Large-scale identification of protein–protein interaction of *Escherichia coli* K-12

Mohammad Arifuzzaman,1,2,9,11 Maki Maeda,1,3,11 Aya Itoh,4 Kensaku Nishikata,5 Chiharu Takita,3 Rintaro Saito,4 Takeshi Ara,4 Kenji Nakahigashi,4 Hsuan-Cheng Huang,6 Aki Hirai,3 Kohei Tsuzuki,4 Seira Nakamura,4 Mohammad Altaf-Ul-Amin,5 Taku Oshima,1,5 Tomoya Baba,1,4 Natsuko Yamamoto,1,2 Tomoyo Kawamura,3 Tomoko Ioka-Nakamichi,3 Masanari Kitagawa,1,8 Masaru Tomita,4 Shigehiko Kanaya,5 Chieko Wada,7,10,12 and Hirotada Mori1,4,12

1Graduate School of Biological Sciences, Nara Institute of Science and Technology, Ikoma, Nara 630-0101, Japan; 2Kyowa Hakko Branch, Japan Bioindustry Association in Tokyo Research Laboratories, Kyowa Hakko Kogyo, Machida-shi, Tokyo 194-8533, Japan; 3CREST, JST (Japan Science and Technology), Kawaguchi, Saitama 332-0012, Japan; 4Institute for Advanced Biosciences, Keio University, Tsuruoka, Yamagata 997-0035, Japan; 5Graduate School of Information Science, Nara Institute of Science and Technology, Ikoma, Nara 630-0101, Japan; 6Institute of Bioinformatics, National Yang-Ming University, Taipei 112, Taiwan, China; 7Institute for Virus Research, Kyoto University, Sakyo, Kyoto 606-8507, Japan; 8Dragon Genomics Center, Takara Bio Inc. Yokkaichi, Mie 512-1211, Japan; 9Department of Biochemistry, BGC Trust Medical College, Chittagong, Bangladesh; 10Graduate School of Biostudies, Konoe, Kyoto University, Sakyo-ku, Kyoto 606-8315, Japan.

These authors contributed equally to this work.

E-mail mhori@gtc.naist.jp; fax. +81-743-72-5669.
E-mail cwada@lif.kyoto-u.ac.jp; fax. +81-75-753-7905.

[Supplemental material is available online at www.genome.org.]

*Escherichia coli* is one of the best characterized organisms and has served as a model system to study many aspects of bacterial physiology and genetics of fundamental and applied interest. Among the 4339 predicted ORFs including previous prediction (Riley et al. 2006) in *E. coli*, nearly 50% are experimentally uncharacterized. In addition to functional analysis of individual ORFs, systematic analyses of relationships between constituent elements, such as gene regulatory networks, protein–protein interactions (PPIs), and metabolic networks, have only recently become feasible.

To date, comprehensive PPI studies have been based on the yeast two-hybrid system that detects binary interactions through activation of reporter gene expression (Fields and Song 1989; Uetz et al. 2000; Ito et al. 2001), and pull-down assays that detect large complexes by copurification of prey proteins through their interactions with bait proteins (Gavin et al. 2002; Ho et al. 2002) or protein chips (Zhu et al. 2001). In *E. coli*, a large-scale protein interaction network was recently carried out by pull-down assay using TAP-tagged bait proteins (Butland et al. 2005). We have already described a comprehensive *E. coli* ORF library (the ASKA library) as a new resource for *E. coli* biology (Kitagawa et al. 2005).

Here, we report the use of this resource in a systematic analysis of PPIs using pull-down assays. With the advent of matrix-assisted laser desorption ionization time of flight (MALDI-TOF) mass spectrometry methods, it is feasible to identify PPIs on a proteome-wide scale.

We have carried out a large-scale identification of protein–protein interactions to gain further understanding of the *E. coli* model cell at the system level. Because *E. coli* is one of the best studied organisms, it should also be an excellent target for systems biology (Kitano 2002) and synthetic biology fields (Silver and Way 2004) approaches.

