Identification, analysis and evaluation of human fusion proteins of bacterial origin

Yu, Yiting

2007

Yu, Y. T. (2007). Identification, analysis and evaluation of human fusion proteins of bacterial origin. Doctoral thesis, Nanyang Technological University, Singapore.

<https://hdl.handle.net/10356/36052>

<https://doi.org/10.32657/10356/36052>

a529973

Identification, Analysis and Evaluation of Human Fusion Proteins of Bacterial Origin

Yu Yiting

School of Mechanical and Aerospace Engineering

A thesis submitted to the Nanyang Technological University
in fulfilment of the requirement for the degree of
Doctor of Philosophy

2007

ACKNOWLEDGEMENTS

I am greatly indebted to Professor Meena Kishore Sakharkar, my supervisor, for giving me an opportunity to work and conduct my research under her guidance and encouragement. Since the very beginning, her wealth of knowledge and brilliance, as well as advice and support during my graduate studies, has been invaluable in my professional development.

I would like to express my sincere gratitude to Professor Pandjassarame Kanguane for his patience, guidance and enthusiasm on this project. The knowledge I gained under his mentorship will carry me far.

I would like to thank the group members in my research team, present and past, Iti Chaturvedi, Zhao Bing, Li Lei, Xue Yuhao, Lee Pern Chern, Kingshuk Ghosh, Nidhi dandona, Lim Yun Ping, Bagavathi S. Perumal, Cui Zhanhua. It has been a great pleasure to work with them.

I would like to give special thanks to some of my great friends in NTU who have made these years both enlightening and enjoyable: Tan Wanloo, Cai Xinle, Cao Changrong, Cui Zhanghua, Zhao Xin, Xing Xiuqing, Qin Wenmin, Che Faxin, Ren Yunxia, Zhang Bingru, Li Yajin, Hou Dongxia, Kwoh Chee Keong.

I have been financially supported by a Research Student Scholarship (three years) from Nanyang Technological University. For this assistance I am very grateful.

From my stay in CANES lab, I wish to thank the faculty and staff for making necessary resources easily obtainable.

Finally, I would like to express my deepest love and appreciation for my mother, father, husband and daughter for their everlasting support and encouragement.

SUMMARY

Gene fusion is identified as a common phenomenon in the formation of new proteins. In order to better understand this evolutionary trend, we first designed a methodology for the identification of fusion proteins (the composite proteins) consisting of two or more fusion partners (the component proteins). We then applied the designed methodology to identify human fusion proteins of bacterial origin in the human genome. We thus obtained a comprehensive list of 141 putative human composite proteins consisting of component proteins (or domains) of bacterial origin. The 141 identified human fusion proteins were visually inspected and manually mapped to metabolic pathways using the KEGG pathway database interface. We thus selected six fusion proteins that function as metabolic enzymes in the citric acid cycle. The connectivity indexes for the six fusion proteins were then calculated using KEGG. The results showed that they produce metabolites with high connectivity for complex networking. To understand the structural and thermodynamics of fusion proteins, we identified a pair of IGPS structures (comprising of HisF and HisH domains or subunits) in the PDB (Protein Data Bank), where one structure is from *S. cerevisiae* (SC) and the other is from *T. thermophilus* (TT). The HisF-HisH structures are fused in SC and unfused in T.T. Thus a HisF-HisH domain-domain interface is formed in SC and a HisF-HisH subunit-subunit interface is formed in TT. Molecular dynamics simulations of these structures show larger interface area, gap volume and radius of gyration for the fused IGPS in SC. compared to the unfused IGPS in TT. However, the gap indices are maintained for the two structures. Thus, fusion proteins with accreted domains have both structural and thermodynamic advantages in organismal evolution.

ABBREVIATION

ACBP	acyl-CoA-binding protein
AIRS	aminoimidazole ribonucleotide synthetase
BLAST	Basic Local Alignment Search Tool
BVT	bacteria-to-vertebrate transfer
DS	dataset
ECH	enoyl-CoA hydratase
ECI	enoyl-CoA isomerase
GAPDH	glyceraldehyde-3-phosphate dehydrogenase
GARS	glycinamide ribonucleotide synthetase
GART	glycinamide ribonucleotide formyltransferase
GCCF	gephyrin cytoskeleton cofactor biosynthesis protein
GK	glutamyl kinase
GluRS	glutamyl-tRNA synthetases
GluProRS	glutamyl-prolyl-tRNA synthetase
GPR	glutamyl phosphate reductase
HMM	hidden Markov model
IGPS	imidazole glycerol phosphate synthase
IHGSC	International Human Genome Sequencing Consortium
KEGG	Kyoto Encyclopedia of Genes and Genomes
MD	molecular dynamics
MOCF	molybdenum cofactor biosynthesis protein signature
MoeA	molybdopterin cofactor
NCBI	National Center for Biotechnology Information
P5CS	pyrroline 5-carboxylate synthetase

PDB	Protein Data Bank
PEI	peroxisomal D3, D2-enoyl-CoA isomerase
PGK	phosphoglycerate kinase
ProRS	prolyl-tRNA synthetases
R_g	Radius of Gyration
RMSD	root mean square deviation
SASA	solvent accessible surface areas
SC	<i>Saccharomyces cerevisiae</i>
TIM	triosephosphate isomerase
TT	<i>Thermus thermophilus</i>

PUBLICATIONS

1. **Yiting Yu**, Kim S, Chua HH, Rodrigues F, Lin D, Sim SH, Derr A, Engels R, Deshazer D, Birren B, Nierman W, Tan P (2006) Genomic patterns of pathogen evolution revealed by comparison of *Burkholderia pseudomallei*, the causative agent of Melioidosis, to avirulent *Burkholderia thailandensis*, BMC Microbiology 6, 46, 2006 (Designated a “Highly Accessed” Article)
2. **Yiting Yu**, Li L, Sakharkar MK, Kanguane P. Insight into gene fusion from molecular dynamics simulation of fused and un-fused IGPS (Imidazole Glycerol Phosphate Synthetase). **Bioinformation** 1, 99-104, 2006
3. Sakharkar MK, **Yiting Yu**, Chow VTK, Kanguane P, Insights to metabolic network evolution by fusion proteins, **Frontiers in Bioscience**, 10: 1070-1078, May, 2005.
4. Sakharkar MK, Perumal BS, Lim YP, Lee PC, **Yiting Yu**, Kanguane P, Alternatively spliced human genes by exon skipping - A database (ASHESdb), **In Silico Biology**, 5, 0021, Dec 2004.
5. **Yiting Yu**, Chaturvedi I, Liew KM, Kanguane P, Sakharkar MK. Can ends justify the means? Digging deep for human fusion genes of prokaryotic origin. **Frontiers in Bioscience** 9, 2964-2971, Sep 2004.
6. Sakharkar MK, **Yiting Yu**, Kanguane P, Can ends justify the means? To fuse or not to fuse. **International Conference on Bioinformatics and its Applications (ICBA'04)** in Fort Lauderdale, Florida, USA, Dec 2004.

TABLE OF CONTENTS

Acknowledgements i

Summary iii

Abbreviation iv

Publications vi

List of figureures..... xi

List of tables xiv

Chapter 1.....1

INTRODUCTION.....1

1.1 Definition of fusion genes..... 1

1.2 Known cases of fusion genes 3

1.2.1 Fusion genes in metabolic pathway 3

1.2.2 Fusion genes and protein-protein interaction 5

1.2.3 Fusion genes in alternative splicing..... 8

1.2.4 Fusion genes in recombination 10

1.3 Fusion gene databases..... 12

1.3.1 Domain fusion database 12

1.3.2 Fusiondb 12

1.4 Advantages of gene fusion 14

1.4.1 Produce novel genes..... 14

1.4.2 Expand substrate specificity 14

1.4.3 Regulatory control by fused catalytic sites..... 15

1.5 Bacterial genes in human genome 17

1.6 Structural rearrangement by fusion 20

1.7 Objectives and scope 22

Chapter 2..... 25

IDENTIFICATION OF FUSION PROTEINS..... 25

2.1 Introduction 25

2.2 Programs used 26

2.3	Description of datasets.....	28
2.3.1	Dataset 1 (ds1)	28
2.3.2	Dataset 2 (ds2)	28
2.4	Identification of fusion proteins	28
2.5	Fusion proteins	34
2.6	Fusion database	38
2.7	Summary	40
Chapter 3		41
FUNCTIONAL EVALUATION OF FUSION PROTEINS		41
3.1	Introduction	41
3.2	Functional inferences to fusion proteins	42
3.2.1	Fusion proteins mimicking operons in bacteria.....	49
3.2.2	Fusion proteins showing alternative splicing	52
3.2.3	Fusion proteins exhibiting multiple functions.....	53
3.2.4	Fusion proteins simulating protein-protein interfaces in bacteria ..	57
3.3	Summary	57
Chapter 4		60
THE ASSOCIATION OF HUMAN FUSION PROTEINS IN METABOLIC NETWORKS		60
4.1	Introduction	60
4.2	Materials and methods	62
4.2.1	Human fusion proteins – definition and dataset source	62
4.2.2	Fusion proteins as metabolic enzymes.....	62
4.2.3	Fusion scenarios linked with metabolic enzymes having fusion structures	65
4.2.4	Prodom domain assignments to fusion proteins	66
4.2.5	Metabolites associated with fusion proteins.....	66
4.2.6	Connectivity index for metabolites of fusion proteins.....	68
4.3	Human fusion proteins.....	68
4.4	Fusion proteins as metabolic enzymes.....	69
4.5	Metabolites of fusion proteins.....	70
4.5.1	Oxaloacetate	72

4.5.2	Acetyl co-A	73
4.5.3	Succinyl co-A	74
4.5.4	Succinate	74
4.5.5	Glutamate	75
4.6	Discussion	76
4.7	Summary	79
Chapter 5		80
INSIGHT TOWARDS GENE FUSION FROM MOLECULAR DYNAMICS		
SIMULATION OF IGPS		80
5.1	Introduction	80
5.2	Materials and methods	83
5.2.1	Starting structures for md simulations	83
5.2.2	Energy minimization	84
5.2.3	Molecular dynamic simulation	87
5.2.4	Analysis of trajectories	87
5.3	Results	92
5.3.1	Structural features of the peptide backbone	92
5.3.2	Structural features of the interface	95
5.3.3	Structural features of the overall shapes	96
5.4	Discussion	102
5.5	Sumarry	104
Chapter 6		106
CONCLUSION		106
6.1	The achievement	106
6.2	The implications	107
6.3	Contribution	109
6.4	The limitations	109
6.5	Future improvement	110
References		111
Appendix 1		129

Appendix 2 132

Appendix 3 141

Appendix 4 151

LIST OF FIGURES

Figure 1.1: A hypothetical fusion scenario is given for human proteins of bacterial origin.....2

Figure 1.2: Fusion genes in metabolic pathways.....4

Figure 1.3: Five examples of *E.coli* proteins predicted to interact by the gene fusion analysis.....7

Figure 1.4: The molecular process for *Kua-UEV* gene fusion.....9

Figure 1.5: Hypothetical scenario of gene fusion to yield a bifunctional aminoacyl-tRNA synthetase..... 11

Figure 1.6: The flowchart of the methods for (a) the Domain Fusion Database and (b) the FusionDB 82.....13

Figure 1.7: Membrane topology and structural domains of DsbD and CcdA..... 16

Figure 1.8: Distribution of the homologues of the predicted human Proteins..... 19

Figure 1.9: Model for the evolution of the β/α barrel scaffold by twofold gene duplication and fusion.....21

Figure 1.10: The project workflow is shown. The workflow contains sections such as (1) data source; (2) identification; (3) validation; (4) pathway association and (5) structural significance..... 24

Figure 2.1: The flowchart of of fusion protein identification. 30

Figure 2.2: The example output of blast program.....31

Figure 2.3: The flowchart for analyzing output of blast..... 33

Figure 2.4: Protein sequences which share significant sequence homology with

human pyrroline-5-carboxylate synthetase (P5CS).....	36
Figure 2.5: Two examples for bacterial domain fusions in human are shown.....	37
Figure 2.6: Screenshot of Fusion database search page, full-page output for a query to NP_001059 and examples of some related information that can be obtained through this page.....	39
Figure 3.1: Functional tree of fusion proteins classification.....	43
Figure 3.2: Schematic presentation of fusion proteins mimicking operons in bacteria.....	51
Figure 3.3: Pathway of de novo purine biosynthesis to IMP.....	55
Figure 3.4: The formation of the human composite protein GARS-AIRS-GART by the fusion of GARS from <i>Bacillus halodurans</i> ; AIRS from <i>Agrobacterium tumefaciens</i> and GART from <i>Mesorhizobium loti</i> is illustrated.....	56
Figure 4.1: Fusion proteins and fusion partners.....	63
Figure 4.2: Fusion proteins and metabolites. A diagram showing citric acid cycle in carbohydrate metabolism is shown.....	64
Figure 5.1: Schematic representation of histidine biosynthetic pathway.....	81
Figure 5.2: Flowchart of two simulation systems for the fused and unfused structures of imidazole glycerol phosphate synthase (IGPS).....	88
Figure 5.3: Schematic representation of the eukaryote <i>S. cerevisiae</i> and the bacteria <i>T. thermophilus</i> protein sequences of the bifunctional enzyme imidazole glycerol phosphate synthase (IGPS).....	93
Figure 5.4: The snapshot picture of IGPS at 0 ns and 5 ns.....	94

Figure 5.5: Root mean square deviation (RMSD) between HisH and HisF in bacterial and yeast IGPS over 5 ns molecular dynamics simulation..... 97

Figure 5.6: Interface area between HisH and HisF in bacterial and yeast IGPS over 5 ns molecular dynamics simulation.....98

Figure 5.7: Gap volume between HisH and HisF in bacterial and yeast IGPS over 5 ns molecular dynamics simulation..... 99

Figure 5.8: Gap index between HisH and HisF in bacterial and yeast IGPS over 5 ns molecular dynamics simulation.....100

Figure 5.9: Radius of gyration of bacterial and yeast IGPS over 5 ns molecular dynamics simulation..... 101

LIST OF TABLES

Table 3.1: Human fusion proteins of bacterial origin..... 44

Table 4.1: Human fusion proteins in carbohydrate metabolism..... 67

Table 4.2: Fusion proteins as metabolic enzymes and associated
Metabolites..... 71

Table 5.1: PDB structures of Imidazole Glycerol Phosphate Synthase
(IGPS).....85

Table 5.2: Percent similarity values calculated for the deduced IGPS amino acid
sequences with the known crystal structures.....86

Table 2.1: List of prokaryotic species used in analysis..... 127

CHAPTER 1

INTRODUCTION

1.1 DEFINITION OF FUSION GENES

Gene fusion is a phenomenon that has generated much curiosity to biologist since its description. As early as 1965, genes encoding components of the tryptophan biosynthesis pathway, which is made up of five enzyme-catalyzed steps, were found to be distributed differently in the genomes of different organisms. In *E. coli*, for example, the two subunits of tryptophan synthetase are encoded by two different genes in contrast with *Neurospora crassa*, where a single gene encodes a product with both activities (DeMoss, 1965). In 1970, two adjacent histidine genes were observed to be fused in laboratory by the mechanism of deletion of translation stop and start codons (Yournon et al., 1970).

Gene fusion refers to two previously separate and independent genes are fused to form a single contiguous gene. A hypothetical gene fusion scenario is shown in Figure 1.1. The N- and C- terminal of one human protein matches two different split genes from two or more bacterial species. The fused human protein is referred as “COMPOSITE” protein and the split proteins in bacteria are referred as “COMPONENT” proteins. The composite protein is also referred as a “Rosetta stone” sequence (Marcotte et al., 1999).

Figure 1. 1

A hypothetical fusion scenario is given for human proteins of bacterial origin. The human fusion protein is formed by the fusion of two or more proteins from one or more bacterial species. They are also called as “composite protein” or “Rosetta Stone protein” and the fusion partners are known as “component proteins”.

1.2 KNOWN CASES OF FUSION GENES

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 (Long, 2000). The fusion genes have been identified in several taxa, eg., in plants (Lumbreras et al., 2000), in fungi (Berthonneau & Mirande, 2000) and in the human lineage (Courseaux & Nahon, 2001). Here we describe several known cases of gene fusion published in the literature.

1.2.1 FUSION GENES IN METABOLIC PATHWAY

Phosphoglycerate kinase (PGK), Glyceraldehyde-3-phosphate dehydrogenase (GAPDH) and triosephosphate isomerase (TIM) catalyze three sequential steps in glycolysis (Figure 1.2). GAPDH catalyzes the conversion of glyceraldehyde-3-phosphate to 1,3-bisphosphoglycerate following its isomerization from dihydroxyacetone phosphate by TIM. PGK takes 1,3-bisphosphoglycerate and transfers one of the phosphates to ADP, creating a new ATP molecule. In mesophilic *Bacillus megaterium*, these three enzymes are encoded by a gene cluster (Schlapfer & Zuber, 1992). However, although coordinately transcribed, they are translated as individual genes.

In the hyperthermophilic bacterium *Thermotoga maritima*, the two glycolytic enzymes PGK and TIM are covalently connected forming a single-chain PGK-TIM fusion protein (Figure 1.2). In contrast to other bacteria, in *T. maritima* the PGK-TIM enzymes are encoded by only one gene containing two open reading frames (ORF);

Figure 1.2

Fusion genes in metabolic pathways. Successive proteins/enzymes in the glycolysis pathway of *B. Megaterium* exists as fused proteins in two other species *T. Maritima* and *P. Oomycote*. PGK, phosphoglycerate kinase; GAPDH, glyceraldehyde-3-phosphate dehydrogenase; TIM, triosephosphate isomerase. This illustration is reproduced from Schlapfer & Zuber (1992); Schurig *et al.* (1995) and Unkles *et al.* (1997)

the *pgk* ORF ends with a stop codon that overlaps the TIM coding sequence *tpi*; the *tpi* gene lacks a separate start codon and promoter elements (Schurig *et al.*, 1995). A frameshift mechanism was proposed to explain the expression of monomeric free PGK, in addition to the bifunctional PGK-TIM enzyme. By comparing both the monomeric PGK and the separated recombinant TIM variant with the natural bifunctional PGK-TIM fusion protein, the linkage of the two enzymes was shown to enhance both the intrinsic stability and the catalytic efficiency of TIM (Beaucamp *et al.*, 1997).

As shown in Figure 1.2, genes encoding TIM and GAPDH are fused and form a single transcriptional unit in *Phylum Oomycota* (Unkles *et al.*, 1997). In *P. Oomycota*, the *tpi* region (encoding TIM) is 5' to the *gpd* region (encoding GAPDH) and a true transcriptional fusion is present, resulting in a mRNA with both *tpi* and *gpd* in the same reading frame. This example shows the two fusion genes that link three nonhomologous genes.

1.2.2 FUSION GENES AND PROTEIN-PROTEIN INTERACTION

The interactions made by a protein, or a protein domain, are an essential part of its function. Interactions between proteins in cells have been studied experimentally on a case-by-case basis for a long time, using a wide variety of biochemical, physical and genetic methods. Edward Marcotte, David Eisenberg and colleagues (1999) described a novel approach to identify protein-protein interactions using gene fusion strategies. It is based on the observation that genes encoding proteins that interact in one organism are sometimes genetically fused (and thus encode a single polypeptide chain) in others. An interface between two interacting component proteins is more

likely to evolve when the proteins are fused into a single chain. They used 4,290 proteins from *Escherichia coli* (as inferred from their ORFs) to search for fused 'homologues' in other organisms (as shown in Figure 1.3). For each triplet proteins, the first protein (the Rossetta Stone protein) is a fusion of the second and third proteins (the component proteins) from *E. coli*. The component proteins of the first three triplets are demonstrated to interact experimentally. For example, the gyrase A and gyrase B subunits of *E. coli* DNA gyrase interact to each other (Sugino et al., 1980) and they are fused into a single chain (topoisomerase II) in yeast. Similarly, the interacting protein α and β subunits of acetate-CoA transferase (Yeh & Ornston, 1981) are fused as succinyl-CoA transferase in human; interacting protein α and ϵ subunits of DNA polymerase III (McHenry & Crow, 1979) are fused as DNA polymerase III α in *B. subtilis*. The component proteins of the final two triplets are not known to interact directly. However, they show pairs of proteins from the same pathway. Histidine biosynthesis HIS2 and HIS10 catalyze two non-sequential steps of the histidine biosynthesis pathway. Γ -glutamyl phosphate reductase and glutamate-5-kinase catalyze the first two steps of the proline biosynthesis pathway. Using two different search methods and the ProDom database (Corpet, 1998), which contains a listing of all known protein domains; they found 6,809 'triplets' (about 1.5 connections per *E. coli* protein). When they applied the same approach to the 5,800 putative proteins of *Saccharomyces cerevisiae*, they found 45,000 such connections.

Figure 1.3

Five examples of *E.coli* proteins predicted to interact by the gene fusion analysis. A triplet of proteins is shown for each case: The second and third proteins are predicted to interact because their homologs are fused in the first protein. The first three predictions are known to interact from experiments. The final two examples show pairs of proteins from the same pathway that is not know to interact directly. This illustration is reproduced from Marcotte et al. (1999).

1.2.3 FUSION GENES IN ALTERNATIVE SPLICING

Pre-mRNA splicing is an essential step in the expression of most eukaryotic genes. Alternative splicing has recently emerged as a major mechanism for regulating the repertoire of gene functions (Gravely, 2001). Many groups have recently presented genomic analyses of alternative splicing suggesting that at least a third, and probably the majority of human genes are alternatively spliced (IHGSC, 2001; Mironov et.al., 1999; Brett et al., 2000; Kan et. al., 2001). Thomson et al. (2000) showed the formation of a fusion gene by alternative splicing.

In *C. elegans* and *D. melanogaster*, *Kua* and *UEV* are in separate loci, and are expressed as independent transcripts. In humans, *Kua* and *UEV* are adjacent genes, expressed either as separate transcripts encoding independent *Kua* and *UEV* proteins, or as a hybrid *Kua-UEV* transcript, encoding a two-domain protein. The schematic molecular process of *Kua-UEV* gene fusion is shown in Figure 1.3. An authentic gene fusion should possess a particular mechanism to override the nonsense codon used to stop translation of the N-terminal protein (eg. mutation). However, *Kua-UEV* employs alternative splicing to skip the exon of *Kua* that contains the *Kua* stop codon and exon of *UEV* that contains a translation initiation codon. Given that many vertebrates genes often contain long UTR regions and an intergenic region, alternative splicing may be an efficient mechanism to avoid the stop codon in upstream gene(s), as represented by the *Kua-UEV* gene.

Figure 1.4

The molecular process for *Kua-UEV* gene fusion (*Kua-UEV* represents the fusion gene). *UEV* and *Kua* are two genes that are located in different loci in *C. elegans* and *D. melanogaster*. In the human genome these two genes are adjacent and a portion of RNA transcripts from the two genes are fused into a single RNA. A chimerical protein KUA-UEV is a fusion protein translated by this fused RNA. This illustration is reproduced from Long (2000).

1.2.4 FUSION GENES IN RECOMBINATION

Historically much evolution has occurred through recombination of existing designs rather than through other kinds of random change. Recombination can result in proteins with new arrangements of domains, which are the evolutionary units of genes and proteins. Gene fusion is one kind of recombination, which results in one long composite protein. The gene encoding bifunctional glutamyl-prolyl-tRNA synthetase was suggested to arise through recombination between genomic sequences encoding the repeated units (Berthonneau & Mirande, 2000).

The genes of glutamyl- and prolyl-tRNA synthetases (GluRS and ProRS) are organized differently in the three kingdoms. In bacteria and archaea, distinct genes encode the two proteins. In higher eukaryotes, the two polypeptides are carried by a single polypeptide chain (glutamyl-prolyl-tRNA synthetase) to form a bifunctional protein (Cerini et al., 1991). According to the phylogenetic analysis of the extant motifs, Berthonneau and Mirande (2000) proposed an evolutionary scenario to account for the emergence of this bifunctional protein GluProRS. They proposed that GluRS and ProRS acquired homologous nucleotide sequences encoding repeated units (Figure 1.5A). These homologous DNA segments would have been the target of a reciprocal translocation event leading to the fusion of GluRS and ProRS genes to generate a bifunctional polypeptide (Figure 1.5B). Modern GluProRSs possess three (human and hamster) or six (fly) repeated units in their linker region. They suggested that if an GluProRS with six units predated the species with three repeats, the deletion of genetic material might have occurred via a cross-over mechanism (Figure 1.5C).

Figure 1.5

Hypothetical scenario of gene fusion to yield a bifunctional aminoacyl-tRNA synthetase. (A) Two independent loci carrying a GluRS and a ProRS of the *C. elegans* type are schematized. The repeated units are indicated by gray or black boxes. The homologous nucleotide sequences are represented by black boxes. (B) Reciprocal translocation through homologous sequences encoding the repeated units indicated in black led to a bifunctional synthetase. (C) Cross-over events might have occurred through the six repeated units to give GluProRS polypeptides with three repeated units. This illustration is reproduced from Berthonneau & Mirande, 2000.

1.3 FUSION GENE DATABASES

1.3.1 DOMAIN FUSION DATABASE

In 2003, the first database for gene fusion events was set up by Ikuar laboratories from Toronto University (Truong & Ikura, 2003). The purpose of this Domain Fusion Database is to disseminate information about domain fusion events detected in yeast and human. Beginning with Pfam's domain prediction in the SWISS-PROT+TrEMBL database, they took use of relational algebra operations using SQL to identify putative functional linkages (Figure 1.6a). They identified 235 and 189 putative fusion proteins in *H. sapiens* and *S. cerevisiae*, respectively. The limitation of this database is it depends on the accuracy of domain assignments of the SWISS-PROT+TrEMBL database. This database is currently available in the public domain <http://calcium.uhnres.utoronto.ca/pi>.

1.3.2 FUSIONDB

FusionDB constitutes a resource dedicated to in-depth analysis of bacterial and archaeal gene fusion events (Suhre & Claverie, 2004). The putative gene fusion events were checked from 51 genomes for which Cluster of Orthogonal Groups (COG) annotation of their genes were available in all 89 fully sequenced bacterial and archaeal genomes (Figure 1.6b). By using BLAST comparisons, a total of 49225 putative fusion genes were identified. The limitation of FusionDB is it only checked gene fusion events in the prokaryotic genomes without taking eukaryotes into consideration. This database is currently available in the public domain <http://igs-server.cnrs-mrs.fr/FusionDB/>.

Figure 1.6

Flowcharts for the two databases: (a) the Domain Fusion Database (Truong & Ikura, 2003) used adopted relational algebra operations to find fusion proteins in humans and yeast. (b) the FusionDB (Suhre & Claverie, 2004) searched fusion proteins between bacteria which have COG (Cluster of Orthogonal Groups) annotation.

