German Centre for
Integrative Biodiversity Research (iDiv)
Halle-Jena-Leipzig

Promoter-based prediction of gene clusters in eukaryotic genomes

Ekaterina Shelest

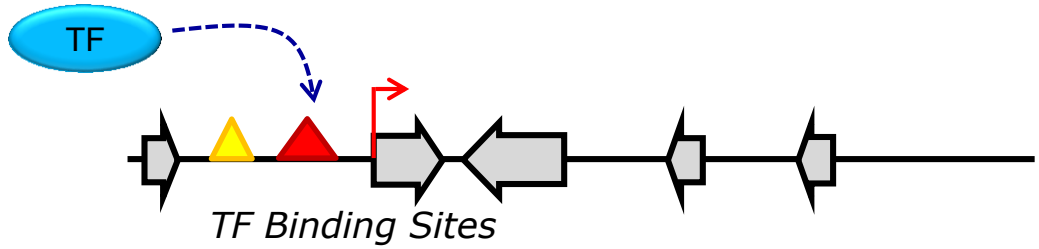
09.03.2018 Göttingen

iDiv is a research centre of the DFG Deutsche Forschungsgemeinschaft

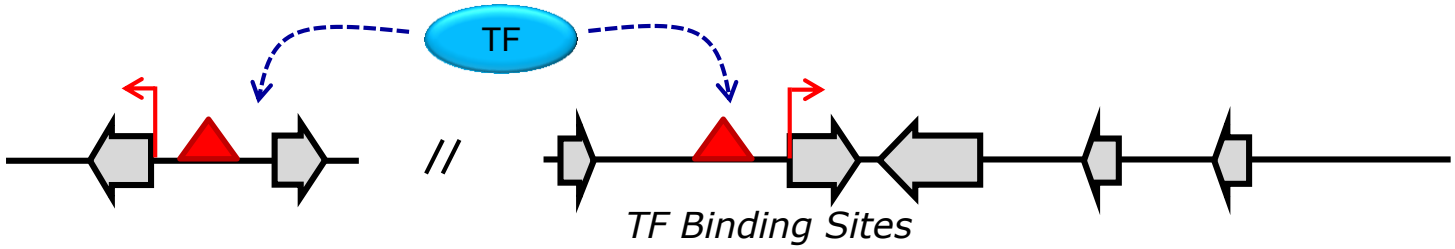
www.idiv.de

Part 1.

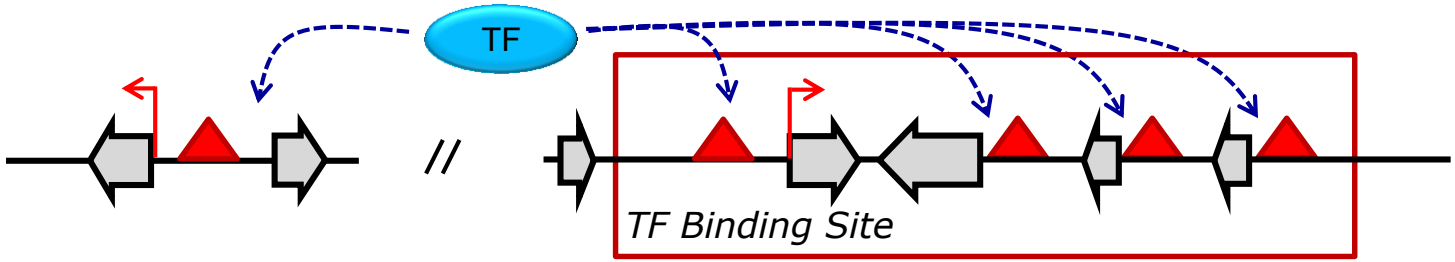
Gene clusters and their discovery


TF
TF Binding Sites

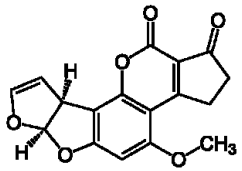
From promoter models to gene clusters



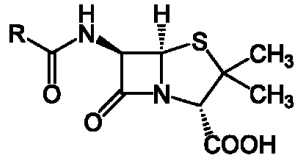
Co-regulated genes



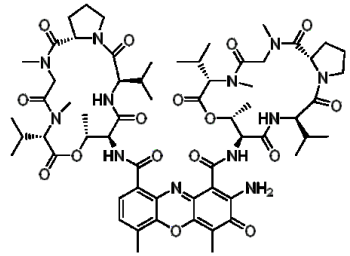
Co-regulated genes and co-localized genes: ***Gene cluster***



Aflatoxin, one of the most potent carcenogens

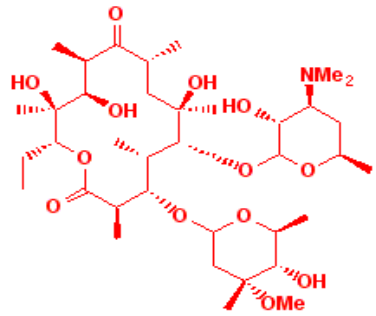


penicillin



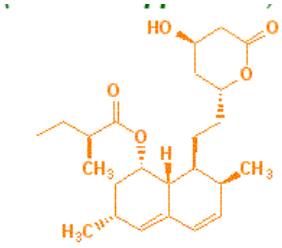
Actinomycin D

Non-ribosomal peptide

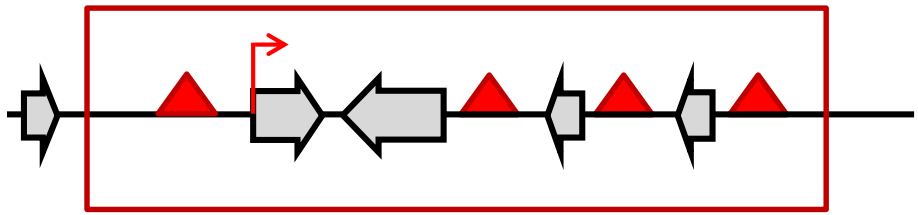


erythromycin A (antibacterial)

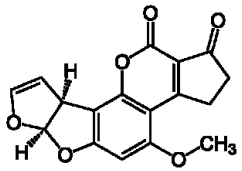
Polyketide



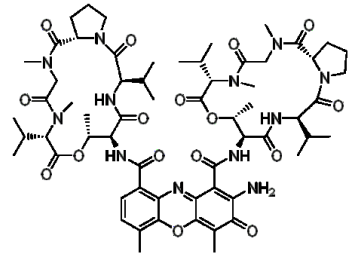
lovastatin (anticholesterol)



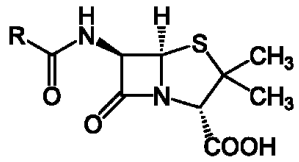
Secondary metabolite gene clusters



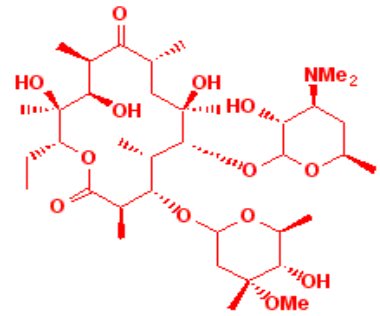
Aflatoxin, one of the most potent carcenogens



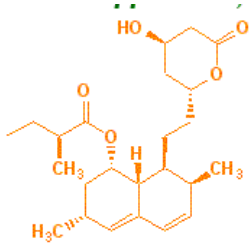
Actinomycin D



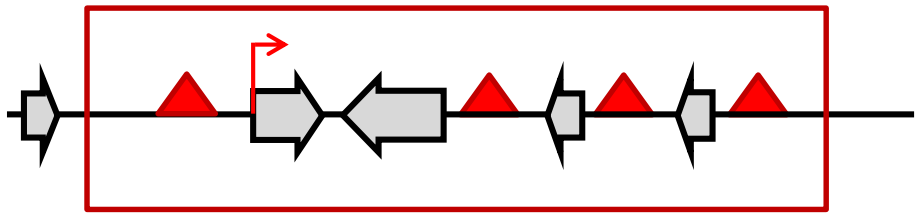
penicillin



erythromycin A (antibacterial)



lovastatin (anticholesterol)



Secondary metabolite gene clusters

Fungi:

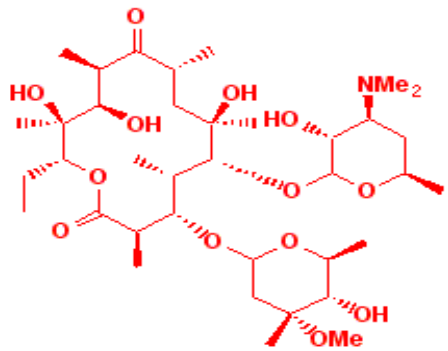
- Hundreds of substances described but molecular (genetic) basis is unknown
- Filamentous fungi: on average ~ 40 clusters per genome, most with unknown products

There are many classes of compounds that are classified as SMs:

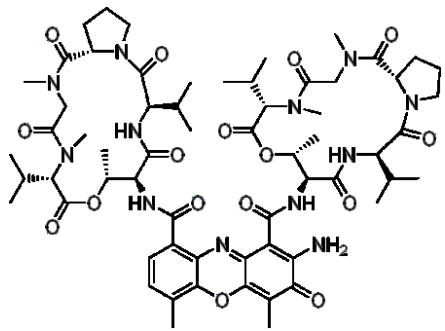
- Polyketides
- Non-ribosomal peptides
- Ribosomally synthesized and post-translationally modified peptides
- Terpenoids
- Alcaloids,
- Etc.

There are many classes of compounds that are classified as SMs:

- **Polyketides** ← OF INTEREST
- **Non-ribosomal peptides** ← OF INTEREST
- Ribosomally synthesized and post-translationally modified peptides
- Terpenoids
- Alcaloids,
- Etc.



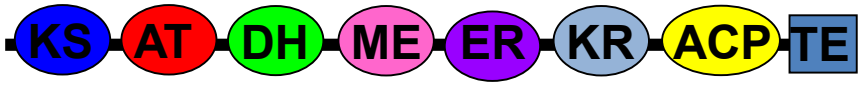
erythromycin A
Polyketide



Actinomycin D
Non-ribosomal peptide

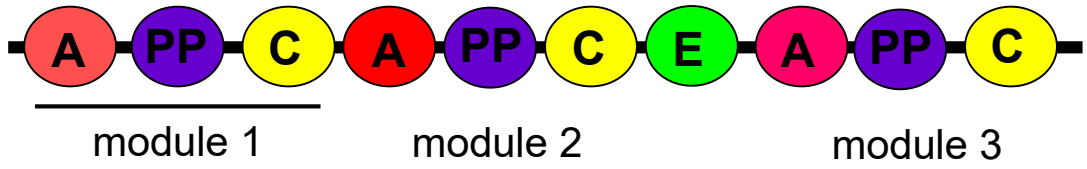
Multi-domain megasynthases

Polyketide synthase (PKS)



KS, Ketosynthase domain;
AT, acetyltransferase domain;
ACP (PP), acyl carrier protein;
KR, ketoacyl reductase domain;
ER, enoyl reductase domain;
DH, dehydratase domain;
ME, methyltransferase domain;
TE, thiolesterase.

Non-ribosomal peptide synthetase (NRPS)



A, adenylation domain;
T (PP), thiolation or peptidyl carrier domain (with a swinging phosphopantetheine group);
C, condensation domain;
E, epimerization domain;
T, thioesterase domain.

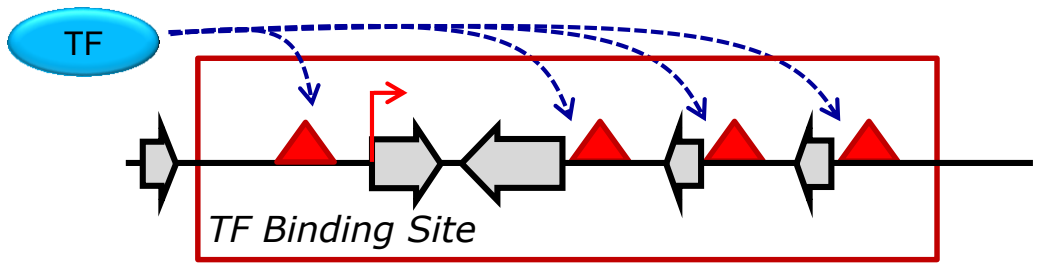
Large size *and* ***typical set of domains =>*** *easy detection in genomes!*

Problems with detection and prediction of (SM) clusters

1. No unambigious definition
2. Pathways (and products) are mostly unknown, so it is hard to predict the set of genes involved in a cluster.
3. Most of clusters are silent under laboratory conditions.
4. Clusters are not necessarily conserved.
5. There are no marker genes except for synthases (PKSs, NRPSs, etc.). Some genes (P450, transporters, transcription factors) are often but not always found in clusters.

What to rely on?

