Accurate Prediction of Orthologous Gene Groups in Microbes

Hongwei Wu*[1, 2], Fenglou Mao*[1], Victor Olman[1] and Ying Xu[1, 2]
*These authors equally contribute to this work.
[1]Department of Biochemistry and Molecular Biology and Institute of Bioinformatics
University of Georgia, Athens, GA 30622
[2]Computational Biology Institute, Oak Ridge National Laboratory, Oak Ridge, TN 37831
Email: {hongweiw, fenglou, olman, xyn}@csbl.bmb.uga.edu

Abstract

We present a new computational method for the prediction of orthologous gene groups for microbial genomes based on the prediction of co-occurrences of homologous genes. The method is inspired by the observation that homologous genes are highly likely to be orthologous if their neighboring genes are also homologous. Based on co-occurrences of homologous genes, we have grouped the (predicted) operons of 77 selected sequenced microbial genomes so that operons of the same group are highly likely to be functionally similar or related. We then cluster the homologous genes in the same operon group so that genes of the same cluster are highly likely to be similar in terms of their sequences and functions, i.e., they are predicted to be orthologous genes. By comparing our predicted orthologous gene groups with the COG assignments and NCBI annotations, we conclude that our method is promising to provide more accurate and specific predictions than the existing methods.

Supplementary materials: *http://csbl.bmb.uga.edu/~fenglou/GFDB/suppl.html*

1. Introduction

In the past few years, we have witnessed a rapidly widening gap between the number of genes that have been identified through worldwide efforts in genome sequencing and bioinformatics prediction and the number of genes that have been experimentally studied. Computational methods are clearly becoming the only technique for gene function characterization that could possibly keep up with the sequencing efforts and computational gene identification. One of the very basic techniques for gene function prediction is through identification of orthologous genes. Bidirectional Best BLAST Hits (BDBH) and its more sophisticated derivations, particularly Cluster of Orthologous Groups (COG) [1], are among the popular techniques for orthology prediction. Though very successful, COG has a number of limitations, including:

- *The classification provided by COG is often not specific enough.* On one hand, the same COG number may be assigned to genes of similar yet distinct functions. To name a few, COG0642 contains different types of sensor proteins, e.g., *baeS* (sensor for drugs), *phoR* (sensor for phosphorus assimilation), *envZ* (osmolarity sensor protein), *phoQ* (resistance sensor to environments), and *creC* (catabolite repression sensor kinase for *phoB*); and COG0745 contains different types of regulator proteins that are associated with these sensor genes, e.g., *baeR*, *phoB*, *ompR*, *phoP* and *creB*. On the other hand, the same gene may be assigned with multiple COG numbers. For example, we have found that the 11 genes in Test Case 1 (see Section 3) are simultaneously assigned with COG1226 (kef-type K+ transport systems, predicted NAD-binding component) and COG0569 (trk-type K+ transport systems, NAD-binding component).
- *COG does not provide predictions of functional relationship among different COG groups*. For example, without referring to their detailed annotations, it is hardly possible to relate COG0674, COG1013, COG1144 and COG1014 together, which are the alpha, beta, delta and gamma subunits of the ferredoxin oxidoreductase, respectively; or to relate COG2025 and COG2086 together, which are the alpha and beta domains of the electron transfer flavoprotein, respectively.

Observing that orthologous genes often co-occur with other orthologous genes in the same neighborhoods (e.g., in the same operons), we have developed a new computational method for prediction of orthologous genes for microbial genomes, based on the prediction of co-occurrences of homologous genes. Our preliminary study on 77 selected microbial genomes has shown that our method is very promising to overcome the aforementioned limitations of the COG assignments. The ultimate goal of our study is to build a new classification system of orthologous gene groups for all microbial genomes.

2. Materials and Methods

The basic idea of our method is that the prediction of orthologous genes should be supported by both sequence similarity and functional similarity/relatedness. While it is relatively straightforward to check the similarity level between two sequences, it is challenging to determine to what extent two genes are functionally similar or related through computational methods. We predict genes' functional similarity/relatedness based on the prediction of functional similarities/relatedness among corresponding *operons*, which is in turn based on the prediction of *homologous gene co-occurrences* and *homologous gene co-occurrence triangles* (defined below). For the study presented in this paper, we have selected 77 microbial genomes (as summarized in Table 1) in such a way that each genome belongs to a different *genus*; and, two genes of different genomes are considered to be homologous if and only if their bi-directional BLASTP [2] searches both have e-value smaller than 10^{-6}.

There have been numerous efforts devoted to the prediction of operons through computational methods, e.g. [3,7]. For the study presented in this paper, we have used our own operon prediction program JPOP [3]. JPOP can reach a prediction accuracy level of 83.3% when benchmarked against the known operons of *Escherichia coli* K12. The predicted operons for the 77 selected genomes are summarized in Table 1.

Table 1. The number of genes, the number of predicted operons (not including the single-gene operons), and the number of genes covered by the predicted operons for the 77 selected genomes.

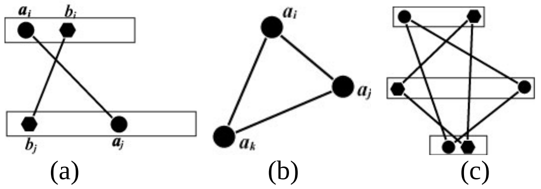
Genome	No. of genes	No. of predicted operons	No. of covered genes
Aeropyrum pernix K1	1841	193	519
Agrobacterium tumefaciens str. C58 chromosome circular (Ceron)	5293	307	849
Aquifex aeolicus VF5	1560	315	881
Archaeoglobus fulgidus DSM 4304	2420	420	1190
Bacillus anthracis A2012	5852	715	1915
Bacteroides thetaiotaomicron VPI-5482	4778	456	1158
Bifidobacterium longum NCC2705	1727	217	563
Candidatus Blochmannia floridanus	583	87	280
Bordetella bronchiseptica RB50	4994	790	2374
Borrelia burgdorferi B31	1640	140	437
Bradyrhizobium japonicum USDA 110	8317	1147	3139
Brucella melitensis 16M	3198	343	821
Buchnera aphidicola str. APS (Acyrthosiphon pisum)	574	101	300
Campylobacter jejuni subsp. jejuni NCTC 11168	1634	321	1047
Caulobacter crescentus CB15	3737	529	1417
Chlamydophila caviae GPIC	1005	148	391
Chlorobium tepidum TLS	2252	297	786
Chromobacterium violaceum ATCC 12472	4407	615	1726
Clostridium acetobutylicum ATCC824	3848	517	1472
Corynebacterium glutamicum ATCC 13032	2993	402	1083
Coxiella burnetii RSA 493	2010	263	753
Deinococcus radiodurans R1	3182	360	857
Enterococcus faecalis V583	3113	389	1068
Escherichia coli K12	4242	594	1709
Fusobacterium nucleatum subsp. nucleatum ATCC 25586	2067	347	1053
Gloeobacter violaceus PCC 7421	4430	441	1107
Haemophilus influenzae Rd KW20	1657	331	944
Halobacterium sp. NRC-1	2622	249	649
Helicobacter hepaticus ATCC 51449	1875	287	829
Lactobacillus plantarum WCFS1	3009	379	1057
Lactococcus lactis subsp. lactis Il1403	2421	337	912
Leptospira interrogans serovar Lai str. 56601	4727	363	965
Listeria innocua Clip11262	3043	494	1475
Mesorhizobium loti MAFF303099	7275	907	2562
Methanobacterium thermoautotrophicum str. Delta H	1873	339	1054
Methanocaldococcus jannaschii DSM 2661	1785	287	742

Methanopyrus kandleri AV19	1687	248	732
Methanosarcina acetivorans str. C2A	4540	438	1144
Mycobacterium tuberculosis H37Rv	3927	572	1555
Mycoplasma penetrans HF-2	1037	130	377
Nanoarchaeum equitans Kin4-M	536	62	147
Neisseria meningitidis serogroup A strain Z2491	2065	268	710
Nitrosomonas europaea ATCC 19718	2461	365	1041
Nostoc sp. PCC 7120	6055	417	986
Oceanobacillus iheyensis HTE831	3500	473	1356
Pasteurella multocida Pm70	2015	377	1126
Photorhabdus luminescens subsp. laumondii TTO1	4683	552	1591
Pirellula sp. 1	7325	448	1071
Porphyromonas gingivalis W83	1909	234	609
Prochlorococcus marinus str. MIT 9313	2265	234	605
Pseudomonas aeruginosa PA01	5567	907	2636
Pyrobaculum aerophilum str. IM2	2605	321	816
Pyrococcus abyssi GE5	1896	336	952
Ralstonia solanacearum GMI1000	5116	613	1555
Rickettsia conorii str. Malish 7	1374	163	409
Salmonella typhimurium LT2	4527	678	1984
Shewanella oneidensis MR-1	4472	526	1436
Shigella flexneri 2a str. 301	4180	689	1821
Sinorhizobium meliloti 1021	6205	541	1342
Staphylococcus aureus subsp. aureus Mu50	2748	416	1183
Streptococcus pneumoniae TIGR4	2094	358	1028
Streptomyces coelicolor A3(2)	8154	818	2148
Sulfolobus solfataricus P2	2977	399	1040
Synechococcus sp. WH 8102	2517	306	789
Thermoanaerobacter tengcongensis strain MB4T	2588	390	1238
Thermoplasma volcanium GSS1	1499	250	671
Thermosynechococcus elongatus BP-1	2475	302	752
Thermotoga maritima MSB8	1858	365	1227
Treponema pallidum	1036	144	392
Tropheryma whipplei str. Twist	808	139	397
Ureaplasma parvum serovar 3 str. ATCC 700970	614	93	276
Vibrio parahaemolyticus RIMD 2210633	4832	493	1232
Wigglesworthia glossinidia endosymbiont of Glossina brevipalpis	611	114	322
Wolinella succinogenes DSM 1740	2044	374	1162
Xanthomonas axonopodis pv. citri str. 306	4312	546	1457
Xylella fastidiosa 9a5c	2832	316	871
Yersinia pestis strain CO92	4067	617	1718

We first provide a few definitions here. Two homologous gene pairs, (a_i, a_j) and (b_i, b_j), with a_i and b_i being from the *i*-th genome, and a_j and b_j from the *j*-th genome, are called a *homologous gene co-occurrence*, if a_i and b_i are in the same operon O_i and a_j and b_j are in the same operon O_j. A triple of homologous genes (a_i, a_j, a_k) is called to form a *homologous gene triangle*, if (a_i, a_j), (a_i, a_k) and (a_j, a_k) are all homologous pairs. Two homologous gene triangles, (a_i, a_j, a_k) and (b_i, b_j, b_k), are called to form a *homologous co-occurrence triangle*, if (a_i, a_j) and (b_i, b_j), (a_i, a_k) and (b_i, b_k), as well as (a_j, a_k) and (b_j, b_k) all form homologous gene co-occurrences. These three definitions are illustrated in Figure 1.

We describe the operons and their relationships (in terms of homologous gene co-occurrences) by using a graph representation where operons are represented as nodes and homologous gene co-occurrences between operons are represented as edges connecting the nodes. In this representation, two homologous co-occurrence triangles are called *related* if they share a common edge. Each transitive closure of this *related* relationship defines an *operon group*. We then describe the genes and their relationships (in terms of homology) in each operon group by using a graph representation where genes are represented as nodes and homologous relationships are represented as edges. We consider two homologous gene triangles as *related* if they share a common edge. We call each transitive closure of this *related* relationship a *homologous gene group*. In the graph representation of each homologous gene group (with nodes for genes and edges for homologous relationships), we consider each *densely connected* cluster (sub-graph) to be an *orthologous gene group*, where *density* is controlled by the granularity parameter chosen during the clustering (as explained below).

While operon groups and homologous gene groups can be determined non-parametrically, identification of orthologous gene groups requires a cutoff value to be given, which controls the connection densities of clusters. By using the Markov clustering algorithm [http://micans.org/mcl/] with different granularity levels ranging from 2.0 to 5.0 [4], we have predicted orthologous groups with different connection densities, which reflects a natural hierarchical classification of genes. We have observed that the prediction is very consistent with our general understanding about *orthology* when the granularity level 5.0 is used; hence, we have only included in this paper the results for this particular choice of the granularity level. The genes in the same orthologous group are predicted to have the same function, which means that the functions of the genes belonging to the same orthologous group are all known once one of them is known. We believe that our prediction of orthologous groups is cleaner and more effective for predictions of gene functions than the concept of *paralogs*.



(a) (b) (c)

Figure 1. A schematic illustration of (a) homologous gene co-occurrence, (b) homologous gene triangle, and (c) homologous co-occurrence triangle, where a box represents an operon.

3. Results and Discussions

Our prediction method leads to the clustering of genes at three different resolution levels --- operon groups, homologous gene groups and orthologous gene groups, respectively. These groups represent a hierarchical classification of genes, in terms of their functional relatedness. Proteins included in the same operon group are functionally related, e.g., they work together in the same biological process. We have observed that for most cases proteins are included in the same homologous group (but not in the same orthologous gene group) either due to Rosetta-stone proteins (when genes correspond to different domains of a protein complex and there is a gene fusion occurring in some genomes) or due to paralogy.

For the 45,432 genes of the 77 selected genomes that are predicted to be part of some operons, we have obtained 1,011 operon groups, 3,177 homologous gene groups and 5,636 orthologous gene groups. In this paper, we discuss three examples to demonstrate the effectiveness of our method.

Test case 1: The trk-type K+ transport system has two components, a NAD-binding component and a membrane component. We have detected homologous gene co-occurrences of these two components in 22 genomes in which these two genes are predicted to be in the same operon (see Figure 2). We have clustered 23 proteins into one orthologous gene group corresponding to the NAD-binding component (denoted as Group 1) and 25 proteins into another orthologous gene group corresponding to the membrane component (denoted as Group 2), as summarized in the supplementary materials.

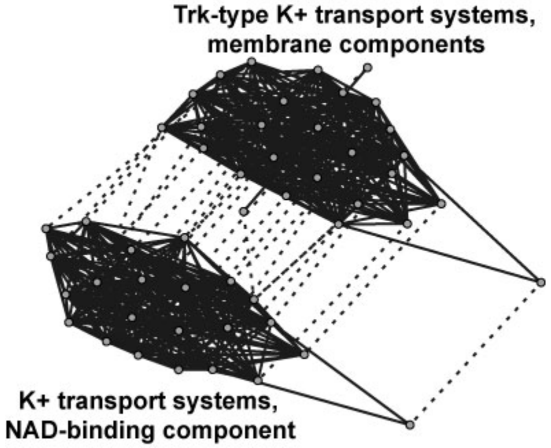

Trk-type K+ transport systems, membrane components
K+ transport systems, NAD-binding component

Figure 2. The operon group consisting of trk-type K+ transport system proteins. A solid link represents that the two connected genes are homologous, and a dashed link represents that the two connected genes are in the same operon.

We have observed from the COG assignments of these genes that (1) 11 genes of Group 1 are simultaneously assigned with two COG numbers, COG1226 (kef-type K+ transport systems predicted NAD-binding component) and COG0569 (trk-type K+ transport systems NAD-binding component), and the other 12 genes of Group 1 are assigned with COG0569; and (2) all the 25 genes in Group 2 are assigned with COG0168 (trk-type K+ transport systems membrane component). While our prediction for Group 2 is consistent with the COG assignments, we believe that all genes in Group 1 should be assigned with COG0569 rather than COG1226 for the following reason: for each gene in Group 1 we can find an accompanying trk-type K+ transport membrane component in the same operon, which provides a strong evidence for the genes in Group 1 to be orthologous. We have also observed from the

NCBI annotations that seven genes in Group 2 are annotated as Na+ transport system proteins. We believe that these NCBI annotations are incorrect, because these seven genes are always within the same operons as the genes for the K+ transport proteins, as supported by the COG assignments.

We have performed multiple sequence alignment (see the supplementary materials) to verify our prediction for both Groups 1 and 2. The proteins within the same group are perfectly aligned except for two proteins in Group 2, 23099118 and 23099119. These two genes correspond to the N- and C-terminal part of the membrane component, respectively, and their combination is perfectly aligned to all the other proteins in Group 2, indicating that our prediction is supported by the multiple sequence alignment.

One operon duplication event, one gene duplication event and one gene fission event have been identified through our prediction. For *Shewanella oneidensis* MR-1, we have found two sets of trk-type K+ transport genes, {24371657, 24371658} and {24375763, 24375764}, which we believe to represent an operon duplication event. We have found that two adjacent genes of *Deinococcus radiodurans* R1, 15806670 and 15806671, both correspond to the K+ membrane component, which we believe to represent a gene duplication event. Also, we have found that both 23099118 and 23099119 of *Oceanobacillus iheyensis* HTE831 are the fission results of the membrane component protein.

Test case 2: The electron transfer flavoprotein has two domains, alpha and beta. In most genomes they are encoded by two different genes, but in *Sulfolobus solfataricus* P2 they are fused into one gene (15899533). We have predicted that the alpha- and beta-domain genes as well as the fused gene all belong to the same homologous group (see Figure 3). This homologous group consists of 106 genes covering 39 genomes, among which 53 are annotated as electron transfer flavoprotein alpha-subunit (alpha-annotated), 52 are annotated as electron transfer flavoprotein beta-subunit (beta-annotated), and the remaining one is annotated as electron transfer flavoprotein alpha and beta-subunit (alpha-beta-annotated). We have observed that (1) within the same genome the alpha- and beta-genes are always in the same operon; and, (2) the alpha-beta annotated gene of *Sulfolobus solfataricus* P2 (15899533) is homologous to most alpha- as well as to beta-genes, as shown in Figure 3.

We can therefore infer from this prediction that the alpha- and beta-annotated genes are functionally closely related. By applying the clustering algorithm, we have obtained three separate orthologous gene groups (see the supplementary materials) corresponding to the alpha-, beta- and alpha-beta annotated genes, respectively. Our prediction for the gene fusion is comparable to the method in [5 6] that predicts gene fusion events through sequence alignment. Compared to the method in [5 6], our prediction is promising to be highly accurate in predicting both orthologous gene groups as well as gene fusion events, because we have incorporated both sequence similarities and functional relatedness/similarities of genes into the prediction.

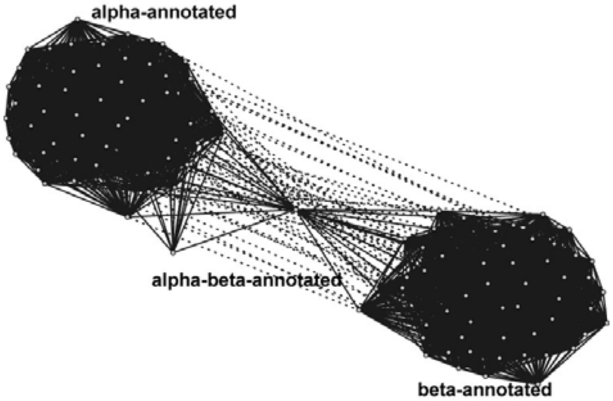


Figure 3. The homologous group for the electron transfer flavoprotein, where the two densely connected clusters correspond to the alpha and beta domains, respectively. The protein between the two clusters is the Rosetta-stone protein.

We have also been able to predict that five COG groups --- COG0674, COG1013, COG1014, COG1144 and COG4231 --- are functionally related/similar, since they belong to the same homologous group (see Figure 4). COG0674, COG1013, COG1014 and COG1144 correspond to the alpha, beta, delta and gamma subunits of the ferredoxin oxidoreductase/paralogs, respectively; and COG4231 corresponds to the indolepyruvate oxidoreductase alpha subunit. By applying the clustering algorithm, we have clustered the 195 proteins of this homologous group into 13 orthologous groups (as summarized in the supplementary materials). We have been able to assign very specific annotations to the 10 large orthologoous groups by referring to their consensus NCBI annotations, which correspond to (1) the 2-oxoacid ferredoxin oxidoreductase alpha subunit, (2) the 2-oxoacid ferredoxin oxidoreductase beta subunit, (3) the pyruvate ferredoxin oxidoreductase alpha subunit, (4) the pyruvate ferredoxin oxidoreductase beta subunit, (5) the pyruvate ferredoxin oxidoreductase gamma subunit, (6) the pyruvate ferredoxin oxidoreductase delta subunit, (7) the 2-ketoglutarate ferredoxin oxidoreductase gamma subunit, (8) the indolepyruvate oxidoreductase alpha subunit, (9) the indolepyruvate oxidoreductase beta subunit, and (10) the 2-oxoisovalerate oxidoreductase beta subunit, respectively. We have also been able to identify two

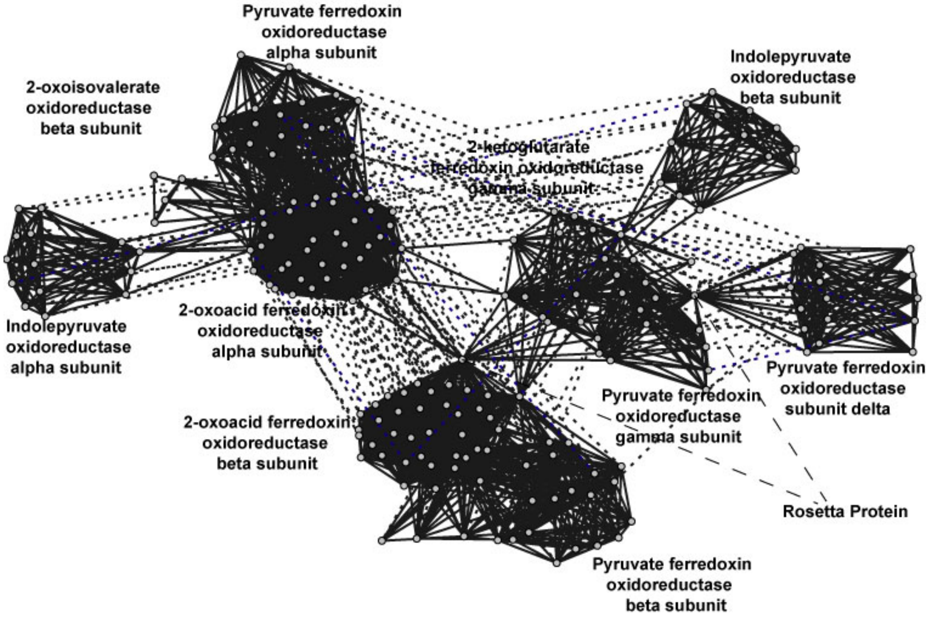

Pyruvate ferredoxin oxidoreductase alpha subunit
2-oxoisovalerate oxidoreductase beta subunit
Indolepyruvate oxidoreductase beta subunit
2-ketoglutarate ferredoxin oxidoreductase gamma subunit
Indolepyruvate oxidoreductase alpha subunit
2-oxoacid ferredoxin oxidoreductase alpha subunit
Pyruvate ferredoxin oxidoreductase subunit delta
2-oxoacid ferredoxin oxidoreductase beta subunit
Pyruvate ferredoxin oxidoreductase gamma subunit
Rosetta Protein
Pyruvate ferredoxin oxidoreductase beta subunit

Figure 4. The homologous group for the ferredoxin oxidoreductase and its paralogs.

proteins of *Methanobacterium thermoautotrophicum* str. Delta H, 15678732 and 15679732, each of which is included in a single-gene orthologous group, as Rosetta-stone proteins. When referring to the COG assignments of these proteins, we have observed that (1) some genes corresponding to the indolepyruvate oxidoreductase alpha subunit are assigned with COG4231 while the other such annotated genes are assigned with COG0674; and, (2) all the beta subunit genes are assigned with COG1014, all the delta subunit genes are assigned with COG1144, and all the gamma subunit genes are assigned with COG1014. We believe that our annotations for these genes are more accurate than their COG assignments, because COG is trying to distinguish between indolepyruvate oxidoreductase proteins and their paralogs (as revealed by the fact that the indolepyruvate oxidoreductase alpha subunit genes are assigned with a different COG number than the other alpha subunit genes), while it fails to distinguish between indolepyruvate oxidoreductase beta subunit genes and their paralogs. Also, our predictions are more consistent with the NCBI annotations than the COG assignments are.

Test case 3 We have predicted two homologous gene groups that belong to the same operon group and correspond to the sensor and regulator genes of the sensor-regulator two-component systems, respectively. We have clustered the sensor genes (360 genes from 52 genomes) into multiple orthologous gene groups corresponding to *baeS phoR envZ phoQ creC colS rstB kdpD cpxA* and some unknown functions respectively; and, the transcription regulators genes (360 genes from 52 genomes) that are associated with these sensor genes into multiple orthologous gene groups corresponding to *baeR phoB ompR phoP creB colR rstA kdpE cpxR* and some unknown functions, respectively (see the supplementary materials). Our clustering of the sensor genes and the regulator genes seem to be significantly better than their COG assignments, as explained below.

Most sensor genes are assigned into two different COG groups, COG0642 and COG2205; in contrast, their associated regulator genes are assigned into just one COG group, COG0745. We believe that the COG assignments of these genes, especially of these regulator genes, are not clear enough to make high-resolution function predictions. Through literature search we know *baeS* is the sensor gene of bacteria for drugs, *phoR* is the sensor gene for low phosphorous concentration, *envZ* is the sensor for environment osmolarity, *phoQ* is the sensor for low Mg2+ environments, *creC* is the sensor for carbon catabolite repression, *colS* plays an important role in the root-colonizing ability, and *cpxA* is the sensor for various cell envelope stresses. These two-component (sensor and regulator) systems are playing different roles though sometimes their functions overlap to some extent. However, assigning all these genes, especially all these regulator genes, into the same group is clearly not specific enough. By applying our method we have not only been able to predict all the sensor genes into one homologous group and all regulator genes into another homologous group, but have also been able to further cluster the sensor genes and their associated regulator genes into different orthologous groups (as summarized in the supplementary materials). We have also predicted several sensor and regulator orthologous gene groups that we believe worthy of experimental investigations.

4. Summary

We have developed a new method for the prediction of orthologous gene groups for microbial genomes based on the prediction of homologous gene co-occurrences. Besides the orthologous gene groups we have also predicted operon groups and homologous groups, where an operon group consists of a group of operons whose genes work together in the same biological process, and a homologous gene group consists of a group of genes that correspond to different domains of a protein complex or a group of paralogous genes. This hierarchical structure of prediction allows us to identify functional links across different orthologous genes, and makes it possible to predict component genes of specific biological pathways or networks. We have observed that many of our predicted orthologous gene groups are consistent with COG assignments though some of our predictions are more specific than COG assignments.

The coverage rate of our method for the prediction of orthologous gene groups of microbial genomes, however, is bounded by the coverage rate of the operon prediction method. In our future study, we plan to generalize the concept of operons in order to increase the coverage rate of our method.

Acknowledgement

This work was supported in part by the US Department of Energy's Genomes to Life Program under project "Carbon Sequestration in *Synechococcus sp*: From Molecular Machines to Hierarchical Modeling" (http://www.genomes-to-life.org). The work is also supported in part by National Science Foundation (#NSF/DBI-0354771 #NSF/ITR-IIS-0407204).

References

[1] Tatusov R.L. E.V. Koonin and D.J. Lipman *A genomic perspective on protein families.* Science 1997. 278 (5338): p. 631-7.

[2] Altschul S.F. et al. *Gapped BLAST and PSI-BLAST: a new generation of protein database search programs.* Nucleic Acids Res. 1997. 25 (17): p. 3389-402.

[3] X. Chen Z.S. Ying Xu and T Jiang *Computational Prediction of Operons in Synechococcus sp WH8102.* Proceedings of 15th International Conference on Genome Informatics 2004: p. 211 - 222.

[4] von Dongen S. *Graph Clustering by Flow Simulation.* Ph.D. dissertation University of Utrecht 2000.

[5] Marcotte E.M. et al. *Detecting protein function and protein-protein interactions from genome sequences.* Science 1999. 285 (5428): p. 751-3.

[6] Suhre K. and J.M. Claverie *FusionDB: a database for in-depth analysis of prokaryotic gene fusion events.* Nucleic Acids Res. 2004. 32 (Database issue): p. D273-6.

[7] Ermolaeva M. D. O. White and S. L. Salzberg *Prediction of operons in microbial genomes.* Nucleic Acids Res. 2001. 29 (5): p. 1216-21.