



Chapter 02. Working with Annotations

Python Programming for Bioinformatics

Robert C. Chi

Agenda

- **Introduction of Annotations**
- **Sequence Records**
 - Bio.SeqRecord.SeqRecord
- **References**
 - Bio.SeqFeature.Reference
- **Features, Locations, Positions**
 - Bio.SeqFeature.SeqFeature
- **Sequences + Features**
 - Bio.Seq.Seq
- **Summary**





INTRODUCTION OF ANNOTATIONS

What is “Annotations”?

• FASTA

.id / .description

SeqRecord X1

```
>gil2765657|emblZ78532.1|CCZ78532 C.californicum 5.8S rRNA gene and ITS1 and ITS2 DNA
CGTAACAAGGTTTCGGTAGGTAACCTGCGGAAGGATCATTGTTGAGACAACAGAATATATGATCGAGTG
AATCTGGAGGACCTGGTAACTCAGCTCGTGGCACTGCTTTTGTGCGTGACCCCTGCTTTGTTGTTGG
GCCTCCTCAAGAGCTTTCATGGCAGGTTTGAACCTTGTAGTACGGTGCAGTTTGCGCCAAGTCATATAAAGC
ATCACTGATGAATGACATATTGTGCAGAAAAATCAGAGGGGCGAGTATGCTACTGAGCATGCCAGTGAAT
TTTTATGACTCTCGCAACGGATATCTTGGCTCTAACATCGATGAAGAACGCAGCTAAATGGGATAAGTGG
TGTGAATTGCAGAAATCCCGTGAACCATCGAGTCTTTTCGCGGATTCGATCGAGGCCATCAGGCTAAG
GGCAGCCTGCGCTGGGCGTGTGTTGCGTCTCTGCGCAATCTCGTTGGCATATCGCTAAGCTGG
CATTATACGGATGGAATGATTGGCCCTTGTGCTAGGTGCGGTGGGTCGAAGGATTGTTGCTTTGATG
GGTAGGAATGTGGCACGAGGTGGAGAATGCTAACAGTCATAAAGCTGCTATTTGAATCCCCATGTTGTT
GTATTTTTTGAACCTACACAAGAACCCTAATTGAACCCCAATGGAGCTAAAATAACCATTTGGCGAGTTGA
TTTCCATTCCAGATGGACCCAGGTGACGGCGGGGCCACCCGCTGAGTTGAGGC
```

Seq

Definition of Annotations:

- The metadata other than sequences.
- e.g., ID, Name, Reference, Features...etc.

• GenBank

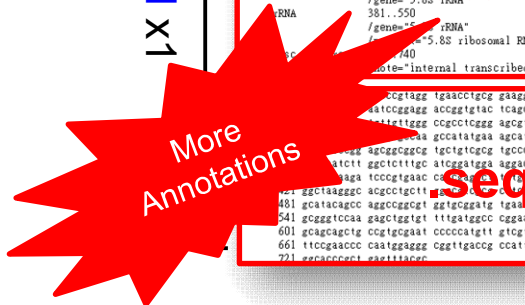
```
LOCUS       Z78533             740 bp    DNA     linear   PLN 30-NOV-2006
DEFINITION  C.californicum 5.8S rRNA gene and ITS1 and ITS2 DNA
ACCESSION   Z78533
VERSION     Z78533.1  GI:2765658
KEYWORDS    5.8S ribosomal RNA; 5.8S rRNA gene; internal transcribed spacer;
            ITS1; ITS2.
SOURCE      Cyrtopidium irapeanum
ORGANISM    Cyrtopidium irapeanum
            Eukaryota; Viridiplantae; Streptophyta; Embryophyta; Tracheophyta;
            Spermatophyta; Magnoliophyta; Liliopsida; Asparagales; Orchidaceae;
            Cyrtopidiaceae; Cyrtopidium
REFERENCE   1
AUTHORS    Cox,A.V., Pridgen,A.M., Albert,V.A. and Chase,M.W.
TITLE      Phylogenetics of the slipper orchids (Cyrtopidiaceae:
            Orchidaceae): nuclear rDNA ITS sequences
JOURNAL    Botanical Journal of Linnean Society 147: 1-14 (2004)
REFERENCES 1-14
AUTHORS    Cox,A.V.
TITLE      Direct Submission
JOURNAL    Submitted (19-ADJ-1996) Cox A.V., Royal Botanic Gardens, Kew,
            Richmond, Surrey TW9 3AB, UK
FEATURES   Location/Qualifiers
            source          1..740
                /organism="Cyrtopidium irapeanum"
                /mol_type="genomic DNA"
                /db_xref="taxon:49711"
            misc_feature     1..380
                /note="5.8S rRNA"
            gene            381..550
                /gene="5.8S rRNA"
            rRNA            381..550
                /gene="5.8S rRNA"
                /note="5.8S ribosomal RNA"
            tRNA            1..740
                /note="internal transcribed spacer 2"
```

