for reagents and S. Arriola for technical assistance. We thank W. Lim, H. Luecke, D. Morgan, B. Shoichet, H. Madhani, and E. O'Shea for advice on the manuscript and members of the Taunton and Shokat laboratories for many helpful discussions. Molecular interaction data have been deposited in the Biomolecular Interaction Network Database (BIND) with accession code 216037.

Supporting Online Material
www.sciencemag.org/cgi/content/full/308/5726/1318/DC1
Materials and Methods
Figs. S1 to S3
References and Notes

6 December 2004; accepted 23 February 2005
10.1126/science1108367

Global Topology Analysis of the *Escherichia coli* Inner Membrane Proteome

Daniel O. Daley,[1]* Mikaela Rapp,[1]* Erik Granseth,[2] Karin Melén,[2] David Drew,[1] Gunnar von Heijne[1,2]†

The protein complement of cellular membranes is notoriously resistant to standard proteomic analysis and structural studies. As a result, membrane proteomes remain ill-defined. Here, we report a global topology analysis of the *Escherichia coli* inner membrane proteome. Using C-terminal tagging with the alkaline phosphatase and green fluorescent protein, we established the periplasmic or cytoplasmic locations of the C termini for 601 inner membrane proteins. By constraining a topology prediction algorithm with this data, we derived high-quality topology models for the 601 proteins, providing a firm foundation for future functional studies of this and other membrane proteomes. We also estimated the overexpression potential for 397 green fluorescent protein fusions; the results suggest that a large fraction of all inner membrane proteins can be produced in sufficient quantities for biochemical and structural work.

Integral membrane proteins account for the coding capacity of 20 to 30% of the genes in typical organisms (*1*) and are critically important for many cellular functions. However, owing to their hydrophobic and amphiphilic nature, membrane proteins are difficult to study, and they account for less than 1% of the known high-resolution protein structures (*2*). Overexpression, purification, biochemical analysis, and structure determination are all far more challenging than for soluble proteins, and membrane proteins have rarely been considered in proteomics or structural genomics contexts to date.

In the absence of a high-resolution three-dimensional structure, an important cornerstone for the functional analysis of any membrane protein is an accurate topology model. A topology model describes the number of transmembrane spans and the orientation of the protein relative to the lipid bilayer. Topology models are usually produced by either sequence-based prediction or time-consuming experimental approaches. We have previously shown that topology prediction can be greatly improved by constraining it with an experimentally determined reference point, such as the location of a protein's C terminus (*3*). For *E. coli* proteins, reference points can be obtained most easily through the use of topology reporter proteins such as alkaline phosphatase (PhoA) and green fluorescent protein (GFP). PhoA and GFP have opposite activity profiles: PhoA is active only in the periplasm of *E. coli* (*4*), whereas GFP is fluorescent only in the cytoplasm (*5*). When fused in parallel to the C terminus of a membrane protein, PhoA and GFP can accurately report on which side of the membrane the C terminus is located (*6*, *7*). Here, we have applied the PhoA/GFP fusion approach to derive topology models for almost the entire *E. coli* inner membrane proteome.

Bioinformatic analysis of the *E. coli* proteome using the hidden Markov model topology predictor TMHMM (*1*) indicates that approximately 1000 of the 4288 predicted genes encode integral inner membrane proteins. We focused on the 737 genes that encode proteins longer than 100 residues with at least two predicted transmembrane helices. The second criterion was necessary to ensure that secreted proteins, whose hydrophobic signal sequence is often mistakenly predicted as a transmembrane helix, were not included.

Of the 737 selected genes, 714 were suitable for cloning into a standard set of *phoA* and *gfp* fusion vectors (*8*). We were able to obtain both fusions for 573 genes and one fusion for an additional 92 genes (Fig. 1, inset). By determining appropriate cutoffs (*8*), the C-terminal location (C_{in}, C_{out}) could be assigned for 502 of the 665 cloned proteins by comparison of whole-cell GFP fluorescence and PhoA activity or, in a small number of cases, by either activity alone (Fig. 1).

To assign the location of the C terminus for the remaining proteins, we used the basic local alignment search tool (BLAST) (*9*) to search for homologs to the unassigned proteins among the 502 assigned proteins, imposing a strict E-value cutoff (10^{-4}) and the requirement that the BLAST-alignment should extend to within 25 residues of the C terminus of both proteins. We were able to assign C-terminal locations for an additional 99 proteins in this way, bringing the total number of assignments to 601 of the 737 proteins in the initial data set (table S1). Obviously, the same homology-based assignment scheme can be used to transfer the experimental data to other membrane proteomes.

