Bioinformatics Resources
- Genbank -

Lecture & Exercises
Prof. B. Rost, Dr. L. Richter, J. Reeb
Institut für Informatik I12
National Center for Biotechnology Information, NCBI

- first ideas in the middle of the 80s
- division of the National Library of Medicine (NLM) inside the National Institutes of Health (NIH)
- political mission
- founded in 1988
- David Lipman

NCBI’s political mission as defined by the bill:

1. design, develop, implement, and manage automated systems for the collection, storage, retrieval, analysis, and dissemination of knowledge concerning human molecular biology, biochemistry, and genetics;

2. perform research into advanced methods of computer-based information processing capable of representing and analyzing the vast number of biologically important molecules and compounds;

3. enable persons engaged in biotechnology research and medical care to use systems developed under paragraph (1) and methods described in paragraph (2); and

4. coordinate, as much as is practicable, efforts to gather biotechnology information on an international basis.
Selected NCBI Accomplishments

1990  Blast
1992  GenBank at NCBI
1994  NCBI website
1995  Genomes
1996  OMIM
1997  PubMed

1999  Human Genome
2000  PubMed Central
2003  Entrez Gene / DTDs
2005  NIH Public Access
2007  Genome Reference Consortium
2008  1000 Genomes Project
NCBI Resources

- NCBI currently hosts a vast bunch of resources

- grouped according to various criteria
  - meta data, project-centric
  - method oriented
  - topic oriented

- sorted in the sections:
  databases, downloads, submissions, tools, howtos
Genbank’s Origin

- Walter Goad, Los Alamos National Laboratory
- Los Alamos Sequence Database 1979
- Creation and release of GenBank in 1982
- End of 1982: 2000 sequences
- Move to NCBI in 1992

Minutes from 20th anniversary of GenBank in 2002

“.... Among them is a memo on Los Alamos National Laboratory stationery dated May 9, 1980, that reads: Monday, May 12 at 10:30 Steve Simon invites you for cake and coffee to celebrate 100,000 bases now in the DNA sequence library.”

taken from https://www.genomeweb.com/genbank-turns-20
Growth of GenBank and WGS

doubling approx. every 18 months, diagram for release 207, Apr. 2015


BioinfRes SS 15
Growth of GenBank and WGS

References for GenBank

- PMID: 24217914
- part of the International Nucleotide Sequence Database Collaboration (INSDC) together with EMBL Nucleotide Sequence Database (EMBL-Bank), part of the European Nucleotide Archive (ENA) and the DNA Data Bank of Japan (DDBJ)
## Most Growing Divisions

<table>
<thead>
<tr>
<th>Division</th>
<th>Description</th>
<th>Release 197 (8/2013)</th>
<th>Annual Increase (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>TSA*</td>
<td>Transcriptome shotgun data</td>
<td>8.6333123.935</td>
<td>49.9</td>
</tr>
<tr>
<td>PHG</td>
<td>Phages</td>
<td>119.812.712</td>
<td>42.5</td>
</tr>
<tr>
<td>VRL</td>
<td>Viruses</td>
<td>1.757.202.472</td>
<td>22.9</td>
</tr>
<tr>
<td>BCT</td>
<td>Bacteria</td>
<td>10.281.048.518</td>
<td>21.8</td>
</tr>
<tr>
<td>ENV</td>
<td>Environmental samples</td>
<td>3.743.277.434</td>
<td>10.9</td>
</tr>
<tr>
<td>INV</td>
<td>Invertebrates</td>
<td>2.737.140.464</td>
<td>9.8</td>
</tr>
<tr>
<td>PLN</td>
<td>Plants</td>
<td>5.963.882.822</td>
<td>8.8</td>
</tr>
<tr>
<td>GSS</td>
<td>Genome survey sequences</td>
<td>23.726.384.753</td>
<td>8.1</td>
</tr>
<tr>
<td>VRT</td>
<td>Other vertebrates</td>
<td>3.068.956.026</td>
<td>6.3</td>
</tr>
<tr>
<td>MAM</td>
<td>Other mammals</td>
<td>911.342.025</td>
<td>5.6</td>
</tr>
<tr>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
</tr>
<tr>
<td>TOTAL</td>
<td>All GenBank sequences</td>
<td>654.613.333.676</td>
<td>45.1</td>
</tr>
</tbody>
</table>