1.4 ADVANTAGES OF GENE FUSION

1.4.1 PRODUCE NOVEL GENES

Origin of new genes with novel functions is a fundamental biological process in nature. It is now clear that fusion of genes is one of the dominant mechanisms for the creation of new genes. Thomson et al. (2000) presented a clear example of new genes with novel functions can originate from fusion of two old genes in humans (Figure 1.4). Experiments showed that UEV is a nuclear protein, whereas both Kua and Kua-UEV localize to cytoplasmic structures. This indicates that the Kua domain determines the cytoplasmic localization of Kua-UEV. Therefore, the addition of a Kua domain to UEV in the fused Kua-UEV protein confers new biological properties to UEV. Gene fusion may be a critical step toward creating a new gene with novel function.

1.4.2 EXPAND SUBSTRATE SPECIFICITY

The realization that many enzymes have broad substrate specificity fueled much of the growth in biocatalysis over the last twenty years. By acquiring additional active domain, fusion genes can expand their substrate specificity. For example, the multidomain transmembrane protein DsbD catalyzes the transfer of electrons from the cytoplasm to the periplasm of *Escherichia coli*. Most bacterial species do not have DsbD, but instead their genomes encode a much smaller protein, CcdA, which resembles the central hydrophobic domain of DsbD. DsbD and CcdA are functional homologs. DsbD is organized into three domains: an N-terminal domain with an immunoglobulin (Ig) fold; a central hydrophobic core and a C-terminal periplasmic

domain with an apparent thioredoxin-like fold (Figure 1.7). While DsbD transfers reducing potential to periplasmic protein disulfide bond isomerases and to the cytochrome *c* thiorreduction pathway, CcdA appears to be involved only in cytochrome *c* biogenesis. Katzen and colleagues (2002) provided strong evidence that, by the acquisition of additional thiol-redox active domains, DsbD expanded its substrate specificity. Hence, the fusion of a thioredoxin-like and a thiol-redox active module resulted in the expansion of the substrate specificity of a transmembrane enzymatic core.

1.4.3 REGULATORY CONTROL BY FUSED CATALYTIC SITES

Catalytic efficiency is generally improved in multi functional enzymes. For example, GTP-cyclohydrolase-II-dihydroxybutanone-phosphate-synthase (GTP2-DHBP) catalyzes reactions in a convergent pathway for riboflavin synthesis. The fusion protein might be a kinetic solution to ensure a stoichiometric supply of products. Kinetic advantages caused by proximity effects are evident in several cases, including hydroxymethyl-dihydropterin-pyrophosphokinase-dihydropteroate-synthase (HPPK-DHPS) and dehydroquinase-dehydratase-shikimate-dehydrogenase (DHQase-SORase). HPPK-DHPS from pea leaves has relatively much higher DHPS activity, consistent with a kinetic improvement by proximity of the two domains (Mouillon et al., 2002). However, the bifunctional arrangement might be more important for relieving intrinsic product inhibition of HPPK and positioning DHPS as a key

Figure 1.7

Membrane topology and structural domains of DsbD (transmembrane protein) and CcdA (central hydrophobic domain of DsbD). The hydrophobic core of DsbD is homologous to CcdA. The two additional domains at the N-terminus and C-terminus of DsbD expand the substrate specificity. This illustration is reproduced from Katzen et al., 2000.

regulatory enzyme. DHQase–SORase produces 3-dehydroshikimic acid, then shikimic acid. Because SORase is about nine times more active than DHQase (Hermann, 1995), the fusion protein would have a kinetic improvement. However, dehydroshikimate is also an intermediate in a pathway for quinate metabolism. Thus, this functional arrangement might be more useful for preferential routing of carbon to aromatic amino acid synthesis.

Pairs of proteins are generally observed to be interacting and/or functionally associated if they are found as functional proteins in another genome while existing as components in a larger fused protein in another genome (Enright, 1999; Marcotte, 1999). Both Marcotte *et al.* (1999) and Ouzounis and colleagues (Enright, 1999) performed earlier studies on domain and gene fusion events, respectively, that were aimed at predicting new interacting and functionally associated proteins. Their studies were done during a period when the pace of the genome projects was just beginning to pick up. They reported a list of predicted interacting pairs of proteins based on observations of stand-alone genes in one genome found as unique components of a fused gene (Rosetta stone sequence) in another genome.

1.5 BACTERIAL GENES IN HUMAN GENOME

To much curiosity, the initial analysis of the human genome draft sequence claimed that between 113 and 223 genes (around 1% of human genes, Figure 1.8) have been transferred from bacteria to humans (or to one of our vertebrate ancestors) over the course of evolution (IHGSC, 2001). These genes were identified by their highly significant sequence similarity to bacterial genes in BLAST searches, and by their lack

of matches among non-vertebrate eukaryote genes. This claim provoked many challenges and debate.

Shortly after publication of the IHGSC (2001), several investigators acknowledged the shortcomings of the original analysis and applied more rigorous methods to test the bacteria-to-vertebrate transfer (BVT) hypothesis. Salzberg *et al.* (2001) were the first to respond. They used the original set of all sequenced human proteins as queries in a BLASTP search of all complete prokaryotic and eukaryotic genomes. From the thousands of human proteins that matched bacterial proteins with expect (E) values of 10^{-10} or better, they identified 41 proteins from the IHGSC (2001) set, and 46 from the Celera (Venter *et al.*, 2001) set for which bacterial proteins were as potential BVTs. Salzberg *et al.* produced two main findings. First, increasing the numbers of eukaryotic genomes quickly reduced the number of putative interdomain transfers, suggesting that apparent BVT might be the result of sampling bias. Second, they found that gene loss in several different lineages can also account for the lack of non-vertebrate eukaryotic homologs of many human proteins. Together, these findings led Salzberg *et al.* to suggest gene loss from some eukaryotes as a more tenable explanation for the IHGSC's BLAST results.

Stanhope *et al.* (2001) started from a subset of the 113-member BVT list to make their disagreement. They took use of 28 genes whose occurrence in the human genome had been confirmed by PCR and constructed phylogenetic trees for each protein. Although Stanhope *et al.* (2001) analyzed only 28 proteins from the original BVT list, their trees explicitly refute the BVT hypothesis for 16. By this method, they

Figure 1.8

Distribution of the homologues of the predicted human proteins. For each protein, a homologue to a phylogenetic lineage was considered present if a search of the NCBI (National Center for Biotechnology Information) nonredundant protein sequence database, using the gapped BLASTp (Basic Local Alignment Search Tool) program, gave a random expectation (*E*) value of ≤ 0.001 . This picture is reproduced from IHGSC (International Human Genome Sequencing Consortium), 2001.

suggested only one vertebrate-to-bacteria transfer, and indicate that none of the remaining proteins are clear cases of BVT. Though many of the possible BVTs identified by IHGSC were later rejected by several methods mentioned above, it has no doubt that the BVT has received considerable attention.

1.6 STRUCTURAL REARRANGEMENT BY FUSION

Some researchers investigated the structure features of the proteins involved in gene fusion events recently. Hua et al. (2002) did a survey of fusion events within 30 genomes and subsequent structure annotations to the component proteins at a superfamily level using hidden Markov models (HMMs). The fusion map was also constructed to delineate evolutionary relationships among superfamilies. This map showed an evolutionary scenario that might have led to the evolution of some multidomain proteins from a series of domain fusion events. The results showed that proteins with the class α/β fold are frequently involved in fusion events, around 86% of the total 676 assigned single-domain fusion pairs including at least one component protein belonging to the α/β fold class. Since most pairs of component proteins with known functions appear to be metabolic enzymes (Enright et al., 1999; Tsoka & Ouzounis, 2000), the result indicates a potential correlation between α/β folds and enzymes, especially for single-domain proteins, as was suggested by Hegyi & Gerstein (1999).

Figure 1.9

Model for the evolution of the β/α barrel scaffold by twofold gene duplication and fusion. The first gene duplication generates two identical half-barrels that are then fused and adapted into an ancestral β/α barrel. A second gene duplication step leads to the diversification of the ancestral β/α barrel into two enzymes with distinct catalytic activities. This illustration is reproduced from Lang et al., 2000.

Another paper presented structural data on a common protein (Lang et al., 2000). The x-ray structures of the two histidine biosynthetic enzymes HisA and HisF show that both enzymes are folded as β/α barrel with a twofold repeat pattern (Thoma et al., 1999). Lang and coworkers (2000) compared atomic structures of these two enzymes and their result led to a model of the evolution of the HisA and HisF β/α barrels by two successive gene duplications (Figure 1.9). In this model, an ancestral half-barrel sequence is duplicated and fused into a complete β/α barrel.

1.7 OBJECTIVES AND SCOPE

To what extent do the genes exist both in prokaryote and in humans? Is the function of fused genes simply an addition of function in pre-existing component genes? What is the role of fusion genes in the metabolic networks in humans? In order to address these questions, we will investigate the prokaryotic gene fusions in humans in this thesis.

Few cases of gene fusion events have been demonstrated experimentally. To complement these often time-consuming experimental methods, we aim to develop an improved computational method for predicting putative fusion genes. We will then investigate the possible rationale for the evolution of complex metabolic networks with the origin and formation of fusion proteins. Furthermore, we will study one specific fusion example by examining its molecular dynamic behavior.

To date, most gene fusion analysis has compared complete genomes between every protein of the query genome to every protein of the reference (Enright et al., 1999; Marcotte et al., 1999; Tsoka & Ouzounis, 2000; Enright & Ouzounis, 2001).

However, redundancy is very common in sequence databases and redundant data often do not supply biological information. Therefore, we used two sets of non-redundant protein sequences representing the human and bacterial genomes. The specific aims of the projects are outlined below and summarized in Figure 1.10.

- (i) Selection of datasets.
- (ii) Identification of fusion proteins.
- (iii) Validation and classification of fusion proteins.
- (iv) Database development.
- (v) Association of fusion proteins in the citric acid cycle.
- (vi) Significance of fusion protein using molecular dynamic simulation.

The hypothesis underlying this study is that a fusion gene in humans 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 or a more general functional association such as between sequential members in a metabolic pathway. 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. 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.

Figure 1.10

The project workflow is shown. The workflow contains sections such as (1) data source; (2) identification; (3) validation; (4) pathway association and (5) structural significance.

CHAPTER 2

IDENTIFICATION OF FUSION PROTEINS

2.1 INTRODUCTION

Functional and physical associations between fusion partners and fusion products have been discussed previously. 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. To check the fusion genes in a wet-lab can only get a relatively small number of cases and is time-consuming. With the availability of rapidly increasing genomic information, a computerized method for gene fusion detection is required. Previously two methodologies have been presented to check the fusion genes: the reference-against-query comparative genomics (Enright et al., 1999; Marcotte et al., 1999; Tsoka & Ouzounis, 2000) and the relational algebra (Suhre & Claverie, 2004). The limitations of the first method are that it checks gene fusion events in a redundant dataset which may cause bias of results and the component genes are searched within one species. The relational algebra method offers an alternative approach to performing gene fusion analysis, however, the key drawback of this method is its reliance on the accurate construction of a profile HMM for all domains. Here, we design an improved methodology in order to check the gene fusion events in human genome on the basis of completely sequenced bacterial genomes available.

2.2 PROGRAMS USED

CD-HIT (Li et.al., 2001) is a fast and flexible program for clustering large protein databases at different sequence identity levels. CD-HIT stands for Cluster Database at High Identity with Tolerance. Since redundancy is very common in sequence databases and these redundant data often do not supply biological information, it is practical to consider only a set of representatives from the complete database. For example, a non-redundant protein database (NR), where proteins with identical sequences are combined to single entries. CD-HIT can cluster the homologous sequences to reduce the size of large protein database. It uses a 'longest sequence first' list removal algorithm to remove sequences above a certain identity threshold. Additionally the algorithm implements a very fast heuristic to find high identity segments between sequences, and so can avoid many costly full alignments. This software can be downloaded freely from the website http://bioinformatics.org/project/?group_id=350.

BLAST (Basic Local Alignment Search Tool) is a set of similarity search programs designed to explore all of the available sequence databases regardless of whether the query is protein or DNA. By far, BLAST is the most popular tool for searching sequence databases. The BLAST programs have been designed for speed, with a minimal sacrifice of sensitivity to distant sequence relationships (Altschul et al., 1997). The scores assigned in a BLAST search have a well-defined statistical interpretation, making real matches easier to distinguish from random background hits. BLAST uses a heuristic algorithm which seeks local as opposed to global alignments; therefore, it is able to detect relationships among sequences which share isolated regions of similarity (Altschul et al., 1997). Several types of BLAST are available. Blastp uses a query protein to find similar sequences in a

database of protein sequences. It calculates pairwise sequence similarity, and returns a list of similar proteins. Tblastn uses a query protein sequence to search nucleotide sequence databases that have been translated in all possible reading frames. Searching nucleotide databases directly removes the need for proteins to be annotated. The Blast software can be downloaded from the ftp of NCBI (<ftp://ftp.ncbi.nih.gov/blast>).

Multiple sequence alignment techniques (Chenna et al., 2003) compare multiple homologous genes (genes that have similar sequences) to derive conserved segments. Multiple alignments are used to find diagnostic patterns to characterize protein families; to detect or demonstrate homology between new sequences and existing families of sequences; to help predict the secondary and tertiary structures of new sequences. Multiple sequence technique uses the integration of pair-wise alignment between two homologs and the notion of distance between two nucleotide sequences or between two amino-acid sequences. The notion of distance can be derived either as an edit distance – number of mismatches derived after pair-wise alignment of two sequences, or as the evolutionary distance between two microorganisms given by an evolutionary tree. The technique is based upon progressive pair-wise comparison to make intermediate alignments between nearest neighbors – homologs having shortest distance, and has been implemented as a greedy algorithm. One commonly used program for progressive multiple sequence alignment is Clustal W (Thompson et al., 1994). This software is freely downloadable from <ftp://ftp.ebi.ac.uk/pub/software/unix/clustalw/>.

2.3 MATERIALS AND METHODS

2.3.1 DATASETS CONSTRUCTION

Flat files containing 37,490 protein sequences in fasta format (Figure 2.1) derived from the draft human genome were downloaded from NCBI (National Center for Biotechnology Information, ftp://ftp.ncbi.nih.gov/genomes/H_sapiens/protein) to a SGI Origin3000 workstation. This is our dataset 1 (DS1). The paralogs in the human genome data are removed at 40% sequence identity using the clustering program CD-HIT (Li et al., 2001). This process produced 26,673 unique human sequences. A stringent 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 40% (Abagyan and Batalov, 1997). The protein sequences derived from 71 completely sequenced bacterial genomes (<ftp://ftp.ncbi.nih.gov/genomes/bacteria>) form DS2 (Figure 2.1). This includes 16 archaea and 55 eubacteria. The list of bacterial species whose sequences are used in this study is given in Table 2.1 (Appendix 1). 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 bacterial sequences.

2.3.2 IDENTIFICATION OF FUSION PROTEINS

To identify human fusion proteins with bacterial origins, we searched the non-redundant 26,673 human sequences against the non-redundant 102,135 bacterial protein sequences using BlastP program (Altschul et.al., 1997) with a cutoff E-value at 1×10^{-10} (Figure 2.1). For each sequence in the database, BLAST will compare it with the query. BLAST first seeks from the sequence pair, equal length

sequence segments, which have maximal aggregate similarity score that cannot be increased by extension or trimming. Such locally optimal alignments are called "highscoring segment pairs" or HSP's. There are two major parameters in the blast program, the S score and E value. Briefly, the S score is a measure of the similarity of the query to the sequence in the database and the E-value is a measure of the reliability of the S score. So the definition of E-value is the probability due to chance, that there is another alignment with a similarity greater than the given S score. The equation of E-value is:

$$E = Kmn e^{-\lambda S} \quad (2.1)$$

Where the parameters K and λ are the natural scales for the search space and the scoring system respectively. Parameters "m" and "n" represent the sequence length of the query and the length of the database r . S is the score for given HSP. In most cases, E value between 0.001 and 0.01 can be used to infer homology reliably (Grey zone is between 0.01 – 1), but lower (more conservative) values are required when hundreds or thousands of searches are performed. The lower the E value, the more significant the score is. The typical threshold for a good E-value from a BLAST search is 10^{-5} or lower. As in our study, to reduce the false positive result, we chose a very stringent cut-off value (E-value is 1×10^{-10}), in which 1×10^{-10} matches would be expected to occur by chance.

Figure 2.1

Flowchart of of fusion protein identification. This figure shows two datasets are constructed from purged sequences of human and bacteria with CD-HIT at 40% identity. The Dataset 1 is blastp against the Dataset 2 with cutoff e value less than 10^{-10} . Queries with 0 or 1 hit are discarded and queries with more than 2 hits are checked for gene fusion. The query is defined as putative fusion protein if the bacterial proteins match against different region of the query and with overlapping region less than 50% of the shorter bacterial protein length.


```
BLASTP 2.2.3 [Apr-24-2002]

Reference: Altschul, Stephen F., Thomas L. Madden, Alejandro A. Schaffer,
Jinghui Zhang, Zheng Zhang, Webb Miller, and David J. Lipman (1997),
"Gapped BLAST and PSI-BLAST: a new generation of protein database search
programs", Nucleic Acids Res. 25:3389-3402.

Query= gi|4501843|ref|NP_001076.1| alpha-1-antichymotrypsin,
precursor; alpha-1-antichymotrypsin; actichymotrypsin [Homo sapiens]
(433 letters)

Database: prokar
102,135 sequences; 30,239,194 total letters

Searching.....done

Sequences producing significant alignments:
Score E
(bits) Value

ref|NP_484821.1| similar to serine protease inhibitor [Nostoc sp... 143 4e-34
gb|AAM32371.1| Serine protease inhibitor [Methanosarcina mazei G... 139 5e-33

>ref|NP_484821.1| similar to serine protease inhibitor [Nostoc sp. PCC 7120]
Length = 374

Score = 143 bits (361), Expect = 4e-34
Identities = 107/379 (28%), Positives = 170/379 (44%), Gaps = 10/379 (2%)

Query: 46 DLGLASANVDFAFSLYKQLVLKALD-KNVIFSPLSISTALAFSLGAHNNTLLEILKASS 104
D + AN F F L+ ++ + + KN+ SP S++ ALA GA +T + K
Sbjct: 3 DTKIVDANNKFGFKLFSEIAKQDGNAKNISVSPTSVAIALAMTYNGASGSTQKAMAKTLE 62

Query: 105 SPHGDLRLQKFTQSFQHLRAPSISSSDELQLSMGNAMFVKEQLSLDRFTEDAKRLYGSE 164
L + + L + + +QL++ N+++V + +SL F + + Y ++
Sbjct: 63 LQGISL--PELNSGYASLNKLLQNPGENVQLTIANSLWVNQDVSLRPDFVQTTRDFYQAK 120

Query: 165 AFATDFQDSAAAKKLINDYVKNTRGKITDLIKDPDSQTMVLVNYIFFKAKWEMPFDPQ 224
DFQ+ A IN +V+ T+GKI ++ + + + L+N I+F KW FD
Sbjct: 121 VTNLDFQNVTAVENT-INSWVQENTKGKIDKIVDNIEPNQALFLINAIYFNGKWSNAPDKS 179

Query: 225 DTHQSRFYLSKKKWVMPMMSLHHLTIPIYFRDEELSCTVVVELKYTGNASALFILPDQD-K 283
T TP FY + + PMMS Y+ + + G S LP ++
Sbjct: 180 QTFPQTFYTASGQQKQHPMMS-QTGEYRYETPQFQSVSLPYGQDGKVSLEYIFLPTKNSN 238

Query: 284 MEEVEAMLLPETLKRWRDSLEFREIGELYLPKFSISRDNLNDILLQLGIEEAFTSKADL 343
++ L E ++W RE G + LP+F D LND L LG+ EAF++KA+
Sbjct: 239 LKAFSQNLNAENWEKWPQFRNRE-GSIRLPRFQTDYDITLNDALKSLGMAEAFSNKANF 297

Query: 344 SGITGARNLAXXXXXXXXXXXFEEGTEASAATAVKITLLSALVETRTI-VRFNRPFLMI 402
SGI NL EEGTEASAAT+V + S + + +RPF
Sbjct: 298 SGI--GDNLTISQVRHKTIVIEVNEEGTEASAATSVGMVATSLREPQQPFKMIIVDRPFFYA 355

Query: 403 IVPTDTQNIFFMSKVTNPS 421
I ++NI FM V P+
Sbjct: 356 IRDNQSKNILFMGAVVEPN 374
```

Figure 2.2

The example output of blast program. Human protein (accession No. NP_001076) is the query sequence, which has two matches with the bacterial sequences in the database.

As shown in Figure 2.2, the file consists of three parts: the header information at the beginning followed by a summary of the alignments, the alignments, and then some additional summary parameters and statistics at the end. In blast output files, the human protein is defined as Query and the bacterial protein is defined as Subject. This sample output shows two bacterial subject hits from DS2 matched the human query protein from DS1. We next parse the BLAST output file by the program *AnalyseResult.pl* written by Perl language. This program parses the alignments and statistics lines and retrieves the numbers we are interested in, such as the protein length, the protein name, start and end position of matching, score, e-value and identity. For example, the information retrieved from the above sample blast is shown in Figure 2.3. We then check for fusion proteins based on the positions where different hits match with the query sequences.

Since only query proteins with more than two hits are possibly the fusion protein templates, we first parse those queries with zero or one hit. The queries with more than two hits are stored, then these sequences are further analyzed for the presence of composite proteins (fusion proteins) consisting of two or more component proteins (or protein-parts) of bacterial origin (Figure 1.1). The putative human fusion proteins are checked by program *check_fusion.pl* written by Perl language. Figure 2.2 gives the flowchart of this program. The composite proteins are identified using the following two criteria:

- (a) the bacterial proteins show similarity to distinct parts of one human protein, where we name them as alignment 1 and alignment 2;
- (b) if there is a overlapping region between alignment 1 and 2, the maximum overlap is half of the length (50%) of the shorter component proteins.

Figure 2.3

The flowchart for analyzing output of blast. Figures are retrieved and ready for identification of gene fusion. Q is the query (human) sequence, S is the subject (bacterial) sequence, Acc is the Accession Number, PL is the protein length. For the homologous region, Q_start and Q_end represent the start and end position in the human protein. Similaly, S_start and S_end are the associated start and end positions in the bacteria protein.

2.5 RESULTS

Although a number of fusion proteins are reported in literature across several phylogenetic distances, a comprehensive list for human fusion proteins of bacterial origin is still not available. We compared a set of unique human protein sequences (26,673) with a set of unique bacterial protein sequences (102,135). Out of the unique human proteins, 23,814 ($\approx 89\%$) sequences have no homologs with bacteria and 527 ($\approx 2\%$) sequences have only one hit. Those with zero or one hit were discarded (Figure 2.1) since they could not be used as the putative fusion protein template. The rest are 2332 ($\approx 9\%$) human proteins with at least two matching bacterial sequences, which were further identified for the gene fusion event. Finally, our results produced a comprehensive set of 141 human composite proteins consisting of component proteins from bacteria. These fusion genes may have arisen by the fusion of two or more component genes of prokaryotic origin through gene transfer. The results of comparative genome analysis show that human fusion protein formation of bacterial origin is a real, but relatively rare, evolutionary phenomenon.

For example, after analyzing the output result of blast, we found that the N-terminal part of human bifunctional protein pyrroline-5-carboxylate synthetase (P5CS) is homologous to glutamate 5-kinase (GK) of the archaea *M. mazei* and seven eubacteria: *V. cholerae*, *B. subtilis*, *C. tepidum*, *S. pneumoniae*, *S. meliloti*, *B. halodurans*, *A. aeolicus* (Figure 2.4). And the C-terminal part of this human fusion protein shows homology to gamma-glutamyl phosphate reductase (GPR) from six bacteria: *S. sp*, *M. mazei*, *X. fastidiosa*, *C. jejuni*, *C. glutamicum* and *N. sp* (Figure 2.4). Therefore, we suggest that this bifunctional protein is derived from fusion of GK and GPR.

Proteins are made up of domains. Many mosaic genes are formed by the fusion of domains rather than the entire genes. In the evolution of eukaryotes, domain fusion takes the form of domain accretion, which is step-by-step addition of new domains to pre-existing cores. Despite the huge number of genes in an organism, the protein coding genes are thought to be made from a limited number of basic protein motifs. Evolution has reused these motifs, combining them to form different tertiary structures with distinct biochemical functions. Therefore, we can study gene fusion from the protein domain level, which represent one of the most useful levels to understand protein function.

The domain architectures of component proteins are different for different fusion case. In Figure 2.5a, we show that the composite human protein peroxisomal D3, D2-enoyl-CoA isomerase (PEI) is formed by the fusion of two component proteins: (1) acyl-CoA binding protein (ACBP) and (2) enoyl-CoA hydratase/isomerase (ECH/ECI). The putative human fusion protein, PEI is made up of three domains, namely, ACBP, ECH and ECI. The ACBP domain is homologous to the *R. solanacearum* ACBP domain and the ECH domain is homologous to the *P. aeruginosa* ECH domain. Thus, this example clearly demonstrates that the putative multi-domain protein is formed by the fusion of two or more bacterial domains which were originally single domain in nature. It is also possible that the component protein itself is a multi-domain protein. For example, the human composite protein Gephyrin is shown to be formed by fusion of molybdopterin biosynthesis enzyme from *C. acetobutylicum* and Molybdenum cofactor biosynthesis protein A from *A. aeolicus* (Figure 2.5b). It is possible that the human composite protein have accreted the ability to perform the function of both the component proteins.

Figure 2.4

Protein sequences which share significant sequence homology with human pyrroline-5-carboxylate synthetase (P5CS). Human P5CS was used as query and homology search was done using BLASTp program as described in the Material and Methods section. Regions of the proteins bearing homology to human P5CS are shown by line diagrams. The N-terminal of P5CS is homologous to glutamate 5-kinase and the C-terminal to Gamma-glutamyl phosphate reductase.

Figure 2.5

Two examples for bacterial domain fusions in humans are shown. Domains in each proteins are shown by rectangular boxes and are assigned using Prodom. (a) The formation of human peroxisomal D3,D2-enoyl-CoA isomerase by the fusion of two single domain proteins: ACBP from *R. solanacearum*, and ECH from *P. aeruginosa* (b) The formation of human Gephyrin by the fusion of one single domain protein Molybdopterin biosynthesis enzyme from *C. acetobutylicum* and N-terminal and central- domains of Molybdenum cofactor biosynthesis protein A from *A. aeolicus*. Domain name abbreviations: ACBP: acyl-CoA-binding protein; ECH: enoyl-CoA hydratase; ECI: enoyl-CoA isomerase; MOCF_1/2: molybdenum cofactor biosynthesis protein signature 1/2; GCCF: Gephyrin cytoskeleton cofactor biosynthesis protein; MoeA_N/C: molybdopterin cofactor N-terminal/Central- region.

2.6 FUSION DATABASE

Gene fusion events have been proposed to represent valuable ‘Rosetta stone’ information for the identification of potential protein–protein interactions and metabolic or regulatory networks (Galperin & Koonin, 2000; Sali, 1999). For the purpose of digging deep for the enormous wealth of fusion genes, an online database FUSION is constructed. The aim of this FUSION database is to help researchers to identify interesting fusion genes for further experimental study.

