2014 3rd International Conference on Environment Energy and Biotechnology
IPCBEE vol.70 (2014) © (2014) IACSIT Press, Singapore
DOI: 10.7763/IPCBEE. 2014. V70. 2

A Novel Algorithm to Reconstruct Phylogenies Using Gene Sequences and Expression Data

Krzysztof Bartoszek[1+] and Pietro Li ó[2]

[1] Department of Mathematics, Uppsala University, Uppsala, Sweden.
[2] Computer Laboratory, University of Cambridge, Cambridge, United Kingdom.

Abstract. Phylogenies based on single loci should be viewed with caution and the best approach for obtaining robust trees is to examine numerous loci across the genome. It often happens that for the same set of species trees derived from different genes are in conflict between each other. There are several methods that combine information from different genes in order to infer the species tree. One novel approach is to use information from different "omics". Here we describe a phylogenetic method based on an Ornstein–Uhlenbeck process that combines sequence and gene expression data. We test our method on genes belonging to the histidine biosynthetic operon. We found that the method provides interesting insights into selection pressures and adaptive hypotheses concerning gene expression levels.

Keywords: phylogeny, evolution, multi-omics, Ornstein-Uhlenbeck process

1. Introduction

Natural selection acts as a driving force at virtually all levels of biological organization. Phylogenetic analysis is a powerful methodology to investigate not only history but also selection mechanisms and function of biological networks at all levels [1]. Phylogenetic trees based on single loci should be viewed with caution and the best approach is to examine numerous loci across the genome. Due to lateral gene transfer it is not surprising that bacterial phylogenetic analyses of single loci have yielded contradictory trees. When DNA sequence data are compiled across multiple loci, either by concatenation into a single sequence and constructing a single phylogeny or by conducting multi–locus significance tests on independent phylogenies, the support becomes robust. There are different hypotheses concerning what types of genetic changes are likely to underlie important phenotypic differences. The "protein evolution" hypothesis proposes that key changes have occurred in coding regions and that these have resulted in important modifications to the encoded proteins. The "gene regulatory evolution" hypothesis suggests that differences between species largely depend on the gene expression pattern. Studies of gene expression differences among species often employ micro–arrays to analyse gene regulatory changes. Differences in gene expression patterns between species — similarly to differences in gene sequences – might accrue in a clock-like fashion and therefore many expression differences are likely to be neutral and have no functional importance [2].

Phylogenetic methods have already been applied to gene expression data [3], [4]. Phylogenetic comparative methods usually require knowledge of the phylogeny connecting the species [there are however exceptions to this i.e. "tree–free" methods, [5]-[7]. This poses a problem to applying them to data directly derived from sequences. The phylogeny is estimated from genetic data and the majority of current phylogenetic comparative methods assume that the phenotypic data does not influence the tree, i.e. the genes controlling it are independent from those responsible for the speciation process [again there are exceptions to this [8]-[11]. In the case of sequence derived data this can be difficult to meet. There can be many dependencies inside the genome due to e.g. linkage disequilibrium. Combining gene sequences with inter–species gene expression levels could allow us to construct tests for the correct phylogeny. The gene

[+] Corresponding author. Tel.: +46184713226.
E-mail address: krzysztof.bartoszek@math.uu.se.

expression levels are something certain and it should be safe to assume they evolved according to the species tree. The conflicting gene trees on the other hand describe only the most probable evolution of the given gene. Hopefully in the ensemble is a gene that followed the speciation process. Hence we should hope that this tree would best explain the observed expression pattern.

Functionally related genes (i.e. genes forming a pathway) in bacterial chromosomes which are organized into polycistronic co-transcribed operons can be an interesting case study. Recent works suggest that the clustering of genes into operons would both reduce the molecular noise associated to gene expression and favour the lateral transferring of an entire cluster of functionally related genes. Kovacs and collaborators have suggested that the order of genes within an operon may affect the noise [12]. In the next section we describe a phylogenetic methodology model evolution as an Ornstein–Uhlenbeck process to analyse a bacterial gene cluster.

