BIOLOGY REPORTS

Published: 24 March 2009

Bioinformatic detection of horizontally transferred DNA in bacterial genomes

Morgan GI Langille and Fiona SL Brinkman*

Address: Department of Molecular Biology and Biochemistry, Simon Fraser University, Burnaby, BC, Canada V5A 1S6

*Corresponding author: Fiona SL Brinkman (brinkman@sfu.ca)

F1000 Biology Reports 2009,**1**:25 (doi: 10.3410/B1-25)

Abstract

We highlight a selection of recent research on computational methods and associated challenges surrounding the prediction of bacterial horizontal gene transfer. This research area continues to face controversy, but is becoming more critical as the importance of horizontal gene transfer in medically and ecologically important prokaryotic evolution is further appreciated.

Introduction and context

Horizontal gene transfer (HGT) is an important driving force in prokaryotic evolution that allows bacteria to quickly share genes not only from similar strains, but also from distantly related species, including phages [1]. This enables bacteria to adapt to changing environmental pressures, but can also lead to problems with treating bacterial illnesses, due to the exchange of antibiotic resistance genes or virulence factors [2].

Although HGT has been shown to be widespread across bacterial strains, the rate of HGT is still debated. Some argue that HGT is so prevalent among bacteria that the ability to reconstruct a tree of life should be seriously reconsidered [3], while other recent research indicates that HGT may not be very prevalent [4]. Overall, methods for the prediction of HGT in bacteria continue to improve, and many would agree that construction of species trees, despite the prevalence of HGT, is worthwhile if appropriate methodologies are applied [5].

Most HGT computational prediction methods can be roughly grouped into two main categories: compositional methods, which identify anomalous sequence signatures within a prokaryotic genome suggestive of a region of HGT, and phylogenetic methods, which analyse the incongruence of a gene tree versus its associated species tree. We briefly highlight here some of the recent improvements in such methodologies and the challenges still faced.

Major recent advances

Sequence composition methods

Sequence composition methods depend on different species having differences in their genome signatures. These methods identify HGT by searching for genomic regions that have an abnormal sequence composition (G+C, dinucleotide bias, and so on) compared to the rest of the genome.

An in-depth study of HGT in the *Salmonella* lineage indicated that ancient horizontally transferred gene sequences tended to share a greater similarity in sequence composition with their host compared to more recently acquired genes [6], clearly supporting the idea that transferred genes ameliorate to their host genome over time [7]. Notably, however, very recently acquired prophage elements tended to have sequence compositions that were more similar to the host genome, not representing amelioration but rather specialization and adaptation to their hosts [6]. Although this study may suggest that more sensitive measures of sequence composition are needed to better predict HGT events, these methods must be carefully designed so that they do not result in an increase in false positives. For example, a recent study of large viruses

further supported previous work indicating that many genes with atypical sequence composition were not horizontally acquired and, instead, the anomalous sequence composition was likely related to certain functions and gene features such as expression level [8]. A recent comparison of several HGT prediction programs showed that these sequence-composition-based methods can predict very different classes of genes, warning that the use of a single method could give biased results [9]. Even with these disadvantages, detecting HGT by sequence composition is still an attractive method, since it usually does not require more than the query genome for analysis.

New research is also producing more intelligent methods; one such method takes into account that single transfer events often include multiple genes and uses the genome location of putative HGTs to further refine predictions [10]. Similarly, there have been many methods focused on the prediction of genomic islands (large regions of HGT) and the accuracy of such genomic island predictors has been increased through the coupling of sequence composition analysis with the identification of additional gene features such as the presence of mobility genes (e.g. integrases and transposases) or tRNAs and direct repeats (known integration sites) [11–15]. This research direction is likely to continue as more sophisticated composition-based methods are developed that also examine other sequence features.

Phylogenetic methods

Many HGT prediction methods look for incongruence between gene trees and an associated species tree. Such methods could increasingly benefit from having a more universal 'tree of life' to use as a species tree reference. One notable study attempted to build such a tree by identifying genes that were present in all species that did not show signs of HGT [16]. Identifying genes that have never been horizontally transferred is a difficult problem that remains controversial, however, some studies have suggested that particular genes could be more resistant to HGT and could therefore be better candidates for construction of a reference tree [17]. However, genes subject to HGT can still provide valuable phylogenetic information, and one study actually embraced using HGTs for tree construction, demonstrating that ancient gene transfers to the ancestor of red algae and green plants can act as informative events that support a common origin of these two groups [18]. Another method that tries to construct a large genome tree, using a selected list of genes that are shared across most genomes, is AMPHORA (a pipeline for AutoMated PHylogenOmic inference) [19]. An automated pipeline was developed that uses 31 'marker' genes, a hidden Markov model (HMM)-based multiple alignment program, and maximum likelihood to construct an organism tree for 578 species. The construction of these large trees is likely to lead to new insights and aid other analyses, but it is appreciated that they do not fully reflect bacterial evolution due to their lack of representation of HGTs. Therefore, complementing these approaches are phylogenetic methods that incorporate or predict HGT events [20–23]. These tools allow for reticulate evolutionary events, such as HGT, and result in a network-like phylogenetic tree that is often represented as a rooted, directed, acyclic graph; this is the same structure that is used by the Gene Ontology project [24]. A software package called PhyloNet [23] was recently published and includes many tools to carry out prediction of HGT and tree construction that should be useful for many researchers. It makes use of a recently created eNewick ('extended Newick') format for containing network-like trees [25], which is based on the well established, classic Newick format.

