Published in final edited form as:

J Proteome Res. 2008 June ; 7(6): 2234–2245. doi:10.1021/pr800055r.

Protein disorder is positively correlated with gene expression in *E. coli*

Oleg Paliy^{1,*}, Shawn M. Gargac¹, Yugong Cheng², Vladimir N. Uversky^{2,3}, and A. Keith Dunker²

¹Department of Biochemistry and Molecular Biology, Wright State University, Dayton, OH 45435

²Center for Computational Biology and Bioinformatics and Department of Biochemistry and Molecular Biology, Indiana University School of Medicine, Indianapolis, IN 46202

³Institute for Biological Instrumentation, Russian Academy of Sciences, 142290 Pushchino, Moscow Region, Russia

Abstract

We considered on a global scale the relationship between the predicted fraction of protein disorder and RNA and protein expression in *E. coli*. Fraction of protein disorder correlated positively with both measured RNA expression levels of *E. coli* genes in three different growth media and with predicted abundance levels of *E. coli* proteins. Though weak, the correlation was highly significant. Correlation of protein disorder with RNA expression did not depend on the growth rate of *E. coli* cultures and was not caused by a small subset of genes showing exceptionally high concordance in their disorder and expression levels. Global analysis was complemented by detailed consideration of several groups of proteins.

Keywords

disordered proteins; gene expression; microarrays; CAI; protein abundance; PONDR

Introduction

In contrast to many proteins with distinct three-dimensional structure, intrinsically disordered proteins lack a well-defined folded structure under physiological conditions¹. Often only a portion of the protein is maintained in the disordered state, whereas the rest of the molecule forms a defined three-dimensional shape. In most cases such a disordered state is crucial for the protein function, with many disordered proteins being involved in protein-protein or protein-DNA interactions^{2,3}. Examples of disordered proteins and proteins with disordered regions include RNA and protein chaperones^{4,5}, transcriptional regulators^{6,7}, signal transduction proteins^{8,9}, ion channels¹⁰, and motor proteins¹¹. Because specific amino acids occur often in known disordered proteins and are considered disorder-promoting, predictions of a potential existence of disordered regions in any protein can be made^{1,12}. Indeed, a number of predictors of protein disorder have been developed¹³ and used to estimate the fractions of protein disorder in proteomes of many different species^{14,15}.

The functional distribution of predicted disordered proteins was recently studied^{14,16–19}. Of 710 Swiss Protein functional keywords associated with at least 20 different proteins, 302 of

*Corresponding author. Address for correspondence: 260 Diggs Lab, Wright State University, 3640 Col. Glenn Hwy, Dayton, OH 45435. Fax (937) 775-3730, oleg.paliy@wright.edu.

the indicated functions were likely carried out by structured proteins and 238 by disordered proteins or disordered regions. Most of the structure-associated functions involved enzyme catalysis or membrane transport, while the disorder-associated functions typically involved signaling, regulation, and control^{16, 18, 19}. Possible connections between protein disorder and other protein and gene features have not yet been considered. In this report we examined on a global scale the relationship between predicted fraction of protein disorder and measured RNA expression and predicted protein abundance levels in *E. coli*.

Materials and methods

Microarray experiments

Affymetrix *E. coli* Antisense GeneChips were used to obtain RNA expression values for *E. coli* genes. *E. coli* K12 cells were grown in three different media: i) LB rich medium, ii) N⁺C⁻ minimal medium²⁰ with 0.4% glycerol and 10 mM ammonium chloride as carbon and nitrogen sources, respectively (designated MM1), and in iii) N⁺C⁻ minimal medium with 0.4% glycerol and 2.5 mM arginine as carbon and nitrogen sources, respectively (designated MM2). Doubling times of cultures in logarithmic phase of growth were 22 min (LB), 66 min (MM1), and 220 min (MM2). Strains were adapted to growth in the appropriate medium by preculturing to the early stationary phase from single colonies. Experimental cultures were incubated in a water bath shaker at 37°C in 500-ml Erlenmeyer flasks containing 1/10 volume of appropriate medium. Starting OD₆₀₀ of the experimental cultures was 0.05; cells were harvested in exponential phase (OD₆₀₀ 0.4–0.5). RNA was stabilized by addition of 1/10 culture volume of cold 95% ethanol – 5% acidic phenol. Total RNA was extracted with a hot phenol-chloroform method²¹. cDNA synthesis, fragmentation, and labeling were performed according to the instructions of the microarray manufacturer (Affymetrix, Inc., USA). Labeled cDNA was hybridized to *E. coli* Antisense GeneChip. Hybridization intensity data were obtained from the scanned array images, and all chips were processed in Affymetrix Gene Chip Operating Software (GCOS) using global normalization approach. Custom probeset mask was employed to limit normalization scaling to *E. coli* mRNA probesets only^{22, 23}. Two biological replicates for each growth condition surveyed were used, and the signal values for each gene were averaged (geometric mean) for each set of replicates. Signal values are shown in normalized arbitrary units.

Protein disorder prediction

A dataset of protein disorder predictions for all *E. coli* K12 proteins (based on *E. coli* MG1655 genome sequence) has been described previously¹⁷. The dataset included prediction of the disorder fraction for each *E. coli* protein (fraction of amino acids predicted to form disordered regions in each protein) and the disorder-order classification of *E. coli* proteins based on the consensus of the charge-hydropathy predictor²⁴ and the cumulative distribution function of the output scores of the PONDR VL-XT algorithm²⁵. In the disorder-order classification, each member of the disordered group is mostly or entirely disordered¹⁷. For consideration of highly expressed disordered proteins, three separate disorder prediction algorithms – PONDR VL-XT²⁵, VSL1²⁶, and VL3²⁷ were run as described in the algorithm references.

Protein abundance prediction

Codon adaptation index (CAI)^{28, 29} was used as a measure of predicted protein abundance. The dataset of CAI values for all *E. coli* genes was taken from <http://www.evolvingcode.net/codon/cai/cais.php> and is based on the calculations of Carbone *et. al.*³⁰.

Statistical analysis

Statistical analysis was carried out in `MatLAB` v7.1. Because all datasets were non-normally distributed and had different value ranges, rank correlation was used to assess the relationships among the data. To test the significance of the calculated Spearman and Kendall Tau correlation coefficients, a permutation algorithm was run to assess the chance of obtaining an observed correlation randomly. Specifically, for each comparison of two datasets, one of the two datasets was left unchanged in the permutation analysis, whereas the values in the second dataset were randomly rearranged, and the correlation coefficient between this permuted dataset and the first dataset was computed. A total of 100,000 permutations were run for each dataset comparison considered, and the distribution of correlation coefficients among these permutation comparisons was obtained.

Results and discussion

Relationship between protein disorder and gene expression in *E. coli*

In order to evaluate the relationship between protein disorder and the gene expression levels in *E. coli*, we measured RNA levels in *E. coli* cells grown in three different conditions: LB rich medium and N⁺C⁻ minimal media supplemented with glycerol as a carbon source and either ammonium chloride (medium designated MM1) or arginine (medium designated MM2) as a nitrogen source. The growth conditions were chosen to provide cell cultures with a wide range of growth rates (doubling time [DT] of 22–220 min). Since the expression of many highly abundant proteins depends on growth rate^{21, 31}, the use of three different growth conditions with widely different DTs of the cultures allowed us to assess if the correlation between predicted protein disorder and RNA expression is dependent on the cell growth rate. As expected³², the average RNA levels in *E. coli* cells depended on culture DT: average RNA expression for *E. coli* genes in three media was LB>MM1>MM2 (Table 1). The disparity was attributed to large differences in the expression levels of ribosomal and transfer RNAs and of mRNA genes coding translational proteins and enzymes of energy metabolism³². Another notable difference among cells grown in different media was the number of genes for which RNA was detected. Whereas there were 2700–3000 RNA species present in *E. coli* cells grown in minimal media, only 2100 genes were detected at the RNA level in LB, mostly because *E. coli* cells did not express many biosynthetic enzymes in rich LB medium.

We used non-parametric rank correlation to evaluate the relationship between mRNA expression (number of mRNAs of each protein-coding gene) and the predicted fraction of protein disorder (fraction of amino acids predicted to form disordered regions in each protein; designated FD) in *E. coli*. Fraction of protein disorder displayed positive correlation with the mRNA expression levels in all three media independently of the rank correlation algorithm used (Table 2). Though weak, the correlation was however highly significant as assessed by permutation analysis ($p < 10^{-5}$). PONDR VL-XT cumulative distribution function scores were combined with charge-hydropathy plots²⁴ to develop a consensus protein disorder classification that indicates proteins that are mostly or entirely disordered; for *E. coli*, 190 such proteins were found¹⁷. The average expression levels of the corresponding genes for these proteins were much higher in each medium compared to the average expression level of *E. coli* gene in that medium. For example, the average mRNA expression in LB was 6600 for disordered proteins, but only 1900 for ordered proteins (average for all protein-coding genes was 2200).

To evaluate these relationships further, we have plotted a distribution of mRNA expression values as a function of protein disorder (Figure 1). All protein-coding *E. coli* genes were sorted by their disorder fraction, partitioned into ten classes (bins) of equal size, and the average mRNA expression in each class was then calculated. In most cases, average mRNA expression

was progressively higher with a higher fraction of predicted disordered regions in a protein (Figure 1). The figure also shows that the observed correlation was not due to a small number of proteins/genes skewing overall results. For example, we have noticed that ribosomal proteins had an average predicted disorder fraction of 0.50 and were some of the most highly expressed genes at the RNA level (average mRNA expression values in LB – 34000, in MM1 – 26700, and in MM2 – 15100; compare to averages for the whole transcriptome in Table 1). To test if the presence of these proteins can account for the majority of the observed positive relationship between FD and gene expression, we have removed these proteins from our datasets and re-ran the correlation analysis. In all cases, a statistically significant positive relationship was still observed between FD and mRNA expression ($p < 10^{-5}$), although the correlation coefficients were slightly lower than those when a complete dataset was considered. For LB, the new Spearman (R_s) and Kendall Tau (R_τ) coefficients were 0.211 and 0.144, respectively (compare these with R_s and R_τ values for complete datasets in Table 2); for MM1: $R_s=0.283$ and $R_\tau=0.194$, and for MM2: $R_s=0.277$ and $R_\tau=0.190$.

