Global analysis of predicted proteomes: Functional adaptation of physical properties

Christopher G. Knight*†‡, Rees Kassen*§, Holger Hebestreit†¶, and Paul B. Rainey*‖

*Department of Plant Sciences, University of Oxford, South Parks Road, Oxford OX1 3RB, United Kingdom; †Glycobiology Institute, Department of Biochemistry, University of Oxford, South Parks Road, Oxford OX1 3QU, United Kingdom; §Department of Biology and Center for Advanced Research in Environmental Genomics, University of Ottawa, Ottawa, ON, Canada K1N 6N5; ¶RMF Dictagene SA, 4, Chemin de la Vulliette, CH-1000 Lausanne 25, Switzerland; and ‖School of Biological Sciences, University of Auckland, P.O. Box 92019, Auckland, New Zealand

Edited by Samuel Karlin, Stanford University, Stanford, CA, and approved April 8, 2004 (received for review November 7, 2003)

The physical characteristics of proteins are fundamentally important in organismal function. We used the complete predicted proteomes of >100 organisms spanning the three domains of life to investigate the comparative biology and evolution of proteomes. Theoretical 2D gels were constructed with axes of protein mass and charge (pI) and converted to density estimates comparable across all types and sizes of proteome. We asked whether we could detect general patterns of proteome conservation and variation. The overall pattern of theoretical 2D gels was strongly conserved across all life forms. Nevertheless, coevolved replicons from the same organism (different chromosomes or plasmid and host chromosomes) encode proteomes more similar to each other than those from different organisms. Furthermore, there was disparity between the membrane and nonmembrane subproteomes within organisms (proteins of membrane proteomes are on the average more basic and heavier) and their variation across organisms, suggesting that membrane proteomes evolve most rapidly. Experimentally, a significant positive relationship independent of phylogeny was found between the predicted proteome and Biolog profile, a measure associated with the ecological niche. Finally, we show that, for the smallest and most alkaline proteomes, there is a negative relationship between proteome size and basicity. This relationship is not adequately explained by AT bias at the DNA sequence level. Together, these data provide evidence of functional adaptation in the properties of complete proteomes.

In silico studies on the evolution and biology of proteomes encoded by sequenced genomes have focused either on functional annotation (1, 2) or total amino acid composition (3, 4). The first relies on homology-based extrapolation, which is necessarily limited and biased by subjects of interest to molecular biologists. The second can overlook many of the biologically interesting features of the encoded proteins. The middle ground, namely the analysis of simple properties of all of the proteins predicted from fully sequenced genomes, has received far less attention.

Analyses at levels above sequence composition but below function have proven fruitful in both DNA sequence analysis and practical proteomics. For DNA sequences this has included analyses of "genome signature" (5) or codon bias to predict highly expressed genes (6). Practical proteomics relies on mass spectrometry and 2D gel electrophoresis that analyze simple physical properties of proteins and peptides: mass and charge. Mass and charge may also be predicted from raw protein sequence. Although either property may be modified posttranslationally, estimates from raw sequence are typically precise and accurate (7) in a way that functional annotation cannot be. These properties are also related to the biological role of proteins in a way that proteome-wide amino acid compositions cannot be: Charge affects the function of proteins (e.g., positively charged histones binding the negatively charged DNA backbone). Protein mass may itself be a target for natural selection in minimizing metabolic costs of production (8).

The mass and charge (pI) of proteins predicted from a complete genome may be represented in a "theoretical 2D gel" (Fig. 1*A*). These plots have been used to assess the performance of practical 2D gels (9). However, biological information is also embodied in these distributions: e.g., the link between the acidic pI distribution of the predicted proteome of the bacterium *Halobacterium* strain NRC-1 and its high salt environment (10). Similarly, basic pI distributions have been linked with thermophily in the archaea (11) and adaptation to an acidic environment in *Helicobacter pylori* (12).

Through the comparison of theoretical 2D gels from >100 sequenced genomes, we demonstrate biologically significant interspecific variation. We partition this variation by subcellular location, demonstrating both different characteristics and different variation in membrane and nonmembrane subproteomes. We also demonstrate and define experimentally a relationship with Biolog profile, a measure of ecological niche, for a range of bacteria. Finally we reveal a relationship between proteome size and pI for the smallest proteomes that is not explained by current hypotheses.

Materials and Methods

Data Set. All 103 complete proteome sets available through the European Bioinformatics Institute (www.ebi.ac.uk/proteome) in early 2003, subdivided into the smallest available subset greater than individual proteins (usually chromosome or plasmid), were used (Table 1 and Table 2, which is published as supporting information on the PNAS web site). We define the proteome of an organism as the collection of all proteins from all subsections of the genome, including plasmids but excluding plastids, if any.

Sequence Analysis. pI was calculated by using a standard iterative algorithm (13, 14). Genetic distances were calculated from small subunit RNA sequences obtained from the National Center for Biotechnology Information, aligned by using CLUSTALW (15), and analyzed with PHYLIP (16). Other analyses were carried out in R1.5.0 (17) by using scripts written in S and PERL and in JMP 5.01 (18).

