Proteins: Structure, Function, and Genetics

Your article (04272002) from "Proteins: Structure, Function, and Genetics" is available for download
=====
RE: Your article (04272002) from "Proteins: Structure, Function, and Genetics" is available for download

Proteins: Structure, Function, and Genetics Published by John Wiley & Sons, Inc.

Dear Sir or Madam,

PDF page proofs for your article are ready for review.

Please refer to this URL address

http://rapidproof.cadmus.com/RapidProof/retrieval/index.jsp

Login: your e-mail address
Password: ----

The site contains 1 file. You will need to have Adobe Acrobat Reader software to read these files. This is free software and is available for user downloading at http://www.adobe.com/products/acrobat/readstep.html.

This file contains:

Author Instructions Checklist
Adobe Acrobat Users - NOTES tool sheet
Reprint Order form
Copyright Transfer Agreement
Return fax form
A copy of your page proofs for your article

After printing the PDF file, please read the page proofs carefully and:

1) indicate changes or corrections in the margin of the page proofs;
2) answer all queries (footnotes A,B,C, etc.) on the last page of the PDF proof;
3) proofread any tables and equations carefully;
4) check that any Greek, especially "mu", has translated correctly.

Within 48 hours, please return the following to the address given below:

1) original PDF set of page proofs,
2) Reprint Order form,
3) Return fax form

Return to:

Marc Nadeau
Journals Editorial/Production
John Wiley & Sons, Inc.
111 River Street
Hoboken, NJ 07030

U.S.A.

Your article will be published online via our EarlyView service within a few days of correction receipt.
Your prompt attention to and return of page proofs is crucial to faster publication of your work.
If you experience technical problems, please contact Doug Frank (e-mail: frankd@cadmus.com, phone: 800-238-3814 (X615).

If you have any questions regarding your article, please contact me. PLEASE ALWAYS INCLUDE YOUR ARTICLE NO. (04272002) WITH ALL CORRESPONDENCE.

This e-proof is to be used only for the purpose of returning corrections to the publisher.

Sincerely,

Marc Nadeau
Associate Production Editor
John Wiley & Sons, Inc.
E-mail: mnadeau@wiley.com
Tel: 201.748.6716
Fax: 201.748.6052

WILEY
Publishers Since 1807

111 RIVER STREET, HOBOKEN, NJ 07030

*****IMMEDIATE RESPONSE REQUIRED*****

Your article will be published online via Wiley's EarlyView® service (www.interscience.wiley.com) shortly after receipt of corrections. EarlyView® is Wiley's online publication of individual articles in full text HTML and/or pdf format before release of the compiled print issue of the journal. Articles posted online in EarlyView® are peer-reviewed, copyedited, author corrected, and fully citable via the article DOI (for further information, visit www.doi.org). EarlyView® means you benefit from the best of two worlds--fast online availability as well as traditional, issue-based archiving.

Please follow these instructions to avoid delay of publication.

□ **READ PROOFS CAREFULLY**

- This will be your only chance to review these proofs. **Please note that once your corrected article is posted online, it is considered legally published, and cannot be removed from the Web site for further corrections.**
- Please note that the volume and page numbers shown on the proofs are for position only.

□ **ANSWER ALL QUERIES ON PROOFS** (Queries for you to answer are attached as the last page of your proof.)

- Mark all corrections directly on the proofs. Note that excessive author alterations may ultimately result in delay of publication and extra costs may be charged to you.

□ **CHECK FIGURES AND TABLES CAREFULLY** (Color figure proofs will be sent under separate cover.)

- Check size, numbering, and orientation of figures.
- All images in the PDF are downsampled (reduced to lower resolution and file size) to facilitate Internet delivery. ----- These images will appear at higher resolution and sharpness in the printed article.
- Review figure legends to ensure that they are complete.
- Check all tables. Review layout, title, and footnotes.

□ **COMPLETE REPRINT ORDER FORM**

- Fill out the attached reprint order form. It is important to return the form even if you are not ordering reprints. You may, if you wish, pay for the reprints with a credit card. Reprints will be mailed only after your article appears in print. This is the most opportune time to order reprints. If you wait until after your article comes off press, the reprints will be considerably more expensive.

RETURN □**PROOFS**
□**REPRINT ORDER FORM**
□**CTA (If you have not already signed one)**

RETURN IMMEDIATELY AS YOUR ARTICLE WILL BE POSTED ONLINE SHORTLY AFTER RECEIPT; FAX PROOFS TO 201-748-6052

QUESTIONS?

Marc Nadeau, Associate Production Editor
John Wiley & Sons, Inc.
111 River Street
Hoboken, NJ 07030
Phone: 201-748-6716
E-mail: mnadeau@wiley.com
Please refer to journal acronym and article production number

Softproofing for advanced Adobe Acrobat Users - NOTES tool

NOTE: ACROBAT READER FROM THE INTERNET DOES NOT CONTAIN THE NOTES TOOL USED IN THIS PROCEDURE.

Acrobat annotation tools can be very useful for indicating changes to the PDF proof of your article. By using Acrobat annotation tools, a full digital pathway can be maintained for your page proofs.

The NOTES annotation tool can be used with either Adobe Acrobat 3.0x or Adobe Acrobat 4.0. Other annotation tools are also available in Acrobat 4.0, but this instruction sheet will concentrate on how to use the NOTES tool. Acrobat Reader, the free Internet download software from Adobe, DOES NOT contain the NOTES tool. In order to softproof using the NOTES tool you must have the full software suite Adobe Acrobat Exchange 3.0x or Adobe Acrobat 4.0 installed on your computer.

Steps for Softproofing using Adobe Acrobat NOTES tool:

1. Open the PDF page proof of your article using either Adobe Acrobat Exchange 3.0x or Adobe Acrobat 4.0. Proof your article on-screen or print a copy for markup of changes.

2. Go to File/Preferences/Annotations (in Acrobat 4.0) or File/Preferences/Notes (in Acrobat 3.0) and enter your name into the “default user” or “author” field. Also, set the font size at 9 or 10 point.

3. When you have decided on the corrections to your article, select the NOTES tool from the Acrobat toolbox and click in the margin next to the text to be changed.

