Can Pathogenic and Nonpathogenic Bacteria Be Distinguished by Sensory Protein Abundance?

Subhrajit Bhar,^a Tungadri Bose,^a Sharmila S. Mande^a

^aBio-Sciences R&D Division, TCS Innovation Labs, Tata Consultancy Services Limited, Pune, India

ABSTRACT Signal transduction systems are essential for microorganisms to respond to their ever-changing environment. They can be distinguished into one-component systems, two-component systems, and extracytoplasmic-function \( \sigma \) factors. Abundances of a few signal-transducing proteins, termed herein as sensory proteins (SPs), have previously been reported to be correlated with the genome size and ecological niche of certain Gram-positive bacteria. No such reports are available for Gram-negative bacteria. The current study attempts to investigate the relationship of the abundances of SPs to genome size in \textit{Escherichia coli}, and the bacterial pathotypes or phylotypes. While the relationship between SP abundance and genome size could not be established, the sensory protein index (SPI), a new metric defined herein, was found to be correlated with \textit{E. coli} virulence. In addition, significant association was observed among the distribution of SPs and \textit{E. coli} pathotypes. Results indicate that such associations might be due to genomic rearrangements to best utilize the resources available in a given ecological niche. Overall, the study provides an in-depth analysis of the occurrence of different SPs among pathogenic and nonpathogenic \textit{E. coli} strains. Possibilities of using the SPI as a marker for identifying pathogenic strains from among an organism complex are also discussed.

IMPORTANCE Sensory proteins (SPs) act as sensors and actuators for a cell and participate in important mechanisms pertaining to bacterial survival, adaptation, and virulence. Therefore, bacterial species residing in similar ecological niches or those sharing common pathotypes are expected to exhibit similar SP signatures. We have investigated profiles of SPs in different species of \textit{Escherichia coli} and present in this article the sensory protein index (SPI), a metric for quantifying the abundance and/or distribution of SPs across bacterial genomes, which could indicate the virulence potency of a bacterium. The SPI could find use in characterizing uncultured strains and bacterial complexes, as a biomarker for disease diagnostics, evaluating the effect of therapeutic interventions, assessing effects of ecological alterations, etc. Grouping the studied strains of \textit{E. coli} on the basis of the frequency of occurrence of SPs in their genomes could potentially replicate the stratification of these strains on the basis of their phylotypes. In addition, \textit{E. coli} strains belonging to the same pathotypes were also seen to share similar SP signatures. Furthermore, the SPI was seen to be an indicator of pathogenic potency of \textit{E. coli} strains. The SPI metric is expected to be useful in the (pathogenic) characterization of hereto uncultured strains which are routinely sequenced in host microbiome analysis projects, or from among an ensemble of microbial organisms constituting a biospecimen. Thus, the possibilities of using the SPI as a biomarker for diagnosis of a disease or the outcome of a therapeutic intervention cannot be ruled out. Further, SPIs obtained from longitudinal ecological samples have the potential to serve as key indicators of environmental changes. Such changes in the environment are often detrimental to the resident biome and methods for timely detection of environmental changes hold huge socioeconomic benefits.
Signal transduction systems are essential mechanisms through which a bacterium acclimatizes with its ever-changing environment (1–3; https://www.mechanobio.info/pathogenesis/what-are-two-component-regulatory-systems/). These systems in bacteria can be broadly classified into three major categories, namely, (i) one-component systems (OCSs), (ii) two-component systems (TCSs), and (iii) extracytoplasmic-function $\sigma$ factors (ECFs) (3). OCSs consist of a single protein with an “input” domain for sensing and an “output” domain for eliciting response (4). The response is usually mediated through nucleic acid binding, protein modification, or enzymatic reactions (4). TCSs are typically composed of a sensory kinase and a response regulator which function through a multistep process. The sensory kinase (mostly histidine kinases) first undergoes autophosphorylation on receiving environmental signals. The signal is then transmitted to the response regulators through phosphorylation of an invariant aspartate residue. Subsequently, the response regulators induce a suitable response through the expression of effector genes (5, 6). In contrast, ECFs are components of the RNA polymerase holoenzyme that determines promoter specificity. ECF activity is regulated through membrane-associated anti-$\sigma$ factors (ASFs). Upon receiving specific stimuli, the ASFs are inactivated and ECF recruits the RNA polymerase core enzyme, thereby allowing the transcription initiation from alternative (ECF-specific) target promoters (3).