SeqRecord X1

.id / .description

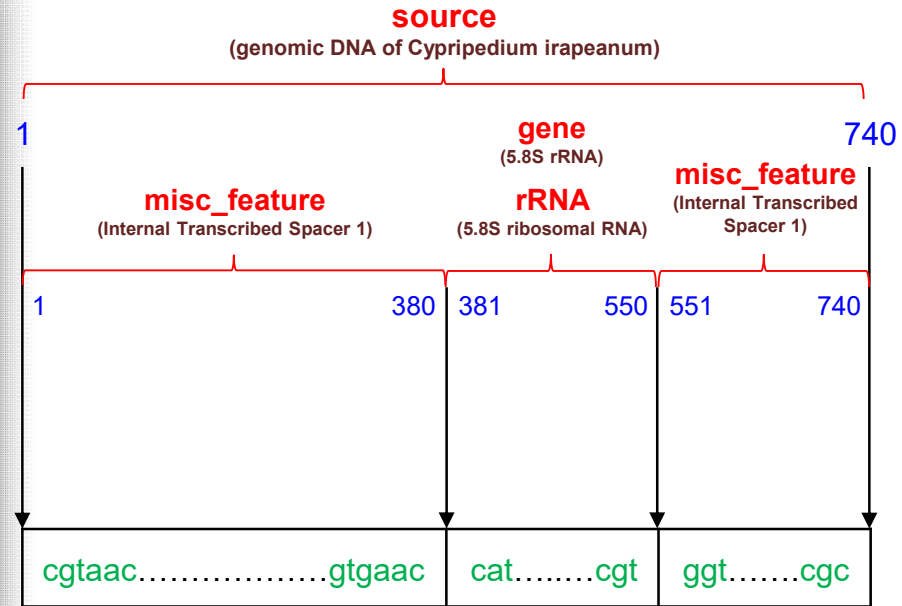
annotations [references]

features



The Gem of GenBank File

| | FEATURES | Location/Qualifiers |
|----------|-------------------------|------------------------------------------------------------------------------------------------|
| Features | 1 source | 1..740 /organism="Cypridium irapeanum" /mol_type="genomic DNA" /db_xref="taxon:49711" |
| | 2 misc_feature | 1..380 /note="internal transcribed spacer 1" |
| | 3 gene | 381..550 /gene="5.8S rRNA" |
| | 4 rRNA | 381..550 /gene="5.8S rRNA" /product="5.8S ribosomal RNA" |
| | 5 misc_feature | 551..740 /note="internal transcribed spacer 2" |
| Sequence | ORIGIN | |
| | | 1 cgtaacaagg ttccgtagg tgaacctcgc gaaggatcat tgatgagacc gtggaataaa |
| | | 61 cgatcgagtg aatccggagg accggtgtac tcagctcacc gggggcattg ctcccgtggt |
| | | 121 gacctgatt tgtgttggg ccgcctcggg agcgtccatg gcgggttga acctctagcc |
| | | 181 cggcgcagtt tgggcgcaa gccatatgaa agcatcaccg gcgaatggca ttgtcttccc |
| | | 241 caaaaccggg agcggcggcg tctgtcgcg tccccaatga attttgatga ctctcgcaaa |
| | | 301 cgggaatctt ggcctcttgc atcggatgga aggaacgagc gaaatgcgat aagtgtgtg |
| | | 361 aattgcaaga tcccgtgaac catcgagtct ttgaaacga agttgcgcc gaggccatca |
| | | 421 ggctaagggc acgctcgtt gggcgtcgcg cttcgtctct ctctgccaa tgcctgcccg |
| | | 481 gcatacagcc aggcggcgt ggtcgggat taaaagatt gccctctgt cctaggtgcg |
| | | 541 gcgggtccaa gagctggtt ttgatggcc cggaaaccgg caagaggagg acggatgctg |
| | | 601 gcagcagctg ccgtgcgaat cccccatgt tctgtcttg tcggacaggc aggagaaccc |
| | | 661 ttccgaacc caatggagg cggttgacc ccattcggat tgaacccag gtcaggcggg |
| | 721 ggcaccgct gagtttacg | |



Annotation related Objects

GenBank File

```

LOCUS       Z78533                740 bp    DNA     linear   PLN 30-NOV-2006
DEFINITION  C. irapeanum 5.8S rRNA gene and ITS1 and ITS2 DNA.
ACCESSION   Z78533
VERSION     Z78533.1 GI:2765658
KEYWORDS    5.8S ribosomal RNA; 5.8S rRNA gene; internal transcribed spacer;
            ITS1; ITS2.
SOURCE      Cyripedium irapeanum
ORGANISM    Cyripedium irapeanum
            Eukaryota; Viridiplantae; Streptophyta; Embryophyta; Tracheophyta;
            Spermatophyta; Magnoliophyta; Liliopsida; Asparagales; Orchidaceae;
            Cyripedioideae; Cyripedium.
REFERENCE   1
AUTHORS     Cox,A.V., Pridgeon,A.M., Albert,V.A. and Chase,M.V.
TITLE       Phylogenetics of the slipper orchids (Cyripedioideae:
            Orchidaceae): nuclear rDNA ITS sequences
JOURNAL     Unpublished
REFERENCE   2 (bases 1 to 740)
AUTHORS     Cox,A.V.
TITLE       Direct Submission
JOURNAL     Submitted (19-AUG-1996) Cox A.V., Royal Botanic Gardens, Kew,
            Richmond, Surrey TW9 3AB, UK
FEATURES             Location/Qualifiers
     source            1..740
                     /organism="Cyripedium irapeanum"
                     /mol_type="genomic DNA"
                     /db_xref="taxon:49711"
     misc_feature      1..380
                     /note="internal transcribed spacer 1"
     gene              381..550
                     /gene="5.8S rRNA"
     rRNA              381..550
                     /gene="5.8S rRNA"
                     /product="5.8S ribosomal RNA"
     misc_feature      551..740
                     /note="internal transcribed spacer 2"
ORIGIN
1  cgtacaagg tttccgtag gaaacctcg gaagatcat tga1gagcc g1gaataaa
61  cgatcgagt aatccggag acccgtgtc tcagctcac gzzgcat1g ctcccg1gt
121  gaccctgatt tgtt1tgg ccgcctcgg agcctcatt gczgatt1ga acctctagc
181  cggcgcagt1 tggcgccaa gccat1atga agcatcacg gcaat1ggca ttgct1tcc
241  caaaaccgg agcgcgcg cgctctcgc tgcocaa1ga att1tgatga ctctgc0aa
301  cggaa1ctt gctct1tgc atcgatga agacacgc gaa1tgcgt aagt1gt1g
361  aattgcaaga tccctgaac catcgactt tt1gaagca ag1tgcgcc gaggccatc
421  gct1aaggc acgcctgct ggcctcgc ctctctct ctcttccaa tct1gc0cc
481  gcatcacgc agc0cgctg cgt1cggatg tgaagaatt gccoc1t1g cctagt1gc
541  gczg1tcaa gact1gtgt ttgatggc cgaacc0cg caaga1z1tg accgat1ct
601  gczgacg1g ccgtcgaa1 ccccat1t1 gct1gct1g tczgacagc agga1aac1
661  ttcgaa0cc ca1tggagg czg1t1gac ccat1cgtt g1gac0ccg t1cagc0gg
721  gcacc0ct gatt1tacc
    
```