Relationship between protein disorder and protein abundance in *E. coli*

To compare the fraction of protein disorder with protein abundance levels, we used codon adaptation index (CAI) as a measure of predicted protein expression in *E. coli* cells³³. Correlation analyses revealed a weak positive relationship between the fraction of protein disorder and CAI index of *E. coli* proteins (Table 2). Similarly to the gene expression – protein disorder comparison, average CAI value was progressively higher with a higher fraction of predicted disordered regions in a protein (Figure 1). As a control, we assessed the relationship between FD and protein length: no correlation was evident ($R_s=-0.077$ and $R_\tau=-0.050$).

Differences in disorder-expression relationship as a function of different growth conditions and protein function

We have noticed that *E. coli* cells grown in LB displayed lower correlation between their gene expression levels and predicted fraction of protein disorder compared with cells grown in minimal media (Table 2). This trend was not growth-rate dependant since there was no difference in R_s and R_τ values for the cells grown in minimal media despite differences in the doubling times of the cultures (66 and 220 min). One potential explanation of these findings could have been the difference in the number of RNAs detected in cells grown in different media: Affymetrix GeneChips have detected reliably only 2100 *E. coli* genes in LB medium, whereas 2700–3000 genes were detected in minimal media. To test this hypothesis, we have compiled a truncated dataset that only contained *E. coli* genes detected in LB (present and marginally present genes as defined by MAS 5.0 algorithm³⁴) and have calculated the R_s and R_τ coefficients between gene expression in LB and protein disorder fraction. The correlation coefficients were virtually the same as those for the complete dataset ($R_s=0.234$, $R_\tau=0.161$), indicating that RNA presence or absence in LB-grown *E. coli* cells could not explain observed differences in FD correlation among media profiled.

However, when we considered genes unique to each medium, a notable difference in correlation rates was found. Specifically, we have examined the correlation between FD and mRNA expression for *E. coli* genes that were expressed only in LB (122 genes), only in MM1 (107 genes), or only in MM2 (226 genes). A subset of genes expressed in all three media was used for comparison (2419 genes total). The correlation between FD and mRNA values in LB was notably lower for LB-unique genes ($R_s=0.127$) than for genes expressed in all three conditions ($R_s=0.228$). At the same time, there was only small difference in R_s for MM1 and MM2 unique genes ($R_s=0.261$ and $R_s=0.239$, respectively) compared to those expressed in three media ($R_s=0.271$). Genes that were expressed in both MM1 and MM2 media displayed the highest concordance between FD and mRNA values ($R_s=0.466$ for FD-mRNA_{MM1} comparison and $R_s=0.416$ for FD-mRNA_{MM2} comparison) and included a number of those

coding for biosynthetic enzymes³². Among the genes unique to LB, several enzymes that had low fraction of amino acids in disordered regions were expressed at relatively high level in that medium which led to poor overall correlation. These included *sdaA* (b1814, serine degradation), *gatZABCD* (b2091-95, galactitol degradation), *treC* (b4239, trehalose degradation), *deoC* (b4381, deoxyribonucleoside degradation), and *fucl* (b2802, degradation of sugars). These findings are consistent with earlier observations indicating that the main differences between gene expression patterns in rich and minimal media comprise the enzymes of degradative and biosynthetic pathways: whereas many of the former are expressed in rich media but not in minimal growth conditions, the opposite is observed for biosynthetic functions³².

In order to further evaluate the FD-mRNA correlation among different media profiled, we have used MultiFun functional classification of *E. coli* proteins³⁵ to consider this relationship as a function of protein role. Because two large functional groups of genes – “carbon compound utilization” (470 total members) and “cell structure” (1180 total members) – had low average FDs of the corresponding proteins (FD of 0.23 and 0.19, respectively; average for proteome is 0.24) and were expressed at a significantly higher level in LB than in MM1 and MM2 media (in LB – 3400 and 8600 on average, respectively, for these two groups; in MM2 – 1400 and 3200, respectively), they accounted for a reduction in the overall rank correlation between FD and mRNA_{LB}. On the other hand, flagellar proteins had a relatively high disorder level (0.37) and were highly expressed only in the minimal medium with ammonium chloride as the nitrogen source. They were not expressed in LB medium likely due to catabolite repression of flagella operons by glucose³⁶, and they were expressed at low levels in minimal medium with arginine as the nitrogen source (a poor source of nitrogen for *E. coli*) because flagella synthesis is down-regulated in poor growth conditions²¹. Proteins of translation, replication, and cell division were classified as disordered more often, and contained a higher fraction of disordered amino acids than average for the proteome. They were expressed at a high level in all three media and had high CAI values. At the same time, proteins participating in transport and metabolism were mostly ordered. We have observed only a slightly higher fraction of protein disorder (compared with an average for *E. coli* proteome) for several functional groups of *E. coli* proteins that were reported previously to contain increased representation of disordered members in other species^{14, 37} – these included chaperones, transcriptional regulators, and signal transduction systems. Considering protein localization, membrane proteins were predominantly ordered (average FD – 0.18), whereas proteins with known cytoplasmic localization had a slightly higher disorder fraction (average FD – 0.29) than average for proteome.

Disorder-expression relationship for essential *E. coli* proteins

We have considered a subset of *E. coli* proteins that are known from the literature to be essential for the survival and growth of this bacterium. The list of such proteins was taken from the PEC database (<http://www.shigen.nig.ac.jp/ecoli/pec/index.jsp>) and a total of 216 essential proteins were considered. Essential proteins had on average a much higher fraction of disorder (FD=0.36), had a higher number of proteins classified as completely disordered (19% vs. 2% for *E. coli* proteome), and were expressed at a higher level in all three media than an average *E. coli* gene, in part because the majority of translational proteins were in the list. Recently described features of protein-protein interaction networks provide a possible explanation for the observed correlations between protein intrinsic disorder, gene transcription levels, and protein essentiality. Protein-protein interaction networks contain a few proteins (hubs) that interact with many partners, and many proteins (ends) that interact with a very few partners^{38–40}. Deletion of a hub protein is lethal more often than is the deletion of an end protein⁴¹, so there exist a likely correlation between protein hubs and essentiality. Indeed, Arifuzzaman and colleagues³⁹ showed experimentally that protein-protein connectivity was significantly

higher for essential *E. coli* gene candidates (17.8 protein partners on average) than for non-essential proteins (6.7 partners), suggesting that hub proteins are indeed the more essential ones. Similar observations were made by Yellaboina *et al.* for the predicted *E. coli* protein interaction network⁴². While not shown previously for *E. coli*, in several other species hub proteins are indicated to contain more disorder than non-hub proteins^{43–45}, because disorder has several advantages for protein-protein interactions, especially for enabling one protein to bind to multiple partners as is required for hubs⁴⁶. Our data provides evidence that a similar trend exists in the proteome of *E. coli* as well, providing a link between protein essentiality and protein disorder in this bacterium.

Structural and functional annotation of highly expressed proteins possessing high levels of predicted disorder

To better understand the function-disorder relationship for highly expressed *E. coli* proteins, we carried out manual literature mining and considered in detail a group of such proteins that had high levels of predicted intrinsic disorder. Our goal was to identify for each protein example the functions associated with at least one experimentally validated intrinsically disordered region and to see whether these experimentally determined disordered regions coincide with regions of predicted disorder. The list of highly expressed disordered proteins was generated by filtering for all *E. coli* genes/proteins in the top 10% of RNA expression in each of three media, top 10% in the CAI index, and top 10% in the fraction of protein disorder. Filtered subsets were superimposed to generate the final set where the genes are in the top 10% of each individual dataset. This procedure produced a list of 51 highly expressed proteins predicted to contain high level of protein disorder. Among these, 30 were ribosomal proteins. The assembly of the ribosome, which involves the sequential binding of numerous proteins *via* multiple pathways leading to large-scale changes in the conformation of the associated RNA and proteins, represents an extreme case involving dramatic structural changes induced by protein-RNA interactions^{47, 48}. Many ribosomal proteins have been shown to be disordered prior to binding to rRNA and to acquire ordered structure during ribosome formation^{24, 49}. The remaining 21 proteins/genes are listed in Table 3. Predictions of disordered regions in these proteins based on three different predictors (PONDR VL-XT²⁵, VSL1²⁶, and VL3²⁷) are provided in Figure 2. We did not find relevant functional or structural information for three proteins from this list: putative α -helix protein YeeX (b2007), putative outer membrane protein SlyB (b1641), and DNA-binding protein HU-beta (b0440). Some important functional and structural properties of the other highly expressed disordered proteins are outlined below. Although many of these proteins contain multiple regions of predicted disorder, analysis of any given protein was considered sufficient if functional and structural information was found for at least one disordered segment.

AceF (b0115) - dihydrolipoyl acetyltransferase component of pyruvate

dehydrogenase—The pyruvate dehydrogenase (PDH) complex catalyzes the oxidative decarboxylation of pyruvate, transferring the resultant acetyl group to coenzyme A. The PDH complex contains three enzymes: pyruvate decarboxylase (E1p), dihydrolipoyl acetyltransferase (E2p), and dihydrolipoyl dehydrogenase (E3)⁵⁰. In the PDH complex from *E. coli*, E2p forms a cubic core with E1 and E3 proteins non-covalently bound to that core^{50, 51}. The independently folded lipoyl domains (residues 2–74, 106–177, and 207–278) form the N-terminal part of the E2 chain, whereas the C-terminal half of the protein contains subunit binding and catalytic domains. NMR analysis of the E2p fragment containing the third lipoyl domain (residues 205–295) revealed that this domain is composed of two four-stranded β -sheets⁵². Three highly flexible regions were identified, the surface loop linking β -strands 1 and 2 (residues 214–219), the lipoyl-lysine β -turn (residues 244–247), and the C-terminal tail (residue 283–295)⁵². Figure 2A shows that these regions are predicted to be disordered.