Given the crucial role of signal-transducing proteins (7) (henceforth referred to as “sensory proteins” or SPs) for the survival of bacteria, we anticipated a possible association between phenotype (or ecological preference) and abundance of SPs among closely related organisms. A similar idea had previously been echoed in a study involving different strains of *Escherichia coli*, wherein the authors concluded that genome recombination frequency was correlated with the preference for environmental niche and pathogenic ferocity of *E. coli* (8). Related to this, another study reported a linear correlation between the genome sizes and the number of SPs encoded in the genome of different Actinobacteria species (3). However, no correlations between genome sizes and/or SP abundance with respect to phenotypes (pathotype/phylotype) or preferred ecological niche have been reported. Considering that the genomic reorganization events are known to be more pronounced among certain groups of bacteria belonging to the same phylum, we speculated that probing the association between genome sizes and SP abundances at a genus (or family) level (instead of phylum level, as reported in earlier literature) could provide more information. Consequently, we investigated the pattern of distribution of SPs among different *E. coli* strains, with the following objectives: (i) investigate distribution of SPs among the genomes of closely related organisms and (ii) explore any existing correlation among the distribution of SPs and the pathogenicity and/or ecological preference of the studied strains. The choice of *E. coli* as the model organism for the study was based on the following. *E. coli* as a species comprises pathogenic as well as nonpathogenic (also commensal) strains of bacteria. Further, different strains of *E. coli* are known to be inhabitants of diverse ecological niches and are therefore exposed to various levels of stress (9). Since SPs have roles in mitigating environmental stress as well as in exhibiting pathogenicity (3, 7), it was assumed that *E. coli* strains residing in different environmental niches and/or having various virulence profiles would harbor differences in their SP repertoire, and would therefore provide an ideal model for this study.

In the current study, data from 62 fully sequenced strains of *E. coli* were inspected to investigate possible associations between the abundance of SPs in these strains and their phenotypic properties. Several phylotypes and pathotypes of *E. coli* were seen to bear distinct SP signatures. In addition, abundances of SPs across the genomes of different *E. coli* strains were seen to be correlated with the pathogenic potential of the strains. Overall, this study evaluates the significance of SPs for classifying *E. coli* strains...
on the basis of their phenotype. The potential of using SPs for stratifying uncultured/uncharacterized strains of bacteria is also discussed.

RESULTS

Sensory proteins are conserved among E. coli pathotypes. SPs act as sensors and actuators for the bacterial cell. Upon sensing a physiological condition, the bacterial cells actuate their battery of secretion systems, toxins, etc., to optimally utilize available nutrients from the environment, as well as to cope with adversity in their surroundings. Consequently, it was speculated that the occurrence of SPs might be associated with the pathological phenotype of the studied E. coli strains (3, 8). To understand the relationship between SPs and E. coli pathotypes (Table S1 in the supplemental material), the SP data for the studied strains were plotted as a heatmap and compared for their specific occurrence patterns across different pathotypes (Fig. 1). The E. coli pathotype data were obtained from literature (8, 10, 11). Overall, the SP distribution among the E. coli strains seemed to correlate with their pathotypes. In other words, different pathotypes of E. coli (e.g., nonpathogenic/commensal, enteroaggregative E. coli [EAggEC], extraintestinal pathogenic E. coli [ExPEC]/uropathogenic E. coli [UPEC], Shiga toxin producing E. coli [STEC], etc.) appeared to possess distinct SP signatures.

For example, the genes for copper/silver efflux (cusS, pcoS) were found to be unique to the EAggEC pathotype. Most of the other studied strains of E. coli lacked cusS and pcoS genes. Similarly, SPs belonging to the D3QK97 family were found to be exclusive to certain STEC/enterohemorrhagic E. coli (EHEC) and enteropathogenic E. coli (EPEC) groups. Further, the gene for AtoS was seen to be present in all of the studied ExPEC/UPEC, adherent-invasive E. coli (AIEC), and avian pathogenic E. coli (APEC) groups, as well as in some of the EPEC and enterotoxigenic E. coli (ETEC) groups. In contrast, the presence of pgtB was not found to be associated with any pathotype. The pgtB gene was found to be carried in the genomes of strains 536, CFT073, S88 (ExPEC/UPEC), and APEC O1 among the studied E. coli strains. Additionally, although most of the ExPEC/UPEC and AIEC pathotypes seemed to lack the yqeI, tap, and trg genes, they were found to encode DctB.

