Current Biology, Vol. 13, 129–133, January 21, 2003, ©2003 Elsevier Science Ltd. All rights reserved. PII S0960-9822(03)00009-5

Trait-to-Gene: A Computational Method for Predicting the Function of Uncharacterized Genes

Mitchell Levesque,[1,4,*] Dennis Shasha,[2] Wook Kim,[3] Michael G. Surette,[3] and Philip N. Benfey[1,4]
[1]New York University
1009 Main Building
100 Washington Square East
New York, New York 10003
[2]Department of Computer Science
Courant Institute of Mathematical Sciences
New York University
251 Mercer Street
New York, New York 10012
[3]Department of Microbiology and Infectious Diseases
University of Calgary
3330 Hospital Drive Northwest
Calgary, Alberta T2N-4N1
Canada

Summary

The function of unknown genes is often inferred from comparisons to well-characterized homologs. In this paper, we show that, even if all of the homologs of a gene are unannotated, its function may be deduced through phylogenetic profiling. We have designed a series of algorithms that make functional predictions of genes based on orthology and set theory, but our approach to predicting gene function requires no previous knowledge of homolog function. With this technique, we successfully identified 94% of the clusters of orthologous groups that are known to be involved in flagella development or function. As a test, we removed the function of three putative flagellar genes that had been previously uncharacterized in *Bacillus subtilis*. We observed a motility phenotype for two of these three genes. Thus, these algorithms allow for high-throughput functional prediction of genes beyond that provided by simple orthology-based annotation endeavors.

Results and Discussion

The availability of multiple genome databases and comparative genomic techniques offers new opportunities to complement conventional gene identification strategies. In particular, the use of phylogenetic profiles to predict gene function has demonstrated some success in identifying new members of developmental and biochemical pathways [1, 2]. We have developed a new approach to predicting gene function that goes beyond previous methods by establishing threshold values for phylogenetic profiles, by comparing sets of genes in potentially complementary pathways, and by testing the predictions through gene knockouts in the model organism *Bacillis subtilis*.

*Correspondence: mpl4@duke.edu
[4]Present address: Department of Biology, Duke University, Box 91000, Durham, North Carolina 27708.

Algorithm Design and Results

We wrote four set-theoretic algorithms that make functional predictions of genes based on phenotypic traits and gene orthology (Figure 1). Orthologous gene sets were obtained from the Cluster of Orthologous Group (COG) database (available at the National Center for Biotechnology Information) [3]. The algorithms use Boolean combinations of COGs and traits to predict gene function. To test the algorithms, we chose flagella formation and function as the proof of principle character, since it is genetically well characterized and is consistently described for most sequenced bacteria [4–7].

Perfect Match Algorithm

The Perfect Match algorithm identifies the exact subset of COGs that equals the subset of genomes with a trait. It is designed to find those genes that are potentially necessary to confer a particular trait and have been lost in organisms that lack the trait. When we tested this algorithm for the flagella trait (which is present in 9 of the 21 bacteria under consideration), 100% of the reported COGs had annotations consistent with their role in flagellar biogenesis (Figure 2). The algorithm reported that 13 out of 3415 COGs in the database were involved in flagella function. Although this represents only 42% of all of the known bacterial flagellar COGs (31 at the time of this analysis), the algorithm's precision was perfect. Perhaps because empirical work has identified many of the most highly conserved genes involved in flagella development, the Perfect Match algorithm predictions did not reveal any new COGs to test. Thus, we would expect these genes to encode the proteins that are required for flagellar formation in bacteria and have not been coopted for other functions in nonflagellar species. However, it is expected that some flagellar genes may either have been replaced by other genes in bacteria with flagella or recruited by nonflagellar bacteria for other purposes. Thus, other algorithms were designed that would allow for the identification of COGs that were not a perfect subset of the genomes with flagella.

One-Needed Algorithm

The One-needed algorithm identifies genes that are necessary for a certain trait but for which other genes may have been recruited in some genomes. This algorithm finds a set of COGs such that the genomes associated with each COG are a subset of the genomes having the trait of interest and the union of the genomes associated with the COGs equals the genomes having the trait of interest (Figure 1). These COGs complement each other's phylogenetic profiles within the set of genomes with a trait. The genes represented by the COGs are potentially functionally convergent in the sense that they may have no evolutionary relationship and yet their products may perform similar functions in different genomes. However, we added the stipulation that a COG must cover at least two unique genomes to complement the

Perfect Match	One-needed
$\{G_c\} = \{G_t\}$	$\{G_{c'}\} \bigcap \{G_{c''}\} \bigcap \{G_{c'''}\} \ldots = \{G_t\}$
Similarity Measure	**All-needed**
$\frac{\left\|\{G_c\} \bigcap \{G_t\}\right\|}{\left\|\{G_c\} \bigcup \{G_t\}\right\|} \geq X$	$\{G_{c'}\} \bigcup \{G_{c''}\} \bigcup \{G_{c'''}\} \ldots = \{G_t\}$

Figure 1. Algorithms Used in This Analysis

G_c represents the set of genomes with COG_c. G_t represents the set of genomes with trait t. X is a threshold value that is adjusted between 0 and 1 to set the stringency of the algorithm. A value of 1 for X generates the same results as the Perfect Match algorithm.

phylogenetic profile of another COG (see the Experimental Procedures in the Supplementary Material available with this article online).

The results of the One-needed algorithm are a pairing of potentially convergent COGs. For flagella, the annotations gave no indication that any of the pairs were characterized as having similar functions to each other. However, reliance on annotations alone will not provide sufficient evidence to either accept or reject the hypotheses generated by the One-needed algorithm. Thus, further experimentation is required to determine if the hypothesis of functional convergence is accurate. In examining the One-needed data, it is interesting to note that, for 11 out of the 19 COG pairs, one member was from a known flagellar COG (Figure 2). The reason is that the One-needed algorithm finds the COGs with the best, but incomplete, representation of the genomes with flagella and then matches other COGs that complete the phylogenetic profile of the first COG to equal the subset of genomes with the trait. For this reason, we used the output of this algorithm to help select candidates for knockout experiments (see below).

All-Needed Algorithm

The All-needed algorithm identifies those genes that are present in all genomes with a trait as well as in other genomes without the trait. This is accomplished by finding the intersection of the sets of COGs that exactly equals the set of genomes with a trait. In this case, the genomes without the trait of interest need not share all of the COGs, although they may share some. Such an

	Similarity Measure					
Perfect	0.8	0.75	0.7	0.65	One-needed	Similarity intersected with One-needed
1256	1256	1256	1256	1256	1261 1671	1261
1291	1291	1291	1291	1291	1261 1440 1447 1455	1283*
1344	1344	1344	1344	1344	1261 1315 1406	1334
1345	1345	1345	1345	1345	2747 1864	1639*
1360	1360	1360	1360	1360	2747 2604	1858
1516	1516	1516	1516	1516	1334 3015	2063
1558	1558	1558	1558	1558	1334 1626	2257
1580	1580	1580	1580	1580	1334 3157	2747
1677	1677	1677	1677	1677	1334 2113	3034
1815	1815	1815	1815	1815	1334 2717	
1843	1843	1843	1843	1843	2063 1582 1647 1728	
1868	1868	1868	1868	1868	1639* 0833	
1749	1749	1749	1749	1749	1639* 1969	
	1157	1157	1157	1157	2257 3031	
	1298	1298	1298	1298	1283* 2143	
	1338	1338	1338	1338	1564 1858 3034	
	1377	1377	1377	1377	1774 1858 3034	
	1536	1536	1536	1536	1699 1858 3034	
	1684	1684	1684	1684	1582 1858 30341647 1728	
	1766	1766	1766	1766		
	1886	1886	1886	1886		
	1987	1987	1987	1987		
	1419	1419	1419	1419		
	1706	1706	1706	1706		
	1317	1317	1317	1317		
		1334	1334	1334		
		1639*	1639*	1639*		
		2063	2063	2063		
		1191	1191	1191		
		1220	1220	1220		
			1551	1551		
			1221	1221		
				0667		
				0673		
				0762		
				0846		
				1261		
				1283*		
				1508		
				1739		
				1858		
				3034		
				2257		
				2747		
				0191		
				0444		
				1124		
				0455		
				0612		

Figure 2. Clusters of Orthologous Groups that Are Predicted to Have a Role in Flagella Biogenesis

Blue font denotes a COG with a flagellar annotation. Black font represents a COG that has no known flagellar role, and the red COGs are those that were knocked out in *B. subtilis*. The red asterisk designates those COGs that have a gene that, when knocked out in *B. subtilis*, results in an altered motility phenotype.

Table 1. Knocked Out *Bacillis subtilis* Genes Predicted to Be Involved in Flagella Assembly/Activity

COG ID	Gene	Size of ORF (bp)	COG Description
COG1283	*yqeW*	933	Na^+/phosphate symporter
COG2257	*ylqH*	282	Uncharacterized BCR homologous to the cytoplasmic domain of flagellar protein FlhB
COG3434	*yuxH*	1230	Previously COG1639 as uncharacterized BCR, now predicted signal transduction proteins containing EAL and modified HD-GYP domains

analysis could theoretically allow us to find genes that are required for a particular trait but that have been retained in species that don't exhibit that character. The application of this algorithm to the flagella trait did not appear to be useful in the identification of relevant genes (data not shown). However, as with all of the algorithms, the results might be quite different for other characters.

Similarity Measure Algorithm

The Similarity Measure algorithm permitted us to set a threshold for similarity in either traits or COG composition. This is accomplished by dividing the intersection of the subset of genomes represented by each COG and the subset of genomes with a particular trait by the union of these sets (Figure 1).

We set the initial threshold at 0.8 and then lowered it by 0.05 unit intervals to determine the effect on the algorithm's ability to correctly predict flagella COGs (Figure 2). With a threshold of 1, this algorithm reports the same COGs as the Perfect Match algorithm, as expected. However, as the threshold is lowered, additional COGs are reported that are either not in every genome with the trait of interest or are in some genomes without the trait. While this approach enabled us to approximate the appropriate threshold for the flagella analysis, it is predicated on functional annotation of target genes and presumably would have to be adjusted to different levels for each trait.

At several threshold levels, interesting aspects of flagella gene usage became apparent. For example, for the 12 new COGs that are reported by the Similarity Measure algorithm at a threshold of 0.8 (and are not reported by Perfect Match), 10 are present in a genome without flagella (*Chlamydia trachomatis*) and 2, COG1419 and COG1706, are represented in all bacteria with flagella, except for *E. coli* (COG1419) or *B. subtilis* (COG1706). All 12 COGs had flagellar annotations at the time of analysis. The reason for the presence in *C. trachomatis* of the ten COGs found primarily in organisms with flagella may be that *C. trachomatis* is an obligate intracellular parasite. It has been noted that some pathogenic bacteria use a type III secretory apparatus to export pathogenic compounds into the eukaryotic hosts. This is the same machinery that is used to export flagellar components by organisms with flagella [8–10]. Of the ten flagellar COGs that are reported with a threshold level of 0.8 and are represented in the *C. trachomatis* genome, nine have roles in the type III secretory systems.

When the threshold value was lowered to 0.75, the Similarity Measure algorithm predicted 30 flagellar COGs, of which 27 (90%) had flagellar annotations (Figure 2). Of the five new COGs that appeared as a result of adjusting the threshold from 0.8 to 0.75, each varied from a perfect score by only two elements. For instance, COGs 2063, 1639, and 1334 are present in all but two genomes with flagella. Three of the five COGs that have a nonflagellar annotation and, to our knowledge, no previously documented role in flagella biogenesis were the first to be reported by our algorithm. One of these COGs, COG1639, had an annotation of "uncharacterized bacterial conserved region" at the time of analysis, but the annotation and numerical designation have since changed to COG3434, with an annotation of "predicted signal transduction protein containing EAL and modified HD-GYP domains."

Nineteen additional COGs are reported when the Similarity Measure is reduced to 0.65 (Figure 2). At this setting, 29 (59%) of the 49 predicted COGs had flagellar annotations. This is 94% of the 31 known flagellar COGs.

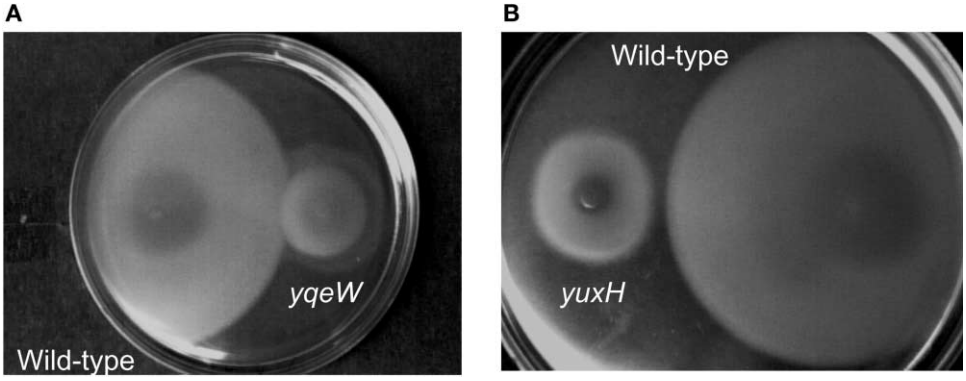

Figure 3. Motility Assay

Motility was assayed in swim plates by inoculating with sterile toothpicks and incubating overnight at 37°C in swim media (LB and 0.25% agar). Similar results were seen at 20°C over 4 days and at 30°C for 2 days (data not shown). *B. subtilis 168* (wild-type) is the control strain that was compared to the mutants (A) *yqeW* and (B) *yuxH*.

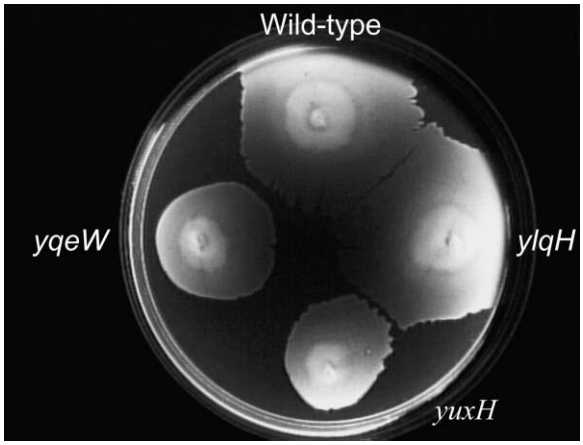

Figure 4. Motility Comparison

Motility on swarm plates (Tryptose Broth with 0.6% agar) was compared between the wild-type (*B. subtilis 168*) and the mutants. Sterile toothpicks were used to inoculate the swarm plates with overnight cultures, and the plates were incubated at 30°C for days.

As expected, lower threshold levels resulted in a greater number of presumed false positives. However, this conjecture can be confirmed only by experimentation, and it is possible that some of these predicted COGs could be involved in flagella development.

B. subtilis Knockout Experiments

Although we initially used the flagella trait to test the recall rate of our algorithms, it was surprising to us that the algorithm predicted that novel, highly conserved genes may be involved in this well-characterized system. By taking the intersection of the outputs of our two most successful algorithms (the One-needed and Similarity Measure algorithms), we generated a short list of nine COGs that included four known flagellar COGs and five newly predicted flagellar COGs (Figure 2). To test the ability of our approach to identify novel members of a pathway, we knocked out the genes in the flagellar organism *B. subtilis* that corresponded to three of those newly predicted COGs (Table 1).

As an assay for flagella function, motility in swim plates was compared between the wild-type and the mutants (see the Experimental Procedures). The wild-type and the *ylqH* (COG2257 deletion) strains took over the whole plate when grown at 37°C, so no significant difference was seen between this mutant and wild-type (data not shown). However, the *yqeW* (COG1283 deletion) and *yuxH* (COG1639 deletion) strains did not swim out as far from the inoculation point compared to the wild-type strain (Figure 3). Motility was also monitored on swarm plates. Once again, at 30° for 3 days, there was no visible difference between the wild-type and the *ylqH* strains, but the *yqeW* and *yuxH* strains did not swarm out as far from the inoculation point compared to the wild-type (Figure 4).

Direct microscopic observation was used (see the Experimental Procedures) to compare three determinants of the wild-type and mutant strains' motility. Again, there were no significant differences between *ylqH* and the wild-type strains (Table 2). Although *yqeW* mutants exhibited similar tumbling frequency to that of the wild-type, it was much slower in speed (Table 2). Conversely, the *yuxH* mutant strain maintained similar speed but displayed much lower tumble frequency than the wild-type. In addition, no significant difference in the growth rates or cell lengths between the mutants and the wild-type strains was observed (data not shown). Because neither *yqeW* nor *yuxH* belong to an operon, the knockouts should have no polar effects on downstream genes. Although the *ylqH* gene is likely part of a polycistronic operon with *ylqG* and possibly *rnh*, polar effects are not an issue since it is located at the 3′ end of the operon. Therefore, the phenotypes associated with the *yqeW* and *yuxH* mutants are solely due to disruption of the respective target loci.

The results of these assays suggest that the genes *yqeW* and *yuxH* in *B. subtilis*, which correspond to COGs 1283 and 3434 (previously 1639), respectively, are involved in flagellar function. COG1283 (*yqeW*) has the annotation "Na^+/phosphate symporter," and the role of sodium-driven motors in establishing ion gradients to power flagellar rotation has been shown in the Vibrionaceae family of bacteria [11, 12]. However, the annotation is based on homology to a Na^+/phosphate symporter, and there is no evidence for either sodium-dependent substrate accumulation or substrate-dependent sodium accumulation. Thus, this gene product may be necessary for proper flagella function in *B. subtilis*, but, at this time, its cellular role remains unclear. To our knowledge, the role of *yqeW* in flagellar function of *B. subtilis* has not been previously demonstrated. The second gene, *yuxH*, has an annotation of "predicted signal transduction proteins containing EAL and modified HD-GYP domains." To date, proteins in this family have not been investigated thoroughly, and, to our knowledge, the role of this signal transduction pathway has not been previously shown to be involved in bacteria motility [13]. However, some evidence suggests that the functional domains of these proteins may also be involved in the regulation of virulence factors in some pathogenic bacteria [13]. Finally, according to our analysis, knockouts of the last gene, *ylqH*, did not confer a motility phenotype (Table 2). This gene belongs to COG2257, which has the

Table 2. Chemotaxis Assay on Chemotaxis Buffer

	yqeW::pMutin4	*yuxH*::pMutin4	*ylqH*::SPC	168 (wild-type)
Tumble Frequency (tumbles/min ± SE)	10.88 ± 0.70	6.94 ± 0.57	13.94 ± 1.29	15.19 ± 1.01
Speed (cm/s ± SE)	1.82 ± 0.09	3.04 ± 0.12	3.20 ± 0.13	3.03 ± 0.13
Proportion of Motile Cells (%, n)	47.1, 136	54.5, 176	77.2, 184	81.9, 199

The chemotaxis assay was performed on chemotaxis buffer (0.1 mM EDTA, 50 μM $CaCl_2$, 0.05% glycerol, 5 mM sodium lactate, 0.3 mM ammonium sulfate, 20 mM potassium phosphate, [pH 7.0]).

annotation "uncharacterized BCR homologous to the cytoplasmic domain of flagellar protein FlhB." It is possible that, since *ylqH* only has homology to one domain of a flagellar protein, this domain may have been coopted into another protein that has a nonflagellar role. This is one potential limitation of the technique, since our algorithms are based on the COG database, which splits apart domains into different COGs. Alternatively, this gene may have a redundant role in the *B. subtilis* genome that is substituted by another endogenous gene product.

In conclusion, our algorithms were successful at recalling 29 out of the 31 known COGs involved in flagella development and appeared to infer the function of at least 2 new genes in this well-studied genetic system.

These results suggest that, as more fully sequenced genomes become available, these algorithms may provide an excellent resource for identifying novel genes in other traits of developmental, medical, or evolutionary significance.

Supplementary Material

Supplementary Material including the Experimental Procedures, Figure S1, which displays two growth curves for wild-type and mutant strains, and Table S1, which displays the character matrix described in the text, is available at http://images.cellpress.com/supmat/supmatin.htm.

Acknowledgments

We thank Ken Birnbaum for his help in interpreting the data and editing the manuscript. The work presented here was supported by grants from the National Science Foundation AT2010 project and Canadian Institutes of Health Research grant MOP-49606; M.L. was supported by a National Science Foundation Predoctoral Fellowship.

Received: July 24, 2002
Revised: October 31, 2002
Accepted: November 5, 2002
Published: January 21, 2003

References

1. 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.
2. McGuire, A.M., Hughes, J.D., and Church, G.M. (2000). Conservation of DNA regulatory motifs and discovery of new motifs in microbial genomes. Genome Res. *10*, 744–757.
3. Tatusov, R.L., Natale, D.A., Garkavtsev, I.V., Tatusova, T.A., Shankavaram, U.T., Rao, B.S., Kiryutin, B., Galperin, M.Y., Fedorova, N.D., and Koonin, E.V. (2001). The COG database: new developments in phylogenetic classification of proteins from complete genomes. Nucleic Acids Res. *29*, 22–28.
4. Aldridge, P., and Hughes, K.T. (2002). Regulation of flagellar assembly. Curr. Opin. Microbiol. *5*, 160–165.
5. Harshey, R.M., and Toguchi, A. (1996). Spinning tails: homologies among bacterial flagellar systems. Trends Microbiol. *4*, 226–231.
6. DeRosier, D.J. (1998). The turn of the screw: the bacterial flagellar motor. Cell *93*, 17–20.
7. Macnab, R.M. (1999). The bacterial flagellum: reversible rotary propeller and type III export apparatus. J. Bacteriol. *181*, 7149–7153.
8. Subtil, A., Blocker, A., and Dautry-Varsat, A. (2000). Type III secretion system in *Chlamydia* species: identified members and candidates. Microbes Infect. *2*, 367–369.
9. Galan, J.E., and Collmer, A. (1999). Type III secretion machines: bacterial devices for protein delivery into host cells. Science *284*, 1322–1328.
10. Aizawa, S.-I. (2001). Bacterial flagella and type III secretion systems. FEMS Microbiol. Lett. *202*, 157–164.
11. McCarter, L.L. (2001). Polar flagellar motility of the *Vibrionaceae*. Microbiol. Mol. Biol. Rev. *65*, 445–462.
12. Larsen, S.H., Adler, J., Gargus, J.J., and Hogg, R.W. (1974). Chemomechanical coupling without ATP the source of energy for motility and chemotaxis in bacteria. Proc. Natl. Acad. Sci. USA *71*, 1239–1243.
13. Galperin, M.Y., Nikolskaya, A.N., and Koonin, E.V. (2001). Novel domains of the prokaryotic two-component signal transduction systems. FEMS Microbiol. Lett. *203*, 11–21.