Software review

Detecting horizontal gene transfer with T-REX and RHOM programs

Abstract

As the Human Genome Project and other genome projects experience remarkable success and a flood of biological data is produced by means of high-throughout sequencing techniques, detection of horizontal gene transfer (HGT) becomes a promising field in bioinformatics. This review describes two freeware programs: T-REX for MS Windows and RHOM for Linux. T-REX is a graphical user interface program that offers functions to reconstruct the HGT network among the donor and receptor hosts from the gene and species distance matrices. RHOM is a set of command-line driven programs used to detect HGT in genomes. While T-REX impresses with a user-friendly interface and drawing of the reticulation network, the strength of RHOM is an extensive statistical framework of genome and the graphical display of the estimated sequence position probabilities for the candidate horizontally transferred genes.

Keywords: *horizontal gene transfer, lateral gene transfer, phylogeny, freeware*

INTRODUCTION

As the Human Genome Project and other genome projects experience remarkable success and a flood of biological data is produced by means of high-throughout sequencing techniques, detection of horizontal gene transfer (HGT) or lateral gene transfer (LGT) has become a very 'hot' field in bioinformatics. HGT is defined as the interspecific transfer of genes via routes other than sexual contact and hybridisation (eg transformation, transduction or conjugation).[1] Syvanen and Clarence suggested by that the term 'horizontal gene transfer' represents the transfer of gene across distinct species between kingdoms and the term 'lateral gene transfer' describes the gene transfer between distinct species within a kingdom. [2] In this review, 'horizontal gene transfer' or HGT is adopted to indicate the meanings of both terms.

Since Akiba, Ochiai and coworkers found that antibiotic-resistant plasmids could move among different bacterial species and spread drug-resistant genes, it has been widely accepted and recognised that HGT can also influence the evolution of plant and animal kingdoms.[3,4] During the past two decades, with the development of high-throughput sequencing technique and computer-assisted genome annotation system, a vast amount of sequence data has been produced. Through the comparisons of nucleotide and amino acid sequences between genomes, many unusual sequence conservations have been found and can be explained by HGT events.[2]

A number of quantitative models and sequence-based algorithms for detecting HGT have been developed. Two types of algorithms are commonly used: supervised and unsupervised analyses. A supervised analysis detects the possible HGT events by comparing a gene tree with its corresponding species tree. There are several ways to implement this strategy: anomalous phylogenetic distribution, incongruent trees and abnormal sequence similarity (ie greatest similarity with a gene from a distantly

related species based on BLAST or FastA results).[5] Biologists prefer anomalous phylogenetic distribution and incongruent tree methods because they can solve the gene tree/species tree problems and detect the donor and acceptor species of the horizontally transferred genes (HTGs). For example, Bergthorsson *et al.* detected a wide spread of mitochondrial HGT among flowering plants.[6] Moreover, Mower *et al.* described two cases of HGT from parasitic flowering plants to corresponding host flowering plants. [7] Not until recently had the implementation of the supervised analysis become a main workflow style via combining several programs. Some of the tasks in the workflow can now be carried out using the program T-REX developed by Makarenkov,[8,9] including multiple sequence alignment, phylogenetic reconstruction, calculation of distance matrices and detection of HGT events.

In an unsupervised analysis, DNA or protein sequence information has been used in various ways to identify genomic features, indicating that the evolutionary history of genes within a particular region differs from that of most genes in a particular genome vertically transmitted.[10] These genes are recognised as the candidate HTGs. However, depending on the evolutionary models, different algorithms have been brought forward. For example, it was estimated that about 17 per cent of the *Escherichia coli* genome is obtained by HGT events based on the analysis of GC content and codon usage bias indexes of opening reading frames (ORFs).[11,12] Karlin and Burge came up with a genome signature based on the dinucleotide composition as an index of genomic features.[13] Recently, the distance-based measure between genes and genome signatures has also been used to detect HGT.[14] Nakamura *et al.* estimated the average posterior probability (horizontal transfer index) using Bayesian method to separate inherent and exogenous genes.[15] When Markov models trained with typical ORFs are used to identify genes in a genome, the genes that cannot be recognised by the models may be candidate HTGs.[16] A new statistical segmentation method based on a hidden Markov model (HMM) has been used to detect the bacterial chromosome heterogeneities, including HTGs. On the basis of this method, a set of programs (called RHOM) were developed by Nicolas, and its C++ sources can be downloaded freely.[17,18]