* not distributed with the release; there specific project server sections
## Top Organisms (Rel. 207)

<table>
<thead>
<tr>
<th>Organism</th>
<th>Entries</th>
<th>Non-WGS base pair</th>
</tr>
</thead>
<tbody>
<tr>
<td><em>Homo sapiens</em></td>
<td>20.921.637</td>
<td>17.714.786.437</td>
</tr>
<tr>
<td><em>Rattus norvegicus</em></td>
<td>2.193.812</td>
<td>6.526.236.496</td>
</tr>
<tr>
<td><em>Bos taurus</em></td>
<td>2.227.298</td>
<td>5.410.360.312</td>
</tr>
<tr>
<td><em>Zea mays</em></td>
<td>4.177.175</td>
<td>5.201.714.457</td>
</tr>
<tr>
<td><em>Sus scrofa</em></td>
<td>3.297.029</td>
<td>4.895.127.638</td>
</tr>
<tr>
<td><em>Danio rerio</em></td>
<td>1.727.668</td>
<td>3.133.901.682</td>
</tr>
<tr>
<td><em>Triticum aestivum</em></td>
<td>1.796.780</td>
<td>1.927.718.314</td>
</tr>
<tr>
<td>...</td>
<td>...</td>
<td>...</td>
</tr>
<tr>
<td><em>Oryza sativa Japonica Group</em></td>
<td>1.376.410</td>
<td>1.265.556.227</td>
</tr>
<tr>
<td>...</td>
<td>...</td>
<td>...</td>
</tr>
<tr>
<td><em>Arabidopsis thaliana</em></td>
<td>2.578.785</td>
<td>1.202.100.008</td>
</tr>
</tbody>
</table>
| ...                       | ...       | ...
## Distribution of Sequence Files (Rel. 207)

<table>
<thead>
<tr>
<th>Division</th>
<th>Number of Files</th>
</tr>
</thead>
<tbody>
<tr>
<td>BCT</td>
<td>178</td>
</tr>
<tr>
<td>CON</td>
<td>317</td>
</tr>
<tr>
<td>ENV</td>
<td>81</td>
</tr>
<tr>
<td>EST</td>
<td>478</td>
</tr>
<tr>
<td>HTG</td>
<td>142</td>
</tr>
<tr>
<td>INV</td>
<td>126</td>
</tr>
<tr>
<td>PAT</td>
<td>219</td>
</tr>
<tr>
<td>PLN</td>
<td>107</td>
</tr>
<tr>
<td>TSA</td>
<td>175</td>
</tr>
<tr>
<td>VRL</td>
<td>34</td>
</tr>
</tbody>
</table>

Release 207 consists of 2333 text files in total.
Database Files

- GenBank comes in a set of compressed text files available via FTP
- 2333 ASCII files (listed in division plus additional list files) in the range of 0.7-520 MB
- Uncompressed ~709 GB
- Each file consists of two portions
Database Files

- Part 1: highly conserved database file headers

<p>| | | | | | | | | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>10</td>
<td>20</td>
<td>30</td>
<td>40</td>
<td>50</td>
<td>60</td>
<td>70</td>
<td>79</td>
</tr>
</tbody>
</table>

GBBCT1.SEQ Genetic Sequence Data Bank April 15 2015

NCBI-GenBank Flat File Release 207.0

**Bacterial Sequences (Part 1)**

<p>| | | | | | | | | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>51396 loci, 92682287 bases, from 51396 reported sequences</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

- Part 1: sequence entries for that division described in the header
The GenBank Flat File Format

- a sequence entry consists of many records (lines)
- each record consists of two parts
  - Part 1: columns 1-10 / Entry Field Name
  - Part 2: remaining line with the content
Part 1/1

- a keyword, beginning in column 1 of the record (e.g., REFERENCE is a keyword)
- a subkeyword beginning in column 3, with columns 1 and 2 blank (e.g., AUTHORS is a subkeyword of REFERENCE)
- or a subkeyword beginning in column 4, with columns 1, 2, and 3 blank (e.g., PUBMED is a subkeyword of REFERENCE)
Part 1/2

