[Spine][http://vfsmspineagent.fsm.northwestern.edu/index_age.html]

Spine用于识别细菌或其他小基因组物种的保守核心基因序列

Usage

`perl spine.pl –f genome_files.txt`

-f 输入序列文件，可以为fasta序列(fasta)，包含genebank序列和注释文件的genebank格式文件(gbk)，或分开的fasta文件伴有对应的gff3格式注释文件(comb)

```
path/to/file1<tab>unique_identifier<tab>fasta or gbk or comb
path/to/file2<tab>unique_identifier<tab>fasta or gbk or comb
```

```
1 /Data_analysis/k.pnuemoniae_divergetn_mic/filtered_cef_avi_analysis_confined/38377_prokka/38377.gbk     38377   gbk
2 /Data_analysis/k.pnuemoniae_divergetn_mic/filtered_cef_avi_analysis_confined/38588_prokka/38588.gbk     38588   gbk
3 /Data_analysis/k.pnuemoniae_divergetn_mic/filtered_cef_avi_analysis_confined/42395_prokka/42395.gbk     42395   gbk
4 /Data_analysis/k.pnuemoniae_divergetn_mic/filtered_cef_avi_analysis_confined/36170_prokka/36170.gbk     36170   gbk
5 /Data_analysis/k.pnuemoniae_divergetn_mic/filtered_cef_avi_analysis_confined/39401_prokka/39401.gbk     39401   gbk
6 /Data_analysis/k.pnuemoniae_divergetn_mic/filtered_cef_avi_analysis_confined/38218_prokka/38218.gbk     38218   gbk
```

-a/--pctcore 输入基因组序列被认定核心基因组区域的最小百分比，默认为100

-g/--maxdist 核心基因组segments之间的最大距离，相邻segments之间小于该距离将合并成combined fragments，之间使用N填充；否则就分为两个或多个fragments，默认10

-r/--refs 作为主要输出的参考基因组序列，参考基因组序列将作为主要的backbone序列来源；通过整数值来对应输入文件的基因组顺序，可使用逗号隔开多个输入值，例如:1,3，序列1位具有最好的优先性，序列3位第二优先参考序列。默认参考基因组的优势排序同输入文件中基因组的排序，第一个基因组拥有最好的参考优势

--mini 仅输出backbone序列，来源于参考基因组。当仅需要backbone序列去获得accessory sequences时，开采用该选型以节约时间。默认针对所有包含的基因组，输出core和accessory序列集

--pangenome 输出pangenome序列，默认不输出

-o/--prefix 输出结果前缀

-p/--pctid 判定为同源性的最小的一致性区域，默认85

-s/--minout 输出的最小的核心区域的大小，单位base，默认10

输出文件

statistics.txt

```
1 Spine version: 0.3.2
2 inputs: --pctcore 0 --refs 1,2,3,4,5,6 --maxdist 10 --pctid 85 --minout 10
3
```

coords.txt 基因组序列的坐标位置，包含：".accessory_coords.txt", ".core_coords.txt", "backbone_coords.txt", "pan genome_coords.txt"

	contig_id	contig_length	start	stop	source_gen	out_seq_id
2	rec1	1865238	1	267	38377	backbone_0001_length_267
3	rec1	1865238	394	87421	38377	backbone_0002_length_87028
4	rec1	1865238	127426	325713	38377	backbone_0003_length_198288
5	rec1	1865238	325725	326065	38377	backbone_0004_length_341
6	rec1	1865238	326101	353396	38377	backbone_0005_length_27296
7	rec1	1865238	365499	391521	38377	backbone_0006_length_26023
8	rec1	1865238	391719	463681	38377	backbone_0007_length_71963

