Pathway Tools Omics Viewer

The Pathway Tools Omics Viewer uses the Metabolic Overview for an organism to illustrate the results of high-throughput experiments in a global metabolic pathway context. Genes (in the case of a gene expression experiment) and proteins (in the case of a proteomics experiment) that are involved in metabolism are mapped to reaction steps in the Metabolic Overview, and the range of data values levels in a given experimental dataset is mapped to a spectrum of colors. Reaction steps in the Metabolic Overview are colored according to the corresponding data value. Similarly, for metabolomics experiments, compound nodes are colored according to the data value for the corresponding compound. This facility enables the user to see instantly which pathways are active or inactive under some set of experimental conditions.

The Omics Viewer can be used for:

- **Microarray Expression Data:** Reaction lines (and protein icons, where present) are color-coded according to the relative or absolute expression level of the gene that codes for the enzyme that catalyzes that reaction step. The Omics Viewer allows a scientist to interpret the results of gene-expression experiments in a pathway context.
- **Proteomics Data:** Reaction lines (and protein icons, where present) are color-coded according to the concentration of the enzyme that catalyzes that reaction step.
- **Metabolomics Data:** Compound icons are color-coded according to the concentration of the compound.
- **Reaction Flux Data:** Reaction lines are color-coded according to reaction flux values.
- **Other Experimental Data:** Any experiment, high-throughput or otherwise, in which data values are assigned to genes, proteins, reactions or metabolites can be viewed in a pathway context using the Omics Viewer.

  The Omics Viewer can show absolute data values (such as the concentration of a metabolite or protein, or the absolute expression level of a gene), or it can be used to compare two sets of experimental data by computing a ratio and mapping the ratios onto a color spectrum.

  The superposition of multiple sets of experimental data on the metabolic overview can also be animated to show, for example, how gene expression levels of enzymes change with time over the course of an experiment.

  Omics Dataset File Format

  Experimental data is imported from a file provided by the user that is stored on the user's computer. Each line of the file contains data for a single gene, protein, reaction or metabolite, and is of the form:
  <name-or-ID> <data-column1>...<data-columnN>
  Columns are separated by the tab character. Lines that start with # or ; are taken to be comment lines and are ignored by the program.

<name-or-ID> can be either a common name for an object (the BioCyc data typically includes extensive synonym lists, and every attempt is made to match a name to the appropriate target), or the BioCyc internal ID for the object. Gene IDs from sequencing projects (such as the *E. coli* B-numbers) are generally acceptable and unambiguous. For protein or reaction data, EC numbers may be used. You must specify whether the entities in the <name-or-ID> column are genes, proteins, reactions, compounds, or a mixture.

The numbers in the data columns can represent either absolute or relative values. If the data values represent absolute numbers, you may choose to visualize either a single column of absolute data values (select "Absolute" and one data column), or the ratio of two data columns as relative data values (select "Relative" and two data columns). If the data values themselves represent relative numbers, then you need supply only a single column number, and select "Relative". An entry (a row of data for a gene or other object) may contain any number of data columns (for example, if you wish to compile measurements from several experiments or time points into a single file), but only those data columns specified will be visualized at a time -- all other columns will be ignored.

Color Scales

The color scale used depends on the type and, by default, the range of the data. Thus, a particular color may correspond to one gene expression level for one dataset, and a different gene expression level for another dataset, depending on the range of values or the supplied maximum cutoff value for each dataset. We use the spectrum from yellow/green to red, with yellow representing the lowest expression levels or ratios in the dataset, blue representing values in the middle, and red representing the highest values. Reactions for which no data was provided are drawn in black. The legend for mapping colors to data values is shown in the key, which is drawn to the right of the overview for a single experiment, or to the left for an animation.

A maximum cutoff value is chosen. By default, this is computed from the data. Alternatively, the user may supply a maximum cutoff value to use. Supplying the same maximum cutoff value for multiple experiments ensures that the same color scale is used for each one, so that the displays are directly comparable.

The minimum cutoff value is determined based on the maximum cutoff value and the other parameters. For absolute data values, we use a minimum cutoff value of zero. For relative data values that are not logs, we use the inverse of the maximum cutoff. For relative data values that are logs, we use the negative of the maximum cutoff. The color spectrum is then mapped evenly along a log scale between the maximum cutoff and the minimum cutoff.