Theoretical 2D Gel Analysis. Two-dimensional normal kernel density estimates were used to convert the scatter plot of a theoretical 2D gel (log molecular mass vs. pI) into an estimate of spot density at every point on a 200 × 200 grid from pI 2 to 14 and molecular mass 10^2 to 10^7 Da. Integral to this approach is the choice of a smoothing parameter (bandwidth, h). Rather than fixing this arbitrarily, an optimized value was found in each dimension (19). The similarity of any two theoretical 2D gels was

This paper was submitted directly (Track II) to the PNAS office.

See Commentary on page 8257.

‡To whom correspondence should be addressed. E-mail: christopher.knight@plants.ox.ac.uk.

www.pnas.org/cgi/doi/10.1073/pnas.0307270101

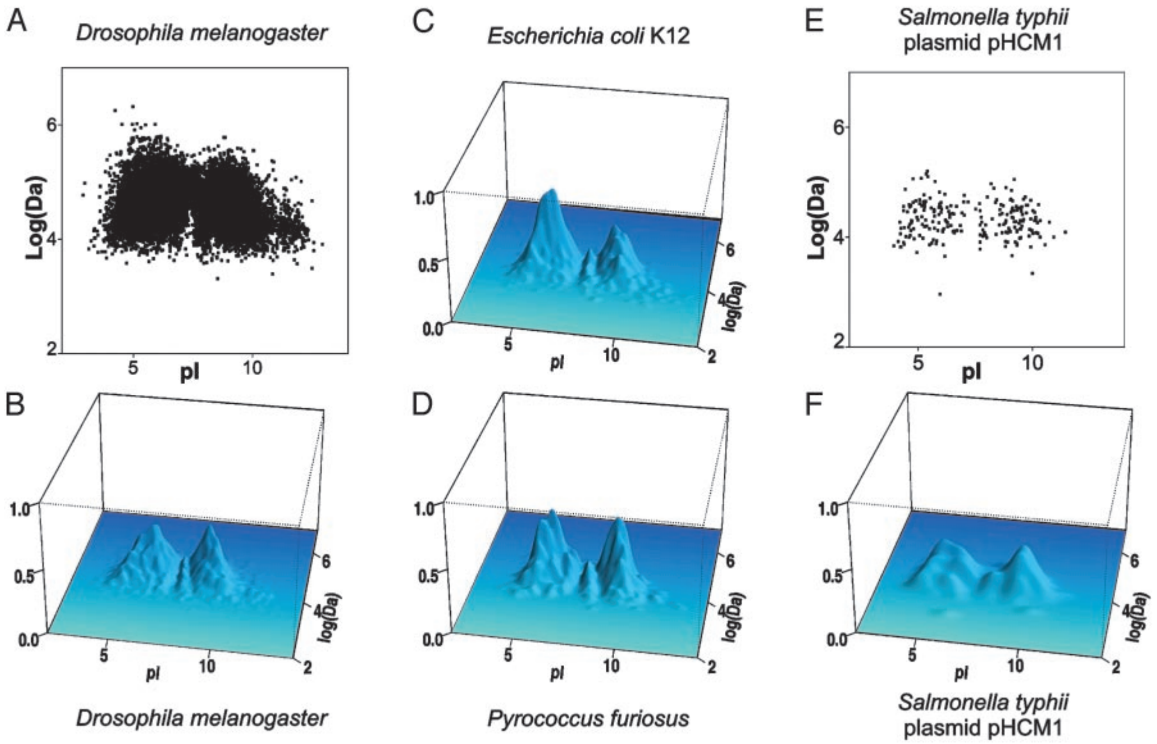

Fig. 1. Theoretical 2D gels. (*A*) Scatter-plot (theoretical 2D gel) of the fruitfly *Drosophila melanogaster.* (*B*) The corresponding density plot as used in the analyses. Similar patterns are shown by bacteria (*C*), archaea (*D*), and plasmids containing orders of magnitude fewer proteins (*E* and *F*).

measured as the rank correlation of all points on the grid for each plot. Although still sensitive to pattern, the rank correlation is less sensitive than alternatives to variations in the spikiness of the distribution caused by proteomes of different sizes being smoothed differently.

Comparative Analysis. Rank correlations were arcsine transformed before analysis. We used randomizations to test whether an observed difference was greater than would be expected by chance alone. The observed value of the difference was compared to the distribution of values obtained from the randomization. Two-tailed tests were used to obtain *P* values. Fig. 3 shows three differences tested in this way: (*i*) The list of eukaryotic chromosomes was reassigned to organisms at random, each organism getting the appropriate number of chromosomes. The pairwise correlations among all chromosomal proteomes were then split into two groups, those within and those between the newly constituted organisms. The difference between the averages of those two groups was recorded. This procedure was repeated 10,000 times. (*ii*) For organisms containing one or more plasmids, the list of chromosome origins (i.e., a list of organisms) was paired at random with a similar list of plasmid origins. Each pairwise correlation between a chromosome and a plasmid proteome was then assigned to one of two groups based on whether their origins were paired or not. The difference between the average of these two groups was recorded. This procedure was repeated 10,000 times. (*iii*) All 24 permutations of pairings between the list of mitochondrial hosts and the list of mitochondria were made. In each case, the correlation between a chromosomal and a mitochondrial proteome was assigned to one of two groups based on whether the host and mitochondrion were paired or not. The difference between the average of these two groups was recorded in each case.

