Genomic evidence that the intracellular proteins of archaeal microbes contain disulfide bonds

Parag Mallick*†‡§, Daniel R. Boutz†‡, David Eisenberg*†‡§, and Todd O. Yeates†‡§¶

†Department of Chemistry and Biochemistry and ‡Department of Energy Center for Genomics and Proteomics, §Molecular Biology Institute, and *Howard Hughes Medical Institute, University of California, Los Angeles, CA 90095-1569

Contributed by David Eisenberg, May 23, 2002

Disulfide bonds have only rarely been found in intracellular proteins. That pattern is consistent with the chemically reducing environment inside the cells of well-studied organisms. However, recent experiments and new calculations based on genomic data of archaea provide striking contradictions to this pattern. Our results indicate that the intracellular proteins of certain hyperthermophilic archaea, especially the crenarchaea *Pyrobaculum aerophilum* and *Aeropyrum pernix*, are rich in disulfide bonds. This finding implicates disulfide bonding in stabilizing many thermostable proteins and points to novel chemical environments inside these microbes. These unexpected results illustrate the wealth of biochemical insights available from the growing reservoir of genomic data.

Growing genomic databases are creating opportunities for new kinds of computational analyses and novel discoveries. The multitude of completely sequenced genomes, some 50 from varied microbes, offers special advantages for comparative studies (1–5). One fruitful but challenging line of genomic inquiry concerns the connection between linear protein sequences and their functional or three-dimensional structural properties. In the present study, the comparison of sequences to structure across multiple genomes leads to a surprising revelation about disulfide bonds in certain microbes.

The abundant disulfide bonds in extracellular proteins, such as antibodies, have long been known to stabilize these proteins in their oxidizing and sometimes variable and disruptive environments (6). In contrast, the chemical environment inside a typical cell is reducing, with a reduction potential around −200 to −300 mV for *Escherichia coli* (7, 8). As a result, cysteine residues in intracellular proteins are generally found in their reduced form with free sulfhydryl groups. Although rare, there are examples of disulfide bonds in intracellular proteins. They are found mainly in proteins that catalyze oxidation-reduction (redox) processes (9–11). Examples include thioredoxin and glutathione reductase, in which a disulfide bond forms during part of a catalytic cycle, and Hsp33 and OxyR, in which a disulfide bond forms as part of a redox-sensing mechanism. But intracellular disulfide bonds such as these are rare and generally transiently formed or marginally stable, rather than being essential for structural integrity.

Given the rarity of intracellular protein disulfide bonds, it was surprising when the recent crystal structure of adenylosuccinate lyase (12) from the hyperthermophilic archaeon *Pyrobaculum aerophilum* (13) revealed a protein chain stabilized by three disulfide bonds (Fig. 1 *Inset*). Other disulfide bonds had been identified in a few proteins from other thermophiles (14–18) but had not prompted wider scrutiny. To investigate the possibility (12) that some entire microorganisms, such as *P. aerophilum*, might be rich in intracellular protein disulfide bonds, a computational genomics study was undertaken, and its positive finding was then validated experimentally.

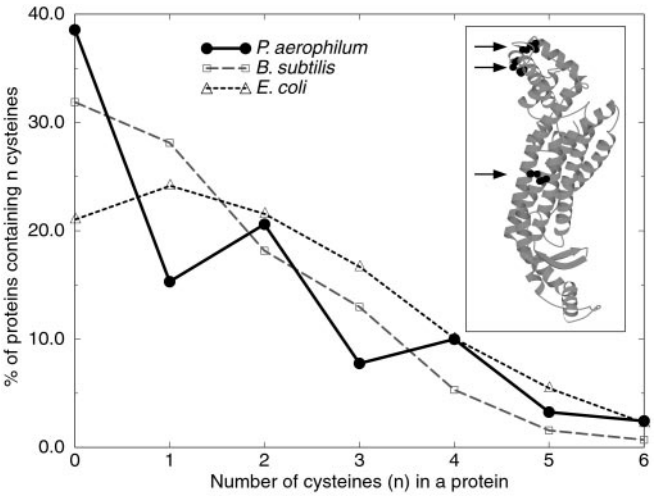

Fig. 1. An abundance of intracellular disulfide bonds in *P. aerophilum* is suggested by the preference for even numbers of cysteine residues in its proteins (bold line). For controls, corresponding plots are shown for *B. subtilis* (dashed line) and *E. coli* (dotted line). In this plot, data are drawn from proteins ranging in size from 100 to 200 amino acids, as the phenomenon is most evident for small to medium-ized proteins. *Inset* shows the structure of adenylosuccinate lyase from *P. aerophilum* (12). Its three disulfide bonds led to the initial suggestion that some hyperthermophilic archaea might be rich in intracellular protein disulfide bonds.

Materials and Methods

Genome Sequence Databases. Predicted protein coding sequences of the genomes used in this study were obtained from the National Center for Biotechnology Information (ftp.ncbi.nih.gov). Of the 50 microbes whose complete genome sequences are available in the public domain, many are strains of the same organism and were thus excluded from the representative sample of 25 microbes presented in this analysis.

