Structural Bioinformatics-Based Design of Selective, Irreversible Kinase Inhibitors

Michael S. Cohen, *et al.*

Science **308**, 1318 (2005);
DOI: 10.1126/science1108367

Science (print ISSN 0036-8075; online ISSN 1095-9203) is published weekly, except the last week in December, by the American Association for the Advancement of Science, 1200 New York Avenue NW, Washington, DC 20005. Copyright 2005 by the American Association for the Advancement of Science; all rights reserved. The title *Science* is a registered trademark of AAAS.

References and Notes

1. G. W. Kreutzberg, *Trends Neurosci.* **19**, 312 (1996).
2. G. Stoll, S. Jander, *Prog. Neurobiol.* **58**, 233 (1999).
3. W. J. Streit, *Glia* **40**, 133 (2002).
4. W. J. Streit, *J. Neurosci. Res.* **77**, 1 (2004).
5. C. Nolte, T. Moller, T. Walter, H. Kettenmann, *Neuroscience* **73**, 1091 (1996).
6. N. Stence, M. Waite, M. E. Dailey, *Glia* **33**, 256 (2001).
7. W. Denk, K. Svoboda, *Neuron* **18**, 351 (1997).
8. S. Jung *et al.*, *Mol. Cell. Biol.* **20**, 4106 (2000).
9. Materials and methods are available as supporting material on *Science* Online.
10. A. Lehmenkuhler, E. Sykova, J. Svoboda, K. Zilles, C. Nicholson, *Neuroscience* **55**, 339 (1993).
11. A. Nimmerjahn, F. Kirchhoff, J. N. D. Kerr, F. Helmchen, *Nat. Methods* **1**, 31 (2004).
12. D. van Rossum, U. K. Hanisch, *Metab. Brain Dis.* **19**, 393 (2004).
13. J. Grutzendler, N. Kasthuri, W. B. Gan, *Nature* **420**, 812 (2002).
14. J. T. Trachtenberg *et al.*, *Nature* **420**, 788 (2002).
15. J. Hirrlinger, S. Hulsmann, F. Kirchhoff, *Eur. J. Neurosci.* **20**, 2235 (2004).
16. G. Raivich *et al.*, *Brain Res. Rev.* **30**, 77 (1999).
17. F. Capani, M. H. Ellisman, M. E. Martone, *Brain Res.* **923**, 1 (2001).
18. F. Vilhardt, *Int. J. Biochem. Cell Biol.* **37**, 17 (2005).
19. We thank S. Erdogan for help with the analysis, J. N. D. Kerr and G. W. Kreutzberg for comments on the manuscript, S. Jung and D. R. Littman for providing the green fluorescent microglia mouse line, and B. Sakmann for generous support. This work was supported by a predoctoral fellowship of the Boehringer Ingelheim Fonds to A.N.

Supporting Online Material
www.sciencemag.org/cgi/content/full/1110647/DC1
Materials and Methods
SOM Text
Figs. S1 to S3
References and Notes
Movies S1 to S12

3 February 2005; accepted 17 March 2005
Published online 14 April 2005;
10.1126/science.1110647
Include this information when citing this paper.

Structural Bioinformatics-Based Design of Selective, Irreversible Kinase Inhibitors

Michael S. Cohen, Chao Zhang, Kevan M. Shokat, Jack Taunton*

The active sites of 491 human protein kinase domains are highly conserved, which makes the design of selective inhibitors a formidable challenge. We used a structural bioinformatics approach to identify two selectivity filters, a threonine and a cysteine, at defined positions in the active site of p90 ribosomal protein S6 kinase (RSK). A fluoromethylketone inhibitor, designed to exploit both selectivity filters, potently and selectively inactivated RSK1 and RSK2 in mammalian cells. Kinases with only one selectivity filter were resistant to the inhibitor, yet they became sensitized after genetic introduction of the second selectivity filter. Thus, two amino acids that distinguish RSK from other protein kinases are sufficient to confer inhibitor sensitivity.

