COMPARISON OF PROTEIN AND MRNA PROFILES OF ESCHERICHIA COLI: DATA VISUALIZATION AND ANALYSIS OF SPECIFIC GENE GROUPS

Oleg Paliy,[1,2] Brian Thomas,[3] Rebecca Corbin,[4] Feng Yang,[4] Jeffrey Shabnowitz, [4] Mark Platt, [4] Charles E. Lyons, Jr., [4] Karen Root, [4] Donald Hunt, [4, 5] and Sydney Kustu[2]

Department of Biochemistry and Molecular Biology, Wright State University, Dayton, Ohio, 45435,[1] Department of Plant & Microbial Biology,[2] and College of Natural Resources,[3] University of California, Berkeley, CA 94720, Department of Chemistry, University of Virginia, Charlottesville, VA 22901,[4] and Department of Pathology, University of Virginia, Charlottesville, VA 229084[5].

ABSTRACT

Despite recent progress in protein identification in whole cell lysates, many laboratories interested in global gene expression depend on assessment of mRNA rather than protein. Hence knowledge of the relationship between the two remains important. Though it has been explored in eukaryotic cells, there are few studies of this relationship in bacteria. In addition, previous studies have generally not considered illustrative examples.

We previously detected with high reliability about one quarter of the proteins of *E. coli* (1147 proteins) in cells grown under a single condition in minimal medium and compared these proteins to global mRNA levels. To understand the relationship between protein detection and mRNA abundance in greater depth we here consider it for specific gene groups (translation apparatus, energy metabolism, motility and chemotaxis, cofactor biosynthesis, transcriptional regulators, and membrane and membrane associated proteins). We also present a data visualization tool that facilitates comparison of whole cell mRNA and protein profiles. In most instances protein detection was associated with a high level of mRNA, as well as with greater protein length and solubility. We failed to detect cognate mRNA for only 34 of the proteins we identified.

INTRODUCTION

Physiological studies of organisms whose genome sequences have been determined have been greatly advanced by development of new techniques to sample cell composition globally at mRNA, protein, and metabolite levels. Availability of global data has led to introduction of the term "systems biology". Affymetrix GeneChip microarrays allow the comparison of mRNA levels under different growth conditions and also provide statistical estimates of mRNA abundance and which genes are transcribed under a single condition [1-4]. Analysis of complex mixtures of tryptic peptides by mass spectrometry provides a powerful method for determining the protein composition of cells [5-7]. Availability of both types of data has allowed comparisons between them and several such comparisons have been made for the yeast *Saccharomyces cerevisiae* [8-10].

We recently made general comparisons between proteins and mRNAs detected in *E. coli* strain MG1655 (CGSC 6300) grown in minimal medium with glycerol as carbon source [11]. Globally there appeared to be a positive relationship between protein detection, performed under conditions that favored proteins of greatest abundance, and mRNA levels. Here we extend our earlier studies by presenting a simple visualization approach to facilitate comparisons between protein and mRNA data and consider examples of specific genes and operons for which the biological literature allows more meaningful analysis.

RESULTS

Data visualization

We created a simple visualization tool that allowed us to display protein and mRNA presence calls in genome order (Fig. 1, http://coli.berkeley.edu/protein_profile/), as we did previously for DNA microarray data [12-15]. The resulting genome image facilitates analysis of the data, particularly in terms of operon organization. This tool allows displaying the protein and mRNA detection either as separate squares, each corresponding to a detected protein or mRNA, or as a vertical rectangle for cases where both protein and mRNA were detected for a particular gene. The image map on the web site allows the gene ID number [16], gene name, and gene description to be displayed above the image. Clicking on a spot of interest transfers the user to the *E. coli* Entry Point database (http://coli.berkeley.edu/

Figure 1 - Genome image of protein and mRNA presence calls for *E. coli* MG1655 grown in minimal medium with glycerol as carbon source and NH_4Cl as nitrogen source.

Genes are arranged in their order on the chromosome of *E. coli* (according to the original *E. coli* annotation [16]) beginning with Blattner (b)0001 and progressing from left to right. There are 100 genes / row. To assist in viewing, each 10 genes are marked with a tick and a narrow vertical line and the background for the rows alternates between light and dark gray. Green bars indicate genes for which protein and mRNA were both detected. Yellow squares indicate the 34 genes for which protein but not mRNA was detected (see text), whereas blue squares indicate genes for which RNA but not protein was detected. Boxes correspond to some of the examples given in Table 1 and discussed in the text. Red boxes denote operons or clusters of operons with relatively abundant protein products. They are (in b number order): the *trp* operon (b1260-01265; b1265 is the *trp* leader); the *his* operon (b2018-2026; b2018 is the *his* leader); the *nuo* operon (b2276-2288), a cluster of ribosomal protein operons (b3294-3321), and the *atp* operon (b3731-3739). White boxes denote operons or clusters of operons with less abundant protein products. They are (in b number order): a cluster of murein-*fts* operons (b0081-0095); the *lac* operon and *lac* regulatory gene (b0342-0345; b0345 is *lacI*); and a cluster of flagellar (*fli*) operons (b1937-1950). At our web site (Protein data display: http://coli.berkeley.edu/protein_profile/), a cursor can be used to determine the b number, name, and description of each gene in the image. Links to the *E. coli* Entry Point (http://coli.berkeley.edu/ecoli/) facilitate obtaining additional information.

