Glossary of Commonly used Annotation Terms

Akela – a general use server for the annotation group as well as other groups throughout TIGR.

Annotation Notebook – a link from the gene list page that is associated with all TIGR role ids. The notebook is used by the annotator to note genes of interest for the Principle Investigator in charge of the project.

asm_feature – a table in the microbial database that stores "features" associated with a particular assembly.

asmbl_id – the numerical identification associated with an assembly in the database.

assembler – the program that "puts together" the raw sequence that is generated in the lab.

assembly – sequence that has been assembled together and loaded into the database.

ATV – The java program that enables the tree viewer for the gene curation page.

autoannotate – the program that assigns annotation to newly defined open reading frames before they are manually curated.

BA – bioinformatics analyst

BAMBOO – this program orients and scaffolds assemblies together based on their paired reads. It then creates a graphical display showing the information.

BCP – bulk curation page

BE – bioinformatics engineer

belvu – a viewer for multiple sequence alignments.

BER – BLAST-extend-repraze

BER2multi – The program that makes a multiple protein alignment from the matching accessions in the BER.

BITS – The incident ticket tracking system that is used by TIGR. Bugs, analysis requests, etc., are submitted to this system.

BRC – Bioinformatics Resource Center

Btab – an acronym for blast matches formatted in a tabulated file.

Btab_skim – The name for the table at the bottom of the gene curation page. Also known as BER skim.

CHADO – the database schema modules developed by the GMOD schema project for building a model organism relational database.

Characterized match – a protein from the BER search results that is experimentally characterized and a good match to our query protein.

Characterized table – the table in the database that stores the accessions of the characterized proteins we have collected.

Closure – the process during the sequencing phase of a genome project in which the gaps in the assembled genomic sequence are "closed".

CMR – Comprehensive Microbial Resource.

Codon – the 3 base pair nucleotide sequence that codes for an amino acid.

Codon usage – The term describing the tendency of organisms to use certain codons and not others for a specific amino acid.

COG – clusters of orthologous groups.

Command line – lingo normally used when running a program from the prompt in a terminal window.

Common – a table in the database that stores "common" information concerning genome projects, including GO ontologies, and Genome Property information

Condor – the system that runs large search jobs by allocating the jobs to multiple computers.

CHP – conserved hypothetical protein

Consistency checks – the "checks" done to a genome project after annotation has been completed. This is done to make sure that the genome is ready for publication.

Contig – the term given to a piece of assembled raw sequence.

Coords – the term describing the coordinate boundaries of an open reading frame.

CVS – a version control system used to update programs. All versions of a certain program are stored here.

Database - a systematically arranged collection of computer data, structured so that it can be automatically retrieved or manipulated.

Database field – an attribute of what the database table describes; a column of the table.

Database table – a unit of data storage in a database, it has a name and consists of one two or more defined fields which are populated with items of data, generally a particular table will store data relevant to a particular topic.

Degenerate – an open reading frame with multiple point mutations, frameshifts, deletions, etc.

Domain – the term describing homology to only part of a protein, and not the whole protein.

EC – Enzyme Commission

egad – a database no longer in use at TIGR, it used to house our collection of all proteins in existence, it has been replaced by PANDA, however, there still reside in egad several tables needed for annotation: the HMM tables and the roles table in particular.

emacs – a commonly used text editor.

End3 – a term used to describe the 3' end coordinate of an open reading frame.

End5 – a term used to describe the 5' end coordinate of an open reading frame.

Equivalog – the term used to describe a hidden markov model that describes the exact function of a protein.

Feat_name – the field in the database that holds the "feature name".

Feat_type – the field in the database that describes the type of "feature".

flaps – a general use server for the annotation group as well as other groups throughout TIGR.

frog – a general use server for the annotation group as well as other groups throughout TIGR.

FS – frameshift

Funny characters – a term to describe characters that do not belong in the DNA or protein sequences, generally these are characters indicating ambiguity in the nucleotide sequence, they should all be resolved prior to the start of annotation.