Results and Discussion

To purify protein–protein interaction partners or protein complexes of *E. coli* systematically, we developed the strategy depicted in Figure 1. Each His-tagged ORF clone (Kitagawa et al. 2005) was transformed into wild-type *E. coli* K-12 W3110. Transformants were grown to the early log phase (OD600 = 0.3), and bait protein expression was induced by adding 0.1 mM IPTG. After 2 h, proteins from mildly sonicated cells were purified on a Ni2+-NTA column identified by MALDI-TOF MS (matrix-assisted laser desorption ionization time of flight mass spectrometry). An extended analysis of these interacting networks by bioinformatics and experimentation should provide new insights and novel strategies for *E. coli* systems biology.
Blue, digested by trypsin in gel slice, analyzed by MALDI-TOF MS, and identified by the PS1 identification program.

Of 4339 His-tagged bait proteins, 2667 proteins were successfully overproduced and purified. Of these, 2337 bait proteins copurified with an associated partner protein(s) and 330 bait proteins purified without any partners (Fig. 2B; Supplemental Table 2). The remaining 1568 bait proteins failed to be overproduced or purified (Fig. 2B; Supplemental Table 3). Of these, 928 were predicted to be membrane protein by SOSUI software (http://sosui.proteome.bio.tuat.ac.jp/sosuiFrame0.html); the remaining 640 are cytosolic proteins. Some of the latter may show toxicity or aggregation when overproduced. In addition, 104 His-tagged bait proteins failed to be identified by mass spectrometry when overproduced as bait proteins and were removed from further analysis (Supplemental Table 9). In the end, 16,050 interacting partners were identified for the 2667 bait proteins examined in our experiments.

To reduce identification of false positive interaction, we first carefully performed control experiments (n = 16) with exactly the same procedure described above, except by using the pCA24N vector to identify proteins with an intrinsic affinity for the Ni²⁺-NTA resin. Several proteins (TreC, GlmS, TufA, TufB, KatE, CstC, HisB, Crp, PykF, RplM) were frequently found (in at least 10 of 16 experiments) to be purified without any bait protein (Supplemental Table 5). In addition, 1906 (1895 plus 11 in least 10 of 16 experiments) to be purified without any partners (Fig. 2B; Supplemental Table 3). Of these, 928 were predicted to be membrane protein by SOSUI software (http://sosui.proteome.bio.tuat.ac.jp/sosuiFrame0.html); the remaining 640 are cytosolic proteins. Some of the latter may show toxicity or aggregation when overproduced. In addition, 104 His-tagged bait proteins failed to be identified by mass spectrometry when overproduced as bait proteins and were removed from further analysis (Supplemental Table 9). In the end, 16,050 interacting partners were identified for the 2667 bait proteins examined in our experiments.

We also identified 2644 (2633, self-identification, plus 11 in Fig. 2A) protein bands in SDS-PAGE gels that were identical to the His-tagged bait proteins. Whether these represented their endogenous gene products, as would be indicative of multimeric proteins or degradation products of His-tagged proteins, could not be distinguished. In total, we identified 11,511 protein partners in heterologous protein complexes in our analysis (Fig. 2; Supplemental Table 1).

To assess the quality of our results, we compared our data with the DIP database, catalogs of experimentally identified interaction between proteins (Xenarios et al. 2000). There are 447 PPIs in the DIP database with gene names for E. coli K-12. The bait proteins in our experiment matched one or both of 372-protein pairs in the DIP database. Of these, we successfully detected 58 PPIs (15.6%) in our screening. In a different study, 101 of 447 PPIs (22.6%) were detected using matrix approach (Bader and Hogue 2002).