cgi-bin/ecoli/coli_entry.pl) [14], where useful information about the gene can be retrieved easily. In our comparison, cognate mRNAs were not detected (called “absent” by the Affymetrix algorithm) for only 34 proteins out of the total list of 1147 proteins (yellow boxes, Fig. 1) because none of the genes for these proteins had a high mRNA signal on the array (5-650; average for the transcriptome was 2000). Three of the proteins (products of *mcrB* [b0149], *ftsX* [b3462], and *alr* [b4053]) are known to be required for murein synthesis and cell division (see below) and one for fatty acid biosynthesis (product of *fabF* [b1095]) [16]. We happen to know independently that the GlnG regulatory protein (=NtrC; product of b3868) is also present [12, 17, 18]. Expression of the genes for these five proteins should have been detected at the mRNA level. We estimate that the limit for protein detection in our experiments was at 50-100

protein copies per cell, whereas Affymetrix microarrays can detect 1 molecule of RNA in a complex mixture of 100,000 distinct RNA molecules [11, 19]. Another 9 proteins whose cognate mRNAs were "absent" were designated ORFs.

Gene examples

To make our understanding of the protein profile concrete, we looked at a number of specific examples (Table 1, Fig. 1; see also supplementary material).

Abundant proteins

We began with abundant proteins we expected to find on the protein list: ribosomal proteins, enzymes of glycerol and central carbon metabolism, and amino acid biosynthetic enzymes. Ribosomal proteins, which are among the most abundant in the cell (~15,000 copies under our growth conditions [20]), were well-represented in the total list of proteins. Of the 55 ribosomal proteins, 49 were identified and mRNA was detected for all 55 genes. The ribosomal proteins that were missed had very few predicted tryptic peptides (1 to 4, whereas the average was 6.0). Nine of the 12 proteins involved in glycerol catabolism were detected, including all of those known to be required for growth, and expression of all 12 genes was detected at the mRNA level. The glycerol facilitator (GlpF) and glycerol phosphate permease (GlpT) were detected only in the membrane sample [11]. Most proteins categorized as glycolytic (14 out of 18), gluconeogenic (all 4) or as components of the tricarboxylic acid cycle (15 out of 17) [21] were detected and their mean mRNA signal intensities, which can be considered approximations of mRNA levels (Affymetrix Inc., Santa Clara, CA, technical note, 2001), were high. Those not detected are not required for the process involved (products of *fruK* and *fumB*), or would not have been detected due to low mRNA levels or small numbers of tryptic fragments (*gapC_1*, *gapC_2*, *fruL* and *farR*). Membrane-bound components of succinate dehydrogenase were detected only in the membrane sample. All the enzymes required for synthesis of the amino acids histidine and tryptophan were detected, as were the corresponding mRNAs. As expected, expression of the regulatory leader regions [22] for the *his* and *trp* operons, *hisL* and *trpL*, was detected at the mRNA but not the protein level. This was true generally for leader regions (including *fruL* mentioned above).

Proteins of low abundance

We next looked for proteins that were not expected to be abundant: flagellar proteins, proteins of the lactose degradative operon, and transcriptional regulators. Flagellar proteins are poorly expressed in strain MG1655 (CGSC 6300) [23, 24]. Of 40 proteins classified as flagellar proteins [21], the only one detected was the product of *fliY* (b1920), which lies at the edge of a cluster of flagellar operons. The *fliY* gene had an mRNA signal intensity of 5500, whereas the mean signal intensity for all other flagellar genes was <200 (the average for an *E. coli* protein was 2000). It has been reported that *fliY* may not be a flagellar gene [25], and there is direct evidence that its product is a periplasmic binding protein for cystine [26, 27]. Products of the lactose utilization operon were not detected and mRNA was detected (unreliably) only for *lacZ* (Table 1 legend). Of 160 known DNA-binding transcriptional regulators [28], we detected only 37 (23%), whereas 124 were considered expressed at the mRNA level. The average mRNA signal intensity for genes corresponding to regulators detected at the protein level was eight-fold higher than that for genes corresponding to undetected regulators.

We considered examples of proteins utilized for synthesis of co-factors that are required for growth in minimal medium because we thought their expression levels might not be high. All of the genes whose products are thought to be required for synthesis of NAD, pyridoxine, riboflavin, thiamine, and biotin (35 total) ([29] and J. Cronan, personal communication) were expressed at the mRNA level, and half of their protein products were present in the total list. The average mRNA signal intensities for these five groups of genes were between 1300 and 2900 (Table 1). It is known that some of the enzymes involved in co-factor synthesis have low turnover numbers (J. Cronan, personal communication), and hence both transcripts and proteins may be more abundant than anticipated.

Finally, we considered a group of proteins whose expression levels were expected to vary widely within the group – proteins required for cell division (*fts* gene products). Of the 12 Fts proteins [30, 31], most of which are membrane bound or associated, five (FtsI, N, X, Z, and ZipA) were detected and 11 were expressed at the mRNA level. FtsZ is by far the most abundant of the Fts proteins [32] and is soluble. FtsI and ZipA are also relatively abundant [33, 34] and have large soluble domains [30]. Several Fts proteins that were not detected are thought to be present at much lower levels or are of unknown abundance ([30] and D. Weiss, personal communication). One of these, FtsW, is an intrinsic membrane protein and FtsL and FtsB(=YgbQ) are small transmembrane proteins [30].