Despite these promising advances, limitations of phylogenetic based HGT prediction methods still exist that must be considered; transfers between sister branches in a tree (often very closely related species) can't usually be detected, and sparsely distributed genes may not be detected if the gene tree is consistent (or inconclusive) with the species tree. Future research is likely to try to overcome or at least minimize these limitations, either through increased species sampling or by combining the power of phylogenetic and sequence composition based approaches.

Metagenomics

Prediction of HGTs in metagenomic datasets is somewhat limited due to the novelty of this type of genomic data, the fact that the organism sources of the sequences are unknown, and use of short sequence read lengths that can lower the statistical power of HGT prediction methods. However, one recent study has designed novel composition and phylogenetic methods for the detection of HGT in several environmental samples [26]. They show that their composition method and phylogenetic method detect different levels of HGT at 0.8–1.5% and 2–8%, respectively. The authors note that these differences are likely to be due to the types of HGT detected by each method, illustrating just how far we still have to go in HGT prediction and the significant potential there is to improve HGT prediction by integrating approaches.

Future directions

Regions of HGT are being repeatedly found to contain virulence genes or other genes of medical and/or

ecological importance, so improved prediction of such regions from primary sequence data will continue to be of significant interest. Considering that science is still working out the details of how genes are transferred by conjugation [27], and we are unsure of the origins of most regions of predicted HGT, we should not be surprised that prediction of HGT still has a long way to go. New computational methods are likely to be developed that improve on algorithm design by inclusion of new biological insights gained from increased sampling of our genetic world, or by better statistical modelling. The role of phages and other vehicles of HGT, in particular, may help shape some predictive methods [28]. Prediction of the more specific boundaries of regions of HGT is one research area that needs more focus. More accurate bioinformatic methods are becoming even more important now, and should be a major goal, as the number of completed microbial genomes increases dramatically and the number of sequences from metagenomic studies and next-generation sequencing eclipses all other sequence data combined. Research that provides unbiased analysis and reviews of the accuracy of HGT methods should be encouraged so that researchers can utilize those methods that work best for their data (akin to what has been done for phylogenomics methods [29] and genomic island prediction methods [12]). As sequence coverage of our genetic world continues to grow and HGT prediction methods continue to improve, hopefully the origins of many HGT events will become clearer, and we will better understand these events that have played such a pivotal role in bacterial adaptation.

Abbreviations

HGT, horizontal gene transfer.

Competing interests

The authors declare that they have no competing interests.

Acknowledgements

We would like to acknowledge the reviewers of this article, including Dr Robert Beiko, who waived his anonymity and contributed useful suggestions.

References

1. Koonin EV, Wolf YI: **Genomics of bacteria and archaea: the emerging dynamic view of the prokaryotic world.** *Nucleic Acids Res* 2008, **36:**6688-719.

2. Wright GD: **The antibiotic resistome: the nexus of chemical and genetic diversity.** *Nat Rev Microbiol* 2007, **5:**175-86.

3. Doolittle WF, Bapteste E: **Pattern pluralism and the Tree of Life hypothesis.** *Proc Natl Acad Sci U S A* 2007, **104:**2043-9.

4. Choi I, Kim S: **Global extent of horizontal gene transfer.** *Proc Natl Acad Sci U S A* 2007, **104:**4489-94.

   F1000 Factor 6.0 *Must Read*
   Evaluated by John Jaenike 16 Mar 2007

5. Galtier N, Daubin V: **Dealing with incongruence in phylogenomic analyses.** *Philos Trans R Soc Lond B Biol Sci* 2008 [Epub ahead of print].

   F1000 Factor 3.0 *Recommended*
   Evaluated by Nicola Mulder 28 Oct 2008

6. Vernikos GS, Thomson NR, Parkhill J: **Genetic flux over time in the *Salmonella* lineage.** *Genome Biol* 2007, **8:**R100.

7. Lawrence JG, Ochman H: **Amelioration of bacterial genomes: rates of change and exchange.** *J Mol Evol* 1997, **44:**383-97.

8. Monier A, Claverie J, Ogata H: **Horizontal gene transfer and nucleotide compositional anomaly in large DNA viruses.** *BMC Genomics* 2007, **8:**456.

   F1000 Factor 3.0 *Recommended*
   Evaluated by Arcady Mushegian 09 Apr 2008

9. Ragan MA, Harlow TJ, Beiko RG: **Do different surrogate methods detect lateral genetic transfer events of different relative ages?** *Trends Microbiol* 2006, **14:**4-8.