Almost all the nonpathogenic strains lacked the protein-coding genes for PgtB, CusS, PcoS, DctB, and D3Q97 domain-containing proteins in their genome. Further, two distinct subclasses were observed in the nonpathogens. The first subclass comprised strains KO11FL, SE11, SE15, W, and HS. Most of them lacked the genes corresponding to DctB, AtoS and FecR. The second subclass comprised strains K-12, DH1, and BL21 and lacked an SP belonging to the DUF3816 family.

The above findings showed the presence of distinct pathotype-specific SPs in various strains of E. coli. This indicated the potential relevance of SPs in determining pathogenic bacteria which could in turn be of importance in designing novel microbiocidal strategies against these pathogens.

Distribution of sensory proteins among E. coli phylogroups. In order to understand the relationship between the phylogroups of the studied 62 E. coli strains and the 60 SP classes considered in this study (Table S2), the E. coli strains were first clustered into five groups based on their SP distribution (discussed in the Materials and Methods section). The cluster count of five was chosen to match the number of phylogroups which represented all the studied strains of E. coli (8, 10, 11). The overlap between the members of the five clusters obtained through hierarchical clustering of the E. coli strains based on their SP distributions and the five E. coli phylogroups was evident from the analysis (Table 1). SP signatures for each of the studied phylogroups have been provided in Fig. S1. The results seemed to indicate that strains belonging to a given phylogroup also contained a similar distribution of SPs (class). In other words, each phylogroup was characterized (or dominated) by a specific set of SPs.

All members of phylogroups E and B2 mapped to the members of cluster 1 and 2, respectively. Notably, all members of phylogroup E were STEC/EHEC and EPEC pathogens. Similarly, members of phylogroup B2, all of which clustered together based on their SP distribution, were seen to be dominated by ExPEC/UPEC pathogens. However,
FIG 1 Depiction of the sensory protein signatures for the studied strains of *Escherichia coli*.
no differences were noted for the other phylogroups. All studied members of phylogroups B1 and D and most members of phylogroup A (14 out of 16) mapped to cluster 3. Interestingly, while phylogroup A mostly consisted of commensal and/or nonpathogenic strains, phylogroups B1 and D majorly consisted of STEC/EHEC and ExPEC/UPEC pathogens. The remaining members of phylogroup A, namely, P12b and MDS42, mapped to clusters 4 and 5, respectively. Therefore, like the distribution of SPs by *E. coli* pathotype, the distribution of SPs in *E. coli* was also found to distinguish the phylogroups. This highlighted the importance of the SPs in adapting to specialized ecological niches.

*E. coli* strains may have evolved through acquisition of sensory proteins to adapt to niche ecologies. In order to adapt to a given environmental niche or to attain certain phenotypes, bacterial genomes often undergo transformation through events like horizontal gene transfer, gene duplication, and gene deletion (12). To investigate the possibility or probability of the occurrence of such transformation events (with respect to SPs) in *E. coli*, the studied strains were clustered on the basis of (i) 16S rRNA gene sequences and (ii) distribution of SPs among these organisms (see the Materials and Methods). Next, a comparative analysis of the two dendrograms was performed to identify possible events of genome modification postspeciation (Fig. 2). The 16S rRNA

### TABLE 1 Analysis of the overlap between clusters of *Escherichia coli* obtained through hierarchical clustering based on distribution of sensory protein genes and the known *E. coli* phylogroups

<table>
<thead>
<tr>
<th>Cluster from SP data</th>
<th>PhylogroupA (n = 16)</th>
<th>PhylogroupB1 (n = 16)</th>
<th>PhylogroupB2 (n = 18)</th>
<th>PhylogroupD (n = 5)</th>
<th>PhylogroupE (n = 7)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Cluster1 (n = 7)</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>7</td>
</tr>
<tr>
<td>Cluster2 (n = 18)</td>
<td>0</td>
<td>0</td>
<td>18</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Cluster3 (n = 35)</td>
<td>14</td>
<td>16</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Cluster4 (n = 1)</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Cluster5 (n = 1)</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
</tbody>
</table>

