SeMPI– a genome-based Secondary Metabolite Prediction and Identification web server

Paul Zierep, Natàlia Padilla, Dimitar Yonchev, Kiran Telukunta, Dennis Klementz and Stefan Günther
Pharmaceutical Bioinformatics, Institute of Pharmaceutical Science, Germany
paul.zierep@pharmazie.uni-freiburg.de

The secondary metabolism of bacteria, fungi and plants yields a vast number of bio-active substances. The constantly increasing amount of published genomic data provides the opportunity for an efficient identification of gene clusters by genome mining. Conversely, for many natural products with resolved structures, the encoding gene clusters have not been identified yet. Structural elucidation of the actual secondary metabolite is still challenging, especially due to as yet unpredictable post-modifications. Here we introduce SeMPI, a web server providing a prediction and identification pipeline for natural products synthesized by polyketide synthases (PKS) of type I modular.

Chain prediction

>New Gene Sequence
VAMVFPGQGAQWQGMARDLLRESQVFAD ...

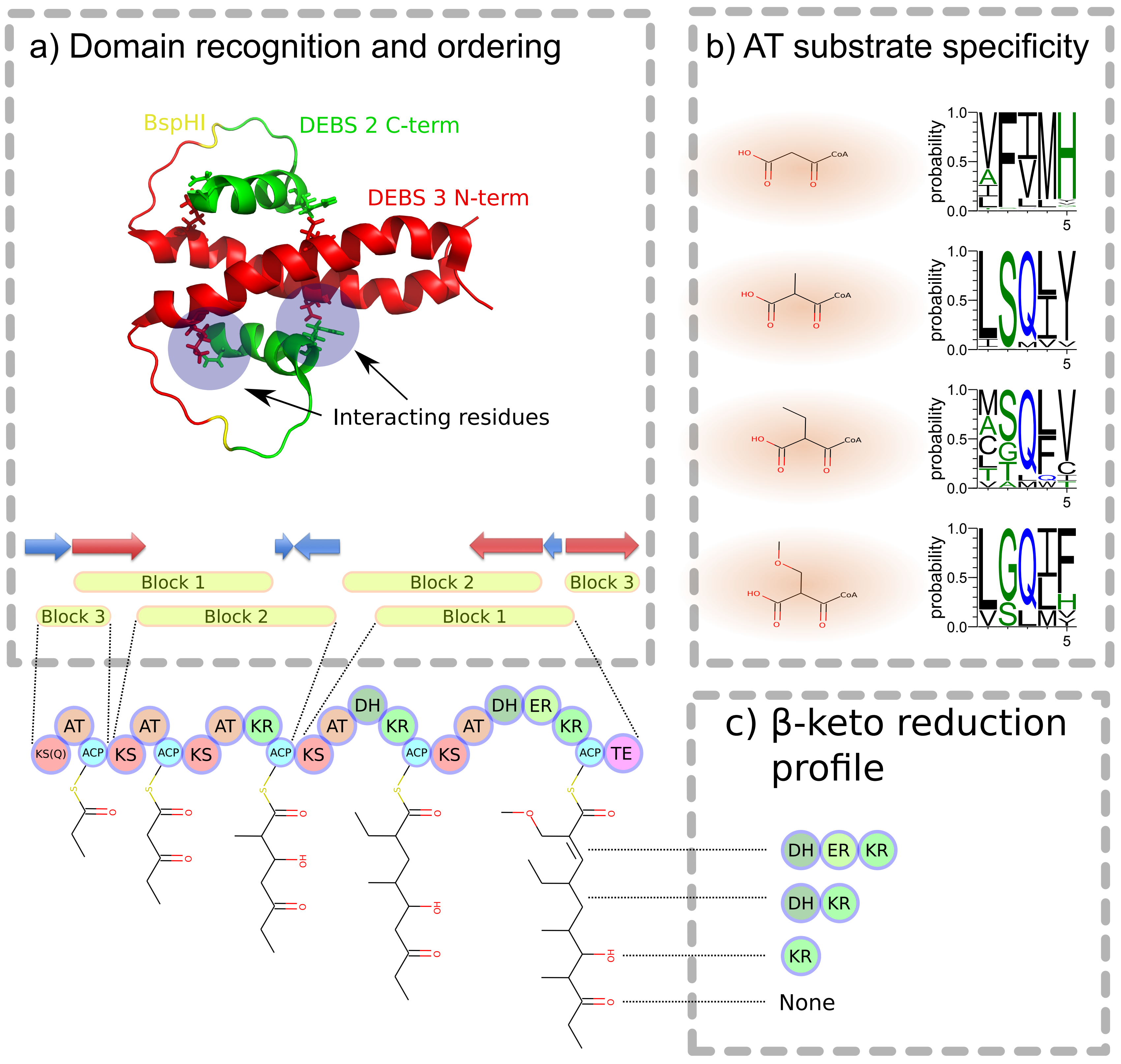
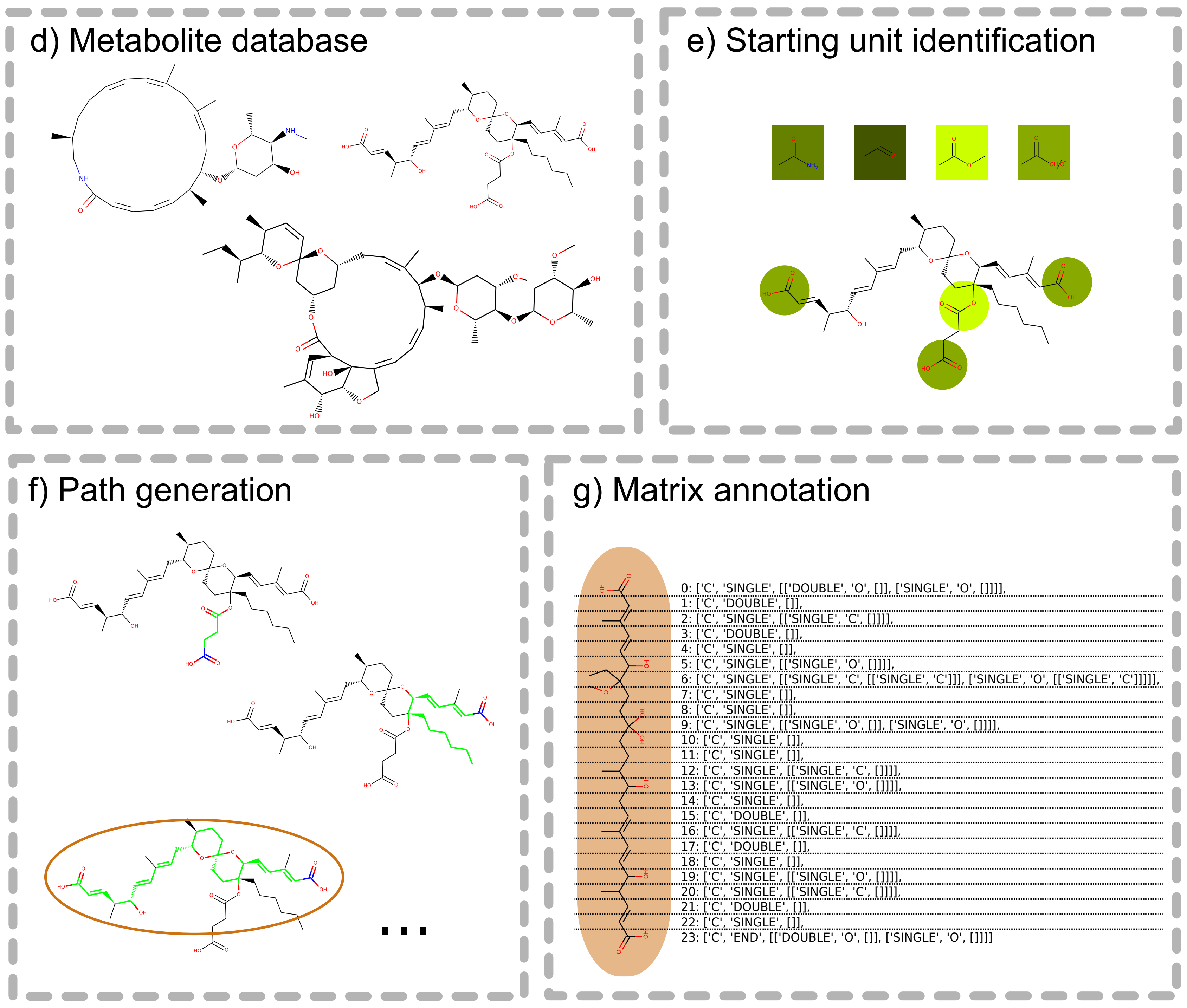