HlpA (Skp, b0178) - periplasmic molecular chaperone for outer membrane proteins—This 17 kDa protein (also known as OmpH) is an important player in the chaperone system that is found in the periplasm and which assists in the folding of outer membrane proteins and in their insertion into the outer membrane⁵³. Skp has been characterized as a molecular chaperone that interacts with unfolded proteins emerging in the periplasm from the Sec translocation machinery^{54, 55}, and Skp is required for efficient release of translocated proteins from the plasma membrane⁵⁵. Analysis of the X-ray crystal structure showed that the protein monomer is composed of two domains, the small association domain (amino acids 19–41 and 133–161) and a tentacle-shaped coiled α -helical domain formed by amino acids 42 to 132⁵⁶. This coiled α -helical domain possesses significant conformational flexibility⁵⁶. Figure 2B shows that the vast majority of this domain is predicted to be intrinsically disordered. Interestingly, proteins that form flexible fibers upon association, such as collagen or coiled-coil α -helices, are of low complexity and are indeed disordered as monomers⁵⁷ with a defined structure occurring only upon the formation of protein complexes. Thus, it is not surprising that coiled-coil loops within single proteins are usually predicted to be disordered if they are long enough.

SucB (b0727) - dihydrolipoyl succinyltransferase chain of 2-oxoglutarate dehydrogenase—Three classes of the dehydrogenase multi-enzyme complexes have been characterized, specific for pyruvate (described above), 2-oxoglutarate, and branched-chain α -keto acids derived from leucine, isoleucine, and valine. The 2-oxoglutarate dehydrogenase is one of the enzymes of the TCA cycle responsible for the conversion of 2-oxoglutarate to succinyl-CoA. Each dehydrogenase complex is composed of multiple copies of three enzymes: a substrate-specific decarboxylase-dehydrogenase (E1), a dihydrolipoamide acyltransferase (E2) specific for each type of complex, and dihydrolipoamide dehydrogenase (E3). The E2 proteins are modular structures comprising one to three lipoyl domains, a peripheral subunit-binding domain, and a large core domain^{50, 58}. The domains are joined together by long (25 to 30 residues) polypeptide segments, which form flexible extended linkers^{59, 60}. An NMR analysis of the E2 lipoyl domain (residues 1–92) of the *E. coli* 2-oxoglutarate dehydrogenase revealed that the C-terminal fragment of this domain (residues 81–92) lacks a defined structure. This C-terminal region corresponds to the beginning of the flexible inter-domain linker (residues 81–105)⁶¹, which is predicted to be highly disordered (Figure 2C).

Pal (b0741) - peptidoglycan-associated lipoprotein—The structural peptidoglycan (PG) layer is located in the periplasm of gram-negative bacteria between a lipopolysaccharide-containing outer membrane and an inner cytoplasmic membrane. A number of proteins, including the outer membrane proteins Pal and OmpA, and the inner membrane flagellar motor protein MotB, interact non-covalently with the cell wall through a periplasmic PG-binding domain. Pal stabilizes the outer membrane by providing a non-covalent link to the PG layer⁶². In addition to this structural function Pal participates in an active mechanism of transport of molecules through the membranes⁶³. The N-terminal fragment of Pal (residues 1–35) from *Haemophilus influenzae*, which is missing in a crystal structure of *E. coli* Pal (PDB ID: 1oap), has been demonstrated by NMR analysis to be highly flexible⁶⁴. This observation is further illustrated by Figure 2D, which shows that the N-terminal fragment of Pal is predicted to be intrinsically disordered.

InfA (b0884) - translational initiation factor IF1—Initiation factor IF1 is a monomeric protein consisting of 71 residues⁶⁵. IF1 is a highly conserved element of the prokaryotic translational apparatus and has also been found in the chloroplasts of plants. Several functions have been reported for IF1, including the enhancement of the rate of 70S ribosome dissociation and subunit association⁶⁶; the stimulation of the activity of initiation factors IF2 and IF3 in the formation of the 30S initiation complex⁶⁷; and the modulation of the interaction of IF2

with the ribosome⁶⁸. NMR analysis revealed that IF1 is an ordered protein consisting of a five-stranded β -sheet, arranged as a β -barrel⁶⁹. While the turns connecting strands I and II and strands II and III are well defined, the region connecting strands III and IV (residues 37–49) and the short loop connecting strands IV and V (residues 57–62) have considerable flexibility on the nanosecond time scale. Similarly, five residues at both N- and C-termini possess high flexibility⁶⁹. Figure 2E shows that these regions are predicted to be disordered by PONDR VL-Xt.

AcpP (b1094) - acyl carrier protein—Acyl carrier protein is a cytosolic protein of 77 amino acids in *E. coli*⁷⁰. Its holo form (holo-Acp) is a fundamental component of fatty acid biosynthesis. Acp is one of the most abundant soluble proteins in *E. coli*, with a cellular concentration of approximately 0.1 mM. Far-UV circular dichroism (CD) analysis revealed that the apo-Acp is devoid of any ordered structure at low ionic strength and at physiological pH as indicated by its small negative mean residue ellipticity at 220 nm^{71, 72}. In agreement with these experimental data, Figure 2F shows that Acp is predicted to be almost completely disordered by two algorithms.

Interestingly, even though Acp was shown to be completely disordered in solution^{71, 72}, the crystal structure of the *E. coli* acyl carrier protein was reported (PDB entry 1t8k)⁷³. The explanation for this apparent contradiction is in the conditions used to crystallize the protein. Far-UV CD analysis was performed under the low ionic strength conditions, whereas large excess of zinc acetate (0.1–0.5 M) was employed to crystallize apo-Acp⁷³. In the latter study, the metal-ion-mediated crystal-packing interactions were used to improve the success rate of protein crystallization signals⁷³. In fact, specific and non-specific metal ion binding is frequently used to stabilize and even crystallize highly charged proteins, which otherwise are intrinsically disordered due to the strong electrostatic repulsion and because of the low content of hydrophobic residues that preclude them from the formation of compact, collapsed conformations typical for the globular proteins²⁴. Acp is a good example of such protein.

HNS (b1237) - DNA-binding protein HNS—HNS is a 15.6 kDa protein that is present at high copy number (up to 10⁵ molecules) in cells of all enterobacteria⁷⁴. This protein interacts with DNA in the bacterial nucleoid in order to facilitate chromosome condensation and modulate gene expression^{75–77}. HNS is composed of two distinct domains: the N-terminal domain, which mediates protein oligomerization; and the C-terminal domain, which is required for non-sequence-specific DNA binding⁷⁸. Oligomerization of HNS is driven by the first 64 residues, which form typical coiled-coil structure. The structure of this fragment is highly concentration-dependent being mostly disordered at low protein concentration and being dominated by α -helices in the coiled-coil conformation at high protein concentrations⁷⁸. The N- and C-terminal domains are separated by a flexible linker that provides the C-terminal domain with a freedom of movement relative to the oligomeric N-terminal domain⁷⁸. Figure 2G shows that both the C-terminal region and the linker region of HNS are predicted to be highly disordered.

PtsI (b2416) - enzyme I of the phosphotransferase system (PTS)—PTSs are widely distributed in bacteria and catalyze sugar phosphorylation/transport, acting both as a sugar “kinase” and as a translocase. Enzyme I (EI) is the first protein in the phosphotransfer sequence that accepts the phosphoryl group from phosphoenolpyruvate (PEP). Proteolytic cleavage of EI produces two domains⁷⁹: the N-terminal domain (residues 1–230) containing the His-189 residue that transfers the phosphoryl group^{80, 81}, and the C-terminal domain (residues 261–575) that binds PEP in the presence of Mg²⁺^{79, 82} and mediates dimerization of the EI⁸³. Crystallographic analysis of the EI dimer revealed that a 30-residue linker (residues 231–260) forms an α -helix. Packing of this helix against the rest of the structure is loose, suggesting potential flexibility that may promote large-scale rearrangement of the domains in the course

of the phosphoryl group transfer⁸⁴. Figure 2H illustrates that this linker region is predicted to be disordered.

GcvH (b2904) - glycine cleavage complex H protein—The glycine-cleavage system (GCS) is a multi-enzyme complex consisting of four different components designated P-protein (200 kDa), H-protein (a monomeric lipoamide-containing protein, 14 kDa), T-protein (41 kDa), and L-protein (100 kDa). This complex catalyzes the oxidative cleavage of glycine in a multistep reaction. GCS-H plays a pivotal role in this process by interacting with the three other enzymes through the lipoyl-lysine arm - a lipoic acid that is covalently bound to lysine. We observed a striking difference in the expression level of *gcvH* in LB (rich medium containing amino acids) and in minimal media (Table 3). NMR analysis of the mitochondrial GCS revealed that the mobility of the major C-terminal helix of the H-protein is dramatically affected by the loading of the methylamine group from the glycine molecule onto the H-protein^{85, 86}. In addition to the flexibility of this helix, two other regions of H-protein were shown to be affected by the lipoylation: the lipoylation site (residues 60–66) and the interaction region (residues 33–36), and the 79–84 region in the "backside" of the protein. Figure 2I shows that these regions are predicted to be flexible by PONDR VL-XT.

NusA (b3169) - transcription pausing factor—RNA synthesis in *E. coli* is catalyzed by RNA polymerase (RNAP), a multi-protein enzyme with an $\alpha_2\beta\beta'$ -subunit core⁸⁷. After initiation of transcription, the essential transcription factor NusA associates with the RNAP core enzyme, where it modulates transcriptional pausing, termination, and antitermination⁸⁸. The crystal structures of NusA factors from *Thermotoga maritima*^{89, 90} and *Mycobacterium tuberculosis*⁹¹ have been solved. Both proteins show a common domain organization, with an amino-terminal RNAP-binding domain followed by three RNA-binding domains. The structures of *Thermotoga maritima* NusA determined by two different groups^{89, 90} are almost identical with the exception of three regions that show large structural differences (residues 1–19, 6.4 Å difference for mainchain atoms; residues 98–108, 4.3 Å difference; residues 123–132, 5.3 Å difference)⁹⁰. Figure 2J shows that the first and the last of these regions are predicted to contain significant amounts of intrinsic disorder in *E. coli*. NusA from *E. coli*, *Chlamydia*, and *Treponema* contain an additional carboxy-terminal region NusACTD that is comprised of 160 residues⁹². NusACTD serves as a multipurpose protein-protein interaction site^{93, 94}. NMR analysis of the NusACTD from *E. coli* (residues 339–495) revealed that this domain includes two structurally similar subdomains separated by highly flexible linker (417–430)⁹³. This linker is predicted to be intrinsically disordered (Figure 2J).

