

Guideline for annotating genomes with GenDB 2.0 - Webinterface

May 13, 2004

Contents

Contents	1
1 System Requirements and Login	3
1.1 Requirements for using the GenDB 2.0 webfrontend	3
1.2 Logging in	4
2 Annotation	5
2.1 First steps with GenDB 2.0	5
2.2 Annotating with GenDB 2.0	5
3 Glossary	17

1 System Requirements and Login

1.1 Requirements for using the GenDB 2.0 webfrontend

Compatible browser

To run GenDB please use Netscape (at least version 7.0), or Mozilla (at least version 1.1), and preferably under UNIX/Linux environment. **Do not use Internet Explorer!!**

Browser settings

Although there are three possible screen resolutions to be used (800x600, 1024x768 or 1280x1024), the larger one (1280x1024) is favored.

Activating Java Script

From the *Edit* context menu of your browser window, select:

- Preferences: Advanced: - enable Java Script.
- Preferences: Advanced: Scripts & Plug-ins:
 - enable Java Script for navigator;
 - activate all options checking all boxes in the “Allow scripts to:” window.

Activating Cookies

From the *Edit* context menu of your browser window, select:

- Preferences: Privacy & Security: Cookies: - enable Cookies for the originating web site only.

Accepting pop up windows

From the *Edit* context menu, select:

- Preferences: Privacy & Security: Pop up windows: block unrequested pop up windows ***must be deactivated!***

1.2 Logging in

Follow the next instructions to log into GenDB 2.0:

- Start the webfrontend of GenDB 2.0 by entering the following URL:
<https://www.cebitec.uni-bielefeld.de/software/genadb/>
- **Never** save your password. To avoid any mistakes, disable automatic password storage (Preferences: Privacy & Security: Remember passwords *must be deactivated!*)
- Log in (username and password)
- When you log in the first time, change your preset password. You can ask the respective project leaders for passwords and accounts for annotators.
- Select your project from the pull down menu: GenDB_XYZ (annotator)

After logging in, the GenDB 2.0 main window is launched, showing the “Contig View” (Figure 1.1).

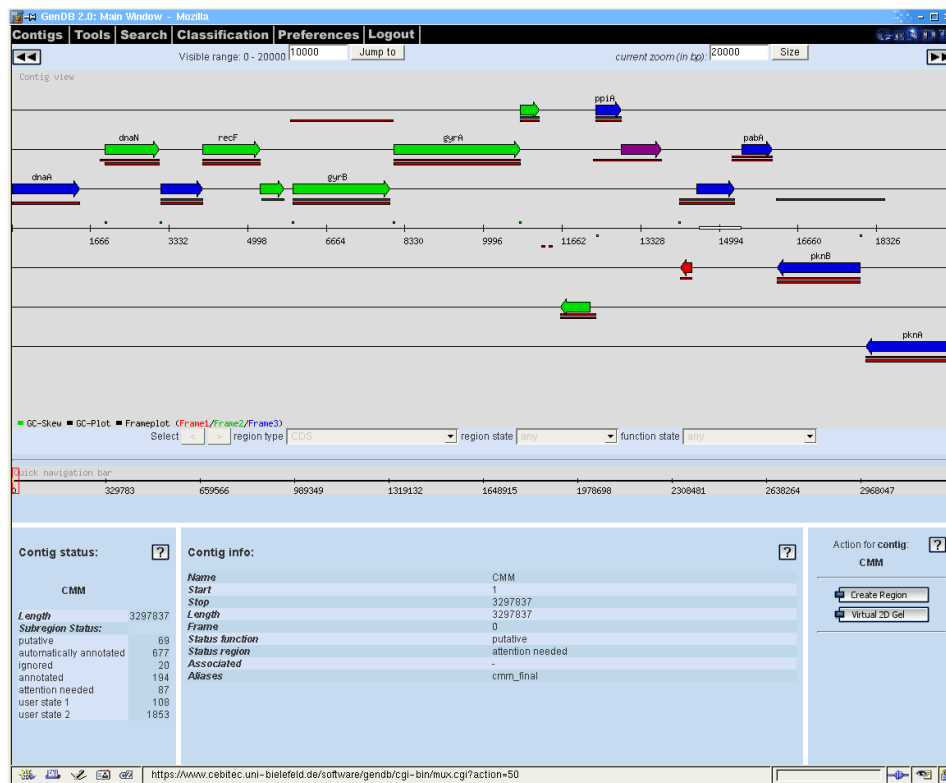


Figure 1.1: GenDB Main Window (“Contig View”).

2 Annotation

2.1 First steps with GenDB 2.0

Before you begin annotating a genome with GenDB 2.0, we suggest to take a course of instruction by any experienced user or GenDB developer. Afterwards, it is recommended that you “play around” with the program for some time to get practice and provide confidence to the program.

If there are still questions, each window has been provided with an **online-help** (with the symbol [?]), which describes the utilization and function of the respective window (note: the online help documentation is still under construction (April 2004)).

2.2 Annotating with GenDB 2.0

This part of the annotation guideline provides practical hints to annotators. To differentiate these practical hints from general information about annotation with GenDB 2.0, the former ones are highlighted **green**.

Selecting a contig

In genome projects where the genomic sequence still consists of more than one contig and/or more than one replicon, the user has to initially select the contig to be annotated. This selection can be done in the “Main Window” of GenDB, in the upper menu bar by clicking the button “Contig”.

Region/CDS

Step 1: Region (CDS) selection. The selection happens either by directly clicking a CDS in the “Contig View” window, or by searching in the “Search” window (see upper menu bar “Search”), or even by searching in the GO, KEGG or COG windows (see upper menu bar “Classification”).