The location of the C terminus for 71 of the 601 proteins was already known from published topology models (table S1) and was used to check the quality of our data. For all but two proteins, ArsB and YccA, our C-terminal assignment agreed with the published assignment. In the case of ArsB, the previous study (*10*) did not include experimental information on the location of the C terminus, and we suggest that our assignment is correct. For YccA, the reported experimental data on the location of the C terminus (*11*) contradicts our result; further studies will be required to resolve this discrepancy. In any case, it appears that

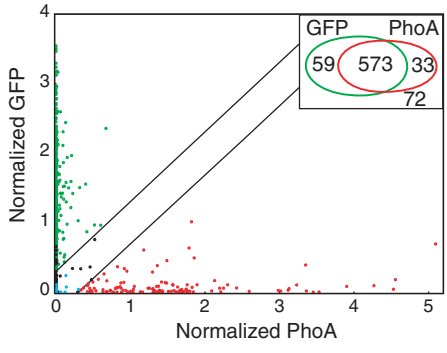

Fig. 1. Normalized PhoA and GFP activities. Cutoff lines for the assignment of C_{in} (cytoplasmic) and C_{out} (periplasmic) orientations are shown in black. Green and red dots: proteins assigned as C_{in} and C_{out}, respectively, based on the experimental data. Black and blue dots: proteins assigned as C_{in} and C_{out}, respectively, based on sequence homology to proteins with experimentally assigned C-terminal locations. (**Inset**) Venn diagram showing the number of proteins for which none, one, or both PhoA (red) and GFP (green) fusions were obtained.

[1]Department of Biochemistry and Biophysics, Stockholm University, SE-106 91 Stockholm, Sweden. [2]Stockholm Bioinformatics Center, AlbaNova, SE-106 91 Stockholm, Sweden.

*These authors contributed equally to this work.
†To whom correspondence should be addressed. E-mail: gunnar@dbb.su.se

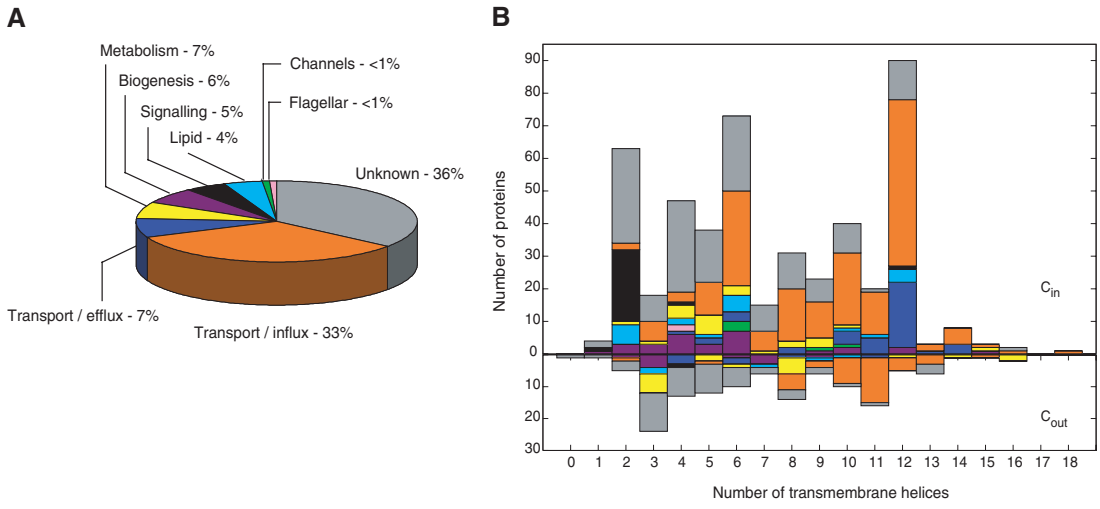

Fig. 2. Functional categorization of the *E. coli* inner membrane proteome. (**A**) The fractions of the inner membrane proteome (737 proteins) assigned to different functional categories. (**B**) The number of proteins with assigned C-terminal location in each functional category for different topologies (601 proteins in total). C_{in} topologies are plotted upward, C_{out} downward. For C_{in} proteins, even numbers of transmembrane helices are three times as common as odd numbers; for C_{out} proteins, odd and even numbers of transmembrane helices are roughly equal.

