4574–4582 *Nucleic Acids Research, 2002, Vol. 30 No. 20*

GeneCensus: genome comparisons in terms of metabolic pathway activity and protein family sharing

J. Lin, J. Qian, D. Greenbaum, P. Bertone, R. Das, N. Echols, A. Senes, B. Stenger and M. Gerstein*

Department of Molecular Biophysics and Biochemistry, Yale University, PO Box 208114, New Haven, CT 06520, USA

Received February 6, 2002; Revised and Accepted August 8, 2002

ABSTRACT

We present a prototype of a new database tool, GeneCensus, which focuses on comparing genomes globally, in terms of the collective properties of many genes, rather than in terms of the attributes of a single gene (e.g. sequence similarity for a particular ortholog). The comparisons are presented in a visual fashion over the web at GeneCensus.org. The system concentrates on two types of comparisons: (i) trees based on the sharing of generalized protein families between genomes, and (ii) whole pathway analysis in terms of activity levels. For the trees, we have developed a module (TreeViewer) that clusters genomes in terms of the folds, superfamilies or orthologs—all can be considered as generalized 'families' or 'protein parts'—they share, and compares the resulting trees side-by-side with those built from sequence similarity of individual genes (e.g. a traditional tree built on ribosomal similarity). We also include comparisons to trees built on whole-genome dinucleotide or codon composition. For pathway comparisons, we have implemented a module (PathwayPainter) that graphically depicts, in selected metabolic pathways, the fluxes or expression levels of the associated enzymes (i.e. generalized 'activities'). One can, consequently, compare organisms (and organism states) in terms of representations of these systemic quantities. Development of this module involved compiling, calculating and standardizing flux and expression information from many different sources. We illustrate pathway analysis for enzymes involved in central metabolism. We are able to show that, to some degree, flux and expression fluctuations have characteristic values in different sections of the central metabolism and that control points in this system (e.g. hexokinase, pyruvate kinase, phosphofructokinase, isocitrate dehydrogenase and citric synthase) tend to be especially variable in flux and expression. Both the TreeViewer and PathwayPainter modules connect to other information sources related to individual-gene or organism properties (e.g. a single-gene structural annotation viewer).

INTRODUCTION

Advances in sequencing technology have created the opportunity to perform large-scale genome comparisons. Presently, there are many systems focusing on specific types of comparisons for many genomes [e.g. COG (1), PENDANT (2), or KEGG (3), WIT (4), MUMmer (5)]. Conversely, there are other systems analyzing single genomes from many perspectives [e.g. Flybase (6) MIPS (7), or SGD (8), ECOCYC (9)]. We present here a new prototype tool (outlined graphically in Fig. 1) that compares multiple genomes through multiple and some novel criteria.

Our approach to genome comparison is two-fold. Our first view, TreeViewer, displays genome-wide comparisons through tree building based upon different characteristics of the genomes (10). These characteristics include broad statistics, such as fold and gene content and amino acid composition. The trees that we provide can be compared against other information, and dynamically reconfigured based on different genomic characteristics. Our second viewer, PathwayPainter, provides the user with an extensive comparison of genomes in terms of their mRNA expression, flux (11) and percent identity (PID), in three major metabolic pathways: TCA, glycolysis and pentose phosphate. Both of these views are linked to additional modules representing more traditional analysis formats. These include modules that examine open reading frames (ORFs), organisms, and various compositions of genomes.

In general, it is relatively difficult to integrate disparate information sources into one comprehensive database; it is difficult to determine which data sets will be useful under different circumstances. We present some useful examples in the Gene Exploring section on how one can extract biologically relevant and novel information from our database. Nevertheless, these demonstrations, using specific features within GeneCensus, do not provide a reason for the inclusion or exclusion of any specific features in the database.

*To whom correspondence should be addressed. Tel: +1 203 432 6105; Fax: +1 360 838 7861; Email: mark.gerstein@yale.edu

FEATURE OVERVIEW: TREEVIEWER

The comparative TreeViewer (outlined graphically in Fig. 2) is an online interface for displaying previously computed trees, and also acts as a tool for comparing trees built using different methods. The organisms included in the tree server provide for diverse phylogenetic comparisons. They encompass all three kingdoms of life (Eukarya, Bacteria, Archaea), diverse environments (normal to extreme), and a wide range of genome sizes (0.6–97 Mb).

The architecture of the tree server is two-dimensional. The first dimension, based on the methods in which the trees are built, include: gene occurrence, fold composition, dinucleotide frequency, COGs, metabolic pathways and traditional ribosomal trees. The second dimension provides information with which we can compare the trees. These characteristics include: taxonomy, fold composition, COGs, AT-content, genome size and superfamily occurrence. Thus, one has the ability to select which tree to build, and in which context that tree will be compared. In addition, one has the ability to compare many trees with the traditional ribosomal tree; each of these options can be selected via the light blue user interface elements. The techniques and procedures for building these trees are explained in Lin and Gerstein (10). The trees that are currently available can be subdivided into the following self-explanatory sections.

Ribosome

These trees are built through comparing the similarity of the ribosomal RNA. This traditional method (12) for phylogenetic analysis is based on the small subunit ribosomal RNA (SSU rRNA). For comparison, the trees based on the large subunit (LSU) are also provided.

Folds

These trees are built based on the presence or absence of folds in different organisms, as determined by Hegyi *et al.* (13). In addition, we compare trees based upon the subdivision of folds into classes (all alpha, all beta, alpha + beta and alpha/beta). A further comparison in this category differentiates between the distance-based and parsimony techniques for tree building.

Superfamilies

Superfamilies are less broad structural groupings than folds, and because of their greater number, they have been found to be more differentiating, producing trees similar to the traditional phylogeny. The data were collected using a similar approach to Hegyi and Gerstein (14).

COGs

We also compare the genomes based on the occurrence of orthologous genes based on COGs, clusters of orthologous genes (1). Trees were built for the three major types of COGs (i.e. metabolism, cellular processes, information storage and processing), as well as for the smaller functional categories. We represent these categories using single letters in the user interface.

Composition

These trees are built on the simple composition of the amino acids and dinucleotides. The trees marked *raw* are based on the absolute number of amino acids and dinucleotides. These values are used to generate a vector and the calculated distance. For the other trees, the numbers calculated are normalized by the total number, producing percentages, which were used to generate a distance matrix for tree construction.

ORFs

This set of trees is composed of trees built on the sequence similarity of homologous genes. The genes chosen for this comparison were present in the genomes only once; thus paralogous genes were not a factor.

Enzymes

The sequence similarity of individual enzymes in the three central metabolic pathways (the TCA cycle, glycolysis and the pentose phosphate pathway) was used to construct these trees.