These regions have been predicted and generated by GenDB via the gene prediction tools *Glimmer* and *Critica*. Some genes may have been missed by the predictions, because both tools are prone to errors. This leads to “holes” in the gene prediction, which often present a GC content and codon usage deviation relative to the rest of the genome.

Hints for annotators: position and length of those “holes” should be reported to the super-annotators ¹. If needed, they can generate regions manually.

Possible frame shifts are also partly predicted by *Critica* (boxes on the middle lines between the frames). But these frame shift predictions may also occur because of operations.

Hint for annotators: any well-founded frame shift suspicion should also be notified to the super-annotator (region number or approx. bp position).

Observations/report

Step 2: After the selection of a region (CDS), the corresponding observations or report should be opened (menu “Action for Region”, at the bottom right of the “Main Window”) and analyzed (the windows are launched by clicking the corresponding menu items). Both windows weight the tool outputs (observation/report) from “good” to “bad”.

Tools in observations

HMMPfm, InterPro and other HMMs: in the DB column, the hits of the tools link to the corresponding matrices, which give more information and contain IPR or GO numbers for the annotation.

Hints for annotators: the HMM matches have to be checked regarding the reliability of the model (in description), the e-value (below 1e-10, but at least below 1e-4), the presence of highly conserved rests, i.e., the active centers of enzymes. This is possible and relatively easy via Pfam result and the link to the description.

TMHMM, HTH, SignalP and other tools for prediction of topology and signal peptides: observations of these tools are, for instance:

- transmembrane helices predicted by TMHMM;
- transport proteins in the membrane/periplasmatic protein = secretion signal predicted by SignalP;
- DNA binding via HTH = helix-turn-helix motif predicted by HTH.

Psi-Blast against COG: this tool delivers the **COG categories**.

Hints for annotators: COG categories can not be given randomly, because they rely on the homologies within their members.

¹The super-annotator is a person (or a small group of persons) who better knows the organism and its genome. He is the project’s *annotation leader*, i.e, he is informed by the annotators about any problems with the annotation. The assignment of this role is project-specific.

Psi-Blast against COG: this tool delivers the **EC number**.

Hints for annotators: it can happen in this computation that several good hits lead to different EC numbers, i.e., that different enzymes are categorized with the same number. The decision which EC number will be used for the annotation is up to the annotator, but has to be validated via the KEGG trees (see the “KEGG window” in GenDB).

BlastP and BlastN: these tools deliver the Blast analyzes at the amino acid sequence level (BlastP) and the nucleic acid sequence level (BlastN).

Hints for annotators: the Blast hits’ credibility have to be checked, for instance by comparing them to other hits and/or via DB and results of the hits.

Further Information on Blast homologies (items 3 and 4)

- Hits that have been characterized by wet-lab experiments are better. Here we mention the Swissprot Database, for instance.
- The alignment should be checked with respect to homology and sequence coverage.
- Evaluation does not mean that the best hits are merely adopted, but the rest also has to be checked for additional experimental evidence, if possible.
- It should also be checked if the homology encloses the whole sequence length or only the region of conserved domains (domains have to be noted as a database entry).
- Exceptional homologies should be revised in detail (keyword: horizontal gene transfer)
- If available, also neighbor genes of the analyzed CDS should be taken in consideration. In some cases this could lead to a better function assignment (e.g. operons).
- Inspecting if the gene order is also conserved in other organisms could lead to interesting and helpful results. Such an analysis can be done with the String database (<http://www.string.embl.de>).

Hits have different levels

Hits of level 1 (dark green), 2 (light green) and 3 (yellow) are significant, while hits with lower levels are probably only interesting for the annotation of “hypothetical proteins”.

The display of the observations is configurable

The observations are configurable concerning the tools' output and the display of results. The configuration is done in the "Main Window", in the item "Preferences" (upper menu bar). By clicking on the submenu "Config Dialog" a new window is launched, called "Preferences Window" (Figure 2.1). Here the tools' settings can be done in the submenu "Tool Results". In turn, the tools can be configured in regard to the amount of output results (-1 = all; "number" = corresponding amount). A further configuration possibility is in the "Observations Window" itself: sorting tool results; amount of displayed levels; ascending or descending sorting of results; maximal amount of output results (-1 = all; "number" = corresponding amount).

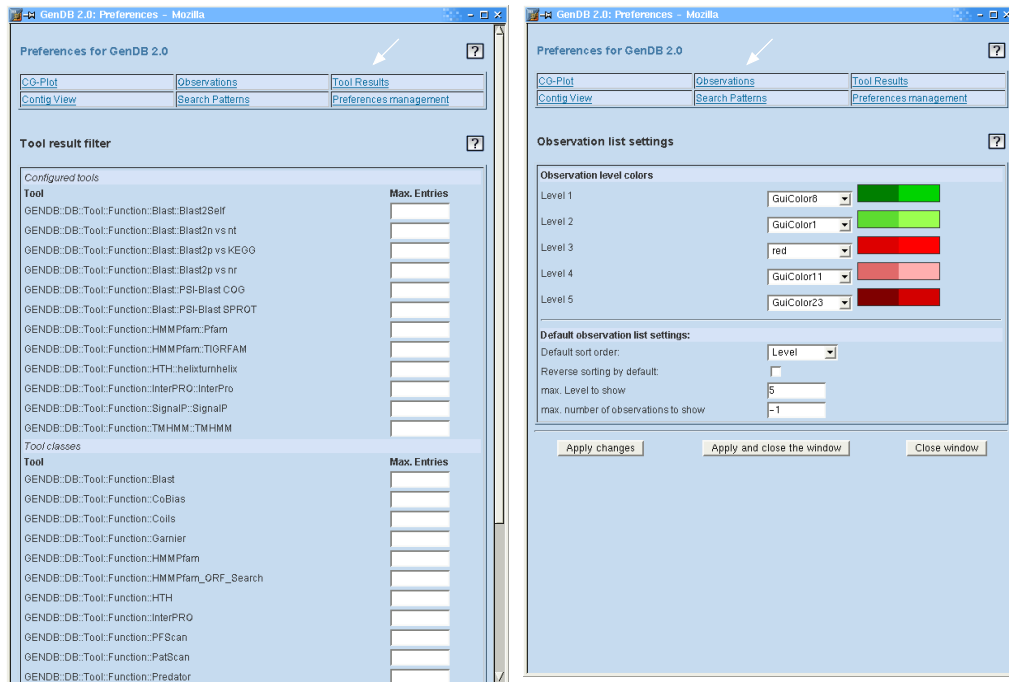


Figure 2.1: The configuration window for "Tool result" and "Observations".

