

PEP2PATH TEST DATA SETS

1. The stendomycin sequence tag

As a first way to test the capabilities of Pep2Path, you can test the stendomycin extended amino acid tag V-V-T(S)-T(S)-A-I(L)-V-G to search the GenBank-wide Pep2Path database at different taxonomic levels, as is shown in the manuscript.

To search for such a extended tag, you can call `nrp2path.py` as follows:

```
python nrp2path.py --db genbank.ppd  
Val,Val,Thr/Ser,Thr/Ser,Ala,Ile/Leu,Val,Gly
```

When searching throughout the entire set of NRPS gene clusters without taxonomic specification, this will yield the following results (top 5 shown), with the biosynthetic gene cluster (BGC) for stendomycin (GG657754_c10) as the best match with a Nrp2Path score of 7.54, compared with scores between 7.47 and 5.79 for hits ranked 2nd to 5th:

When using the command

```
python nrp2path.py --db genbank.ppd --taxonomy Streptomyces  
Val,Val,Thr/Ser,Thr/Ser,Ala,Ile/Leu,Val,Gly
```

to delimit the search to *Streptomyces* BGCs, the difference between the stendomycin BGC and the runners-up becomes much more pronounced:

Sequence tag 1: Val,Val,Thr/Ser,Thr/Ser,Ala,Ile/Leu,Val,Gly					
Rank	Accession	Cluster nr	Description	Nrp2Path Score	Colinearity Score
	Sequence tag	Matching AA	Matching domains	NRPSPredictor2 code predictions	NRPSPredictor2 SVM predictions
1.	GG657754	c10	Streptomyces hygroscopicus ATCC 53653 genomic scaffold supercont1.1, whole genome shotgun sequence.	7.53553517418	1.0
			Val,Val,Thr/Ser,Thr/Ser,Ala,Ile/Leu,Val,Gly	gly, val, ile/leu, ala, thr, thr, val, val	
32_A1	EFL21631_A1	EFL21631_A2,EFL21631_A3,EFL21631_A4,EFL21631_A5,EFL21631_A6,EFL21631_A7,EFL21631_A8	gly, val, N/A, ala, thr, thr, val, val	gly, val, val, ala, thr, thr, val, val	
2.	AM746336	c1	Streptomyces collinus kirromycin biosynthesis gene cluster, strain Tu 365.	5.49641118813	0.2
			Val,Val,Thr/Ser,Thr/Ser,Ala,Ile/Leu,Val,Gly	thr, ser, ala, ile, val, gly	
	CAN89656_A1	CAN89656_A2,CAN89663_A1,CAN89638_A1,CAN89662_A1,CAN89633_A1	CAN89656_A1, CAN89656_A2, CAN89663_A1, CAN89638_A1, CAN89662_A1, CAN89633_A1	thr, ser, N/A, gly, val, gly	thr, ser, ser, gly, val, gly
3.	CM000950	c17	Streptomyces pristinaespiralis ATCC 25486 chromosome, whole genome shotgun sequence. 4.85734270255	0.25	0.25
			Val,Val,Thr/Ser,Thr/Ser,Ala,Ile/Leu,Val,Gly	ile/leu, ala, ser, thr, val	EDY67138_A1,EFH32277_A1,EFH32236_A1,EDY67106_A1,EDY67106_A2
			N/A, N/A, ser, thr, abu	met, ser, ser, thr, abu	
4.	AGSW01000262	c1	Streptomyces sp. W007 contig00280, whole genome shotgun sequence.	4.59308354132	0.0
			Val,Val,Thr/Ser,Thr/Ser,Ala,Ile/Leu,Val,Gly	gly, val, ile, ala, thr, thr	

```

EHM24052_A1,EHM24082_A1,EHM24059_A1,EHM24066_A1,EHM24083_A1,EHM24087_A1
gly,N/A,ile,N/A,thr,N/A    gly,dpg,ile,pro,thr,tyr
5.   AOP201000014      c1      Streptomyces aurantiacus JA 4570 Seq14, whole genome shotgun
sequence.        4.48304176461  0.6      Val,Val,Thr/Ser,Thr/Ser,Ala,Ile/Leu,Val,Gly
gly,Val,ile,ala,ser,ser
EPH46598_A1,EPH46610_A1,EPH46597_A1,EPH46597_A2,EPH46596_A1,EPH46596_A2
gly,N/A,gly,ala,N/A,ser  val,Val,gly,ala,ser,ser

```

The size of the tag can be varied, by taking a contiguous substring out of the original tag, e.g. T(S)-T(S)-A-I(L). Depending on the specific tag and tag size chosen, you will get different search results. Figure 2 of the manuscript shows the average result for all possible tags of sizes 2-8 when searching either at the genome level (within *Streptomyces hygroscopicus* ATCC 53653) or at the genus level (within *Streptomyces*) using the --taxonomy option available from the command-line.