Control Sets of Proteins. Sets of proteins known either to contain or to not contain disulfide bonds were extracted from the Protein Data Bank (PDB) by using the PDB keyword SSBOND. These sets are available at http://www.doe-mbi.ucla.edu/~parag/FOLD_RECOGNITION/DISULFIDES. An additional set of protein structures known to contain metals was derived from the Metalloprotein Database, which is available at http://metallo.scripps.edu/.

Identification of Intracellular Proteins. Only proteins believed to be intracellular were included in the analysis. A protein was predicted to be intracellular if it had a (max SignalP positives) <2 (19) and a MOMENT prediction of 0 transmembrane segments (20).

Sequence-Structure Assignment. The genome sequences of the 25 studied microbes were matched with corresponding protein

Abbreviations: TCEP, Tris-(2-carboxyethyl)phosphine; CPM, 7-diethylamino-3-(4′-maleimidylphenyl)-4-methylcoumarin; redox, oxidation-reduction.

¶To whom reprint requests should be addressed. E-mail: yeates@mbi.ucla.edu.

structures whenever possible. Structures were initially matched to genome sequences by using BLAST. If a reliable ($E < 0.000001$) match was not found by BLAST, the search was continued with PSI-BLAST (5, 21). If a reliable ($E < 0.000001$) match was not found by PSI-BLAST, the search was continued by using the method of Sequence Derived Properties (22). In addition, PSI-BLAST matches were verified by using the Method of Sequence Derived Properties.

Identification of Potential Metalloproteins. Proteins whose cysteines were believed to bind metals were excluded from the analysis, because the spatial proximity of cysteine residues in such cases tended to complicate the subsequent analysis. These proteins were either identified by PROSITE (24) patterns of metal-binding sites (e.g., CXXC-CXXC) or by sequence-structure mapping of cysteines onto metal-binding sites. The set of PROSITE motifs used is available at http://www.doe-mbi.ucla.edu/~parag/FOLD_RECOGNITION/DISULFIDES.

Using Sequence-Structure Mapping to Identify Potential Disulfide Bonds. For a pair of cysteines to be classified as a potential disulfide bond, their C-α atoms must be less than 8 Å apart when mapped onto a homologous structure. It was additionally required that each cysteine in the pair be spatially closer to the other member of the pair than to any other cysteine, and that the two cysteines be greater than four residues apart. The upper separation limit of 8 Å was chosen to minimize false negative and false positive predictions of disulfide bonds (data not shown) in calculations on sets of control proteins known to contain or not contain disulfide bonds.

Estimation of Fraction of Cysteines Expected to Form Disulfide Bonds. Within a given set of protein sequences (such as from a complete genome), the value **f** for the fraction of the cysteines expected to form disulfide bonds (reported in Table 1) was obtained from

$$P_{\text{obs}} = f * P_1 + (1 - f) * P_2,$$

in which P_{obs} is the fraction of those cysteines, within the test set of proteins, which satisfied the criterion of being potentially disulfide bonded when mapped onto a homologous structure; P_1 is the fraction of those cysteines known to form disulfide bonds, which passed the criterion for a potential disulfide bond following structural mapping, and P_2 is the fraction of cysteines known to not form disulfide bonds that satisfied the same criterion following structural mapping. The values obtained from the control sets for P_1 and P_2 were 0.70 and 0.08, respectively. To ensure that the calculations on proteins from the control sets were parallel with the calculations on the genomic protein sequences, close homologs of the control proteins were removed from the library of three-dimensional structures, and thus control proteins, like the genomic sequences, were mapped onto structures of sufficiently distant homologs. A detailed examination showed that the majority of false positives were due to metalloproteins within the control set.

The Triangle Method. BLAST was used to identify proteins from *E. coli* with homologs in *P. aerophilum.* These pairs of corresponding proteins were then mapped onto the same three-dimensional structure as described above. Cases were identified for which cysteines (unique to one species) were mapped into close proximity. The method is illustrated in Fig. 3.

Fluorescent Labeling of Free Cysteines. *E. coli* and *P. aerophilum* cells were treated identically throughout the procedure. The fluorescent sulfhydryl-modifying reagent 7-diethylamino-3-(4′-maleimidylphenyl)-4-methylcoumarin (CPM; Molecular Probes) was dissolved in ethanol. Twenty-milligram cell pellets were resuspended in 0.1 ml of Tris/NaCl buffer (0.1 M Tris/0.1 M NaCl, pH 7.0) with 0.5 mg/ml of CPM and incubated for 30 min on ice. The suspension volume was increased to 500 μl with Tris/NaCl buffer and SDS to a final concentration of 1%. The CPM concentration was adjusted to 0.3 mg/ml. Cells were lysed by sonication, and the protein supernatant was clarified by centrifugation at 20,000 $\times$ *g* for 12 min, then divided in two fractions. One fraction was treated with 10 mM Tris(2-carboxyethyl)phosphine (TCEP) to reduce any disulfide bonds present. Both fractions were heated to 65°C for 4 min, then incubated for 20 min at 25°C. The CPM concentration for each sample was adjusted to 0.5 mg/ml followed by incubation at 25°C for 20 min. SDS loading buffer ($\times$2) (nonreducing) was added, and samples were resolved on a 10% SDS-polyacrylamide gel. CPM labeling of protein bands was analyzed by UV excitation and imaged on an AlphaImager (model 2200; Alpha Innotech, San Leandro, CA).