*Numbers in boldface indicate the maximum correspondence between an obtained cluster and a phylogroup.*
A dendrogram was seen to comprise two distinct clusters, indicating a major speciation event in the past. The first lineage (cluster) comprised strains from the O157, O111, and O55 serotypes of *E. coli*, along with the nonpathogenic K-12 and DH1 strains. The second (larger and more diverse) cluster was constituted of O104, O26, O83, BL21, DI4, SE11, and KO11 among others. In contrast, no such distinctions (i.e., branching) were observed in the SP-based dendrogram. Compared to the 16S rRNA dendrogram, pathogenic and nonpathogenic strains were seen to form several smaller subclusters in the SP-based dendrogram. This suggested that *E. coli* strains which belonged to different lineages but exhibited similar phenotypes (or resided in similar ecological niches) might have undergone convergent genomic modifications, at least in the context of SPs.

For example, in the current analysis, O55 and O157 serotypes of *E. coli* were seen to group with (nonpathogenic) DH1 and K-12 on the basis of similarity in the 16S rRNA gene. However, it was apparent from the heatmap depicting occurrence patterns of SPs across different *E. coli* pathotypes (Fig. 1) as well as the SP-based dendrogram (Fig. 2) that the genomes of O55 and O157 encoded certain SPs (such as D3QK97-like histidine kinases) which were found to be absent in all other strains considered in the study. It was found that D3QK97-like histidine kinases in O55 and O157 closely resembled those in *Shigella dysenteriae* and *Shigella boydii* (Fig. S2 and File S1), thereby indicating that the gene was possibly acquired from pathogenic non-*E. coli* enteric strains through events of horizontal gene transfer. Data presented in Fig. 1 also indicate that, like most other pathogenic *E. coli* strains, O55 and O157 have retained proteins such as energy-coupling-factor transporters (DUF3816-family proteins) which were specifically absent in all the nonpathogenic DH1 and K-12 strains. Several other nonpathogenic strains of *E. coli*, including the BL21, P12b, REL606, and BW2952 strains, were also seen to lack genes for DUF3816-family proteins (Fig. 1). However, the genomes of other nonpathogenic *E. coli* strains like ED1a, IA11, W, LY180, KO11FL, and SE11 encoded DUF3816-family proteins (Fig. 1). Notably, results from the 16S rRNA dendrogram (Fig. 2) indicated that most of the latter set of organisms had a common ancestry. Also, O55 and O157 were seen to lack the sensory homologs of atoS, fecR, and safA genes which were found to be present in the DH1 and K-12 strains (Fig. 1). In yet another instance, the *E. coli* strains W, KO11FL, SE11, LY180, and UMN026 were found to be closely associated in the 16S rRNA dendrogram, but UMN026 segregated from the rest in the SP-based dendrogram (Fig. 2). The genome of the UMN026 strain lacked the safA gene but included atoS and fecR, which differentiated it from its neighbors in the 16S rRNA dendrogram (Fig. 1 and 2). Pertinently, while UMN026 is pathogenic in nature, all the other strains (W, KO11FL, SE11, and LY180) are nonpathogenic/laboratory strains of *E. coli* (8, 10, 11).

In another observation, the *E. coli* strains LF82, NRG 857C, and IHE3034 were found to form a close cluster in the SP-based dendrogram. However, from the 16S rRNA gene sequence data it was evident that although LF82 and IHE3034 shared a common lineage, NRG 857C belonged to a distinctly different lineage (Fig. 2). Comparative analysis of the SPs encoded by the genome of NRG 857C and its neighbors in the 16S rRNA (SE15 and NA114), as well as the SP-based dendrogram (IHE3034 and LF82), revealed the occurrence of fecR and safA genes in the genome of SE15 or NA114 which were absent in NRG 857C (along with IHE3034 and LF82). To evaluate when NRG 857C (or its ancestors) probably lost fecR and safA genes in the course of evolution, a deeper analysis was performed. The SP compositions of all the neighbors of NRG 857C (in the 16S rRNA dendrogram) were scanned to check the occurrence of fecR and safA genes on their genomes (Fig. S3). The safA gene was found to be present in almost all the neighbors of NRG 857C (Fig. 2), indicating that the safA gene was probably lost from the genome of NRG 857C after its speciation from the common ancestor of NRG 857C, SE15, and NA114. The above observations indicate that different strains of *E. coli* might have undergone genetic changes to acquire or lose SPs to adapt to specific niche ecologies or to develop specific phenotypes.
Sensory protein abundances may be correlated with *E. coli* pathogenicity. Results presented so far have indicated that SP signatures in *E. coli* are associated with the phenotype of the strain. Hence, efforts were made to understand whether the abundance of SPs across *E. coli* genomes was also correlated with the pathogenic potential of the strain. To answer this question, a metric called sensory protein index (SPI) was computed in accordance with the procedure described in the Materials and Methods section.