Bio.SeqIO



Bio.SeqRecord.SeqRecord x1

```

.name = 278533
.description = C. irapeanum 5.8S...
.id = 278533.1
.....
    
```

Bio.SeqFeature.Reference x2

```

REFERENCE 1
AUTHOR  Cox,A.V., Pridgeon,A.M., Albert,V.A. and Chase,M.V.
TITLE   Phylogenetics of the slipper orchids (Cyripedioideae:
        Orchidaceae): nuclear rDNA ITS sequences
JOURNAL Unpublished
    
```

Bio.SeqFeature.SeqFeature x5

```

source      1..740
            /organism="Cyripedium irapeanum"
            /mol_type="genomic DNA"
            /db_xref="taxon:49711"
    
```

Bio.Seq.Seq x1

```

1  cgtacaagg tttccgtag gaaacctcg gaagatcat tga1gagcc g1gaataaa
61  cgatcgagt aatccggag acccgtgtc tcagctcac gzzgcat1g ctcccg1gt
121  gaccctgatt tgtt1tgg ccgcctcgg agcctcatt gczgatt1ga acctctagc
181  cggcgcagt1 tggcgccaa gccat1atga agcatcacg gcaat1ggca ttgct1tcc
241  caaaaccgg agcgcgcg cgctctcgc tgcocaa1ga att1tgatga ctctgc0aa
301  cggaa1ctt gctct1tgc atcgatga agacacgc gaa1tgcgt aagt1gt1g
361  aattgcaaga tccctgaac catcgactt tt1gaagca ag1tgcgcc gaggccatc
421  gct1aaggc agc0cgctg ggcctcgc ctctctct ctcttccaa tct1gc0cc
481  gcatcacgc agc0cgctg cgt1cggatg tgaagaatt gccoc1t1g cctagt1gc
541  gczg1tcaa gact1gtgt ttgatggc cgaacc0cg caaga1z1tg accgat1ct
601  gczgacg1g ccgtcgaa1 ccccat1t1 gct1gct1g tczgacagc agga1aac1
661  ttcgaa0cc ca1tggagg czg1t1gac ccat1cgtt g1gac0ccg t1cagc0gg
721  gcacc0ct gatt1tacc
    
```

Bio.SeqFeature.FeatureLocation

Bio.SeqFeature.Position Bio.SeqFeature.Position

1 .. 740

● Blue: Annotation related Objects

Brief of this Section

- **Annotation-related Objects**
 - Parsing: **Bio.SeqIO**
 - Records: **Bio.SeqRecord.SeqRecord** (+ **.attributes**)
 - References: **Bio.SeqFeature.Reference**
 - Features: **Bio.SeqFeature.SeqFeature**
 - Locations: **Bio.SeqFeature.FeatureLocation**
 - Positions: **Bio.SeqFeature.Position**
 - Sequences: **Bio.Seq.Seq**

SUMMARY





Bio.SeqRecord.SeqRecord

SEQUENCE RECORDS

Parse Records from GenBank

```
1 # Create a list to store all parsed SeqRecords
2 seq_records = []
3
4 # Parse one Record a time and add into seq_records
5 from Bio import SeqIO
6 for rec in SeqIO.parse("ls_orchid.gbk", "genbank"):
7     seq_records.append(rec)           [0] [1] [2] ... [n]
8     ↑ seq_records = [rec, rec, rec, ...] + [rec]
9 # Show how many records have been parsed
10 print("Total Number of Records:", len(seq_records))
11 print()
12
13 # Show the first record to prove the parsing was successful
14 print("--- First Record ---")
15 print(seq_records[0])
```

Total Number of Records: 94

```
--- First Record ---
ID: Z78533.1
Name: Z78533
....
Seq('CGTAAC...CGC')
```

Practice

- **Parse Records from GenBank**

- **Write** and **Run** the following codes on a Colab page called “[AnnotationObjects.ipynb](#)”:

- Download the GenBank File

```
[8] 1 import os
    2 if not os.path.isfile("ls_orchid.gb"):
    3     os.system("wget https://raw.githubusercontent.com/biopython/biopython/master/Doc/examples/ls_orchid.gb")
```

- Get the List of All SeqRecords

```
1 # Create a list to store all parsed SeqRecords
2 seq_records = []
3
4 # Parse one Record a time and add into seq_records
5 from Bio import SeqIO
6 for rec in SeqIO.parse("ls_orchid.gb", "genbank"):
7     seq_records.append(rec)
8
9 # Show how many records have been parsed
10 print("Total Number of Records:", len(seq_records))
11 print()
12
13 # Show the first record to prove the parsing was successful
14 print("--- First Record ---")
15 print(seq_records[0])
```