Recently, data for more than 5000 E. coli PPIs were reported by a large-scale pull-down analysis using TAP- or SPA-tagged bait protein expressed from the chromosome under control of the native promoter (Butland et al. 2005). Their strategy kept cellular stoichiometries of target proteins; however, 2–4% cultures were required as starting material to overcome sensitivity problems. We used an overproduction system with multicopy plasmid clones to focus on the identification of potential interactors. Our system, therefore, lost stoichiometry between bait and prey proteins but avoided sensitivity problems. Other differences are the position and size of tag fragments, larger TAP or SPA tags at C terminus versus smaller His tag at N terminus for our study. A total of 521 bait proteins were used in common; these detected 1168 prey proteins by the TAP/SPA system (Butland et al. 2005) and 995 prey proteins in our study, consisting of 5030 and 3088 interactions, respectively. Among these, 218 (4% or 7% inter-
actions are in common. In spite of the different strategies, this overlap is similar to the comparison of the large-scale PPIs in yeast (Ito et al. 2002) and of value to apply for further biochemical and physiological analyses based on the result of both E. coli studies.

We also evaluated our results using overall topological properties. Although over one-half of the bait proteins had five or fewer interacting partners, some proteins, notably GroL and NadE, had hundreds. The frequency follows a power law, i.e., $\sim k^{-2.23}$, where $k$ is the number of interacting partners (Fig. 3A). The emergent property of this interaction network is that the average distance of protein pairs in the network will be short and will not be greatly influenced by random deletion of proteins, reflecting how the network evolved (Barabasi and Albert 1999). To evaluate statistically, we constructed a random network having the same number of edges as randomly choosing pairs of proteins from our result. The distribution was quite different from our observation (Fig. 3A); however, creating a random network by randomly shuffling bait proteins and subsequently prey proteins revealed similar topological properties (data not shown).

We recently identified 303 essential gene candidates in E. coli as genes that we were unable to delete from the chromosome (Baba et al. 2006). Using this information, we compared the number of interacting partners for essential and nonessential genes (Fig. 3B); we found that the connectivity is significantly higher for essential genes (17.8 in average) than for nonessential genes (6.7). This property is consistent with the observation in S. cerevisiae (Yu et al. 2004).

These results provide further support that the network of protein interactions identified in this study is sufficiently reliable and worthwhile to apply for further investigation of a global protein–protein interaction map in E. coli.

High-throughput protein–protein interaction screening increases the chance of identifying artificial partners that physically interact but that are not in close proximity in the cell, because they are in distinct subcellular locations or expressed at times. To reduce the number of false positives and to aid in extracting the substantial meaningful interactions, we classified interacting proteins by biological function according to the KEGG pathway database (http://www.genome.ad.jp/kegg/pathway.html; Kanehisa et al. 2002). Proteins were first classified into KEGG functional categories and interactions of proteins belonging to the same pathway were then analyzed by a single-linkage clustering algorithm as a way to find functional units (Fig. 4). In this way, we identified 107 functional units involved in metabolic pathways, transcriptional and translational machinery, recombination, ABC transport, or flagella assembly (Supplemental Table 6). To characterize the interaction of proteins not classified into KEGG pathway, we identified proteins that interact with at least two proteins belonging to the same pathway and defined these groups as functional components (Fig. 4). Sixty functional components, consisting of three or four proteins each, were identified. Of these, 42 components are involved in different metabolic pathways, 17 in RNA synthesis, and one in recombination (Supplemental Table 6). Therefore, protein–protein interactions in E. coli metabolic pathways were analyzed with respect to being in a functional unit.
Pyruvate dehydrogenase (PDH) is an enzyme complex for conversion of pyruvate to acetyl coenzyme A that funnels glycolytic carbon into the citric acid cycle. PDH is comprised of pyruvate dehydrogenase (E1P), lipoate acetyltransferase (E2P), and lipoyl dehydrogenase (E3) encoded by aceE, aceF, and lipdA, respectively (Quail et al. 1994). We identified a novel interaction of PDH with PdhR (Supplemental Table 6), which is a metabolite responsive repressor of the pdhB-aceE-aceF-lpdA operon. These data suggest that PdhB may function as a sensor of the enzymatic activity or cellular concentration of PDH.