2. Multivariate Phylogenetic Adaptive Framework

Felsenstein [13] proposed the first [however see also [8] continuous phylogenetic comparative methods procedure, independent contrasts. This assumes a Brownian motion evolving on the phylogenetic tree and we observe the tree and contemporary phenotypic sample. Already in [14] it was observed that such a model does not allow for modelling adapting traits [15], [16]. A Brownian motion will mean that the phenotype randomly oscillates around the ancestral state. Therefore an Ornstein–Uhlenbeck process was proposed [17] with different regimes on the phylogeny to model a phenotype adapting to different conditions (e.g. habitats). This was further developed to a trait evolving towards a randomly changing environment [18]. Ornstein–Uhlenbeck models have been also applied to study evolutionary rates [19], [20]. One can naturally take the evolving phenotype to be measured gene–expression levels and apply the aforementioned methods. However it would be more interesting to consider how expression levels of different genes co–evolve. In [21] an Ornstein–Uhlenbeck model for a multiple, say k, co-adapting traits is presented, $dY(t) = -\mathbf{A}(Y(t) - \theta(t))\,dt + \mathbf{\Sigma} dB(t)$, where $\mathbf{A}$, $\mathbf{\Sigma}$ are k×k matrices, θ is a vector step function over the phylogeny and B(t) is a k–dimensional standard Wiener process. A maximum–likelihood estimation procedure for this model is implemented in the mvSLOUCH R [21], [22] package, available on CRAN (http://cran.r-project.org/web/packages/mvSLOUCH/). One can further incorporate into the estimation procedure measurement error (or intra–species variability). This is an important factor to keep in mind as micro–array experiments can be very noisy and measurement variance can have a profound effect on a phylogenetic analysis [23]. We use capital letter for the protein's name and small letter for the gene's name.

3. Analysis of Histidine Operon Gene Expression

To illustrate our approach we study the sequence and gene expression data of the bacterial Histidine biosynthetic operon in seven species, *Streptomyces coelicolor* (Strept), *Pseudomonas aeruginosa* (Pseudo), *Mycobacterium tuberculosis* (Mycoba), *Helicobacter pylori* (Helico), *Bacillus subtilis* (Bacill), *Escherichia coli* (Escher) and *Salmonella enterica serovar Typhimurium* (Salmon). The operon consists of the following genes: GDCBHAFIE. This case study has been chosen for the many interesting characteristics of the Histidine operon [24]: 1) in enteric bacteria such as E. coli and Salmonella, all structural genes for the enzymes responsible for histidine synthesis are in the histidine operon in the same order as the reactions of the pathway; 2) it has been shown it is an ancient operon; 3) in Escherichia coli it contains internal promoters, pointing to an evolutionary history of mini operons aggregation; 4) it is contained in a single operon in some bacteria (such as in *E. coli* and *S. typhimurium*), some genes are dispersed in more than one cluster in other bacteria (such as *P. aeruginosa* and *H. pylori*); 5) it contains duplicated genes (His A and F), and, in some bacteria, bifunctional (IE) enzymes. We consider four tree gene trees, based on the hisD, hisE, hisF and hisG genes of the operon. The trees were constructed from amino acid sequences using PhyML [25], [26] and *Thermotoga* (Thermo) as an outgroup. These trees are illustrated in Fig. 1. We can see that they are different and we would like to use phenotypic (gene expression in this case) data to study which one is closer to the species tree. We downloaded gene expression data from the Colombos database [27] via the R package Rcolombos. The data are contrasts between different conditions for the species' genes from a collection of experiments. It was a problem to find between species data that would correspond to each other, i.e. a pair of conditions that was measured in all species. Therefore we choose as one condition a wild type "without any treatment". We tried to find these to be as similar as possible to each other, albeit if one checks the accession codes in Tab. 1 one can see that these were in different time points of colony growths. We did not choose a second contrast but rather took an average of all contrasts that contained this one chosen "condition". Our phenotype can be understood the average change of gene expression level. In order to evaluate differences due to the gene positioning inside the operon, we considered hisD, hisF, hisG and hisIE genes. There are not enough species to run a full four dimensional analysis so instead we ran all possible six bivariate combinations. Even in this case we can see a large variability in the results. This variability is due to the high-dimensional numerical optimization of the likelihood function. Using six pairs on the one hand does not allow us to catch all interactions between genes but on the other hands makes conclusions about the chosen phylogenetic tree more robust. In our study we also take into account the variance of these contrasts which can be treated as measurement error.

