# Annotation of Plasmid sequences

Plasmids are used in Molecular Biology and Genetics to introduce selected genetic information into cells and organisms. A typical plasmid contains at least one replication origin (to ensure autonomous replication in *E. coli*) and a selection marker (which enables to select cells which contain the plasmid). In addition the plasmid may contain additional genes, promoters and other features enabling them to be used for specific experiments.

Aim of this project is to analyse a nucleotide sequence of a plasmid (fasta or genbank file) and to annotate the plasmid by detecting all genetic features. To achieve this goal three approaches need to be combined:

> Common features: Based on a number of annotated plasmid sequences in genebank format common features of plasmids are extracted and a list of these features with annotations and sequences is build up ("learning"). The to be annotated plasmid sequences will then be analysed whether the feature is present.
>
> Protein coding genes: Protein coding genes not overlapping with the common features identified above will be identified through a BLAST request.
>
> Primer binding sites: Primers are used for sequencing and PCR amplification. A list of common primers and its sequences will be provided. Primer binding sites shall be detected on the plasmid sequences.
>
> Special translated features: For genetic engineering proteins may contain additional peptide sequences with special function, e.g. epitopes, tag etc. These peptides shall be recognized.

All annotations (common features, additional protein coding genes and primer binding sites) are then stored together with the sequence in a genbank file.

Hints:

Common features:
The extraction and categorization of the common features is the most demanding part of the project. You will be provided with a short and a long learning file. The short file vectors_100.gb contains 100 vectors and the file vectors.gb contains 3576 plasmids *records* with 33293 *features*. The short learning file is for development purposes only.

Filtering: Valid plasmid sequences have more than 1500 bases. *Records* with shorter sequences can be discarded as they only contain partial sequences. Only sequence *features* of the *types* ['promoter', 'CDS', 'polyA_signal','rep_origin', 'primer_bind', 'terminator', 'protein_bind', 'misc_binding', 'misc_recomb', 'oriT', 'LTR', 'misc_signal', 'enhancer', 'mobile_element', 'RBS', 'sig_peptide', '-10_signal', '-35_signal', 'mRNA', 'tRNA', 'rRNA'] need to be analysed.

Clustering: In a first step features with identical sequences can be grouped and identified. Manual inspection is required to give the correct "consensus" standard_name *feature.qualifier* and note *feature.qualifier*. Output generated by the script will facilitate this curation step.

In a second step features of the same *type* and similar annotation (to be assessed by the keys in *feature.qualifiers* note, product, db_xref, standard_name, gene) shall be sequence compared either using the multialignment server MUSCLE or by using a pairwise sequence alignment with *Bio.pairwise.*

For usage of MUSCLE, see this code:

```
from bioservices import *

sequences='>test_a\nAGAGAGAGAG\n\n>test_b\nAGAAAGAA\n\n>test_c\nAGAGGAGAG\n\n'

m=MUSCLE(verbose=False)
jobid=m.run(frmt="fasta", sequence=sequences, email="georg.lipps@fhnw.ch")
while m.getStatus(jobid)==u'RUNNING':
print "Status: ",m.getStatus(jobid)

result=m.getResult(jobid, "sequence")
print "sequence:"
print result
result=m.getResult(jobid, "aln-fasta")
print "aln-fasta:"
print result
result=m.getResult(jobid, "phylotree")
print "phylotree:"
print result
result=m.getResult(jobid, "pim")
print "pim:"
print result
result=m.getResult(jobid, "out")
print "out:"
print result
```

The near identical features can then be merged (PSSM !) and further manual inspection is required to give the correct “consensus” standard_name *feature.qualifier* and note *feature.qualifier*.

Hauptaufgabe

And the end of this process there will be a manually curated list of sequences and PSSMs describing common features found on plasmids. For every feature the number of instances in the learning file is also known (i.e. their importance).

Protein coding genes:
Via blastx it is possible to detect coding regions for known proteins of the plasmid. A single blastx request will probably only yield hits to the longest coding region (giving the highest scores). Therefore it has to be assured that the complete circular sequence is investigated and including the sequences around the beginning and end of the sequence file. Potential coding regions shall be checked whether they are actually part of an open reading frame and thus would give rise to a protein.

Primer binding sites:
For sequencing and PCR it is required that 15 bases of the 3’ end of the primer match perfectly with one of the strands of the plasmid. A list of common primers is found in file common_primer.mfasta. The annotation should indicate when there is only a partial match to a primer.