Instead of the extended amino acid tag, one might also want to use the full original mass shift sequence, from which a larger number of tags are generated (including more rare amino acids). This also removes the need for a manual interpretation step. The mass shift sequence obtained for stendomycin by Kersten et al. is '99-99-83-83-71-113-99-57'. Hence, you can run nrp2path with, e.g., the following command:

```
python nrp2path.py --db genbank.ppd --taxonomy Streptomyces
99,99,83,83,71,113,99,57
```

This will yield the following results:

Rank	Accession	Cluster nr	Description	Nrp2Path Score	Colinearity Score
	Sequence tag	Matching AA	Matching domains	NRPSPredictor2	code predictions
	NRPSPredictor2	SVM predictions			
1.	GG657754 c10	Streptomyces hygroscopicus ATCC 53653 genomic scaffold supercont1.1, whole genome shotgun sequence.	12.1692710096	1.0	99,99,83,83,71,113,99,57
32_A1	gly,Val,ile/leu/val,ala,thr,thr,Val,Val	EFL21631_A1,EFL21631_A2,EFL21631_A3,EFL21631_A4,EFL21631_A5,EFL21631_A6,EFL21631_A7,EFL21631_A8			
2.	GG657754 c7	Streptomyces hygroscopicus ATCC 53653 genomic scaffold supercont1.1, whole genome shotgun sequence.	9.43698943739	0.571428571429	99,99,83,83,71,113,99,57
50_A4	iva,iva,thr,thr/ser,ala,Val,ala,gly	EFL21149_A1,EFL21153_A1,EFL21152_A1,EFL21150_A1,EFL21150_A2,EFL21150_A3,EFL21150_A4			
3.	JX508597 c1	Streptomyces griseoviridis culture-collection NRRL:2427 griseoviridin and viridogrisein Sgv gene cluster, complete sequence.	9.31003856982	0.333333333333	
4.	AGN74898_A1,AGN74886_A1,AGN74892_A1,AGN74880_A1,AGN74895_A1,AGN74885_A1,AGN74885_A2	cys,N/A,gly,pip,ser,gly,N/A	cys,tyr,gly,sal,ser,gly,Val		
0.857142857143	99,99,83,83,71,113,99,57	gly,Val,Val,gly,thr,thr/ser,Val,iva			
00_A4	AGP60631_A2,AGP60631_A3,AGP60631_A4,AGP60631_A5,AGP60700_A1,AGP60700_A2,AGP60700_A3,AGP60700_A4	ala,Val,ala,N/A,thr,N/A,Val,N/A	ala,Val,ala,Tyr,thr,glu,Val,Tyr		
5.	NZ_ALNP01000029 c1	Streptomyces sp. AA0539 contig00029, whole genome shotgun sequence.	8.58449212445	0.857142857143	99,99,83,83,71,113,99,57
44_A1,WP_019435944_A2,WP_019435944_A3	val,Val,thr,thr,ala,leu/val,Val,Val	WP_020699878_A1,WP_020699878_A2,WP_020699878_A3,WP_020699878_A4,WP_020699878_A5,WP_019435944_A1,WP_019435944_A2,WP_019435944_A3	gly	val,Val,thr,thr,Val,N/A,Val,Val	
				val,Val,thr,thr,Val,Val,Val	

Note that the score is different from the score for the previously used tag "Val,Val,Thr/Ser,Thr/Ser,Ala,Ile/Leu,Val,Gly". The reason for this is that there is a larger number of redundant amino acids in the tag: the mass shift tag 99,99,83,83,71,113,99,57 in fact translates to "Val/Ala/Iva/Ile/Leu,Val/Ala/Iva/Ile/Leu,Thr/Dhb/Ser,Thr/Dhb/Ser,Ala/B-Ala/Gly,Abu/Pro/Ile/Leu/Val,Val/Ala/Iva/Ile/Leu,Gly". The consequence is that in the new situations more tag positions are calculated with baseline P(A) values, as the amino acid is not considered an observation.

2. Eighteen recently discovered nonribosomal peptides.

A second data set that can be used to test Pep2Path is constituted by the benchmarking dataset of 18 recently discovered nonribosomal peptides outlined in Supplementary Table 1 of the manuscript.