FEATURE OVERVIEW: PATHWAYPAINTER

PathwayPainter provides a multi-organism representation of three major metabolic pathways and their component enzymes. (See Fig. 3 for a graphical outline.) It is divided into two views: (i) the Pathway View provides a macro-view—a schematic of each metabolic cycle flanked by the flux or various expression values for each of the enzymes; (ii) the Enzyme View provides a micro-view; the data (expression, flux, PID) is presented with reference to each individual enzyme.

Pathway View

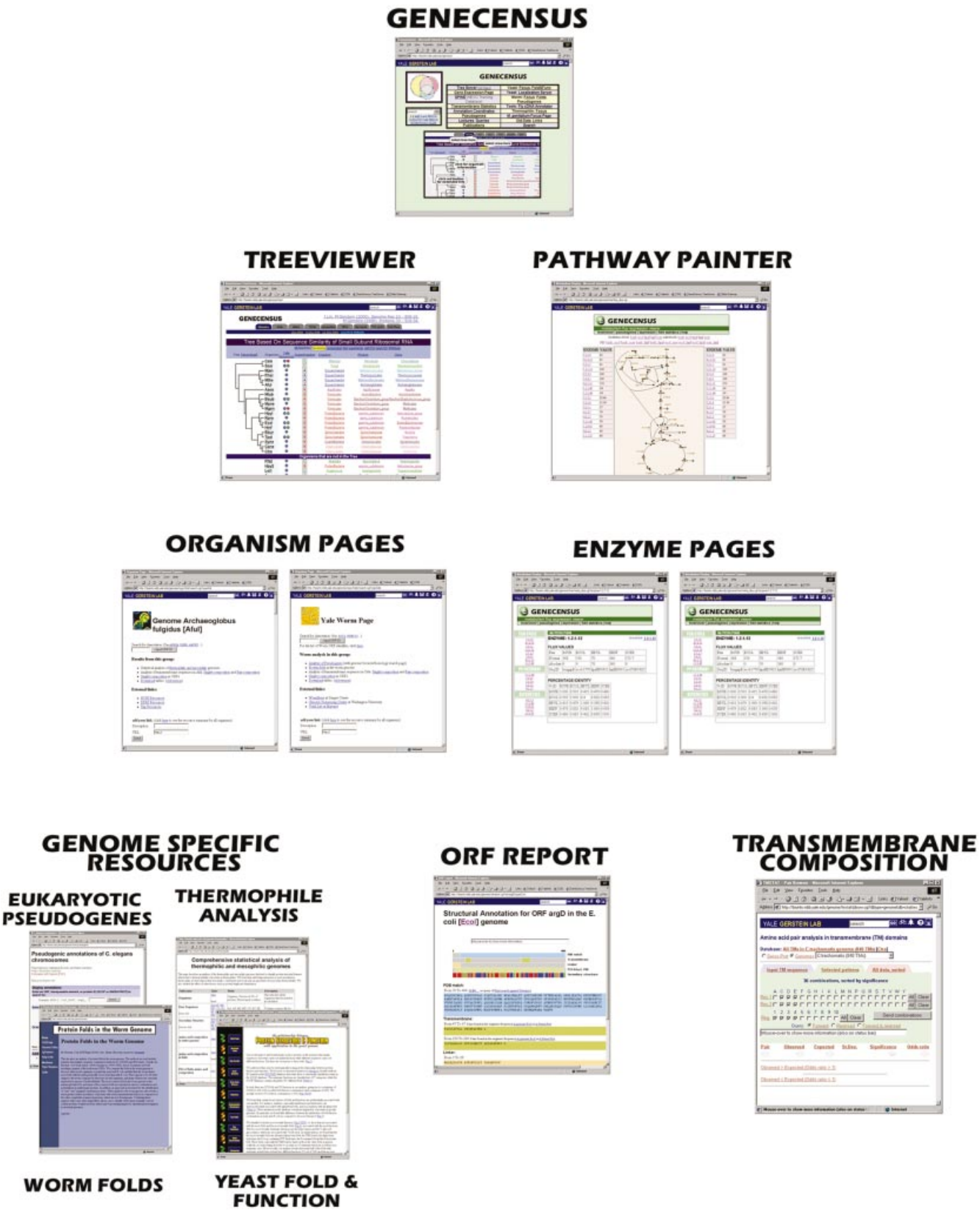
The Pathway View allows the user to compare information for each enzyme in the form of: (i) flux values (normalized, absolute and standard deviation); (ii) average and standard deviation of gene expression change (DNA and cDNA arrays) (explained below); (iii) PID (between orthologous enzymes in the pathway for multiple organisms). Presently, the flux and PID information is available for *Escherichia coli*, *Saccharomyces cerevisiae*, *Bacillus subtilis*, *Haemophilus influenzae* and *Helicobacter pylori*. Two sets of information can be independently selected to display the data of choice, and these sets are labeled right column and left column. An overview map of the pathways is available in the center column for reference.

Flux analysis

Flux, a measure of the rate at which metabolites are processed to become output, is calculated at a steady state (15). Determination of flux provides critical information for rational pathway modification and metabolic engineering (16,17). While there are many published maps of pathways illustrating the processing of basic metabolites, they provide little in terms of describing pathway fluxes under diverse conditions.

We obtained raw absolute flux values for three organisms (*S.cerevisiae*, *B.subtilis*, *E.coli*) (18–20) (These are reported as 'absolute' fluxes on the website). For two organisms (*H.influenzae* and *H.pylori*), we calculated theoretical relative flux values using stoichiometric analysis. We describe this

calculation here. Our first step involves reconstructing the map of the central metabolic pathway in the two organisms using information from the KEGG metabolic database (21). It is known that *H.influenzae* (22) and *H.pylori* (23) have incomplete TCA cycles. We decomposed the reconstructed pathway into elementary modes using the METATOOL software (24).



Each elementary mode consists of a minimal set of enzymes that could operate at steady state with all irreversible reactions proceeding in the appropriate direction and further reduced to omit extraneous metabolites not necessary for the net reaction (25). One should note that there is more than one elementary mode that can connect two chemical species. In order to choose the best elementary mode that represents the most efficient routes of chemical conversion, we optimized the process by using a combined objective function of maximization of ATP and minimization of glucose use. We obtained the ratio of the end products of glucose metabolism produced from earlier studies. For example, *H.influenzae* produces succinate and acetate as the end products of glucose metabolism in the ratio of 4.3:1 (22). These ratios act as constraints in the optimization process. We used the LINDO software to carry out this optimization.

Finally, we merged the results for all five organisms and normalized the flux values to make them comparable. Normalization of values is done with respect to glucose intake (i.e. the entry of glucose in the metabolic pathway is considered to represent 100% relative flux). We computed the relative flux that inputs into various pathway routes as a fraction of this initial amount. Therefore, even though the actual pathway flux can vary from one organism to another, normalized fluxes are comparable in a relative sense. We report these final normalized values on the website.