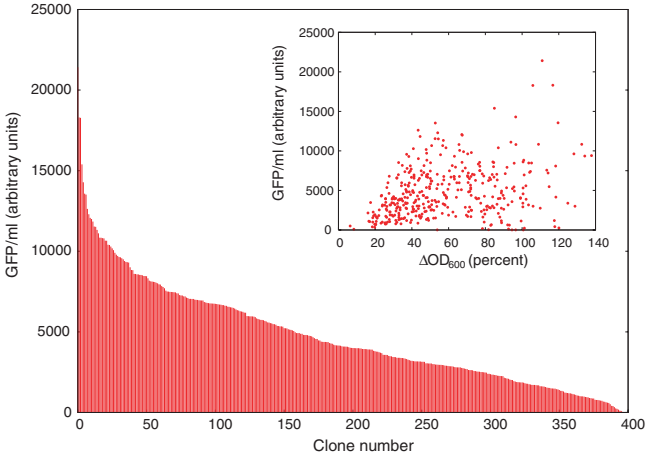

Fig. 3. GFP fluorescence for 397 C_{in} proteins. (**Inset**) Scatter plot with GFP activity plotted against the change in OD_{600} seen 2 hours after induction of protein expression compared with nontransformed cells grown under the same conditions [$100 \times \Delta OD_{600}$ (transformed cells) / ΔOD_{600} (nontransformed cells)].

the error rate in our C-terminal assignments is on the order of 1% or less.

Using the experimentally determined C-terminal locations as constraints for the TMHMM topology predictor (*3*), we generated experimentally based topology models for the 601 proteins, including 46 models from our previously published work (*6*, *7*) (table S1 and www.sbc.su.se/~erikgr/tmhmm/index.html).

In the absence of experimental data, TMHMM alone predicts the correct C-terminal location for only 78% of the 601 proteins. By providing unambiguous C-terminal locations, the inclusion of experimental data thus leads to a major improvement in the overall quality of the topology models (illustrated in fig. S1). This is also reflected in the TMHMM reliability score (*3*); the score increases for 526 proteins and decreases for 75 proteins upon fixing of the C terminus.

To obtain a global view of the topologies within the proteome and how they relate to protein function, proteins were sorted according to known or predicted functional categories (Fig. 2). The most obvious trend is the predominance of N_{in}-C_{in} topologies (57% of all proteins), which suggests that pairs of closely spaced transmembrane helices ("helical hairpins") may be a basic building block in membrane proteins. The largest functional category is transport proteins, many with 6 or 12 transmembrane helices. Most proteins with unknown function have ≤6 transmembrane helices, pointing to a systematic lack of studies of the smaller inner membrane proteins.

We have previously identified a case in which a gene duplication event has led to the formation of two separately expressed homologous proteins (YdgQ and YdgL) with opposite orientations in the membrane (*12*). To identify new instances of this kind, we searched for families of homologs that include pairs of proteins with the same number of predicted transmembrane helices but oppositely assigned C-terminal locations. Only the YdgE and YdgF proteins, both members of the small multidrug resistance (SMR) family of transporters (*13*), were found (fig. S2). The *ydgE* and *ydgF* genes overlap each other on the *E. coli* chromosome, and the two proteins catalyze drug efflux only when coexpressed (*14*), which suggests that they form an antiparallel heterodimer (or higher oligomer) in the inner membrane.

EmrE, another member of the SMR family, has been suggested to adopt a dual topology in the inner membrane, where one N_{in}-C_{in} molecule forms an antiparallel homodimer with one N_{out}-C_{out} molecule (*15*–*17*). EmrE is assigned as C_{out} in our data set, but also has GFP fluorescence above background. Notably, EmrE contains very few positively charged residues, evenly distributed between the different loops, and thus lacks a clear "positive-inside" bias (*18*). We searched for additional candidate dual topology proteins with a weak charge bias and above-background PhoA and GFP activities. Five additional proteins emerged as possible dual topology proteins: SugE (a member of the SMR family), CrcB, YdgC, YnfA, and YbfB (fig. S2). Strikingly, all these proteins are small (around 100 residues) and have three or four strongly predicted transmembrane helices.