*.fasta 输出基因组片段的核酸序列文件

```
1 >39401_core_0001_length_423
2 ttgcctggcggcactagcgcggtggtcccacctgaccccatgccgaactcagaagtgaaacgccgtagcgccgatggtagtgtggggtctcc
3 >39401_core_0002_length_5908
4 tagcgggtcagagcaacgatccctaacgtgaggccggctaccgcggccgcacccccggcgattgcgcctgcggtttcccagttgtcttctgt
5 >39401_core_0003_length_28794
6 taataaaaaaggcgctatcccatgccgagtagcgccttttttattcaataacatagctgaaatgtatcagttcatgccgtatttttcagttt
7 >39401_core_0004_length_66864
8 accgggtttctggtcagattcccggccggtccggtatttagcgattaacgtggctccccaaacggggagccgtttagttggtggccgggtat
```

loci.txt 核心基因组中的coding序列，包含".accessory_loci.txt", ".core_loci.txt", "pangenome_loci.txt", "backbone_loci.txt"

	locus_id	gen_contig_id	gen_contig_start		gen_contig_stop	strand	out_seq_id	out_seq_start	out_seq_stop	pct_locus	overhangs	product
2	GHLKHLGD_00004	rec1	592	812	-	39401_core_0002_length_5908	1	221	99.55	1,0		hypothetical protein
3	GHLKHLGD_00005	rec1	1106	4216	-	39401_core_0002_length_5908	515	3625	100.00	0,0		Multidrug export protein AcrF
4	GHLKHLGD_00006	rec1	4229	5368	-	39401_core_0002_length_5908	3638	4777	100.00	0,0		Multidrug export protein AcrE
5	GHLKHLGD_00007	rec1	5747	6397	+	39401_core_0002_length_5908	5156	5806	100.00	0,0		HTH-type transcriptional regulator AcrR
6	GHLKHLGD_00031	rec1	22480	22776	-	39401_core_0003_length_28794	66	362	100.00	0,0		DNA-binding protein Fis
7	GHLKHLGD_00032	rec1	22801	23766	-	39401_core_0003_length_28794	387	1352	100.00	0,0		putative tRNA-dihydrouridine synthase
8	GHLKHLGD_00033	rec1	24124	25005	-	39401_core_0003_length_28794	1710	2591	100.00	0,0		Ribosomal protein L11 methyltransferase
9	GHLKHLGD_00034	rec1	25017	26468	-	39401_core_0003_length_28794	2603	4054	100.00	0,0		Sodium/pantothenate symporter

position_counts.txt

[AGEnt][https://github.com/egonozer/AGEnt]

AGEnt用于从core genome中识别accessory genomic elements

Usage

```
perl AGEnt.pl -r core_genome.fasta -q query_genome.fasta
```

-q fasta格式或者genebank格式的输入query序列文件。如果使用注释后的genebank格式文件，AGEnt将会提取CDS坐标来将gene分至core或acessory groups。如果时genebank文件，CDS坐标必须拥有"locus_id"标签用于gene信息提取。针对RAST输出，不含"locus_id"标签，可使用[gbk_reformat][http://vfsmspineagent.fsm.northwestern.edu/download.html]程序

Genbank Reformat prepares Genbank files to be used in SPINE or AGEnt. Performs a number of functions:

- **1:** Joins multi-part sequences (i.e. contigs or chromosomes / plasmids) into a single sequence. *This is no longer necessary for Spine or AGEnt. You can leave the setting as 'No'.*
- **2:** Adds `locus_tag` tags to features where they were not included (i.e. as output by RAST).

WARNING: Genbank files that already have `locus_tag` tags on any features will not have any new tags added.

- **3:** Orders features by coordinates rather than by feature type. Again, RAST does this.

-r core/referenc序列，fasta格式或genebank格式

-b 同时输出core 序列和坐标，默认仅输出accessory 序列和坐标

-c 包含query genome中的基因名称和坐标文件路径，将输出一个文件将基因分配到core和accessory中，默认文件格式为"glimmer"格式：

```
>contig_name_1
orf_ID_1<space(s)>start_coordinate<space(s)>stop_coordinate
orf_ID_2<space(s)>start_coordinate<space(s)>stop_coordinate
>contig_name_2
orf_ID_3<space(s)>start_coordinate<space(s)>stop_coordinate
etc...
```

