Motif-Based Method for the Genome-Wide Prediction of Eukaryotic Gene Clusters

Thomas Wolf, Vladimir Shelest, and Ekaterina Shelest*

Leibniz Institute for Natural Product Research and Infection Biology e. V.
Hans-Knöll-Institute (HKI),
Research group Systems Biology / Bioinformatics,
Beutenbergstrasse 11a, 07745 Jena, Germany
ekaterina.shelest@hki-jena.de

Abstract. Genomic clustering of functionally interrelated genes is not unusual in eukaryotes. In such clusters, co-localized genes are co-regulated and often belong to the same pathway. However, biochemical details are still unknown in many cases, hence computational prediction of clusters' structures is beneficial for understanding their functions. Yet, in silico detection of eukaryotic gene clusters (eGCs) remains a challenging task. We suggest a novel method for eGC detection based on consideration of cluster-specific regulatory patterns. The basic idea is to differentiate cluster from non-cluster genes by regulatory elements within their promoter sequences using the density of cluster-specific motifs' occurrences (which is higher within the cluster region) as an additional distinguishing feature. The effectiveness of the method was demonstrated by successful re-identification of functionally characterized clusters. It is also applicable to the detection of yet unknown eGCs. Additionally, the method provides valuable information about the binding sites for cluster-specific regulators.

Keywords: eukaryotic gene clusters, transcription regulation, secondary metabolites, transcription factor binding sites.

1 Introduction

Genomic clustering (co-localization) of functionally interrelated genes in conjunction with co-regulation, although less present than in prokaryotes, has been found in a great variety of eukaryotic species, from yeast to vertebrates [1,2].

The term "gene cluster" can imply various interpretations. In this work, we consider as clusters the sets of co-localized and co-regulated genes, the products of which are presumably functionally connected (e. g., they can belong to the same biochemical or signaling pathway). Thus, the co-localization and co-regulation are the main characteristics of such eGCs and they form the basis of our approach.

* Corresponding author.

A. Petrosino, L. Maddalena, P. Pala (Eds.): ICIAP 2013 Workshops, LNCS 8158, pp. 389–398, 2013.

Clusters of co-expressed genes have been found in higher eukaryotes, such as drosophila and human [3,4]. It has been shown that genes belonging to the same metabolic pathways are localized significantly closer to each other than it can be expected by chance [2]. This was demonstrated for diverse metabolic pathways from the KEGG database. A relatively well investigated class of eGCs represent the clusters of secondary metabolite genes, which are found in fungi, plants, and protists [5]. Secondary metabolites (SMs) are pharmaceutically important substances (e. g., antibiotics, antimycotics, toxins). The genes responsible for their synthesis, modifications, transport, etc., are often organized in clusters [6]. These clusters are characterized by modest sizes (normally not more than 20 genes) and tight co-localization: the genes are immediately adjacent to each other, although the insertions of non-cluster genes are also possible. The expression of SM clusters is often governed by specific regulators [7] and in many cases the specific transcription factor (TF) is embedded in the cluster [8]. Moreover, non-cluster specific (broad) TFs are also involved in the regulation of SM clusters [9] (Fig. 1).

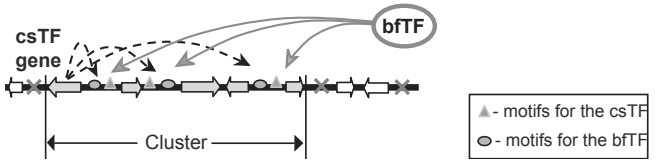

Fig. 1. Regulation of a gene cluster by cluster specific and broad function transcription factors (csTF and bfTF, correspondingly)

There are several ways to predict eGCs genome-wide. One of the first methods was suggested by Lee and Sonnhammer [2] who linked the gene annotations from KEGG with the localization information. The same approach was used in some follow-up works, e. g., in [10], where the authors suggested a method to identify all possible clusters of genes annotated to the same GO term. These methods predict any clusters regardless to their functions and specific features. In the particular cases, like SM clusters, such approaches will give imprecise predictions, mostly because the assignments of genes to pathways are partly or completely unknown.