Table 1. The abundance of disulfide bonds in intracellular proteins from various microbial genomes, as estimated from genomic calculations

Organism	°C	G	A	*f*
P. aerophilum	104		✓	0.44
A. pernix	100		✓	0.40
Pyrococcus abyssi	102			0.31
Pyrococcus horikoshii	102			0.28
A. aeolicus	93		✓	0.17
P.Methanobacterium thermoautotrophicum	90			0.15
T. maritima	90			0.13
P.Methanococcus jannaschii	86			0.13
Archaeoglobus fulgidus	92			0.11
Synechocystis PCC6803				0.08
Ureaplasma urealyticum		+		0.07
Mycobacterium tuberculosis		+		0.07
Neisseria meningitides				0.06
Mycoplasma genitalium		+		0.06
Rickettsia prowazekii				0.06
E. coli				0.05
Haemophilus influenzae				0.05
Mycoplasma pneumoniae		+		0.04
Borrelia burgdorferi				0.03
Treponema pallidum				0.03
Helicobacter pylori				0.03
Chlamydia pneumoniae				0.02
Chlamydia trachomatis				0.02
Bacillus subtilis		+		0.01

Abundance, reported as *f* in column 5, is defined as the fraction of the total number of intracellular cysteines that are expected to form disulfide bonds. The archaea have been shaded and are seen to have the highest values of *f*. The maximal growth temperature for thermophiles is given in column 2. Gram positive bacteria (G) are noted with a ''+'' in column 3. Aerobic thermophiles (A) are denoted with a check in column 4. The uncertainty of *f*, 2% on average, was estimated by using counting statistics and standard binomial theory. Estimated intracellular disulfide abundances are notably above zero even for some mesophilic eubacteria. In these cases, many of the disulfides can be traced to redox-related proteins or periplasmic proteins not successfully removed by the automatic filtering procedure.

Results

Unusual Patterns of Cysteine Occurrence in *P. aerophilum*. Two different algorithmic approaches were developed to assess the number of disulfide bonds present in the proteins encoded by a given genome. The first approach, based on protein sequence data alone, reasoned that an organism rich in disulfide bonds might contain a detectable abundance of proteins with an even,

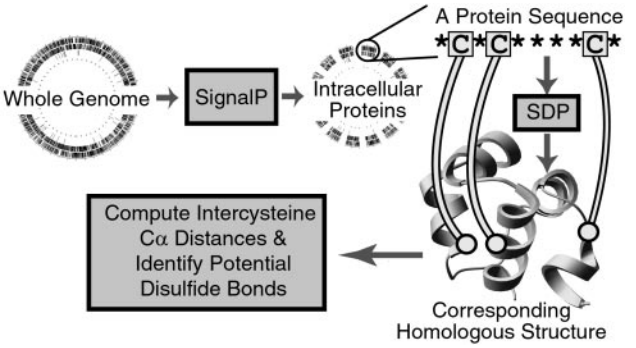

Fig. 2. A sequence-structure mapping method for estimating disulfide abundance in a set of protein sequences. For each protein sequence in a genome, a homologous structure is identified if one is available in the Protein Data Bank, by using BLAST, PSI-BLAST and the Method of Sequence Derived Properties (22, 23). Each sequence is mapped onto its homologous three-dimensional protein structure. The spatial proximity of each pair of cysteines is then examined to determine whether the two could potentially form a disulfide bond. The total disulfide abundance for a genome is estimated statistically by comparison with results on sets of control proteins as described in *Materials and Methods*.

rather than an odd, number of cysteines. As shown in Fig. 1, this simple expectation was borne out by the amino acid sequences of the intracellular proteins from *P. aerophilum*. In *P. aerophilum* (but not in other well-studied microbes), proteins with an even number of cysteines clearly outnumber those with an odd number.