Expression levels

PathwayPainter encompasses information from a variety of gene expression experiments, corresponding to data collected with both DNA and cDNA arrays. In particular, we have collected various microarray data sets from the Stanford Microarray Database (26) and extracted expression data for each enzyme in the following three pathways: (i) TCA cycle, (ii) glycolysis and (iii) pentose phosphate. While the determination of gene expression levels using high-throughput experimentation is a growing field, we present here mainly data sets derived from yeast experiments, the most common organism for expression analysis to date. The dynamic nature of GeneCensus will allow us to provide additional microbial expression data sets as they become available.

In the current version of GeneCensus we focus on six individual experiments: (i) cell-cycle experimentation of yeast cells synchronized by alpha factor arrest (27); (ii) a second cell-cycle experiment where yeast cells were similarly synchronized via the arrest of a cdc15 temperature-sensitive mutant (27); (iii) a yeast diauxic shift experiment measuring the temporal program of gene expression following a metabolic shift from fermentation to respiration (28); (iv) an assay measuring the change in yeast expression during sporulation (29); (v) an experiment capturing the cellular response of *E.coli* following exposure to UV radiation (30); and (vi) a profile of gene expression in the germ line of *Caenorhabditis elegans* (31).

Enzyme View

The Enzyme View of PathwayPainter details the absolute and normalized flux levels for the enzyme in each genome. Additionally, it allows for the visualization of sequence similarity between the organisms compared with the specific enzyme for which the flux is being measured. Below the PID table, another table outlines the expression values of that specific enzyme in yeast and *E.coli* under multiple conditions. Finally, links are provided to the TreeViewer wherein the user can view trees based on that particular enzyme.

In relation to gene expression, for each enzyme, we report the following. (i) Raw unscaled values R as available from the various sites as either copies per cell, $\log_2$(ratio of mRNA levels), or normalized transcript level divided by mean value, depending on the specific data set. We represent these as $P(i, t)$, which represent the expression of gene i at time t. The calculated ratios are thus $R(i, t) = P(i, t)/P(i, r)$ in which $P(i, r)$ is the reference state. (ii) Multi-experiment scaled expression value M derived from multiple experiments (32), which provides a standard of comparison for expression data. This is derived from scaling together various microarray and SAGE data sets and is on an absolute scale in copies per cell (32). (iii) Average expression ratio change C over of the length of the profile, which is calculated by $C = \langle R(i, t) \rangle$. This measures the degree of variability in expression in a particular experiment for a given enzyme. (iv) Expression ratio fluctuation E was calculated using the standard deviation of expression ratios (i.e. $E = \sqrt{\langle (R(i,t) - \langle R(i,t) \rangle_t)^2 \rangle_t}$. Note that enzymes consistently expressed under most conditions will show minimal standard deviations that are closely correlated between experiments.

MODULES

In addition to the TreeViewer and the PathwayPainter we also provide further subsidiary modules. These are the ORF, Organism and Composition viewers.

ORFViewer

The ORFViewer module provides various resources related to a given ORF. For example, protein structural annotations are graphically represented on a plot for every ORF. These include: (i) PDB PSI-BLAST matches; (ii) regions of low complexity [identified with the SEG program using standard parameters $K(1) = 3.4$, $K(2) = 3.75$ and a window size of 45 residues (33)]; (iii) transmembrane segments [defined using the GES hydrophobicity scale (34)]; (iv) linker regions (e.g. low complexity regions); and (v) uncharacterized regions.

Figure 1. (Opposite) A pictorial overview of GeneCensus through screenshots. The top image shows the homepage, which, in addition to linking to pages in GeneCensus, also provides links to multiple other bioinformatics resources, including pages on gene expression, protein interactions and pseudogenes. GeneCensus bifurcates into two semi-independent modules, the TreeViewer and PathwayPainter, shown on the second level of the diagram. Information relevant to TreeViewer can be accessed through the secondary modules (as seen on the third tier) such as the OrganismViewer. Similarly, enzyme-specific data, as opposed to genome-specific enzyme data, can be viewed. The fourth and final level of GeneCensus provides more specific information for many of the ORFs in the ORFViewer, as well as some smaller modules with less generalized information. These include information on: (i) transmembrane proteins, (ii) pseudogenes, (iii) thermophile analysis, and (iv) data on folds for both the worm and yeast genome.

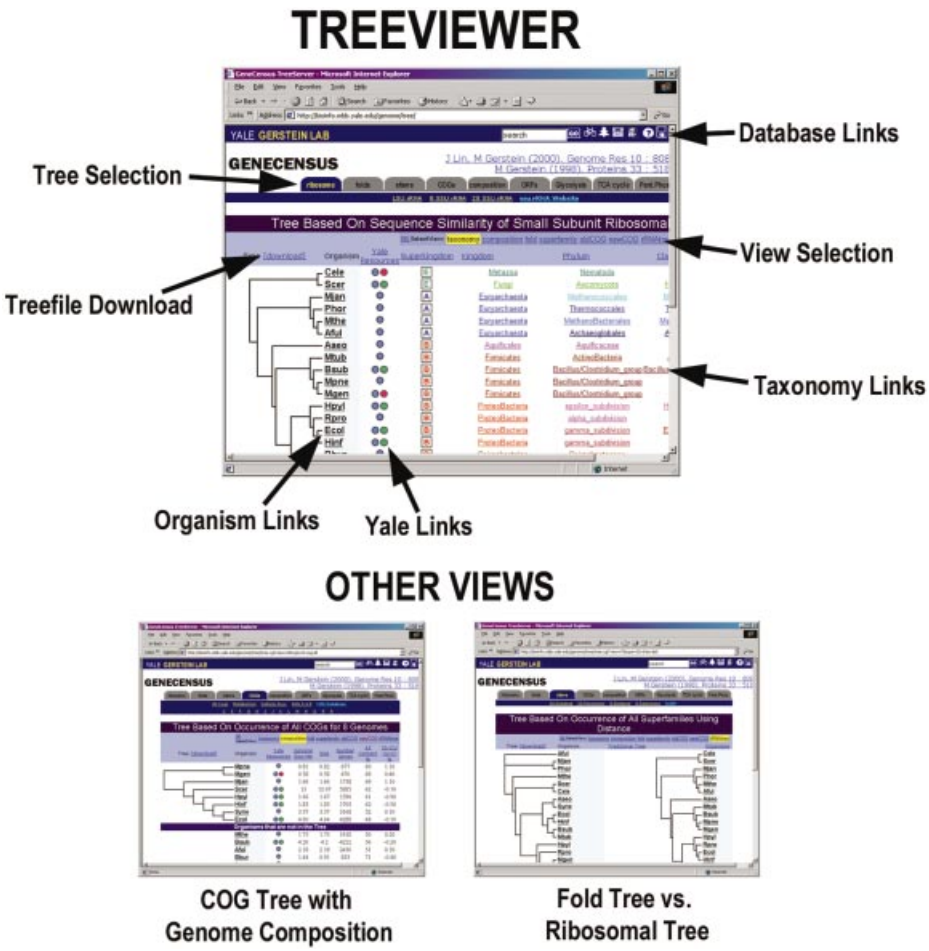


Figure 2. An annotated close-up of the TreeViewer module. The figure highlights the important parts of the web page format. The top bar, which is maintained throughout the site, provides a search option, a help file and links to PartsList (36), NESGC (41) and Molecular Motions Database (48). To manipulate the view of the data on the web page we provide a menu bar to select which type of tree to view and a second menu bar to determine in which secondary dimension to view the tree. In addition, there are multiple color-coded links next to each organism—green for metabolic pathways, blue for the organism page and red for other Yale pages associated with that organism. For examples of the multiple views, we present a COG tree viewed through genome composition and a fold tree viewed in comparison to the traditional ribosomal tree.