4. Enter your corrections into the NOTES text box window. Be sure to clearly indicate where the correction is to be placed and what text it will effect. If necessary to avoid confusion, you can use your TEXT SELECTION tool to copy the text to be corrected and paste it into the NOTES text box window. At this point, you can type the corrections directly into the NOTES text box window. **DO NOT correct the text by typing directly on the PDF page.**

5. Go through your entire article using the NOTES tool as described in Step 4.

6. When you have completed the corrections to your article, go to File/Export/Annotations (in Acrobat 4.0) or File/Export/Notes (in Acrobat 3.0). Save your NOTES file to a place on your harddrive where you can easily locate it. **Name your NOTES file with the article number assigned to your article in the original softproofing e-mail message.**

7. **When closing your article PDF be sure NOT to save changes to original file.**

8. To make changes to a NOTES file you have exported, simply re-open the original PDF proof file, go to File/Import/Notes and import the NOTES file you saved. Make changes and re-export NOTES file keeping the same file name.

9. When complete, attach your NOTES file to a reply e-mail message. Be sure to include your name, the date, and the title of the journal your article will be printed in.

John Wiley & Sons, Inc.
Publishers Since 1807

REPRINT BILLING DEPARTMENT • 111 RIVER STREET • HOBOKEN, NJ 07030
PHONE: (201) 748-8789; FAX: (201) 748-6326
E-MAIL: reprints @ wiley.com

PREPUBLICATION REPRINT ORDER FORM

Please complete this form even if you are not ordering reprints. This form **MUST** be returned with your corrected proofs and original manuscript. Your reprints will be shipped approximately 4 weeks after publication. Reprints ordered after printing are substantially more expensive.

JOURNAL: *PROTEINS: Structure, Function, and Genetics* VOLUME_____ ISSUE_____

TITLE OF MANUSCRIPT____

MS. NO.___________ NO. OF PAGES______ AUTHOR(S)___________________________

REPRINTS 8 1/4 X 11

No. of Pages	100 Reprints	200 Reprints	300 Reprints	400 Reprints	500 Reprints
	$	$	$	$	$
1-4	336	501	694	890	1,052
5-8	469	703	987	1,251	1,477
9-12	594	923	1,234	1,565	1,850
13-16	714	1,156	1,527	1,901	2,273
17-20	794	1,340	1,775	2,212	2,648
21-24	911	1,529	2,031	2,536	3,037
25-28	1,004	1,707	2,267	2,828	3,388
29-32	1,108	1,894	2,515	3,135	3,755
33-36	1,219	2,092	2,773	3,456	4,143
37-40	1,329	2,290	3,033	3,776	4,528

** REPRINTS ARE ONLY AVAILABLE IN LOTS OF 100. IF YOU WISH TO ORDER MORE THAN 500 REPRINTS, PLEASE CONTACT OUR REPRINTS DEPARTMENT AT (201)748-8789 FOR A PRICE QUOTE.

□ Please send me __________ reprints of the above article at...................... $____________

Please add appropriate State and Local Tax {Tax Exempt No.__________________} $____________
Please add 5% Postage and Handling.. $____________
TOTAL AMOUNT OF ORDER** .... $____________

***International orders must be paid in U.S. currency and drawn on a U.S. bank*

Please check one: □ Check enclosed □ Bill me □ Credit Card
If credit card order, charge to: □ American Express □ Visa □ MasterCard □ Discover
Credit Card No.____________________________ Signature__________________________ Exp. Date___________

Bill To:
Name______________________________________
Address/Institution_________________________
____
____
____

Ship To:
Name______________________________________
Address/Institution_________________________
____
____
____

Purchase Order No. ____________________________ Phone ________________ Fax ________________
E-mail: ________________________________

A Method Using Active-Site Sequence Conservation to Find Functional Shifts in Protein Families: Application to the Enzymes of Central Metabolism, Leading to the Identification of an Anomalous Isocitrate Dehydrogenase in Pathogens

Rajdeep Das and Mark Gerstein*
AQ: 1 *Department of Molecular Biophysics and Biochemistry, Yale University, New Haven, Connecticut*

ABSTRACT **We have introduced a method to identify functional shifts in protein families. Our method is based on the calculation of an active-site conservation ratio, which we call the "ASC ratio." For a structurally based alignment of a protein family, this ratio is the average sequence similarity of the active-site region compared to the full-length protein. The active-site region is defined as all the residues within a certain radius of the known functionally important groups. Using our method, we have analyzed enzymes of central metabolism from a large number of genomes (35). We found that for most of the enzymes, the active-site region is more AQ: 2 highly conserved than the full-length sequence. However, for three tricarboxylic acid (TCA)-cycle enzymes, active-site sequences are considerably more AQ: 3 diverged (than full-length ones). In particular, we were able to identify in six pathogens a novel isocitrate dehydrogenase that has very low sequence similarity around the active site. Detailed sequence–structure analysis indicates that while the active-site structure of isocitrate dehydrogenase is most likely similar between pathogens and nonpathogens, the unusual sequence divergence could result from an extra domain added at the N-terminus. This domain has a leucine-rich motif similar one in the *Yersinia pestis* cytotoxin and may therefore confer additional pathogenic functions. Proteins 2003;00:000–000.**
© 2003 Wiley-Liss, Inc.

Key words: metabolic enzyme; genome; active site; sequence variation

INTRODUCTION

With the completion of a large number of genome sequences, a major problem for biology is functional annotation on a genome scale—determining a function for all the proteins encoded by a genome.[1,2] Unfortunately, it is very hard to do large-scale functional annotation purely on the basis of sequence.[3] Sequences diverge beyond the point of obvious recognition in terms of functional and structural homologues.[4,5] Furthermore, divergence on the sequence over its entire length has to do with many factors, and it is hard to abstract the specific bit of divergence that relates to functional conservation. This is where structure can play a major role in guiding people to the functionally important residues associated with the active site. Also with the advent of structural genomics, we are now confronted with the production of a large number of crystal and NMR structures. It is very useful to use these structures in a systematic fashion to refine our understanding of protein function. This is the overall aim of this work. We tried to develop a method for assessing active-site sequence conservation in comparison to full-length conservation, and we applied this to a number of pathway enzymes and achieved novel results. AQ: 4

Pathway enzymes have been a topic of wide scientific interest in both the pre- and postgenomic eras. Studies have focused on several aspects of pathway analyses. Particularly in the postgenomic era, metabolic pathways have been studied using sequence information. Such genomic analyses of pathways have been performed in many ways, and various metabolic databases have been constructed.[6–11] 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 the substrate or intermediate.[12] Previous studies analyzed protein families in terms of sequence–structure relationships.[13] In this article, we have analyzed a number of enzymes of the central metabolic pathways [i.e., glycolysis, pentose phosphate pathway, and tricarboxylic acid (TCA) cycle] in 35 organisms in terms of sequence variability around active sites. AQ: 5