Estimating Disulfide Bond Abundance by Sequence-Structure Mapping. To obtain more direct numerical estimates for the abundance of protein disulfide bonds, a second approach was devised to take advantage of the wealth of available protein structural data (25). If a sufficient number of three-dimensional protein structures were known from every organism of interest, then a direct observation of disulfide abundances would be possible. With the present database of structures, this is not possible. Instead, it was necessary to rely on the availability of "homologous" structures to estimate the likelihood that the cysteines in a set of intracellular protein sequences from a particular organism participate in disulfide bonds. If two cysteines in a query sequence (whose structure *is not* known) are near each other in space when mapped onto a similar protein (whose structure *is* known), then one might infer that the two cysteines of the query protein form a disulfide bond (Fig. 2). However, difficulties that complicate this approach include inaccuracies that may arise from limitations of the mapping algorithm and from structural divergence of the two proteins. To circumvent these obstacles, a statistical method was devised in which the algorithm was first applied to two sets of control proteins: one whose cysteines were known to be involved in disulfide bonds, and one whose cysteines were known to be in the free sulfhydryl form. The use of control sets of proteins made it possible to judge whether two cysteines might be involved in a disulfide bond. By comparing the results of calculations on a set of proteins from a particular genome (or from a eukaryotic cell organelle) to the results from the control calculations, it was possible to obtain numerical estimates of disulfide abundances (see *Materials and Methods*). As a further point of comparison, a Triangle Method, also based on sequence-structure mapping (Fig. 3, Table 2), was devised to identify a set of protein pairs, each composed of an *E. coli* protein that cannot form a disulfide bond and a homologous protein from *P. aerophilum* that can possibly form a disulfide bond.

Control for Metalloprotein False Positives. As a precaution, measures were taken to prevent metalloproteins, which often contain multiple cysteine residues in proximity, from being falsely identified as disulfide-containing proteins. All of the protein sequences from *P. aerophilum* and from all other microbial genomes were scanned for the presence of ProSite motifs for metal-binding sites. Protein sequences containing such motifs were excluded from further analysis. In any case, protein sequences from *P. aerophilum* showed no higher abundance of metal-binding motifs and no unusual tendency to align to metalloprotein structures.

Organisms Found with Abundant Intracellular Disulfide Bonds. An analysis of the protein coding regions of 25 fully sequenced genomes revealed several genomes with a striking fraction of their total number of cysteines expected to participate in intracellular disulfide bonds (Table 1). Nine of the 25 genomes analyzed are predicted to contain greater than 10% of their intracellular cysteines within disulfide bonds. This group of nine genomes includes all seven archaea examined and both of the thermophilic eubacteria. Among this group, four organisms stand out. Two species of *Pyrococcus* have approximately 30% of their cysteines in disulfide bonds, whereas two other archaea, *P. aerophilum* and *Aeropyrum pernix*, are predicted to have an

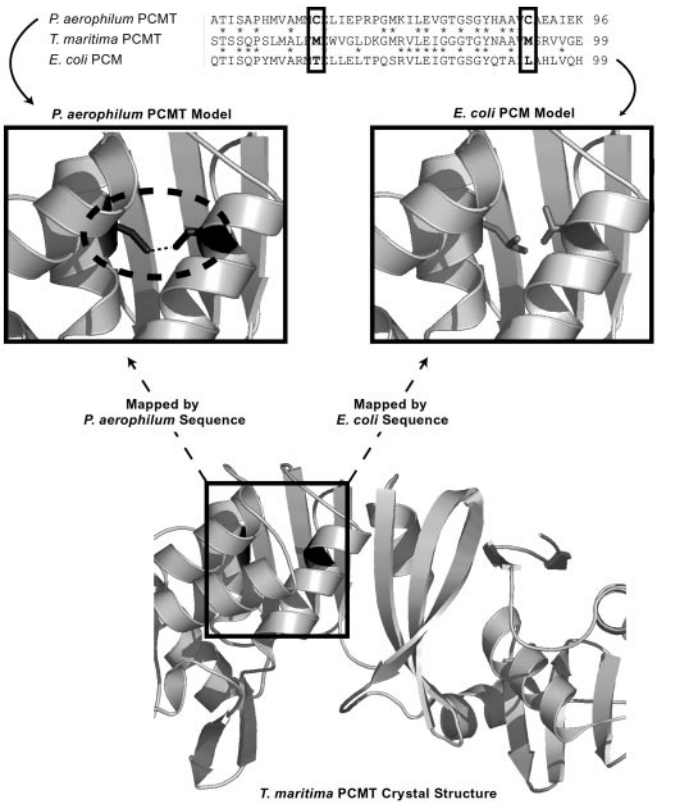

Fig. 3. The Triangle Method for searching for possible disulfide bonds in proteins from *P. aerophilum*. Triplets of similar proteins are found by BLAST, one of which is from *P. aerophilum*, a second from *E. coli*, and the third of known structure. Here the protein of known structure is PCMT from *Thermotoga maritima* (*Lower Center*). The three sequences align convincingly (fragment of alignment shown at the top). Two additional cysteine residues present in the *P. aerophilum* sequence are indicated by boxes. The model on the upper left shows that the *P. aerophilum* sequence, when mapped on the *T. maritima* structure, places these two cysteine residues (circled, shown in black) close enough to form a disulfide bond. In the *E. coli* model (*Upper Right*), neither of the corresponding residues is cysteine, and no disulfide bond can form here or anywhere else in the *E. coli* enzyme.