(Solution [URL](#) of this Practice)

Attributes of SeqRecord

```
1 LOCUS ① Z78533 740 bp ② DNA ③ linear ④ PLN 30-NOV-2006
2 ⑤ DEFINITION C.irapeanum 5.8S rRNA gene and ITS1 and ITS2 DNA.
3 ⑥ ACCESSION Z78533 ⑦ ⑧
4 VERSION Z78533.1 GI:2765658
5 ⑨ KEYWORDS 5.8S ribosomal RNA; 5.8S rRNA gene; internal transcribed spacer;
6 ITS1; ITS2.
7 ⑩ SOURCE Cypripedium irapeanum
8 ⑪ ORGANISM Cypripedium irapeanum
9 ⑫ Eukaryota; Viridiplantae; Streptop
10 Spermatophyta; Magnoliophyta; Lili
11 Cypripedioideae; Cypripedium.
```

```
1 # Take the first SeqRecord as Example
2 rec = seq_records[0]
3
4 # Show the attributes of SeqRecord
5 ① print("Locus Name:", rec.name)
6 ② print("Sequence Type:", rec.annotations['molecule_type'])
7 ③ print("Sequence Topology:", rec.annotations['topology'])
8 ④ print("Published Date:", rec.annotations['date'])
9 ⑤ print("Sequence Definition:", rec.description)
10 ⑥ print("Accession No.:", rec.annotations['accessions'])
11 ⑦ print("Sequence Version:", rec.annotations['sequence_version'])
12 ⑧ print("NCBI GenInfo Identifier (G.I.):", rec.annotations['gi'])
13 ⑨ print("Keywords:", rec.annotations['keywords'])
14 ⑩ print("Sequence Source:", rec.annotations['source'])
15 ⑪ print("Organism:", rec.annotations['organism'])
16 ⑫ print("Taxonomy:", rec.annotations['taxonomy'])
```

Practice

- **Get Attributes of SeqRecord**

- **Write** and **Run** the following codes on a Colab page called “[AnnotationObjects.ipynb](#)”:

- Show Attributes of SeqRecord

```
[17] 1 # Take the first SeqRecord as Example
      2 rec = seq_records[0]
      3
      4 # Show the attributes of SeqRecord
      5 print("Locus Name:", rec.name)
      6 print("Sequence Type:", rec.annotations['molecule_type'])
      7 print("Sequence Topology:", rec.annotations['topology'])
      8 print("Published Date:", rec.annotations['date'])
      9 print("Sequence Definition:", rec.description)
     10 print("Accession No.:", rec.annotations['accessions'])
     11 print("Sequence Version:", rec.annotations['sequence_version'])
     12 print("NCBI GenInfo Identifier (G.I.):", rec.annotations['gi'])
     13 print("Keywords:", rec.annotations['keywords'])
     14 print("Sequence Source:", rec.annotations['source'])
     15 print("Organism:", rec.annotations['organism'])
     16 print("Taxonomy:", rec.annotations['taxonomy'])
```

```
Locus Name: Z78533
Sequence Type: DNA
Sequence Topology: linear
Published Date: 30-NOV-2006
Sequence Definition: C.irapeanum 5.8S rRNA gene and ITS1 and ITS2 DNA
Accession No.: ['Z78533']
Sequence Version: 1
NCBI GenInfo Identifier (G.I.): 2765658
Keywords: ['5.8S ribosomal RNA', '5.8S rRNA gene', ...]
Sequence Source: Cypripedium irapeanum
```



(Solution [URL](#) of this Practice)

Operation: Get Length

▼ Length of SeqRecord

✓
0s



```
1 # SeqRecord can get the length directly  
2 print("Length of SeqRecord:", len(rec))
```

Length of SeqRecord: 740

Practice

- **Get the Length of Sequence in SeqRecord**
 - **Write** and **Run** the following codes on a Colab page called “[AnnotationObjects.ipynb](#)”:

Length of SeqRecord

```
[ ] 1 # SeqRecord can get the length directly  
    2 print("Length of SeqRecord:", len(rec))
```

Length of SeqRecord: 740

(Solution [URL](#) of this Practice)



Operation: Slicing

```
FEATURES          Location/Qualifiers
  gene            381..550
                  /gene="5.8S rRNA"

  rRNA            381..550
                  /gene="5.8S rRNA"
                  /product="5.8S ribosomal RNA"

1 cgtaacaagg tttccgtagg tgaacctgcg gaaggatcat tgatgagacc gtggaataaa
61 cgatcgagtg aatccggagg accggtgtac tcagctcacc gggggcattg ctcccgtggt
121 gaccctgatt tgttgttggg ccgcctcggg agcgtccatg gcgggttga acctctagcc
181 cggcgcagtt tgggcgccaa gccatatgaa agcatcaccg gcgaatggca ttgtcttccc
241 caaaaaccgg agcggcgggc tgctgtcgcg tgcccaatga attttgatga ctctcgcaaa
301 cgggaatcct ggctctttgc atcggatgga aggacgcagc gaaatgcat aagtgggtgtg
361 aattgcaaga tcccgtgaac catcagatct tttgaacgca agttgcgccc gaggccatca
421 ggctaagggc acgcctgctt gggcgtcgcg cttcgtctct ctctcgcaaa tgcttgcccg
481 gcatacagcc aggccggcgt ggtgcggatg tgaagattg gcccccttgg cctaggtgcg
541 gggggtccaa gagctgggtt tttgatggcc cggaaaccgg caagaggtgg acggatgctg
601 gcagcagctg cctgtgcaat cccccatggt gtcgtgcttg tcggacagggc aggagaaccc
661 ttccgaaccc caatggaggg cggttgaccg ccattcggat gtgacccagc gtcaggcggg
721 ggcaccgct  gagtttacg
```