In many cases, several genes or proteins, each with their own expression level or concentration, will map to a single reaction. This is because the reaction might be catalyzed by an enzyme complex made up of several gene products, or the reaction might be catalyzed by several isozymes, each with its own gene or genes. Since a reaction can only be colored a single color, we must choose which data value to use. For absolute data

values, we choose the maximum. For relative data values, we choose the value whose log has the greatest deviation from zero, under the assumption that the user is primarily interested in identifying the entities whose behavior differ most between the two datasets.

Omics Viewer Results

After you submit your dataset to the Pathway Tools, the Omics Viewer returns several results:

1. The Overview Diagram, colorized with experimental data.
2. The color key for the Overview.
3. For single experiments, some basic statistics computed from the data file. The program counts and lists gene/protein/metabolite names that could not be resolved, or for which data was missing or malformed. Since, for example, not all genes will code for enzymes, and therefore not all will correspond to reactions in the Metabolic Overview, we compile separate statistics for only those that are represented in the Overview and for the dataset as a whole. The statistics that we compute and tabulate are: number of values, minimum, maximum and median values, and mean and standard deviation of the natural logs of the values. These statistics are not computed when generating animations
4. A histogram that shows the distribution of values in the dataset. This histogram is displayed directly beneath the color key. The data value range is divided into 50 intervals, using the same criteria that we use for assigning colors. The number of data values in each interval is shown on the histogram, colored appropriately. To the left of the vertical axis is the histogram for the entities that are represented in the overview. To the right of the axis is the histogram for all other entities in the dataset.

2. GENOSTAR

Genostar is a bioinformatics solution provider based in Grenoble, France. The company was founded in 2004 following the "Genostar consortium" that was created in 1999 as a public-private consortium by Genome Express, Hybrigenics,INRIA (Institut National de Recherche en Informatique et Automatique / French National Institute for Research in Computer Science and Control) and The Pasteur Institute.

Metabolic Pathway Builder is a bioinformatics environment dedicated to microbial research. This streamlined bioinformatics solution covers sequence assembly, mapping, annotation transfer and identification of protein domains, comparative genomics, structural searches, metabolic pathway analysis, modeling and simulation of biological networks. Genostar's software is platform independent and can thus be used for both Mac OS X, Windows, and Linux.

Sequence assembly

- Mapping of an ensemble of sequences on a reference sequence
    - between a reference sequence and contigs, between two sequences or between two sets of sequences
    - finding of exact matches with minimum length using MUMmer
    - detection of specific regions and SNPs
    - creation of an assembled sequence relative to reference sequences

Genomic annotation

- Gene prediction: ab-initio gene prediction using a Hidden Markov model based method
- BlastX
- Automatic annotation transfer using BlastP

Proteic annotation

Metabolic Pathway Builder integrates several methods dedicated to proteic annotation:

- Pfam domain prediction using HMMER
- Several EMBOSS methods (antigenic, 2D structure prediction)

Expression Data Solution (EDS)

Genostar's Expression Data Solution (EDS) connects microarray data to genes, gene products and biochemical reactions, based on keywords and annotations. This software solution allows to:

- Assign expression values to the gene names and IDs
- Identify co-expressed genes and visually analyze the reactions and metabolic pathways in which they are involved
- Identify and perform analysis on co-regulated genes in terms of genomic localization, functional annotation and metabolism
- Colorize CDSs of interest in genomic maps according to their expression values and highlight the corresponding reactions in interactive metabolic KEGG maps
- Analyze the significance of functional data of a collection or sub-collection of CDSs (GO, KEGG and more): Fisher test
- Collect and visualize all functional data in exportable tables and maps

Genostar's MicroB database is constructed of perfectly integrated and rigorously cross-checked genomic, proteic, biochemical and metabolic data approximately 1100 bacterial and archaeal organisms.