Membrane and membrane-associated proteins

We looked at additional membrane proteins or proteins associated with membranes: the F_1F_0 ATPase and NADH dehydrogenase I, abundant proteins known to be present in cells grown in minimal medium, and products of the 23 experimentally studied ATP-binding cassette (ABC) transport operons [35] for which we detected at least one protein product. Six of the nine gene products that constitute the ATP synthase were detected and expression of all nine genes, which constitute a single operon (*atp*), was detected at the mRNA level. Two of the three proteins that were missed had only one or three predicted tryptic peptides, whereas the average number for proteins of the operon was 10.6. Ten of the 13 gene products that constitute NADH dehydrogenase I were detected, several only in the membrane sample. Expression of all 13 genes,

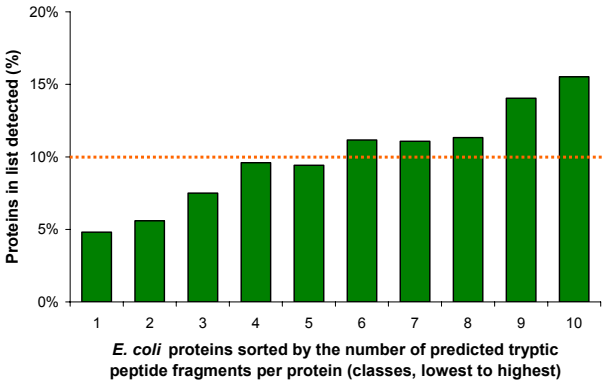

Figure 2 - Detection of *E. coli* proteins as a function of codon usage bias and the number of tryptic peptide fragments per protein.

To get the X axis, all 4290 proteins were first sorted from lowest to highest number of tryptic peptides per protein. They were then divided into 10 equal classes (429 each). Class 1 contained the 10% of proteins with the smallest number of tryptic peptides and so on. The proportion (percent) of all detected proteins in each class was plotted on the Y axis. The dotted line represents the hypothetical percent of detected proteins in each class (10%) if there is no relationship between protein detection and the number of tryptic peptides per protein. The ranges of the number of predicted tryptic peptides for classes 1-10 were as follows: (1) 0-4, (2) 4-6, (3) 6-7, (4) 7-9, (5) 9-11, (6) 11-13, (7) 13-15, (8) 15-19, (9) 19-25, (10) 25-80.

which constitute a single operon (*nuo*), was detected at the mRNA level. The three proteins that were missed were transmembrane subunits and were short relative to those detected (average of 4.3 vs. 14.8 predicted tryptic peptides).

ABC transport systems are usually composed of three different types of proteins: soluble periplasmic binding proteins, trans-membrane transport proteins, and inner-membrane associated ATP-binding proteins. Twenty soluble periplasmic binding protein components of ABC transporters were detected in the cell extract (the total number of periplasmic components in the 23 operons considered was 21). The genes corresponding to them generally had the highest mRNA signal intensities in their operons. By contrast, about half of the membrane-associated ATP-binding components coded by these operons and only one quarter of the integral membrane proteins (much shorter than the others) were detected. With one or two exceptions, membrane and membrane-associated components were detected exclusively in the membrane sample. Thus, preparation of the membrane sample was essential for detection of membrane components of ABC transport operons and useful for detection of other membrane and membrane-associated proteins discussed above.

General observations

Consideration of the above examples indicated that protein detection showed a positive relationship to mRNA level for the corresponding gene, a relationship that pertains globally [11]. Protein detection also shows a positive relationship to protein hydrophilicity and protein length (expressed as the number of predicted tryptic peptides). Our ability to detect *E. coli* proteins was progressively better with a higher number of predicted tryptic peptides per protein (Fig. 2). The positive relationship was especially strong for short proteins (0-12 tryptic peptides) but became gradually weaker for longer proteins (Fig. 3) because detection of only a single tryptic peptide was sufficient to call the protein "present" [11]. We showed previously that there was not a positive linear relationship between the number of tryptic peptides per protein and mRNA signal intensity for the corresponding gene [11].

Comparison of the protein list to the lists of proteins identified on 2D gels

Neidhardt and colleagues pioneered the use of two-dimensional (2D) gel electrophoresis to determine the protein composition of *E. coli* [39], an approach that has been intensively pursued by others [40-43]. We extracted from the SWISS-2DPAGE protein database (http://us.expasy.org/ch2d/) a list of proteins identified on 2D gels from cells grown under a variety of conditions, not just the single growth condition we used for our experiments. A search for *E. coli* proteins yielded a list of 336 unique protein names (as of August 2004). We detected 86% of these unique proteins and detected the cognate mRNAs for 96% of them (Table 2). As was true for proteins on our total list [11], cognate mRNAs for proteins identified on 2D gels had a higher average signal intensity than those for all *E. coli* proteins or all expressed proteins (Tables 1 and 2). Although the average signal intensity of cognate mRNAs for proteins we detected from the 2D gel list was higher than that for proteins we failed to detect, the average signal intensity for proteins we failed to detect was nevertheless higher than overall averages. The proteins identified on 2D gels were longer than the average *E. coli* protein, and those we detected were longer than those we failed to detect. Proteins identified on 2D gels were five to six fold low in membrane proteins and more than threefold low in proteins of unknown function.

Recently, a large-scale analysis of *E. coli* 2D gel maps was carried out by Lopez-Campistrous and colleagues [44], who were able to identify 575 *E. coli* proteins. Among these, 450 (78%) were present in our protein list, and mRNAs were detected for 94% of them. Properties of the 575 proteins and their corresponding mRNAs were similar to those from the SWISS-2DPAGE protein list, except that the number of membrane proteins and proteins of unknown function was increased (Table 2). Those proteins among the 575 that we did not detect had somewhat low mRNA levels (average signal of 1500 versus 2000 for *E. coli* protein-coding genes).

