CNSA Sequence Submission Instructions

The following example is the reference format for submitting sequence data to CNSA. This format refers to the format of [GenBank Flat File Format(GBFF)](https://www.ncbi.nlm.nih.gov/Sitemap/samplerecord.html).It is recommended that you use [NCBI Sequin](https://www.ncbi.nlm.nih.gov/Sequin/index.html) to annotate the sequence in GBFF format to CNSA. If a sample has multiple sequences, please put all the sequences of the sample in a file. A single file should contain less than 10,000 sequences.

For example:

LOCUS Seq1 5028 bp DNA 21-JUN-1999

DEFINITION Saccharomyces cerevisiae TCP1-beta gene, partial cds, and Axl2p

(AXL2) and Rev7p (REV7) genes, complete cds.

ACCESSION

DBLINK

VERSION

KEYWORDS .

SOURCE Saccharomyces cerevisiae (baker's yeast)

ORGANISM Saccharomyces cerevisiae

Eukaryota; Fungi; Ascomycota; Saccharomycotina; Saccharomycetes;

Saccharomycetales; Saccharomycetaceae; Saccharomyces.

REFERENCE 1 (bases 1 to 5028)

AUTHORS Torpey,L.E., Gibbs,P.E., Nelson,J. and Lawrence,C.W.

TITLE 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)

AUTHORS Roemer,T., Madden,K., Chang,J. and Snyder,M.

TITLE 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)

AUTHORS Roemer,T.

TITLE Direct Submission

JOURNAL Submitted (22-FEB-1996) Terry Roemer, Biology, Yale University, New Haven, CT, USA

COMMENT

FEATURES Location/Qualifiers

source 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"

/translation="SSIYNGISTSGLDLNNGTIADMRQLGIVESYKLKRAVVSSASEAAEVLLRVDNIIRARPRTANRQHM"

gene 687..3158

/gene="AXL2"

CDS 687..3158

/gene="AXL2"

/note="plasma membrane glycoprotein"

/codon\_start=1

/function="required for axial budding pattern of S.

cerevisiae"

/product="Axl2p"

/protein\_id="AAA98666.1"

/translation="MTQLQISLLLTATISLLHLVVATPYEAYPIGKQYPPVARVNESF

TFQISNDTYKSSVDKTAQITYNCFDLPSWLSFDSSSRTFSGEPSSDLLSDANTTLYFN

VILEGTDSADSTSLNNTYQFVVTNRPSISLSSDFNLLALLKNYGYTNGKNALKLDPNE

VFNVTFDRSMFTNEESIVSYYGRSQLYNAPLPNWLFFDSGELKFTGTAPVINSAIAPE

TSYSFVIIATDIEGFSAVEVEFELVIGAHQLTTSIQNSLIINVTDTGNVSYDLPLNYV

YLDDDPISSDKLGSINLLDAPDWVALDNATISGSVPDELLGKNSNPANFSVSIYDTYG

DVIYFNFEVVSTTDLFAISSLPNINATRGEWFSYYFLPSQFTDYVNTNVSLEFTNSSQ

DHDWVKFQSSNLTLAGEVPKNFDKLSLGLKANQGSQSQELYFNIIGMDSKITHSNHSA

NATSTRSSHHSTSTSSYTSSTYTAKISSTSAAATSSAPAALPAANKTSSHNKKAVAIA

CGVAIPLGVILVALICFLIFWRRRRENPDDENLPHAISGPDLNNPANKPNQENATPLN

NPFDDDASSYDDTSIARRLAALNTLKLDNHSATESDISSVDEKRDSLSGMNTYNDQFQ

SQSKEELLAKPPVQPPESPFFDPQNRSSSVYMDSEPAVNKSWRYTGNLSPVSDIVRDS

YGSQKTVDTEKLFDLEAPEKEKRTSRDVTMSSLDPWNSNISPSPVRKSVTPSPYNVTK

HRNRHLQNIQDSQSGKNGITPTTMSTSSSDDFVPVKDGENFCWVHSMEPDRRPSKKRL

VDFSNKSNVNVGQVKDIHGRIPEML"

gene complement(3300..4037)

