Whole Proteome pI Values Correlate with Subcellular Localizations of Proteins for Organisms within the Three Domains of Life

Russell Schwartz, Claire S. Ting and Jonathan King

Genome Res. 2001 11: 703-709
Access the most recent version at doi:10.1101/gr.158701

Article

Whole Proteome pI Values Correlate with Subcellular Localizations of Proteins for Organisms within the Three Domains of Life

Russell Schwartz,[1,2,3,4] Claire S. Ting,[1,2] and Jonathan King[1]

[1]*Department of Biology, Massachusetts Institute of Technology, Cambridge, Massachusetts 02139, USA*

Isoelectric point (pI) values have long been a standard measure for distinguishing between proteins. This article analyzes distributions of pI values estimated computationally for all predicted ORFs in a selection of fully sequenced genomes. Histograms of pI values confirm the bimodality that has been observed previously for bacterial and archaeal genomes (Van Bogelen et al. 1999) and reveal a trimodality in eukaryotic genomes. A similar analysis on subsets of a nonredundant protein sequence database generated from the full database by selecting on subcellular localization shows that sequences annotated as corresponding to cytosolic and integral membrane proteins have pI distributions that appear to correspond with the two observed modes of bacteria and archaea. Furthermore, nuclear proteins have a broader distribution that may account for the third mode observed in eukaryotes. On the basis of this association between pI and subcellular localization, we conclude that the bimodal character of whole proteome pI values in bacteria and archaea and the trimodal character in eukaryotes are likely to be general properties of proteomes and are associated with the need for different pI values depending on subcellular localization. Our analyses also suggest that the proportions of proteomes consisting of membrane-associated proteins may be currently underestimated.

Predictions of the complete assemblage of proteins an organism is capable of expressing have been made from annotated genome sequences and have yielded information important for comparative genomics, as well as experimental applications. Recently, Van Bogelen et al. (1999) reviewed the value of two-dimensional polyacrylamide gel electrophoresis (2D PAGE) for correlating protein expression with cellular state. These investigators presented a computer-generated analog of a second dimension gel that was produced by plotting calculated pI values against calculated molecular weights for all proteins in an entire proteome. Using the proteome predicted for *Escherichia coli* MG1655 (Blattner et al. 1997), they identified a distinct bimodal pattern to this plot, with peaks centered around pI 5.5 and pI 9 (Van Bogelen et al. 1999). Although the causes for and generality of this bimodality were not established in this particular study, these investigators suggested that it may be a result of the relationship between intracellular pH values and protein pI values (Van Bogelen et al. 1999). Because proteins are generally least soluble near their isoelectric points (Arakawa and Timasheff 1985) and the cytoplasm has a pH near neutrality, the biomodality of protein pIs, with peak values greater than or less than pH 7.0, may be a general property of prokaryotic proteomes.

We hypothesized that the bimodality observed for *E. coli* may also be caused by the requirement for different protein pI values due to subcellular localization. To test this hypothesis, we developed a computer program to calculate the approximate mass and pI values for polypeptide sequences, as described originally by Van Bogelen et al. (1999). Although these calculated pI values are necessarily inexact, because they are derived absent any information about the influence of protein fold on ionizable groups, previous pI estimates from one-dimensional sequence data using a similar methodology were found to be in reasonably close agreement with experimentally determined pI values (Sillero and Ribeiro 1989). Our program was applied to the predicted ORFs from the completed genomes of organisms belonging to each of the three domains (bacteria, archaea, or eukarya) established from comparative ribosomal RNA gene sequencing (Woese 1987). The representative organisms and their relevant characteristics are summarized in Table 1.

RESULTS

Figure 1 shows examples of scatter plots simulating 2D PAGE gels for *E. coli* K12, *Methanococcus jannaschii* (Bult et al. 1996), and *Drosophila melanogaster* (Adams et al. 2000). Figure 2 shows histograms of estimated pI values for denatured states of proteins derived from predicted ORFs for *E. coli* K12, *Synechocystis* sp. strain PCC 6803 (Kaneko et al. 1996), *Thermotoga maritima* (Nel-

[2]**These authors contributed equally to this work.**
[3]**Present address: Massachusetts Institute of Technology, 77 Massachusetts Avenue, Room 68-322, Cambridge, MA 02139, USA.**
[4]**Corresponding author.**
E-MAIL rss@alum.mit.edu; FAX (617) 252-1843.
Article published on-line before print: *Genome Res.,* 10.1101/gr.158701. Article and publication are at www.genome.org/cgi/doi/10.1101/gr.158701.

Table 1. Fully-Sequenced Genomes of the Representative Organisms Included in This Study.