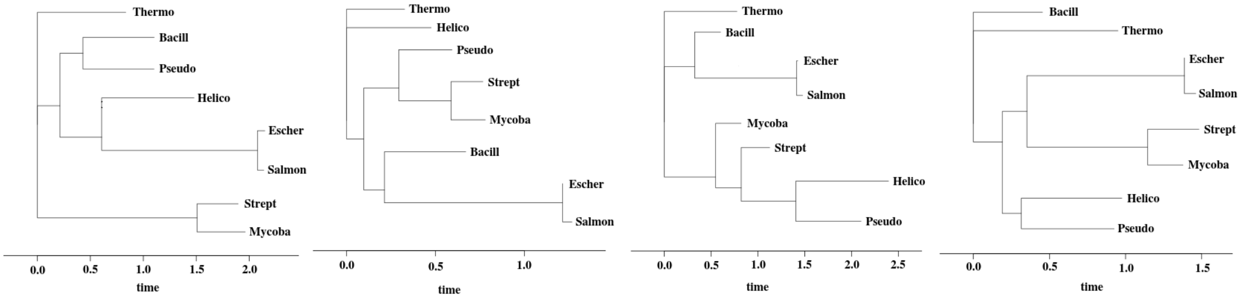

Fig. 1: From the left: HisD gene tree, HisIE gene tree, HisF tree, HisG tree.

Table 1: Comparative data used to find support for the species tree. Data from the Colombos database.

Species	Accession	hisD		hisF		hisG		hisI	
		mean	var	mean	var	mean	Var	mean	Var
Escher	GSM991216	-0.251	0	0.001	0.000	-0.142	0	0.102	0
Salmon	GSM460281	-0.475	0.121	-0.443	0.269	-0.252	0.183	-0.857	0.208
Strept	GSM564847	-0.578	0.089	NA	0	-0.470	0.092	-0.050	0.014
Mycoba	GSM183531	0.611	0.263	-0.767	0.112	0.783	0.176	1.257	0.934
Bacill	GSM672549	2.087	5.148	1.882	4.591	1.881	4.964	1.892	4.432
Pseudo	GSM821495	-0.050	0.020	NA	0	0.186	0.089	0.219	0.059
Helico	-	NA	0	NA	0	NA	0	NA	0

We estimated the parameters of a bivariate Brownian motion and Ornstein–Uhlenbeck process evolving on hisD, hisE, hisF and hisG trees for all possible six pairs of genes. Using the AIC_c we found that in all six cases, the pair of genes best supported a Brownian motion (i.e. **A**=**0**) evolving on the HisG tree. We summarize the results in Tab. 2. The larger support for the HisG tree with respect to the other trees could be explained by several facts. It is noteworthy that hisG is the first gene of the histidine operon; the hisG gene codes for the ATP phosphoribosyltransferase (EC:2.4.2.17) which catalyses the first step in the biosynthesis of histidine in bacteria. The enzyme is at the equilibrium between an active dimeric form, an inactive hexameric form and higher aggregates which make its evolution more constrained. The best fitting model was a Brownian motion one which does not assume any stabilizing selection. This could be the consequence of the chosen phenotype. This result suggests that, although the genes in an operon are co-transcribed, they have different evolutionary dynamics and possibly different selection pressure. We studied an average gene expression contrast and found it was best explained by a Brownian motion. However we should expect that the same contrast for all species should be under a selection pressure and favour an Ornstein-Uhlenbeck model.