Phosphorylation of serine, threonine, and tyrosine residues is a primary mechanism for regulating protein function in eukaryotic cells. Protein kinases, the enzymes that catalyze these reactions, regulate essentially all cellular processes and have thus emerged as therapeutic targets for many human diseases (*1*). Small-molecule inhibitors of the Abelson tyrosine kinase (Abl) and the epidermal growth factor receptor (EGFR) have been developed into clinically useful anticancer drugs (*2*, *3*). Selective inhibitors can also increase our understanding of the cellular and organismal roles of protein kinases. However, nearly all kinase inhibitors target the adenosine triphosphate (ATP) binding site, which is well conserved even among distantly related kinase domains. For this reason, rational design of inhibitors that selectively target even a subset of the 491 related human kinase domains continues to be a daunting challenge.

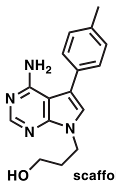
Structural and mutagenesis studies have revealed key determinants of kinase inhibitor selectivity, including a widely exploited selectivity filter in the ATP binding site known as the "gatekeeper." A compact gatekeeper (such as threonine) allows bulky aromatic substituents, such as those found in the Src family kinase inhibitors, PP1 and PP2, to enter a deep hydrophobic pocket (*4*–*6*). In contrast, larger gatekeepers (methionine, leucine, isoleucine, or phenylalanine) restrict access to this pocket. A small gatekeeper provides only partial discrimination between kinase active sites, however, as ~20% of human kinases have a threonine at this position. Gleevec, a drug used to treat chronic myelogenous leukemia, exploits a threonine gatekeeper in the Abl kinase domain, yet it also potently inhibits the distantly related tyrosine kinase, c-KIT, as well as the platelet-derived growth factor receptor (PDGFR) (*7*).

We therefore sought a second selectivity filter that could be discerned from a primary sequence alignment. Among the 20 amino acids, cysteine has unique chemical reactivity and is commonly targeted by electrophilic inhibitors. In the case of cysteine protease inhibitors, the reactive cysteine is not a selectivity filter, because it is found in every cysteine protease and is essential for catalysis. Electrophilic, cysteine-directed inhibitors of the EGFR kinase domain have also been reported (*8*), but here again, the cysteine does not act as a selectivity filter, because neither the electrophile nor the reactive cysteine is required for potent, selective inhibition by these compounds. In this report, we describe the rational design of selective kinase inhibitors that require the simultaneous presence of a threonine gatekeeper and a reactive cysteine, which are uniquely found in the C-terminal kinase domain of p90 ribosomal protein S6 kinases (RSKs).

We used a kinomewide sequence alignment (*1*, *9*) to search for cysteines that, together with a threonine gatekeeper, could form a covalent bond with an inhibitor in the ATP pocket. We focused on the conserved glycine-rich loop, which interacts with the triphosphate of ATP and is one of the most flexible structural elements of the kinase domain (*10*). A cysteine near this solvent-exposed loop is likely to have a lower pK_a and therefore to be more reactive than a cysteine buried in the hydrophobic pocket. Out of 491 related kinase domains in the human genome (*1*), we found 11 with a cysteine at the C-terminal end of the glycine-rich loop (Fig. 1A), a position usually occupied by valine. We next examined the gatekeeper in these

Table 1. Half-maximal inhibitory concentrations (IC$_{50}$ in μM) for fmk and the pyrrolo[2,3-*d*]pyrimidine scaffold against the kinase activities of wild-type (WT) and mutant RSK2 CTDs. RSK2 CTDs were expressed in *E. coli* as His$_6$-tagged proteins and activated by incubation with bacterially expressed His$_6$-ERK2 and ATP. Kinase assay conditions: 30-min inhibitor pretreatment, 1 nM RSK2 CTD, 0.1 mM ATP, 0.1 mM "CTD-tide" substrate (*14*). WT and mutant CTDs had similar kinase activities.

	WT	C436V	T493M
fmk	0.015 ± 0.001	>10	3.4 ± 0.3
scaffold	1.2 ± 0.08	0.43 ± 0.14	>30



scaffold

Program in Chemistry and Chemical Biology, and Department of Cellular and Molecular Pharmacology, University of California, San Francisco, CA 94143–2280, USA.