同时可采用-q/-c指定输入序列及注释信息

```
-q /path/to/chrom_I.fasta,/path/to/chrom_II.fasta,/path/to/plasmid.fasta
```

```
-c /path/to/chrom_I.gff3,/path/to/chrom_II.gff3,/path/to/plasmid.gff3
```

-f 由-c选项指定的ORF坐标文件格式：'glimmer', 'genebank', prodigal', 'gff'，默认为glimmer

-m 将query序列判定为core时，query和reference之间最小的比对一致性，默认为85

-o 输出文件前缀，默认为'output'

-p 输出序列的前缀，默认同-o指定前缀

-s 最小输出片段的大小，单位base，默认10

输出文件

statistics.txt，coords.txt，'*.fasta'，loci.txt同spine和ClustAGE

[ClustAGE][https://sourceforge.net/projects/clustage/files/]

ClustAGE将一组来自细菌或其他小的基因组物种的accessory genomic element(AGEs)聚类并识别最小的accessory genomic elements，同时识别这些elements在所提供基因组的序列中的分布情况。

Usage

`perl ClustAGE.pl -f age_files.txt`

`ClustAGE.pl -f ClustAGE_input.txt --annot ClustAGE_annot.txt -p --graph_se -o clustage_graphs`

-f accessory genome elements fasta 文件

```
/path/to/accessory_elements_1.fasta<tab>genome_name_1<tab>(optional)rank
/path/to/accessory_elements_2.fasta<tab>genome_name_2<tab>(optional)rank
```

其中rank时指定strain的数值，可以是实际数字或者相对数字，同时该rank值可选。如果rank值设定为"R"，那么改序列将指定为reference，且属于改基因组的序列不会用于作为bin representatives，但是该基因组针对bin representatives的比对序列将会报出

--annot 可选输入文件，该注释信息将包含在输出文件中

```
/path/to/gen1.accessory_loci.txt<tab>genome_name_1
/path/to/gen2.accessory_loci.txt<tab>genome_name_2
```

该注释文件应为Spine或AGEnt的输出格式文件, "loci.txt"

`locusID<tab>Source contig ID<tab>Source start<tab>Source stop<tab>Strand<tab>Accessory sequence ID<tab>Accessory start<tab>Accessory stop<tab>% of gene<tab>Overlap<tab>Gene product`

同时该注释文件genome_name应该和-f参数指定输入文件的genome_name一样

--age 为用作bin representatives的fasta格式的AGE序列，如果指定该文件，那么-f文件将不会用与识别新的bin representatives，且-f指定文件将比对到--age文件

-e/--evalue 最大的BLAST e-value cutoff，默认1e-6

-i/--pctid 最小的核酸序列一致性(%)，默认85%

--dustoff 取消默认使用blast针对低复杂度序列进行过滤。该选项用于含有大量低复杂度序列的物种，默认(dust masking on)

-a/--maxaign 用于报告的最小数目的blast比对。太小的值将导致不正确结果，尤其是在比较大量序列数目时，默认100000

-x/--min_age 最小的acessory size，单位bp。为用于ClustAGE指定bin representative的最短的可能序列，默认为200

-o/--out 输出文件前缀，默认out

-p/--graph 输出AGEs的图像，包含AGE在输入输入文件中的分布信息

--graph_se 输出图像包含subelement dividers信息

输出文件

AGEs.key.txt 为acessory genomic element representative的特征

bin_id	source_id	source_genome	source_length	bin_start		bin_stop		bin_length
bin1	39401_accessory_0063_length_214277	39401	214277	1		214277	214277	
bin2	38588_accessory_0125_length_107588	38588	107588	1		107588	107588	
bin3	42395_accessory_0019_length_95949	42395	95949	1		95949	95949	
bin4	42395_accessory_0128_length_110018	42395	110018	1		90527	90527	
bin5	42395_accessory_0048_length_56182	42395	56182	1		56182	56182	
bin6	38218_accessory_0010_length_54848	38218	54848	1		54848	54848	
bin7	38377_accessory_0133_length_51057	38377	51057	1		51057	51057	
bin8	42395_accessory_0016_length_46119	42395	46119	1		46119	46119	