The tricarboxylic acid (TCA) cycle is responsible for the oxidation of acetyl coenzyme A (CoA), which is derived mainly from pyruvate produced by glycolysis. We have identified the SucA–SucB complex that forms 2-ketoglu-tarate dehydrogenase and the succinate thiokinase complex (SucC–SucD; Buck et al. 1985). We detected an interaction between SucB and SucC in addition to the interactions of SucA–SucB and SucC–SucD, suggesting that a SucA–SucB–SucC–SucD complex may exist. The biological importance of such large complexes is unknown. In the TCA cycle, SucA–SucB and SucC–SucD catalyze separate reactions in a pathway; the interaction of SucA–SucB and SucC–SucD in a large complex may be important to prevent release of substrate from one reaction for the subsequent step.

We also identified a SucC–SdhA (succinate dehydrogenase) interaction, suggestive of formation of a SdhA–SuccD or SdhA–SucABC complex, for mediating conversion of 2-ketoglu-tarate to fumarate in the TCA cycle. E. coli has three distinct fumarases encoded by fumA, fumB, and fumC (Bell et al. 1989; Guest 1992; Guest and Russell 1992). FumA and FumB are 90% identical and belong to a novel class of fumarases that are dimeric, oxygen-sensitive, iron-sulfur proteins; our analysis identified the FumA–FumB interaction (Supplemental Table 6).

RNA polymerase (RNAP) is a multi-subunit complex for transcription, consisting of α, β, and β′ core subunits and one of several σ factors for promoter selection. Our analysis identified many known interactions, including all σ factors except σ32, which regulates most of the genes in stationary phase, NusA, NusG, and Rho, enhancing of RNAP pausing (Lau et al. 1983; Sigmund and Morgan 1988) and Rho-dependent termination (Nehrke et al. 1993; Kainz and Gourue 1998). Network structure was drawn using Cytoscape 2.1 (http://www.cytoscape.org/) in Figure 5. Additionally, we identified several novel interactions including RNAP with RhlE, a putative ATP-dependent RNA helicase, and RNAP with VacB, a 3′–5′ exoribonuclease (Fig. 5A). Earlier reports suggest that RhlE (Li et al. 1998) and VacB (Cheng et al. 1998) are important for RNA maturation. E. coli cold shock protein CspA, a homolog of eukaryotic Y-box DNA-binding proteins (Wolffe 1994; Sommerville and Ladomery 1996; Yamanaka et al. 1998) that binds strongly to the 110-bp segment of the hns promoter in the presence of RNAP (Brandi et al. 1994), was also identified in interaction with RpoB and RpoC (Fig. 5A), suggesting involvement in the formation of an open complex with RNAP. These data suggest that CspA may act as a transcriptional modulator.

Yet, we failed to identify interactions between RNAP and transcriptional factors, possibly due to their presence in low concentration of transient interactions with RNAP, or high affinity to target DNA. We did observe interactions between some chaperones (DnaK, ClpB, Tig) and RNAP (Supplemental Table 1). The DnaK (Hsp70) chaperone molecule has been reported to assist the proper folding and assembly of RNAP (Kimura and Ishihama 1996). Our data suggest that not only DnaK but also other chaperones (like ClpB and Tig) may assist the assembly of RNAP as well.