Another group of cluster-detection methods relies on expression data (microarrays, etc.) [11]. These methods are reliable as long as the data is good, as they provide relatively solid evidence for co-regulation. However, many eGCs, for instance, in fungal genomes are silent under laboratory conditions [6] and it is challenging to experimentally determine the conditions for the cluster induction. Thus, the application of such methods to cryptic clusters is limited.

Some methods have been developed specifically for particular cluster types, e. g., for the SM clusters. Most of the methods developed so far for the detection of SM clusters are similarity based [12,13,14]. Due to the limited number of known clusters that can serve as a template, and also to the possible incorrect assignments of genes to clusters, similarity based methods are error-prone

and tend to overestimate the clusters' lengths. Additionally, these methods do not differentiate closely located (adjacent) eGCs, interpreting them as a single cluster.

These limitations could be circumvented by consideration of sequence characteristics of the cluster regions: GC content and averaged DNA curvative profile [15]. However, not all clusters are characterized by a conserved curvative pattern, which means that a substantial part of them would be skipped by the method if applied to a genome-wide search.

We suggest a novel approach to predict gene clusters based on the density of transcription factor binding site (TFBS) occurrences. In contrast to related tools, our method is not similarity-based. The main idea is that the cluster-specific TFBSs should be enriched in the cluster in comparison to other parts of the genome. Yet, their occurrence outside the cluster is not excluded. We characterize promoters by cluster-specific motif occurrences and consider the density of the motifs as the main feature of the cluster region. The method is applicable to any clusters of co-regulated genes. We demonstrate its usefulness on the example of SM clusters.

2 Results

The presumable co-regulation of the cluster genes presupposes that their promoters share at least one common motif to bind the regulating TF. Ideally, this common motif should be specific to the cluster but not to the surrounding genes (since they are not co-regulated). As the cluster-specific TF (csTF) is assumed not to have ubiquitous functions, its TFBSs should not be widely distributed across the genome. On the other hand, the cluster genes are not necessarily adjacent; "alien" genes inside the cluster may occur (e. g., in [11]). Thus, our requirements for the cluster genes are the following: (i) genes are co-localized; (ii) promoters share at least one common motif; (iii) there can be "gap" genes that do not share the common motif with the rest of the cluster. These requirements allow us to formulate the algorithm to find clusters in a genomic sequence. We call our approach the motif density method (MDM).

2.1 Motif Density Method

The basic idea of the method is that the binding sites for csTFs are enriched in the region of the cluster. Note that we do not exclude their occurrence outside the cluster. Most important, the cluster-specific motifs should be observed in consecutive promoters.

To start the cluster predictions, we need to specify the so-called "anchor" genes. These can be the genes that are already assigned to the pathway in question. In the case of the SM clusters, polyketide synthases (PKSs) or non-ribosomal peptide synthetases (NRPSs) can serve as the anchor genes. PKSs and NRPSs are characterized by a specific set of domains and large size, which makes them relatively easy to detect in genomes.

Step 1: Motif Search. On the first step, all anchor genes are searched and marked in the genomes. Next, an interim set of genes around the anchor gene of interest is selected. Since we do not know how the anchor gene is located relative to the presumable cluster (in the middle or on the edge), we consider several gene sets around the anchor gene not to miss the correct motif: 4/6/8 genes upstream, 4/6/8 genes downstream, and 2 genes up- and downstream the anchor gene. The common motifs are predicted by MEME [16] in the corresponding promoter sequences (−1000/+50 bp around the transcription start site or the whole intergenic region if it is shorter than 1000 bp). Occurrence in the anchor gene promoter is the prerequisite for the further consideration. The best-scoring motif (the one with the lowest score as defined by MEME) out of all considered promoter sets is then searched in all promoter sequences genome-wide.

