Localization, Annotation, and Comparison of the *Escherichia coli* K-12 Proteome under Two States of Growth*

Ana Lopez-Campistrous‡, Paul Semchuk‡, Lorne Burke‡, Taunya Palmer-Stone‡, Stephen J. Brokx§, Gordon Broderick‡, Drell Bottorff‡, Sandra Bolch‡, Joel H. Weiner‡§¶, and Michael J. Ellison‡§

Here we describe a proteomic analysis of *Escherichia coli* in which 3,199 protein forms were detected, and of those 2,160 were annotated and assigned to the cytosol, periplasm, inner membrane, and outer membrane by biochemical fractionation followed by two-dimensional gel electrophoresis and tandem mass spectrometry. Represented within this inventory were unique and modified forms corresponding to 575 different ORFs that included 151 proteins whose existence had been predicted from hypothetical ORFs, 76 proteins of completely unknown function, and 222 proteins currently without location assignments in the Swiss-Prot Database. Of the 575 unique proteins identified, 42% were found to exist in multiple forms. Using DIGE, we also examined the relative changes in protein expression when cells were grown in the presence and absence of amino acids. A total of 23 different proteins were identified whose abundance changed significantly between the two conditions. Most of these changes were found to be associated with proteins involved in carbon and amino acid metabolism, transport, and chemotaxis. Detailed information related to all 2,160 protein forms (protein and gene names, accession numbers, subcellular locations, relative abundances, sequence coverage, molecular masses, and isoelectric points) can be obtained upon request in either tabular form or as interactive gel images. *Molecular & Cellular Proteomics* 4:1205–1209, 2005.

Large scale proteomic analyses of experimental model organisms provide valuable resources to a broad range of investigators working on both general and specific aspects of cell function. Recently there has been renewed interest in the bacterium *Escherichia coli* as a proof-of-concept model for systems-based approaches (1–3). Here we describe a proteomic analysis of *E. coli* with an eye toward creating a resource of potential value to system approaches as well as problem-based approaches.

*E. coli* is probably the best understood of the simple model organisms and the most amenable to experimental analysis. Its genome has been fully sequenced (4), and the availability of complete genome sequence data bases facilitates the proteomic analysis of *E. coli* using MS. The proteome of an *E. coli* cell is estimated to have 4,285 proteins (5) with pI values ranging from 3.38 to 13.0 and molecular masses between 1.59 to 248 kDa (6–8). These proteins are distributed among four well defined subcellular compartments: 1) the cytosol (2,885 known and predicted species), 2) the inner membrane (670 known and predicted species), 3) the outer membrane (87 known and predicted species), and 4) the periplasm, which separates the two membranes (138 known and predicted species).

Recent “gel-free” proteomic approaches have coupled orthogonal chromatographic approaches with MS/MS (9) where one or more peptide tags serve as proxies for the identity, state, and abundance of a given protein. In the present work, we used the conventional and established method of 2D-PAGE (10) for protein separation and analysis based on several considerations. 1) The approach is both accepted and accessible. 2) With the ability to separate thousands of proteins on a single gel, the resolution of 2D-PAGE remains unchallenged. 3) With 2D-PAGE, protein mobility is highly sensitive to modification, making it possible to assess both the integrity and modification state of individual species. 4) The method lends itself readily to either absolute or relative quantification of intact protein species using several complementary approaches. 5) Gel visualization as the initial phase of analysis provides an immediate qualitative evaluation of the quality and global outcome of an experiment.

In the present study we biochemically fractionated *E. coli* into its subcellular components and created high resolution annotated two-dimensional electrophoresis protein gels of the whole cell, inner membrane, outer membrane, and the intervening periplasmic space at proteomic scale. From these gels, we collated the identity, location, abundance, modification state, apparent isoelectric point, and molecular mass for 2,160 protein spots corresponding to 424 known and 151

---

1 The abbreviation used is: 2D, two-dimensional.
putative genes. Building on these results, we used DIGE to determine the major differences in protein expression that occur when cells are grown in the presence or absence of amino acids.

RESULTS AND DISCUSSION

Proteomic Characterization of E. coli by Compartment—Shown in Fig. 1 and Supplemental Fig. 1 are two-dimensional electrophoresis gels of E. coli proteins derived from either whole cells or cells that had been biochemically fractionated into their periplasm, inner membrane, and outer membrane components. Circled spots correspond to protein species identified by tryptic digestion and LC/MS/MS analysis resulting in an average sequence coverage of ~20% for each polypeptide. The results of this analysis have been statistically summarized in Supplemental Tables 1 and 2. Detailed information for each entry is listed in Supplemental Table 3 and includes: protein and gene name, National Center for Biotechnology Information (NCBI) accession number, subcellular location (Swiss-Prot), subcellular location, and abundance (this work). All the information provided herein for each protein entry and their isoforms has been made available at www.projectcybercell.ca, which also includes the sequence coverage for all identified spots and their predicted and measured molecular mass and pI. The data can be accessed interactively in either tabular form or by selecting a given protein spot from any of the six displayed gels.