MATERIALS AND METHODS

Organisms and Databases Used

In this article, we have analyzed 18 enzymes of the central metabolic pathways, in 35 organisms, in terms of sequence variability around active sites. A ribosomal tree, shown in Figure 1(A), lists all the organisms studied in the F1

*Correspondence to: Mark Gerstein, Department of Molecular Biophysics and Biochemistry, 266 Whitney Avenue, Yale University, P.O. Box 208114, New Haven CT 06520. E-mail: mark.gerstein@yale.edu

Received 6 December 2002; Accepted 19 August 2003

Published online 00 Month 2003 in Wiley InterScience (www.interscience.wiley.com). DOI: 10.1002/prot.10639

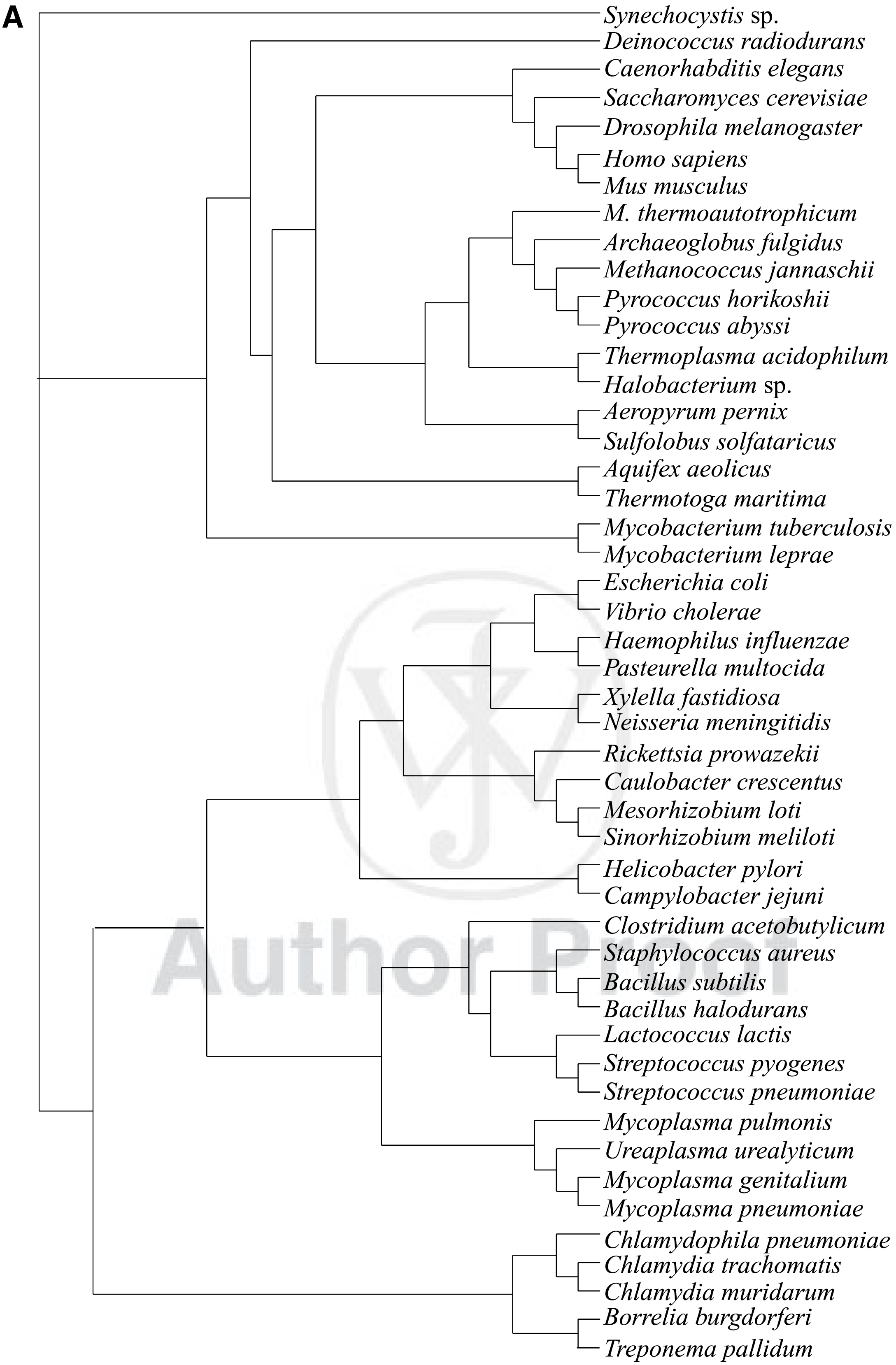


Fig. 1. (**A**) Phylogenetic tree based on rRNA. Phylogenetic tree of 35 organisms studied, based on rRNA sequences. Organisms represent all three kingdoms of life (i.e., archaea, bacteria, and eukarya). (**B**) Average pairwise similarities are shown for the enzymes in central metabolic pathway. For heteromeric enzymes, the subunit that has an 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 if we encountered other problems, as shown by asterisk.

