Breakdown of console commands

I. Contents

II. Introduction

For the majority of use-cases, SoFIA's command line consists of three main parts, 1) the target file containing the entities to be annotated, 2) the requested entities and 3) the files containing the provided resources. For detailed information about the command line refer to the documentation on the homepage (https://www.github.com/childsish/sofia). Each following section details the command lines used to produce the examples provided in the main article.

III. Variant annotation

The full command line:

```
> sofia.py aggregate\
      BRCA.vcf:chromosome_id=ucsc\
   -e chromosome\
      position\
      gene_id\
      amino_acid_variant\
      variant_effect\
      variant.info[AF]:1000genomes:resource=1000genomes\
      pathway_id\
   -r gencode.gtf:gene_id=hugo\
      GRCh37.fasta\
      1000genomes.vcf:1000genomes:chromosome_id=ensembl\
      kegg.txt:format=gene_pathway_map:gene_id=entrez
   -m chromosome_id=chromosome_id.txt\
      gene_id=gene_id.txt
```

The file name with the target entities is passed to SoFIA as the first positional argument. The requested entities are defined after the `-e` flag. The provided resources are defined after the `-r` flag. Each requested entity becomes a column in the output. If identifiers need to be converted, the identifier mappings are defined after the `-m` flag. Entities are pre-defined by the template and a full list of entities defined by the template can be obtained with the command '`sofia.py info entity`'. To aid repetition, requested entities and provided resources can be specified in a text file and passed to the framework using the flags `-E` (file of requested entities) and `-R` (file of provided resources).

Target entities:

```
BRCA.vcf:chromosome_id=ucsc
```

This file contains a list of variants derived from the BRCA variant dataset from TCGA. We also specify an extra attribute using '`:`' separated fields. This attributes lets SoFIA know that the chromosome identifiers in the resource follow the UCSC style (ie. starting with the characters 'chr').

Requested entities:

```
chromosome
```

The name of the chromosome.

```
position
```

The position of the variant on the chromosome (1-indexed).

```
gene_id
```

The name of the gene the variant is in.

```
amino_acid_variant
```

The amino acid change wrought by the variant.

```
variant_effect
```

The effect of the variant on the protein product. Terms conform to the standard defined by the sequence ontology.

```
variant.info[AF]:1000genomes:resource=1000genomes
```

The allele frequency of the variant in the 1000 genomes project. This request must specify that the 1000 genomes resource should be used as both the target entities and the 1000 genomes project can provide a variant. This requested entity line can be further broken down into three parts separated by the delimiter '`:`'.

1. Requested entity and property access.

   `variant` the base entity being requested.

   `.info` the 'info' property of the entity.

   `[AF]` the 'AF' field of the 'info' property.

2. Header. This will appear in the header row of the output for the appropriate column.

   `1000genomes` the column name to appear in the header row.

3. Requested attributes. The entity must obtain these attributes during resolution.

   `resource=1000genomes` the variant must use the resource named 1000genomes.

```
pathway_id
```

The pathway(s) in which the variant can be found.

Provided resources

```
1000genomes.vcf:1000genomes:chromosome_id=ensembl
```

Data from the 1000 genomes project. A `variant` entity requires this resource and specifically requests it in the attributes. This provided resource line can be further broken down into three parts separated by the delimiter '`:`'.

1. The name of the resource file; full or relative path.

   `1000genomes.vcf` a collections of variants from the 1000 genomes project.

2. A name for the resource that can be referenced by the entities.

   `1000genomes` entities can refer to the resource with this name.

3. Resource attributes. Entities derived from this resource will gain the listed attributes unless converted.

   `chromosome_id=ensembl` chromosome identifiers follow the Ensembl style

```
gencode.gtf:gene_id=hugo
```