INSTALLATION AND COMPILATION OF T-REX AND RHOM

The T-REX program including its C++ sources files can be downloaded freely from the website of the author Dr Makarenkov. Currently, there are Windows, Macintosh and DOS versions, but the DOS and Macintosh versions do not contain the modules for HGT reconstruction. The Windows version is updated frequently (the latest version 4.01 is from January, 2005), and it is applicable to various popular versions of Windows Operating System (OS) (Windows 9x/ME/NT/2000/XP). The downloaded file is a zipped file (7.63 M), including an executable file 'installer.exe', which can guide the user to complete the installation of the T-REX program. This program also includes several important methods for constructing phylogenetic trees.

RHOM has no binary executable files for download. However, its C++ sources are freely available for UNIX/Linux at the website of the Institut National De La Recherche Agronomique (INRA) in France and is easily compiled.[18] In the Linux operating system (Red Hat release 9), the command 'make' will create an executable file named 'rhom.em' in the parent directory. However, the rhom.em takes a long time to analyse a bacterial genome. It requires ~40 h and 256 MB of active memory on a PC Intel Celeron 2.0 GHz to analyse a genome of 4.2 MB. In order to display the RHOM results in a graph, the Seq.grap program is necessary, which can be obtained by compiling its

C++ sources according to the README file.[19]

After installation and compilation, users can run T-REX through the menus with the graphical interface. The specific parameters must be given in the command-line for RHOM.

T-REX AND RHOM IN DETAIL

The main functions of T-REX are to reconstruct a phylogenetic tree, a reticulogram and an HGT reticulogram. There is a great difference between construction of the HGT reticulogram and other functions in input files and procedures.

A T-REX analysis can detect the HGT events based on two matrices simultaneously, ie a gene distance matrix and a species distance matrix. The incongruence between the two matrices may be attributed to HGT, and thus could be used to detect the HGT events.[9] A flowchart of the procedure is shown in Figure 1. When both distance matrices have been loaded into the program, an optimisation criterion based on the least squares or the Robinson–Foulds tree topological distance is recommended for HGT reconstruction. After the analysis, the possible HGT events can be viewed in one window or in different windows according to the user's option. In addition, the priority of these HGT events based on their likelihood can also be numbered on the tree. Figure 2 shows an example of the horizontal gene transfer of *rbc*L gene in eubacteria and plasmids. The gene and species distance matrices are provided by the authors with the T-REX program and built from the *rbc*L and 16S rRNA sequences, respectively.

RHOM is an HMM-based freeware that uses an expectation maximisation (EM) algorithm to give the maximum likelihood estimation of the parameters. There are two folders (SOURCES and TestRHOM) in the downloaded zipped file. The C++ sources are in the SOURCES folder and the examples in the TestRHOM folder could be used to evaluate the program.

The binary executable file rhom.em has four options: the option '−h' is used to display the syntax of program 'rhom.em', and the result of which is 'Usage: . /rhom.em −model <model_desc>-seq <seq_list>-em <em_param>'; other three options are '−model', '-sq' and '−em' followed by a filename for each, as described below.

(1) The option '-model' is used to specify a model description file (eg the file 'model.desc' in the TestRHOM folder). There are three lines in that file corresponding to three parameters: model, type and nstate. The user must specify the 'model' and 'nstate' for their data sets. Parameter 'model' defines the dimensions of the model ranging from 0 to 3; 'nstate' defines the number of hidden states ranging from 2 to 8. In order to detect HGT, the parameter 'model' is specified as two and the 'nstate' as three.[17]

(2) The option '-seq' is used to specify a list file (eg the file 'fic_seq.txt' in the TestRHOM folder). There is no sequence in the list file, only a list of the paths and names of sequence files. If there are no more paths in front of the filename of the sequence files, the default of the sequence file and the list file in the program should be in the same folder. As to the format of the sequence file, the FastA format, Genbank format, EMBL format, and a self-defined format are all available in RHOM. The self-defined format is similar to the Fasta format, which begins with the name of the sequence, and ends with '#' in the title line, then is followed by lines of sequence data. Note that LOCUS and ORIGIN have been considered in Genbank format, and ID and SQ in EMBL format only.