The region editor

The region editor (Figure 2.2) permits the user to change the start codon position of a region, in case the prediction is suspicious. Indications to an incorrect start codon prediction can be:

- different results from Critica and Glimmer (e.g., region acquired through horizontal gene transfer)

- proteins with high homology show different starts (Blast results)
- other observations point to other starts

In these cases, the start codon should be corrected and the resulting changes in the observation should be checked.

Criteria for choosing the start codon

- Shine-Delgarno/RBS predicted by Critica (AG rich region sequence 5-11 bp upstream the start: e.g. GGAGG)
- no overlaps with other CDSs (except operons)
- same start codon as high homologous regions (approx. the same position of homologous regions in Blast results). In this case caution is demanded, as databases can also contain incorrect entries!
- in signal peptides: usual length between 20 and 30 amino acids (quality of signal peptides can be tested through SignalP)

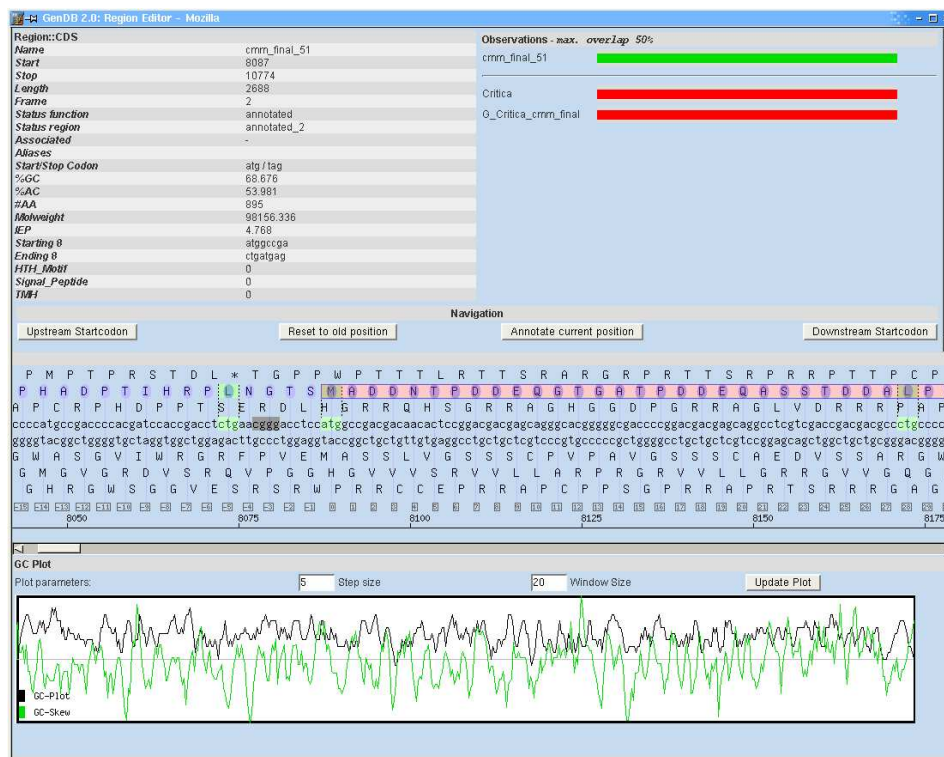


Figure 2.2: The “Region Editor”.

Annotating a region

After selecting a region to be annotated, the annotation dialog (Figure 2.3) can be opened by clicking the button “Annotate Region” in the “Action for Region” window (in the “Main window”).

Hints for annotators: which fields give which information and which fields should be filled out?

The completion of the “Annotation Dialog” is introduced by the example of annotating the gene dnaA:

The screenshot shows the 'Annotation Dialog' window. The left sidebar has two main sections: 'Functional' and 'Regional'. The 'Functional' section lists several annotations, with the most recent one, '12:47:33: daniela', highlighted in bold. The 'Regional' section also lists annotations, with '3.3.2004 13:17:30: Critica' highlighted. The main area is titled 'Annotation Detail:' and contains the following fields:

- Set regional status to:** annotated_2
- Set functional status to:** annotated
- Date:** 27.4.2004 12:47:33
- Genename:** gyrA
- EC-Number:** 5.99.1.3
- Gene Product:** DNA gyrase subunit A
- Description:** DNA gyrase subunit A (EC 5.99.1.3). DNA gyrase negatively supercoils closed circular double-stranded DNA in an ATP-dependent manner and also catalyzes the interconversion of other topological isomers of double-stranded DNA rings including catenanes and knotted rings.
- Function:** DNA gyrase (topoisom)
- Confidence:** High confidence in function and specificity
- Comment:** Annotation derived from meta auto annotator. Derived by combining results from PSI-Blast SPROT, Blast2p vs KEGG, InterPro, PSI-Blast COG, TIGRFAM.

At the bottom, there are two tables: 'Observations used for this annotation' and 'Additional observations'. The 'Observations used for this annotation' table lists several observations with their descriptions, tools, and databases. The 'Additional observations' table lists observations that are not selected for the current annotation.