Step 2: Transforming the Genomic Sequence into the Sequence of Promoters. Counting Occurrences in Frames. On this step, we switch to consideration of promoters as units characterized by the number of occurrences of a particular motif. The order of units follows the order of the corresponding promoters in the genomic sequence. Now instead of the real genomic sequence we consider a string of numbers, which represent the motifs' occurrences in a unit. For instance, if 1 motif was found in the first promoter, 2 motifs in the second, and 0 in the third and fourth promoters, the string will be 1-2-0-0. This number string is scanned by a sliding window (frame) with the step of one unit counting the cumulative number of found motifs per frame. The highest number of occurrences per frame should be obtained for the window coinciding with the cluster. Consideration of different frame lengths allows us to determine the real cluster length.

Step 3: Scoring. To select the optimal frame we apply a scoring system. As the "gap" genes are allowed in the cluster, we allow gaps ("empty" promoters) in the frames but introduce a gap penalty. In this way, we do not forbid the occurrence of small gaps, which are indeed common in clusters, but larger gaps are scored with a penalty that is growing depending on the gap length. The promoters with motifs, on the contrary, add a positive value to the score depending on the number of motifs found.

Let us consider a frame with the length l. In this frame, each promoter i is characterized by the number of found motifs m_i. The consecutive promoters without motifs ($m_i = 0$) form a gap, which is characterized by its length d, the number of gaps in the frame being n.

Then the score S of a frame is calculated as:

$$S = \sum_{i=1}^{l} m_i - \sum_{j=1}^{n} P^{d_j} \quad , \tag{1}$$

where P is the gap penalty and is an adjustable parameter.

The scores are calculated for different frame lengths (normally from 3 to 30, because this is the usual size of the known clusters).

Step 4: Visualization and Selection of the Optimal Frame. The frames are characterized by their score, position, and length. To visualize all characteristics at once, we apply the heat maps (Fig. 2).

2.2 Effectiveness of the Approach

To demonstrate the effectiveness of MDM, we applied it to the re-identification of several functionally characterized SM clusters with known borders. We selected two clusters with characterized regulatory patterns (TFBSs) in order to see if our motif predictions match the real motifs. These chosen examples are the aflatoxin cluster in *Aspergillus flavus* and violaceol cluster in *Aspergillus nidulans*. The latter was also of special interest because it is located in close vicinity to another eGC (orsellinic acid cluster). It was tempting to see if our method is able to separate the two clusters.

The other examples are clusters with characterized products and different patterns of regulation. For instance, the asperfuranone is subject to inter-cluster cross-talk (see Discussion for more details). For all clusters, we compared the predictions of MDM to those of the SMURF tool (Table 1). The gap penalty was set to 1.3 for all examples.

Aflatoxin Cluster in *A. flavus*. Aflatoxin is produced by different Aspergilli [17] and its production is regulated by the csTF AflR, along with several broad function TFs (depending on conditions). The binding sites for AflR have the consensus sequence TCG(N_5)CGA. In *A. flavus*, the cluster spans 21 genes with 15 promoters (AFL2G_07210 to AFL2G_07230). The analysis was run on the genomic sequence from the Broad Institute website [18].

