# A reality check for alignments and trees

William Martin[1], Mayo Roettger[1] and Peter J. Lockhart[2]

[1] Institute of Botany III, University of Düsseldorf, Düsseldorf, Germany
[2] Allan Wilson Centre for Molecular Ecology and Evolution, Massey University, Palmerston North, New Zealand

**Making multiple sequence alignments is one of the more commonplace procedures in modern biology. Multiple alignments are typically generated by feeding sequences into the alignment program from the N-terminus to the C-terminus. Recent results show that if the same sequences are processed from the C- to the N-terminus, a different alignment is often obtained. Because phylogenetic trees are built from alignments, the resulting trees can also differ. The new findings highlight sequence alignment as a crucial step in molecular evolutionary studies and provide straightforward measures to assess alignment reliability.**

Nowhere in biology is evolutionary thinking more deeply engrained than in genomics. From the process of assembling genome sequences, to trees of genomes based upon hundreds of genes, genomics is about using the comparative tools of molecular evolution to address and inform biology. Most of the procedures germane to molecular evolution start with a multiple alignment, a matrix of rows (sequences) and columns (sites) that, ideally, has the homologous positions of each sequence written one under the other so that they can be compared [1]. Given such an alignment, biologists can then go about the business of inferring gene functions by comparison to known sequences, or inferring evolutionary trees that might illuminate past events in the history of life. Indeed, for all organisms that have not left physical traces of their past in the fossil record, gene sequences and molecular evolution are our main – if not sole – sources of information about their history. This information comes from sequence alignments, so alignments are important. Different alignments can suggest different evolutionary histories [1,2]. But how can we tell good alignments from bad ones? In addition, how can we tell the good parts of an alignment from the bad parts? Biologists and mathematicians have been troubled by such issues, and the specific problem of assessing alignment quality has been somewhat of an open question [3,4]. Computational biologists have long been working to solve problems related to refining and assessing sequence alignment quality [5–10] but the problems are hard [1–10] and perhaps underappreciated by the evolutionary and genomics communities.

Amid these developments, Landan and Graur [11] have unearthed a 'skeleton in the closet' of current sequence alignment practice that highlights the importance of assessing alignment quality. Their approach to demonstrate this is as simple as it is elegant and effective. Existing alignment programs typically read the sequences in from left to right, as in reading English. Landan and Graur use existing alignment programs but they instruct them also to read the sequences in from right to left, as in reading Hebrew. This produces two multiple alignments in parallel, which they call the 'heads' and the 'tails' alignment. Even though the heads and tails alignments start with identical information and are processed by identical algorithms, the two resulting alignments can contain highly different collections of site patterns (columns), and, accordingly, they can – and in many cases do – produce highly different trees.

The heads versus tails difference comes from an arbitrary step in the sequence alignment procedure, in which the computer has to decide among equally optimal solutions (paths) whether or not to insert a gap in one of the sequences. Specialists refer to the decision between accepting the different, equally possible paths as either taking the high road or the low road [12]. It was recognized long ago that the arbitrary decision to take the high road or the low road will give different alignments, and that the difference is a measure of the uncertainty of the alignment [12]. However, the issue was not pursued in any depth with regard to the day-to-day practice of generating alignments for phylogeny. Hence, it became largely forgotten, and the phylogenomics community perhaps now needs to recall the influence that this arbitrary choice can have on the multiple alignment result [11].

Heads versus tail alignments provide a simple means to evaluate the importance of this effect with different alignment programs. For typical real sequence data from genomes, Landan and Graur [11] found that <50% of the columns were identical in about half of the heads versus tails alignment comparisons. The consequence is that trees generated from the heads versus tails alignments typically differed by 30% of their branches or more, which is a substantial difference in tree topology.

So, what are the implications of these new findings? Are all previously published trees in need of reinspection? Do we need to start over again? The situation might not be so bad because for highly conserved or recently diverged sequences, the effect of heads versus tails alignments seems to be relatively minor. However, the more divergent sequences are, the more pronounced the differences between the heads versus tails alignment become. For typical real sequence data as biologists encounter it in day-to-day gene and genome comparisons, the differences can be highly significant [11], as the example in Figure 1 illustrates. The trees can differ too [11], as shown in Figure 2. So, for those studying ancient gene families or deeply diverged groups, there is ample cause for general concern

*Corresponding author:* Martin, W. (w.martin@uni-duesseldorf.de).
Available online 7 September 2007.