3. The metabolic SearcH And Reconstruction Kit (metaSHARK)

The metabolic SearcH And Reconstruction Kit (metaSHARK) is a new fully automated software package for the detection of enzyme-encoding genes within unannotated genome data and their visualization in the context of the surrounding metabolic network. The gene detection package (SHARKhunt) runs on a Linux system and requires only a set of raw DNA sequences (genomic, expressed sequence tag and/or genome survey sequence) as input. Its output may be uploaded to our web-based visualization tool (SHARKview) for exploring and comparing data from different organisms. We first demonstrate the utility of the software by comparing its results for the raw*Plasmodium falciparum* genome with the manual annotations available at the PlasmoDB and PlasmoCyc websites. We then apply SHARKhunt to the unannotated genome sequences of the coccidian parasite *Eimeria tenella* and observe that, at an *E*-value cut-off of 10^{-20}, our software makes 142 additional assertions of enzymatic function compared with a recent annotation package working with translated open reading frame sequences. The ability of the software to cope with low levels of sequence coverage is investigated by analyzing assemblies of the *E.tenella* genome at estimated coverages from 0.5× to 7.5×. Lastly, as an example of how metaSHARK can be used to evaluate the genomic evidence for specific metabolic pathways, we present a study of coenzyme A biosynthesis in *P.falciparum* and *E.tenella*.

Existing semi-automated enzyme annotation software starts with a set of predicted proteins from an annotated genome and, by a variety of text mining and/or sequence comparison methods, constructs a list of the enzymatic functions that are asserted to be present. Our software (SHARKhunt) differs in that it requires only a set of DNA sequences [finished chromosomes, contigs, genome survey sequences or expressed sequence tags (ESTs)] as input, and hence can be applied to extract new knowledge of metabolic capabilities from preliminary data produced by unannotated and ongoing genome sequencing projects. Models derived from sets of known enzymes are used to search through the DNA sequences to find regions with significant similarity to the model sequences. The confidence of each functional assertion is measured by an *E*-value score, and the full set of predictions is output in various formats for consultation, human curation or further automated network analysis. Results may be uploaded to an online visualization tool (SHARKview), which permits users to browse freely around the KEGG metabolic network, run BLAST searches on their predicted gene sequences and compare data from different organisms or different sources of annotation.

Preparation of data

The SHARKhunt search protocol described below requires both a set of PSI-BLAST polypeptide profiles and a set of associated HMMER profile hidden Markov models (HMMs)

We use the set of PRIAM PSI-BLAST profiles (July 2004 version) and the protein sequence data from which they were derived as the basis for our profile HMMs. This set constitutes 2562 PSI-BLAST profiles covering a total of 1967 enzymatic functions, defined by Enzyme Commission (EC) number. There are more profiles than functions in the PRIAM data because some functions are represented by more than one homologous family of proteins. For each PRIAM profile containing more than one sequence, a HMM is generated by taking the original set of protein sequences constituting the profile, constructing a multiple alignment using

MUSCLE and passing this to the HMMbuild program (part of the HMMER package, available from http://hmmer.wustl.edu).

The SHARKhunt search protocol

The search method used in SHARKhunt is an automated version of a protocol that we have previously demonstrated to be effective in searching genomic DNA for distant homologues of a set of model sequences . The Wise2 package is used to aligning the profile HMMs with genomic DNA and produce predicted polypeptide sequences for the resulting hypothetical gene fragments. Although the Wise2 algorithm is very powerful, it is too slow to be used to search through a whole genome. Hence, we apply a preliminary filtering step, where PSI-TBLASTN is used to search the genome for regions with some similarity to each original PRIAM profile (hits with *E*-value < 1.0). These regions are then extracted from the genome and passed to Wise2 for further analysis with the appropriate HMM (Figure 1). Any resulting Wise2 hits are assessed by using PSI-BLAST to compare the predicted polypeptide translation with the original PRIAM profile, and the *E*-values and locations of predicted coding regions are output.

The SHARKhunt search protocol is a two-stage, profile-based method that aims at detecting all regions of a genome homologous to an enzyme model. A preliminary PSI-TBLASTN search identifies regions with some similarity to a PRIAM profile. These regions .In cases where several different profiles produce hits to the same region of DNA, we take the function of that region to be the same as the best hit (i.e. the lowest *E*-value). Some enzymes require more than one functional unit in order to operate, or are represented by more than one homologous family of proteins. This has already been taken into account in the generation of the PRIAM profiles and their associated logical AND/OR rules . To assert the presence of an enzymatic function in the genome, the hits must agree with the relevant PRIAM rule. The SHARKhunt package (including the programs MUSCLE, HMMer, PSI-BLAST and Wise2 and the necessary PRIAM profile data) is available to download from the metabolic SearcH And Reconstruction Kit (metaSHARK) website (http://bioinformatics.leeds.ac.uk/shark/). The package comprises two executable scripts: SHARKhunt invokes a Java program to run the profile searches. The output files created by this program include a FASTA format library of predicted polypeptide sequences, a GFF (gene feature format, seehttp://www.sanger.ac.uk/Software/formats/GFF/) file containing the locations of the predicted gene structures on the DNA sequence, a list of EC numbers representing the functions detected within the input DNA, together with their PSI-BLAST *E*-value confidence scores, and an eXtensible Markup Language (XML) file, which may be uploaded to the metaSHARK website (see above URL) for easy browsing and visualization of the results.SHARKmodel allows users to create their own PSI-BLAST and HMMer profile models from sets of homologous polypeptide sequences for customized searches using the SHARKhunt protocol.