FUSION database can be accessed freely at URL <http://sege.ntu.edu.sg/wester/fusion>. It can be fully searchable in a variety of ways: (i) by accession number (Display a NRIC Refseq number); (ii) by gene name; (iii) by GI number; (iiii) by key word search with AND/OR operators; or (v) by entering an amino acid sequence in FASTA format (Figure 2.6). In cases where a putative human fusion protein has paralogous proteins, output in full-page mode contains identity percentage, chromosome information and multiple alignments are provided (Figure 2.6). FUSION database constitutes a resource dedicated to in-depth analysis of gene fusion events of bacterial and archaeal origin. Together, this database provides the complete set of information required for in-depth analysis of non-homology-based gene function attribution. While many further experimental approaches are needed for studying gene fusion and its functional importance, these features may help many researchers make a direct step from finding novel fusion genes in this database and to investigate them rapidly in the lab. For instance, biologists may predict operon structures in bacterial genomes; predict protein-protein interaction; propose metabolic and regulatory networks.

2.7 SUMMARY

Identification of human fusion proteins of bacterial origin is both non-trivial and computationally intensive. Here, we described a comparative genomics approach to identify human fusion proteins. The identified fusion proteins are stored in a database called FUSION. The FUSION database constitutes a resource dedicated to in-depth analysis of gene fusion events of bacterial and archaeal origin. Such events can provide the 'Rosetta stone' in the search for potential protein-protein interactions, as well as metabolic and regulatory networks. Moreover, the FUSION database extends the notion of gene fusion from that of genes in one genome to that of genes from different genomes. Multiple sequence alignments are available for paralogous human fusion proteins. Fusion database is fully searchable with access to NCBI accession number, GI number and text data at all levels. Together, this database provides the complete set of information required for in-depth analysis of non-homology-based gene function attribution.

CHAPTER 3

FUNCTIONAL EVALUATION OF FUSION PROTEINS

3.1 INTRODUCTION

Gene fusion has been identified across various phylogenetic groups and this suggests that there exist processes other than vertical inheritance during evolution (Genereux & Logsdon., 2003). Fusion genes gain added advantage in higher organism by coupling biochemical/signal transduction reactions through tight regulation of fusion partners, compared to individual fusion partners in lower organisms (Tsoka & Ouzounis, 2001). Thus, fusion genes produce proteins with novel or enhanced function. Yanai and colleagues used gene fusion to establish links between fusion genes and functional network of their involvement (Yanani et al., 2001). As previously introduced in Chapter 1, gene fusion has also been used to illustrate protein-protein interactions (Marcotte et al., 1999), novel gene function (Long, 2001), enhanced substrate specificity (Katzen et al., 2002) and multi-functional enzyme specificity (Berthonneau & Mirande, 2000). An interesting relational algebra approach has also been demonstrated to identify fusion proteins across different phylogenetic distances (Truong & Ikura, 2003). Therefore, identification and characterization of fusion genes in the human genome will shed light into its evolutionary biology. Herein in this chapter, we report these genes of which many are found to mimic bacterial operons and simulate protein-protein interfaces. Few others are also known to exhibit multiple functions and alternative splicing.

3.2 FUNCTIONAL INFERENCES TO FUSION PROTEINS

The experiment in chapter 2 identified 141 human proteins consisting of two or more fusion partners of bacterial origin. 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 bacterial systems (Table 3.1). Our interest is to infer accreted functions by fusion proteins compared to their fusion partners in a bacterial system. Hence, these fusion proteins are grouped into four categories using functional inferences (Figure 3.1). These four groups are classified as follows:

- (a) Fusion proteins mimicking operons in bacteria. In this category, the component proteins have homologous gene pairs which located in one operon of any bacteria species.
- (b) Fusion proteins with multiple functions. The composite proteins have more than one biological function while their associated component proteins are usually mono-functional protein.
- (c) Fusion proteins exhibiting alternative splicing. The composite proteins have several variants arising from alternative splicing while the component proteins are homologous to the variants.
- (d) Fusion proteins simulating bacterial protein-protein interfaces. The component proteins are identified to interact to each other so the protein-protein interface transformed to domain-domain interface.

Figure 3.1

Functional tree of fusion proteins classification.

Table 3.1: Human fusion proteins of bacterial origin

Human Fusion Proteins				Fusion Partners of Bacterial Origin						Comments				
S/N	RefSeq Acc.	Ch	Pro. length	Protein	Pathway	N terminal		C terminal		Protein length	Match region	Protein	Pathway	
						Pro. length	Match region	Protein	Pathway					
Fusion proteins mimicking bacterial operons														
1	NP_004276	1	791	Glucose dehydrogenase / 6-phospho gluconolactonase	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 <i>P. putida</i> (Petruschka, 2002) Bifunctional enzyme (Clarke, 2001)
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 <i>E. coli</i>)	411	440..750	3-hydroxyacyl-CoA dehydrogenase	Lipid Metabolism; Amino Acid Metabolism; Carbohydrate Metabolism (in <i>E. coli</i>)	fadAB operon in <i>E. coli</i> (Yang, 1983) Trifunctional protein (Orti, 1999)
3	NP_004332	2	2225	Carbamoyl phosphate synthetase 2 / Aspartate trans carbamylase/ Dihydro-orotate 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 <i>S. typhimurium</i> (Kilstrup, 1988) Trifunctional protein (Chen, 1989)
4	NP_001150	2	1338	Aldehyde oxidase 1	Amino Acid Metabolism; of Cofactors and Vitamins	489	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 <i>S. acidocaldarius</i> (Thia-Toong, 2002)
5	NP_001957	3	723	Enoyl-CoA hydratase/ 3-hydroxyacyl CoA dehydrogenase	Lipid Metabolism; Amino Acid Metabolism; Carbohydrate Metabolism	297	5..277	carbon-monoxide dehydrogenase small subunit	Energy Metabolism (in <i>M. loli</i>)	799	580..1316	carbon monoxide dehydrogenase, large subunit	Energy Metabolism (in <i>M. loli</i>)	cdh operon in <i>M. soehngenii</i> (Eggen, 1991)
														fadAB operon in <i>E. coli</i> (Yang, 1983) Bifunctional protein (Hoeftler, 1994)

6	NP_00105	3	1621	DNA topoisomerase II, β	773	56..727	DNA gyrase, subunit B (gyrB)	490	728..1054	DNA Gyrase Subunit A		Operon in <i>M. tuberculosis</i> (Ummiraman, 2002) Protein-protein interaction (Sugino, 1980)
7	NP_00092	4	1174	DNA directed RNA polymerase II polypeptide B	550	28..535	DNA-directed RNA polymerase, subunit B "	649	565..1172	DNA-dependent RNA polymerase subunit B '	Genetic Information Processing	RNAP operon in <i>M. thermoautotrophicum</i> (Smith, 1997)
8	NP_00543	4	624	3'-phosphoadenosine 5'-phosphosulfate synthetase	186	51..221	adenosine 5'-phosphosulfate kinase	459	273..619	ATP sulfurylase	Nucleotide Metabolism; Metabolism of Amino Acids; Energy Metabolism (in <i>E. coli</i>)	cys operon in <i>E. coli</i> (Leyh, 1992) Bifunctional enzyme (Venkatachalam, 1998)
9	NP_00644	4	425	AIR carboxylase; SAICAR synthetase	233	12..247	Phosphoribosyl aminoimidazole succinocarboxamide (SAICAR) synthetase (purC)	180	267..421	Phosphoribosyl aminoimidazole(AIR) carboxylase (purE)	Nucleotide Metabolism (in <i>Bacillus subtilis</i>)	pur Operon in <i>Bacillus subtilis</i> (Ebbole, 1987) Multifunctional protein (Mimet, 1990)
10	NP_00042	5	520	Succinyl CoA: 3-oxoacid CoA transferase	250	36..284	Acyl CoA:acetate CoA transferase, subunit α	219	302..516	Acyl CoA:acetate CoA-transferase subunit β	Carbohydrate Metabolism (in <i>E. coli</i>)	Operon in <i>C. acetobutylicum</i> (Cary, 1990) Protein-protein interaction (Yeh, 1981)
11	NP_03647	5	1086	nicotinamide nucleotide transhydrogenase	530	48..587	NADP transhydrogenase α subunit	480	618..1081	NADP transhydrogenase β subunit	Metabolism of Cofactors and Vitamins	mnt operon in <i>R. rubrum</i> (Yamaguchi, 1994)
12	NP_00593	6	636	molybdenum cofactor synthesis-step 1 protein isoform	333	61..366	molybdenum cofactor biosynthesis protein A	156	481..629	molybdenum cofactor biosynthesis protein CB		moa operon in <i>E. coli</i> (Rivers, 1993)
13	NP_00903	9	918	sarcosine dehydrogenase	382	63..445	sarcosine oxidase, subunit β	379	483..894	sarcosine oxidase, subunit α	Amino Acid Metabolism	sox operon in <i>Corynebacterium sp.</i> (Chlumsky, 1995)

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 <i>M. thermotautotrophicum</i> (Smith, 1997)
15	NP_002851	10	795	pyrroline-5-carboxylate synthetase	Amino Acid Metabolism; Carbohydrate Metabolism	356	72..381	glutamyl kinase	5- Amino Acid Metabolism	484	363 .. 770	γ -glutamyl phosphate reductase	Amino Acid Metabolism	ProBA operon in <i>T. thermophilus</i> (Kosuge, 1994) Bifunctional enzyme (Aral, 1996)
16	NP_000911	11	1178	pyruvate carboxylase precursor	Carbohydrate Metabolism	477	38..483	pyruvate carboxylase, subunit A	Amino Acid Metabolism; Carbohydrate Metabolism (in <i>S. mutans</i>)	567	563..1178	pyruvate carboxylase, subunit B	Amino Acid Metabolism; Carbohydrate Metabolism (in <i>T. tengcongensis</i>)	operon-like arrangement in <i>M. barkeri</i> (Mukhopadhyay, 2001)
17	NP_001087	17	1105	ATP citrate lyase	Genetic Information Processing	398	41..418	citrate lyase, subunit 1	Carbohydrate Metabolism; Energy Metabolism (in <i>E. coli</i>)	610	496..1089	citrate lyase, subunit 2	Carbohydrate Metabolism; Energy Metabolism (in <i>E. coli</i>)	Probable operon in <i>K. pneumoniae</i> (Bott & Dimroth, 1994)
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 <i>M. thermotautotrophicum</i> (Smith, 1997)
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 (Cerini, 1991)
20	NP_000405	5	736	hydroxysteroid (17-beta) dehydrogenase	Lipid Metabolism	303	1..301	short-chain dehydrogenases	Lipid Metabolism	286	327..607	MaoC like dehydratase		Multifunctional enzyme (Leenders, 1998)

21	NP_005467	9	722	UDP-N-acetylglucosamine-2-acetylmannosaminidase	Metabolism of Complex Carbohydrates	377	11..349	UDP-N-acetylglucosamine 2-epimerase	Metabolism of Complex Carbohydrates	315	412..675	glucose kinase	Carbohydrate Metabolism; Metabolism of Complex Carbohydrates ; Biosynthesis of Secondary Metabolites	Bifunctional enzyme (Lucka, 1999)
22	NP_005947	14	935	methylentetrahydrofolate dehydrogenase/cyclohydrolase; formyltetrahydrofolate synthetase		284	5..293	methylentetrahydrofolate dehydrogenase (FolD)	Carbohydrate Metabolism; Metabolism of Cofactors and Vitamins	599	319..925	Formyltetrahydrofolate synthetase	Carbohydrate Metabolism; Metabolism of Cofactors and Vitamins	Trifunctional protein (Hum, 1988)
23	NP_004095	17	2509	fatty acid synthase	Fatty acid biosynthesis	1610	3..842	polyketide synthase		325	1557..1852	quinone oxidoreductase		Multifunctional enzyme (Wakil, 1989)
24	NP_006648	21	541	formiminotransferase cyclodeaminase	Amino Acid Metabolism; Metabolism of Cofactors and Vitamins	298	1..290	Glutamate formiminotransferase	Amino Acid Metabolism; Metabolism of Cofactors and Vitamins	217	378..523	Formiminotetrahydrofolate cyclodeaminase	Metabolism of Cofactors and Vitamins	Bifunctional enzyme (Beaudet, 1976)
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 (Brodsky, 1997) Trifunctional protein (Brodsky, 1997)
Fusion proteins exhibiting alternative splicing														
26	NP_037361	12	1237	apoptotic protease activating factor isoform a	Cell Growth and Death	1381	109..493	regulatory protein		626	603..1191	WD-40 protein		Alternative splicing (Ogawa, 2003)
27	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 (Brodsky, 1997) Trifunctional protein (Brodsky, 1997)
Fusion proteins simulating bacterial protein-protein interfaces														
										222	806..1004	GART	Nucleotide Metabolism	

28	NP_006587	3	2724	polymerase (DNA directed), theta	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 (Dong, 1996)
29	NP_000246	6	750	Methylmalonyl Co A mutase precursor	Lipid Metabolism	681	42..519	Methylmalonyl Co A mutase, subunit β	Lipid Metabolism	144	611..749	Methylmalonyl Co A mutase, subunit α		Protein-protein interaction (Marsh, 1989)
30	NP_001084	12	2483	acetyl-Coenzyme A carboxylase beta	Metabolism of Complex Carbohydrates	677	256..960	propionyl-CoA carboxylase, subunit α	Metabolism of Complex Carbohydrates	516	1885..2267	Propionyl-CoA carboxylase subunit β	Carbohydrate Metabolism; Metabolism of Complex Carbohydrates; Biosynthesis of Secondary Metabolites	Protein-protein interaction (Haase, 1984)

3.2.1 FUSION PROTEINS MIMICKING OPERONS IN BACTERIA

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 bacteria. In bacteria, genes involved in a related pathway are arranged as operons. These operons facilitate the coordinated regulation of genes, as the clustered genes are transcribed into one messenger RNA (Figure 3.2). Operon transcripts always code for more than one protein, and bacteria handle this by starting translation of the mRNA into amino acids separately at the beginning of each protein-coding section. However, most eukaryotes do not have operons, except for one group of animals—the nematodes (small, unsegmented worms) (Mering & Bork, 2002). Thus these ‘polycistronic’ transcripts are a problem for eukaryotes. 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 (Figure 3.2).

The fusion protein pyrroline 5-carboxylate synthetase (P5CS) catalyzes ATP and NAD(P)H-dependent conversion of L-glutamate to glutamic gamma-semialdehyde (GSA), the metabolic precursor for proline biosynthesis. The phenotypic features for deficiency of P5CS include joint hyperlaxity, skin hyperelasticity, cataract and mental retardation with hyperammonemia and low plasma levels of proline, citrulline and ornithine (Kamoun et al., 1998). The P5CS gene encodes a bifunctional enzyme, with both gamma-glutamyl kinase (gamma-GK) and gamma-glutamyl phosphate reductase (gamma-GPR) activities that catalyzes the first 2 steps in proline biosynthesis (Aral et al, 1996). This protein has high level of sequence similarity to two proteins from the archaeon *Methanosarcina mazei* Goel: to glutamate-5-kinase (GK) in the N-terminal region

and to gamma-GPR in the C-terminal region (Figure 2.2, Figure 2.3). In *T. thermophilus*, these two proteins are separately translated in one operon, GK preceding GPR, involving in the pathway of urea cycle and metabolism of amino group (Kosuge, 1994). This suggests that fusion proteins in humans are formed by the fusion of two or more fusion partners.

A similar example using the human composite protein - succinyl coA transferase suggested that it is formed as a result of gene fusion between *E. coli* Co-A transferase α and Co-A transferase β . Experiments show that these two component proteins interact together in *E. coli*. Thus, genes whose products need to interact closely in *E. coli* have a noticeable tendency to be fused into a single gene in humans, encoding a combined polypeptide in which the proteins have a higher chance of interacting together. However, experimental data for the interactions of component proteins in *Methanosarcina mazei* Goel is required to establish evolutionary gene fusion as an effective route for efficient protein-protein interactions in humans. Sixteen more cases are listed in Table 3.1. Since enzyme genes in operons tend to catalyze successive reactions in metabolic pathways. Alternatively, genes in operons interact with each other either by physical assembly of their products or through participating networked reactions in cells, so that they are likely to be involved in closely related biological processes.

Figure 3.2

Schematic presentation of fusion proteins mimicking operons in bacteria. In bacteria, a four-gene operon, where gene 1 is next to gene 2, is under the control of regulatory elements (promoter, operator) and factors that bind to these elements. Transcription of the cluster results in a single molecule, a multi-gene transcript of mRNA. In humans, genes are transcribed into separate transcripts then translated into different proteins. In this case, protein 1 and 2 in humans are fused by a linker.

3.2.2 FUSION PROTEINS SHOWING ALTERNATIVE SPLICING

Recent genome-wide analyses of alternative splicing indicate that 40-60% of human genes are alternatively spliced, suggesting that alternative splicing is one among the most important processes of human biology (Mironov et.al., 1999; Brett et al., 2000; Kan, et al., 2001). Two fusion proteins are shown to exhibit alternative splicing from this study (Table 3.1). A classified example is the GARS-AIRS-GART gene and this gene produces two spliced variants as shown in Figure 3.4. The two protein isoforms play an essential role in the regulation of a biochemical pathway. The human protein GARS-AIRS-GART consists of three domains: (1) glycinamide ribonucleotide synthetase (GARS); (2) aminoimidazole ribonucleotide synthetase (AIRS) and (3) glycinamide ribonucleotide formyltransferase (GART). The human GARS-AIRS-GART gene encodes proteins for three of the 10 enzymatic steps necessary for the conversion of phosphoribosyl pyrophosphate to inosine monophosphate by the *de novo* purine pathway (Figure 3.3). The three enzyme activities are encoded in a linear, non-overlapping fashion on the GARS-AIRS-GART mRNA (Aimi et al., 1990). The GARS, AIRS and GART domains have homologous domains with origin from *Bacillus halodurans*, *Agrobacterium tumefaciens* and *Mesorhizobium loti*, respectively (Figure 3.4). The GARS-AIRS-GART gene is shown to produce a tri-functional protein (110 Kda) with GARS, AIRS and GART enzymatic activities and also a mono-functional GARS protein (50 Kda) by alternative splicing (Brodsky et al, 1997). They showed that the use of a poly-adenylation site in the intron between the terminal GARS and the first AIRS exons by alternative splicing produced the GARS mono-functional protein. In humans, the GARS message contains an in-frame stop codon consisting of the

TAA at the 5' splice donor site. This results in the expression of a GARS protein which is identical to the GARS domain in the tri-functional protein.

The fusion of GARS, AIRS and GART enzymatic activities during evolution into a single tri-functional protein may have occurred to allow coordinate regulation of the three enzymatic activities. This gene fusion has also been postulated to be due to a selective advantage incurred with the ability to channel labile intermediates between the covalently linked enzyme active sites. A separate monofunctional GARS protein may be necessary to release metabolites from the purine pathway so that they may be used for other metabolic pathways. Alternatively, the separate GARS protein may interact directly with the first enzyme in the *de novo* pathway, phosphoribosylpyrophosphate amidotransferase, to facilitate the transfer and prevent the degradation of the highly labile phosphoribosylamine (Rudolph & Stubbe, 1995). Separate GARS and GARS-AIRS-GART mRNAs have been observed in humans, mouse, chicken and *D. melanogaster* (Brodsky et al., 1997). One more case is listed in Table 3.1.

3.2.3 FUSION PROTEINS EXHIBITING MULTIPLE FUNCTIONS

Multifunctional enzymes have been postulated to have arisen from smaller monofunctional proteins, and each enzymes activity exists as a separate active domain in the multifunctional protein. We also find some of the human fusion polypeptides are mosaic of functional domains, which derived their enzyme activities from the fusion of the smaller genes encoding the monofunctional proteins in bacteria. 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 to one another. 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 (Reed et al., 1974). By sequestering reactive intermediates, their conversion by undesired chemical reactions is prevented as substrates are channeled from one catalytic site to the next (Perham, 1975). A covalently linked multifunctional complex is likely to be more stable than one formed by non-covalently formed protein-protein interfaces containing reaction (or active) centers. Thus, fusion of two or more mono-functional bacterial proteins into a single polypeptide in a higher organism is certainly under selective advantage in the course of protein evolution. Since the enzyme activities in the polypeptides are coupled, also better and more efficient regulation of the overall reaction is possible.

The fusion protein GARS-AIRS-GART exhibits multiple functions in humans (McCarthy and Hardie, 1984). Each of GARS, AIRS and GART proteins are mono-functional and part of the *pur* operon in *B. subtilis* and *E. coli* (Ebbole et al., 1987). The GARS-AIRS is a bifunctional protein in *S. cerevisiae* and GARS-AIRS-GART is tri-functional in *Drosophila* (Ebbole et al., 1987). In humans, 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 3.1.

Figure 3.3
Pathway of de novo purine biosynthesis to IMP. In humans, the second, third and fifth steps of de novo purine biosynthesis are catalyzed by a trifunctional protein with glycinamide ribonucleotide synthetase (GARS), aminoimidazole ribonucleotide synthetase (AIRS) and glycinamide ribonucleotide formyltransferase (GART) enzymatic activities (Schild et al., 1990).

Figure 3.4

The formation of the human composite protein GARS-AIRS-GART by the fusion of GARS from *Bacillus halodurans*; AIRS from *Agrobacterium tumefaciens* and GART from *Mesorhizobium loti* is illustrated. The homologous regions are represented as the amino acid positions of the fusion protein. The other spliced variant GARS is also presented in humans.

3.2.4 FUSION PROTEINS SIMULATING PROTEIN-PROTEIN INTERFACES IN BACTERIA

Some human fusion proteins simulate protein-protein interfaces in bacteria (Table 3.1). 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* as shown in Figure 3.5 (Haase, 1984). 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 protein 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* (Marcotte et al., 1999). 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 3.1.

3.3 SUMMARY

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 humans compared to their fusion partners (which are physically separated) in bacteria. 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 (Long et al., 2000). 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 humans 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 (Marcotte et al., 1999) or a more general functional association such as between sequential members in a metabolic pathway (Reed, 1974; Perham, 1975). 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 (Salzberg et al., 2001). 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 chapter 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.

CHAPTER 4

THE ASSOCIATION OF HUMAN FUSION PROTEINS IN METABOLIC NETWORKS

4.1 INTRODUCTION

Metabolism is the entire network of chemical reactions carried out by living cells. Metabolism of living cells converts substrates into metabolic energy, redox potential and metabolic end products that are essential to maintain cellular function. One biochemical feature that distinguishes the latter type of cell from its putative early precursors is the presence of metabolic pathways.

Fusion proteins in one species consist of two or more fusion partners from one or more species and they exhibit accreted or novel function compared to fusion partners (Yiting *et al.*, 2004). In recent years, several fusion proteins have been identified across distant phylogenetic distances and their accreted function is comprehensively discussed. Fusion proteins exhibit enhanced functional networks (Yanai *et al.*, 2001), substrate specificity (Katzen *et al.*, 2002), multi-functionality (Berthonneau & Mirande, 2000), simulate protein-protein interfaces (Marcotte *et al.*, 1999) and acquire novel function (Long, 2000). Therefore, the formation of fusion protein is evolutionarily selective and functionally incremental or novel. Hence, it is important to establish the mapping between events of fusion and fission across phylogenetically diverged species. However, this mapping is highly combinatorial, information demanding and computationally intensive.

In recent years, databases have been constructed to capture fusion events across distant phylogenies. These databases contain fusion proteins between human and yeast (Truong & Ikura, 2003); human and prokaryotes (Yiting *et al.*, 2004) and within prokaryotes (Suhre & Claverie, 2004). It has been shown that many human proteins of prokaryotic origin mimic operons and exhibit multiple functions (Yiting *et al.*, 2004). Consequently, we identified 6 human proteins of prokaryotic origin that are associated with metabolic pathways. We further probed into their accreted function by establishing connectivity to their metabolites. Thus far, studies on metabolites have focused on their connectivity in networks and little is known about the origin and evolution of metabolic members that regulate network dynamics (Fornie *et al.*, 2004; Blank & Sauer, 2004). Nonetheless, metabolic pathways are connected through an amazing diversity of compounds with different chemical structures and biological activities for a balanced stoichiometry. The material balance is maintained through the networks by regulating a mosaic of metabolites (substrates and products) at levels of entry and exit in a synchronized manner. Such an inter-connected synchronized design could lead to the simultaneous flow of metabolites in numerous directions with optimal kinetic rates (Segre *et al.*, 2002; Stelling *et al.*, 2002). This is achieved through the tight regulation of proteins in networks (Ihmels *et al.*, 2004). Herein, we discuss the possible rationale for the evolution of complex metabolic networks with the origin and formation of fusion proteins. These results emphasize the utility of design principles for metabolic gene regulation providing a means of facilitating cross-talk between pathways using common substrates.

4.2 MATERIALS AND METHODS

4.2.1 HUMAN FUSION PROTEINS – DEFINITION AND DATASET SOURCE

By definition, a human fusion protein should show evidence of fusion partners in one or more prokaryotes (Yiting et al., 2004). In this case, the N terminal and C terminal domains are made of independent fusion partners from the same or different bacterial species (Figure 4.1). It should be noted that the fusion partners have high sequence homology with their corresponding domains in the fused protein. The 141 human fusion proteins of bacterial origin presented in our Chapter 2 are used in this study (Yiting et al., 2004).

4.2.2 FUSION PROTEINS AS METABOLIC ENZYMES

The 141 fusion proteins as identified in Chapter 2 were visually inspected and manually mapped to metabolic pathways using the KEGG (Kyoto Encyclopedia of Genes and Genomes) pathway database interface (Kanehisa & Goto, 2000). The KEGG database contains updated information on metabolic pathways, regulatory networks and molecular complexes. This information is used to identify the location of fusion proteins in the citric acid cycle of metabolic pathways (Figure 4.2). This exercise enabled us to select six fusion proteins (from the list of 141 fusion proteins) that function as metabolic enzymes in the citric acid cycle (Figure 4.2).

Figure 4.1

Fusion proteins and fusion partners. Human fusion protein [black], partners from same bacterial species [red], partners from different bacterial species [blue, pink], partners with non-homologous N or C terminal domains in a bacterial species [green], a similar fused structure in a bacterial species [purple], slanting bars indicate N terminal domains and vertical bars indicate C terminal domains. Each designated fusion pair is shown within two faint lines.

Figure 4.2

Fusion proteins and metabolites. A diagram showing citric acid cycle in carbohydrate metabolism is shown. The six fusion proteins discussed in this study is indicated using an asterisk (*) in red. The metabolites associated with the fusion proteins are indicated using BOLD font. The associated pathways with the metabolites are highlighted in yellow.

4.2.3 FUSION SCENARIOS LINKED WITH METABOLIC ENZYMES HAVING FUSION STRUCTURES

The fusion proteins consist of two or more fusion partners (Figure 4.1). In each of these fusion proteins [black bars in Figure 4.1], the N terminal domain [slanting bars in Figure 4.1] represents one fusion partner and the C terminal domain [vertical bars in Figure 4.1] represents another fusion partner. In Figure 4.1, each designated fusion pair is illustrated within two faint lines.