To compare theoretical 2D gels and Biolog profiles (see below), differences were measured as phylogenetically independent contrasts (20). These contrasts are comparisons between the taxa branching from a node in a dichotomous phylogeny. The phylogeny used is shown in Fig. 7, which is published as supporting information on the PNAS web site. This calculation entails estimating traits (theoretical 2D gel and Biolog profile) for taxa containing several strains. For Biolog profile we started from the smallest taxa and averaged the raw data in each replicate for the two taxa branching from each node. The difference between the Biolog profile of two taxa was calculated as the average across replicates of the Euclidean distance between the profiles, following MacLean and Bell (21). Estimates of the predicted proteomes for taxa containing several strains were constructed, starting from the smallest taxa, by pooling the data for an equal number of proteins from the two taxa branching from each node. Two taxa never had the same number of proteins; thus, all of them were used from the taxon with fewer proteins and a random selection (without replacement) of the same size was taken from the proteins of the other taxon. The resulting predicted proteomes were analyzed and compared as above.

Table 1. Numbers of predicted proteomes used in the analyses

	Eukaryotes	Bacteria	Archaea	Totals
Chromosome	44	84	16	144
Plasmid/plastid	5	53	4	62
Complete*	3	0	0	3
Totals	52	137	20	209

*The human, mouse, and *C. elegans* databases that were used were not subdivided into chromosomes.

Separation of Membrane Proteins. HMMTOP2 (22) was used to identify membrane-spanning domains. We followed the ap-

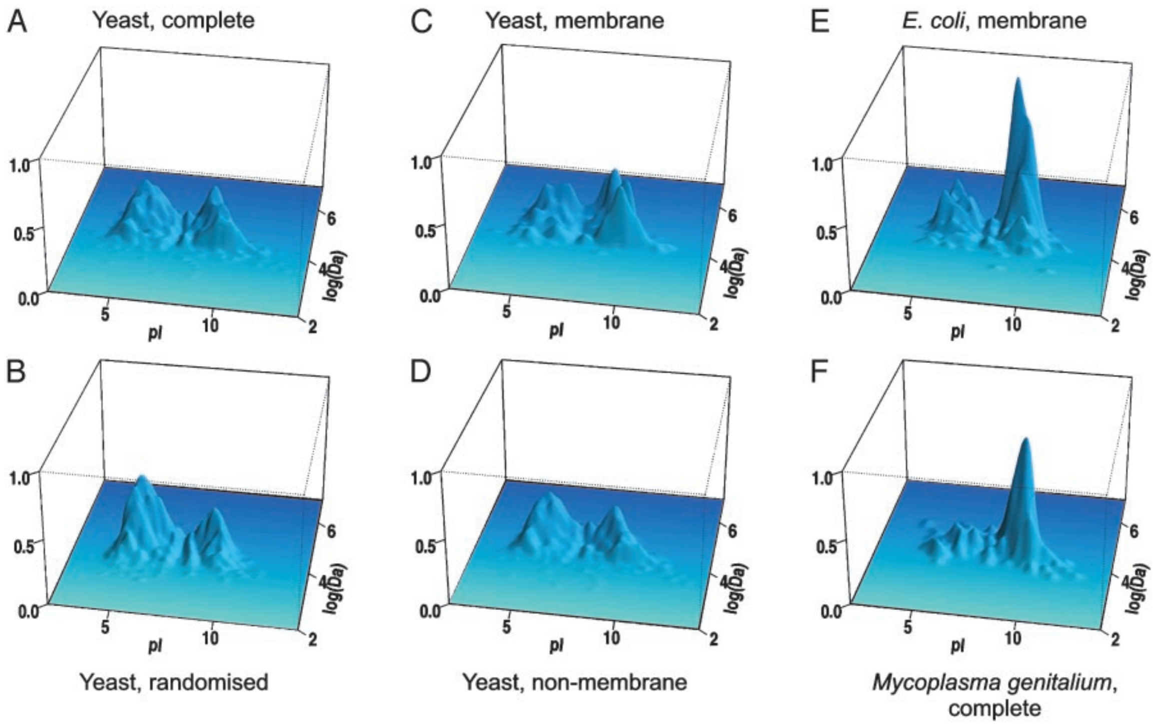

Fig. 2. Alternative theoretical 2D gels. (*A*) Typical two-winged theoretical 2D gel showing yeast (*S.* cerevisiae). (*B*) Randomization of amino acids among yeast proteins. The distribution is similarly bimodal, but significantly different, particularly in its narrower spread of pI. (*C* and *D*) Division of the yeast proteome into membrane and nonmembrane subsections. As with all other proteomes, the membrane proteome is more basic. (*E*) The *E. coli* membrane subproteome showing much greater basic bias than in yeast. (*F*) One of the smallest proteomes showing a similarly basic biased proteome.

proach of Wallin and von Heijne (23) and omitted from subsequent analysis proteins with a single membrane-spanning domain (19 ± 0.6%, 95% confidence interval).