Organism	Domain	Genome size (Mbp)	Number of predicted ORFs	Description	Reference
Pyrococcus abyssi	Archaea	1.77	1765	hyperthermophile; chemoorganotroph	R. Heilig (unpub.)
Methanococcus jannaschii	Archaea	1.66	1767	hyperthermophile; chemolithotroph; anaerobe	Bult et al. 1996
Escherichia coli K12	Bacteria	4.6	4290	facultative aerobe	Blattner et al. 1997
Thermotoga maritima	Bacteria	1.86	1849	hyperthermophile; chemoorganotroph; anaerobe	Nelson et al. 1999
Mycobacterium tuberculosis	Bacteria	4.4	3294	causative agent of tuberculosis	Cole et al. 1998
Helicobacter pylori	Bacteria	1.7	1565	survives under acidic conditions; microaerophile; human pathogen	Tomb et al. 1997
Synechocystis sp. strain PCC6803	Bacteria	3.6	3167	unicellular cyanobacterium	Kaneko et al. 1996
Saccharomyces cerevisiae	Eukarya	12	6294	unicellular fungi	Goffeau et al. 1996
Caenorhabditis elegans	Eukarya	97	22898	multicellular eukaryote	The *C. elegans* Sequencing Consortium 1998
Drosophila melanogaster	Eukarya	180	14118	multicellular eukaryote	Adams et al. 2000

son et al. 1999), *Mycobacterium tuberculosis* (Cole et al. 1998), *Helicobacter pylori* (Tomb et al. 1997), *M. jannaschii*, and *Pyrococcus abyssi* (R. Heilig, unpubl.). In each case, the plots exhibited the same bimodal distribution observed by Van Bogelen et al. (1999), with a lower peak appearing at approximately pI 5 and a higher peak at approximately pI 9. Similar bimodal distributions could not be generated randomly using the single amino acid frequencies unique to a particular prokaryote (Fig. 3, shown for *E. coli* K12), suggesting that a biological basis may underlie the observed pattern. Furthermore, the proportion of proteins predicted to belong to each of these groups (~pI 5 vs. ~pI 9) differed significantly among these prokaryotes.

Interestingly, when these same analyses were extended to three eukaryotic genomes (*Saccharomyces cerevisiae* [Goffeau et al. 1996], *C. elegans* [*C. elegans* Sequencing Consortium 1998], and *D. melanogaster* [Adams et al. 2000]), an apparently trimodal distribution was observed. Figure 4 shows histograms of pI values for the three eukaryotic proteomes examined. In each case, predicted protein sequences were observed to cluster around pI 5 and pI 9, as was previously observed for each of the bacterial and archaeal proteomes. However, for the eukaryotic proteomes, proteins were also observed to cluster in a third region located at approximately pI 7 (Fig. 4).

The analyses were repeated using three subsets of proteins from the SWISS-PROT database of nonredundant protein sequence data, release 38 (Bairoch and Apweiler 2000). The specific subsets examined were selected by scanning for the following strings in their annotations: "SUBCELLULAR LOCATION: CYTOPLASMIC" for the first subset, "SUBCELLULAR LOCATION: INTEGRAL MEMBRANE PROTEIN" for the second, "SUBCELLULAR LOCATION: NUCLEAR" for the third. These three subsets contained 5556, 7031, and 4898 protein sequences, accounting for 5.0%, 6.3%, and 4.4% of the total content of the SWISS-PROT database, respectively.

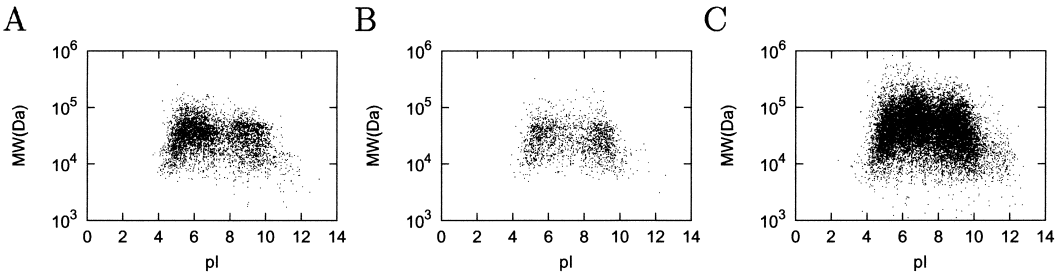

Figure 1 Scatter plots of estimated molecular mass versus pI for (*A*) *Escherichia coli* K12–, (*B*) *Methanococcus jannaschii*–, and (*C*) *Drosophila melanogaster*–predicted ORFs.

