CAN ENDS JUSTIFY THE MEANS? DIGGING DEEP FOR HUMAN FUSION GENES OF PROKARYOTIC ORIGIN

Yu Yiting, Iti Chaturvedi, Liew Kim Meow, Pandjassarame Kanguane and Meena Kishore Sakharkar

School of Mechanical and Production Engineering, Nanyang Center for Supercomputing and Visualization, Nanyang Technological University, Singapore 639798

TABLE OF CONTENTS

1. Abstract
2. Introduction
3. Materials and methods
   3.1. Description of datasets
     3.1.1. Dataset 1 (DS1)
     3.1.2. Dataset 2 (DS2)
   3.2. Identification of fusion proteins
   3.3. Functional inferences to fusion proteins
   3.4. Availability
4. Results and discussion
   4.1. Fusion proteins mimicking operons in prokaryotes
   4.2. Fusion proteins exhibiting multiple functions
   4.3. Fusion proteins showing alternative splicing
   4.4. Fusion proteins simulating protein-protein interfaces in prokaryotes
5. Conclusion
6. Acknowledgements
7. References

1. ABSTRACT

Gene fusion has been described as an important evolutionary phenomenon. This report focuses on identifying, analyzing, and tabulating human fusion proteins of prokaryotic origin. These fusion proteins are found to mimic operons, simulate protein-protein interfaces in prokaryotes, exhibiting multiple functions and alternative splicing in humans. The accredited biological functions for each of these proteins is made available as a database at http://sege.ntu.edu.sg/wester/fusion/

2. INTRODUCTION

Gene fusion is a phenomenon that has generated much curiosity since its description. Fusion genes gain added advantage in higher organisms by coupling biochemical/signal transduction reactions through tight regulation of fusion partners, compared to individual fusion partners in lower organisms (1). Thus, fusion genes produce proteins with novel or enhanced function. Gene fusion is believed to occur by gene transfer and gene fusion. The transfer of genes and bringing together of genes from two genomes into a single gene (gene fusion) has long been identified as a potentially important evolutionary phenomenon (2). The human genome project shows that a small fraction of human genes (<1%) is exclusively homologous to bacterial genes (3). Though, lateral gene transfer (4) and differential loss of genes (5) have been described to account for the presence of bacterial genes in the human genome, the frequencies of these transfers remain a subject of conjecture (6). Functional and physical associations between fusion partners and fusion products have been discussed earlier (2, 7). Two opposing forces work in palindrome: one that shuffles the genome and the other that prevents the shuffle by gene fusion. Thus, fusion genes are treated as one unit, working in synergy to achieve optimal functionality.

Gene fusion has been identified across various phylogenetic groups and this suggests that there exist processes other than vertical inheritance during evolution (8). Yanai and colleagues used gene fusion to establish links between fusion genes and functional network of their involvement (9). Gene fusion has also been used to illustrate protein-protein interactions (7), novel gene function (2), enhanced substrate specificity (10) and multifunctional enzyme specificity (11). An interesting relational algebra approach has also been demonstrated to identify fusion proteins across different phylogenetic distances (12). Therefore, identification and characterization of fusion genes in the human genome will shed light into its evolutionary biology. Herein, we report human fusion genes of which many are found to mimic prokaryotic operons and simulate protein-protein interfaces. Few others are also known to exhibit multiple functions and alternative splicing.

3. MATERIALS AND METHODS

3.1. Description of datasets

3.1.1. Dataset 1 (DS1)

The 37,490 protein sequences derived from the draft human genome obtained from NCBI (ftp://ftp.ncbi.nih.gov/genomes/H_sapiens) form DS1. The paralogs in the human genome data are removed at 40%
## Table 1. Human fusion proteins of prokaryotic origin

<table>
<thead>
<tr>
<th>RefSeq Accession</th>
<th>Chromosome Number</th>
<th>Protein Name</th>
<th>Pathway</th>
<th>Protein Length</th>
<th>Match Region</th>
<th>Protein Name</th>
<th>Pathway</th>
<th>Protein Length</th>
<th>Match Region</th>
<th>Comments</th>
</tr>
</thead>
<tbody>
<tr>
<td>NP_0004276</td>
<td>1</td>
<td>Glucose-1-dehydrogenase (in S. acidocaldarius)</td>
<td>Carbohydrate Metabolism</td>
<td>479</td>
<td>27-510</td>
<td>Glucose-6-phosphate dehydrogenase</td>
<td>Carbohydrate Metabolism (in E. coli)</td>
<td>261</td>
<td>555-751</td>
<td>untreated for in E. coli (29)</td>
</tr>
<tr>
<td>NP_000173</td>
<td>2</td>
<td>Hydroxyl-acyl dehydrogenase (in S. acidocaldarius)</td>
<td>Carbohydrate Metabolism</td>
<td>258</td>
<td>46-278</td>
<td>Enol-CoA hydratase</td>
<td>Carbohydrate Metabolism (in E. coli)</td>
<td>411</td>
<td>440-750</td>
<td>untreated for in E. coli (30)</td>
</tr>
<tr>
<td>NP_004332</td>
<td>2</td>
<td>Carboxylase dehydrogenase</td>
<td>Carbohydrate Metabolism</td>
<td>423</td>
<td>1446-1790</td>
<td>3-Hydroxyacyl-CoA dehydrogenase</td>
<td>Lipid Metabolism (in E. coli)</td>
<td>305</td>
<td>921-2229</td>
<td>untreated for in E. coli (31)</td>
</tr>
<tr>
<td>NP_001150</td>
<td>2</td>
<td>Aldolase</td>
<td>Amino Acid Metabolism</td>
<td>489</td>
<td>9-531</td>
<td>Carbon-mono dehydrogenase</td>
<td>Amino Acid Metabolism (in E. coli)</td>
<td>799</td>
<td>580-1316</td>
<td>untreated for in E. coli (32)</td>
</tr>
<tr>
<td>NP_0001957</td>
<td>3</td>
<td>3-Phospho-adenosine 5'-phosphosulfate synthase</td>
<td>Nucleotide Metabolism</td>
<td>297</td>
<td>5-277</td>
<td>Enol-CoA hydratase</td>
<td>Amino Acid Metabolism (in E. coli)</td>
<td>411</td>
<td>302-705</td>
<td>untreated for in E. coli (33)</td>
</tr>
<tr>
<td>NP_0001059</td>
<td>3</td>
<td>DNA topoisomerase</td>
<td>DNA Topoisomerase</td>
<td>773</td>
<td>56-727</td>
<td>DNA gyrase B subunit</td>
<td>Amino Acid Metabolism (in E. coli)</td>
<td>490</td>
<td>728-1054</td>
<td>untreated for in E. coli (34)</td>
</tr>
<tr>
<td>NP_000929</td>
<td>4</td>
<td>DNA directed RNA polymerase II</td>
<td>Amino Acid Metabolism</td>
<td>550</td>
<td>28-535</td>
<td>DNA-directed RNA polymerase subunit B</td>
<td>Amino Acid Metabolism (in E. coli)</td>
<td>649</td>
<td>565-1172</td>
<td>untreated for in E. coli (35)</td>
</tr>
<tr>
<td>NP_0005434</td>
<td>4</td>
<td>3-Phospho-adenosine 5'-phosphosulfate synthase</td>
<td>Carbohydrate Metabolism</td>
<td>186</td>
<td>51-221</td>
<td>Adenosine 5'-phosphosulfate</td>
<td>Nucleotide Metabolism (in E. coli)</td>
<td>459</td>
<td>273-619</td>
<td>untreated for in E. coli (36)</td>
</tr>
<tr>
<td>NP_0006443</td>
<td>4</td>
<td>AIR carboxylase</td>
<td>Carbohydrate Metabolism</td>
<td>533</td>
<td>12-247</td>
<td>Phospho dihydroxyacetone synthase</td>
<td>Nucleotide Metabolism (in B. subtilis)</td>
<td>180</td>
<td>267-421</td>
<td>untreated for in E. coli (37)</td>
</tr>
<tr>
<td>NP_000427</td>
<td>5</td>
<td>Succinyl CoA:3-oxoacyl-CoA transferase</td>
<td>Metabolism of Fatty Acids</td>
<td>250</td>
<td>36-284</td>
<td>Acyl-CoA synthase transferase</td>
<td>Amino Acid Metabolism (in E. coli)</td>
<td>219</td>
<td>302-516</td>
<td>untreated for in E. coli (38)</td>
</tr>
<tr>
<td>NP_036475</td>
<td>5</td>
<td>Niacinamide nucleotide transhydrogenase</td>
<td>Amino Acid Metabolism</td>
<td>530</td>
<td>48-587</td>
<td>NADP transhydrogenase</td>
<td>Nucleotide Metabolism (in E. coli)</td>
<td>480</td>
<td>618-1081</td>
<td>untreated for in E. coli (39)</td>
</tr>
<tr>
<td>NP_005933</td>
<td>6</td>
<td>Molybdopterin synthesis</td>
<td>Molybdopterin synthesis</td>
<td>133</td>
<td>61-366</td>
<td>Molybdenum cofactor biogenesis protein A</td>
<td>Molybdopterin synthesis</td>
<td>156</td>
<td>481-629</td>
<td>untreated for in E. coli (40)</td>
</tr>
<tr>
<td>NP_000932</td>
<td>9</td>
<td>Sarcosine dehydrogenase</td>
<td>Amino Acid Metabolism</td>
<td>382</td>
<td>63-445</td>
<td>Sarcosine oxidase</td>
<td>Amino Acid Metabolism</td>
<td>179</td>
<td>483-894</td>
<td>untreated for in E. coli (41)</td>
</tr>
<tr>
<td>NP_000896</td>
<td>10</td>
<td>DNA directed RNA polymerase III</td>
<td>Amino Acid Metabolism</td>
<td>907</td>
<td>11-909</td>
<td>DNA-directed RNA polymerase subunit A</td>
<td>Amino Acid Metabolism</td>
<td>380</td>
<td>963-1366</td>
<td>untreated for in E. coli (42)</td>
</tr>
<tr>
<td>NP_000285</td>
<td>10</td>
<td>Pyrrole-5-carboxylate synthetase</td>
<td>Amino Acid Metabolism</td>
<td>356</td>
<td>72-381</td>
<td>Glutamyl-5-kinase</td>
<td>Amino Acid Metabolism</td>
<td>484</td>
<td>563-770</td>
<td>untreated for in E. coli (43)</td>
</tr>
</tbody>
</table>
This process produced 26,673 unique human sequence identity using the clustering program CD-HIT (13). A low measure of 40% sequence identity is used to remove redundant sequences because
homologous proteins share a common fold, even when the overall sequence identity is less than 10% (14).

3.1.2. Dataset 2 (DS2)

The protein sequences for 71 completely sequenced prokaryotic genomes obtained from NCBI (ftp://ftp.ncbi.nih.gov/genomes/Bacteria) form DS2. The list is available online. Sequences from 71 genomes are then merged into one single file (223,676 sequences). The redundant sequences in DS2 are removed as described in DS1. This process produced 102,135 unique prokaryotic sequences (UPS).

3.2. Identification of fusion proteins

The 26,673 UHS are searched against the 102,135 UPS using BLASTP at an E value cutoff of ≤ 10−10. This experiment identified 141 human fusion proteins consisting of two or more fusion partners of prokaryotic origin. The list is available online.

3.3. Functional inferences to fusion proteins

Molecular functions are inferred for many of these fusion proteins using data collected from literature. For 29 of 141 fusion proteins, “accreted functions” are inferred using experimental data for fusion partners in prokaryotic systems (Table 1). These fusion proteins are grouped into four categories using functional inferences.

3.4. Availability

http://sege.ntu.edu.sg/wester/fusion

4. RESULTS AND DISCUSSION

Although a number of fusion proteins are reported in literature across several phylogenetic distances, a comprehensive list for human fusion proteins of prokaryotic origin is not available. We identified 141 fusion proteins of prokaryotic origin in the human genome. These fusion genes may have arisen by the fusion of two or more component genes of prokaryotic origin through gene transfer to attain optimal functional versatility and/or novelty. Our interest is to infer accreted functions by fusion proteins in relation to their fusion partners in a prokaryotic system. Hence, we classified fusion proteins into four categories based on their accreted functions. These categories of fusion proteins are discussed below.

4.1. Fusion proteins mimicking operons in prokaryotes

Interestingly, 18 of the 29 fusion proteins mimic operons (cluster of genes that are juxtaposed next to each other and are transcribed as one unit) in prokaryotes. In prokaryotes, genes involved in a related pathway are arranged as operons. This is also true in the un-segmented worm C. elegans that is shown to have operons (15). Fusion could be a way of co-regulation as efficiently as operons with two or more juxtaposed genes in a single unit. This could be a potent indicator of optimal design. The fusion protein pyrroline-5-carboxylate synthetase (P5CS) catalyzes ATP and NAD(P)H dependent conversion of L-glutamate to glutamic-γ-semialdehyde (GSA) in proline biosynthesis. The P5CS protein is bi-functional with γ-glutamate-5-kinase (γ-GK) and γ-glutamyl phosphate reductase (γ-GPR) activities required for proline biosynthesis (16). N terminal γ-GK and C terminal γ-GPR match prokaryotic GK and GPR proteins, respectively. In T. thermophilus, these two proteins operate as one operon with GK preceding GPR (17). This suggests that fusion proteins in human are formed by the fusion of two or more fusion partners. Seventeen more cases are listed in Table 1.

4.2. Fusion proteins exhibiting multiple functions

In eukaryotes, many multi-functional proteins catalyze successive reactions in biochemical/signal transduction pathways. The reaction rate is maximally optimized in these cases because the subsequent reaction centers (active sites) are physically placed side by side. This facilitates the ease capture of reaction intermediates from one reaction center to another as substrates (circumventing diffusion effects). Clustering of active sites for catalyzing a reaction sequence has several potential advantages: the catalytic activity can be enhanced because the local substrate concentrations are increased significantly (18). By sequestering reactive intermediates, their conversion by undesired chemical reactions is prevented as substrates are channeled from one catalytic site to the next (19). A covalently linked multifunctional protein is likely to be more stable than non-covalently formed protein-protein interfaces containing reaction (or active) centers. Thus, fusion of two or more monofunctional prokaryotic proteins into a single polypeptide in a higher organism is certainly under selective advantage in the course of evolution. The fusion protein GARS-AIRS-GART exhibits multiple functions in human (20). Each of GARS, AIRS and GART proteins are mono-functional and part of the pur operon in B. subtilis and E. coli (21). The GARS-AIRS is a bifunctional protein in S. cerevisiae and GARS-AIRS-GART is tri-functional in Drosophila (21). In human, it is found that GARS-AIRS-GART is tri-functional and is formed by the fusion of three monofunctional enzymes. Thus, fusion proteins in a higher organism exhibit expanded function by physical co-existence of two or more mono-functional fusion partners. Six more cases are listed in Table 1.

4.3. Fusion proteins showing alternative splicing

Recent genome-wide analyses indicate that 40-60% of human genes are alternatively spliced, suggesting that alternative splicing is one of the significant processes of human biology (22, 23, 24). Two fusion proteins are shown to exhibit alternative splicing from this study (Table 1). A classic example is the GARS-AIRS-GART gene that produces two spliced variants, namely: (1) a tri-functional GARS-AIRS-GART; (2) a mono-functional GARS. The mono-functional GARS protein is produced by differential use of an intronic poly-adenylation signal located in the intron separating the last GARS exon from the first AIRS exon. Separate GARS and GARS-AIRS-GART mRNAs have been observed in human, mouse, chicken and D. melanogaster (25). One more case is listed in Table 1.

4.4. Fusion proteins simulating protein-protein interfaces in prokaryotes

Some fusion proteins simulate protein-protein interfaces in prokaryotes. For example, the human fusion protein acetyl coenzyme A carboxylase β simulates the
dimer of propionyl-CoA carboxylase α subunit and propionyl-CoA carboxylase β subunit in Mycobacterium smegmatis (26). Thus, two domains in acetyl-coenzyme A carboxylase β simulate a protein-protein interface formed by propionyl-CoA carboxylase α subunit and propionyl-CoA carboxylase β subunit in Mycobacterium smegmatis. This suggests that fusion events select protein-protein interfaces by fusing two fusion partners into a single polypeptide chain. Marcotte and colleagues identified human fusion proteins succinyl-CoA transferase and 5,6-diaminopimelic acid synthetase made up of fusion components that are known or predicted to interact in E. coli (7). Interestingly, our approach identified these two fusion proteins. It should also be noted that these two proteins not only simulate protein-protein interfaces in E. coli but also mimic operon-like structures in T. thermophila and M. barkeri, respectively. Two more cases are listed in Table 1.

5. CONCLUSION

Modular organization of proteins has been postulated as a widely used strategy for protein evolution. We identified 29 fusion proteins of prokaryotic origin in the human genome. Analysis of fusion proteins suggests that these proteins exhibit enhanced or novel functions in human compared to their fusion partners (which are physically separated) in prokaryotes. These fusion proteins are found to mimic operons and simulate protein-protein interfaces in prokaryotes. They are also found to exhibit multiple functions and alternative splicing in humans. Our findings strongly suggest that, by the acquisition of additional active domains, fusion proteins expand their substrate specificity and evolve functional novelty. It is often thought that the function of fused genes is simply an addition of function in pre-existing component genes. However, this hypothesis is inconsistent with an observed phenomenon of accelerated evolution in chimerical genes. A recent structural analysis of the histidine biosynthesis components HisA and HisF indicate that the protein structure after gene fusion was also subject to structural and functional adaptation (27). In this sense, gene fusion may be one of the critical steps towards creating a new gene with novel or accreted function.

The hypothesis underlying this analysis is that a fusion gene in human can indicate an association between the independent genes in prokaryotes, assuming that orthologous genes have parallel functions in both human and one or more prokaryotes. Linking genes by way of fusion events, as proposed earlier can hint at direct physical interactions between proteins (7) or a more general functional association such as between sequential members in a metabolic pathway (18, 19). One of many possible mechanisms of fusion events is lateral gene transfer and this hypothesis remains as speculation due to lack of sufficient genome data of distant evolutionary origin (6). The idea of gene transfer from a prokaryote to human is intriguing. However, the significant mechanical barriers, as well as constraints to natural selection, warn caveats when considering inter-kingdom gene transfer.

The list of fusion proteins presented in this report will provide some meaningful insights into protein evolution. It should be noted that our analysis is restricted to human fusion genes of prokaryotic origin. About 20% proteins (29 fusion proteins) generated by our analysis are identified to mimic operons, exhibit multiple functions, show alternative splicing and simulate protein-protein interfaces using data obtained rigorously from published literature. However, the experimental verification of accreted function using published report in this study is minimal. Therefore, it is important to verify their accreted functions using experimental data coupled with other stringent and more complete computational procedures. Characterization of this set of genes is undoubtedly critical and this involves case-by-case isolation of their proteins followed by specific functional assays. The data obtained by this analysis is available for download and search at our web site. We propose to extend our quest to identify and characterize fusion proteins across different phylogenetic distances. This exercise may shed some light into the possible mechanism of fusion events between prokaryotes and human.

6. ACKNOWLEDGEMENTS

This project is supported by a research grant from A*STAR (BMRC research grant #01/1/21/19/1191), Singapore and Nanyang Technological University.

7. REFERENCES


9. Yanai, I., A. Derti and C. DeLisi: Genes linked by fusion events are generally of the same functional category: A


13. Li, W. Z., L. Jaroszewski and A. Godzik: Clustering of highly homologous sequences to reduce the size of large protein database. *Bioinformatics* 17, 282-283 (2001)


35. Eggen, R. I., A. C. Geerling, M. S. Jetten and W. M. de Vos: Cloning, expression, and sequence analysis of the


57. Dong, F., S. E. Weitzel and P. H. von Hippel: A coupled complex of T4 DNA replication helicase (gp41) and polymerase (gp43) can perform rapid and processive DNA strand-displacement synthesis. Proc Natl Acad Sci 93, 14456-14461 (1996)


Key Words: Fusion Proteins, Prokaryotic Origin, Accretion

Send correspondence to: Meena Kishore Sakharkar Ph.D, School of Mechanical & Production Engineering, Nanyang Technological University, Singapore 639798, Tel: 65-6790-5836, Fax: 65-6774-4340, E-mail: mmeena@ntu.edu.sg