AccB (b3255) - biotin carboxyl carrier protein (BCCP) of acetyl-CoA carboxylase—A homodimer of the biotin carboxyl carrier protein is one of three functional components of acetyl-CoA carboxylase (ACC) that catalyze the initial step in the *de novo* biosynthesis of long-chain fatty acids through the carboxylation of acetyl-CoA to malonyl-CoA⁹⁵. In the functional cycle of ACC, the BCCP participates in three heterologous protein-protein interactions. First, it serves as the substrate in the biotin ligation reaction, in which the cofactor is linked to a specific lysine residue of apo-BCCP. Second, in its biotinylated form, holo-BCCP interacts with the biotin carboxylase subunit during the incorporation of CO₂ into the ureido group of the biotin ring system. Finally, the carboxylated holo carrier protein interacts with transcarboxylase during the transfer of CO₂ moiety to acetyl-CoA to yield malonyl-CoA. BCCP has two regions of known function. The C-terminal half of the protein is a tightly folded and highly conserved biotin domain^{95, 96}. Upstream of the biotin domain is a flexible linker region of about 42 residues with over half of residues being proline or alanine⁹⁷. The remaining N-terminal residues of BCCP are responsible for protein-protein interactions⁹⁸. The flexibility of the N-terminal fragment was determined in the proteolytic cleavage experiments, which

produced a protease resistant C-terminal domain^{99, 100}. Figure 2K supports these observations and shows that N-terminal region is predicted to be intrinsically disordered.

RpoZ (b3649) - omega (ω) subunit of RNA polymerase—*E. coli* RNA polymerase consists of the α , β , β' , and ω subunits⁸⁷. The ω subunit, encoded by *rpoZ* gene, is required in the restoration of the denatured RNA polymerase core ($\alpha_2\beta\beta'$) to its functionally active form¹⁰¹. The defined role of ω is to hold β' subunit in a non-aggregated form and to recruit β' into the $\alpha_2\beta$ subassembly¹⁰². The ω subunit is located at the surface of the holoenzyme and is accessible for interaction with transcriptional activators¹⁰³. Using a limited proteolysis with the endoproteinase GluC, a protease-sensitive 38-amino-acid domain was identified in the C-terminal portion of the protein indicating likely lack of compact structure making the region accessible to the protease¹⁰². Figure 2L shows that this part of the protein is predicted to be disordered by all three algorithms.

AtpC (b3731), AtpD (b3732), AtpF (b3736) - epsilon (ϵ), beta (β) and b subunits of the F_1F_0 ATP synthase—Membrane-bound F_1F_0 ATP synthase is a ubiquitous multi-subunit enzyme complex that catalyzes the synthesis of ATP from ADP and phosphate using proton gradient as a driving force for the reaction. In *E. coli* the ATP synthase is built up of two different components, the peripheral F_1 part ($\alpha_3\beta_3\gamma\delta\epsilon$) and the membrane-embedded F_0 complex (ab_2c_{12})¹⁰⁴. The F_1 and F_0 components are linked by two 45 Å stalks. A central stalk is formed by the ϵ subunit and by part of the γ subunit, whereas the peripheral stalk is constituted by the hydrophilic portions of the two *b* subunits of the F_0 and the δ subunit of the F_1 ^{105, 106}. The function of the central stalk is to transmit energy between the proton channel in the membrane and the catalytic nucleotide binding sites on the β subunits *via* conformational changes in the stalk-forming proteins. The ϵ subunit of the central stalk exhibits a two domain structure. The N-terminal 84 residues form a 10-stranded β -sandwich, whereas the C-terminal 48 amino acids are arranged as two α -helix hairpins. Although these two domains are relatively tightly associated with little or no flexibility relative to one another, the ϵ subunit was shown to exist in two states depending on whether ATP or ADP occupied the catalytic sites¹⁰⁷. Figure 2M shows that this functional flexibility can be attributed to the intrinsically disordered nature of the linker.

The function of the *b* subunit (the “second stalk” of ATP synthase) is to link the F_1 and F_0 components, thereby keeping the F_1F_0 complex together at the membrane and coupling ATP synthesis to proton translocation¹⁰⁸. The *b* subunit is present as a dimer in ATP synthase, and residues 62–122 are required to mediate dimerization¹⁰⁸. Figure 2O shows that this dimerization region is predicted to be disordered.

In the F_1 component, three α - and three β -subunits are arranged alternately around a central helix formed by C-terminal fragment of γ -subunit¹⁰⁹. Analysis of the crystal structure of F_1 revealed that the β subunits are present in three different conformations: tight, loose, or open, containing 5'-adenylyl- β - γ -imidodiphosphate, ADP, or no nucleotide, respectively. Open conformation differs from the other two by a large hinge motion of the C-terminal domain¹¹⁰. Figure 2N shows that this hinge region is predicted to be disordered.

NusG (b3982) - transcription termination protein NusG—NusG is an abundant transcription elongation factor that modulates Rho-dependent transcription termination by aiding in the recruitment of Rho to the transcription complex¹¹¹. It also participates in the antitermination of transcription of ribosomal RNAs. Based on the crystal structure of NusG from *Aquifex aeolicus* (aaeNusG) a homology model for *E. coli* NusG (ecoNusG) was created¹¹². According to this model ecoNusG has two domains that are equivalent to domain I and domain III of aaeNusG. In place of aaeNusG domain II, ecoNusG contains an extended loop (residues 44–67)¹¹². There are three highly flexible regions in ecoNusG: the seven N-terminal

amino acids, an extended loop replacing domain II and a linker connecting domains I and III (residues 116–133). All these regions are predicted to be intrinsically disordered (Figure 2P).

GroES (b4142) and GroEL (b4143) chaperones—Chaperonins and co-chaperonins are ring-shaped molecular chaperones. *E. coli* GroEL and GroES are the prototypical chaperonin and co-chaperonin pair. Together, they comprise the GroE intracellular machine that binds, unfolds, and refolds nascent and misfolded bacterial proteins^{113, 114}. GroEL is an oligomeric complex of 14 identical 57-kDa subunits, arranged in two 7-membered rings. Each GroEL subunit consists of 547 amino acids arranged in three distinct domains: an apical domain (residues 189–377) that binds non-folded proteins and GroES, an intermediate domain (residues 137–188 and 378–409), and an equatorial domain (residues 2–136 and 410–525) that binds ATP and is involved in inter-ring interactions¹¹⁵. GroEL and GroES undergo cyclic interactions of association and dissociation controlled by an "ATP clock", where ATP hydrolysis is coupled to cycling¹¹⁶. Binding of ATP and GroES to GroEL causes a major conformational change in GroEL involving a rotation and twist of the GroES-binding apical domain of GroEL. Figure 2Q illustrates that the fragments surrounding the hinge regions connecting the apical domain with the intermediate and equatorial domains are predicted to be intrinsically disordered.

In the operation of the GroEL machine, GroES acts both as an allosteric modulator of GroEL and a competitor of substrate¹¹⁷. The GroES 7-mer binds to the GroEL 14-mer *via* mobile loops located on each of the GroES subunits¹¹⁸. The GroES mobile loop is a stretch of ~16 amino acids that exhibits a high degree of flexible disorder in the free protein. Although only a hydrophobic tripeptide in the mobile loop physically interacts with GroEL, the entire mobile loop becomes structurally ordered upon binding to GroEL^{118, 119}. Figure 2R shows that a region corresponding to this flexible loop behaves as a short region of predicted order embedded within the long region of predicted disorder. This is a characteristic pattern of molecular recognition features, MoRFs, which are short disordered regions that undergo disorder-to-order transition upon interaction with their binding partners^{37, 120, 121}.

Throughout our analysis of known structural information for the above proteins we came across several cases where the proteins predicted to be disordered and shown to be disordered in solution were able to fold into a crystallizable structure under specific conditions. The explanation for this apparent contradiction is in the conditions used to crystallize the proteins. Since intrinsically disordered proteins/regions (IDPs/IDRs) are characterized by specific amino acid composition distinct from that of ordered proteins, there are dramatic differences in the folding behavior of these two protein classes. All the information necessary for the intrinsically ordered proteins to fold is encoded in their amino acid sequences¹²², and these proteins do spontaneously fold into unique 3D structure to fulfill their specific biological activities. On the contrary, a portion of the folding information is missing from the IDP sequences and these proteins do not usually fold spontaneously. However, some IDPs/IDRs can fold after a specific interaction with their unique binding partners^{1, 24, 123, 124}. In these cases, the propensity to fold is encoded in the collective properties of the IDP amino acid residues complemented by the properties of its binding partner, and therefore, the complex between the IDP and its binding partner represents a foldable unit. As a result, some protein segments that are predicted to be disordered show up in ordered structures as parts of multi-protein assemblies, or as protein-ligand complexes. Some of these potentially foldable and crystallizable IDRs, known as molecular recognition features (MoRFs), are involved in protein-protein interactions and are characterized by a very specific pattern in the disorder prediction plots and therefore can be predicted with relatively high accuracy^{37, 120, 121} as we described above for GroEL.