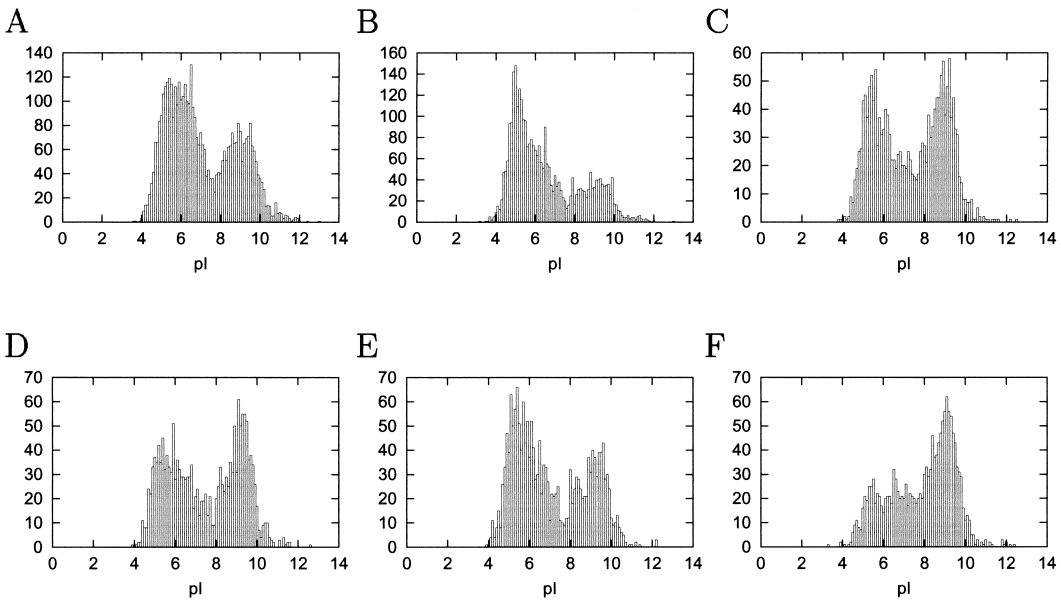

Figure 2 Histograms of pI values at 0.1 unit intervals for (*A*) *Escherichia coli* K12–, (*B*) *Synechocystis* sp. strain PCC 6803–, (*C*) *Methanococcus jannaschii*–, (*D*) *Pyrococcus abyssi*–, (*E*) *Thermotoga maritima*–, and (*F*) *Helicobacter pylori*–predicted ORFs.

Figure 5 shows histograms corresponding to these three subsets of proteins. Although cytoplasmic proteins exhibited a distinct clustering around pI 5 to pI 6, integral membrane proteins were clustered primarily around pI 8.5 to pI 9 (Fig. 5). In contrast, nuclear proteins were almost evenly distributed throughout the pI range (pI 4.5–pI 10), encompassing both cytoplasmic and integral membrane proteins. As expected, analysis of the amino acid frequencies of proteins within each of these three subclasses from the SWISS-PROT database indicated that integral membrane proteins were enriched in nonpolar residues (Leu, Ile) and that cytoplasmic proteins were relatively enriched in charged (acidic) residues (Asp, Glu; Fig. 6). Unexpectedly, however, nuclear proteins were found to be enriched in nonpolar (Pro) and uncharged polar (Ser) residues (Fig. 6). As shown in Table 2, when these same analyses were extended to the proteomes of prokaryotic and eukaryotic organisms, distinct differences were observed in the relative distributions of amino acid frequencies. These differences may contribute in part to the distinct pI profiles of the various proteomes. In general, all organisms were enriched in leucine relative to residues such as cysteine, histidine, methionine, and tryptophan. However, the three eukaryotic organisms possessed a relatively higher proportion of serine residues compared to the prokaryotic organisms (Table 2).

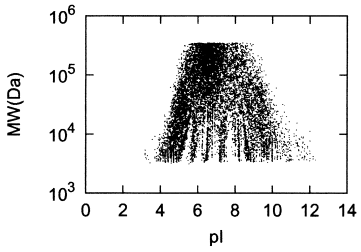

Figure 3 Simulated scatter plot of estimated molecular mass versus pI, produced by generating random sequences with amino acid frequencies selected independently according to their observed proportions in the *Escherichia coli* K12 genome. Sequence lengths were randomly selected such that logarithms of their lengths were uniformly distributed between 1.5 and 3.5.