Although proteins with internal duplications in which the two halves of the protein have opposite orientations in the membrane are quite common among the *E. coli* inner membrane proteins (*19*–*23*), we conclude that homologs with opposite membrane orientations, as well as proteins with a dual topology, are exceedingly rare. Possibly, protein folding and assembly are more efficient when the two oppositely oriented halves are part of a single polypeptide chain than when they are expressed separately.

For the ~80% of the inner membrane proteome with a C_{in} orientation, the whole-cell GFP fluorescence provides a good estimate of the amount of fusion protein inserted into the membrane (*24*). Using standard overexpression conditions (*8*), we tabulated the GFP activity for the 397 proteins assigned as C_{in} (Fig. 3). We also assessed the effect of overexpression on cell growth, as determined by the change in optical density of the cell suspension after induction of membrane pro-

tein synthesis. Although a small number of proteins appeared toxic (Fig. 3, inset), the vast majority had only a limited effect on cell growth. Overexpression levels do not correlate with—and hence cannot be predicted by—obvious sequence characteristics such as codon usage, protein size, hydrophobicity, and number of transmembrane helices (table S2). The C-terminal His_8 tag and the tobacco etch virus (TEV) protease site present in the GFP fusions (*8*) make it possible to use an efficient, standardized purification protocol for the whole clone collection; yields of purified fusion protein are typically ≥1 mg per liter of culture (*25*). This sets a lower limit for what can be expected for individual proteins expressed, for example, without a GFP tag or using other expression vectors and growth conditions (*26*).

In conclusion, by analyzing a library of *E. coli* inner membrane proteins fused to PhoA and GFP, we have derived an experimentally based set of topology models for the membrane proteome and provide a large-scale data set on membrane protein overexpression. Our results provide an important basis for future functional studies of membrane proteomes and will facilitate the identification of well-expressed targets for structural genomics projects.

References and Notes

1. A. Krogh, B. Larsson, G. von Heijne, E. Sonnhammer, *J. Mol. Biol.* **305**, 567 (2001).
2. S. H. White, *Protein Sci.* **13**, 1948 (2004).
3. K. Melén, A. Krogh, G. von Heijne, *J. Mol. Biol.* **327**, 735 (2003).
4. C. Manoil, J. Beckwith, *Science* **233**, 1403 (1986).
5. B. J. Feilmeier, G. Iseminger, D. Schroeder, H. Webber, G. J. Phillips, *J. Bacteriol.* **182**, 4068 (2000).
6. D. Drew *et al.*, *Proc. Natl. Acad. Sci. U.S.A.* **99**, 2690 (2002).
7. M. Rapp *et al.*, *Protein Sci* **13**, 937 (2004).
8. Materials and methods are available as supporting material on *Science* Online.
9. S. F. Altschul, W. Gish, W. Miller, E. W. Myers, D. J. Lipman, *J. Mol. Biol.* **215**, 403 (1990).
10. J. Wu, L. S. Tisa, B. P. Rosen, *J. Biol. Chem.* **267**, 12570 (1992).
11. A. Kihara, Y. Akiyama, K. Ito, *EMBO J.* **18**, 2970 (1999).
12. A. Sääf, M. Johansson, E. Wallin, G. von Heijne, *Proc. Natl. Acad. Sci. U.S.A.* **96**, 8540 (1999).
13. I. T. Paulsen *et al.*, *Mol. Microbiol.* **19**, 1167 (1996).
14. K. Nishino, A. Yamaguchi, *J. Bacteriol.* **183**, 5803 (2001).
15. I. Ubarretxena-Belandia, C. G. Tate, *FEBS Lett.* **564**, 234 (2004).
16. I. Ubarretxena-Belandia, J. M. Baldwin, S. Schuldiner, C. G. Tate, *EMBO J.* **22**, 6175 (2003).
17. C. Ma, G. Chang, *Proc. Natl. Acad. Sci. U.S.A.* **101**, 2852 (2004).
18. G. von Heijne, *EMBO J.* **5**, 3021 (1986).
19. B. van den Berg *et al.*, *Nature* **427**, 36 (2004).
20. D. Fu *et al.*, *Science* **290**, 481 (2000).
21. R. Dutzler, E. B. Campbell, M. Cadene, B. T. Chait, R. MacKinnon, *Nature* **415**, 287 (2002).
22. S. Khademi *et al.*, *Science* **305**, 1587 (2004).
23. T. Shimizu, H. Mitsuke, K. Noto, M. Arai, *J. Mol. Biol.* **339**, 1 (2004).
24. D. Drew, G. von Heijne, P. Nordlund, J. W. L. de Gier, *FEBS Lett.* **507**, 220 (2001).
25. D. Drew *et al.*, *Protein Sci*, in press.
26. S. Eshaghi *et al.*, *Protein Sci.* **14**, 676 (2005).