Conclusions

In this paper we have considered the relationship between genome-wide predictions of protein disorder and the RNA and protein expression in *E. coli*. Both measured RNA levels in *E. coli* cells grown in three different media and predicted levels of protein expression correlated positively with predicted fraction of disordered regions in *E. coli* proteins. This relationship did not depend specifically on the growth rate of the *E. coli* cultures and was not due to a small subset of genes/proteins with very high concordance values between compared metrics. In our consideration of highly expressed proteins predicted to have high fraction of disorder, the disorder predictions matched well with the experimentally elucidated regions of protein flexibility and disorder. It is tempting to speculate that there might be a direct link between protein disorder and protein level in *E. coli* cells, for example, because disordered proteins may carry out the essential control and regulation functions that are needed to respond to the various environmental conditions. Our analysis of essential *E. coli* proteins seems to support such hypothesis. However, another possibility is that disordered proteins might undergo more rapid protein degradation compared to structured proteins, which cells can counter by increasing mRNA levels of the corresponding genes. In this case, higher synthesis and degradation rates could make the levels of these proteins very sensitive to the environment, with slight changes in either production or degradation leading to significant shifts in protein levels. Further studies are needed to collect more experimental examples of disordered proteins, their functions, and their cellular expression levels to corroborate our findings.

Acknowledgements

We are thankful to Benjamin Withman for manuscript proofreading. This work was supported by Wright Brothers Institute grant WBSC9004A and by Wright State University grant 666760 to O.P., by Wright State University work-study scholarship to S.M.G, by the grants R01 LM007688-01A1 (to A.K.D.) and GM071714-01A2 (A.K.D and V.N.U.) from the National Institutes of Health, and by the Programs of the Russian Academy of Sciences for the "Molecular and cellular biology" and "Fundamental science for medicine" (to V.N.U.).

References

1. Dunker AK, Lawson JD, Brown CJ, Williams RM, Romero P, Oh JS, Oldfield CJ, Campen AM, Ratliff CM, Higgs KW, Ausio J, Nissen MS, Reeves R, Kang C, Kissinger CR, Bailey RW, Griswold MD, Chiu W, Garner EC, Obradovic Z. Intrinsically disordered protein. *J Mol Graph Model* 2001;19(1): 26–59. [PubMed: 11381529]
2. Lacy ER, Filippov I, Lewis WS, Otieno S, Xiao L, Weiss S, Hengst L, Kriwacki RW. p27 binds cyclin-CDK complexes through a sequential mechanism involving binding-induced protein folding. *Nat Struct Mol Biol* 2004;11(4):358–364. [PubMed: 15024385]
3. Spolar RS, Record MT Jr. Coupling of local folding to site-specific binding of proteins to DNA. *Science* 1994;263(5148):777–784. [PubMed: 8303294]
4. Ivanyi-Nagy R, Davidovic L, Khandjian EW, Darlix JL. Disordered RNA chaperone proteins: from functions to disease. *Cell Mol Life Sci* 2005;62(13):1409–1417. [PubMed: 15924259]
5. Tompa P, Csermely P. The role of structural disorder in the function of RNA and protein chaperones. *Faseb J* 2004;18(11):1169–1175. [PubMed: 15284216]
6. Liu J, Perumal NB, Oldfield CJ, Su EW, Uversky VN, Dunker AK. Intrinsic disorder in transcription factors. *Biochemistry* 2006;45(22):6873–6888. [PubMed: 16734424]
7. Minezaki Y, Homma K, Kinjo AR, Nishikawa K. Human transcription factors contain a high fraction of intrinsically disordered regions essential for transcriptional regulation. *J Mol Biol* 2006;359(4): 1137–1149. [PubMed: 16697407]
8. Iakoucheva LM, Brown CJ, Lawson JD, Obradovic Z, Dunker AK. Intrinsic disorder in cell-signaling and cancer-associated proteins. *J Mol Biol* 2002;323(3):573–584. [PubMed: 12381310]
9. Romero PR, Zaidi S, Fang YY, Uversky VN, Radivojac P, Oldfield CJ, Cortese MS, Sickmeier M, LeGall T, Obradovic Z, Dunker AK. Alternative splicing in concert with protein intrinsic disorder

enables increased functional diversity in multicellular organisms. *Proc Natl Acad Sci U S A* 2006;103(22):8390–8395. [PubMed: 16717195]

10. Magidovich E, Fleishman SJ, Yifrach O. Intrinsically disordered C-terminal segments of voltage-activated potassium channels: a possible fishing rod-like mechanism for channel binding to scaffold proteins. *Bioinformatics* 2006;22(13):1546–1550. [PubMed: 16601002]
11. Thomas N, Imafuku Y, Kamiya T, Tawada K. Kinesin: a molecular motor with a spring in its step. *Proc Biol Sci* 2002;269(1507):2363–2371. [PubMed: 12495505]
12. Romero, Obradovic, Dunker K. Sequence Data Analysis for Long Disordered Regions Prediction in the Calcineurin Family. *Genome Inform Ser Workshop Genome Inform* 1997;8:110–124.
13. Ferron F, Longhi S, Canard B, Karlin D. A practical overview of protein disorder prediction methods. *Proteins* 2006;65(1):1–14. [PubMed: 16856179]
14. Ward JJ, Sodhi JS, McGuffin LJ, Buxton BF, Jones DT. Prediction and functional analysis of native disorder in proteins from the three kingdoms of life. *J Mol Biol* 2004;337(3):635–645. [PubMed: 15019783]
15. Dunker AK, Obradovic Z, Romero P, Garner EC, Brown CJ. Intrinsic protein disorder in complete genomes. *Genome Inform Ser Workshop Genome Inform* 2000;11:161–171.
16. Xie H, Vucetic S, Iakoucheva LM, Oldfield CJ, Dunker AK, Obradovic Z, Uversky VN. Functional anthology of intrinsic disorder. 3. Ligands, post-translational modifications, and diseases associated with intrinsically disordered proteins. *J Proteome Res* 2007;6(5):1917–1932. [PubMed: 17391016]
17. Oldfield CJ, Cheng Y, Cortese MS, Brown CJ, Uversky VN, Dunker AK. Comparing and combining predictors of mostly disordered proteins. *Biochemistry* 2005;44(6):1989–2000. [PubMed: 15697224]
18. Xie H, Vucetic S, Iakoucheva LM, Oldfield CJ, Dunker AK, Uversky VN, Obradovic Z. Functional anthology of intrinsic disorder. 1. Biological processes and functions of proteins with long disordered regions. *J Proteome Res* 2007;6(5):1882–1898. [PubMed: 17391014]
19. Vucetic S, Xie H, Iakoucheva LM, Oldfield CJ, Dunker AK, Obradovic Z, Uversky VN. Functional anthology of intrinsic disorder. 2. Cellular components, domains, technical terms, developmental processes, and coding sequence diversities correlated with long disordered regions. *J Proteome Res* 2007;6(5):1899–1916. [PubMed: 17391015]
20. Gutnick D, Calvo JM, Klopotoski T, Ames BN. Compounds which serve as the sole source of carbon or nitrogen for *Salmonella typhimurium* LT-2. *J Bacteriol* 1969;100(1):215–219. [PubMed: 4898986]
21. Gyaneshwar P, Paliy O, McAuliffe J, Jones A, Jordan MI, Kustu S. Lessons from *Escherichia coli* genes similarly regulated in response to nitrogen and sulfur limitation. *Proc Natl Acad Sci U S A* 2005;102(9):3453–3458. [PubMed: 15716358]
22. Soupene E, van Heeswijk WC, Plumbridge J, Stewart V, Bertenthal D, Lee H, Prasad G, Paliy O, Charennoppakul P, Kustu S. Physiological studies of *Escherichia coli* strain MG1655: growth defects and apparent cross-regulation of gene expression. *J Bacteriol* 2003;185(18):5611–5626. [PubMed: 12949114]
23. Corbin RW, Paliy O, Yang F, Shabanowitz J, Platt M, Lyons CE Jr, Root K, McAuliffe J, Jordan MI, Kustu S, Soupene E, Hunt DF. Toward a protein profile of *Escherichia coli*: comparison to its transcription profile. *Proc Natl Acad Sci U S A* 2003;100(16):9232–9237. [PubMed: 12878731]
24. Uversky VN, Gillespie JR, Fink AL. Why are "natively unfolded" proteins unstructured under physiologic conditions? *Proteins* 2000;41(3):415–427. [PubMed: 11025552]
25. Li X, Romero P, Rani M, Dunker AK, Obradovic Z. Predicting Protein Disorder for N-, C-, and Internal Regions. *Genome Inform Ser Workshop Genome Inform* 1999;10:30–40.
26. Peng K, Vucetic S, Radivojac P, Brown CJ, Dunker AK, Obradovic Z. Optimizing long intrinsic disorder predictors with protein evolutionary information. *J Bioinform Comput Biol* 2005;3(1):35–60. [PubMed: 15751111]
27. Obradovic Z, Peng K, Vucetic S, Radivojac P, Brown CJ, Dunker AK. Predicting intrinsic disorder from amino acid sequence. *Proteins* 2003;53:566–572. [PubMed: 14579347]
28. Sharp PM, Li WH. The codon Adaptation Index--a measure of directional synonymous codon usage bias, and its potential applications. *Nucleic Acids Res* 1987;15(3):1281–1295. [PubMed: 3547335]
29. Sharp PM, Cowe E, Higgins DG, Shields DC, Wolfe KH, Wright F. Codon usage patterns in *Escherichia coli*, *Bacillus subtilis*, *Saccharomyces cerevisiae*, *Schizosaccharomyces pombe*,