*To whom correspondence should be addressed. E-mail: taunton@cmp.ucsf.edu

kinases. Three closely related paralogs, RSK1, RSK2, and RSK4, have a threonine gatekeeper, whereas the remaining nine kinases, including RSK3, have larger gatekeepers (Fig. 1A). RSK1 and RSK2 are downstream effectors of the Ras-mitogen–activated protein kinase (MAPK) pathway and are directly activated by the MAPKs, ERK1 and ERK2 (*11*, *12*). Mutations in the RSK2 gene cause Coffin-Lowry syndrome, a human disorder characterized by severe mental retardation (*13*). However, the precise roles of RSKs are poorly understood. All RSKs have two kinase domains. The regulatory C-terminal kinase domain (CTD) has the cysteine and threonine selectivity filters.

To exploit both selectivity filters in RSK family kinases, we needed a scaffold that could present an electrophile to the cysteine while occupying the hydrophobic pocket defined by the gatekeeper. Crystal structures of kinases with bound ATP analogs all reveal van der Waals contacts between a conserved valine, analogous to the cysteine we identified in RSKs (Fig. 1A), and the adenine C-8 position. We therefore designed and synthesized cmk and fmk (Fig. 1B), pyrrolopyrimidines that contained a chloromethylketone and a fluoromethylketone, respectively. The *p*-tolyl substituent was designed to occupy the putative hydrophobic pocket. Structurally related pyrazolopyrimidines interact similarly with Src family kinases (*4*–*6*). We hypothesized that the electrophilic halomethylketones would be within striking distance of the key cysteine in RSK1, RSK2, and RSK4 and that kinases with only one of the two selectivity filters would be resistant to the inhibitors.

We first tested the electrophilic pyrrolopyrimidines against the RSK2 CTD in vitro (*14*). Both fmk (Table 1) and cmk (*15*) inhibited RSK2 CTD activity with similar potencies, but we focused on fmk because of its greater chemical stability. To test whether both selectivity filters were required for fmk sensitivity, we expressed two CTD mutants, C436V, in which Cys^{436} was replaced with Val, and T493M, in which Thr^{493} was replaced with Met. Fmk was a potent and selective inhibitor of wild-type (WT) RSK2 [half-maximal inhibitory concentration (IC_{50}) = 15 nM], with greater than 600- and 200-fold selectivity over the C436V and T493M mutants, respectively (Table 1). To test whether fmk forms an irreversible covalent bond with RSK2, we prepared a biotinylated derivative (see Supporting Online Material). Biotin-fmk reacted irreversibly with WT RSK2, but not with the selectivity filter mutants, as shown by denaturing gel electrophoresis and Western blot analysis with streptavidin–horseradish peroxidase (HRP) (Fig. 2A). ERK2, required to activate RSK2 in vitro, was not labeled by biotin-fmk, despite the presence of a solvent-exposed cysteine in its ATP pocket (*16*).

We next tested the selectivity of biotin-fmk in a human epithelial cell lysate containing thousands of potentially reactive proteins.