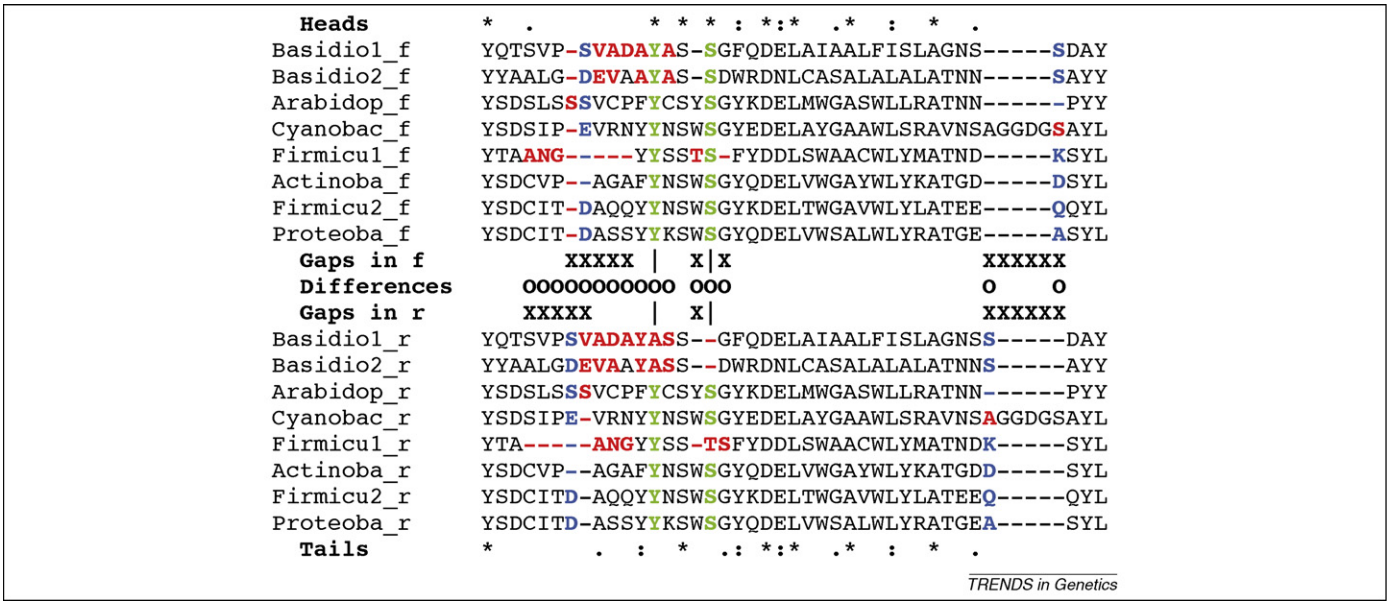


**Figure 1**. Example illustrating the difference between heads and tails alignments for homologues of gi_58270232, an endoglucanase precursor from *Cryptococcus neoformans*. The top alignment is heads (or forward), and the bottom alignment is tails (or reverse), generated as described [11] using MUSCLE [7]; the tails alignment was reversed once more here so that the sites can be easily compared. 'O's mark the columns at which site patterns differ in the two alignments, with the site pattern differences highlighted in red. Two site patterns that are identical, or nearly so, but shifted in position relative to one another in the heads and tails alignment, are highlighted in blue. 'X's mark the sites that contain gaps. A pipe '|' marks sites that are strictly conserved in one alignment but not in the other (highlighted in green). Note that if gapped sites were to be excluded from the heads and tails alignments (a common editing step before tree building), the alignments would still contain different collections of site patterns. Only 52% of the site patterns shown are identical between the two alignments in the region shown. Most of the site patterns that contain gaps differ between the two alignments.

