**EchoLOCATION: an in silico analysis of the subcellular locations of Escherichia coli proteins and comparison with experimentally derived locations**

Richard S.P. Horler¹, Andrew Butcher¹, Nikitas Papangelopoulos¹, Peter D. Ashton², and Gavin H. Thomas¹*

¹Department of Biology (Area 10) and ²Technology Facility, Department of Biology, University of York, PO BOX 373, York YO10 5YW, UK

Associate Editor: Prof. Dmitrij Frishman

**ABSTRACT**

Summary: EchoLOCATION is a database that provides a comprehensive analysis of the subcellular locations of E. coli K-12 proteins. Locations are predicted by integrating data from a range of publicly available algorithms combined with extensive curation of experimental literature. The data can be searched in a variety of ways and can generate lists of subcellular proteomes for analysis. Experimental evidence supports the locations of over 500 envelope proteins (periplasm, inner and outer membrane). From analysis of disagreements between in silico predictions and experimental data we provide an analysis of protein types where subcellular prediction algorithms are currently not accurate.


Contact: ght2@york.ac.uk

**INTRODUCTION**

The bacterium Escherichia coli is a model organism that has been intensively studied in the laboratory and was one of the first organisms to have its complete genome sequenced (Blattner et al., 1997). Although E. coli is a unicellular prokaryote, its proteins can be sorted to a variety of different subcellular compartments (Fig. 1) and often proteins of particular molecular functions are localized to particular compartments. Our E. coli K-12 database, EchoBASE, aims to predict functions for uncharacterised gene products (Misra et al., 2005, Thomas, 1999a) by integrating a wide range of experimental and computational different data. We have developed EchoLOCATION as an additional tool operating within EchoBASE to provide a single source for information about protein locations in E. coli K-12 to aid in experimental and computational studies of this organism. Data from EchoLOCATION was included in the updated community annotation of E. coli K-12 released in 2006 (Riley et al., 2006) and more recently has been used successfully as a starting point for novel functional characterization of E. coli gene products (Ruiz et al., 2008; Lock and Harry, 2008). We herein describe this resource and present an analysis of the accuracy of experimental versus in silico prediction of the subcellular localization of proteins.

**CREATION AND ORGANIZATION OF ECHOLOCATION**

*To whom correspondence should be addressed.
as integral membrane proteins (location B). Outer membrane β-barrels proteins were determined by combining the output of both BBF and BOMP (location H). Proteins predicted to contain a type-II signal peptide (lipoproteins) with no TMHs or other conflicting evidence were marked up as lipoproteins with the inner (location E) or outer (location I) membrane location determined by the amino acid residue at the +2 position from the cleavage site (Seydel et al., 1999). Any protein with a type-I signal peptide and no TMHs was predicted to be periplasmic (location G). If the predicted TMH is within a signal peptide with a high probability of cleavage, then it is classed as periplasmic. Proteins which have one or two TMHs helices which are not removed by signal peptidase are classed as membrane anchored proteins (locations C and D). Finally, there were a small number of proteins where experimental evidence has suggested that they are secreted (location J).

ACCESSING ECHOLOCATION DATA

The EchoLOCATION data is available through the EchoBASE database and is connected to each individual gene page. When viewing a gene page the subcellular location is shown together with any comments for that protein, which links to the detailed EchoLOCATION page (Figure 2). This displays all the results for that protein and an assignment of subcellular location, with either theoretical or experimental support indicated.

Fig. 2. EchoLOCATION display page for the CysZ protein

The topology viewer is a useful tool to allow the quick visualization of the predicted topology of a membrane protein and to see how well TMHMM and HMMTOP algorithms agree. In addition, advanced features are accessible via the EchoLOCATION mainpage. This index page includes a list showing the number of proteins in each subcellular location (both predicted and experimentally confirmed) and a diagrammatic representation of the subcellular locations. From both the list and the diagram it is possible to generate a list of proteins in each specific location. A BLAST interface allows the identification of the subcellular locations of E. coli K-12 homologues of the query sequence, enabling users to transfer location predictions to other Gram-negative bacteria. Finally, the advanced search uses criteria such as number of TMHs, % of the protein predicted to be located within the membrane and molecular weight, to identify proteins which match certain general and location related criteria.