Verification of search method

To confirm that the search protocol is effective in detecting enzyme-encoding genes within eukaryotic DNA, SHARKhunt was run on the genome of the human malaria parasite, *Plasmodium falciparum*. The profile data used for this search were a special 'jacknife' dataset, prepared from the original PRIAM models by removing all sequences from *Plasmodium* spp. and re-making the affected profiles (38 out of 2562 profiles).The results were compared with the enzymatic functions annotated for this organism in the initial genome publication (data available at the PlasmoDB website: http://www.plasmodb.org/) and the subsequent re-annotation of metabolic enzymes prepared by the PlasmoCyc team (http://plasmocyc.stanford.edu/).As a comparison with an existing method based on polypeptide sequence input, searches were also run against*P.falciparum* with the PRIAM software, using the same set of jacknife PSI-BLAST profiles and all predicted polypeptides from the original genome annotation.

Analysis of preliminary genome sequences

As a practical example of how metaSHARK may be used to investigate metabolic pathways within unfinished genomes, we applied the software to the available genomic DNA sequences of another apicomplexan parasite, the coccidian *Eimeri tenella* (http://www.sanger.ac.uk/Projects/E_tenella/). A full description of the *E.tenella* genome project has been given elsewhere. Briefly, more than 855 000 reads with an average read length of 525 bp have yielded ~45 Mb of unique sequence for a coverage of ~7.5×. The data are presently assembled into 8718 contigs and the G+C content of the genome is ~53%. Little annotation has yet been undertaken on this genome, and *ab initio* genefinding is hampered by the current lack of well-characterized gene structures in *Eimeria* spp. As such, it presents us with a good example of an organism for which the metaSHARK system may be of use.

For comparison with SHARKhunt, we ran the PRIAM software on all *E.tenella* open reading frames (ORFs) of 50 amino acids or longer. In addition, the best available *ab initio* gene predictions for this species were obtained using the genefinding program TwinScan2, trained on a set of 350 human-annotated genes from the related coccidian, *Toxoplasma gondii* (Aaron J. Mackey, personal communication). The polypeptide translations of these gene predictions were used in a PRIAM analysis, in parallel with the genomic ORF data. The PRIAM software was also run on all ORFs of at least 50 amino acids found within the clustered (98% similarity) EST+ORESTES data available for *E.tenella* (downloadable from the Sanger Institute website, seehttp://www.sanger.ac.uk/Projects/E_tenella/). These clusters were obtained by using publicly available ESTs from the NCBI and open reading frame-expressed sequence tags (ORESTES) cDNA reads generated by Alda M. Madeira and Arthur Gruber at the University of Sao Paulo,

Brazil (unpublished data). ORESTES reads are cDNA fragments synthesized by a low stringency RT–PCR process using arbitrary 18–25mer primers.

Effect of variation in genome coverage on performance

To investigate how the number of enzymatic functions predicted by SHARKhunt is affected by variations in the coverage of the input genome sequence, we re-ran the *E.tenella* searches using an earlier genome assembly based on an ~4.3× genome coverage (from December 2002), and also on an assembly constructed using the program Phrap (http://www.phrap.org/) from sets of reads representing ~0.5× coverage (all data obtained from the Sanger Institute website).

Reference

The Pathway Tools cellular overview diagram and Omics Viewer by Suzanne M. Paley, Peter D. Karp

Tools for the functional interpretation of metabolomic experiments by Monica Chagoyen, Florencio Pazos -

Briefings in Bioinformatics, Volume 14, Issue 6, November 2013, Pages 737–744