Table S3 provides the computed SPI values for all the studied *E. coli* strains. These *E. coli* strains were then classified into two groups (pathogenic and nonpathogenic) based on their known pathotypes (Table S1). Subsequently, the SPI values obtained using the four methods (see Materials and Methods) were inspected for their ability to segregate the pathogenic and the nonpathogenic strains (Fig. 3). Of the four methods, SPI values obtained using method SPI\(_{M3}\) and SPI\(_{M4}\) were able to distinguish between the pathogenic and nonpathogenic strains with highest accuracies (accuracy values of ~82%).

In order to assess the predictive ability of the two indices (i.e., SPI\(_{M3}\) and SPI\(_{M4}\)) for classifying a strain of *E. coli* which was previously unknown as pathogenic or nonpathogenic, a support vector machine (SVM)-based approach was utilized (see Materials and Methods). The test was performed using hundred-fold cross validation and the prediction statistics were computed (Table S4). In both cases (i.e., SPI\(_{M3}\) and SPI\(_{M4}\)), the model was observed to achieve ~94% sensitivity (in prediction of pathogenic *E. coli*) and over 60% specificity (in prediction of nonpathogenic *E. coli*). Precision, accuracy, and area under the concentration-time curve (AUC) values of ~80% over hundred-fold cross validations established the robustness of these methods for SPI\(_{M3}\) and SPI\(_{M4}\). Overall, findings from this analysis suggest that SPI could be an indicator of the pathogenicity of a given *E. coli* strain.
DISCUSSION

SPs have been shown to play important roles in adaptation and survival of bacteria in various ecological niches (5, 7). Further, SPs are known to function as key mediators of virulence in different species of bacteria (7). In this in silico study, the SP loads in the genomes of different E. coli strains were compared to gauge the extent to which SPs were correlated with phenotypes and phylogeny of these strains. It was noted that different pathotypes of E. coli bear specific SP signatures. Such signatures included the presence of the cusS and pcoS genes in EAggEC, genes for D3QK97-family proteins in STEC/EHEC and EPEC, and the atoS gene in ExPEC/UPEC, AIEC, and APEC. Notably, EHEC and EPEC strains also share commonalities in the virulence gene profiles and EPEC strains are believed to be Stx-negative derivatives of EHEC strains (13, 14). Similarly, AIEC has previously been shown to share several genetic and phenotypic similarities with ExPEC (15). In this context, the current work provides a comparative analysis of SPs among various E. coli pathotypes.

Several instances were noted wherein E. coli strains with common phylogenetic origin demonstrated different phenotypic traits. More often than not, E. coli strains with similar pathotype/phenotype also possessed a similar repertoire of SP genes irrespective of their phylogenetic origins. Evidence of such convergent evolution was observed for the pathogenic as well as the commensal/nonpathogenic strains. While the commensal or nonpathogenic strains seemed to have two distinct phylogenetic origins, the majority of them were found to group together in the SP-based analysis (Fig. 3). It was interesting to note that E. coli strains belonging to phylogroup E and nonpathogenic strains like K-12 and DH1 seemed to have a common evolutionary origin, whereas O55, O157, and Xuzhou21 (belonging to phylogroup E) seemed to acquire certain SPs from other enteropathogens like S. dysenteriae and S. boydii. Such changes could have provided a competitive advantage for survival in a particular environmental niche.