Scenario 1 and scenario 2

As shown in Figure 4.1, the fusion pairs may either come from the same bacterial species [red – scenario 1] or from different bacterial species [blue, pink – scenario 2]. The fusion partners shown in red, blue and pink represents a homologous (by measure of sequence similarity) full length ORF (open reading frame) in one or more bacterial species.

Scenario 3

The fusion partners shown in green (refer to Figure 4.1) have matching regions to either N or C terminal domains of the fused protein. In addition, the fusion partners represented in green also contain either preceding (N terminal region for bacterial protein) or following (C terminal region for the bacterial protein) domains or regions that are non-homologous (indicated by a protruding line from the center of a bar) to regions of a fusion protein. Thus, these categories of fusion partners have both homologous and non-homologous domains with reference to a fusion protein.

4.2.4 PRODOM DOMAIN ASSIGNMENTS TO FUSION PROTEINS

Fusion proteins consist of two or more domains for additive or novel role. Therefore, it is important to document the different domains that constitute a fusion protein (Table 4.1). For this purpose, we used the PRODOM database, which is a comprehensive set of protein domain families automatically generated from the SWISS-PROT and TrEMBL (Servant et al., 2002). PRODOM is freely accessible at <http://www.toulouse.inra.fr/prodom.html>. This automatic assignment exercise indicates that each fusion protein consists of two or more PRODOM domains. This suggests fusion proteins consist of several domain-like units as building blocks.

4.2.5 METABOLITES ASSOCIATED WITH FUSION PROTEINS

The metabolic enzymes having fusion structures use metabolites as substrates and produce metabolites as products (Table 4.2). It is our interest to establish the link between fusion proteins and the associated metabolites (Figure 4.2). For this purpose, we used the KEGG (Kyoto Encyclopedia of Genes and Genomes) interface (Kanehisa & Goto, 2000). KEGG is a suite of databases and associated software, integrating our current knowledge on molecular interaction networks in biological processes (PATHWAY database), the information about the universe of genes and proteins (GENES/SSDB/KO databases), and the information about the universe of chemical compounds and reactions (COMPOUND/GLYCAN/REACTION databases).

4.2.6 CONNECTIVITY INDEX FOR METABOLITES OF FUSION PROTEINS

Each metabolite associated with a fusion protein is connected with many other metabolic enzymes in pathways. Here, we hypothesize that the multifaceted connectivity of these metabolites with other members of the pathway is the driving force for domain accretion in fusion proteins. For this purpose, we define connectivity index for a metabolite associated with a fusion protein. By definition, the connectivity index of a metabolite is defined as its ability to connect (number of known links) with other enzymes/proteins in pathways. The connectivity index was calculated from the KEGG ligand database (Table 4.2).

4.3 RESULTS

4.3.1 HUMAN FUSION PROTEINS

Human fusion proteins of bacterial origin have been identified and their role in biological system is implied (Yiting et al., 2004). These proteins exhibit enhanced or novel functions through additive structural architectures. They are shown to mimic operons, simulate protein-protein interfaces, and perform multiple functions and exhibit alternative splicing (Yiting et al., 2004). We further examined these proteins and identified six of them that are involved in the citric acid cycle (Figure 4.2), suggesting that fusion genes are significantly enriched in the TCA cycle (Man Whitney, $p=0.0052$). These proteins consist of two or more PRODOM domains from two or more fusion partners of bacterial origin (Table 4.1). In these fused proteins, the N and C terminal domains represent physically separated fusion

partners in one or more prokaryotes (Figure 4.1). The origin and mechanism of fusion protein is puzzling. In Figure 4.1, each designated fusion pair is illustrated within two faint lines. A detailed analysis of the fusion pairs suggests a number of fusion scenarios (Figure 4.1). These scenarios (see Materials and Methods) were illustrated as scenario 1 (fusion partners from same bacterial species), scenario 2 (fusion partners from different bacterial species) and scenario 3 (fusion partners have both homologous and non-homologous domains with reference to a fusion protein). This illustration indicates that fusion of partners to form a fused structure takes several possible routes as shown by different scenarios. It should be noted that a similar fusion structure could be found in one or more bacterial species [purple bars in Figure 4.1]. This implies that the fusion structures given in this report are not exclusive fused entities for human proteins. However, these human fusion proteins consist of fusion partners from one or more bacterial species. This suggests that fusion events are not only seen between bacterial and human but also within bacterial species. Hence, the origin and accreted role of human fusion proteins is evolutionarily interesting and functionally puzzling. Therefore, an understanding of their structural and functional evolution is critical.

4.3.2 FUSION PROTEINS AS METABOLIC ENZYMES

We selected six metabolic enzymes with fusion structures from a list of 141 fusion proteins reported elsewhere (Yiting et al., 2004). The KEGG pathway database was used to identify the location of fusion proteins in the citric acid cycle (Figure 4.2). The six metabolic enzymes consisting of fusion structures are given in Table 4.1. This association between fusion proteins and members of metabolic networks is interesting.

It is our interest to establish a detailed understanding of their role in metabolic networks in the light of their domain accretion and gene fusion. As given in Table 4.1, these human fusion proteins have N and C terminal domains matching independent subunits in bacterial genomes. This observation implies that metabolic enzymes with fusion structures have incremental role in pathways. Although, an association between fusion proteins and metabolic networks is realized through this observation, it is important to establish the significance of this relation in specific quantitative terms. Additionally, the automatic PRODOM assignment procedure indicates that fusion protein consists of two or more PRODOM domains, suggesting that domain-like units serve as building blocks in their evolution. In conclusion, fusion proteins acquire complex structural architectures through the modular arrangements of building blocks at multiple layers of organization.

4.3.3 METABOLITES OF FUSION PROTEINS

The six metabolic enzymes having fusion structures use or produce metabolites that have multiple trajectories in pathways (Figure 4.2). These metabolites have high connectivity indicating greater involvement in cross talk between networks (Table 4.2). Hence, these metabolites are associated with a mosaic of reactions in networks. Therefore, a stoichiometric pressure is built on these metabolites and the need to establish a material balance is critical. Herein, we relate the high connectivity index of metabolites with their corresponding fusion proteins in the citric acid cycle. The five metabolites produced by these six enzymes are summarized in Table 4.2. Data in Table 4.2 suggests that these five metabolites have multifaceted role by participating in a mosaic of reactions and pathways.

Table 4.2

Fusion proteins as metabolic enzymes and associated metabolites

Metabolites	Fusion proteins as metabolic enzymes	Number of associated metabolic reactions	Number of associated metabolic pathways	Number of enzymes associated with the metabolite		
				Total	Enzyme uses metabolite as a substrate	Enzyme produces metabolite as a product
Oxaloacetate	Pyruvate carboxylase & ATP citrate lyase	41	9	40	16	24
Acetyl co-A	Pyruvate dehydrogenase complex (E3-binding) & ATP citrate lyase	132	28	98	80	18
Succinate	Succinyl co-A transferase	79	11	40	5	35
Succinyl co-A	Methylmalonyl co-A mutase	26	6	18	13	5
Glutamate	Ornithine amino-transferase	121	11	82	25	57

4.3.3.1 OXALOACETATE

The first metabolite, oxaloacetate is produced by pyruvate carboxylase and ATP citrate lyase (Figure 4.2) and it is associated with 41 reactions in 9 pathways (Table 4.2). It acts as substrate at 16 points and as product at 24 points. Clearly, oxaloacetate is under high stoichiometric pressure and data suggests that this metabolite is produced by two fusion proteins with accreted domains. However, the fusion partners are physically separated in one or more prokaryotes (Figures 4.1A and 4.1B). Hence, the mere observation demonstrating fusion/fission events is greatly intriguing. Pyruvate carboxylase consists of 14 PRODOM domains, where the N and C terminal domains are structurally similar to the physically separated subunits A and B of fusion partners in bacteria (Table 4.1). However, a similar fused structure is also seen in the bacterium *A. tumefaciens*. In *Methanosarcina barkeri*, pyruvate carboxylase exhibits an *operon* like structure (Mukhopadhyay et al., 2001). Similarly, ATP citrate lyase is made up of 9 PRODOM domains, where the N and C terminal domains are structurally similar to the physically separated subunits 1 and 2 of bacterial fusion partners. In *Klebsiella pneumoniae*, ATP citrate lyase also exhibits a unique *operon* like structure (Bott & Dimroth, 1994). Thus, the fused structure is analogous to an *operon* like arrangement. Data in Table 4.2 indicates that these two fusion proteins are associated with oxaloacetate which is severely constraint through several inter-connections in metabolism. In such an environment, oxaloacetate is formed from pyruvate by pyruvate carboxylase and from citrate by ATP citrate lyase. Thus, here we show the possible role of fusion proteins in multifaceted network dynamics for a balanced synthesis of oxaloacetate. Therefore, one or more human fusion proteins could meet the demand for metabolites by fusing two or more fusion

partners that are physically separated in one or more bacterial genomes. This is consistent with a previous report which suggested that biotin carboxylase family enzymes have evolved into a complex multifunctional protein from smaller mono-functional precursors through successive gene fusions (Obermayer & Lynen, 1976). Therefore, the hypothesis is that fusion proteins with modular organization acquire selective incremental functions. The interesting aspect in this analysis is the analogous observation between operon like structures in *M. barkeri* and *K. pneumoniae* and fusion structures in humans.

4.3.3.2 ACETYL COA

The second metabolite, acetyl co-A is produced by ATP citrate lyase and pyruvate dehydrogenase (Figure 4.1). This metabolite is associated with 131 reactions in 27 pathways (Table 4.2). It acts as substrate at 85 points and as product at 18 points. Thus, acetyl co-A is associated with many reactions in metabolic networks. However, acetyl co-A is also product of fusion proteins ATP citrate lyase and pyruvate dehydrogenase. This suggests that the origin of fusion proteins has a critical role in evolution of complex networks. Pyruvate dehydrogenase is made up of 8 PRODOM domains with an N terminal subunit A and a C terminal subunit B and ATP citrate lyase is made up of 9 PRODOM domains with an N terminal subunit 1 and a C terminal subunit 2 (Table 4.2). In such a selection environment, acetyl CoA is formed from pyruvate by pyruvate dehydrogenase and from citrate by ATP citrate lyase. Thus, here we show the possible role of fusion proteins in increasing the production of acetyl CoA for material balance in pathways.

4.3.3.3 SUCCINYL CO-A

The third metabolite, succinyl co-A is produced by methyl malonyl co-A mutase (Figure 4.1). It is involved in 26 reactions in 6 pathways. It acts as substrate at 13 points and as product at 6 points. Undoubtedly, succinyl CoA is also under high stoichiometric pressure in the network. It is made up of 9 PRODOM domains where the N and C terminal domains are analogous to subunits α N-terminus and subunits α C-terminus (Table 4.2). These results indicate that the metabolites of fusion proteins are severely constraint in the network. In such a selection environment, succinyl co-A is formed from L-methylmalonyl CoA by methylmalonyl CoA mutase. It has been shown that methylmalonyl CoA mutase simulates protein-protein interactions in *Propionibacterium shermanii* (Marsh et al., 1989). The fusion partners are physically separated in *Propionibacterium shermanii* and they associate through weak interactions without covalent bonding. Thus, fusion imparts stability to protein-protein associations through the formation of a more stable covalently linked domain-domain interfaces.

4.3.3.4 SUCCINATE

The fourth metabolite, succinate is produced by succinyl co-A transferase (Figure 4.1). It is associated with 79 reactions in 11 pathways (Table 4.2). It also shows that succinate acts as substrate at 5 points and as product at 35 points. Clearly, succinate is under high stoichiometric pressure in the network. Succinyl Co-A transferase consist of two fusion partners (Figure 4.2). It is made up of 6 PRODOM domains with an N terminal subunit α and a C terminal subunit β (Table 4.2). In such a selection environment, succinate is formed from succinyl CoA by succinyl CoA transferase. This fusion protein simulates protein-protein

interactions in *Pseudomonas putida* (Parales & Harwood, 1992) and mimics operon like structure in *C. acetobutylicum* (Yeh et al., 1981).

4.3.3.5 GLUTAMATE

The fifth metabolite, glutamate is produced by *ornithine aminotransferase* (Figure 4.1) and is involved in 121 reactions in 11 pathways (Table 4.2). It acts as substrate at 25 points and as product at 57 points. This protein is made up of 8 PRODOM domains with an N terminal subunit A and a C terminal subunit B (Table 4.2). These data show the possible link between fusion proteins and pathways with reference to oxaloacetate, acetyl co-A, succinyl co-A, succinate and glutamate. The hypothesis is that fusion proteins could meet the demand for metabolites by fusing two or more fusion partners that are physically separated in one or more bacteria. This is consistent with a previous report which suggested that biotin carboxylase family of enzymes have evolved into a complex multifunctional protein from smaller mono-functional precursors through successive gene fusions (Obermayer & Lynen, 1976). Therefore, the fusion proteins with modular organization possibly acquire selective incremental functions for optimized role in complex networks.

4.4 DISCUSSION

Protein evolution is extremely efficient in generating systems that are optimally adapted in cellular environment (Freeland et al., 2000; Heinrich & Schuster, 1998; Stephani et al., 1999; Akashi & Gojobori, 2002). Optimality can be achieved by changing the topology of metabolic networks by tuning enzymatic or regulatory materials (Segre et al., 2002). Here, we show that metabolites like oxaloacetate, acetyl co-A, succinyl co-A, succinate and glutamate are products of fusion proteins.

These metabolites have high connectivity index, suggesting their greater degree of involvement within networks. This observation implies the association of fusion proteins with complex metabolic networks. The association between human fusion proteins and metabolites with high connectivity is intriguing (Table 4.2). Detailed analysis of fusion proteins highlights the transition from a 'protein-protein interface' to either a 'domain-domain interface' or an operon structure (a group of genes all controlled by the same regulatory gene) (Yiting et al., 2004). This evolutionary transition is intriguing and it is important to systematically investigate the functional link between fusion partners and fused proteins using thermodynamics calculations. The transition may be thermodynamically favorable as fusion proteins acquire reduced entropy compared to their physically separated fusion partners. Therefore, it is envisaged that fusion proteins confer selective advantage in the evolution of regulating metabolic dynamics. This is specifically advantageous for multi-enzyme complexes as it selects kinetic advantages over unassociated enzyme components by increasing connectivity with metabolites.

Fusion of proteins catalyzing sequential steps in a biochemical pathway are normally found since close association of the two active sites enhances the efficiency of two consecutive reactions. Fusion proteins have selectively advantageous by decreasing the regulational load in the cell for a particular process (Enright et al., 2001). Other advantages like favoring intermediate channeling, for example, two subunits of SCOT are fused so that intrinsic binding energy for the ADP moiety of the Coenzyme can be channeled to overcome energy barriers such as substrate destabiliation (Fierke & Jencks, 1986). What's more, the unstability of enzyme complex is another possible force for fusion. For instance, if an equilibrium exists between the separate domains and an initial assembled complex

($A+B \rightleftharpoons AB$) and the rate of formation of this initial complex is much less than its rate of dissociation, the two domains might likely to be fused during the course of evolution, like SCOT (Rochet et al., 1997). Thus, the fusion protein could expedite channeling of intermediate, for instance, to overcome the substantial energy barrier that exists in intermediate formation, as well as the purported instability in cells.

It is also reported that fusion of components into a single polypeptide ensures stability between physically connected domain structures and active sites for a balanced stoichiometric production of intermediates in complex networks (McCarthy & Hardie, 1984; Rochet et al., 1997). The physical proximity of multiple active centers in the same metabolic pathways alleviates molecular diffusion and reduces side reactions in cellular environment (West, 1997). Our data for the six metabolic enzymes having fusion structures aligns well with these observations. This enables fusion proteins to catalyze sequential steps in a biochemical pathway because association of two active sites enhances the efficiency of two consecutive reactions. Thus, fused protein architecture illustrates an evolutionary strategy accreted for maintaining complex stoichiometric balance. The development of a fused protein system could represent an evolutionary step creating a more efficient pathway that is able to overcome inherent obstacles.

It is often thought that the function of fused genes is simply an addition of function to pre-existing component genes. However, chimerical genes generate proteins with novel function (Brodsky et al., 1997). A recent structural analysis of the Histidine biosynthesis components HisA and HisF indicates that the protein structure after gene fusion was also subject to structural and functional adaptation (Lang et al., 2000). In this case, gene fusion produced a new protein fold with novel function. Thus, most fusion proteins reveal that they have acquired separate

functional domains from each component protein through domain accretion by gene fusion. Physical connection between fused domains increases structural propensity between active centers for the regulation of material balance (Koonin et al., 2002). Similarly, by studying small-molecule metabolism (SMM) in *E. coli*, Teichmann et al. (2001) suggested that even proteins as fundamental to the functioning of a free-living cell, and also as ancient as the central SMM enzymes, are not all simple single-domain enzymes but are the products of extensive domain combinations. Since fusion proteins help in the evolution of complex networks, even a modest addition of domains could significantly increase numerous new interactions. This strategy helps to maintain equilibrium in a dynamic network with huge nodes. Thus, large networks of molecular interactions are regulated by relatively few genes in some organisms (IHGSC, 2001; Lander et al., 2001). That is, the ‘gene number’ (number of genes in a species genome) is negligible to ‘reaction number’ (number of reactions in a species cell) in higher eukaryotes. Our results illustrate an important evolutionary phenomenon that involves the formation of cellular network dynamics with the help of fusion proteins having multi-domain structural architectures with incremental functions.

It has recently been reported that the connectivity of metabolic networks follows approximately a power law, i.e., the frequency, $P(k)$, of metabolites participating in k reactions is given by $P(k) \sim k^{-\gamma}$, where γ is a constant coefficient (Ma & Zeng, 2003; Arita, 2004). This distribution implies that although most metabolites are involved in only few reactions, some metabolites serve as “hubs” and are involved in many reactions. These hubs are more frequent than would be expected, for example, in random networks. Interestingly, many hubs, such as ATP, NADH, glutamate, coenzyme A, and their derivatives, serve as key compounds in

the transfer of specific biochemical groups, such as phosphate groups, redox equivalents, amino groups, and acetyl groups. These reports suggest that the fusion genes are associated with such “hub” metabolites. Complex features of metabolic networks such as the presence of hubs may result from selection of growth rate if essential biochemical mechanisms are considered.

4.5 SUMMARY

The current analysis provides new insights into relation of fusion proteins and network evolution. Thus, evolution of fusion genes has a key role in the selection and design of multifaceted network associated with complex genomes. This might have enabled fusion proteins to accrete incremental biochemical function for a balanced regulation of metabolic networks. Hence, protein fusion confers a selective advantage in evolution. There also seems to be a tendency for the higher levels of metabolic organization to evolve with preference around central metabolites. Around central hub-metabolites, new proteins evolve, often by enzyme recruitment from existing pathways. We believe that an understanding of fusion scenarios and their association with members of a pathway should enable us to appreciate the role played by them to combine or to share metabolites across networks, creating novel pathways with functional diversity. In reality, there are many more fusion structures of varying phylogeny and a majority of them have not been captured as the networks are extremely large and complex. Although, the insights drawn from a detailed study of six fusion proteins is found interesting and valuable, additional evidence is required to establish a comprehensive relationship between additive biochemical function and network dynamics.

CHAPTER 5

INSIGHT TOWARDS GENE FUSION FROM MOLECULAR

DYNAMICS SIMULATION OF IGPS

5.1 INTRODUCTION

Histidine biosynthesis is an ancient and complex metabolic pathway involving either seven enzyme complexes in eukaryotes, or nine in other kingdoms of life (Fani, 1998). IGPS (Imidazole glycerol phosphate synthase) catalyzes the fifth and sixth steps of the nine-step histidine biosynthetic pathway in microbes, fungi, and plants (Figure 5.1). It forms the imidazole ring of the histidine precursor imidazole glycerol phosphate (Klem and Davisson, 1993; Klem et al., 2001; Alifano et al., 1996). IGPS is a glutamine amidotransferase that catalyzes the formation of IGP (Imidazole glycerol phosphate) and AICAR (5-aminoimidazole-4-carboxamide ribonucleotide) from PRFAR (N^1 -[(5'-phosphoribulosyl)formimino]-5-aminoimidazole-4-carboxamide ribonucleotide). Interestingly, IGPS functions at the junction of histidine biosynthesis and *de novo* purine biosynthesis, since AICAR is the entry point to the latter (Figure 5.1). So, IGPS is a key metabolic enzyme, which links amino acid and nucleotide biosynthesis. Therefore, it is important to study the structural and functional evolution of IGPS. In bacteria, IGPS forms a hetero dimer interface with glutaminase (HisH) subunit and cyclase (HisF) subunit (Klem & Davisson, 1993). The HisH subunit is a triad glutamine amidotransferase, and its partner HisF subunit, is a $(\beta/\alpha)_8$ barrel protein that completes a cyclase reaction. The two subunits must first dock before the

Figure 5.1

Schematic representation of histidine biosynthetic pathway. IGPS (Imidazole glycerol phosphate synthase) catalyzes the fifth and sixth step of the histidine biosynthetic pathway in microbes, fungi, and plants. IGPS catalyzes the bifurcation step of the histidine and *de novo* purine biosynthesis pathways.

reaction takes place. However, in fungi and plants, the two subunits are fused into a single polypeptide with the HisH domain at the N terminal and HisF domain at the C terminal (Kuenzler et al., 1993; Chittur et al., 2000).

IGPS has two active sites, the glutamine binding site and the PRFAR binding site. A glutamine binding site is formed by the Cys82, His178 and Glu180 catalytic triad in the HisH subunit and Gln124 in the HisF subunit in *Thermus thermophilus* (TT, Omi et al., 2002). In *Saccharomyces cerevisiae* (SC), the glutamine binding pocket is formed by Cys83, His193, Glu195 and Gln397 (Chaudhuri et al., 2001). These active site residues are completely conserved among yeast and bacteria. It resides at the interface of HisH and HisF subunits in bacteria. However, in yeast it is formed at the interface of HisH and HisF domains. Within the interface active site, one molecule of glutamine binds to the cysteine of its catalytic triad and the nascent ammonia is formed (Klem & Davisson, 1993). Ammonia is then channeled through the $(\beta/\alpha)_8$ barrel tunnel to the PRFAR binding site of hisF, which located in the C-terminal portion of the barrel. The stability of the interface plays an important role in glutamine catalysis. Recent crystal structures and molecular dynamics simulations strongly suggest that HisF, the synthase domain, mediates the insertion of ammonia by channeling ammonia from one remote active site to the other (Amaro et al., 2003). Thus far, the only bifunctional HisF-HisH structure solved is that of SC (Chaudhuri et al., 2003). The HisH domain from SC joined to the HisF domain through a linker region. Although many structural features of the two enzymes are similar, some structural differences are found only in the IGPS from SC. One specific difference is the unique linker region in the fused structure, which may play a role in interdomain interactions. In each protein, interface plays an important role in the catalysis of the

enzyme, as suggested that there are critical subunit or domain interactions that mediate the catalytic properties for glutamine hydrolysis (Klem & Davission, 1993). Thus, gene fusion retains glutaminase active site in yeast. Nevertheless, the significance of fusion is not clear. Therefore, it is our interest to probe into the structural and thermodynamics of the fused and unfused structures using MD (Molecular dynamics) simulation.

5.2 MATERIALS AND METHODS

5.2.1 STARTING STRUCTURES FOR MD SIMULATIONS

Flat files containing the PDB (Protein Data Bank) protein sequences (17,843) in fasta format derived from known protein structure were downloaded from NCBI (National Center for Biotechnology Information at URL: <ftp://ftp.ncbi.nih.gov/blast/db/FASTA/pdbaa>). We applied all-against-all blast to the PDB protein sequences to check the existence of proteins separated in organism but fused in the other organism at the structure level. As a result, IGPS (Imidazole Glycerol Phosphate Synthase) was identified to be a fusion protein in eukaryote (*Saccharomyces cerevisiae*) while split in prokaryote (*Thermotoga maritima*, *Thermus thermophilus* and *Pyrobaculum aerophilum*). Table 5.1 gives all the fused and unfused IGPS protein sequences with known structures in PDB database. These retrieved sequences shared a high degree of sequence similarity spanning over the entire length of protein sequence (Table 5.2), with the HisF (cyclase) being more conserved than the HisH (glutaminase). As shown in Table 5.1, most of the IGPS protein structures are solved with a binding complex (e.g., substrate analogue). In order to compare the intact fused and unfused structures, we chose

the free enzymes (PDB code 1KA9 and 1OX6 respectively) without complex binding as the templates for MD (molecular dynamic) simulation study. Thus, we used the X-ray structure of IGPS from TT at 2.3 Å resolution (Omi et al., 2002) and IGPS from SC at 2.4 Å resolution (Chaudhuri et al., 2003) as the starting coordinates for the MD simulation. Hydrogen atoms were added to these structures using SYBYL.

5.2.2 ENERGY MINIMIZATION

All molecular mechanics operations were carried out using the Tripos force field (Vinter et al., 1997) in SYBYL software (Molecular Modeling Software Package, Version 6.8, Tripos Associates Inc.) running on a Silicon Graphics Workstation. The total potential energy of a system was obtained as the sum of six contributions: bond stretching, angle bending, torsion, van der Waals, electrostatic, and out-of-plane (for aromatic conjugated systems). Minimizations of the potential energy of the system were carried out using the Simplex algorithm and the Powell torsional gradient algorithm as implemented in SYBYL, terminating when a 0.5 Kcal/mol·Å energy gradient shift was obtained. A distance-dependent constant function dielectric constant of 1.0 was used to compute electrostatic effects. We used a non-bonded distance cutoff of 8 Å. The net atomic charges in the residues were calculated by the Gasteiger-Hucker method (Marsili et al., 1980; Gasteiger & Marsili, 1980). After this exercise we obtained two energy minimized structures which were used as our initial structures for MD simulation.

Table 5.1 PDB structures of Imidazole Glycerol Phosphate Synthase (IGPS)

PDB ID	Source	Kingdom	Compound	Fused(Y) or Unfused(N)	Description	Year
1GPW	<i>Thermotoga maritima</i>	Bacteria	HisF, HisH	N	HisH, HisF_D11N variant complex	2001
1K9V			HisH	N		2001
1KXJ			HisH	N		2002
1THF			HisF	N		2001
1VH7			HisF	N		2003
1KA9	<i>Thermus thermophilus</i>	Bacteria	HisF, HisH	N	Native IGPS	2001
1H5Y	<i>Pyrobaculum aerophilum</i>	Archaea	HisF	N		
1JVN	<i>Saccharomyce s cerevisiae</i>	Eukaryote	HisHF	Y	Complexed with the glutamine analogue acivicin	2001
1OX4			HisHF	Y	Complexed with the glutamine analogue 6- diazo-5-oxo- L-norleucine (DON)	2003
1OX5			HisHF	Y	Complexed with the substrate PRFAR and covalently modified by the glutamine analogue acivicin	2003
1OX6			HisHF	Y	free enzyme	2003

Table 5.2 Percent similarity values calculated for the deduced IGPS amino acid sequences with the known crystal structures.

