Published online April 30, 2004

Nucleic Acids Research, 2004, Vol. 32, No. 8 ***2353–2361***
DOI: 10.1093/nar/gkh555

Identification and functional analysis of 'hypothetical' genes expressed in *Haemophilus influenzae*

Eugene Kolker*, Kira S. Makarova[1], Svetlana Shabalina[1], Alex F. Picone, Samuel Purvine, Ted Holzman, Tim Cherny, David Armbruster[2], Robert S. Munson Jr[2], Grigory Kolesov[3], Dmitrij Frishman[3] and Michael Y. Galperin[1]

BIATECH, 19310 North Creek Parkway, Suite 115, Bothell, WA 98011, USA, [1]National Center for Biotechnology Information, NLM, National Institutes of Health, Bethesda, MD 20894, USA, [2]Children's Research Institute, and The Ohio State University, 700 Children's Drive, Columbus, OH 43205, USA and [3]Technische Universität München, Wissenshaftzentrum Weihenstephan, Am Forum 1, 85354 Freising, Germany

Received February 23, 2004; Revised and Accepted March 30, 2004

ABSTRACT

The progress in genome sequencing has led to a rapid accumulation in GenBank submissions of uncharacterized 'hypothetical' genes. These genes, which have not been experimentally characterized and whose functions cannot be deduced from simple sequence comparisons alone, now comprise a significant fraction of the public databases. Expression analyses of *Haemophilus influenzae* cells using a combination of transcriptomic and proteomic approaches resulted in confident identification of 54 'hypothetical' genes that were expressed in cells under normal growth conditions. In an attempt to understand the functions of these proteins, we used a variety of publicly available analysis tools. Close homologs in other species were detected for each of the 54 'hypothetical' genes. For 16 of them, exact functional assignments could be found in one or more public databases. Additionally, we were able to suggest general functional characterization for 27 more genes (comprising ~80% total). Findings from this analysis include the identification of a pyruvate-formate lyase-like operon, likely to be expressed not only in *H.influenzae* but also in several other bacteria. Further, we also observed three genes that are likely to participate in the transport and/or metabolism of sialic acid, an important component of the *H.influenzae* lipo-oligosaccharide. Accurate functional annotation of uncharacterized genes calls for an integrative approach, combining expression studies with extensive computational analysis and curation, followed by eventual experimental verification of the computational predictions.

INTRODUCTION

The recent progress in genome research started in 1995 with the sequencing of the first complete genome of a cellular life form: the 1.8 Mb genome of *Haemophilus influenzae* strain Rd KW20 (1). Eight years later, the genomes of over 100 organisms from all major phylogenetic lineages have been sequenced, and sequencing of many more is currently under way (2,3). Disparities in the accuracy of genome annotation that were the subject of many heated discussions at the beginning of the genome era (4,5) are largely gone. Still, the so-called '70% hurdle' (6) holds, as functions of only ~50–70% of the genes in any given genome can be predicted with reasonable confidence (3,6). The remaining genes are either (i) homologous to genes of unknown function, and are typically referred to as 'conserved hypothetical' genes, or (ii) do not have any known homologs. Since it is often unclear whether they encode actual proteins, the latter genes are commonly referred to as 'hypothetical', 'uncharacterized', or 'unknown' proteins. As of May 25, 2003, the NCBI protein database contained ~360 000 protein sequences from ~120 completely sequenced microbial genomes; one out of three proteins had no assigned function and one out of ten proteins was annotated as 'conserved hypothetical'. Even for *Escherichia coli* strain K-12, arguably the best studied of all organisms, there are still ~2000 genes that have never been experimentally characterized, almost half of all proteins encoded in its genome (3,7). At the current rate of experimental characterization of new *E.coli* genes, 20–30 per year (7), it will take many decades before the biological function of all these proteins is established.

As we have noted earlier, 'conserved hypothetical' genes pose a major challenge to the efforts toward understanding of complete genomes (8). The very idea that there are important genes whose functions are still obscure is quite unsettling as it reveals that there are still important gaps in our understanding of basic (micro)biology (9,10). Our recent study of the *H.influenzae* proteome identified 15 'conserved hypothetical' proteins that were confidently detected in aerobically grown

*To whom correspondence should be addressed. Tel: +1 425 481 7200, extension 100; Fax: +1 425 481 5384; Email: ekolker@biatech.org

cells (11) and whose genes were found to be essential in transposon mutagenesis studies (12). This prompted us to take a closer look at the expressed genes of *H.influenzae* which were annotated as 'hypothetical'.

The choice of *H.influenzae* was driven by the fact that, as the first sequenced microbial genome, it has become a testbed for many annotation efforts during the past 8 years. Despite these efforts, almost one-third of *H.influenzae* genes are currently annotated as 'hypothetical'. In this study, transcriptome and proteome analyses of *H.influenzae* cells grown under normal conditions resulted in confident identifications for 54 such proteins, all of which turned out to have homologs in other organisms and therefore could be considered 'conserved hypothetical'. We were able to come up with general functional characterization for 43 genes (~80% of the test set). For 16 of these, exact functions were assigned through transfer of the functional annotation of orthologous proteins from other organisms. This work demonstrates that high-throughput transcriptome and proteome expression studies need to be integrated with computational approaches and careful curation.

MATERIALS AND METHODS

Gene expression analysis

Haemophilus influenzae gene expression was measured using a QIAGEN Operon microarray of 70mer DNA fragments representing all predicted open reading frames from the *H.influenzae* strain Rd KW20 genome spotted on Corning's UltraGAPSII slides. The RNA was isolated from *H.influenzae* strain Rd KW20 cells grown overnight in normal anaerobic and aerobic conditions on a rich medium [brain–heart infusion broth (11)]. Hybridizations and labeling followed the protocols of the Brown laboratory (http://cmgm.stanford.edu/pbrown/protocols/index.html) (13). Raw gene expression data were initially processed by Axon Instruments GenePix analysis package, and background-subtracted median-normalized intensities were calculated. Based on 12 replicates of such processed intensities, the expression values and corresponding standard errors were estimated using maximum likelihood analysis (14).

Escherichia coli strain MG1655 gene expression was measured using Affymetrix whole-genome oligonucleotide arrays (15,16). Typically, each *E.coli* gene and both strands of each intergenic region were assayed with 15 probe pairs of 25mer nucleotide sequences (16). The expression values and corresponding standard errors were estimated from duplicated experiments using the expectation maximization approach (15).