We constructed a global chaperone network based on our protein–protein interaction data showing the target specificities of the chaperones (Fig. 5B; Supplemental Table 7). Chaperones, such as DnaK, GroL, and Trigger Factor (Tig), are required for folding of newly synthesized polypeptides (Deuerling et al. 2003). We identified interactions of 776 ORFs with GroL. The molecular weights of most GroL (HSP60) substrates range from 20 to 80 kDa (data not shown), in agreement with earlier results (Houry et al. 1999). Furthermore, 310 DnaK (HSP70) substrates and 40 Tig substrates were identified. The functional relationship between these chaperone systems and the overlap of DnaK, GroL, and Tig is only partially understood; 92 proteins appear to be substrates of both DnaK and GroL. These proteins may require both systems for folding, degradation, or translocation. Interestingly, only six proteins were identified as overlapping targets of DnaK and Tig, in contrast to earlier observations (Deuerling et al. 2003). This inconsistency may be due to the use of different strains; nine proteins were identified as common substrates for Tig and GroL. These data suggest that Tig shares overlapping protein substrates or sequential process with other chaperones under normal growth conditions.

ClpB is unique among the Hsp100 family proteins of E. coli; ClpB cooperates with DnaK in the solubilization and refolding of unfolded proteins (Glover and Lindquist 1998; Goloubinoff et al. 1999; Mogk et al. 1999; Motohashi et al. 1999; Zolkiewski 1999). Of 43 candidate ClpB substrates, 11 proteins also interact with DnaK and 13 with GroL, suggesting that ClpB cooperates with DnaK/GroL (data not shown). YbdQ is an experimentally uncharacterized protein; however, it is a homolog of the universal stress protein UspA in E. coli and is induced following temperature shift (30°C–42°C; data not shown). We identified 83 target proteins of YbdQ, which mostly overlap with DnaK, DnaJ, and GroL. Our identification of multiple overlapping targets provides further insight into the process of protein folding by chaperones.

Prokaryotic genes are often organized into operons, clusters of genes that are transcribed together from the same mRNA. We estimated 653 transcriptional units based on correlation of expression profiles between genes (H. Maeno and S. Kanaya, unpublished), which is consistent with the estimated total of 630–700 operons in E. coli (Salgado et al. 2000). The interaction between proteins encoded by the same operon was analyzed and interactions in 52 operons were identified (Supplemental Table 8).

A total of 798 proteins of unknown function were identified as prey. First of all, these data provide evidence that these proteins are synthesized in a cell (Supplemental Table 4). To further characterize interaction of these prey proteins, we carried out domain analysis of frequently identified proteins (data not shown). As mentioned above, we observed that YbdQ contains the conserved domain of universal stress protein (UspA), YcH and YcgE contain the conserved domain of translational Initiation factor (COG0223, P01253) and helix turn helix (COG0789, PF00376) domains, respectively (http://www.ncbi.nlm.nih.gov/COG/old/xognitor.html). Hence, our data offer insight into the functional prediction of uncharacterized proteins.

In summary, our data from pull-down assay using His-tagged bait proteins represent a global protein–protein interaction map in an E. coli cell and provide a framework for an understanding of cellular systems. The functions of most of these complexes are not known; however, much work over the next few years will expand our knowledge of PPIs and help to analyze the structure and function of protein complexes qualitatively and quantitatively.
Methods

Purification of the bait protein

*E. coli* K-12 W3110 cells producing histidine-tagged bait proteins were grown at 37°C in 5 mL LB containing Cm (50 µg/mL) medium to OD_{600} = 0.3. Samples were taken 2 h after the addition of 0.1 mM IPTG. The cells were collected by centrifugation (10,000 rpm, 3 min, 4°C) and resuspended in 400 µL of cold buffer 1 (50 mM Sodium phosphate [pH 7.0], 200 mM NaCl, Proteinase inhibitor [Roche]). All manipulations were done at 4°C. Crude cell extracts were obtained by sonication (5 × 15 sec, level 3, Astrason ultrasonic processor) and centrifugation (16,000 rpm, 15 min). Cell extracts were loaded onto a 30-µL Nickel (Ni^{2+})-column (prepared according to manufacturer’s instructions [Qiagen] and equilibrated with buffer I), washed three times with 1 mL of buffer 2 (20 mM imidazole and 0.05% n-octyl-β-glucoside [PIERCE]), and proteins were eluted with buffer 3 (50 mM Tris-HCl [pH 6.8], 2% SDS, 0.1% bromophenol blue, 10% glycerol, 100 mM DTT, 6 M Urea and 250 mM imidazole). Eluted proteins were analyzed by 7.5–15% gradient SDS-PAGE (Bio Craft) followed by Coomassie Brilliant Blue (CBB-R250) staining. Electrophoresis was done at 20 mA current for 3 h by model BE-222 (Bio Craft).