HisH					
	<i>Tm1</i>	<i>Tm2</i>	<i>Tt</i>	<i>Pa</i>	<i>Sc</i>
<i>Tm1</i>	100%				
<i>Tm2</i>	98%	100%			
<i>Tt</i>	54%	43%	100%		
<i>Pa</i>	—	—	—	100%	
<i>Sc</i>	44%	43%	47%	—	100%
HisF					
<i>Tm1</i>	100%				
<i>Tm2</i>	95%	100%			
<i>Tt</i>	72%	69%	100%		
<i>Pa</i>	67%	65%	75%	100%	
<i>Sc</i>	49%	47%	48%	50%	100%

The upper triangle are the percent similarities compared separately for the HisF subunit (in bacteria and archaea) and HisF domain (for yeast). Similarly, the lower triangle are the percent similarities compared separately for the HisH subunit (in bacteria and archaea) and HisH domain (for yeast). Two amino acid sequences for *Tm* were deduced in PDB, namely, *Tm1* and *Tm2*. *Tm* : *Thermotoga maritime*; *Tt*: *Thermus thermophilus*; *Pa*: *Pyrobaculum aerophilum*; *Sc*: *Saccharomyces cerevisiae*

5.2.3 MOLECULAR DYNAMIC SIMULATION

The *in vacuo* system was simulated at constant temperature, constant-volume (NVT) ensemble, also referred to as the canonical ensemble. We constructed two simulated systems for the fused and unfused structures respectively. The two systems were run at a temperature of 300 K using a coupling constant of 100 femtosecond. The initial atom velocities were employed from a Maxwell-Boltzmann distribution with scaling velocities. The non-bonded pair list was updated every 25 femtosecond and a 8 Å distance cutoff was applied. During the simulation, the integration step was set up as 1 femtosecond and molecular snapshots were saved every 1000 steps (1 picosecond). A total of 5000 structures were generated and the simulation properties were derived from analyses of these snapshots.

5.2.4 ANALYSIS OF TRAJECTORIES

In order to detect structural differences between the two simulated systems, we performed a comprehensive analysis of their structures in each trajectory. The flexibilities of the different structures were assessed by several properties computed from the simulations: gap volume, gap index, interface area, RMSD and radius of gyration. The work flow of the MD simulation is given in Figure 5.2.

Figure 5.2

Flowchart of two simulation systems for the fused and unfused structures of imidazole glycerol phosphate synthase (IGPS).

RMSD (Root-mean square deviation)

We calculated RMSD for the C α atoms by comparing the difference between the initial structure and the simulated structures in the trajectory. RMSD is calculated from the equation below.

$$\text{RMSD} = \sqrt{\frac{\sum_{i=1}^N (r_i^{\text{trajectory}} - r_i^{\text{initial}})^2}{N}} \quad 5.1$$

where $r_i^{\text{trajectory}}$ is the position of the i^{th} alpha carbon atom of the simulated trajectory structure, r_i^{initial} the position of the i^{th} alpha carbon atom of the initial structure. N is the number of alpha carbon atoms present in the protein, where $N = 550$ for yeast IGPS and $N = 449$ for bacterial IGPS.

SASA (Solvent accessible surface area)

The solvent accessible surface areas (SASA) between domains or subunits are computed using Naccess (Version 2.1.1) program. The program uses the Lee & Richards (1971) method, whereby a probe of given radius is rolled around the surface of the molecule, and the path traced out by its centre is the accessible surface. Typically, the probe has the same radius as water (1.4 Angstroms) and hence the surface described is often referred to as the solvent accessible surface area. The structures of trajectories were picked up for calculation every 10 ps.

Interface area

The interface area for IGPS from TT is defined as the change in its solvent accessible surface area (ΔASA) buried between HisF and HisH subunits. Similarly,

the interface area for IGPS from SC is defined as the change in its solvent accessible surface area (Δ ASA) buried between HisF and HisH domains. The Δ ASA upon complex formation was calculated using the formula mentioned below

$$\Delta\text{ASA} (\text{\AA}) = \text{ASA of HisF} + \text{ASA of HisH} - \text{ASA of HisHF} \quad 5.2$$

Gap volume between domains/subunits

The gap volume gives a measure of the complementarity of the interacting surfaces. The gap volume between protein subunits or protein domains is calculated using a program SURFNET (Laskowski, 1995, available free from <ftp.biochem.ucl.ac.uk>). The SURFNET program generates surfaces and void regions between surfaces from coordinate data supplied in a PDB file. Each pair of subunit/domain atoms are considered in turn, placing a sphere (maximum radius 5.0 angstroms) halfway between the surfaces of the two atoms, such that its surface touches the surfaces of the atoms in the pair. Checks are made to test if any other atoms intercept this sphere and each time an intercept is detected the size of the sphere is reduced accordingly. At any time the size of the sphere falls below a minimum (minimum radius 1.0 Å) the sphere is discarded. If the sphere remains after all checks its size is recorded. The sizes of all the allowable gap-spheres are then used to calculate the gap volume between the two subunits. The structures of trajectories were picked up for calculation every 10 ps.

Gap index

Gap index evaluates the complementation of the protein interacting surfaces as defined by Jones and Thornton (1996), where Gap index (Å) is defined as the

ration of gap volume ($\text{\AA}^3$) between the protein subunits (or domains) to the interface area ($\text{\AA}^2$). The gap index is calculated by the following equation:

$$\text{Gap Index (\AA)} = \frac{\text{Gap Volume (\AA}^3\text{)}}{\text{Interface ASA (\AA}^2\text{)}} \quad 5.3$$

Radius of gyration (R_g)

The Radius of Gyration (R_g) is a measure of the size of the protein. R_g is a radius giving a one dimensional property proportional to the size of a molecule. Its use in analysis can be seen in, for instance, analysing the results of a protein unfolding simulation where we want to see R_g increasing as the protein unfolds. The compactness of a structure can also be understood quantitatively by monitoring the R_g . The radius of gyration is given in equation 5.4.

$$R_g = \sqrt{\frac{\sum_i^N (m_i * r_i)^2}{\sum_i^N m_i}} \quad 5.4$$

where $r_i = R_i - R_{cm}$. R_i is a vector from an arbitrary point to a particular mass m_i and R_{cm} is the vector position of the center of mass of the protein from the arbitrary point. We calculated the radius of gyration of all atoms for structure of each trajectory.

5.3 RESULTS

The bacterial *Thermus thermophilus* IGPS is a herteromeric protein consisting of HisH and HisF subunits, while in the yeast *S. cerevisiae*, IGPS is a bifunctional enzyme resulting from a HisH-HisF gene fusion (Figure 5.3). We performed molecular dynamics (MD) simulations (5 ns in length) analyses on the bacterial and yeast IGPS in vacuo to investigate the differences in the properties of the monomer and the heterodimer from the perspective of dynamic and structural behaviors. The results from a 5 ns MD simulation of IGPS in TT and SC are presented in Figure 5.4.

5.3.1 STRUCTURAL FEATURES OF THE PEPTIDE BACKBONE

In the course of initial energy minimization, the total energy of the TT molecular dynamics simulation assembly reached a minimum at about -4580 kJ/mole in 267 steps; for the SC protein, the total energy reached a minimum at about -6164 kJ/mole in 297 steps. Figure 5.4 are ribbon illustrations of snapshots of the trajectories at 0 ns, 1 ns, 2 ns, 3 ns, 4ns and 5 ns for each IGPS.

The stability and the equilibrium state of the simulated structures can be evaluated by comparing their geometries with the geometry of the starting structure by means of RMSD of alpha carbons. The energy minimized structure was taken as starting structure for the calculation of RMSD. A plot of the RMSD values as a function of the simulation time is shown in Figure 5.5. The large RMSD change in the initial several picoseconds is the result of heating the system from 0 K to 300 K. In the final part of the simulation the RMSD values fluctuate about an almost constant average value. This indicates that the simulation leads to stable

Figure 5.3

Schematic representation of the eukaryote *S. cerevisiae* and the bacteria *T. thermophilus* protein sequences of the bifunctional enzyme imidazole glycerol phosphate synthase (IGPS). In bacteria, IGPS is formed by the glutaminase subunit and the cyclase subunit. In fungi, the IGPS is fused into a single polypeptide including the glutaminase domain and cyclase domain.

Figure 5.4

The snapshot picture of IGPS at 0, 1, 2, 3, 4 and 5 ns. The molecules are rendered as a ribbon diagram with contrasting colors for the glutaminase (bottom) and cyclase (top) domains.

structures. The average calculated RMSD of C α atoms of the unfused and fused structures of IGPS from their respective initial structures are 3.10 ± 0.07 Å and 3.29 ± 0.08 Å respectively during the last 4 ns. Thus, the fused structure from SC has a slightly higher average RMSD than that of the unfused structure from TT.

5.3.2 STRUCTURAL FEATURES OF THE INTERFACE

The structure features of the interface are monitored by three parameters: interface area, gap volume and gap index. The interface area during the MD simulation between HisF and HisH is shown for SC and TT (Figure 5.6). This property shows that the yeast and bacterial IGPS are stable during most of the simulation. The glutaminase-cyclase interface for bacterial IGPS is approximate 3220 Å² and the interface extends over an area of approximately 4636 Å² for yeast IGPS. Results show that the interface area is larger for the fused IGPS in SC than the unfused IGPS in TT

We also calculated the gap volume between HisF and HisH in the fused SC IGPS and in the unfused TT IGPS (Figure 5.7). However, the gap indexes between HisF and HisH in the SC and TT structures are similar over a 5 ns MD simulation (Figure 5.8). The gap volume of yeast IGPS is larger than that of the bacteria over the course of simulation. As shown in Figure 5.7, the mean gap volume is 1957 Å³ for bacterial IGPS and 2851 Å³ for yeast IGPS. The Gap index (Å) is defined as the ratio of gap volume between the inter- or intra-chain domains (Å³) to the interface area (Å²). Gap index evaluates the glutaminase complementation with the cyclase in yeast and bacterial IGPS. In yeast IGPS, the mean gap index is 0.615 ± 0.05 Å and in bacterial IGPS complexes is 0.608 ± 0.07 Å (Figure 5.8). On average, the ratios of gap volume to interface area are very similar in the two proteins.

5.3.3 STRUCTURAL FEATURES OF THE OVERALL SHAPES

The R_g (Radius of gyration) for the fused IGPS (SC) and the unfused IGPS (TT) over a 5 ns simulation is shown in Figure 5.9. The R_g for the fused structure is significantly larger than that of the unfused structure. The overall shapes of the whole molecule for the two proteins were monitored by the calculation of the corresponding radius of gyration (R_g) for each MD run. The radiuses of gyration were calculated for all atoms and presented as functions of MD simulation time in Figure 5.9. The calculated radiuses of gyration also confirm the stability of the two systems during the simulation. The average radius of gyration is 23.02 ± 0.08 Å and 21.26 ± 0.06 Å for yeast and bacterial IGPS respectively. Thus, the comparison of the radiuses of gyration during the 5ns MD simulations suggests that the fused structure promotes the extended conformations of the IGPS overall shape.

Figure 5.5

Root mean square deviation (RMSD) between HisH and HisF in bacterial and yeast IGPS over 5 ns molecular dynamics simulation. The RMSD in yeast is similar with bacteria.

Figure 5.6

Interface area between HisH and HisF in bacterial and yeast IGPS over 5 ns molecular dynamics simulation. Interface area is calculated as ΔASA (change in ASA) upon subunit or domain interaction. The domain-domain interface area for yeast is larger than bacteria.

Figure 5.7

Gap volume between HisH and HisF in bacterial and yeast IGPS over 5 ns molecular dynamics simulation. The gap volume in yeast is larger than bacteria.

Figure 5.8

Gap index between HisH and HisF in bacterial and yeast IGPS over 5 ns molecular dynamics simulation.

Figure 5.9

Radius of gyration of bacterial and yeast IGPS over 5 ns molecular dynamics simulation.

5.4 DISCUSSION

Gene fusion is an important evolutionary phenomenon towards the formation of proteins with new structural architectures. Comparative sequence analysis between closely and distantly related genome species shows evidence for gene fusion/fission. Therefore, it is of great significance to document the selection force generating such proteins with fused structural architectures. However, there is no documentation of structural evidence supporting the dynamics significance of these fused structures in the evolution of homologous proteins. We identified a pair of IGPS X-ray structures for yeast and bacteria from the PDB. IGPS in both bacteria and yeast consist of HisF and HisH as domains or subunits. Interestingly, an interface is formed between HisH and HisF in both SC and TT. However, the nature of the interface is subunit-subunit in TT and domain-domain in SC. Therefore, it is our interest to probe into the structural and thermodynamics importance of fused IGPS compared to the unfused IGPS.

The results given in Figure 5.4 to Figure 5.9 demonstrate structural consequence and dynamics of fused IGPS in SC compared to the un-fused IGPS in TT. IGPS in SC forms a domain-domain interface between HisH and HisF compared to a subunit-subunit interface in TT. The domain-domain interface area in SC is larger than the subunit interface area in TT over a 5 ns molecular dynamics simulation. The larger interface area in SC facilitates better domain-domain interactions compared to subunit interactions in TT (Figure 5.6). The amount of interface area determines the degree of atomic interaction at the interface. Larger HisH and HisF interface in SC imply better interaction between these two domains. Better interaction between HisF and HisH facilitates greater

stability and kinetics in SC. This is assisted largely by the linker segment connecting HisF and HisH domains in SC

The gap volume between HisF and HisH domains from SC IGPS is larger than that between HisF and HisH subunits in IGPS from TT (Figure 5.7). The increased gap volume in SC IGPS may aid in substrate flow into the active sites formed by HisH and HisF domains. However, this flow of substrate is relatively restricted in TT IGPS in exchange for interface stability formed by subunit interaction. Larger gap volume in SC IGPS is partly helped by the linker between HisH and HisF which provides enhanced flexibility for these two domains.

Interestingly, the increased gap volume in SC IGPS do not affect gap index (ratio of gap volume to interface area) in both SC IGPS and TT IGPS (Figure 5.8). This suggests that increased gap volume is proportional to the increased interface in SC compared to that in TT. The gap index provides an alternative measure of geometric complementarity at the interface (Jones & Thornton, 1996). In descriptive sense, larger gap index have poor geometric fitting. In this approach, the gap index is calculated as the ratio of cavity volume to interface area (Jones & Thornton, 1996). The mean gap indices ($0.6 \text{ \AA}$) obtained from the two simulation systems are significantly less than that in high affinity protein complexes ($2.5 \text{ \AA}$) (Jones & Thornton, 1996). The similar gap indices might be the consequence of functional constraint during evolution in SC IGPS and TT IGPS. The IGPS in SC is structurally advantages in providing larger interface area and greater gap volume, whilst maintaining similar Gap index.

Radius of gyration in proteins is a measure of their size and implies their compactness. The radius of gyration given for IGPS from SC and TT in Figure 5.9 describes the unfolding of the structure during simulation. The flexibility rendered

by the linker between HisF and HisH in the case of SC IGPS is shown by the increased radius of gyration compared to that in TT throughout the simulation period over 5 ns. The difference in the average radius of gyration between SC and TT IGPS is about 1.76 Å. This provides the explanation for the increased stability leading to greater kinetics of IGPS caused by the linker in the fused structure of SC IGPS. The comparison of radius of gyration during the 5ns MD simulations suggests that the fused structure promotes the extended conformations of the IGPS. The increased R_g in SC provides more flexibility for SC IGPS by still maintaining similar gap index. Thus, gene fusion plays an important role in evolution towards the formation of proteins with additive structural architecture for enhanced thermodynamics and kinetics.

The raise and fall in interface area, gap volume and gap index in TT during simulation is unusual. This may be due to the high interface movement between the weakly associated subunits. The proposed hypothesis driving the formation of fused proteins by gene fusion is the structural determinant providing increased stability, dynamics and kinetics facilitated during evolutionary selection. This is evident by the structure and dynamics of IGPS as described using interface area, gap volume and radius of gyration in SC and TT.

5.5 SUMMARY

A number of fusion proteins have been identified by comparative genomic analysis using sequence comparison as described in Chapter 1. This suggests that gene fusion is common in evolutionary phylogeny. However, the force driving gene fusion in organismal evolution is not fully evident due to the lack of structural, dynamic, thermodynamic and kinetic data supporting this phenomenon. Despite the growth in structures at PDB, the number of structural pairs illustrating

fusion/fission in distant phylogeny is limited. Here, we show the importance of fused protein by probing the fused IGPS structure in SC as against the unfused IGPS structure in TT using molecular dynamics simulation. The simulation shows the larger interface area, gap volume and radius of gyration in SC IGPS compared to TT IGPS. Thus, forced IGPS in SC have better structural features than unfused IGPS in TT. This finding has meaningful insight towards gene fusion in establishing optimal dynamics and kinetics.

CHAPTER 6

CONCLUSION

6.1 THE ACHIEVEMENT

The recent availability of dozens of completely sequenced genomes provides a unique opportunity to investigate gene fusion phenomenon at the molecular level. In chapter 2 and 3, we identified, analyzed and tabulated a set of fusion genes in human genome with bacterial origin by comparative genomic study. Our results suggest that gene fusion is a real evolutionary phenomenon and new proteins with novel architectures are evolved by fusion of two or more older parts. Analysis of fusion proteins suggests that these proteins exhibit enhanced or novel functions in humans compared to their fusion partners (which are physically separated) in bacteria. 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. The accredited biological functions for each of these proteins are made available in the form of a database at <http://sege.ntu.edu.sg/wester/fusion/>.

In chapter 4, we identified, investigated and discussed the association of six fusion proteins in metabolic networks. These six proteins, which are associated with the citric acid cycle, are selected from the list of fusion proteins identified above. By employing the method of modular analysis, we defined three possible gene fusion scenarios. We further showed that metabolites like oxaloacetate, acetyl co-A, succinyl co-A, succinate and glutamate are products of fusion proteins. These metabolites have high connectivity index, suggesting their greater degree of

involvement within networks. Therefore, our results indicated that evolution of fusion genes has a key role in the selection and design of multifaceted network associated with complex genomes. This might have enabled fusion proteins to accrete incremental biochemical function for a balanced regulation of metabolic networks. Hence, protein fusion may confer a selective advantage in evolution. There also seems to be a tendency for the higher levels of metabolic organization to evolve with preference around central metabolites. We believe that an understanding of fusion scenarios and their association with members of a pathway should enable us to appreciate the role played by them to combine or to share metabolites across networks, creating novel pathways with functional diversity.

In chapter 5, we showed the importance of fused protein by probing the fused IGPS structure in SC as against the unfused IGPS structure in TT using molecular dynamics simulation. The simulation shows the larger interface area, gap volume and radius of gyration in SC IGPS compared to TT IGPS. Thus, forced IGPS in SC have better structural features than unfused IGPS in TT. This finding has meaningful insight towards gene fusion in establishing optimal dynamics and kinetics.

6.2 THE IMPLICATIONS

It has been shown that gene fusion events across genomes can be used for predicting functional associations of proteins, including physical interactions and complex formation (Enright & Ouzounis, 2001). Except for implications in molecular biology, the phenomenon of gene fusion can be utilized in other fields. In the field of biomedical science, it has been suggested that fusion genes can be employed as putative microbial drug targets in *Helicobacter pylori* (Sakharkar et al., 2006). The author performed a match of the 32 identified cases of gene fusion

to the essential gene. Their results showed that almost all of the identified fusion genes are either essential genes or membrane proteins or species-specific genes, which are proposed categories of putative microbial drug targets. Thus, they suggest that gene fusion may serve as a marker for putative microbial drug target identification. Similarly, the human fusion genes identified in this work can be exploited as a strategy to identify essential genes and further filtration of the identified essential genes for the detection of non-homologous human genes represents a promising means of identifying novel drug targets and facilitates the search for new antibiotics. In addition, fusion genes have potential implications for protein engineering. Considerable efforts have been made to create new hybrid enzymes (Kim et al., 1996; Nixon et al., 1997) or modular enzymes (Mootz et al., 2000; Khosla & Harbury, 2001; Weber & Marahiel, 2001) with multiple functions or higher catalytic efficiency by making fusions of protein modules. However, in some cases, although the individual domains or subunits can be recombined in chimeric enzymes by genetic engineering, the resulting chimeras may lose overall stability and even activity, possibly due to a structural mismatch at the fusion interface. The compatibility of the fused domains or modules will be crucial to any attempt at constructing multienzymes. The fusion genes we identified may offer an attractive framework for designing chimeric enzymes following Nature's lead, because component proteins have co-existed stably in the natural fused proteins possibly due to some functional advantages, such as co-regulation, co-localization, substrate channelling, etc. Understanding Nature's strategies and mechanisms for protein evolution may provide insights into the rational design of proteins with novel biological functions.

6.3 CONTRIBUTION

Firstly, we developed an improved method to predict fusion genes computationally and employed this method to identify a list of human fusion genes of prokaryotic origin. The list of fusion proteins presented in this report is the first of its kind and will provide meaningful insights into protein evolution. Secondly, the significance of fusion proteins in cellular networks and their evolutionary mechanism is discussed in this work for the first of time. Our analysis provides new insights into relation of fusion proteins and network evolution. Thirdly, by studying a specific gene fusion case through dynamic simulation, we documented the structure and dynamic differences between fused and unfused proteins. These structural features for the first time demonstrate the evolutionary advantage in generating proteins with novel structural architecture through gene fusion. Thus this work is of great importance to understand the selection forces driving gene fusion during evolution.

6.4 THE LIMITATIONS

The experimental verification of accreted function using published report in this study is minimal. However, we are not able to evaluate the fusion genes identified experimentally due to lack of appropriate facilities though it is important to verify their accreted functions for the putative fusion genes using experimental data. Similarly, although the insights drawn from a detailed study of six fusion proteins on their relation to complex network is found interesting and valuable, additional evidence is required to establish a comprehensive relationship between additive biochemical function and network dynamics.

6.5 FUTURE IMPROVEMENT

It should be noted that our analysis is restricted to human fusion genes of prokaryotic origin. This work can be further extended to identify and characterize fusion proteins across different phylogenetic distances. This exercise will provide us more information on how complex genomic structures are built during evolution. For example, if the same fusion gene is identified in different species, we can construct a molecular phylogenetic tree and investigate the origin of the gene and test if horizontal gene transfer plays an important role in disseminating this gene. As the increasing numbers of fusion proteins are identified and mapped to the metabolic network, we will have a better understanding of the role played by them.

REFERENCES

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

Aimi J., Qiu H., Williams J., Dixon J. E. (1990) De novo purine nucleotide biosynthesis: cloning of human and avian cDNAs encoding the trifunctional glycineamide ribonucleotide synthetase-aminoimidazole ribonucleotide synthetase-glycineamide ribonucleotide transformylase by functional complementation in *E. coli*. *Nucleic Acids Res.* 18: 6665-6672.

Alifano P., Fani R., Lio P., Lazcano A., Bazzicalupo M., Carlomagno M.S., Bruni C.B. (1996) Histidine biosynthetic pathway and genes: structure, regulation, and evolution. *Microbiol Rev.* 60: 44-69.

Altschul S.F., Madden T.L., Schaffer A.A., Zhang J., Zhang Z., Miller W., Lipman D.J. (1997) Gapped BLAST and PSI-BLAST: a new generation of proteins database search programs. *Nucleic Acids Res* 25: 3389-3402.

Amaro R., Tajkhorshid E., Luthey-Schulten Z. (2003) Developing an energy landscape for the novel function of a (beta/alpha)₈ barrel: ammonia conduction through HisF. *Proc Natl Acad Sci U S A.* 100: 7599-7604.

Akashi H., Gojobori T. (2002) Metabolic efficiency and amino acid composition in the proteomes of *Escherichia coli* and *Bacillus subtilis*. *Proc. Natl. Acad. Sci. USA* 99: 3695-3700.

Aral B., Schlenzig J.S., Liu G., Kamoun P. (1996) 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.

Arita M. (2004) The metabolic world of *Escherichia coli* is not small. *Proc Natl Acad Sci U S A*. 101: 1543–1547.

Beaucamp N., Schurig H., Jaenicke R. (1997) The PGK-TIM fusion protein from *Thermotoga maritima* and its constituent parts are intrinsically stable and fold independently. *Biol Chem*. 378: 679-685.

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

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

Blank L.M., Sauer U. (2004) TCA cycle activity in *Saccharomyces cerevisiae* is a function of the environmentally determined specific growth and glucose uptake rates. *Microbiol*. 150, 1085-1093.

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

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

Brodsky G., Barnes T., Bleskan J., Becker L., Cox M., Patterson D. (1997) The human GARS-AIRS-GART gene encodes two proteins which are differentially expressed

during human brain development and temporally overexpressed in cerebellum of individuals with Down syndrome. *Hum Mol Genet.* 6: 2043-2050.

Cary J.W., Petersen D.J., Papoutsakis E.T., Bennett GN. (1990) 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.

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

Chaudhuri B.N., Lange S.C., Myers R.S., Chittur S.V., Davisson V.J., Smith J.L. (2001) Crystal Structure of Imidazole GlycerolPhosphate Synthase: A Tunnel through a $(\beta/\alpha)_8$ Barrel Joins Two Active Sites. *Structure.* 9: 987-997.

Chaudhuri B.N., Lange S.C., Myers R.S., Davisson V.J., Smith J.L. (2003) Toward understanding the mechanism of the complex cyclization reaction catalyzed by imidazole glycerolphosphate synthase: crystal structures of a ternary complex and the free enzyme. *Biochemistry.* 42: 7003-7012.

Chen K.C., Vannais D.B., Jones C., Patterson D., Davidson J.N. (1989) 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.

Chenna R., Sugawara H., Koike Lopez T.R., Gibson T.J., Higgins D.J., Thompson J.D. (2003) Multiple sequence alignment with the Clustal series of programs. *Nucleic Acids Res.* 31: 3497-3500.

Chittur S.V., Chen Y., Davisson V.J. (2000) Expression and purification of imidazole glycerol phosphate synthase from *Saccharomyces cerevisiae*. *Protein Expr. Purif.* 18: 366–377.

Chittur S.V., Klem T.J., Shafer C.M., Davisson V.J. (2001) Mechanism for acivicin inactivation of triad glutamine amidotransferases. *Biochemistry* 40: 876–887.

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

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

Corpet F., Gouzy J., Kahn D. (1998) The ProDom database of protein domain families. *Nucleic Acids Res.* 26: 323-326.

Courseaux A., Nahon J.L. (2001) Birth of two chimeric genes in the Hominidae lineage. *Science.* 291: 1293-1297.

Demoss J.A. (1965) The conversion of shikimic acid to anthranilic acid by extracts of *neurospora crassa*. *J Biol Chem.* 240:1231-1235.

Dong F., Weitzel S.E., von Hippel P.H. (1996) 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 U S A.* 93: 14456-14461.

Ebbole D.J., Zalkin H. (1987) 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.

Eggen R.I., Geerling A.C., Jetten M.S., de Vos W.M. (1991) Cloning, expression, and sequence analysis of the genes for carbon monoxide dehydrogenase of *Methanotrix soehngenii*, *J. Biol. Chem.* 266: 6883-6887.

