File Name:

Description:
Abstract – Rajdeep Das, 2003
Abstract

Comprehensive genomic analysis of thermophilic proteins and metabolic enzymes

Rajdeep Das
2003

We address the question of thermal stability of proteins in thermophiles through comprehensive genome comparison, focusing on the occurrence of salt bridges. One of the main conclusions of this work is that intra-helical salt bridges are more prevalent in thermophiles than mesophiles and thus suggest that they are an important factor stabilizing thermophilic proteins. We also analyze the sequence of metabolic enzymes. Analyses identified a novel isocitrate dehydrogenase in six organisms. We also analyze the distribution of protein folds in spirochetes. In a collaborative project, we develop a web-based tool, GeneCensus, to compare and integrate biological data.
Comprehensive genomic analysis of thermophilic proteins and metabolic enzymes

A Dissertation
Presented to the Faculty of the Graduate School
Of
Yale University
In Candidacy for the Degree of
Doctor of Philosophy

by
Rajdeep Das

Dissertation Director: Mark Gerstein

December 2003
Introduction

With the advent of new genome sequencing technology, genomic sequences from large number of organisms were made available. This opened up a new area of research known as functional genomics. Various studies had been conducted on genome sequences to understand the complex biological processes.

In chapter 1, we tried to understand the stability of proteins in thermophiles by analyzing the sequences in their proteomes. The most comprehensive part of our investigation is to understand the contribution of salt bridge interaction to thermal stability. We determined that the salt bridge interaction is prevalent in the helices of thermophilic proteins (1,2).

In chapter 2, we discussed sequence and structure-based analysis of central metabolic enzymes. Sequence divergence around the active site reflects the functional shift that an organism has undergone responding to a change in physiological environment. Based on this sequence analysis, we were to identify a novel isocitrate dehydrogenase present in six organisms that is similar to a cytotoxin from Yersinia pestis. (6)

In chapter 3, we described a comprehensive genome analysis of two spirochetes, T. pallidum and B. burgdorferi. First, we focused on the occurrence of protein structures in these organisms. We found that there were only a few spirochete-specific folds, relative to those in other types of bacteria. We also analyzed fold occurrence in genomes in relation to their fatty acid biosynthetic pathway. (3)

Finally in chapter 4, we discussed GeneCensus, a collaborative project to develop a web-based tool for comparative analyses of metabolic pathway in terms sequence divergence, relative flux and, gene expression (4,5).
Acknowledgements

It is a great pleasure to acknowledge the many people whose help and support have made this work possible.

First, I would like to thank my thesis advisor, Mark Gerstein, for introducing me to the new field of bioinformatics. He has been very supportive and encouraging during the six years of my PhD. In particular, I thank him for his great persistence during the early years of our efforts.

I also wish to thank Dr. Hedi Hegyi, Ted Johnson, Dov Greenbaum and Paul Bertone for their help. I am also thankful to Dr. Soumitra Basu for generously spending so much time with me, discussing various scientific issues. Also I would like to thank so many of my friends for making my stay in New Haven so enjoyable.

I would like to thank my family in India. I am forever grateful to my parents, sister and brother for their support and affection. And finally, I would like to thank my wife, Manmita for all her love and encouragement and for taking care of so many things through the last so many years.
References


6. Das R., Gerstein M., Comparative analyses of pathway proteins: Identification of novel isocitrate dehydrogenase in pathogens (under review)
Chapter 1

*Thermophilic protein stability from sequence comparison*

**Abstract**

Here we address the question of the thermal stability of proteins in thermophiles through comprehensive genome comparison, focusing on the occurrence of salt bridges. We compare a set of twelve genomes (from four thermophilic archaeons, one eukaryote, six mesophilic eubacteria, and one thermophilic eubacteria). Our results show that thermophiles have a greater content of charged residues than mesophiles, both at the overall genomic level and in alpha helices. Furthermore, we found that in thermophiles the charged residues in helices tend to be preferentially arranged with a 1-4 helical spacing and oriented so that intra-helical charge pairs reinforce with the helix dipole. Collectively, these results imply that intra-helical salt bridges are more prevalent in thermophiles than mesophiles and thus suggest that they are an important factor stabilizing thermophilic proteins. We also find that the proteins in thermophiles appear to be somewhat shorter than those in mesophiles. However, this later observation may have more to do with evolutionary relationships than with physically stabilizing factors. In all our statistics we were careful to take into account and control for biases, which could have, for instance, have arisen from repetitive or duplicated sequences. In particular, we repeated our calculation using a variety of random and directed sampling schemes. One of these involved making a "stratified sample," i.e. a representative cross-section of the genomes derived from a set of 52 orthologous proteins present roughly once in each
genome. For another sample, we focused on the subset of the 52 orthologs that have known 3D structures. This allowed us to determine the frequency of tertiary as well as main-chain salt bridges. Our statistical controls support our overall conclusion about the prevalence of salt bridges in thermophiles in comparison to mesophiles.

**Introduction**

*What are Thermophiles and how do their Proteins Achieve Stability?*

Thermophilic archaea and thermophilic eubacteria thrive in high temperatures. They live in places such as hot springs and deep-sea hydrothermal vents under extreme conditions. It is not well understood how proteins in thermophiles withstand the elevated temperatures at which these organisms live, which denature the proteins from mesophilic proteins. So far, several biophysical studies have been performed to identify the stability factors. These studies suggest about fifteen different physicochemical factors that contribute to thermostability, such as hydrogen bonding, hydrophobic internal packing, helix dipole stabilization, and salt bridge optimization, among others. These factors have been reviewed by several authors (25, 49, 53, 66, 9, 33, 59, 71, 72).

**The Salt Bridge as a Major Factor**

Since the study of thermostability of bacterial ferredoxins by Perutz (47), the three-dimensional structures of a large number of thermophilic proteins have been determined. These structures as well as structural information obtained through homology modelling have revealed that there is a strong correlation between the number of salt bridges and protein thermal stability (28,29, 38, 40, 52, 56, 4, 63, 73). A theoretical study by Elcock
(13) further supports this correlation by showing that since the hydration free energies of charged groups become less favorable at high temperatures, the unfavourable desolvation penalty incurred on forming the salt bridges is reduced in magnitude. Therefore, salt bridges should become more stable at elevated temperatures.

Studies of protein structures have shown that there are several ways in which salt bridges can stabilize proteins. Ion pair networks, helix stabilizing salt bridges, salt bridges buried in a hydrophobic core and surface salt bridges between two subunits are among the most frequently encountered types (48, 28, 40, 42, 56, 68, 75). These salt bridges can be broadly divided into two main classes:

(i) **Intra-helical or local**: This class arises from side chain interaction between charged residues in single helices. Biophysical studies have revealed that intra-helical EK, ER, DK, DR salt bridge pairs with the separations of 3 and 4 specifically stabilize helices (32, 57).

(ii) **Tertiary**: This class results from interaction between non-local charged residues of proteins. The class includes a large variety of salt bridges, such as inter-helical salt bridges, helix-sheet salt bridges and inter-subunit salt bridges.

Figure 1A illustrates local and tertiary salt bridges. We have studied the contribution of both types of salt bridges to protein thermostability. In addition to salt bridges discussed, other electrostatic interactions, such as charge-helix dipole interactions, can also stabilize
proteins. Previous research has shown that negatively charged residue can interact with the positive side of helix dipoles and thereby stabilize thermophilic proteins (46, 3, 66). In our study we also analyzed the results to see whether such interactions are important for protein thermostability.

**A Genome-wide Comprehensive Study: Our Goals and Strategy**

Most of the studies of protein thermostability referred to above involved analysis of only a few proteins. With the advent of fast DNA sequencing technology, complete genome sequences of several organisms are now available (11). As a result, it is now possible to comprehensively study all the proteins in an organism at the sequence level and compare them with the proteins of other organisms. We have performed similar analyses, comparing various aspects of protein structure, such as secondary structural composition and fold usage, between several recently sequenced genomes (20, 21, 22, 27). Similar studies have been carried out by other investigators (17, 54, 74, 41, 14, 1).

In this investigation our aim was to determine the importance of salt bridges, as well as other stability factors such as deamidation and protein length, for protein stability in thermophiles by comprehensive analysis of all protein sequences in their genomes and compare with mesophilic sequences. All of the factors that were studied here are shown in the Figure 1A. We analyzed the genomes of twelve organisms, of which four archaea and one hyper-thermophilic eubacteria are grouped together as thermophiles. The rest,
one eukarya and six eubacteria, are grouped together as mesophiles. These organisms are listed in Table 1.

Our overall strategy of salt bridge analysis is shown in Figure 1B-D. In the first step of our study we calculated the amino acid composition of all twelve genomes. In the second step we focused on the intra-helical salt bridges, calculating the frequency of putative salt bridge pairs in the helices of proteins. Since it is possible that the frequency of intra-helical salt bridges in a genome is biased due to sequence repeats and multiple paralogs specific to the organism, we analyzed a small set of orthologous proteins that are present in all the organisms and performed similar calculations on this set. In the third step we looked into the frequency of tertiary salt bridges in proteins in all the genomes. Within the above orthologous set, we calculated this frequency in only those proteins whose structures are known.

Results

Amino Acid Composition in the Genome, overall and just in Helices

The amino acid composition both in the entire genome and protein helices of the organisms are shown in Figures 2A and 2B.

Since secondary structures of most of the proteins in the genomes are unknown, we used predicted protein secondary structures to calculate the amino acid compositions of helices. Protein secondary structure were predicted using the program GOR(IV) (18, 19, 23). This is a well-established and commonly used method. It is statistically based, so
that the prediction for the state of a particular residue (say, Ala to be in a helix) is based on the frequency of that residue’s occurrence in the state in a database of solved structures (taking into account neighbors at ±1, ±2 and so forth). The GOR method uses only single sequence information compared to current state-of-the-art methods that incorporate multiple sequence information (36, 50, 51). While single sequence predictions are slightly less accurate than multiple sequence methods (65% versus 71%), we feel that using single sequence methods avoids various bias problems that can plague multiple sequence methods – i.e. we can only get multiple sequence information for a biased sample from each genome. Furthermore, we felt that the difference in accuracy between single and multiple sequence methods is not so vital in the overall context of our study, given our focus on averaged results.

It was observed that at both the overall genomic level and in helices, the amounts of glutamate, lysine and arginine (E, K, R) are higher in thermophilic proteins than in mesophilic proteins. This increase in charged residues suggests that in general we can expect to see more salt bridges in thermophiles than in mesophiles. Analysis of amino acid composition shows that the number of negatively charged aspartate residue remained almost the same in all the organisms.