Special translated features:
A list of peptide sequences tags_epitopes.mfasta is provided. It has to checked whether the six frame translation of the plasmid sequence contains any instances of these peptides sequences.

Suggested partition of work:

member 1:
Extraction of sequence identical features. Assemble the annotation for every sequence by analysing the *feature.qualifiers* dictionary. Setup a statistic how often a term is used.

Check whether there is overlap among the set of extracted identical sequences (pairwise sequence alignment).

Blast the set of sequences against nucleotide database and collect top hits.

➔ The assembled output will help to curate the list of entries.

member 2:
Identification of features with similar annotation (look in to *feature.qualifiers* dictionary). Multiple alignment of these sequences. Merging of the nearly identical sequences into a multiple alignment and a PSSM.

Annotation statistics for every cluster via *feature.qualifiers* dictionary.

Blast of typical sequence against nucleotide database.

➔ The assembled output will help to curate the list of entries.

member 3:
Finding matches of primer binding sites in provided plasmid sequence.

BLAST plasmid sequence against protein database and collect possible protein coding regions. Verify they are open reading frames.

Finding instances of special translated features in plasmid sequences.

Write a genbank file with all annotations in feature fields.

Agreeing with member 1 and 2 how the sequence features are provided (for storing objects in files see cPickle). Performing annotation based on identical sequences or nearly identical sequences based on work of member 1 and 2.

Write a genbank file with all annotations in feature fields.

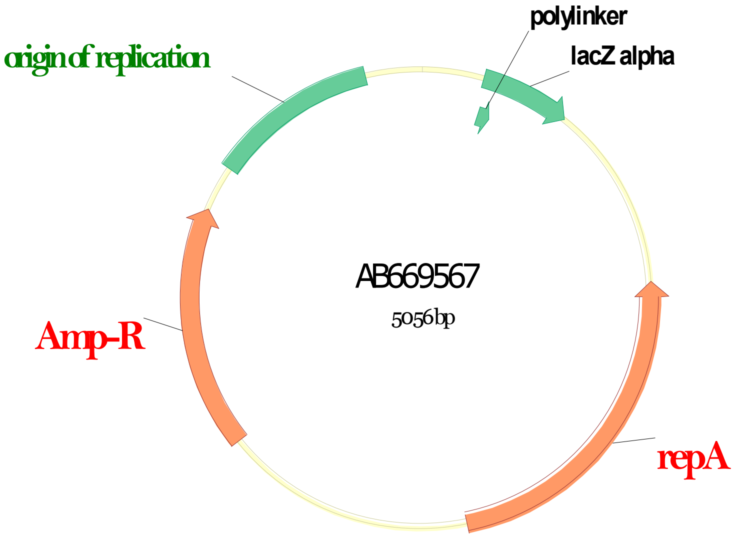
optional:
Output of a graphic of the plasmid with the annotated features.