Enright A.J., Ouzounis C.A. (2001) Functional associations of proteins in entire genomes by means of exhaustive detection of gene fusions. *Genome Biol.* 2: RESEARCH0034.

Fani R., Lio P., Lazcano A. (1995) Molecular evolution of the histidine biosynthetic pathway. *J Mol Evol.* 41: 760-774.

Fernie A.R., Carrari F., Sweetlove L.J. (2004) Respiratory metabolism: glycolysis, the TCA cycle and mitochondrial electron transport. *Curr Opin Plant Biol.* 7: 254-261.

Fierke C.A., Jencks W.P. (1986) Two functional domains of coenzyme A activate catalysis by coenzyme A transferase. Pantetheine and adenosine 3'-phosphate 5'-diphosphate. *J Biol Chem.* 261:7603-7606.

Freeland S.J., Knight R.D., Landweber L.F., Hurst L.D. (2000) Early Fixation of an Optimal Genetic Code. *Mol. Biol. Evol.* 17: 511-518.

Galperin, M.Y. and Koonin, E.V. (2000) Who's your neighbor? New computational approaches for functional genomics. *Nat. Biotechnol.*: 18, 609–613.

Gasteiger J., Marsili, M. (1980) Iterative partial equalization of orbital electronegativity - a rapid access to atomic charges. *Tetrahedron* 36: 3219-3228

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

Gravely B.R. (2001) Alternative splicing: increasing diversity in the proteomic world. *Trends Genet.* 17: 100–107.

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

Hegyi H., Gerstein M. (1999) The relationship between protein structure and function: a comprehensive survey with application to the yeast genome. *J. Mol. Biol.* 288: 147–164.

Hermann K.M. (1995) The shikimate pathway: early steps in the biosynthesis of aromatic compounds. *Plant Cell* 7: 907–919.

Heinrich R., Schuster S. (1998) The modeling of metabolic systems. Structure, control and optimality. *Biosystems* 47: 61–77.

Henze K., Horner D.S., Suguri S., Moore D.V., Sánchez L.B., Müller M., Embley T.M. (2001) Unique phylogenetic relationship of glucokinase and glucosephosphate isomerase of the amitochondriate eukaryotes *Giardia intestinalis*, *Spironucleus barkhanus* and *Trichomonas vaginalis*. *Gene* 281:123–131.

Hoefler G., Forstner M., McGuinness M.C., Hulla W., Hiden M., Krisper P., Kenner L., Ried T., Lengauer C., Zechner R., et al. (1994) 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.

Hua SJ, Guo T, Gough J & Sun ZR. (2002) Proteins with Class α/β Fold Have High-level Participation in Fusion Events. *J. Mol. Biol.* 320:713–719.

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

Ihmels J., Levy R., Barkai N. (2004) Principles of transcriptional control in the metabolic network of *Saccharomyces cerevisiae*. *Nat. Biotechnol.* 22: 86-92.

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

Jones S., Thornton J.M. (1996) Principles of protein-protein interactions. *Proc Natl Acad Sci U S A.* 93: 13-20.

Kamoun P., Aral B., Saudubray J.M. (1998) A new inherited metabolic disease: delta1-pyrroline 5-carboxylate synthetase deficiency. *Bull Acad Natl Med.* 182: 131-137.

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

Kanehisa M., Goto S. (2000) KEGG: Kyoto encyclopedia of genes and genomes. *Nucleic Acids Res.* 28: 27-30.

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

Khosla C., Harbury P.B. (2001) Modular enzymes. *Nature.* 409 : 247-252.

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

Kim Y.G., Cha J., Chandrasegaran S. (1996) Hybrid restriction enzymes: zinc finger fusions to *Fok* I cleavage domain. *Proc Natl Acad Sci U S A.* 93: 1156–1160.

Klem T.J., Davisson V.J. (1993) Imidazole glycerol phosphate synthase: the glutamine amidotransferase in histidine biosynthesis. *Biochemistry* 32: 5177–5186.

Klem T.J., Chen Y., Davisson V.J. (2001) Subunit Interactions and Glutamine Utilization by *Escherichia coli* Imidazole Glycerol Phosphate Synthase *J. Bacteriol.* 183: 989-996.

Knighton D.R., Kan C.C., Howland E., Janson C.A., Hostomska Z., Welsh K.M., Matthews D.A. (1994) Structure of and kinetic channelling in bifunctional dihydrofolate reductase-thymidylate synthase. *Nat. Struct. Biol.* 1: 186–194.

Koonin E.V., Wolf Y.I., Karev G.P. (2002) The structure of the protein universe and genome evolution. *Nature* 420: 218-223.

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

Kribii R., Arro M., Del Arco A., Gonzalez V., Balcells L., Delourme D., Ferrer A., Karst F., Boronat A. (1997) Cloning and characterization of the *Arabidopsis thaliana* *SQS1* gene encoding squalene synthase – involvement of the C-terminal region of the

enzyme in the channeling of squalene through the sterol pathway. *Eur. J. Biochem.* 249: 61–69.

Kuenzler M., Balmelli T., Egli C.M., Paravicini G., Braus G.H. (1993) Cloning, primary structure, and regulation of the HIS7 gene encoding a bifunctional glutamine amidotransferase:cyclase from *Saccharomyces cerevisiae*. *J. Bacteriol.* 175: 5548–5558.

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

Lander E.S., Linton L.M., Birren B., Nusbaum C., Zody M.C., Baldwin J., Devon K., Dewar K., Doyle M., FitzHugh W., Funke R., Gage D., Harris K., Heaford A., Howland J., Kann L., Lehoczky J., LeVine R., McEwan P., McKernan K., Meldrim J., Mesirov J.P. (2001) Initial sequencing and analysis of the human genome. *Nature* 409: 860-921.

Laskowski R.A. (1995) SURFNET: A program for visualizing molecular surfaces, cavities and intermolecular interactions. *J. Mol. Graph.* 13: 323-330.

Lee B., Richards F.M. (1971) The interpretation of protein structures: estimation of static accessibility. *J Mol Biol.* 55: 379-400.

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

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

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

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

Long M. (2001) Evolution of novel genes. *Curr. Opin. Genet. Dev.* 11: 673-680.

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

Lumbreras V., Alba M.M., Kleinow T., Koncz C., Pages M. (2001) Domain fusion between SNF1-related kinase subunits during plant evolution. *EMBO Rep.* 2: 55-60.

Ma H., Zeng A.P. (2003) Reconstruction of metabolic networks from genome data and analysis of their global structure for various organisms. *Bioinformatics.* 19: 270-277.

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

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

Marsili, M., Gasteiger, J. (1980) Pi charge distribution from molecular topology and pi orbital electronegativity. *Croat. Chem. Acta.* 53: 601-614.

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

McHenry C. S., Crow W. (1979) DNA polymerase III of *Escherichia coli*. Purification and identification of subunits. *J Biol Chem.* 254: 1748-1753.

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

Minet M., Lacroute F. (1990) 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.

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

Mootz H.D., Schwarzer D., Marahiel M.A. (2000) Construction of hybrid peptide synthetases by module and domain fusions. *Proc Natl Acad Sci U S A.* 97: 5848–5853.

Mouillon J.M., Ravanel S., Douce R., Rebeille F. (2002) Folate synthesis in higher-plant mitochondria: coupling between the dihydropterin pyrophosphokinase and the dihydropteroate synthase activities. *Biochem. J.* 363: 313–319.

Mukhopadhyay B., Purwantini E., Kreder C.L., Wolfe R.S. (2001) 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.

Nixon A.E., Warren M.S., Benkovic S.J. (1997) Assembly of an active enzyme by the linkage of two protein modules. *Proc Natl Acad Sci U S A*. 94: 1069–1073.

Obermayer M., Lynen F. (1976) Structure of biotin enzymes. *Trends. Biochem. Sci.* 1: 169–171.

Ogawa T., Shiga K., Hashimoto S., Kobayashi T., Horii A., Furukawa T. (2003) 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.

Omi R., Mizuguchi H., Goto M., Miyahara I., Hayashi H., Kagamiyama H., Hirotsu K. (2002) Structure of Imidazole Glycerol Phosphate Synthase from *Thermus thermophilus* HB8: Open-Closed Conformational Change and Ammonia Tunneling. *J. Biochem.* 132: 759-765.

Orii K.E., Orii K.O., Souri M., Orii T., Kondo N., Hashimoto T., Aoyama T. (1999) 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.

Parales R.E., Harwood C.S. (1992) Characterization of the genes encoding beta-ketoadipate: succinyl-coenzyme A transferase in *Pseudomonas putida*. *J Bacteriol*. 174: 4657-4666.

Perham R.N. (1975) Self-assembly of biological macromolecules. *Phil. Trans. R. Soc. Lond., Ser. B* 272: 123-136.

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

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

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

Rochet J.C., Oikawa K., Hicks L.D., Kay C.M., Bridger W.A., Wolodko W.T. (1997) Productive interactions between the two domains of pig heart CoA transferase during folding and assembly. Biochemistry 36: 8807-8820.

Rudolph J., Stubbe J. (1995) Investigation of the mechanism of phosphoribosylamine transfer from glutamine phosphoribosylpyrophosphate amidotransferase to glycineamide ribonucleotide synthetase. Biochemistry 34: 2241-2250.

Sakharkar K.R., Sakharkar M.K., Chow V.T. (2006) Gene fusion in *Helicobacter pylori*: making the ends meet. Antonie Van Leeuwenhoek. 89:169-80.

Sali A. (1999) Functional links between proteins. Nature 402: 23–26.

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

Schlapfer B.S., Zuber H. (1992) Cloning and sequencing of the genes encoding glyceraldehyde-3-phosphate dehydrogenase, phosphoglycerate kinase and triosephosphate isomerase (gap operon) from mesophilic *Bacillus megaterium*:

comparison with corresponding sequences from thermophilic *Bacillus stearothermophilus*. *Gene*. 122: 53-62.

Schild D., Brake A.J., Kiefer M.C., Young D., Barr P.J. (1990) Cloning of three human multifunctional de novo purine biosynthetic genes by functional complementation of yeast mutations. *Proc Natl Acad Sci U S A*. 87: 2916-2920.

Schurig H., Beaucamp N., Ostendorp R., Jaenicke R., Adler E., Knowles J.R. (1995) PGK and TIM from the hyperthermophilic bacterium *T. maritima* form a covalent bifunctional enzyme complex. *EMBO J*. 14: 442 – 451.

Segre D., Vitkup D., Church G.M. (2002) Analysis of optimality in natural and perturbed metabolic networks. *Proc. Natl. Acad. Sci. USA* 99: 15112-15117.

Servant F., Bru C., Carre S., Courcelle E., Gouzy J., Peyruc D., Kahn D. (2002) ProDom: Automated clustering of homologous domains. *Brief. Bioinform.* 3: 246-251.

Smith D.R., Doucette-Stamm L.A., Deloughery C., Lee H., Dubois J., Aldredge T., Bashirzadeh R., Blakely D., Cook R., Gilbert K. et al. (1997) Complete genome sequence of *Methanobacterium thermoautotrophicum* deltaH: functional analysis and comparative genomics. *J Bacteriol.* 179: 7135-7155.

Stanhope M.J., Lupas A., Italia M.J., Koretke K.K., Volker C., Brown J.R. (2001) Phylogenetic analyses do not support horizontal gene transfers from bacteria to vertebrates. *Nature* 411: 940-944.

Stelling J., Klamt S., Bettenbrock K., Schuster S., Gilles E.D. (2002) Metabolic network structure determines key aspects of functionality and regulation. *Nature* 420: 190-193.

Stephani A., Nuno J.C., Heinrich R. (1999) Optimal Stoichiometric Designs of ATP-producing Systems as Determined by an Evolutionary Algorithm. *J. Theor. Biol.* 199: 45-61.

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

Suguri S., Henze K., Sánchez L.B., Moore D.V., Müller M. (2001) Archaeobacterial relationships of the phosphoenolpyruvate carboxykinase gene reveal mosaicism of *Giardia intestinalis* core metabolism. *J. Eukaryot. Microbiol.* 48: 493-497.

Suhre K., Claverie J.M. (2004) FusionDB: a database for in-depth analysis of prokaryotic gene fusion events. *Nucleic Acids Res.* 32(Database issue): D273-276.

Teichmann S.A., Rison S.C., Thornton J.M., Riley M., Gough J., Chothia C. (2001) Small-molecule metabolism: an enzyme mosaic. *Trends Biotechnol.* 19: 482-486.

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

Thoma R., Obmolova G., Lang D.A., Schwander M., Jenö P., Sterner R., Wilmanns M. (1999) Efficient expression, purification and crystallisation of two hyperthermostable enzymes of histidine biosynthesis. *FEBS Lett.* 454: 1-6.

Thomson T.M., Lozano J.J., Loukili N., Carrio R., Serras F., Cormand B., Valeri M., Diaz V.M., Abril J., Burset M., Merino J., Macaya A., Corominas M., Guigo R. (2000) Fusion of the human gene for the polyubiquitination coeffector UEV1 with Kua, a newly identified gene. *Genome Res.* 10: 1743-1756.

Thompson J.D, Higgins D.G., Gibson T.J. (1994) CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. *Nucleic Acids Res.* 22: 4673–4680.

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

Tsoka S., Ouzounis C.A. (2000) Prediction of protein interactions: metabolic enzymes are frequently involved in gene fusion. *Nat Genet.* 26: 141-142.

Unkles S.E., Logsdon J.M. Jr, Robison K., Kinghorn J.R., Duncan J.M. (1997) The *tigA* gene is a transcriptional fusion of glycolytic genes encoding triose-phosphate isomerase and glyceraldehyde-3-phosphate dehydrogenase in oomycota. *J Bacteriol.* 179: 6816-6823.

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

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

Venter J.C., Adams M.D., Myers E.W., Li P.W., Mural R.J., Sutton G.G., Smith H.O., Yandell M., Evans C.A., Holt R.A. et al. (2001) The sequence of the human genome. *Science* 291: 1304–1351.

Vinter J.G., Davis A., Saunders M.R. (1997) Strategic approaches to drug design. I. An integrated software framework for molecular modeling. *J. Comput. Aided Mol. Des.* 1: 31-51.

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

Weber T., Marahiel M.A. (2001) Exploring the domain structure of modular nonribosomal peptide synthetases. *Structure* 9: 3-9.

West I.C (1997) Channelling in intermediary metabolism, ed. Auius, L.S. (H. S. A., London). 9: 13-69.

Yamaguchi M., Hatefi Y. (1994) 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.

Yanai I., Derti A., DeLisi C. (2001) 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.

Yang S.Y., Schulz H. (1983) 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.

Yeh W.K., Ornston L.N. (1981) 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.

Yiting Y., Chaturvedi I., Meow L.K., Kanguane P., Sakharkar M.K. (2004) Can ends justify the means? Digging deep for human fusion genes of prokaryotic origin. *Front Biosci* 9: 2964-2971.

Youno J., Tanemura S. (1970) Restoration of in-phase translation by an unlinked suppressor of a frameshift mutation in *Salmonella typhimurium*. *Nature*. 225: 422-426.

APPENDIX 1

Table 2.1 List of prokaryotic species used in analysis. G(+) and G(-) indicate gram positive and gram negative respectively.

S/N	Name of Species	kingdom	abbre.	Strain	No.of genes
1	<i>Agrobacterium tumefaciens</i>	Bacteria (G-)	At	C58 (Celeron) C58 (U of Washington)	4556 4661
2	<i>Aeropyrum pernix</i>	Archaea	Ap		2694
3	<i>Aquifex aeolicus</i>	Bacteria	Aa		1553
4	<i>Archaeoglobus fulgidus</i>	Archaea	Af		2407
5	<i>Bacillus halodurans</i>	Bacteria (G+)	Bh		4066
6	<i>Bacillus subtilis</i>	Bacteria (G+)	Bs		4100
7	<i>Borrelia burgdorferi</i>	Bacteria (G-)	Bb		850
8	<i>Brucella melitensis</i>	Bacteria (G-)	Bm		3198
9	<i>Buchnera aphidicola</i>	Bacteria (G-)	Ba	str. Sg	545
10	<i>Buchnera sp.</i>	Bacteria (G-)	Bsp		564
11	<i>Campylobacter jejuni</i>	Bacteria (G-)	Cj		1634
12	<i>Caulobacter crescentus</i>	Bacteria (G-)	Cc	CB15	3737
13	<i>Chlamydia muridarum</i>	Bacteria	Cm		909
14	<i>Chlamydia trachomatis</i>	Bacteria	Ctr		894
15	<i>Chlamydophila pneumoniae</i>	Bacteria	Cpn	AR39 CWL029 J138	1110 1052 1069
16	<i>chlorobium tepidum</i>	Bacteria (G-)	Ct	TLS	2252
17	<i>Clostridium acetobutylicum</i>	Bacteria (G+)	Ca		3672
18	<i>Clostridium perfringens</i>	Bacteria (G+)	Cp		2660
19	<i>Corynebacterium glutamicum</i>	Bacteria (G+)	Cg		3040
20	<i>Deinococcus radiodurans</i>	Bacteria	Dr		2936
21	<i>Escherichia coli</i>	Bacteria (G-)	Ec	K12 O157:H7 O157:H7 EDL933 subsp.	4289 5361 5347
22	<i>Fusobacterium nucleatum</i>	Bacteria (G-)	Fn	nucleatum	2068
23	<i>Haemophilus influenzae</i>	Bacteria (G-)	Hi	Rd	1709
24	<i>Halobacterium sp.</i>	Archaea	Hsp	NRC-1	2058
25	<i>Helicobacter pylori</i>	Bacteria (G-)	Hp	26695 J99	1566 1491
26	<i>Lactococcus lactis subsp. lactis</i>	Bacteria (G+)	Ll		2266
27	<i>Listeria innocua</i>	Bacteria (G+)	Li		2968
28	<i>Listeria monocytogenes</i>	Bacteria (G+)	Lm	EGD-e	2846
29	<i>Mesorhizobium loti</i>	Bacteria (G-)	Ml		6746
30	<i>Methanococcus jannaschii</i>	Archaea	Mj		1770
31	<i>Methanopyrus kandleri</i>	Archaea	Mk	AV19	1687
32	<i>Methanosarcina acetivorans</i>	Archaea	Ma	C2A	4540
33	<i>Methanosarcina mazei Goel</i> <i>Methanothermobacter</i>	Archaea	Mm		3371
34	<i>thermautotrophicus</i>	Archaea	Mth	Delta H	1869

S/N	Name of Species	kingdom	abbre.	Strain	No.of genes
35	<u>Mycobacterium leprae</u>	Bacteria (G+)	Mle		1605
36	<u>Mycobacterium tuberculosis</u>	Bacteria (G+)	Mtu	CDC1551 H37Rv	4187 3918
37	<u>Mycoplasma genitalium</u>	Bacteria (G+)	Mg		484
38	<u>Mycoplasma pneumoniae</u>	Bacteria (G+)	Mpn		688
39	<u>Mycoplasma pulmonis</u>	Bacteria (G+)	Mpu		782
40	<u>Neisseria meningitidis</u>	Bacteria (G-)	Nm	MC58 Z2491	2079 2065
41	<u>Nostoc sp.</u>	Bacteria	Nsp	PCC 7120	5366
42	<u>Pasteurella multocida</u>	Bacteria (G-)	Pm		2015
43	<u>Pseudomonas aeruginosa</u>	Bacteria (G-)	Pa		5567
44	<u>Pyrobaculum aerophilum</u>	Archaea	Pae		2605
45	<u>Pyrococcus abyssi</u>	Archaea	Pab		1769
46	<u>Pyrococcus furiosus</u>	Archaea	Pf	DSM 3638	2065
47	<u>Pyrococcus horikoshii</u>	Archaea	Ph		1801
48	<u>Ralstonia solanacearum</u>	Bacteria	Rs		5116
49	<u>Rickettsia conorii</u>	Bacteria	Rc		1374
50	<u>Rickettsia prowazekii</u>	Bacteria	Rp		835
51	<u>Salmonella enterica</u>	Bacteria (G-)	Se	enterica serovar Typhi	4395
52	<u>Salmonella typhi</u>	Bacteria (G-)	Sty		4451
53	<u>Sinorhizobium meliloti</u>	Bacteria (G-)	Sm		3341
54	<u>Staphylococcus aureus</u>	Bacteria (G+)	Sa	subsp. aureus Mu50	2714
55	<u>Streptococcus pneumoniae</u>	Bacteria (G+)	Spn	subsp. aureus MW2	2632
56	<u>Streptococcus pyogenes</u>	Bacteria (G+)	Spy	R6	2043
57	<u>Streptomyces coelicolor</u>	Bacteria (G+)	Sc	TIGR4	2094
58	<u>Sulfolobus solfataricus</u>	Archaea	Ss	M1 GAS	1697
59	<u>Sulfolobus tokodaii</u>	Archaea	St	MGAS315	1865
60	<u>Synechocystis sp.</u>	Bacteria (G-)	Ssp	A3(2)	7512
61	<u>Thermoanaerobacter tengcongensis</u>	Bacteria (G-)	Tt		2977
62	<u>Thermoplasma acidophilum</u>	Archaea	Ta	PCC 6803	2826
63	<u>Thermoplasma volcanium</u>	Archaea	Tv		3167
64	<u>Thermotoga maritima</u>	Bacteria	Tm		2588
65	<u>Treponema pallidum</u>	Bacteria (G-)	Tp		1478
66	<u>Ureaplasma urealyticum</u>	Bacteria (G+)	Uu		1526
67	<u>Vibrio cholerae</u>	Bacteria (G-)	Vc		1846
68	<u>Xanthomonas axonopodis</u>	Bacteria (G-)	Xa		1031
69	<u>Xanthomonas campestris</u>	Bacteria (G-)	Xc	pv. citri str. 306	611
70	<u>Xylella fastidiosa</u>	Bacteria (G-)	Xf	pv. campestris str. ATCC	4312
71	<u>Yersinia pestis</u>	Bacteria (G-)	Yp	33913	4181
				9a5c	2766
					3885

APPENDIX 2

[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, Pandjassaram 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 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%

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

Table 1. Human fusion proteins of prokaryotic origin

SN	Human Fusion Proteins					Fusion Partners of Prokaryotic Origin								Comments
	RefSeq Accession	Chromosome Number	Protein Length	Protein Name	Pathway	N terminal				C terminal				
						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/ 6-phospho gluconolactonase	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 <i>P. putida</i> (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 <i>E. coli</i>)	411	440..750	3-Hydroxyacyl-CoA dehydrogenase	Lipid Metabolism; Amino Acid Metabolism; Carbohydrate Metabolism (in <i>E. coli</i>)	fadAB operon in <i>E. coli</i> (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 (<i>S. typhimurium</i>)	1076	394..1440	Carbamoyl-phosphate synthase, large subunit	Amino Acid Metabolism; Nucleotide metabolism (in <i>S. typhimurium</i>)	carAB operon in <i>S. typhimurium</i> (32) Trifunctional protein (33)
						423	1446..1790	Dihydro-orotase (pyrC)	Amino Acid Metabolism; Nucleotide metabolism (in <i>S. acidocaldarius</i>)	305	1921..2223	Aspartate carbamoyl transferase (pyrB1)	Amino Acid Metabolism; Nucleotide metabolism (in <i>S. acidocaldarius</i>)	pyr operon in <i>S. acidocaldarius</i> (34)
4	NP_001150	2	1338	Aldehyde oxidase I	Amino Acid Metabolism; Metabolism of Cofactors and Vitamins	489	9..531	Carbon-monoxide dehydrogenase small subunit	Energy Metabolism (in <i>M. luti</i>)	799	580..1316	Aarbon monoxide dehydrogenase , large subunit	Energy Metabolism (in <i>M. luti</i>)	cdh operon in <i>M. neoaurum</i> (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 <i>E. coli</i>)	411	302..705	3-Hydroxyacy CoA dehydrogenase	Lipid Metabolism; Amino Acid Metabolism; Carbohydrate Metabolism (in <i>E. coli</i>)	fadAB operon in <i>E. coli</i> (30) Bifunctional protein (36)
6	NP_001059	3	1621	DNA topoisomerase II, β		773	56..727	DNA gyrase, Subunit (gyrB)		490	728..1054	DNA Gyrase Subunit A		Operon in <i>M. tuberculosis</i> (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 <i>M. thermocaldophilus</i> (39)
8	NP_005434	4	624	3'-Phospho-adenosine 5'-phosphosulfate synthetase	Nucleotide Metabolism	186	51..221	Adenosine 5'-phosphosulfate kinase	Nucleotide Metabolism; Metabolism of Other Amino Acids; Energy Metabolism (in <i>E. coli</i>)	459	273..619	ATP sulfurylase	Nucleotide Metabolism; Metabolism of Other Amino Acids; Energy Metabolism (in <i>E. coli</i>)	cys operon in <i>E. coli</i> (40) Bifunctional enzyme (41)
9	NP_006443	4	425	AIR carboxylase; SAICAR synthetase	Amino Acid Metabolism	233	12..247	Phosphoribosylaminoimidazole succinocarboxamide (SAICAR) synthetase (nucC)	Nucleotide Metabolism (in <i>B. subtilis</i>)	180	267..421	Phosphoribosylaminoimidazole(AIR) carboxylase (purE)	Nucleotide Metabolism (in <i>B. subtilis</i>)	pur Operon in <i>B. subtilis</i> (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 <i>E. coli</i>)	219	302..516	Acyl CoA:acetate CoA-transferase subunit β	Carbohydrate Metabolism (in <i>E. coli</i>)	Operon in <i>C. acetobutylicum</i> (43) Protein-protein interaction (44)
11	NP_036475	5	1086	Nicotinamide nucleotide transhydrogenase		530	48..587	NADP transhydrogenase α subunit	Metabolism of Cofactors and Vitamins	480	618..1081	NADP transhydrogenase β subunit	Metabolism of Cofactors and Vitamins	nnt operon in <i>R. rubrum</i> (45)
12	NP_005933	6	636	Molybdenum cofactor biosynthesis-step 1 protein isoforms		333	61..366	Molybdenum cofactor biosynthesis protein A		156	481..629	Molybdenum cofactor biosynthesis protein CB		moa operon in <i>E. coli</i> (46)
13	NP_009032	9	918	Sarcosine dehydrogenase	Genetic Information Processing	382	63..445	Sarcosine oxidase subunit β	Amino Acid Metabolism	379	483..894	Sarcosine oxidase, subunit a	Amino Acid Metabolism	sox operon in <i>Corynebacterium</i> 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 <i>M. thermocaldophilus</i> (39)
15	NP_002851	10	795	Pyrroline-5-carboxylate synthetase	Amino Acid Metabolism; Carbohydrate Metabolism	356	72..381	Glutamyl 5-kinase	Amino Acid Metabolism	484	363 .. 770	?-Glutamyl phosphate reductase	Amino Acid Metabolism	ProBA operon in <i>T. thermophilus</i> (17) Bifunctional enzyme (16)

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

16	NP_000911	11	1178	Pyruvate carboxylase precursor	Carbohydrate Metabolism	477	38..483	Pyruvate carboxylase, subunit A	Amino Acid Metabolism; Carbohydrate Metabolism (in <i>S. mutans</i>)	567	563..1178	Pyruvate carboxylase, subunit B	Amino Acid Metabolism; Carbohydrate Metabolism (in <i>T. tengcongensis</i>)	Operon-like arrangement in <i>M. barkeri</i> (48)
17	NP_001087	17	1105	ATP citrate lyase	Genetic Information Processing	398	41..418	Citrate lyase, subunit 1	Carbohydrate Metabolism; Energy Metabolism (in <i>E. coli</i>)	610	496..1089	Citrate lyase, subunit 2	Carbohydrate Metabolism; Energy Metabolism (in <i>E. coli</i>)	Probable operon in <i>K. pneumoniae</i> (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 <i>M. thermoadaptophilum</i> (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) Trifunctional protein (25)
Fusion proteins exhibiting alternative splicing														
26	NP_037361	12	1237	Apoptotic protease activating factor 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	Trifunctional 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 β	Lipid Metabolism	144	611..749	Methyl malonyl Co A mutase, subunit α		Protein-protein interaction (58)
29	NP_001084	12	2483	Acetyl coenzyme A carboxylase β	Metabolism of Complex Carbohydrates	677	256..960	Propionyl-CoA carboxylase, subunit α	Metabolism of Complex Carbohydrates	516	1885..2267	Propionyl-CoA carboxylase subunit β	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

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

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

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

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-pyrroline-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

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

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. Darnedde, 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-phosphogluconolactonase. 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. Neuhaard: 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

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

- genes for carbon monoxide dehydrogenase of *Methanotrix 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-ketoacid 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-methylenetetrahydrofolate 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*

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

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

APPENDIX 3

[Frontiers in Bioscience 10, 1070-1078, May 1, 2005]

INSIGHTS TO METABOLIC NETWORK EVOLUTION BY FUSION PROTEINS

Meena Kishore Sakharkar¹, Yu Yiting¹, Vincent T. K. Chow², Pandjassaram Kanguane¹¹ NCSV, MPE, Nanyang Technological University, Singapore 639798, ² Department of Microbiology, NUS, Singapore 119260

TABLE OF CONTENTS

1. Abstract
2. Introduction
3. Materials and methods
 - 3.1. Human fusion proteins – definition and dataset source
 - 3.2. Fusion proteins as metabolic enzymes
 - 3.3. Fusion scenarios linked with metabolic enzymes having fusion structures
 - 3.3.1. Scenario 1 and Scenario 2
 - 3.3.2. Scenario 3
 - 3.4. PRODOM domain assignments to fusion proteins
 - 3.5. Metabolites of fusion proteins
 - 3.6. Connectivity index for metabolites of fusion proteins
4. Results
 - 4.1. Human fusion proteins
 - 4.2. Fusion proteins as metabolic enzymes
 - 4.3. Metabolites of fusion proteins
5. Discussion
6. Conclusion
7. Acknowledgements
8. References

1. ABSTRACT

Human fusion proteins consisting of two or more fusion partners of prokaryotic origin exhibit accreted function. Recent studies have elucidated the importance of fusion proteins in complex regulatory networks. The significance of fusion proteins in cellular networks and their evolutionary mechanism is largely unknown. Here, we discuss the association of six fusion proteins with the citric acid cycle. We define possible gene fusion scenarios and show that they produce metabolites with high connectivity for complex networking. Complex networking of metabolites requires proteins with incremental structural architectures and functional capabilities. Such higher order functionality is frequently provided by fusion proteins. Therefore, evolution of fusion proteins capable of producing metabolites with greater connectivity for enhanced cross-talk between pathways is critical for the selection of multiple trajectories in maintaining a stoichiometric balance during regulation. The association of six fusion proteins with the citric acid cycle and their capability to produce metabolites with high connectivity index is intriguing. This suggests that fusion gene products and their evolution have had a key role in the selection of complex multifaceted networks. In addition, we propose that fusion proteins have gained additive biochemical function for a balanced regulation of metabolic networks.