**Use of Log of Odds (LOD) Calculation: Abundance of Local Salt bridges**

Because the content of charged residues in the helices is higher, we expect to see a greater number of intra-helical salt bridges in thermophiles than in mesophiles. In order
to see whether salt bridges are even more numerous than this elevated *a priori* "baseline", we calculated an odds ratio for all possible 400 amino acid pairs in helices. As described more fully in the caption to Figure 1B, this is essentially the observed number of occurrences of a given pair, divided by its expected number if there is no correlation -- e.g. the frequency of EK(3) divided by product of the individual E and K frequencies. Note that here the notation EK(3) implies an EK salt bridge pair with a separation of 3 residues; we used similar notations throughout the text. We then took the logarithm of this odds ratio, arriving at a log odds (LOD) value. LOD values are a measure of relative abundance for each pair in helices. Therefore, a higher LOD value for a particular pair would mean a higher frequency of that pair than other pairs in the genome. We calculated the LOD values for 400 amino acid pairs with a separation from 1 to 6. It is observed that for any salt bridge pair, its LOD value peaks at a separation of either 3 or 4 residues, indicating that these pairs probably represent intra-helical salt bridges, as suggested by the previous biophysical studies. As an illustration of this general result we plotted the LOD values of EK pairs at various separations in helices in Figure 3. Note that the LOD values of EK pairs peak at the separation of 3 for all organisms.

Results of these LOD value calculations shows that the LOD values for salt bridge pairs EK, ER and DR with the separation of 3 and 4 are generally higher in helices of thermophilic proteins than those in mesophilies. In order to see whether the charged residues in strand part of the protein sequences are correlate in a salt-bridge fashion, we performed a similar LOD-value calculation on the strands and compared the result with that for helices. Similarly, we calculated genome-wide LOD values for such pairs by
performing calculations on entire protein sequences. Comparison of the results showed that the LOD values for salt-bridge pairs are higher in helices than in other secondary structural elements and that this is true to a greater degree in thermophilic organisms. Our results thus imply that the charged amino acid residues are not only more numerous in thermophiles than mesophiles, but are also more highly correlated in helices of thermophilic proteins with a salt-bridge separation of 3 and 4.

**Correlation between Temperature and Salt Bridge Frequency**

We computed the LOD values for EK pairs with the separation of 3 and 4, and the result is shown in Figure 4A. The figure showed that the LOD values are higher for thermophiles than for mesophiles. From the figure it is also observed that among thermophilic species, LOD values increase from MT to OT commensurate with the steady increase in physiological temperatures from MT (65°C) to OT (98°C). This correlation of physiological temperature with intra-helical salt bridge frequency suggests that higher temperatures require a greater number of salt bridges to stabilize the helices in proteins.

**Helix Dipole Stabilization**

In our LOD calculation for helices we found that the values for EK(3) and EK(4) pairs are always higher than the corresponding values for KE pairs (data not shown). This variation of LOD value with the orientation of the charged pair is significant in terms of
charge-helix dipole interaction. Since negatively charged glutamate residue can stabilize a helix by interacting with the positive amino end of a helix dipole (46, 12, 3, 66), this observation indicates that thermophilic proteins gain stability from helix dipole stabilization.

Analyses of Control for Biases in the Statistics

The Problem of Bias in Comprehensive Genome-wide Statistics

While doing genome-wide surveys, care must be taken to assess the degree to which calculated statistics could be biased. In this regard, there are a number of specific issues relevant here. Firstly, sequence repeats, e.g. repetitive charged sequences in a set of thermophilic proteins, could skew the results. Secondly, unique protein sequences enriched in salt bridges could be highly duplicated in thermophilic genomes (forming large paralogous families), and this could also influence our results (see, for instance, figure 1C). A similar situation may arise involving only the sequences unique to mesophiles. We therefore needed to test the significance of LOD results and verify our conclusions with statistical controls and alternate procedures.

Rank Statistics

One technique to test the significance of our results is the use of rank statistics. Here the idea is that if we arrange the LOD values of all 400 pairs for each separation in an ordered list and observe that a particular pair -- EK(3), for example -- is at the top of the list, then we can infer that this pair is among the most over-represented in the helices of
the proteins for that organism. Table 3 summarizes the rank statistics for salt-bridge pairs that rank in the top 20 of a possible 400. The results showed that while the ranks of salt-bridge pairs vary greatly among all twelve genomes, the ranks of EK(3) pairs are generally higher for thermophiles in helices compared to mesophiles. MT is an exception to this general trend. In contrast, when the non-helical regions are considered, this distinction lessened.

*Random Resampling*

We directly addressed the problem of sequence repeats by a random resampling procedure. We simulated thermophilic and mesophilic genomes by randomly drawing proteins from two large pools of thermophilic and mesophilic sequences. From these simulated genomes we calculated the LOD values for charged amino acid pairs in helices. Figure 4B showed the distribution of these values for the EK(3) pair. Note the distinct difference in the distributions. Statistical tests are performed to estimate the degree of significance of this difference, and it is found that given the width of the distribution, the chance that any mesophile would have a LOD value similar to a thermophile is less than 5% (for EK(3) and EK(4)). This implies that our LOD calculation results are significant in a statistical sense. (These calculations are described in more detail in the figure caption.)

*Stratified Resampling using Orthologs*

Another way of minimizing biases is through the use of stratified sampling procedures (2). The idea here can most easily be described in terms of a demographic comparison of
a particular characteristic between populations -- for example, height in northern versus southern populations. It is possible that the overall population could be fractionated into further subdivisions using another parameter, potentially linked to height, say age (old vs. young). Our initial analysis of salt bridge statistics is analogous to computing the average height over the entire population irrespective of age. However, the possibility that one population has more of a certain age group than another could potentially skew the statistics (e.g. Northerners are older and taller). To compensate for such bias in the sample we could take a representative sample from every age group and calculate the average height for that strata. This is what we did in stratified sampling to study the salt bridge abundance.

Our strata are sets of orthologous proteins present in each of the 12 genomes. Orthologous proteins evolved from common ancestral genes and usually share the same structure and function (15). Statistics obtained from sets of orthologous proteins can be considered to be relatively free from bias arising from sequence repeats or large paralogous families. In our study we selected 52 sets of orthologous proteins (listed on our website). Our ortholog selection strategy is explained in detail in Figure 1C. It is derived using the cluster-of-orthologous groups (COGs) approach (64). We used only COGs for which we could determine a single best representative for each genome, and we extended the initial COGs assignments (currently 8 genomes) to include all twelve genomes in our study.
On our set of 52 orthologous proteins analyses were performed similar to those performed on the entire genomes. Composition analysis showed a similar trend of increasing amounts of charged residues from mesophile to thermophile, as is observed in the overall genome analysis (Figure 2C). Note in the figure that the hyperthermophilic eubacteria Aquifex aeolicus moved closer in position to the other eubacteria, perhaps indicating that some exclusively archaeal paralogous family is heavily weighted with charged residues. Likewise, we calculated LOD values for our set of 52 orthologs. The results for important salt bridge pairs are shown in Table 2C. Although the LOD values for EK(3) had decreased for both thermophiles and mesophiles, thermophiles still maintain higher average LOD values for EK(3), EK(4), DR(3) and ER(4). This result is important; in spite of involving only 52 groups of proteins, the stratified resampling comparisons shows that the putative salt bridge frequency is clearly higher in thermophiles than mesophiles.

**Study of Tertiary Salt Bridges**

So far, our study of salt bridges had focused only on intra-helical salt bridges. Moreover, these statistics depended on accuracy in the prediction of protein secondary structures. Therefore, to complement our conclusions on intra-helical salt bridge abundance, we studied the tertiary salt bridges in thermophilic and mesophilic proteins of known structure. Here, we followed a procedure similar to that of Schuler & Margalit (58). Since any such study of tertiary salt-bridges requires a knowledge of detailed protein 3D structure, which is unknown for most proteins in these genomes, we used the strategy, schematized in Figure 1D. Where possible, we mapped the sequence of a protein with a
known 3D structure onto a corresponding orthologous group of sequences to identify the putative tertiary salt bridges in the new sequences. This approach rests on the idea that since orthologous proteins have conserved structures, knowledge of one protein structure can be extended to others in the same group. More specifically, we took query sequences from each of our 52 orthologous groups of proteins and compared them with the PDB structural database by pairwise sequence comparison (43, 62). This resulted in a list of 18 PDB structures that map onto corresponding orthologous groups. As listed in Table 4, we classified these 18 orthologous groups of known structure into three categories: (i) Ribosomal proteins, (ii) Amino-acyl tRNA synthetases, and (iii) Other proteins.

Using the strategy outlined in Figure 1D, we obtained rough estimates of the number of salt bridges for each protein in the 18 orthologous groups of known structure. Table 4 showed some summary statistics based on these numbers. It shows that for two categories, ribosomal proteins and tRNA-synthetases, thermophiles have somewhat more tertiary salt bridges than mesophiles. For proteins in the "other" category, the difference between thermophiles and mesophiles is not as significant.

Other Stabilizing Factors

So far we have discussed the role of salt-bridge interactions in thermophilic proteins. It should be noted here that since structures for most of the proteins are unknown, it is not possible for us to study the contribution of other factors, such as the effect of hydrophobic internal packing on protein thermal stability. However, in addition to salt-
bridges, we also studied the effect of two other factors on thermostability of proteins and compared their results with that of the salt-bridges: (i) deamidation (ii) protein length.

**Deamidation**

It is shown that glutamines and asparagines can undergo a deamidation reaction that leads to instability (7). Therefore reduction in the amount of these two amino acids can stabilize proteins. In our amino acid composition study, we observed that compared to mesophiles, the amounts of glutamine (N) and asparagine (Q) are lower in thermophiles. Furthermore, we noticed that among hydrophobic amino acids the amount of valine and isoleucine is higher in thermophiles. In this context it should be noted that amount of proline, which is believed to contribute thermal stability in proteins (45, 26, 73) does not exhibit any bias, and remains almost same both in thermophiles and mesophiles.

**Protein Length and Thermal Stability: The Contradictory Position of Aquifex aeolicus**

It has been argued that shorter protein length increases the compactness of the protein and reduces flexibility. A biophysical study by Nagi & Regan (44) had suggested that there is an inverse correlation between loop length and protein stability. Thompson & Eisenberg (65) in a recent study put forward a thermodynamic argument supporting this correlation, and showed that thermophilic proteins have a higher tendency towards shorter loops and fall off more rapidly than the mesophilic counterparts, by comparing homologous proteins from the genomes of a large number of organisms. In our study, we analyzed the
sequence length distribution of proteins of all organisms to understand how protein length is related to thermostability. Our results, shown in Figure 5A, indicates that the length distributions of thermophiles indeed fall off more rapidly than those of mesophiles. Furthermore, when we fit curves to the length distribution of just thermophilic or just mesophilic proteins, we found that the median (and mode) length is less in the thermophiles than in the mesophiles (Figure 5B). This result is true for both the genome overall and for our sample of 52 orthologs. Therefore, our "first pass" results on protein length appear to support the notion that the proteins in thermophiles are shorter than those in mesophiles.