Peptide	(Sub)phylum	Species	Gene cluster (NCBI accession + antiSMASH gene cluster number)	NRP search tag
Micropeptin	Cyanobacteria	<i>Microcystis Aeruginosa</i> K-139	AB481215_c1	Asp-Thr-Arg-Thr-Ile-Tyr-Ile Ala-Asp-Gly-Phe-Pro-Tyr-Trp-Gly-Leu-Leu
Skyllamycin	Actinobacteria	<i>Streptomyces</i> sp. Acta 2897 <i>Streptomyces pyridomyceticus</i>	JF430460_c1	
Pyridomycin	Actinobacteria	NRRL B-2517	HM436809_c1	Pip-Thr-Ala-Xxx-Xxx
Althiomycin	Deltaproteobacteria	<i>Myxococcus xanthus</i> DK897	FR831800_c1	Gly-Cys-Ser-Cys-Gly
Hormaomycin	Actinobacteria	<i>Streptomyces griseoflavus</i> w-384	HQ542230_c1	Pro-Ala-Thr-Phe-Ala-Phe-Ile-Pro
Koranimine	Firmicutes	<i>Bacillus</i> sp. NK 2003	JF828091_c1	Thr-Leu-Leu-Phe-Val-Phe-Val
Nostophycin	Cyanobacteria	<i>Nostoc</i> sp. Strain 152 <i>Streptomyces roseosporus</i> NRRL	JF430079_c1	Gln-Gly-Pro-Phe-Ile-Pro
Arylomycin	Actinobacteria	11379	NZ_ABYX01000222_c1	Ser-Ala-Gly-Hpg-Ala-Tyr Ser-Leu-Val-Ser-Val-Ser-Tyr-Lys-Ala
Myxoprincomide	Deltaproteobacteria	<i>Myxococcus xanthus</i> DK 1622	NC_008095_c11	
Nodularin	Cyanobacteria	<i>Nostoc</i> sp. 73.1	JF342711_c1	Asp-Arg-Xxx-Glu-Ser
Napsamycin	Actinobacteria	<i>Streptomyces</i> sp. DSM5940	HQ287563_c1	Tyr-Met-Dab-Trp
Cupriachelin	Betaproteobacteria	<i>Ralstonia eutropha</i> H16	NC_008314_c1	Xxx-Asp-Dab-Asp-Gly-Orn
Caerulomycin	Actinobacteria	<i>Actinoalloteichus</i> sp. WH1-2216-6	JF419316_c1	Cys-Thr
Collismycin	Actinobacteria	<i>Streptomyces</i> sp. CS40	HE575208_c1	Cys-Leu
Rhodochelin	Actinobacteria	<i>Rhodococcus jostii</i> RHA1	NC_008268_c8	Thr-Orn
Nocobactin	Actinobacteria	<i>Nocardia farcinica</i> IFM 10152	NC_006361_c3	Lys-Lys-Thr
Paenibactin	Firmicutes	<i>Paenibacillus eigii</i> B69	HQ668144_c1	Dhb-Ala-Thr
Mycobactin	Actinobacteria	<i>Mycobacterium tuberculosis</i> H37Rv	NC_000962_c9	Sal-Thr-Lys

This data set contains both peptides with long sequence tags (for which it is easier to identify their BGCs) and peptides with short sequence tags (for which it is more difficult to identify their BGCs).

Let us look at one example of each.

The skyllamycin molecule consists of 10 amino acids, which together form a sequence tag 'Ala-Asp-Gly-Phe-Pro-Tyr-Trp-Gly-Leu-Leu'. To search for the BGC of this molecule in the entire GenBank database, we use the following command:

```
python nrp2path.py --db genbank.ppd Ala,Asp,Gly,Phe,Pro,Tyr,Trp,Gly,Leu,Leu
```