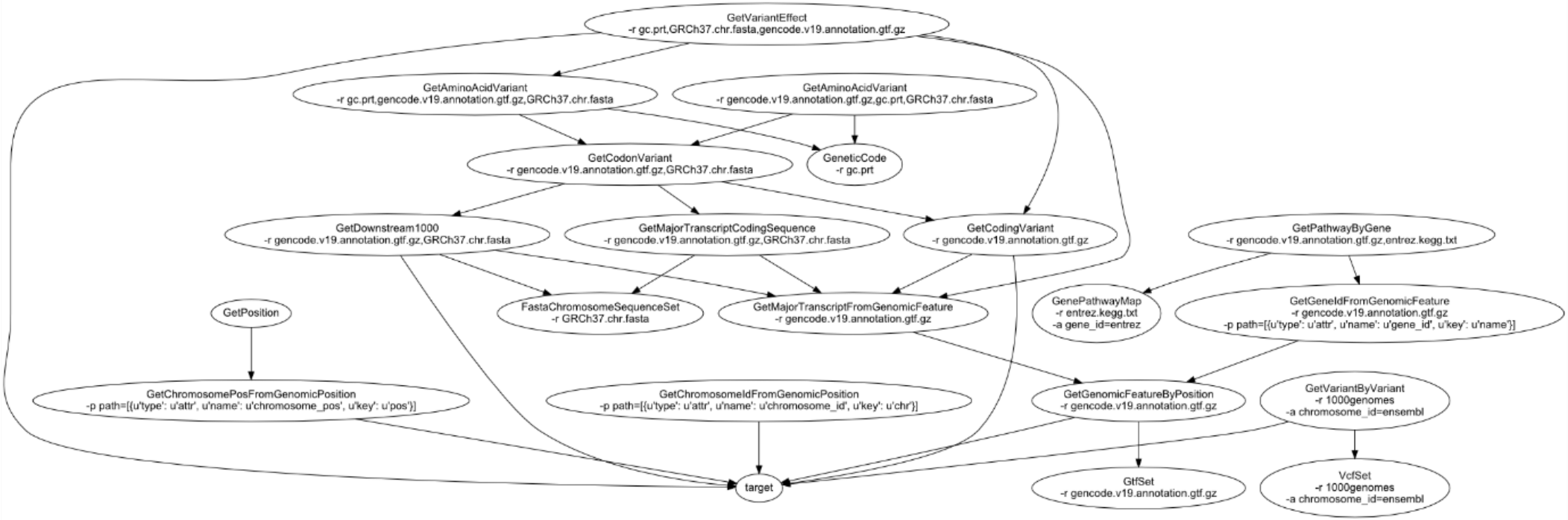
Gene models obtained from the Gencode project. We further specify that the gene identifiers provided are HUGO identifiers. The `gene_id`, `amino_acid_variant`, `variant_effect` entities derive directly from this resource.

```
GRCh37.fasta
```

The human chromosome sequence. The `amino_acid_variant` entity derives directly from this resource.

```
kegg.txt:format=gene_pathway_map:gene_id=entrez
```

A mapping between KEGG pathways and genes. We need to specify that the file format is a gene/pathway mapping as this can not be inferred from the extension. We further specify that the gene identifiers are Entrez identifiers. The `pathway_id` entity derives directly from this resource.


GetVariantEffect
-r gc.prt,GRCh37.chr.fasta,gencode.v19.annotation.gtf.gz
GetAminoAcidVariant
-r gc.prt,gencode.v19.annotation.gtf.gz,GRCh37.chr.fasta
GetAminoAcidVariant
-r gencode.v19.annotation.gtf.gz,gc.prt,GRCh37.chr.fasta
GetCodonVariant
-r gencode.v19.annotation.gtf.gz,GRCh37.chr.fasta
GeneticCode
-r gc.prt
GetDownstream1000
-r gencode.v19.annotation.gtf.gz,GRCh37.chr.fasta
GetMajorTranscriptCodingSequence
-r gencode.v19.annotation.gtf.gz,GRCh37.chr.fasta
GetCodingVariant
-r gencode.v19.annotation.gtf.gz
GetPathwayByGene
-r gencode.v19.annotation.gtf.gz,entrez.kegg.txt
GetPosition
FastaChromosomeSequenceSet
-r GRCh37.chr.fasta
GetMajorTranscriptFromGenomicFeature
-r gencode.v19.annotation.gtf.gz
GenePathwayMap
-r entrez.kegg.txt
-a gene_id=entrez
GetGeneIdFromGenomicFeature
-r gencode.v19.annotation.gtf.gz
-p path=[{u'type': u'attr', u'name': u'gene_id', u'key': u'name'}]
GetChromosomePosFromGenomicPosition
-p path=[{u'type': u'attr', u'name': u'chromosome_pos', u'key': u'pos'}]
GetChromosomeIdFromGenomicPosition
-p path=[{u'type': u'attr', u'name': u'chromosome_id', u'key': u'chr'}]
GetGenomicFeatureByPosition
-r gencode.v19.annotation.gtf.gz
GetVariantByVariant
-r 1000genomes
-a chromosome_id=ensembl
target
GtfSet
-r gencode.v19.annotation.gtf.gz
VcfSet
-r 1000genomes
-a chromosome_id=ensembl