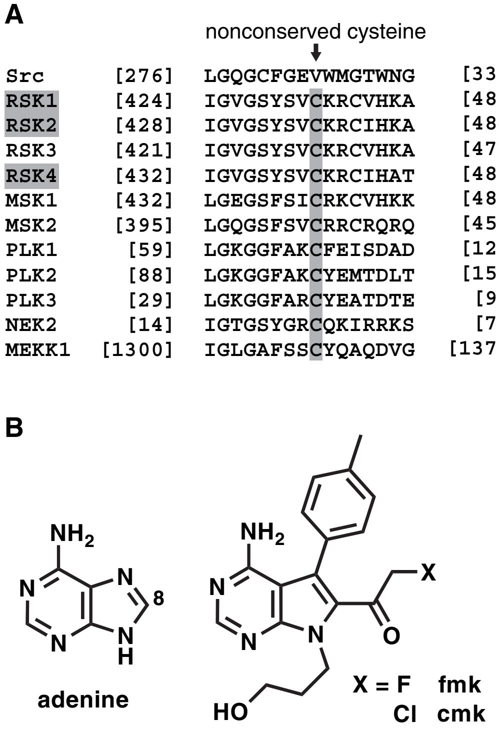

A
nonconserved cysteine
gatekeeper
Src [276] LGQGCFGEVWMGTWNG [333] SEEP---IYIVTEYMSKGSL
RSK1 [424] IGVGSYSVCKRCVHKA [480] DDGKH--VYLVTELMRGGEL
RSK2 [428] IGVGSYSVCKRCIHKA [484] DDGKY--VYVVTELMKGGEL
RSK3 [421] IGVGSYSVCKRCVHKA [477] DDGKF--VYLVMELMRGGEL
RSK4 [432] IGVGSYSVCKRCIHAT [488] DDGRY--VYLVTDLMKGGEL
MSK1 [432] LGEGSFSICRKCVHKK [489] HDQLH--TFLVMELLNGGEL
MSK2 [395] LGQGSFSVCRRCRQRQ [452] HDQLH--TYLVLELLRGGEL
PLK1 [59] LGKGGFAKCFEISDAD [121] EDNDF--VFVVLELCRRRSL
PLK2 [88] LGKGGFAKCYEMTDLT [150] EDKEN--IYILLEYCSRR-S
PLK3 [29] LGKGGFARCYEATDTE [91] EDADN--IYIFLELCSRK-S
NEK2 [14] IGTGSYGRCQKIRRKS [75] IDRTNTTLYIVMEYCEGGDL
MEKK1 [1300] IGLGAFSSCYQAQDVG [1371] CEKSN--YNLFIEWMAGGSV
B
NH2
N
N
N
N
H
8
adenine
NH2
N
N
N
O
X
HO
X = F fmk
Cl cmk

Fig. 1. Structural bioinformatics guides the design of electrophilic inhibitors of RSK family protein kinases. (**A**) Sequence alignment of the 11 human kinases with a cysteine selectivity filter at the C-terminal end of the glycine-rich loop. Of these 11, RSK1, RSK2, and RSK4 are the only kinases with a threonine selectivity filter in the gatekeeper position. Src, which has a threonine gatekeeper but lacks the cysteine, is shown for comparison. (**B**) Chemical structures of adenine and the rationally designed halomethylketone pyrrolopyrimidines, cmk and fmk.