The motif search was performed from scratch in order to confirm the ability of the algorithm to re-identify the real (known) motif. Seven interim sets of promoters around the anchor gene ALF2G_07228 *pksA* (in different arrangements) were submitted to MEME for motif prediction. For each set we could get

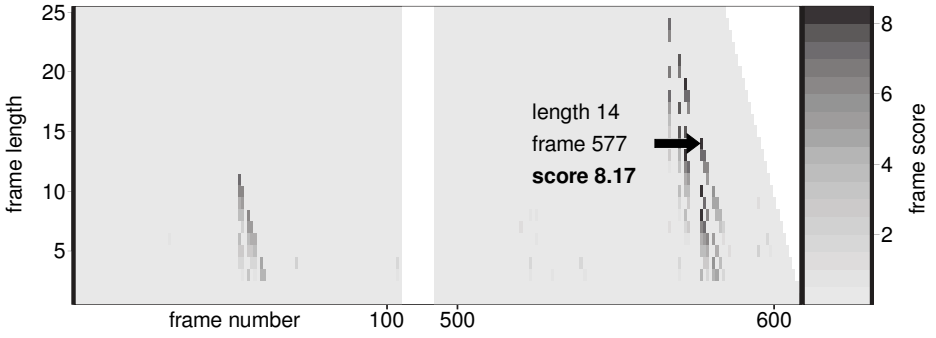

Fig. 2. Heat map for the re-identified aflatoxin cluster (right) and the sub-cluster (left), both on contig 7

common motifs. The motif in the set "8 promoters upstream *pksA*" scored the best and thus was submitted to the genome-wide search. Remarkably, this motif coincided with the AflR TFBS. The AflR motif correctly identified the cluster region with high precision: from AFL2G_07211 to AFL2G_07230 (Fig. 2). The predictions made by the other (non-AflR-like) motifs were much more noisy and failed to detect the cluster.

Violaceol Cluster in *A. nidulans*. The violaceol cluster was described recently [19] and its regulation is yet not well investigated. However, the potential binding sites for the cluster specific regulator were proposed in [19]. MDM was applied to the re-identification of this cluster in the same way as to the aflatoxin cluster, starting with the motif prediction from scratch. The genomic sequence was downloaded from Aspergillus genome database [20]. MDM successfully detected the correct motif (CYCGGAGWWWC) and the correct cluster location (Fig. 3). The length of the cluster is two genes longer than the reported one due to the high number of the csTFBSs in the promoters (Fig. 3). We return to this in the Discussion section. As expected, the orsellinic acid cluster, which is located only five genes apart from the violaceol cluster and which is not regulated by the violaceol csTF, was not detected. In this way, we show the specificity of MDM and its ability to separate closely located clusters.

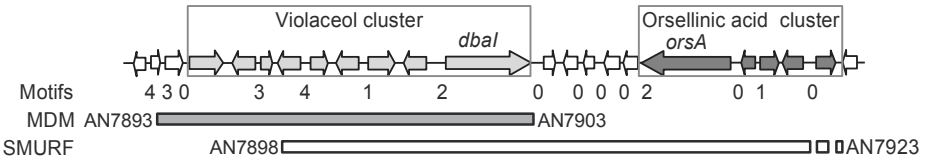

Fig. 3. Re-identification of the violaceol cluster with MDM and SMURF. Coordinates of the real cluster: AN7896 to AN7903.

Asperfuranone Cluster in *A. nidulans*. The regulation of the asperfuranone cluster is a particular case, because the asperfuranone csTF (AfoA) is subject to the regulation by ScpR, the regulator of the NRPS-containing gene cluster *inp*. Under inducing conditions, ScpR triggers AfoA, which in turn induces the expression of the asperfuranone cluster genes (except for AN1031 *afoB*) [21]. Therefore, the *afoA* promoter contains the motif for the ScpR binding [21], whereas the other cluster genes should contain another, not yet described TFBS for AfoA. By the application of MDM we re-identified the cluster nearly perfectly, with expected missing of the *afoA* and *afoB* genes (see also in Discussion).