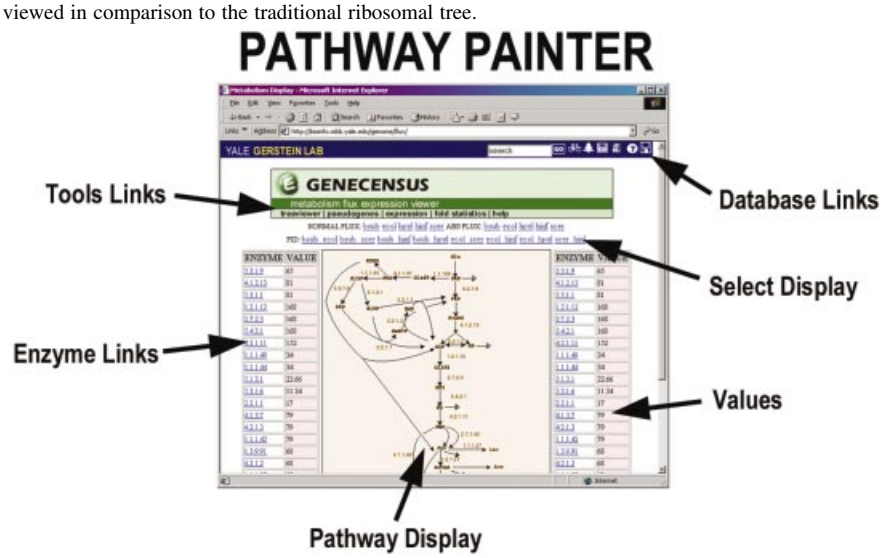


Figure 3. The second major module, the PathwayPainter. As with the TreeViewer, the top menu bar is maintained. This page is built around the metabolic pathway. We present an overview image of two pathways in the center of the page. Flanking the image are the component enzymes, with a value next to each enzyme. Using the menu on the top of the page, the user can select the desired value for those enzymes in that pathway. These include the flux values for multiple organisms, various expression values for yeast and *E.coli*, and PID variability. Additionally, each enzyme links to an enzyme-oriented page which displays the data in its entirety for that specific enzyme.

We also provide organism-specific information. For example, additional information available for yeast includes gene expression data, subcellular localization (35), protein structural comparisons (36), previously unannotated genes (37), transposon tagging, gene disruptions (38) and data from protein chip experiments (39). Organism-specific information for *Mycoplasma genitalium* information includes a breakdown of structural characterizations of the genome and related NESGC construct database entries (40,41).

The proliferation of multiple interchangeable nomenclature schemes remains an outstanding problem in compiling ORF annotation. For example, YJR155W, AADA_YEAST, AAD10, J2245, P47182 and CAA89688.1 all refer to the identical ORF in the yeast genome. As a solution, we include Smartlink in GeneCensus; Smartlink is a translator that integrates all the disparate systems and maps one nomenclature to another.

OrganismViewer

The OrganismViewer integrates varied information about an organism. For each organism, we provide an amino acid composition viewer (see below for more details), links to the source data, and additional links to external resources. These pages were designed as an open system, allowing users to add relevant resource links. We will continue to integrate more studies and resources for each genome.

The OrganismViewer not only presents the research specific to GeneCensus, but also integrates research from other institutions and acts as a central resource and link manager for each genome. For example, the *C.elegans* page provides comprehensive analysis of pseudogenes and protein folds; alternatively, the *S.cerevisiae* page includes studies of the relationship between fold and function, microarray expression data analysis (42), clustering of phenotype patterns (38), subcellular localization of proteins (35), composition of individual ORF features (43) and comparison of the yeast and worm genomes in terms of folds (44). We have also included a statistical analysis of thermophilic and mesophilic genomes (45).

CompositionViewer

In early cryptography, it was discovered that one could break ciphers by analyzing the frequency of letters/symbols and their combinations within a coded text. When the frequency of a letter or letter combination differs from the expected frequency of a randomly inserted letter, it can be interpreted as significant—and possibly deciphered. Similarly, a protein sequence can be decoded and important and significant amino acid combinations discerned by examining the frequency and specific occurrences of entities whose occurrence is higher than they would be if inserted randomly (i.e. the expected occurrence of a given amino acid within a sequence). These combinations may be important for concepts such as binding or protein structure (46). Additionally, determining the individual amino acid composition of the ORFs in thermophilic organisms is essential for understanding their stability in extreme environments.

The CompositionViewer, another module within GeneCensus, provides composition information about genomes similar to that used to build and compare trees in the TreeViewer; the CompositionViewer is a valuable tool for analyzing the aforementioned amino acid pair patterns in a genome. Included within this component are (i) calculations of amino acid composition, (ii) secondary structure, and (iii) a tool for dynamically calculating amino acid composition pairs. Pairs, in terms of amino acid composition, are defined as combinations of an amino acid residue with another residue at separation i, $i + k$. For instance, AL3 corresponds to an alanine residue and a leucine residue at i, $i + 3$ (AxxL). The significance of over- and under-represented pairs is calculated by comparing the observed occurrence of the pair in a database with a random expectation distribution. The random distribution is calculated as the average of any possible internal permutations of all sequences of the database. The significance corresponds to the probability of observing the same or a larger difference between observed and expected occurrences of a pair in random sequences.

For membrane data we used the TMSTAT method (43) to analyze frequently occurring combinations of residues of transmembrane domains. The server returns the most significant over- and under-represented singlets and pairs in each database.