The abundance of SPs across E. coli genomes was also evaluated in this study. E. coli strains did not show any correlations (R² = 0.04; P value = 0.73) between the number of SPs encoded in the genome and the genome length (Fig. S4). The density of occurrence of SPs across E. coli genomes was therefore checked, and referred to as SPI. As the results presented in this study indicated SP signatures in E. coli to be associated with the phenotype of the strain, it was speculated that the SPI across E. coli genomes might be correlated with the pathogenic potential of the strain. In this regard, an earlier study which investigated the distribution of signal-transducing proteins (STPs) across different actinobacterial species indicated a linear relationship between genome length and abundance of STPs encoded in the genome (3). The study did not report any correlation between pathogenicity and the abundance or density of occurrence of STPs across the actinobacterial genomes. In this study, the density of occurrence of SPs (i.e., the SPI) across E. coli genomes seemed to be associated with the pathogenicity status of the strain (Fig. 3).

Grouping the studied strains of E. coli on the basis of the frequency of occurrence of SPs in their genomes could fairly replicate the stratification of these strains on the basis of their phylogenotypes. In addition, E. coli strains belonging to the same pathotypes were also seen to share similar SP signatures. Furthermore, the SPI was seen to be an indicator of the pathogenic potency of E. coli strains. The SPI metric could be useful in the (pathogenic) characterization of hereto uncultured strains which are obtained and routinely sequenced in host microbiome analysis projects, or from among an ensemble of microbial organisms constituting a biospecimen. Thus, the possibilities for using the SPI as a biomarker for diagnosis of a disease or the outcome of a therapeutic intervention cannot be ruled out. Further, SPIs obtained from longitudinal ecological samples have the potential to serve as key indicators of environmental changes. Such changes in the environment are often detrimental to the resident biome and methods for timely detection of environmental changes hold huge socioeconomic benefits.
MATERIALS AND METHODS

Data acquisition and preprocessing. Publicly available information (fasta and protein table files) pertaining to the proteome of 62 fully sequenced *E. coli* strains were retrieved from NCBI Refseq database (ftp://ftp.ncbi.nlm.nih.gov/genomes/archive/old_refseq/Bacteria/). The selected strains ranged from commensal laboratory strains (like K-12) to intestinal and extraintestinal pathogens (e.g., those associated with UTI infections like CFT073), as well as nonhuman pathogens (e.g., avian strains like APEC O1), etc. (Table S1). While the fasta files contained information pertaining to the amino acid sequences of the proteins encoded by the genome (and plasmids) of the studied *E. coli* strains, the protein table files contained the functional annotations of these proteins.

Figure 4 describes the sequence of steps that were followed for identifying SPs encoded by the genome of the studied *E. coli* strains. First, the protein table files were scanned to identify 1,801 well-annotated SPs. Amino acid sequences corresponding to these SPs were also obtained from the fasta files. Given that all the SPs in the studied *E. coli* strains may not be properly annotated, a BLASTp (16) search was performed with the objective of identifying an additional set of poorly characterized SPs. For this purpose, all the protein sequences from the 62 studied *E. coli* strains were searched for sequence similarity against the 1,801 well-annotated SPs obtained in the earlier step. A total of 1,499 BLAST hits, qualifying with an E value cutoff at $1 \times 10^{-5}$ and identity and coverage cutoffs of 95%, were obtained. The resultant 3,300 SPs were further analyzed for occurrence (and architecture) of functional domains using SMART (17, 18) and were subsequently clustered using CD-HIT (sequence identity threshold of 0.95; alignment coverage for longer and shorter sequences at 0.95 in each case) (19, 20). Proteins belonging to each of the clusters were then manually analyzed for the presence/absence of functional domains as well as domain organization. Based on this analysis, some proteins were reclassified. Finally, 3,300 SPs were categorized into 60 classes. The list of the identified SPs along with their annotations and clustering information is available in Table S2.

