ISSN 1672-9145 **Acta Biochimica et Biophysica Sinica** 2005, **37**(8): 561–566 **CN** 31-1940/Q

Short Communication

Identification and Categorization of Horizontally Transferred Genes in Prokaryotic Genomes

Shuo-Yong SHI[1], Xiao-Hui CAI[1], and Da-fu DING[1,2]*

[1] *Key Laboratory of Proteomics, Institute of Biochemistry and Cell Biology, Shanghai Institutes for Biological Sciences, Chinese Academy of Sciences, Graduate School of the Chinese Academy of Sciences, Shanghai 200031, China;*
[2] *School of Life Science & Technology, Shanghai Jiaotong University, Shanghai 200030, China*

Abstract Horizontal gene transfer (HGT), a process through which genomes acquire genetic materials from distantly related organisms, is believed to be one of the major forces in prokaryotic genome evolution. However, systematic investigation is still scarce to clarify two basic issues about HGT: (1) what types of genes are transferred; and (2) what influence HGT events over the organization and evolution of biological pathways. Genome-scale investigations of these two issues will advance the systematical understanding of HGT in the context of prokaryotic genome evolution. Having investigated 82 genomes, we constructed an HGT database across broad evolutionary timescales. We identified four function categories containing a high proportion of horizontally transferred genes: cell envelope, energy metabolism, regulatory functions, and transport/binding proteins. Such biased function distribution indicates that HGT is not completely random; instead, it is under high selective pressure, required by function restraints in organisms. Furthermore, we mapped the transferred genes onto the connectivity structure map of organism-specific pathways listed in Kyoto Encyclopedia of Genes and Genomes (KEGG). Our results suggest that recruitment of transferred genes into pathways is also selectively constrained because of the tuned interaction between original pathway members. Pathway organization structures still conserve well through evolution even with the recruitment of horizontally transferred genes. Interestingly, in pathways whose organization were significantly affected by HGT events, the operon-like arrangement of transferred genes was found to be prevalent. Such results suggest that operon plays an essential and directional role in the integration of alien genes into pathways.

Key words genome evolution; horizontal gene transfer; function and pathway analysis

Horizontal gene transfer, the transfer of genes between different species, has been widely recognized as one of the major driving forces in prokaryotic genome evolution [1,2]. However, it has been pointed out that most analyses of HGT have focused on either studying the phylogenetic relations of individual gene families, which may be actively transferred, or estimating the global fraction of individual genomes where genes are introduced by HGT [3]. To date, there are inadequate data to answer two important issues: what types of genes are subject to be transferred and to what extent the transferred genes have been successfully integrated into existing biological pathways.

As to the first question, Nakamura *et al.* recently reported that biological functions of transferred genes are biased to three categories: cell envelope, cellular process, and regulatory process, suggesting that the transferability of genes depends heavily on their functions [4]. However, their work was based on a set of horizontally transferred (HT) genes detected by DNA composition analysis, which is limited in identifying recent HGT events due to the amelioration processes (adaptation to host genome features) [5]. Therefore, their conclusion may not be generalized to all HGT genes.

Received: February 3, 2005 Accepted: May 7, 2005
This work was supported by the grants from the National High Technology Research and Development Program of China (No. 2002AA234021) and the Knowledge Innovation Program of the Chinese Academy of Sciences (KJCX1-08, KSCX2-2-07)
*Corresponding author: Tel, 86-21-54921254; Fax, 86-21-54921011; E-mail, dingdafu@server.shcnc.ac.cn

DOI: 10.1111/j.1745-7270.2005.00075.x

Pathway organization and evolution is a long-standing, interesting and important question. Schmidt *et al.* investigated the enzyme structure distribution and concluded that the overall topology of metabolic networks is preserved through the evolution process [6]. It is interesting to measure whether HGT, a force different from other vertical evolutionary events, will also comply with pathway organization rules. Some previous individual pathway analyses indicated that HGT might improve pathway capacities and thus provide recipient species readily-made responses to environmental challenge [3,7,8]. All of those results imply the active role of HGT in pathway evolution. In contrast, some other reports argued that selective barriers exist and preclude the recruitment of transferred genes. As a result, in most cases, alien homologues will have little chance to improve the fitness of networks [9,10]. A systematic investigation is necessary to measure the impact of HGT on pathway organization and evolution.