- blank characters, indicating that this record is a continuation of the information under the keyword or subkeyword above it
- a code, beginning in column 6, indicating the nature of an entry (feature key) in the FEATURES table
Part 1/3

- a number, ending in column 9 of the record:
  - This number occurs in the portion of the entry describing the actual nucleotide sequence and designates the numbering of sequence positions

- two slashes (//) in positions 1 and 2, marking the end of an entry
Part 2

- The second part of each sequence entry record contains the information appropriate to its keyword.
- In positions 13 to 80 for keywords.
- In positions 11 to 80 for the sequence.
Entry Field Types (incomplete)

- **Locus**: A short mnemonic name for the entry, chosen to suggest the sequence's definition; **mandatory** keyword/Exactly one record.

- **Definition**: A concise description of the sequence; **mandatory** keyword/One or more records

- **Accession**:
  - the primary accession number is a unique, **unchanging identifier** assigned to each GenBank sequence record.
  - to be used for citations from GenBank
  - **mandatory** keyword/One or more records.
Entry Field Types (incomplete)

- **Version:**
  - compound identifier consisting of the primary accession number and a numeric version number associated with the current version of the sequence data in the record
  - optionally followed by an integer identifier (a "GI") assigned to the sequence by NCBI
  - **mandatory** keyword/exceptly one record
Entry Field Types (incomplete)

- **DBLINK**: provides cross-references to resources that support the existence a sequence record; **optional** keyword/one or more records

- **Keywords**: short phrases describing gene products and other information about an entry; **mandatory** keyword in all annotated entries/one or more records
Entry Field Types (incomplete)

- **Source**: Common name of the organism or the name most frequently used in the literature; **mandatory** keyword in all annotated entries/one or more records/includes one subkeyword

- **Organism**: Formal scientific name of the organism (first line) and taxonomic classification levels (second and subsequent lines); **mandatory** subkeyword in all annotated entries/two or more records
Entry Field Types (incomplete)

- **Reference:**
  - Citations for all articles containing data reported in this entry
  - includes *seven subkeywords* and may repeat
  - *mandatory* keyword/one or more records

- **Journal:** lists the journal name, volume, year, and page numbers of the citation; *mandatory* subkeyword/one or more records

- optional subkeywords: Authors, Consortium, Title, Medline, Pubmed, Remark
Entry Field Types (incomplete)

- **Features**: table containing information on portions of the sequence that code for proteins and RNA molecules; sites of biological significance; optional keyword/one or more records

- **Origin**:
  - specification of how the first base of the reported sequence is operationally located within the genome
  - **mandatory** keyword/exactly one record
  - followed by sequence data (multiple records)

- **//**: entry termination symbol; **mandatory** at the end of an entry/exactly one record
## Detailed Locus Format

<table>
<thead>
<tr>
<th>Columns</th>
<th>Contents</th>
</tr>
</thead>
<tbody>
<tr>
<td>01-05</td>
<td>'LOCUS'</td>
</tr>
<tr>
<td>06-12</td>
<td>spaces</td>
</tr>
<tr>
<td>13-28</td>
<td>Locus name</td>
</tr>
<tr>
<td>29-29</td>
<td>space</td>
</tr>
<tr>
<td>30-40</td>
<td>Length of sequence, right-justified</td>
</tr>
<tr>
<td>41-41</td>
<td>space</td>
</tr>
<tr>
<td>42-43</td>
<td>bp</td>
</tr>
<tr>
<td>44-44</td>
<td>space</td>
</tr>
<tr>
<td>45-47</td>
<td>spaces, ss- (single-stranded), ds- (double-stranded), or ms- (mixed-stranded)</td>
</tr>
<tr>
<td>48-53</td>
<td>NA, DNA, RNA, tRNA (transfer RNA), rRNA (ribosomal RNA), mRNA (messenger RNA), uRNA (small nuclear RNA), left justified</td>
</tr>
<tr>
<td>54-55</td>
<td>space</td>
</tr>
<tr>
<td>56-63</td>
<td>'linear' followed by two spaces, or 'circular'</td>
</tr>
<tr>
<td>64-64</td>
<td>space</td>
</tr>
<tr>
<td>65-67</td>
<td>The division code</td>
</tr>
<tr>
<td>68-68</td>
<td>space</td>
</tr>
<tr>
<td>69-79</td>
<td>Date, in the form dd-MMM-yyyy (e.g., 15-MAR-1991)</td>
</tr>
</tbody>
</table>
Accession Format

