BIOINFORMATICS
Electronic edition http://www.bioinformatics.oupjournals.org

VOLUME 19
NUMBER 11
JULY 22 2003
PAGES 1412–1416

DOI: 10.1093/bioinformatics/btg174

GeneTRACE–reconstruction of gene content of ancestral species

*Victor Kunin and Christos A. Ouzounis**

Computational Genomics Group, The European Bioinformatics Institute, EMBL Cambridge Outstation, Cambridge CB10 1SD, UK

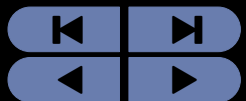
Received on November 11, 2002; revised on January 22, 2003; accepted on February 20, 2003.

*To whom correspondence should be addressed.

Abstract

While current computational methods allow the reconstruction of individual ancestral protein sequences, reconstruction of complete gene content of ancestral species is not yet an established task. In this paper, we describe GeneTrace, *an efficient linear-time algorithm that allows the reconstruction of evolutionary history of individual protein families as well as the complete gene content of ancestral species. The performance of the method was validated with a simulated evolution program called SimulEv. Our results indicate that given a set of correct phylogenetic profiles and a correct species tree, ancestral gene content can be reconstructed with sensitivity and selectivity of more than 90%. SimulEv simulations were also used to evaluate performance of the reconstruction of gene content-based phylogenetic trees, suggesting that these trees may be accurate at the terminal branches but suffer from long branch attraction near the root of the tree.*

Contact: *ouzounis@ebi.ac.uk*



GO BACK
CLOSE FILE

Introduction

While computational sequence analysis methods were developed to reconstruct ancestral DNA and protein sequences (Pupko *et al.*, 2000), these studies generally focus on the reconstruction of sequences of single genes or proteins, rather than complete genomes. In this study, we present a method aimed to reconstruct the gene content of ancestral prokaryotes and assess its performance.

To reconstruct the gene content of ancestral species, we suggest a framework for the inference of presence or absence of individual protein families at any node on a phylogenetic tree. Our approach is based on the following assumptions:

(1) When most of the clade members contain a representative of a protein family, the observed distribution pattern would normally result from vertical gene descent. The common ancestor of the clade is thus assumed to contain the corresponding family.

(2) If a protein family is present in most of the descendants of a particular ancestor, but is not found in some subclade, the observed gene absence would normally result from gene loss.

(3) Protein family distribution interspersed in distantly related clades would be indicative of horizontal gene transfer.

GO BACK
CLOSE FILE

Previously, these gene distribution patterns, also called phylogenetic profiles, have been used to predict protein function (Pellegrini *et al.*, 1999), to build gene content-based phylogenetic trees (Snel *et al.*, 1999) and to deduce the cell localization of gene products (Marcotte *et al.*, 2000). Recently, a first approach to reconstruct the gene content of ancestral species using phylogenetic profiles was reported (Snel *et al.*, 2002). While this method was based on similar assumptions, the details of algorithm, threshold calibration and reliability of the generated predictions have not been reported. Another complication was the usage of orthologs, which are not always easily discernible.

Other studies using phylogenetic profiles with reference to phylogenetic reconstruction include attempts for functional predictions from genome data using either parsimony (Liberles *et al.*, 2002) or kernel methods (Vert, 2002). While a by-product of these two applications was the reconstruction of ancestral states, no indication of performance or accuracy of ancestral reconstruction were given.

Here, we describe an implementation and rigorous testing of an improved linear-time algorithm, specifically designed to trace the history of protein families using phylogenetic profiles. To avoid the issues of ortholog definition, we have decided to use phylogenetic profiles that contain information about the presence or absence of an entire protein family. Family information was obtained from entire genome sequences clustered with the TRIBE-MCL algorithm (Enright *et al.*, 2002). However, the algorithm could utilize other types of phylogenetic profiles, for example Clusters of Orthologous Genes

(Tatusov *et al.*, 1997). Finally, to assess the performance of the algorithm, we have also developed a new method for the simulation of genome evolution, described herein.