EXPERIMENTAL DATA COLLECTION

The search for experimentally determined subcellular locations of proteins were limited to those for proteins in a non “default” location; that is proteins which would need to contain specific targeting signals, e.g. signal peptides, and so would result in a periplasmic or membrane localization. A total of 1165 proteins were investigated from those identified within the target subcellular locations and experimentally determined subcellular locations for over 500 of these expressed proteins were extracted from literature. The literature used experiments ranging from proteome wide experiments, such as a global study of lipoproteins (Gonnet et al., 2004) to experiments on individual proteins, such as the periplasmic nitrate reductase (Thomas et al., 1999b). The sources of experimental data were diverse and required manual curation and analysis. For many proteins, data from a number of different experiments were considered to produce the most accurate curated predictions. We were conservative in our interpretation of some data sources so that we did not make non-cytoplasmic assertions of subcellular location unless the evidence was convincing and usually supported by multiple lines of evidence. For example, some proteomics data from subcellular fractions were clearly contamind by highly abundant cytoplasmic proteins (Link et al., 1997). Due to the poor representation of integral membrane proteins in proteomics studies, we curated a paper that reported results of a genome-wide study to predict the topology of integral membrane proteins (Daley et al., 2005). However, as membrane localization was not demonstrated directly in this study we included only the 100 proteins with > 3 predicted TMHs that produced PhoA positive signals.

ACCURACY OF ECHOLOCATION

The accuracy of EchoLOCATION was determined by comparing the results of the manually curated algorithms outputs with the experimentally determined locations. Of the > 500 experimentally confirmed locations over 93% agreed with the locations assigned by EchoLOCATION. The remaining 7% of the experimentally confirmed locations were in conflict with the predicted locations, these consist of 38 proteins summarized in Table 1. The level of accuracy of EchoLOCATION is remarkably high, although one must consider that much of the experimental data used in EchoLOCATION to support subcellular locations will have been used in the training sets for many of the algorithms used as primary source data for EchoLOCATION. Given this relatively high level of correct predictions, we considered proteins where the EchoLOCATION predictions disagreed with the experimental literature were worth further examination to establish the reasons for the mispredictions and to discover if any obvious patterns emerge from this analysis.
The 38 proteins with mispredicted subcellular locations fall into 8 different types, which raise a variety of issues about the accuracy of the algorithms used, the experimental methods used and our definitions of different subcellular compartment used in EchoLOCATION. There are a number of proteins that are located in macromolecular structures outside of the outer membrane that are components of the flagellum (FlhCD, FlgKLED) or type I pili (FimAFGHI) and are targeted to these locations by specialised biogenesis pathways (Capitani et al., 2006; Macnab, 2003) that are not recognised in any of the subcellular prediction algorithms used in this study. Although these proteins sit outside of the outer membrane, they are not truly ‘extracellular’ in that they do not diffuse away from the cell, and so we have defined the sublocation of ‘cell surface appendage’. Consequently the only truly ‘secreted’ proteins from E. coli K-12 in EchoBASE are YebF (Zhang et al., 2006) and HlyE (Ralph et al., 1998) and the later is cryptic in K-12 strains.