NCBI Index (1-based): [381 : 550]

Python Index (0-based): [380 : 549]

Upper bound Exclusive: [380 : 550]

```
1 # Set the Start and End Position
2 startIndex = 381 - 1 # Python is 0-based
3 endIndex = 550 # Upperbound is exclusive
4
5 # Slice from the SeqRecord
6 sub_rRNA = rec[startIndex:endIndex]
7 print(sub_rRNA)
8 print(sub_rRNA.features)
```



```
ID: Z78533.1
Name: Z78533
Description: C.irapeanum 5.8S rRNA gene and ITS1 and ITS2 DNA
Number of features: 2
/molecule_type=DNA
Seq('CATCGAGTCTTTTGAACGCAAGTTGCGCCCGAGGCCATCAGGCTAAGGGCACGC...CAA')
[SeqFeature(FeatureLocation(ExactPosition(0), ExactPosition(170), strand=1), type='gene'),
SeqFeature(FeatureLocation(ExactPosition(0), ExactPosition(170), strand=1), type='rRNA')]
```

Slice + Annotations Preserved !!

Practice

- **Slice a SeqRecord**

- **Write** and **Run** the following codes on a Colab page called “[AnnotationObjects.ipynb](#)”:

- Slicing of SeqRecord

```
1 # Set the Start and End Position
2 startIndex = 381 - 1 # Python is 0-based
3 endIndex = 550 # Upperbound is exclusive
4
5 # Slice from the SeqRecord
6 sub_rRNA = rec[startIndex:endIndex]
7 print(sub_rRNA)
8 print(sub_rRNA.features)
```

```
ID: Z78533.1
Name: Z78533
Description: C.irapeanum 5.8S rRNA gene and ITS1 and ITS2 DNA
Number of features: 2
/molecule_type=DNA
Seq('CATCGAGTCTTTTGAACGCAAGTTGCGCCCGAGGCCATCAGGCTAAGGGCACGC...CAA')
[SeqFeature(Location(ExactPosition(0), ExactPosition(170), strand=1), type='gene'),
 SeqFeature(Location(ExactPosition(0), ExactPosition(170), strand=1), type='rRNA')]
```



(Solution [URL](#) of this Practice)

Operation: Reverse Complement

```
1 # Reverse Complement
2 print("Original 5' to 3':")
3 print(rec)
4
5 print("Reverse Complement 5' to 3':")
6 print(rec.reverse_complement())
```

Original 5' to 3':
ID: Z78533.1
Name: Z78533
Description: C.irapeanum 5.8S rRNA gene and ITS1 and ITS2 DNA
Number of features: 5
/molecule_type=DNA
...
Seq(CGTAAC.....CGC)

Reverse Complement 5' to 3':
ID: <unknown id>
Name: <unknown name>
Description: <unknown description>
Number of features: 5
Seq(GCGTAA.....ACG)

- **Not all the Annotations Preserved !!**
But **.features** were kept!
- **No .complement()** support currently

Practice

- **Reverse Complement of a SeqRecord**

- **Write** and **Run** the following codes on a Colab page called “[AnnotationObjects.ipynb](#)”:

- ▼ Reverse Complement of SeqRecord

```
[17] 1 # Reverse Complement
      2 print("Original 5' to 3':")
      3 print(rec)
      4
      5 print("Reverse Complement 5' to 3':")
      6 print(rec.reverse_complement())
```

(Solution [URL](#) of this Practice)



Operation: Drop Nucleotides

- **Drop** a Single Nucleotide at **index=2** (0-based)

CATCGAGTCTTTTGAACGCAAGTTGCGCCCGAGGCCATCAGGCTAAGGGCACGC...CAA

CA_CGAGTCTTTTGAACGCAAGTTGCGCCCGAGGCCATCAGGCTAAGGGCACGC...CAA

0~1 ~~2~~ 3~

```
1 # Drop the Single Nucleotide at index=2 (0-based)
2 print("Original rRNA:", sub_rRNA)
3 modified_rRNA = sub_rRNA[:2] + sub_rRNA[3:]
4 print("Modified rRNA:", modified_rRNA)
```

Modified rRNA: ID: Z78533.1
Name: Z78533
Description: C.irapeanum 5.8S rRNA gene and ITS1 and ITS2 DNA
Number of features: 0
/molecule_type=DNA
Seq('C**A**CGAGTCTTTTGAACGCAAGTTGCGCCCG...CAA')

Practice

- **Drop Nucleotides of SeqRecord**

- **Write** and **Run** the following codes on a Colab page called "[AnnotationObjects.ipynb](#)":

- ▼ Drop Nucleotides of SeqRecord

```
✓ ▶ 1 # Drop the Single Nucleotide at index=2 (0-based)
    2 print("Original rRNA:", sub_rRNA)
    3 modified_rRNA = sub_rRNA[:2] + sub_rRNA[3:]
    4 print("Modified rRNA:", modified_rRNA)
```

(Solution [URL](#) of this Practice)