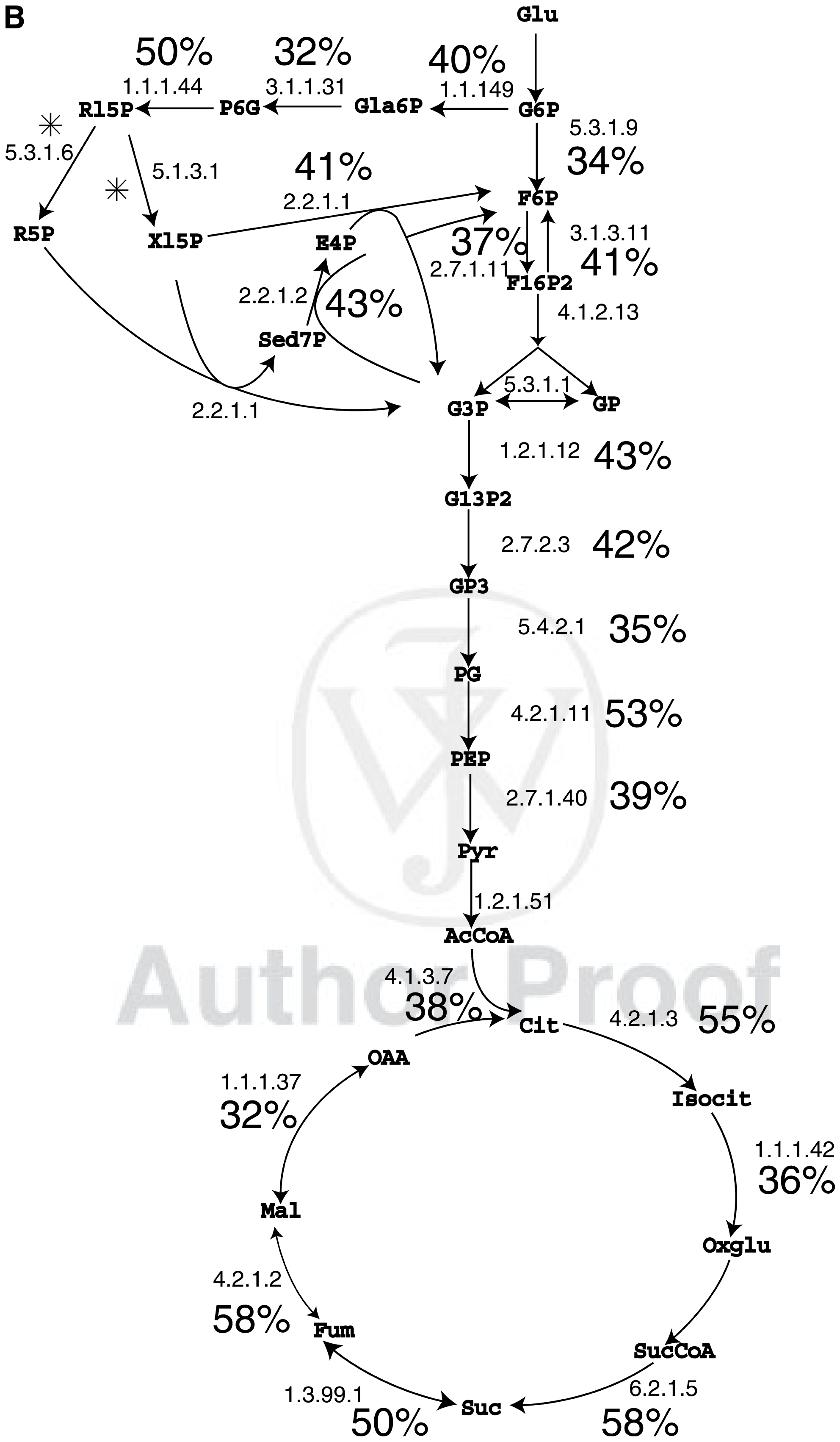


Figure 1. (Continued.)

TABLE I. Active Residues of the Enzymes

Enzyme	EC No.	PDB ID	Active residues	Reference
Malate dehydrogenase	1.1.1.37	1IB6	R81, R87, N119, H177	20
Isocitrate dehydrogenase	1.1.1.42	1HQS	R110, R120, R144, N106, C118, S104, T30, Y151	22
Fumarate reductase	1.3.99.1	1QLA	R301a, R404a, H369a, F141a	17
Fructose-1,6-biphosphatase	3.1.3.11	1FRP	R243a, K274a	23

analyses: All of the three major kingdoms of life (i.e., archaea, eubacteria, and eukarya) are represented here.

We collected all the sequences of central metabolic enzymes from the KEGG database for 35 organisms.[11] We only selected those sequences that were clearly annotated as functionally active sequence. In some cases, enzymes have multiple sequences and we selected only those isoforms that can be found in at least 10 other organisms for AQ: 6 comparison. We began by selecting one representative structure from the Protein Data Bank (PDB) for each of the 14 enzymes.[14] A list of the chosen PDB structures is T1 shown in Table I for four enzymes that we discuss in this article.

Structural Identification of Active Site

We developed a new method to calculate the active-site AQ: 7 conservation sequence, which we call ASC ratio. Below, we describe the calculation of our ASC ratio, which is essentially the ratio of the active-site sequence similarity to the overall full-length protein, and then we describe how it is employed.

The first step in this procedure is the structural identification of the active-site neighborhood. We got the central position of this from the literature, using the position of biochemically identified functionally important residues.

Table I shows the specific active-site residues that were considered for the enzymes in our study. Next, we define an active-site environment as all the residues that fall within a radius of 10 Å from the active-site residues. An average of about 90 residues falls within this radius. We have used the multipurpose program MOLEMAN to determine the residues in the active-site sphere.[15] Given a certain distance, the program can determine all residues that fall within that radius from the selected residue.

Calculation of ASC Ratio

For each enzyme in question, we gathered all homologs from the KEGG database (see above) and constructed a multiple alignment. We used the program CLUSTAL to AQ: 8 generate multiple alignments of the sequences.[16] Once we determined the residues in the active-site sphere in each representative structure, we mapped those residues onto the sequences from other organisms, using multiple sequence alignment. Finally, we calculated the pairwise sequence similarity between all members of the enzyme for these active-site residues. This part of our analysis is somewhat similar to three-dimensional (3D) cluster analyses used earlier to study a group of protein families.[17] The F2 overall strategy is illustrated in Figure 2. In addition to the active sites, we also computed pairwise full-length sequence similarity from our alignments. 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. Therefore, such residues are likely to be more conserved than the residues in the balance of the sequence. In order to identify sequences for which the active site is modified, we calculated a ratio of active-site similarity to full-length similarity. We call this the ASC ratio, denoted by the symbol R. Since the active-site residues are more likely to be conserved than the residues in the remaining sequences, the ratio of the two quantities, our ASC ratio, is expected to be more than 1. However, if the residues in the active-site sphere are modified, the pairwise active-site similarity becomes lower than the pairwise full-length similarity, and the ASC ratio will be less than 1. We have computed all the ratios from the pairwise similarity values and plotted their distribution. When an organism's enzyme sequences are all extensively modified, the ASC ratio values will tend to cluster at the point where the ASC ratio is less than 1, allowing for easy identification of those organisms.

Detailed Characterization of Functional Shift: Structural Comparison and Identification of New Sequence Motifs

Based on our initial analysis, we determine the enzymes and associated organisms for which we observed modified active sites. In this step, we tried to analyze these enzymes in terms of structure- and sequence-based characterization. For sequence motif detection, we searched Genbank database/PDB sequence database against a query sequence using standard BLAST. We have modeled the structure using the program MODELER.[18] We used the built-in alignment generation option of the program, MALIGN, to generate the alignment for modeling. Finally, we used the program STRUCTAL to compare the homology models.[19]

RESULTS AND DISCUSSION

Enzymes in the central metabolism vary greatly in average pairwise sequence identity. Some of the enzymes in the pathway are highly conserved, with an average sequence identity $\sim$60%, and some are less conserved. This is shown in Figure 1(B). For most of the central metabolic enzymes, we have 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 for which we have we found large sequence variation near the active site (i.e., a large ASC ratio); these we discuss below.