Determination of Biolog Profile. Eleven bacteria were used: two strains of *Agrobacterium tumefaciens* C58 (sequenced by the University of Washington, Seattle, and Cereon Genomics, Cambridge, MA), two independently archived versions of the sequenced strain of *Bacillus subtilis* 168 (1A1 and 1A700 from the *Bacillus* Genetic Stock Center), *Caulobacter crescentus* CB15, *Escherichia coli* K-12, *Lactococcus lactis lactis* IL1403, *Pseudomonas aeruginosa* PA01, *Pseudomonas putida* KT2440, *Shewanella oneidensis* MR-1, and *Rhizobium meliloti*. Each strain's ability to metabolize 95 different carbon sources was assayed by using Biolog GN2 plates (Hayward, CA) by following protocol established by MacLean and Bell (21). Growth was assessed after 24 h as absorbance at 600 nm. The Biolog profile comprised these values corrected to the blank well containing no substrate. Three independent replicates were performed for each strain.

Results and Discussion

Pattern: Predicted Proteomes Show Broad Similarities and Evolutionarily Relevant Variation. The distribution of predicted proteins on a theoretical 2D gel (pI vs. log molecular mass, not including predictions of abundance) was calculated for 103 fully sequenced genomes. This distribution shows broadly similar "butterfly" patterns in all three domains of life: a unimodal mass distribution with large acidic and basic "wings" and a lower "body" peak at pI ≈ 8 (Fig. 1 *B*–*D* and Fig. 8, which is published as supporting information on the PNAS web site). This pattern is visible over more than two orders of magnitude variation in total protein numbers (Fig. 1, compare *A* and *B* with *E* and *F*).

Proximately, the bimodality in pI distribution is due to the preponderance of strongly acidic and basic residues (Asp, Glu, Lys, Tyr, and Arg) over ones with a pK close to 7 (His and Cys) as shown by Kawashima et *al* (11). The bimodality may also be related to the difficulty of maintaining protein structure and solubility near cytoplasmic pH (9). These proximate explanations raise a deeper question: Are theoretical 2D gels best understood as the aggregate properties of their component amino acids, or is the distribution of amino acids among proteins important? To test these alternatives we made at least 2,000 amino acid sequence randomizations for each of three representative proteomes: yeast (*Saccharomyces cerevisiae*), a eukaryote with balanced acidic and basic wings of its theoretical 2D gel; *Halobacterium* strain NRC-1, an archaeon with an acidic proteome; and *Buchnera aphidicola* (*Schizaphis gramium*), a bacterium with a basic proteome. Total protein numbers and lengths were retained, but the complete set of the amino acids of the proteome was reassigned to proteins at random. The randomized distributions appeared bimodal like the true distributions (Fig. 2, compare *A* with *B*). However, for each organism the correlation of the true distribution with every single one of the randomized distributions was lower than the average correlation among an independent group of 50 randomized proteomes. This result suggests that the distribution of proteins in a theoretical 2D gel is not merely a function of global amino acid composition. Interestingly, the principal difference between the randomized and true distributions was the spread of the pI distribution, which was wider in the true distribution than in any of the randomizations (Fig. 2, compare *A* with *B*), a result consistent with functional specialization of proteins according to charge.

To what extent are genetic differences among species reflected in the divergence of their 2D gels? A broad scale comparison of simple features of theoretical 2D gels (average and spread of protein mass and pI, positions of the acidic and basic wings of the distributions) across the domains of life is shown in Figs. 9 and

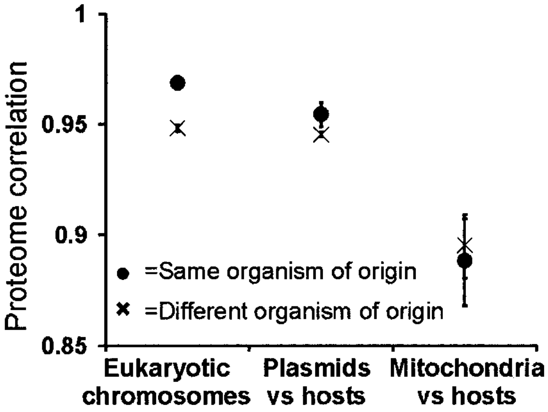

Fig. 3. Proteome subsets compared within and between organisms. Points are means and SE bars of comparisons between all pairs of proteomes. For the plasmid and mitochondrial data, all comparisons between proteomes from a particular pair of organisms were averaged before inclusion.