(3) The option '-em' is used to specify the EM parameter file and to give the initial values of EM. In RHOM, the initial values which are necessary to run EM can be given in two ways: the user can specify them, or a multiple random initial value can be used and then the optimal value will be estimated by the

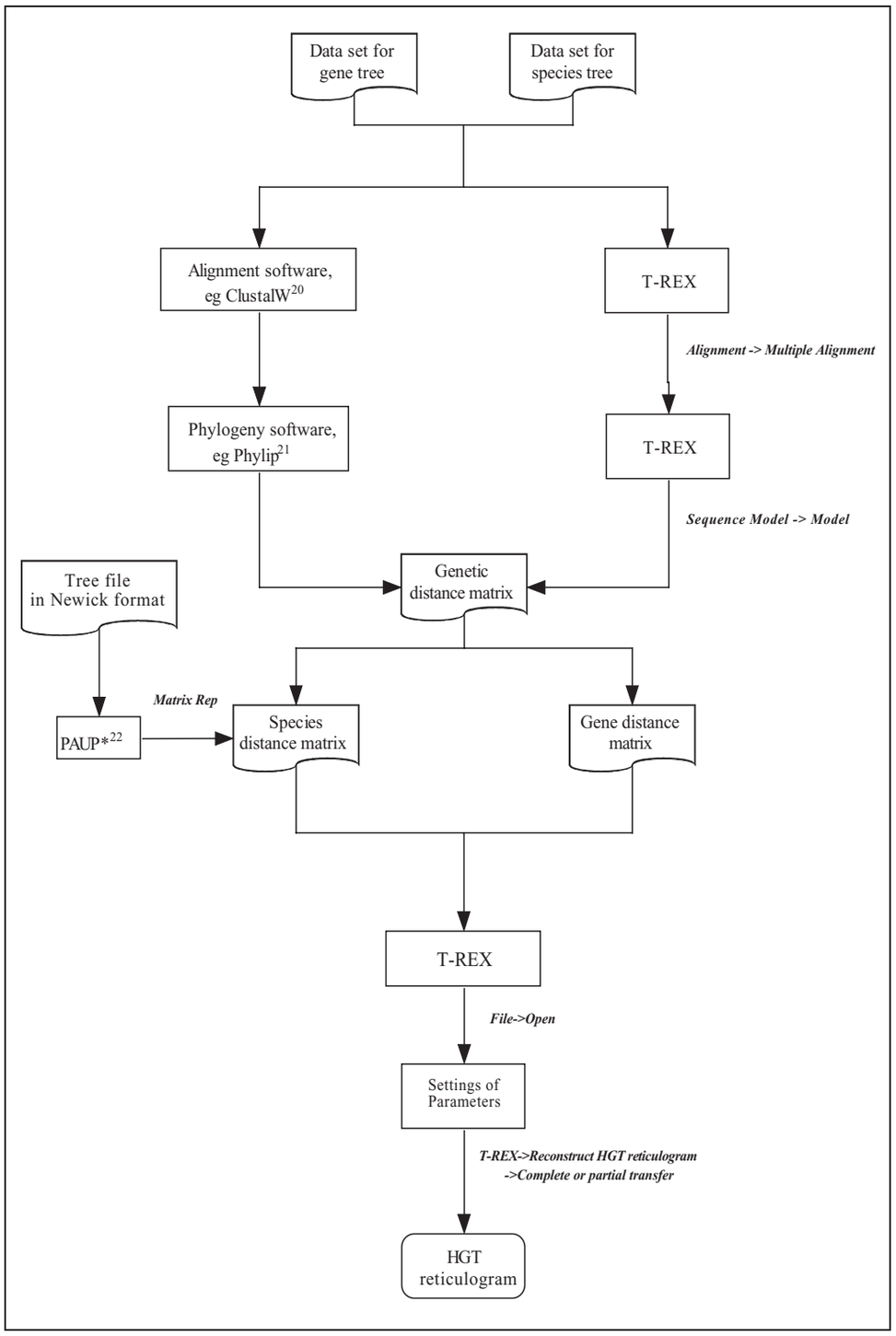

Figure 1: A workflow chart of reticulogram construction for detecting HGT with the T-REX programs

program. Most parameters defined by the two ways are different, but two, 'niter' and 'eps', are common. The parameter 'niter' is an integer and is defined as the maximum value of EM; 'eps' is a real number and is used to define the terminal standard. It is recommended to give random initial values for EM and accept the default settings specified by the program for a new user.

As an example the complete genome of *Bacillus subtilis* is analysed with RHOM.

