Resource

Cell

Expansion of Biological Pathways Based on Evolutionary Inference

Yang Li,[1,2,6] Sarah E. Calvo,[1,3,6] Roee Gutman,[4] Jun S. Liu,[2,*] and Vamsi K. Mootha[1,3,5,*]
[1]Howard Hughes Medical Institute and Department of Molecular Biology, Massachusetts General Hospital, Boston, MA 02114, USA
[2]Department of Statistics, Harvard University, Cambridge, MA 02138, USA
[3]Broad Institute, Cambridge, MA 02141, USA
[4]Department of Biostatistics, Brown University, Providence, RI 02912, USA
[5]Department of Systems Biology, Harvard Medical School, Boston, MA 02115, USA
[6]Co-first author
*Correspondence: jliu@stat.harvard.edu (J.S.L.), vamsi@hms.harvard.edu (V.K.M.)
http://dx.doi.org/10.1016/j.cell.2014.05.034

SUMMARY

The availability of diverse genomes makes it possible to predict gene function based on shared evolutionary history. This approach can be challenging, however, for pathways whose components do not exhibit a shared history but rather consist of distinct "evolutionary modules." We introduce a computational algorithm, clustering by inferred models of evolution (CLIME), which inputs a eukaryotic species tree, homology matrix, and pathway (gene set) of interest. CLIME partitions the gene set into disjoint evolutionary modules, simultaneously learning the number of modules and a tree-based evolutionary history that defines each module. CLIME then expands each module by scanning the genome for new components that likely arose under the inferred evolutionary model. Application of CLIME to ~1,000 annotated human pathways and to the proteomes of yeast, red algae, and malaria reveals unanticipated evolutionary modularity and coevolving components. CLIME is freely available and should become increasingly powerful with the growing wealth of eukaryotic genomes.

INTRODUCTION

Biological pathways and complexes represent the fruits of extensive pruning, expansion, and mutation that have occurred over evolutionary timescales. For example, mitochondria represent a defining feature of all eukaryotes, yet an estimated one-half of the organelle's ancestral machinery has been lost (Vafai and Mootha, 2012), and the remaining machinery varies significantly across eukaryotic taxa, with many new lineage-specific innovations. Similarly, cilia were likely present in the last common eukaryotic ancestor, though most plants and fungi lost this organelle completely, whereas nematodes have specifically lost motile cilia. Charting the evolutionary history of modern-day pathways and complexes can help to define the taxonomic distribution of pathways and thereby highlight model organisms for experimental studies. Such evolutionary analyses may also teach us about the environmental niches within which they evolved. Importantly, correlated gains and losses can help to predict the function of unstudied genes and also reveal alternative functions even for genes considered to be well characterized.

Pioneering work introduced the concept of "phylogenetic profiling" to chart the phylogenetic distribution of genes and relate them to each other (Pellegrini et al., 1999). In this approach, a binary vector of presence and absence of a given gene across sequenced organisms is used to predict function of genes sharing a similar profile, based on the Hamming distance (Hamming, 1950). A number of different computational methods have been developed (Kensche et al., 2008) and have been applied successfully to predict components for prokaryotic protein complexes (Pellegrini et al., 1999); phenotypic traits like pili, thermophily, and respiratory tract tropism (Jim et al., 2004); cilia (Li et al., 2004); mitochondrial complex I (Ogilvie et al., 2005); and small RNA pathways (Tabach et al., 2013).

Although many phylogenetic profiling algorithms are now available, several features limit their utility (Kensche et al., 2008). First, most existing methods compare an input gene to a query gene one at a time—which cannot take advantage of patterns only discernible by analyzing a collection of input genes. Second, most methods do not explicitly model errors in a gene's phylogenetic profile, each of which may be individually noisy due to the inherent challenges of genome assembly, gene annotation, and detection of distant homologs (Trachana et al., 2011). Third, with a few notable exceptions (Barker and Pagel, 2005; von Mering et al., 2003; Vert, 2002; Zhou et al., 2006), most existing algorithms do not take into account the phylogenetic tree of the input species but assume independence across species and hence are highly sensitive to the choice of organisms selected. Available tree-based methods are computationally intensive and not readily scalable to large genomes (Barker et al., 2007; Barker and Pagel, 2005).

Because most existing phylogenetic profiling methods are designed to operate on single genes, they cannot be readily extended to biological pathways, where each member may have different phylogenetic profiles. Our previous experience