bin_id为representative序列的唯一ID，对应AGE.fasta中的ID

AGEs.fasta 为representative AGE序列中的核酸序列

```
>bin1 39401_accessory_0063_length_214277
aactgcatcccaaaagttgaactccaactgagtaaaggtgcagtacattctgggcggggagcagtcaatgatacaaaatactgaagataaaacgaa
>bin2 38588_accessory_0125_length_107588
ttttttttttgagggaaggatatgcctgagcaaaaaatggaaatcacggacgagcagttcgctaatttcaccggaacgatgtttaacacgtcgttta
>bin3 42395_accessory_0019_length_95949
tttgctgttttcatgcatccttgcgaatcattatatgatgttggttcaacaggttagcgtgaaaactcttcccggtgcattttgattttaccctcc
>bin4 42395_accessory_0128_length_110018
gcctatccacccgttaaacaaaatgattaatatttaaaaatactcgttgaacacccgtgccgtaataatgaagtacattaaaaacgttatccggtc
>bin5 42395_accessory_0048_length_56182
```

AGEs.annotations.txt representative accessory regions内的基因注释内容

```
bin_id  annotation(s)
bin1    GHLKHLGD_02461[100.00%]"Putative ATP-dependent DNA helicase YjcD",GHLKHLGD_02462[100
bin2    MJNCPELE_04319[100.00%]"hypothetical protein",MJNCPELE_04320[100.00%]"hypothetical p
bin3    EDNNINJN_00720[100.00%]"Prophage integrase IntS",EDNNINJN_00721[100.00%]"hypothetica
bin4    EDNNINJN_04557[100.00%]"hypothetical protein",EDNNINJN_04558[100.00%]"hypothetical p
bin5    EDNNINJN_01706[100.00%]"Putative SOS response-associated peptidase YedK",EDNNINJN_01
bin6    IALAIEND_00291[100.00%]"Tyrosine-protein kinase YwqD",IALAIEND_00292[100.00%]"hypoth
bin7    LLFCEEMA_04813[100.00%]"Protein UmuC",LLFCEEMA_04814[100.00%]"Antitoxin HigA1",LLFCE
```

bin_id对应AGEs.fasta和AGEs.key.txt文件名称；annotation(s)对应所包含基因信息（对应为each entry)和百分率(对应该基因核酸长度): PA2185[100%]"non-heme catalase KatN", PA2186[100%]"heyothetical protein"

Subelements.key.txt AGEs的subelements的特性，subelements为AGEs基于AGE在输入文件中的分布的subdivisions

subelement	bin_id	source_id	source_genome	start	stop	length	avg_rank	num_genomes	38218	38588	36170	39401	42395	38377
bin1_se00001	bin1	39401_accessory_0063_length_214277		39401	1	76464	76464	NA	1	0	0	0	1	0	0
bin1_se00002	bin1	39401_accessory_0063_length_214277		39401	76465	76760	296	NA	2	0	1	0	1	0	0
bin1_se00003	bin1	39401_accessory_0063_length_214277		39401	76761	78217	1457	NA	1	0	0	0	1	0	0
bin1_se00004	bin1	39401_accessory_0063_length_214277		39401	78218	81819	3602	NA	2	0	1	0	1	0	0
bin1_se00005	bin1	39401_accessory_0063_length_214277		39401	81820	97128	15309	NA	1	0	0	0	1	0	0
bin1_se00006	bin1	39401_accessory_0063_length_214277		39401	97129	97327	199	NA	2	0	1	0	1	0	0
bin1_se00007	bin1	39401_accessory_0063_length_214277		39401	97328	212894	115567	NA	1	0	0	0	1	0	0
bin1_se00008	bin1	39401_accessory_0063_length_214277		39401	212895	213700	806	NA	2	0	1	0	1	0	0
bin1_se00009	bin1	39401_accessory_0063_length_214277		39401	213701	214159	459	NA	1	0	0	0	1	0	0