GC skew – the tool that measures the 3rd base pair wobble for a particular reading frame. (I think this def. referes to "third posisition GC skew" and the plain GC skew is something different - but I'm not completely sure.....)

GCP – Gene Curation Page

GENBANK accession builder – The tool used by an annotator to create files to be submitted to GenBank for a particular genome.

Gene cluster – A group of genes adjacent to one another, and often, but not always, oriented in the same direction, that have a similar function (e.g. transport) or are part of a biochemical pathway.

Gene Symbol – The abbreviation given to a particular gene. It is stored in the ident table.

Genome Properties - The Genome Properties system consists of a suite of "Properties" which are carefully defined attributes of prokaryotic organisms whose status can be described by numerical values or controlled vocabulary terms for individual completely sequenced genomes. Evaluation of these properties may take place via manual curation or by computer algorithms.

Genome Viewer – The graphical interface used by an annotator that allows the user to view the genomic organization in a selected genome.

getdb – the command that allows the user to search for information about a database by feeding the program either the database abbreviation or the organism name.

getph – the command that allows the user to search for a person at TIGR by either their phone number or Name.

getquota – the command that allows the user to check the quota and used space in a particular directory.

GIP – genome initiation protocol

glimmer – The program used at TIGR to find open reading frames in whole genomic sequence.

GO – gene ontology

Gold standard annotation – the top standard of annotation done by our annotation team at TIGR.

Helpdesk – the system used at TIGR to report bugs, problems, etc.

HMM – hidden Markov model

Homolog - A gene similar in structure and evolutionary origin to a gene in another species.

Hypotheticals with praze – hypothetical proteins that have results in the BER table.

ident – a table in the database that stores "identity information" for an ORF.

Interevidence region analysis – a program run by TIGR annotators to find any genes missed by glimmer that are in intergenic regions.

Intergenic region – the region between two genes that does not have a coding ORF associated with it.

InterPro – a searchable database of protein families, domains, and functional sites that is maintained by the European Bioinformatics Institute.

Isology – sequence similarity of aligned nucleic acids or amino acids; the similarity may be due to homology or convergence.

Library – one of the stages of the genome sequencing process where a genomic DNA sample from the organism is fragmented and cloned into a plasmid or phage library for sequencing.

Locus – the final public identifier for a gene, has associated location and annotation

Main Role – the highest level of a TIGR role category, each TIGR main role has several sub roles

MANATEE – the web-based annotation tool used by all TIGR annotators.

MGAT – multi-genome annotation tool; it is a web-based annotation tool that allows the annotator to annotate related genes in multiple genomes simultaneously.

Multiple alignment – an alignment of multiple DNA or protein sequences.

Mummer – The software package used at TIGR that aligns two selected genomes or chromosomes against one another.

niaa – non-identical amino acid database; the internal TIGR protein database. It consists of the all of the TIGR and public database protein information.

nraa – non-redundant amino acid database; this database was replaced by niaa.

Nucmer – one of the mummer packages. It compares genomic information based on nucleic acid sequence.

Oligomer skew – a way to find the origin of replication of a genome. It is based on the fact that oligomers of DNA and their compliments in a genome are sometimes symmetrical around one point. That point is the origin of replication.

Omniome – the term used for the CMR database. It contains all of the sequenced bacterial genomes.

Omnium – the name of the CMR database.

ORF – open reading frame

ORF management – the process that the annotators follow to resolve overlapping genes, find missing genes, and confirm start sites.

orf_attribute – one of the tables in the microbial database that stores "attributes" associated with each ORF.

ortholog – genes in two different species that evolved from a common ancestor. Orthologs retain function over the course of evolution.

Overlaps – the term used to describe genes that "overlap" in a genome.

Pairwise alignment – the alignment of two DNA or amino acid sequences to each other.