2. INTRODUCTION

Fusion proteins in one species consist of two or more fusion partners from one or more other species and they exhibit accreted function compared to fusion partners (1). In recent years, several fusion proteins have been

identified across distant phylogenetic distances and their accreted function is comprehensively discussed. They exhibit enhanced functional networks (2), substrate specificity (3), multi-functionality (4), simulate protein-protein interfaces (5) and acquire novel function (6). Therefore, the formation of fusion protein is evolutionarily selective and functionally critical. Hence, it is important to establish the mapping between events of fusion and fission across different species. However, this mapping is highly combinatorial, information demanding and computationally intensive.

In recent years, databases have been constructed to capture fusion events across distant phylogenies. These databases contain fusion proteins between human and yeast (7), human and prokaryotes (1) and within prokaryotes (8). It has been shown that many human proteins of prokaryotic origin mimic *operons* (a group of genes controlled by the same regulatory gene) and exhibit multiple functions (1). Subsequently, we identified 6 human proteins of prokaryotic origin that are associated with metabolic pathways. We further probed their accreted function by establishing connectivity to metabolites. Thus far, studies on metabolism have focused on their connectivity in networks and little is known about the origin and evolution of metabolic members that regulate network dynamics (9-10). Nonetheless, metabolic pathways are connected through an amazing diversity of compounds with different chemical structures and biological activities for a balanced stoichiometry. Therefore, the material balance is maintained through the networks by regulating a mosaic of metabolites (substrates and products) at levels of entry and

Network evolution and fusion proteins

Figure 1. Fusion proteins and fusion partners. Human fusion protein [black], partners from same bacterial species [red], partners from different bacterial species [blue, pink], partners with non-homologous N or C terminal domains in a bacterial species [green], a similar fused structure in a bacterial species [purple], slanting bars indicate N terminal domains and vertical bars indicate C terminal domains. Each designated fusion pair is shown within two faint lines.

exit in a synchronized manner. Such an inter-connected synchronized design could lead to the simultaneous flow of metabolites in numerous directions with optimal kinetic rates (11-12). This is achieved through the tight regulation of proteins in networks through selective protein architectures (13). Evolution of protein architectures by gene fusion is proposed as a key mechanism for network development. Herein, we discuss the possible rationale for the evolution of complex metabolic networks with the origin and formation of fusion proteins.

3. MATERIALS AND METHODS

3.1 Human fusion proteins – definition and dataset source

By definition, a human fusion protein should show evidence of fusion partners in one or more prokaryotes (1). In this case, the N terminal and C terminal

domains are made of independent fusion partners from the same or different bacterial species (Figure 1). It should be noted that the fusion partners have high sequence homology with their corresponding domains in the fused protein. The 141 human fusion proteins of prokaryotic origin presented in our previous report are used in this study (1).

3.2 Fusion proteins as metabolic enzymes

The 141 fusion proteins (1) were visually inspected and manually mapped to metabolic pathways using the KEGG (Kyoto Encyclopedia of Genes and Genomes) pathway database interface (15). The KEGG database contains updated information on metabolic pathways, regulatory networks and molecular complexes. This information is used to identify the location of fusion proteins in the citric acid cycle of metabolic pathways (Figure 2). This exercise enabled us to select six fusion

Network evolution and fusion proteins

Figure 2. Fusion proteins and metabolites. A diagram showing citric acid cycle in carbohydrate metabolism. The six fusion proteins discussed in this study are indicated using an asterisk (*) in red. The metabolites associated with the fusion proteins are indicated using BOLD font. The associated pathways with the metabolites are highlighted in yellow.

[Frontiers in Bioscience 10, 1070-1078, May 1, 2005]

proteins (from the list of 141 fusion proteins) that function as metabolic enzymes in the citric acid cycle (Figure 2).

3.3. Fusion scenarios linked with metabolic enzymes having fusion structures

The fusion proteins consist of two or more fusion partners (Figure 1). In each of these fusion proteins [black bars in Figure 1], the N terminal domain [slanting bars in Figure 1] represents one fusion partner and the C terminal domain [vertical bars in Figure 1] represents another fusion partner. In Figure 1, each designated fusion pair is illustrated within two faint lines.

3.3.1. Scenario 1 and scenario 2

As shown in Figure 1, the fusion pairs may either come from the same bacterial species [red – scenario 1] or from different bacterial species [blue, pink – scenario 2]. The fusion partners shown in red, blue and pink represents a highly homologous (by measure of sequence similarity) full length ORF (open reading frame) in one or more bacterial species.

3.3.2. Scenario 3

The fusion partners shown in green (refer to Figure 1) have matching regions to either N or C terminal domains of the fused protein. In addition, the fusion partners represented in green also contain either preceding (N terminal region for bacterial protein) or following (C terminal region for the bacterial protein) domains or regions that are non-homologous (indicated by a protruding line from the center of a bar) to regions of a fusion protein. Thus, these categories of fusion partners have both homologous and non-homologous domains with reference to a fusion protein.

3.4 PRODOM domain assignments to fusion proteins

Fusion proteins consist of two or more domains for additive or novel role. Therefore, it is important to document the different domains that constitute a fusion protein (Table 1). For this purpose, we used the PRODOM database, which is a comprehensive set of protein domain families automatically generated from the SWISS-PROT and TrEMBL (14). This automatic assignment exercise indicates that each fusion protein consists of two or more PRODOM domains. This suggests fusion proteins consist of several domain-like units as building blocks.

3.5. Metabolites associated with fusion proteins

The metabolic enzymes having fusion structures use metabolites as substrates and produce metabolites as products (Table 2). It is our interest to establish the link between fusion proteins and the associated metabolites (Figure 2). For this purpose, we used the KEGG interface (15).

3.6. Connectivity index for metabolites of fusion proteins

Each metabolite associated with a fusion protein is connected with many other metabolic enzymes in pathways. Here, we hypothesize that the multifaceted connectivity of these metabolites with other members of the pathway is the driving force for domain accretion in fusion

proteins. For this purpose, we define connectivity index for a metabolite associated with a fusion protein. By definition, the connectivity index of a metabolite is defined as its ability to connect (number of known links) with other enzymes/proteins in pathways. The connectivity index was calculated from the KEGG ligand database (Table 2).

4. RESULTS

4.1. Human fusion proteins

Human fusion proteins of prokaryotic origin have been identified and their role in biological system is implied (1). These proteins exhibit enhanced or novel functions through additive structural architectures. They are shown to mimic operons, simulate protein-protein interfaces, perform multiple functions and exhibit alternative splicing (1). We further examined these proteins and identified six of them that are involved in the citric acid cycle (Figure 2). These proteins consist of two or more PRODOM domains from two or more fusion partners of bacterial origin (Table 1). In these fused proteins, the N and C terminal domains represent physically separated fusion partners in one or more prokaryotes (Figure 1). The origin and mechanism of fusion protein is puzzling. In Figure 1, each designated fusion pair is illustrated within two faint lines. A detailed analysis of the fusion pairs suggests a number of fusion scenarios (Figure 1). These scenarios (see Materials and Methods) were illustrated as scenario 1 (fusion partners from same bacterial species), scenario 2 (fusion partners from different bacterial species) and scenario 3 (fusion partners have both homologous and non-homologous domains with reference to a fusion protein). This illustration indicates that fusion of partners to form a fused structure takes several possible routes as shown by different scenarios. It should be noted that a similar fusion structure could be found in one or more bacterial species [purple bars in Figure 1]. This implies that the fusion structures given in this report are not exclusive fused entities for human proteins. However, these human fusion proteins consist of fusion partners from one or more bacterial species. This suggests that fusion events are not only seen between bacterial and human but also within bacterial species. Hence, the origin and accreted role of human fusion proteins is evolutionarily interesting and functionally puzzling. Therefore, an understanding of their structural and functional evolution is critical.

4.2. Fusion proteins as metabolic enzymes

We selected six metabolic enzymes with fusion structures from a list of 141 fusion proteins reported elsewhere (1). The KEGG pathway database was used to identify the location of fusion proteins in the citric acid cycle (Figure 2). The six metabolic enzymes consisting of fusion structures are given in Table 1. This association between fusion proteins and members of metabolic networks is interesting. It is our interest to establish a detailed understanding of their role in metabolic networks in the light of their domain accretion and gene fusion. As given in Table 1, these human fusion proteins have N and C terminal domains matching independent subunits in bacterial genomes. This observation implies that metabolic enzymes with fusion structures have incremental role in

Network evolution and fusion proteins

Table 1. Human fusion proteins in carbohydrate metabolism

Human fusion (composite) proteins					Fusion partners (component) of prokaryotic origin									
RefSeq Accession	PL	Protein name	Prodom domain		N terminal					C terminal				
			#	ID	PL	MR	Protein	#	Prodom domains	PL	MR	Protein	#	Prodom domains
1	NP_000911	1178	Pyruvate carboxylase precursor	14	477	38..483	Pyruvate carboxylase, subunit A	5	PD000820 PD000180 PD489184 PD000755 PD206296	567	563..1178	Pyruvate carboxylase, subunit B	8	PD002904 PD429392 PD327958 PD414942 PD328972 PD414942 PD440923 PD000268
2	NP_001087	1105	ATP citrate lyase	9	398	41..418	Citrate lyase, subunit 1	2	PD003397 PD010086	610	496..1089	Citrate lyase, subunit 2	4	PD337294 PD014373 PD002034 PD013870
3	NP_003468	501	Pyruvate dehydrogenase	8	473	57..136	Biotin/lipoyl attachment	2	PD055009 PD000268	177	324..500	FAD-dependent pyridine nucleotide-disulphide oxidoreductase	3	PD510157 PD001115 PD564191
4	NP_000427	520	Succinyl CoA transferase	6	250	36..284	B-etoadipate: succinyl CoA transferase, subunit α	5	PD515695 PD477798 PD583163 PD588193 PD003187	219	302..516	B-ketoadipate: succinyl-CoA transferase subunit β	1	PD004976
5	NP_000246	750	Methylmalonyl Co-A mutase	9	681	42..519	Methylmalonyl Co A mutase, subunit α, N-terminus	7	PD188365 PD191238 PD004447 PD004310 PD005702 PD537237 PD592542	144	611..749	Methylmalonyl Co A mutase, subunit α, C-terminus	2	PD002527 PD514726
6	NP_000265	439	Ornithine amino-transferase	8	426	58..330	Ornithine-oxo-acid aminotransferase	4	PD528736 PD082173 PD000465 PD455654	181	261..425	Probable ornithine aminotransferase (C-terminal part)	4	PD000465 PD066084 PD236513 PD001337

PL = protein length; MR = matching region in the fusion protein; # = number of PRODOM domains; ID = PRODOM domain identifier

pathways. Although, an association between fusion proteins and metabolic networks is realized through this observation, it is important to establish the significance of this relation in specific quantitative terms. Additionally, the automatic PRODOM assignment procedure indicates that fusion protein consists of two or more PRODOM domains, suggesting that domain-like units serve as building blocks in their evolution. In conclusion, fusion proteins acquire complex structural architectures through the modular arrangements of building blocks at multiple layers of organization.

4.3. Metabolites of fusion proteins

The six metabolic enzymes having fusion structures use or produce metabolites that have multiple

trajectories in pathways (Figure 2). These metabolites have high connectivity indicating greater involvement in cross talk between networks (Table 2). Hence, these metabolites are associated with a mosaic of reactions in networks. Therefore, a stoichiometric pressure is built on these metabolites and the need to establish a material balance is critical. Herein, we relate the high connectivity index of metabolites with their corresponding fusion proteins in the citric acid cycle. The five metabolites produced by these six enzymes are summarized in Table 2. Data in Table 2 suggests that these five metabolites have multifaceted role by participating in a mosaic of reactions and pathways.

The first metabolite, oxaloacetate is produced by pyruvate carboxylase and ATP citrate lyase (Figure 2) and

[Frontiers in Bioscience 10, 1070-1078, May 1, 2005]

Table 2. Fusion proteins as metabolic enzymes and associated metabolites

Metabolites	Fusion proteins as metabolic enzymes	Number of associated metabolic reactions	Number of associated metabolic pathways	Number of enzymes associated with the metabolite		
				Total	Enzyme uses metabolite as a substrate	Enzyme produces metabolite as a product
Oxaloacetate	Pyruvate carboxylase & ATP citrate lyase	41	9	40	16	24
Acetyl co-A	Pyruvate dehydrogenase complex (E3-binding) & ATP citrate lyase	132	28	98	80	18
Succinate	Succinyl co-A transferase	79	11	40	5	35
Succinyl co-A	Methylmalonyl co-A mutase	26	6	18	13	5
Glutamate	Ornithine amino-transferase	121	11	82	25	57

it is associated with 41 reactions in 9 pathways (Table 2). It acts as substrate at 16 points and as product at 24 points. Clearly, oxaloacetate is under high stoichiometric pressure and data suggests that this metabolite is produced by two fusion proteins with accreted domains. However, the fusion partners are physically separated in one or more prokaryotes (Figures 1A and 1B). Hence, the mere observation demonstrating fusion/fission events is greatly intriguing. Pyruvate carboxylase consists of 14 PRODOM domains, where the N and C terminal domains are structurally similar to the physically separated subunits A and B of fusion partners in bacteria (Table 1). However, a similar fused structure is also seen in the bacterium *A. tumefaciens*. In *Methanosarcina barkeri*, pyruvate carboxylase exhibits an *operon* like structure (16). Similarly, ATP citrate lyase is made up of 9 PRODOM domains, where the N and C terminal domains are structurally similar to the physically separated subunits 1 and 2 of bacterial fusion partners. In *Klebsiella pneumoniae*, ATP citrate lyase also exhibits a unique *operon* like structure (17). Thus, the fused structure is analogous to an *operon* like arrangement. Data in Table 2 indicates that these two fusion proteins are associated with oxaloacetate which is severely constraint through several inter-connections in metabolism. In such an environment, oxaloacetate is formed from pyruvate by pyruvate carboxylase and from citrate by ATP citrate lyase. The second metabolite, acetyl co-A is produced by ATP citrate lyase and pyruvate dehydrogenase (Figure 1). This metabolite is associated with 132 reactions in 28 pathways (Table 2). It acts as substrate at 80 points and as product at 18 points. Thus, acetyl co-A is associated with many reactions in metabolic networks. However, acetyl co-A is also product of fusion proteins ATP citrate lyase and pyruvate dehydrogenase. Pyruvate dehydrogenase is made up of 8 PRODOM domains with an N terminal subunit A and a C terminal subunit B and ATP citrate lyase is made up of 9 PRODOM domains with an N terminal subunit 1 and a C terminal subunit 2 (Table 2).

The third metabolite, succinyl co-A is produced by methyl malonyl co-A mutase (Figure 1). It is involved in 26 reactions in 6 pathways. It acts as substrate at 13 points and as product at 5 points. The enzyme methyl malonyl co-A mutase is made up of 9 PRODOM domains where the N and C terminal domains are analogous to subunits A and B (Table 2). These results indicate that metabolites of fusion proteins are severely constraint in the network. It has been shown that methylmalonyl co-A mutase simulates protein-protein interactions in *Propionibacterium shermanii* (19). The fusion partners are physically separated in

Propionibacterium shermanii and they associate through weak interactions without covalent bonding. Thus, fusion imparts stability to protein-protein associations through the formation of a more stable covalently linked domain-domain interfaces. The fourth metabolite, succinate is produced by succinyl co-A transferase (Figure 1). It is associated with 79 reactions in 11 pathways (Table 2). It also shows that succinate acts as substrate at 5 points and as product at 35 points. Succinyl co-A transferase consist of two fusion partners (Figure 2). It is made up of 6 PRODOM domains with an N terminal subunit A and a C terminal subunit B (Table 2). This fusion protein is shown to simulate protein-protein interactions in *Pseudomonas putida* (20) and mimic operon like structure in *C. acetobutylicum* (21). The fifth metabolite, glutamate is produced by *ornithine aminotransferase* (Figure 1) and is involved in 121 reactions in 11 pathways (Table 2). It acts as substrate at 25 points and as product at 57 points. This protein is made up of 8 PRODOM domains with an N terminal subunit A and a C terminal subunit B (Table 2). These data show the possible link between fusion proteins and pathways with reference to oxaloacetate, acetyl co-A, succinyl co-A, succinate and glutamate. The hypothesis is that fusion proteins could meet the demand for metabolites by fusing two or more fusion partners that are physically separated in one or more bacteria. This is consistent with a previous report which suggested that biotin carboxylase family of enzymes have evolved into a complex multifunctional protein from smaller mono-functional precursors through successive gene fusions (18). Therefore, the fusion proteins with modular organization possibly acquire selective incremental functions for optimized role in complex networks.

5. DISCUSSION

Protein evolution is extremely efficient in generating systems that are optimally adapted in cellular environment (22-25). Optimality can be achieved by changing the topology of metabolic networks by tuning enzymatic or regulatory materials (11). Here, we show that metabolites like oxaloacetate, acetyl co-A, succinyl co-A, succinate and glutamate are products of fusion proteins. These metabolites have high connectivity index, suggesting their greater degree of involvement within networks. This observation implies the association of fusion proteins with complex metabolic networks. The association between human fusion proteins and metabolites with high connectivity is intriguing (Table 2). Detailed analysis of fusion proteins highlights the transition from a 'protein-protein interface' to either a 'domain-domain interface' or

Network evolution and fusion proteins

an operon structure (a group of genes all controlled by the same regulatory gene) (1). This evolutionary transition is intriguing and it is important to systematically investigate the functional link between fusion partners and fused proteins using thermodynamics calculations. The transition may be thermodynamically favorable as fusion proteins acquire reduced entropy compared to their physically separated fusion partners. Therefore, it is envisaged that fusion proteins confer selective advantage in the evolution of regulating metabolic dynamics. This is specifically advantageous for multi-enzyme complexes as it selects kinetic advantages over unassociated enzyme components by increasing connectivity with metabolites. It is also reported that fusion of components into a single polypeptide ensures stability between physically connected domain structures and active sites for a balanced stoichiometric production of intermediates in complex networks (26-27). The physical proximity of multiple active centers in the same metabolic pathways alleviates molecular diffusion and reduces side reactions in cellular environment (28). Our data for the six metabolic enzymes having fusion structures aligns well with these observations. This enables fusion proteins to catalyze sequential steps in a biochemical pathway because association of two active sites enhances the efficiency of two consecutive reactions. Thus, fused protein architecture illustrates an evolutionary strategy accreted for maintaining complex stoichiometric balance. It is often thought that the function of fused genes is simply an addition of function to pre-existing component genes. However, chimerical genes generate proteins with novel function (29). A recent structural analysis of the Histidine biosynthesis components HisA and HisF indicates that the protein structure after gene fusion was also subject to structural and functional adaptation (30). In this case, gene fusion produced a new protein fold with novel function. Thus, most fusion proteins reveal that they have acquired separate functional domains from each component protein through domain accretion by gene fusion. Physical connection between fused domains increases structural propensity between active centers for the regulation of material balance (31). Since fusion proteins help in the evolution of complex networks, even a modest addition of domains could significantly increase numerous new interactions. This strategy helps to maintain equilibrium in a dynamic network with huge nodes. Thus, large networks of molecular interactions are regulated by relatively few genes in some organisms (32). That is, the 'gene number' (number of genes in a species genome) is negligible to 'reaction number' (number of reactions in a species cell) in higher eukaryotes. Our results illustrate an important evolutionary phenomenon that involves the formation of cellular network dynamics with the help of fusion proteins having multi-domain structural architectures with incremental functions.

6. CONCLUSION

The current analysis provides new insights into the relation of fusion proteins and network evolution. Thus, evolution of fusion genes has a key role in the selection and design of multifaceted network associated with complex genomes. This might have enabled fusion proteins to

accrete incremental biochemical function for a balanced regulation of metabolic networks. Hence, protein fusion confers a selective advantage in evolution. We believe that an understanding of fusion scenarios and their association with members of a pathway should enable us to appreciate the role played by them to combine or to share metabolites across networks, creating novel pathways with functional diversity. In reality, there are many more fusion structures of varying phylogeny and a majority of them have not been captured as the networks are extremely large and complex. Although, the insights drawn from a detailed study of six fusion proteins is found interesting and valuable, additional evidence is required to establish a comprehensive relationship between additive biochemical function and network dynamics.

7. ACKNOWLEDGEMENTS

This work was supported by Nanyang Technological University, National University of Singapore and A*STAR-BMRC, Singapore (Grant # 01/1/21/19/191 and Grant # 03/1/22/19/242).