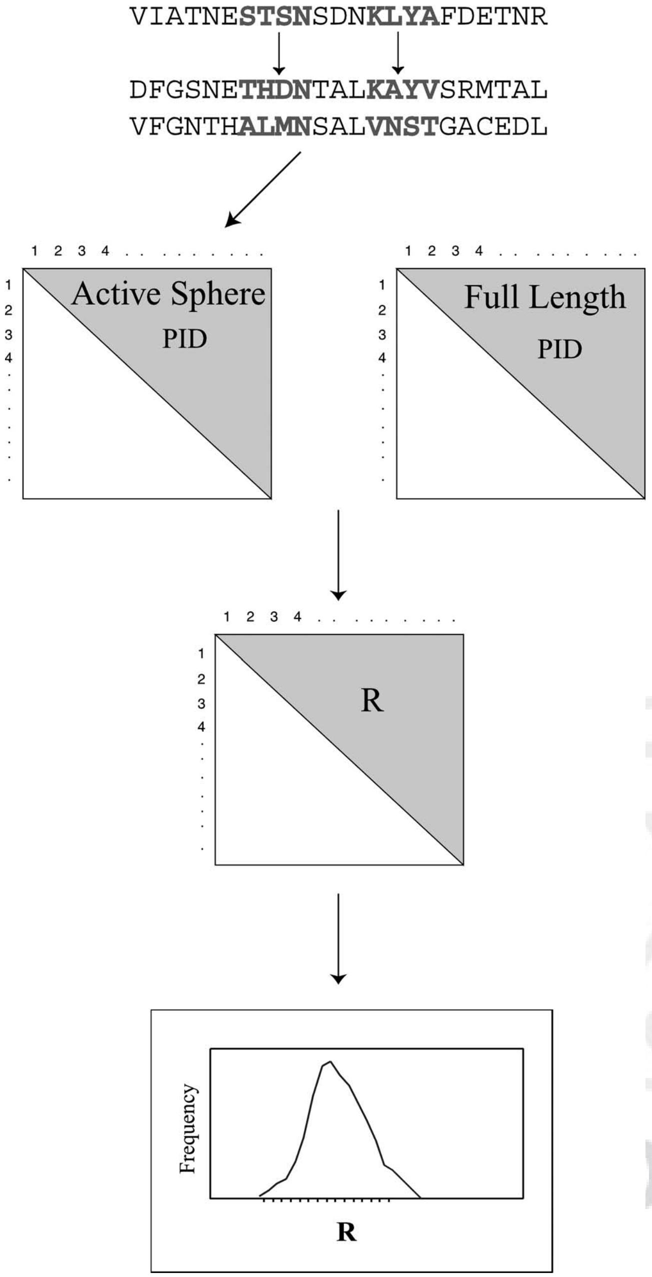


Fig. 2. Strategy of structure-based sequence comparison. We determined the residues that fell within the sphere of a 10-Å radius around the active residues using MOLEMAN.[10] Then we mapped those residues onto sequences from other organisms, using multiple sequence alignment. The red-colored residues in the sequence are the residues that fall 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 shown.

Example 1: Isocitrate Dehydrogenase

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

$$\text{Isocitrate} + \text{NADP+} \rightarrow \text{2-Oxoglutarate} + CO_2 + \text{NADPH}$$

Distribution of ASC ratios shows that 6 organisms have AQ: 9 modified isocitrate dehydrogenase; therefore, the distribution of ASC ratios falls in the region where the value is less than 1. Figure 3(A) shows the distribution of the ASC F3 ratios and the corresponding list of 6 organisms in the box. They are all known pathogens. We have compared 33 sequences of isocitrate dehydrogenase from different organisms. The probability of pathogenic sequences, as a sample of the entire distribution of ASC ratios, having a mean ASC ratio of 0.52 was tested using normal statistics. The result shows that the probability p is very significant (i.e., less than 0.00001).

Homology modeling–based structural comparison of the active site

We compared the structure of the active site by comparing model structures of 4 pathogenic isocitrate dehydrogenases with the representative structure. We have modeled the structure of the 30-residue cluster using homology modeling, as described in the Methods section. Table II T2 shows the root-mean-square deviation (RMSD) difference between the modeled isocitrate dehydrogenase and the representative structure. The pairwise RMSD values are quite low and indicate that the active site of the pathogenic isocitrate dehydrogenase is most likely similar to the known structure. It should be noted here that we are comparing modeled structures derived from low homology and are thus likely to observe a very large deviation. Therefore, a small RMSD would mean a similarity in structures.

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 the 6 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* (a bubonic plague pathogen). The stretch of sequences is extremely rich in leucine. Although the antihost function of this effector is unknown, YopM is believed to be an important cytotoxin for bacterial virulence.[20] Figure 4 F4 shows the sequence alignment of the sequence YopM with the N-terminal domain of 4 pathogenic isocitrate dehydrogenases. Other evidence that the modified isocitrate dehydrogenase may have additional function comes from the observation that *Mycobacterium tuberculosis* has two isocitrate dehydrogenase sequences, Rv3339c and Rv0066c, the first of which is a standard isocitrate dehydrogenase sequence, and second of which is the modified sequence. AQ: 10 Therefore, the presence of the two sequences may indicate an extra function of modified sequence.

It should be noted here that our phylogenetic clustering of the organisms based on protein sequences also groups