Operation: Reorganize two Genes

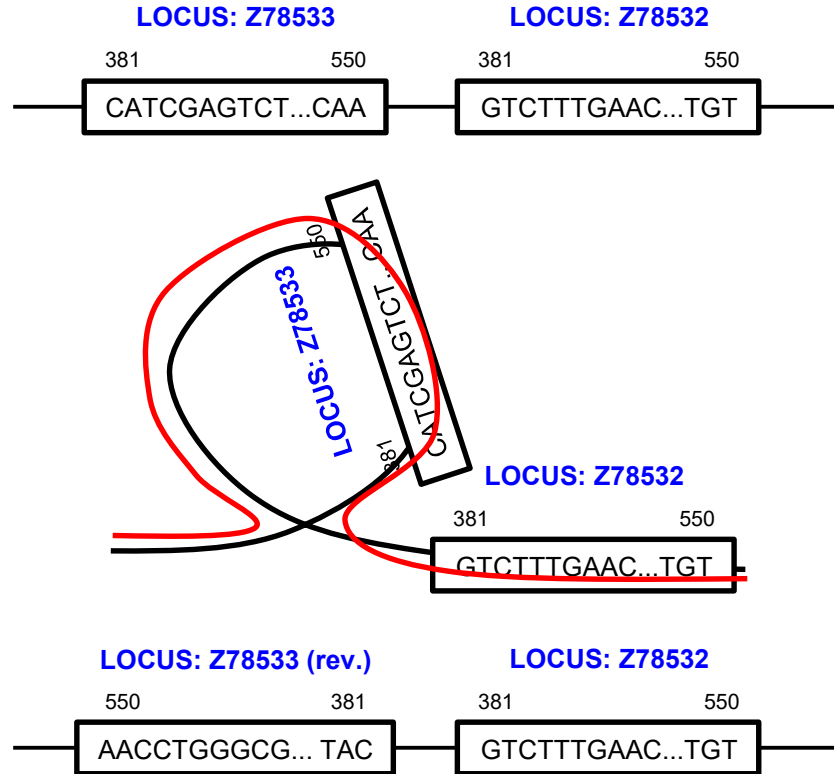
GenBank File

```
LOCUS       Z78533                740 bp    DNA     linear   PLN 30-NOV-2006
DEFINITION  C.irapeanum 5.8S rRNA gene and ITS1 and ITS2 DNA.
...
FEATURES             Location/Qualifiers
     gene             381..550
                     /gene="5.8S rRNA"
     rRNA             381..550
                     /gene="5.8S rRNA"
                     /product="5.8S ribosomal RNA"
ORIGIN
1  cgtaacaagg ttccgtagg tgaacctgcg gaagatcat tgatgagacc gtggaataaa
61  cgatcgagtg aatccggagg accggtgtac tcagctcacc gggggcattg ctcccgtggt
121 gaccctgatt tgttgttggg ccgctcggg  agcgtccatg gcgggtttga acctctagcc
```

Locus ID: Z78533

```
//
LOCUS       Z78532                753 bp    DNA     linear   PLN 30-NOV-2006
DEFINITION  C.californicum 5.8S rRNA gene and ITS1 and ITS2 DNA.
...
FEATURES             Location/Qualifiers
     gene             381..550
                     /gene="5.8S rRNA"
     rRNA             381..550
                     /gene="5.8S rRNA"
                     /product="5.8S ribosomal RNA"
ORIGIN
1  cgtaacaagg ttccgtagg tgaacctgcg gaagatcat tgttgagaca acagaataa
61  tgatcgagtg aatctggagg acctgtggta actcagctcg tctgtgcact gcttttgtcg
121 tgaccctgct ttgttgttgg gcctcctcaa gacgtttcat ggcaggtttg aactttagta
```

Locus ID: Z78532



Operation: Reorganize two Genes

```
1 # Reorganize two genes
2 sub_rRNA1 = seq_records[0][380:550]
3 print("rRNA 1 (5'-3'): -----")
4 print(sub_rRNA1)
5 print()
6
7 sub_rRNA2 = seq_records[1][380:550]
8 print("rRNA 2 (5'-3'): -----")
9 print(sub_rRNA2)
10 print()
11
12 # Reverse the sub_rRNA1
13 sub_rRNA1 = sub_rRNA1[::-1]
14 print("rRNA 1 (3'-5', Reversed): -----")
15 print(sub_rRNA1)
16 print()
17
18 # Join two rRNA
19 new_rRNA = sub_rRNA1 + sub_rRNA2
20 print("New rRNA : -----")
21 print(new_rRNA)
22 print()
```

```
rRNA 1 (5'-3'): -----
ID: Z78533.1
Name: Z78533
Description: C.irapeanum 5.8S rRNA gene and ITS1 and ITS2 DNA
Number of features: 2
/molecule_type=DNA
Seq('CATCGAGTCTTTTGAACGCAAGTTGCGCCCAGGCCATCAGGCTAAGGGCACGC...CAA')
```

slice
record[0]

```
rRNA 2 (5'-3'): -----
ID: Z78532.1
Name: Z78532
Description: C.californicum 5.8S rRNA gene and ITS1 and ITS2 DNA
Number of features: 2
/molecule_type=DNA
Seq('GTCTTTGAACGCAAGTTGCGCTCGAGGCCATCAGGCTAAGGGCACGCCTGCCTG...TGT')
```

slice
record[1]

```
rRNA 1 (3'-5', Reversed): -----
ID: Z78533.1
Name: Z78533
Description: C.irapeanum 5.8S rRNA gene and ITS1 and ITS2 DNA
Number of features: 0 features lost
/molecule_type=DNA
Seq('AACCTGGCGGCGTGGATCCGTGTTCCCGGTTAGAAAAGTGTAGGCGTGGTGCG...TAC')
```

reverse
rRNA 1

```
New rRNA : -----
ID: <unknown id>
Name: <unknown name>
Description: <unknown description>
Number of features: 2
/molecule_type=DNA
Seq('AACCTGGCGGCGTGGATCCGTGTTCCCGGTTAGAAAAGTGTAGGCGTGGTGCG...TGT')
```