8. REFERENCES

1. Yiting Y., I. Chaturvedi, L.K. Meow, P. Kanguane & M.K. Sakharkar: Can ends justify the means? Digging deep for human fusion genes of prokaryotic origin. *Front Biosci* 9, 2964-2971 (2004)
2. Yanai I., A. Derti. & 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. USA* 98, 7940-7945 (2001)
3. Katzen F., M. Deshmukh, F. Daldal & J. Beckwith: Evolutionary domain fusion expanded the substrate specificity of the transmembrane electron transporter DsbD. *EMBO J.* 21, 3960-3969 (2002)
4. Berthonneau E. & M. Mirande: A gene fusion event in the evolution of aminoacyl-tRNA synthetases. *FEBS Lett.* 470, 300-304 (2000)
5. Marcotte E.M., M. Pellegrini, H.L. Ng, D.W. Rice, T.O. Yeates & D. Eisenberg: Detecting protein function and protein-protein interactions from genome sequences. *Science* 285, 751-753 (1999)
6. Long M: A new function Evolved from Gene Fusion. *Genome Res.* 10, 1655-1657 (2000)
7. Truong K. & M. Ikura: Domain fusion analysis by applying relational algebra to protein sequence and domain databases. *BMC Bioinformatics.* 4, 16 (2003)
8. Suhre K. & J.M. Claverie: FusionDB: a database for in-depth analysis of prokaryotic gene fusion events. *Nucleic Acids Res.* 1, D273-D276 (2004)
9. Fernie A.R., F. Carrari & L.J. Sweetlove: Respiratory metabolism: glycolysis, the TCA cycle and

Network evolution and fusion proteins

- mitochondrial electron transport. *Curr Opin Plant Biol.* 7, 254-261 (2004)
10. Blank L.M. & U. Sauer: TCA cycle activity in *Saccharomyces cerevisiae* is a function of the environmentally determined specific growth and glucose uptake rates. *Microbiology.* 150, 1085-1093 (2004)
11. Segre D., D. Vitkup & G.M. Church: Analysis of optimality in natural and perturbed metabolic networks. *Proc. Natl. Acad. Sci. USA* 99, 15112-15117 (2002)
12. Stelling J., S. Klamt, K. Bettenbrock, S. Schuster & E.D. Gilles: Metabolic network structure determines key aspects of functionality and regulation. *Nature* 420, 190-193 (2002)
13. Ihmels J., R. Levy & N. Barkai: Principles of transcriptional control in the metabolic network of *Saccharomyces cerevisiae*. *Nat. Biotechnol.* 22, 86-92 (2004)
14. Servant F., C. Bru, S. Carrere, E. Courcelle, J. Gouzy, D. Peyruc & D. Kahn: ProDom: automated clustering of homologous domains. *Brief Bioinform.* 3, 246-325 (2002)
15. Kanehisa M., S. Goto, S. Kawashima, Y. Okuno & M. Hattori: The KEGG resource for deciphering the genome. *Nucleic Acids Res.* 32, D277-D280 (2004)
16. Mukhopadhyay B., E. Purwantini, C.L. Kreder, & 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)
17. Bott M. & P. Dimroth: *Klebsiella pneumoniae* genes for citrate lyase and citrate lyase ligase: localization, sequencing, and expression. *Mol Microbiol.* 14, 347-356 (1994)
18. Obermayer M. & F. Lynen: Structure of biotin enzymes. *Trends. Biochem. Sci.* 1, 169-171 (1976)
19. Marsh E.N., S.E. Harding, & P.F. Leadlay: Subunit interactions in *Propionibacterium shermanii* methylmalonyl-CoA mutase studied by analytical ultracentrifugation. *Biochem. J.* 260, 353-358 (1989)
20. Parales R.E. & C.S. Harwood: Characterization of the genes encoding beta-ketoadipate: succinyl-coenzyme A transferase in *Pseudomonas putida*. *J Bacteriol.* 174, 4657-4666 (1992)
21. Yeh W.K. & 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)
22. Freeland S.J., R.D. Knight, L.F. Landweber & L.D. Hurst: Early Fixation of an Optimal Genetic Code. *Mol. Biol. Evol.* 17, 511-518 (2000)
23. Heinrich R. & S. Schuster: The modeling of metabolic systems. Structure, control and optimality. *Biosystems* 47, 61-77 (1998)
24. Stephani A., J.C. Nuno & R. Heinrich: Optimal Stoichiometric Designs of ATP-producing Systems as Determined by an Evolutionary Algorithm. *J. Theor. Biol.* 199, 45-61 (1999)
25. Akashi H. & T. Gojobori: Metabolic efficiency and amino acid composition in the proteomes of *Escherichia coli* and *Bacillus subtilis*. *Proc. Natl. Acad. Sci. USA* 99, 3695-3700 (2002)
26. McCarthy A.D. & D.G. Hardie: Fatty acid synthase: an example of protein evolution by gene fusion. *Trends Biochem. Sci.* 9, 60-63 (1984)
27. Rochet J.C., Oikawa, K., Hicks, L.D., Kay, C.M., Bridger, W.A. & Wolodko, W.T. Productive interactions between the two domains of pig heart CoA transferase during folding and assembly. *Biochemistry* 36, 8807-8820 (1997)
28. West I.C: Channelling in intermediary metabolism, ed. Auius, L.S. (H. S. A., London). 9, 13-69 (1997)
29. Brodsky G., T. Barnes, J. Bleskan, L. Becker, M. Cox & 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)
30. Lang D., R. Thoma, M. Henn-SAX, R. Sterner & M. Wilmanns: Structural evidence for evolution of the beta/alpha barrel scaffold by gene duplication and fusion. *Science* 289, 1546-1550 (2000)
31. Koonin E.V., Y.I. Wolf & G.P. Karev: The structure of the protein universe and genome evolution. *Nature* 420, 218-223 (2002)
32. Lander E.S., L.M. Linton, B. Birren, C. Nusbaum, M.C. Zody, J. Baldwin, K. Devon, K. Dewar, M. Doyle, W. FitzHugh, R. Funke, D. Gage, K. Harris, A. Heaford, J. Howland, L. Kann, J. Lehoczy, R. LeVine, P. McEwan, K. McKernan, J. Meldrim, J.P. Mesirov, C. Miranda, W. Morris, J. Naylor, C. Raymond, M. Rosetti, R. Santos, A. Sheridan, C. Sougnez, N. Stange-Thomann, N. Stojanovic, A. Subramanian, D. Wyman, J. Rogers, J. Sulston, R. Ainscough, S. Beck, D. Bentley, J. Burton, C. Clee, N. Carter, A. Coulson, R. Deadman, P. Deloukas, A. Dunham, I. Dunham, R. Durbin, L. French, D. Grafham, S. Gregory, T. Hubbard, S. Humphray, A. Hunt, M. Jones, C. Lloyd, A. McMurray, L.

Network evolution and fusion proteins

Matthews, S. Mercer, S. Milne, J.C. Mullikin, A. Mungall, R. Plumb, M. Ross, R. Shownkeen, S. Sims, R.H. Waterston, R.K. Wilson, L.W. Hillier, J.D. McPherson, M.A. Marra, E.R. Mardis, L.A. Fulton, A.T. Chinwalla, K.H. Pepin, W.R. Gish, S.L. Chissoe, M.C. Wendl, K.D. Delehaunty, T.L. Miner, A. Delehaunty, J.B. Kramer, L.L. Cook, R.S. Fulton, D.L. Johnson, P.J. Minx, S.W. Clifton, T. Hawkins, E. Branscomb, P. Predki, P. Richardson, S. Wenning, T. Slezak, N. Doggett, J.F. Cheng, A. Olsen, S. Lucas, C. Elkin, E. Uberbacher, M. Frazier, R.A. Gibbs, D.M. Muzny, S.E. Scherer, J.B. Bouck, E.J. Sodergren, K.C. Worley, C.M. Rives, J.H. Gorrell, M.L. Metzker, S.L. Naylor, R.S. Kucherlapati, D.L. Nelson, G.M. Weinstock, Y. Sakaki, A. Fujiyama, M. Hattori, T. Yada, A. Toyoda, T. Itoh, C. Kawagoe, H. Watanabe, Y. Totoki, T. Taylor, J. Weissenbach, R. Heilig, W. Saurin, F. Artiguenave, P. Brottier, T. Bruls, E. Pelletier, C. Robert, P. Wincker, D.R. Smith, L. Doucette-Stamm, M. Rubenfield, K. Weinstock, H.M. Lee, J. Dubois, A. Rosenthal, M. Platzer, G. Nyakatura, S. Taudien, A. Rump, H. Yang, J. Yu, J. Wang, G. Huang, J. Gu, L. Hood, L. Rowen, A. Madan, S. Qin, R.W. Davis, N.A. Federspiel, A.P. Abola, M.J. Proctor, R.M. Myers, J. Schmutz, M. Dickson, J. Grimwood, D.R. Cox, M.V. Olson, R. Kaul, C. Raymond, N. Shimizu, K. Kawasaki, S. Minoshima, G.A. Evans, M. Athanasiou, R. Schultz, B.A. Roe, F. Chen, H. Pan, J. Ramser, H. Lehrach, R. Reinhardt, W.R. McCombie, M. de la Bastide, N. Dedhia, H. Blocker, K. Hornischer, G. Nordsiek, R. Agarwala, L. Aravind, J.A. Bailey, A. Bateman, S. Batzoglou, E. Birney, P. Bork, D.G. Brown, C.B. Burge, L. Cerutti, H.C. Chen, D. Church, M. Clamp, R.R. Copley, T. Doerks, S.R. Eddy, E.E. Eichler, T.S. Furey, J. Galagan, J.G. Gilbert, C. Harmon, Y. Hayashizaki, D. Haussler, H. Hermjakob, K. Hokamp, W. Jang, L.S. Johnson, T.A. Jones, S. Kasif, A. Kasprzyk, S. Kennedy, W.J. Kent, P. Kitts, E.V. Koonin, I. Korf, D. Kulp, D. Lancet, T.M. Lowe, A. McLysaght, T. Mikkelsen, J.V. Moran, N. Mulder, V.J. Pollara, C.P. Ponting, G. Schuler, J. Schultz, G. Slater, A.F. Smit, E. Stupka, J. Szustakowski, D. Thierry-Mieg, J. Thierry-Mieg, L. Wagner, J. Wallis, R. Wheeler, A. Williams, Y.I. Wolf, K.H. Wolfe, S.P. Yang, R.F. Yeh, F. Collins, M.S. Guyer, J. Peterson, A. Felsenfeld, K.A. Wetterstrand, A. Patrinos, M.J. Morgan, J. Szustakowski, P. de Jong, J.J. Catanese, K. Osoegawa, H. Shizuya, S. Choi & Y.J. Chen: Initial sequencing and analysis of the human genome. *Nature* 409, 860-921 (2001)

Key Words: Gene Fusion, Fusion Partners, Functional Accretion, Networks, Connectivity, Metabolites

Send correspondence to: Meena Kishore Sakharkar Ph.D, Nanyang Technological University, Singapore 639798; Tel: (+65) 6 790 5836; Fax: (+65) 6 774 4340; E-mail: mmeena@ntu.edu.sg

<http://www.bioscience.org/current/vol10.htm>

Galley

APPENDIX 4

Insight into gene fusion from molecular dynamics simulation of fused and un-fused IGPS (Imidazole Glycerol Phosphate Synthetase)

Yu Yiting, Li Lei, Meena Kishore Sakharkar, Pandjassaram Kanguane*

School of Mechanical and Aerospace Engineering, Nanyang Technological University, 50, Nanyang Avenue, Singapore 639798; Address correspondence to Dr. Pandjassaram Kanguane, School of Mechanical and Aerospace Engineering, Nanyang Technological University, 50, Nanyang Avenue, Singapore 639798, E-mail: mpandjassaram@ntu.edu.sg, Phone: +65 6790 4957 ; Fax: +65 6774 4340

received February 4, 2005; accepted February 27, 2006; published online February 28, 2006

Abstract:

Gene fusion produces proteins with novel structural architectures during evolution. Recent comparative genome analysis shows several cases of fusion/fission across distant phylogeny. However, the selection forces driving gene fusion are not fully understood due to the lack of structural, dynamics and kinetics data. Available structural data at PDB (protein databank) contains limited cases of structural pairs describing fused and un-fused structures. Nonetheless, we identified a pair of IGPS (imidazole glycerol phosphate synthetase) structures (comprising of HisF - glutaminase unit and HisH - cyclase unit) from *S. cerevisiae* (SC) and *T. thermophilus* (TT). The HisF-HisH structural units are domains in SC and subunits in TT. Hence, they are fused in SC and un-fused in TT. Subsequently, a domain-domain interface is formed in SC and a subunit-subunit interface in TT between HisF and HisH. Our interest is to document the structure and dynamics differences between fused and un-fused IGPS. Therefore, we probed into the structures of fused IGPS in SC and un-fused IGPS in TT using molecular dynamics simulation for 5ns. Simulation shows that fused IGPS in SC has larger interface area between HisF-HisH and greater radius of gyration compared to un-fused IGPS in TT. These structural features for the first time demonstrate the evolutionary advantage in generating proteins with novel structural architecture through gene fusion.

Keywords: gene fusion; fused proteins; evolution; molecular dynamics; interface; domains; subunits

Background:

Proteins with novel structural architectures are generated by gene fusion in one species' compared to another species. [1, 2] Proteome wide comparative analyses within and across kingdoms showed a large number of fused structures. [3] Proteins created by gene fusion are shown to have enhanced role in pathways by Yanai *et al.* [4], simulate protein subunit interaction by Marcotte *et al.* [5], novel function by

Long [6], enhanced substrate specificity by Katzen *et al.* [7] and enzyme multi-functionality by Berthonneau and Mirande. [8] These reports indicate the existence of several isolated cases of fused protein as a result of gene fusion in evolutionary history. However, the advantage (structure, dynamics and kinetics) of producing fused proteins in one species compared to the un-fused protein orthologs in another species is not fully understood.

Figure 1: A schematic representation of histidine biosynthetic pathway is given. IGPS (imidazole glycerol phosphate synthase) catalyzes the fifth and sixth step of the histidine biosynthetic pathway in microbes, fungi, and plants. IGPS catalyzes the bifurcation step of the histidine and *de novo* purine biosynthesis pathways.

Figure 2: A schematic representation of the 555 residue long IGPS from SC and TT is shown. The HisH and HisF domains in SC are fused by a 33 residue long linker (206-238). However, the HisH and HisF subunits are un-fused in TT and the linker is absent.

Despite the availability of several protein structures at the PDB (protein databank), cases of structural pairs describing fused and un-fused protein structures are limited. Nevertheless, we identified a pair of IGPS (imidazole glycerol phosphate synthetase) structures (comprising of HisF - glutaminase unit and HisH - cyclase unit) from *S. cerevisiae* (SC) and *T. thermophilus* (TT). IGPS catalyzes the fifth and sixth steps of the histidine biosynthetic pathway in microbes, fungi, and plants. It forms the imidazole ring of the histidine precursor imidazole glycerol phosphate. [9, 10] IGPS is a glutamine

amido-transferase that catalyzes the formation of IGP (Imidazole glycerol phosphate) and AICAR (5-aminoimidazole-4-carboxamide ribonucleotide) from PRFAR (N^1 -((5'-phosphoribosyl) formimino) - 5-aminoimidazole-4-carboxamide ribonucleotide). Interestingly, IGPS functions at the junction of histidine biosynthesis and *de novo* purine biosynthesis, since AICAR is the entry point to the latter (Figure 1). Thus, IGPS is a key metabolic enzyme, which links amino acid and nucleotide biosynthesis pathways. IGPS has different structural architectures in SC and TT. In TT, IGPS forms a hetero-dimer interface with glutaminase (HisH) and cyclase (HisF) subunits. [1] In SC, the two subunits are fused into a single polypeptide an N terminal HisH domain and a C terminal HisF domain forming an interface between HisH-HisF domains. [11] The conserved glutamine binding site in IGPS is at the interface of HisH and HisF in both TT and SC. [12, 13] Thus, the stability of the interface plays an important role in glutaminase catalysis. The subunit interaction in TT and domain interaction in SC mediate the catalytic activity of glutamine hydrolysis. [9] Thus, the fused protein retains the glutaminase active site and a small linker connects HisF and HisH in SC. However, the structure, dynamics and kinetics advantages of this arrangement in fused proteins are not known. Therefore, it is our interest to probe into the structure and dynamics properties of the fused (SC - IGPS) and un-fused (TT - IGPS) structures using molecular dynamics simulation.

Methodology:

Initial IGPS structures for simulation:

We used the IGPS structures for SC (PDB code: 1OX6 - resolution 2.4 Å) [14] and TT (PDB code: 1KA9 - resolution 2.3 Å) [12] downloaded from PDB. Hydrogen atoms were added to these structures using SYBYL 6.8 (Tripos Associates Inc.).

Molecular dynamics simulation:

All molecular mechanics calculations were carried out using the TRIPOS force field [15] in SYBYL (Molecular Modeling Software Package, Version 6.8, Tripos Associates Inc.) running on a Silicon Graphics Workstation. The energy function used in the force field was defined as the sum of six contributions (bond stretching, angle bending, torsion, van der Waals, electrostatic and planarity (for aromatic conjugated systems). Minimizations of the potential energy of the system were carried out using the Simplex algorithm and the Powell torsional gradient algorithm as implemented in SYBYL, terminating when a 0.5 Kcal/mol-Å energy gradient shift was obtained. A distance dependent dielectric constant of 1.0 was used to compute electrostatic effects. The non-bonded cutoff distance used was 8 Å and the net

atomic charges in the residues were calculated by the Gasteiger-Hucker method. [16, 17] The *in vacuo* system was simulated at constant temperature, constant volume (NVT) ensemble which is referred to as the canonical ensemble. The system was run at a temperature of 300 K using a coupling constant of 100 femtosecond. The initial atom velocities were employed from a Maxwell-Boltzmann distribution with scaling velocities. The non-bonded pair list was updated every 25 femtosecond and an 8 Å cut-off was applied. During the simulation, the integration step was set up as 1 femtosecond and molecular snapshots were saved for every 1000 steps (1 pico-second). A total of 5000 structures were generated and the simulation properties were derived from the analyses of these snapshots.

Analysis:

We performed a comprehensive analysis of structures in each trajectory to detect structural differences between the two simulated systems. The flexibilities of the different structures were assessed by computing gap volume, gap index, interface area and radius of gyration.

Results:

Figure 2 illustrates the fused and un-fused IGPS structures in SC and TT, respectively. A small linker connects HisH (glutaminase) and HisF (cyclase) in SC and thus IGPS is fused in SC. However, this linker is absent in TT and HisH – HisF are un-fused in TT. The HisH domain in SC has 47% similarity to the HisH subunit in TT. Similarly, the HisF domain in SC has 48% similarity to the HisF subunit in TT. The HisH and HisF units are homologous and have similar structures in SC and TT.

Figure 3 shows the structural snapshots of TT IGPS and SC IGPS at 0 and 5 ns simulation. The HisH and HisF interface in TT and SC is also visualized in Figure 3. The linker connecting HisH and HisF in SC is labeled and this linker is absent in TT. Thus, the

interface is formed by HisH and HisF domains in SC and HisH and HisF subunits in TT. This demonstrates an evolutionary transition from a subunit-subunit interface in TT to a domain-domain interface in SC.

Figure 4 shows the interface area (change in solvent accessible surface area upon interface formation between HisH and HisF calculated using NACCESS implemented using Lee and Richard algorithm [18]) in TT IGPS and SC IGPS for structures generated over a 5 ns simulation. The interface area between HisH and HisF is significantly larger ($> 1000 \text{ \AA}^2$) in fused SC IGPS compared to the un-fused TT IGPS throughout the simulation period.

Figure 3: Snapshots of IGPS at 0 ns and 5 ns for SC and TT are shown. The molecules are rendered as a ribbon diagram with contrasting colors for the glutaminase (bottom) and cyclase (top) domains. The figure shows the bacterial IGPS is a heterodimer and the yeast IGPS is a monomer. The C-terminal cyclase domain of yeast IGPS has a longer loop at the top of the barrel than that of the bacteria.

Table 1: Residue conservation at the interface of IGPS in TT and SC

HisF (TT SC)	Total	Interior	Interface	Surface
(a) No. of conserved residue	113	14	20	61
(b) No. of residues	317	34 43	42 43	241 231
(a)/(b)	35%	41% 33%	47%	26%
HisH (TT SC)	Total	Interior	Interface	Surface
(c) No. of conserved residue	69	9	17	23
(d) No. of residues	205	36 41	36 53	133 111
(c)/(d)	34%	25% 22%	47% 32%	17% 21%

Data shows that interface residues are more conserved than surface residues for HisF and HisH between TT and SC. The number of conserved residues for HisF is 113 ($> 95 = (14+20+61)$) and the remaining 18 conserved residues are located at different regions (interior/interface/surface) in the two structures from TT and SC. This explanation holds true for the HisH structures in TT and SC.

Figure 4: Interface area between HisH and HisF is given for IGPS from SC and TT over a 5 ns molecular dynamics simulation. The domain-domain interface area in SC is larger than TT throughout the simulation period.

Figure 5 shows the gap volume (calculated using SURFNET [19]) between HisH and HisF in SC IGPS and TT IGPS for structures generated over a 5 ns simulation. Similar to interface area, the gap volume is consistently larger in SC IGPS compared to TT IGPS throughout the simulation period.

Figure 5: Gap volume between HisH and HisF is given for IGPS from SC and TT over a 5 ns molecular dynamics simulation. The domain-domain gap volume in SC is larger than TT throughout the simulation period.

Figure 6 shows the gap index (ratio of gap volume to interface area) between HisH and HisF in SC IGPS and TT IGPS for structures generated over a 5 ns simulation. Unlike interface area and gap volume,

gap index is steadily similar throughout the simulation period.

Figure 6: Gap index (ratio of volume to interface area) between HisH and HisF is given for IGPS from SC and TT over a 5 ns molecular dynamics simulation. The gap index is similar for the interface between HisH and HisF from SC and TT.

Figure 7 shows the radius of gyration for SC IGPS and TT IGPS for structures generated over a 5 ns simulation. Similar to interface area and gap volume, the radius of gyration for SC IGPS is considerably larger compared to TT IGPS throughout the simulation period.

Figure 7: Radius of gyration (measure of unfolding and flexibility) for IGPS from SC and TT is given over a 5 ns molecular dynamics simulation. The radius of gyration is larger for SC IGPS is larger than TT IGPS throughout the simulation.

Table 2: Structural properties of IGPS in TT and SC is given for initial and final structures

Parameters	Initial crystal structure		Final structure after simulation (5 ns)		Difference between initial crystal and final structures	
	TT	SC	TT	SC	TT	SC
Interface area (Å ²)	2691.5	3940.3	1652.7291	2617.3474	-1039	-1323
Gap volume (Å ³)	3606	3952	3363.6	4627.1	-242	675
Gap index (Å)	0.746	0.997	0.491357207	0.565656	-0.256	-0.432
Radius of gyration (Å)	25.52	21.59	21.25	22.91	-4.27	1.32

Discussion:

Gene fusion is an important evolutionary phenomenon for the formation of proteins with new structural architectures. [1-8] Comparative sequence analysis between closely and distantly related species shows evidence for gene fusion/fission. [1, 2, 3] Therefore, it is of great significance to document the selection force generating such proteins with fused structural architectures. However, there is no documentation for structural evidence supporting the dynamics of these fused structures in the evolution of orthologous proteins.

The interface residues between HisF and HisH in TT and SC are more conserved than surface residues (Table 1). The interface residues similarities imply catalytic conservation at the interface. The structural properties for IGPS in TT and SC are given for initial and final structures (Table 2). The interface area, gap volume and gap index are greater in SC than TT in both initial and final structures. These values increased relatively due to simulation in both SC and TT. However, the radius of gyration in TT is larger than SC for the initial structure unlike the final structure (Table 2). Interestingly, the radius of gyration increased in SC and decreased in TT due to simulation.

The results given in Figure 3 to Figure 7 demonstrate the structure dynamics of fused IGPS in SC compared to the un-fused IGPS in TT. The IGPS in SC forms a domain-domain interface between HisH and HisF compared to a subunit-subunit interface in TT. The transition from a subunit-subunit interface in TT to a domain-domain interface in SC is interesting. The domain-domain interface area in SC is larger than the subunit interface area in TT over a 5 ns molecular dynamics simulation. The interface area in SC is 1400 Å² greater than in TT. The larger interface area in SC facilitates better domain-domain interactions compared to subunit interactions in TT (Figure 4). The amount of interface area determines the degree of atomic interaction at the interface. Larger HisH and HisF interface in SC imply better interaction between these two domains. Better interaction between HisF and HisH facilitates greater stability and kinetics in SC. This is assisted largely

by the linker segment connecting HisF and HisH domains in SC.

The gap volume between HisF and HisH domains from SC IGPS is larger than that between HisF and HisH subunits in IGPS from TT (Figure 5). The increased gap volume in SC IGPS may aid in substrate flow into the active sites formed by HisH and HisF domains. However, this flow of substrate is relatively restricted in TT IGPS in exchange for interface stability formed by subunit interaction. Larger gap volume in SC IGPS is partly helped by the linker between HisH and HisF which provides enhanced flexibility for these two domains. Interestingly, the increased gap volume in SC IGPS does not affect gap index (ratio of gap volume to interface area) in both SC IGPS and TT IGPS (Figure 6). This suggests that increased gap volume is proportional to the increased interface in SC compared to that in TT.

Radius of gyration in proteins is a measure of their size and implies their compactness. The radius of gyration for IGPS from SC and TT given in Figure 7 describes the unfolding of the structure during simulation. The flexibility rendered by the linker between HisF and HisH in the case of SC IGPS is shown by the increased radius of gyration compared to that in TT throughout the simulation period over 5 ns. The difference in the average radius of gyration between SC and TT IGPS is about 1.76 Å. This provides the explanation for the increased stability leading to greater kinetics of IGPS caused by the linker in the fused structure of SC IGPS.

The raise and fall in interface area, gap volume and gap index in TT during simulation is unusual. This may be due to the high interface movement between the weakly associated subunits. The proposed hypothesis driving the formation of fused proteins by gene fusion is the structural determinant providing increased stability, dynamics and kinetics facilitated during evolutionary selection. This is evident by the structure and dynamics of IGPS as described using interface area, gap volume and radius of gyration in SC and TT.

Conclusion:

A number of fusion proteins have been identified by comparative genome analysis using sequence comparison. This suggests that gene fusion is common in evolutionary phylogeny. However, the selection force driving gene fusion in organism evolution is not fully evident due to the lack of structure, dynamics and kinetics data supporting this phenomenon. Despite the growth in structures at PDB, the number of structural pairs illustrating fusion/fission in distant phylogeny is limited. Here, we show the importance of fused protein by probing

the fused IGPS structure in SC as against the un-fused structure in TT using molecular dynamics simulation. The simulation shows larger interface area and radius of gyration in SC IGPS compared to TT IGPS. Thus, fused IGPS in SC have better structural features than un-fused IGPS in TT. This finding provides meaningful insight for gene fusion in establishing optimal dynamics and kinetics. This is an extremely interesting one and is likely to become more important as the international structural genomics efforts increase significantly their production of structures

References:

- [1] Y. Yiting *et al.*, *Front. Biosci.* **9**, 2964-2971 (2004) [PMID: 15353329]
- [2] M.K. Sakharkar *et al.*, *Front. Biosci.* **10**, 1070-1078 (2005) [PMID: 15769606]
- [3] K. Truong & M. Ikura, *BMC Bioinformatics* **4**, 16 (2003) [PMID: 12734020]
- [4] I. Yanai *et al.*, *Proc. Natl. Acad. Sci. U.S.A.* **98**, 7940-7945 (2001) [PMID: 11438739]
- [5] E. Marcotte *et al.*, *Science* **285**, 751-753 (1999) [PMID: 10427000]
- [6] Long M. *Genome Res.* **10**, 1655-1657 (2000) [PMID: 11076848]
- [7] F. Katzen *et al.*, *EMBO J.* **21**, 3960-3969 (2002) [PMID: 12145197]
- [8] E. Berthonneau & M. Mirande, *FEBS Lett.* **470**, 300-304 (2000) [PMID: 10745085]
- [9] T.J. Klem & V.J. Davisson, *Biochemistry* **32**, 5177-5186 (1993) [PMID: 8494895]
- [10] T.J. Klem *et al.*, *Bacteriol.* **183**, 989-996 (2001) [PMID: 11208798]
- [11] S.V. Chittur *et al.*, *Protein Expr. Purif.* **18**, 366-377 (2000) [PMID: 10733892]
- [12] R. Omi *et al.*, *J. Biochem.* **132**, 759-765 (2002) [PMID: 12417026]
- [13] B.N. Chaudhuri *et al.*, *Structure* **9**, 987-997 (2001) [PMID: 11591353]
- [14] B.N. Chaudhuri *et al.*, *Biochemistry* **42**, 7003-7012 (2003) [PMID: 12795595]
- [15] J.G. Vinter *et al.*, *J. Comput. Aided. Mol. Des.* **1**, 31-51 (1987) [PMID: 3505586]
- [16] M. Marsili & J. Gasteiger, *Croat. Chem. Acta.* **53**, 601-614 (1980)
- [17] J. Gasteiger & M. Marsili, *Tetrahedron* **36**, 3219-3228 (1980)
- [18] B. Lee, & F.M. Richards, *J. Mol. Biol.* **55**, 379-400 (1971) [PMID: 5551392]
- [19] R.A. Laskowski, *J. Mol. Graph.* **13**, 323-330 (1995) [PMID: 8603061]

Citation: Yiting *et al.*, Bioinformation 1(3): 99-104 (2006)

Edited by K. Gunasekaran