Figure 1. The gene cluster boundaries are based on the antiSMASH 3.0 [1] annotation. Domain signatures are identified using profiles Hidden Markov Models and ordered based on interacting residues, determined by alignment of the docking domains with 6-deoxyerythronolide B synthase (DEBS) (a). Acetyltransferase (AT) specificity is predicted using position probability matrices (b) and β-keto modification is based on presence of reductive domains (c). This information is used to predict the basic PK-chain.

Database screening



```
0: ['C', 'SINGLE', [['DOUBLE', 'O', []], ['SINGLE', 'O', []]]],
1: ['C', 'DOUBLE', []],
2: ['C', 'SINGLE', [['SINGLE', 'C', []]]],
3: ['C', 'DOUBLE', []],
4: ['C', 'SINGLE', []],
5: ['C', 'SINGLE', [['SINGLE', 'O', []]]],
6: ['C', 'SINGLE', [['SINGLE', 'C', [['SINGLE', 'C']]], ['SINGLE', 'O', [['SINGLE', 'C']]]]],
7: ['C', 'SINGLE', []],
8: ['C', 'SINGLE', []],
9: ['C', 'SINGLE', [['SINGLE', 'O', []], ['SINGLE', 'O', []]]],
10: ['C', 'SINGLE', []],
11: ['C', 'SINGLE', []],
12: ['C', 'SINGLE', [['SINGLE', 'C', []]]],
13: ['C', 'SINGLE', [['SINGLE', 'O', []]]],
14: ['C', 'SINGLE', []],
15: ['C', 'DOUBLE', []],
16: ['C', 'SINGLE', [['SINGLE', 'C', []]]],
17: ['C', 'DOUBLE', []],
18: ['C', 'SINGLE', []],
19: ['C', 'SINGLE', [['SINGLE', 'O', []]]],
20: ['C', 'SINGLE', [['SINGLE', 'C', []]]],
21: ['C', 'DOUBLE', []],
22: ['C', 'SINGLE', []],
23: ['C', 'END', [['DOUBLE', 'O', []], ['SINGLE', 'O', []]]]
```