- Drosophila melanogaster* and *Homo sapiens*; a review of the considerable within-species diversity. *Nucleic Acids Res* 1988;16(17):8207–8211. [PubMed: 3138659]
30. Carbone A, Zinovyev A, Kepes F. Codon adaptation index as a measure of dominating codon bias. *Bioinformatics* 2003;19(16):2005–2015. [PubMed: 14594704]
 31. Gyaneshwar P, Paliy O, McAuliffe J, Popham DL, Jordan MI, Kustu S. Sulfur and nitrogen limitation in *Escherichia coli* K-12: specific homeostatic responses. *J Bacteriol* 2005;187(3):1074–1090. [PubMed: 15659685]
 32. Tao H, Bausch C, Richmond C, Blattner FR, Conway T. Functional genomics: expression analysis of *Escherichia coli* growing on minimal and rich media. *J Bacteriol* 1999;181(20):6425–6440. [PubMed: 10515934]
 33. Karlin S, Barnett MJ, Campbell AM, Fisher RF, Mrazek J. Predicting gene expression levels from codon biases in alpha-proteobacterial genomes. *Proc Natl Acad Sci U S A* 2003;100(12):7313–7318. [PubMed: 12775761]
 34. Liu WM, Mei R, Di X, Ryder TB, Hubbell E, Dee S, Webster TA, Harrington CA, Ho MH, Baid J, Smekens SP. Analysis of high density expression microarrays with signed-rank call algorithms. *Bioinformatics* 2002;18(12):1593–1599. [PubMed: 12490443]
 35. Serres MH, Goswami S, Riley M. GenProtEC: an updated and improved analysis of functions of *Escherichia coli* K-12 proteins. *Nucleic Acids Res* 2004;32(Database issue):D300–D302. [PubMed: 14681418]
 36. Adler J, Templeton B. The effect of environmental conditions on the motility of *Escherichia coli*. *J Gen Microbiol* 1967;46(2):175–184. [PubMed: 4961758]
 37. Oldfield CJ, Cheng Y, Cortese MS, Romero P, Uversky VN, Dunker AK. Coupled folding and binding with alpha-helix-forming molecular recognition elements. *Biochemistry* 2005;44(37):12454–12470. [PubMed: 16156658]
 38. Goh KI, Oh E, Jeong H, Kahng B, Kim D. Classification of scale-free networks. *Proc Natl Acad Sci U S A* 2002;99(20):12583–12588. [PubMed: 12239345]
 39. Arifuzzaman M, Maeda M, Itoh A, Nishikata K, Takita C, Saito R, Ara T, Nakahigashi K, Huang HC, Hirai A, Tsuzuki K, Nakamura S, Altaf-Ul-Amin M, Oshima T, Baba T, Yamamoto N, Kawamura T, Ioka-Nakamichi T, Kitagawa M, Tomita M, Kanaya S, Wada C, Mori H. Large-scale identification of protein-protein interaction of *Escherichia coli* K-12. *Genome Res* 2006;16(5):686–691. [PubMed: 16606699]
 40. Butland G, Peregrin-Alvarez JM, Li J, Yang W, Yang X, Canadien V, Starostine A, Richards D, Beattie B, Krogan N, Davey M, Parkinson J, Greenblatt J, Emili A. Interaction network containing conserved and essential protein complexes in *Escherichia coli*. *Nature* 2005;433(7025):531–537. [PubMed: 15690043]
 41. Jeong H, Mason SP, Barabasi AL, Oltvai ZN. Lethality and centrality in protein networks. *Nature* 2001;411(6833):41–42. [PubMed: 11333967]
 42. Yellaboina S, Goyal K, Mande SC. Inferring genome-wide functional linkages in *E. coli* by combining improved genome context methods: comparison with high-throughput experimental data. *Genome Res* 2007;17(4):527–535. [PubMed: 17339371]
 43. Ekman D, Light S, Bjorklund AK, Elofsson A. What properties characterize the hub proteins of the protein-protein interaction network of *Saccharomyces cerevisiae*? *Genome Biol* 2006;7(6):R45. [PubMed: 16780599]
 44. Haynes C, Oldfield CJ, Ji F, Klitgord N, Cusick ME, Radivojac P, Uversky VN, Vidal M, Iakoucheva LM. Intrinsic disorder is a common feature of hub proteins from four eukaryotic interactomes. *PLoS Comput Biol* 2006;2(8):e100. [PubMed: 16884331]
 45. Dosztanyi Z, Chen J, Dunker AK, Simon I, Tompa P. Disorder and sequence repeats in hub proteins and their implications for network evolution. *J Proteome Res* 2006;5(11):2985–2995. [PubMed: 17081050]
 46. Dunker AK, Cortese MS, Romero P, Iakoucheva LM, Uversky VN. Flexible nets. The roles of intrinsic disorder in protein interaction networks. *Febs J* 2005;272(20):5129–5148. [PubMed: 16218947]
 47. Recht MI, Williamson JR. RNA tertiary structure and cooperative assembly of a large ribonucleoprotein complex. *J Mol Biol* 2004;344(2):395–407. [PubMed: 15522293]

48. Klein DJ, Moore PB, Steitz TA. The roles of ribosomal proteins in the structure assembly, and evolution of the large ribosomal subunit. *J Mol Biol* 2004;340(1):141–177. [PubMed: 15184028]
49. Uversky VN. Natively unfolded proteins: a point where biology waits for physics. *Protein Sci* 2002;11(4):739–756. [PubMed: 11910019]
50. Perham RN. Domains, motifs, and linkers in 2-oxo acid dehydrogenase multienzyme complexes: a paradigm in the design of a multifunctional protein. *Biochemistry* 1991;30(35):8501–8512. [PubMed: 1888719]
51. de Kok A, Hengeveld AF, Martin A, Westphal AH. The pyruvate dehydrogenase multi-enzyme complex from Gram-negative bacteria. *Biochim Biophys Acta* 1998;1385(2):353–366. [PubMed: 9655933]
52. Jones DD, Stott KM, Howard MJ, Perham RN. Restricted motion of the lipoyl-lysine swinging arm in the pyruvate dehydrogenase complex of *Escherichia coli*. *Biochemistry* 2000;39(29):8448–8459. [PubMed: 10913250]
53. Missiakas D, Betton JM, Raina S. New components of protein folding in extracytoplasmic compartments of *Escherichia coli* SurA, FkpA and Skp/OmpH. *Mol Microbiol* 1996;21(4):871–884. [PubMed: 8878048]
54. Harms N, Koningstein G, Dontje W, Muller M, Oudega B, Luirink J, de Cock H. The early interaction of the outer membrane protein phoe with the periplasmic chaperone Skp occurs at the cytoplasmic membrane. *J Biol Chem* 2001;276(22):18804–18811. [PubMed: 11278858]
55. Schafer U, Beck K, Muller M. Skp, a molecular chaperone of gram-negative bacteria, is required for the formation of soluble periplasmic intermediates of outer membrane proteins. *J Biol Chem* 1999;274(35):24567–24574. [PubMed: 10455120]
56. Walton TA, Sousa MC. Crystal structure of Skp, a prefoldin-like chaperone that protects soluble and membrane proteins from aggregation. *Mol Cell* 2004;15(3):367–374. [PubMed: 15304217]
57. Romero P, Obradovic Z, Dunker AK. Folding minimal sequences: the lower bound for sequence complexity of globular proteins. *FEBS Lett* 1999;462(3):363–367. [PubMed: 10622726]
58. Reed LJ, Hackert ML. Structure-function relationships in dihydrolipoamide acyltransferases. *J Biol Chem* 1990;265(16):8971–8974. [PubMed: 2188967]
59. Green JD, Perham RN, Ullrich SJ, Appella E. Conformational studies of the interdomain linker peptides in the dihydrolipoyl acetyltransferase component of the pyruvate dehydrogenase multienzyme complex of *Escherichia coli*. *J Biol Chem* 1992;267(33):23484–23488. [PubMed: 1429691]
60. Turner SL, Russell GC, Williamson MP, Guest JR. Restructuring an interdomain linker in the dihydrolipoamide acetyltransferase component of the pyruvate dehydrogenase complex of *Escherichia coli*. *Protein Eng* 1993;6(1):101–108. [PubMed: 8433963]
61. Ricaud PM, Howard MJ, Roberts EL, Broadhurst RW, Perham RN. Three-dimensional structure of the lipoyl domain from the dihydrolipoyl succinyltransferase component of the 2-oxoglutarate dehydrogenase multienzyme complex of *Escherichia coli*. *J Mol Biol* 1996;264(1):179–190. [PubMed: 8950276]
62. Ray MC, Germon P, Vianney A, Portalier R, Lazzaroni JC. Identification by genetic suppression of *Escherichia coli* TolB residues important for TolB-Pal interaction. *J Bacteriol* 2000;182(3):821–824. [PubMed: 10633120]
63. Lloubes R, Cascales E, Walburger A, Bouveret E, Lazdunski C, Bernadac A, Journet L. The Tol-Pal proteins of the *Escherichia coli* cell envelope: an energized system required for outer membrane integrity? *Res Microbiol* 2001;152(6):523–529. [PubMed: 11501670]
64. Parsons LM, Lin F, Orban J. Peptidoglycan recognition by Pal, an outer membrane lipoprotein. *Biochemistry* 2006;45(7):2122–2128. [PubMed: 16475801]
65. Pon CL, Wittmann-Liebold B, Gualerzi C. Structure--function relationships in *Escherichia coli* initiation factors. II. Elucidation of the primary structure of initiation factor IF-1. *FEBS Lett* 1979;101(1):157–160. [PubMed: 376343]
66. Grunberg-Manago M, Dessen P, Pantaloni D, Godefroy-Colburn T, Wolfe AD, Dondon J. Light-scattering studies showing the effect of initiation factors on the reversible dissociation of *Escherichia coli* ribosomes. *J Mol Biol* 1975;94(3):461–478. [PubMed: 1100842]