Table 2. A selected set of 10 proteins from *P. aerophilum* predicted by the Triangle Method to have disulfide bonds, contrasted with homologs from *E. coli* that do not have disulfide bonds

PA ID	*E. coli* ID	PA/EC *E* value	PDB match	PA/PDB *E* value	EC/PDB *E* value
pag5_1383	ecoli-1786220	7.00E-24	3hdha	0	0
pag5_2460	ecoli-1786809	2.00E-24	1qdla	0	0
pag5_2582	ecoli-1788869	2.00E-10	1rhd	0	1.43E-08
pag5_2887	ecoli-1790398	5.00E-40	1aosa	0	0
pag5_3221	ecoli-1786236	9.00E-27	1yub	7.89E-10	1.33E-08
pag5_3343	ecoli-1788184	1.00E-29	1bs2a	0	0
pag5_3383	ecoli-1790383	1.00E-10	1dik	2.59E-14	0
pag5_3453	ecoli-1790288	5.00E-10	1e5ka	0	4.16E-09
pag5_3596	ecoli-1789825	3.00E-18	1qmhb	0	0
pag5_63	ecoli-1786319	1.00E-22	1g291	5.97E-14	0

The first two columns give the identification numbers (ID) for the genome sequences. The third column gives the BLAST *E* values between the two sequences, showing high similarities (low scores). Columns 5 and 6 give the BLAST *E* values between the *P. aerophilum* (PA) and *E. coli* (EC) sequences, respectively, and the sequence of the protein of known structure identified in column 4. Again, the *E* values are small, indicating structural similarity among all three proteins. In each case, the *P. aerophilum* protein contains at least one pair of cysteine residues whose Cα atoms lie in proximity when mapped onto the Protein Data Bank (PDB) structure. This suggests that each *P. aerophilum* protein may form a disulfide bond in an oxidizing environment. The corresponding *E. coli* proteins cannot form disulfide bonds, because they lack spatially proximal cysteine residues.

extraordinary fraction of their cysteines in disulfide bonds, 40 and 44%, respectively. Structural studies already provide confirmation in specific cases (26).

Fluorescent Labeling of Free Cysteines Confirms the Presence of Disulfide Bonds Within the Proteins of *P. aerophilum*. To experimentally verify our computational predictions, *P. aerophilum* was chosen as a test organism. Proteins from cell lysate were denatured in either the presence or absence of the reducing agent TCEP, and free cysteine residues were then fluorescently labeled by modification with the thiol-specific reagent CPM. TCEP does not contain a free thiol and therefore has limited reactivity with CPM, eliminating the need to remove the reducing agent before labeling. CPM gains fluorescence on reacting with the free sulfhydryl groups present on cysteines not involved in disulfide bonds. To get an accurate profile of free cysteines present under natural conditions, proteins were exposed to CPM before and throughout denaturation to prevent any naturally free cysteines from forming nonspecific disulfide bonds. Additional CPM was added after reduction to label cysteines that were originally involved in disulfide bonds and therefore inaccessible to labeling before reduction. Comparison of the banding pattern of labeled *P. aerophilum* proteins from nonreduced and reduced samples after electrophoresis reveals many bands labeled in the reduced sample that are unlabeled in the nonreduced sample (Fig. 4). The absence of corresponding fluorescent bands in the nonreduced sample indicates the presence of disulfide bonds in these proteins. This experimental result clearly supports the conclusions of the computational analysis.

Various experimental controls were also performed. Parallel experiments on *E. coli* cells showed very little difference between the labeling of reduced and nonreduced samples. In both species, the protein composition of reduced and nonreduced samples was found to be nearly identical by Coomassie staining (not shown). Finally, a *P. aerophilum* sample that was reduced but then not subsequently exposed to fresh label was indistinguishable on a gel from a nonreduced sample, showing that the differing appearance of labeled bands in gels of reduced and nonreduced samples is not somehow the result of changes in mobility.

Eukaryotic Organelle Disulfide Abundances. To extend our analysis to eukaryotic cells, we performed a similar prediction of disulfide abundance for proteins from different subcellular locations within yeast [as defined by the Yeast Protein Database (27)]. Despite the small sample sizes, disulfide abundances could be estimated for a few subcellular compartments. We found both extracellular proteins and proteins localized to the lysosome showed a high disulfide fraction of greater than 60%, consistent with our present knowledge of organellar and extracellular chemical environments. The nucleus and cytoplasm showed only