Protein expression analysis

Protein expression studies of *H.influenzae* strain Rd KW20 used cells grown under the same conditions in the same medium as in the gene expression studies. Cells were resuspended in phosphate-buffered saline and disrupted by passing through SLM Instruments French pressure cell at 15 000 psi. Soluble and membrane fractions were separated by ultracentrifugation, processed and proteolysed by trypsin (11). Soluble and membrane peptide mixtures were injected onto a 10 cm × 100 μm capillary column by the FAMOS autosampler, connected on-line via a Brechbuhler electrospray ionization source to a ThermoFinnigan LCQ DECA XP ion trap mass spectrometer. We used a standard top-down data dependent ion selection approach to tandem mass spectrometry (MS/MS), optimized for protein coverage by employing multiple narrowly overlapping *m/z* window ranges (11). Combining statistical models for peptide (17) and protein (18) identifications with expert verification allowed us to compile the resulting set of high (>90%) confidence protein identifications (11).

Functional characterization of genes

The genes of interest were compared against the SWISS-PROT database (http://www.expasy.org/sprot) (19), the Conserved Domain Database (CDD) (http://www.ncbi.nlm.nih.gov/Structure/cdd/cdd.shtml) (20) and Clusters of Orthologous Groups of Proteins database (COG) (http://www.ncbi.nlm.nih.gov/COG) (21,22) using their respective search tools (owing to the inclusion of COG data into the latest release of CDD, these two databases are referred to here as CDD/COG), and against the NCBI protein database (http://www.ncbi.nlm.nih.gov) and the Protein Data Bank (PDB) (http://www.rcsb.org) using PSI-BLAST (23) with default parameters run until convergence. Additional structural characterization of *H.influenzae* proteins was performed by sequence similarity searches against a collection of SCOP-based profiles (24) using IMPALA (25). In addition, species-specific databases at the Munich Information Center for Protein Sequences (http://mips.gsf.de), the Institut Pasteur (http://genolist.pasteur.fr/Colibri/) and the University of Miami (http://bmb.med.miami.edu/EcoGene/EcoWeb) were checked for functional assignments of the genes in question.

Conserved operons were delineated using a combination of STRING (http://www.bork.embl-heidelberg.de/STRING), SNAPper (http://pedant.gsf.de/snapper) and ERGO (http://ergo.integratedgenomics.com/ERGO) tools (26–28). Protein–protein interaction data were obtained from the Database of Interacting Proteins (DIP) (http://dip.doe-mbi.ucla.edu) (29) and the Hybrigenics PIMRider (30). Putative promoter regions of the studied *H.influenzae* genes were compared with upstream regions of all *H.influenzae* and *E.coli* genes using the OWEN program (31). Finally, the recently enabled global search of all NCBI databases (http://www.ncbi.nlm.nih.gov/Entrez) was performed to ensure the complete retrieval of all the available information. All findings were manually reassessed.

RESULTS

Identification of 'hypothetical' genes expressed in *H.influenzae*

Since genome-wide expression experiments typically study patterns of gene and/or protein expression in response to changes in environmental conditions, such as starvation, ultraviolet radiation, acid or various kinds of stresses, many uncharacterized genes thus become known as 'stress-induced'. Clearly, this does not add much to the understanding of their functions, but at least takes them out of the 'hypothetical' category, proving that these genes code for real proteins that can actually be expressed (32).

Table 1. Gene and protein expression analysis of *H.influenzae*

Original genomic annotation	Total no. of genes	Total no. of expressed genes		
		Transcriptomic data	Proteomic data	Both data methods
All genes	1705	1653	414	401
All hypothetical	556	523	56	54
Conserved	348	337	49	47
Non-conserved	208	186	7	7

This study concentrated on the 'hypothetical' genes from *H.influenzae* that are expressed under normal growth conditions (i.e. during anaerobic or aerobic growth in a rich medium), rather than 'stressed' conditions. The test protein set was selected based on four independent criteria (Table 1). Among all *H.influenzae* genes that (i) showed statistically reliable expression levels in microarray experiments, we have identified those whose products (ii) were confidently detected in liquid chromatography–tandem mass spectrometry (LC–MS/MS) protein expression analyses. We further focused on those proteins that (iii) were originally (1) annotated as 'hypothetical' and (iv) were still listed as such in GenBank on May 25, 2003. The resulting data set included 54 such proteins (Table 1); the complete list is shown in Table 1S (Supplementary Material). A great majority of these genes were originally annotated as 'conserved hypothetical', based on the presence of close homologs in other microorganisms. For the remaining genes, close homologs were initially unavailable, but appeared in the database by the time of this study. This means that all *H.influenzae* genes, confidently identified in both transcriptomic and proteomic experiments and further characterized below, can now be referred to as 'conserved hypothetical'.

A comparison of the gene expression data for this set and their *E.coli* orthologs (where available), obtained in our earlier studies (15,16), reaffirmed that these genes were also expressed in *E.coli* (Table 1S). In addition, seven of these genes (HI0017, HI0119, HI0315, HI0700, HI0847, HI1053 and HI1349) were predicted by Karlin *et al*. (33) to be highly expressed based on their codon frequencies; for several other genes, high expression was predicted for their *E.coli* and *Vibrio cholerae* orthologs (33). These observations substantiated that the *H.influenzae* test set consisted of genes which, while apparently uncharacterized, coded for actual proteins and were worthy of a detailed study.

Interestingly, about one-third of these 54 *H.influenzae* genes were found to be essential in transposon mutagenesis studies, another third were found to be non-essential and the rest produced inconclusive results (Table 1S). This trend was unchanged for the 10 *H.influenzae* genes that did not have *E.coli* orthologs. These data and the apparent discrepancies in gene essentiality between *H.influenzae* and *E.coli* (Table 1S) reiterate the fact that the *H.influenzae* genome is not just a subset of *E.coli* genes (34,35).

Annotation of 'conserved hypothetical' genes in public databases

A close look at the genes of interest revealed that for some of them updated annotations were already available in curated sequence databases such as SWISS-PROT and CDD/COG (19,20,22). Indeed, SWISS-PROT entries for 10 proteins of the current set offered at least some functional annotation (Table 2S, category I, in Supplementary Material) based primarily on experimental characterization of their homologs in *E.coli* and other organisms.