67. Pon CL, Gualerzi CO. Mechanism of protein biosynthesis in prokaryotic cells. Effect of initiation factor IF1 on the initial rate of 30 S initiation complex formation. *FEBS Lett* 1984;175(2):203–207. [PubMed: 6383865]
68. Stringer EA, Sarkar P, Maitra U. Function of initiation factor 1 in the binding and release of initiation factor 2 from ribosomal initiation complexes in *Escherichia coli*. *J Biol Chem* 1977;252(5):1739–1744. [PubMed: 320213]
69. Sette M, van Tilborg P, Spurio R, Kaptein R, Paci M, Gualerzi CO, Boelens R. The structure of the translational initiation factor IF1 from *E.coli* contains an oligomer-binding motif. *Embo J* 1997;16(6):1436–1443. [PubMed: 9135158]
70. Lambalot RH, Walsh CT. Holo-[acyl-carrier-protein] synthase of *Escherichia coli*. *Methods Enzymol* 1997;279:254–262. [PubMed: 9211277]
71. Schulz H. Increased conformational stability of *Escherichia coli* acyl carrier protein in the presence of divalent cations. *FEBS Lett* 1977;78(2):303–306. [PubMed: 328303]
72. Schulz H. On the structure-function relationship of acyl carrier protein of *Escherichia coli*. *J Biol Chem* 1975;250(6):2299–2304. [PubMed: 234965]
73. Qiu X, Janson CA. Structure of apo acyl carrier protein and a proposal to engineer protein crystallization through metal ions. *Acta Crystallogr D Biol Crystallogr* 2004;60(Pt 9):1545–1554. [PubMed: 15333924]
74. Esposito D, Petrovic A, Harris R, Ono S, Eccleston JF, Mbabaali A, Haq I, Higgins CF, Hinton JC, Driscoll PC, Ladbury JE. H-NS oligomerization domain structure reveals the mechanism for high order self-association of the intact protein. *J Mol Biol* 2002;324(4):841–850. [PubMed: 12460581]
75. Drlica K, Rouviere-Yaniv J. Histone-like proteins of bacteria. *Microbiol Rev* 1987;51(3):301–319. [PubMed: 3118156]
76. Higgins CF, Dorman CJ, Stirling DA, Waddell L, Booth IR, May G, Bremer E. A physiological role for DNA supercoiling in the osmotic regulation of gene expression in *S. typhimurium* and *E. coli*. *Cell* 1988;52(4):569–584. [PubMed: 2830029]
77. Hulton CS, Seirafi A, Hinton JC, Sidebotham JM, Waddell L, Pavitt GD, Owen-Hughes T, Spassky A, Buc H, Higgins CF. Histone-like protein H1 (H-NS), DNA supercoiling, and gene expression in bacteria. *Cell* 1990;63(3):631–642. [PubMed: 2171779]
78. Smyth CP, Lundback T, Renzoni D, Siligardi G, Beavil R, Layton M, Sidebotham JM, Hinton JC, Driscoll PC, Ladbury JE. Oligomerization of the chromatin-structuring protein H-NS. *Mol Microbiol* 2000;36(4):962–972. [PubMed: 10844682]
79. LiCalsi C, Crocenzi TS, Freire E, Roseman S. Sugar transport by the bacterial phosphotransferase system. Structural and thermodynamic domains of enzyme I of *Salmonella typhimurium*. *J Biol Chem* 1991;266(29):19519–19527. [PubMed: 1655788]
80. Alpert CA, Frank R, Stuber K, Deutscher J, Hengstenberg W. Phosphoenolpyruvate-dependent protein kinase enzyme I of *Streptococcus faecalis*: purification and properties of the enzyme and characterization of its active center. *Biochemistry* 1985;24(4):959–964. [PubMed: 3922407]
81. Weigel N, Waygood EB, Kukuruzinska MA, Nakazawa A, Roseman S. Sugar transport by the bacterial phosphotransferase system. Isolation and characterization of enzyme I from *Salmonella typhimurium*. *J Biol Chem* 1982;257(23):14461–14469. [PubMed: 6754728]
82. Seok YJ, Lee BR, Zhu PP, Peterkofsky A. Importance of the carboxyl-terminal domain of enzyme I of the *Escherichia coli* phosphoenolpyruvate: sugar phosphotransferase system for phosphoryl donor specificity. *Proc Natl Acad Sci U S A* 1996;93(1):347–351. [PubMed: 8552636]
83. Zhu PP, Szczepanowski RH, Nosworthy NJ, Ginsburg A, Peterkofsky A. Reconstitution studies using the helical and carboxy-terminal domains of enzyme I of the phosphoenolpyruvate:sugar phosphotransferase system. *Biochemistry* 1999;38(47):15470–15479. [PubMed: 10569929]
84. Teplyakov A, Lim K, Zhu PP, Kapadia G, Chen CC, Schwartz J, Howard A, Reddy PT, Peterkofsky A, Herzberg O. Structure of phosphorylated enzyme I, the phosphoenolpyruvate:sugar phosphotransferase system sugar translocation signal protein. *Proc Natl Acad Sci U S A* 2006;103(44):16218–16223. [PubMed: 17053069]
85. Guilhaudis L, Simorre JP, Blackledge M, Neuburger M, Bourguignon J, Douce R, Marion D, Gans P. Investigation of the local structure and dynamics of the H subunit of the mitochondrial glycine

- decarboxylase using heteronuclear NMR spectroscopy. *Biochemistry* 1999;38(26):8334–8346. [PubMed: 10387079]
86. Guilhaudis L, Simorre JP, Bouchayer E, Neuburger M, Bourguignon J, Douce R, Marion D, Gans P. Backbone and sequence-specific assignment of three forms of the lipoate-containing H-protein of the glycine decarboxylase complex. *J Biomol NMR* 1999;15(2):185–186. [PubMed: 10605092]
 87. Cramer P. Multisubunit RNA polymerases. *Curr Opin Struct Biol* 2002;12(1):89–97. [PubMed: 11839495]
 88. Liu K, Zhang Y, Severinov K, Das A, Hanna MM. Role of *Escherichia coli* RNA polymerase alpha subunit in modulation of pausing, termination and anti-termination by the transcription elongation factor NusA. *Embo J* 1996;15(1):150–161. [PubMed: 8598198]
 89. Worbs M, Bourenkov GP, Bartunik HD, Huber R, Wahl MC. An extended RNA binding surface through arrayed S1 and KH domains in transcription factor NusA. *Mol Cell* 2001;7(6):1177–1189. [PubMed: 11430821]
 90. Shin DH, Nguyen HH, Jancarik J, Yokota H, Kim R, Kim SH. Crystal structure of NusA from *Thermotoga maritima* and functional implication of the N-terminal domain. *Biochemistry* 2003;42(46):13429–13437. [PubMed: 14621988]
 91. Gopal B, Haire LF, Gamblin SJ, Dodson EJ, Lane AN, Papavinasasundaram KG, Colston MJ, Dodson G. Crystal structure of the transcription elongation/anti-termination factor NusA from *Mycobacterium tuberculosis* at 1.7 Å resolution. *J Mol Biol* 2001;314(5):1087–1095. [PubMed: 11743725]
 92. Mah TF, Kuznedelov K, Mushegian A, Severinov K, Greenblatt J. The alpha subunit of *E. coli* RNA polymerase activates RNA binding by NusA. *Genes Dev* 2000;14(20):2664–2675. [PubMed: 11040219]
 93. Eisenmann A, Schwarz S, Prasch S, Schweimer K, Rosch P. The *E. coli* NusA carboxy-terminal domains are structurally similar and show specific RNAP- and lambdaN interaction. *Protein Sci* 2005;14(8):2018–2029. [PubMed: 15987884]
 94. Mah TF, Li J, Davidson AR, Greenblatt J. Functional importance of regions in *Escherichia coli* elongation factor NusA that interact with RNA polymerase, the bacteriophage lambda N protein and RNA. *Mol Microbiol* 1999;34(3):523–537. [PubMed: 10564494]
 95. Roberts EL, Shu N, Howard MJ, Broadhurst RW, Chapman-Smith A, Wallace JC, Morris T, Cronan JE Jr, Perham RN. Solution structures of apo and holo biotinyl domains from acetyl coenzyme A carboxylase of *Escherichia coli* determined by triple-resonance nuclear magnetic resonance spectroscopy. *Biochemistry* 1999;38(16):5045–5053. [PubMed: 10213607]
 96. Athappilly FK, Hendrickson WA. Structure of the biotinyl domain of acetyl-coenzyme A carboxylase determined by MAD phasing. *Structure* 1995;3(12):1407–1419. [PubMed: 8747466]
 97. Cronan JE Jr. Interchangeable enzyme modules. Functional replacement of the essential linker of the biotinylated subunit of acetyl-CoA carboxylase with a linker from the lipoylated subunit of pyruvate dehydrogenase. *J Biol Chem* 2002;277(25):22520–22527. [PubMed: 11956202]
 98. Choi-Rhee E, Cronan JE. The biotin carboxylase-biotin carboxyl carrier protein complex of *Escherichia coli* acetyl-CoA carboxylase. *J Biol Chem* 2003;278(33):30806–30812. [PubMed: 12794081]
 99. Fall RR, Vagelos PR. Acetyl coenzyme A carboxylase. Molecular forms and subunit composition of biotin carboxyl carrier protein. *J Biol Chem* 1972;247(24):8005–8015. [PubMed: 4565671]
 100. Fall RR, Nervi AM, Alberts AW, Vagelos PR. Acetyl CoA carboxylase: isolation and characterization of native biotin carboxyl carrier protein. *Proc Natl Acad Sci U S A* 1971;68(7):1512–1515. [PubMed: 4934522]
 101. Mukherjee K, Chatterji D. Studies on the omega subunit of *Escherichia coli* RNA polymerase—its role in the recovery of denatured enzyme activity. *Eur J Biochem* 1997;247(3):884–889. [PubMed: 9288911]
 102. Ghosh P, Ishihama A, Chatterji D. *Escherichia coli* RNA polymerase subunit omega and its N-terminal domain bind full-length beta' to facilitate incorporation into the alpha2beta subassembly. *Eur J Biochem* 2001;268(17):4621–4627. [PubMed: 11531998]