This immediately leads to the identification of a very good hit which stands out from the lower-ranking hits (score of 5.48 compared to other scores <4.32):

```
Sequence tag 1: Ala,Asp,Gly,Phe,Pro,Tyr,Trp,Gly,Leu,Leu
1. JF430460 c1 Streptomyces sp. Acta 2897 skyllamycin biosynthetic gene cluster,
complete sequence. 5.4825289332 1.0 ala,asp,gly,phe,pro,tyr,trp,gly,leu,leu
ala,asp,gly,phe,pro,tyr,trp,gly,leu,leu
AEA30272_A2, AEA30272_A3, AEA30272_A4, AEA30273_A1, AEA30273_A2, AEA30273_A3, AEA30273_A4, AEA30274_A1, AEA30274_A2, AEA30274_A3 ala,asn,gly,phe,pro,N/A,N/A,gly,N/A,N/A
ala,asp,gly,phe,pro,phe,trp,gly,tyr,tyr
2. ANKQ01000001 c4 Microcystis aeruginosa TAIHU98 contig1, whole genome shotgun
sequence. 4.31423570852 0.6 ala,asp,gly,phe,pro,tyr,trp,gly,leu,leu
asp,gly,phe,pro,tyr,trp
ELP56809_A1, ELP56284_A1, ELP56284_A2, ELP57196_A1, ELP57196_A2, ELP55719_A1
N/A, val, tyr, pro, tyr, tyr leu, val, tyr, pro, tyr, tyr
3. NZ_AOUG01000020 c1 Burkholderia pseudomallei NCTC 13392 NCTC13392_28, whole genome
shotgun sequence. 4.13566642421 0.0 ala,asp,gly,phe,pro,tyr,trp,gly,leu,leu ala,asp,gly
WP_017844181_A1, WP_017844174_A1, WP_017844180_A1 ala,asp,gly ala,asp,gly
4. CP003782 c7 Burkholderia pseudomallei MSHR305 chromosome 1, complete sequence.
4.13566642421 0.0 ala,asp,gly,phe,pro,tyr,trp,gly,leu,leu ala,asp,gly
AGR68580_A1, AGR67582_A1, AGR69939_A1 ala,asp,gly ala,asp,gly
5. CP003782 c10 Burkholderia pseudomallei BPC006 chromosome II, complete sequence.
4.13566642421 0.0 ala,asp,gly,phe,pro,tyr,trp,gly,leu,leu ala,asp,gly
AFR19519_A1, AFR19491_A1, AFR19517_A1 ala,asp,gly ala,asp,gly
```

...

On the contrary, when we use the very short and much more unspecific ‘Cys-Thr’ sequence tag of caerulomycin as a query for a GenBank-wide search, using the command

```
python nrp2path.py --db genbank.ppd Cys,Thr
```

we get a wide range of almost equally possible hits. Here are the top five hits for this search, with scores between 4.40 and 4.53:

```
Sequence tag 1: Cys,Thr
1.    CP001630 c11      Actinosynnema mirum DSM 43827, complete genome.      4.52672830483      0.0
      cys,thr cys,thr ACU37567_A2,ACU37564_A1  cys,thr cys,thr
2.    CAGD01000003   c1      Brevibacillus laterosporus GI-9 WGS project CAGD00000000 data,
contig BLGIcontig03, whole genome shotgun sequence.4.48198383548      0.0      thr,cys thr,cys
      CCF13030_A1,CCF12997_A1  thr,cys thr,cys
3.    AFRV01000004   c1      Brevibacillus laterosporus LMG 15441 BRLAcontig4c, whole genome
shotgun sequence.4.48198383548      0.0      thr,cys thr,cys EGP33003_A1,EGP32975_A1  thr,cys
      thr,cys
4.    KF264553 c1      Uncultured bacterium esnapd14 genomic sequence.      4.45274478818      0.0
      cys,thr cys,thr AGS49688_A2,AGS49687_A1  cys,thr cys,thr
5.    AJGF01000005   c2      Mycobacterium abscessus M93 Contig05, whole genome shotgun
sequence.      4.39663391462      0.0      cys,thr cys,thr EIC67550_A2,EIC67544_A1  cys,thr
      cys,thr
```

If we have taxonomic information on the organism from which the peptide was derived, this helps in identifying the correct gene cluster nonetheless. In this case, we know that the molecule derives from the genus *Actinoalloteichus*, so we can use this to do a more specific search:

```
python nrp2path.py --db genbank.ppd --taxonomy Actinoalloteichus Cys,Thr
```