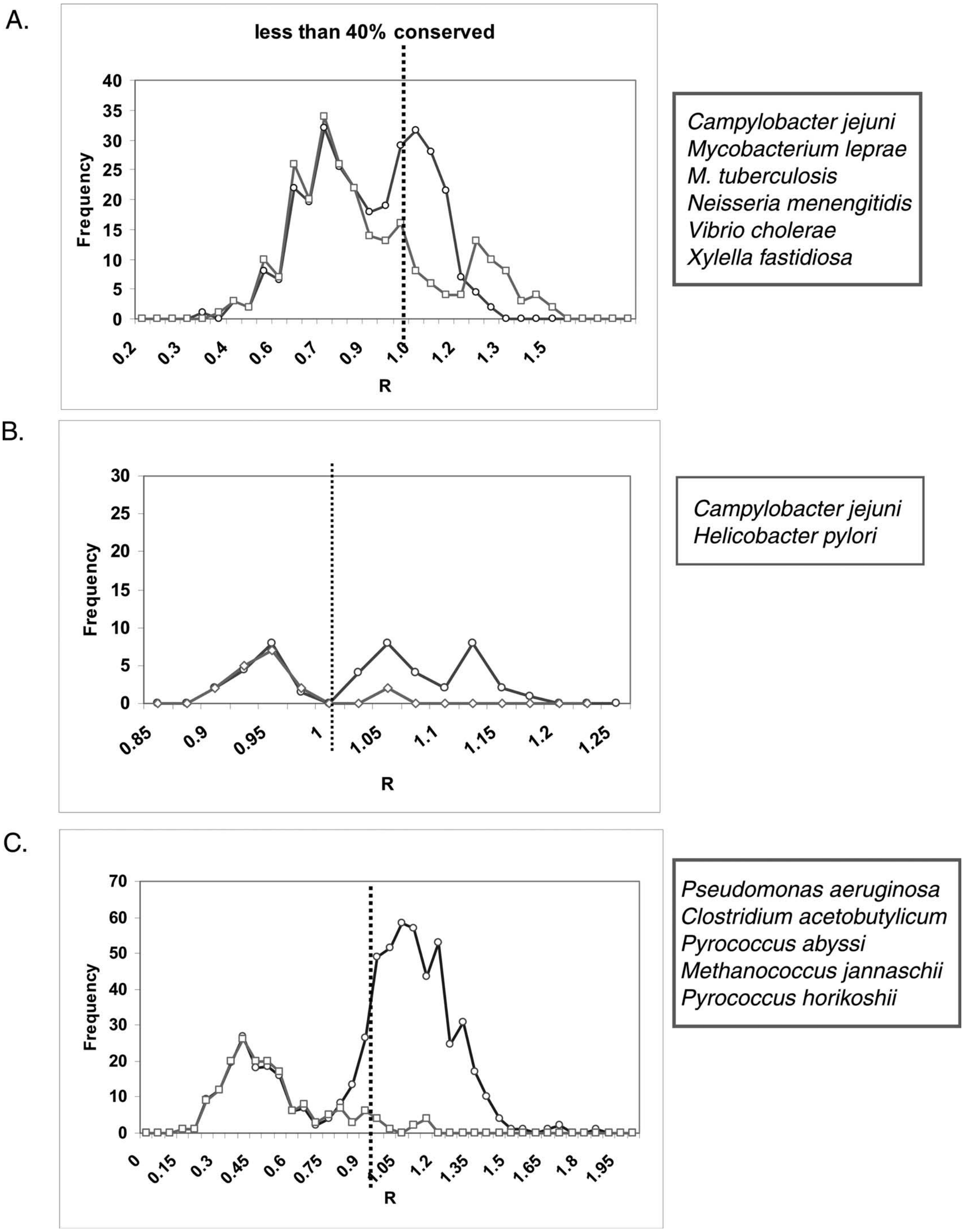
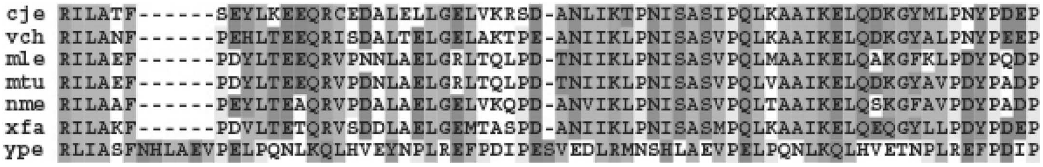


Fig. 3. Distribution of *R*-values for three enzymes in the TCA cycle. The *x* axis shows *R*-values and the *y* axis shows frequency. Overall distribution is shown in the blue line, and the red line represents the distribution that corresponds to the 6 organisms with modified sequences. The dashed bar represents the line where *R* is 1; the active-site similarity is equal to the overall sequence similarity. **A**, **B**, and **C**, respectively, represent the distributions corresponding to isocitrate dehydrogenase, fumerate reductase, and malate dehydrogenase.



```
cje RILATF------SEYLKEEQRCEDALELLGELVKRSD-ANLIKTPNISASIPQLKAAIKELQDKGYMLPNYPDEP
vch RILANF------PEHLTEEQRISDALTELGELAKTPE-ANIIKLPNISASVPQLKAAIKELQDKGYALPNYPEEP
mle RILAEF------PDYLTEEQRVPNNLAELGRLTQLPD-TNIIKLPNISASVPQLMAAIKELQAKGFKLPDYPQDP
mtu RILAEF------PDYLTEEQRVPDNLAELGRLTQLPD-TNIIKLPNISASVPQLVAAIKELQDKGYAVPDYPADP
nme RILAAF------PEYLTEAQRVPDALAELGELVKQPD-ANVIKLPNISASVPQLTAAIKELQSKGFAVPDYPADP
xfa RILAKF------PDVLTETQRVSDDLAELGEMTASPD-ANIIKLPNISASMPQLKAAIKELQEQGYLLPDYPDEP
ype RLIASFNHLAEVPELPQNLKQLHVEYNPLREFPDIPESVEDLRMNSHLAEVPELPQNLKQLHVETNPLREFPDIP
```

Fig. 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*.

TABLE II. Comparison of Representative Structure and Modeled Structure for Isocitrate Dehydrogenase Based on C-α Atoms

Organisms	RMSD
Campylobacter jejuni	0.76
Mycobacterium leprae	0.23
Mycobacterium tuberculosis	0.54
Neisseria meningitidis	1.25

the pathogenic organisms in one cluster, as shown in F5 Figure 5. Interestingly the structure of similar isocitrate dehydrogense has been solved recently showing similarity in the active site.[21]

Example 2: Malate Dehydrogenase

$$\text{Malate} + \text{NAD+} \rightarrow \text{Oxaloacetate} + \text{NADH}$$

The distribution of ASC ratios for malate dehydrogenase is shown in Figure 3(C). This figure shows that 5 organisms—*Pseudomonas aeruginosa*, *Clostridium acetobutylicum*, *Pyrococcus abyssi*, *Methanococcus jannaschii*, and *Pyrococcus horikoshii*—all have modified malate dehydrogenase sequences. However, previous studies show that *M. jannaschii* has two sequences for malate dehydrogenase, MJ0490 and MJ1425.[22] Although two sequences are annotated to be 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. In our analyses, we have used sequences from 35 organisms. The probability of a sample of sequences of this size having a mean ASC ratio of 0.45 (mean ASC ratio for the modified organisms) was tested using normal statistics and is less than 0.00001.

Example 3: Fumarate Reductase

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

$$\text{Fumarate} + \text{FADH2} \rightarrow \text{Succinate} + \text{FAD}$$

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 *Helicobacter pylori* and *Campylobacter jejuni* compared with the rest of the organisms. These two mucosal pathogens are known to have fumarate respiration.[23,24] The distribution of ASC ratios is shown in Figure 3(B). In order to see if there is any structural difference in the active site of the two pathogenic fumarate reductases, we modeled the active-site structures for the two organisms. Table II shows that the pairwise RMSD from that of the known structure is low enough to argue that the active-site structure in the two pathogenic fumarate reductases is similar to the known structure.