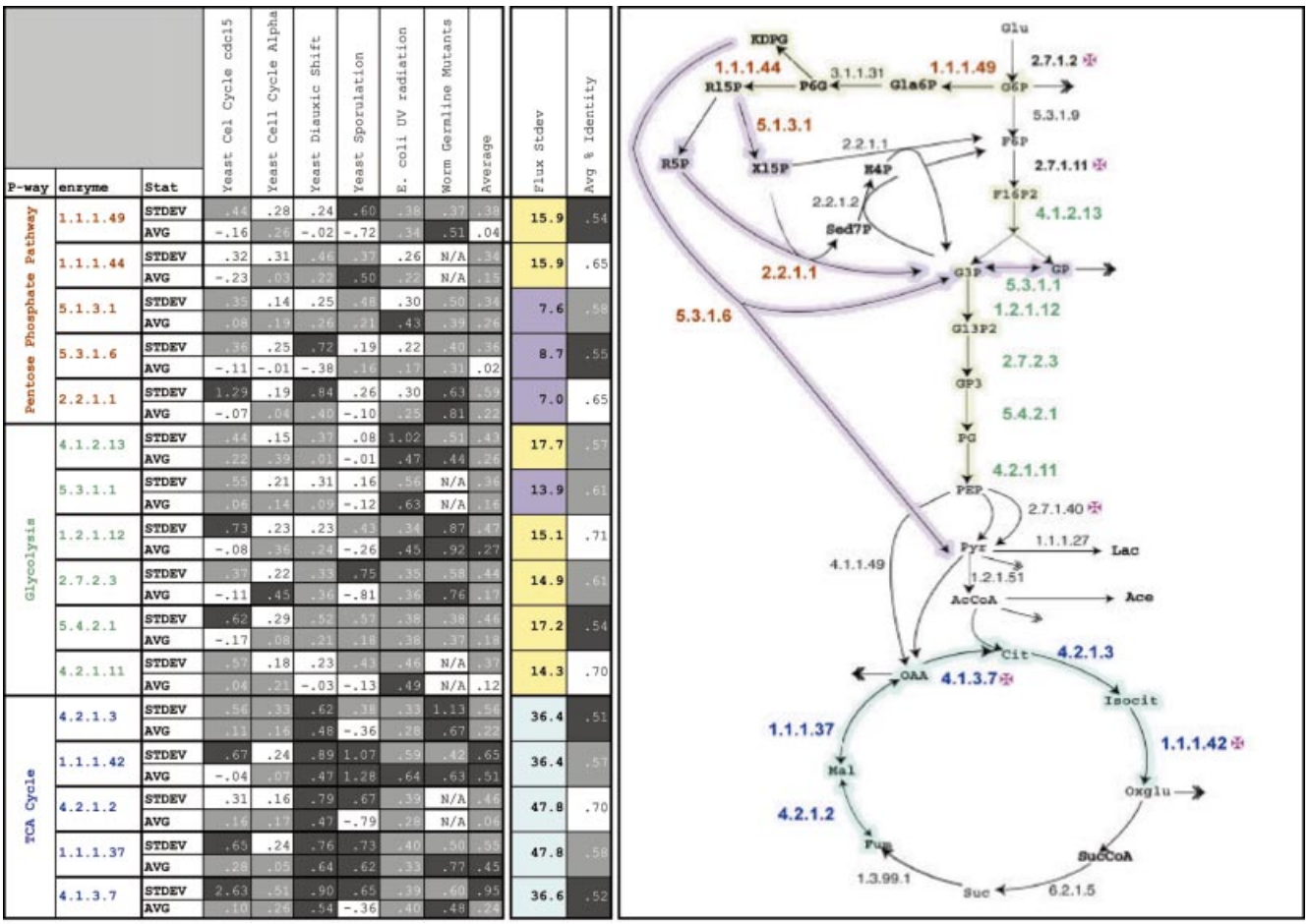
This viewer is accessed either through the organism viewer or directly at http://bioinfo.mbb.yale.edu/genome/tmstat/comp.cgi.

GENE EXPLORING: PRACTICAL RESULTS USING GENECENSUS

PathwayPainter illustration

Given the difficulty of describing our database without actually providing a manual, we attempt to provide directions for using the data, and a illustration of the utility of the system by presenting some biologically relevant qualitative conclusions that can be extracted from our database.

In Figure 4, we illustrate the scientific conclusions one can derive from PathwayPainter through comparing the variation in expression, flux and PID of enzymes over many different experiments and organisms. We present some of our findings here. Additionally, we show how these data may be utilized to determine which enzymes can best be used as internal controls for normalizing microarrays. The data presented in the figure include: (i) the average expression change C, which is the average of all the expression ratios from all the time points between two conditions; (ii) expression fluctuation E, which represents the standard deviation of the expression ratios from all experimental time points; (iii) flux variation F, which indicates the standard deviation of flux from the different organisms; and (iv) sequence similarity S, which is the average percentage similarity of orthologous pairs. Notably, the experiments can be compared not only with the average expression change C, but also with expression fluctuation E. Both values are important for a clear understanding of the expression ratio profiles, since enzymes may have very different average expression change and expression fluctuation values. For example, in the cdc15-arrested cell-cycle experiment, both citric synthase (4.1.3.7) and glycolaldehyde-transferase (2.2.1.1) exhibit low average expression change but very high expression fluctuation.



P-way	enzyme	Stat	Yeast Cel Cycle cdc15	Yeast Cell Cycle Alpha	Yeast Diauxic Shift	Yeast Sporulation	E. coli UV radiation	Worm Germline Mutants	Average	Flux Stdev	Avg % Identity
Pentose Phosphate Pathway	1.1.1.49	STDEV	.44	.28	.24	.60	.38	.37	.38	15.9	.54
		AVG	-.16	.26	-.02	-.72	.34	.51	.04		
	1.1.1.44	STDEV	.32	.31	.46	.37	.26	N/A	.34	15.9	.65
		AVG	-.23	.03	.27	.50	.22	N/A	.15		
	5.1.3.1	STDEV	.35	.14	.25	.48	.30	.50	.34	7.6	.56
		AVG	.08	.19	.26	.21	.43	.39	.26		
	5.3.1.6	STDEV	.36	.25	.72	.19	.22	.40	.36	8.7	.55
		AVG	-.11	-.01	-.38	.16	.17	.31	.02		
	2.2.1.1	STDEV	1.29	.19	.84	.26	.30	.63	.59	7.0	.65
		AVG	-.07	.04	.40	-.10	.25	.81	.22		
Glycolysis	4.1.2.13	STDEV	.44	.15	.37	.08	1.02	.51	.43	17.7	.57
		AVG	.22	.39	.01	-.01	.47	.44	.26		
	5.3.1.1	STDEV	.55	.21	.31	.16	.56	N/A	.36	13.9	.61
		AVG	.06	.14	.09	-.12	.63	N/A	.16		
	1.2.1.12	STDEV	.73	.23	.23	.43	.34	.87	.47	15.1	.71
		AVG	-.08	.36	.24	-.26	.45	.92	.27		
	2.7.2.3	STDEV	.37	.22	.33	.75	.35	.58	.44	14.9	.61
		AVG	-.11	.45	.36	-.81	.36	.76	.17		
	5.4.2.1	STDEV	.62	.29	.52	.57	.38	.38	.46	17.2	.54
		AVG	-.17	.08	.21	.18	.38	.37	.18		
	4.2.1.11	STDEV	.57	.18	.23	.43	.46	N/A	.37	14.3	.70
		AVG	.04	.21	-.03	-.13	.49	N/A	.12		
TCA Cycle	4.2.1.3	STDEV	.56	.33	.62	.38	.33	1.13	.58	36.4	.51
		AVG	.11	.14	.48	-.36	.28	.67	.27		
	1.1.1.42	STDEV	.67	.24	.89	1.07	.59	.42	.65	36.4	.57
		AVG	-.04	.07	.47	1.28	.64	.63	.51		
	4.2.1.2	STDEV	.31	.16	.79	.67	.39	N/A	.46	47.8	.70
		AVG	.16	.17	.47	-.79	.28	N/A	.06		
	1.1.1.37	STDEV	.65	.24	.76	.73	.40	.50	.55	47.8	.58
		AVG	.28	.05	.64	.62	.33	.77	.45		
	4.1.3.7	STDEV	2.63	.51	.90	.65	.39	.60	.95	36.6	.52
		AVG	.10	.24	.54	-.36	.40	.48	.24		