Figure 2.3: The “Annotation Dialog”.

History The history is a list of already created annotations (either automatic or manually from other users) regarding function and/region of the gene in question. The latest generated annotation is highlighted in bold (in the beginning of the whole annotation

there is usually only the automatic annotation shown - Metanor).

Navigation (“>>” or “<<” arrows in the upper area of the “Annotation Dialog” window) the arrows permit to jump to the next or latter region, respectively.

Set regional status to after checking the region, one of the following stati should be assigned to this region:

- ***attention needed:*** the status of this region is still not confident and needs to be revised.
- ***ignored:*** ignore this region.
- ***putative:*** additional status. Do not use this one.
- ***annotated_1, annotated_2 or annotated_3:*** the function of these stati can be defined independently in each project.
- ***finished:*** this region is assured and does not need to be reprocessed.

Set functional status to after checking the function, one of the following stati should be assigned to this region:

- ***putative:*** initial status. There has been neither automatic nor manual annotation.
- ***automatically annotated:*** this region has been automatically annotated by Metanor - in the beginning of the manual annotation, most of the regions have this function status.
- ***ignored:*** ignore this region.
- ***annotated:*** the function of this region has been assigned by an annotator. The annotation of the region is concluded.
- ***attention needed:*** the functional status of this region is still not confident and needs to be revised.
- ***user state 1, user state 2:*** the function of these stati can be defined independently in each project.

“ADD NEW” and “ACCEPT” buttons: both buttons conclude an annotation process. Though the conclusion is done in different ways **IMPORTANT! Note the different function of both buttons!!**

“ADD NEW” A new annotation is generated. Regional and functional status are adopted from the corresponding fields.

“ACCEPT” A new annotation is generated. The user points out to the system that he/she inputs in all fields. The consequence is that the regional status is set to “finished” and the functional status to “annotated”.

Gene name In this field the user can enter a gene name. *The gene name should only be entered if it is concerned to the gene in question, with the corresponding function and specificity.* A requirement for placing the gene name is that it should be based on several observations, establish a relationship with an operon (if it is the case), and further criteria. The used criteria should be entered in the field “Descriptions” (see below). The denomination of a gene occurs with 4 letters. In case of paralogous genes, the name should be added by a number. The counting starts at the dnaA gene (CDS1). The next paralogous gene will have a “1” added to its name, and so on. In some projects the denomination of genes can be based on the corresponding *E. coli* homologous - if there is an absence of the gene in question in this organism, the next related one should be taken. **No one should conceive a gene name him/herself!!**

GO numbers/Go info Gene Ontology. By clicking on the “GO number” button in the “GO Info” window the user gets further information about that number.

EC number Enzyme Commission Number. The EC-number is usually assigned by the auto-annotator, based on Psi-Blast against KEGG results, although sometimes multiple good hits can show up with different numbers. The decision on which number should be taken for the annotation is up to the annotator. The annotator should try to validate the EC-number based on the KEGG metabolic pathway. *EC numbers only exist for enzymes!!*

Gene product In this field the user should enter a detailed (but not too many details!) description of the gene product. Many predefined descriptions can be accessed via the pull-down menu “Choose predefined gene product here:”. If possible, a gene product should be always described based on specific evidences (for instance, *conserved hypothetical protein* with a GTP-binding domain predicted by HMMs = *putative GTP-binding protein*). In case of hybrid/ or multifunctional proteins, each matching functional description should be entered.

In the pull-down menu “Choose predefined gene product here:” there are the following standard terms predefined:

- ***hypothetical protein predicted by...***: standard denomination of a gene product that does not have significant homologies (level 1-3 over more than 60% of the amino acid sequence) to known gene products.

- ***conserved hypothetical protein:*** standard denomination of a gene product that presents significant homologies to at least one hypothetical or conserved hypothetical protein from another genome.
- ***putative secreted protein:*** standard denomination of a gene product that has a signal peptide predicted by the SignalP tool and that could not be characterized by any further function (e.g., a *conserved hypothetical protein* with a signal peptide).
- ***putative membrane protein:*** standard denomination of a gene product that has TM-domains predicted by the TMHMM tool. Here is caution demanded, as sometimes secretion signal peptides are predicted as TM-domain, which is often unfounded.
- ***no similarity:*** please use hypothetical protein
- ***none:*** please use hypothetical protein

Further possible denominations for gene products that do not belong to the predefined standard terminology:

- ***putative [name]:*** denomination for a gene product that has weak evidence pointing to a specific function.
- ***[Name], probable:*** denomination for a gene product that presents weak, but significant homologies to a group of proteins with specific function.
- ***[Name]-family probable:*** denomination for a gene product that belongs to a family with Pfam/InterPro-HMM hits (< e- ??) and/or present more significant homologies (>30% identity over at least 80% of the amino acid sequence) or even conserved domains (preferentially in respect to SWISSPROT).
- ***[Name]:*** denomination for a gene product that presents a high level of confidentiality in its function and experimental specificity. The gene product should be clearly classified via HMMs and/or homologies ((preferentially in respect to SWISSPROT). The terms may also be combined (e.g., *XYZ-family outer membrane lipoprotein* or *outer membrane lipoprotein-sorting protein Lola*)

Descriptions The descriptions are made up of several entry fields and, by submitting the sequence, they are transferred to a public database (e.g., EMBL database).

- **Text field:** this completion depends on the project and is established in the beginning of the project.
- **Experimental:** this item should be activated by presence of experimental evidences and is exported to the “evidence” field in the EMBL database.