IV. Gene model comparison

The full command line:

```
> sofia.py aggregate\
       BRCA.vcf:chromosome_id=ucsc\
    -e chromosome\
       position\
       gene_id:gene_id=hugo:resource=refseq\
       amino_acid_variant:resource=refseq\
       variant_effect:resource=refseq\
       gene_id:gene_id=hugo:resource=gencode\
       amino_acid_variant:resource=gencode\
       variant_effect:resource=gencode\
    -r GRCh37.fasta:chromosome_id=ucsc\
       gencode.gtf:gencode:chromosome_id=ucsc\
       refseq.gff:refseq:chromosome_id=ncbi\
    -m chromosome_id=chromosome_id.txt
```

Target entities:

```
BRCA.vcf:chromosome_id=ucsc
```

This file is described in the previous section.

Requested entities:

```
chromosome
position
```

These entities do not differ from the previous section.

```
gene_id:gene_id=hugo:resource=refseq
```

The name of the gene the variant is in. The attributes specify that the gene identifier must also be a HUGO identifier and that the entity is derived from the '`refseq`' resource.

```
amino_acid_variant:resource=refseq
```

The amino acid change wrought by the variant. The attributes specify that the entity is derived from the '`refseq`' resource.

```
variant_effect:resource=refseq
```

The effect of the variant on the protein product. Terms conform to the standard defined by the sequence ontology. The attributes specify that the entity is derived from the '`refseq`' resource.

```
gene_id:gene_id=hugo:resource=gencode
```

The name of the gene the variant is in. The attributes specify that the gene identifier must also be a HUGO identifier and that the entity is derived from the '`gencode`' resource.

```
amino_acid_variant:resource=gencode
```

The amino acid change wrought by the variant. The attributes specify that the entity is derived from the '`gencode`' resource.

```
variant_effect:resource=gencode
```

The effect of the variant on the protein product. Terms conform to the standard defined by the sequence ontology. The attributes specify that the entity is derived from the '`gencode`' resource.

Provided resources:

```
GRCh37.fasta:chromosome_id=ucsc
```

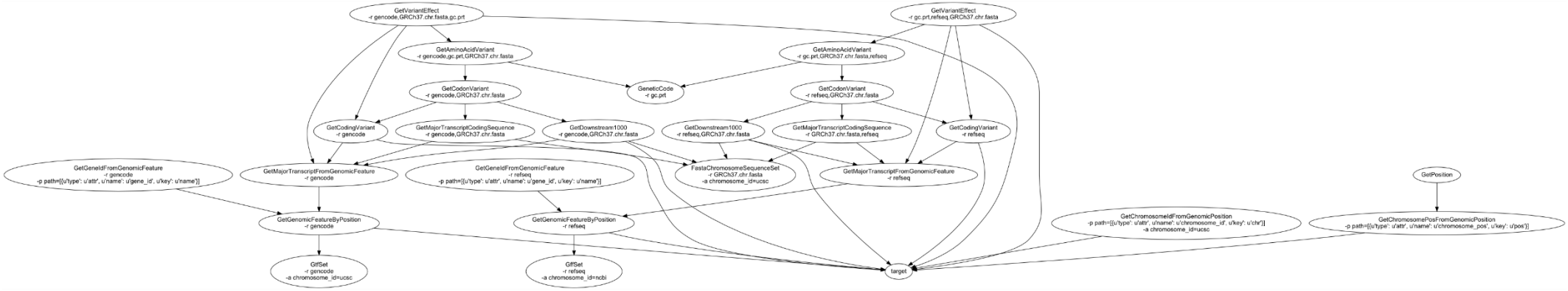
The human chromosome sequence. The `amino_acid_variant` entity derives directly from this resource. The attributes specify that the chromosome identifiers follow the UCSC style.

```
gencode.gtf:gencode:chromosome_id=ucsc
```

