**DRAFT:**

**Definition for antiSMASH specific domain feature for storage in BioSQL**

For storing the antiSMASH annotation in the database I need to have separate domain features containing the annotations and results of the domain specific analyses (at the moment is this mainly the nrpspks module); I would like to implement a general active site detection module and maybe also a trans-AT KS domain classification module which also would add data to this feature.

Therefore, I would propose defining a new “aSDomain” feature containing all the domain-annotation/prediction we can generate. This feature would also be directly exported to BioSQL allowing domain specific queries

In the EMBL/Genbank Exporters, this feature can be easily converted to a standard CDS\_motif or misc\_feature.

**Feature**

aSDomain start..end

**Qualifiers**

Mandatory:

/domain="Condensation" # Domain type (MBiG

keyword or separate

MBiG qualifier??)

/asDomain\_id=locus\_tag.number # unique domain id

/locus\_tag="test\_1" # parent locus tag

/translation="AGDA..." # Translation

optional:

/domain\_subtype=”Cglyc” # Domains subtype (if

applicable)

/description="blablabla" # Description of domain

(e.g. PFAM desc line)

/evalue=1.23 # Evalue

/score=123 # Score

/detection="HMMer" # Tool used to identify

domain

/database="Pfam-A" # Database that was

searched

/database\_version=1.23 # Database version that

was searched

/note="blablabla" # Unspecified notes

/activesite\_XXX # Annotation from the

active site prediction

module; all actions

defined in a XML

configuration file

output for example:

/activesite\_aa\_pos="23,24,25" # list of active site

aa-positions

/acitvesite\_aa="D,A,S" # amino acids at active

site positions

/activesite\_reference="A,B,C" # reference amino acids

/activesite\_activity="active" # activity prediction

(e.g. active, if aa

matches reference,

dubious if not )

/activesite\_specificity="D-configuration" # specificity prediction

base on active site aa

/specificity="<tool>: result" # specificity

predictions based on

other tools

output for example:

/specificity="NRPSpredictor2 SVM: hpg"

/specificity="Stachelhaus code: hpg"

/specificity="Minowa: hpg"

/specificity="consensus: hpg"

One could optionally also add experimental evidences and references to the prediction methods/ amino acids…

**Implementation in existing antiSMASH code:**

For a short time solution it would be trivial to modify the fullhmmer module to generate such entries for all PFAM hits.

For database prototype I already implemented a workaround in the specific\_analysis.py script in the nrpspks module, which adds a CDS\_motif for the domains recognized by this module in addition to the annotation in the sec\_met feature. Also, here the information could easily be added and the syntax could be adapted accordingly.

**Future plans (and basically why I need input from you):**

When the active-site prediction module is implemented, it would be desirable that the nrpspks prediction tools also directly add their information to the asDomain feature during processing (and not as implemented in my workaround) as last step with the information extracted from the sec\_met feature annotation. Thus the asDomain feature would replace the very complex pksnrpsvars dictionary/list/data structure.