## Example of a plasmid record in genbank:

```
LOCUS       AB669567                5056 bp    DNA     circular SYN 16-JUN-2012
DEFINITION  Cloning vector pUC-TTrepT DNA, complete sequence.
ACCESSION   AB669567
VERSION     AB669567.1  GI:379698694
KEYWORDS    .
SOURCE      Cloning vector pUC-TTrepT
  ORGANISM  Cloning vector pUC-TTrepT
            other sequences; artificial sequences; vectors.
REFERENCE   1
  AUTHORS   Vieira,J. and Messing,J.
  TITLE     The pUC plasmids, an M13mp7-derived system for insertion
            mutagenesis and sequencing with synthetic universal primers
  JOURNAL   Gene 19 (3), 259-268 (1982)
   PUBMED   6295879
REFERENCE   2
  AUTHORS   Fujita,A., Misumi,Y. and Koyama,Y.
  TITLE     Two versatile shuttle vectors for Thermus thermophilus-Escherichia
            coli containing multiple cloning sites, lacZalpha gene and
            kanamycin or hygromycin resistance marker
  JOURNAL   Plasmid 67 (3), 272-275 (2012)
   PUBMED   22252135
REFERENCE   3  (bases 1 to 5056)
  AUTHORS   Fujita,A.
  TITLE     Direct Submission
  JOURNAL   Submitted (08-SEP-2011) Contact:Atsushi Fujita National Institute
            of Advanced Science and Technology, Biomedical Research Institute;
            1-1-1 Higashi, Tsukuba, Ibaraki 305-8566, Japan URL
            :http://www.aist.go.jp/
FEATURES             Location/Qualifiers
     source          1..5056
                     /organism="Cloning vector pUC-TTrepT"
                     /mol_type="other DNA"
                     /db_xref="taxon:1085938"
                     /note="derivative of pUC13 containing a new MCS and T.
                     thermophilus repA gene; constructed for the E. coli-T.
                     thremophilus shuttle vector"
     misc_feature    216..533
                     /note="E. coli beta-galactosidase gene (LacZ) alpha
                     peptide"
     misc_feature    233..283
                     /note="polylinker
                     HindIII-NotI-SalI-NruI-AflII-EcoRV-Acc65I-EcoRI"
     misc_feature    690..3062
                     /note="derived from pYK225 (derivative of pTT8)"
     gene            complement(1197..2360)
                     /gene="repA"
     CDS             complement(1197..2360)
                     /gene="repA"
                     /codon_start=1
                     /transl_table=11
                     /product="putative RepA protein"
                     /protein_id="BAL70402.1"
                     /db_xref="GI:379698695"
                     /translation="MVLRAYAALRGLSPEALRAHLLAPPLRPERAREAFQRPYLAHFA
                     QTLPRYPYATDDPKEGVRIYKRENALKRVHVQVGHYPHAVLRLVVDVDLPWPQVEERI
                     HALPPSLVLVNPRSGHFHAWYELDPIPLTPPPGREGSLKGALALLAEVEALLEAYYGA
                     DPGYNGLLSRNPFLHPPEWTWGGGKRWSLRDLHRELRGLLPSGTRRRVDPGLASYGRN
                     NALFDRLRAEAYAHVALFRGVPGGEEAFRAWVEQRAHALNQSLFRDHPKGPLDPREVH
                     HTAKSVAKWTYRNYRGARVYPVSSTGRPDRSRLSPQARALIPPLQGQELQEAVREGGR
                     RRGSRRRQEAEEKLTEALKRLQARGERVTARALAREAGVKPHTASKWLKRMRE"
     gene            3255..4115
                     /gene="Amp-R"
     CDS             3255..4115
                     /gene="Amp-R"
                     /codon_start=1
                     /transl_table=11
                     /product="beta-lactamase"
                     /protein_id="BAL70403.1"
                     /db_xref="GI:379698696"
                     /translation="MSIQHFRVALIPFFAAFCLPVFAHPETLVKVKDAEDQLGARVGY