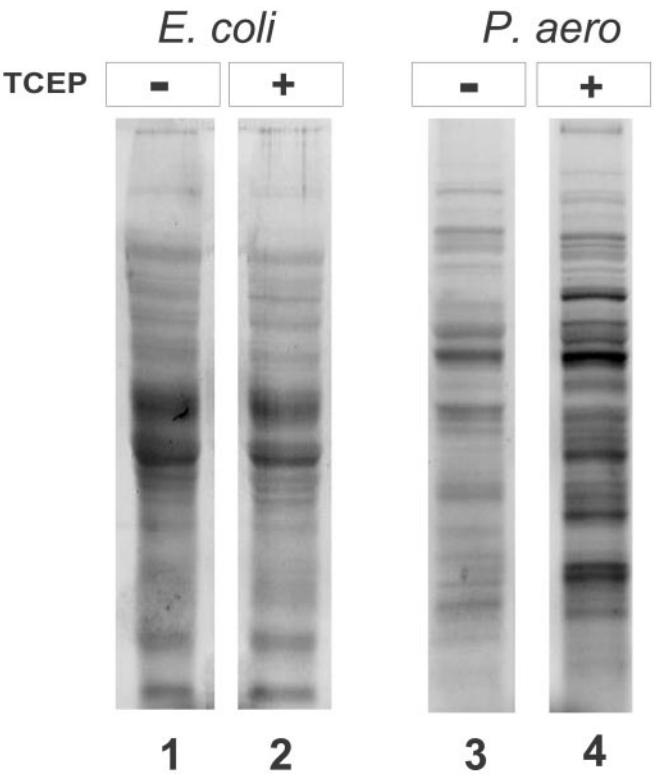

Fig. 4. A fluorescent image of an SDS/PAGE gel showing abundant disulfide bonds in the proteins of *P. aerophilum*. Cysteine residues in their free sulfhydryl form were selectively labeled with the fluorescent reagent, CPM, either in the absence of (lanes 1 and 3) or after (lanes 2 and 4) reduction of disulfides with TCEP. *E. coli* was chosen as a control organism and shows very few protein disulfide bonds, as expected (lanes 1 and 2). The negative of the image is shown for improved visualization of labeled bands.

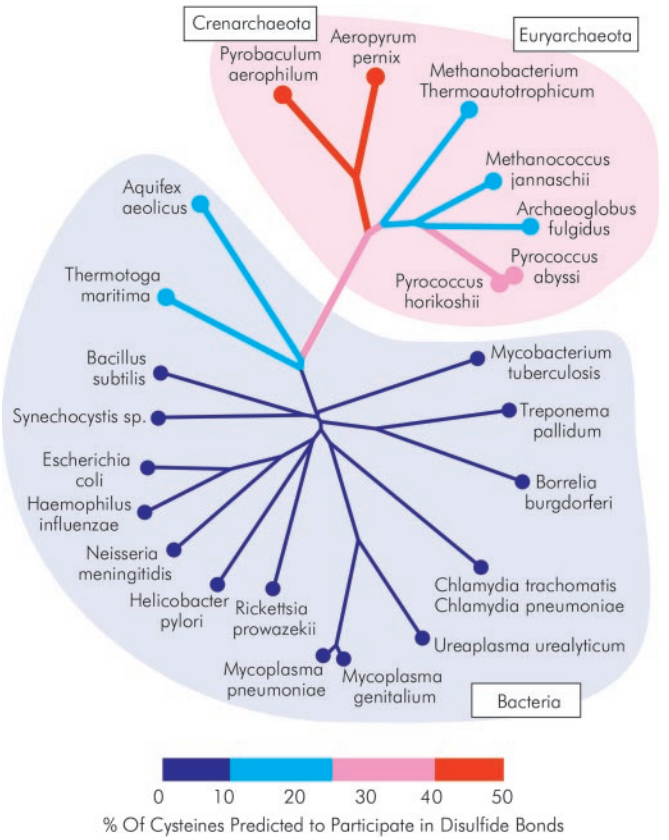

Fig. 5. An unrooted phylogenetic tree showing the 25 genomes whose disulfide bond abundances were estimated computationally. The thickness and color of each branch illustrate the predicted disulfide content of the microbe at the branch's terminus according to the key (bottom); the coloring of internal branches reflects an average over multiple species. Archaeal organisms lie within the pink region. Bacterial genomes lie within the light blue region. The crenarchaeal branches of the tree (deep red, thickest) have the greatest predicted disulfide content, exceeding 40%. The majority of bacteria have low predicted disulfide contents ranging from 0 to 10% (deep blue, thin lines). The thermophilic eubacteria, *Aquifex* and *Thermotoga*, have predicted disulfide contents in the 10–20% range, whereas the euryarchaea have values ranging from 10 to 30%. Other methods of construction can lead to phylogenetic trees that differ in detail from the one shown here without affecting the salient features.

trace amounts of cysteines in disulfide bonds, and the mitochondrion showed a slightly greater abundance of disulfides.