- **COG:** (Cluster of Orthologous Groups of Proteins) functional classification of the concerning CDS. The COG number is usually supplied by the automatic annotator and should only be changed in exceptional cases. COGs can also be delivered via the Observations: tool Psi-Blast against COG; via the “Get” button all fields are automatically completed, i.e., the COG tree is opened for the manual completion. COGs are based on homology, so that several COGs or no COGs can exist for a specific function. The COGs should not be randomly chosen for the annotation, the most strong homologous cluster should be annotated in this field (see Observations). Additionally, several COGs appear in different classes of function. If this is correct, then do not change it, because via “Get” only one class of function can be imported. It may happen that wrong or meaningless COGs appear here, do not use them. If there is no Psi-Blast against COG result of level 1 or 2, then the annotation gets the number: COG0000 (No funcat 0).
- **Function:** in general, the function corresponds to the COG term. If no appropriate COG exists, other specifications can be entered, deduced from GO, TIGR roles, or Monica Riley categories, etc. This entry will be exported to the EMBL database, and should be filled out whenever it is possible. Which of those existing schemes are going to be utilized is determined in the beginning of the annotation, and has to be consistently established for all annotators.
- **Confidence:** the confidence describes the reliability of the annotation. There are also predefined descriptions:
 - *hypothetical protein:* same as in “Gene Product”.
 - *conserved hypothetical protein:* same as in “Gene Product”.
 - *specificity unclear:* this description should only be utilized if absolute necessary (e.g., for transport systems which substrate is unknown). Otherwise please select “Family Membership”.
 - *function unclear:* this description should only be utilized if absolute necessary. Otherwise please select “Family Membership”.
 - *family membership:* this description is utilized if there has been a family membership prediction (or some specific function), for instance through homology to other members of the family or through InterPro/Pfam - but no unique specificity (function) can be assigned.
 - *high confidence in function and specificity:* this description is utilized if a specific function and specificity can be assigned with certainty; for instance if a given enzyme from the intermediary metabolism can be detected either by homology searches, or HMMs and Psi-Blast against COG and KEGG, or even if there is experimental evidence. Criteria for the selection of *high confidence in function and specificity:* multiple full-length hits with more than

30% amino acid identity + at least one match to an experimentally characterized protein (+ HMM matches [not all proteins have HMM matches]) + all conserved active centers and binding sites, etc. If this category is placed, corresponding publication or other references should be checked for experimental or other well-founded evidences rather than only an unconfirmed automatic annotation. GENES OF THIS CATEGORY SHOULD HAVE A NAME ASSIGNED.

- **Comment:** the content of this field is **not** being exported to public databases with the sequence submission. This means that project-oriented comments can be inserted in this field, such as hints for other annotators, etc. Before the manual annotation begins, this field usually contains a comment from the auto-annotator Metanor.
- **Observations:** all observation references that have been used for the annotation can be edited in this window, if specific observations do not support the annotation. New observations can also be added as “supporting evidence”.

Hint: by filling out the fields please be sure they do not contain redundant information, like inserting twice the description of the gene product (e.g. also in “Descriptions”).

Hints for annotators: sometimes it helps to have a look at the neighboring regions of the gene to be annotated. If this gene is part of an operon, further indications for function and specificity can be deduced.

Criteria for an operon:

- only small intergenic regions
- flanked by regulator/terminator sequence
- function in the same context

Example: ABC-transporter gene cluster = periplasmatic substrate binding protein, membrane permease and ATP binding protein.

The conclusion of the annotation is done by clicking the “ADD-NEW” and “ACCEPT” buttons.

After closing the session, the user should always log out (“Logout” button in the upper menu bar from the “Main Window”). Otherwise, the user will be automatically logged out after 30 minutes.

Gene Ontology (GO)

GO numbers (for more information see www.geneontology.org) allow a hierarchical classification of genes relative to molecular function (e.g., GO3677 DNA binding, GO3688

DNA replication origin binding, GO5524 ATP binding), biological process (e.g., GO6270 DNA replication initiation, GO6275 regulation of DNA replication) and cellular component. GO numbers are found in the descriptions of Pfam and InterPro (Observations: tool *InterPro* -> link in column DB). GO numbers can also be searched for directly in this page. Assigned GO numbers should be checked through the link next to the field - it also features the corresponding “GO - Trees”.

3 Glossary

Accession number An Accession number is a unique identifier given to a sequence when it is submitted to one of the DNA repositories (GenBank, EMBL, DDBJ). The initial deposition of a sequence record is referred to as version 1. If the sequence is updated, the version number is incremented, but the Accession number will remain constant.

Blast Basic Local Alignment Search Tool (Altschul et al., J Mol Biol 215:403-410; 1990). A sequence comparison algorithm that is optimized for speed and used to search sequence databases for optimal local alignments to a query. See the BLAST chapter (Chapter 15) or the tutorial or the narrative guide to BLAST.

blastn Nucleotide X nucleotide BLAST. blastn takes nucleotide sequences in FASTA format, GenBank Accession numbers, or GI numbers and compares them against the NCBI Nucleotide databases.

blastp proteinprotein BLAST. blastp takes protein sequences in FASTA format, GenBank Accession numbers, or GI numbers and compares them against the NCBI Protein databases.

BLOSUM 62 Blocks Substitution Matrix. A substitution matrix in which scores for each position are derived from observations of the frequencies of substitutions in blocks of local alignments in related proteins. Each matrix is tailored to a particular evolutionary distance. In the BLOSUM 62 matrix, for example, the alignment from which scores were derived was created using sequences sharing no more than 62

Boolean This term refers to binary algebra that uses the logical operators AND, OR, XOR, and NOT; the outcomes consist of logical values (either TRUE or FALSE). The keyword boolean indicates that the expression or constant expression associated with the identifier takes the value TRUE or FALSE. The logical-AND (&&) operator produces the value 1 if both operands have nonzero values; otherwise, it produces the value 0. The logical-OR (??) operator produces the value 1 if either of its operands has a nonzero value. The logical-NOT (!) operator produces the value 0 if its operand is true (nonzero) and the value 1 if its operand is FALSE (0). The exclusive OR (XOR) operator yields TRUE only if one of its operands are TRUE and the other is FALSE. If both operands are the same (either TRUE or FALSE), the operation yields FALSE.