We further utilized protein pI values to estimate how representative the proteins with known or predicted functions are of those with no predicted function. Figure 7 shows plots comparing distributions of pI values between sequences that have been assigned at least tentative functions and those that have not for *H. pylori*, *P. abyssi*, *M. jannaschii*, and *D. melanogaster*. These specific organisms were chosen because their predicted ORFs were annotated in formats that facilitated such analysis. In each case, the bimodality or trimodality of the graph is the same for the two data sets. We also estimated the proportions of the full proteomes contained in the upper peak of each graph. We judged this upper peak to begin to predominate at approximately pH 7.5. The percentage of residues with an estimated pI of 7.5 or higher is 38% for *E. coli*, 28% for *Synechocystis*, 49% for *M. jannaschii*, 50% for *P. abyssi*, 39% for *T. maritima*, and 62% for *H. pylori*. It is more difficult to make accurate estimates for the eukarya because of the presence of the third mode; however, again using pH 7.5 as an estimate of where the second and third peaks of the histogram of Figure 4 make equal contributions to the total frequency, we arrive at estimates of 48% of *S. cerevisiae*, 53% of *C. elegans*, and

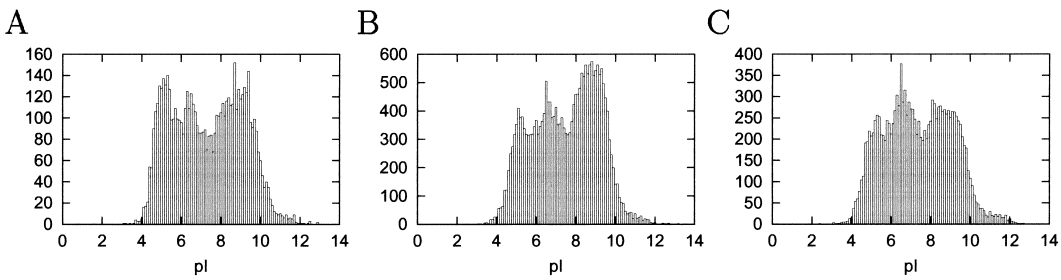

Figure 4 Histograms of pI for predicted ORFs of (*A*) *Saccharomyces cerevisiae*, (*B*) *Caenorhabditis elegans*, and (*C*) *Drosophila melanogaster*.

47% of *D. melanogaster* sequences lying in the highest mode.

To assess the accuracy of our computed pI values, we calculated the pI values for a set of proteins with pI values that have been determined experimentally. We selected proteins from the SWISS-2D PAGE database of proteins isolated in 2D PAGE gels (Hoogland et al. 1998, 1999, 2000). From these proteins, we selected those that had corresponding annotations in the SWISS-PROT database and had translations containing no ambiguous residues. We then calculated the difference between measured pI value (using the average of all values for a given sequence recorded in the SWISS-2D PAGE database) and computationally calculated pI value for each protein. Among all proteins examined, the average calculated value was 0.51 units above the average measured value, with a standard deviation of 0.76 units. Among those annotated as "SUBCELLULAR LOCATION: CYTOPLASMIC" in SWISS-PROT (a sample size of 169), the calculated values exceeded the measured values by an average of 0.36 units, with a standard deviation of 0.36 units. Among those annotated as "SUBCELLULAR LOCATION: NUCLEAR" in SWISS-PROT (a sample size of 17), the calculated values exceeded the measured values by an average of 0.31 units, with a standard deviation of 0.36 units. Among those annotated as "SUBCELLULAR LOCATION: INTEGRAL MEMBRANE" in SWISS-PROT (a sample size of 4), the calculated values exceeded the measured values by an average of 0.46 units, with a standard deviation of 0.45 units.

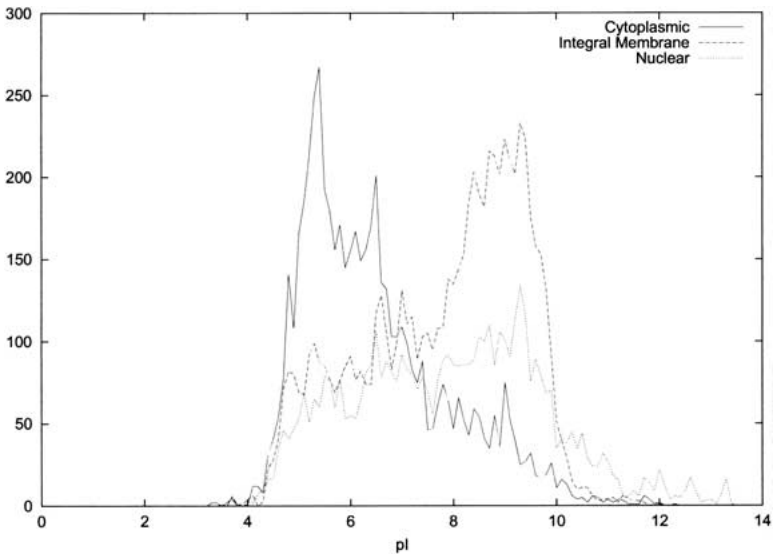

Figure 5 Histograms of proteins separated by calculated pI values for cytoplasmic, membrane, and nuclear proteins, as extracted from SWISS-PROT based on annotation.