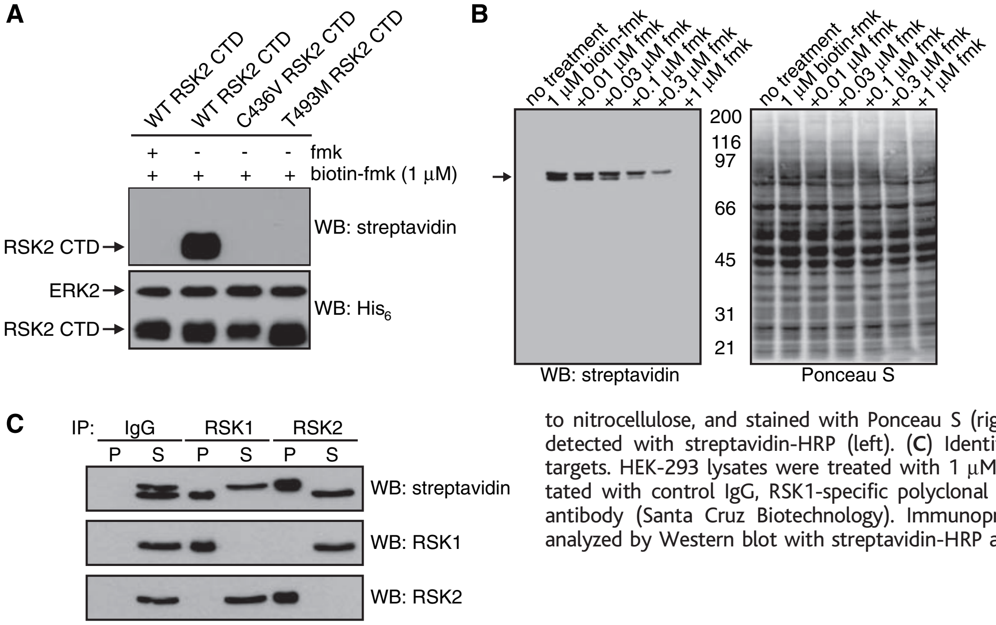


Fig. 2. Selective, irreversible targeting of RSK family kinases by fmk. (**A**) Covalent labeling of WT, but not mutant RSK2 by biotin-fmk. RSK2 with a hexahistidine tag (His_6-RSK2) CTDs were treated with 1 μM biotin-fmk in the presence of His_6-ERK2 for 1 hour. Proteins were resolved by SDS–polyacrylamide gel electrophoresis (SDS-PAGE) and detected by Western blot with streptavidin-HRP or antibodies to His_6. (**B**) Targeting of two ~90-kD proteins in human epithelial cell lysates by biotin-fmk. HEK-293 cell lysates were treated with the indicated concentrations of unlabeled fmk and then with 1 μM biotin-fmk. Proteins were resolved by SDS-PAGE, transferred to nitrocellulose, and stained with Ponceau S (right). Proteins labeled with biotin-fmk were detected with streptavidin-HRP (left). (**C**) Identification of RSK1 and RSK2 as biotin-fmk targets. HEK-293 lysates were treated with 1 μM biotin-fmk. Proteins were immunoprecipitated with control IgG, RSK1-specific polyclonal antibodies, or a RSK2-specific monoclonal antibody (Santa Cruz Biotechnology). Immunoprecipitates (P) and supernatants (S) were analyzed by Western blot with streptavidin-HRP and antibodies to RSK1 and RSK2.

Biotin-fmk (1 μM) reacted with only two proteins, and labeling was abolished by pretreatment with 1 μM fmk (Fig. 2B). These ~90-kD proteins were shown to be RSK1 and RSK2 by quantitative immunodepletion with specific antibodies (Fig. 2C). A cell-permeable, fluorescent derivative of fmk was also found to be highly selective toward RSK1 and RSK2 when added to cells growing in culture (*15*).

The only known substrate of the RSK2 CTD is Ser^{386} of RSK2 itself (*17*, *18*). Phosphorylation of Ser^{386} creates a docking site for phosphoinositide-dependent kinase 1 (PDK1), which then phosphorylates and activates the N-terminal kinase domain (NTD) (*19*). The activated NTD then phosphorylates downstream substrates. Treatment of serum-starved COS-7 cells with EGF induced Ser^{386} phosphorylation of endogenous RSK2, which was inhibited by fmk with a half-maximal effective concentration (EC_{50}) of ~200 nM (Fig. 3A). Thus, the CTD appears to be the primary kinase responsible for EGF-stimulated Ser^{386} phosphorylation, consistent with results obtained with kinase-inactive mutants (*17*, *18*). Fmk (10 μM) had no effect on EGF-stimulated phosphorylation of ERK1 or ERK2 (Fig. 3B), the MAP kinases directly upstream of RSK2. This result further highlights the selectivity of fmk, as the signaling pathway leading to ERK activation involves at least three protein kinases (EGFR, Raf, and MEK), two of which (EGFR and Raf) have threonine gatekeepers, as well as potentially reactive cysteines, in their ATP binding pockets.

We next tested whether inhibition of the CTD by fmk could block signaling downstream of RSK2. In cells transfected with WT RSK2, treatment with EGF induced phosphorylation of histone H3 at Ser^{10}, which was completely blocked by fmk (Fig. 3C). By contrast, H3 phosphorylation was unaffected by up to 10 μM fmk in cells expressing the RSK2 gatekeeper mutant, T493M. Mutation of the cysteine selectivity filter to valine (C436V) also conferred complete resistance to fmk (fig. S1). Similar to RSK family kinases, the mitogen- and stress-activated kinases, MSK1 and MSK2, have two kinase domains. The CTD of MSK1 has a cysteine analogous to Cys^{436} of RSK2 (Fig. 1A), but unlike RSK2, MSK1 has a methionine gatekeeper. Fmk had no effect on H3 phosphorylation mediated by WT MSK1 (Fig. 3D). By contrast, an MSK1 mutant with a threonine gatekeeper was potently inhibited. Thus, despite having only ~40% sequence identity to RSK2, MSK1 became equally sensitive to fmk once the second selectivity filter was introduced. Fmk had no effect on phorbol ester–stimulated H3 phosphorylation in nontransfected fibroblasts (fig. S2), consistent with a dominant role for endogenous MSK1 and MSK2 in this pathway (*20*).