/gene="REV7"

CDS complement(3300..4037)

/gene="REV7"

/codon\_start=1

/product="Rev7p"

/protein\_id="AAA98667.1

/translation="MNRWVEKWLRVYLKCYINLILFYRNVYPPQSFDYTTYQSFNLPQ

FVPINRHPALIDYIEELILDVLSKLTHVYRFSICIINKKNDLCIEKYVLDFSELQHVD

KDDQIITETEVFDEFRSSLNSLIMHLEKLPKVNDDTITFEAVINAIELELGHKLDRNR

RVDSLEEKAEIERDSNWVKCQEDENLPDNNGFQPPKIKLTSLVGSDVGPLIIHQFSEK

LISGDDKILNGVYSQYEEGESIFGSLF"

ORIGIN

1 gatcctccat atacaacggt atctccacct caggtttaga tctcaacaac ggaaccattg

61 ccgacatgag acagttaggt atcgtcgaga gttacaagct aaaacgagca gtagtcagct

121 ctgcatctga agccgctgaa gttctactaa gggtggataa catcatccgt gcaagaccaa

181 gaaccgccaa tagacaacat atgtaacata tttaggatat acctcgaaaa taataaaccg

241 ccacactgtc attattataa ttagaaacag aacgcaaaaa ttatccacta tataattcaa

301 agacgcgaaa aaaaaagaac aacgcgtcat agaacttttg gcaattcgcg tcacaaataa

361 attttggcaa cttatgtttc ctcttcgagc agtactcgag ccctgtctca agaatgtaat

421 aatacccatc gtaggtatgg ttaaagatag catctccaca acctcaaagc tccttgccga

481 gagtcgccct cctttgtcga gtaattttca cttttcatat gagaacttat tttcttattc

541 tttactctca catcctgtag tgattgacac tgcaacagcc accatcacta gaagaacaga

601 acaattactt aatagaaaaa ttatatcttc ctcgaaacga tttcctgctt ccaacatcta

661 cgtatatcaa gaagcattca cttaccatga cacagcttca gatttcatta ttgctgacag

721 ctactatatc actactccat ctagtagtgg ccacgcccta tgaggcatat cctatcggaa

781 aacaataccc cccagtggca agagtcaatg aatcgtttac atttcaaatt tccaatgata

841 cctataaatc gtctgtagac aagacagctc aaataacata caattgcttc gacttaccga

901 gctggctttc gtttgactct agttctagaa cgttctcagg tgaaccttct tctgacttac

961 tatctgatgc gaacaccacg ttgtatttca atgtaatact cgagggtacg gactctgccg

1021 acagcacgtc tttgaacaat acataccaat ttgttgttac aaaccgtcca tccatctcgc

1081 tatcgtcaga tttcaatcta ttggcgttgt taaaaaacta tggttatact aacggcaaaa

1141 acgctctgaa actagatcct aatgaagtct tcaacgtgac ttttgaccgt tcaatgttca

1201 ctaacgaaga atccattgtg tcgtattacg gacgttctca gttgtataat gcgccgttac

1261 ccaattggct gttcttcgat tctggcgagt tgaagtttac tgggacggca ccggtgataa

1321 actcggcgat tgctccagaa acaagctaca gttttgtcat catcgctaca gacattgaag

1381 gattttctgc cgttgaggta gaattcgaat tagtcatcgg ggctcaccag ttaactacct

1441 ctattcaaaa tagtttgata atcaacgtta ctgacacagg taacgtttca tatgacttac

1501 ctctaaacta tgtttatctc gatgacgatc ctatttcttc tgataaattg ggttctataa

1561 acttattgga tgctccagac tgggtggcat tagataatgc taccatttcc gggtctgtcc

1621 cagatgaatt actcggtaag aactccaatc ctgccaattt ttctgtgtcc atttatgata

1681 cttatggtga tgtgatttat ttcaacttcg aagttgtctc cacaacggat ttgtttgcca

1741 ttagttctct tcccaatatt aacgctacaa ggggtgaatg gttctcctac tattttttgc

1801 cttctcagtt tacagactac gtgaatacaa acgtttcatt agagtttact aattcaagcc

1861 aagaccatga ctgggtgaaa ttccaatcat ctaatttaac attagctgga gaagtgccca