DISCUSSION

On the basis of our database analyses, we propose that the bimodal character of whole-proteome pI values in the bacteria and archaea we examined and the trimodal character in eukaryotes are likely to be general properties of proteomes, determined by the need for different pI values, depending on subcellular localization. Integral membrane proteins and cytosolic proteins are found in large numbers in all proteomes (Fig. 5). The two major protein clusters (~pI 5 and ~pI 9) we have observed in the full proteomes of organisms belonging to the domains bacteria, archaea, and eukarya most likely corresponded to these two classes of proteins. Nuclear proteins (Fig. 5) produced a peak similar in position to the third mode observed only in the eukaryotic proteomes (Fig. 4). Although other classes of proteins (i.e., not nuclear, cytoplasmic, or integral membrane) must eventually be accounted for within the observed distributions, our analyses indicate that pI calculations may provide a way of assigning tentative subcellular localizations to proteins that have been identified in sequenced genomes but have not been characterized further. In addition, the fact that the distributions of pI values differ notably from one organism to another suggests the potential value of this measure to comparative genomics.

The comparatively high pI values of integral membrane proteins are consistent

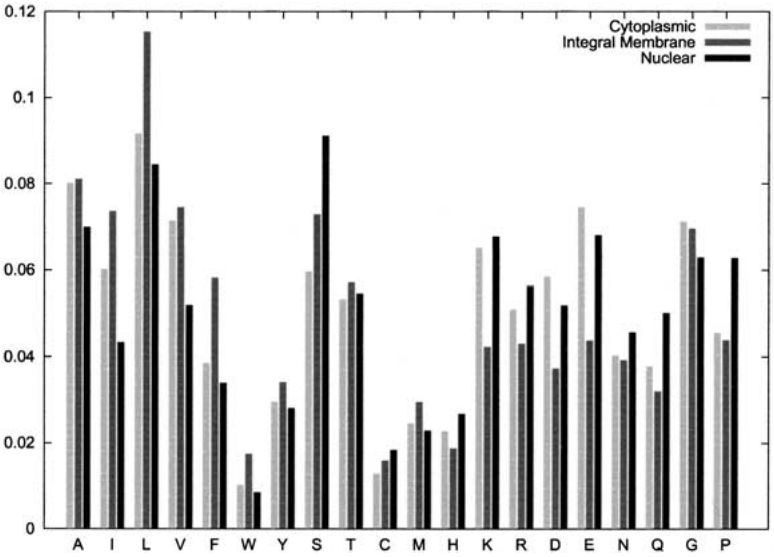

Figure 6 Single amino acid frequences for cytoplasmic, integral membrane, and nuclear proteins from SWISS-PROT.

with the fact that most biomembranes have negatively charged surfaces (Gennis 1989). A slight bias toward basic residues in the regions of membrane proteins lying near the surface of the membrane would be expected to promote favorable electrostatic interactions and help to stabilize the proteins in the membranes. We do not have any well-supported hypotheses for why cytosolic proteins should have pI values generally below 7 nor why whatever causes this effect in cytosolic proteins does not act on nuclear proteins.

Approximately half of the predicted proteins of the genomes sequenced to date have no known function. The similarity of pI distributions for predicted proteins of assigned and unassigned functions suggests that known proteins are to some extent representative of the entire proteome by this measure. If our conclusion linking pI to subcellular localization is correct, however, then membrane proteins appear to be disproportionately overrepresented among the unknown proteins compared to known proteins, although the degree to which this is true varies significantly between the organisms examined. This suggests a possible need for rethinking approaches to isolating and identifying the remaining proteins of unknown function. Our estimates of numbers of proteins assigned to the pI peak we identify with membrane proteins lead to an estimated total membrane protein content of 38% for *E. coli*, 28% for *Synechocystis*, 49% for *M. jannaschii*, 50% for *P. abyssi*, 39% for *T. maritima*, 62% for *H. pylori*, 48% for *S. cerevisiae*, 53% for *C. elegans*, and 47% for *D. melanogaster*. Due to the overlap in calculated pI values for cytoplasmic and membrane proteins (Fig. 5), these estimates are rough approximations. Although they are within the range of past estimates (~35%) of the proportion of membrane proteins based on transmembrane prediction methods (Frishman and Mewes 1997), they are approximately double other estimates made using different prediction approaches (18% to 29%; Kihara and Kanehisa 2000).