CDS coding region, coding sequence. CDS refers to the portion of a genomic DNA sequence that is translated, from the start codon to the stop codon, inclusively, if complete. A partial CDS lacks part of the complete CDS (it may lack either or both the start and stop codons). Successful translation of a CDS results in the synthesis of a protein.

CGI Common Gateway Interface. A mechanism that allows a Web server to run a program or script on the server and send the output to a Web browser.

cluster A group that is created based on certain criteria. For example, a gene cluster may include a set of genes whose similar expression profiles are found to be similar according to certain criteria, or a cluster may refer to a group of clones that are related to each other by homology.

COGs Clusters of Orthologous Groups (of proteins) were delineated by comparing protein sequences from completely sequenced genomes. Each COG consists of individual proteins or groups of paralogs from at least three lineages and thus corresponds to an ancient conserved domain.

consensus sequence The nucleotides or amino acids found most commonly at each position in the sequences of homologous DNAs, RNAs, or proteins.

contig A contiguous segment of the genome made by joining overlapping clones or sequences. A clone contig consists of a group of cloned (copied) pieces of DNA representing overlapping regions of a particular chromosome. A sequence contig is an extended sequence created by merging primary sequences that overlap. A contig map shows the regions of a chromosome where contiguous DNA segments overlap. Contig maps provide the ability to study a complete and often large segment of the genome by examining a series of overlapping clones, which then provide an unbroken succession of information about that region.

definition line A sequence in FASTA format begins with a single-line description, followed by lines of sequence data. The definition line or description line is distinguished from the sequence data by a "greater than" (>) symbol in the first column (see example); also DEFLINE, as in a flatfile.

Domain A "domain" refers to a discrete portion of a protein assumed to fold independently of the rest of the protein and which possesses its own function.

draft sequence Draft sequence refers to DNA sequence that is not yet finished but is generally of high quality (i.e., an accuracy of greater than 90%). Draft sequence data are mostly in the form of 10,000 base pair-sized fragments, the approximate chromosomal locations of which are known. The following keywords are associated with draft sequence: phase 0, light-pass coverage of a clone, generally only 1 coverage; phase 1, 410 coverage of a BAC clone (order and orientation of the fragments are unknown); and phase 2, 410 coverage of a BAC clone (order and orientation of the fragments are known). Phase 3 refers to the completely finished sequence.

E-value Expect value. The E-value is a parameter that describes the number of hits one can "expect" to see by chance when searching a database of a particular size. It decreases exponentially with the score (S) that is assigned to a match between two sequences. Essentially, the E-value describes the random background noise that exists for matches between sequences. For example, an E-value of 1 assigned to a hit can be interpreted as meaning that in a database of the current size, one might expect to see one match with a similar score simply by chance. This means that the lower the E-value, or the closer it is to "0", the higher is the "significance" of the match. However, it is important to note that searches with short sequences can be virtually identical and have relatively high E-value. This is because the calculation of the E-value also takes into account the length of the query sequence. This is because shorter sequences have a high probability of occurring in the database purely by chance.

EC number A number assigned to a type of enzyme according to a scheme of standardized enzyme nomenclature developed by the Enzyme Commission of the Nomenclature Committee of the International Union of Biochemistry and Molecular Biology (IUBMB). EC numbers may be found in ENZYME, the Enzyme nomenclature database, maintained at the ExPASy molecular biology server.

EMBL European Molecular Biology Laboratory

Entrez Entrez is a retrieval system for searching several linked databases. It provides access to the following NCBI databases: PubMed, GenBank, Protein, Structure, Genome, PopSet, OMIM, Taxonomy, Books, ProbeSet, 3D Domains, UniSTS, SNP, and CDD. (See the Entrez chapter or the Entrez web page.)

EST Expressed Sequence Tag. ESTs are short (usually approximately 300-500 base pairs), single-pass sequence reads from cDNA. Typically, they are produced in large batches. They represent the genes expressed in a given tissue and/or at a given developmental stage. They are tags (some coding, others not) of expression for a given cDNA library. They are useful in identifying full-length genes and in mapping.

ExPASy Expert Protein Analysis System is a proteomics server of the Swiss Bioinformatics Institute (SIB).

FASTA The first widely used algorithm for similarity searching of protein and DNA sequence databases. The program looks for optimal local alignments by scanning the sequence for small matches called "words". Initially, the scores of segments in which there are multiple word hits are calculated ("init1"). Later, the scores of several segments may be summed to generate an "initn" score. An optimized alignment that includes gaps is shown in the output as "opt". The sensitivity and speed of the search are inversely related and controlled by the "k-tup" variable, which specifies the size of a "word" (Pearson and Lipman). Also refers to a format for a nucleic acid or protein sequence.

finished sequence High-quality, low-error DNA sequence that is free of gaps. To qualify as a finished sequence, only a single error out of every 10,000 bases (i.e., an accuracy of 99.999%) is allowed.

gap A gap is a space introduced into an alignment to compensate for insertions and deletions in one sequence relative to another. To prevent the accumulation of too many gaps in an alignment, introduction of a gap causes the deduction of a fixed amount (the gap score) from the alignment score. Extension of the gap to encompass additional nucleotides or amino acid is also penalized in the scoring of an alignment. (See the figure for more information.)