Methods

Reconstruction algorithm

We present a method called GeneTrace that allows the inference of the most likely evolutionary scenario that led to the observed present-day distribution of protein families (Fig. 1). The GeneTrace input consists of phylogenetic profiles of protein families and an evolutionary tree including all organisms involved. Inner nodes on this tree represent ancestral organisms (Fig. 1a). Two types of events are considered: protein family gain and loss. The algorithm consists of the following stages (Fig. 1):

(1) For each inner node, the minimal number of potential changes that are required to obtain the observed family distribution is calculated for both possible cases: gene family presence and absence at the node (Fig. 1b). Both gene acquisition and loss are penalised by a single point. The calculation proceeds from terminal nodes of the tree towards the root. For each node down the tree, the penalty is equal to the sum of the penalties of its daughter nodes. These penalties are transformed into assignments of family presence or absence at the node in any of the following cases (Fig. 1c):

- If the descendants of the node exhibit a uniform pattern—either family presence or absence, the corresponding pattern is assigned to the node.

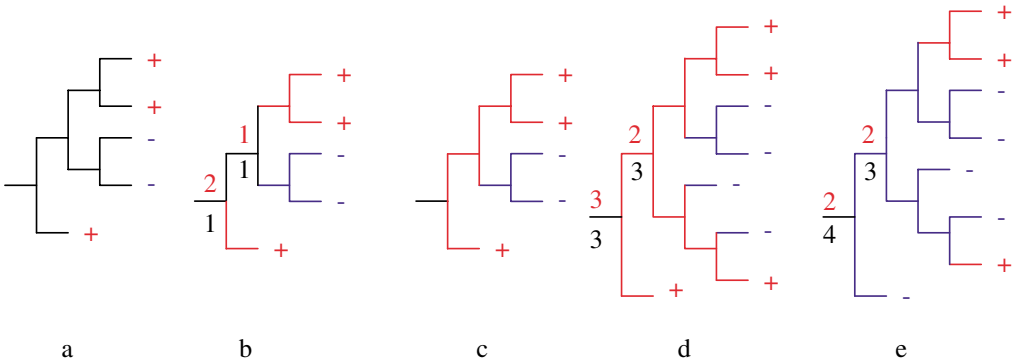


Fig. 1. The flow of the GENETRACE algorithm. Gene presence is marked by red colour, and gene absence is marked by blue. The input consists of a trusted species tree and phylogenetic profile. Plus (+) shows gene presence and minus (−) absence in an extant species (terminal nodes on the tree). (a) The input data. (b) Unambiguous cases are resolved, and the number of independent changes required to obtain the given data is calculated for both gene presence (red numbers) and absence (blue numbers) for each internal node. (c) A putative scenario for evolutionary history of the gene is suggested, based on the Gain threshold (see text). (d) When the difference of potential gains and losses is between the Gain and Loss thresholds, the final assignment is dependent on the subtree neighbourhood.

- If the difference between the number of potential gains and losses is larger than a threshold value called the GAIN threshold, and the family presence is observed on at least two daughter subtrees, family presence

GO BACK
CLOSE FILE

is assigned to the node.

- If the difference between number of potential gains and losses is smaller than a threshold value called the LOSS threshold, family absence is assigned to the node.

(2) Starting from the root of the tree, unassigned nodes inherit the parental assignment (Fig. 1d,e). The parent of the root is assumed not to contain any genes, thus delaying the first assignment to the first evidence of family presence.

The algorithm infers the presence and absence of the family on the nodes of evolutionary tree, and generates a list of nodes where family gain and loss is predicted to occur. Horizontal gene transfer is inferred if more then one family gain is reported.

GAIN and LOSS thresholds are different, as they stand for family gain and loss at the node of interest. The GAIN threshold is conceptually analogous to the HGT penalty described earlier (Snel *et al.*, 2002). This threshold stands for the assessment of the probability for multiple gene loss events versus HGT events. Family gain is assumed if the cost of all losses is smaller than the allowed penalty for horizontal transfer. The number of suspected horizontal transfers would be the number of family gains minus one (accounting for family genesis).