10. Azad RK, Lawrence JG: **Detecting laterally transferred genes: use of entropic clustering methods and genome position.** *Nucleic Acids Res* 2007, **35:**4629-39.

11. Merkl R: **SIGI: score-based identification of genomic islands.** *BMC Bioinformatics* 2004, **5:**22.

12. Langille MGI, Hsiao WWL, Brinkman FSL: **Evaluation of genomic island predictors using a comparative genomics approach.** *BMC Bioinformatics* 2008, **9:**329.

13. Rajan I, Aravamuthan S, Mande SS: **Identification of compositionally distinct regions in genomes using the centroid method.** *Bioinformatics* 2007, **23:**2672-7.

14. Vernikos GS, Parkhill J: **Interpolated variable order motifs for identification of horizontally acquired DNA: revisiting the *Salmonella* pathogenicity islands.** *Bioinformatics* 2006, **22:**2196-203.

15. Chatterjee R, Chaudhuri K, Chaudhuri P: **On detection and assessment of statistical significance of genomic islands.** *BMC Genomics* 2008, **9:**150.

16. Ciccarelli FD, Doerks T, von Mering C, Creevey CJ, Snel B, Bork P: **Toward automatic reconstruction of a highly resolved tree of life.** *Science* 2006, **311:**1283-7.

   F1000 Factor 8.1 *Exceptional*
   Evaluated by John Jaenike 07 Mar 2006, Michael Wagner 09 Mar 2006, Fiona Brinkman 08 Jun 2006

17. Sorek R, Zhu Y, Creevey C, Francino M, Bork P, Rubin E: **Genome-wide experimental determination of barriers to horizontal gene transfer.** *Science* 2007, **318:**1449-52.

   F1000 Factor 4.8 *Must Read*
   Evaluated by William Martin 21 Dec 2007, Julian Parkhill 15 Jan 2008

18. Huang J, Gogarten J: **Ancient horizontal gene transfer can benefit phylogenetic reconstruction.** *Trends Genet* 2006, **22:**361-6.

   F1000 Factor 3.0 *Recommended*
   Evaluated by Nicolas Galtier 07 Nov 2006

19. Wu M, Eisen JA: **A simple, fast, and accurate method of phylogenomic inference.** *Genome Biol* 2008, **9:**R151.

   F1000 Factor 3.2 *Recommended*
   Evaluated by Yuri Wolf 30 Jan 2009, Jacques Ravel 05 Feb 2009

20. Makarenkov V: **T-REX: reconstructing and visualizing phylogenetic trees and reticulation networks.** *Bioinformatics* 2001, **17:**664-8.

21. MacLeod D, Charlebois RL, Doolittle F, Bapteste E: **Deduction of probable events of lateral gene transfer through comparison of phylogenetic trees by recursive consolidation and rearrangement.** *BMC Evol Biol* 2005, **5:**27.

F1000 Factor 3.0 *Recommended*
Evaluated by Jeffrey Lawrence 13 Apr 2005

22. Beiko RG, Hamilton N: **Phylogenetic identification of lateral genetic transfer events.** *BMC Evol Biol* 2006, **6:**15.

23. Than C, Ruths D, Nakhleh L: **PhyloNet: a software package for analyzing and reconstructing reticulate evolutionary relationships.** *BMC Bioinformatics* 2008, **9:**322.

24. Ashburner M, Ball CA, Blake JA, Botstein D, Butler H, Cherry JM, Davis AP, Dolinski K, Dwight SS, Eppig JT, Harris MA, Hill DP, Issel-Tarver L, Kasarskis A, Lewis S, Matese JC, Richardson JE, Ringwald M, Rubin GM, Sherlock G: **Gene ontology: tool for the unification of biology. The Gene Ontology Consortium.** *Nat Genet* 2000, **25:**25-9.

25. Morin MM, Moret BM: **NETGEN: generating phylogenetic networks with diploid hybrids.** *Bioinformatics* 2006, **22:**1921-3.

26. Tamames J, Moya A: **Estimating the extent of horizontal gene transfer in metagenomic sequences.** *BMC Genomics* 2008, **9:**136.

27. Babic A, Lindner A, Vulic M, Stewart E, Radman M: **Direct visualization of horizontal gene transfer.** *Science* 2008, **319:**1533-6.

F1000 Factor 3.0 *Recommended*
Evaluated by Tom Rapoport 06 May 2008

28. Zaneveld JR, Nemergut DR, Knight R: **Are all horizontal gene transfers created equal? Prospects for mechanism-based studies of HGT patterns.** *Microbiology* 2008, **154:**1-15.

29. Dutilh BE, van Noort V, van der Heijden RT, Boekhout T, Snel B, Huynen MA: **Assessment of phylogenomic and orthology approaches for phylogenetic inference.** *Bioinformatics* 2007, **23:**815-24.