Gene models obtained from the Gencode project. The attributes specify that the chromosome identifiers follow the UCSC style. `gene_id`, `amino_acid_variant`, `variant_effect` entities derive directly from this resource.

```
refseq.gff:refseq:chromosome_id=ncbi
```

Gene models obtained from the Refseq. The attributes specify that the chromosome identifiers are NCBI identifiers. `gene_id`, `amino_acid_variant`, `variant_effect` entities derive directly from this resource.


GetVariantEffect
-r gencode,GRCh37.chr.fasta,gc.prt
GetVariantEffect
-r gc.prt,refseq,GRCh37.chr.fasta
GetAminoAcidVariant
-r gencode,gc.prt,GRCh37.chr.fasta
GetAminoAcidVariant
-r gc.prt,GRCh37.chr.fasta,refseq
GeneticCode
-r gc.prt
GetCodonVariant
-r gencode,GRCh37.chr.fasta
GetCodonVariant
-r refseq,GRCh37.chr.fasta
GetCodingVariant
-r gencode
GetMajorTranscriptCodingSequence
-r gencode,GRCh37.chr.fasta
GetDownstream1000
-r gencode,GRCh37.chr.fasta
GetDownstream1000
-r refseq,GRCh37.chr.fasta
GetMajorTranscriptCodingSequence
-r GRCh37.chr.fasta,refseq
GetCodingVariant
-r refseq
GetGeneIdFromGenomicFeature
-r gencode
-p path=[{u'type': u'attr', u'name': u'gene_id', u'key': u'name'}]
GetMajorTranscriptFromGenomicFeature
-r gencode
GetGeneIdFromGenomicFeature
-r refseq
-p path=[{u'type': u'attr', u'name': u'gene_id', u'key': u'name'}]
FastaChromosomeSequenceSet
-r GRCh37.chr.fasta
-a chromosome_id=ucsc
GetMajorTranscriptFromGenomicFeature
-r refseq
GetPosition
GetGenomicFeatureByPosition
-r gencode
GetGenomicFeatureByPosition
-r refseq
GetChromosomeIdFromGenomicPosition
-p path=[{u'type': u'attr', u'name': u'chromosome_id', u'key': u'chr'}]
-a chromosome_id=ucsc
GetChromosomePosFromGenomicPosition
-p path=[{u'type': u'attr', u'name': u'chromosome_pos', u'key': u'pos'}]
GffSet
-r gencode
-a chromosome_id=ucsc
GffSet
-r refseq
-a chromosome_id=ncbi
target

V. Sequence feature calculation

The full command line:

```
> sofia.py aggregate\
       ecoli.gbk\
    -e gene_id\
       codon_adaptation_index\
       effective_number_of_codons\
       translation_start_mfe\
       number_of_pest_sequences\
       number_of_upstream_orfs\
    -r ecoli.gbk
```

Target entities:

```
ecoli.gbk
```

A GenBank file downloaded directly from the NCBI database. As a target, the genes defined in the file are annotated. The accession number at time of download was U00096.3.

Requested entities:

```
gene_id
```

The name of a gene in the GenBank file.

```
codon_adaptation_index
```

The codon adaptation index.

```
effective_number_of_codons
```

The number of effective codons.

```
translation_start_mfe
```

The translation start minimum free energy.

```
number_of_pest_sequences
```

The number of PEST sequences.