However, when we looked at the sequence lengths in further detail, we found that the situation is more complicated. The distribution of protein lengths for *Aquifex aeolicus*, a hyperthermophilic eubacteria, is more similar to those of mesophilic eubacteria than to the other thermophilic organisms, which are all archaea. Furthermore, yeast appeared to have distinctly longer proteins than those in either of the prokaryotic kingdoms. The distribution of protein lengths therefore appears to be more related to kingdom than to environment -- reflecting historical contingency rather than chemical necessity. This result is illustrated in Figure 6A, which shows how "phylogenetic composition" of proteins with a given length became progressively less archaenal and more eukaryotic as we moves to longer proteins. This result is further borne out in the table of Figure 6B, where it can be seen how average protein lengths correlates with kingdom. In this table we include average protein length for *C. elegans*, the other known eukaryotic genome, to illustrate that long sequences are characteristic of other eukaryotes beside yeast.
Conclusion

From the comparison of our results on amino-acid composition and LOD statistics, we argue that the occurrence of excess intra-helical salt bridges in thermophiles reflects its origin in two factors. First, thermophiles have a higher content of charged amino acids than the mesophiles. Secondly, these charged residues are more preferentially arranged with a 1,4 salt-bridge spacing in thermophilic helices than in mesophilic helices. Since the results determined with orthologous groups of proteins are similar to those obtained with whole genomes, we infer that sequence repeats and/or paralogous sequence families did not skew the observed abundance of intra-helical salt bridges in thermophiles. Our results also show that thermophilic proteins have higher occurrence of tertiary salt bridges than mesophilic proteins. Thus we conclude that the salt-bridge interactions probably play a vital role in stabilizing thermophilic proteins. Our study also shows that, in addition to salt-bridges, there are other factors that can contribute to protein thermal stability. Reduction of deamidation by decreasing the amounts of glutamine and asparagine in proteins probably also confers stability to thermophilic proteins. Though we examined the contribution of protein sequence lengths to stability, we found that they are only loosely connected with the protein thermostability. Therefore, among all the three factors that we studied here, we found that while the extent of contribution to thermostability varied for each factor, salt-bridge contribution is most consistent with the increasing physiological temperatures and one of the most important factors for protein thermostability.
Reference


hyperthermophile Thermotoga maritima: possible determinants of protein stability. Biochemistry 36(20), 6009-16.


35. Kawarabayasi, Y., Sawada, M., Horikawa, H., Haikawa, Y., Hino, Y., Yamamoto, S., Sekine, M., Baba, S., Kosugi, H., Hosoyama, A., Nagai, Y., Sakai, M., Ogura, K., Otsuka, R., Nakazawa, H., Takamiya, M., Ohfuku, Y., Funahashi, T., Tanaka,


Figure 1.
Figure 3.
A. 

![Graph showing LOD values for different groups with bars indicating variations.]

B. 

![Graph showing frequency distribution of LOD values with two bell curves, one for Thermophilic and one for Mesophilic.]

Figure 4.
Figure 5.
Figure 6
<table>
<thead>
<tr>
<th>Organism</th>
<th>Category</th>
<th>Genome ID</th>
<th># of Proteins</th>
<th>Physiological condition</th>
</tr>
</thead>
<tbody>
<tr>
<td><em>Pyrococcus horikoshii</em> (Strain OT3)</td>
<td>archaea</td>
<td>OT</td>
<td>2061</td>
<td>98°C, anaerobe</td>
</tr>
<tr>
<td>(Kawarabayasi et al., 1998)</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><em>Aquifex aeolicus</em> (Deckert et al., 1998)</td>
<td>eubacteria, gram negative</td>
<td>AA</td>
<td>1522</td>
<td>95°C</td>
</tr>
<tr>
<td><em>Methanococcus janaschii</em> (Bult et al., 1996)</td>
<td>archaea</td>
<td>MJ</td>
<td>1735</td>
<td>85°C, anaerobe</td>
</tr>
<tr>
<td><em>Archaeoglobus fulgidus</em> (Klenk et al., 1997)</td>
<td>archaea</td>
<td>AF</td>
<td>2409</td>
<td>83°C, anaerobe</td>
</tr>
<tr>
<td><em>Methanobacterium thermoautotrophicum</em> (Smith et al., 1997)</td>
<td>archaea</td>
<td>MT</td>
<td>1869</td>
<td>65°C, anaerobe</td>
</tr>
<tr>
<td><em>Haemophilus influenzae</em> (Fleischmann et al., 1995)</td>
<td>eubacteria, gram negative</td>
<td>HI</td>
<td>1680</td>
<td>mesophilic temp.</td>
</tr>
<tr>
<td><em>Mycoplasma genitalium</em> (Fraser et al., 1995)</td>
<td>eubacteria, gram positive</td>
<td>MG</td>
<td>470</td>
<td>mesophilic temp.</td>
</tr>
<tr>
<td><em>Mycoplasma pneumoniae</em> (Himmelreich et al., 1996)</td>
<td>eubacteria, gram positive</td>
<td>MP</td>
<td>677</td>
<td>mesophilic temp.</td>
</tr>
<tr>
<td><em>Helicobacter pylori</em> (Tomb et al., 1997)</td>
<td>eubacteria, gram negative</td>
<td>HP</td>
<td>1590</td>
<td>mesophilic temp.</td>
</tr>
<tr>
<td><em>Escherichia coli</em> (Blattner et al., 1997)</td>
<td>eubacteria, gram negative</td>
<td>EC</td>
<td>4288</td>
<td>mesophilic temp.</td>
</tr>
<tr>
<td><em>Synechocystis sp.</em> (Kaneko et al., 1996)</td>
<td>cyanobacteria</td>
<td>SS</td>
<td>3168</td>
<td>mesophilic temp.</td>
</tr>
<tr>
<td><em>Saccharomyces cerevisiae</em> (Goffeau et al., 1997)</td>
<td>eukaryote, fungus</td>
<td>SC</td>
<td>6218</td>
<td>mesophilic temp.</td>
</tr>
</tbody>
</table>

Table 1.
### A. LOD values for helix

<table>
<thead>
<tr>
<th>Spacing</th>
<th>Pair</th>
<th>AF</th>
<th>MJ</th>
<th>MT</th>
<th>OT</th>
<th>AA</th>
<th>EC</th>
<th>HI</th>
<th>HP</th>
<th>MG</th>
<th>MP</th>
<th>SC</th>
<th>SS</th>
</tr>
</thead>
<tbody>
<tr>
<td>3</td>
<td>EK</td>
<td>57</td>
<td>59</td>
<td>36</td>
<td>64</td>
<td>61</td>
<td>44</td>
<td>45</td>
<td>50</td>
<td>43</td>
<td>35</td>
<td>36</td>
<td>36</td>
</tr>
<tr>
<td></td>
<td>ER</td>
<td>48</td>
<td>25</td>
<td>42</td>
<td>39</td>
<td>36</td>
<td>38</td>
<td>27</td>
<td>36</td>
<td>27</td>
<td>37</td>
<td>25</td>
<td>32</td>
</tr>
<tr>
<td></td>
<td>DR</td>
<td>45</td>
<td>33</td>
<td>53</td>
<td>48</td>
<td>48</td>
<td>48</td>
<td>40</td>
<td>32</td>
<td>-3</td>
<td>20</td>
<td>30</td>
<td>40</td>
</tr>
<tr>
<td>4</td>
<td>EK</td>
<td>61</td>
<td>60</td>
<td>50</td>
<td>63</td>
<td>62</td>
<td>44</td>
<td>48</td>
<td>46</td>
<td>43</td>
<td>42</td>
<td>45</td>
<td>46</td>
</tr>
<tr>
<td></td>
<td>ER</td>
<td>47</td>
<td>27</td>
<td>48</td>
<td>46</td>
<td>43</td>
<td>36</td>
<td>31</td>
<td>23</td>
<td>33</td>
<td>25</td>
<td>27</td>
<td>33</td>
</tr>
<tr>
<td></td>
<td>DR</td>
<td>38</td>
<td>38</td>
<td>48</td>
<td>44</td>
<td>44</td>
<td>44</td>
<td>47</td>
<td>48</td>
<td>56</td>
<td>60</td>
<td>36</td>
<td>42</td>
</tr>
</tbody>
</table>

### B. LOD values for entire genome

<table>
<thead>
<tr>
<th>Spacing</th>
<th>Pair</th>
<th>AF</th>
<th>MJ</th>
<th>MT</th>
<th>OT</th>
<th>AA</th>
<th>EC</th>
<th>HI</th>
<th>HP</th>
<th>MG</th>
<th>MP</th>
<th>SC</th>
<th>SS</th>
</tr>
</thead>
<tbody>
<tr>
<td>3</td>
<td>EK</td>
<td>43</td>
<td>47</td>
<td>23</td>
<td>14</td>
<td>48</td>
<td>27</td>
<td>27</td>
<td>38</td>
<td>27</td>
<td>21</td>
<td>22</td>
<td>22</td>
</tr>
<tr>
<td></td>
<td>ER</td>
<td>37</td>
<td>17</td>
<td>30</td>
<td>22</td>
<td>25</td>
<td>26</td>
<td>22</td>
<td>27</td>
<td>15</td>
<td>25</td>
<td>14</td>
<td>21</td>
</tr>
<tr>
<td></td>
<td>DR</td>
<td>16</td>
<td>8</td>
<td>18</td>
<td>1</td>
<td>10</td>
<td>18</td>
<td>12</td>
<td>9</td>
<td>2</td>
<td>4</td>
<td>-1</td>
<td>11</td>
</tr>
<tr>
<td>4</td>
<td>EK</td>
<td>40</td>
<td>40</td>
<td>28</td>
<td>14</td>
<td>40</td>
<td>21</td>
<td>24</td>
<td>31</td>
<td>19</td>
<td>25</td>
<td>23</td>
<td>24</td>
</tr>
<tr>
<td></td>
<td>ER</td>
<td>31</td>
<td>17</td>
<td>31</td>
<td>-7</td>
<td>27</td>
<td>24</td>
<td>19</td>
<td>12</td>
<td>17</td>
<td>14</td>
<td>12</td>
<td>21</td>
</tr>
<tr>
<td></td>
<td>DR</td>
<td>9</td>
<td>4</td>
<td>12</td>
<td>24</td>
<td>5</td>
<td>16</td>
<td>17</td>
<td>22</td>
<td>12</td>
<td>13</td>
<td>4</td>
<td>13</td>
</tr>
</tbody>
</table>