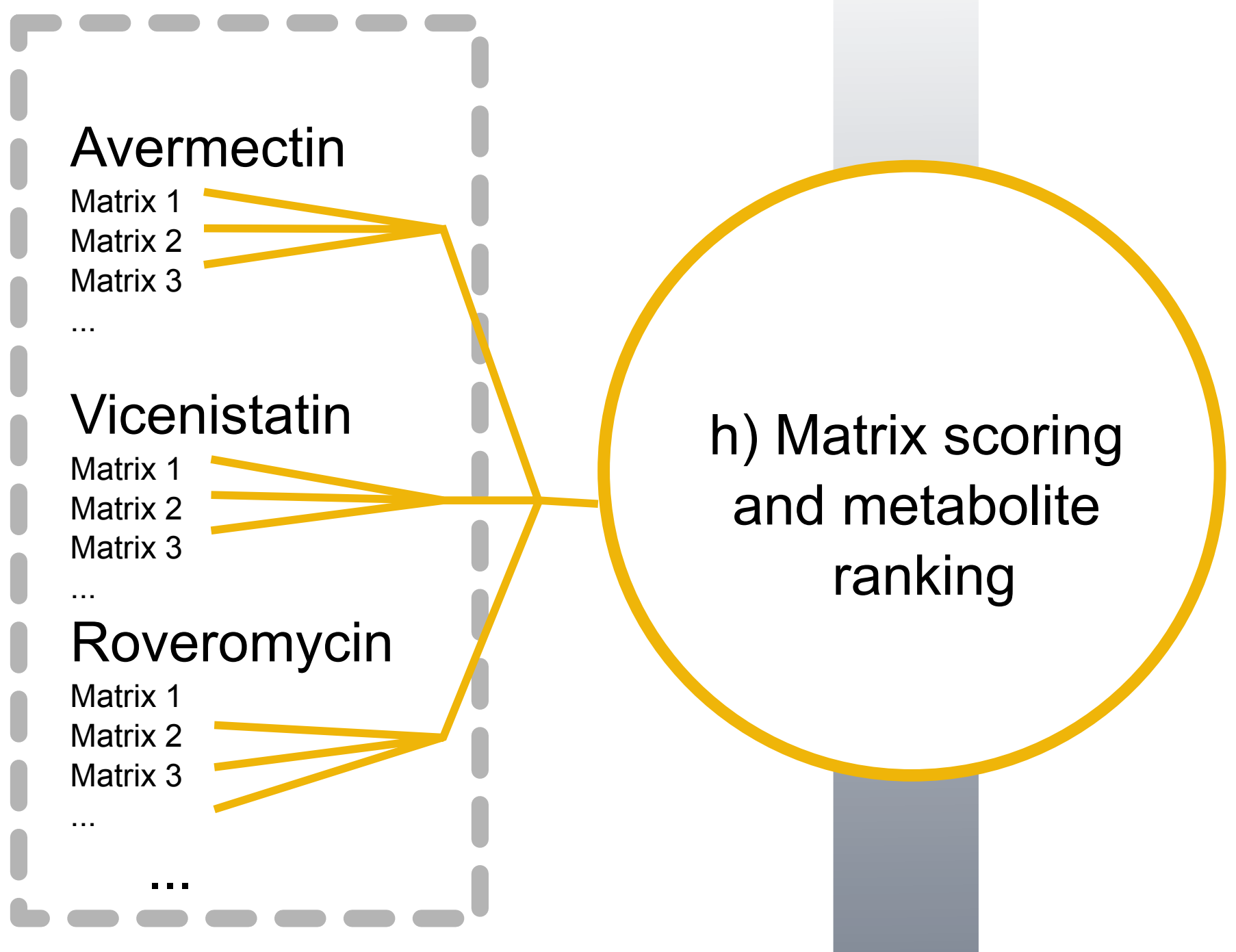


Figure 2. A database [2] with more then 4000 diverse molecules is used for chain comparison (d). Based on conserved starting units (e) a set of paths, representing the putative initial biosynthesized carbon chain of the metabolite is computed for each molecule (f). All paths a stored in a specific format, referred as matrix (g) and compared with the predicted initial chain (h).