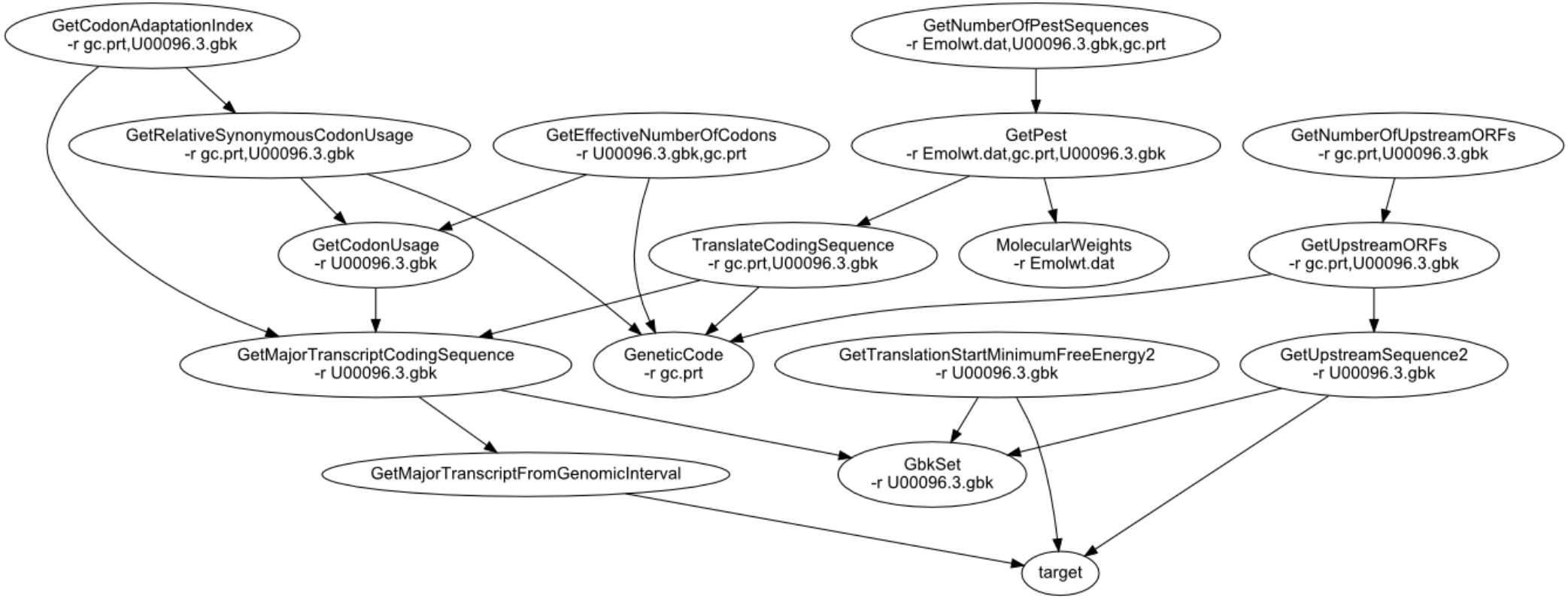
```
number_of_upstream_orfs
```

The number of ORFs upstream of the start codon.

Provided resources:

```
ecoli.gbk
```

The GenBank file again. As a resource, the genomic sequence found in the GenBank file is used to generate the coding sequences needed to calculate the requested entities.


GetCodonAdaptationIndex
-r gc.prt,U00096.3.gbk
GetNumberOfPestSequences
-r Emolwt.dat,U00096.3.gbk,gc.prt
GetRelativeSynonymousCodonUsage
-r gc.prt,U00096.3.gbk
GetEffectiveNumberOfCodons
-r U00096.3.gbk,gc.prt
GetPest
-r Emolwt.dat,gc.prt,U00096.3.gbk
GetNumberOfUpstreamORFs
-r gc.prt,U00096.3.gbk
GetCodonUsage
-r U00096.3.gbk
TranslateCodingSequence
-r gc.prt,U00096.3.gbk
MolecularWeights
-r Emolwt.dat
GetUpstreamORFs
-r gc.prt,U00096.3.gbk
GetMajorTranscriptCodingSequence
-r U00096.3.gbk
GeneticCode
-r gc.prt
GetTranslationStartMinimumFreeEnergy2
-r U00096.3.gbk
GetUpstreamSequence2
-r U00096.3.gbk
GetMajorTranscriptFromGenomicInterval
GbkSet
-r U00096.3.gbk
target