

# Genome Biology Ontology + Gatekeeper

Jasper Koehorst

Laboratory of Systems and Synthetic Biology



# Current formats

- Not designed
  - To store computational annotation meta-data
  - For semantic data mining
  - To query / ask questions
- Therefore
  - No database system like query interface
  - No data provenance of predictions is included

```
LOCUS      SCU49845      5028 bp      PLN      21-JUN-1999
DEFINITION Saccharomyces cerevisiae TCP1-beta gene, partial cds, and Axl2p
            (AXL2) and Rev7p (REV7) genes, complete cds.
ACCESSION  U49845
VERSION    U49845.1      GI:1293613
KEYWORDS   .
SOURCE     Saccharomyces cerevisiae (baker's yeast)
            Saccharomyces cerevisiae
            Eukaryota; Fungi; Ascomycota; Saccharomycotina; Saccharomycetes;
            Saccharomycetales; Saccharomycetaceae; Saccharomyces.
REFERENCE  1 (bases 1 to 5028)
            Torpey,L.E., Gibbs,P.E., Nelson,J. and Lawrence,C.W.
            Cloning and sequence of REV7, a gene whose function is required for
            DNA damage-induced mutagenesis in Saccharomyces cerevisiae
JOURNAL    Yeast 10 (11), 1503-1509 (1994)
PUBMED     7871890
REFERENCE  2 (bases 1 to 5028)
            Roemer,T., Madden,K., Chang,J. and Snyder,M.
            Selection of axial growth sites in yeast requires Axl2p, a novel
            plasma membrane glycoprotein
JOURNAL    Genes Dev. 10 (7), 777-793 (1996)
PUBMED     8846915
REFERENCE  3 (bases 1 to 5028)
            Roemer,T.
            Direct Submission
JOURNAL    Submitted (22-FEB-1996) Terry Roemer, Biology, Yale University, New
            Haven, CT, USA
FEATURES   source
            Location/Qualifiers
            1..5028
            /organism="Saccharomyces cerevisiae"
            /db_xref="taxon:4932"
            /chromosome="IX"
            /map="9"
            CDS
            <1..206
            /codon_start=3
            /product="TCP1-beta"
            /protein_id="AAA98665.1"
            /db_xref="GI:1293614"
            /translation="SSIYNGISTSGLDLNNGTIADNRQLGIVESYKLKRAVVSASEA
            AEVLLRVDNIIRARFRFANRQH"
```

```
0 ##gff-version 3.2.1
1 ##sequence-region ctg123 1 1497228
2 ctg123 . gene      1000 9000 . + . ID=gene00001;Name=EDEN
3 ctg123 . TF_binding_site 1000 1012 . + . ID=tfbs00001;Parent=gene00001
4 ctg123 . mRNA      1050 9000 . + . ID=mRNA00001;Parent=gene00001;Name=EDEN. 1
5 ctg123 . mRNA      1050 9000 . + . ID=mRNA00002;Parent=gene00001;Name=EDEN. 2
6 ctg123 . mRNA      1300 9000 . + . ID=mRNA00003;Parent=gene00001;Name=EDEN. 3
7 ctg123 . exon      1300 1500 . + . ID=exon00001;Parent=mRNA00003
8 ctg123 . exon      1050 1500 . + . ID=exon00002;Parent=mRNA00001,mRNA00002
9 ctg123 . exon      3000 3902 . + . ID=exon00003;Parent=mRNA00001,mRNA00003
10 ctg123 . exon      5000 5500 . + . ID=exon00004;Parent=mRNA00001,mRNA00002,mRNA00003
11 ctg123 . exon      7000 9000 . + . ID=exon00005;Parent=mRNA00001,mRNA00002,mRNA00003
12 ctg123 . CDS       1201 1500 . + 0 ID=cds00001;Parent=mRNA00001;Name=edenprotein.1
13 ctg123 . CDS       3000 3902 . + 0 ID=cds00001;Parent=mRNA00001;Name=edenprotein.1
14 ctg123 . CDS       5000 5500 . + 0 ID=cds00001;Parent=mRNA00001;Name=edenprotein.1
15 ctg123 . CDS       7000 7600 . + 0 ID=cds00001;Parent=mRNA00001;Name=edenprotein.1
16 ctg123 . CDS       1201 1500 . + 0 ID=cds00002;Parent=mRNA00002;Name=edenprotein.2
17 ctg123 . CDS       5000 5500 . + 0 ID=cds00002;Parent=mRNA00002;Name=edenprotein.2
18 ctg123 . CDS       7000 7600 . + 0 ID=cds00002;Parent=mRNA00002;Name=edenprotein.2
19 ctg123 . CDS       3301 3902 . + 0 ID=cds00003;Parent=mRNA00003;Name=edenprotein.3
20 ctg123 . CDS       5000 5500 . + 1 ID=cds00003;Parent=mRNA00003;Name=edenprotein.3
21 ctg123 . CDS       7000 7600 . + 1 ID=cds00003;Parent=mRNA00003;Name=edenprotein.3
22 ctg123 . CDS       3391 3902 . + 0 ID=cds00004;Parent=mRNA00003;Name=edenprotein.4
23 ctg123 . CDS       5000 5500 . + 1 ID=cds00004;Parent=mRNA00003;Name=edenprotein.4
24 ctg123 . CDS       7000 7600 . + 1 ID=cds00004;Parent=mRNA00003;Name=edenprotein.4
```