103. Dove SL, Hochschild A. Conversion of the omega subunit of *Escherichia coli* RNA polymerase into a transcriptional activator or an activation target. *Genes Dev* 1998;12(5):745–754. [PubMed: 9499408]
104. Altendorf K, Stalz W, Greie J, Deckers-Hebestreit G. Structure and function of the F(o) complex of the ATP synthase from *Escherichia coli*. *J Exp Biol* 2000;203(Pt 1):19–28. [PubMed: 10600669]
105. Gogol EP, Lucken U, Capaldi RA. The stalk connecting the F1 and F0 domains of ATP synthase visualized by electron microscopy of unstained specimens. *FEBS Lett* 1987;219(2):274–278. [PubMed: 2886365]
106. Wilkens S, Capaldi RA. ATP synthase's second stalk comes into focus. *Nature* 1998;393(6680):29. [PubMed: 9590688]
107. Wilkens S, Capaldi RA. Solution structure of the epsilon subunit of the F1-ATPase from *Escherichia coli* and interactions of this subunit with beta subunits in the complex. *J Biol Chem* 1998;273(41):26645–26651. [PubMed: 9756905]
108. Del Rizzo PA, Bi Y, Dunn SD, Shilton BH. The "second stalk" of *Escherichia coli* ATP synthase: structure of the isolated dimerization domain. *Biochemistry* 2002;41(21):6875–6884. [PubMed: 12022893]
109. Abrahams JP, Leslie AG, Lutter R, Walker JE. Structure at 2.8 Å resolution of F1-ATPase from bovine heart mitochondria. *Nature* 1994;370(6491):621–628. [PubMed: 8065448]
110. Bianchet MA, Hulihan J, Pedersen PL, Amzel LM. The 2.8-Å structure of rat liver F1-ATPase: configuration of a critical intermediate in ATP synthesis/hydrolysis. *Proc Natl Acad Sci U S A* 1998;95(19):11065–11070. [PubMed: 9736690]
111. Pasmán Z, von Hippel PH. Regulation of rho-dependent transcription termination by NusG is specific to the *Escherichia coli* elongation complex. *Biochemistry* 2000;39(18):5573–5585. [PubMed: 10820031]
112. Steiner T, Kaiser JT, Marinkovic S, Huber R, Wahl MC. Crystal structures of transcription factor NusG in light of its nucleic acid- and protein-binding activities. *Embo J* 2002;21(17):4641–4653. [PubMed: 12198166]
113. Braig K. Chaperonins. *Curr Opin Struct Biol* 1998;8(2):159–165. [PubMed: 9631288]
114. Wang JD, Weissman JS. Thinking outside the box: new insights into the mechanism of GroEL-mediated protein folding. *Nat Struct Biol* 1999;6(7):597–600. [PubMed: 10404205]
115. Grallert H, Buchner J. Review: a structural view of the GroE chaperone cycle. *J Struct Biol* 2001;135(2):95–103. [PubMed: 11580259]
116. Todd MJ, Viitanen PV, Lorimer GH. Dynamics of the chaperonin ATPase cycle: implications for facilitated protein folding. *Science* 1994;265(5172):659–666. [PubMed: 7913555]
117. Braig K, Otwinowski Z, Hegde R, Boisvert DC, Joachimiak A, Horwich AL, Sigler PB. The crystal structure of the bacterial chaperonin GroEL at 2.8 Å. *Nature* 1994;371(6498):578–586. [PubMed: 7935790]
118. Landry SJ, Zeilstra-Ryalls J, Fayet O, Georgopoulos C, Gierasch LM. Characterization of a functionally important mobile domain of GroES. *Nature* 1993;364(6434):255–258. [PubMed: 8100614]
119. Shewmaker F, Maskos K, Simmerling C, Landry SJ. The disordered mobile loop of GroES folds into a defined beta-hairpin upon binding GroEL. *J Biol Chem* 2001;276(33):31257–31264. [PubMed: 11395498]
120. Mohan A, Oldfield CJ, Radivojac P, Vacic V, Cortese MS, Dunker AK, Uversky VN. Analysis of molecular recognition features (MoRFs). *J Mol Biol* 2006;362(5):1043–1059. [PubMed: 16935303]
121. Vacic V, Oldfield CJ, Mohan A, Radivojac P, Cortese MS, Uversky VN, Dunker AK. Characterization of molecular recognition features, MoRFs, and their binding partners. *J Proteome Res* 2007;6(6):2351–2366. [PubMed: 17488107]
122. Anfinsen CB. Principles that govern the folding of protein chains. *Science* 1973;181(96):223–230. [PubMed: 4124164]
123. Wright PE, Dyson HJ. Intrinsically unstructured proteins: re-assessing the protein structure-function paradigm. *J Mol Biol* 1999;293(2):321–331. [PubMed: 10550212]
124. Dyson HJ, Wright PE. Coupling of folding and binding for unstructured proteins. *Curr Opin Struct Biol* 2002;12(1):54–60. [PubMed: 11839490]

***E. coli* proteins sorted by the fraction of amino acids in their disordered regions (classes, lowest to highest)**

Figure 1. Measured RNA expression and predicted protein abundance as a function of predicted fraction of protein disorder in *E. coli*

To get the X axis, all *E. coli* proteins for which disorder prediction was available (4234 proteins total) were first sorted from lowest to highest fraction of protein disorder (FD). They were then divided into 10 equal classes (423 each). Class 1 contained the 10% of proteins with the smallest FD values and so on. The average RNA expression levels for all genes in each class were plotted as columns using the left Y axis scale for three different media: LB (black columns), MM1 (light grey columns), and MM2 (dark grey columns). Error bars show the standard error of the mean. The average CAI values, which represent codon usage bias for each gene, are plotted as filled diamonds using the right Y axis scale. The ranges of FD values for classes 1–10 were as follows: (1) 0.00–0.07, (2) 0.07–0.11, (3) 0.11–0.15, (4) 0.15–0.18, (5) 0.18–0.22, (6) 0.22–0.25, (7) 0.25–0.30, (8) 0.30–0.35, (9) 0.35–0.42, (10) 0.42–1.00.

Figure 2. PONDRL VL-XT (red curves), VSL1 (green curves) and VL3 (blue curves) predictions for 18 highly expressed *E. coli* proteins with high content of predicted disorder

In each panel the X-axis represents the position of amino acids in a protein, while the Y-axis represents the disorder propensity for a given residue. Residues with scores greater than 0.5 are predicted to be disordered and those with scores less than 0.5 are predicted to be ordered. Gray shaded areas correspond to the disordered regions discussed in the text. **A)** AceB (b0115); **B)** HlpA (b0178); **C)** SucB (b0727); **D)** Pal (b0741); **E)** InfA (b0884); **F)** AcpP (b1094); **G)** HNS (b1237); **H)** PtsI (b2416); **I)** GcvH (b2904); **J)** NusA (b3169); **K)** AccB (b3255); **L)** RpoZ (b3649); **M)** AtpC (b3731); **N)** AtpD (b3722); **O)** AtpF (b3736); **P)** NusG (b3982); **Q)** GroES (b4142); **R)** GroEL (b4143).

Table 1

Global properties of compared datasets

Dataset	Arithmetic Mean ^f
Fraction of disordered amino acids ^a	0.24 [*]
RNA expression level (LB ^b)	3700 ^{**}
RNA expression level (MM1 ^c)	2400 ^{**}
RNA expression level (MM2 ^d)	2100 ^{**}
CAI ^e	0.31 [*]

^a designated as FD in text^b LB — Luria-Bertani rich medium^c MM1 — glycerol-ammonium based N⁻C⁻ minimal medium^d MM2 — glycerol-arginine based N⁻C⁻ minimal medium^e CAI — codon adaptation index^f represent “per gene/protein” average of each dataset for all *E. coli* genes or proteins^{*} numbers are rounded to the second decimal place^{**} numbers are rounded to the nearest hundred

Table 2

Global relationship between fraction of protein disorder and gene expression and predicted protein abundance in *E. coli*

Comparison	Spearman correlation, R_s	Kendall Tau correlation, R_τ
FD ^a — gene expression (LB ^b)	0.234 [*]	0.160 [*]
FD — gene expression (MM1 ^c)	0.304 [*]	0.209 [*]
FD — gene expression (MM2 ^d)	0.298 [*]	0.205 [*]
FD — CAI ^e	0.203 [*]	0.140 [*]

^aFD — fraction of protein disorder

^bLB — Luria-Bertani rich medium

^cMM1 — glycerol-ammonium based N[−]C[−] minimal medium

^dMM2 — glycerol-arginine based N[−]C[−] minimal medium

^eCAI — codon adaptation index

^{*} results are statistically significant with $p < 10^{-5}$

Table 3
List of highly expressed disordered proteins in *E. coli*^a

B number	Gene name	FD ^{b,*}	Gene expression (LB) ^{c,**}	Gene expression (MM1) ^{d,**}	Gene expression (MM2) ^{e,**}	CAI ^{f,*}
b0115	aceF	0.48	9500	5500	4400	0.61
b0178	hlpA	0.42	11800	18000	20700	0.61
b0440	hupB	0.42	5100	7700	9600	0.53
b0727	sucB	0.57	62500	28200	17000	0.55
b0741	pal	0.52	15600	14300	11900	0.66
b0884	infA	0.50	3800	7300	5600	0.45
b1094	acpP	0.74	59600	43100	32000	0.64
b1237	hns	0.67	43500	42600	38300	0.55
b1641	slyB	0.45	11100	13100	15200	0.47
b2007	yeeX	0.49	6500	7100	11100	0.49
b2416	ptsI	0.42	15000	12100	17600	0.46
b2904	gcvH	0.53	37300	5400	5500	0.52
b3169	musA	0.47	12700	12100	6300	0.52
b3255	aceB	0.57	17500	15400	11100	0.52
b3649	rpoZ	0.77	4800	6300	6300	0.57
b3731	atpC	0.50	4100	4500	5800	0.46
b3732	atpD	0.46	4800	4800	4900	0.63
b3736	atpF	0.65	23000	22800	13900	0.46
b3982	musG	0.49	3800	7100	7000	0.55
b4142	groS	0.43	110800	26100	14700	0.49
b4143	groL	0.46	24500	8800	6300	0.78

^aThe list comprises all genes/proteins that had top 10% values in each of following datasets: gene expression in each of three media profiled, CAI index, FD. The list contained a total of 51 proteins, with 30 of them being ribosomal proteins. The non-ribosomal proteins are listed in the table. Average genome values for each metric are shown in Table 1.

^bFD — fraction of protein disorder

^cLB — Luria-Bertani rich medium

^dMM1 — glycerol-ammonium based N⁺C⁻ minimal medium

^eMM2 — glycerol-arginine based N⁺C⁻ minimal medium

f_{CAI} — codon adaptation index

* numbers were rounded to the second decimal place

** numbers were rounded to the nearest hundred