It is unlikely that the overall character of the pI plots for bacteria, archaea, and eukaryotes is a product of errors in the calculations of pI values. Although our comparisons of calculated and experimentally determined pI values for a set of proteins reveal a systematic bias toward overestimating pI values, the magnitude of

Table 2. Amino Acid Frequencies by Organism as a Percentage of Total Numbers of Residues

	Ala	Cys	Asp	Glu	Phe	Gly	His	Ile	Lys	Leu	Met	Asn	Pro	Gln	Arg	Ser	Thr	Val	Trp	Tyr
Escherichia coli	9.5	1.2	5.1	5.7	3.9	7.4	2.3	6.0	4.4	10.6	2.9	4.0	4.4	4.4	5.5	5.8	5.4	7.1	1.5	2.8
Synechocystis	8.5	1.0	5.0	6.0	4.0	7.4	1.9	6.3	4.2	11.4	2.0	4.0	5.1	5.6	5.1	5.8	5.5	6.7	1.6	2.9
Mycobacterium jannaschii	5.4	1.3	5.5	8.7	4.3	6.3	1.4	10.5	10.4	9.5	2.3	5.3	3.4	1.4	3.8	4.5	4.0	6.8	0.7	4.4
Pyrococcus abyssi	6.7	0.6	4.6	8.8	4.4	7.3	1.5	8.5	7.8	10.3	2.4	3.3	4.3	1.7	5.7	5.0	4.2	8.1	1.2	3.8
Helicobacter pylori	6.8	1.1	4.8	6.9	5.4	5.8	2.1	7.2	8.9	11.2	2.3	5.9	3.3	3.7	3.5	6.8	4.4	5.6	0.7	3.7
Thermotoga maritima	5.9	0.7	5.0	8.9	5.2	6.9	1.6	7.2	7.6	10.0	2.4	3.6	4.0	2.0	5.5	5.7	4.5	8.6	1.1	3.6
Saccharomyces cerevisiae	5.5	1.3	5.8	6.5	4.5	4.9	2.2	6.6	7.3	9.5	2.1	6.1	4.4	3.9	4.5	9.1	5.9	5.6	1.0	3.4
Caenorhabditis elegans	6.3	2.1	5.3	6.5	4.9	5.3	2.3	6.2	6.5	8.7	2.6	4.9	4.9	4.1	5.2	8.0	5.8	6.2	1.1	3.2
Drosophila melanogaster	7.5	1.9	5.2	6.4	3.5	6.2	2.7	4.9	5.6	9.2	2.4	4.7	5.5	5.2	5.6	8.3	5.6	5.9	1.0	3.0

Frequencies were computed based on the sequences of all predicted ORFs for the respective genomes, including those whose identifications are only tentative.

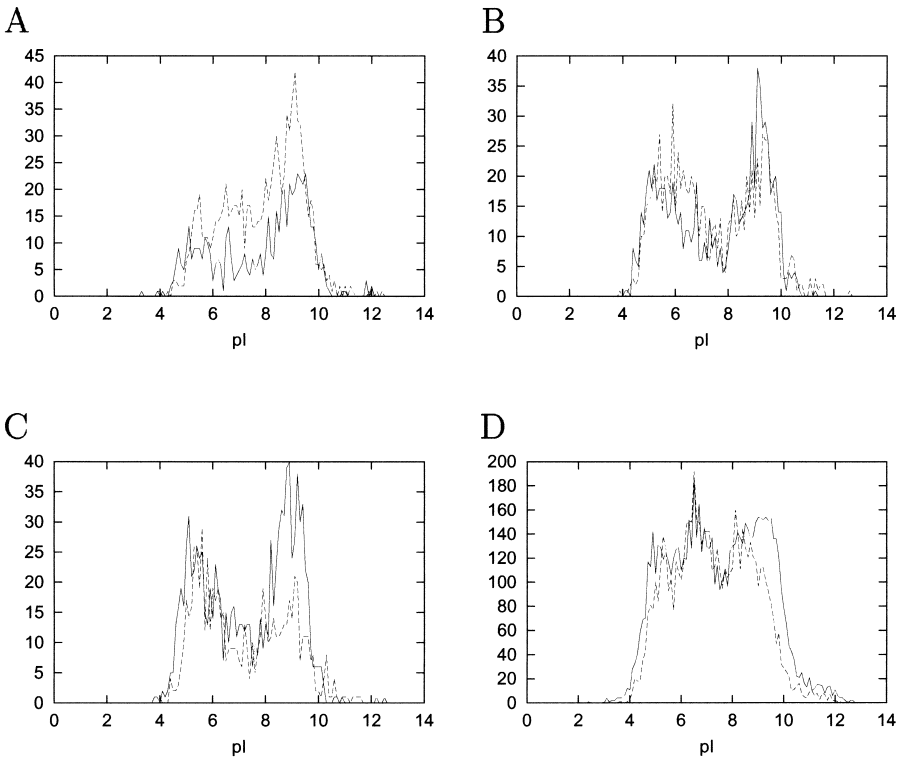

Figure 7 Comparison of identified vs. unidentified proteins for (*A*) *Helicobacter pylori*, (*B*) *Pyrococcus abyssi*, (*C*) *Methanococcus jannaschii*, and (*D*) *Drosophila melanogaster*. Dashed lines represent proteins identified, possibly on the basis of homology; solid lines represent proteins for which no function has been identified (unannotated proteins in the *D. melanogaster* database and proteins annotated as hypothetical in the other three databases).