Aspyridon Cluster in *A. nidulans*, Gliotoxin Cluster in *A. fumigatus*, and WYK-1 cluster in *A. oryzae*. We applied MDM to the re-identification of three more clusters. In all three cases we detected the clusters, although

Table 1. Comparision of SM gene cluster predictions beetween SMURF and MDM

Method	Cluster Start	End	Reference
	Aflatoxin *(Aspergillus flavus)*		
Experimental	AFL2G_07210	AFL2G_07230	[17]
MDM	AFL2G_07211	AFL2G_07230	
SMURF	AFL2G_07219	AFL2G_07248	
	Asperfuranone *(Aspergillus nidulans)*		
Experimental	AN1029	AN1036	[21]
MDM	AN1032	AN1036	
SMURF	AN1029	AN11288[1]	
	Aspyridon *(Aspergillus nidulans)*		
Experimental	AN8408	AN8415	[22]
MDM	AN8401	AN8421	
SMURF	AN8415	AN9243	
	Gliotoxin *(Aspergillus fumigatus)*		
Experimental	AFU6G_09630	AFU6G_09745	[23]
MDM	AFU6G_09630	AFU6G_09785[2]	
SMURF	AFU6G_09580	AFU6G_09740	
	Violaceol *(Aspergillus nidulans)*		
Experimental	AN7896	AN7903	[19]
MDM	AN7893	AN7903	
SMURF	AN7898	AN7923	
	WYK-1 *(Aspergillus oryzae)*		
Experimental	AO090001000009	AO090001000019	[24]
MDM	AO090001000009	AO090001000018	
SMURF	AO090001000009	AO090001000031	

not ideally. The results are presented in Table 1 and discussed in detail in the Discussion section.

3 Discussion

Computational prediction of eukaryotic clusters is especially important when precise information about the corresponding pathways is missing. In such cases, the predicted cluster's structure can point at the involvement of particular enzymes in the pathway and thus be beneficial for the understanding of the pathway's functioning.

Neither of the so far published tools has used the promoter information for the cluster prediction. Since the co-regulation is the basic idea of the cluster

[1] AN11288 is located 2 genes upstream AN1036.
[2] AFU6G_09785 is located 4 genes upstream AFU6G_09745.

definition, we consider the neglect of the promoter information as an oversight. We developed an approach that not only allows to reliably predict the eGCs but also provides information about the potential regulators of the cluster (by description of their TFBSs).

We compared the performance of our method with that of SMURF, the most prominent similarity based approach to SM cluster predictions. SMURF fails to detect the correct borders for most of the clusters and mixes the violaceol cluster with the orsellinic acid cluster reporting them as a single eGC (Fig. 3). MDM gives better or comparable predictions for all examined eGCs and solves the problem of the two adjacent clusters. In the aflatoxin cluster prediction, only one gene of 21 (AFL2G_07210) is missing because the bidirectional promoter between AFL2G_07210 and AFL2G_07209 does not contain the AflR TFBS. This may be reasonable, as AFL2G_07209 does not belong to the cluster and AFL2G_07210 has no assigned cluster function [17]. In the violaceol cluster, two promoters upstream the cluster also shared the specific motif. This does not contradict the experimental data, as the corresponding genes show slight expression under cluster-inducing conditions [19]. In fact, their involvement in the cluster under some specific conditions is not excluded and the function of the csTFBSs deserves additional examination. It remains problematic how to predict clusters with such mosaic regulation. We aim to address this problem in the next versions of MDM.

As mentioned above, the asperfuranone cluster is an interesting case, because its regulator AfoA is induced by a csTF of another cluster. AfoA is shown to induce all cluster genes except for *afoB* [21]. Our findings confirm this experimental result, since the promoter of *afoB* apparently does not contain the AfoA binding motif.

The prediction of the aspyridon cluster by MDM is not perfect, however, it covers the whole cluster, although adding several extra genes up- and downstream of it. Given that SMURF does not find the cluster at all, we consider this result rather good. For the gliotoxin cluster, the left border is found perfectly but on the right side MDM predicts four more genes as cluster members. In such cases (when the promoters have a potential TFBS for a cluster-specific regulator) we cannot exclude a possibility that the cluster is actually longer and those genes can be expressed under some specific conditions. This could be a subject of further experimental investigation. The MDM prediction of the WYK-1 cluster is missing one gene. However, compared to the SMURF result (12 genes more) the prediction of the MDM is closer to the real cluster borders.