GenBank GenBank is a database of nucleotide sequences from more than 100,000 organisms. Records that are annotated with coding region features also include amino acid translations. GenBank belongs to an international collaboration of sequence databases that also includes EMBL and DDBJ. [See the GenBank chapter (Chapter 1) or the GenBank web page.]

homologous The term refers to similarity attributable to descent from a common ancestor. Homologous chromosomes are members of a pair of essentially identical chromosomes, each derived from one parent. They have the same or allelic genes with genetic loci arranged in the same order. Homologous chromosomes synapse during meiosis.

HTML Hypertext Markup Language. HTML is derived from SGML. It is a text-based mark-up language and is used to primarily display information using a web browser and to link pieces of information via hyperlinks. The tags used in an HTML document provide information only on how the content is to be displayed but do not provide information about the content they encompass.

locus In a genomic context, locus refers to position on a chromosome. It may, therefore, refer to a marker, a gene, or any other landmark that can be described.

LocusID Each new LocusLink record is assigned a unique identifying number—a LocusID (although coding regions on genomic sequences found by gene prediction software are an exception to this).

LocusLink LocusLink provides a single query interface to curated sequence and descriptive information about genetic loci. LocusLink issues a stable ID (LocusID) for each locus and presents information on official nomenclature, aliases, sequence Accession numbers, phenotypes, EC numbers, OMIM numbers, UniGene clusters, map information, and relevant web sites. LocusLink is a collaborative effort among NCBI, Human Gene Nomenclature Committee, OMIM, and others. LocusLink currently contains human, mouse, rat, zebrafish, and fruit fly loci; organisms can be searched together or separately.

MegaBLAST MegaBLAST is a program for aligning sequences that differ slightly as a result of sequencing or other similar "errors". When larger word size is used, it is up to 10 times faster than more common sequence-similarity programs. MegaBLAST is also able to efficiently handle much longer DNA sequences than the blastn program of the traditional BLAST algorithm. It uses the GREEDY algorithm for a nucleotide sequence alignment search.

mFAST A Multi-FASTA format.

minimal tiling path An ordered list or map that defines the minimal set of overlapping clones needed to provide complete coverage of a chromosome or other extended segment of DNA (compare with tiling path).

MMDB Molecular Modeling Database. MMDB is a database of three-dimensional biomolecular structures derived from X-ray crystallography and nuclear magnetic resonance (NMR) spectroscopy.

MMDB-ID Molecular Modeling Database Accession number.

NCBI National Center for Biotechnology Information

NMR Nuclear Magnetic Resonance. NMR is a spectroscopic technique used for the determination of protein structure.

nr-PDB non-redundant Protein Data Bank

ortholog Orthology describes genes in different species that derive from a single ancestral gene in the last common ancestor of the respective species.

orthology Orthology describes genes in different species that derive from a common ancestor, i.e., they are direct evolutionary counterparts.

paralog A paralog is one of a set of homologous genes that have diverged from each other as a consequence of gene duplication.

paralogy Paralogy describes the relationship of homologous genes that arose by gene duplication.

PDB Protein Data Bank. The PDB is a database for 3D macromolecular structure data.

Pfam Pfam is a database housing a large collection of multiple sequence alignments and hidden Markov models covering many common protein domains.

PHRAP A computer program that assembles raw sequence into sequence contigs (see above) and assigns to each position in the sequence an associated "quality score", on the basis of the PHRED scores of the raw sequence reads. A PHRAP quality score of X corresponds to an error probability of approximately $10^{-X}/10$. Thus, a PHRAP quality score of 30 corresponds to 99.9% accuracy for a base in the assembled sequence.

PHRED A computer program that analyses raw sequence to produce a "base call" with an associated "quality score" for each position in the sequence. A PHRED quality score of X corresponds to an error probability of approximately $10^{-X}/10$. Thus, a PHRED quality score of 30 corresponds to 99.9% accuracy for the base call in the raw read. phyletic pattern Pattern of presenceabsence of a cluster of orthologs (COG) in different species.

PHYLIP PHYLogeny Inference Package A package of programs for various computer platforms to infer phylogenies or evolutionary trees, freely available from the Web.

PIR Protein Information Resource

PNG Portable Network Graphics. An extensible file format for the lossless, well-compressed storage of raster images (images that are composed of horizontal lines of pixels, such as those created by a computer screen). Compression of image, media, and application files is necessary to reduce the transmission time across the web. The technique of lossless compression reduces the size of the file without sacrificing any original data, and the image after expansion is exactly as it was before compression. PNG overcomes the patent issues of GIF (Graphic Interchange Format) and can replace many common uses of TIFF (Tagged Image File Format). Several features such as indexed color, grayscale, and truecolor are supported, as well as an optional alpha-channel. PNG is designed to work well in online viewing applications and is supported as an image standard by the WWW.

PRF Protein Research Foundation

PROW Protein Reviews On the Web. An online resource that features PROW Guides authoritative, short, structured reviews on proteins and protein families. The Guides provide approximately 20 standardized categories of information (abstract, biochemical function, ligands, references, etc.) for each protein.

pseudogene A sequence of DNA that is very similar to a normal gene but that has been altered slightly so that it is not expressed. Such genes were probably once functional but, over time, acquired one or more mutations that rendered them incapable of producing a protein product.

PSI-BLAST Position-Specific Iterated BLAST. PSI-BLAST (Altschul et al., J Mol Biol 215:403-410; 1990) is used for iterative protein sequence similarity searches using a position-specific score matrix (PSSM). It is a program for searching protein databases using protein queries to find other members of the same protein family. All statistically significant alignments found by BLAST are combined into a multiple alignment, from which a PSSM is constructed. This matrix is used to search the database for additional significant alignments, and the process may be iterated until no new alignments are found.

PSSM Position-Specific Score Matrix. The PSSM gives the log-odds score for finding a particular matching amino acid in a target sequence.