Pyrrolopyrimidines inhibit Src family kinases such as Fyn (*21*), which raises the possibility that they might be susceptible to reversible inhibition by cmk or fmk. WT Fyn was weakly inhibited by both compounds, with IC_{50} values of ~4 μM (compared with an IC_{50} value of 15 nM for fmk against RSK2). By contrast, cmk and fmk potently inhibited a Fyn mutant in which Val^{285} was replaced with Cys (IC_{50} values of 1 nM and 100 nM, respectively). Similarly, at concentrations that completely inhibited RSK2 in cells, neither cmk nor fmk reversed the cellular phenotypes induced by v-Src (with a Thr in the gatekeeper position). By contrast, cmk (and less potently, fmk) promoted morphological reversion and reduced global tyrosine phosphorylation in cells expressing a v-Src mutant in which Val^{281} was replaced with Cys (fig. S3).

In this study, we have rationally designed halomethylketone-substituted inhibitors whose molecular recognition by protein kinases requires the simultaneous presence of two selectivity filters: a cysteine following the glycine-rich loop and a threonine in the gatekeeper position. We estimate that ~20% of human kinases have a solvent-exposed cysteine in the ATP pocket. Because of the structural conservation of the pocket, it should be possible to predict the orientation of these cysteines. In addition, there are many reversible kinase inhibitors whose binding modes have been characterized by x-ray crystallography. The integration of both types of information should allow the design of scaffolds that exploit selectivity filters other than the gatekeeper, as well as the appropriate sites for attaching electrophilic substituents.

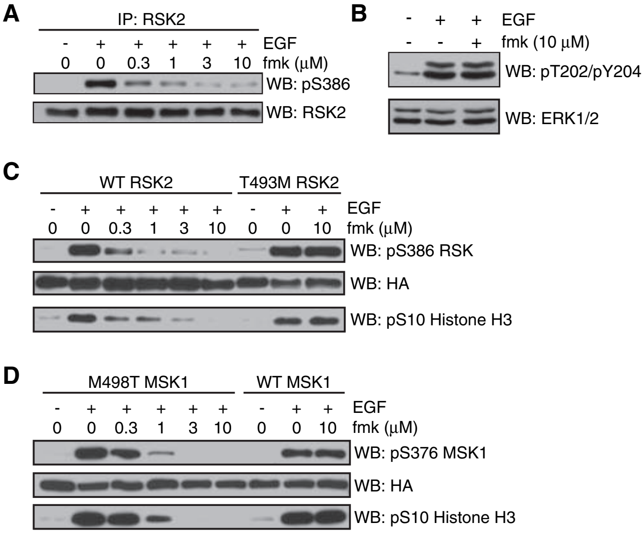


Fig. 3. Effect of fmk on EGF-activated RSK2 or MSK1. (**A**) Inhibition of EGF-stimulated RSK2 autophosphorylation by fmk. COS-7 cells were deprived of serum for 20 hours then treated with the indicated concentrations of fmk. Cells were stimulated for 10 min with EGF (1 ng/mL). RSK2 was immunoprecipitated and analyzed by Western blot with antibodies to phospho-Ser^{386} RSK (Cell Signaling Technology) and total RSK2. (**B**) Failure of fmk to inhibit EGF-stimulated activation of ERK1 and ERK2. Serum-starved COS-7 cells were treated with or without 10 μM fmk, then stimulated with EGF as in (A). Doubly phosphorylated and total ERK1 and ERK2 were detected by Western blot (both antibodies from Cell Signaling Technology). (**C**) Inhibition of EGF-stimulated H3 phosphorylation in cells expressing WT RSK2, but not T493M RSK2. COS-7 cells were transfected with hemagglutinin (HA)-tagged WT or T493M RSK2. Twenty-four hours after transfection, cells were serum-starved for 3 hours then treated with the indicated concentrations of fmk. Cells were stimulated with EGF (150 ng/mL) for 25 min and subsequently lysed in Laemmli sample buffer. Proteins were resolved by SDS-PAGE and detected by Western blot with antibodies specific for phospho-Ser^{386} RSK, the HA epitope (Roche), or phospho-Ser^{10} H3 (Upstate). (**D**) Inhibition of MSK1 by fmk after mutation of Met^{498} to Thr. COS-7 cells were transfected with HA-tagged WT or M498T MSK1. MSK1 autophosphorylation and H3 phosphorylation were assessed as in (C).