} **information lost**

join

Practice

- **Reorganize two Genes of SeqRecord**

- **Write** and **Run** the following codes on a Colab page called “[AnnotationObjects.ipynb](#)”:

```
1 # Reorganize two genes
2 sub_rRNA1 = seq_records[0][380:550]
3 print("rRNA 1 (5'-3'): -----")
4 print(sub_rRNA1)
5 print()
6
7 sub_rRNA2 = seq_records[1][380:550]
8 print("rRNA 2 (5'-3'): -----")
9 print(sub_rRNA2)
10 print()
11
12 # Reverse the sub_rRNA1
13 sub_rRNA1 = sub_rRNA1[::-1]
14 print("rRNA 1 (3'-5', Reversed): -----")
15 print(sub_rRNA1)
16 print()
17
18 # Join two rRNA
19 new_rRNA = sub_rRNA1 + sub_rRNA2
20 print("New rRNA : -----")
21 print(new_rRNA)
22 print()
```

(Solution [URL](#) of this Practice)



Operation: Print in Other Formats

```
1 # Print the first SeqRecord in FASTA format
2 print(seq_records[0].format("fasta"))
```



```
>Z78533.1 C.irapeanum 5.8S rRNA gene and ITS1 and ITS2 DNA
CGTAACAAGGTTTCCGTAGGTGAACCTGCGGAAGGATCATTGATGAGACCGTGAATAAA
CGATCGAGTGAATCCGGAGGACCGGTGTACTIONACCTACCCGGGGGATTGCTCCCGTGGT
GACCTGATTTGTTGTTGGGCCGCCTCGGGAGCGTCCATGGCGGGTTTGAACCTCTAGCC
CGGCGCAGTTTTGGGCGCCAAGCCATATGAAAGCATCACCGGCGAATGGCATTGTCTTCCC
CAAACCCGGAGCGGCGGCGTGTCTGCGGTGCCAATGAATTTTATGACTCTCGCAA
CGGGAATCTTGGCTCTTTGCATCGGATGGAAGGACGCAGCGAAATGCGATAAGTGGTGTG
AATTGCAAGATCCCGTGAACCATCGAGTCTTTTGAACGCAAGTTGCGCCGAGGCCATCA
GGCTAAGGGCACGCCTGCTTGGGCGTCGCGCTTCGTCTCTCCTGCCAATGCTTGCCCG
GCATACAGCCAGGCCGGCGTGGTGCAGGATGTGAAAGATTGGCCCTTGTGCCTAGGTGCG
GCGGGTCCAAGAGCTGGTGTGTTGATGGCCCGAACC CGGAAGAGGTGGACGGATGCTG
GCAGCAGCTGCCGTGCCAATCCCCATGTTGTCGTGCTTGTCCGACAGGCAGGAGAACC
TTCCGAACCCCAATGGAGGGCGGTTGACCGCCATTCGGATGTGACCCAGGTCAGGCGGG
GGCACCCGCTGAGTTTACGC
```

Supported Format

| Format name | Read | Write | Index | Notes |
|-------------|------|-------|-------|--------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
| abi | 1.58 | No | N/A | Reads the ABI "Sanger" capillary sequence traces files, including the PHRED quality scores for the base calls. This allows ABI to FASTQ conversion. Note each ABI file contains one and only one sequence (so there is no point in indexing the file). |
| abi-trim | 1.71 | No | N/A | Same as "abi" but with quality trimming with Mott's algorithm. |
| ace | 1.47 | No | 1.52 | Reads the contig sequences from an ACE assembly file. Uses Bio.Sequencing.Ace internally |
| cif-atom | 1.73 | No | No | Uses Bio.PDB.MMCIFParser to determine the (partial) protein sequence as it appears in the structure based on the atomic coordinates. |
| cif-seqres | 1.73 | No | No | Reads a macromolecular Crystallographic Information File (mmCIF) file to determine the complete protein sequence as defined by the <code>_pdbx_poly_seq_scheme</code> records. |

Full List: <https://bit.ly/3c5q4Fs>

(Note: Some format only support for read, not write)

Practice

- **Print SeqRecord in Other Formats**

- **Write** and **Run** the following codes on a Colab page called “[AnnotationObjects.ipynb](#)”:

- Print SeqRecord in Other Formats

```
✓ ▶ 1 # Print the first SeqRecord in FASTA format
    2 print(seq_records[0].format("fasta"))
```

(Solution [URL](#) of this Practice)



Brief of this Section

- **Bio.SeqRecord.SeqRecord**

- Attributes: `.id`, `.name`, `.description`, `.annotations[]`...

- Objects: `.features[]`, `.seq...`

- Operations:

- Get Length: `len(seq_rec)`

- Slicing: `seq_rec[0:20]` + Annotation Preserved

- Reverse Complement: `seq_rec.reverse_complement()` + features kept

- Drop Nucleotides: `seq_rec[:2]` + `seq_rec[3:]`

- Reorganize: `seq_rec[0][380:550:-1]` + `seq_rec[1][380:550]`

- Output Format: `seq_rec.format("fasta")`





Bio.SeqFeature.Reference

REFERENCES

What is “Reference”

```
LOCUS       Z78533                740 bp    DNA     linear   PLN 30-NOV-2006
DEFINITION  C.irapeanum 5.8S rRNA gene and ITS1 and ITS2 DNA.
ACCESSION   Z78533
VERSION     Z78533.1  GI:2765658
KEYWORDS    5.8S ribosomal RNA; 5.8S rRNA gene; internal transcribed spacer;
            ITS1; ITS2.
SOURCE      Cypripedium irapeanum
  ORGANISM  Cypripedium irapeanum
            Eukaryota; Viridiplantae; Streptophyta; Embryophyta; Tracheophyta;
            Spermatophyta; Magnoliophyta; Liliopsida; Asparagales; Orchidaceae;
            Cypripedioideae; Cypripedium

REFERENCE   1
  AUTHORS   Cox,A.V., Pridgeon,A.M., Albert,V.A. and Chase,M.W.
  TITLE     Phylogenetics of the slipper orchids (Cypripedioideae:
            Orchidaceae): nuclear rDNA ITS sequences
  JOURNAL   Unpublished
-----
REFERENCE   2 (bases 1 to 740)
  AUTHORS   Cox,A.V.
  TITLE     Direct Submission
  JOURNAL   Submitted (19-AUG-1996) Cox A.V., Royal Botanic Gardens, Kew,
            Richmond, Surrey TW9 3AB, UK
```