The CDD/COG database provided annotations for 25 additional genes (Table 2S, category II). Some of these annotations were unequivocal, based on experimental data for *H.influenzae* proteins or their orthologs, while others were somewhat vague, reflecting the presence of conserved sequence motifs or subtle sequence similarities to previously characterized proteins. For example, the HI0241 protein is a close homolog of a well-characterized *E.coli* protein YajC (36,37). Therefore HI0241 and its *E.coli* ortholog can be reasonably characterized as 'preprotein translocase subunit YajC' (Table 2S). Curiously, these proteins are still annotated as 'hypothetical' in the SWISS-PROT and PIR databases, although updated annotations are already available in the CDD/COG and Pfam (38) databases.

All SWISS-PROT and CDD/COG assignments listed in Table 2S were manually reassessed and, with one exception (HI0719, see below), found to be appropriate. In one case (HI0227, a member of COG2731, annotated as 'beta subunit of beta-galactosidase'), the CDD/COG annotation could not be verified, and this protein was considered uncharacterized (see below).

Sequence analysis of *H.influenzae* 'conserved hypothetical' genes

Detailed sequence analysis of the remaining proteins using exhaustive PSI-BLAST searches allowed us to provide tentative characterizations for five more genes (Table 2 and Table 2S, category III) that lacked either SWISS-PROT or CDD/COG annotation. Several of these predictions had interesting biological implications. For example, PSI-BLAST searches revealed that the product of the HI0521 (*yjjI*) gene is distantly related to pyruvate-formate lyase (Table 2). Remarkably, this gene forms a predicted operon with the HI0520 (*yjjW*) gene that encodes a homolog of pyruvate-formate lyase activating enzyme PflA. The presence of similar lyase-activating enzyme gene pairs in a variety of diverse microorganisms provides an additional degree of confidence in annotating the HI0521 protein as 'glycyl radical enzyme, related to pyruvate-formate lyase'. In addition, OWEN searches (31) identified similar upstream regions preceding the HI0521-HI0520 operon and the *E.coli yjjI-yjjW* operon, suggesting that these genes are similarly regulated. A comparison of HI0521 protein against the known structure of pyruvate-formate lyase (Fig. 1) showed conservation of the principal catalytic residues and secondary structure elements,

Table 2. Sequence-based functional characterization of 'conserved hypothetical' genes

Gene name *H.influenzae*	*E.coli*	SWISS-PROT no.	Size (aa)	COG no.	Updated annotation, fold
HI0105	*ybgI*	Q57354	279	0327	NIF3 homolog, the central CutA-like domain is distantly related to the nitrogen regulator GlnB
HI0146	*yiaO*	P44542	329	1638	Periplasmic component of a TRAP-typedicarboxylate transport system
HI0147	n/a	P44543	633	3090, 1593	A fusion of two subunits of a TRAP-type dicarboxylate transport system
HI0148	*yjhT*	P44544	379	3055	Cell surface adhesion protein with a Kelch-like seven-bladed beta-propeller fold
HI0367	*yfgA*	Q57065	303	1426	Predicted transcriptional regulator with an N-terminal *xre*-type HTH domain
HI0396	*ycfD*	P44683	404	2850	Cupin superfamily metalloenzyme, possible dioxygenase
HI0467	*yicC*	P44726	287	1561	Uncharacterized stress-induced protein
HI0520	*yjjW*	P44743	262	1180	Pyruvate formate-lyase activating enzyme
HI0521	*yjjI*	P44744	514	n/a	Glycyl radical enzyme, distantly related to pyruvate-formate lyase

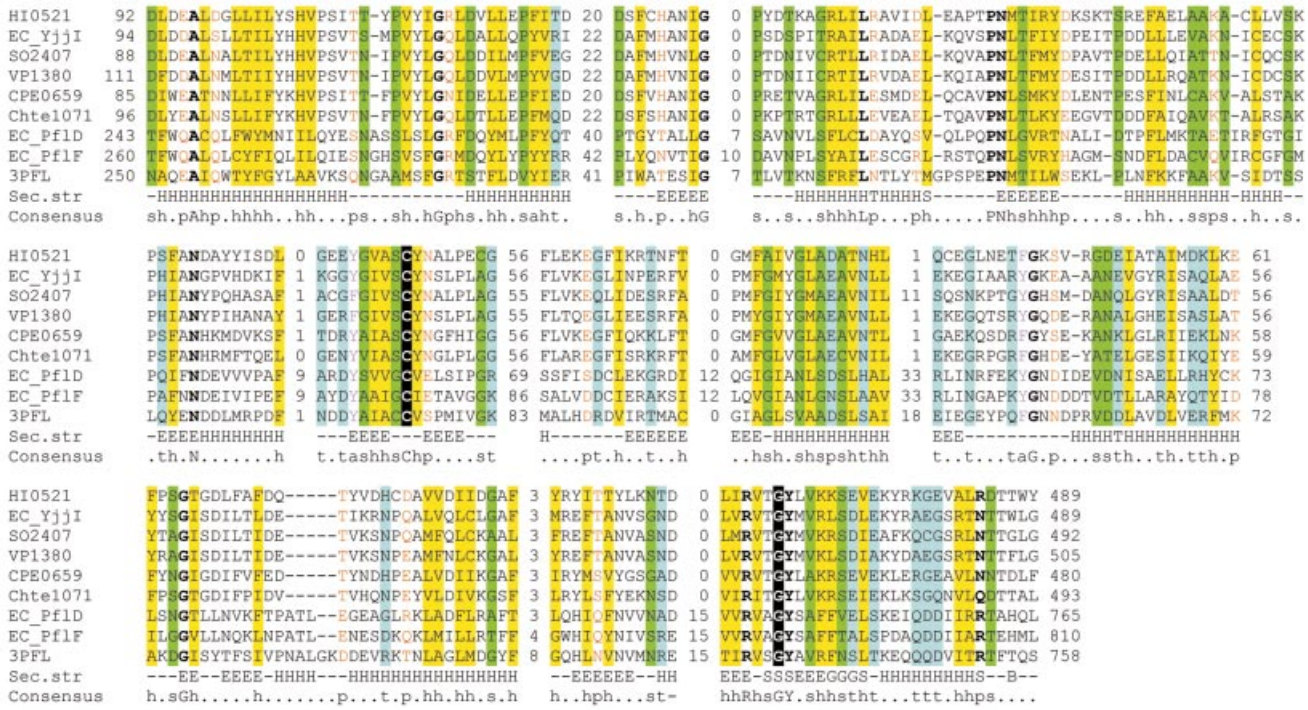

Figure 1. Sequence alignment of HI0521 and related proteins with pyruvate-formate lyases.

thus confirming that these proteins are related. Further experimental analysis of the HI0521 protein would be needed to establish its natural substrate.