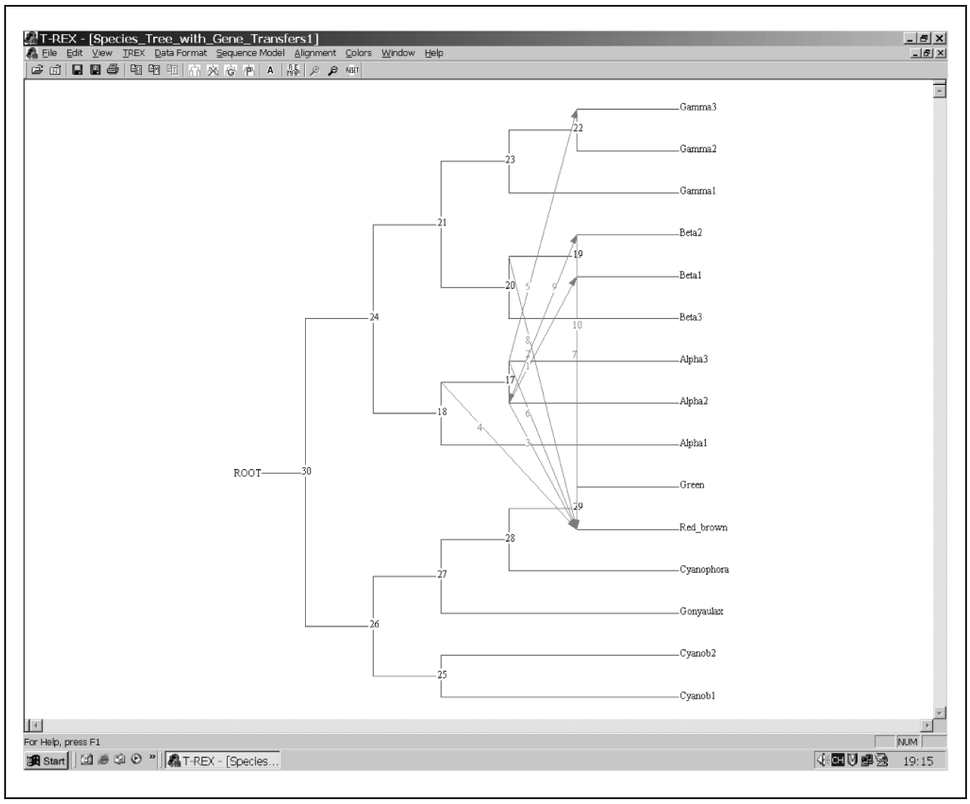

Figure 2: A scanshot of T-REX. The optimisation criterion based on the Robinson–Foulds tree topological distance is used to detect HGT events. Numbers represent the priority of the HGT events based on their likelihood. Thus, the transfer between alpha2 and beta1 proteobacteria is the most significant among the ten detected events. Arrows (→) in the HGT reticulogram graph indicate the directions of the HGT events

Its annotation file is retrieved from NCBI (NC_000964). Although the result file named NC_000964.gbk.rhom is about 175 MB in size and is difficult to be displayed by a standard software, Seq.graph program can transfer the results into a GIF-format file as shown in Figure 3. The probability of each position and annotation of the *B. subtilis* are marked in one graph. The *PBSX* prophage sequences obtained from the bacteriophage are indicated in blue lines.