References and Notes

1. G. Manning, D. B. Whyte, R. Martinez, T. Hunter, S. Sudarsanam, *Science* **298**, 1912 (2002).
2. B. J. Druker *et al.*, *Nat. Med.* **2**, 561 (1996).
3. A. J. Barker *et al.*, *Bioorg. Med. Chem. Lett.* **11**, 1911 (2001).
4. Y. Liu *et al.*, *Chem. Biol.* **6**, 671 (1999).
5. T. Schindler *et al.*, *Mol. Cell* **3**, 639 (1999).
6. X. Zhu *et al.*, *Structure Fold. Des.* **7**, 651 (1999).
7. E. Buchdunger *et al.*, *J. Pharmacol. Exp. Ther.* **295**, 139 (2000).
8. D. W. Fry *et al.*, *Proc. Natl. Acad. Sci. U.S.A.* **95**, 12022 (1998).
9. O. Buzko, K. M. Shokat, *Bioinformatics* **18**, 1274 (2002).
10. I. Tsigelny *et al.*, *Biopolymers* **50**, 513 (1999).
11. T. W. Sturgill, L. B. Ray, E. Erikson, J. L. Maller, *Nature* **334**, 715 (1988).
12. M. Frodin, S. Gammeltoft, *Mol. Cell. Endocrinol.* **151**, 65 (1999).
13. A. Hanauer, I. D. Young, *J. Med. Genet.* **39**, 705 (2002).
14. C. A. Chrestensen, T. W. Sturgill, *J. Biol. Chem.* **277**, 27733 (2002).
15. M. S. Cohen, J. Taunton, unpublished data.
16. B. J. Canagarajah, A. Khokhlatchev, M. H. Cobb, E. J. Goldsmith, *Cell* **90**, 859 (1997).
17. T. A. Vik, J. W. Ryder, *Biochem. Biophys. Res. Commun.* **235**, 398 (1997).
18. K. N. Dalby, N. Morrice, F. B. Caudwell, J. Avruch, P. Cohen, *J. Biol. Chem.* **273**, 1496 (1998).
19. M. Frodin, C. J. Jensen, K. Merienne, S. Gammeltoft, *EMBO J.* **19**, 2924 (2000).
20. A. Soloaga *et al.*, *EMBO J.* **22**, 2788 (2003).
21. A. F. Burchat *et al.*, *Bioorg. Med. Chem. Lett.* **10**, 2171 (2000).
22. This work was supported by the Searle Scholars Foundation (J.T.), the NIH (K.M.S.), and the ARCS Foundation (M.S.C.). We thank Y. Feng, T. Alber, K. Shah, T. Sturgill, C. Bjorbaek, M. Frodin, and M. Cobb

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

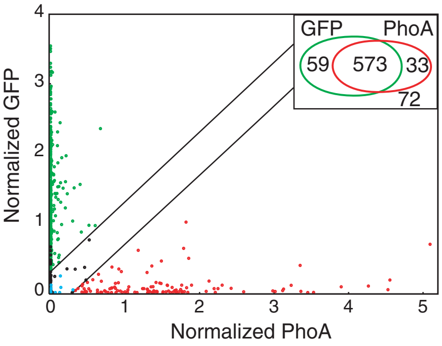


Fig. 1. Normalized PhoA and GFP activities. Cutoff lines for the assignment of C_{in} (cytoplasmic) and C_{out} (periplasmic) orientations are shown in black. Green and red dots: proteins assigned as C_{in} and C_{out}, respectively, based on the experimental data. Black and blue dots: proteins assigned 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