The results of the re-identification of the known clusters show that there is space for the improvement of our approach. In many cases, MDM predictions are not perfect. Yet, in the great majority they are better than those made by the similarity-based method, which underscores the higher potential of the motif-based approach.

The genome-wide detection of the csTFBSs can help to discover other genes and even additional clusters regulated by the csTF. Regulatory cross-talk between the clusters has already been described in fungi [21]. In our examples, we

could detect a second peak on the heat map for the AflR motif (Fig. 2). The peak corresponds to a frame in a distant location on the same contig. There is no SM synthase gene in this cluster-like stretch, however, the genes are typical for SM clusters (monooxygenases, methyltransferase, MFS transporters, etc.). There can be two explanations for that: either this is a sub-cluster that is in some way involved in the aflatoxin biosynthetic pathway, or these are the remainings of a damaged cluster that has lost the synthase. In any case, this intriguing sub-cluster deserves further investigation.

To our knowledge, MDM is the first attempt to consider the promoter information in the eGC prediction. We show the high potential of this approach on the examples of the SM clusters, however, the method can be applied to the detection of any eGCs analogous to the SM clusters.

Acknowledgements. This work was financially supported by the Pakt für Wissenschaft und Forschung (2009-2012) and by the International Leibniz Research School for Microbial and Molecular Interactions (ILRS), as part of the excellence graduate school Jena School for Microbial Communication (JSMC), supported by the Deutsche Forschungsgemeinschaft.

References

1. Blumenthal, T.: Gene clusters and polycistronic transcription in eukaryotes. Bioessays 20, 480–487 (1998)
2. Lee, J.M., Sonnhammer, E.L.: Genomic gene clustering analysis of pathways in eukaryotes. Genome Res. 13, 875–882 (2003)
3. Caron, H., van Schaik, B., van der Mee, M., Baas, F., Riggins, G., van Sluis, P., Hermus, M.C., van Asperen, R., Boon, K., Voute, P.A., Heiqsterkamp, S., van Kampen, A., Versteeg, R.: The human transcriptome map: Clustering of highly expressed genes in chromosomal domains. Science 291, 1289–1292 (2001)
4. Spellman, P.T., Rubin, G.M.: Evidence for large domains of similarly expressed genes in the Drosophila genome. J. Biol. 1, 5 (2002)
5. Sasso, S., Pohnert, G., Lohr, M., Mittag, M., Hertweck, C.: Microalgae in the postgenomic era: a blooming reservoir for new natural products. FEMS Microbiol. Rev. 36, 761–785 (2012)
6. Brakhage, A.A., Schroeckh, V.: Fungal secondary metabolites - strategies to activate silent gene clusters. Fungal Genet. Biol. 48, 15–22 (2011)
7. Keller, N.P., Hohn, T.M.: Metabolic Pathway Gene Clusters in Filamentous Fungi. Fungal Genet. Biol. 21, 17–29 (1997)
8. Brakhage, A.A.: Regulation of fungal secondary metabolism. Nat. Rev. Microbiol. 11, 21–32 (2013)
9. Hoffmeister, D., Keller, N.P.: Natural products of filamentous fungi: enzymes, genes, and their regulation. Nat. Prod. Rep. 24, 393–416 (2007)
10. Yi, G., Sze, S.H., Thon, M.R.: Identifying clusters of functionally related genes in genomes. Bioinformatics 23, 1053–1060 (2007)
11. Schroeckh, V., Scherlach, K., Nützmann, H.W., Shelest, E., Schmidt-Heck, W., Schuemann, J., Martin, K., Hertweck, C., Brakhage, A.A.: Intimate bacterial-fungal interaction triggers biosynthesis of archetypal polyketides in *Aspergillus nidulans*. Proc. Natl. Acad. Sci. USA 106, 14558–14563 (2009)