When family presence is observed on the parental tree node, assigning family absence to a node would imply gene loss. If some descendants of this node appear to have the protein family, the loss would be followed by regaining. In

such a scenario, the introduction of the LOSS threshold brings an additional requirement for higher amounts of gene loss for assigning gene absence and allows a more parsimonious version of events. The described system comprised of two walks on the tree and two thresholds allows considering subtree neighbourhoods (Fig. 1d and e), and thus is superior to the previous model (Snel *et al.*, 2002).

The algorithm is implemented as a set of programs in Perl programming language. The performance is acceptable: analysis of phylogenetic profiles of 51 complete bacterial proteomes consisting of 12 762 protein families takes approximately 17 minutes on Ultra 5 Sun workstation.

Accuracy analysis

Our current collection of sequenced genomes is a small sample from the rich world of micro-organisms, and new data is constantly arriving from genome projects. We thus aimed to find how stable these predictions are, given the constant arrival of new data. To estimate the confidence levels of predictions, we applied the jackknife procedure for each family analysed, by removing at random half of the available genomes and recalculating the predictions. The procedure is iterated 100 times, and the fraction of positive family assignments is taken as a confidence level for the initial assignment.

Simulated evolution

As sequencing ancient genomes in order to verify the performance of the algorithm is impossible, other methods are required. One of the possible verification methods is simulated evolution and comparison between real and reconstructed data. We designed a program we call SimulEv (SIMULated EVolution) as the test case for the performance of the evolutionary reconstruction.

The SimulEv simulation is started from a hypothetical ancestral organism with predefined number of genes or protein families. These ‘genes’ do not contain any sequence or other information beyond gene identifiers. The initial genome multiplies, giving rise to two daughter genomes. Daughter genomes are mutated by random addition and removal of genes from the genomes. Genes can be added by gene genesis (synthesis of a new type of gene) or horizontal transfer (gain of an existing type of gene). The extent of each of processes is set by a parameter, multiplied by a random value at each generation. During the first stages of evolution, when the number of genomes is still small, horizontal transfer would be virtually meaningless, thus is not allowed. When the number of genomes reaches a certain threshold, horizontal transfer is permitted.

As the number of genomes grows exponentially through each generation, a constraint to keep this number constant is required. This step is a simple representation of selection. Currently the selective pressure is applied only to extreme genome sizes, removing all the genomes with number of genes below the minimal and higher then the maximal threshold. To keep the number of genomes constant, remaining genomes are selected at random.

The resulting system is designed to emulate the three major processes in genome evolution: gene genesis, gene loss and horizontal gene transfer. Knowledge of the exact order of events in the simulated evolutionary scenario allows the rigorous testing of computational methods addressing genome evolution and gene content phylogeny reconstruction.

Availability

The programs are available from the authors upon request.

Results

Gene content phylogeny

We used SimulEv (see methods) to produce simulated evolutionary scenarios. We first aimed to find how well gene content-based tree reconstruction (Snel *et al.*, 1999) would restore the topology of the real tree. To achieve this, we generated phylogenetic profiles with multiple simulation experiments under various evolutionary scenarios using SimulEv. The trees generated by the reconstruction method were then compared to the correct trees, as recorded by SimulEv.

A typical example of such simulation is shown in Fig. 2. This example illustrates the generic properties of tree reconstruction as observed in many evolutionary scenarios. First, the branching order of terminal branches is usually reconstructed correctly. This leads to the conclusion that gene content trees are suitable for genome reconstruction at short to medium evolutionary distances, as was previously reported elsewhere (Korbel *et al.*, 2002). Branch length is also estimated with adequate precision at these evolutionary distances.

