Development of fusion and duplication finder BLAST (fdfBLAST): a systematic tool to detect differentially distributed gene fusions and resolve trifurcations in the tree of life

Submitted by **Guy Leonard**

to the **University of Exeter** as a thesis for the degree of **Doctor of Philosophy in Biological Sciences** in **December 2010**

I certify that all material in this thesis, which is not, my own work has been identified and that no material has previously been submitted and approved for the award of a degree by this or any other University.

Signature:

Guy Leonard 20/07/2011

Abstract

The construction of a tree of life and the placing of taxa into their correct phylogenetic context is the underpinning of modern evolutionary biology. However, many parts of the tree are currently unresolved due to conflicts within the sequence data. These sources of conflict include: horizontal gene transfer (HGT), hidden paralogy, and the effects of methodological artefacts such as Long Branch attraction (LBA). These limitations are further compounded by absence of key taxa that are yet to be sampled. Therefore, whilst phylogenetic methods are fundamentally useful for the reconstruction of the tree of life, given their current limitations, additional strategies are needed in order to fully resolve the tree of life. Gene fusions represent a potential source of evolutionary synapomorphies useful for resolving contentious branching relationships in the tree of life. I therefore, built a program to analyse whole genome datasets for the presence of differentially distributed gene fusion events (shared derived characters - SDCs). These putative SDCs can then be polarised with the help of traditional phylogenetic techniques and used as synapomorphies on the tree of life. Having constructed this program and tested it on established fusion datasets, I analysed five sets of four genomes from across the tree of life (the Deuterostomia, Fungi, Vertebrata, Viridiplantae and Discicristata). I used this data to identify the relative rates of gene fusion events. Previous studies have suggested that fission events occurred more often than gene fusion events. However, our analysis broadly suggests the opposite (albeit with a higher rate of fissions in the Deuterostomia). This result has direct implications for the use of gene fusions as evolutionary informative synapomorphies because the identification of a lower rate of reversion suggests that

these characters are less likely to be homoplasious and therefore represent useful tools for polarising evolutionary relationships. Six phylogenetically informative synapomorphies were recovered, three in the Discicristata which resolve the monophyly of the Kinetoplastida and four in the Fungi, one of which represented a HGT event and was independently discovered and previously published. Thus, this thesis reports the development and testing of a new tool to identify differentially distributed gene fusion events. The datasets analysed demonstrate that the program can be used to find phylogenetically informative gene fusion characters that can help resolve the tree of life in conjunction with traditional phylogenetic methods.

Table of Contents