Benchmark

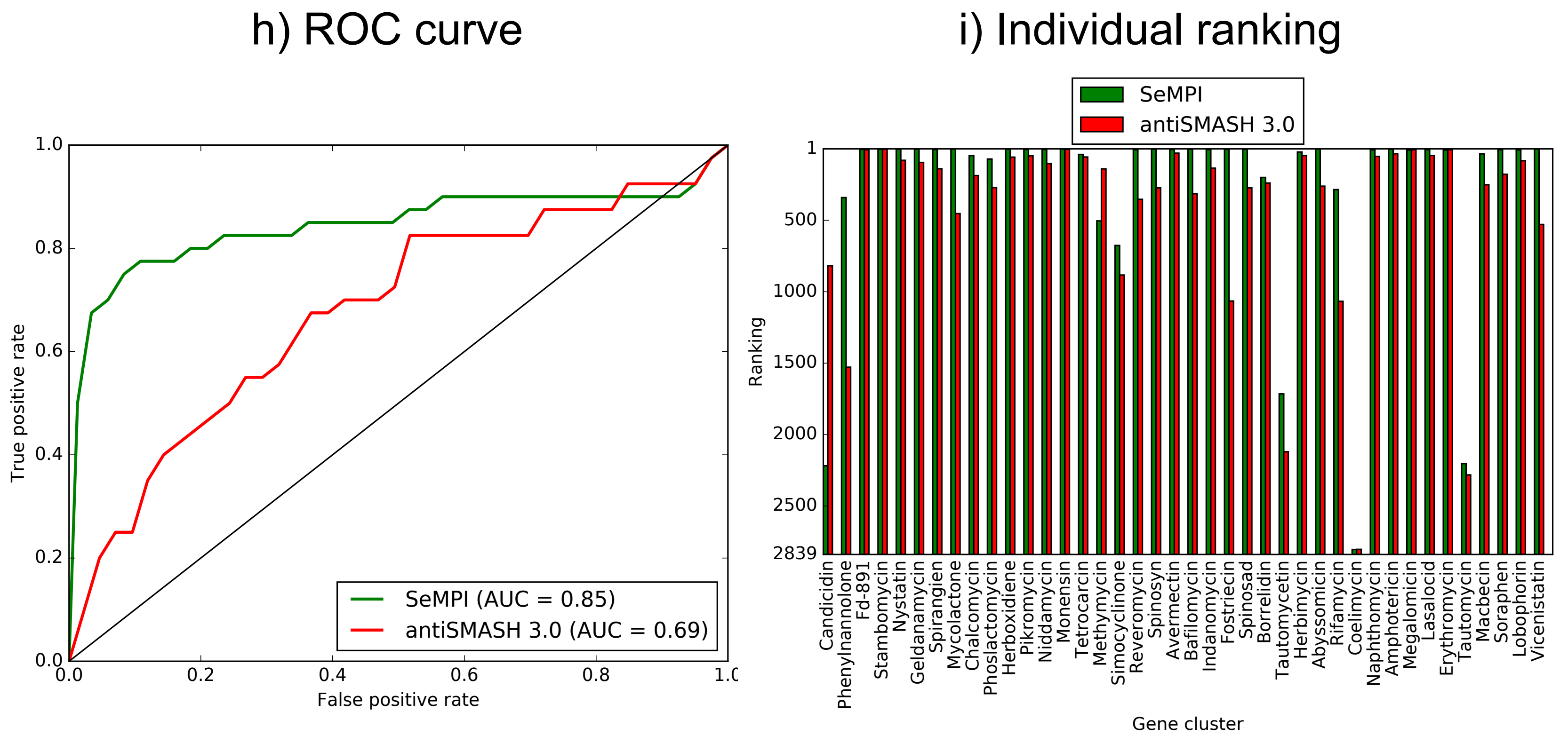
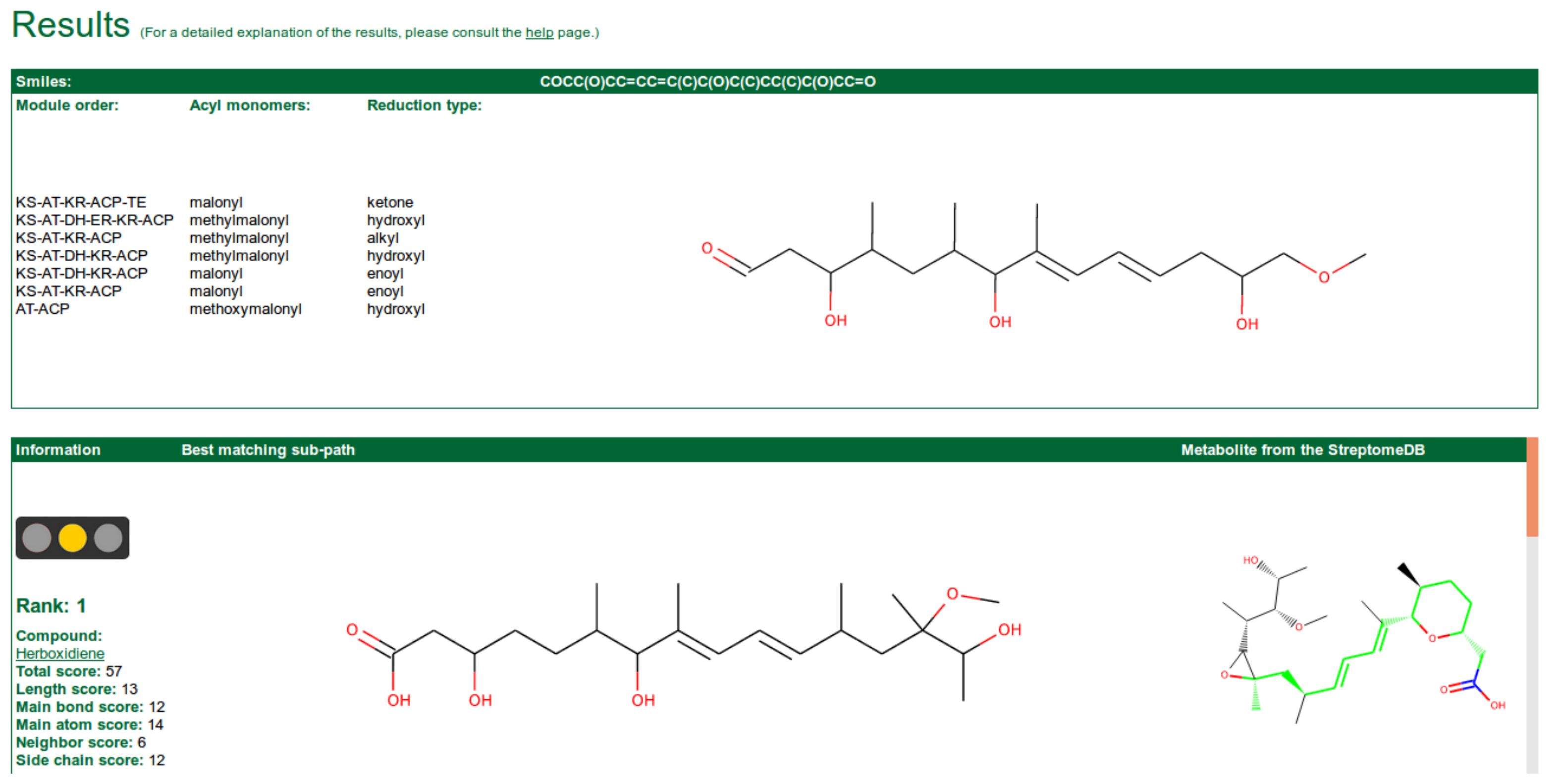