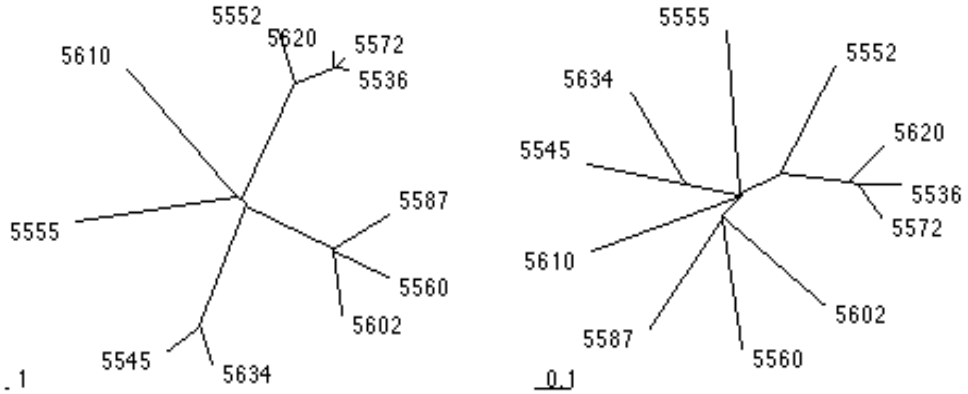
However, long branch attraction phenomena are often observed near the root, where gene content-based trees fail to reconstruct the correct branching order or branch length (Fig. 2). Incorrect nodes are readily distinguished, because they usually have low bootstrap values (not shown). The inability of this approach to reconstruct deep branching patterns correctly may explain the differences observed in deep prokaryotic phylogeny between trees based on the small rRNA subunit (Maidak *et al.*, 2001) and gene content (Snel *et al.*, 1999).

Table 1. The performance of the GENETRACE method as tested with SimulEv simulated evolution

Generations number	Gene genesis	HGT	Starting genome size	TRUE	False Negatives	False Positives	% Sensitivity	% Selectivity
30	5	25	520	27310	1232	3361	95.7	89.0
30	5	25	2950	146530	204	3758	99.9	97.5
30	5	25	1500	76449	382	3900	99.5	95.1
30	10	20	520	27118	1247	1741	95.6	94.0
30	10	20	1500	74576	491	2039	99.3	97.3
30	10	20	2950	146679	270	2681	99.8	98.2
30	15	15	520	26918	1411	1378	95.0	95.1
30	15	15	1500	74753	450	1576	99.4	97.9
30	15	15	2950	145852	204	1392	99.9	99.1
30	20	10	520	25956	1555	683	94.3	97.4
30	20	10	1500	74607	560	1137	99.3	98.5
30	20	10	2950	146463	300	987	99.8	99.3
30	25	5	520	26021	1850	400	93.4	98.5
30	25	5	1500	75839	1090	420	98.6	99.4
30	25	5	2950	146921	330	507	99.8	99.7
50	5	25	520	26840	2658	3297	91.0	89.1
50	5	25	1500	75604	926	4232	98.8	94.7
50	5	25	2950	144246	570	5113	99.6	96.6
50	10	20	520	25485	2963	2257	89.6	91.9
50	10	20	1500	74897	1434	3440	98.1	95.6
50	10	20	2950	145530	603	3383	99.6	97.7
50	15	15	520	26632	2676	1290	90.9	95.4
50	20	10	520	26914	2901	932	90.3	96.7
50	25	5	520	26389	2894	366	90.1	98.6
100	5	25	520	28835	4003	3866	87.8	88.2
100	10	20	520	27077	3878	2582	87.5	91.3

Columns are divided by a vertical line. On the left, the input parameters for SimulEv are given: number of generations, the extent of gene genesis and HGT (see text) and starting genome size, respectively. In all examples, the extent of gene loss is constant and has been set to 30, counterbalanced by the sum of gene genesis and HGT. On the right, the performance measures for GENETRACE are shown, including the number of true cases, false (negative and positive) cases and the percent sensitivity and selectivity measures calculated from the corresponding values of the false cases, respectively.