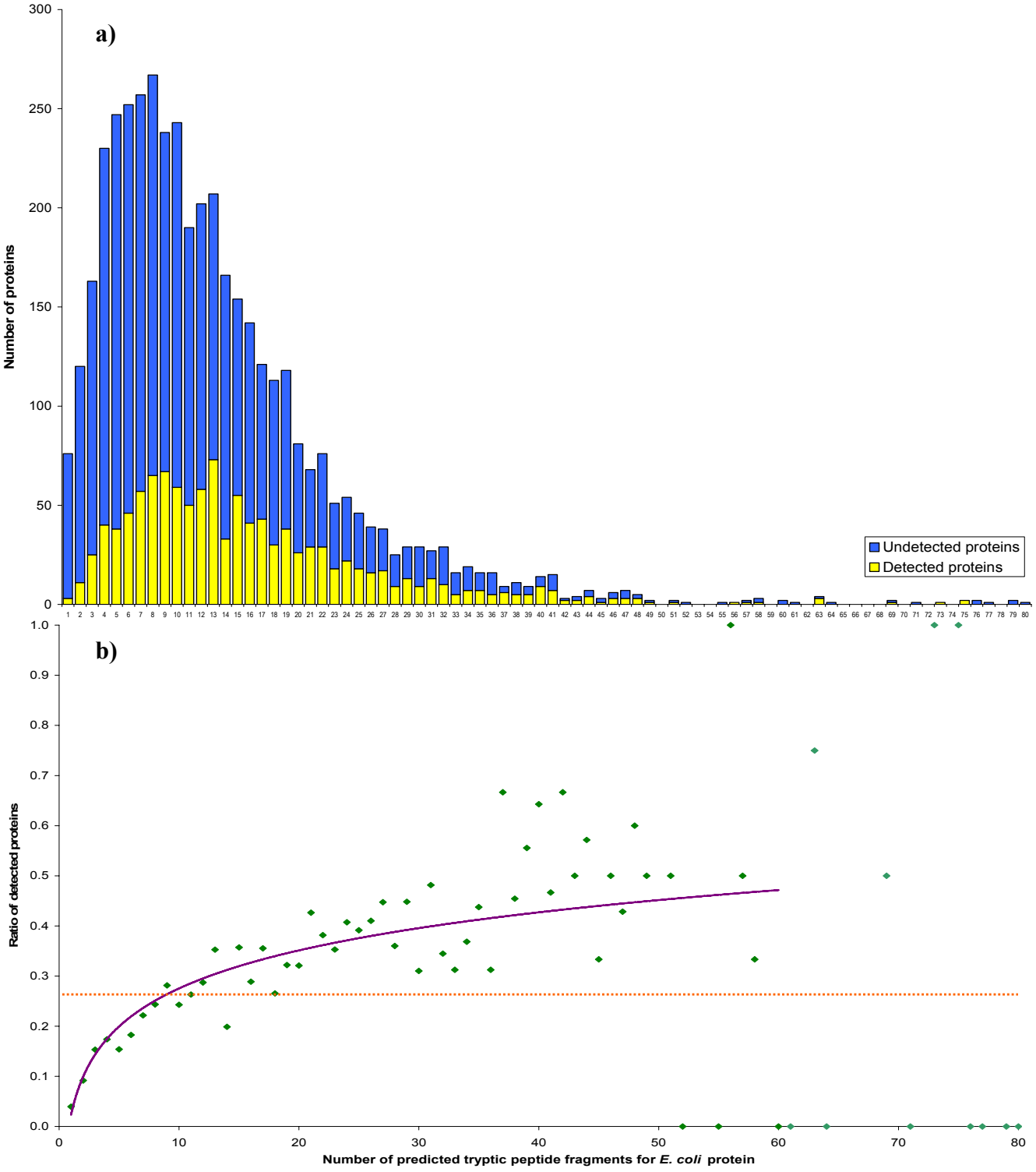

Figure 3 - Protein detection as a function of the number of predicted tryptic peptide fragments per *E. coli* protein.

(a) The X axis shows the number of predicted tryptic peptide fragments per protein. The numbers of detected (yellow) and undetected (blue) proteins are plotted on the Y axis as stacked columns. Each vertical combination of yellow and blue bars represents all *E. coli* proteins with that particular number of predicted tryptic peptides. **(b)** The X axis is as in (a). The Y axis shows a ratio of detected proteins to all proteins for each group of proteins with a particular number of predicted tryptic peptides. The purple line represents a log-fit curve of the data (only values for proteins with 1-60 tryptic peptide fragments were used to produce the curve). The dotted orange line represents the hypothetical case if there is no relationship between protein detection (26.7% overall) and the number of tryptic peptide fragments per protein.

Table 1 - Protein examples

Functional category and description[a]	No. of proteins	Average mRNA signal[b]	Proteins detected	mRNAs detected[c]
Ribosomal	55	28100	49	55
Glycerol catabolism	12	6700	9	12
Glycolysis and gluconeogenesis	22	6200	18	22
TCA cycle	17	12700	15	16
Histidine biosynthesis[d]	8	4700	8	8
Tryptophan biosynthesis[d]	5	3900	5	5
Flagella	40	300	1	12
Lactose catabolism	4	200	0	2[e]
Transcriptional regulators[f]	160	1600	37	124
Co-factor biosynthesis[g]	35	1900	19	35
Cell division (Fts)[h]	12	2500	5	11
F_1F_0 ATPase	9	18000	6	9
NADH dehydrogenase I	13	7600	10	13
ABC transporters[i]				
-Periplasmic	21	6600	20	21
-ATP-binding	23	1900	11	23
-Membrane	31	1500	8	28
All *E. coli* proteins	4291	2000[j]	1147	2826