In all, 2,160 species were identified representing 575 different ORFs. Among these are 151 proteins (and their locations) whose existence had been predicted from hypothetical ORFs and 76 proteins (and their locations) of unknown function. In the case of the whole cell component, focusing the first dimension over the narrow pH ranges 4.5–5.5 and 5.5–6.7 significantly increased spot resolution, resulting in the identification of 113 protein entries not found at the broader pH range (Supplemental Table 1).

Overall the efficiency of compartmental fractionation achieved here is evident from the dramatically different gel patterns observed for each compartment. An estimate of the purity of each compartment (inner membrane, 94%; periplasm, 87%; outer membrane, 87%) is derived from the sum of spot intensities corresponding to known or predicted members of a given compartment relative to the total spot intensity of the gel. These values are likely underestimates based on two considerations. 1) A small but significant fraction of inner and outer membrane proteins are known to co-localize at Bayer adhesion sites, points of contact between both membranes (11). 2) The location of a significant number of proteins that contribute to the total gel intensities have not been reported or predicted previously (Supplemental Table 2).

The assignment of proteins to compartments was determined by a conservative estimate of their relative enrichment from gel to gel. A total of 459 protein entries were found to be unique to one gel or another and were therefore immediately assigned to a corresponding compartment (Supplemental Table 1). Another 112 proteins partitioned sufficiently to a given compartment to warrant assignment (Supplemental Figs. 2–4). Of the protein entries in Swiss-Prot that presently have no designated location, 222 proteins have been assigned here. Other location assignments made here are in generally good agreement with those reported by Swiss-Prot (Supplemental Table 2) but not entirely free of discrepancy. The artifactual association of proteins and/or membrane during cell lysis and fractionation probably account in part for some of these discrepancies. For example, the ribosomal protein RplO and the recombination protein RecA are clearly cytosolic in light of their functions yet appear to co-localize with the outer membrane. Interestingly both proteins are known to form polymers (12, 13), a reasonable explanation for their (and possibly other’s) co-elution with the denser outer membrane fraction.

On the other hand, it is equally likely that some of the
seemingly controversial assignments made here have biological meaning. There are for example 12 proteins that Swiss-Prot has classified as cytosolic that we assigned to the inner membrane (AceE, AceF, FabZ, FruB, GlpD, ManX, Rne, RpoE, SdaB, ThrA, TrpD, and UspA). These discrepancies can be reasonably reconciled if these proteins were peripherally bound to the inner membrane on the cytoplasmic side. Similarly there are four proteins that are classified as periplasmic by Swiss-Prot that we assigned to either the inner or outer membrane (DacB, DcrB, HybA, and YraP). These apparent differences could be reasonably accommodated if these proteins were peripherally associated with the side of either the inner or outer membrane that faces the periplasmic space.

Our search of the available literature indicates that the assignments made here are likely correct for at least eight of these proteins (AceE and AceF (14), FabZ (15), FruB (16), GlpD (17), SdaB (18), DcrB (19), and HybA (20)).

Modified Protein Forms—Deviations between the observed and predicted Mr and/or pl of any given protein are often indicative of modifications such as chain cleavage or the covalent modification of amino acids. Supplemental Fig. 5 (A and B) illustrates the overall extent of protein modification as a correlation of the observed and expected Mr and pl. Of the 575 different protein entries compiled from the six reference gels (Fig. 1 and Supplemental Fig. 1), 241 (42%) were found to exist in more than one form at an average of 7.5 forms (spots) per entry (or 3.5 forms when all 575 entries are considered). For the 241 entries that were subject to modification, 70% of the modified forms varied only in their pl, whereas 22% varied only in their Mr. Only 8% of the variation could be attributed to changes in both pl and Mr. We conclude from these results that the majority of these forms arise from the modification of individual amino acids with particularly dramatic examples being OppA (distributed between 15 spots ranging over 7.4 pH units), AtaA (12 spots over 1.1 pH units), and GapA (14 spots over 2.8 pH units). A more comprehensive analysis of this type of modification reveals that they result from the deamidation of asparagine and glutamine to their acidic counterparts, aspartate and glutamate.2

Eighteen forms were identified that exhibited significant differences from the expected mass (Supplemental Fig. 2A, not labeled). Of these, four showed a higher than expected mass, whereas 14 exhibited a lower than expected mass. The nature of the higher than expected masses remains unresolved. Of the smaller mass variants, six were identified as truncated versions of full-length entries based on the extent of protein sequence covered by their tryptic fragments. Forms of GyrB, PrsA, and YacE carried C-terminal deletions, whereas Lon and AlaS carried N-terminal deletions, and Ppr appeared to be deleted from both ends of the predicted protein sequence. Although it is reasonable to conclude from these findings that these fragments arise from proteolytic cleavage, we note that in no instance has the sequence of the predicted protein been confirmed experimentally.3

When taken globally, the results of the present study reveal that the extent of E. coli protein modification is considerably greater than reported previously. Using a similar approach to that described here, Link et al. (22) reported that 18% of 223 unique proteins identified existed as modified forms compared with the 42% of 575 proteins identified in the present study, a difference of ~2.3-fold. Outside of the difference in sample population, differences in sample preparation and electrophoretic running conditions can also account for this discrepancy.