The largest group of mispredictions involved proteins that have been experimentally determined to be associated with the inner membrane, but which lack recognizable TMHs. These fall into a number of categories, which highlight some important areas where in silico methods could potentially be improved. The first class of proteins in this group are those that can be isolated from membrane fractions as they form stable complexes with inner membrane proteins, for example YedY which interacts with YedZ (Loschi et al., 2004), and hence is seen in inner membrane proteomes. There will be many additional examples of this type of membrane association from cytoplasmic proteins that also form tight interactions with membrane proteins, like the ATP-binding cassette components of ABC transporters, and these interactions are probably impossible to predict by any global in silico method. The second type of proteins in this group include protein like GlpC, that are found strongly associated with the inner membrane but lack any clear transmembrane helices and probably contain amphipathic helices that function to anchor them to the membrane. Similarly the DacABCD proteins have been shown experimentally to associate with the inner membrane by a C-terminal amphipathic helix (Harris et al., 2002) after export to the periplasm. Other proteins like Iap and CysQ, which have no obvious binding partner but are isolated from inner membrane fractions may also use amphipathic helices for association and despite recent attempts to write algorithms to predict these amphipathic helices (Sapay et al., 2006), there is still a general lack of experimental knowledge about proteins of this type which is important precursor for building robust training sets for new in silico approaches.

There are some examples of lipoproteins where the in silico methods could not resolve between inner and outer membrane locations and which have been resolved experimentally. MetQ, YehR, YecK have been found in the inner membrane (Merlin et al., 2002; Juncker et al., 2003), while RlpB has been detected in the outer membrane (Wu et al., 2006). The other examples of mispredictions are due to generally very accurate and successful current algorithms failing to correctly place proteins that fall near their boundaries/cut-off values. For example, the DegQ protein was predicted to be an inner membrane protein on the basis that HMMTOP predicted two transmembrane helices in

Table 1: List of proteins with experimentally determined subcellular location different to that predicted by EchoLOCATION. Fully referenced comments are available through the EchoLOCATION WWW site.
the middle of the protein, although TMHMM only predicted a single TMH which is at the N-terminus of the protein and which corresponds to its signal peptide. However, experimental evidence suggests that DegQ is a periplasmic protein and here HMMTOP has overpredicted TMHs. We examined this for the 777 integral membrane proteins predicted in EchoLOCATION and found there were 217 proteins (28%) where HMMTOP predicted more TMHs than TMHMM, while only 51 (7%) when TMHMM predicted more helices than HMMTOP.

We also observed signal peptide predictions for Sec-dependent substrates where proteins with scores just over the borderline have been experimentally confirmed to be cytoplasmic and in other cases the signal peptides are not cleaved and function as N-terminal membrane anchors, as observed in Gram positive bacteria (Zhou et al., 2008). We saw a similar pattern with predictions of lipoproteins, especially with predictions for outer membrane lipoproteins. The SrlD and PanE proteins have both been demonstrated experimentally to be cytoplasmic and their scores in LipOP are lowest and third lowest, respectively, of all the 99 proteins initially assigned to the outer membrane lipoproteins (Supplementary Fig. 1). By inference, the protein with the second lowest score, YdeK, is very likely to also be a false positive prediction and is most probably a cytoplasmic protein, which is a similar conclusion to that reached by Gommet al., (2004) in their focused study of E. coli lipoproteins. A final consideration is that some proteins have different subcellular locations depending on different physiological conditions or due to temporal or cell cycle-dependent changes. For example, the GlnK protein is cytoplasmic, but during ammonium shock becomes membrane associated by forming a complex with the ammonium transporter AmtB (Coutts et al., 2002). In these cases, we have described all the locations in the comments field and the primary location in the subcellular location field.

In conclusion, our analysis has demonstrated that current in silico algorithms for prediction of the subcellular location of E. coli K-12 proteins are generally very accurate, with only around 8% of proteins with experimentally determined locations differing from the in silico predictions. To improve the overall accuracy will require development of specialized prediction tools for detecting relatively rare features such as proteins targeted as flagella components, protein with non-cleaved N-terminal membrane anchors and proteins with amphipathic helices that allow them to associate with the membrane. However, it is also still clear that manual annotation of subcellular location by using a combination of experimental and in silico data gives the most accurate predictions for this model bacterium.

ACKNOWLEDGMENTS

We thank Louise Fairweather for some initial data collection.

REFERENCE LIST

We thank Louise Fairweather for some initial data collection.