10, which are published as supporting information on the PNAS web site. There is little systematic differentiation between domains, except the known phenomenon that eukaryotes tend to average heavier proteins (24). To explore the effect of phylogenetic relatedness on proteome complement in more detail, we considered the relationship of genetic distance between strains to the similarity of their theoretical 2D gels. For pairwise combinations of organisms in our database, genetic distance was estimated by using small subunit RNA sequences (Table 3, which is published as supporting information on the PNAS web site). It is striking that there seems to be very little quantitative relationship between this and 2D gel similarity (Fig. 11 and Tables 3 and 4, which are published as supporting information on the PNAS web site). The most closely related organisms can show proteomes as different as is typical between domains of life. Only the maximum correlations observed between proteomes shows any clear effect of phylogeny: highest between the most closely related organisms (r = 1.000 for two strains of *Staphylococcus aureus*) and lower between domains (r = 0.981 between eukaryotes and archaea). Thus, some effect of phylogeny was seen, but it was minor relative to overall proteome variation. This result suggests that in practical proteomics, realized 2D gels, with the additional effects of posttranslational modifications and expression differences, would show negligible phylogenetic influence across species.

To determine whether 2D gel variation is related to the biology of the organisms and is thus evolutionarily relevant, we compared coevolved and noncoevolved DNA replicons. If theoretical 2D gels show evolutionarily relevant variation, coevolved replicons (chromosomes from the same organism or host chromosomes and their plasmids) should be more similar than those that have not coevolved. Such reasoning has been used to demonstrate the biological relevance of DNA signatures (5). We thus compared the predicted proteomes from every eukaryotic chromosome against every other one. Similarly, we compared the predicted proteomes from all plasmids with all host chromosomes. Proteomes from individual eukaryotic chromosomes were significantly more similar to the proteomes of coevolved chromosomes (i.e., from the same organism) than others (Fig. 3, the observed difference was more extreme than 10,000 randomizations, implying $P < 0.001$). In the same way, proteomes encoded by plasmids (with a median of only 64 proteins) were more similar to the proteomes of the host with which they had coevolved than to others ($P = 0.002$, randomization test).

This pattern demonstrates that theoretical 2D gels have evolved in parallel in coevolved replicons. It is notable that, for the four mitochondrially encoded proteomes in this data set, there is no evidence of such a relationship. Mitochondrial proteomes, although small, are much less similar to their hosts than are plasmids (Fig. 3) and actually slightly less similar on average to their own host chromosomes than others, although this is not a significant effect ($P = 0.5$, permutation test). This pattern is the same as seen for genome signature (5), suggesting that very different constraints act on proteins coded within mitochondria.

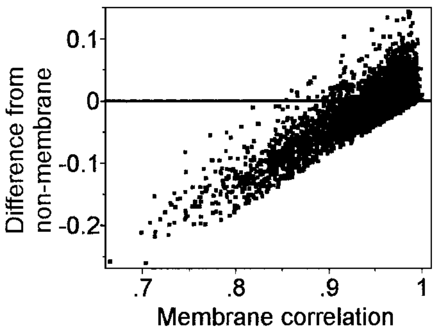

Fig. 4. Two comparisons were made independently for each pair organisms, one between membrane subproteomes and one between nonmembrane subproteomes. Each spot corresponds to the results for a pair of organisms: the x axis is the similarity of the membrane subproteomes; the y axis is the difference between the two subproteome comparisons.

Deconstructing the Pattern: Membrane Proteins Differ More in Disparate Proteomes. Subcellular localization is crucial to protein function; e.g., membrane proteins are the principal mediators between a cell and its environment. Each proteome was partitioned into membrane and nonmembrane subproteomes. On average, 25.4% (± 0.4% SE) were confirmed as membrane proteins. Extremes were the *Guillardia theta* nucleomorph with 43% membrane proteins and *Xylella fastidiosa* with only 18%. Like other recent authors (25, 26), we fail to replicate Wallin and von Heijne's (23) result (obtained using only 14 proteomes) that larger proteomes have higher proportions of membrane proteins. In fact, among unicellular organisms the correlation, although small, is significant and negative ($r = -0.28$, $n = 98$, $P = 0.005$). This finding is consistent with organisms tending to minimize the number of heavy and, hence, costly to produce proteins (8), given that in all cases except yeast, membrane proteomes average heavier than nonmembrane proteomes. Membrane proteomes also invariably averaged more basic than corresponding nonmembrane proteomes (Fig. 2, compare *C* with *D*), which confirms the pattern seen by Schwartz *et al.* (27). However the effect is quite small in many proteomes, and only rarely (Fig. 2*E*) is there a clear relationship between the basic wing of theoretical 2D gels and membrane proteins. As Schwartz *et al.* (27) suggest, this effect could be due to the basic residues commonly found on either side of membrane spanning helices.

