

# More on ClustalW parsing


```
from Bio import Clustalw
from Bio.Alphabet import IUPAC
from Bio.Align.FormatConvert import FormatConverter
```

```
"""The function code covers a clustalw alignment to a fasta file
and then writes out the fasta format to a file"""
```

```
def clustal2fasta(cl_filename, fa_filename, datatype=0):
    "Args: existing clustal file, name fasta file, datatype 0=DNA, 1=RNA 2=protein"
    #parse clustaw file
    if datatype==0:
        alignment=Clustalw.parse_file(cl_filename,IUPAC.ambiguous_dna)
    elif datatype==1:
        alignment=Clustalw.parse_file(cl_filename,IUPAC.ambiguous_rna)
    elif datatype==2:
        alignment=Clustalw.parse_file(cl_filename,IUPAC.protein)
    else:
        print "Wrong datatype argument."
        return 0
    #convert alignment to fasta
    converter=FormatConverter(alignment)
    fasta_align=converter.to_fasta()
    print fasta_align #print to screen to see if it worked

    #Write fasta to file
    handle = open(fa_filename, "w")
    handle.write(str(fasta_align))
    handle.close()
```

Change fasta object  
to string for writing



# Genbank Records

## Parsing problem



*Original first line: LOCUS AF308428 922 bp DNA linear INV 12-MAR-2002*

LOCUS AF308428 922 bp DNA INV 12-MAR-2002

DEFINITION Dendroctonus ponderosae elongation factor 1 alpha (ef-1a) gene,  
partial cds.

ACCESSION AF308428

VERSION AF308428.1 GI:12007231

KEYWORDS .

SOURCE mountain pine beetle.

ORGANISM [Dendroctonus ponderosae](#)

Eukaryota; Metazoa; Arthropoda; Tracheata; Hexapoda; Insecta;  
Pterygota; Neoptera; Endopterygota; Coleoptera; Polyphaga;  
Cucujiformia; Phytophaga; Scolytidae; Dendroctonus.

REFERENCE 1 (bases 1 to 922)

AUTHORS Sequeira,A.S., Normark,B.B. and Farrell,B.D.

TITLE Evolutionary assembly of the conifer fauna: distinguishing ancient  
from recent associations in bark beetles

JOURNAL Proc. R. Soc. Lond., B, Biol. Sci. 267 (1460), 2359-2366 (2000)

```

FEATURES             Location/Qualifiers
    source            1..922
                        /organism="Dendroctonus ponderosae"
                        /isolate="hlt18"
                        /db_xref="taxon:77166"
    mRNA              join(<1..600,672..>922)
                        /gene="ef-1a"
                        /product="elongation factor 1 alpha"
    gene               <1..>922
                        /gene="ef-1a"
    CDS                join(<1..600,672..>922)
                        /gene="ef-1a"
                        /codon_start=1
                        /product="elongation factor 1 alpha"
                        /protein_id="AAG45087.1"
                        /db_xref="GI:12007232"
                        /translation="GSFKYAWVLDKDKAERERGITIDIALWKFETSKYYVTIIDAPGH
RDFIKNMITGTSQADCAVLIVAAGTGEFEAGISKNGQTRHALLAFTLGVKQLVVGVN
KMDSTEPPYSES RFEEIKKEVSSYIKKIGYNPAAVAFVPISGWHGDNMLEASAKMPWF
KGWNVERKEGKAEGKTLIDALDAILPPSRPTDKPLRLPLQDVYKIGGIGTVPVGRIET
GVLKPGMVVMFAPANITTEVKS VEMHHEALQEAVPGDNVGFNVKNVSVKELRRGYVAG
DSKNNPP"

```

BASE COUNT 262 a 211 c 215 g 234 t

ORIGIN

```

1 ggttcttca agtacgcgtg ggtcttgac aaactgaaag ccgaacgtga acgtggatt
61 accattgata ttgctctatg gaaattcgaa acatccaagt actatgtcac catcattgat
121 gctcctgggc acagagattt catcaagaac atgatcactg gaaccttca ggccgattgt
181 gccgtcctga tcgtagctgc tggactggg gaattgaag ctggcatctc caaaaatgga
241 caaaccagag aacacgctct gcttgccttc acccttgggt tgaacaact tgcctgggga
301 gtcaataaaa tggactccac cgaaccccca tacagcgaaa gccgtttcga ggaattaag

```

.  
.