[a] Unless otherwise stated, according to Riley and Labedan [21].
[b] For genes corresponding to all proteins in the group, rounded to the nearest hundred.
[c] Affymetrix presence calls (see [11]).
[d] Leader peptides not included.
[e] The mRNA for *lacI* was called present in all three experiments, whereas *lacZ* mRNA was present in one, marginal in one, and absent in one [11].
[f] Known DNA-binding transcriptional regulators, according to Perez-Rueda and Collado-Vides [28].
[g] NAD, pyridoxine, riboflavin, and thiamine, according to Koonin and Galperin [29], and biotin, according to J. Cronan (personal communication).
[h] [30, 31], D. Weiss, personal communication.
[i] Known ABC transporters according to Paulsen (http://66.93.129.133/transporter/wb/index2.html; [35]). This is not the entire category. Rather, if any protein product of an experimentally studied ABC transport operon was detected, we considered all members of that operon. Proportions of proteins detected were essentially the same for predicted ABC transport operons.
[j] For the 2826 protein coding genes called present at the mRNA level, the average mRNA signal was 2900.

Table 2 - *E. coli* proteins detected on 2D gels

Profile element	Swiss-2DPAGE [a]	Lopez-Campistrous *et al.* [b]
Proteins identified on 2D gels	336	575
Number detected in this work[c]	291	450
Number with mRNA detected[c]	322	543
Average mRNA signal[d]	7100	5300
Average number of tryptic peptides predicted[e]	16.3	16.0
Membrane proteins[c]	22	76
Proteins of unknown function[f]	29	83

[a] *E. coli* proteins in the Swiss-2DPAGE database (http://us.expasy.org/ch2d/) as of August 2004 (complete list available on web site).
[b] List of *E. coli* proteins identified on 2D gels by Lopez-Campistrous *et al* [44].
[c] See [11].
[d] Rounded to the nearest hundred. The average signals for proteins detected and undetected were 7700 and 3200, respectively, for Swiss-2DPAGE list; and 6300 and 1500, respectively, for Lopez-Campistrous *et al.* list. The average signal for an *E. coli* protein was 2000.
[e] The average numbers for genes corresponding to proteins detected and undetected were 16.9 and 12.5, respectively, for Swiss-2DPAGE list; and 16.3 and 14.7, respectively, for Lopez-Campistrous *et al.* list. The average *E. coli* protein has 13.1 peptides.
[f] Open reading frames [21].

Table 3 - Minimal gene sets

Profile element	Mushegian-Koonin MGS[a]	Gil *et al.* MGS[b]	*Buchnera aphidicola* BAp genome[c]
Genes/proteins in the set	255	206	582
Genes/proteins with *E. coli* orthologues	243	203	574
Proteins we detected[d]	184	160	387
mRNAs we detected[d]	236	200	540
Average mRNA signal[e]	10300	11400	6900

[a] A minimal gene set as defined by Mushegian and Koonin [45] by comparing the genomes of *Mycoplasma genitalium* and *Haemophilus influenzae*.
[b] A minimal gene set as defined by Gil *et al.* [47] by comparing genomes of five completely sequenced endosymbionts.
[c] Protein-coding genes for *Buchnera aphidicola* BAp [48, 49].
[d] See [11]
[e] Rounded to the nearest hundred. The average signals for genes corresponding to proteins detected and undetected in the minimal lists were 12200 and 4200, respectively, for Mushegian and Koonin list; and 13200 and 4900, respectively, for Gil *et al.* list; and those for genes corresponding to proteins detected and undetected in the *Buchnera* list were 9000 and 2800, respectively. The average signal for an *E. coli* protein was 2000.

Comparison of the protein and mRNA lists to minimal gene sets and to the genome of *Buchnera spp*

By comparing the genomes of *Haemophilus influenzae* (about 1700 genes) and *Mycoplasma genitalium* (about 470 genes) Mushegian and Koonin [45] identified a set of 255 orthologous genes [46]. In a more recent study, Gil and co-workers [47] determined a common core set of 206 genes by comparing five sequenced genomes of endosymbionts. We were interested in whether we detected proteins and mRNAs from these sets, which were inferred to be essential, and in whether genes from these sets were expressed at a level higher than average for *E. coli*. Likewise, we considered a list of *E. coli* genes orthologous to those of *Buchnera aphidicola* [48, 49], an obligate endosymbiont that is closely related to *E. coli* [48]. The genome of *B. aphidicola* contains only about 580 protein-coding genes and apparently evolved from the genome of the last common ancestor of *Buchnera* and *E. coli* through gene loss [49]. Many *Buchnera* genes are considered essential for cellular function [48].

We detected high percentages of proteins in both minimal gene sets and the *Buchnera* / *E. coli* orthologous gene list (Table 3). Properties of proteins in the minimal and *Buchnera* lists and their cognate mRNAs differed from those of all *E. coli* proteins in the ways described above for proteins in the 2D gel lists. Proteins we detected or failed to detect in each list also differed from each other in the ways described above (legend to Table 3). A number of proteins we failed to detect in the *Buchnera* list have already been discussed above (e.g. Fts proteins and members of their operons required for murein synthesis; components of the F_1F_0 ATPase and NADH dehydrogenase I). In addition, more than half of the 155 proteins that we failed to detect in the *Buchnera* list were proteins of unknown or putative function (44 and 11 proteins, respectively) or were "flagellar" proteins (27 proteins). Our failure to detect the latter is due to a peculiarity of the MG1655 (CGSC 6300) strain used in these experiments (see above). *Buchnera* does not have flagella and lacks, in particular, genes coding for flagellar filament proteins [48]. It has been speculated that the proteins designated "flagellar" in *Buchnera* may constitute a type III secretion system [48].