Structural genomics data

Haemophilus influenzae is a subject of an ongoing structural genomics project (39) that has selected as targets for detailed structural studies many 'hypothetical' proteins, including 24 proteins from our list (see http://s2f.carb.nist.gov/). Structures of six of them have already been deposited in PDB (Table 3). Unfortunately, in four of these six cases, the 3D structure provided little or no clue to the protein function. In one case, a conserved His residue was found in the protein in phosphorylated form (PDB: 1mwq, M.A. Willis *et al*., in preparation), which suggests the participation of this protein (HI0828) in cellular regulatory mechanisms. In another notable case, a model of the native tetrameric structure of the HI1034 protein was shown to dock well with DNA, suggesting a DNA- or RNA-binding function (40). Although these results (two general functional predictions per six structures) are somewhat disappointing, they seem to reflect accurately the current contribution of structural genomics to the understanding of protein function on the genome scale (41).

Genome context analysis

For the HI0442 protein, neither sequence similarity searches nor determination of the 3D structure could give clear clues to the functional role (Table 3). However, a functional assignment for HI0442 and its orthologs in other bacteria could still be made based on the co-localization and apparent co-expression of this gene with the *recR* gene, almost universally conserved in bacteria (42). The RecR gene product participates in the repair of recombinational DNA damage, which could also be the function of HI0422 protein. However,

Table 3. 'Conserved hypothetical' genes characterized through structural genomics and genome context

Gene name	SWISS-PROT no.	Size (aa)	COG no.	PDB entry	Co-localized genes	Functional predictions based on structure and genome context
HI0227	P44583	155	2731	1jop	*nanRATEK*	Sugar (*N*-acetylneuraminate?)-binding protein, cupin superfamily
HI0315	P44634	246	0217	1kon, 1lfp	*ruvCAB, mutT*	Predicted component of the Holliday junction resolvasome
HI0442	P44711	121	0718	1j8b	*recR, dnaX*	A component of the DNA replication and/or repair system(s)
HI0719	P44839	130	0251	1j7h	*tdcB*	Putative regulator of *pur* operon and Ile biosynthesis; mammalian tumor-associated antigen
HI0828	P44887	98	2350	1mwq	*ispA, yciA, slt, bolA*	Predicted regulator of cell division and/or morphogenesis
HI1034	P44096	163	1666	1in0	*apbA, serB*	DNA- or RNA-binding protein

detection of this protein in the cells grown in rich media under no apparent stress and conserved (albeit not universal) co-localization with the *dnaX* gene, encoding a subunit of DNA polymerase III, both suggest that HI0422 might be involved in normal housekeeping functions such as DNA replication (11).

Likewise, the structures of HI0315 (YebC) homologs from *Aquifex aeolicus* (43) and *E.coli* failed to suggest the function for this protein, encoded in every bacterium except *Buchnera*. Nevertheless, the conserved association of the *yebC* gene with *ruvABC* genes, encoding three subunits of the Holliday junction resolvasome, coupled with the universal presence of this gene in all bacteria (except *Buchnera*) leave little doubt that HI0315 protein is somehow involved in the same process. Since the *yebC* gene appears to be non-essential in *H.influenzae* (Table 1S), it could encode an auxiliary component of the resolvasome that plays a role in DNA replication or, less likely, DNA transcription.

Similar genome-context-based approaches comprise a very important resource that becomes ever more powerful with the growth in the number of sequenced genomes (44). Indeed, consistent application of such methods to the selected set of *H.influenzae* genes (Table 3S in Supplementary Material) allowed us to improve functional assignments for six proteins (Table 3). Several new assignments were noteworthy. A BLAST search identified HI0146 and HI0147 as components of a TRAP-family dicarboxylate transporter, but could not predict its substrate specificity. The genome context analysis (Fig. 1S in Supplementary Material) offered a clue to their function. HI0146–HI0148 genes form a divergent operon with the six-gene operon. This operon includes *N*-acetylneuraminate lyase (*nanA*) and *N*-acetylmannosamine-6-phosphate epimerase *(nanE)* genes, and is responsible for the utilization of *N*-acetylneuraminic acid. In *H.influenzae* this can be either catabolized or used for sialylation of the surface lipopolysaccharide in the outer cell membrane (45,46). It is conceivable that TRAP-type transporter in *H.influenzae*, *V.cholerae, Pasteurella multocida* and several other organisms is a non-orthologous replacement of the MFS superfamily transporter NanT found in *E.coli*. Based on conserved association of these genes with *nan* genes in diverse bacteria (Fig. 1S), the *Fusobacterium nucleatum* homolog of HI0146 was recently annotated as an *N*-acetylneuraminate-binding protein and the homolog of HI0147 annotated as *N*-acetylneuraminate transporter (47). The HI0148 protein contains a typical leader peptide and seven Kelch-like repeats, and can be predicted to have a beta-propeller fold. Given its conserved association with the *nan* genes, it is reasonable to suggest that this secreted protein might also be involved in *N*-acetylneuraminate uptake and/or incorporation into lipo-oligosaccharide. The HI0227 gene is also associated with the *nan* gene cluster in various organisms (Fig. 1S), suggesting that its product might also participate in the metabolism of *N*-acetylneuraminic acid or its derivatives.

The recently determined structure of HI0227 protein (PDB: 1jop) (A. Teplyakov *et al*., in preparation) shows that it is a member of the cupin superfamily of double-stranded beta-helix fold proteins. This superfamily includes, among others, dTDP-4-dehydrorhamnose 3,5-epimerase (RmlC, PDB: 1dzr), yeast phosphomannose isomerase (PDB: 1pmi) and the arabinose-binding domain of the transcriptional regulator AraC (PDB: 2arc), which is consistent with a possible involvement of HI0227 in carbohydrate metabolism. However, since deletion of the *E.coli yhcH* gene (HI0227 ortholog) did not affect cell growth on *N*-acetylneuraminate as a sole carbon source (48), the function of this protein remains to be determined. An experimental testing of the likely interaction of HI0227 with *N*-acetylneuraminic acid would be a logical next step and should shed light on the function(s) of this protein.