We emphasized that the reliable HGT dataset across broad evolutionary timescales was preliminary for all above analyses. Various methods based on different models have been designed to detect a specific HGT dataset with different types and ages [11]. The gene transfer events will be more robustly delineated when all of the available methods are used; application of a variety of methods will provide the best information about the scope of gene transfer across broad timescale [12,13].

In this article, we attempt for the first time to design a stringent procedure in combination with different methods to address two basic issues mentioned above.

Experimental Procedures

We retrieved the complete sequences of 82 prokaryote genomes from the National Center for Biotechnology Information ftp site (ftp://www.ncbi.nih.gov/genomes/bacteria/). We parsed the downloaded GenBank files and then constructed a non-redundant protein database.

In general, there are three types of approaches for HGT identification: abnormally high BLAST hits to genes of distant taxa or anomalous phylogenetic profile distribution; phylogenetic analysis; and atypical DNA composition analysis. Each method has its advantages and caveats [11]. To achieve a more conservative HGT set across broad timescales in evolution, we designed a combined method for HGT identification. The scheme of pipeline is available in supplement file (http://www.abbs.info/sd/pipeline.pdf).

The first method we used draws on unexpectedly high BLAST hits to a distant taxon. The suspicion of HGT usually emerges when a protein/gene sequence from a particular organism shows stronger similarity to homologs from distant taxa. However, the simple best-match method is prone to erroneous results [14,15], so we designed a more rigorous approach to avoid false positives. All protein sequences from each genome were compared with our protein database using the BLASTP program [16] (*E*-value cut off 1e–10) and the results were parsed to search for paradoxical phyletic distribution of homologs. We considered an open-reading frame (ORF) to be horizontally derived if all of the following conditions were met.

(1) The *E*-value of the best BLAST hit in *H* should be less than 1e–20 and the first five BLAST hits must be homologs from a distant taxon. We set the distant taxon on two levels: non-self phylum or non-self superkingdom.

(2) All of its homologs are from a distant taxon or the *E*-value of the closest homolog from a distant taxon (E_1) is significantly lower than the *E*-value of the closest homolog not from the distant taxon (E_{nd}). We set the cutoff as **Equation (1)**:

$$[\log(E_1)/\log(E_{nd})]\geq 2 \quad (1)$$

(3) In the taxonomy tree constructed from our 82 species, there are 10 different phylums which include less than five species members. Therefore, only 23 species from those phylums will be suitable for identifying the candidates of HT genes originated from non-self superkingdoms to avoid sampling bias [15].

Orthologs are genes in different species that evolved from a common ancestral gene by speciation [17]. When the orthologs of one gene are mainly distributed in a non-self superkingdom or a non-self phylum, it is evident that HGT occurred. In this article, we used the operational definition of ortholog as follows. Given a protein *P*, its ortholog in another genome must be the best hit of *P* in this genome by BLAST search (*E*-value cut off 1e–10 and more than 60% of residues should be included in BLAST alignment), and *vice versa*. If the corresponding phylogenetic orthologs distribution of *P* is unusual, the gene encoding protein *P* may be horizontally transferred in evolution.

Both of the two methods described above are based on BLAST search. Although a stringent cutoff was set to remove some false positives, these methods are still approximate approaches and need more phylogenetic validation. Recently, Frickey *et al.* developed a suite of programs, PhyloGenie (http://protevo.eb.tuebingen.mpg.de/miscpages/phylogenie/), which provides a novel alignment routine to achieve good alignment and also automates the steps from seed sequence to phylogenetic inference

[18]. HGT candidates identified by the two methods mentioned above were re-checked by PhyloGenie.

As stated before, phylogenetic analysis does not work well in identifying transferred genes among close species [11]. Thus, there are legitimate reasons for using atypical DNA composition analysis as a alternative approach to infer HGT. Garcia-Vallve *et al.* constructed an HGT database by analyzing G+C content, codon and amino acid usage of 82 genomes [19]. Recently, Tsirigos *et al.* also developed a new composition analysis method for HGT identification [20]. Previous research has pointed out that different methods based on DNA composition analysis have different advantages and caveats [21]. To achieve a more conservative HGT set, the agreement results of the two methods were taken as HT candidates.