Figure 3. Based on the 40 gene clusters from the test data-set, SeMPI reached an AUC-value of 0.85, compared to antiSMASH with an AUC-value of 0.69 (l). In the individual ranking of the gene cluster products among natural products from the StreptomeDB, SeMPI ranked 27 correct gene cluster products (antiSMASH: 5) within the first ten of 2839 possible ranks (m).

Website output



Results (For a detailed explanation of the results, please consult the help page.)

Smiles: COCC(O)CC=CC=C(C)C(O)C(C)CC(C)C(O)CC=O

Module order:	Acyl monomers:	Reduction type:
KS-AT-KR-ACP-TE	malonyl	ketone
KS-AT-DH-ER-KR-ACP	methylmalonyl	hydroxyl
KS-AT-KR-ACP	methylmalonyl	alkyl
KS-AT-DH-KR-ACP	methylmalonyl	hydroxyl
KS-AT-DH-KR-ACP	malonyl	enoyl
KS-AT-KR-ACP	malonyl	enoyl
AT-ACP	methoxymalonyl	hydroxyl

Information | Best matching sub-path | Metabolite from the StreptomeDB

Rank: 1
Compound: Herboxidiene
Total score: 57
Length score: 13
Main bond score: 12
Main atom score: 14
Neighbor score: 6
Side chain score: 12

Figure 4. Besides comprehensive information about the predicted chain, SeMPI also lists matching molecules form the database, based on a ranking of their best scoring paths. This provides researchers with extensive data about the investigated gene cluster and might help to identify the actual secondary metabolite or to indicate the possibility for a not-yet annotated compound.

Scan & Visit SeMPI

References

[1] Weber, T., Blin, K., Duddela, S., Krug, D., Kim, H.U., Bruccoleri, R., Lee, S.Y., Fischbach, M.A., Muller, R., Wohlleben, W. *et al.* (2015) antiSMASH 3.0-a comprehensive resource for the genome mining of biosynthetic gene clusters. Nucleic Acids Res., 43, W237-243.

[2] Klementz, D., Doring, K., Lucas, X., Telukunta, K.K., Erxleben, A., Deubel, D., Erber, A., Santillana, I., Thomas, O.S., Bechthold, A. *et al.* (2016) StreptomeDB 2.0--an extended resource of natural products produced by streptomycetes. Nucleic Acids Res., 44, D509-514.

Pharmazeutische Bioinformatik
www.pharmaceutical-bioinformatics.de

DFG
C-factors RTG1976
UNI FREIBURG