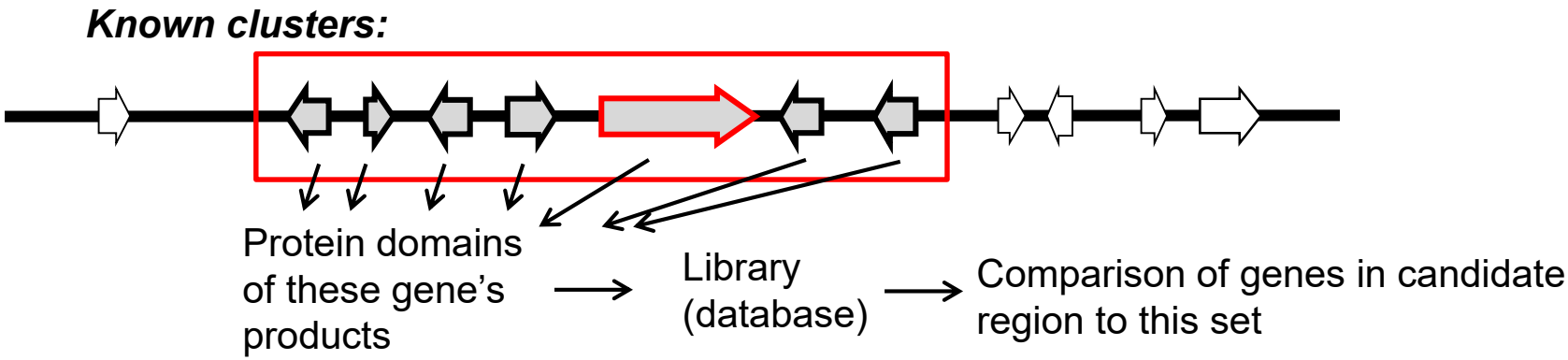
- either genes/proteins or regulation



Methods developed so far are based on:

- Gene / protein annotation
- Protein similarity (antiSMASH, SMURF, etc.)
- Expression data (Andersen et al, PNAS 2013)

Protein similarity-based methods (antiSMASH, SMURF, etc.)

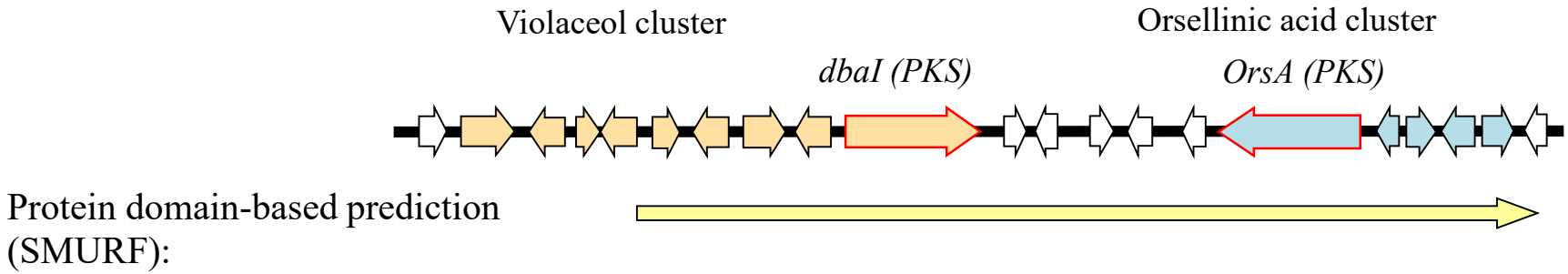


BUT:

- there are no marker genes except the anchors;
- many products and pathways (hence genes) are unknown

Issues with protein-based tools:

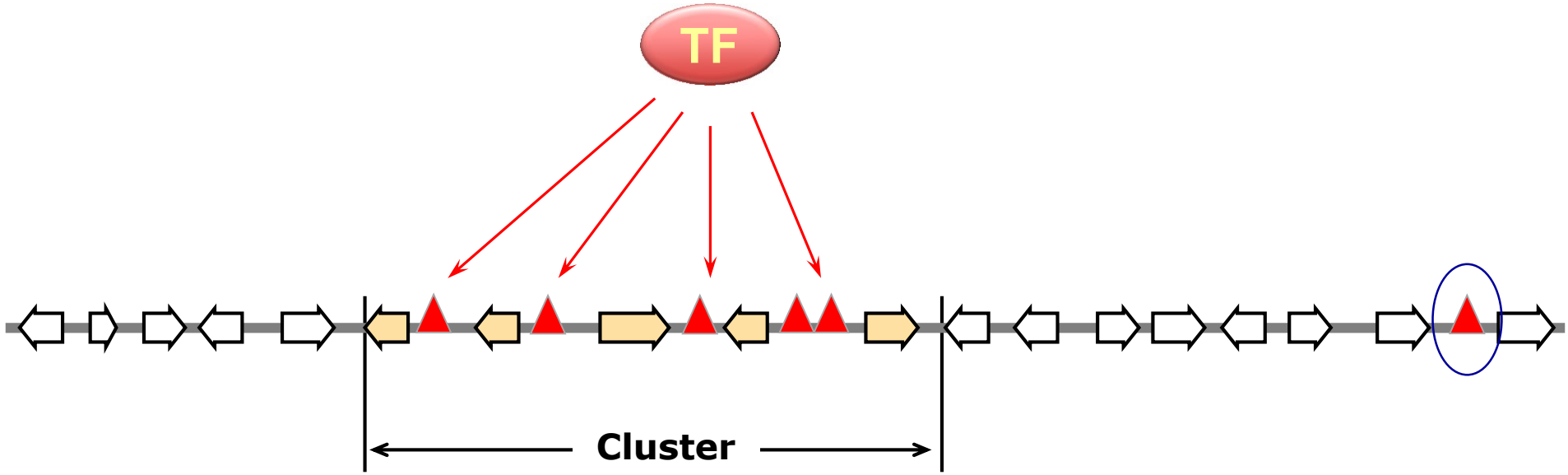
- Over-estimation of cluster lengths
- Prediction of "alien" genes as cluster genes
- No way to differentiate closely located clusters



No methods based on regulation information!

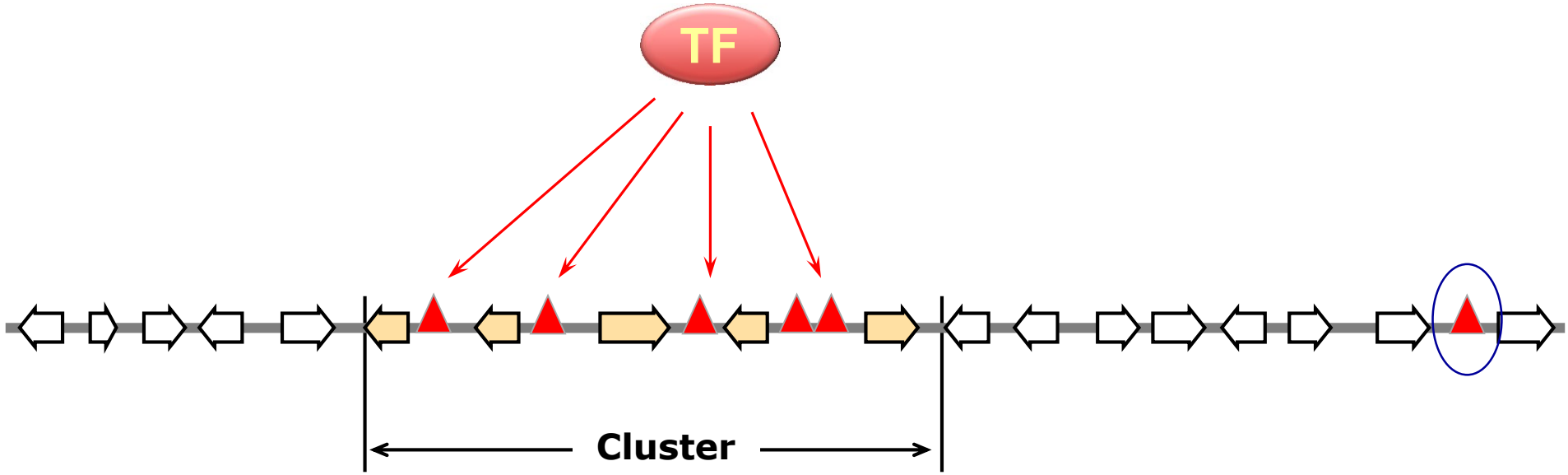
Definition:

Cluster definition: **Co-regulated** and co-localized genes



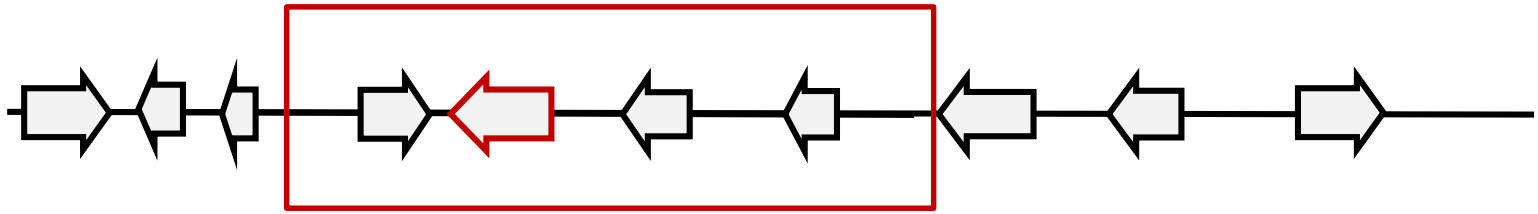
Basic idea:

To detect co-localized shared motifs (TFBSs) in the vicinity of the main biosynthetic enzymes (PKSs and NRPSs)

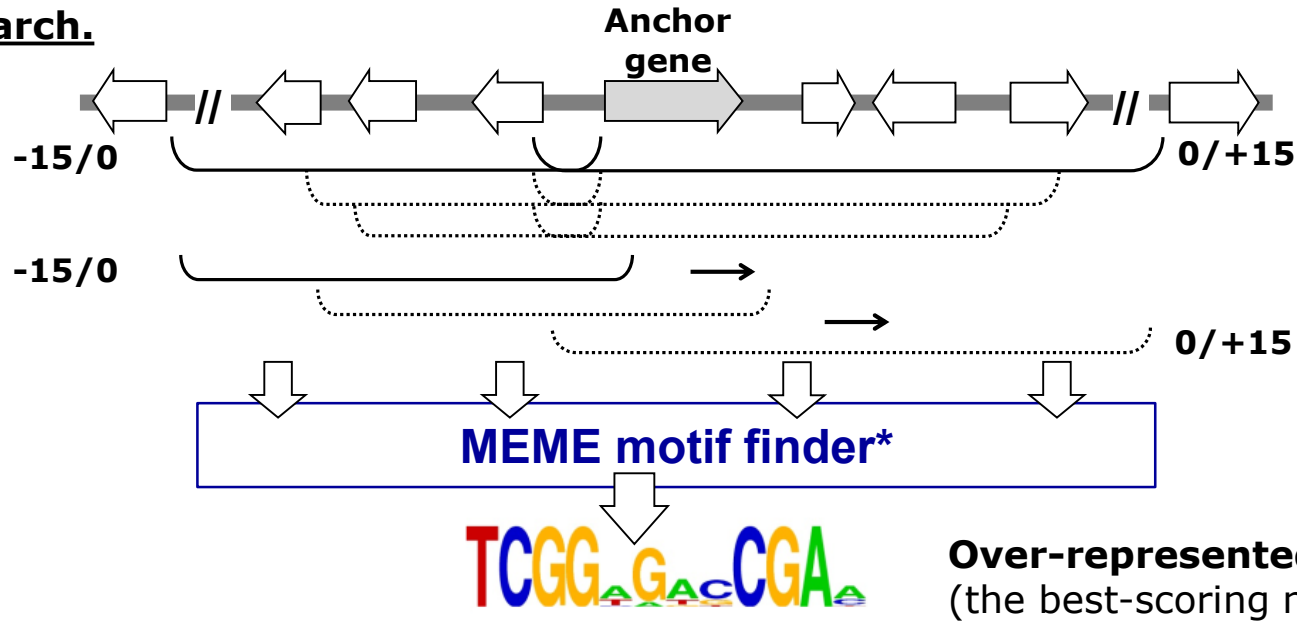


Promoter-based method for gene cluster prediction:

CASSIS – Cluster ASSociation by Islands of Sites



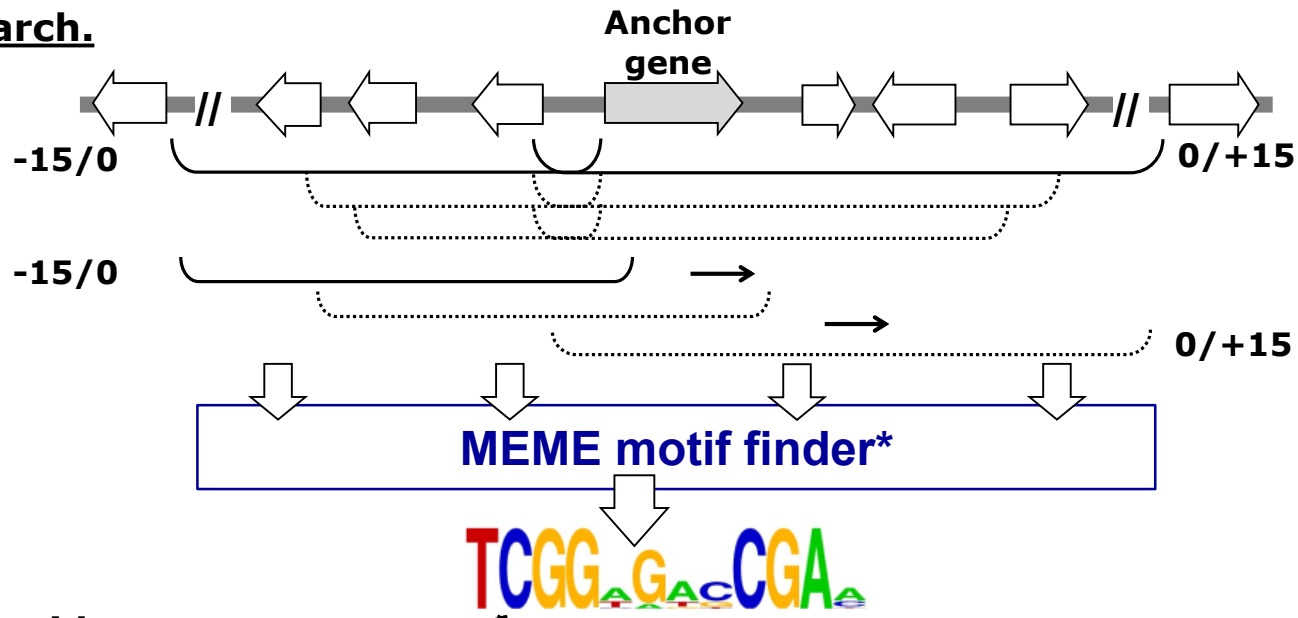
Step 1: Motif search.



interim sets of promoters

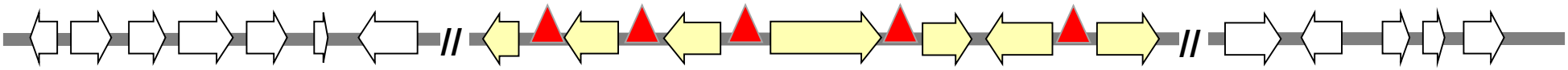
Over-represented motifs
(the best-scoring motif for each frame)

Step 1: Motif search.

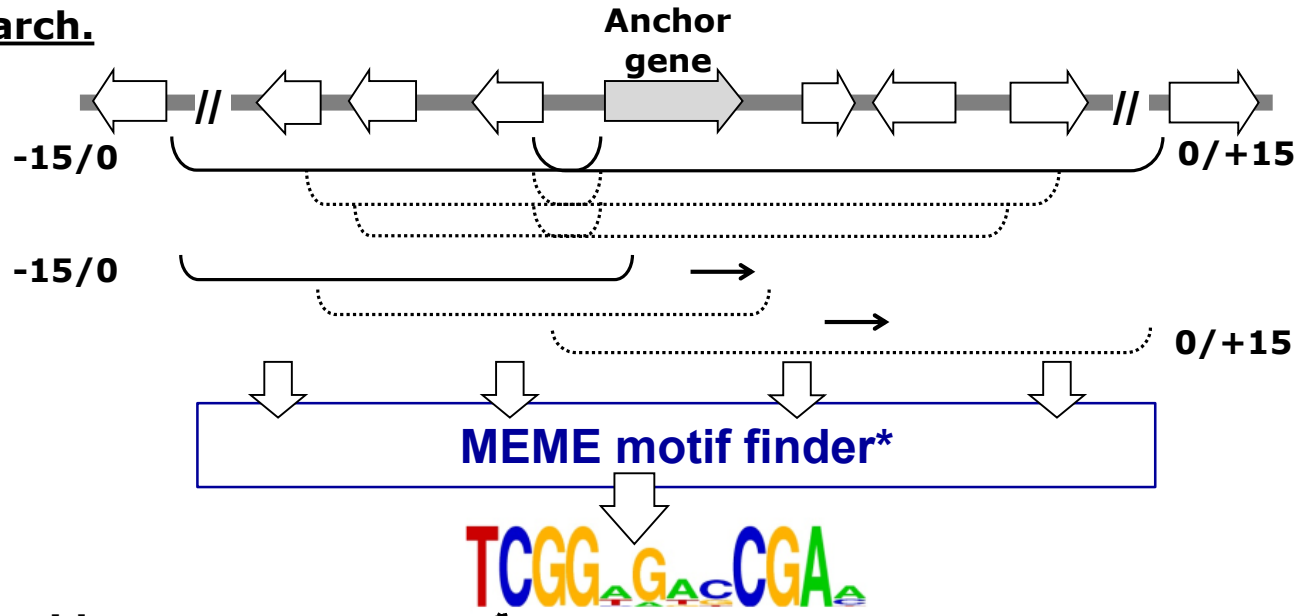


interim sets of promoters

Step 2: Genome-wide motif search

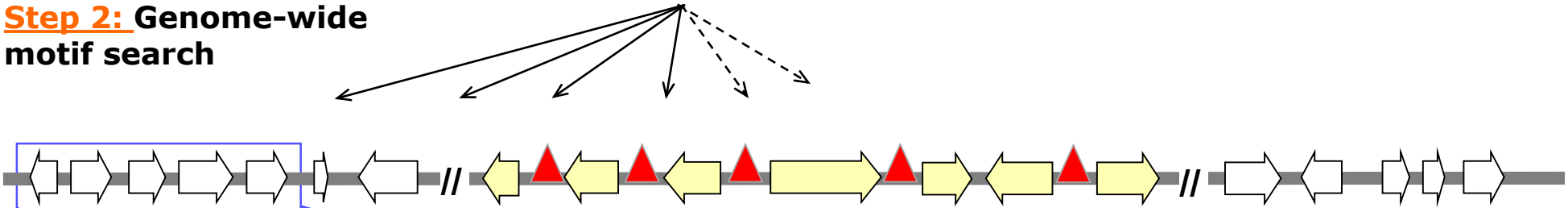


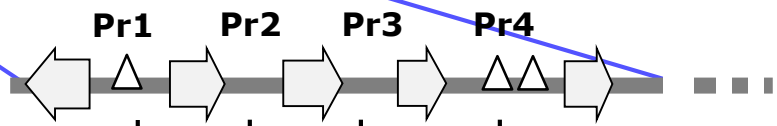
Step 1: Motif search.



interim sets of promoters

Step 2: Genome-wide motif search





Step 3: Transforming genomic sequence into a number string

1 0 0 Registering found motifs

1 0 0 1 0 0 0 0 1 0 0 0 … 0 0 **1 1 1 1 0 1 1 1** 0 0 **…**

„Island" of numbers = Cluster

Step 4: Searching for "islands" of sites

Step 4: Defining the cluster borders: set of rules

1. „Gap rule"

CASSIS scans the number string immediately upstream and downstream of the anchor promoter until it hits the first "zero" value (promoter without binding site).

Gap rule: ≤ 2 zero-promoters

Is based on observations of real-life clusters (>30 known eukaryotic SM clusters).

Adjustable parameters and their estimation

What can influence the search:

1. MEME and FIMO searches. Refining the latter by adjusting the e-value and p-value cut-offs can be crucial for the whole cluster prediction.
2. Intrinsic CASSIS parameters:
   (i) the proportion of promoters with the motif in the genome (reflecting the genome- wide motif frequency);
   (ii) the maximal allowed number of "zero" promoters ("gaps") within the cluster (Gap rule)

All these parameters are estimated using a training set of experimentally verified SM clusters.
For the Ascomycete training set, the parameter values were:

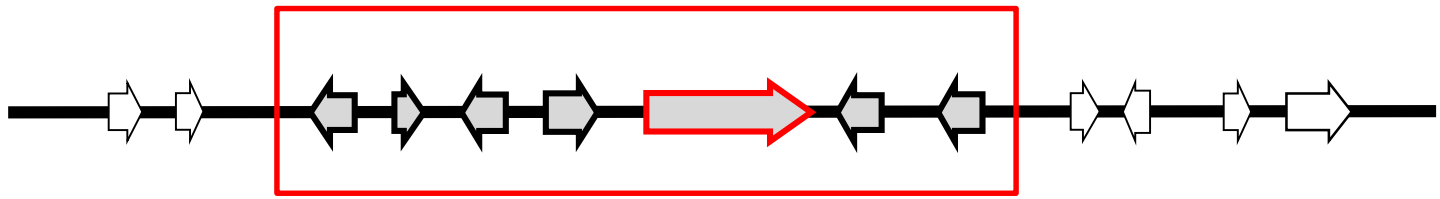
- frequency 14%;
- gap of 2 zero-promoters.

CASSIS is applicable to detection of *any* clusters as long as their genes are co-regulated and co-localized.

The type of a cluster is defined by its anchor gene.

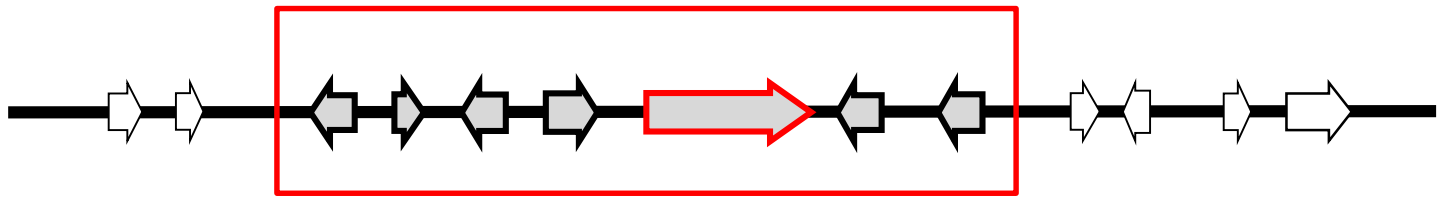
How to find a Gene Cluster in a genome?

1. Find an anchor gene
2. Find other genes



How to find a Secondary Metabolite Gene Cluster in a genome?

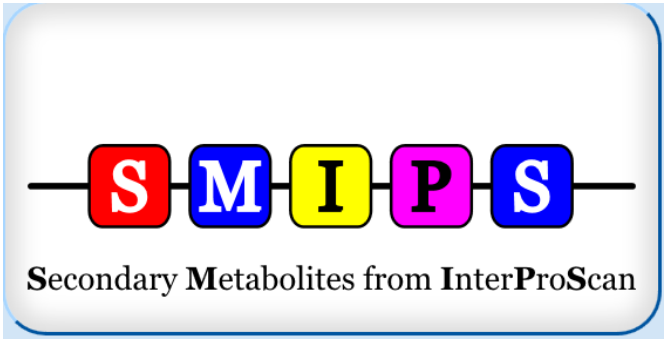
1. Find an anchor gene -> **SMIPS**
2. Find other genes (define the borders) -> **CASSIS**



SMIPS tool

Based on the prediction of the protein domains

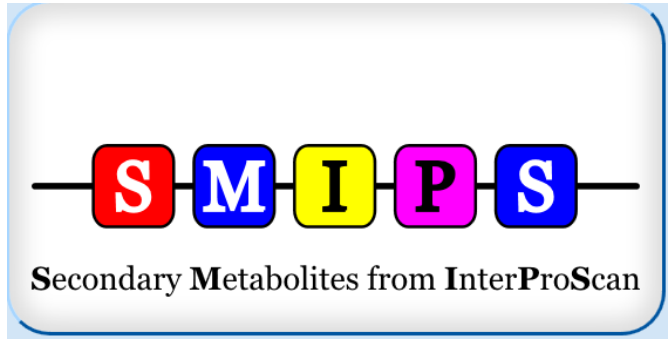
(InterProScan)



SMIPS tool

Based on the prediction of the protein domains

(InterProScan)



Genome-wide protein domain predictions (InterProScan)

List of typical anchor gene domains

→ Predictions of anchor genes

SMIPS

Input:

Protein sequences or InterProScan tables

Output:

Genome-wide predictions of anchor genes (PKSs, NRPSs, DMATs (dimethylallyl tryptophan synthases))

CASSIS

Input:

Genome sequence; feature tables (.gff and alike); anchor gene(s)

Output:

Cluster borders predictions.

Additional information: Shared motifs for each cluster.

Leibniz Institute for Natural Product Research and Infection Biology
Hans Knöll Institute

Member of the Leibniz Association

Go to the tools:

SMIPS Secondary Metabolites from InterProScan | CASSIS Cluster Assignment by Islands of Sites

The CASSIS suite

detection of secondary metabolite gene clusters in eukaryotic genomes

CASSIS

CASSIS ("cluster assignment by islands of sites") is a tool to predict secondary metabolite gene clusters around a given anchor/backbone gene. A gene cluster is a small group of genes, which are tightly co-localized, co-regulated, and participate in the same metabolic pathway.