KEGG protein pathways provide organism-specific molecular interaction network information including metabolic pathways, regulatory pathways, and molecular complexes. We mapped the protein dataset to the KEGG organism-specific protein pathways according to the pathway information relating to each record in the KEGG protein database. What should be noted is that the so-called "organism-specific pathways" are computer-driven to some extent. Therefore, certain pathways represented both in isolation and as a subgraph of larger pathways were merged by recursive procedures to avoid partial pathway duplication. All the proteins in the 82 genomes were mapped to the pathway database KEGG. In one genome containing N genes, there are M genes that are horizontally transferred. If we consider a pathway containing K members, any of these members either belongs to HT genes or does not; for example, H of the A genes are horizontally transferred and K–H are not. It is important to establish the probability of this happening by chance. This probability is appropriately modeled by hypergeometric distribution with parameters N, M and K [22]. Using the cumulative hypergeometric probability distribution, we can calculate the chance of finding at least H horizontally transferred genes in this specific pathway by chance, so that the p-value is given by **Equation (2)**.

$$p\text{-value} = 1 - \sum_{i=0}^{H-1} \frac{\binom{M}{i}\binom{N-M}{K-H}}{\binom{N}{K}} \quad (2)$$

This test measures whether a pathway is enriched with HT genes to a greater extent than expected by chance. When the p-value is less than 0.05 (significant level 0.05) and that pathway has more than three transferred genes, we considered that the organization of this pathway was significantly affected by HGT.

Results and Discussion

We applied our combined method to examine 82 completely sequenced genomes. The statistics of HGT number and proportion in each genome is available (http://www.abbs.info/sd/hgtstatistictable.htm). In this article, different methods were combined to collect an HGT set across broad timescales in evolution. The outputs of different methods are compared in **Table 1**.

As expected, it is clear that one individual method failed to detect all the candidates of HGT. What should be noted is that the overlap number between the two methods is small. Such observation is in agreement with previous research. Ragan observed that different approaches, when applied to the same genomic sequence, recognized significantly different subsets of genes as being subject to HT and the overlap of different methods is less than expected by chance under a statistical model [12]. This is mainly due to the null hypothesis and assumptions used by different methods (**Table 1**). Thus, a multifaceted approach is needed for HGT identification and HGT might better be

Table 1 Comparison of different methods to identify horizontal gene transfer candidates

Method	Description	Output (n)	Common (n)
BLAST based on phylogenetic validation	Biased towards those genes which have been transferred across large phylogenetic distances, regardless of their time of arrival in a genome, but fails to detect transfers between more closely-related organisms.	4896	197
DNA composition analysis	Limited in recent HGT but can identify transfer among close relatives.	8211	

We combined two different methods to achieve a conservative HGT dataset. The power and limitation of each method is described in column 2. The number of HGT candidates identified by each method is listed in column 3. Column 4 indicates the number of agreement of different approaches.

represented as the union, not the intersection, of approaches addressing compatible null hypotheses. All HGT candidates were stored in database and the flat file can be downloaded from following website (http://www.abbs.info/sd/hgtdb.xls). Other related information, such as the location, function and taxonomy of the corresponding genomes, and links to databases UNIPRO, TIGR, GO and KEGG, was integrated into our database.

There are a total of 8893 HT proteins mapped to the function categories defined by the TIGR (The Institute of Genomic Research) microbial database [23]. Three uninformative categories were omitted: disrupted reading frame, unclassified and unknown function. **Fig. 1** shows the distribution of all HT genes in each category.

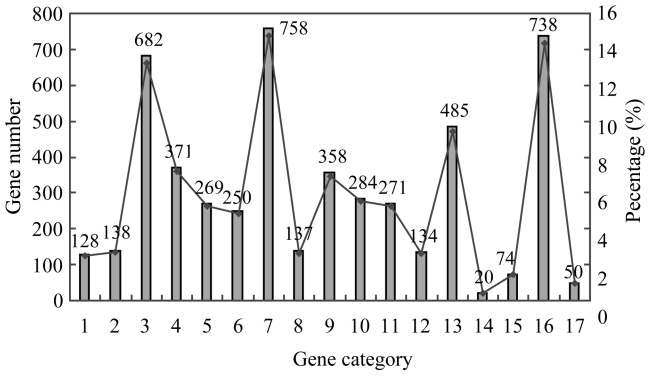

Fig. 1 Distribution of horizontally transferred genes in function categories

1, amino acid biosynthesis; 2, biosynthesis of cofactors, prosthetic groups, and carriers; 3, cell envelope; 4, cellular processes; 5, central intermediary metabolism; 6, DNA metabolism; 7, energy metabolism; 8, fatty acid and phospholipid metabolism; 9, other categories; 10, protein fate; 11, protein synthesis; 12, purines, pyrimidines, nucleosides, and nucleotides; 13, regulatory functions; 14, signal transduction; 15, transcription; 16, transport and binding proteins; 17, viral functions.