Pairwise match – analogous to pairwise alignment, indicates that a protein matches another based on a pairwise alignment.

PANDA – the internal TIGR database that contains all nucleic acid and protein information.

Paralog – proteins that are related by sequence similarity within a genome.

Paralogous family – a family of paralogs that is found in a genome. At TIGR we have a paralogous family program that generates these families.

Pathema – the database that holds all of the data associated with the Bioinformatics Resource Center. This is a NIAID funded program that creates a core resource for pathogens.

PFAM – Protein families database of alignments and HMMs.

PI – principal investigator or isoelectric point, depending on the context

PM – point mutation

promer – one of the mummer packages. It compares genomic sequence based on amino acid information.

PROSITE - PROSITE is a database of protein families and domains. It consists of biologically significant sites, patterns and profiles that help to reliably identify to which known protein family (if any) a new sequence belongs.

psearch – the system at TIGR that runs and monitors parallel jobs. Parallel jobs are a number of individual, but related, processes that are run on one or more of a system's nodes. A node in this case is a computer attached to TIGR's network.

Random – a stage in genome sequencing where the pieces of genomic DNA are being sequenced from the libraries. The "random" sequences will then be assembled together.

RBS – ribosome binding site

Role notes – the notes associated with a role category that annotators have put there to help others assign genes to the category when they are working on it.

Role id – the identification number assigned to TIGR role categories.

Scaffold – a term used to describe the orientation of genomic assemblies to one another. The program bamboo generates these scaffolds.

SGC – small genome control

SGC_logs – the log file generated by the small genome control process. It is created on a nightly basis for genome projects that are in annotation.

signalP – the program used at TIGR that searches for signal peptides in a protein sequence. Each protein is scored, and based on these scores the annotator can predict whether the protein is secreted or not. This tool is built by CBS, www.cbs.dtu.dk/index.shtml.

Small genome database – databases that contain the genomic information for bacterial genome sequencing projects.

SQL – structured query language; the programming language used to navigate and manipulate relational databases.

sqsh – the program used at TIGR to interact with our relational databases.

start site – the codon that corresponds to the beginning of the coding region of an open reading frame, the three start codons in bacteria are: ATG, GTG, and TTG

stop site – the codon that corresponds to the end of a gene or open reading frame, the three stop codons in bacteria are: TAA, TAG, TGA

Subfamily – the HMM isology that describes a type of TIGRFAM HMM. A subfamily HMM describes a specific family of proteins. This type of isology is more specific than the superfamily HMM isology.

subrole – the identification number that corresponds to roles that are subsets of a TIGR main role category.

Superfamily – a HMM isology that refers to a type of TIGRFAM HMM. A superfamily HMM describes a large family of related proteins. It is much more general that a subfamily isology HMM.

Swiss-prot – an external public database of DNA and proteins sequences that is maintained by the Swiss institute of Bioinformatics and the European institute of Bioinformatics.

SYBIL – the test database here at TIGR or the software package used to view comparative genomic data.

TI – a type of GO term that is created by a TIGR annotator when there is not a GO term available to describe the gene that is being annotated. This term will be submitted to GO, and given a GO term at a later date.

TIGRFAM – The library of HMMs that are created and maintained here at TIGR.

TMHMM – the program used at TIGR to identify transmembrane spanning regions of a protein. This tool is built by CBS, www.cbs.dtu.dk/index.shtml.

Turbo annotation – the type of annotation where the annotation team only annotates certain TIGR categories, so that a draft manuscript can be written. The genome is then finished when the manuscript has been submitted to a journal.

UFMG – the unfinished microbial genomes webpage. It contains information concerning all unfinished microbial sequencing projects ongoing at TIGR. It also contains a BLAST page where an outside user may BLAST a sequence against an unfinished genome project.

Workflow – a production pipeline that is used here at TIGR. It can be used to run processes, scripts, etc.

xmen – a general use server for the annotation group as well as other groups throughout TIGR.