1921 agaatttcga caagctttca ttaggtttga aagcgaacca aggttcacaa tctcaagagc

1981 tatattttaa catcattggc atggattcaa agataactca ctcaaaccac agtgcgaatg

2041 caacgtccac aagaagttct caccactcca cctcaacaag ttcttacaca tcttctactt

2101 acactgcaaa aatttcttct acctccgctg ctgctacttc ttctgctcca gcagcgctgc

2161 cagcagccaa taaaacttca tctcacaata aaaaagcagt agcaattgcg tgcggtgttg

2221 ctatcccatt aggcgttatc ctagtagctc tcatttgctt cctaatattc tggagacgca

2281 gaagggaaaa tccagacgat gaaaacttac cgcatgctat tagtggacct gatttgaata

2341 atcctgcaaa taaaccaaat caagaaaacg ctacaccttt gaacaacccc tttgatgatg

2401 atgcttcctc gtacgatgat acttcaatag caagaagatt ggctgctttg aacactttga

2461 aattggataa ccactctgcc actgaatctg atatttccag cgtggatgaa aagagagatt

2521 ctctatcagg tatgaataca tacaatgatc agttccaatc ccaaagtaaa gaagaattat

2581 tagcaaaacc cccagtacag cctccagaga gcccgttctt tgacccacag aataggtctt

2641 cttctgtgta tatggatagt gaaccagcag taaataaatc ctggcgatat actggcaacc

2701 tgtcaccagt ctctgatatt gtcagagaca gttacggatc acaaaaaact gttgatacag

2761 aaaaactttt cgatttagaa gcaccagaga aggaaaaacg tacgtcaagg gatgtcacta

2821 tgtcttcact ggacccttgg aacagcaata ttagcccttc tcccgtaaga aaatcagtaa

2881 caccatcacc atataacgta acgaagcatc gtaaccgcca cttacaaaat attcaagact

2941 ctcaaagcgg taaaaacgga atcactccca caacaatgtc aacttcatct tctgacgatt

3001 ttgttccggt taaagatggt gaaaattttt gctgggtcca tagcatggaa ccagacagaa

3061 gaccaagtaa gaaaaggtta gtagattttt caaataagag taatgtcaat gttggtcaag

3121 ttaaggacat tcacggacgc atcccagaaa tgctgtgatt atacgcaacg atattttgct

3181 taattttatt ttcctgtttt attttttatt agtggtttac agatacccta tattttattt

3241 agtttttata cttagagaca tttaatttta attccattct tcaaatttca tttttgcact

3301 taaaacaaag atccaaaaat gctctcgccc tcttcatatt gagaatacac tccattcaaa

3361 attttgtcgt caccgctgat taatttttca ctaaactgat gaataatcaa aggccccacg

3421 tcagaaccga ctaaagaagt gagttttatt ttaggaggtt gaaaaccatt attgtctggt

3481 aaattttcat cttcttgaca tttaacccag tttgaatccc tttcaatttc tgctttttcc

3541 tccaaactat cgaccctcct gtttctgtcc aacttatgtc ctagttccaa ttcgatcgca

3601 ttaataactg cttcaaatgt tattgtgtca tcgttgactt taggtaattt ctccaaatgc

3661 ataatcaaac tatttaagga agatcggaat tcgtcgaaca cttcagtttc cgtaatgatc

3721 tgatcgtctt tatccacatg ttgtaattca ctaaaatcta aaacgtattt ttcaatgcat

3781 aaatcgttct ttttattaat aatgcagatg gaaaatctgt aaacgtgcgt taatttagaa

3841 agaacatcca gtataagttc ttctatatag tcaattaaag caggatgcct attaatggga

3901 acgaactgcg gcaagttgaa tgactggtaa gtagtgtagt cgaatgactg aggtgggtat

3961 acatttctat aaaataaaat caaattaatg tagcatttta agtataccct cagccacttc

4021 tctacccatc tattcataaa gctgacgcaa cgattactat tttttttttc ttcttggatc

4081 tcagtcgtcg caaaaacgta taccttcttt ttccgacctt ttttttagct ttctggaaaa