27. Supported by grants from the Swedish Research Council, the Marianne and Marcus Wallenberg Foundation, the Swedish Foundation for Strategic Research, and the Swedish Cancer Foundation to G.v.H., by the Swedish Knowledge Foundation to K.M., and by a European Molecular Biology Organization Long-Term Fellowship to D.O.D.

Supporting Online Material

www.sciencemag.org/cgi/content/full/308/5726/1321/DC1
Materials and Methods
Figs. S1 and S2
Tables S1 and S2
References

13 January 2005; accepted 14 March 2005
10.1126/science.1109730

Firearm Violence Exposure and Serious Violent Behavior

Jeffrey B. Bingenheimer,[1]* Robert T. Brennan,[2] Felton J. Earls[2]

To estimate the cause-effect relationship between exposure to firearm violence and subsequent perpetration of serious violence, we applied the analytic method of propensity stratification to longitudinal data on adolescents residing in Chicago, Illinois. Results indicate that exposure to firearm violence approximately doubles the probability that an adolescent will perpetrate serious violence over the subsequent 2 years.

Within the past few decades, the popular notion that violence begets violence has come under scientific scrutiny. Early research by psychologists, criminologists, and others focused on the impact of being physically abused as a child on subsequent delinquency, community violence, and spouse and child abuse. Simple comparisons of violent offenders and nonoffenders showed that the former were more likely to report having been abused during childhood (*1*, *2*). More carefully controlled prospective studies comparing abused and nonabused children confirmed these basic relationships (*3*) and provided insights into the cognitive and neurological mechanisms involved (*4*, *5*).

Recently, interest has expanded to encompass exposure to violence occurring in community settings such as neighborhoods and schools. This change was spurred in part by elevated rates of violent crime, including firearm homicide, in American cities in the early 1990s (*6*, *7*). In several studies conducted around that time, urban children and adolescents reported alarmingly high levels of exposure to community violence, both as witnesses and as victims (*8–10*). These findings raised troubling questions about the possible developmental ramifications of such widespread experience with violence.

Numerous recent investigations have revealed statistical associations between children's and adolescents' self-reports of exposure to community violence and concurrent or subsequent assessments of violence and aggression (*11–14*). Available estimates of these associations, however, do not adequately control for the possibility that a common set of personal characteristics and environment circumstances may jointly influence who is exposed to community violence and who becomes a perpetrator of violent acts. The extent to which these statistical associations are attributable to cause-effect relationships therefore remains uncertain (*15*).

The randomized experiment is the scientific gold standard for causal inference, but in the instance of community violence is neither technically nor ethically feasible. We used the method of propensity score stratification (*16–18*) to approximate a randomized experiment in which exposure to firearm violence was the treatment variable and subsequent perpetration of serious violence was the outcome. This method is based upon counterfactual thinking and the framework of potential outcomes described by Rubin (*19*) and others (*20*). Investigators in economics (*21*), medicine (*22*), and other fields (*23*) are increasingly using propensity score matching and stratification to improve the credibility of estimates of cause-effect relationships obtained from observational data.

Propensity stratification views exposure allocation as a process involving both systematic and random components. First, personal and environmental characteristics of the individual determine systematically her or his probability π of exposure, called the propensity score. The individual then participates in a lottery in which the exposure is assigned with probability π, or nonexposure is assigned with probability $1 - \pi$. In theory, comparing individuals with identical propensity scores but different realized exposures is analogous to conducting a randomized experiment, and therefore provides a valid basis for measuring a cause-effect relationship between expo-

[1]Department of Health Behavior and Health Education, 1420 Washington Heights, University of Michigan School of Public Health, Ann Arbor, MI 48109–2029, USA. [2]Department of Social Medicine, Harvard Medical School, 1430 Massachusetts Avenue, 4th Floor, Cambridge, MA 02138, USA.

*To whom correspondence should be addressed. E-mail: bartbing@umich.edu