                     IELDLNSGKILESFRPEERFPMMSTFKVLLCGAVLSRIDAGQEQLGRRIHYSQNDLVE
                     YSPVTEKHLTDGMTVRELCSAAITMSDNTAANLLLTTIGGPKELTAFLHNMGDHVTRL
                     DRWEPELNEAIPNDERDTTMPVAMATTLRKLLTGELLTLASRQQLIDWMEADKVAGPL
                     LRSALPAGWFIADKSGAGERGSRGIIAALGPDGKPSRIVVIYTTGSQATMDERNRQIA
                     EIGASLIKHW"
     misc_feature    4276..4861
                     /note="origin of replication (ori) in E. coli"
ORIGIN
        1 gcgcccaata cgcaaaccgc ctctccccgc gcgttggccg attcattaat gcagctggca
       61 cgacaggttt cccgactgga aagcgggcag tgagcgcaac gcaattaatg tgagttagct
      121 cactcattag gcaccccagg ctttacactt tatgcttccg gctcgtatgt tgtgtggaat
      181 tgtgagcgga taacaatttc acacaggaaa cagctatgac catgattacg ccaagcttgg
      241 cggccgcgtc gactcgcgac ttaagatatc cggtaccgaa ttcactggcc gtcgttttac
      301 aacgtcgtga ctgggaaaac cctggcgtta cccaacttaa tcgccttgca gcacatcccc
      361 ctttcgccag ctggcgtaat agcgaagagg cccgcaccga tcgcccttcc caacagttgc
      421 gcagcctgaa tggcgaatgg cgcctgatgc ggtattttct ccttacgcat ctgtgcggta
      481 tttcacaccg catatggtgc actctcagta caatctgctc tgatgccgca tagttaagcc
      541 agccccgaca cccgccaaca cccgctgacg cgccctgacg ggcttgtctg ctcccggcat
      601 ccgcttacag acaagctgtg accgtctccg ggagctgcat gtgtcagagg ttttcaccgt
      661 catcaccgaa acgcgcgaga cgaaagggcg gctatggagg ggttctccct gtctacgctt
      721 gagcgcccga gggccggtcc agcccccggg ccagagcggg ctccggaccc tgcccccaga
      781 aggaccctct cacccgtcct aaggccccta gaaggccggg gggcgggggg aggggtggga
      841 ttggtgcccc ccgcccaccc cccgccttaa aagccgtttt aggggcggtc ctcgaggggg
      901 aaaatcagtg tttttgccgt cctggacgat gaaaattgcg tgcgggcggg gtaaatatgt
      961 gagttatctc actttctctc ggatctcaag gtgtctaccc caagaaaaac aaaatttttg
     1021 cggcttccag gaaaaagaag gggacttta gccgtcaact acgctaagga gactctctag
     1081 gttctccaat acagcccctt cggggctggc gggggttgca cccccgcacc cccgcctgta
     1141 taccctgata gccaccagct tggggcacat ttttggggtt tgtgccgtcc tggacgctac
     1201 tcccgcatcc tcttcaacca cttggaggcg gtatggggct tgacccccgc ctcccgggcc
     1261 agggccctgg ccgtgacccg ctcccccgg gcctggaggc gcttcagggc ctccgtgagc
     1321 ttctcctcgg cctcctgcct gcgccgggat ccgcgccgcc ttccgccctc ccgcaccgcc
     1381 tcctggagct cctggccctg gaggggcggg atcagggccc gggcctgggg agagaggcgg
     1441 ctccggtccg gcctccccgt ggaggagacc gggtagaccc tcgcccccg gtagttccgg
     1501 taggtccact tggccacgct cttcgccgtg tggtggacct cccgggggtc aaggggcccc
     1561 ttggggtggt cccggaagag ggactggttc agggcgtggg ccctctgctc cacccaggcc
     1621 cggaaggcct cctccccccc ggggacgccc cggaagaggg ccacgtgggc gtaggcctcc
     1681 gcccgcaggc ggtcaaacag ggcgttgttc cgcccgtagg aggccaggcc cgggtccacc
     1741 cgcctccggg tcccggaggg aaggagcccc cggagctccc ggtggaggtc ccgcaggctc
     1801 caccgcttcc ccccgcccca ggtccactcc gggggtgga ggaaggggtt tcgggagagg
     1861 agaccgttgt agcccgggtc cgccccgtag taggcctcca gcagggcctc cacctccgcg
     1921 agaagggcca gggcccctt caggctcccc tcccgcccgg gcgggggcgt gagggggatg
     1981 gggtccagct cgtaccaggc gtggaagtgg cccgatctcg ggttgaccag gaccagggag
     2041 gggggaggg cgtggatccg ctcctccacc tggggccagg ggaggtccac gtccaccacc
     2101 agccgcaaga cggcgtgggg gtagtggccc acctggacgt ggacccgctt cagggcgttc
     2161 tcccgcttgt agatgcgcac cccctccttg gggtcgtccg tggcgtaggg gtagcggggg
     2221 agagtctggg cgaagtgggc gaggtagggc cgctggaagg cctcccgggc ccgctccggg
     2281 cggaggggag gggccaggag gtgggcgcgg agggcctccg gggagaggcc gcgcagggcg
     2341 gcgtaggcgc gaagaaccac ctcccccagg tcggggttta tctggaagag tcctgcgatt