DISCUSSION

It has long been estimated that *E. coli* expresses at least a quarter of its genome under a single growth condition [20]. It is now clear that this is a minimum estimate. Considering the list of 1147 proteins detected by high pressure liquid chromatography-tandem mass spectrometry (HPLC-MS/MS) in the context of the operons containing their cognate genes indicates that *E. coli* probably expresses at least one-third of its proteins (~1600) in minimal glycerol medium [11]. The community is now poised to begin determining global differences in the amounts of *E. coli* proteins under different growth conditions [44, 50, 51], an undertaking that will help with the biological interpretation of the failure to detect a particular protein under any one condition. For example, although we failed to detect more than half of the Fts proteins, these are essential for growth under all conditions ([30, 31] and D. Weiss, personal communication), and hence must be present. By contrast, enzymes of the *lac* operon are in some sense absent in cells grown on glycerol.

The "genome image" visualization approach shown in Fig. 1 can help biologists interpret mRNA and protein profiles of cells more easily. Such data visualization allows researchers to make quick qualitative assessments of these profiles, especially for bacteria and archaea. In organisms belonging to these two kingdoms of life genes of common function often form multi-gene expression units (operons) and hence are adjacent on the chromosome. Genes in operons are seen as strings of the same color on a genome image (see footnote to Fig. 1) and thus their qualitative behavior is readily determined. As quantitative comparisons between protein levels in different samples become available the program can be further extended to incorporate them together with comparisons of the corresponding mRNA levels.

CONCLUSIONS

Considering specific examples of genes, operons, and groups of genes enriched understanding of protein and mRNA detection in *E. coli* grown under a single condition and global relationships between them. A simple visualization tool facilitated qualitative comparisons between protein and mRNA profiles across operons.

METHODS

Data analysis

The data set was as described in [11]. The protein composition of a whole cell lysate of *E. coli* strain MG1655 (CGSC6300) [24] grown in minimal-glycerol medium was acquired from a trypsin-digested protein extract using HPLC-MS/MS. Affymetrix *E. coli* Antisense GeneChips were used to obtain mRNA levels and mRNA presence calls under the same conditions.

Gene and protein names, ID numbers, and descriptions were taken from the *E. coli* Entry Point database (http://coli.berkeley.edu/cgi-bin/ecoli/coli_entry.pl). These were as defined by Blattner *et al.* [16]. Gene functional categories were as originally defined by Riley and Labedan [21]. The list of *E. coli* proteins identified by 2D electrophoresis was downloaded from the SWISS-2DPAGE database (http://us.expasy.org/ch2d/). All data comparisons were performed in EXCEL with the help of VISUAL BASIC scripts. The complete gene lists used here are available at the web site.

ACKNOWLEDGEMENTS

We thank John Cronan, Kelly Hughes, and David Weiss for information on co-factor synthesis, flagellar

function, and cell division, respectively, Julio Collado-Vides for providing a list of known transcriptional regulators, and Francisco J. Silva and Andres Moya for providing a list of the genes of *Buchnera*.

This work was supported by National Institutes of Health grants GM37537 to DH and GM38361 to SK and by Wright Brothers Institute grant WBSC9004A to OP.