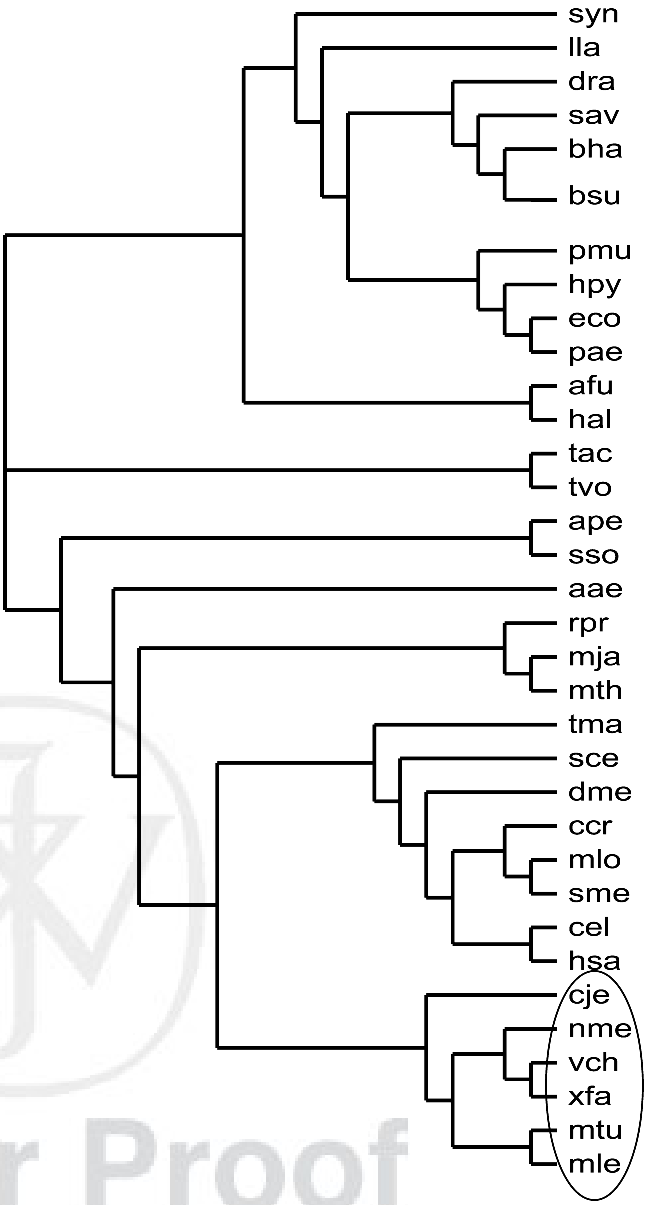


Fig. 5. Phylogenetic clustering based on isocitrate dehydrogenase sequences of the organisms. Although organisms cluster differently in the ribosomal tree, the 6 most diverged isocitrate dehydrogenase sequences cluster together.

For both enzymes, isocitrate dehydrogenase and fumarate reductase, we conclude that the active-site structures of the enzymes in the suspect organisms are most likely to be similar to the known structure. Thus, 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 the enzymatic function in the organisms. The number of sequences in fumarate reductase was 10. The probability of a sample of sequences of this size having a mean ASC ratio of 0.94 (mean ratio for the modified organisms) was tested using normal statistics and is a marginal 0.065. This can be a result of small size of the data set.

In addition, we have also observed a few glycolytic enzymes for which there are a group of organisms with modified active sites. Fructose-1,6-biphosphatase is one such enzyme that catalyzes conversion of fructose-1,6-biphosphatase to fructose-6-phosphate. We observed that four organisms—*Lactobacillus lactis*, *Clostridium acetobutylicum*, *Bacillus subtilis*, and *Staphylococcus aureus*—have very diverged sequences around the active site. Similarly, archaeal glyceralde-3-phosphate dehydrogenases have very modified sequences in the active site. However, we were not able to characterize these enzymes further, either structure- or sequencewise.

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 1(B). In particular, the two enzymes (EC Nos.: 11142 and 11137) for which we observed significant sequence variation around the active site have low-average pairwise sequence similarity. It is possible that the observed sequence variation in the active site for these two enzymes can be a result of a bad local sequence alignment in that particular region. This will lead to low pairwise sequence similarity for the active-site sequence. We therefore performed a statistical test to verify this possibility. We calculated the average pairwise sequence similarity for a large number of random clusters, each comprising ~90 residues. To generate a random cluster, we selected a random residue in the representative sequence with known structure and then determined all the residues that are present within the sphere of a 10 Å radius. Using this procedure, we have generated large number of random clusters. If the alignments in the active-site sequences were particularly bad, active-site sequence similarity would be lower than the average sequence similarity values for these random clusters. For example, in *Vibrio 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 1.2×10^{-12}. Similarly for malate dehydrogenase, results show that the *P*-values are less than 0.01 for the organisms with modified active site. Therefore, we conclude that the residues in the active-site sphere most likely represent the active site in the enzyme and do not represent random residues, as would be expected in case of bad alignment. However, in the case of fumarate reductase, *H. pylori* and *C. jejuni* have average active-site similarity of 36% and less that average pairwise identity of random clusters (40%), and have a *P*-value of less than 2.87×10^{-7} to occur
AQ: 11 by chance. Clearly, the smaller number of organisms (i.e., 10) in our analyses biases the results in fumarate reductase. However, the biological significance of this result is not clear.

CONCLUSIONS

In this article, we have introduced a method to identify functional shifts in protein families based on the calculation of an active-site conservation (ASC) ratio. For a structurally based alignment of a protein family, this ratio is the average sequence similarity for the active-site region compared to the protein's full-length. We have analyzed AQ: 12
the sequence variation of enzymes around active sites in a large number of organisms. For most of the organisms, the results showed that sequences are highly conserved around the active site. However, there are 3 enzymes in the TCA cycle for which we have observed that sequences are extremely divergent. Homology modeling showed that diverged sequences might have an active-site structure similar to the known structure. Most interestingly, 6 pathogenic organisms have unique isocitrate dehydrogenase that has sequence similarity to a cytotoxin in *Y. pestis* that is linked to bacterial pathogenicity.