12. Khaldi, N., Seifuddin, F.T., Turner, G., Haft, D., Nierman, W.C., Wolfe, K.H., Fedorova, N.D.: SMURF: Genomic mapping of fungal secondary metabolite clusters. Fungal Genet. Biol. 47, 736–741 (2010)
13. Medema, M.H., Blin, K., Cimermancic, P., de Jager, V., Zakrzewski, P., Fischbach, M.A., Weber, T., Takano, E., Breitling, R.: antiSMASH: rapid identification, annotation and analysis of secondary metabolite biosynthesis gene clusters in bacterial and fungal genome sequences. Nucleic Acids Res 39, W339–W346 (2011)
14. Fedorova, N.D., Moktali, V., Medema, M.H.: Bioinformatics approaches and software for detection of secondary metabolic gene clusters. Methods Mol. Biol. 944, 23–45 (2012)
15. Do, J.H., Miyano, S., The, G.C.: window-averaged DNA curvature profile of secondary metabolite gene cluster in *Aspergillus fumigatus* genome. Appl. Microbiol. Biotechnol. 80, 841–847 (2008)
16. Bailey, T.L., Boden, M., Buske, F.A., Frith, M., Grant, C.E., Clementi, L., Ren, J., Li, W.W., Noble, W.S.: MEME SUITE: tools for motif discovery and searching. Nucleic Acids Res. 37, W202–W208 (2009)
17. Amaike, S., Keller, N.P.: Aspergillus flavus. Annu. Rev. Phytopathol. 49, 107–133 (2011)
18. Aspergillus Comparative Sequencing Project, Broad Institute of Harvard and MIT, `http://www.broadinstitute.org/`
19. Gerke, J., Bayram, O., Feussner, K., Landesfeind, M., Shelest, E., Feussner, I., Braus, G.H.: Breaking the silence: protein stabilization uncovers silenced biosynthetic gene clusters in the fungus *Aspergillus nidulans*. Appl. Environ. Microbiol. 78, 8234–8244 (2012)
20. Arnaud, M.B., Chibucos, M.C., Costanzo, M.C., Crabtree, J., Inglis, D.O., Lotia, A., Orvis, J., Shah, P., Skrzypek, M.S., Binkley, G., Miyasato, S.R., Wortman, J.R., Sherlock, G.: The Aspergillus Genome Database, a curated comparative genomics resource for gene, protein and sequence information for the Aspergillus research community. Nucleic Acids Res. 38, D420–D427 (2010)
21. Bergmann, S., Funk, A.N., Scherlach, K., Schroeckh, V., Shelest, E., Horn, U., Hertweck, C., Brakhage, A.A.: Activation of a silent fungal polyketide biosynthesis pathway through regulatory cross talk with a cryptic nonribosomal peptide synthetase gene cluster. Appl. Environ. Microbiol. 76, 8143–8149 (2010)
22. Bergmann, S., Schümann, J., Scherlach, K., Lange, C., Brakhage, A., Hertweck, C.: Genomics driven discovery of PKS-NRPS hybrid metabolites from *Aspergillus nidulans*. Nat. Chem. Biol. 3, 213–217 (2007)
23. Gardiner, D.M., Howlett, B.: Bioinformatic and expression analysis of the putative gliotoxin biosynthetic gene cluster of *Aspergillus fumigatus*. FEMS Microbiol. Lett. 248, 241–248 (2005)
24. Imamura, K., Tsuyama, Y., Hirata, T., Shiraishi, S., Sakamoto, K., Yamada, O., Akita, O., Shimoi, H.: Identification of a Gene Involved in the Synthesis of a Dipeptidyl Peptidase IV Inhibitor in *Aspergillus oryzae*. Appl. Environ. Microbiol. 78, 6996–7002 (2012)