Identification of prey proteins

Protein bands were excised as ~1-mm² pieces with a razor blade and transferred into buffer 3 (0.4 M ammonium carbonate and 50% acetonitrile). The gel was washed with Milli-Q water briefly and two times with 100 µL washing buffer (50% acetonitrile, 0.2 M ammonium hydrogen carbonate). After shrinking with 100% acetonitrile, the gel pieces were digested overnight at 37°C with the protease Trypsin (Sigma) (65 µL, 0.003 µg/µL, 10 mM HCl) in the reaction buffer (0.2 M ammonium bicarbonate). The resulting tryptic peptides were extracted from the gel with the elution buffer (50% acetonitrile, 0.3% trifluoroacetic acid) and desalted with ZipTip_{C18} (Millipore). Less than 1µL of the sample was spotted to the sample plate of mass spectrometry (Voyager DE-PRO, Applied Biosystems) with matrix solution α-cyano-4-hydroxy cinnamic acid (α-CHCA) (<5 mg/mL, 50% acetonitrile, 0.3% trifluoroacetic acid). Bait and prey proteins were identified by Data Explorer and/or Proteomics solution 1 (PS1) (Applied Biosystems). ORF database used for protein identification by PS1 was constructed by ourselves (GenoBase, http://ecoli.aist-nara.ac.jp).

Figure 5. Subgraph of protein–protein interactions. (A) RNAP interacting protein network. Networks of proteins that interact directly (blue nodes) with each of the RNAP subunits, RpoA, RpoB, and RpoC (red nodes). (B) Molecular chaperone and their target network properties. Red nodes represent major chaperone proteins and blue nodes represent their targets. The largest cluster shows the GroL interaction network and the second largest one shows the network of DnaK.
Comparison with published PPI data

To match our data with those from the use of the TAP/SPA system (Butland et al. 2005), protein names were compared in-batch, using values in Supplemental Table 1 and data provided by the authors. First, our 2337 and their 648 baits were compared to find 521 common baits; those interactions that do not result from the common baits were then excluded. Interactions excluded from our data are identified in Supplemental Table 1. Of the remaining 3088 and 5030 interactions from our and their experiments with the common baits were used to find 218 common interactions (Supplemental Table 1).

Acknowledgments

We thank Barry L. Wanner for critical reading and suggestions for improvement. This work was supported by a grant from CREST, JST (Japan Science and Technology) and in part from NEDO (New Energy and Industrial Technology Development Organization).

References


Large-scale identification of protein–protein interaction of
*Escherichia coli* K-12

Mohammad Arifuzzaman, Maki Maeda, Aya Itoh, et al.

*Genome Res.* 2006 16: 686-691
Access the most recent version at doi:10.1101/gr.4527806

Supplemental Material  http://genome.cshlp.org/content/suppl/2006/04/10/gr.4527806.DC1

References  This article cites 41 articles, 14 of which can be accessed free at: http://genome.cshlp.org/content/16/5/686.full.html#ref-list-1

License

Email Alerting Service  Receive free email alerts when new articles cite this article - sign up in the box at the top right corner of the article or click here.

To subscribe to *Genome Research* go to: http://genome.cshlp.org/subscriptions

Cold Spring Harbor Laboratory Press