If theoretical 2D gels are biologically important, not only should they be significantly different for different subcellular locations within an organism, but their patterns of variation between organisms should differ according to subcellular location. All pairwise comparisons of membrane subproteomes are compared to all pairwise comparisons of nonmembrane subproteomes in Fig. 4. Across organisms, neither comparison (either of membrane or nonmembrane subproteomes) was consistently more similar. However, for comparisons between proteomes with disparate membrane subproteomes (left side of the graph of Fig. 4), the nonmembrane subproteomes were less divergent. This finding is consistent with more rapid functional evolution

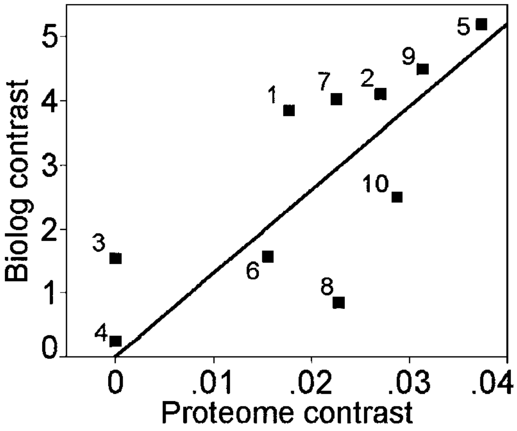

Fig. 5. Phylogenetically independent relationship between theoretical 2D gels and Biolog profile, a proxy for ecological niche. Each point corresponds to comparisons (independent contrasts) between two taxa for differentiation in theoretical 2D gel (defined as 1 − the proteome correlation used elsewhere) and Biolog profile differentiation. The taxa compared in each point are those branching from the numbered node in the phylogeny shown in Fig. 7. The line is a least-squares fit through the origin.

among membrane proteins, which could result from the relative conservation of the internal vs. the external environment of cells or a broader range of potential interactions possible with external environments.

Relation of Theoretical 2D Gels to Biology. Having demonstrated biologically relevant variation in theoretical 2D gels beyond that attributable to phylogeny, we asked about the nature of that biological variation. We hypothesize that it could be related directly to the ecology of the organisms in question, which has been demonstrated for the extremophile *Halobacterium* strain NRC-1 (10) but never more generally. In the absence of any simple general measure of ecological niche, we used as a proxy the ability of strains to grow in a Biolog plate which comprises 95 different environments, each containing a different carbon substrate. The choice of substrates and conditions was arbitrary; the key parameter was how the profile of growth differed between organisms. For a subset of 11 bacteria used in this study, we measured Biolog profile and related interstrain differences in this profile to differences in the theoretical 2D gels (pairwise comparisons in Tables 5 and 6, which are published as supporting information on the PNAS web site). Despite the lack of an overall relationship between the theoretical 2D gel and genetic distance (Fig. 11), these strains showed significant correlations of both pairwise theoretical 2D gel correlations and Biolog profile differences with genetic distance (rank correlations = 0.35 and −0.37, respectively; $P = 0.01$). Independent contrasts were thus used to control for phylogeny (Fig. 5). It is clear that, having accounted for this phylogenetic effect, there is a positive relationship between the degree of divergence of these strains' predicted proteomes and the divergence of their Biolog profile ($P < 0.0001$, $n = 10$ for a regression forced through the origin). This finding suggests that the form of an organism's theoretical 2D gel is related to its ecology.

The existence of a relationship between theoretical 2D gels and the ability to grow in different environments is surprising. Because the difference between environments was the available metabolic substrate, the relationship could be due to variation in substrate assimilation (primarily by using membrane proteins) or to variation in central metabolism (primarily by using nonmembrane proteins). We tested these alternatives by considering membrane and nonmembrane subproteomes separately. We observed a positive correlation with Biolog profile in both membrane and nonmembrane subproteomes; however, it is stronger in the nonmembrane subproteome (rms error for regression through the origin = 1.77 for the membrane subproteome but only 1.28 for the nonmembrane subproteome). This finding suggests that central metabolic processes are the principal mediators between variation in this measure of ecological niche and variation in proteomes.

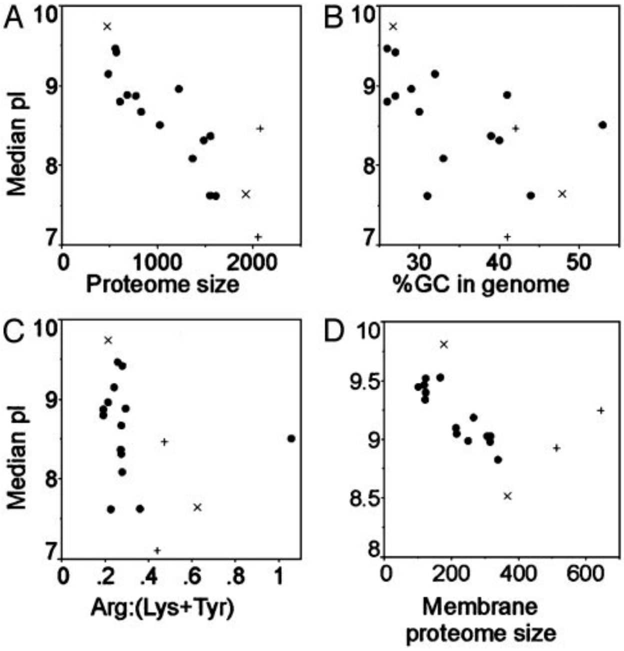

Fig. 6. Relationships of proteome pI among the smallest, most basic proteomes (●, bacteria; ×, eukaryotes; +, archaea). (*A*) Relationship with size across complete proteomes. Only the organisms shown in this graph feature in subsequent graphs. (*B*) Relationship with total DNA compositional bias. (*C*) Relationship with the ratio of arginine (the basic amino acid with high GC in its codons) to lysine and tyrosine (the basic amino acids with high AT in their codons). (*D*) The relationship with proteome size among membrane proteomes.