the overestimation is several standard deviations below the size of the gaps between the peaks identified in the pI histograms. It should be noted, however, that the availability of experimentally determined pI values is sparse, particularly with respect to nuclear and membrane proteins. The expansion of this database over the next several years will permit further refinement of these comparisons between calculated and experimentally measured protein pI values.

During this study, we found the lack of consistency in the design of protein sequence databases to be a significant obstacle. This difficulty is important to address, for it will undoubtedly hinder similar future computational analyses of protein sequence databases. The lack of consistency resulted from the use of different formats for databases constructed by different groups, thereby necessitating the development of software tools to convert them to a common minimal format for analysis. A subtler but more intractable problem was the lack of consistency in annotation formats even within a single database. For example, the analysis of proteins by localization was significantly hindered by the fact that only 27% of the sequences in the SWISS-PROT database had a subcellular localization annotation. Furthermore, even among these particular proteins there was a lack of consensus regarding how much information to provide and in what format. Although these inconsistencies may not be serious obstacles for scientists interested in manually examining a few sequences, they are a major problem for conducting large-scale computational analyses of protein sequence databases. The recent explosion in available genomic and proteomic data has created an opportunity for exploring important questions that were inaccessible only a few years ago. These efforts, however, will be undermined considerably by the database inconsistencies we have observed. We suggest to the community as a whole the benefits of adopting open and universal standards for the format of sequence databases. This will advance the analysis of sequence databases and the development of computational tools for use in such research.

Table 3. pKa Values of Ionizable Groups in Proteins Used for the Estimation of pI

Terminal carboxyl	3.1
Aspartic and glutamic acid	4.4
Histidine	6.5
Terminal amino	8.0
Cysteine	8.5
Tyrosine or lysine	10.0
Arginine	12.0

The values were taken from Stryer (1995).

METHODS

Molecular masses were estimated by summing residue masses for all residues in a polypeptide chain and adding the additional contributions of the N-terminal hydrogen and C-terminal hydroxide potentially present in a full polypeptide chain. pI values were estimated from each amino acid sequence on the assumption of fixed pKa values for all ionizable groups, as given in Table 3. A bisection search was applied to locate the pH for which the net charge of the polypeptide was zero. This means that for any given pH, we calculated the charge of a polypeptide at that pH by summing over all ionizable groups the average charge for that group at the pH being examined. We then sampled those values at the endpoints of a region initially covering the pH range 0 to 14.0 and

successively subdivided the region, recursing on the portion of the region that has one endpoint positively charged and the other negatively charged. The process was repeated until the region was reduced to a range of 10^{-6} pI value, at which point the midpoint of the region was considered to be the pI of the polypeptide being examined. The aforementioned computations were performed using code written in the C programming language. From the results of these calculations, scatter plots were created of all proteins in a given database, with pI plotted along the *X*-axis and the logarithm of mass plotted along the *Y*-axis, as was done originally by Van Bogelen et al. (1999). In addition, we generated histograms of the numbers of sequences in each interval of 0.1 pI value between 0 and 14.

Selection of sequences based on annotation to select proteins with specific subcellular localizations or to separate proteins of known function from proteins of unknown function was done with code written in the Perl programming language applied to the relevant amino-acid-sequence databases. Determination of amino acid compositions by genome was also done with code written in the Perl programming language applied to the individual amino-acid-sequence databases. All graphics presented in this paper were generated with Gnuplot (Linux version 3.7).

ACKNOWLEDGMENTS

R.S. was supported by NIH grant 7-T32-HG0039–05, a Training Grant in Genomic Sciences. C.S.T. was supported by NIH grant GM 17,980. We thank Peter Thumfort for carefully reading this manuscript and suggesting improvements.

The publication costs of this article were defrayed in part by payment of page charges. This article must therefore be hereby marked "advertisement" in accordance with 18 USC section 1734 solely to indicate this fact.