In several other cases, two or more 'hypothetical' genes that are non-neighbors in *H.influenzae* were found to share gene neighborhoods in other organisms. These shared genome contexts (indicated as categories in Table 3S) probably reflect certain functional tasks that need to be carried out in the growing cell of *H.influenzae*, much like sialylation of the lipo-oligosaccharide in the previous example. Although we still lack a basic understanding of these tasks, the good news is that their number appears to be very limited (Table 3S). One such group includes HI0467, which is currently annotated only as 'Uncharacterized stress-induced protein', yet nonetheless is also expressed in non-stressed aerobic cells and is essential for growth of *H.influenzae* (Table 1S). It is apparently co-expressed with guanylate kinase and the omega chain (RpoZ) of the RNA polymerase (Table 3S); it is tempting to speculate that HI0467 is somehow involved in transcriptional regulation by guanyl nucleotides.

Another potential approach to identifying functionally related genes relies on the identification of similar promoter regions, which suggests their co-regulation by the same transcriptional regulators (49). For the genes studied here, this approach was unsuccessful (with one exception, see above), as similarities between the upstream intergenic regions of uncharacterized genes were typically too low to indicate their possible co-regulation with reasonable confidence.

Phyletic patterns

Phylogenetic patterns of gene distribution among particular lineages of organisms with completely sequenced genomes can sometimes provide helpful clues to functional annotation (3,8,50,51). Unfortunately, most of the genes (proteins) of the current set had similar, but non-informative, phylogenetic distributions, i.e. they were encoded mostly in the gamma-proteobacteria. An organism-by-organism listing (the phyletic pattern) showed, for example, that HI0017 and HI1681 were encoded exclusively in the representatives of three gamma-proteobacterial families, namely Pasteurellaceae, Enterobacteriaceae and Vibrionaceae. However, this particular phyletic pattern is found in as many as 39 CDD/COGs in the current collection. Therefore suggesting that these two proteins have a common functionality would be highly speculative.

Protein–protein interactions

Studies of protein–protein interactions comprise another powerful high-throughput approach to gaining an insight into protein function (52). Unfortunately, recent analyses of the protein–protein interactions data reported a high rate of false-positive interactions (53). In addition, protein–protein interactions typically offer only vague clues as to the exact function of the protein in question. Therefore, as with other approaches, functional characterizations can be obtained from genome-wide protein interaction data only through painstaking curation. For example, the functional assignment of the HI0315 (YebC) protein, discussed above, could be strengthened by the experimental observation that the *Helicobacter pylori* YebC homolog, HP0162, interacts with HP1046 (YhbC). YhbC is another 'conserved hypothetical' gene, whose gene forms a conserved operon with *nusA* gene. Since that NusA protein directly interacts with the RNA-polymerase, regulating initiation and termination of transcription (54), it may be that YebC and YhbC are also involved in these processes which include DNA unwinding. This example shows that even reliable protein–protein interaction data do not offer a surefire way to assign function to 'hypothetical' proteins.

Uncharacterized conserved genes

It should be noted that despite our best efforts we could not come up with any clues to the functions of at least seven genes from the original list: HI0246, HI0668, HI0700, HI0847, HI1168, HI1236 and HI1709. For two more, HI0370 and HI1681, a conserved gene neighborhood (e.g. co-expression of HI1681-like genes with methylglyoxal synthase) is clearly insufficient to come up with general or even partial functional predictions. Furthermore, a careful look at Tables 2 and 3 reveals that the annotations for some entries seem somewhat inadequate. For example, although HI1514 is correctly annotated in SWISS-PROT and other databases as a phage *Mu*-encoded protein, its exact function is still unknown; it might participate in protein–protein interactions or even have some enzymatic function.

DISCUSSION

At the very beginning of the genome sequencing era, Walter Gilbert and colleagues presciently warned of a 'database explosion' stemming from the rapidly increasing amount of incoming DNA sequences (55). Luckily, this threat has not materialized thus far due to the swift growth in computational power and storage capacity. Nonetheless, while we have managed to cope with data accumulation, our capacity to comprehend these data is far from perfect. While at least 50–70% of proteins encoded in any genome are homologous to proteins already present in databases (3,6), every newly sequenced genome brings about hundreds to thousands of novel genes that have never been seen before, and whose very existence in the living cell is uncertain, let alone their function. Since it is unrealistic to expect that all of these genes or even a significant fraction of them will be studied experimentally any time soon, the functions of these genes need to be deduced through an integrative approach, combining high-throughput methods, computational analyses and curation (10).

As noted previously (8), when a gene is annotated as 'conserved hypothetical', this does not necessarily mean that the function of its product is completely unknown or that its very existence is questionable. Indeed, certain exceptions notwithstanding, if a protein is encoded in several different genomes, it is not really hypothetical any more. In addition, a general prediction of its function can be often made based on a conserved sequence motif, subtle sequence similarity to a previously characterized protein or the presence of diagnostic structural features (3). Many 'conserved hypothetical' genes can be confidently predicted to be ATPases, GTPases, methyltransferases, metalloproteases, DNA- or RNA-binding proteins or membrane transporters (22). Additional hints regarding the gene function(s) could come from genome context analysis, phyletic patterns, domain fusions and protein–protein interaction data (3,42,52).

In this study, we have taken a close look at 54 *H.influenzae* 'conserved hypothetical' genes (Tables 1 and 1S) whose expression was confidently detected both at the mRNA and protein levels. For 40 of these genes (Tables 2 and 2S), at least a general functional description has been provided using primarily homology-based searches. For 16 of these, the exact function was already known or could be assigned based on the experimental data from close homologs. Structural genomics data (Table 3) led to two more general functional predictions that were not made from sequence analysis alone. Genome context analysis resulted in five general functional predictions (Tables 3 and 3S) and improved the homology-based annotations for several more genes. It has lent support to the prediction that HI0520 can function as a formate acyltransferase (Fig. 1) and allowed substrate specificity predictions for HI0146 and HI0148 (Fig. 1S). Although promoter search methods were unhelpful in this work, they should become more powerful with the accumulation of additional genome sequences or when applied to eukaryotes, where a single promoter recognition site may be expected to occur in many places throughout the genome.