CASSIS utilizes a so-called "motif-based" prediction method. It is mainly based on the hypothesized co-regulation of cluster genes. Hence, CASSIS searches for transcription factor binding sites shared by promoter sequences of putative cluster genes.

The motif-based method applied by CASSIS is complementary to similarity-based methods, such as those exploited by antiSMASH or SMURF.

SMIPS

SMIPS ("secondary metabolites by InterProScan") is a tool for genome-wide prediction of anchor/backbone genes. Anchor genes encode enzymes, which play a major role in the biosynthesis of secondary metabolites. SMIPS identifies three most common classes of the anchor genes: polyketide synthases (PKS), non-ribosomal peptide synthetases (NRPS), and dimethylallyltryptophan synthases (DMATS).

Results

- **Cross-validation**
- **Comparison with other tools**

Results

- **Cross-validation (LOO)**

Table 1. Benchmark results of the leave-one-out cross-validation for CASSIS.

Characteristics	CASSIS performance[a]
Sensitivity	0.84 ± 0.0010
Specificity	0.98 ± 0.0002
Precision	0.71 ± 0.0010
Accuracy	0.96 ± 0.0002
FDR	0.29 ± 0.0010
F_1-score	0.73 ± 0.0008

[a] Average for all 38 LOO experiments. Error is the standard error of the mean. See Supplementary Table S1 for the list of used clusters.

Results

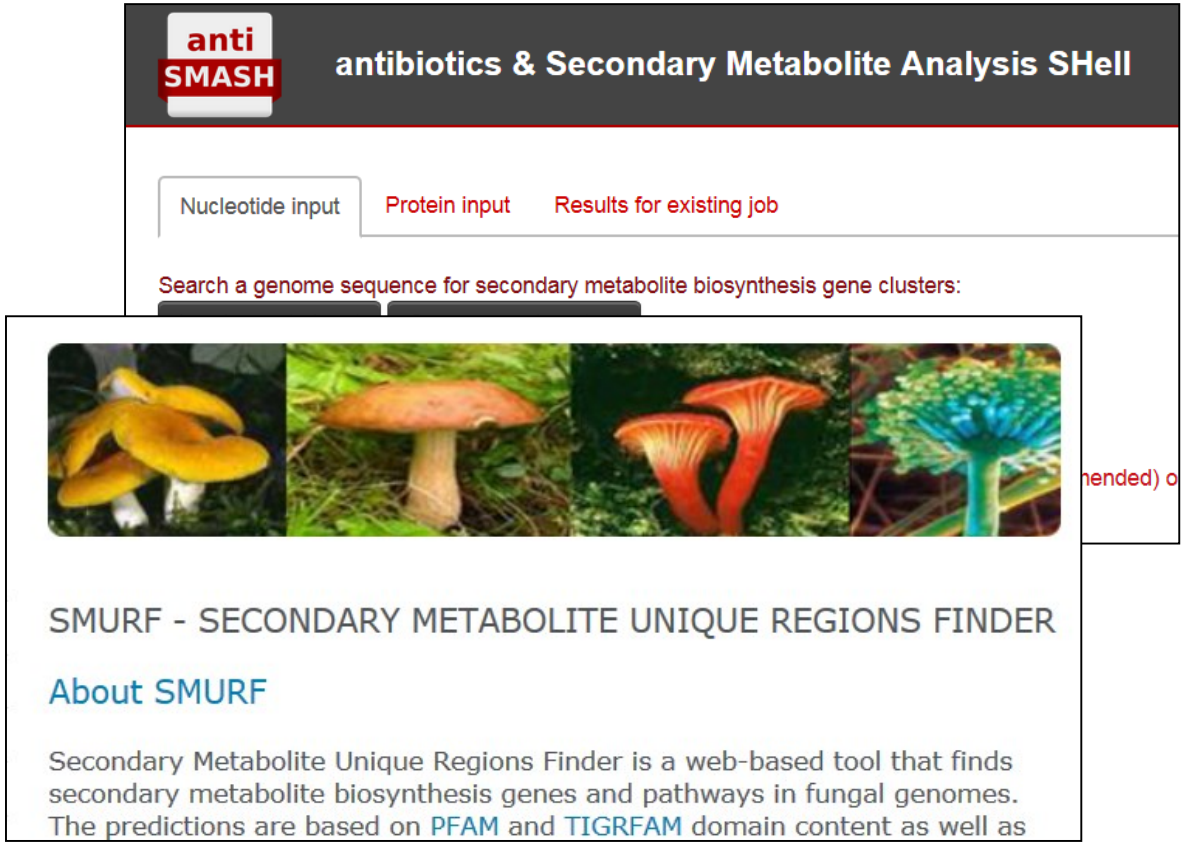
Cross-validation (LOO)

Table 1. Benchmark results of the leave-one-out cross-validation for CASSIS.

Characteristics	CASSIS performance[a]
Sensitivity	0.84 ± 0.0010
Specificity	0.98 ± 0.0002
Precision	0.71 ± 0.0010
Accuracy	0.96 ± 0.0002
FDR	0.29 ± 0.0010
F_1-score	0.73 ± 0.0008

[a] Average for all 38 LOO experiments. Error is the standard error of the mean. See Supplementary Table S1 for the list of used clusters.

Comparison with other tools



Results

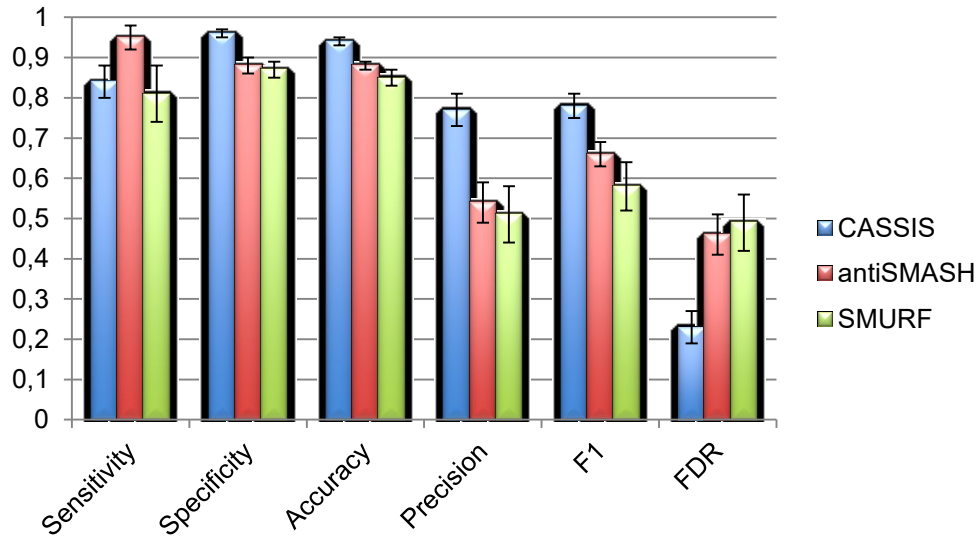
Cross-validation (LOO)

Table 1. Benchmark results of the leave-one-out cross-validation for CASSIS.

Characteristics	CASSIS performance[a]
Sensitivity	0.84 ± 0.0010
Specificity	0.98 ± 0.0002
Precision	0.71 ± 0.0010
Accuracy	0.96 ± 0.0002
FDR	0.29 ± 0.0010
F_1-score	0.73 ± 0.0008

[a] Average for all 38 LOO experiments. Error is the standard error of the mean. See Supplementary Table S1 for the list of used clusters.

Comparison with other tools



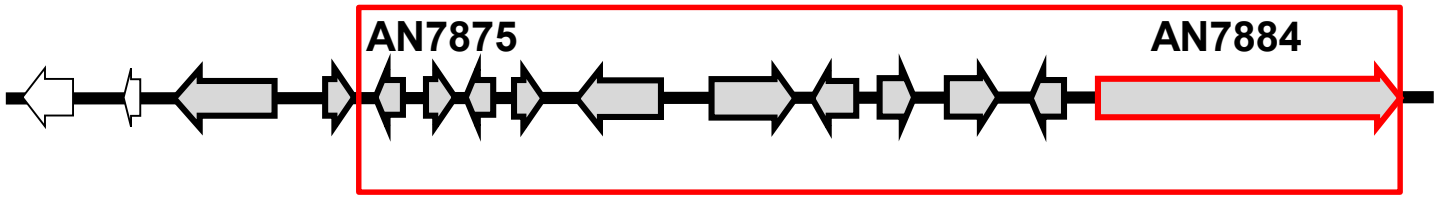
Comparison of CASSIS with the similarity-based antiSMASH and SMURF tools: Re-identification of the 12 test clusters not used for the tools' training.

⇒ CASSIS integration into the antiSMASH (made in 2017)
⇒ Users can have 2 types of prediction (protein-based and promoter-based)

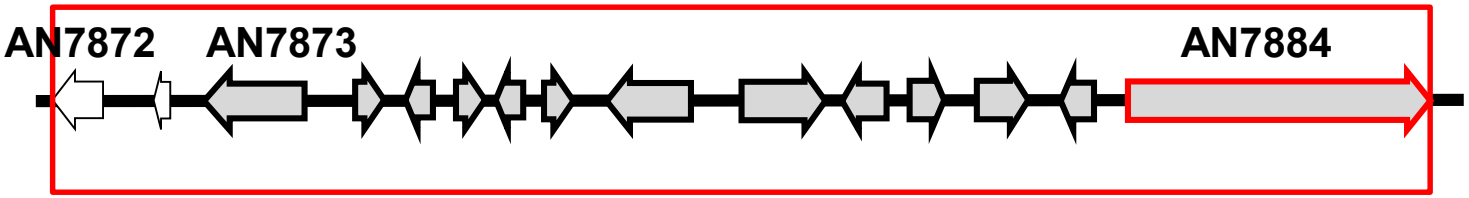
Examples. Stories of application

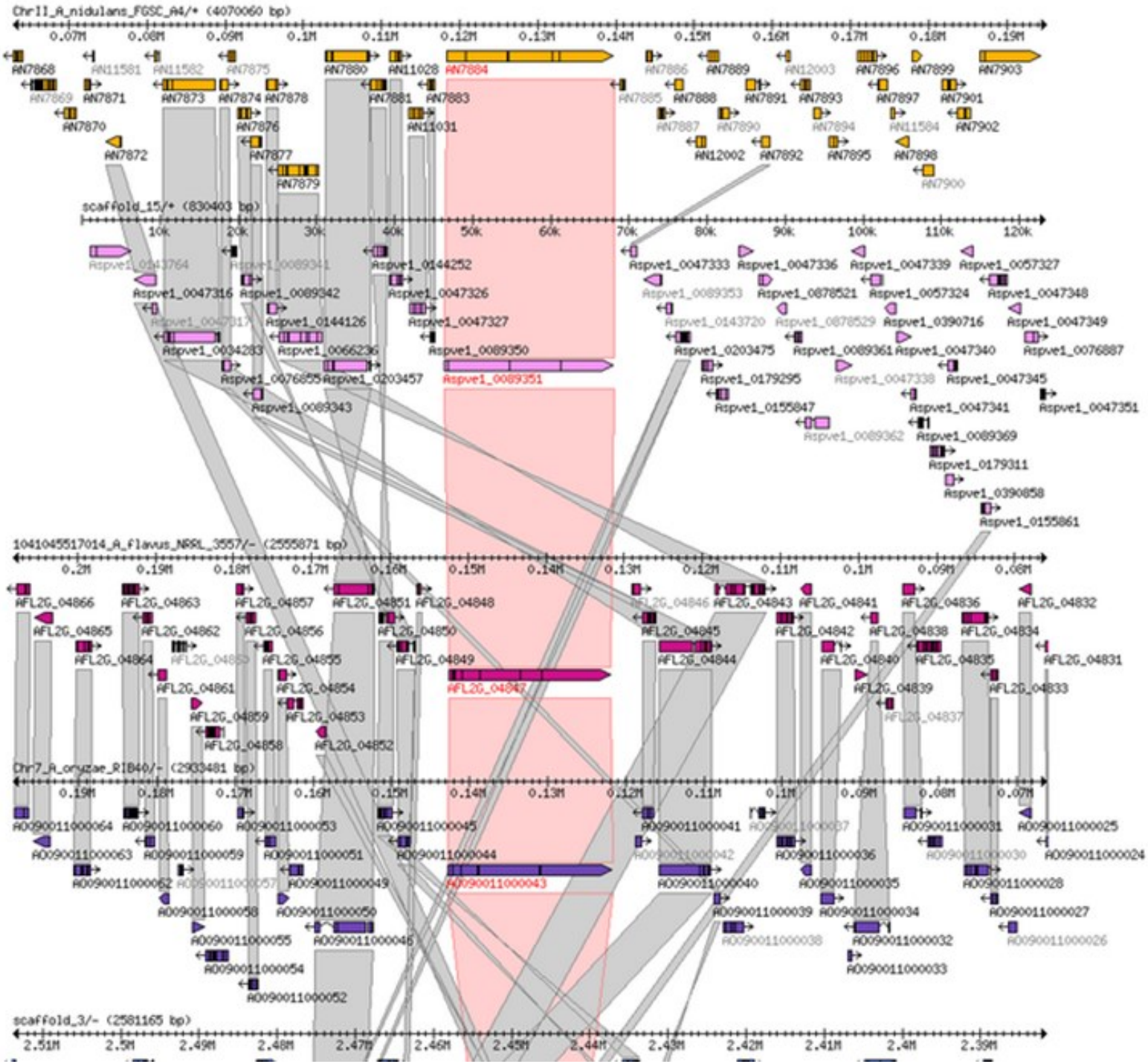
AN7884 was not characterized until recently