GO BACK
CLOSE FILE



a b

Fig. 2. Comparison of correct tree as reported by SimulEv (a), and a tree reconstructed with gene content phylogenetic reconstruction (b). Numbers represent simulated 'extant' genomes. Note the correctness of terminal branches order and the wrong deep branching. The SimulEv parameters used were: 60 generations, 100 loss, 50 HGT, 50 gene genesis 51 genomes, 1500 genes.

Specificity/selectivity analysis of GeneTrace

We subsequently aimed to analyse the performance of the GENETRACE method. The difficulty was that SimulEv simulations are subject to numerous

GO BACK
CLOSE FILE

parameters, such as the size of the founder genome, the minimal and the maximal allowed genome sizes, the ratio between the HGT and the gene genesis, the amount of genome turnover per generation and the number of generations. Despite the fact that we have previously reported the estimation of relative contribution of gene loss, gene genesis and horizontal gene transfer in the evolution of prokaryotes (Kunin and Ouzounis, submitted), we used these estimations only as parameters to the GENETRACE reconstruction and not as the input parameters to SimulEv. Instead, we tested a larger parameter space to investigate whether the GENETRACE predictions are correct for any SimulEv parameters, and thus under various evolutionary scenarios. To avoid continuous growth or shrinking of simulated genomes, the parameters were set to keep the sum of average gene gain and loss to be constant and set to zero.

To assess the reliability of GENETRACE reconstruction, we executed SimulEv with various parameters, and stored the gene content at every node of the simulated evolutionary tree. Parameters of GENETRACE were fixed to the GAIN penalty of 2 and LOSS penalty of 4. The correct tree and the phylogenetic profiles of the top level (the 'extant species') of the SimulEv simulation were submitted to GENETRACE. The phylogenetic profiles restored by GENETRACE were compared to the correct phylogenetic profiles as reported by SimulEv. The selectivity (number of correct hits from total—or a measure of false positives) and the sensitivity (fraction of genes reported in the node by SimulEv detected by GENETRACE—or a measure of false negatives) were calculated (Table 1). The results suggest that on any SimulEv input parameters both the selectivity and the sensitivity of GENETRACE were above 90%.

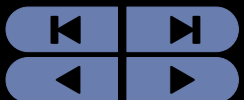
SimulEv simulations aimed to find the validity of our approach on exactly known, hypothetical evolutionary scenarios. However, we expect that the input data coming from genome projects could be of different quality in several aspects. First, automatic clustering methods may produce errors leading to incorrect phylogenetic profiles. Depending on sequence divergence, these methods can exclude genuine orthologs from the cluster, or include functionaly divergent proteins into a family cluster. Second, the trees generated by phylogenetic reconstruction data may also contain errors, limiting the scope of conclusions. As most trees generated by current methods are unlikely to reproduce accurate deep branching for the region close to the root, we recommend using several trees derived from different methods, and find consensus predictions before drawing any conclusions. However, assuming correct input information, GENETRACE is expected to provide at least 90% of accurate predictions (Table 1).

Discussion

We have been exploring the calibration of the observed relative frequency of processes forming gene content of prokaryotic genomes (Kunin and Ouzounis, submitted). To achieve this, we tracked the evolutionary history of 12 762 protein families obtained from 51 complete genomes using Tribe-MCL clustering (Enright *et al.*, 2002) and both 16S rRNA and gene content-based trees.

Our results indicate that gene loss is observed two or three times more frequently then horizontal gene transfer, suggesting a possible value range for the threshold of GENETRACE. Large-scale genomic analysis with this threshold confirmed previous reports of reductive evolution of pathogenic bacteria (Moran, 2002) and gain of multiple genes by metabolically versatile organisms such as *Pseudomonas aureginosa* and *Mesorhisobium loti*. The parameters were then verified on a set of strain-specific genes of *Helicobacter pylori*, suggesting the accuracy of the predictions to be more than 80% (Kunin and Ouzounis, submitted).