4141 gtttatatta gttaaacagg gtctagtctt agtgtgaaag ctagtggttt cgattgactg

4201 atattaagaa agtggaaatt aaattagtag tgtagacgta tatgcatatg tatttctcgc

4261 ctgtttatgt ttctacgtac ttttgattta tagcaagggg aaaagaaata catactattt

4321 tttggtaaag gtgaaagcat aatgtaaaag ctagaataaa atggacgaaa taaagagagg

4381 cttagttcat cttttttcca aaaagcaccc aatgataata actaaaatga aaaggatttg

4441 ccatctgtca gcaacatcag ttgtgtgagc aataataaaa tcatcacctc cgttgccttt

4501 agcgcgtttg tcgtttgtat cttccgtaat tttagtctta tcaatgggaa tcataaattt

4561 tccaatgaat tagcaatttc gtccaattct ttttgagctt cttcatattt gctttggaat

4621 tcttcgcact tcttttccca ttcatctctt tcttcttcca aagcaacgat ccttctaccc

4681 atttgctcag agttcaaatc ggcctctttc agtttatcca ttgcttcctt cagtttggct

4741 tcactgtctt ctagctgttg ttctagatcc tggtttttct tggtgtagtt ctcattatta

4801 gatctcaagt tattggagtc ttcagccaat tgctttgtat cagacaattg actctctaac

4861 ttctccactt cactgtcgag ttgctcgttt ttagcggaca aagatttaat ctcgttttct

4921 ttttcagtgt tagattgctc taattctttg agctgttctc tcagctcctc atatttttct

4981 tgccatgact cagattctaa ttttaagcta ttcaatttct ctttgatc

//

|  |  |
| --- | --- |
| **FIELD** | **COMMENTS** |

|  |  |  |
| --- | --- | --- |
| **LOCUS** | The LOCUS field contains a number of different data elements, including locus name, sequence length, and modification date. Each element is described below. |  |

|  |  |  |
| --- | --- | --- |
| **locus name** | You can define a locus name for your sequence. There are no restrictions on the name. The locus name in this example is Seq1. |  |
| **Sequence Length** | Number of nucleotide base pairs (or amino acid residues) in the sequence record. In this example, the sequence length is 5028 bp.  There is no maximum limit on the size of a sequence that can be submitted to CNSA. |  |

|  |  |  |
| --- | --- | --- |
| **Molecule Type** | The type of molecule that was sequenced. In this example, the molecule type is DNA.  Each CNSA record must contain contiguous sequence data from a single molecule type. The various molecule types can include genomic DNA, genomic RNA, precursor RNA, mRNA (cDNA), ribosomal RNA, transfer RNA, small nuclear RNA, and small cytoplasmic RNA. |  |

|  |  |  |
| --- | --- | --- |
| **Modification Date** | The date in the LOCUS field is the **date of last modification**. The sample record shown here was last modified on 21-JUN-1999. |  |

|  |  |  |
| --- | --- | --- |
| **DEFINITION** | Brief description of sequence; includes information such as source organism, gene name/protein name, or some description of the sequence's function (if the sequence is non-coding). If the sequence has a coding region (CDS), description may be followed by a completeness qualifier, such as "complete cds". |  |