Discussion

Which Organisms Have Abundant Intracellular Disulfide Bonds? The crenarchaea are richest in disulfide bonds. In a phylogenetic tree [constructed from aligned 16s-RNA sequences from the Ribosomal Database Project (28)], the archaeal branch of the tree clearly shows a greater abundance of disulfides than the eubacterial branch (Fig. 5; Table 1). Furthermore, there is a clear distinction between the crenarchaeal and euryarchaeal subbranches of the archaea, because the disulfide abundances of the crenarchaeota *P. aerophilum* and *A. pernix* are substantially higher than the euryarchaea. The observed difference between these two archaeal subbranches might be due in part to the oxygen-tolerant nature of the two crenarchaeal microbes (29, 30). Note that intracellular disulfides do not appear to be solely an archaeal phenomenon. The disulfide content of the two thermophilic eubacteria, *Aquifex aeolicus* and *Thermotoga maratima*, groups these organisms more closely with the thermophilic archaea than with the other (mesophilic) eubacteria. Even within the thermophilic organisms, there is a gross correlation between disulfide abundance and maximal growth temperature (Table 1). These observations support a hypothesis that intracellular disulfides are a result of selective pressure for thermostable proteins. When more genomic information becomes available about mesophilic archaea, especially mesophilic anaerobic crenarchaea, it may be possible to say whether the presence of intracellular disulfide bonds is a characteristic of all archaea or an adaptation to high temperature.

Biochemical Implications of Intracellular Disulfide Bonds. The apparent abundance of intracellular disulfide bonds in some organisms raises numerous questions. What is the redox nature of these intracellular environments? Do the disulfide bonds stabilize these proteins? How did the disulfide bonds evolve? As discussed above, disulfide bonds in intracellular proteins had been thought to be rare, in agreement with the reductive nature of the cytoplasm in organisms that have been well characterized. However, the distinct cellular environments of thermophilic microorganisms have only begun to be elucidated. For example, the well-studied glutathione system of redox regulation operates only in the purple bacteria and the cyanobacteria. The archaea (31) and many other microbes use a variety of other thiol compounds (32, 33) and unusual cofactors, few of which have been studied in detail. Furthermore, some key redox-related enzymes, such as catalase, are notably absent from several thermophilic microbes (34, 35). Because of potential differences in redox environment or redox mechanisms, disulfide bonds in the intracellular proteins of these microbes could be energetically stabilizing. The recently determined sequences and three-dimensional structures of numerous proteins from moderate and extreme thermophiles have shed light on several factors that are used in different combinations and to different degrees by various proteins to increase their thermal stability. Stability mechanisms that have been implicated most often are increased ion pairing, hydrogen bonding, and compactness (36–38). The present study argues strongly that disulfide bonds are another important mechanism for achieving thermostability.

Conclusion

The microbial results presented above argue for the existence of intracellular disulfide bonds within several archaeal and thermophilic genomes. Furthermore, the pattern of disulfide abundance across the tree of life argues that these intracellular disulfide bonds play a role in thermostability. The present study illustrates the power of integrating genomic data with protein structure and function to illuminate the chemistry and biology of unusual organisms. Additional genomic data and further experimental studies will be required to explore more fully the significance of the abundant disulfide bonds revealed in these unusual microbes.

We thank Jeffrey Miller, Sorel Fitz-Gibbon, Tom Graeber, Evan Gamble, Rob Grothe, and Jamil Momand for helpful discussions, and Sarah Yohannan (Univ. of California, Los Angeles) for *P. aerophilum* cells. This work was supported by the U.S. Department of Energy Office of Biological and Environmental Research, the National Science Foundation (P.M.), and the National Institutes of Health (T.O.Y. and D.R.B.).

1. Marcotte, E. M., Pellegrini, M., Ng, H. L., Rice, D. W., Yeates, T. O. & Eisenberg, D. (1999) *Science* **285,** 751–753.
2. Lin, J. & Gerstein, M. (2000) *Genome Res.* **10,** 808–818.
3. Fraser, C. M., Eisen, J., Fleischmann, R. D., Ketchum, K. A. & Peterson, S. (2000) *Emerg. Infect. Dis.* **6,** 505–512.
4. Sunyaev, S., Lathe, W. & Bork, P. (2001) *Curr. Opin. Struct. Biol.* **11,** 125–130.
5. Wolf, Y. I., Brenner, S. E., Bash, P. A. & Koonin, E. V. (1999) *Genome Res.* **9,** 17–26.
6. Thornton, J. M. (1981) *J. Mol. Biol.* **151,** 261–287.
7. Gilbert, H. F. (1990) *Adv. Enzymol. Relat. Areas Mol. Biol.* **63,** 69–172.