Early works have highlighted both the existence of internal promoters (so that the transcription of some genes could start independently from the main promoter) and the presence of non-complete/interrupted transcription due to early detachment of the RNA polymerase [28], [29]. The internal promoters divide the genes in three sets: hisG and hisD; hisF; hisIE. The study presented here is limited in the number of species and genes. The phenotype we took was an average contrast for the six pairs of genes. We took all available contrasts with a wild type. We tried to choose the wild types so that they would match each other as well as possible between the species. This however is not perfect as e.g. the time points of colony growths are not the same in all experiments. The comparative models are also high dimensional so with such a small data set Ornstein–Uhlenbeck models will be heavily punished. To be both powerful and robust, phylogenetic inference needs to include more "omics" and we are delighted to report the use of multivariate phylogenetic inference methods based on Ornstein–Uhlenbeck models. This approach provides an improved understanding of the biological processes that shape life at the "omics" level. Secondly, this method provides an improved ability to compare genes and the differences between phenotypes. This could lead to better feature prediction (for example the presence of internal promoters or termination sites) and to the genetic design for synthetic biology. Our study here is limited in terms of space and is intended to illustrate the potentialities of the method. Work in progress consists in testing larger phylogenies and publicly available gene expression data sets.

Table 2: Results of the phylogenetic comparative analysis.

		(hisD,hisF)		(hisD,hisG)		(hisD,hisI)		(hisF,hisG)		(hisF,hisI)		(hisG,hisI)	
log-likelihood		10.562		13.338		10.687		16.351		7.482		12.095	
AIC$_c$		-5.124		-10.675		-5.374		-16.701		1.037		-8.190	
Y(0)	gene 1	-0.113		0.132		-0.180		0.089		0.170		-0.091	
	gene 2	0.276		0.046		0.406		-0.121		0.465		0.553	
$\Sigma\Sigma^T$	gene 1	0.410	0.816	1.196	0.588	0.067	0.286	0.791	0.195	0.702	1.513	0.052	0.457
	gene 2	0.816	1.625	0.588	0.289	0.286	1.228	0.195	0.048	1.513	3.261	0.457	3.994

4. Acknowledgements

K. B. was supported by Svenska institutets Visbyprogrammet scholarship nr. 1142/2013. P.L. thanks the European ''MIMOmics'' (305280) project. We thank Anna Stokowska for valuable comments on the manuscript.

5. References

[1] J. P. Huelsenbeck, B. Rannala, and J. P. Masly. Accommodating phylogenetic uncertainty in evolutionary studies. *Science*, **88**:2349–2350, 2000.

[2] P. Khaitovich, G. Weiss, M. Lachmann, I. Hellmann, W. Enard, B. Muetzel, U. Wirkner, W. Ansorge and S. Pääbo. A neutral model of transcriptome evolution. *PLos Biol.*, **2**:682–689, 2004.

[3] C. W. Dunn, X. Luo and Z. Wu. Phylogenetic analysis of gene expression. *ArXiv e-prints*, 2013.

[4] K. H. Eng, H. C. Bravo and S. Keleş. A phylogenetic mixture model for the evolution of gene expression. *Mol. Biol. Evol.*, **26**(10):2363–2372, 2009.

[5] F. Bokma. Detection of punctuated equilibrium from molecular phylogenies. *J. Evol. Biol.*, **15**:1048–1056, 2002.

[6] F. W. Crawford and M. A. Suchard. Diversity, disparity, and evolutionary rate estimation for unresolved Yule trees. *Syst. Biol.*, **62**(3):439–455, 2013.

[7] S. Sagitov and K. Bartoszek. Interspecies correlation for neutrally evolving traits. *J. Theor. Biol.*, **309**:11–19, 2012.

[8] A. W. F. Edwards. Estimation of the branch points of a branching diffusion process. *J. Roy. Stat. Soc. B*, **32**

(2):155–174, 1970.

[9] R. G. FitzJohn. Quantitative traits and diversification. *Syst. Biol.*, **59** (6):619–633, 2010.

[10] J. P. Huelsenbeck and B. Rannala. Detecting correlation between characters in a comparative analysis with uncertain phylogeny. *Evolution,* **57** (6):1237–1247, 2003.