### C. LOD values for 52 COG proteins

<table>
<thead>
<tr>
<th>Spacing</th>
<th>Pair</th>
<th>AF</th>
<th>MJ</th>
<th>OT</th>
<th>MT</th>
<th>AA</th>
<th>EC</th>
<th>HI</th>
<th>HP</th>
<th>MG</th>
<th>MP</th>
<th>SC</th>
<th>SS</th>
</tr>
</thead>
<tbody>
<tr>
<td>3</td>
<td>EK</td>
<td>48</td>
<td>57</td>
<td>55</td>
<td>38</td>
<td>61</td>
<td>38</td>
<td>36</td>
<td>45</td>
<td>40</td>
<td>23</td>
<td>41</td>
<td>26</td>
</tr>
<tr>
<td></td>
<td>ER</td>
<td>38</td>
<td>-27</td>
<td>14</td>
<td>24</td>
<td>36</td>
<td>30</td>
<td>43</td>
<td>17</td>
<td>-5</td>
<td>51</td>
<td>-5</td>
<td>21</td>
</tr>
<tr>
<td></td>
<td>DR</td>
<td>33</td>
<td>65</td>
<td>37</td>
<td>31</td>
<td>45</td>
<td>47</td>
<td>29</td>
<td>29</td>
<td>6</td>
<td>20</td>
<td>37</td>
<td>35</td>
</tr>
<tr>
<td>4</td>
<td>EK</td>
<td>58</td>
<td>60</td>
<td>57</td>
<td>26</td>
<td>62</td>
<td>41</td>
<td>65</td>
<td>35</td>
<td>46</td>
<td>24</td>
<td>62</td>
<td>33</td>
</tr>
<tr>
<td></td>
<td>ER</td>
<td>29</td>
<td>15</td>
<td>39</td>
<td>34</td>
<td>43</td>
<td>37</td>
<td>38</td>
<td>9</td>
<td>-2</td>
<td>12</td>
<td>38</td>
<td>5</td>
</tr>
<tr>
<td></td>
<td>DR</td>
<td>54</td>
<td>37</td>
<td>76</td>
<td>23</td>
<td>32</td>
<td>55</td>
<td>52</td>
<td>44</td>
<td>60</td>
<td>70</td>
<td>52</td>
<td>73</td>
</tr>
</tbody>
</table>

**Table 2.**
### HELIX

<table>
<thead>
<tr>
<th>Sep</th>
<th>Pair</th>
<th>AF</th>
<th>MJ</th>
<th>MT</th>
<th>OT</th>
<th>AA</th>
<th>EC</th>
<th>HI</th>
<th>HP</th>
<th>MG</th>
<th>MP</th>
<th>SS</th>
<th>SC</th>
</tr>
</thead>
<tbody>
<tr>
<td>3</td>
<td>EK</td>
<td>4</td>
<td>5</td>
<td>-</td>
<td>4</td>
<td>4</td>
<td>9</td>
<td>7</td>
<td>7</td>
<td>13</td>
<td>-</td>
<td>-</td>
<td>13</td>
</tr>
<tr>
<td>3</td>
<td>ER</td>
<td>10</td>
<td>-</td>
<td>13</td>
<td>18</td>
<td>-</td>
<td>12</td>
<td>-</td>
<td>14</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>3</td>
<td>DR</td>
<td>13</td>
<td>-</td>
<td>-</td>
<td>13</td>
<td>12</td>
<td>8</td>
<td>10</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>10</td>
</tr>
<tr>
<td>4</td>
<td>EK</td>
<td>5</td>
<td>9</td>
<td>12</td>
<td>9</td>
<td>7</td>
<td>12</td>
<td>9</td>
<td>13</td>
<td>11</td>
<td>15</td>
<td>10</td>
<td>10</td>
</tr>
<tr>
<td>4</td>
<td>ER</td>
<td>11</td>
<td>14</td>
<td>-</td>
<td>14</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>4</td>
<td>DR</td>
<td>-</td>
<td>-</td>
<td>13</td>
<td>13</td>
<td>-</td>
<td>9</td>
<td>10</td>
<td>11</td>
<td>9</td>
<td>7</td>
<td>11</td>
<td>18</td>
</tr>
<tr>
<td>3</td>
<td>DK</td>
<td>9</td>
<td>13</td>
<td>8</td>
<td>8</td>
<td>16</td>
<td>2</td>
<td>5</td>
<td>6</td>
<td>11</td>
<td>10</td>
<td>4</td>
<td>9</td>
</tr>
<tr>
<td>4</td>
<td>DK</td>
<td>10</td>
<td>-</td>
<td>16</td>
<td>19</td>
<td>-</td>
<td>6</td>
<td>11</td>
<td>17</td>
<td>-</td>
<td>13</td>
<td>15</td>
<td>9</td>
</tr>
</tbody>
</table>

### GENOME

<table>
<thead>
<tr>
<th>Sep</th>
<th>Pair</th>
<th>AF</th>
<th>MJ</th>
<th>MT</th>
<th>OT</th>
<th>AA</th>
<th>EC</th>
<th>HI</th>
<th>HP</th>
<th>MG</th>
<th>MP</th>
<th>SS</th>
<th>SC</th>
</tr>
</thead>
<tbody>
<tr>
<td>3</td>
<td>EK</td>
<td>4</td>
<td>5</td>
<td>9</td>
<td>3</td>
<td>3</td>
<td>3</td>
<td>5</td>
<td>3</td>
<td>9</td>
<td>-</td>
<td>7</td>
<td>-</td>
</tr>
<tr>
<td>3</td>
<td>ER</td>
<td>6</td>
<td>-</td>
<td>4</td>
<td>11</td>
<td>11</td>
<td>4</td>
<td>7</td>
<td>9</td>
<td>-</td>
<td>-</td>
<td>9</td>
<td>-</td>
</tr>
<tr>
<td>4</td>
<td>EK</td>
<td>4</td>
<td>8</td>
<td>11</td>
<td>9</td>
<td>6</td>
<td>12</td>
<td>10</td>
<td>6</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>4</td>
<td>ER</td>
<td>9</td>
<td>-</td>
<td>7</td>
<td>14</td>
<td>15</td>
<td>3</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
</tbody>
</table>

Table 3
Table 1. List of Organisms.

The table lists the twelve organisms whose sequences are used in calculation. Column three shows the two-letter abbreviations for the genomes of the organisms listed in the first column. The fourth column lists the number of open reading frames found in the genome. The last column shows the physiological temperatures of thermophiles. For mesophiles we referred to 'mesophilic temperatures', which range from 10 to 45 °C. Data-files of predicted proteins were taken from the websites referred to in the papers above, with the exception of OT for which predicted proteins were from the analysis of Suckow et al. (1998).

Table 2: LOD values (values are in percentages).

Table 2A and 2B show LOD values of salt bridge pairs in helix and genome. Since in helices salt bridge pairs at the separation of 3 and 4 are known to stabilize proteins, we have listed their LOD values separately. LOD values of the salt bridge pairs in strands are not shown here, as it is obvious from the whole genome results. Table 2C lists the LOD values of the ion pairs for 52 orthologous proteins. Note that LOD values for the slat bridge pairs remained high even in the small set of 52 orthologous proteins.

Table 3. Rank statistics of salt bridge pairs.

The ranks of other salt bridge pairs (ER(3), EK(4) and DR(3)) were not remarkably different between thermophiles and mesophiles. A similar study on the predicted strand
sequence did not show any significant ranking for salt bridge pairs (results are not shown).

**Table 4. Statistics with tertiary salt bridges.**

This table summarizes the results of tertiary salt bridge counts. Column one shows the COG identifiers for the orthologous groups that are selected. Second column gives the functional class for each of this group and the fourth column lists the PDB identifiers for homologous proteins with known structures. Third column represents our category. For every protein, we calculated the average number of salt bridges present in thermophiles and in mesophiles as shown in columns 5 and 6. Column 7 shows the difference between the two. Based on this difference we set up a scoring scheme that qualitatively describes the relative abundance of tertiary salt bridges. If the difference is > 1.0, a positive (+) sign is assigned showing a predominance of salt bridges in thermophiles; if the difference is < -1.0, a negative (-) sign is assigned showing a predominance of salt bridges in mesophiles; for any other value of difference, no sign is assigned to either thermophiles or mesophiles, thus showing no bias for salt bridges. Note that in the two main categories (ribosomal proteins and tRNA synthetases thermophiles) thermophiles have a higher amount of tertiary salt bridges than mesophiles.

**Figure 1A.** Three factors that are studied for their contribution to protein thermostability.

(i) Local salt bridges (ii) tertiary salt bridges (iii) protein length.
**Figure 1B.** Determining the position of local (intra-helical) and tertiary salt bridges.

The box in the figure represents a protein sequence with known structure and each $\pm$ combination connected with dotted lines represents a salt bridge pair as observed in the structure. Since the EK pair involves an interaction between the charged residues in two separate secondary structural elements, it is defined as a tertiary salt bridge. Similarly the DK pair occurring within a helix is termed a local salt bridge. Using this definition we calculated the LOD values for the intra-helical amino acid pair as follows. The odds ratio $R$ for any particular pair, say XY, at a separation $i$ is defined by,

$$
R[XY(i)] = \frac{\text{Observed number of occurrences for XY pair separated by } i}{\text{Expected number of occurrences for XY pair separated by } i}
$$

The LOD value is the log (base 10) of the ratio. The observed number of occurrences for any salt bridge pair is the simple count for that pair in a genome. The expected number of occurrences for that pair is calculated as follows: Given the frequencies of two amino acids X and Y in the helices as $P(X)$ and $P(Y)$, the probability of an XY pair occurring, assuming the occurrence of X and Y is completely uncorrelated, is:

$$
P(X,Y) = P(X) \, P(Y)
$$
If the total number of all amino acid pairs is N, the expected number of occurrences of the XY pair is calculated as,

\[ N(\text{XY}) = N \cdot P(X) \cdot P(Y) \]

**Figure 1C. Method of Stratified Resampling based on Orthologous Relationship.**

This diagram illustrates our strategy of stratified resampling. We selected only those sequences from each genome (assumed size ~2000 proteins) that were linked by simple orthologous relationships. That is, only those corresponding proteins, which are present in all twelve genomes, are selected, and then only a single representative is actually counted. The dashed lines (-) in the figure show the sequences that are missing for any orthologous group and are thus discarded from our calculation. Using this procedure, we have been able to filter out the effect of paralogous sequences as well as sequence repeats that may bias our results.

More specifically, to identify orthologs we followed a five-step procedure:

(i) We started with the COGs classification at the NCBI (Tatusov, *et al.*, 1997). This currently contains 864 orthologous groups that are present in varying degrees in 8 of the first genomes sequenced (a subset of the twelve genomes used in this study).