REFERENCES

1. Selinger DW, Cheung KJ, Mei R, Johansson EM, Richmond CS, Blattner FR, Lockhart DJ, Church GM: **RNA expression analysis using a 30 base pair resolution Escherichia coli genome array**. *Nat Biotechnol* 2000, **18**(12):1262-1268.
2. Wassarman KM, Repoila F, Rosenow C, Storz G, Gottesman S: **Identification of novel small RNAs using comparative genomics and microarrays**. *Genes Dev* 2001, **15**(13):1637-1651.
3. Hubbell E, Liu WM, Mei R: **Robust estimators for expression analysis**. In: *Bioinformatics.* vol. 18; 2002: 1585-1592.
4. Liu WM, Mei R, Di X, Ryder TB, Hubbell E, Dee S, Webster TA, Harrington CA, Ho MH, Baid J *et al*: **Analysis of high density expression microarrays with signed-rank call algorithms**. In: *Bioinformatics.* vol. 18; 2002: 1593-1599.
5. Gygi SP, Rist B, Gerber SA, Turecek F, Gelb MH, Aebersold R: **Quantitative analysis of complex protein mixtures using isotope-coded affinity tags**. *Nat Biotechnol* 1999, **17**(10):994-999.
6. Florens L, Washburn MP, Raine JD, Anthony RM, Grainger M, Haynes JD, Moch JK, Muster N, Sacci JB, Tabb DL *et al*: **A proteomic view of the Plasmodium falciparum life cycle**. In: *Nature.* vol. 419; 2002: 520-526.
7. Lipton MS, Pasa-Tolic L, Anderson GA, Anderson DJ, Auberry DL, Battista JR, Daly MJ, Fredrickson J, Hixson KK, Kostandarithes H *et al*: **Global analysis of the Deinococcus radiodurans proteome by using accurate mass tags**. In: *Proc Natl Acad Sci U S A.* vol. 99; 2002: 11049-11054.
8. Futcher B, Latter GI, Monardo P, McLaughlin CS, Garrels JI: **A sampling of the yeast proteome**. *Mol Cell Biol* 1999, **19**(11):7357-7368.
9. Griffin TJ, Gygi SP, Ideker T, Rist B, Eng J, Hood L, Aebersold R: **Complementary profiling of gene expression at the transcriptome and proteome levels in Saccharomyces cerevisiae**. *Mol Cell Proteomics* 2002, **1**(4):323-333.
10. Ghaemmaghami S, Huh WK, Bower K, Howson RW, Belle A, Dephoure N, O'Shea EK, Weissman JS: **Global analysis of protein expression in yeast**. In: *Nature.* vol. 425; 2003: 737-741.
11. Corbin RW, Paliy O, Yang F, Shabanowitz J, Platt M, Lyons CE, Jr., Root K, McAuliffe J, Jordan MI, Kustu S *et al*: **Toward a protein profile of Escherichia coli: comparison to its transcription profile**. *Proc Natl Acad Sci U S A* 2003, **100**(16):9232-9237.
12. Zimmer DP, Soupene E, Lee HL, Wendisch VF, Khodursky AB, Peter BJ, Bender RA, Kustu S: **Nitrogen regulatory protein C-controlled genes of Escherichia coli: scavenging as a defense against nitrogen limitation**. *Proc Natl Acad Sci U S A* 2000, **97**(26):14674-14679.
13. Wendisch VF, Zimmer DP, Khodursky A, Peter B, Cozzarelli N, Kustu S: **Isolation of Escherichia coli mRNA and comparison of expression using mRNA and total RNA on DNA microarrays**. In: *Anal Biochem.* vol. 290; 2001: 205-213.
14. Zimmer DP, Paliy O, Thomas B, Gyaneshwar P, Kustu S: **Genome image programs: visualization and interpretation of Escherichia coli microarray experiments**. In: *Genetics.* vol. 167; 2004: 2111-2119.
15. Gyaneshwar P, Paliy O, McAuliffe J, Popham DL, Jordan MI, Kustu S: **Sulfur and nitrogen limitation in Escherichia coli K-12: specific homeostatic responses**. *J Bacteriol* 2005, **187**(3):1074-1090.
16. Blattner FR, Plunkett G, 3rd, Bloch CA, Perna NT, Burland V, Riley M, Collado-Vides J, Glasner JD, Rode CK, Mayhew GF *et al*: **The complete genome sequence of Escherichia coli K-12**. In: *Science.* vol. 277; 1997: 1453-1474.
17. Magasanik B: **Regulation of transcription of the glnALG operon of Escherichia coli by protein phosphorylation**. *Biochimie* 1989, **71**(9-10):1005-1012.
18. Reitzer LJ, Magasanik B: **Transcription of glnA in E. coli is stimulated by activator bound to sites far from the promoter**. *Cell* 1986, **45**(6):785-792.
19. **GeneChip arrays provide optimal sensitivity and specificity for microarray expression analysis**. In. Affymetrix, Santa Clara, Ca: Affymetrix Technical Note.
20. Neidhardt FC, Ingraham JL, Schaechter M: **Physiology of the bacterial cell: a molecular approach**: Sinauer Associates, Sunderland, MA; 1990.
21. Riley M, Labedan B: **E. coli gene products: Physiological functions and common ancestries**. In: *Escherichia coli and Salmonella: Cellular and Molecular Biology.* Edited by Neidhardt F, Curtiss R, Lin ECC, Ingraham J, Low KB, Magasanik B, Reznikoff W, Riley M, Schaechter M, Umbarger HE. Washington, D.C.: ASM Press; 1996: 2118-2202.
22. Landick R, Turnbough CL, Jr., Yanofsky C: **Transcription Attenuation**. In: *Escherichia coli and Salmonella: Cellular and Molecular Biology.* Edited by Neidhardt F, Curtiss R, Lin ECC, Ingraham J, Low KB, Magasanik B, Reznikoff W, Riley M, Schaechter M, Umbarger HE. Washington, D.C.: ASM Press; 1996: 1263-1286.
23. Lehnen D, Blumer C, Polen T, Wackwitz B, Wendisch VF, Unden G: **LrhA as a new transcriptional key regulator of flagella, motility and chemotaxis genes in Escherichia coli**. *Mol Microbiol* 2002, **45**(2):521-532.
24. Soupene E, van Heeswijk WC, Plumbridge J, Stewart V, Bertenthal D, Lee H, Prasad G, Paliy O, Charernnoppakul P, Kustu S: **Physiological studies of Escherichia coli strain MG1655: growth defects and apparent cross-regulation of gene expression**. *J Bacteriol* 2003, **185**(18):5611-5626.
25. Ikebe T, Iyoda S, Kutsukake K: **Structure and expression of the fliA operon of Salmonella typhimurium**. *Microbiology* 1999, **145 (Pt 6)**:1389-1396.
26. Butler JD, Levin SW, Facchiano A, Miele L, Mukherjee AB: **Amino acid composition and N-terminal sequence of purified cystine binding protein of Escherichia coli**. *Life Sci* 1993, **52**(14):1209-1215.
27. Quadroni M, Staudenmann W, Kertesz M, James P: **Analysis of global responses by protein and peptide fingerprinting of proteins isolated by two-dimensional gel electrophoresis. Application to the sulfate-starvation response of Escherichia coli**. *Eur J Biochem* 1996, **239**(3):773-781.
28. Perez-Rueda E, Collado-Vides J: **The repertoire of DNA-binding transcriptional regulators in Escherichia coli K-12**. *Nucleic Acids Res* 2000, **28**(8):1838-1847.