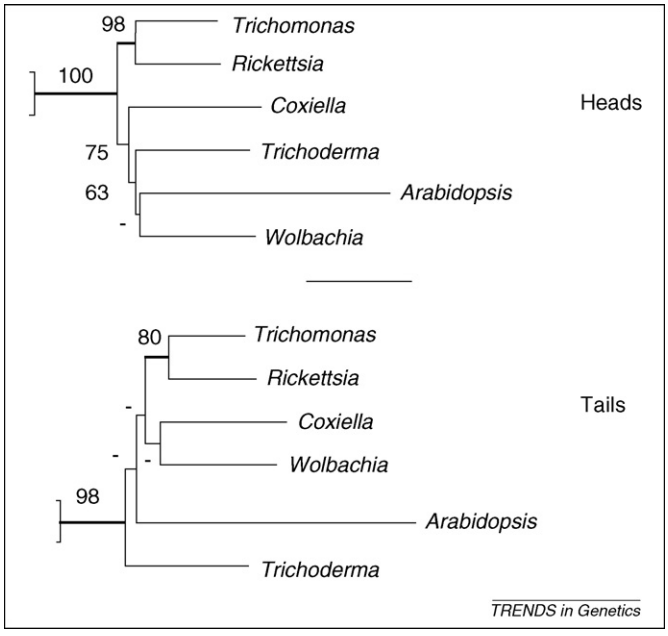


**Figure 2**. Heads and tails alignments can produce different trees, as illustrated here with alignments [7] for homologues of gi_42567717, an ankyrin repeat family protein from *Arabidopsis*. The two trees shown were constructed using the PHYML program [16] from the two alignments after gapped sites were removed. Rate variation across sites was modelled assuming a discrete $\gamma$ distribution of eight rate categories, with the $\alpha$ parameter estimated from the data. The bootstrap proportions shown on the maximum likelihood optimal trees were obtained using neighbour-joining and Dayhoff distances. The outgroup sequences are from the cyanobacteria *Nostoc* PCC 7120 and *Anabaena variabilis*. The scale bar in the centre of the figure indicates one substitution per site. Bootstrap proportions lower than 50 are shown as a dash. In this example, the total length of the heads alignment without gaps is 388 sites, and only 40% of the site patterns among sites that lack gaps in both the heads and tails alignments are identical. The two branches that are the same in the two trees are highlighted with thicker lines.

that many, or possibly even most, of the branches of many published trees might have been heavily dependent upon an arbitrary decision concerning the direction in which the sequences were fed into the alignment program. It might be that the influence of an arbitrary step in the pipeline of building trees from genome data has been insufficiently explored.

Phylogeny buffs are currently keen to know how much 'support' there is for a particular branch in a tree, usually estimated in the currency of bootstrap proportions or other measures, such as Bayesian support values [13–15]. Independent of such support values, we will now need to be asking in addition: does the reverse alignment produce the same result? If not, then the conclusion would be that the result hinges upon an arbitrary step in the alignment procedure and is hence likely to be an artefact of how the alignment was generated. If we take the new findings of Landan and Graur [11] seriously, and all indicators suggest that we should, it means that molecular evolutionists should at the very least multiply all of their future work by two: a heads and a tails alignment analysis to see whether the two results are consistent.

## References

1 Kumar, S. and Filipski, A. (2007) Multiple sequence alignment: in pursuit of homologous DNA positions. *Genome Res.* 17, 127–135
2 Morrison, D.A. and Ellis, J.T. (1997) Effects of nucleotide sequence alignment on phylogeny estimation: a case study of 18S rDNAs of apicomplexa. *Mol. Biol. Evol.* 14, 428–441
3 Batzoglou, S. (2005) The many faces of sequence alignment. *Brief. Bioinformatics* 6, 6–22
4 Morrison, D.A. (2006) Multiple sequence alignment for phylogenetic purposes. *Aust. Syst. Bot.* 19, 479–539