Analyzing distribution of sensory proteins among different lineages of *E. coli*. In order to inspect the effect of evolutionary events (such as gene duplication, gene deletion, and/or horizontal gene transfer) on the distribution of SPs among *E. coli* strains of different lineages, an analysis was performed in which strains of *E. coli* were hierarchically clustered based on (i) their phylogenetic lineages (16S rRNA sequences) and (ii) the distribution of SPs in their genome (including plasmids). The dendrograms were constructed using the “hclust” function from the “stats” package in R. The two hierarchically clustered dendrograms were subsequently compared to gauge possible events of gene duplication, gene deletion, and horizontal gene transfers (with respect to SPs) in the course of evolution.
Analyzing distribution of sensory proteins among different *E. coli* subtypes. Occurrence of SPs across different strains of *E. coli* was analyzed in order to evaluate any correlations between the SP distribution and the bacterial phenotypes. During this analysis, it was assumed that the occurrence of a single copy of a given class of SP would be adequate for its optimal (and desired) functioning and consequently the data pertaining to the distribution of SPs (Table S5) among different strains of *E. coli* were transformed into a binary matrix. Association of SP distribution was then compared against two *E. coli* phenotypic categories, i.e., (i) phylotype and (ii) pathotype. The studied strains of *E. coli* were segregated into five phylogroups and eight pathotypes in accordance with earlier published data (8, 10, 11). Details of the phylogroups and pathotypes for each of the studied *E. coli* strains are provided in Table S1. For the phylogroup analysis, the SP distribution data (as a binary matrix) were used to hierarchically cluster the 62 studied strains of *E. coli* into 5 groups (k = 5) using the “rect.hclust” function from “stats” package in R. For the pathotype analysis, the SP signatures (i.e., presence/absence of SPs) for the studied *E. coli* strains were compared to check for intra- and interpathotype differences.

Computing sensory protein index. Previous studies have hinted at the possibility of a correlation between the genome size and abundance of encoded SP for organisms belonging to a particular environmental niche (5). To ascertain if any direct association existed between the SPs and observable phenotypes (phenotype/phylogenotype) of the studied strains, the abundance of SPs in the genome (including plasmids) of each of the studied *E. coli* strains was computed. The sensory protein index (SPI) was computed using the following methods. (i) SPI\textsubscript{IM1}: the total number of SPs encoded by the genome (and plasmid) of an organism was considered the SP abundance. (ii) SPI\textsubscript{IM2}: the total number of nucleotide bases which encoded SPs in the genome of an organism (including plasmids) was considered the SP abundance. (iii) SPI\textsubscript{IM3}: SP abundance was computed as a ratio of total count of SPs encoded by an organism to the total genomic size (including plasmids) of that organism, where SPI\textsubscript{IM3} equals the total count of sensory proteins encoded by the genome/the genome size (in Mb). (iv) SPI\textsubscript{IM4}: SP abundance was computed as a ratio of total number of bases covered by all the encoded SPs of an organism to the total size of the genome (including plasmids) of that organism, where SPI\textsubscript{IM4} equals the total length of bases encoding sensory proteins in the genome/the genome size (in Mb).

Statistical validation for SPI. A hierarchical clustering was performed on SPI\textsubscript{IM3} and SPI\textsubscript{IM4} and in each case was divided into two clusters using the “hclust” and “cutree” functions from packages “stats” and “dendextend,” respectively. The accuracy of the clustered output was determined and exhibited the potential of the two indices in determination of pathogenicity in *E. coli*. Further, to add confidence to the predictive ability of the SPI SVM (support vector machine) algorithm, we employed the “svm” function in the “e1071” package in R with a “linear” kernel. While models were built on the train set of randomly sampled 80% data, predictions were made on the remaining 20% of the data (the test set). The process was cross-validated to 100-fold. The mean of the observed statistics (over a hundred-fold cross validations) was evaluated to identify the potential of the indices to differentiate *E. coli* strains into pathogenic and nonpathogenic groups.

**SUPPLEMENTAL MATERIAL**

Supplemental material is available online only.

**SUPPLEMENTAL FILE 1**, XLS file, 0.02 MB.

**SUPPLEMENTAL FILE 2**, XLS file, 0.1 MB.

**SUPPLEMENTAL FILE 3**, XLS file, 0.02 MB.

**SUPPLEMENTAL FILE 4**, XLS file, 0.04 MB.

**SUPPLEMENTAL FILE 5**, XLS file, 0.05 MB.

**SUPPLEMENTAL FILE 6**, PDF file, 0.2 MB.

**ACKNOWLEDGMENTS**

S.B. and T.B. designed the study and performed the analysis. S.B., T.B., and S.S.M. interpreted the data and wrote the manuscript. S.S.M. supervised the project. S.B., T.B., and S.S.M. were employed by Tata Consultancy Services Ltd.

**REFERENCES**