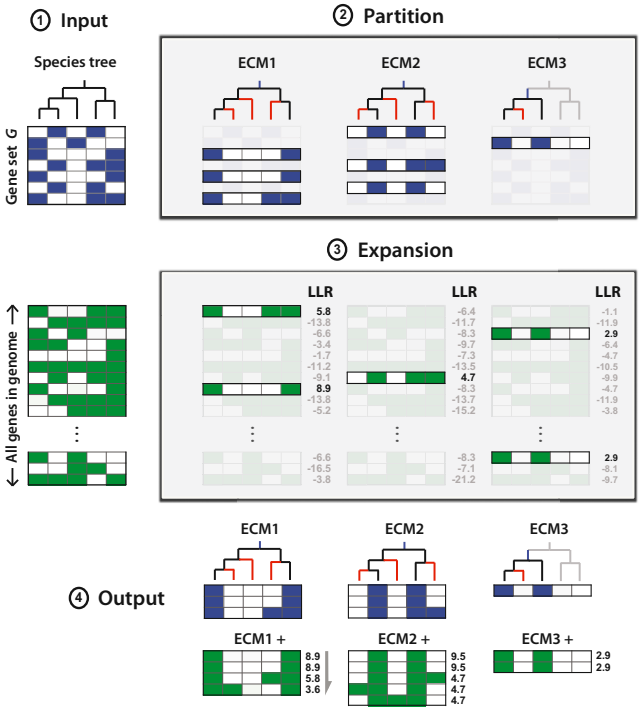

Figure 1. Schematic Overview of CLIME

CLIME partitions an input set of genes into evolutionarily conserved modules (ECMs) and predicts additional genes sharing the same inferred model of evolution. Input: species tree, an input gene set (*G*), and a phylogenetic matrix (*X*) for all genes in a reference organism showing presence (green) or absence (white) across all extant species in the tree. For display purposes, a separate blue/white matrix shows the profiles of genes in *G*, which are a subset of *X*. Partition: input genes *G* are partitioned into *K* distinct ECMs, using a Bayesian mixture of HMMs to simultaneously infer the number of ECMs and the shared evolutionary history of each ECM. Each ECM is modeled by a tree-structured HMM with an inferred gain branch (blue) and branch-specific probabilities of gene loss (red). Expansion: each ECM is expanded by identifying genes within the genome that are more likely to have evolved from the ECM's model of evolutionary history compared to a null model of evolution, scored by the log-likelihood ratio (LLR). Output: *K* disjoint ECM clusters and associated ECM+ expansions.
See also Figures S1, S2, S3, S4, and S5.

with mitochondrial complex I illustrates this point (Pagliarini et al., 2008). Human complex I is a macromolecular machine consisting of 44 structural subunits. We observed that these subunits did not share a single, common history of gains and losses across eukaryotic evolution but clustered into several distinct evolutionary modules. One "ancestral" module consisted of 14 core subunits that were present in bacteria and in humans yet lost independently four times in eukaryotic evolution, whereas other modules consisted of recent animal or vertebrate innovations. By first identifying the "ancestral" module, we could scan the human genome to identify additional genes sharing the same evolutionary history. Five of these genes have since been shown to encode complex I assembly factors that are mutated in inherited complex I deficiencies (Mimaki et al., 2012).

Our previous analysis suggested that biological pathways, as we conceive of them, represent mosaics of gene modules, each sharing a coherent pattern of evolutionary gains and losses. If such modules can be detected accurately, they can then be "expanded" to identify new components. The major challenge in accurate detection is that the number and histories of modules have to be inferred simultaneously.

Here, we introduce a method that generalizes this approach in a statistically principled manner, using a Bayesian mixture of tree-based hidden Markov models. Our method, called *cl*ustering by *i*nferred *m*odels of *e*volution (CLIME), first partitions an input gene set into modules of genes that exhibit coherent evolutionary histories and then expands each module with new genes sharing the same evolutionary history. CLIME is distinct from existing approaches in that it (1) is a tree-based method for partitioning an input set of related genes, (2) automatically learns the number of distinct evolutionary modules in the input set, and (3) leverages information from the entire input gene set to more reliably predict new genes that have arisen with a shared pattern of evolutionary gains and losses.

We systematically applied CLIME to over 1,000 human complexes and pathways, two human cellular organelles (cilia and mitochondria), and three entire genomes (red algae, yeast, and the malaria parasite). The results, the software, and an online analysis portal are freely available at http://www.gene-clime.org.

RESULTS

CLIME: An Algorithm for Clustering Genes Based on Inferred Models of Evolution

The CLIME algorithm partitions genes based on inferred models of evolution (Figure 1). CLIME accepts three user-defined inputs: (1) a binary species tree; (2) a phylogenetic profile matrix, *X*, defining the presence or absence of all genes in a given organism across all species in the tree; and (3) an input gene set *G*. CLIME partitions the input set *G* into disjoint evolutionarily conserved modules (ECMs) using a Bayesian mixture model to infer simultaneously the number of ECMs, the evolutionary model for each ECM, and gene's membership for each ECM. The algorithm next creates an ECM expansion set, ECM+, that includes other genes in the genome that are likely to have arisen under the ECM's inferred model of evolution compared to a null model.

CLIME models the evolution of an individual gene using a tree-based hidden Markov model (HMM), with the assumption that each gene has a single gain event in evolution followed by zero or more loss events on the species tree (Figures 2A and 2B). CLIME does not consider branch lengths, only the tree topology. For each gene *g*, the HMM of evolution is based on the presence/absence profile across *S* living species (X_g, the observed states). The HMM contains 2*S*-1 hidden states (H_g) corresponding to the true presence/absence of that gene in all living and extinct species (Figure 2B). The model includes a user-defined observation error parameter ε (default 0.01) representing the probability that the observed data are errors compared to the true hidden presence/absence (e.g., incomplete genome assembly/annotation). CLIME infers a tree-based HMM to model the evolution of each gene separately, as well as to model the evolution of each ECM. The evolutionary model of each gene *g* is represented by a single gain branch (λ_g) and a vector of branch-specific loss probabilities of its ECM (θ_k)—inferred at the preprocessing step and partition step,