- six or eight characters
- six character format:
  - single uppercase letter
  - 5 digits
- eighth character format:
  - two uppercase letters
  - 6 digits
- primary accession number always the first one
Features (Incomplete)

- authoritative source: http://www.insdc.org/documents/feature-table
- feature table contains information about:
  - gene and gene products
  - regions of biological significance
  - can enumerate differences between various reports
  - provides cross-references to other data collections
  - allows hierarchical relation between the features
Layout

- first line of the feature table is a header
- includes the keyword ‘FEATURES’ and the column header ‘Location/Qualifiers’
- each feature consists of:
  - descriptor line containing a feature key and a location
  - a continuation line for the location may follow
  - feature qualifiers may follow the descriptor line
  - key: column 6-20, location starts in column 22
  - qualifiers on subsequent lines at column 22 starting with a ‘/’
A Few Frequent Features

- **CDS**: sequence coding for amino acids in protein (includes stop codon)
- **exon**: region that codes for part of spliced mRNA
- **gene**: region that defines a functional gene, possibly including upstream (promotor, enhancer, etc) and downstream control elements, and for which a name has been assigned
- **mRNA**: messenger RNA
- ....... > 60 features currently
Location and Qualifiers

- Location:
  - a location can be: a single base, a span of bases, a site between two bases, a join of sequences, ...
  - examples: 23, 23..56, 23^24, join (23..56, 87..110)

- Qualifiers:
  - format: from column 22 /qualifier_name[=value]
  - types: free text, enumeration or controlled vocabulary, citations, sequences, feature labels
Database Cross References /db_xref

- **Qualifier**: /db_xref="database:identifier"
- **Definition**: database cross-reference: pointer to related information in another database
- **Scope**: all feature keys
- **Example**: /db_xref="Swiss-Prot:P12345"
- currently > 120 databases available
### Anatomy of a Genbank Flat File

**LOCUS** 6CHH845 5028 bp  DDB  DDB  IDM  11-JUN-1993

**DEFINITION** Saccharomyces cerevisiae TCP1-beta gene, partial cds. and Axl2p (Axl2) and Smp3p (Smp3) genes, complete cds.

**ACCESSION** U19482

**VERSION** 016946.1 @1293613

**SOURCE**  
- Saccharomyces cerevisiae (baker's yeast)
- \[Saccharomyces cerevisiae\]
- \[Saccharomyces cerevisiae\]
- \[Saccharomyces cerevisiae\]

**ORIGIN**  
- Yeast 10 (10), 1555-1550 (1994)

**REMARKS**  
- Torrey, L.E., Glick, J.E., Nelson, J. and Laurence, C.V.

**TITLE** Cloning and sequence of \(\textit{REV}\), a gene whose function is required for DNA damage-induced autophagy in \(\textit{Saccharomyces cerevisiae}\)