REFERENCES

1. Vitkup D, Melamud E, Moult J, Sander C. Completeness in structural genomics. Nat Struct Biol 2001:8;559–566.
2. Venclovas C, Zelma A, Fidelis K, Moult J. Comparison of performance in successive CASP experiments. Proteins 2001;5:163–170.
3. Teichmann SA, Chothia C, Gerstein M. Advances in structural genomics. Curr Opin Struct Biol;1999:9:390–399.
4. Wilson C, Kreychman J, Gerstein M. Assessing annotation transfer for genomics: quantifying the relations between protein sequence, structure and function through traditional and probabilistic scores. J Mol Biol 2000;297:233–249.
5. Wood TC, Pearson WR. Evolution of protein sequences and structures. J Mol Biol 1999;291:977–995.
6. Bono H, Ogata H, Goto S, Kanehisa M. Reconstruction of amino acid biosynthesis pathways from the complete genome sequence. Genome Res 1998;8:203–210.
7. Galperin MY, Koonin EV. Functional genomics and enzyme evolution: homologous and analogous enzymes encoded in microbial genomes. Genetica 1999;106:159–170.
8. Dandekar T, Schuster S, Snel B, Huynen M, Bork P. Pathway alignment: application to the comparative analysis of glycolytic enzymes. Biochem J 1999;343:115–124.
9. Forst CV, Schulten K. Evolution of metabolisms: a new method for the comparison of metabolic pathways using genomics information. J Comput Biol 1999;6:343–360.
10. Ogata H, Fujibuchi W, Goto S, Kanehisa M. A heuristic graph comparison algorithm and its application to detect functionally related enzyme clusters. Nucleic Acids Res 2000;28:4021–4028.
11. Kanehisa M. Goto S, KEGG: Kyoto encyclopedia of genes and genomes. Nucleic Acids Res 2000;28:27–30.
12. Naylor GJ, Gerstein M. Measuring shifts in function and evolutionary opportunity using variability profiles: a case study of the globins. J Mol Evol 2000;51:223–233.
13. Dym O, Eisenberg D. Sequence–structure analysis of FAD-containing proteins. Protein Sci 2001;10:1718–1728.
14. Sussman JL, Lin LD, Jiang J, Manning NO, Prilusky J, Ritter O, Abola EE. Protein Data Bank (PDB): database of three-dimensional structural information of biological macromolecules. Acta Crystallogr D Biol Crystallogr 1998;54:1078–1084.
15. Kleywegt GJ, Jones TA. Phi/psi-chology: Ramachandran revisited. Structure 1996;4:1395–1400.
16. Higgins DG, Thompson JD, Gibson TJ. Using CLUSTAL for multiple alignments. Methods Enzymol 1996;266:383–402.
17. Landgraf R, Xanarios I, Eisenberg D. Three-dimensional cluster analysis identifies interfaces and functional residue clusters in proteins. J Mol Biol 2001;307:1487–1502.
18. Sali A, Blundell TL. Comparative protein modeling by satisfaction of spatial restraints. J Mol Biol 1993;234:779–815.
19. Gerstein M, Levitt M. Comprehensive assessment of automatic structural alignment against a manual standard, the SCOP classivication of proteins. Protein Sci 1998;10:445–456.
20. Evdokimov AG, Anderson DE, Routzahn KM, Waugh DS. Unusual molecular architecture of the *Yersinia pestis* cytotoxin yopm: a leucine-rich repeat protein with the shortest repeating unit. J Mol Biol 2001;312:807–821.
21. Yasutake Y, Watanabe S, Yao M, Takada Y, Fukunaga N, Tanaka

I. Structure of monomeric isocitrate dehydrogenase: evidence of a protein monomerization by a domain duplication. Structure 2002; 10:1637–1648.

22. Graupner M, Xu H, White RH. Identification of an archaeal 2-hydroxy acid dehydrogenase catalyzing reactions involved in coenzyme biosynthesis in methanoarchaea. J Bacteriol 2000;182: 3688–3692.

23. Ge Z. Potential of fumarate reductase as a novel therapeutic target in *Helicobacter pylori* infection. Expert Opin Ther Targets 2002;6:135–146.

24. Sellars MJ, Hall SJ, Kelly DJ. Growth of *Campylobacter jejuni* supported by respiration of fumarate, nitrate, nitrite, trimethylamine-*n*-oxide, or dimethyl sulfoxide requires oxygen. J Bacteriol 2002;184:4187–4196.

xx. Lancaster CR, Kroger A, Auer M, Michel H. Structure of fumarate reductase from *Wolinella succinogenes* at 2.2 Å resolution. Nature 1999;402:377–385. **AQ: 13**

xx. Bell JK, Yennawar HP, Wright SK, Thompson JR, Viola RE, Banaszak LJ. Structural analyses of a malate dehydrogenase with a variable active site. J Biol Chem 2001;276:31156–31162. **AQ: 13**

xx. Singh SK, Matsuno K, LaPorte DC, Banaszak LJ. Crystal structure of *Bacillus subtilis* isocitrate dehydrogenase at 1.55 Å: insights into the nature of substrate specificity exhibited by *Escherichia coli* isocitrate dehydrogenase kinase/phosphatase. J Biol Chem 2001;276:26154–26163. **AQ: 13**

xx. Xue Y, Huang S, Liang JY, Zhang Y, Lipscomb WN. Crystal structure of fructose-1,6-bisphosphatase complexed with fructose 2,6-bisphosphate, amp, andZn2+ at 2.0-Å resolution: aspects of synergism between inhibitors. Proc Natl Acad Sci USA 1994;91: 12482–12486. **AQ: 13**

AQ1: If running head is not OK, please provide one that is 45 or fewer characters.

AQ2: Sentence beginning "We found that. . ." as meant?

AQ3: Sentence beginning "However,. . ." as meant?

AQ4: Sentence beginning "We tried. . ." as meant?

AQ5: Sentence beginning "In this article. . ." as meant?

AQ6: Sentence beginning "In some cases. . ." as meant?

AQ7: In sentence beginning "We developed. . .", is wording change OK, so that the term matches the acronym?

AQ8: Journal style requires that references be cited in numerical order. Ref. 26 has been changed to ref. 16, and succeeding references have been renumbered.

AQ9: CO2 as meant?

AQ10: Sentence beginning "Other evidence. . ." as meant?

AQ11: In sentence beginning "However. . .", the meaning is unclear. Please clarify. Are you saying that the final P-value is too small to occur by chance?

AQ12: Sentence beginning "For a. . ." as meant?

AQ13: References numbered xx are not cited in body of article. Please cite in article or delete from reference list. If cited, please cite in numerical order and renumber references in both text citations and in the reference list.