In summary, the integrated analysis performed in this work resulted in the assignment of precise function(s) to 16 genes and general functions to 27 more genes (43 out of 54, ~80% total), and gave some clues to the potential functions (partial functional characterizations) of three more genes. Some of these general functional predictions are sufficiently precise to allow relatively straightforward experimental verification. The results of this pilot project show several trends listed below

that might be important for further attempts at large-scale functional analyses of the uncharacterized genes in other organisms.

Typically, sequencing and initial annotation will leave 30–50% of genes uncharacterized. Gene and protein expression studies of the cells grown under normal or stressed conditions can help determine when (if ever) these genes are expressed and sometimes suggest their functions. A major bottleneck of such studies is the reliability of the data, which has to be ensured through rigorous statistical analysis (11,35,53,56).

Publicly available data and tools powerfully complement the expression studies. However, most databases are slow in incorporating experimental data. Reliable annotation of a given gene might be available in some databases but not others, begging for serious development in the genome annotation databases. Of course, one has to beware of poorly justified and misleading annotations that are quite common in public databases (3,57,58), generated mostly through automated annotation pipelines (3,59). For the time being, manual curation is vital to sort reliable annotations from spurious ones.

Sequence similarity searches using BLAST, PSI-BLAST or other tools can frequently provide only a general functional assignment (as in Table 2). This should not be perceived as a drawback of these methods; a general or even partial functional assignment is much better than none, as it can be used for designing experimental approaches geared towards estimations of the exact function of a given gene.

Although for many genes the exact functional annotation is still unattainable, detailed sequence analysis can sometimes suggest a general biochemical function, while genome context methods can provide useful clues as to the cellular role of the protein in question. Combining this evidence can lead to reasonably good experimentally testable hypotheses. An important caveat is that both gene neighborhood and phyletic profiling methods critically depend on the correct estimation of orthologous relationships and typically fail in cases of protein families that contain numerous paralogs.

Structural genomics projects were originally expected to provide clear clues as to the functions of uncharacterized genes and allow their functional annotation with minimal experimental follow-up. The findings of this work and others (41,60) indicate that, at least for 'conserved hypothetical' proteins, function can rarely be gleaned from structure alone. Rather, 3D structures tend to narrow the list of possible functions and occasionally lead to testable hypotheses, not unlike the results of sequence similarity searches.

Computational methods, including sophisticated sequence analysis, phyletic patterns, domain fusions, structural threading and gene neighborhoods can and should be used for prediction of the likely biochemical properties of these genes. However, the ultimate biological function(s) for members of new conserved protein families can be established through direct experimentation.

Experimental studies *per se* do not always result in (exact) functional assignments. A case in point is HI0719 (YjgF), a protein that has been shown to affect expression of *pur* and *ile* genes, inhibit translation and even serve as mammalian tumor-associated antigen (61). Despite several years of intensive experimental studies, resulting in the demonstration of endoribonuclease activity in its rat homolog and determination of the 3D structures of its yeast and *Bacillus subtilis* homologs, the actual function of this protein still remains enigmatic.

In conclusion, 'conserved hypothetical' genes pose a challenge not just to functional genomics, but also to general (micro)biology. As the list of the 'conserved hypothetical' proteins keeps growing at an escalating pace, integrative studies that combine protein and gene expression analyses, computational biology, comparative genomics, mutational analysis and curation can help in identifying the most intriguing genes in every genome (8,10). This in turn will create reasonable and verifiable hypotheses for further experimental and computational work towards better understanding of the functioning of (bacterial) cells.

SUPPLEMENTARY MATERIAL

Supplementary Material is available at NAR Online.

ACKNOWLEDGEMENTS

We greatly appreciate helpful discussions with Dayle Daines, Samuel Karlin, Natali Kolker, Eugene Koonin, Aleksey Kondrashov, David Lipman, Jan Mrazek, Andrei Osterman, Bernhard Palsson, Richard Roberts, Arnold Smith, Tatiana Tatusova, Alexey Teplyakov and Brian Tjaden. We thank Dayle Daines and Arnold Smith for their help with the RNA samples. This work was supported by the NIH's R01 grants DC03915 and DC005980 (R.M.) and the DOE's OBER and OSCAR *Genomics: GTL* grant DE-FG08–01ER63218 (E.K.).

REFERENCES

1. Fleischmann,R.D., Adams,M.D., White,O., Clayton,R.A., Kirkness,E.F., Kerlavage,A.R., Bult,C.J., Tomb,J.-F., Dougherty,B.A., Merrick,J.M. *et al*. (1995) Whole-genome random sequencing and assembly of *Haemophilus influenzae* Rd. *Science*, **269**, 496–512.
2. Drell,D. (2002) The Department of Energy microbial cell project: a 180 degrees paradigm shift for biology. *OMICS*, **6**, 3–9.
3. Koonin,E.V. and Galperin,M.Y. (2002) *Sequence–Evolution–Function. Computational Approaches in Comparative Genomics*. Kluwer Academic, Boston, MA.
4. Koonin,E.V., Mushegian,A.R., Galperin,M.Y. and Walker,D.R. (1997) Comparison of archaeal and bacterial genomes: computer analysis of protein sequences predicts novel functions and suggests a chimeric origin for the archaea. *Mol. Microbiol.*, **25**, 619–637.
5. Brenner,S.E. (1999) Errors in genome annotation. *Trends Genet.*, **15**, 132–133.
6. Bork,P. (2000) Powers and pitfalls in sequence analysis: the 70% hurdle. *Genome Res.*, **10**, 398–400.
7. Thomas,G.H. (1999) Completing the *E.coli* proteome: a database of gene products characterised since the completion of the genome sequence. *Bioinformatics*, **15**, 860–861.
8. Galperin,M.Y. (2001) Conserved 'hypothetical' proteins: new hints and new puzzles. *Comp. Funct. Genomics*, **2**, 14–18.
9. Frazier,M.E., Johnson,G.M., Thomassen,D.G., Oliver,C.E. and Patrinos,A. (2003) Realizing the potential of the genome revolution: the genomes to life program. *Science*, **300**, 290–293.
10. Roberts,R.J. (2004) Identifying protein function—a call for community action. *PLoS Biol.*, **2**, 293–294.
11. Kolker,E., Purvine,S., Galperin,M.Y., Stolyar,S., Goodlett,D.R., Nesvizhskii,A.I., Keller,A., Xie,T., Eng,J.K., Yi,E. *et al*. (2003) Initial proteome analysis of model microorganism *Haemophilus influenzae* Rd strain KW20. *J. Bacteriol.*, **185**, 4593–4602.
12. Akerley,B.J., Rubin,E.J., Novick,V.L., Amaya,K., Judson,N. and Mekalanos,J.J. (2002) A genome-scale analysis for identification of genes required for growth or survival of *Haemophilus influenzae*. *Proc. Natl Acad. Sci. USA*, **99**, 966–971.