We analysed the genomic region with CASSIS:



+ Synteny prediction:




ChrII_A_nidulans_FGSC_A4/+ (4070060 bp)
AN7868
AN11581
AN11582
AN7875
AN7880
AN11028
AN7884
AN7886
AN7889
AN12003
AN7896
AN7899
AN7903
AN7869
AN7871
AN7873
AN7874
AN7878
AN7881
AN7883
AN7885
AN7888
AN7891
AN7893
AN7897
AN7901
AN7870
AN7876
AN11031
AN7887
AN7890
AN7894
AN11584
AN7902
AN7872
AN7877
AN12002
AN7892
AN7895
AN7898
AN7879
AN7900
scaffold_15/+ (830403 bp)
Aspve1_0144252
Aspve1_0047333
Aspve1_0047336
Aspve1_0047339
Aspve1_0057327
Aspve1_0047316
Aspve1_0089342
Aspve1_0047326
Aspve1_0089353
Aspve1_0078521
Aspve1_0057324
Aspve1_0047348
Aspve1_0144126
Aspve1_0047327
Aspve1_0143720
Aspve1_0078529
Aspve1_0390716
Aspve1_0047349
Aspve1_0034283
Aspve1_0066236
Aspve1_0089350
Aspve1_0203475
Aspve1_0089361
Aspve1_0047340
Aspve1_0076887
Aspve1_0076855
Aspve1_0203457
Aspve1_0089351
Aspve1_0179295
Aspve1_0047338
Aspve1_0047345
Aspve1_0089343
Aspve1_0155847
Aspve1_0047341
Aspve1_0047351
Aspve1_0089362
Aspve1_0089369
Aspve1_0179311
Aspve1_0390858
Aspve1_0155861
1041045517014_A_flavus_NRRL_3557/- (2555871 bp)
AFL2G_04866
AFL2G_04863
AFL2G_04857
AFL2G_04851
AFL2G_04848
AFL2G_04846
AFL2G_04843
AFL2G_04841
AFL2G_04836
AFL2G_04832
AFL2G_04865
AFL2G_04862
AFL2G_04856
AFL2G_04850
AFL2G_04845
AFL2G_04842
AFL2G_04838
AFL2G_04834
AFL2G_04864
AFL2G_04855
AFL2G_04849
AFL2G_04844
AFL2G_04840
AFL2G_04835
AFL2G_04831
AFL2G_04861
AFL2G_04854
AFL2G_04847
AFL2G_04839
AFL2G_04833
AFL2G_04859
AFL2G_04853
AFL2G_04837
AFL2G_04858
AFL2G_04852
Chr7_A_oryzae_RIB40/- (2903481 bp)
AO090011000064
AO090011000060
AO090011000053
AO090011000045
AO090011000041
AO090011000037
AO090011000031
AO090011000025
AO090011000063
AO090011000059
AO090011000051
AO090011000044
AO090011000042
AO090011000036
AO090011000030
AO090011000024
AO090011000062
AO090011000057
AO090011000049
AO090011000043
AO090011000040
AO090011000035
AO090011000028
AO090011000058
AO090011000050
AO090011000039
AO090011000034
AO090011000027
AO090011000055
AO090011000046
AO090011000038
AO090011000032
AO090011000026
AO090011000054
AO090011000033
AO090011000052
scaffold_3/- (2581165 bp)

2016:

GDCh **Communications** Angewandte International Edition Chemie

Natural Products

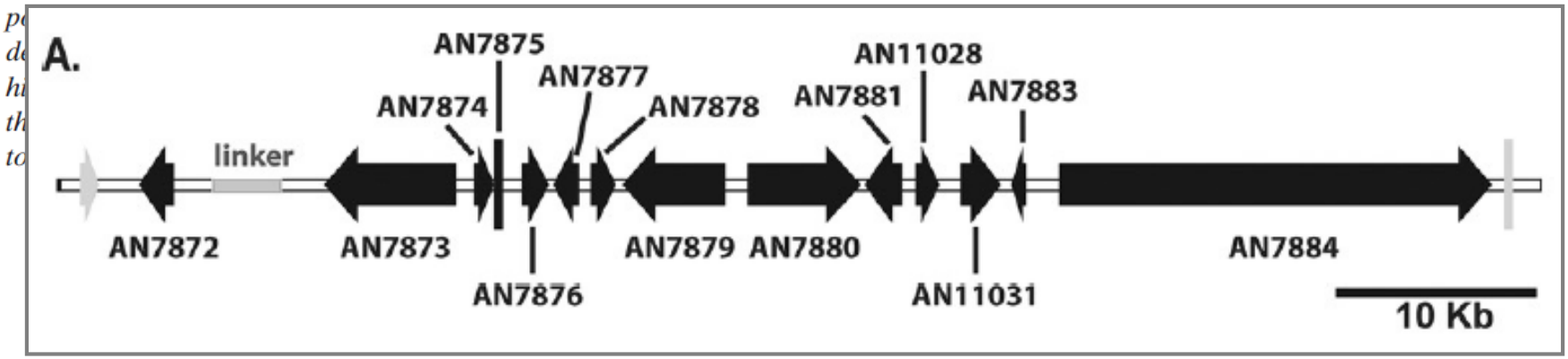
International Edition: DOI: 10.1002/anie.201507097
German Edition: DOI: 10.1002/ange.201507097

Development of Genetic Dereplication Strains in *Aspergillus nidulans* Results in the Discovery of Aspercryptin

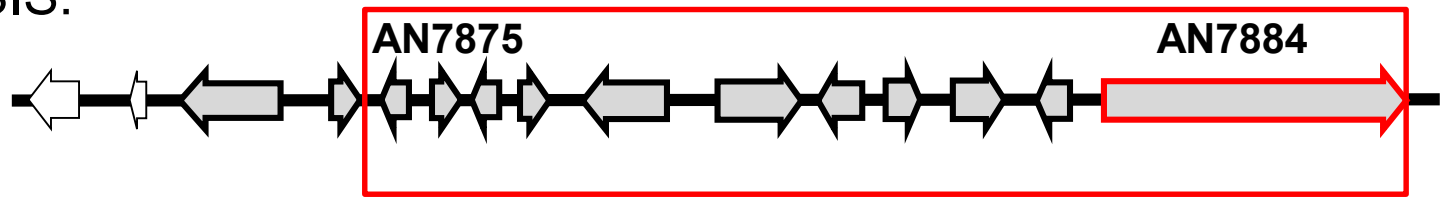
*Yi-Ming Chiang, Manmeet Ahuja, C. Elizabeth Oakley, Ruth Entwistle, Anabanadam Asokan, Christoph Zutz, Clay C. C. Wang, and Berl R. Oakley**

Abstract: *To reduce the secondary metabolite background in Aspergillus nidulans and minimize the rediscovery of com-*

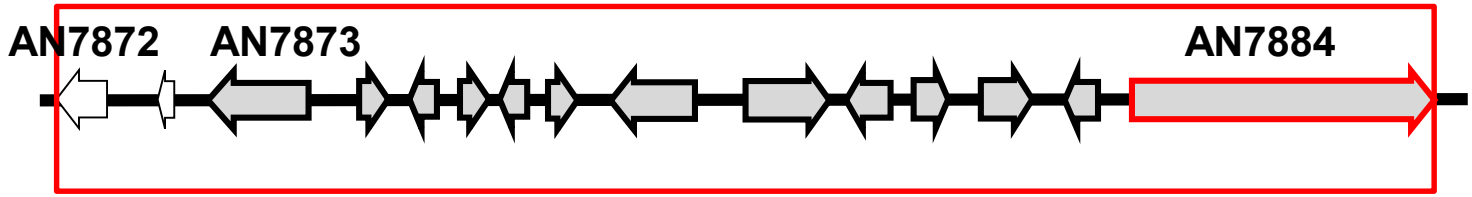
the major known SM biosynthetic pathways in *A. nidulans*, thereby reducing the complexity of SM profiles such that



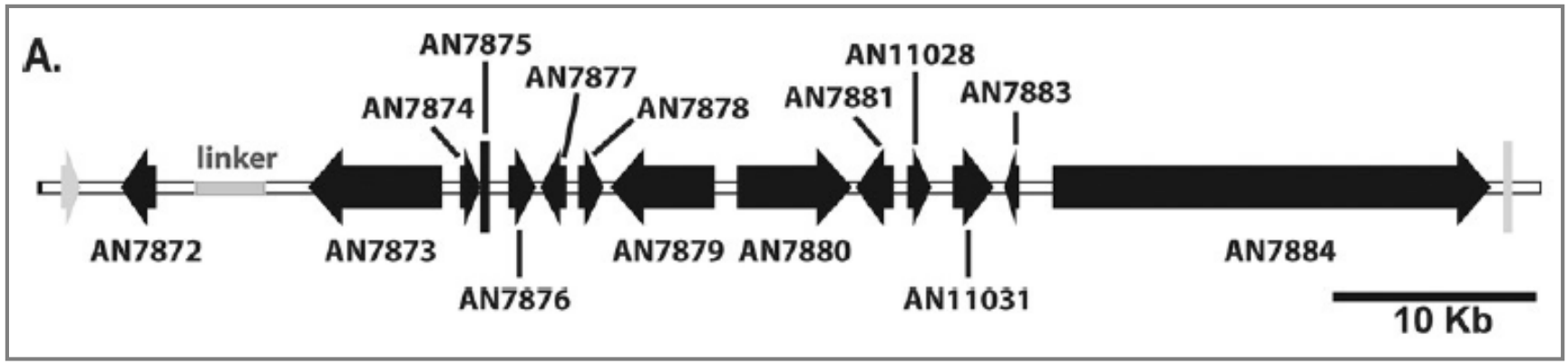
CASSIS:



+ Synteny prediction:



Synteny is a powerful tool!



Systems Biology/ Bioinformatics group, Hans Knöll Institute, Jena:

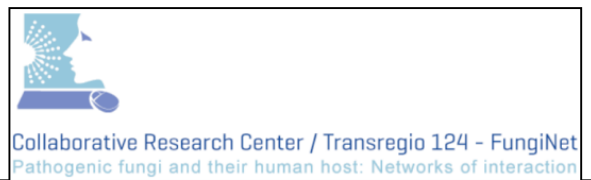
Vladimir Shelest

Thomas Wolf

Alina Burmistrova

Experimental work:

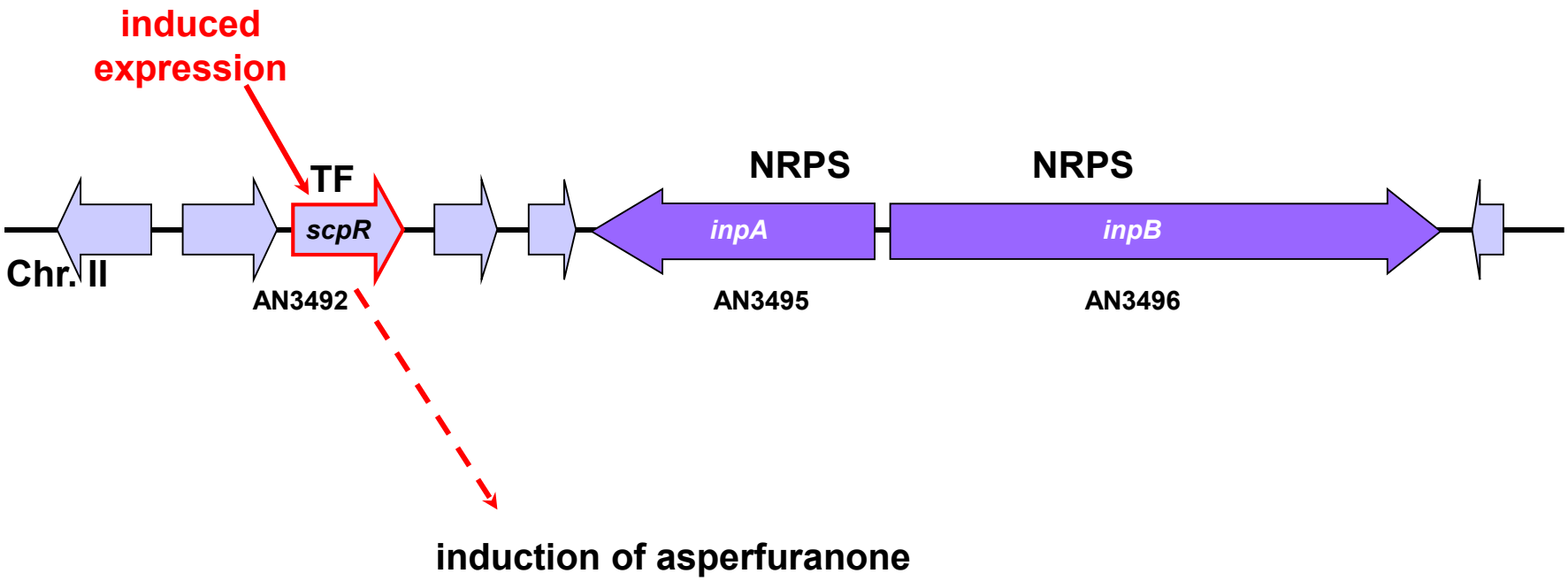
Applied Molecular Microbiology lab, Hans Knöll Institute, Jena



Thank you for your attention!