PubMed A retrieval system containing citations, abstracts, and indexing terms for journal articles in the biomedical sciences. It includes literature citations supplied directly to NCBI by publishers as well as URLs to full text articles on the publishers' web sites. PubMed contains the complete contents of the MEDLINE and PREMEDLINE

databases. It also contains some articles and journals considered out of scope for MEDLINE, based on either content or on a period of time when the journal was not indexed and, therefore, is a superset of MEDLINE.

Reciprocal best hits Reciprocal best hits are proteins from different organisms that are each other's top BLAST hit, when the proteomes from those organisms are compared to each other. For example, proteins AZ in organism 1 are compared against proteins AAZZ in organism 2. If protein A has a best hit to protein RR, and RR's best hit, when it is compared to all the proteins in organism 1, also turns out to protein A, then A and RR are reciprocal best hits. However, if RR's best hit is to B rather than to A, then A and RR are not reciprocal best hits.

RefSeq RefSeq is the NCBI database of reference sequences; a curated, non-redundant set including genomic DNA contigs, mRNAs and proteins for known genes, and entire chromosomes.

RepeatMasker Program that screens DNA sequences for interspersed repeats and low-complexity DNA sequences.

RPS-BLAST Reverse Position-Specific BLAST. A program used to identify conserved domains in a protein query sequence. It does this by comparing a query protein sequence to position-specific score matrices (PSSMs) that have been prepared from conserved domain alignments. RPS-BLAST is a "reverse" version of position-specific iterated BLAST (PSI-BLAST); however, RPS-BLAST compares a query sequence against a database of profiles prepared from ready-made alignments, whereas PSI-BLAST builds alignments starting from a single protein sequence.

SMART Simple Modular Architecture Research Tool. A tool to allow automatic identification and annotation of domains in user-supplied protein sequences. For example, the SWISS-PROT database is an extensively annotated and nonredundant collection of protein sequences. SWISS-PROT annotations have been mined for SMART-derived annotations of alignments.

SSAHA Sequence Search and Alignment by Hashing Algorithm. SSAHA is a software tool for very fast matching and alignment of DNA sequences and is used for searching databases containing large amounts (gigabases) of genome sequence. It achieves its fast search speed by converting sequence information into a "hash table" data structure, which can then be searched very rapidly for matches (Ning et al., *Genome Res* 11:1725-1729; 2001).

substitution matrix A substitution matrix containing values proportional to the probability that amino acid *i* mutates into amino acid *j* for all pairs of amino acids. Such matrices are constructed by assembling a large and diverse sample of verified pairwise alignments of amino acids. If the sample is large enough to be statistically significant, the resulting matrices should reflect the true probabilities of mutations occurring through a period of evolution. (See also BLOSUM 62.)

SWISS-PROT SWISS-PROT is a curated protein sequence database that provides a high level of annotation (such as the description of protein function, domain structures, post-translational modifications, variants, etc.), a minimal level of redundancy, and high level of integration with other databases.

synteny On the same strand. The phrase "conserved synteny" refers to conserved gene order on chromosomes of different, related species.

Tax BLAST BLAST Taxonomy Reports page. Tax BLAST groups BLAST hits by source organism, according to information in NCBI's Taxonomy database. Species are listed in order of sequence similarity with the query sequence, the strongest match listed first.

taxID Taxonomy Identifier. The taxID is a stable unique identifier for each taxon (for a species, a family, an order, or any other group in the taxonomy database). The taxID is seen in the GenBank records as a "source" feature table entry; for example, /db_xref="taxon:<9606>" is the taxID for *Homo sapiens*, and the line is therefore found in all recent human sequence records.

TIGR The Institute for Genomic Research

tiling path An ordered list or map that defines a set of overlapping clones that covers a chromosome or other extended segment of DNA.

UNIX UNIX is an operating system that was developed by Dennis Ritchie and Kenneth Thompson at Bell Labs more than 30 years ago. It allows multitasking and multiuser capabilities and offers portability with other operating systems. It comes with hundreds of programs that are of two types: integral utilities, such as the command line interpreter; and tools such as email, which are not necessary for the operation of UNIX but provide additional capabilities to the user. It is functionally organized at three levels: the kernel, which schedules tasks and manages storage; the shell, which connects and interprets user's commands, calls programs from memory, and executes them; and tools and applications, which offer additional functionality to the operating system, such as word processing and business applications. UNIX was registered by Bell Laboratories as

a trademark for computer operating systems. Today, this mark is owned by The Open Group.

URL Uniform Resource Locator. The address of a resource on the Internet. URL syntax is in the form of protocol://host/localinfo, where "protocol" specifies the means of fetching the object (such as HTTP, used by WWW browsers and servers to exchange information, or FTP), "host" specifies the remote location where the object resides, and "localinfo" is a string (often a file name) passed to the protocol handler at the remote location. Also called Uniform Resource Identifier (URI).

weight An assignment of importance to a term in a search query. If a term in a search query is found to match a word in a document, that word is given a "weight". The exact weight of the word will depend on the emphasis given to the word by the author or its position in the document. For example, a word that occurs in a chapter title will have a higher weight than the same word if it occurs in the body of the chapter. Similarly, words that occur in data collections are also assigned weights, depending on how frequently the terms occur in the collection.

WGS sequence Whole Genome Shotgun sequence. In this semi-automated sequencing technique, high-molecular-weight DNA is sheared into random fragments, size selected (usually 2, 10, 50, and 150 kb), and cloned into an appropriate vector. The clones are then sequenced from both ends. The two ends of the same clone are referred to as mate pairs. The distance between two mate pairs can be inferred if the library size is known and has a narrow window of deviation. The sequences are aligned using sequence assembly software. Proponents of this approach argue that it is possible to sequence the whole genome at once using large arrays of sequencers, which makes the whole process much more efficient than the traditional approaches.