**JOURNAL** Yeast 10 (1), 1555-1550 (1994)

**PERMID** 7811890

**REFERENCE**  
- 2 (bases 1 to 528)

**AUTHORS** Boehm, F., Hadden, D., Chang, J. and Snyder, M.

**T1** Selection of axial growth sites in yeast requires Axl2p, a novel \[plasma membrane glycoprotein\]

**JOURNAL** Genes Dev. 10 (7), 777-792 (1996)

**PERMID** 8649315

**REFERENCE**  
- 3 (bases 1 to 528)

**AUTHORS** Boehm, F.

**TITLE** Recent Submission

**JOURNAL** Submitted (22-2-1996) Terry Boehr, Biology, Yale University, New Haven, CT, USA

**FEATURES**  
- Location/Qualifiers
  - Location 1-528 /organism="Saccharomyces cerevisiae"
  - Location 1-528 /db_xref="taxon:4929"
  - Location 1-528 /chromosome="1X"
  - Location 1-528 /map="X"  
- **CDS**
  - 51-206 /codon_start=1 /product="PCP1-beta" /protein_id="AAAS666-1.1" /db_xref="gi:11290614" /translation=""  
  - 93...3158 /gene="REV1" /product="REV1" /protein_id="AAAS666-1.1" /db_xref="gi:11290615" /translation=""
- **Gene**
  - 1-528 /gene="REV1" /note="plasma membrane glycoprotein" /codon_start=1 /function="required for axial budding pattern of S. cerevisiae" /product="Axl2p" /protein_id="AAAS666-1.1" /db_xref="gi:11290615" /translation="XGQGVLQKLRPIRIELLEDIELLWEQKVKFYYRQATKTVCTAMASKSEVLPKPSKQDKRPRPQK"
Anatomy of a Genbank Flat File

**Locus line**

```
LOCUS SCU88415  RNA, DNA  PMN  23-NOV-1998
DEFINITION Saccharomyces cerevisiae, taxon, gene, partial cDNA, and A15P (AALP) and Rev7p (REV7) gene, GenBank
ACCESSION U99845
VERSION U99845.1 02-1293683
KEYWORDS .
SOURCE Saccharomyces cerevisiae
ORIGIN Saccharomyces cerevisiae
  Drosophila: Fungi; Neurospora; Saccharomyces; 
  Rebecronut; Saccharomyces; 
  Saccharomyces; 
  source: Saccharomyces.
REFERENCE 1 (bases 1 to 1926)
  Turner, L.E., Gilbe, P.F., Nelson, J. and Lawrence, C.V.
  TITLE Cloning and sequence of REV7, a gene whose function is required for 
  DNA damage-induced mutagenesis in Saccharomyces cerevisiae 
  Yeast 10 (11), 1553-1559 (1994)
  JOURNAL 7571090
  PUBLISHED 1996-02-19
  REVISION 2 (bases 1 to 1926)
  Authors: Helen, T., Naden, P., Chang, J. and Snyder, M.
  TITLE Selecting of axial growth sites in yeast requires A15P, a novel 
  plasma membrane glycoprotein 
  JOURNAL Genes Dev. 10 (7), 777-792 (1996)
  FORMED 856135
  REVISION 3 (bases 1 to 5208)
  Authors: Helen, T.
  TITLE Direct Submission 
  JOURNAL Submitted (22-SEP-1996) Terry Helen, Biology, Yale University, New 
  Haven, CT, USA
  FEATURES .
  LOCATION/Predictors
  .
  .
  .
  .
  .
  .
  .
  .
  .
  .
```

BioinfRes SS 15

**ROSTLAB.**
Anatomy of a Genbank Flat File

Accession Number, Version and GI number
### Anatomy of a Genbank Flat File

#### Feature table with annotations:

<table>
<thead>
<tr>
<th>Feature</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>gene</td>
<td>1350</td>
</tr>
<tr>
<td>CDS</td>
<td>352</td>
</tr>
<tr>
<td>feature</td>
<td>required</td>
</tr>
<tr>
<td>function</td>
<td>for axial budding of S. cerevisiae</td>
</tr>
</tbody>
</table>
Useful Resources from NCBI

- **Materials:**
  - Electronic bookshelf
- **NCBI manuals**
- **text books**
Useful Resources from NCBI

- Processes, e.g. Prokaryotic Genome Annotation Pipeline
- designed for bacterial and archaeal genomes
- multi-level process including protein-coding gene prediction and functional genome unit like rRNAs, tRNAs, small RNAs, pseudogenes control regions, repeats, insertion elements a.s.f.
- combination of *ab-initio* prediction and homology based methods
Useful Resources from NCBI

- reference databases: RefSeq
- comprehensive, integrated, non-redundant, well-annotated set of sequences, including genomic DNA, transcripts, and proteins
- stable reference for genome annotation, esp. subset of RefSeqGene
- reference sequences
- reference coordinates
- accessible via BLAST, Entrez and FTP
RefSeq

- created by:
  - Eukaryotic Genome Annotation Pipeline
  - Prokaryotic Genome Annotation Pipeline
  - Manual curation
  - Submission to INSDC members
- reflect current knowledge of sequences data and biology
- format consistency
- Accession number contains an “_”
RefSeq Growth