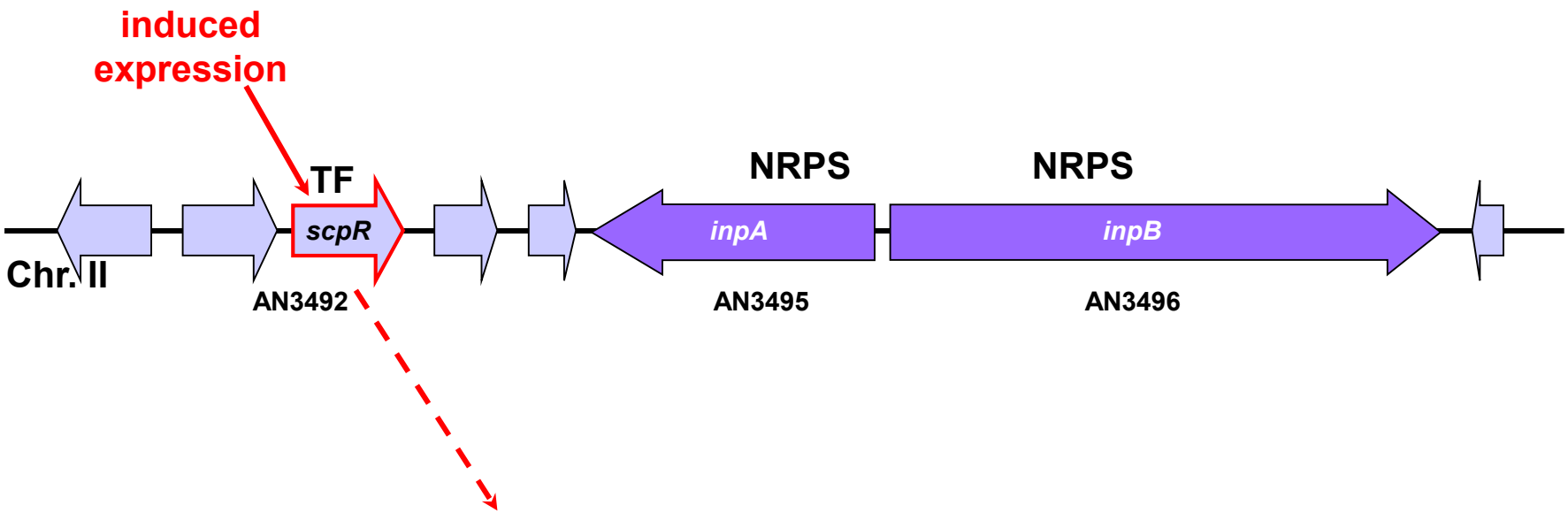
Inter-cluster cross-regulation

HKI Jena, 2010: Activation of the silent NRPSs AN3495/3496



S. Bergmann et al., 2010

HKI Jena, 2010: Activation of the silent NRPSs AN3495/3496

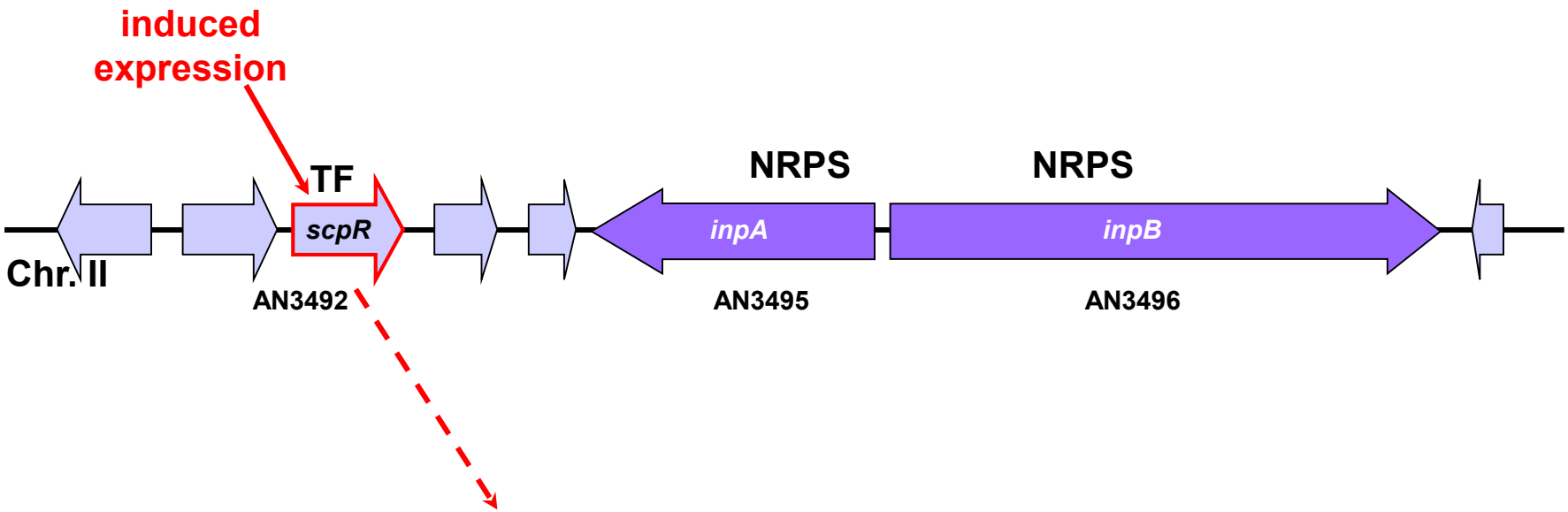


induction of asperfuranone

But asperfuranone
is a polyketide!

S. Bergmann et al., submitted

HKI Jena, 2010: Activation of the silent NRPSs AN3495/3496



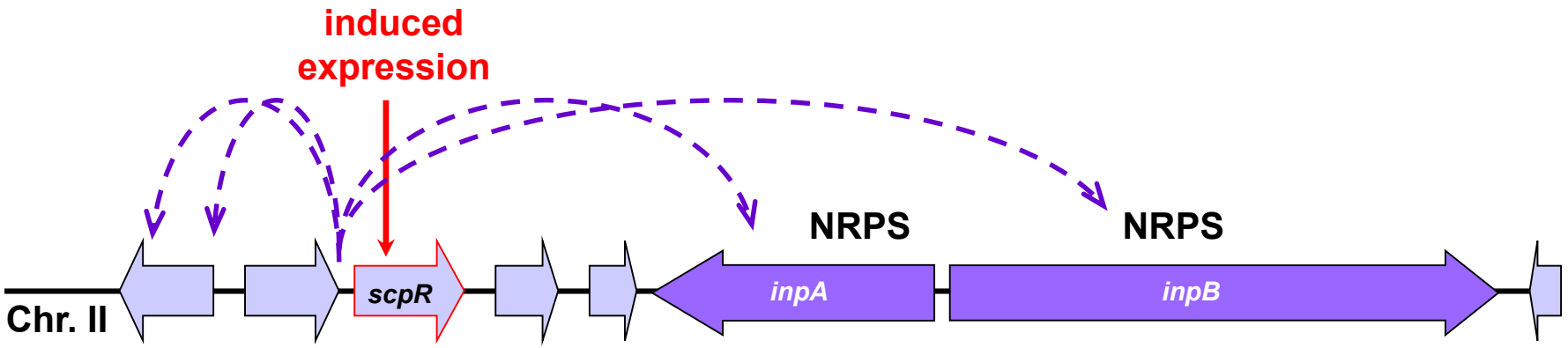
induction of asperfuranone

But asperfuranone
is a polyketide!

⟷ induction of a PKS cluster!

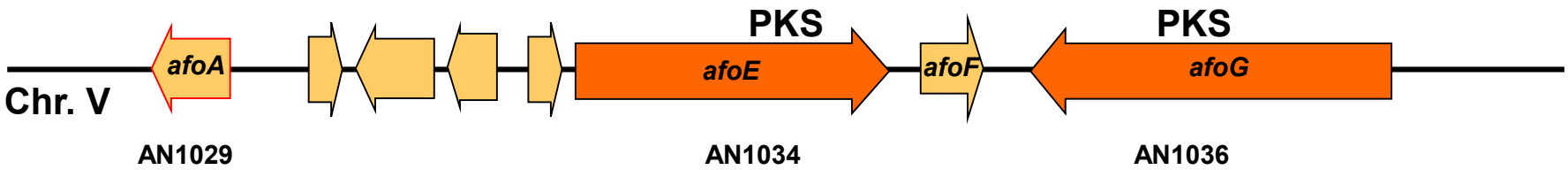
S. Bergmann et al., 2010

Regulatory cross-talk between the clusters

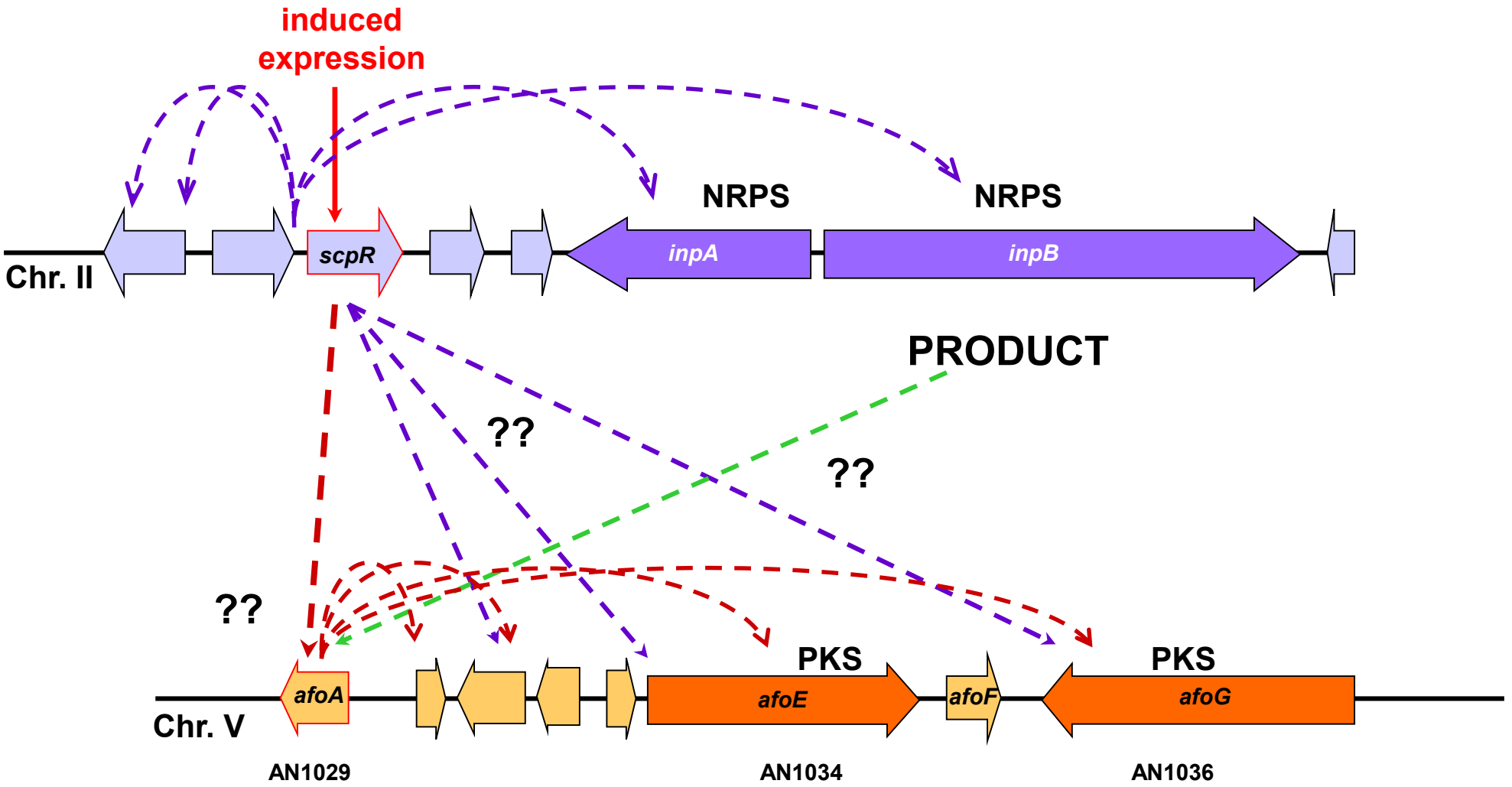


?? HOW?