subelements.fasta subelements序列的核酸序列文件，默认对应输出至少100bp的长度序列，可有参数--min_se_seq选项调整

```
>bin1_se00001
aactgcatcccaaaagttgaactccaactgagtaaaggtgcagtacattctgggcggggagcagtcaatgatacaaaatactgaagataaaa
>bin1_se00002
accattgatcatgctgaaactgccgaggaaaatatcgatgccgttacacccgacgctggctttcgggaaactgacattgacgaggttggtat
>bin1_se00003
cagtagccacttaagccccttcactttctgaaataccggtattaacccgatcggttgttggcatgtcgtagccgctaagtttttcatccat
>bin1_se00004
tgcatcaggagtgacagctgatgcgaaagattacattttctggcagtgagtattttctcaactctcgccttacggtcagggttttttatcat
>bin1_se00005
aatggaaaatacgacgcattgccattcatccagagcattaagcggtttaaatcggtggaactttatattgccagtctcgaaactaaagcccc
```

subelements.annotations.txt 包含在subelement区域的基因信息

```
subelement      annotation(s)
bin1_se00001    GHLKHLGD_02461[100.00%]"Putative ATP-dependent DNA helicase YjcD",GHLKHLGD_0
bin1_se00002    GHLKHLGD_02538[18.58%]"hypothetical protein",GHLKHLGD_02539[0.10%]"hypotheti
bin1_se00003    GHLKHLGD_02539[99.90%]"hypothetical protein",GHLKHLGD_02540[10.13%]"hypothet
bin1_se00004    GHLKHLGD_02540[89.87%]"hypothetical protein",GHLKHLGD_02541[100.00%]"hypothe
bin1_se00005    GHLKHLGD_02544[28.52%]"hypothetical protein",GHLKHLGD_02545[100.00%]"hypothe
bin1_se00006    -
bin1_se00007    GHLKHLGD_02559[100.00%]"hypothetical protein",GHLKHLGD_02560[100.00%]"hypoth
bin1_se00008    GHLKHLGD_02675[70.08%]"hypothetical protein"
bin1_se00009    GHLKHLGD_02675[29.92%]"hypothetical protein"
```

subelements.csv 输入文件中的subelement分析信息

```
genome,rank,bin1_se00001,bin1_se00002,bin1_se00003,bin1_se00004,bin1_se00005,bin1_se00006,bi
36170,NA,0,0,0,0,0,0,0,0,0,0,0,0,1,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0
38218,NA,0,0,0,0,0,0,0,0,0,0,0,0,1,0,0,0,0,0,0,0,0,0,0,1,0,1,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0
39401,NA,1,1,1,1,1,1,1,1,1,1,0,0,1,0,0,0,0,0,0,0,0,0,0,1,0,1,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0
42395,NA,0,0,0,0,0,0,0,0,0,0,0,0,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1
38377,NA,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,1,0,1,0,1,0,1,0,1,0,0,0,0,0,0,0,1,1,0,1,0,0,1,0,0,1,0,1,1
38588,NA,0,1,0,1,0,1,0,1,0,1,1,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,1,0,0,1,0,1,1,1,0,0,0,0,1
```

subelements.alignments.txt 包含在每个输入genome的subelements的来源

	subelement	38218_contig	38218_start	38218_stop	38588_contig	38588_start	
1	bin1_se00001	-	0	0	-	0	0	3940
2	bin1_se00002	-	0	0	38588_accessory_0139_length_38348	9927	1022
3	bin1_se00003	-	0	0	-	0	0	3940
4	bin1_se00004	-	0	0	38588_accessory_0139_length_38348	10222	1381
5	bin1_se00005	-	0	0	-	0	0	3940
6	bin1_se00006	-	0	0	38588_accessory_0139_length_38348	-14416	-146
7	bin1_se00007	-	0	0	-	0	0	3940
8	bin1_se00008	-	0	0	38588_accessory_0144_length_29700	128	128
9	bin1_se00008	-	0	0	38588_accessory_0217_length_805 -1	-805	-

graphs folder 包含输出的AGE的分布图