(ii) We restricted our attention initially to the 110 COGs present in all 8 genomes.

(iii) Then we dealt with the issue of those COGs represented by multiple proteins in certain genomes (i.e. paralogs). To compensate for this effect, we chose only those
COGs that had a maximum of ten sequences in total. In the few cases when we had paralogs, to pick a best representative, we consulted the dendograms on the COGs website.

(iv) To enlarge a COGs cluster to the 12 genomes used here, we performed pairwise sequence comparison using the FASTA program (version 2.0) (Lipman & Pearson, 1985) where the COGs sequences were used as queries against the four additional genomes not part of the original COGs study (i.e. AA, OT, AF, and MT). We used an ‘e-value’ threshold of 0.01 in these comparisons. The e-value describes the number of errors per query expected in a single database scan, so a value of 0.01 means that about one out of a hundred cluster linkages will be in error.

(v) Finally, we kept only those COGs that had easy-to-find members in the extra four genomes.

Application of the whole procedure resulted in the list of 52 COGs that we used in our study. A subset of 18 of these had homologs in the PDB structure databank and was used for the tertiary salt bridge study. The rhombus in the last column of the figure indicates these.

**Figure 1D.** Determination of tertiary salt bridges by an Indirect Method of Structure Mapping
To determine the positions of the salt bridges in a protein of unknown structure, where possible, we mapped its sequence onto a homologous protein of known structure in the PDB. All the salt bridges in the protein with known structure were determined by a program that takes coordinates of a protein and gives a list of hydrogen bonds as an output occurring in it (Gerstein, 1992). The list of hydrogen bonds considered here involved only side-chain/side-chain and side-chain/main-chain interactions between amino acids, as the main-chain/main-chain hydrogen bonds are mostly involved in forming secondary structural elements. Next we aligned all twelve sequences in each orthologous group with the corresponding PDB sequence by multiple sequence alignment using CLUSTALW (Higgins et al., 1996). Then for every salt bridge pair in the PDB protein, a corresponding amino acid pair was determined in the similar position in other proteins. It has been observed that in some proteins, the amino acid pair corresponding to a salt bridge is conserved, whereas in others it is replaced either by a non-ionic pair or by a complementary salt bridge pair.

Figure 2. Amino acid composition in genome, helix and 52 orthologous proteins

In the figures the blackened area represents the portion of charged residues E, D, K and R in a helix. This area increases from mesophiles to thermophiles. On the contrary, the amounts of amine residues, N and Q decrease in thermophilic helices. Also note that among hydrophobic groups (AILV) there is an increase in the contents of L and V. Each column shows the total content of any amino acid for all 12 genomes; thus the totals add up to 100%.
Figure 3. LOD values of the EK pair in helix as a function of separation

LOD values for EK-pair peak at a separation of either 3 or 4, suggesting that the pair at these positions represents salt bridge pair.

Figure 4A. LOD values of EK salt bridge pairs with separation of 3 and 4

LOD values increase with the increase in physiological temperatures shown along the horizontal axis. For mesophiles, they are indicated by a range from 10 to 45 °C.

Figure 4B. Distribution of LOD values of EK(3) for randomly generated mesophilic and thermophilic genomes

The figure shows the distribution curves of EK(3) LOD values for randomly generated thermophilic and mesophilic genomes. The difference of the two means of the distribution is $|\Delta| = 0.18$. The sample variance for thermophiles is $s_x^2 = 0.0038$ and that for mesophiles, $s_y^2 = 0.0059$. We performed a standard double-blind experiment to test the significance of the difference of means. We calculated the Z score as follows:

$$Z = \frac{\langle X \rangle - \langle Y \rangle}{\sigma}$$

where $\langle X \rangle$ and $\langle Y \rangle$ are the means of thermophilic and mesophilic distributions, respectively, $\sigma^2 = \frac{s_x^2}{n_x} + \frac{s_y^2}{n_y}$, and $n_x$ and $n_y$ are the number of observations for each distribution (500 here). Results show that the probability that the two distributions will have same mean is less than 5 percent.
**Figure 5A.** Length distribution of proteins in twelve organisms

We used an extreme value distribution for the fit curve shown by the bold line:

Frequency at any protein length $x$ is given by, $y = \exp(c-b(x-a)) - \exp(-b(x-a))$ where $a = 211.0$, $b = 0.007142$, and $c = 0.2277$. Note that some sequences longer than 983 amino acids are not shown in the graph. Two letter abbreviations are defined in Table 1. It is evident from the figure that at shorter protein lengths thermophiles exceed the fit curve while mesophiles are below it, but at the longer protein length mesophiles exceed the fit curve and thermophiles go under.

**Figure 5B.** Comparison of thermophilic and mesophilic fit curves for length distribution both for overall genome sequences and orthologous proteins.

We used the same extreme value distribution for the fit curve as in Figure 6. Only the fit curves are shown here.

**Figure 6A.** Length distribution of proteins in terms of overall percentage composition at various lengths

Amount of protein at various protein lengths is shown for different genomes. In the figure vertical axis represents fraction of total amount of proteins present at various protein length for all the twelve organisms. Percentage content of proteins with longer protein length increases in yeast and decreases in archaea.

**Figure 6B.** Average protein length in twelve organisms
In eukaryote category we included average protein length of protein in *C. elegans* genome (CESC, 1998). Shaded genomes represent the thermophiles. Overall averages for each category are given on the top of every category-column. Note that the average protein length for archaea is shorter than that for either of the other forms of life.
Chapter 2

Sequence and structure-based genomic analyses of central metabolic enzymes

Abstract

We have analyzed enzymes of central metabolism from a large number of genomes (55) in terms of sequence variability around active sites. We find that, for most of the enzymes, active-sites sequences are more highly conserved than full-length sequences. However, for three TCA-cycle enzymes, active-site sequences are considerably more diverged than full-length sequence. In particular, we have been identified a novel isocitrate dehydrogenase in six pathogens that has very low sequence similarity around its active site with respect to other organisms. Detailed sequence-structure analysis indicates that while the active site structure of isocitrate dehydrogenase is similar between pathogens and non-pathogens, the unusual sequence divergence probably results from an extra domain at the N-terminus. This domain has a leucine-rich motif similar to Yersinia pestis cytotoxin and therefore may have additional pathogenic functions.

Introduction

Pathway enzymes have been a topic of wide scientific interest in both the pre- and post-genomic eras. In the post-genomic era, pathways have been studied using sequence information in many ways, and several metabolic databases have been constructed (1-6). In contrast to overall sequence conservation, sequence variability of an enzyme near the
functional site may reflect a functional shift. This functional shift can occur in many ways, such as a change in the binding affinity of a substrate or intermediate (7). Previous studies analyzed protein families in terms of sequence-structure relationships (8). In the present investigation, we analyzed 18 enzymes of the central metabolic pathways (i.e. glycolysis, pentose phosphate pathway, and TCA cycle) in 55 organisms in terms of sequence variability around active sites. The goal of our study is to identify organisms with an unusual active site that may indicate modified metabolic characteristics for those organisms. We also attempted to characterize the novel proteins identified in terms of sequence motifs and model structure comparison. A ribosomal tree, shown in Figure 1A, lists all the organisms studied in the analyses: all of the three major kingdoms of life i.e., archaea, eubacteria and eukarya are represented.

Methods

Analyses consisted of two steps.

**Step I: Identification of novel enzyme by active site comparison**

The first step is the identification of organisms with modified enzymes. We began by selecting one representative structure from the PDB for each of the 14 enzymes (9). A list of the PDB structures is chosen shown in Table 1 for the four enzymes discussed in this investigation. Information regarding their active site residues is obtained from the literature. Table 1 also shows the active site residues that are considered in this study; we defined an active site environment as all the residues that fall within a radius of $10 \text{ Å}$ from active site residues. An average of approximately 90 residues fell within this radius. They were identified using the program MOLEMAN (10). Once the residues in the active
sphere were identified in each representative structure, they were mapped onto the sequences from other organisms by multiple sequence alignment. Finally, the pairwise sequence similarity was calculated between all enzymes for these active site residues. This part of the analysis resembles the 3-dimensional cluster analyses used earlier to study a group of protein families (11). The overall strategy of calculating active site similarity is illustrated in Figure 2. In addition to active sites, we also computed pairwise full-length sequence similarity. The basis of our study is the general understanding that residues that form the active site environment are under selective pressure, since they are critical to the enzymatic function, and are therefore likely to be more conserved than the residues in the balance of the sequence. In order to identify sequences in which the active site is modified, we calculated a ratio of active site similarity to full-length similarity. Since the active site residues are more likely to be conserved than the residues in the remaining sequences, the ratio of the two quantities, R, is expected to be more than 1. However, if the residues in the active site sphere are modified, the pairwise active site similarity should be lower than the pairwise full-length similarity, and R will be less than 1. If a sequence is highly diverged from other sequences in an enzyme class, all the pairwise R-values corresponding to that sequence will be less than 1. So among all the values of ratio matrix, those R-values will have a distribution that is clustered in one side, allowing for easy identification of those organisms.
Step II: Characterization of novel enzyme: Identification of new sequence motifs and Structural comparison

Based on our initial analysis we determined the enzymes for which we observed modified active sites and associated organisms. The step-II we analyzed these enzymes in terms of sequence-based characterization.

Result and Discussion

The enzymes in central metabolism vary greatly in average pairwise sequence identity. Some of the enzymes in the pathway are highly conserved with an average sequence identity ~60% and some are less conserved. This is shown in Figure 1B. For most of the central metabolic enzymes, we found that the sequences around an active site are more conserved than the rest of the sequences. However, there are three TCA cycle enzymes and one glycolytic enzyme that have large sequence variation near active sites; these we discuss below.

Isocitrate dehydrogenase

The most interesting result is observed for isocitrate dehydrogenase. This is a TCA cycle enzyme, which catalyzes the following reaction:

\[
\text{Isocitrate} + \text{NADP}^+ \rightarrow \text{2-Oxoglutarate} + \text{CO}_2 + \text{NADPH}
\]

We compared 33 sequences of isocitrate dehydrogenase from different organisms. The distribution of R shows that six organisms have a modified isocitrate dehydrogenase;
therefore the distribution of R fell in the region where the value is less than 1. Figure 3A shows the distribution of the R-values and the corresponding list of six organisms. They are all known pathogens. The probability of pathogenic sequences as a sample of the entire distribution of R-values, having a mean R-value of 0.52, is tested using normal statistics. The result shows that the probability p is very significant i.e. less than 0.00001.