REFERENCES

Adams, M.D., Celniker, S.E., Holt, R.A., Evans, C.A., Gocayne, J.D., Amanatides, P.G., Scherer, S.E., Li, P.W., Hoskins, R.A., Galle, R.F., et al. 2000. The genome sequence of *Drosophila melanogaster*. *Science* **287:** 2185–2195.

Arakawa, T. and Timasheff, S.N. 1985. Theory of protein solubility. *Meth. Enzymol.* **114:** 49–77.

Bairoch, A. and Apweiler, R. 2000. The SWISS-PROT protein sequence database and its supplement TrEMBL in 2000. *Nucleic Acids Res.* **28:** 45–48.

Blattner, F.R., Plunkett, G., Bloch, C.A., Perna, N.T., Burland, V., Riley, M., Collado-Vides, J., Glasner, J.D., Rode, C.K., Mayhew, G.F., et al. 1997. The complete genome sequence of *Escherichia coli* K-12. *Science* **277:** 1453–1474.

Bult, C.J., White, O., Olsen, G.J., Zhou, L., Fleischmann, R.D., Sutton, G.G., Blake, J.A., FitzGerald, L.M., Clayton, R.A., Gocayne, J.D., et al. 1996. Complete genome sequence of the methanogenic archaeon, *Methanococcus jannaschii*. *Science* **273:** 1058–1073.

C. elegans Sequencing Consortium. 1998. Genome sequence of the nematode *C. elegans*: A platform for investigating biology. *Science* **282:** 2012–2018.

Cole, S.T., Brosch, R., Parkhill, J., Garnier, T., Churcher, C., Harris, D., Gordon, S.V., Eiglmeier, K., Gas, S., Barry, C.E., et al. 1998. Deciphering the biology of *Mycobacterium tuberculosis* from the complete genome sequence. *Nature* **393:** 537–544.

Frishman, D. and Mewes, H.W. 1997. Protein structural classes in five complete genomes. *Nature Struct. Biol.* **4:** 626–628.

Gennis, R.B. 1989. In *Biomembranes: Molecular structure and function,* p. 252. Springer-Verlag New York.

Goffeau, A., Barrell, B.G., Bussey, H., Davis, R.W., Dujon, B., Feldmann, H., Galibert, F., Hoheisel, J.D., Jacq, C., Johnston, M., et al. 1996. Life with 6000 genes. *Science* **274:** 546–567.

Hoogland, C., Sanchez, J.-C., Tonella, L., Bairoch, A., Hochstrasser, D.F., and Appel, R.D. 1998. Current status of the SWISS-2DPAGE database. *Nucleic Acids Res.* **26:** 332–333.

———. 1999. The SWISS-2DPAGE database: what has changed during the last year. *Nucleic Acids Res.* **27:** 289–291.

Hoogland, C., Sanchez, J.-C., Tonella, L., Binz, P.-A., Bairoch, A., Hochstrasser, D.F., and Appel, R.D. 2000. The 1999 SWISS-2DPAGE database update. *Nucleic Acids Res.* **28:** 286–288.

Kaneko, T., Sato, S., Kotani, H., Tanaka, A., Asamizu, E., Nakamura, Y., Miyajima, N., Hirosawa, M., Sugiura, M., Sasamoto, S., et al. 1996. Sequence analysis of the genome of the unicellular cyanobacterium *Synechocystis* sp. strain PCC6803. II. Sequence determination of the entire genome and assignment of potential protein-coding regions. *DNA Res.* **3:** 109–136.

Kihara, D., and Kanehisa, M. 2000. Tandem clusters of membrane proteins in complete genome sequences. *Genome Res.* **10:** 731–743.

Nelson, K.E., Clayton, R.A., Gill, S.R., Gwinn, M.L., Dodson, R.J., Haft, D.H., Hickey, E.K., Peterson, J.D., Nelson, W.C., Ketchum, K.A., et al. 1999. Evidence for lateral gene transfer between Archaea and Bacteria from genome sequence of *Thermotoga maritima*. *Nature* **399:** 323–329.

Sillero, A., and Ribeiro, J.M. 1989 Isoelectric points of proteins: Theoretical determination. *Anal. Biochem.* **179:** 319–325.

Stryer, L. 1995. In *Biochemisty*, 3rd ed., p. 23. W. H. Freeman and Company, New York.

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. The complete genome sequence of the gastric pathogen *Helicobacter pylori*. *Nature* **388:** 539–547.

Van Bogelen, R.A., Schiller, E.E., Thomas, J.D., and Neidhardt, F.C. 1999. Diagnosis of cellular states of microbial organisms using proteomics. *Electrophoresis* **20:** 2149–2159.

Woese, C. 1987. Bacterial evolution. *Microbiol. Rev.* **51:** 221–271.

Received August 4, 2000; accepted in revised form February 16, 2001.