USING THE PROGRAMS

The main tasks of the two T-REX and RHOM programs are different; nevertheless, they are complementary in detecting HGT events. Detection of HGT plays an important role in characterising the phylogeny and genome-wide evolution of bacteria, and adaptation of the metabolism pathway to the environment. Using RHOM to detect HGT does not require any prior knowledge of a particular microbial genome apart from the HMM parameters. Therefore, it is appropriate for analysing a newly sequenced genome of a microorganism of uncertain phylogenetic relationship.

However, the genome-based methods focus only on whether or not a gene is an HTG; they cannot distinguish which genome it is. The genes that transferred into the receptor genome would evolve by a process called ingression. In some cases, a HTG would replace the original gene and thus cause the loss of the original gene. This process is called xenologous gene displacement.[23] The evolutionary history of these xenologous genes could only be reconstructed using some specific software like T-REX.

After the annotation of a newly sequenced genome, RHOM can be used to describe the genome in detail with a statistical frame. Although the analysis is time-consuming, it is acceptable for single genome analysis. Seq.graph program is then used to handle the RHOM results and transfer them to a GIF graph, which describes the typical and atypical regions

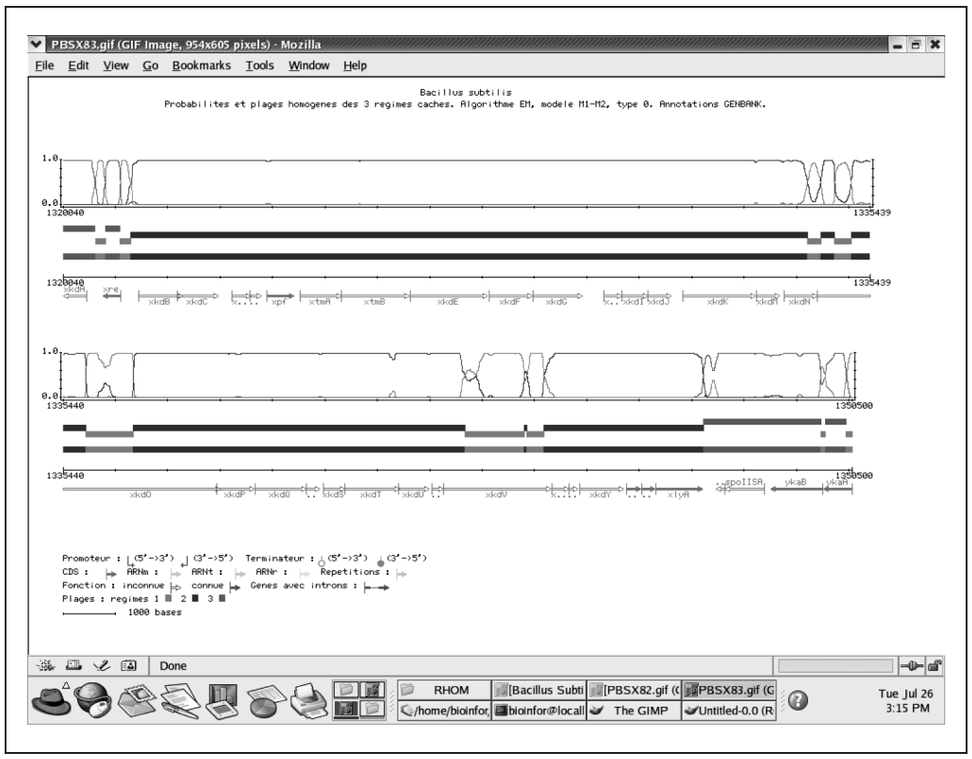

Figure 3: A scanshot of the GIF file created by seq.graph based on the results of the RHOM. A detection of an atypical segment using a three-state *M1–M2* model (1320–1350 kb), containing the *PBSX* prophage. Most of the *PBSX* prophage sequence is marked with blue lines. Red and green lines indicate other gene classes

of the genome. According to annotations of the genome, the sequences of the candidate HTGs are retrieved and BLAST or FastA is used to detect the orthologous genes for each HTG. Then, T-REX will be used to detect HGT events. Because of the large number of candidate HTGs, an integrated workflow may be a practical procedure for the analysis.