5 Edgar, R.C. and Sjolander, K. (2003) SATCHMO: sequence alignment and tree construction using hidden Markov models. *Bioinformatics* 19, 1404–1411
6 Thompson, J.D. *et al.* (2003) RASCAL: rapid scanning and correction of multiple sequence alignments. *Bioinformatics* 19, 1155–1161
7 Edgar, R.C. (2004) MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Res.* 32, 1792–1797
8 Lassmann, T. and Sonnhammer, E.L. (2005) Automatic assessment of alignment quality. *Nucleic Acids Res.* 33, 7120–7128
9 Chakrabarti, S. *et al.* (2006) State of the art: refinement of multiple sequence alignments. *BMC Bioinformatics* 7, 499
10 Schwartz, A.S. and Pachter, L. (2007) Multiple alignment by sequence annealing. *Bioinformatics* 23, e24–e29
11 Landan, G. and Graur, D. (2007) Heads or tails: a simple reliability check for multiple sequence alignments. *Mol. Biol. Evol.* 24, 1380–1383
12 States, D.J. and Boguski, M.S. (1991) Similarity and homology. In *Sequence Analysis Primer* (Gribskov, M. and Devereux, J., eds), pp. 89–157, Stockton Press
13 Douady, C.J. *et al.* (2003) Comparison of Bayesian and maximum likelihood bootstrap measures of phylogenetic reliability. *Mol. Biol. Evol.* 20, 248–254
14 Simmons, M.P. *et al.* (2004) How meaningful are Bayesian support values? *Mol. Biol. Evol.* 21, 188–199
15 Semple, C. and Steel, M.A. (2003) *Phylogenetics.*. Oxford University Press
16 Guindon, S. and Gascuel, O. (2003) A simple, fast, and accurate algorithm to estimate large phylogenies by maximum likelihood. *Syst. Biol.* 52, 696–704

0168-9525/$ – see front matter 
doi:10.1016/j.tig.2007.08.007

Genome Analysis

# Chromatin remodelling is a major source of coexpression of linked genes in yeast

**Nizar N. Batada, Araxi O. Urrutia and Laurence D. Hurst**

Department of Biology & Biochemistry, University of Bath, Bath, BA2 7AY, UK

**In diverse organisms, neighbouring genes in the genome tend to be positively coexpressed more than expected by chance. When the similarity of transcription regulation is controlled for, adjacent genes have much higher coexpression rates than unlinked genes, supporting a role for chromatin modelling. Consequently, many incidences of low-to-moderate level coexpression of linked genes might well be spurious rather than an indication of functional coordination. These results have implications for gene therapy and for understanding gene order evolution, suggesting that chromosomal proximity alone is adequate to achieve some level of coexpression.**

## Introduction

In eukaryotic genomes, neighbouring genes commonly have similar expression profiles [1]. In yeast, for example, adjacent genes are coexpressed to a significantly higher level than expected [2–5], coexpression being measured as the correlation in mRNA levels over time. That many coexpressed pairs are on opposite strands and divergently oriented (←→) (Figure S1 in the supplementary material online) has led to the suggestion that bidirectional promoters [5,6], residing between the genes, might explain much coexpression, although only a few well-characterized examples are known [7,8]. Instances are known in other taxa (as shown, for example, by Wright *et al.* [9]) and are conjectured to explain many instances of coexpression in diverse species [6,10–12].

Most highly coexpressed genes pairs in yeast, however, are on the same DNA strand [5,6] (→→ or ←←) (Figure S1 in the supplementary material online). Furthermore, on average, genes in close proximity in the genome show coexpression, even if they are not immediate neighbours [3,4,6]. Such exceptions cannot be the result of bidirectional promoters but might be due to transcription control similarity (TCS) [6] (e.g. tandem duplicates where 5′ control regions are also duplicated). Other sequence level explanations are viable (e.g. transcriptional read-through [13]).

Are such explanations at the sequence level sufficient to account for all positive coexpression? The finding of longer range (e.g. tens of genes) correlation of expression in *Drosophila* [14], human [15,16] and yeast [3,4] has led to the suggestion that chromatin modification might also have a role. In humans, for example, the silencing of spans of genes seems to be modified at the level of chromatin [17,18]. However, the signals of coexpression seen across large blocks might be explained by signals derived from clusters of bidirectional promoters or transcriptional read-through [13].

## Chromatin-level regulation could explain coexpression

Recent single mRNA molecule experiments indicate that chromatin modification might indeed explain much coexpression of neighbouring genes [19]. Importantly, Raj *et al.* [19] have shown that two reporter genes adjacent to each other fire in a coordinated fashion but fire independently if unlinked, despite the same transcriptional regulation. To see why, consider a simple null model in which gene expression is only possible when the genes are in open chromatin, which is assumed to span several genes [20]. If chromatin is frequently opening and then closing, possibly stochastically, then genes in close proximity on the chromosome will simultaneously be amenable to transcription, whereas those unlinked will have less coordinated transcription. Common transcription factor bindings sites

*Corresponding author:* Hurst, L.D. (l.d.hurst@bath.ac.uk).
Available online 5 September 2007.