[Issues](#) [Advance Articles](#) [Publish](#) [Purchase](#) [Alerts](#) [About](#)



**Volume 3, Issue 4**  
November 1987

## An access interface for the MS-DOS diskette format of GenBank(R), a gene sequence database

Michael J Weise

Bioinformatics (1987) 3(4): 313-317. DOI: <https://doi.org/10.1093/bioinformatics/3.4.313>  
Published: 01 November 1987 [Article history](#)

[Cite](#) [Share](#) [Tools](#)

### Article Contents

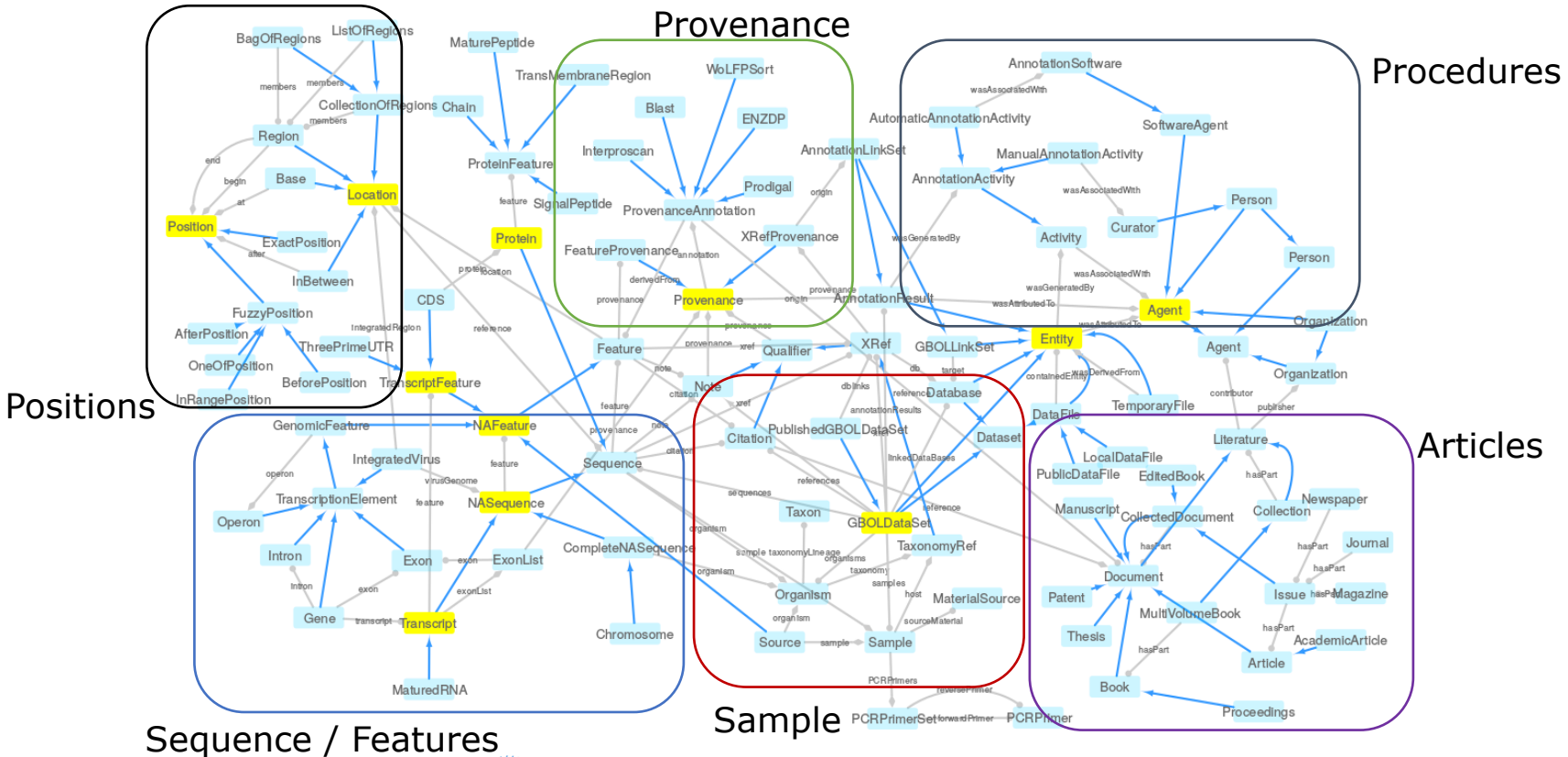
Comments (0)