Global Changes in Protein Expression Resulting from Amino Acid Starvation—We examined the consequences of amino acid depletion on protein expression by DIGE. Protein samples derived from whole cells grown exponentially in the presence and absence of amino acids were first covalently modified with fluorescent dyes that exhibited different fluorescent properties (Cy5, with amino acids; Cy3, without amino acids), then combined, and run on the same gel (Supplemental Fig. 6). The relative contribution of protein to each spot was then determined by deconvolution of the fluorescent spectra of each dye.

A total of 29 spots were identified that exhibited significant differences (>2-fold) in abundance between the two conditions corresponding to 23 different proteins (Supplemental Table 4). Notably most of these proteins (18) showed elevated levels of expression in the absence of amino acids, consistent with the greater level of metabolic activity that would be required for amino acid biosynthesis and precursor uptake.

Fig. 2 is a low resolution metabolic map that positions the major changes in protein levels according to pathway. Most of the proteins up-regulated in the absence of amino acids are central metabolic enzymes involved in carbon/amino acid synthesis and transport. Also included in this category are binding proteins involved in chemotaxis (MalE, MglB, ArgT, and AldA) and proteins of the Leu transport system (LivK and LivJ). Previous work has shown that these chemotaxis proteins are induced during glucose-limited growth (23, 24); therefore their increase in the absence of amino acids may reflect a similar enhancement for nutrient sensing. On the other hand, induction of the Leu transport system is important for the retention of endogenously synthesized amino acids (25), clearly an advantageous adaptation to conditions where amino acids are limiting. Proteins with reduced levels of expression in the absence of amino acids (GuaB, RplL, Ssb, YncE, and YbiL) are relatively few and may reflect the down-regulation of global processes (DNA/RNA synthesis and translation) expected from cells that grow and divide more slowly.

In addition to the present study, there have been three


3 Swiss-Model Repository, a database for theoretical protein models (swissmodel.expasy.org/repository/).
sizable efforts to characterize the *E. coli* proteome that together account for 1302 unique protein entries, 29% of the known *E. coli* ORFs (4, 5). The results of all four studies have been statistically summarized in Table I. In addition to this study, two of these studies used 2D-PAGE and are therefore more easily compared. Link et al. (22) identified and assigned 218 different proteins to the four compartments examined here, whereas the Swiss 2D-PAGE data base (26–28) identified 336 different proteins from unfractionated whole cells. Combined both of these data sets overlap with our own by 71%. The most recent study generated a high precision list (403 entries) and a lower precision list (743 entries) of unique proteins by HPLC followed by tandem MS of tryptic peptides mixtures derived from either soluble (cytosol plus periplasm) or total membrane fractions (outer plus inner) of *E. coli* (21). These data sets exhibit less overlap with our study than the 2D-PAGE approaches noted above (55% for the high precision list, 38% for the lower precision list).

It is apparent from these comparisons that although HPLC/MS/MS outperforms 2D-PAGE both in terms of speed and volume of protein assignments, 2D-PAGE is unprecedented for other kinds of information such as protein integrity, modification state, absolute and relative abundances, pI, and *M*<sub>r</sub>. The fact that these approaches remain complementary underscores the two key challenges faced by all facets of proteomic research: the need for significant improvements in sensitivity and resolution.

* This work was supported by the Alberta Science Research Authority, Canada Western Economic Diversification, and the Canada Foundation for Innovation. The costs of publication of this article were defrayed in part by the payment of page charges. This article must therefore be hereby marked “advertisement” in accordance with 18 U.S.C. Section 1734 solely to indicate this fact.

**S** The on-line version of this article (available at [http://www.mcponline.org](http://www.mcponline.org)) contains supplemental material.

¶ Holds a Canada Research Chair in Membrane Biochemistry.

To whom correspondence should be addressed: Inst. for Biomolecular Design, 3–67 Medical Sciences Bldg., University of Alberta, Edmonton, Alberta T6G 2H7, Canada. Tel.: 780-492-6352; Fax: 780-492-9394; E-mail: mike.ellison@ualberta.ca.

**REFERENCES**

complement experimentation and trial design. The Scientist 18, 33–36