Now, the top five contains three copies of the actual caerulomycin gene cluster in different *Actinoalloteichus* species, with equal scores (4.15). The next hits, which correspond to altogether unrelated gene clusters, have scores of just 2.39 and 0.12:

```
Sequence tag 1: Cys,Thr
1.    NZ_AVGX02000004 c1      Actinoalloteichus spitiensis RMV-1378 Contig4, whole genome
shotgun sequence.4.15293228248      1.0      cys,thr cys,thr WP_016696705_A1,WP_016696706_A1
      cys,thr cys,thr
2.    JQ687072 c1      Actinoalloteichus cyanogriseus strain NRRL B-2194 methyltransferase
(caeG2), transporter (caeH3), transcriptional regulator (caeI2), ABC transporter (caeH1), ABC
transporter (caeH2), acyl-CoA dehydrogenase (caeB5), methyltransferase (caeG1), aminotransferase
(caeC), FAD-dependent oxidoreductase (caeB6), NrpS (caeA1), L-lysine 2-amino transferase (caeP1),
FAD-dependent oxidoreductase (caeP2), PKS/NrpS (caeA2), NrpS (caeA3), acyl-CoA dehydrogenase
(caeB1), thioesterase (caeA4), LuxR family two component transcriptional regulator (caeI1),
amidohydrolase (caeD), AMP-dependent ligase (caeF), aldehyde dehydrogenase (caeB2), FAD-dependent
oxidoreductase (caeB3), F420-dependent NADP oxidoreductase (caeB4), transcriptional regulator
(caeI3), and monooxygenase (caeB7) genes, complete cds.      4.15293228248      1.0      cys,thr
      cys,thr AFK24516_A1,AFK24517_A1  cys,thr cys,thr
3.    JF419316 c1      Actinoalloteichus sp. WH1-2216-6 putative formate acetyltransferase gene,
partial cds; 50S ribosomal protein and ubiquinone/menaquinone biosynthesis methyltransferase
genes, complete cds; caerulomycin A biosynthetic gene cluster, complete sequence; and F420-
dependent NADP oxidoreductase coenzyme gene, partial cds.      4.15293228248      1.0      cys,thr
      cys,thr AFD30954_A1,AFD30953_A1  cys,thr cys,thr
4.    NZ_AVGX02000207 c1      Actinoalloteichus spitiensis RMV-1378 Contig207, whole genome
shotgun sequence.2.39341063725      1.0      cys,thr cys,thr WP_016698708_A1,WP_016698709_A1
      N/A,thr trp,thr
5.    NZ_AVGX02000192 c1      Actinoalloteichus spitiensis RMV-1378 Contig192, whole genome
shotgun sequence.0.119351846221      N/A      cys,thr cys      WP_016698577_A1  leu      val
```

Table 1 outlines the results of a procedure in which all possible (sub)tags of sizes 2-8 from the set of 18 NRPs were used on 50 randomly generated search spaces of sizes 5 to 100.

3. Nine recently discovered RiPPs.

Finally, the set of search tags for nine *Streptomyces* RiPPs outlined in Table 3 of the manuscript offer a way to test the ripp2path script.

Peptide	Search tag	Genome
SSV-2083	I(L)GA(C)GTA(C)WI(L)A(C)V	<i>Streptomyces sviceus</i> ATCC 20983
SGR-1832	AVAQ(K)FVI(L)Q(K)GSTI(L)	<i>Streptomyces griseus</i> IFO 13350
SCO-2138	VHFVGWI(L)	<i>Streptomyces coelicolor</i> A3(2)
SLI-2138	GI(L)VHFVGWI(L)	<i>Streptomyces lividans</i> TK24
SWA-2138	I(L)AGI(L)VHFI(L)GWI(L)	<i>Streptomyces</i> sp. E14 (WASP)
SRO15-2005	YWSRRI(L)(L)	<i>Streptomyces roseosporus</i> NRRL 15998
SRO15-2212	VVI(L)S(C)T	<i>Streptomyces roseosporus</i> NRRL 15998
SRO15-3108	AS(C)ATVTI(L)	<i>Streptomyces roseosporus</i> NRRL 15998
SAL-2242	VTI(L)S(C)T	<i>Streptomyces albus</i> J1074

The genomes of the corresponding *Streptomyces* strains have been included within the ‘RiPP_genomes’ subdirectory. One can run ripp2path on these peptide search tags, by e.g. using the command

```
python ripp2path.py --tags Val,His,Phe,Val,Gly,Trp,Ile/Leu --seq  
TestData/RiPP_genomes/S_coelicolor.gbk
```

for the SCO-2138 peptide and its VHFVGWI(L) search tag.

This leads to a unique match in the *Streptomyces coelicolor* A3(2) genome, as can be seen in the following results:

Rank	Accession sequence	Start position(s) % Identity	Strand	Query tag sequence	Matching
1	AL645882	2625565 +		VHFVGWL	100
2	AL645882	2625565 +		VHFVGWL	85
3	AL645882	4858516 -		VGFVGWI	85
4	AL645882	2233450 -		WHFVGQL	71
5	AL645882	5710464 -		VWFIGWL	71
6	AL645882	1290558 -		VVGVGWL	71
7	AL645882	6742392 -		VVFGLWL	71
8	AL645882	2505024 -		VVEVGWL	71
9	AL645882	7271690 +		VTVVGWL	71
10	AL645882	7455987 -		VTFVGAL	71
...					

The right column in Table 3 of the manuscript shows the number of 100% matches obtained with each of the nine RiPP extended sequence tags in the table at the top of this page.