The correlation of theoretical 2D gels with the Biolog profile implies a broad relationship, but much smaller scaled relationships also exist. We see one in the very smallest proteomes, belonging exclusively to parasitic organisms. As has been observed (28), these tiny proteomes can be very basic on average (Fig. 2*F*). For instance, the two *Buchnera* strains each have a median pI of >9 (compared with *E. coli,* which is closely related, but has a median pI of only 6.2). The usual interpretation is that the large AT bias in the DNA of these organisms causes greater inclusion of the basic amino acids lysine and tyrosine, which are coded by strongly AT biased codons, the ultimate cause being inefficient DNA repair (29). However, comparisons among these tiny basic proteomes, rather than between tiny basic proteomes and less extreme proteomes, reveals a quantitative relationship between proteome size and basicity (Fig. 6*A*).

To test the origin of this correlation, we examined the relationship of proteome basicity and DNA AT bias in the expectation that if this were the cause of the correlation a similar quantitative relationship would be found. However, among these organisms, AT bias in the genome is only weakly correlated with basicity (Fig. 6*B*). A more specific measure, directly relevant to basicity, is the ratio of arginine (the basic amino acid with GC biased codons) to lysine and tyrosine (the basic amino acids with AT biased codons). Fig. 6*C* shows that most of these organisms have a similar, low value of the ratio, perhaps reflecting a functional limit. The recent genome sequence of the obligate symbiotic hyperthermophile *Nanoarchaeum equitans* reveals a tiny proteome: It is phylogenetically distant from the strains used in this study, has a large set of DNA repair enzymes, and shows no evidence of ongoing genome reduction (30). Despite these characteristics, *N. equitans* conforms to the relationships in Fig. 6, with a proteome size of 563, a median pI of 8.9, and a DNA compositional bias of 32% GC. Thus, there is no evidence that the quantitative relationship between small proteome size and proteome basicity originates in DNA codon bias and poor DNA repair.

The processes leading to the relationship between the size and basicity of the tiniest parasitic proteomes remain to be determined. Whatever these processes are, given the clear adaptive significance of acidic proteome pI (10), these basic pI distributions likely also have a relationship with function. One interpretation of the current results is that, in the same way that proteome minimization is an adaptive and ongoing process (31), increasing proteome pI is an adaptive process occurring in parallel. It is not yet clear whether raised pI is occurring by means of the selection of proteins in the proteome or within particular proteins. Which alternative is correct may be decided by using homology relationships that, although of great interest, are beyond the scope of this paper. Thus, changes in average pI might be due to differential amplification of acidic or basic families of proteins or protein domains; e.g., information processing domains are particularly likely to contain charge clusters (32) and may be preferentially retained during genome reduction of intracellular parasites, which could cause a pI shift in their proteomes. Alternatively, pI changes may originate in amino acid changes in homologous proteins. In that case, comparisons between surface and interior sections of proteins, as have been effective in halophilic bacteria (33), should show differences in the degree of basicity. We see preliminary evidence for this latter hypothesis: In other intracellular parasites, raised host cell pH has been shown (34). If raised host cell pH is widespread, raised proteome pI could well be a protein-level adaptation to enable the parasite's proteins to function in such a basic environment. Although the relationship between basicity and proteome size is present in both membrane and nonmembrane subproteomes, in the bacteria at least, the relationship is more distinct among membrane subproteomes (Fig. 6*D*), suggesting that it may be primarily related to these organisms' interaction with their environment or host.

In this work, we have demonstrated ways that simple protein properties may relate to function across complete proteomes. The few previous studies that considered such simple protein properties for complete proteomes, independent of functional annotation, have focused on membrane protein identification (25); low complexity sequences (35), including charge clusters (32); or charge distributions (11, 27). There are many physical properties readily predictable from raw protein sequences; several, such as hydrophobicity or stability, are crucial to protein function. Understanding how these relate to function at the scale of complete proteomes will bring new insight into the workings of evolution at a scale relevant to both whole organism and molecular biology.

We thank Chris Brock for programming expertise, Nicole Zitzmann for continuing support, Elisabeth Gasteiger for help with the pI algorithm, Craig MacLean for Biolog profile help, Dawn Field for discussions, and anonymous referees for valuable comments. C.G.K. is supported by the Environmental Genomics program of the National Environment Research Council, and R.K. is supported by the Natural Sciences and Engineering Research Council of Canada and St. Hugh's College, Oxford.