Figure 4. A cross-section of the results that can be seen with the PathwayPainter module. Enzymes were chosen from three metabolic pathways: citric acid cycle (blue), glycolysis (green) and pentose phosphate pathway (red); the information presented includes expression, flux and sequence similarity data. We present expression data, the relative expression of a gene in relation to a control, from six experiments: yeast diauxic shift, yeast sporulation, *E.coli* UV response, *C.elegans* mutant germline, and two yeast cell-cycle expression sets. We summarized the data by calculating the standard deviation and the average for each enzyme profile in each experiment, as well as combined statistics for all the experiments. Values in the top quartile were shaded black, in the middle two, gray, and in the bottom quartile, white. Sequence similarities of the enzymes were calculated by averaging the percentage sequence identity between orthologous genes. These are shaded in the same fashion as the expression values. For the flux values, we calculated the standard deviation of the flux values for all organisms examined. Values in the top quartile are colored aqua, middle two quartiles, yellow, and bottom quartile, purple. We show a schematic of all three pathways, with enzyme numbers color coded by pathway. The arrows representing the reaction are colored by the degree of flux variation; this seems to correlate closely with the pathways. TCA shows the greatest flux values and the pentose phosphate pathway comprises the lowest. The pink crosses label all the irreversible control points in the metabolic system. The average PID of the enzyme seems to have little correlation with the expression or flux values. Clearly, the figure also shows a relationship between flux and the enzyme's resources, including placement in the overall pathway structure.

We make five key observations: (i) experiment comparison, (ii) subsystem analysis, (iii) enzyme comparisons (control point characteristics), (iv) PID and (v) normalization.

Experiment comparison. In Figure 4, PathwayPainter is used to compare six expression experiments. There are some notable global differences between these experiments. Both the *E.coli* and the worm expression sets show higher average expression change *C*, reflecting the changes in worm development and the effects of UV on *E.coli*. Conversely, the cell-cycle experiments show smaller average expression changes, reflecting the more constant state of housekeeping genes (e.g. metabolic pathway enzymes) within the cell. As expected in the diauxic shift experiment, the TCA cycle enzymes have high values for average expression changes *C* and expression fluctuations *E*; this substantiates previous observations that the change in medium for yeast increases the expression of TCA enzymes (28). In the yeast sporulation experiments, the positive and negative values in the average expression change capture the up- and down-regulation of different enzymes in the system.

Subsystem analysis. Different pathways or subsystems of central metabolism exhibit specific trends and characteristics. Figure 4 highlights one of these characteristics by coloring the arrows according to their flux variation *F*, with the highest values depicted in blue, median values in green and the lowest in red. From the schematic, as well as the table, one can see that the TCA cycle has all of the highest flux variation *F*, indicating that that the TCA cycle changes the most in metabolite processing. Biologically, this confirms the notion that the TCA cycle functions very differently depending on

environmental factors, such as aerobic and anaerobic conditions. The flux variation for glycolytic enzymes is near the average, except for triphosphate isomerase (5.3.1.1), which provides a shunt in the pathway. Similarly, the expression fluctuation, *E*, also correlates well with the division into subsystems; most of the highest values (>0.47) belong to the TCA cycle, the lowest values (<0.38) belong to the pentose phosphate pathway and the middle values belong to glycolysis. Expression fluctuation and flux variation clusters similarly to pathway divisions and correlate well with each other.

Enzyme comparisons. PathwayPainter also allows for multiple comparisons of specific enzymes. Expression variation is particularly evident at branch points and control points (where reactions are essentially irreversible). These points represent those enzymes with the greatest expression fluctuations (both C and E). For example, isocitrate dehydrogenase (1.1.1.42) and citric synthase (4.1.3.7), which have the two highest average expression changes, are both important control points in central metabolism. We performed additional analyses on other control point proteins and found that their average expression changes were very high as well. For example, phosphofructokinase (2.7.1.11), hexokinase (2.7.1.2) and pyruvate kinase (2.7.1.40) have values of 1.1, 0.6 and 0.6, respectively, representing an almost 100% increase compared to many of the other enzymes. In situations where the cell is perturbed, the response is reflected in the change of expression in the control point enzymes; in the worm development, *E.coli* response to UV, and yeast sporulation experiments, an increase in average expression change *C* at branch points (4.1,3.7, 1.1.1.37) is observed.

By comparing expression variability of an enzyme across multiple data sets, we have shown that many of the important metabolic control points are most variable in terms of expression. This variability indicates the intricate regulation of these enzymes.

Percentage identity. We found that sequence similarity *S* does not correlate strongly with either flux variation *F* or average expression change *C*. We conclude that sequence identity is not a predictor of expression or flux values.

Normalization. The expression-variability data in PathwayPainter can also be used to assess whether a group of genes can be applied to the normalization of microarray data between experiments in different organisms and under different experimental conditions. Presently, there are many efforts underway to determine robust methods to normalize microarray data through internal controls (47). For example, attempts have been made to establish normalization based on housekeeping genes, that is, those genes thought to be consistently expressed in the vast majority of conditions. We propose that the detailed study of the expression variability in metabolic enzymes shown by the PathwayPainter module can be useful in determining which enzymes could potentially be used as constants in microarray normalization approaches.

TreeViewer illustration

Informative pan-genomic analyses can be performed with the GeneCensus TreeViewer module. For example, the traditional ribosomal tree groups Gram-positive bacteria into one homogeneous group. However, further analysis using other types of trees subdivides them in informative ways. In particular, a number of the trees available in GeneCensus show that the bacteria *B.subtilis* and *Mycobacterium tuberculosis* tend to cluster independently of the other Gram-positive bacteria *M.genitalium* and *Mycoplasma pneumoniae*. This is a product of the radically differing size and gene compositions spanning the Gram-positive class. Further analysis of Gram-positive bacteria using the composition module shows that these two organisms have a high percentage of guanine as opposed to the other Gram-positive bacteria. Thus, while we may link them together due to the high peptoglycan content in their cellular walls (i.e. resulting in a Gram-positive stain), using the multiple modules in GeneCensus, we show that they differ radically in many other genomic properties.