|  |  |  |
| --- | --- | --- |
| **ACCESSION** | The unique identifier assigned by CNSA for a sequence record. After the data is submitted, the accession number will be filled by the system. Please leave it blank. |  |
| **VERSION** | The version number assigned by CNSA for a sequence record. After the data is submitted, the version number will be filled by the system. Please leave it blank. |  |
| **DBLINK** | CNSA data accessions associated with this sequence. After the data is submitted, the accession number will be filled by the system. Please leave it blank. |  |
| **KEYWORDS** | Word or phrase describing the sequence. If no keywords are included in the entry, the field contains only a period. |  |

|  |  |  |
| --- | --- | --- |
| **SOURCE** | Free-format information including an abbreviated form of the organism name, sometimes followed by a molecule type. |  |

|  |  |  |
| --- | --- | --- |
| **ORGANISM** | The formal scientific name for the source organism (genus and species, where appropriate) and its lineage, based on the phylogenetic classification scheme used in the [CNGBdb Organism Database](https://db.cngb.org/search/organism/). The complete lineage will be available in the Taxonomy Database. |  |

|  |  |  |
| --- | --- | --- |
| **REFERENCE** | Publications by the authors of the sequence that discuss the data reported in the record. References are automatically sorted within the record based on date of publication, showing the oldest references first.  Some sequences have not been reported in papers and show a status of "unpublished" or "in press". When an accession number and/or sequence data has appeared in print, sequence authors should send the complete citation of the article to datasubs@cngb.org and the CNSA staff will revise the record. |  |

|  |  |  |
| --- | --- | --- |
| **AUTHORS** | List of authors in the order in which they appear in the cited article. |  |

|  |  |  |
| --- | --- | --- |
| **TITLE** | Title of the published work or tentative title of an unpublished work.  Sometimes the words "**Direct Submission**" instead of an article title. This is usually true for the last citation in the REFERENCE field because it tends to contain information about the submitter of the sequence, rather than a literature citation. The last citation is therefore called the "**submitter block**". Additional information is provided below, under the header **Direct Submission**. |  |

|  |  |  |
| --- | --- | --- |
| **JOURNAL** | Abbreviation of the journal name. |  |

|  |  |  |
| --- | --- | --- |
| **PUBMED** | PubMed Identifier (PMID). |  |

|  |  |  |
| --- | --- | --- |
| [**Direct Submission**](https://www.ncbi.nlm.nih.gov/Sitemap/samplerecord.html#SubmitterBlockA) | Contact information of the submitter, such as institute/department and postal address. This is always the last citation in the References field. It is required in all records.  The Authors subfield contains the submitter name(s), Title contains the words "Direct Submission", and Journal contains the address.  The date in the Journal subfield is the date on which the author prepared the submission. |  |

|  |  |  |
| --- | --- | --- |
| **COMMENT** | You can add other information such as Assembly Method, Sequencing Technology, etc. |  |
| **FEATURES** | Information about genes and gene products, as well as regions of biological significance reported in the sequence. These can include regions of the sequence that code for proteins and RNA molecules, as well as a number of other features.  The **location of each feature** is provided as well, and can be a single base, a contiguous span of bases, a joining of sequence spans, and other representations. If a feature is located on the complementary strand, the word "complement" will appear before the base span. If the "**<**" symbol precedes a base span, the sequence is partial on the 5' end (e.g., CDS  <1..206).  If the "**>**" symbol follows a base span, the sequence is partial on the 3' end (e.g., CDS  435..915>). |  |

|  |  |  |
| --- | --- | --- |
| **source** | Mandatory feature in each record that summarizes the length of the sequence, scientific name of the source organism, and Taxon ID number. Can also include other information such as map location, strain, clone, tissue type, etc., if provided by submitter. |  |