Accessions complete

- Species
- Total Accessions
- Nucleotides
- Transcripts
- Proteins

BioinfRes SS 15
Databases Accessible via Entrez

<table>
<thead>
<tr>
<th>Literature</th>
<th>Genes</th>
</tr>
</thead>
<tbody>
<tr>
<td>Books</td>
<td>EST</td>
</tr>
<tr>
<td>MoSH</td>
<td>Gene</td>
</tr>
<tr>
<td>NLM Catalog</td>
<td>GEO DataSets</td>
</tr>
<tr>
<td>PubMed</td>
<td>GEO Profiles</td>
</tr>
<tr>
<td>PubMed Central</td>
<td>HomoloGene</td>
</tr>
<tr>
<td>PubMed Health</td>
<td>PopSet</td>
</tr>
<tr>
<td><strong>Health</strong></td>
<td>UniGene</td>
</tr>
<tr>
<td>ClinVar</td>
<td><strong>Proteins</strong></td>
</tr>
<tr>
<td>dbGaP</td>
<td>Conserved Domains</td>
</tr>
<tr>
<td>GTR</td>
<td>Protein</td>
</tr>
<tr>
<td>MedGen</td>
<td>Protein Clusters</td>
</tr>
<tr>
<td>OMIM</td>
<td>Structure</td>
</tr>
<tr>
<td><strong>Genomes</strong></td>
<td><strong>Chemicals</strong></td>
</tr>
<tr>
<td>Assembly</td>
<td>BioSystems</td>
</tr>
<tr>
<td>BioProject</td>
<td>PubChem BioAssay</td>
</tr>
<tr>
<td>BioSample</td>
<td>PubChem Compound</td>
</tr>
<tr>
<td>Clone</td>
<td>PubChem Substance</td>
</tr>
</tbody>
</table>
| dbVar | **S**
| Epigenomics | **E**
| Genome | **N**
| GSS | **D**
| Nucleotide | **X**
| Probe | **O**
| **R**


BioinfRes SS 15
Computation: Blast at NCBI
### Sequences producing significant alignments:

<table>
<thead>
<tr>
<th>Description</th>
<th>Max score</th>
<th>Total score</th>
<th>Query cover</th>
<th>E value</th>
<th>Accession</th>
</tr>
</thead>
<tbody>
<tr>
<td>Yersinia enterocolitica mshE, mshA, mshC and mshH genes</td>
<td>1441</td>
<td>1441</td>
<td>100%</td>
<td>0.0</td>
<td>X77217.1</td>
</tr>
<tr>
<td>Vibrio cholerae O1 biotype El Tor strain 0167 chromosome I, complete sequence</td>
<td>1408</td>
<td>1408</td>
<td>100%</td>
<td>0.0</td>
<td>CP000642.1</td>
</tr>
<tr>
<td>Vibrio cholerae strain 2016EL-2175 chromosome 1, complete sequence</td>
<td>1408</td>
<td>1408</td>
<td>100%</td>
<td>0.0</td>
<td>CP007691.1</td>
</tr>
<tr>
<td>Vibrio cholerae MS56 DNA, complete genome chromosome I</td>
<td>1408</td>
<td>1408</td>
<td>100%</td>
<td>0.0</td>
<td>AP004524.1</td>
</tr>
<tr>
<td>Vibrio cholerae 0524 chromosome I, complete sequence</td>
<td>1408</td>
<td>1408</td>
<td>100%</td>
<td>0.0</td>
<td>CP003391.1</td>
</tr>
<tr>
<td>Vibrio cholerae O1 strain 2010T-1768 chromosome 1, complete sequence</td>
<td>1408</td>
<td>1408</td>
<td>100%</td>
<td>0.0</td>
<td>CP003669.1</td>
</tr>
<tr>
<td>Vibrio cholerae VO-1/2 chromosome 1, complete sequence</td>
<td>1408</td>
<td>1408</td>
<td>100%</td>
<td>0.0</td>
<td>CP001485.1</td>
</tr>
<tr>
<td>Vibrio cholerae O95 chromosome I, complete sequence</td>
<td>1408</td>
<td>1408</td>
<td>100%</td>
<td>0.0</td>
<td>CP001295.1</td>
</tr>
<tr>
<td>Vibrio cholerae O1/6,6-2 chromosome I, complete sequence</td>
<td>1408</td>
<td>1408</td>
<td>100%</td>
<td>0.0</td>
<td>CP001293.1</td>
</tr>
<tr>
<td>Vibrio cholerae O95 chromosome 2, complete genome</td>
<td>1408</td>
<td>1408</td>
<td>100%</td>
<td>0.0</td>
<td>CP000697.1</td>
</tr>
<tr>
<td>Vibrio cholerae O1 biotype El Tor strain N16681 chromosome I, complete sequence</td>
<td>1408</td>
<td>1408</td>
<td>100%</td>
<td>0.0</td>
<td>AE000852.1</td>
</tr>
<tr>
<td>Vibrio cholerae LMA3994 chromosome I, complete sequence</td>
<td>1391</td>
<td>1391</td>
<td>100%</td>
<td>0.0</td>
<td>CP000565.1</td>
</tr>
<tr>
<td>Vibrio cholerae strain 1154-74, complete genome</td>
<td>1325</td>
<td>1325</td>
<td>100%</td>
<td>0.0</td>
<td>CP010811.1</td>
</tr>
<tr>
<td>Vibrio cholerae strain 10432-62, complete genome</td>
<td>1286</td>
<td>1286</td>
<td>100%</td>
<td>0.0</td>
<td>CP010812.1</td>
</tr>
<tr>
<td>Synthetic construct Vibrio cholerae dpo FLJ498027.01B mshE gene, complete sequence</td>
<td>808</td>
<td>808</td>
<td>57%</td>
<td>0.0</td>
<td>DQ117326.1</td>
</tr>
<tr>
<td>Synthetic construct Vibrio cholerae dpo FLJ498045.01B mshF gene, complete sequence</td>
<td>272</td>
<td>272</td>
<td>18%</td>
<td>0.0</td>
<td>DQ117327.1</td>
</tr>
<tr>
<td>Synthetic construct Vibrio cholerae dpo FLJ498030.01B mshG gene, complete sequence</td>
<td>213</td>
<td>213</td>
<td>16%</td>
<td>0.0</td>
<td>DQ117325.1</td>
</tr>
<tr>
<td>Vibrio cholerae DAI-2 genome (bad)</td>
<td>200</td>
<td>200</td>
<td>14%</td>
<td>3e-47</td>
<td>AE096272.1</td>
</tr>
</tbody>
</table>

![Diagram](image-url)
### V cholerae mshB, mshA, mshC and mshD genes

<table>
<thead>
<tr>
<th>Query</th>
<th>Subject</th>
<th>Score</th>
<th>Expect</th>
<th>Identities</th>
<th>Gaps</th>
<th>Strand</th>
<th>Plus/Minus</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1</td>
<td>0</td>
<td>0.0</td>
<td>870/870(100%)</td>
<td>0/870(0%)</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

#### Related Information

**Vibrio cholerae C1 bovar EL Tor strain F1147 chromosome I, complete sequence**

Sequence ID: gbi:CP008482.1  Length: 3040756  Number of Matches: 1
External Interfaces to Entrez / API

- there are a number of APIs to access the various services from NCBI, described at:
- basic searching:
  - esearch.fcgi?db=<database>&term=<query>
  - Input: Entrez database (&db); any Entrez text query (&term)
  - Output: List of UIDs matching the Entrez query
• storing results:
  - esearch.fcgi?
    db=<database>&term=<query>&usehistory=y
  - input: any Entrez text query (&term); Entrez database (&db); &usehistory=y
  - output: web environment (&WebEnv) and query key (&query_key) parameters specifying the location on the Entrez history server of the list of UIDs matching the Entrez query
• Associating Search Results with Existing Search Results:
  - esearch.fcgi?
    db=<database>&term=<query1>&usehistory=y
  - esearch.fcgi?
    db=<database>&term=<query2>&usehistory=y&WebEnv=$web1
  - Input: Any Entrez text query (&term); Entrez database (&db); &usehistory=y; Existing web environment (&WebEnv) from a prior E-utility call
  - Output: Web environment (&WebEnv) and query key (&query_key) parameters specifying the location on the Entrez history server of the list of UIDs matching the Entrez query
E-utility Webinar

- https://www.youtube.com/watch?v=iCFVVexp3Oo