alignment generated using the *A. fumigatus* dimethylallyl tryptophan synthase FtmPT2 as a seed sequence as previously described (Sonnhammer et al., 1998). Characterized or partially characterized seed members include several dimethylallyltryptophan synthases, a brevianamide F prenyltransferase, the LtxC enzyme involved in lyngbyatoxin biosynthesis, and a probable dimethylallyl tyrosine synthase.

2.2. Identification of putative decorating enzymes

To define protein domains commonly present in SM decorating enzymes, transporter, and transcriptional regulators; we examined the domains detected in the 22 *A. fumigatus* clusters we used as a training set. The list of clusters included two genetically characterized *A. fumigatus* clusters involved in biosynthesis of fumitremorgin (Grundmann et al., 2008; Kato et al., 2009; Maiya et al., 2006) and melanin (Fujii et al., 2004; Tsai et al., 1999) and 10 clusters predicted based on expression data: *A. fumigatus* clusters *Pes1*, siderophore, fumigaclavine, pseurotin, the gliotoxin-like polyketide (McDonagh et al., 2008; Perrin et al., 2007), and gliotoxin (Gardiner and Howlett, 2005). The rest of the 22 clusters were predicted manually based on genes' name and their proximity to the adjacent backbone gene (Perrin et al., 2007). Some domains were present almost exclusively in clusters, while others were evenly distributed throughout the entire genome (Supplementary Table 2). The final 27 SM-defining domains were selected as domains most likely to be found in a cluster based on their distribution.

2.3. Identification of putative SM clusters

Once all putative backbone genes are identified in a genome, the SMURF algorithm then evaluates their adjacent genes to test whether they are part of an SM gene cluster (Supplementary Fig. 1). A window of ±20 genes on each side of a backbone gene is scanned for the 27 SM-defining domains using HMMer. The number 20 was established empirically based on the training set of 22 *A. fumigatus* clusters. Genes in the window are tagged as "SM domain positive" if they contain at least one of these domains, or "SM domain negative" if they do not. Then the boundaries of any putative cluster are defined by the algorithm that evaluates each gene by walking rightwards from the backbone gene until it reaches as a stop signal, which is defined below. The last gene on the rightwards walk before the stop signal is given the label alpha. After that SMURF carries out an identical walk leftwards from the backbone gene, until a stop signal is encountered defining a left-limit gene beta. The interval between alpha and beta is the preliminary extent of the cluster.

The algorithm requires two key parameters: d, the maximum intergenic distance (in base pairs) permitted between two adjacent genes in the same cluster; and y, the maximum number of SM domain negative genes, which is allowed within a cluster. By a trial-and-error process, we identified the parameters $d = 3814$ bp and $y = 10$ genes as optimal based on the training set of 22 clusters. A stop signal is defined as either an intergenic distance that is larger than the limit d, or a cumulative number of negative genes between the backbone gene and the current position that is larger than y (Supplementary Fig. 1).

To take into account the intergenic distances, the SMURF algorithm trims each cluster to ensure that the interval between alpha and beta is less than y. Then, additional genes are trimmed at both ends of the cluster until the algorithm reaches the first backbone or SM domain positive gene on each side. In some instances, SMURF predicts overlapping clusters, in which case the two clusters are merged into one.