13. Courcelle,J., Khodursky,A., Peter,B., Brown,P.O. and Hanawalt,P.C. (2001) Comparative gene expression profiles following UV exposure in wild-type and SOS-deficient *Escherichia coli*. *Genetics*, **158**, 41–64.
14. Ideker,T., Thorsson,V., Siegel,A.F. and Hood,L.E. (2000) *J. Comput. Biol.*, **7**, 805–817.
15. Tjaden,B., Haynor,D.R., Stolyar,S., Rosenow,C. and Kolker,E. (2002) Identifying operons and untranslated regions of transcripts using *Escherichia coli* RNA expression analysis. *Bioinformatics*, **18**, S337–S344.
16. Tjaden,B., Saxena,R.M., Stolyar,S., Haynor,D.R., Kolker,E. and Rosenow,C. (2002) Transcriptome analysis of *Escherichia coli* using high-density oligonucleotide probe arrays. *Nucleic Acids Res.*, **30**, 3732–3738.
17. Keller,A., Nesvizhskii,A.I., Kolker,E. and Aebersold,R. (2002) Empirical statistical model to estimate the accuracy of peptide identifications made by MS/MS and database search. *Anal. Chem.*, **74**, 5383–5392.
18. Nesvizhskii,A.I., Keller,A., Kolker,E. and Aebersold,R. (2003) A statistical model for identifying proteins by tandem mass spectrometry *Anal. Chem.*, **75**, 4646–4658.
19. Boeckmann,B., Bairoch,A., Apweiler,R., Blatter,M.C., Estreicher,A., Gasteiger,E., Martin,M.J., Michoud,K., O'Donovan,C., Phan,I. *et al*. (2003) The SWISS-PROT protein knowledgebase and its supplement TrEMBL in 2003. *Nucleic Acids Res.*, **31**, 365–370.
20. Marchler-Bauer,A., Anderson,J.B., DeWeese-Scott,C., Fedorova,N.D., Geer,L.Y., He,S., Hurwitz,D.I., Jackson,J.D., Jacobs,A.R., Lanczycki,C.J. *et al*. (2003) CDD: a curated Entrez database of conserved domain alignments. *Nucleic Acids Res.*, **31**, 383–387.
21. Tatusov,R.L., Koonin,E.V. and Lipman,D.J. (1997) A genomic perspective on protein families. *Science*, **278**, 631–637.
22. Tatusov,R.L., Galperin,M.Y., Natale,D.A. and Koonin,E.V. (2000) The COG database: a tool for genome-scale analysis of protein functions and evolution. *Nucleic Acids Res.*, **28**, 33–36.
23. Altschul,S.F., Madden,T.L., Schaffer,A.A., Zhang,J., Zheng,Z., Miller,W. and Lipman,D.J. (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res.*, **25**, 3389–3402.
24. LoConte,L., Brenner,S.E., Hubbard,T.J., Chothia,C. and Murzin,A.G. (2002) SCOP database in 2002: refinements accommodate structural genomics. *Nucleic Acids Res.*, **30**, 264–267.
25. Schaffer,A.A., Wolf,Y.I., Ponting,C.P., Koonin,E.V., Aravind,L. and Altschul,S.F. (1999) IMPALA: matching a protein sequence against a collection of PSI-BLAST-constructed position-specific score matrices. *Bioinformatics*, **15**, 1000–1011.
26. vonMering,C., Huynen,M., Jaeggi,D., Schmidt,S., Bork,P. and Snel,B. (2003) STRING: a database of predicted functional associations between proteins. *Nucleic Acids Res.*, **31**, 258–261.
27. Kolesov,G., Mewes,H.W. and Frishman,D. (2002) SNAPper: gene order predicts gene function. *Bioinformatics*, **18**, 1017–1019.
28. Overbeek,R., Larsen,N., Walunas,T., D'Souza,M., Pusch,G., Selkov,E., Jr, Liolios,K., Joukov,V., Kaznadzey,D., Anderson,I. *et al*. (2003) The ERGO genome analysis and discovery system. *Nucleic Acids Res.*, **31**, 164–171.
29. Xenarios,I., Salwinski,L., Duan,X.J., Higney,P., Kim,S.M. and Eisenberg,D. (2002) DIP, the Database of Interacting Proteins: a research tool for studying cellular networks of protein interactions. *Nucleic Acids Res.*, **30**, 303–5.
30. Wojcik,J., Boneca,I.G. and Legrain,P. (2002) Prediction, assessment and validation of protein interaction maps in bacteria. *J. Mol. Biol.*, **323**, 763–770.
31. Ogurtsov,A.Y., Roytberg,M.A., Shabalina,S.A. and Kondrashov,A.S. (2002) OWEN: aligning long collinear regions of genomes. *Bioinformatics*, **18**, 1703–1704.
32. Seputiene,V., Motiejunas,D., Suziedelis,K., Tomenius,H., Normark,S., Melefors,O. and Suziedeliene,E. (2003) Molecular characterization of the acid-inducible *asr* gene of *Escherichia coli* and its role in acid stress response. *J. Bacteriol.*, **185**, 2475–2484.
33. Karlin,S., Mrazek,J., Campbell,A. and Kaiser,D. (2001) Characterizations of highly expressed genes of four fast-growing bacteria. *J. Bacteriol.*, **183**, 5025–5040.
34. Kolker,E., Purvine,S., Picone,A., Cherny,T., Akerley,B.J., Munson,R.S., Jr, Palsson,B.O., Daines,D.A. and Smith,A.L. (2002) *H.influenzae* consortium: Integrative study of *H.influenzae*—human interactions. *OMICS*, **6**, 341–348.
35. Raghunathan,A., Price,N., Galperin,M.Y., Makarova,K.S., Purvine,S., Picone,A.F., Cherny,T., Xie,T., Munson,R. Jr, Tyler,R. *et. al*. (2003) *In silico* metabolic model and protein expression of *Haemophilus influenzae* Rd strain KW20 in rich medium. *OMICS*, **8**, 25–42.
36. Taura,T., Akiyama,Y. and Ito,K. (1994) Genetic analysis of SecY: additional export-defective mutations and factors affecting their phenotypes. *Mol. Gen. Genet.*, **243**, 261–269.