     2401 ttagcgaggg tgtccgggat ttgcgcttcc ggccgggtgt ggtgctccat ctcttgcctc
     2461 cttccccccg gccgggttag gatggaggcc tcctggggag gagaggcctc ggcggttaac
     2521 ccggccgagg tctgtttcat tcatgccccc cattctagcg acaagccccg ggaccaagcg
     2581 gtagtcccca ggtcaatgcc cccagaaccg ccaccagggc cgcctccgcg gggctccag
     2641 ggccttcagg gcctcccgga ggcgggcgtt ctcctcccgc agggccctta gctccccctc
     2701 cacccgctcc agccgctcgg ccagggcccg gaggagggcc agggcctccc cctcggaggg
     2761 gagggcgagg gccgggaagt ctccctggac ccgggccagg gcctcctcca gggggaggcc
     2821 ttcccgcagg tgggcctccc gggccgcccg gaggcgggcc agggcctcct tgggccagag
     2881 ccgtcccccc cggggatccc ggggcagggg acctaccaac cgctcccaca gggcggcgta
     2941 gcgccggagg gtggccgggg agacccccag ggcccgggcg gccaaggccg gggcaggag
     3001 ggtcgggtct tccccttccg ggtgctccac gaccccagtc tacccaggcg ctcagggtga
     3061 gcggcctcgt gatacgccta tttttatagg ttaatgtcat gataataatg gtttcttaga
     3121 cgtcaggtgg cacttttcgg ggaaatgtgc gcggaacccc tatttgttta tttttctaaa
     3181 tacattcaaa tatgtatccg ctcatgagac aataaccctg ataaatgctt caataatatt
     3241 gaaaaaggaa gagtatgagt attcaacatt tccgtgtcgc ccttattccc ttttttgcgg
     3301 cattttgcct tcctgttttt gctcacccag aaacgctggt gaaagtaaaa gatgctgaag
     3361 atcagttggg tgcacgagtg ggttacatcg aactggatct caacagcggt aagatccttg
     3421 agagttttcg ccccgaagaa cgttttccaa tgatgagcac ttttaaagtt ctgctatgtg
     3481 gcgcggtatt atcccgtatt gacgccgggc aagagcaact cggtcgccgc atacactatt
     3541 ctcagaatga cttggttgag tactcaccag tcacagaaaa gcatcttacg gatggcatga
     3601 cagtaagaga attatgcagt gctgccataa ccatgagtga taacactgcg gccaacttac
     3661 ttctgacaac gatcggagga ccgaaggagc taaccgcttt tttgcacaac atgggggatc
     3721 atgtaactcg ccttgatcgt tgggaaccgg agctgaatga agccatacca aacgacgagc
     3781 gtgacaccac gatgcctgta gcaatggcaa caacgttgcg caaactatta actggcgaac
     3841 tacttactct agcttcccgg caacaattaa tagactggat ggaggcggat aaagttgcag
     3901 gaccacttct gcgctcggcc cttccggctg gctggtttat tgctgataaa tctggagccg
     3961 gtgagcgtgg gtctcgcggt atcattgcag cactggggcc agatggtaag ccctcccgta
     4021 tcgtagttat ctacacgacg gggagtcagg caactatgga tgaacgaaat agacagatcg
     4081 ctgagatagg tgcctcactg attaagcatt ggtaactgtc agaccaagtt tactcatata
     4141 tactttagat tgatttaaaa cttcattttt aatttaaaag gatctaggtg aagatccttt
     4201 ttgataatct catgaccaaa atcccttaac gtgagttttc gttccactga gcgtcagacc
     4261 ccgtagaaaa gatcaaagga tcttcttgag atcctttttt tctgcgcgta atctgctgct
     4321 tgcaaacaaa aaaaccaccg ctaccagcgg tggtttgttt gccggatcaa gagctaccaa
     4381 ctctttttcc gaaggtaact ggcttcagca gagcgcagat accaaatact gtccttctag
     4441 tgtagccgta gttaggccac cacttcaaga actctgtagc accgcctaca tacctcgctc
```

```
4501 tgctaatcct gttaccagtg gctgctgcca gtggcgataa gtcgtgtctt accgggttgg
4561 actcaagacg atagttaccg gataaggcgc agcggtcggg ctgaacgggg ggttcgtgca
4621 cacagcccag cttggagcga acgacctaca ccgaactgag atacctacag cgtgagctat
4681 gagaaagcgc cacgcttccc gaagggagaa aggcggacag gtatccggta agcggcaggg
4741 tcggaacagg agagcgcacg agggagcttc caggggaaa cgcctggtat ctttatagtc
4801 ctgtcgggtt tcgccacctc tgacttgagc gtcgattttt gtgatgctcg tcaggggggc
4861 ggagcctatg gaaaaacgcc agcaacgcgg cctttttacg gttcctggcc ttttgctggc
4921 cttttgctca catgttcttt cctgcgttat cccctgattc tgtggataac cgtattaccg
4981 cctttgagtg agctgatacc gctcgccgca gccgaacgac cgagcgcagc gagtcagtga
5041 gcgaggaagc ggaaga
```

Simulated output of a *de novo* annotation of the plasmid sequence:


polylinker
lacZ alpha
origin of replication
AB669567
5056bp
Amp-R
repA