CONCLUSION: GENECENSUS, A COMPARATIVE GENOMICS DATABASE AND TOOL

GeneCensus is part of the new generation of tools that help researchers navigate this post-genomic world. Overall, GeneCensus provides many levels of information. Whether one is interested in comparing genomes based on whole genome or pathway properties, looking at flux or expression information in specific organisms, or studying specific genes or pathways, the fluid incorporation of many different sources of data in GeneCensus allows researchers to view, or dynamically calculate, their information of interest, as well as investigate related data ranging from the amino acid level to multiple organism comparisons. Researchers can thus gain global system views and put their research interest into perspective within the vast sea of genomic data now available; through GeneCensus they can integrate varied data sources in an effort to actualize genome annotation. Given the high degree of false negatives and positives in many of the high-throughput genomic experiments, much of the data cannot, on its own merits, provide information for annotation. Fortunately, the noise that accompanies all of these experiments is not systematic, and the integration of the various data sets in, for instance, the phylogenetic analyses, allows users to cross-validate and improve the accuracy of their results.

ACKNOWLEDGEMENT

M.G. acknowledges support from the NIH (P01 GM54160-02).

REFERENCES

1. Wheeler,D.L., Church,D.M., Lash,A.E., Leipe,D.D., Madden,T.L., Pontius,J.U., Schuler,G.D., Schriml,L.M., Tatusova,T.A., Wagner,L. *et al.* (2002) Database resources of the National Center for Biotechnology Information: 2002 update. *Nucleic Acids Res.*, **30**, 13–16.
2. Wixon,J. and Kell,D. (2000) The Kyoto encyclopedia of genes and genomes—KEGG. *Yeast*, **17**, 48–55.
3. Frishman,D., Albermann,K., Hani,J., Heumann,K., Metanomski,A., Zollner,A. and Mewes,H.W. (2001) Functional and structural genomics using PEDANT. *Bioinformatics*, **17**, 44–57.
4. Overbeek,R., Larsen,N., Pusch,G.D., D'Souza,M., Selkov,E.,Jr, Kyrpides,N., Fonstein,M., Maltsev,N. and Selkov,E. (2000) WIT: integrated system for high-throughput genome sequence analysis and metabolic reconstruction. *Nucleic Acids Res.*, **28**, 123–125.