901 actccaagaa caaccaccc ag

//

# REVIEW: FASTA RECORD PARSER

```
def fasta_parser(fasta_file):
    "One argument (fasta_file) is a string file name for a fasta text file"
    parser = Fasta.RecordParser() #Create a specific BioPython fasta parser
    file = open(fasta_file, 'r')
    iterator = Fasta.Iterator(file,parser) #iterator goes through file


    """This while loop iterates through each line of the file until no
    more entries are found. Returns the data to self.fasta_data"""
    while 1:
        cur_record = iterator.next() #Gets each record one at a time

        if cur_record is None:
            break #Ends while loop at break

        print cur_record.title
        print cur_record.sequence
```

# GENBANK RECORD PARSER

```
def genbank_parser(genbank_file):  
    "One argument (fasta_file) is a string file name for a fasta text file"  
    feature_parser = GenBank.FeatureParser() #Create a specific BioPython Genbank parser  
    gb_handle = open(genbank_file, 'r')  
    gb_iterator = GenBank.Iterator(gb_handle,feature_parser) #iterator goes through file  
  
    """This while loop iterates through each line of the file until no  
    more entries are found. Returns the data to self.fasta_data"""  
    while 1:  
        cur_entry = gb_iterator.next() #Gets each record one at a time  
  
        if cur_entry is None:  
            break #Ends while loop at break  
  
        #Do something with the record  
        print cur_entry.seq
```



## GENBANK PARSER CONTINUED...

```
"""Continuing the function from the previous slide"""  
print "Extracting data from %s Genbank file" % cur_entry.id  
    # loop through all of the features for the entry  
    print "Name: ", cur_entry.name  
    print "Descrip: ", cur_entry.description  
    print "Annotat: ", cur_entry.annotations  
    for feature in cur_entry.features:  
        print "Feature: ", feature.type  
        print "Qual", feature.qualifiers  
        for sub_feature in feature.sub_features:  
            print "Sub_feature: ", sub_feature  
            print sub_feature.qualifiers  
    print "Seq: ", cur_entry.seq[1:10]
```

**ID:** AF308428.1

**Name:** AF308428

**Descrip:** Dendroctonus ponderosae elongation factor 1 alpha (ef-1a) gene, partial cds.

**Annotat:** {'gi': '12007231', 'organism': 'Dendroctonus ponderosae',  
'taxonomy': ['Eukaryota', 'Metazoa', 'Arthropoda', 'Tracheata', 'Hexapoda', 'Insecta',  
'Pterygota', 'Neoptera', 'Endopterygota', 'Coleoptera', 'Polyphaga', 'Cucujiformia',  
'Phytophaga', 'Scolytidae', 'Dendroctonus'], 'keywords': [], 'data\_file\_division': 'INV',  
'date': '12-MAR-2002',  
'references': [<Bio.SeqFeature.Reference instance at 01F9E74C>],  
'source': 'mountain pine beetle'}

**Feature 1** source

**Qualifiers** {'organism': ['Dendroctonus ponderosae'], 'isolate': ['hlt18'], 'db\_x

**Feature 2** mRNA

**Qualifiers** {'gene': ['ef-1a'], 'product': ['elongation factor 1 alpha']}

**Sub\_feature 1** type: mRNA

location: (<0..600)

ref: None:None

strand: 1

qualifiers:

**Seq:** Seq('GTTCTTTCA', IUPACAmbiguousDNA())

**Ouput from**

**print feature.qualifiers**

**for the fourth feature of the genbank file**

----- **output** -----

**Feature 4 CDS**

**Qualifiers** {'protein\_id': ['AAG45087.1'], 'codon\_start': ['1'], 'translation': ['GSFKYAWVLDKL  
TEVKSVMHHEALQEAVPGDNVGFNVKNVSVKELRRGYVAGDSKNNPP'],  
'gene': ['ef-1a'], 'product': ['elongation factor 1 alpha'], 'db\_xref': ['GI:12007232']}

----- **gb file** -----

```
CDS      join(<1..600,672..>922)
         /gene="ef-1a"
         /codon_start=1
         /product="elongation factor 1 alpha"
         /protein_id="AAG45087.1"
         /db_xref="GI:12007232"
         /translation="GSFKYAWVLDKKAERERGITIDIALWKFETSKYYVTIIDAPGH
RDFIKNMITGTSQADCAVLIVAAGTGEFEAGISKNGQTREHALLAFTLGVKQLVVGVN
KMDSTEPPYSES RFEEIKKEVSSYIKKIGYNPAAVAFVPISGWHGDNMLEASAKMPWF
KGWNVERKEGKAEGKTLIDALDAILPPSRPTDKPLRLPLQDVYKIGGIGTVPVGRIET
GVLKPGMVVMFAPANITTEVKSVMHHEALQEAVPGDNVGFNVKNVSVKELRRGYVAG
DSKNNPP"
```

**try:** `prot = feature.qualifiers["translation"]`