**Sequence Motifs**

Finally, we analyzed the sequence of the modified isocitrate dehydrogenase in terms of sequence motifs. It is interesting to note that all the pathogenic sequences are ~300 residues longer than the class sequence length. It is possible that there may be a domain addition to all of the six pathogenic sequences. From the multiple sequence alignment, it is observed that the extra stretch of sequence occurs on the N-terminal side of the enzyme. When we searched for similar sequences in the sequence database, we observed a 90 residue-long stretch of sequence in this extra domain with high sequence similarity to a leucine rich domain of an effector, YopM of Yersinia pestis (the bubonic plague pathogen). The stretch of sequences is extremely rich in leucine. Although the anti-host function of this effector is unknown, YopM is believed to be an important cytotoxin for bacterial virulence (12). Figure 4 shows the sequence alignment of the sequence YopM with the N-terminal domain of four pathogenic isocitrate dehydrogenases. Other evidence that the modified isocitrate dehydrogenase may have additional function comes from the observation that *M. tuberculosis* has two isocitrate dehydrogenase sequences Rv3339c and Rv0066c. First one is a standard isocitrate dehydrogenase sequence and second one is the modified sequence. Therefore, the presence of the two sequences may
indicate an extra function of modified sequence. Interestingly, phylogenetic clustering of
the organisms based on protein sequences also groups the pathogenic organisms in one
cluster as shown in Figure 5.

It should be noted here that six sequences annotated to be isocitrate dehydrogenase in
GenBank is based on sequence similarity of Azotobacter vinelandii’s isocitrate
dehydrogenase. It is entirely possible that these sequences in six organisms do not have
any isocitrate dehydrogenase activity and their true functions can only be determined
experimentally. Therefore our assertion that they represent modified isocitrate
dehydrogenase can only be validated by experimental evidence.

*Malate dehydrogenase*

\[
\text{Malate} + \text{NAD}^+ \rightarrow \text{Oxaloacetate} + \text{NADH}
\]

Malate dehydrogenase sequences from 35 organisms are analyzed. The distribution of R
for malate dehydrogenase is shown in Figure 3C. This figure shows that five organisms--
Pseudomonas aeruginosa, Clostridium acetobutylicum, Pyrococcus abyssi,
Methanococcus jannaschii and Pyrococcus horikoshii-- all have modified malate
dehydrogenase sequences. However, previous studies showed that Methanococcus
jannaschii has two sequences for malate dehydrogenase, MJ0490 and MJ1425 (13).
Although two sequences are annotated as malate dehydrogenase, MJ1425 is linked to
methylpterin biosynthesis. Therefore it is likely that the enzyme in these organisms has a
dual role: catalyzing conversion of malate and biosynthesis of the cellular component..
The probability of a sample of sequences of this size having a mean R-value of 0.45 (mean R for the modified organisms) is tested using normal statistics and it is less than 0.00001.

_Fumarate reductase_

Fumarate reductase, usually associated with anaerobic respiration, catalyzes the following reaction:

\[
\text{Fumarate} + \text{FADH}_2 \rightarrow \text{Succinate} + \text{FAD}
\]

This enzyme consists of three subunits: A, B, and C. Subunit-A binds to fumarate, and B and C bind to the Fe-S cluster and membrane, respectively. Comparison of the active site residue cluster shows that while the Fe-S cluster-binding B and flavin-binding C subunits are similar across organisms, the binding site of the fumarate in subunit A is different for _H. pylori_ and _C. jejuni_ compared with the rest of the organisms. These two mucosal pathogens are known to have fumarate respiration (14, 15). The distribution of R-values is shown in Figure 3B. Structural difference in the active site of the two pathogenic fumarate reductases is analyzed by modeling the active site structures of the two organisms.

For both enzymes (isocitrate dehydrogenase and fumarate reductase), we concluded that the sequence variability observed near the functional site probably has a role in altering
the binding affinity of the substrate in the active site and is possibly one of the mechanisms for controlling of enzymatic function in the organisms. The number of sequences in fumarate reductase is 10. The probability of a sample of sequences of this size having a mean R-value of 0.94 (mean R for the modified organisms) is tested using normal statistics and it is a marginal 0.065. This could be a result of small size of the dataset.

In addition, we also observed a few glycolytic enzymes with modified active sites. Fructose-1,6-biphosphatase is one such enzyme that catalyzes the conversion of fructose-1,6-biphosphatase to fructose-6-phosphate. We observed that four organisms (L. lactis, C. acetobutylicum, B. subtilis and S. aureus) have very diverged sequence around active site. Similarly, archaeal glyceralde-3-phosphate dehydrogenases have very modified sequence in active site. However, we are not able to characterize these enzymes further either structure- or sequence-wise.

**Statistical test to verify the possibility of bad alignment**

It should be pointed out that the average pairwise sequence similarity is quite low for some of the enzymes we studied as shown in Figure 1B. In particular we observed significant sequence variation around the active site in two enzymes (EC #: 11142 and 11137) that have low average pairwise sequence similarity. It is possible that the observed sequence variation in the active site for these two enzymes can be the result of bad sequence alignment of the sequences with low homology. We therefore performed a statistical test to verify this possibility. We calculated the average pairwise sequence
similarity for a large number of clusters each comprising ~90 residues. If the alignment is bad, average sequence similarity values calculated for active site cluster and that of any random cluster should be very similar. For example, in *V. cholerae*, an organism with modified isocitrate dehydrogenase, the average pairwise sequence identity is 21% for random clusters and that of active site cluster is 27%. Although the values are low, calculation shows that active site identity of 27% has a P-value of 1.2x10^{-12} of occurring by chance. Similarly for malate dehydrogenase, P-values are less than 0.01 for the organisms with a modified active site. Therefore, we conclude that the residues in the active site sphere most likely represented an active site in the enzyme and did not represent random residues as would be expected in the case of a bad alignment. However, in the case of fumarate reductase, *H. pylori* and *C. jejuni* have average active site similarity of 36% that is less than average pairwise identity of random clusters (40%) and had a P-value of less than 2.87x10^{-7} to occur by chance. Clearly, the smaller number of organisms (i.e. 10) in the analyses may have biased the results for fumarate reductase. Biological significance of this result is not clear.

**Conclusion**

In this investigation we analyzed the sequence variation of enzymes around active sites in a large number of organisms. For most of the organisms, the results show that sequences are highly conserved around the active site. However, there are three enzymes in the TCA cycle for which we observed that the sequence is extremely divergent. Most interestingly, six pathogenic organisms have unique isocitrate dehydrogenase that has sequence similarity to a cytotoxin in *Yersinia pestis* that is linked to bacterial pathogenicity.
Reference

1. Bono H, Ogata H, Goto S, Kanehisa M. Reconstruction of amino acid
biosynthesis pathways from the complete genome sequence. Genome Res 1998;
8(3): 203-10.

2. Galperin MY, Koonin EV. Functional genomics and enzyme evolution.
Homologous and analogous enzymes encoded in microbial genomes. Genetica
1999; 106(1-2): 159-70.

3. Dandekar T, Schuster S, Snel B, Huynen M, Bork P. Pathway alignment:
343 Pt 1: 115-24.

4. Forst CV, Schulten K. Evolution of metabolisms: A new method for the
1999; 6(3-4): 343-60.

5. Ogata H, Fujibuchi W, Goto S, Kanehisa M. A heuristic graph comparison
algorithm and its application to detect functionally related enzyme clusters.


7. Naylor GJ, Gerstein M. Measuring shifts in function and evolutionary opportunity
223-33.


Figure 1A
Active Sphere
PID

Full Length
PID

R

Frequency
R
Figure 3

A.

less than 40% conserved

Campylobacter jejuni
Mycobacterium leprae
M tuberculosis
Neisseria meningitidis
Vibrio cholerae
Xylella fastidiosa

B.

Campylobacter jejuni
Helicobacter pylori

C.

Pseudomonas aeruginosa
Clostridium acetobutylicum
Pyrococcus abyssi
Methanococcus jannaschii
Pyrococcus horikoshii
Figure 4.
Figure 1A. Phylogenetic tree based on rRNA

Phylogenetic tree of the 55 organisms studied. Tree is based on rRNA sequences. Organisms represent all three kingdoms of life i.e. archaea, bacteria and eukarya.

1B. Average pairwise similarities are shown for the enzymes in central metabolic pathway. For heteromeric enzymes, the subunit that has active site is considered for the pairwise similarity calculation. We did not calculate the values for the enzymes for which we did not have enough data or information regarding active sites or we encountered other problems as shown by asterisk sign (*).

Figure 2. Strategy of structure based sequence comparison

We determined the residues that fell in the sphere of 10Å° radius around the active residues using MOLEMAN (10). Then we mapped those residues onto sequences from other organisms, using multiple sequence alignment. The red-colored residues in the figure in the sequence are the residues that fell in the active site sphere of the known structure; these are mapped onto the sequences of other organisms. Corresponding residues in other sequences are also shown in red. We computed pairwise sequence similarity matrices for the active sphere residues and for the overall full-length sequence proteins as shown in the figure.

Figure 3. Distribution of R-values for three enzymes in TCA cycle

The x-axis shows R-values and the y-axis shows frequency. Overall distribution is shown in the blue line and the red line represents the distribution that corresponds to the six organisms with modified sequences. The dashed bar in the figure represented the line
where R is 1; the active site similarity is equal to the overall sequence similarity. A, B and C, respectively, represented the distributions corresponding to isocitrate dehydrogenase, fumarate reductase and malate dehydrogenase.

Figure 4. Sequence alignment of novel isocitrate dehydrogenase with the cytotoxin from Yersinia Pestis

Organisms are abbreviated as follows: Cje: Campylobacter jejuni; Hpy: Helicobacter pylori; Vch: Vibrio cholerae; Mle: Mycobacterium leprae; Mtu: M. tuberculosis; Xfa: Xylella fastidiosa.

Figure 5. Phylogenetic clustering based on isocitrate dehydrogenase sequences of the organisms

Although organisms clustered differently in the ribosomal tree, the six most diverged isocitrate dehydrogenase sequences cluster together.

Table 1. Active residues of the enzymes

Table shows the active residues identified for the enzymes. EC# and PDB ID show the enzyme classification number of the enzyme and PDB ID of the representative structure.
Chapter 3

A Study of the Protein Structures, Functions and Metabolic Pathways in Treponema pallidum and Borrelia burgdorferi

ABSTRACT

We performed a comprehensive genome analysis on two spirochetes, T. pallidum and B. burgdorferi. First, we focused on the occurrence of protein structures in these organisms. We found that there are only a few spirochete-specific folds, relative to those in other types of bacteria. The most common fold, by far, in the spirochetes is the P-loop NTP hydrolase, followed by the TIM barrel. These folds also happen to be amongst the most multifunctional of the known folds. We also surveyed the membrane-protein structures in T. pallidum and found a notable large family with twelve transmembrane (TM) helices, reflecting the prevalence of 12-TM transporters in bacteria. We found that the lipid biosynthesis pathway is absent from the spirochetes. Among five folds that are used in the biosynthetic pathway, only NAD(P) binding Rossmann fold has a very low abundance in the genome compared to other bacterial genomes.