It appears clear that four main categories have significantly high proportions of horizontally transferred genes: cell envelope (13.2%), energy metabolism (14.7%), regulatory functions (9.4%), and transport and binding proteins (14.3%). There are also two categories showing relatively high distribution: cellular processes (7.2%) and other categories (6.9%). Note that "other categories" includes genes responsible for prophage functions, transposon functions and plasmid functions and is usually pathogenicity-related. It is worth pointing out that energy metabolism, which was shown to have the lowest proportion of horizontally transferred genes in the work of Nakamura *et al.* [4], is among these four categories. To further investigate this discrepancy, we compared our HGT dataset with the results of Nakamura *et al.*, and found that genes related to energy metabolism were mainly identified by methods not used by Nakamura *et al.*. We postulate that the difference is due to an underestimate of ancient HGT events in the work of Nakamura *et al*. Moreover, many previous researchers also observed that transferred genes involved in energy metabolism may provide an organism with a new adaptive capability, such as the ability to utilize a new source of carbon and energy [2]. Thus the transfer of large numbers of energy metabolism genes is not unexpected.

Similar to the work of Nakamura *et al.*, we also found that some operational gene categories, including amino acid biosynthesis, biosynthesis of cofactors or carriers, intermediary metabolism, fatty acid and phospholipid metabolism, have a low proportion of horizontally transferred genes. Based on these results, Nakamura *et al.* suggested that operational genes, previously considered generally transferable, should be further classified into two groups, according to their transferability, that is, the highly transferable genes and the lowly transferable genes. However, it seems that such conclusions do not clearly explain the abnormal function categories distribution in some genomes (http://www.abbs.info/sd/funcdistri.xls). For example, as shown in the supplement, the amino acid biosynthesis gene category has a comparably high proportion to one "highly" transferable gene group in four species; in fact, the highly transferable categories also sometimes show low proportions in certain genomes. We postulated that horizontal gene transfer is a process dominated by selection pressure from the demand of function evolution in individual organisms. According to the "genetic annealing model" [24], HGT will become progressively less likely to happen as the workings of cellular components improve. In addition, further study also pointed out that selective barriers in an optimized system make neutral evolution of alien genes unlikely [9]. Taken together, we believe, after the formation of a modern organism, the rate of transfer slows down and only when the organism is in emergency, such as a change of environment, do transfer events become active. In fact, our research and previous study [25] both observed that HGT is always related to adaptations to environmental challenges or responses to new niche expansion. Commonly, such evolutionary requirements involve the following aspects: novel metabolic capabilities, antibiotic

or resistance characteristics, virulence or pathogenity-related function, new cell surface structure and the change of regulation mechanisms. All of these are related to the observed function categories with high proportions of horizontally transferred genes. Therefore, we suggest that the biased function distribution should be ascribed to the guide of evolutionary demand of recipent genomes, reflecting the adaptive response to enviroment challenges. Furthermore, function demands are different in different organisms. Not all above-mentioned functions are in need and, in addition, some species may have particular requirements, such as the alternative amino acid and biosynthesis route. This may explain why some organisms did not show classical function category distribution. In general, it is more reliable to analyze the biased function distribution in the evolutionary context of specific species.