[11] W. P. Maddison, P. E. Midford and S. P. Otto. Estimating a binary character's effect on speciation and extinction. *Syst. Biol.*, **56** (5):701–710, 2007.

[12] K. Kovacs, LD Hurst, B. Papp. Stochasticity in protein levels drives colinearity of gene order in metabolic operons of Escherichia coli. *PloS Biol* 7: e1000115, 2009.

[13] J. Felsenstein. Phylogenies and the comparative method. *Am. Nat.*, **125** (1):1–15, 1985.

[14] T. F. Hansen. Stabilizing selection and the comparative analysis of adaptation. *Evolution*, **51**(5):1341–1351, 1997.

[15] T. F. Hansen and S. H. Orzack. Assessing current adaptation and phylogenetic inertia as explanations of trait evolution: the need for controlled comparisons. *Evolution*, **59** (10):2063–2072, 2005.

[16] J. Felsenstein. Phylogenies and quantitative characters. *Annu. Rev. Ecol. Syst.*, **19**:445–471, 1988.

[17] M. A. Butler and A. A. King. Phylogenetic comparative analysis: a modelling approach for adaptive evolution. *Am. Nat.*, **164** (6):683–695, 2004.

[18] T. F. Hansen, J. Pienaar and S. H. Orzack. A comparative method for studying adaptation to a randomly evolving environment. *Evolution,* **62**:1965–1977, 2008.

[19] S. Aris-Brosou and Z. Yang. Effects of models of rate evolution on estimation of divergence dates with special reference to the Metazoan 18S ribosomal RNA phylogeny. *Syst. Biol.* **51** (5):703–714, 2002.

[20] S. Aris-Brosou and Z. Yang. Bayesian models of episodic evolution support a late Precambrian explosive diversification of the Metazoa. *Mol. Biol. Evol.*, **20** (12):1947–1954, 2003.

[21] K. Bartoszek, J. Pienaar, P. Mostad, S. Andersson and T. F. Hansen. A phylogenetic comparative method for studying multivariate adaptation. *J. Theor. Biol.*, **314**:204–215, 2012.

[22] R Core Team. R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria, 2013. URL http://www.R-project.org.

[23] T. F. Hansen and K. Bartoszek. Interpreting the evolutionary regression: the interplay between observational and biological errors in phylogenetic comparative studies. *Syst. Biol.*, **61** (3):413–425, 2012.

[24] P. Alifano, R. Fani, P. Li ò, A. Lazcano, M. Bazzicalupo, MS Carlomagno, CB Bruni. Histidine biosynthetic pathway and genes: structure, regulation, and evolution. *Microbiol Rev*. **60**:44-69, 1996.

[25] S. Guindon and O. Gascuel. A simple, fast, and accurate algorithm to estimate large phylogenies by maximum likelihood. *Syst. Biol.*, **52**(5):696–704, 2003.

[26] S. Guindon, J. F. Dufayard, V. Lefort, M. Anisimova, W. Hordijk and O. Gascuel. New algorithms and methods to estimate maximum–likelihood phylogenies: assessing the performance of PhyML 3.0. *Syst. Biol.*, **59** (3):307–321, 2010.

[27] P. Meysman, P. Sonego, L. Bianco, Q. Fu, D. Ledezma-Tejeida, S. Gama-Castro, V.iebens, J. Michiels, K. Laukens, K. Marchal, J. Collado-Vides and K. Engelen. COLOMBOS v2.0: an ever expanding collection of bacterial expression compendia. *Nucleic.Acids Res.*, 2013.

[28] J. F. Atkins and J. C. Loper. Transcription Initiation in the Histidine Operon of *Salmonella typhimurium. Proc Natl Acad Sci U S A,* **65**: 925–932, 1970.

[29] B. Ely and Z. Ciesla. Internal Promoter P2 of the Histidine Operon of *Salmonella typhimurium. J. Bacteriol,* **120** (2): 984–986, 1974.