INTRODUCTION

In the last five years there have been a number of organisms, spanning all three forms of life, whose genomes have been sequenced. Among these organisms, Borrelia burgdorferi (4) and Treponema pallidum (5), two pathogenic prokaryotic spirochetes, have generated a lot of interest among the scientific community. In this investigation we analyzed the occurrence of various protein-folds in these two genomes by comparing them to the
structural domains in SCOP (8). Such analyses of structure and fold families are important, as structural knowledge is essential to understand the precise molecular mechanism of the proteins. Comparing two or more complete, evolutionarily related genomes in terms of domains also allowed us to define the essential building blocks of a functioning organism, while the differences can highlight the unique, often pathogenesis-related features of a pathogen microbe.

Lipid metabolism in these organisms is also intriguing. Although spirochetes, when compared to bacteria, are relatively rich in lipids (22% of their dry weight), they lack all the enzymes of fatty acid biosynthesis. Studies have shown that they do not synthesize long chain fatty acids and need an external supply of fatty acids for their survival (7). In this investigation we also discussed our how the folds used in lipid metabolism are distributed in the spirochetes.

FOLD USAGE IN SPIROCHETES

We performed a structural analysis of the two genomes, using a combination of FastA and PSI-blast searches with the domain sequences of SCOP 1.39, a structural database, as queries, against the two genomic sequences (1, 9). We found structural matches for less than one-fourth of the ORFs, more precisely 289 of the total of 1638 ORFs in B. burgdorferi and 252 of the total of 1031 ORFs in T. pallidum.

The most common folds are largely the same as in other prokaryotes, as shown in Table 1. The most common fold by far in both organisms is the P-loop containing NTP
hydrolase, while the second most common fold, the TIM-barrel, has only about one-third as many matches. Both of these folds belong to the alpha/beta fold class, the most common structural class in bacteria. However, in contrast to most bacteria, including E. coli, there are also two abundant alpha-helical folds in the list, the alpha-alpha superhelix and the long helix oligomers. As shown in the table, in E. coli these two folds rank very low (62 and 70, respectively). Interestingly, both folds rank also quite high in yeast and in mycoplasmas (data not shown). Two of the folds, the class II aaRS synthetases and the Ferredoxin-like fold, belong to the alpha+beta class, while the rest of the folds in the table are also members of the alpha/beta class. None of the ten most abundant folds in either genome belong to the all-beta class, a highly abundant class in higher organisms.

A more general functional distribution of the most common folds is shown in Figure 1. It shows the 5 most commonly occurring folds in the Swissprot database and the number of different - mostly enzymatic - functions they are associated with (6). The most versatile fold, the TIM-barrel has as many as 16 different functions associated with it (it is commonly believed that this number is achieved as a result of convergent evolution, i.e. the fold evolved independently several times). Interestingly, the large majority of the folds, about 70% have only one enzymatic function associated with them (note however, that this analysis is based on single-domain proteins with enzymatic functions).

In B. burgdorferi and T. pallidum we identified 123 and 126 folds, respectively. The two organisms share more than a hundred of their folds, 106 altogether. As shown in Figure 2, there are only a few spirochete-specific folds: compared with E. coli and yeast T.
*pallidum* has no specific folds, and only two compared with *E. coli* and *B. subtilis*. These are the 6-bladed beta-propeller (2sil) and a specific Cysteine-proteinase (1gcb). (The codes in parentheses are PDB identifiers with the corresponding chains (11). The Borrelia also has two special folds, one of them an endonuclease (1smn, chain A), the other an outer-surface protein (1osp, chain O), occurring exclusively in this organism, in as many as 9 copies, which probably indicated an important pathogenic role for this fold.

**Folds in fatty acid biosynthesis pathway**

Fatty acid biosynthetic pathway is absent in spirochetes. There are five folds that are used in the fatty acid biosynthesis. Theses are NAD(P)-binding Rossmann fold, thiolase-like fold, barrel-sandwich hybrid, ClpP/crotonase fold, 6-phosphogluconate dehydrogenase fold. Our fold distribution results show that only NAD(P)-binding Rossmann fold which is abundant in bacterial genomes is characteristically under represented in spirochetes. However for other four folds, their occurrences are comparable to other bacterial genomes.

Figure 3 shows the phylogenetic tree of 10 completely sequenced organisms, including the two spirochetes (*T. pallidum* and *B. burgdorferi*). The tree is constructed based on the fold content of the organisms, in good agreement with the traditional phylogenetic trees derived from pairwise sequence similarity of their ribosomal genes. It shows that *T. pallidum* and *B. burgdorferi* are the most closely related to each other among the 10 organisms.
We also analyzed the two genomes for transmembrane helix content. The distribution of the ORFs with transmembrane helices with respect to the number of helices per ORF in *T. pallidum* is shown in Figure 4. Interestingly, the distribution shows local maximums at 7 and 12. The significance of the 7-transmembrane proteins in higher organisms is well documented, and the local maximum at 12 indicated the important role of the 12-TM transporter proteins in bacteria.

REFERENCES


1998. Protein Data Bank (PDB): database of three-dimensional structural
54(1 (Pt 6)), 1078-84.
Figure 1.
Figure 4.
<table>
<thead>
<tr>
<th>Fold Name</th>
<th>ID</th>
<th>Number in T. pallidum</th>
<th>Number in B. burgdorferi</th>
<th>Number in E. coli</th>
</tr>
</thead>
<tbody>
<tr>
<td>P-loop NTP hydrolase</td>
<td>3.24</td>
<td>43</td>
<td>67</td>
<td>72(4)</td>
</tr>
<tr>
<td>TIM Barrel</td>
<td>3.1</td>
<td>14</td>
<td>20</td>
<td>93(1)</td>
</tr>
<tr>
<td>like Ribonuclease H</td>
<td>3.47</td>
<td>13</td>
<td>12</td>
<td>50(6)</td>
</tr>
<tr>
<td>class II aaRS synthetases</td>
<td>4.61</td>
<td>13</td>
<td>10</td>
<td>15(28)</td>
</tr>
<tr>
<td>alpha-alpha superhelix</td>
<td>1.91</td>
<td>11</td>
<td>6</td>
<td>6(62)</td>
</tr>
<tr>
<td>like Ferredoxin</td>
<td>4.34</td>
<td>11</td>
<td>6</td>
<td>82(2)</td>
</tr>
<tr>
<td>adenine NT alpha hydrolase</td>
<td>3.17</td>
<td>7</td>
<td>10</td>
<td>16(27)</td>
</tr>
<tr>
<td>FAD/NAD(P)-binding domain</td>
<td>3.4</td>
<td>6</td>
<td>4</td>
<td>38(7)</td>
</tr>
<tr>
<td>Periplasmic binding proteins</td>
<td>3.82</td>
<td>7</td>
<td>9</td>
<td>36(9)</td>
</tr>
<tr>
<td>like II</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>long helix oligomers</td>
<td>1.105</td>
<td>6</td>
<td>6</td>
<td>5(70)</td>
</tr>
</tbody>
</table>

Table 1.
Table 1. The 10 most frequently occurring folds in T. pallidum and B. burgdorferi. The numbers indicate how many times the folds are found in these two organisms, and also in E.coli. The numbers in parentheses for E.coli indicate the ranking of each fold in this organism.

Figure 1. The functional distribution of the most diverse folds occurring in Swissprot. The X-axis indicates the number of functions, and the Y-axis denoted the number of folds that have been associated with that many functions. A schematic is shown for each of the 5 functionally most diverse folds.

Figure 2. Two Venn-diagrams, showing the number of shared and unique folds among T.pallidum, E.coli and yeast (on the left-hand side), and among T.pallidum, E.coli and Bacillus subtilis.

Figure 3. A phylogenetic tree of 10 completely sequenced organisms, including T.pallidum and B. burgdorferi. The tree is constructed based on the pairwise similarity between the fold contents of the genomes.

Figure 4. Distribution of the ORFs with transmembrane helices in T.pallidum with respect to the number of helices per ORF. The highest number of TM-helices is found to be 19, detected only in one ORF.
Chapter 4

GeneCensus: A web-based tool for comparative analyses of Metabolic pathway in terms of relative flux and sequence divergence

Abstract

GeneCensus is a new database tool that focused 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 http://GeneCensus.org. The system concentrates on two types of comparisons: (1) trees based on the sharing of generalized protein families between genomes, and (ii) whole pathway analyses in terms of activity levels.
Introduction

Advances in sequencing technology have created the opportunity to perform large-scale genome comparison. Presently there are many systems focusing on specific types of comparisons for many genome (COG(1), PENDANT(2), or KEGG(3), WIT(4), MUMmer(5)). We present here a new prototype tool 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 based upon different characteristics of the genomes (1). These characteristics include bread statistics, such as fold and gene content and amino acid composition. The trees that we provide can be compared against other information, and dynamically reconfigure based on different genomic characteristics. Our second viewer, Pathway Painter, provides the user with an extensive comparison of genomes in terms of their mRNA expression, flux (2) and percent identity (PID), in three major metabolic pathways: TCA, glycolysis and pentose phosphate pathway. 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. The overall view of GeneCensus is shown in Figure 1.

Feature Overview: TreeViewer

Comparative TreeViewer is outlined graphically in Figure 2. It 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 were built, includes: 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 includes: taxonomy, fold composition, COGs, AT-content, genome size and superfamily occurrence. Thus, we can 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 & Gerstein (1). The trees that are currently available can be subdivided into the following self-explanatory sections:

(i) Ribosome - These trees are built comparing the similarity of the ribosomal RNA. This traditional method (5) 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.

(ii) Folds. These trees are built based on the presence or absence of folds in different organisms, as determined by Hegyi et al. (6). 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.

(iii) **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 was collected using a similar approach to Hegyi & Gerstein (7).

(iv) **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.

(v) **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.

(vi) **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.

(vii) **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 macroview—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 the information for each enzyme in the form of: (i) relative flux values (normalized, absolute and standard deviation); (ii) average and standard deviation of gene expression change (DNA and cDNA arrays); (iii) PID (between orthologous enzymes in the pathway for multiple organisms). We present the flux and PID information for E. coli, S cerevisia, B subtilis, H influenzae, H 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.