In this article, we mapped our horizontally transferred gene candidates onto the connectivity structure map of organism-specific pathways listed in KEGG. We found that, among the total of 5150 protein pathways in KEGG, 1421 pathways (approximately 28%) involved horizontally transferred genes. The full list of those pathways is available (http://www.abbs.info/sd/allhgtpath.doc). We found that, with the increase of involved transferred members, the number of related pathways sharply decreases. Of particular note is that only 1095 pathways (77%) have two or less horizontally transferred members. To some extent, the recruitment of one or two transferred genes into one pathway may only affect a few specific steps, if not a single step, through the replacement of original function members or the introduction of novel preferred proteins. They may affect only a very small number of members of the whole pathway structure. To evaluate the extent to which alien genes are integrated in each pathway, we used hypergeometric probability distribution to model the chance probability of finding at least H horizontally transferred genes in one specific pathway by chance. Through this approach, we found that only 114 pathways (2%) may be seriously affected by HGT. A list of these 114 pathways is also available (http://www.abbs.info/sd/sighgtpath.doc). It seems that HGT influences only specific pathways and, in most cases, pathways organization conserve well. Such results further support previous research conclusions [10]. This phenomenon can be well explained from an evolutionary point of view. Considering a pathway is a complex system, such an observation is not unexpected. After long-standing co-evolution, closely functional or structural interactions exist between original members, making a pathway a fine-tuned machine. Thus, in the evolutionary process after the formation of modern species, most pathways will hold specific organizational properties and the massive incorporation of alien genes are constrained to keep the original optimal pathway organization. Above all, we suggest that HGT, an important evolutionary source, may provide many opportunities for pathway improvement and innovation. However, in most cases, the process of integrating transferred genes into existing pathways is under strong selection pressure because of the co-evolution of tightly function-related host genes. The overall pathway organization structure will preserve well throughout evolution history.

Another interesting observation is that, in the 114 pathways significantly affected by HGT, operon structure may have an important influence on the integration of alien genes into existing pathways. The operons were identified by approaches similar to previous work [26,27]. When one operon has a significantly high proportion (more than 60%) of HT genes, it may possibly play a significant role in bringing the related HT genes into the corresponding pathway. Among the total 221 operons in 114 pathways, 77 (35%) have such characteristic. For comparison, random tests (100,000 replicates) were used by randomly picking genes with the same number of HT genes in each pathway. The expected value of random occurrence is 20 and the standard deviation is 2.72, which indicates that HT genes are more likely to be arranged in operon structure (significant level 0.001). Such results indicate that operon structure may play an essential and directional role in bringing HT genes into pathways. In fact, the integration of alien genes in operon form has its selective advantage, which can bring the recipient pathway new functional units and lighten the damage to original pathway organization at the same time.

To further understand the impact of HGT on pathway evolution, we also classified the 114 pathways according to the KEGG pathway categories to detect what type of pathways are more likely to integrate horizontally transferred genes. As expected, the most mobile pathways are related to the metabolism category (97 pathways), which is in good agreement with the observation that genes in the metabolic function group are more apt to be transferred. Such transfer may be preferred by the adaptive value of improving the metabolic capability in some pathways. There are also some pathways involved in genetic information processing including the protein secretion system (mainly the type III secretion system), aminoacyl tRNA biosynthesis and ribosome complex (14 pathways). The type III secretion system has been proved to be pathogenetic-related which is often the important part of the

pathogenetic island [28]. But it is surprising that some subunits of ribosome complex are also horizontally transferred. We speculated that this transfer occurred before the formation of translation machinery, that is, at the "progenotes time" as the "genetic annealing model" stated [24]. At that time, the horizontal gene transfer is pervasive, even in the primitive translation machine. The other pathways are involved in environmental information processing (4 pathways). All of those are related to the ABC transporters system: a hot spot of gene transfer as observed in many studies.

In this paper, we try to answer two basic questions about HGT. The clarification of these two issues may shed more light on the understanding of the behavior and mode of HGT in prokaryotic genome evolution. In the future, more accurate data such as the validated list of transferred genes and a more accurate pathway graph will certainly refine these results.

Acknowledgement

We thank Dr. Haixu TANG (Indiana University School of Informatics, Bloomington, USA) for assisting in the preparation of this manuscript.

