

## A Distance Measure for Phylogenetic Analysis of Genomes

Minh Duc Cao<sup>1</sup> Trevor I. Dix<sup>1,2</sup> Lloyd Allison<sup>1</sup>  
{minhduc,trevor,lloyd}@infotech.monash.edu.au

<sup>1</sup>Clayton School of Information Technology, Monash University, Australia

<sup>2</sup>Faculty of Information & Communication Technologies, Swinburne University of Technology, Australia

### 1. Introduction

It is well known that species phylogenetic analysis based on single genes or parts of the genomes are often inconsistent because of factors such as variable rates of evolutions and horizontal gene transfer. The availability of more and more sequenced genomes allows phylogeny construction from complete genomes that is less sensitive to such inconsistency. For such long sequences, construction methods like *maximum parsimony* [2] and *maximum likelihood* [4] are often not possible due to their intensive computation requirement. Other methods such as the *neighbour joining* [6] method, require a measure of distances between any two genomes. We present in this study a measure of genetic distance between genomes based on information theory [7, 9]. This method uses the *expert model* [3], a biological oriented compression algorithm, to estimate the information content of sequences. We demonstrate that our distance measure can be applied to build the phylogenetic tree of a number of *Plasmodium* parasites from their genomes.

### 2. Method and Results

Similar to [1], our work is based on the premise that if two sequences are related, one sequence must tell something useful about the other, whose information content can be measured by lossless compression. The information content  $\mathcal{I}_X$  of sequence  $X$  can be approximated by the length of the encoded message obtained by compressing  $X$  using the expert model. If a sequence  $Y$  related to  $X$  is available, the expert model can compress sequence  $X$  based on the background knowledge from sequence  $Y$  to give a measure of *conditional information content* of  $X$  given  $Y$ ,  $\mathcal{I}_{X|Y}$ . The more related  $Y$  to  $X$ , the more information we can save for compressing  $X$ . In other words, the difference between the information content of  $X$  and the conditional information content of  $X$  given  $Y$  is a measure of similarity of the two sequences. The quantity is defined as the *shared information*:  $\mathcal{I}_{X,Y} = \mathcal{I}_X - \mathcal{I}_{X|Y}$ .

In theory,  $\mathcal{I}_{X,Y}$  should be equal to  $\mathcal{I}_{Y,X}$  as they both present the shared information of the two sequences. However, this is not always

2 M.D. Cao, I.T. Dix, & L. Allison

the case in practise due to arithmetic errors in compression. We therefore take the average of the two as the similarity measure of the two sequences. As the two genomes may have different lengths, and they may have nucleotide composition bias, we normalise the similarity measure by a factor of  $\mathcal{I}_X + \mathcal{I}_Y$ :

$$\begin{aligned} S_{X,Y} &= \frac{\mathcal{I}_X - \mathcal{I}_{X|Y} + \mathcal{I}_Y - \mathcal{I}_{Y|X}}{\mathcal{I}_X + \mathcal{I}_Y} \\ &= 1 - \frac{\mathcal{I}_{X|Y} + \mathcal{I}_{Y|X}}{\mathcal{I}_X + \mathcal{I}_Y} \end{aligned} \quad (1)$$

In other words, the distance measure is:

$$D_{X,Y} = \frac{\mathcal{I}_{X|Y} + \mathcal{I}_{Y|X}}{\mathcal{I}_X + \mathcal{I}_Y} \quad (2)$$

We applied the distance measure to build the phylogenetic tree of eight malaria parasites of the *Plasmodium* genus, namely *P. berghei*, *P. yoelii*, *P. chabaudi* (rodent malaria), *P. falciparum*, *P. vivax*, *P. knowlesi*, *P. reichenowi* (primate malaria) and *P. galinaceum* (bird malaria). Their genomes were obtained from PlasmoDB release 5.4 [10]. Pairwise distances were computed as Eq. 2. The tree was generated using the *neighbour joining* [6] method and is shown in Fig. 1. The tree is consistent with some of earlier work [5, 8].

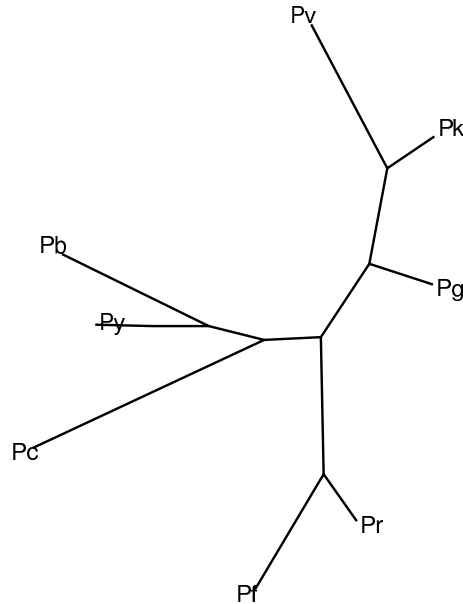


Fig. 1. Tree constructed by Neighbour Joining.

## References

- [1] L. Allison, C. S. Wallace, and C. N. Yee. Finite-state models in the alignment of macromolecules. *Journal of Molecular Evolution*, 35(1):77–89, 1992.
- [2] J. Camin and R. Sokal. A method for deducing branching sequences in phylogeny. *Evolution*, 19:311–326, 1965.
- [3] M. D. Cao, T. I. Dix, L. Allison, and C. Mears. A simple statistical algorithm for biological sequence compression. *Data Compression Conference*, pp 43–52, 2007.
- [4] J. Felsenstein. Evolutionary trees from DNA sequences: a maximum likelihood approach. *Journal of Molecular Biology*, 76(6):368–376, 1981.
- [5] M. C. Leclerc, J. P. Hugot, P. Durand, and F. Renaud. Evolutionary relationships between 15 *Plasmodium* species from new and old world primates (including humans): an 18s rDNA cladistic analysis. *Parasitology*, 129(16):677–684, 2004.
- [6] N. Saitou and M. Nei. The neighbor-joining method: A new method for reconstructing phylogenetic trees. *Mol. Biol. Evol*, 4(4):406–425, 1987.
- [7] C. E. Shannon. A mathematical theory of communication. *The Bell System Technical Journal*, 27:379–423, 1948.
- [8] M. E. Siddall and J. R. Barta. Phylogeny of *Plasmodium* species: Estimation and inference. *The Journal of Parasitology*, 78(3):567–568, 1992.
- [9] C. S. Wallace and D. M. Boulton. An information measure for classification. *Computer Journal*, 11(2):185–194, August 1968.
- [10] PlasmoDB. <http://www.plasmodb.org/common/downloads/release-5.4/>, accessed 2008.