Relative Flux Analysis

Relative flux, a measure of the ratio at which metabolites are processed to become output along various branches, is calculated at a steady state (15). Determination of relative 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 percent 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 cDNA microarrays and Affymetrix Gene Chips. In particular, we have collected various microarray datasets 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 *C. 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 percent identity 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:

(i) **Raw unscaled values** *R* as available from the various sites as either copies per cell, log₂(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* 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 GeneChip and SAGE data sets and is on 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. \( E = \sqrt{\langle (R(i,t) - \langle R(i,t) \rangle)^2 \rangle} \)). Note that enzymes consistently expressed under most conditions will show minimal standard deviations that are closely correlated between experiments.

**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 an illustration of the utility of the system by presenting some biologically relevant qualitative conclusions that can be extracted from the database.

In Figure 4 illustrates the scientific conclusions one can derive from PathwayPainter, comparing the variation in expression, flux, and percent identity of enzymes over many different experiments and organisms. We present some of our findings here, as well as showing how this 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) relative flux variation $F$, which indicates the standard deviation of relative 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.

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

(i) 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 had high values for average
expression changes C and expression fluctuations E; this substantiated previous observations that the change in medium for yeast increases the expression of TCA enzymes (3). In the yeast sporulation experiments, the positive and negative values in the average expression change captured the up- and down-regulation of different enzymes in the system.

(ii) Subsystem Analysis

Different pathways or subsystems of central metabolism exhibits 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; pentose phosphate metabolism also had F values. Similarly, the expression fluctuation, E, also correlates well with the division into subsystems; most of the highest values (> 0.47) belonged to the TCA cycle, the lowest values (<0.38) belongs to the pentose phosphate pathway, and the middle values belongs to glycolysis. Expression fluctuation and flux variation clusters similarly to pathway divisions and correlate well with each other.
(iii) 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) had 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 show that many of the important metabolic control points are the most variable in terms of expression. This variability indicates the intricate regulation of these enzymes.

(iv) 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.
(v) **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 (4). 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 grouped 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 *M. tuberculosis* tend to cluster independently of the other gram positive bacteria *M. genitalium* and *M. 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 show that the two bacteria under study 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 datasets in, for instance, the phylogenetic analyses, allows the users to cross-validate and improve the accuracy of their results.
References


Figure 2

TREEVIEWER

- Database Links
- View Selection
- Taxonomy Links
- Organism Links
- Yale Links

OTHER VIEWS

- COG Tree with Genome Composition
- Fold Tree vs. Ribosomal Tree
<table>
<thead>
<tr>
<th>P-Way Enzyme</th>
<th>Stat</th>
<th>Yeast Cell Cycle Alpha</th>
<th>Yeast Cell Cycle Shift</th>
<th>Yeast Sporulation</th>
<th>E. coli UV radiation</th>
<th>Homodimerization</th>
<th>Average</th>
<th>Flux Stdv</th>
<th>Avg % Identity</th>
</tr>
</thead>
<tbody>
<tr>
<td>3.1.1.49</td>
<td>STDV</td>
<td>1.28</td>
<td>1.30</td>
<td>1.30</td>
<td>1.26</td>
<td>N/A</td>
<td>1.28</td>
<td>.15</td>
<td>65</td>
</tr>
<tr>
<td></td>
<td>AVG</td>
<td>-.16</td>
<td>-.02</td>
<td>-.22</td>
<td>.26</td>
<td>N/A</td>
<td>.04</td>
<td>.65</td>
<td></td>
</tr>
<tr>
<td>1.1.1.44</td>
<td>STDV</td>
<td>.32</td>
<td>.31</td>
<td>.26</td>
<td>N/A</td>
<td></td>
<td>1.59</td>
<td>.65</td>
<td></td>
</tr>
<tr>
<td></td>
<td>AVG</td>
<td>.23</td>
<td>.23</td>
<td>.50</td>
<td>N/A</td>
<td></td>
<td>.74</td>
<td>.65</td>
<td></td>
</tr>
<tr>
<td>5.1.3.1</td>
<td>STDV</td>
<td>.14</td>
<td>.25</td>
<td>.30</td>
<td>N/A</td>
<td></td>
<td>7.6</td>
<td>.65</td>
<td></td>
</tr>
<tr>
<td></td>
<td>AVG</td>
<td>.18</td>
<td>.25</td>
<td>.50</td>
<td>N/A</td>
<td></td>
<td>4.1</td>
<td>.65</td>
<td></td>
</tr>
<tr>
<td>5.3.1.6</td>
<td>STDV</td>
<td>.25</td>
<td>.19</td>
<td>.22</td>
<td>N/A</td>
<td></td>
<td>8.7</td>
<td>.65</td>
<td></td>
</tr>
<tr>
<td></td>
<td>AVG</td>
<td>.21</td>
<td>.19</td>
<td>.43</td>
<td>N/A</td>
<td></td>
<td>3.8</td>
<td>.65</td>
<td></td>
</tr>
<tr>
<td>2.2.1.11</td>
<td>STDV</td>
<td>.19</td>
<td>.26</td>
<td>.30</td>
<td>N/A</td>
<td></td>
<td>7.0</td>
<td>.65</td>
<td></td>
</tr>
<tr>
<td></td>
<td>AVG</td>
<td>.07</td>
<td>.25</td>
<td>.50</td>
<td>N/A</td>
<td></td>
<td>5.2</td>
<td>.65</td>
<td></td>
</tr>
<tr>
<td>4.1.2.13</td>
<td>STDV</td>
<td>.15</td>
<td>.08</td>
<td>.02</td>
<td>N/A</td>
<td></td>
<td>17.7</td>
<td>.65</td>
<td></td>
</tr>
<tr>
<td></td>
<td>AVG</td>
<td>-.19</td>
<td>-.01</td>
<td>.47</td>
<td>N/A</td>
<td></td>
<td>17.7</td>
<td>.65</td>
<td></td>
</tr>
<tr>
<td>5.3.1.1</td>
<td>STDV</td>
<td>.21</td>
<td>.31</td>
<td>.16</td>
<td>N/A</td>
<td></td>
<td>13.9</td>
<td>.65</td>
<td></td>
</tr>
<tr>
<td></td>
<td>AVG</td>
<td>.21</td>
<td>.31</td>
<td>.50</td>
<td>N/A</td>
<td></td>
<td>9.4</td>
<td>.65</td>
<td></td>
</tr>
<tr>
<td>1.2.1.13</td>
<td>STDV</td>
<td>.23</td>
<td>.23</td>
<td>.87</td>
<td>N/A</td>
<td></td>
<td>15.1</td>
<td>.65</td>
<td></td>
</tr>
<tr>
<td></td>
<td>AVG</td>
<td>.23</td>
<td>.23</td>
<td>.50</td>
<td>N/A</td>
<td></td>
<td>10.2</td>
<td>.65</td>
<td></td>
</tr>
<tr>
<td>2.7.2.3</td>
<td>STDV</td>
<td>.22</td>
<td>.72</td>
<td>.33</td>
<td>N/A</td>
<td></td>
<td>14.9</td>
<td>.65</td>
<td></td>
</tr>
<tr>
<td></td>
<td>AVG</td>
<td>.22</td>
<td>.72</td>
<td>.50</td>
<td>N/A</td>
<td></td>
<td>9.4</td>
<td>.65</td>
<td></td>
</tr>
<tr>
<td>5.4.2.1</td>
<td>STDV</td>
<td>.07</td>
<td>.29</td>
<td>.02</td>
<td>N/A</td>
<td></td>
<td>17.2</td>
<td>.65</td>
<td></td>
</tr>
<tr>
<td></td>
<td>AVG</td>
<td>-.17</td>
<td>.29</td>
<td>.50</td>
<td>N/A</td>
<td></td>
<td>12.5</td>
<td>.65</td>
<td></td>
</tr>
<tr>
<td>4.2.1.11</td>
<td>STDV</td>
<td>.18</td>
<td>.23</td>
<td>.87</td>
<td>N/A</td>
<td></td>
<td>14.3</td>
<td>.65</td>
<td></td>
</tr>
<tr>
<td></td>
<td>AVG</td>
<td>.18</td>
<td>.23</td>
<td>.50</td>
<td>N/A</td>
<td></td>
<td>9.4</td>
<td>.65</td>
<td></td>
</tr>
<tr>
<td>4.3.1.9</td>
<td>STDV</td>
<td>.62</td>
<td>.76</td>
<td>.57</td>
<td>N/A</td>
<td></td>
<td>36.4</td>
<td>.65</td>
<td></td>
</tr>
<tr>
<td></td>
<td>AVG</td>
<td>.48</td>
<td>.76</td>
<td>.77</td>
<td>N/A</td>
<td></td>
<td>26.5</td>
<td>.65</td>
<td></td>
</tr>
<tr>
<td>4.1.1.42</td>
<td>STDV</td>
<td>.24</td>
<td>.23</td>
<td>.79</td>
<td>N/A</td>
<td></td>
<td>36.4</td>
<td>.65</td>
<td></td>
</tr>
<tr>
<td></td>
<td>AVG</td>
<td>.24</td>
<td>.23</td>
<td>.50</td>
<td>N/A</td>
<td></td>
<td>26.6</td>
<td>.65</td>
<td></td>
</tr>
<tr>
<td>4.2.1.2</td>
<td>STDV</td>
<td>.31</td>
<td>.36</td>
<td>.67</td>
<td>N/A</td>
<td></td>
<td>47.8</td>
<td>.70</td>
<td></td>
</tr>
<tr>
<td></td>
<td>AVG</td>
<td>.31</td>
<td>.36</td>
<td>.87</td>
<td>N/A</td>
<td></td>
<td>47.8</td>
<td>.70</td>
<td></td>
</tr>
<tr>
<td>1.1.1.37</td>
<td>STDV</td>
<td>.65</td>
<td>.74</td>
<td>.77</td>
<td>N/A</td>
<td></td>
<td>47.8</td>
<td>.70</td>
<td></td>
</tr>
<tr>
<td></td>
<td>AVG</td>
<td>.65</td>
<td>.74</td>
<td>.97</td>
<td>N/A</td>
<td></td>
<td>47.8</td>
<td>.70</td>
<td></td>
</tr>
<tr>
<td>4.1.3.8</td>
<td>STDV</td>
<td>.61</td>
<td>.90</td>
<td>.85</td>
<td>N/A</td>
<td></td>
<td>36.6</td>
<td>.50</td>
<td></td>
</tr>
<tr>
<td></td>
<td>AVG</td>
<td>.61</td>
<td>.90</td>
<td>.85</td>
<td>N/A</td>
<td></td>
<td>36.6</td>
<td>.50</td>
<td></td>
</tr>
</tbody>
</table>
Figure 1.
We present 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.
We provide 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
tree viewed in comparison to the traditional ribosomal tree.

**Figure 3**

We present our 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 percent identity variability. Additionally, each enzyme links to an enzyme
oriented page which displays the data in its entirety for that specific enzyme.

**Figure 4**

Figure 4 presents 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. In the center 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 percent identity 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.