29. Koonin EV, Galperin MY: **Sequence-evolution-function: Computational approaches in compara-tive genomics**: Kluwer Academic Publishers, Boston, USA; 2003.
30. Errington J, Daniel RA, Scheffers DJ: **Cytokinesis in bacteria**. *Microbiol Mol Biol Rev* 2003, **67**(1):52-65, table of contents.
31. Gill DR, Hatfull GF, Salmond GP: **A new cell division operon in Escherichia coli**. *Mol Gen Genet* 1986, **205**(1):134-145.
32. Lu C, Stricker J, Erickson HP: **FtsZ from Escherichia coli, Azotobacter vinelandii, and Thermotoga maritima--quantitation, GTP hydrolysis, and assembly**. *Cell Motil Cytoskeleton* 1998, **40**(1):71-86.
33. Dougherty TJ, Kennedy K, Kessler RE, Pucci MJ: **Direct quantitation of the number of individual penicillin-binding proteins per cell in Escherichia coli**. *J Bacteriol* 1996, **178**(21):6110-6115.
34. Hale CA, de Boer PA: **Direct binding of FtsZ to ZipA, an essential component of the septal ring structure that mediates cell division in E. coli**. *Cell* 1997, **88**(2):175-185.
35. Dassa E, Hofnung M, Paulsen IT, Saier MH, Jr.: **The Escherichia coli ABC transporters: an update**. *Mol Microbiol* 1999, **32**(4):887-889.
36. Carbone A, Zinovyev A, Kepes F: **Codon adaptation index as a measure of dominating codon bias**. *Bioinformatics* 2003, **19**(16):2005-2015.
37. Sharp PM, Li WH: **The codon Adaptation Index--a measure of directional synonymous codon usage bias, and its potential applications**. *Nucleic Acids Res* 1987, **15**(3):1281-1295.
38. Ma J, Campbell A, Karlin S: **Correlations between Shine-Dalgarno sequences and gene features such as predicted expression levels and operon structures**. In: *J Bacteriol.* vol. 184; 2002: 5733-5745.
39. VanBogelen RA, Abshire KZ, Pertsemlidis A, Clark RL, Neidhardt FC: **Gene-Protein Database of Escherichia coli K-12, Edition 6**. In: *Escherichia coli and Salmonella: Cellular and Molecular Biology.* Edited by Neidhardt F, Curtiss R, Lin ECC, Ingraham J, Low KB, Magasanik B, Reznikoff W, Riley M, Schaechter M, Umbarger HE. Washington, D.C.: ASM Press; 1996: 2067-2117.
40. Champion KM, Nishihara JC, Joly JC, Arnott D: **Similarity of the Escherichia coli proteome upon completion of different biopharmaceutical fermentation processes**. In: *Proteomics.* vol. 1; 2001: 1133-1148.
41. Link AJ, Robison K, Church GM: **Comparing the predicted and observed properties of proteins encoded in the genome of Escherichia coli K-12**. In: *Electrophoresis.* vol. 18; 1997: 1259-1313.
42. Loo RR, Cavalcoli JD, VanBogelen RA, Mitchell C, Loo JA, Moldover B, Andrews PC: **Virtual 2-D gel electrophoresis: visualization and analysis of the E. coli proteome by mass spectrometry**. In: *Anal Chem.* vol. 73; 2001: 4063-4070.
43. Tonella L, Hoogland C, Binz PA, Appel RD, Hochstrasser DF, Sanchez JC: **New perspectives in the Escherichia coli proteome investigation**. In: *Proteomics.* vol. 1; 2001: 409-423.
44. Lopez-Campistrous A, Semchuk P, Burke L, Palmer-Stone T, Brokx SJ, Broderick G, Bottorff D, Bolch S, Weiner JH, Ellison MJ: **Localization, annotation, and comparison of the Escherichia coli K-12 proteome under two states of growth**. *Mol Cell Proteomics* 2005, **4**(8):1205-1209.
45. Mushegian AR, Koonin EV: **A minimal gene set for cellular life derived by comparison of complete bacterial genomes**. In: *Proc Natl Acad Sci U S A.* vol. 93; 1996: 10268-10273.
46. Koonin EV: **How many genes can make a cell: the minimal-gene-set concept**. In: *Annu Rev Genomics Hum Genet.* vol. 1; 2000: 99-116.
47. Gil R, Silva FJ, Pereto J, Moya A: **Determination of the core of a minimal bacterial gene set**. *Microbiol Mol Biol Rev* 2004, **68**(3):518-537.
48. Shigenobu S, Watanabe H, Hattori M, Sakaki Y, Ishikawa H: **Genome sequence of the endocellular bacterial symbiont of aphids Buchnera sp. APS**. *Nature* 2000, **407**(6800):81-86.
49. Silva FJ, Latorre A, Moya A: **Why are the genomes of endosymbiotic bacteria so stable?** In: *Trends Genet.* vol. 19; 2003: 176-180.
50. Champion MM, Campbell CS, Siegele DA, Russell DH, Hu JC: **Proteome analysis of Escherichia coli K-12 by two-dimensional native-state chromatography and MALDI-MS**. In: *Mol Microbiol.* vol. 47; 2003: 383-396.
51. Gygi SP, Rist B, Gerber SA, Turecek F, Gelb MH, Aebersold R: **Quantitative analysis of complex protein mixtures using isotope-coded affinity tags**. In: *Nat Biotechnol.* vol. 17; 1999: 994-999.
52. Glasner JD, Liss P, Plunkett G, 3rd, Darling A, Prasad T, Rusch M, Byrnes A, Gilson M, Biehl B, Blattner FR *et al*: **ASAP, a systematic annotation package for community analysis of genomes**. In: *Nucleic Acids Res.* vol. 31; 2003: 147-151.