DISCUSSION

T-REX is a user-friendly graphical program and most operations and parameter settings can be chosen through menus. The speed of detecting HGTs is fast and the display of the results is self-explanatory. T-REX is used to solve the incongruent phylogenetic tree problems. However, this program does not provide methods for constructing phylogenetic trees, such as maximum parsimony, maximum likelihood and Bayesian methods. In T-REX, the format of the output file is not plain text (ASCII) and the record results are stored in an application-specific format. For example, the tree file created by T-REX with an extension name (.TRE) cannot be read by other tree-displaying programs such as TreeView.[24]

RHOM provides an extensive statistical framework of genome and does not need any prior knowledge to detect HGT events. For example, the ability of RHOM to detect three HTGs that have been detected experimentally in *B. subtilis* provides a biological validation of the program.[17] The results of RHOM are handled by seq.graph and a graph of the genome is produced. Valuable predictions can be made by use of the graph and the annotation of the genome. However, RHOM is not easy for users to operate and it also needs a great deal of time to deal with the calculations based on complicated modes.

The common feature of the two programs is to present results in a diagram. Biologists prefer this procedure because it is brief, but comprehensive. Although the two programs implement both supervised and unsupervised analyses, they still cannot meet all the need of detection of HGT. These programs can detect those HGT events that occurred recently, whereas

genes that have been transferred into a genome for a long time or for which the DNA composition was originally close to the host cannot be detected. Furthermore, false positive results of HGT are always inevitable. To solve this problem, more accurate models, algorithms and independent experimental approaches that can circumvent PCR are needed.[25,26]

Since nucleotide and protein sequences were first used to detect HTGs, more than ten algorithms have been elaborated.[9] Unfortunately, few authors have distributed their programs freely and even in most available programs, such as T-REX and RHOM, the function to detect HGT is just one part of their tasks. The supervised and unsupervised analyses have not been implemented in a single program. After scanning a genome, unsupervised methods will be used to analyse the evolutionary history of the HTGs and the evolution of metabolic pathway extensively. Therefore, a multiple-functional, multiple-platform and user-friendly freeware to detect HGT is still urgently needed.

Acknowledgments

We would like to thank the Bioinformatics Working Group in School of Life Sciences of Fudan University for useful discussion and Mr Guillaume Neault of Queen's University for critical reading of the manuscript. This work was supported in part by the National Basic Research Project (973) (2003CB715900) and High-Tech Project (863) (2002AA23104).

Zuofeng Li, Li Wang and Yang Zhong
School of Life Sciences,
Fudan University,
Shanghai 200433, China
and Shanghai Center for Bioinformation Technology,
Shanghai 201203, China
Tel: 86 21 55664436
E-mail: yangzhong@fudan.edu.cn