There are several directions to improve our method. First, the current implementation does not implement correction for branch length, providing the most parsimonious scenario instead. Phylogenetic studies of DNA and protein sequences suggest that maximum likelihood methods are usually superior to parsimony methods that do not consider branch length (Wiens and Servedio, 1999). Furthermore, evolution of individual clades or certain protein families may differ from the mean and require parameters specific to a clade or a protein family. Finally, it would be interesting to compare our approach with methods using ortholog-based phylogenetic profiles, when they become available.

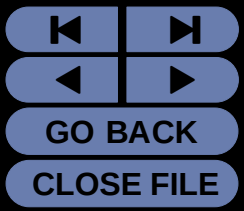


Acknowledgements

We thank members of the Computational Genomics Group for discussions. This work was supported by the European Molecular Biology Laboratory. C. O. thanks the UK Medical Research Council and IBM Research for additional support.

Note added in Proof

After submission of this manuscript, an independently developed algorithm for the reconstruction of ancestral states has also appeared (Mirkin *et al.*, 2003).



References

Enright,A.J., Van Dongen,S. and Ouzounis,C.A. (2002) An efficient algorithm for large-scale detection of protein families. *Nucleic Acids Res.*, **30**, 1575–1584. MEDLINE Abstract

Korbel,J.O., Snel,B., Huynen,M.A. and Bork,P. (2002) SHOT: a web server for the construction of genome phylogenies. *Trends Genet.*, **18**, 158–162. MEDLINE Abstract

Liberles,D.A., Thorén,A., von Heije,G. and Elofsson,A. (2002) The use of phylogenetic profiles for gene predictions. *Curr. Genomics*, **3**, 131–137.

Maidak,B.L., Cole,J.R., Lilburn,T.G., Parker,C.T. and Jr.Saxman,P.R. *et al.* (2001) The RDP-II (Ribosomal Database Project. *Nucleic Acids Res.*, **29**, 173–174. MEDLINE Abstract

Marcotte,E.M., van Der Bliek,A.M. and Eisenberg,D. (2000) Localizing proteins in the cell from their phylogenetic profiles. *Proc. Natl Acad. Sci. USA*, **97**, 12115–12120. MEDLINE Abstract

Mirkin,B.G., Fenner,T.I., Galperin,M.Y. and Koonin,E.V. (2003) Algorithms for computing parsimonious evolutionary scenarios for genome evolution, the last universal common ancestor and dominance of horizontal gene transfer in the evolution of prokaryotes. *BMC Evol. Biol.*, **3**.

Moran,N.A. (2002) Microbial minimalism: genome reduction in bacterial pathogens. *Cell*, **108**, 583–586. MEDLINE Abstract

Pellegrini,M., Marcotte,E.M., Thompson,M.J., Eisenberg,D. and Yeates,T.O. (1999) Assigning protein functions by comparative genome analysis: protein phylogenetic profiles. *Proc. Natl Acad. Sci. USA*, **96**, 4285–4288. MEDLINE Abstract

Pupko,T., Pe'er,I., Shamir,R. and Graur,D. (2000) A fast algorithm for joint reconstruction of ancestral amino acid sequences. *Mol. Biol. Evol.*, **17**, 890–896. MEDLINE Abstract

Snel,B., Bork,P. and Huynen,M.A. (1999) Genome phylogeny based on gene content. *Nat. Genet.*, **21**, 108–110. MEDLINE Abstract

Snel,B., Bork,P. and Huynen,M.A. (2002) Genomes in flux: the evolution of archaeal and proteobacterial gene content. *Genome Res.*, **12**, 17–25. MEDLINE Abstract

Tatusov,R.L., Koonin,E.V. and Lipman,D.J. (1997) A genomic perspective on protein families. *Science*, **278**, 631–637. MEDLINE Abstract

Vert,J.-P. (2002) A tree kernel to analyse phylogenetic profiles. *Bioinformatics*, **18**, S276–S284. MEDLINE Abstract

Wiens,J.J. and Servedio,M.R. (1999) Phylogenetic analysis and intraspecific variation: performance of parsimony, likelihood, and distance methods. *Syst. Biol.*, **48**, 228–253.