1. Karlin, S., Brocchieri, L., Trent, J., Blaisdell, B. E. & Mrazek, J. (2002) *Theor. Popul. Biol.* **61,** 367–390.
2. Kanapin, A., Batalov, S., Davis, M. J., Gough, J., Grimmond, S., Kawaji, H., Magrane, M., Matsuda, H., Schonbach, C., Teasdale, R. D., *et al.* (2003) *Genome Res.* **13,** 1335–1344.
3. Tekaia, F., Yeramian, E. & Dujon, B. (2002) *Gene* **297,** 51–60.
4. Dumontier, M., Michalickova, K. & Hogue, C. W. (2002) *BMC Bioinformatics* **3,** 39.
5. Campbell, A., Mrazek, J. & Karlin, S. (1999) *Proc. Natl. Acad. Sci. USA* **96,** 9184–9189.
6. Karlin, S. & Mrazek, J. (2000) *J. Bacteriol.* **182,** 5238–5250.
7. Link, A. J., Robison, K. & Church, G. M. (1997) *Electrophoresis* **18,** 1259–1313.
8. Seligmann, H. (2003) *J. Mol. Evol.* **56,** 151–161.
9. Urquhart, B. L., Cordwell, S. J. & Humphery-Smith, I. (1998) *Biochem. Biophys. Res. Commun.* **253,** 70–79.
10. Kennedy, S. P., Ng, W. V., Salzberg, S. L., Hood, L. & DasSarma, S. (2001) *Genome Res.* **11,** 1641–1650.
11. Kawashima, T., Amano, N., Koike, H., Makino, S., Higuchi, S., Kawashima-Ohya, Y., Watanabe, K., Yamazaki, M., Kanehori, K., Kawamoto, T., *et al.* (2000) *Proc. Natl. Acad. Sci. USA* **97,** 14257–14262.
12. Tomb, J. F., White, O., Kerlavage, A. R., Clayton, R. A., Sutton, G. G., Fleischmann, R. D., Ketchum, K. A., Klenk, H. P., Gill, S., Dougherty, B. A., *et al.* (1997) *Nature* **388,** 539–547.
13. Altland, K. (1990) *Electrophoresis* **11,** 140–147.
14. Bjellqvist, B., Basse, B., Olsen, E. & Celis, J. E. (1994) *Electrophoresis* **15,** 529–539.
15. Thompson, J. D., Higgins, D. G. & Gibson, T. J. (1994) *Nucleic Acids Res.* **22,** 4673–4680.
16. Felsenstein, J. (1989) *Cladistics* **5,** 164–166.
17. Ihaka, R. & Gentleman, R. (1996) *J. Comput. Graph. Stat.* **5,** 299–314.
18. SAS Institute (2002) *JMP Version 5 Statistics and Graphics Guide* (SAS Publishing, Cary, NC).
19. Sheather, S. J. & Jones, M. C. (1991) *J. R. Stat. Soc. B* **53,** 683–690.
20. Felsenstein, J. (1985) *Am. Nat.* **125,** 1–15.
21. MacLean, R. C. & Bell, G. (2003) *Proc. R. Soc. London Ser. B* **270,** 1645–1650.
22. Tusnady, G. E. & Simon, I. (2001) *Bioinformatics* **17,** 849–850.
23. Wallin, E. & von Heijne, G. (1998) *Protein Sci.* **7,** 1029–1038.
24. Zhang, J. (2000) *Trends Genet.* **16,** 107–109.
25. Krogh, A., Larsson, B., von Heijne, G. & Sonnhammer, E. L. (2001) *J. Mol. Biol.* **305,** 567–580.
26. Liu, J. & Rost, B. (2001) *Protein Sci.* **10,** 1970–1979.
27. Schwartz, R., Ting, C. S. & King, J. (2001) *Genome Res.* **11,** 703–709.
28. Shigenobu, S., Watanabe, H., Hattori, M., Sakaki, Y. & Ishikawa, H. (2000) *Nature* **407,** 81–86.
29. Moran, N. A. (2002) *Cell* **108,** 583–586.
30. Waters, E., Hohn, M. J., Ahel, I., Graham, D. E., Adams, M. D., Barnstead, M., Beeson, K. Y., Bibbs, L., Bolanos, R., Keller, M., *et al.* (2003) *Proc. Natl. Acad. Sci. USA* **100,** 12984–12988.
31. van Ham, R. C., Kamerbeek, J., Palacios, C., Rausell, C., Abascal, F., Bastolla, U., Fernandez, J. M., Jimenez, L., Postigo, M., Silva, F. J., *et al.* (2003) *Proc. Natl. Acad. Sci. USA* **100,** 581–586.
32. Karlin, S., Mrazek, J. & Gentles, A. J. (2003) *Curr. Opin. Struct. Biol.* **13,** 344–352.
33. Fukuchi, S., Yoshimune, K., Wakayama, M., Moriguchi, M. & Nishikawa, K. (2003) *J. Mol. Biol.* **327,** 347–357.
34. Rodriguez-Cabezas, N., Gonzalez, M. A., Lazuen, J., Cifuentes, J., Soler-Diaz, A. & Osuna, A. (1998) *Int. J. Parasitol.* **28,** 1841–1851.
35. Nandi, T., Dash, D., Ghai, R., Rao, C. B., Kannan, K., Brahmachari, S. K., Ramakrishnan, C. & Ramachandran, S. (2003) *J. Biomol. Struct. Dyn.* **20,** 657–668.