Bio.SeqIO



Bio.SeqRecord.SeqRecord

.annotations['references']

Bio.SeqFeature.Reference

Bio.SeqFeature.Reference

Attributes of References

```
1 # Get the first SeqRecord
2 rec = seq_records[0]
3
4 # Iterate each reference
5 for ref in rec.annotations['references']:
6     print("Author:", ref.authors)
7     print("Title:", ref.title)
8     print("Journal:", ref.journal)
9     print("Medline ID:", ref.medline_id)
10    print("PubMed ID:", ref.pubmed_id)
11    print("Comments:", ref.comment)
12    print()
```

```
Author: Cox,A.V., Pridgeon,A.M., Albert,V.A. and Chase,M.W.
Title: Phylogenetics of the slipper orchids (Cypripedioideae: Orchidaceae): nuclear rDNA ITS sequences
Journal: Unpublished
Medline ID:
PubMed ID:
Comments:
```

```
Author: Cox,A.V.
Title: Direct Submission
Journal: Submitted (19-AUG-1996) Cox A.V., Royal Botanic Gardens, Kew, Richmond, Surrey TW9 3AB, UK
Medline ID:
PubMed ID:
Comments:
```

Practice

- **Show the Attributes of Bio.SeqFeature.Reference**
 - **Write** and **Run** the following codes on a Colab page called “[AnnotationObjects.ipynb](#)”:

```
1 # Get the first SeqRecord
2 rec = seq_records[0]
3
4 # Iterate each reference
5 for ref in rec.annotations['references']:
6     print("Author:", ref.authors)
7     print("Title:", ref.title)
8     print("Journal:", ref.journal)
9     print("Medline ID:", ref.medline_id)
10    print("PubMed ID:", ref.pubmed_id)
11    print("Comments:", ref.comment)
12    print()
```

(Solution [URL](#) of this Practice)





Bio.SeqFeature.SeqFeature

FEATURES, LOCATIONS, POSITIONS

What is "Features"

- The **Annotations** to explain **sub-sequences**

```
LOCUS       Z78533                740 bp    DNA     linear   PLN 30-NOV-2006
DEFINITION  C.irapeanum 5.8S rRNA gene and ITS1 and ITS2 DNA.
```

| FEATURES | Location/Qualifiers |
|--------------|------------------------------------------------------------------------------------------------|
| source | 1..740 /organism="Cypridium irapeanum" /mol_type="genomic DNA" /db_xref="taxon:49711" |
| misc_feature | 1..380 /note="internal transcribed spacer 1" |
| gene | 381..550 /gene="5.8S rRNA" |
| rRNA | 381..550 /gene="5.8S rRNA" /product="5.8S ribosomal RNA" |
| misc_feature | 551..740 /note="internal transcribed spacer 2" |

```
ORIGIN
1  cgtaacaagg ttccgtagg tgaacctgcg gaaggatcat tgatgagacc gtggaataaa
61  cgatcgagtg aatccggagg accggtgtac tcagctcacc gggggcattg ctccccgtgtg
121 gaccctgatt tgttgttggg ccgcctcggg agcgtccatg gcgggtttga acctctagcc
...
```

Bio.SeqIO



Bio.SeqRecord.SeqRecord

.features[]

Bio.SeqFeature.SeqFeature

Bio.SeqFeature.SeqFeature

Bio.SeqFeature.SeqFeature

Bio.SeqFeature.SeqFeature

Bio.SeqFeature.SeqFeature

Attributes of SeqFeature

| FEATURES | Location/Qualifiers |
|----------|-----------------------------------|
| ① source | ② 1..740 |
| | /organism="Cypripedium irapeanum" |
| | ③ /mol_type="genomic DNA" |
| | /db_xref="taxon:49711" |



```
1 # Get the first SeqRecord
2 rec = seq_records[0]
3
4 # Iterate each feature
5 for feat in rec.features:
6 ① print("Feature Type:", feat.type)
7 ② print("Feature Location:", feat.location)
8 ③ print("Feature Qualifiers:", feat.qualifiers)
9  print()
```

.type (string)

- The type of feature.

.location (Bio.SeqFeature.FeatureLocation)

- The location of a feature.
- A compound object containing:
 - `.start`: The start position (e.g., 1)
 - `.end`: The end position (e.g., 740)

.qualifiers (OrderedDict)

- A dictionary stores the additional information.
 - `Key`: The name of information.
 - `Value`: The information itself.

Practice

- **Show the Attributes of Bio.SeqFeature.SeqFeature**
 - **Write** and **Run** the following codes on a Colab page called “[AnnotationObjects.ipynb](#)”:

```
1 # Get the first SeqRecord
2 rec = seq_records[0]
3
4 # Iterate each feature
5 for feat in rec.features:
6     print("Feature Type:", feat.type)
7     print("Feature Location:", feat.location)
8     print("Feature Qualifiers:", feat.qualifiers)
9     print()
```

(Solution [URL](#) of this Practice)