[<Previous](#) [Next>](#)

#### Abstract

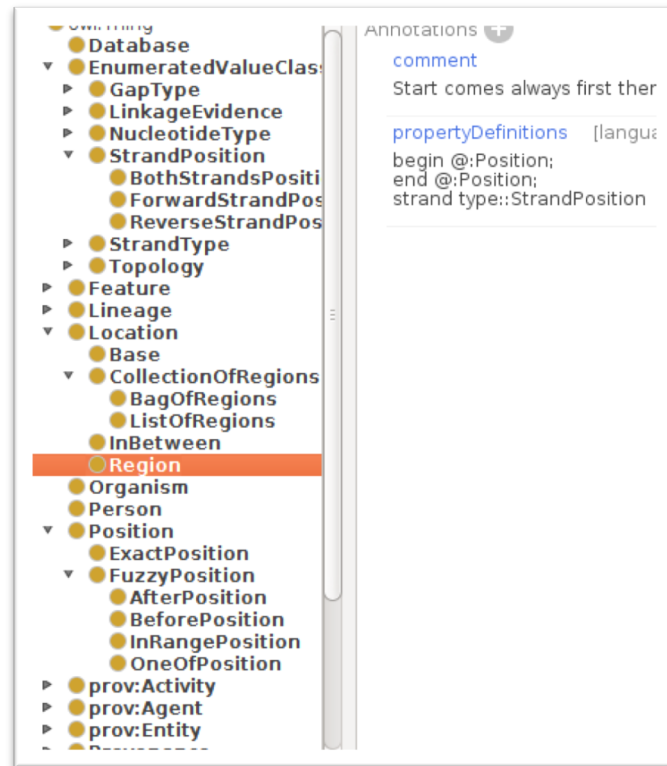
An interface program has been developed for users of MS-DOS computers and the GenBank(R) gene sequence files in their diskette format. With the program a user is able to produce keyword, author and entry name listings of GenBank items or to select GenBank sequences for viewing, printing or decoding. The decode option uncompresses sequence data and yields a character file which has the format used on GenBank magnetic tapes. Program options are chosen by selecting items from command menus. While the program is designed primarily for hard disk operation, it also allows users of diskette-based computers to work with GenBank files.

# Overview of the types in GBOL



# Code generation: EMPUSA

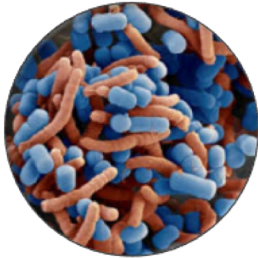
- Linked data graph is free format: Ontology defines structure but does not enforce it.
  - **NEED TO MANTAIN CONSISTENCY**
- **From Ontology (protégé file)**
  - **OWL + ShEx**
- API: Java + R
  - Instance validation included
- > 80.000 lines of code generated
- HTML documentation (website)
- OWL compatible file



# Semantic Annotation Platform with Provenance

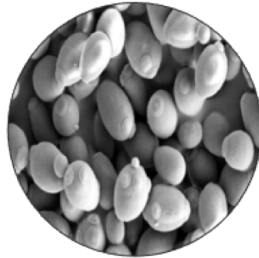
## Conversion types

- EMBL / GenBank
- FASTA
- GFF
- QTL
- VCF
- ...



## Genetic elements

- Gene prediction
- tRNA/rRNA
- Crispr
- ...



## Functional annotation

- BLAST
- Enzyme predictions
- Domain annotation
- Signal peptides
- Transmembrane
- Localization
- ...