37. Duong,F. and Wickner,W. (1997) The SecDFyajC domain of preprotein translocase controls preprotein movement by regulating SecA membrane cycling. *EMBO J.*, **16**, 2756–2768.
38. Bateman,A., Birney,E., Cerruti,L., Durbin,R., Etwiller,L., Eddy,S.R., Griffiths-Jones,S., Howe,K.L., Marshall,M. and Sonnhammer,E.L. (2002) The Pfam protein families database. *Nucleic Acids Res.*, **30**, 276–280.
39. Eisenstein,E., Gilliland,G.L., Herzberg,O., Moult,J., Orban,J., Poljak,R.J., Banerjei,L., Richardson,D. and Howard,A.J. (2000) Biological function made crystal clear—annotation of hypothetical proteins via structural genomics. *Curr. Opin. Biotechnol.*, **11**, 25–30.
40. Teplyakov,A., Obmolova,G., Bir,N., Reddy,P., Howard,A.J. and Gilliland,G.L. (2003) Crystal structure of the YajQ protein from *Haemophilus influenzae* reveals a tandem of RNP-like domains. *J. Struct. Funct. Genomics*, **4**, 1–9.
41. Frishman,D. (2003) What we have learned about prokaryotes from structural genomics. *OMICS*, **7**, 211–224.
42. Lim,K., Tempczyk,A., Parsons,J.F., Bonander,N., Toedt,J., Kelman,Z., Howard,A., Eisenstein,E. and Herzberg,O. (2003) Structure of the YibK methyltransferase from *Haemophilus influenzae* (HI0766): a cofactor bound at a site formed by a knot. *Proteins*, **50**, 375–379.
43. Shin,D.H., Yokota,H., Kim,R. and Kim,S.H. (2002) Crystal structure of conserved hypothetical protein Aq1575 from *Aquifex aeolicus*. *Proc. Natl Acad. Sci. USA*, **99**, 7980–7985.
44. Overbeek,R., Fonstein,M., D'Souza,M., Pusch,G.D. and Maltsev,N. (1999) The use of gene clusters to infer functional coupling. *Proc. Natl Acad. Sci. USA*, **96**, 2896–2901.
45. Plumbridge,J. and Vimr,E. (1999) Convergent pathways for utilization of the amino sugars *N*-acetylglucosamine, *N*-acetylmannosamine and *N*-acetylneuraminic acid by *Escherichia coli*. *J. Bacteriol.*, **181**, 47–54.
46. Vimr,E., Lichtensteiger,C. and Steenbergen,S. (2000) Sialic acid metabolism's dual function in *Haemophilus influenzae*. *Mol. Microbiol.*, **36**, 1113–1123.
47. Kapatral,V., Ivanova,N., Anderson,I., Reznik,G., Bhattacharyya,A., Gardner,W.L., Mikhailova,N., Lapidus,A., Larsen,N., D'Souza,M. *et al*. (2003) Genome analysis of *F.nucleatum sub spp vincentii* and its comparison with the genome of *F.nucleatum* ATCC 25586. *Genome Res.*, **13**, 1180–1189.
48. Kalivoda,K.A., Steenbergen,S.M., Vimr,E.R. and Plumbridge,J. (2003) Regulation of sialic acid catabolism by the DNA binding protein NanR in *Escherichia coli*. *J. Bacteriol.*, **185**, 4806–4815.
49. Laikova,O.N., Mironov,A.A. and Gelfand,M.S. (2001) Computational analysis of the transcriptional regulation of pentose utilization systems in the gamma subdivision of Proteobacteria. *FEMS Microbiol. Lett.*, **205**, 315–322.
50. Galperin,M.Y. and Brenner,S.E. (1998) Using metabolic pathway databases for functional anotation. *Trends Genet.*, **14**, 332–333.
51. Pellegrini,M., Marcotte,E.M., Thompson,M.J., Eisenberg,D. and Yeates,T.O. (1999) Assigning protein functions by comparative genome analysis: protein phylogenetic profiles. *Proc. Natl Acad. Sci. USA*, **96**, 4285–4288.
52. Xenarios,I., Fernandez,E., Salwinski,L., Duan,X.J., Thompson,M.J., Marcotte,E.M. and Eisenberg,D. (2001) DIP: The Database of Interacting Proteins: 2001 update. *Nucleic Acids Res.*, **29**, 239–241.
53. Sprinzak,E., Sattath,S. and Margalit,H. (2003) How reliable are experimental protein–protein interaction data? *J. Mol. Biol.*, **327**, 919–923.
54. Gusarov,I. and Nudler,E. (2001) Control of intrinsic transcription termination by N and NusA: the basic mechanisms. *Cell*, **107**, 437–449.
55. Bhatia,U., Robison,K. and Gilbert,W. (1997) Dealing with database explosion: a cautionary note. *Science*, **276**, 1724–1725.
56. Holzman,T. and Kolker,E. (2004) Statistical analysis of global gene expression data: some practical considerations. *Curr. Opin. Biotechnol.*, **15**, 52–57.
57. Galperin,M.Y. and Koonin,E.V. (2000) Who's your neighbor? New computational approaches for functional genomics. *Nat. Biotechnol.*, **18**, 609–613.

58. Nahum,L.A. and Riley,M. (2001) Divergence of function in sequence-related groups of *Escherichia coli* proteins. *Genome Res.*, **11**, 1375–1381.
59. Galperin,M.Y. and Frishman,D. (1999) Towards automated prediction of protein function from microbial genomic sequences. *Methods in Microbiology*, Vol. 28, *Automation*, Academic Press, London, pp. 245–263.
60. Volz,K. (1999) A test case for structure-based functional assignment: the 1.2 A crystal structure of the yjgF gene product from *Escherichia coli*. *Protein Sci.*, **8**, 2428–2437.
61. Parsons,L., Bonander,N., Eisenstein,E., Gilson,M., Kairys,V. and Orban,J. (2003) Solution structure and functional ligand screening of HI0719, a highly conserved protein from bacteria to humans in the YjgF/YER057c/UK114 family. *Biochemistry*, **42**, 80–89.