|  |  |  |
| --- | --- | --- |
| **Taxon** | A stable unique identification number for the taxon of the source oganism. A taxonomy ID number is assigned to each taxon (species, genus, family, etc.) in the [CNGBdb Taxonomy Database](https://db.cngb.org/search/organism/). See also the Organism field, above. |  |
| **CDS** | Coding sequence; region of nucleotides that corresponds with the sequence of amino acids in a protein (location includes start and stop codons). The CDS feature includes an amino acid translation. Authors can specify the nature of the CDS by using the qualifier "/evidence=experimental" or "/evidence=not\_experimental".  Submitters are also encouraged to annotate the mRNA feature, which includes the 5' untranslated region (5'UTR), coding sequences (CDS, exon), and 3' untranslated region (3'UTR). |  |

|  |  |  |
| --- | --- | --- |
| **<1..206** | Base span of the biological feature indicated to the left, in this case, a CDS feature. (The CDS feature is described above, and its base span includes the start and stop codons.) **Features can be complete, partial on the 5' end, partial on the 3' end, and/or on the complementary strand.** Examples:   1. **complete** feature is simply written as ***n*..*m*** Example:    687..3158  The feature extends from base 687 through base 3158 in the sequence shown 2. **<**     indicates **partial on the 5' end** Example:    <1..206  The feature extends from base 1 through base 206 in the sequence shown, and is partial on the 5' end 3. **>**     indicates **partial on the 3' end** Example:    4821..5028>  The feature extends from base 4821 through base 5028 and is partial on the 3' end 4. **(complement)**  indicates that the feature is on the complementary strand Example:    complement(3300..4037)  The feature extends from base 3300 through base 4037 but is actually on the complementary strand. It is therefore read in the opposite direction on the reverse complement sequence. (For an example, see the third CDS feature in the sample record shown on this documentation. In this case, the amino acid translation is generated by taking the reverse complement of bases 3300 to 4037 and reading that reverse complement sequence in its 5' to 3' direction.) |  |

|  |  |  |
| --- | --- | --- |
| **protein\_id** | A protein sequence identification number, similar to the Version number of a nucleotide sequence. After annotation, if the protein sequence already has an accession number of another database, fill in that accession number such as AAA98666.1 in the above example. If the sequence is a newly discovered sequence, leave it blank. CNSA will assign an accession number to the protein sequence. |  |
| **translation** | The amino acid translation corresponding to the nucleotide coding sequence (CDS). In many cases, the translations are conceptual. Note that authors can indicate whether the CDS is based on experimental or non-experimental evidence. |  |

|  |  |  |
| --- | --- | --- |
| **gene** | A region of biological interest identified as a gene and for which a name has been assigned. The base span for the gene feature is dependent on the furthest 5' and 3' features. |  |

|  |  |  |
| --- | --- | --- |
| **complement** | Indicates that the feature is located on the complementary strand. |  |

|  |  |  |
| --- | --- | --- |
| **Other Features** | Examples of other records that show a variety of biological features; a graphic format is also available for each sequence record and visually represents the annotated features:   * **AF165912** (gene, promoter, TATA signal, mRNA, 5'UTR, CDS, 3'UTR) [GenBank flat file](http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?cmd=Search&db=Nucleotide&term=AF165912%5bpacc%5d&doptcmdl=GenBank) * **AF090832** (protein bind, gene, 5'UTR, mRNA, CDS, 3'UTR) [GenBank flat file](http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?cmd=Search&db=Nucleotide&term=AF090832%5bpacc%5d&doptcmdl=GenBank) * **L00727** (alternatively spliced mRNAs) [GenBank flat file](http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?cmd=Search&db=Nucleotide&term=L00727%5bpacc%5d&doptcmdl=GenBank) |  |

|  |  |
| --- | --- |
| **ORIGIN** | The ORIGIN shows sequence data, 10 bases as one block, and 60 bases on one line. |