5. Delcher,A.L., Phillippy,A., Carlton,J. and Salzberg,S.L. (2002) Fast algorithms for large-scale genome alignment and comparison. *Nucleic Acids Res.*, **30**, 2478–2483.
6. Gelbart,W.M., Crosby,M., Matthews,B., Rindone,W.P., Chillemi,J., Russo Twombly,S., Emmert,D., Ashburner,M., Drysdale,R.A., Whitfield,E. *et al.* (1997) FlyBase: a *Drosophila* database. The FlyBase consortium. *Nucleic Acids Res.*, **25**, 63–66.
7. Mewes,H.W., Frishman,D., Guldener,U., Mannhaupt,G., Mayer,K., Mokrejs,M., Morgenstern,B., Munsterkotter,M., Rudd,S. and Weil,B. (2002) MIPS: a database for genomes and protein sequences. *Nucleic Acids Res.*, **30**, 31–34.
8. Dwight,S.S., Harris,M.A., Dolinski,K., Ball,C.A., Binkley,G., Christie,K.R., Fisk,D.G., Issel-Tarver,L., Schroeder,M., Sherlock,G. *et al.* (2002) Saccharomyces Genome Database (SGD) provides secondary gene annotation using the Gene Ontology (GO). *Nucleic Acids Res.*, **30**, 69–72.
9. Karp,P.D., Riley,M., Saier,M., Paulsen,I.T., Collado-Vides,J., Paley,S.M., Pellegrini-Toole,A., Bonavides,C. and Gama-Castro,S. (2002) The EcoCyc Database. *Nucleic Acids Res.*, **30**, 56–58.
10. Lin,J. and Gerstein,M. (2000) Whole-genome trees based on the occurrence of folds and orthologs: implications for comparing genomes on different levels. *Genome Res.*, **10**, 808–818.
11. Holms,H. (2001) Flux analysis: a basic tool of microbial physiology. *Adv. Microb. Physiol.*, **45**, 271–340.
12. Woese,C.R. and Fox,G.E. (1977) Phylogenetic structure of the prokaryotic domain: the primary kingdoms. *Proc. Natl Acad. Sci. USA*, **74**, 5088–5090.
13. Hegyi,H., Lin,J., Greenbaum,D. and Gerstein,M. (2002) Structural genomics analysis: characteristics of atypical, common, and horizontally transferred folds. *Proteins*, **47**, 126–141.
14. Hegyi,H. and Gerstein,M. (2001) Annotation transfer for genomics: measuring functional divergence in multi-domain proteins. *Genome Res.*, **11**, 1632–1640.
15. Stephanopoulos,G., Aristidou,A. and Nielsen,J. (1998) *Metabolic Engineering: Principles and Methodologies*, 1st Edn. Academic Press, San Diego.
16. Stephanopoulos,G. and Gill,R.T. (2001) After a decade of progress, an expanded role for metabolic engineering. *Adv. Biochem. Eng. Biotechnol.*, **73**, 1–8.
17. Schilling,C.H., Schuster,S., Palsson,B.O. and Heinrich,R. (1999) Metabolic pathway analysis: basic concepts and scientific applications in the post-genomic era. *Biotechnol. Prog.*, **15**, 296–303.
18. Schilling,C.H., Edwards,J.S. and Palsson,B.O. (1999) Toward metabolic phenomics: analysis of genomic data using flux balances. *Biotechnol. Prog.*, **15**, 288–295.
19. Sauer,U., Hatzimanikatis,V., Hohmann,H.P., Manneberg,M., van Loon,A.P. and Bailey,J.E. (1996) Physiology and metabolic fluxes of wild-type and riboflavin-producing *Bacillus subtilis*. *Appl. Environ. Microbiol.*, **62**, 3687–3696.
20. Nissen,T.L., Schulze,U., Nielsen,J. and Villadsen,J. (1997) Flux distributions in anaerobic, glucose-limited continuous cultures of *Saccharomyces cerevisiae*. *Microbiology*, **143**, 203–218.
21. Kanehisa,M., Goto,S., Kawashima,S. and Nakaya,A. (2002) The KEGG databases at GenomeNet. *Nucleic Acids Res.*, **30**, 42–46.
22. Tuyau,J.E., Sims,W. and Williams,R.A. (1984) The acid end-products of glucose metabolism of oral and other haemophili. *J. Gen. Microbiol.*, **130**, 1787–1793.
23. Kelly,D.J. (1998) The physiology and metabolism of the human gastric pathogen *Helicobacter pylori*. *Adv. Microb. Physiol.*, **40**, 137–189.
24. Pfeiffer,T., Sanchez-Valdenebro,I., Nuno,J.C., Montero,F. and Schuster,S. (1999) METATOOL: for studying metabolic networks. *Bioinformatics*, **15**, 251–257.
25. Schuster,S., Dandekar,T. and Fell,D.A. (1999) Detection of elementary flux modes in biochemical networks: a promising tool for pathway analysis and metabolic engineering. *Trends Biotechnol.*, **17**, 53–60.
26. Sherlock,G., Hernandez-Boussard,T., Kasarskis,A., Binkley,G., Matese,J.C., Dwight,S.S., Kaloper,M., Weng,S., Jin,H., Ball,C.A. *et al.* (2001) The Stanford Microarray Database. *Nucleic Acids Res.*, **29**, 152–155.
27. Spellman,P.T., Sherlock,G., Zhang,M.Q., Iyer,V.R., Anders,K., Eisen,M.B., Brown,P.O., Botstein,D. and Futcher,B. (1998) Comprehensive identification of cell cycle-regulated genes of the yeast *Saccharomyces cerevisiae* by microarray hybridization. *Mol. Biol. Cell*, **9**, 3273–3297.
28. DeRisi,J.L., Iyer,V.R. and Brown,P.O. (1997) Exploring the metabolic and genetic control of gene expression on a genomic scale. *Science*, **278**, 680–686.
29. Chu,S., DeRisi,J., Eisen,M., Mulholland,J., Botstein,D., Brown,P.O. and Herskowitz,I. (1998) The transcriptional program of sporulation in budding yeast. *Science*, **282**, 699–705.
30. Courcelle,J., Khodursky,A., Peter,B., Brown,P.O. and Hanawalt,P.C. (2001) Comparative gene expression profiles following UV exposure in wild-type and SOS-deficient *Escherichia coli*. *Genetics*, **158**, 41–64.
31. Reinke,V., Smith,H.E., Nance,J., Wang,J., Van Doren,C., Begley,R., Jones,S.J., Davis,E.B., Scherer,S., Ward,S. *et al.* (2000) A global profile of germline gene expression in *C. elegans*. *Mol. Cell*, **6**, 605–616.
32. Greenbaum,D., Jansen,R. and Gerstein,M. (2002) Analysis of mRNA expression and protein abundance data: an approach for the comparison of the enrichment of features in the cellular population of proteins and transcripts. *Bioinformatics*, **18**, 585–596.
33. Wootton,J.C. and Federhen,S. (1996) Analysis of compositionally biased regions in sequence databases. *Methods Enzymol.*, **266**, 554–571.
34. Engelman,D.M., Steitz,T.A. and Goldman,A. (1986) Identifying nonpolar transbilayer helices in amino acid sequences of membrane proteins. *Annu. Rev. Biophys. Biophys. Chem.*, **15**, 321–353.
35. Drawid,A. and Gerstein,M. (2000) A Bayesian system integrating expression data with sequence patterns for localizing proteins: comprehensive application to the yeast genome. *J. Mol. Biol.*, **301**, 1059–1075.
36. Qian,J., Stenger,B., Wilson,C.A., Lin,J., Jansen,R., Teichmann,S.A., Park,J., Krebs,W.G., Yu,H., Alexandrov,V. *et al.* (2001) PartsList: a web-based system for dynamically ranking protein folds based on disparate attributes, including whole-genome expression and interaction information. *Nucleic Acids Res.*, **29**, 1750–1764.
37. Kumar,A., Harrison,P.M., Cheung,K.H., Lan,N., Echols,N., Bertone,P., Miller,P., Gerstein,M.B. and Snyder,M. (2002) An integrated approach for finding overlooked genes in yeast. *Nat. Biotechnol.*, **20**, 58–63.
38. Ross-Macdonald,P., Coelho,P.S., Roemer,T., Agarwal,S., Kumar,A., Jansen,R., Cheung,K.H., Sheehan,A., Symoniatis,D., Umansky,L. *et al.* (1999) Large-scale analysis of the yeast genome by transposon tagging and gene disruption. *Nature*, **402**, 413–418.
39. Zhu,H., Klemic,J.F., Chang,S., Bertone,P., Casamayor,A., Klemic,K.G., Smith,D., Gerstein,M., Reed,M.A. and Snyder,M. (2000) Analysis of yeast protein kinases using protein chips. *Nature Genet.*, **26**, 283–289.
40. Balasubramanian,S., Schneider,T., Gerstein,M. and Regan,L. (2000) Proteomics of *Mycoplasma genitalium*: identification and characterization of unannotated and atypical proteins in a small model genome. *Nucleic Acids Res.*, **28**, 3075–3082.
41. Bertone,P., Kluger,Y., Lan,N., Zheng,D., Christendat,D., Yee,A., Edwards,A.M., Arrowsmith,C.H., Montelione,G.T. and Gerstein,M. (2001) SPINE: an integrated tracking database and data mining approach for identifying feasible targets in high-throughput structural proteomics. *Nucleic Acids Res.*, **29**, 2884–2898.
42. Gerstein,M. and Jansen,R. (2000) The current excitement in bioinformatics-analysis of whole-genome expression data: how does it relate to protein structure and function? *Curr. Opin. Struct. Biol.*, **10**, 574–584.
43. Senes,A., Gerstein,M. and Engelman,D.M. (2000) Statistical analysis of amino acid patterns in transmembrane helices: the GxxxG motif occurs frequently and in association with beta-branched residues at neighboring positions. *J. Mol. Biol.*, **296**, 921–936.
44. Gerstein,M., Lin,J. and Hegyi,H. (2000) Protein folds in the worm genome. *Pac. Symp. Biocomput.*, 30–41.
45. Das,R. and Gerstein,M. (2000) The stability of thermophilic proteins: a study based on comprehensive genome comparison. *Funct. Integr. Genomics*, **1**, 76–88.
46. Bussemaker,H., Li,H. and Siggia,E. (2000) Building a dictionary for genomes: identification of presumptive regulatory sites by statistical analysis. *Proc. Natl Acad. Sci. USA*, **97**, 10096–10100.
47. Yang,M.C., Ruan,Q.G., Yang,J.J., Eckenrode,S., Wu,S., McIndoe,R.A. and She,J.X. (2001) A statistical method for flagging weak spots improves normalization and ratio estimates in microarrays. *Physiol. Genomics*, **7**, 45–53.
48. Krebs,W.G. and Gerstein,M. (2000) The morph server: a standardized system for analyzing and visualizing macromolecular motions in a database framework. *Nucleic Acids Res.*, **28**, 1665–1675.