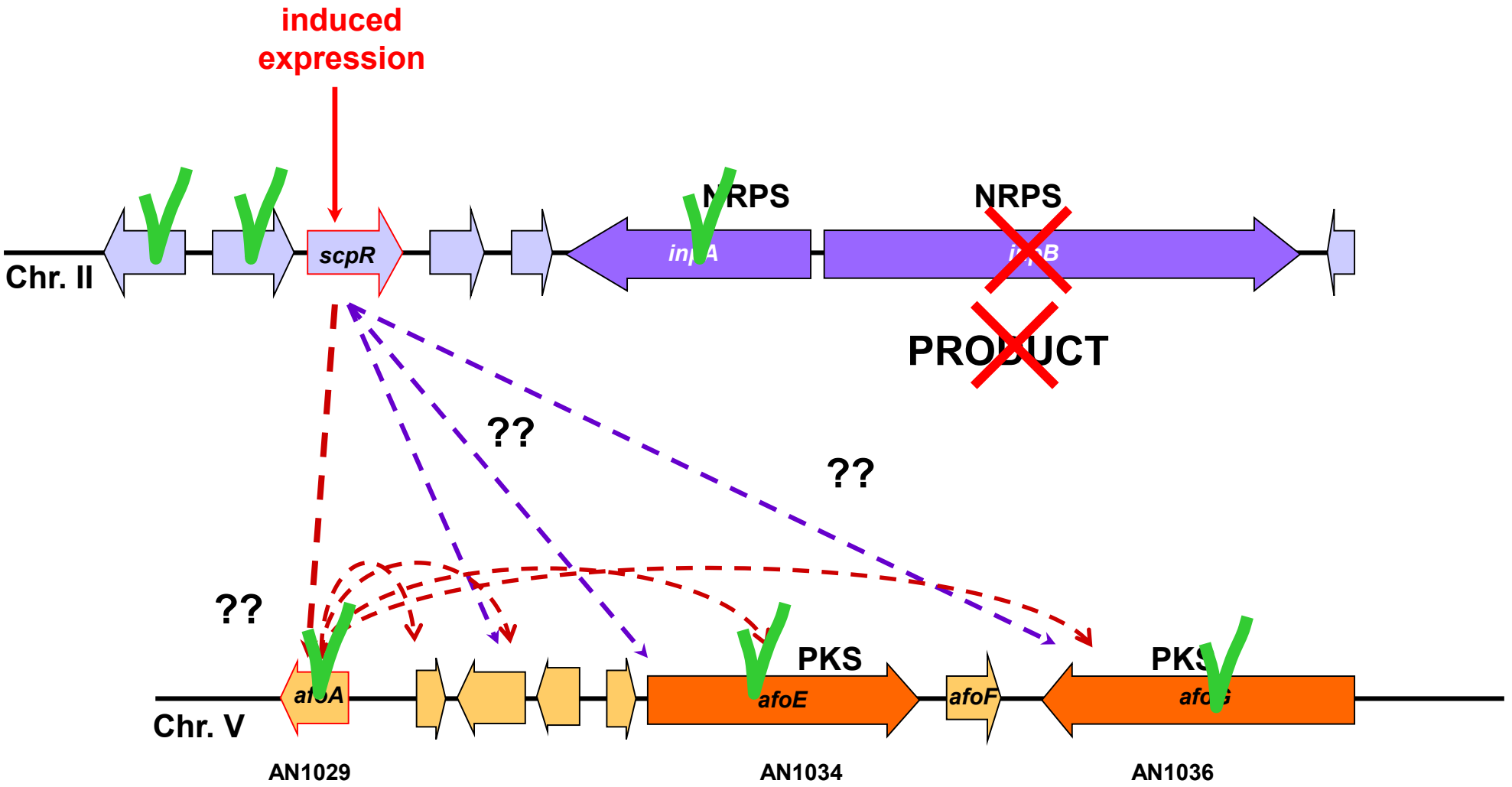
asperfuranone biosynthetic cluster:



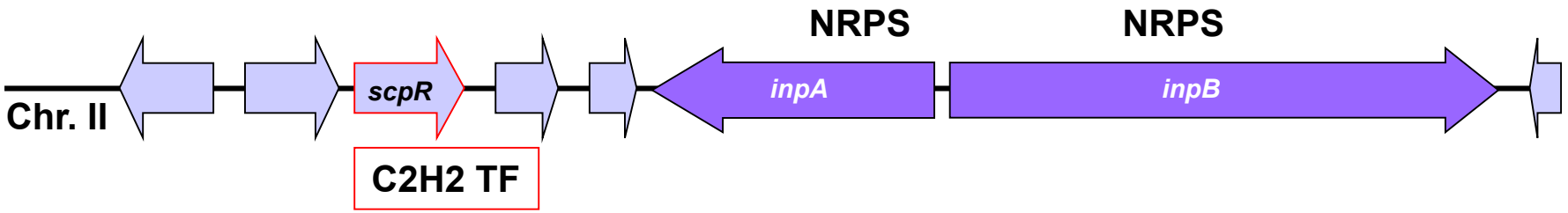
Regulatory cross-talk between the clusters

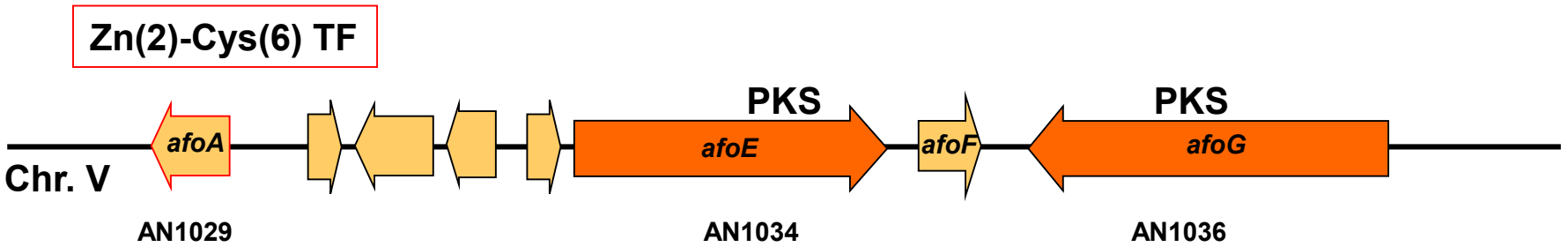


Regulatory cross-talk between the clusters

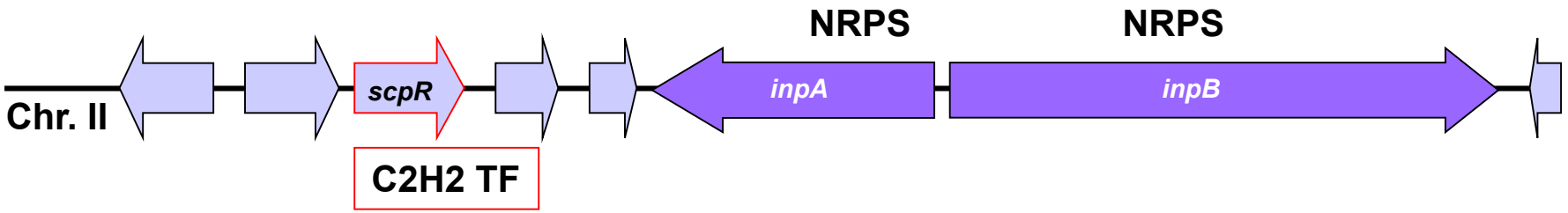


Regulatory cross-talk between wet-lab and bioinformatics

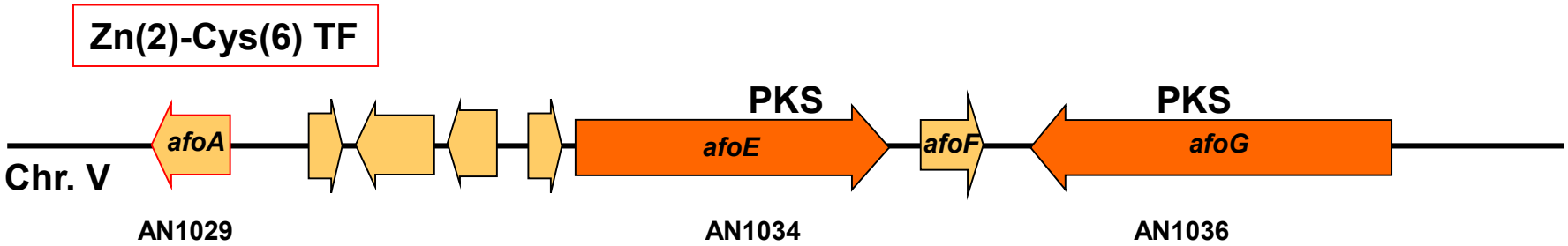




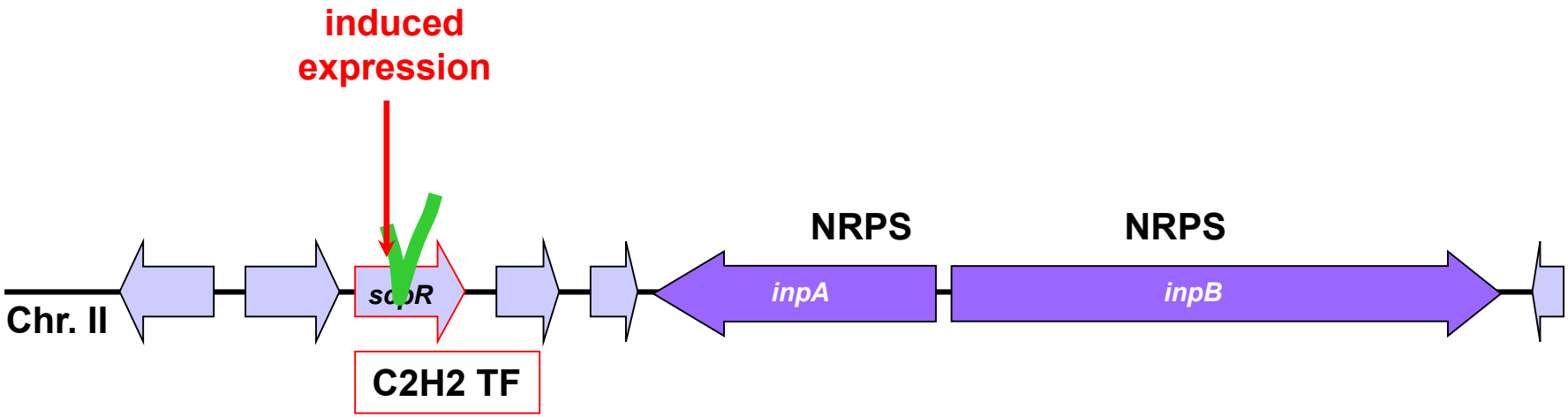
Regulatory cross-talk between wet-lab and bioinformatics



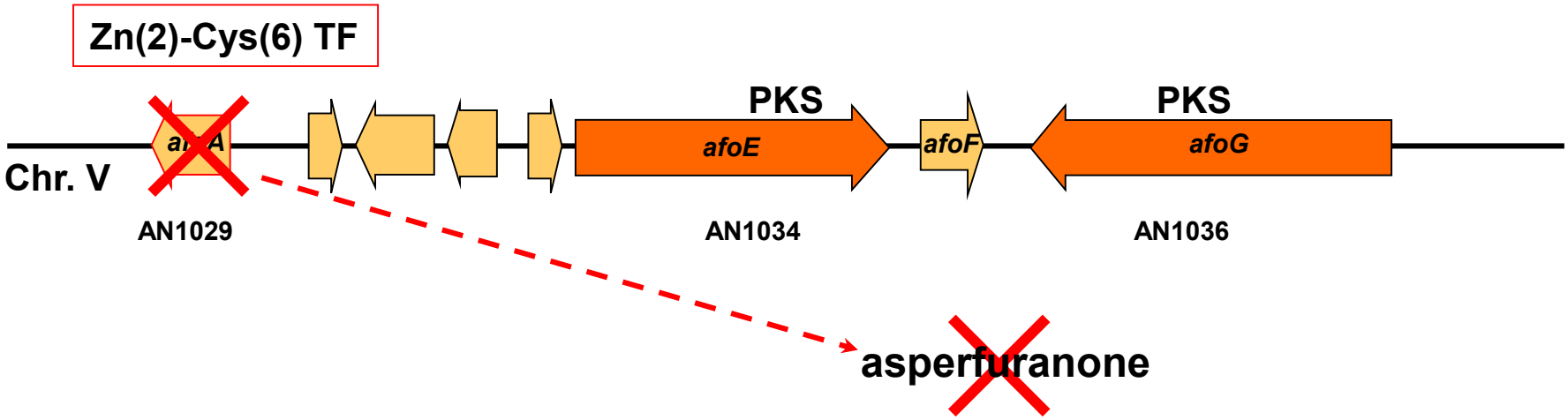
Suggestion from the bioinformatics side:
deletion of the *afoA* TF (AN1029)