References

1 Doolittle WF. Phylogenetic classification and the universal tree. Science 1999, 284: 2124–2129
2 Gogarten JP, Doolittle WF, Lawrence JG. Prokaryotic evolution in light of gene transfer. Mol Biol Evol 2002, 19: 2226–2238
3 Boucher Y, Douady CJ, Papke RT, Walsh DA, Boudreau ME, Nesbo CL, Case RJ *et al*. Lateral gene transfer and the origins of prokaryotic groups. Annu Rev Genet 2003, 37: 283–328
4 Nakamura Y, Itoh T, Matsuda H, Gojobori T. Biased biological functions of horizontally transferred genes in prokaryotic genomes. Nat Genet 2004, 36: 760–766
5 Lawrence JG, Ochman H. Amelioration of bacterial genomes: Rates of change and exchange. J Mol Evol 1997, 44: 383–397
6 Schmidt S, Sunyaev S, Bork P, Dandekar T. Metabolites: A helping hand for pathway evolution? Trends Biochem Sci 2003, 28: 336–341
7 Jain R, Rivera MC, Moore JE, Lake JA. Horizontal gene transfer in microbial genome evolution. Theor Popul Biol 2002, 61: 489–495
8 Koonin EV, Makarova KS, Aravind L. Horizontal gene transfer in prokaryotes: Quantification and classification. Annu Rev Microbiol 2001, 55: 709–742
9 Kurland CG, Canback B, Berg OG. Horizontal gene transfer: A critical view. Proc Natl Acad Sci USA 2003, 100: 9658–9662
10 Hong SH, Kim TY, Lee SY. Phylogenetic analysis based on genome-scale metabolic pathway reaction content. Appl Microbiol Biotechnol 2004, 65: 203–210
11 Eisen JA. Horizontal gene transfer among microbial genomes: new insights from complete genome analysis. Curr Opin Genet Dev 2000, 10: 606–611
12 Ragan MA. On surrogate methods for detecting lateral gene transfer. FEMS Microbiol Lett 2001, 201: 187–191
13 Lawrence JG, Ochman H. Reconciling the many faces of lateral gene transfer. Trends Microbiol 2002, 10: 1–4
14 Koski LB, Golding GB. The closest BLAST hit is often not the nearest neighbor. J Mol Evol 2001, 52: 540–542
15 Salzberg SL, White O, Peterson J, Eisen JA. Microbial genes in the human genome: Lateral transfer or gene loss? Science 2001, 292: 1903–1906
16 Altschul SF, Madden TL, Schaffer AA, Zhang J, Zhang Z, Miller W, Lipman DJ. Gapped BLAST and PSI-BLAST: A new generation of protein database search programs. Nucleic Acids Res 1997, 25: 3389–3402
17 Fitch WM. Homology a personal view on some of the problems. Trends Genet 2000, 16: 227–231
18 Frickey T, Lupas AN. PhyloGenie: Automated phylome generation and analysis. Nucleic Acids Res 2004, 32: 5231–5238
19 Garcia-Vallve S, Guzman E, Montero MA, Romeu A. HGT-DB: A database of putative horizontally transferred genes in prokaryotic complete genomes. Nucleic Acids Res 2003, 31: 187–189
20 Tsirigos A, Rigoutsos I. A new computational method for the detection of horizontal gene transfer events. Nucleic Acids Res 2005, 33: 922–933
21 Dufraigne C, Fertil B, Lespinats S, Giron A, Deschavanne P. Detection and characterization of horizontal transfers in prokaryotes using genomic signature. Nucleic Acids Res 2005, 33: e6
22 Chang CF, Wai KM, Patterton HG. Calculating the statistical significance of physical clusters of co-regulated genes in the genome: The role of chromatin in domain-wide gene regulation. Nucleic Acids Res 2004, 32: 1798–1807
23 Peterson JD, Umayam LA, Dickinson T, Hickey EK, White O. The comprehensive microbial resource. Nucleic Acids Res 2001, 29: 123–125
24 Woese C. The universal ancestor. Proc Natl Acad Sci USA 1998, 95: 6854–6859
25 Lawrence JG. Gene transfer, speciation, and the evolution of bacterial genomes. Curr Opin Microbiol 1999, 2: 519–523
26 Strong M, Mallick P, Pellegrini M, Thompson MJ, Eisenberg D. Inference of protein function and protein linkages in *Mycobacterium tuberculosis* based on prokaryotic genome organization: A combined computational approach. Genome Biol 2003, 4: R59
27 Rogozin IB, Makarova KS, Murvai J, Czabarka E, Wolf YI, Tatusov RL, Szekely LA *et al*. Connected gene neighborhoods in prokaryotic genomes. Nucleic Acids Res 2002, 30: 2212–2223

Edited by
Jun YU