References

1. Mallet, J. (2005), 'Hybridization as an invasion of the genome', *Trends Ecol. Evol.*, Vol. 20, pp. 229–237.
2. Syvanen, M. and Clarence, I. K. (2002), 'Horizontal Gene Transfer', Academic Press, London.
3. Akiba, T., Koyama, K., Ishiki, Y. *et al.* (1960), 'On the mechanism of the development of multiple-drug-resistant clones of *Shigella*', *Jpn J. Microbiol.*, Vol. 4, pp. 219–227.
4. Ochiai, K., Yamanaka, T., Kimura, K. *et al.* (1959), 'Studies on the inheritance of drug resistance between *Shigella* strains and *Escherichia coli* strains', *Jpn Med. J.*, Vol. 1861, pp. 34–46.
5. Philippe, H. and Christophe, J. D. (2003), 'Horizontal gene transfer and phylogenetics', *Curr. Opin. Microbiol.*, Vol. 6, pp. 498–505.
6. Bergthorsson, U., Adams, K. L., Thomason, B. *et al.* (2003), 'Widespread horizontal transfer of mitochondrial genes in flowering plants', *Nature*, Vol. 424, pp. 197–201.
7. Mower, J. P., Stefanovic, S., Young, G. J. *et al.* (2004), 'Gene transfer from parasitic to host plants', *Nature*, Vol. 432, pp. 165–166.
8. T-REX (URL: http://www.labunix.uqam.ca/~makarenv/trex.html).
9. Makarenkov, V. (2001), 'T-REX: Reconstructing and visualizing phylogenetic trees and reticulation networks', *Bioinformatics*, Vol. 17, pp. 664–668.
10. Ochman, H., Lawrence, J. G. and Groisman, E. A. (2000), 'Lateral gene transfer and the nature of bacterial innovation', *Nature*, Vol. 405, pp. 299–304.
11. Lawrence, J. G. and Ochman, H. (1997), 'Amelioration of bacterial genomes: Rates of changes and exchange', *J. Mol. Evol.*, Vol. 44, pp. 383–397.
12. Koski, L.B., Morton, R.A. and Golding, G. B. (2001), 'Codon bias and base composition are poor indicators of horizontally transferred genes', *Mol. Biol. Evol.*, Vol. 18, pp. 404–412.
13. Karlin, S. and Burge, C. (1995), 'Dinucleotide relative abundance extremes: A genomic signature', *Trends Genet.*, Vol. 11, pp. 283–290.
14. Dufraigne, C., Fertil, B., Lespinats, S. *et al.* (2005), 'Detection and characterization of horizontal transfers in prokaryotes using genomic signature', *Nucleic Acids Res.*, Vol. 33, p. e6.
15. Nakamura, Y., Itoh, T., Matsuda, H. *et al.* (2004), 'Biased biological functions of horizontally transferred genes in prokaryotic genomes', *Nature Genet.*, Vol. 36, pp. 760–766.
16. Ragan, M. A. (2001), 'On surrogate methods for detecting lateral gene transfer', *FEMS Microbiol. Lett.*, Vol. 201, pp. 187–191.
17. Nicolas, P., Bize, L., Muri, F. *et al.* (2002),

'Mining *Bacillus subtilis* chromosome heterogeneities using hidden Markov models', *Nucleic Acids Res.*, Vol. 30, pp. 1418–1426.

18. RHOM (URL: http://ssb2.jouy.inra.fr/ssb/rhom/).
19. SEQ.GRAPH (URL: http://ssb2.jouy.inra.fr/ssb/software/seq.graph/availability.html).
20. Thompson, J. D., Gibson, T. J., Plewniak, F. *et al.* (1997), 'The Clustal-X windows interface: Flexible strategies for multiple sequence alignment aided by quality analysis tools', *Nucleic Acids Res.*, Vol. 24, pp. 4876–4882.
21. PHYLIP (URL: http://evolution.genetics.washington.edu/phylip.html).
22. Swofford, D. L. (1998), 'PAUP*. Phylogenetic analysis using parsimony (*and other methods)', Version 4, Sinauer Associates, Sunderland, MA.
23. Koonin, E. V., Makarova, K.S. and Aravind, L. (2001), 'Horizontal gene transfer in prokaryotes: quantification and classification', *Annu. Rev. Microbiol.*, Vol. 55, pp. 709–742.
24. Page, R. D. M. (1996), 'TREEVIEW: An application to display phylogenetic trees on personal computers', *Comput. Appl. Biosci.*, Vol. 12, pp. 357–358.
25. Ragan, M. A. (2001), 'Detection of lateral gene transfer among microbial genomes', *Curr. Opin. Genet. Dev.*, Vol. 11, pp. 620–626.
26. Martin, W. (2005), 'Lateral gene transfer and other possibilities', *Heredity*, Vol. 94, pp. 565–566.