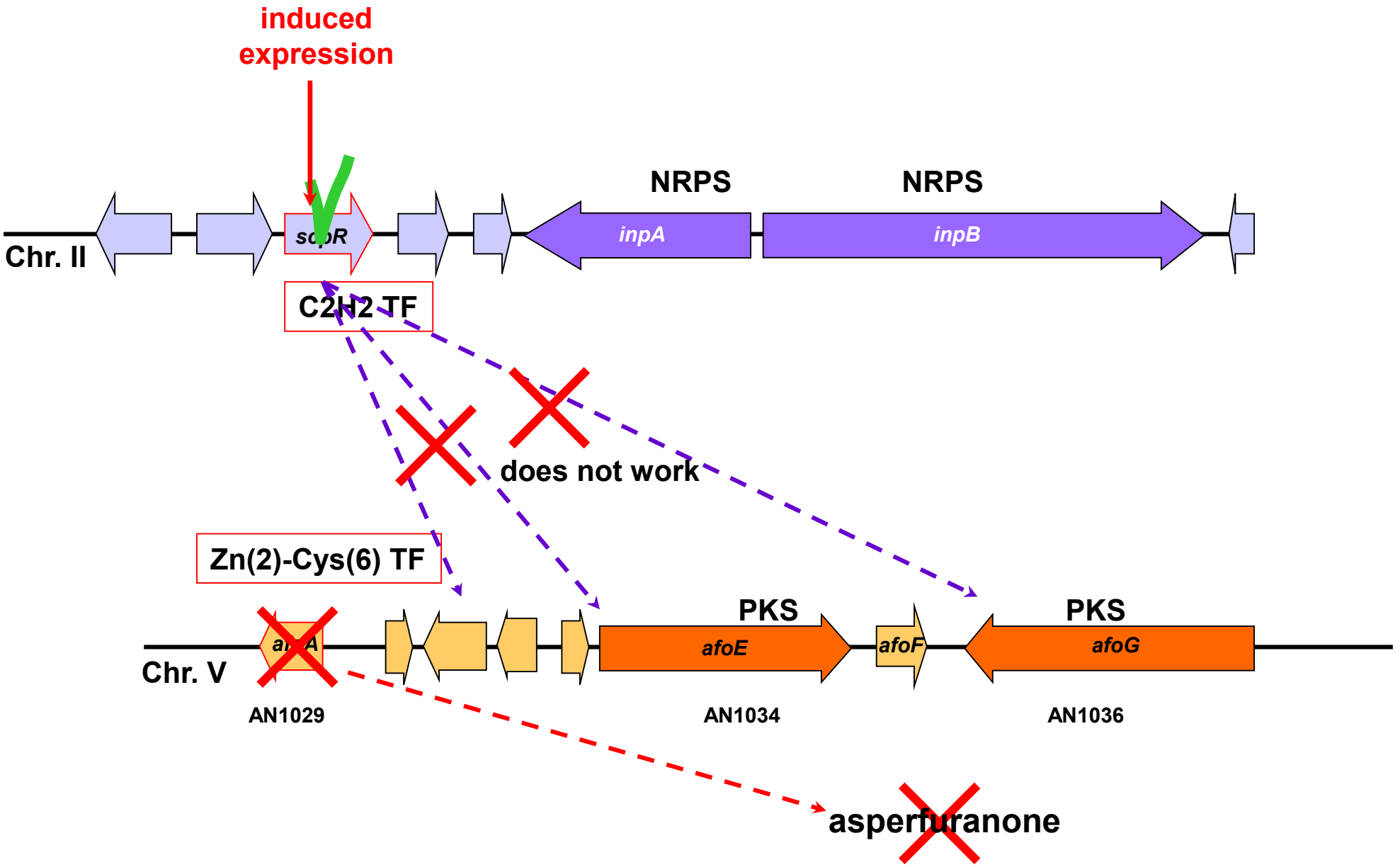
Regulatory cross-talk between wet-lab and bioinformatics



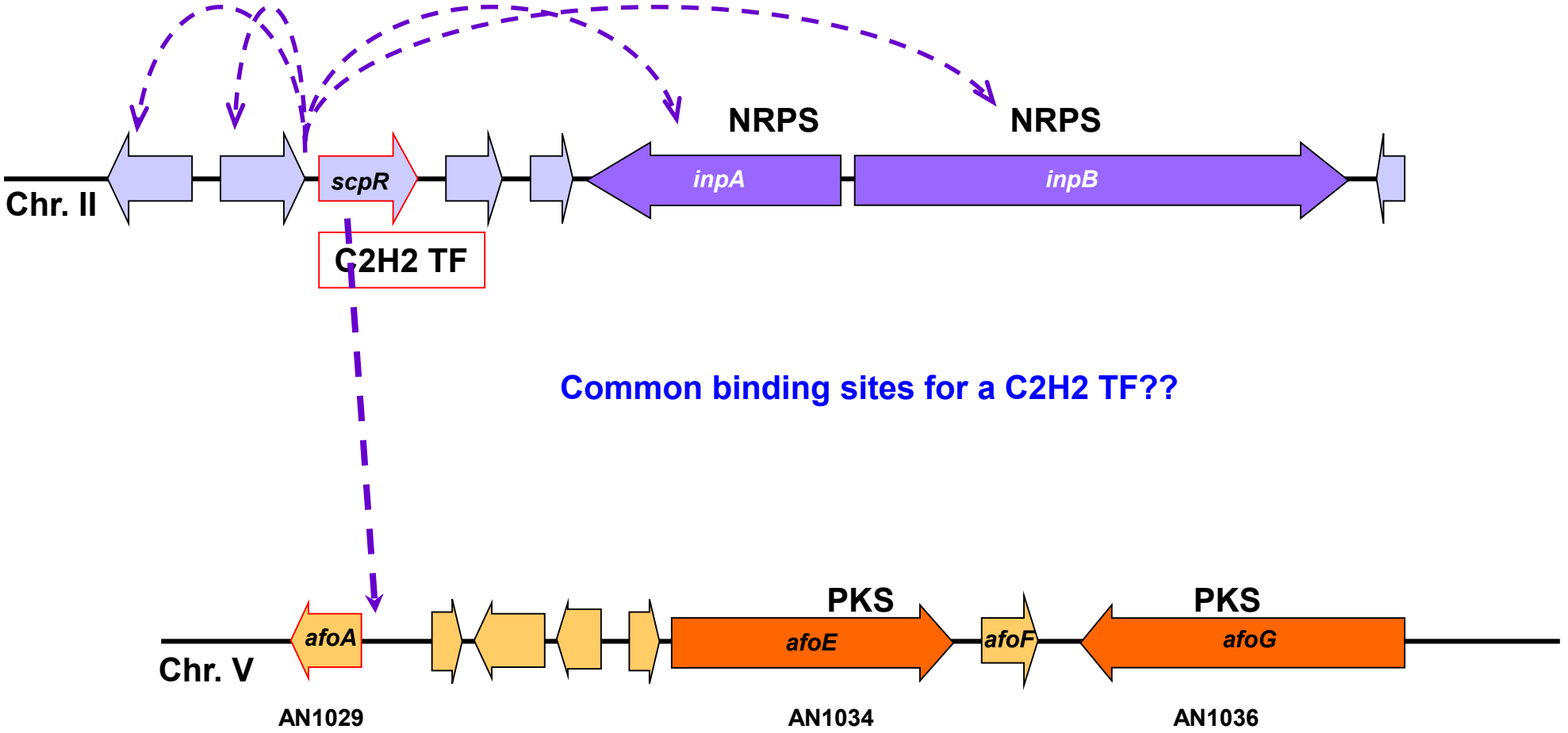
Suggestion from the bioinformatics side:
deletion of the *afoA* TF (AN1029)



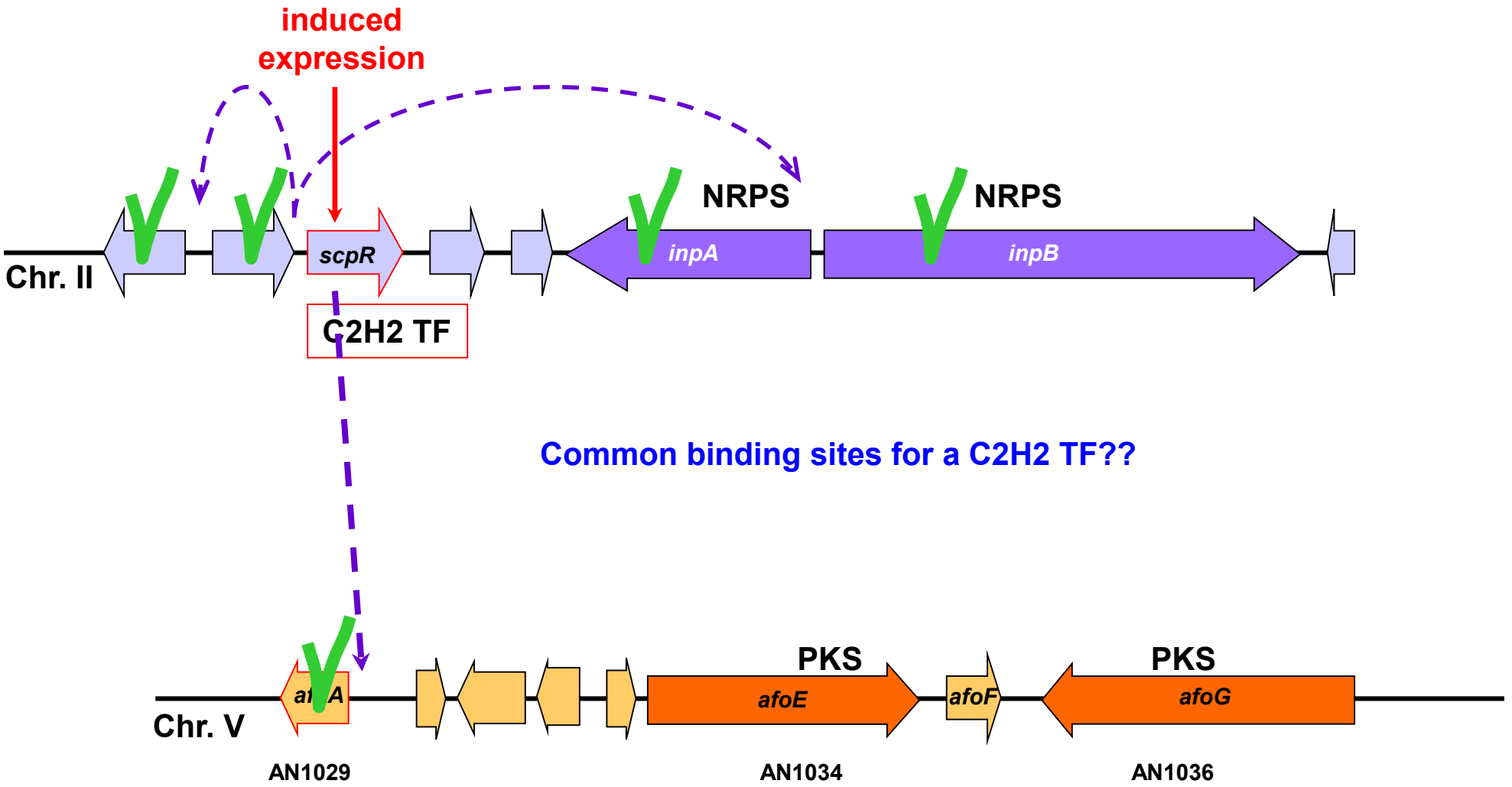
Regulatory cross-talk between wet-lab and bioinformatics



Regulatory cross-talk between wet-lab and bioinformatics



Regulatory cross-talk between wet-lab and bioinformatics



Regulatory cross-talk between wet-lab and bioinformatics

The MEME search in the upstream sequences:

Sequence name	Strand	Start	P-value	Site	
AN1029-30_interg1371	+	915	1.12e-07	AGAACGTGGT	**CTAAAGGATTGA** GCTGACGATG
AN3496.4/AN3495.4	-	630	3.39e-07	TAACGATTAG	**CAAAAGGATTGA** CTAAATCAAG
AN1029-30_interg1371	-	1002	5.65e-07	AGCCACTAGC	**CTAAAGGAATCA** GACCTTTAAT
AN3491.4/AN3490.4	+	476	1.03e-06	CATCACCCGT	**CCAAAGGATGCA** CCAAGGAACA

This motif is very similar to the one of RME1, a yeast zinc-finger transcription factor with Cys2His2 domain: