[Frontiers in Bioscience 9, 2964-2971, September 1, 2004]

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

Yu Yiting, Iti Chaturvedi, Liew Kim Meow, Pandjassarame Kangueane 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

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 multi-functional 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

S/N	*Human Fusion Proteins*					*Fusion Partners of Prokaryotic Origin*								*Comments*
						N terminal				*C terminal*				
	RefSeq Accession	**Chromosome Number**	**Protein Length**	**Protein Name**	**Pathway**	**Protein Length**	**Match Region**	**Protein Name**	**Pathway**	**Protein Length**	**Match Region**	***Protein Name***	***Pathway***	
Fusion proteins mimicking prokaryotic operons														
1	NP_004276	1	791	Glucose 1 dehydrogenase/ 6phospho glucono-lactonase	Carbohydrate Metabolism	479	27..510	Glucose-6-phosphate dehydrogenase	Carbohydrate Metabolism; Amino Acids Metabolism	261	555..751	6-phospho gluconolactonase	Carbohydrate Metabolism	zwf-pgl-eda-operon in *P. putida* (28) Bifunctional enzyme (29)
2	NP_000173	2	763	Hydroxyl-acyl dehydrogenase, (subunit A)	Lipid Metabolism; Amino Acid Metabolism; Carbohydrate Metabolism	258	46..278	Enoyl-CoA hydratase	Lipid Metabolism; Amino Acid Metabolism; Carbohydrate Metabolism (in *E. coli*)	411	440..750	3-Hydroxyacyl-CoA dehydrogenase	Lipid Metabolism; Amino Acid Metabolism; Carbohydrate Metabolism (in *E. coli*)	fadAB operon in *E. coli* (30) Trifunctional protein (31)
3	NP_004332	2	2225	Carbamoyl phosphate synthetase 2 / Aspartate trans carbamylase/ Dihydro-orotase trifunctional protein	Amino Acid Metabolism; Nucleotide Metabolism	374	2..351	Carbamoyl-phosphate synthetase small subunit	Amino Acid Metabolism; Nucleotide metabolism (*S. typhimurium*)	1076	394..1440	Carbamoyl-phosphate synthase, large subunit	Amino Acid Metabolism; Nucleotide metabolism (in *S. typhimurium*)	carAB operon in *S. typhimurium* (32) Trifunctional protein (33)
						423	1446..1790	Dihydro-orotase (pyrC)	Amino Acid Metabolism; Nucleotide metabolism (in *S. acidocaldarius*)	305	1921..2223	Aspartate carbamoyl transferase (pyrBI)	Amino Acid Metabolism; Nucleotide metabolism (in *S. acidocaldarius*)	pyr operon in *S. acidocaldarius* (34)
4	NP_001150	2	1338	Aldehyde oxidase 1	Amino Acid Metabolism; Metabolism of Cofactors and Vitamins	489	9..531	Carbon-monoxide dehydrogenase small subunit	Energy Metabolism (in *M. loti*)	799	580..1316	Aarbon monoxide dehydrogenase, large subunit	Energy Metabolism (in *M. loti*)	cdh operon in *M. soehngenii* (35)
5	NP_001957	3	723	Enoyl-CoA, hydratase/ 3-hydroxyacyl CoA dehydrogenase	Lipid Metabolism; Amino Acid Metabolism; Carbohydrate Metabolism	297	5..277	Enoyl-CoA hydratase	Lipid Metabolism; Amino Acid Metabolism; Carbohydrate Metabolism (in *E. coli*)	411	302..705	3-Hydroxyacyl CoA dehydrogenase	Lipid Metabolism; Amino Acid Metabolism; Carbohydrate Metabolism (in *E. coli*)	fadAB operon in *E. coli* (30) Bifunctional protein (36)
6	NP_001059	3	1621	DNA topoisomerase II, ß		773	56..727	DNA gyrase, Subunit B (gyrB)		490	728..1054	DNA Gyrase Subunit A		Operon in *M. tuberium* (37) Protein-protein interaction (38)
7	NP_000929	4	1174	DNA directed RNA polymerase II polypeptide B		550	28..535	DNA-directed RNA polymerase, subunit B "	Genetic Information Processing	649	565..1172	DNA-dependent RNA polymerase subunit B '	Genetic Information Processing	RNAP operon in *M. thermoautotrophicum* (39)
8	NP_005434	4	624	3'-Phospho-adenosine 5'-phosphosulfate synthethase	Nucleotide Metabolism	186	51..221	Adenosine 5'-phosphosulfate kinase	Nucleotide Metabolism; Metabolism of Other Amino Acids; Energy Metabolism (in *E. coli*)	459	273..619	ATP sulfurylase	Nucleotide Metabolism; Metabolism of Other Amino Acids; Energy Metabolism (in *E. coli*)	cys operon in *E. coli* (40) Bifunctional enzyme (41)
9	NP_006443	4	425	AIR carboxylase; SAICAR synthetase	Amino Acid Metabolism	233	12..247	Phospho ribo-sylaminoimidazole succino-carboxamide (SAICAR) synthetase (purC)	Nucleotide Metabolism (in *B. subtilis*)	180	267..421	Phosphoribosylaminoimidazole(AIR) carboxylase (purE)	Nucleotide Metabolism (in *B. subtilis*)	pur Operon in *B. subtilis* (21) Multifunctional protein (42)
10	NP_000427	5	520	Succinyl CoA: 3-oxoacid CoA transferase	Metabolism of Cofactors and Vitamins	250	36..284	Acyl CoA:acetate CoA transferase, subunit a	Carbohydrate Metabolism (in *E.coli*)	219	302..516	Acyl CoA:acetate CoA-transferase subunit ß	Carbohydrate Metabolism (in *E.coli*)	Operon in *C. acetobutylicum* (43) Protein-protein interaction (44)
11	NP_036475	5	1086	Nicotinamide nucleotide transhydrogenase		530	48..587	NADP transhydrogenase a subunit	Metabolism of Cofactors and Vitamins	480	618..1081	NADP transhydrogenase ß subunit	Metabolism of Cofactors and Vitamins	nnt operon in *R. rubrum* (45)
12	NP_005933	6	636	Molybdenum cofactor synthesis-step 1 protein isoforms		333	61..366	Molybdenum cofactor biosynthesis protein A		156	481..629	Molybdenum cofactor biosynthesis protein CB		moa operon in *E. coli* (46)
13	NP_009032	9	918	Sarcosine dehydrogenase	Genetic Information Processing	382	63..445	Sarcosine oxidase subunit ß	Amino Acid Metabolism	379	483..894	Sarcosine oxidase, subunit a	Amino Acid Metabolism	sox operon in *Corynebacterium sp.* (47)
14	NP_008986	10	1391	DNA directed RNA polymerase III	Amino Acid Metabolism	907	11..909	DNA-directed RNA polymerase subunit A'	Genetic Information Processing	380	963..1366	DNA-dependent RNA polymerase, subunit A"	Genetic Information Processing	RNAP operon in *M. thermoautotrophicum* (39)
15	NP_002851	10	795	Pyrroline-5-carboxylate synthetase	Amino Acid Metabolism; Carbohydrate Metabolism	356	72..381	Glutamayl 5-kinase	Amino Acid Metabolism	484	363 .. 770	?-Glutamyl phosphate reductase	Amino Acid Metabolism	ProBA operon in *T. thermophilus* (17) Bifunctional enzyme (16)

16	NP_000911	11	1178	Pyruvate carboxylase precursor	Carbohydrate Metabolism	477	38..483	Pyruvate carboxylase, subunit A	Amino Acid Metabolism; Carbohydrate Metabolism (*in S. mutans*)	567	563..1178	Pyruvate carboxylase, subunit B	Amino Acid Metabolism; Carbohydrate Metabolism (in *T. tengcongensis*)	Operon-like arrangement in *M. barkeri* (48)
17	NP_001087	17	1105	ATP citrate lyase	Genetic Information Processing	398	41..418	Citrate lyase, subunit 1	Carbohydrate Metabolism; Energy Metabolism (*in E. coli*)	610	496..1089	Citrate lyase, subunit 2	Carbohydrate Metabolism; Energy Metabolism (*in E. coli*)	Probable operon in *K. pneumoniae* (49)
18	NP_000928	17	1970	DNA directed RNA polymerase II polypeptide A		895	18..896	DNA-directed RNA polymerase subunit A'	Genetic Information Processing	451	1056..1479	DNA-directed RNA polymerase subunit A"	Genetic Information Processing	RNAP operon in *M. thermoautotrophicum* (39)
Fusion proteins with Multiple functions														
19	NP_004437	1	1440	Glutamyl-prolyl tRNA synthetase	Amino Acid Metabolism; Metabolism of Cofactors and Vitamins; Genetic Information Processing	555	117..612	Glutamyl-tRNA synthetase	Amino Acid Metabolism; Metabolism of Cofactors and Vitamins; Genetic Information Processing	480	947..1408	Prolyl-tRNA synthetase	Amino Acid Metabolism; Genetic Information Processing	Bifunctional enzyme (50)
20	NP_000405	5	736	Hydroxysteroid (17-β) dehydrogenase	Lipid Metabolism	303	1..301	Short-chain dehydrogenases	Lipid Metabolism	286	327..607	MaoC like dehydratase		Multifunctional enzyme (51)
21	NP_005467	9	722	UDP-N-acetyl-glucosamine 2-epimerase/N-acetyl-mannosamine kinase	Metabolism of Complex Carbohydrates	377	11..349	UDP-N acetyl glucosamine 2-epimerase	Metabolism of Complex Carbohydrates	315	412..675	Glucose kinase	Carbohydrate Metabolism; Metabolism of Complex Carbohydrates; Biosynthesis of Secondary Metabolites	Bifunctional enzyme (52)
22	NP_005947	14	935	Methylenetetrahydrofolate dehydrogenase/ cyclohydrolase; formyltetrahydrofolate synthetase		284	5..293	Methylenetetrahydrofolate dehydrogenase (FolD)	Carbohydrate Metabolism; Metabolism of Cofactors and Vitamins	599	319..925	Formyltetrahydrofolate synthetase	Carbohydrate Metabolism; Metabolism of Cofactors and Vitamins	Trifunctional protein (53)
23	NP_004095	17	2509	Fatty acid synthase	Fatty acid biosynthesis	1610	3..842	Polyketide synthase		325	1557..1852	Quinone oxidoreductase		Multifunctional enzyme (54)
24	NP_006648	21	541	Formimino-transferase cyclo-deaminase	Amino Acid Metabolism; Metabolism of Cofactors and Vitamins	298	1..290	Glutamate formimino transferase	Amino Acid Metabolism; Metabolism of Cofactors and Vitamins	217	378..523	Formimino tetrahydrofolate cyclodeaminase	Metabolism of Cofactors and Vitamins	Bifunctional enzyme (55)
25	NP_000810	21	1010	GARS-AIRS-GART	Nucleotide Metabolism; Metabolism of Cofactors and Vitamins	429	5..427	GARS	Nucleotide Metabolism	356	435..777	AIRS	Nucleotide Metabolism	Alternative splicing (25) Trifuncitonal protein (25)
Fusion proteins exhibiting alternative splicing														
26	NP_037361	12	1237	Apoptotic protease activatingfactor isoform a	Cell Growth and Death	1381	109..493	Regulatory protein		626	603..1191	WD-40 repeat protein		Alternative splicing (56)
25	NP_000810	21	1010	GARS-AIRS-GART	Nucleotide Metabolism; Metabolism of Cofactors and Vitamins	429	5..427	GARS	Nucleotide Metabolism	356	435..777	AIRS	Nucleotide Metabolism	Alternative splicing (25)
										222	806..1004	GART	Nucleotide Metabolism	Trifuncitonal protein (25)
Fusion proteins simulating prokaryotic protein-protein interfaces														
27	NP_006587	3	2724	Polymerase (DNA directed), θ	Amino Acid Metabolism; Metabolism of Cofactors and Vitamins; Genetic Information Processing	720	221..963	DNA helicase	Amino Acid Metabolism; Metabolism of Cofactors and Vitamins; Genetic Information Processing	893	2031..2627	DNA-directed DNA polymerase I	Amino Acid Metabolism; Genetic Information Processing	Protein-protein interaction (57)
28	NP_000246	6	750	Methyl malonyl Co A mutase precursor	Lipid Metabolism	681	42..519	Methyl malonyl Co A mutase, subunit ß	Lipid Metabolism	144	611..749	Methyl malonyl Co A mutase, subunit a		Protein-protein interaction (58)
29	NP_001084	12	2483	Acetyl co-enzyme A carboxylase β	Metabolism of Complex Carbohydrates	677	256..960	Propionyl-CoA carboxylase, subunit a	Metabolism of Complex Carbohydrates	516	1885..2267	Propionyl-CoA carboxylase subunit ß	Carbohydrate Metabolism; Metabolism of Complex Carbohydrates; Biosynthesis of Secondary Metabolites	Protein-protein interaction (26)

sequence identity using the clustering program CD-HIT (13). This process produced 26,673 unique human sequences (UHS). 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 $\leq 10^{-10}$. 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 easy 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 mono-functional 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 mono-functional 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 co-enzyme A carboxylase β simulates the

dimer of propionyl co-A carboxylase α subunit and propionyl co-A carboxylase β subunit in *Mycobacterium smegmatis* (26). Thus, two domains in acetyl co-enzyme A carboxylase β simulate a protein - protein interface formed by propionyl co-A carboxylase α subunit and propionyl co-A 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 Co-A transferase and δ-1-pyroline-5-carboxylate 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. thermophilus* 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

1. Tsoka, S. and C. A. Ouzounis: Functional versatility and molecular diversity of the metabolic map of Escherichia coli. *Genome Res* 9, 1503-1510 (2001)

2. Long, M: A new function Evolved from Gene Fusion. *Genome Res* 10, 1655-1657 (2000)

3. International Human Genome Sequencing Consortium Initial sequencing and analysis of the human genome. *Nature* 409, 860-921 (2001)

4. Ponting, C. P: Plagiarized bacterial genes in the human book of life. *Trends Genet* 17, 235-237 (2001)

5. Andersson, J. O., W. F. Doolittle and C. L. Nesbø: Are there bugs in our genome? *Science* 292, 1848-1850 (2001)

6. Salzberg, S. L., O. White, J. Peterson and J. A. Eisen: Microbial genes in the human genome: lateral transfer or gene loss? *Science* 292, 1903-1906 (2001)

7. Marcotte, E. M, M. Pellegrini, H. L. Ng, D. W. Rice, T. O. Yeates and D. Eisenberg: Detecting protein function and protein-protein interactions from genome sequences. *Science* 285, 751-753 (1999)

8. Genereux, D. P. and J. M. Jr. Logsdon: Much ado about bacteria-to-vertebrate lateral gene transfer. *Trends Genet* 19, 191-195 (2003)

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

systematic analysis of 30 microbial genomes. *Proc Natl Acad Sci* 98, 7940-7945 (2001)

10. Katzen, F., M. Deshmukh, F. Daldal and J. Beckwith: Evolutionary domain fusion expanded the substrate specificity of the transmembrane electron transporter DsbD. *EMBO J* 21, 3960-3969 (2002)

11. Berthonneau, E. and M. Mirande: A gene fusion event in the evolution of aminoacyl-tRNA synthetases. *FEBS Lett* 470, 300-304 (2000)

12. Truong, K. and M. Ikura: Domain fusion analysis by applying relational algebra to protein sequence and domain databases. *BMC Bioinformatics* 4, 16 (2003)

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)

14. Abagyan, R. A. and S. Batalov: Do aligned sequences share the same fold? *J Mol Biol* 273, 355-368 (1997)

15. Mering, C. V. and P. Bork: Teamed up for transcription. *Nature* 417, 797-798 (2002)

16. Aral, B., J. S. Schlenzig, G. Liu and P. Kamoun: Database cloning human delta 1-pyrroline-5-carboxylate synthetase (P5CS) cDNA: a bifunctional enzyme catalyzing the first 2 steps in proline biosynthesis. *C R Acad Sci III* 319, 171-178 (1996)

17. Kosuge, T., K. Tabata and T. Hoshino: Molecular cloning and sequence analysis of the proBA operon from an extremely thermophilic eubacterium Thermus thermophilus. *FEMS Microbiol Lett* 123, 55-61 (1994)

18. Reed, L. J: Multienzyme complexes. *Acc Chem Res* 7, 40-46 (1974)

19. Perham, R. N: Self-assembly of biological macromolecules. *Philos Trans R Soc Lond B Biol Sci* 272, 123-136 (1975)

20. McCarthy, A. D. and D. G. Hardie: Fatty acid synthase: an example of protein evolution by gene fusion. *Trends Biochem Sci* 9, 60-63 (1984)

21. Ebbole, D. J. and H. Zalkin: Cloning and characterization of a 12-gene cluster from Bacillus subtilis encoding nine enzymes for *de novo* purine nucleotide synthesis. *J Biol Chem* 262, 8274-8287 (1987)

22. Mironov, A. A., J. W. Fickett and M. S. Gelfand: Frequent alternative splicing of human genes. *Genome Res* 9, 1288-1293 (1999)

23. Brett, D., J. Hanke, G. Lehmann, S. Haase, S. Delbruck, S. Krueger, J. Reich and P. Bork: EST comparison indicates 38% of human mRNAs contain possible alternative splice forms. *FEBS Lett* 474, 83-86 (2000)

24. Kan, Z., E. C. Rouchka, W. R. Gish and D.J. States: Gene structure prediction and alternative splicing analysis using genomically aligned ESTs. *Genome Res* 11, 889-900 (2001)

25. Brodsky, G., T. Barnes, J. Bleskan, L. Becker, M. Cox and D. Patterson: The human GARS-AIRS-GART gene encodes two proteins which are differentially expressed during human brain development and temporally over-expressed in cerebellum of individuals with Down syndrome. *Hum Mol Genet* 6, 2043-2050 (1997)

26. Haase, F. C., H. Beegen and S. H. Allen: Propionyl-coenzyme A carboxylase of Mycobacterium smegmatis. An electron microscopic study. *Eur J Biochem* 140, 147-151 (1984)

27. Lang, D., R. Thoma, M. Henn-SAX, R. Sterner and M. Wilmanns: Structural evidence for evolution of the beta/alpha barrel scaffold by gene duplication and fusion. *Science* 289, 1546-1550 (2000)

28. Petruschka, L., K. Adolf, G. Burchhardt, J. Dernedde, J. Jurgensen and H. Herrmann: Analysis of the zwf-pgl-eda-operon in Pseudomonas putida strains H and KT2440. *FEMS Microbiol Lett* 215, 89-95 (2002)

29. Clarke, J. L., D. A. Scopes, O. Sodeinde and P. J. Mason: Glucose-6-phosphate dehydrogenase-6-phospho-gluconolactonase. A novel bifunctional enzyme in malaria parasites. *Eur J Biochem* 268, 2013-2019 (2001)

30. Yang, S. Y. and H. Schulz: The large subunit of the fatty acid oxidation complex from Escherichia coli is a multifunctional polypeptide. Evidence for the existence of a fatty acid oxidation operon (fad AB) in Escherichia coli. *J Biol Chem* 258, 9780-9785 (1983)

31. Orii, K. E., K. O. Orii, M. Souri, T. Orii, N. Kondo, T. Hashimoto and T. Aoyama: Genes for the human mitochondrial trifunctional protein alpha- and beta-subunits are divergently transcribed from a common promoter region. *J Biol Chem* 274, 8077-8084 (1999)

32. Kilstrup, M., C. D., Lu, A. Abdelal and J. Neuhard: Nucleotide sequence of the carA gene and regulation of the carAB operon in Salmonella typhimurium. *Eur J Biochem* 176, 421-429 (1988)

33. Chen, K. C., D. B. Vannais, C. Jones, D. Patterson and J. N. Davidson: Mapping of the gene encoding the multifunctional protein carrying out the first three steps of pyrimidine biosynthesis to human chromosome 2. *Hum Genet* 82, 40-44 (1989)

34. Thia-Toong, T. L., M. Roovers, V. Durbecq, D. Gigot, N. Glansdorff and D. Charlier: Genes of *de novo* pyrimidine biosynthesis from the hyperthermoacidophilic crenarchaeote Sulfolobus acidocaldarius: novel organization in a bipolar operon. *J Bacteriol* 184, 4430-4441 (2002)

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

genes for carbon monoxide dehydrogenase of Methanothrix soehngenii. *J Biol Chem* 266, 6883-6887 (1991)

36. Hoefler, G., M. Forstner, M. C. McGuinness, W. Hulla, M. Hiden, P. Krisper, L. Kenner, T. Ried, C. Lengauer, R. Zechner, H. W. Moser and G. L. Chen**:** cDNA cloning of the human peroxisomal enoyl-CoA hydratase: 3-hydroxyacyl-CoA dehydrogenase bifunctional enzyme and localization to chromosome 3q26.3-3q28: a free left Alu Arm is inserted in the 3' noncoding region. *Genomics* 19, 60-67 (1994)

37. Unniraman, S., M. Chatterji and V. Nagaraja: DNA gyrase genes in Mycobacterium tuberculosis: a single operon driven by multiple promoters. *J Bacteriol* 184, 5449-5456 (2002)

38. Sugino, A., N. P. Higgins and N. R. Cozzarelli: DNA gyrase subunit stoichiometry and the covalent attachment of subunit A to DNA during DNA cleavage. *Nucleic Acids Res* 8, 3865-3874 (1980)

39. Smith, D. R., L. A. Doucette-Stamm, C. Deloughery, H. Lee, J. Dubois, T. Aldredge, R. Bashirzadeh, D. Blakely, R. Cook, K. Gilbert, D. Harrison, L. Hoang, P. Keagle, W. Lumm, B. Pothier, D. Qiu, R. Spadafora, R. Vicaire, Y. Wang, J. Wierzbowski, R. Gibson, N. Jiwani, A. Caruso, D. Bush and JN Reeve: Complete genome sequence of Methanobacterium thermoautotrophicum deltaH: functional analysis and comparative genomics. *J Bacteriol* 179, 7135-7155 (1997)

40. Leyh, T. S., T. F. Vogt and Y. Suo: The DNA sequence of the sulfate activation locus from Escherichia coli K-12. *J Biol Chem* 267, 10405-10410 (1992)

41. Venkatachalam, K. V., H. Akita and C. A. Strott: Molecular cloning, expression, and characterization of human bifunctional 3'-phosphoadenosine 5'-phosphosulfate synthase and its functional domains. *J Biol Chem* 273, 19311-19320 (1998)

42. Minet, M. and F. Lacroute: Cloning and sequencing of a human cDNA coding for a multifunctional polypeptide of the purine pathway by complementation of the ade2-101 mutant in Saccharomyces cerevisiae. *Curr Genet* 18, 287-291 (1990)

43. Cary, J. W., D. J. Petersen, E. T. Papoutsakis and G. N. Bennett: Cloning and expression of Clostridium acetobutylicum ATCC 824 acetoacetyl-coenzyme A:acetate/butyrate:coenzyme A-transferase in Escherichia coli. *Appl Environ Microbiol* 56, 1576-1583 (1990)

44. Yeh, W. K. and L. N. Ornston: Evolutionarily homologous alpha 2 beta 2 oligomeric structures in beta-ketoadipate succinyl-CoA transferases from Acinetobacter calcoaceticus and Pseudomonas putida. *J Biol Chem* 256, 1565-1569 (1981)

45. Yamaguchi, M. and Y. Hatefi: Energy-transducing nicotinamide nucleotide transhydrogenase: nucleotide sequences of the genes and predicted amino acid sequences of the subunits of the enzyme from Rhodospirillum rubrum. *J Bioenerg Biomembr* 26, 435-445 (1994)

46. Rivers, S. L., E. McNairn, F. Blasco, G. Giordano and D. H. Boxer: Molecular genetic analysis of the moa operon of Escherichia coli K-12 required for molybdenum cofactor biosynthesis. *Mol Microbiol* 8, 1071-1081 (1993)

47. Chlumsky, L. J., L. Zhang and M. S. Jorns: Sequence analysis of sarcosine oxidase and nearby genes reveals homologies with key enzymes of folate one-carbon metabolism. *J Biol Chem* 270, 18252-18259 (1995)

48. Mukhopadhyay, B., E. Purwantini, C. L. Kreder and R. S. Wolfe: Oxaloacetate synthesis in the methanarchaeon Methanosarcina barkeri: pyruvate carboxylase genes and a putative Escherichia coli-type bifunctional biotin protein ligase gene (bpl/birA) exhibit a unique organization. *J Bacteriol* 183, 3804-3810 (2001)

49. Bott, M. and P. Dimroth: Klebsiella pneumoniae genes for citrate lyase and citrate lyase ligase: localization, sequencing, and expression. *Mol Microbiol* 14, 347-356 (1994)

50. Cerini, C., P. Kerjan, M. Astier, D. Gratecos, M. Mirande and M. Semeriva: A component of the multisynthetase complex is a multifunctional aminoacyl-tRNA synthetase. *EMBO J* 10, 4267-4277 (1991)

51. Leenders, F., V. Dolez, A. Begue, G. Moller, J. C. Gloeckner, Y. de Launoit, and J. Adamski: Structure of the gene for the human 17beta-hydroxysteroid dehydrogenase type IV. *Mamm Genome* 9, 1036-1041 (1998)

52. Lucka, L., M. Krause, K. Danker, W. Reutter and R. Horstkorte: Primary structure and expression analysis of human UDP-N-acetyl-glucosamine-2-epimerase/N-acetyl-mannosamine kinase, the bifunctional enzyme in neuraminic acid biosynthesis. *FEBS Lett* 454, 341-344 (1999)

53. Hum, D. W., A. W. Bell, R. Rozen and R. E. MacKenzie: Primary structure of a human trifunctional enzyme. Isolation of a cDNA encoding methylenetetrahydrofolate dehydrogenase-methenyltetrahydrofolate cyclohydrolase-formyltetrahydrofolate synthetase. *J Biol Chem* 263, 15946-15950 (1988)

54. Wakil, S. J: Fatty acid synthase, a proficient multifunctional enzyme. *Biochemistry* 28, 4523-4530 (1989)

55. Beaudet, R. and R. E. Mackenzie: Formiminotransferase cyclodeaminase from porcine liver. An octomeric enzyme containing bifunctional polypeptides. *Biochim Biophys Acta* 453, 151-161 (1976)

56. Ogawa, T., K. Shiga, S. Hashimoto, T. Kobayashi, A. Horii and T. Furukawa: APAF-1-ALT, a novel alternative splicing form of APAF-1, potentially causes impeded ability of undergoing DNA damage-induced apoptosis in the LNCaP human prostate cancer cell line. *Biochem*

Biophys Res Commun 306, 537-543 (2003)

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)

58. Marsh, E. N., S. E. Harding and P. F. Leadlay: Subunit interactions in Propionibacterium shermanii methylmalonyl-CoA mutase studied by analytical ultracentrifugation. *Biochem J* 260, 353-358 (1989)

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