8. Hwang, C., Sinskey, A. J. & Lodish, H. F. (1992) *Science* **257,** 1496–1502.
9. Prinz, W. A., Aslund, F., Holmgren, A. & Beckwith, J. (1997) *J. Biol. Chem.* **272,** 15661–15667.
10. Choi, H., Kim, S., Mukhopadhyay, P., Cho, S., Woo, J., Storz, G. & Ryu, S. (2001) *Cell* **105,** 103–113.
11. Jakob, U., Muse, W., Eser, M. & Bardwell, J. C. (1999) *Cell* **96,** 341–352.
12. Toth, E. A., Worby, C., Dixon, J. E., Goedken, E. R., Marqusee, S. & Yeates, T. O. (2000) *J. Mol. Biol.* **301,** 433–450.
13. Fitz-Gibbon, S. T., Ladner, H., Kim, U. J., Stetter, K. O., Simon, M. I. & Miller, J. H. (2002) *Proc. Natl. Acad. Sci. USA* **99,** 984–989.
14. Jiang, Y., Nock, S., Nesper, M., Sprinzl, M. & Sigler, P. B. (1996) *Biochemistry* **35,** 10269–10278.
15. Maes, D., Zeelen, J. P., Thanki, N., Beaucamp, N., Alvarez, M., Thi, M. H., Backmann, J., Martial, J. A., Wyns, L., Jaenicke, R. & Wierenga, R. K. (1999) *Proteins* **37,** 441–453.
16. DeDecker, B. S., O'Brien, R., Fleming, P. J., Geiger, J. H., Jackson, S. P. & Sigler, P. B. (1996) *J. Mol. Biol.* **264,** 1072–1084.
17. Chiu, H. J., Johnson, E., Schroder, I. & Rees, D. C. (2001) *Structure (Cambridge, U.K.)* **9,** 311–319.
18. Meyer, J., Clay, M. D., Johnson, M. K., Stubna, A., Munck, E., Higgins, C. & Wittung-Stafshede, P. (2002) *Biochemistry* **41,** 3096–3108.
19. Nielsen, H., Brunak, S. & von Heijne, G. (1999) *Protein Eng.* **12,** 3–9.
20. Eisenberg, D., Schwarz, E., Komaromy, M. & Wall, R. (1984) *J. Mol. Biol.* **179,** 125–142.
21. Altschul, S. F., Madden, T. L., Schaffer, A. A., Zhang, J., Zhang, Z., Miller, W. & Lipman, D. J. (1997) *Nucleic Acids Res.* **25,** 3389–3402.
22. Mallick, P., Goodwill, K. E., Fitz-Gibbon, S., Miller, J. H. & Eisenberg, D. (2000) *Proc. Natl. Acad. Sci. USA* **97,** 2450–2455.
23. Fischer, D. & Eisenberg, D. (1996) *Protein Sci.* **5,** 947–955.
24. Falquet, L., Pagni, M., Bucher, P., Hulo, N., Sigrist, C. J., Hofmann, K. & Bairoch, A. (2002) *Nucleic Acids Res.* **30,** 235–238.
25. Berman, H. M., Westbrook, J., Feng, Z., Gilliland, G., Bhat, T. N., Weissig, H., Shindyalov, I. N. & Bourne, P. E. (2000) *Nucleic Acids Res.* **28,** 235–242.
26. Cort, J. R., Mariappan, S. V., Kim, C. Y., Park, M. S., Peat, T. S., Waldo, G. S., Terwilliger, T. C. & Kennedy, M. A. (2001) *Eur. J. Biochem.* **268,** 5842–5850.
27. Costanzo, M. C., Crawford, M. E., Hirschman, J. E., Kranz, J. E., Olsen, P., Robertson, L. S., Skrzypek, M. S., Braun, B. R., Hopkins, K. L., Kondu, P., *et al.* (2001) *Nucleic Acids Res.* **29,** 75–79.
28. Maidak, B. L., Cole, J. R., Lilburn, T. G., Parker, C. T., Jr., Saxman, P. R., Farris, R. J., Garrity, G. M., Olsen, G. J., Schmidt, T. M. & Tiedje, J. M. (2001) *Nucleic Acids Res.* **29,** 173–174.
29. Volkl, P., Huber, R., Drobner, E., Rachel, R., Burggraf, S., Trincone, A. & Stetter, K. O. (1993) *Appl. Environ. Microbiol.* **59,** 2918–2926.
30. Kawarabayasi, Y., Hino, Y., Horikawa, H., Yamazaki, S., Haikawa, Y., Jin-no, K., Takahashi, M., Sekine, M., Baba, S., Ankai, A., *et al*. (1999) *DNA Res.* **6,** 83–101, 145–152.
31. Fahey, R. C. & Sundquist, A. R. (1991) *Adv. Enzymol. Relat. Areas Mol. Biol.* **64,** 1–53.
32. Spies, H. S. & Steenkamp, D. J. (1994) *Eur. J. Biochem.* **224,** 203–213.
33. Sundquist, A. R. & Fahey, R. C. (1989) *J. Biol. Chem.* **264,** 719–725.
34. Graham, D. E., Kyrpides, N., Anderson, I. J., Overbeek, R. & Whitman, W. B. (2001) *Methods Enzymol.* **330,** 40–123.
35. Deckert, G., Warren, P. V., Gaasterland, T., Young, W. G., Lenox, A. L., Graham, D. E., Overbeek, R., Snead, M. A., Keller, M., Aujay, M., *et al.* (1998) *Nature (London)* **392,** 353–358.
36. Jaenicke, R. & Bohm, G. (1998) *Curr. Opin. Struct. Biol.* **8,** 738–748.
37. Thompson, M. J. & Eisenberg, D. (1999) *J. Mol. Biol.* **290,** 595–604.
38. Scandurra, R., Consalvi, V., Chiaraluce, R., Politi, L. & Engel, P. C. (2000) *Front. Biosci.* **5,** D787–D795.