BIOINFORMATICS ORIGINAL PAPER

Vol. 24 no. 5 2008, pages 606–612
doi:10.1093/bioinformatics/btn005

Genome analysis

Reconstructing ancestral genome content based on symmetrical best alignments and Dollo parsimony

Onur Sakarya[1,2], Kenneth S. Kosik[2,3,*] and Todd H. Oakley[4,*]

[1]Department of Computer Science, [2]Neuroscience Research Institute, [3]Department of Molecular, Cellular and Developmental Biology and [4]Department of Ecology, Evolution and Marine Biology, University of California, Santa Barbara, CA 93106, USA

Received on October 3, 2007; revised on December 13, 2007; accepted on January 3, 2008

Advance Access publication January 9, 2008

Associate Editor: Alex Bateman

ABSTRACT

Motivation: Gene duplications and losses (*GDLs*) are important events in genome evolution. They result in expansion or contraction of gene families, with a likely role in phenotypic evolution. As more genomes become available and their annotations are improved, software programs capable of rapidly and accurately identifying the content of ancestral genomes and the timings of *GDLs* become necessary to understand the unique evolution of each lineage.

Results: We report EvolMAP, a new algorithm and software that utilizes a species tree-based gene clustering method to join all-to-all symmetrical similarity comparisons of multiple gene sets in order to infer the gene composition of multiple ancestral genomes. The algorithm further uses Dollo parsimony-based comparison of the inferred ancestral genes to pinpoint the timings of *GDLs* onto evolutionary intervals marked by speciation events. Using EvolMAP, first we analyzed the expansion of four families of G-protein coupled receptors (GPCRs) within animal lineages. Additional to demonstrating the unique expansion tree for each family, results also show that the ancestral eumetazoan genome contained many fewer GPCRs than modern animals, and these families expanded through concurrent lineage-specific duplications. Second, we analyzed the history of *GDLs* in mammalian genomes by comparing seven proteomes. In agreement with previous studies, we report that the mammalian gene family sizes have changed drastically through their evolution. Interestingly, although we identified a potential source of duplication for 75% of the gained genes, remaining 25% did not have clear-cut sources, revealing thousands of genes that have likely gained their distinct sequence identities within the descent of mammals.

Availability: Query server, source code and executable are available at http://kosik-web.mcdb.ucsb.edu/evolmap/index.htm

Contact: kosik@lifesci.ucsb.edu, oakley@lifesci.ucsb.edu

Supplementary information: Supplementary data are available at *Bioinformatics* online.

1 INTRODUCTION

A robust understanding of gene family evolution can leverage knowledge of gene function in one species to allow predictions in other species. Gene family history may also indicate specific mutational events responsible for phenotypic changes during evolution. At the heart of gene function prediction is distinguishing orthologous and paralogous gene pairs. A pair of orthologous genes are derived from a speciation event (Fitch, 1970; Ingram, 1961) whereas a pair of paralogous genes are derived from a mutational event that copied a gene within an ancestral genome (a gene duplication event). Orthologous genes often retain similar functions, whereas paralogous genes are more likely to diverge in function (Koonin, 2005). Gains or losses in gene copy number in genomes may be associated with the evolution of phenotypic traits (Ohno, 1970; Plachetzki and Oakley, 2007). Primarily for these reasons, gene family evolution has received considerable attention in bioinformatics, where numerous methods have been developed.

Most methods to understand gene history can be classified as similarity-based or tree-based methods. 'Similarity-based' methods primarily rely on the relative similarity between pairs of genes, without explicitly utilizing information from a phylogeny of the species containing the genes of interest. For example, a common approach to understanding gene family history uses the symmetrical best scores from all-to-all BLAST alignments, which have been effectively used for detection of orthologs from two given genomes (Li *et al*., 2003; O'Brien *et al*., 2005; Wall *et al*., 2003). When only two species are involved, their species-level relationships are known. Once genes from more than two species are analyzed, reliable orthology detection becomes more challenging because gene duplications and losses (*GDLs*) are common during evolution, and gene similarity is correlated with, but does not always reflect common ancestry. Different similarity-based algorithms, including COG/KOG (Tatusov *et al*., 2003), InParanoid/MultiParanoid (Alexeyenko *et al*., 2006; Remm *et al*., 2001), OrthoMCL (Li *et al*., 2003) and Roundup (Deluca *et al*., 2006) deal with these challenges in different ways. Other ortholog identification methods use 'tree-based' approaches, including RIO (Zmasek and Eddy, 2002), Orthostrapper (Storm and Sonnhammer, 2002), TreeFam (Li *et al*., 2006) and OrthologID (Chiu *et al*., 2006). These methods explicitly consider both gene and species history, in some cases simultaneously, to estimate reconciled trees (Durand *et al*., 2006). A challenge for tree-based methods is that they rely on correct estimation of both

*To whom correspondence should be addressed.

gene tree and species tree and require extensive amounts of computer time, often making genome scale analyses impractical or impossible.

Two other recent approaches, implemented in CAFE and SYNERGY use elements of both similarity-based and tree-based approaches (De Bie *et al.*, 2006; Wapinski *et al.*, 2007). CAFE primarily uses information about the numbers of genes in precomputed orthologous families, and does not simultaneously consider gene family phylogeny or similarity among genes. SYNERGY employs a novel algorithm to first find families of sufficiently similar genes and then to count *GDLs* by considering their gene family phylogeny trees.

In this contribution, we present EvolMAP software, which implements a new algorithm to estimate the gene content of hypothetical ancestral species and to find the timings of gene duplication and loss events, relative to evolutionary transitions marked by speciation events. Unlike De Bie *et al.* (2006) but like the tree-based approaches mentioned earlier, our algorithm simultaneously considers information on species history and gene history. However, unlike many of the tree-based methods above, our approach does not perform the computationally costly calculations required to explicitly estimate both species phylogeny and gene phylogeny.

Below, we explain our approach in detail, report availability of software for making the calculations, and analyze multiple datasets to allow comparison with previous methods. First, we analyze the history of G-protein coupled receptors (GPCRs) in animals. GPCRs constitute a diverse superfamily of proteins involved in signal transduction pathways and are found in many copies in animal genomes. An important question is whether those genes existed in high copy numbers in the early animal ancestors or were expanded subsequently in separate lineages. Through our analyses, we demonstrate the unique expansion tree for each GPCR family and find that eumetazoan and bilaterian ancestors had much fewer GPCRs than modern animals. Second, we reanalyze the history of mammalian *GDLs* which was studied previously under a gene family framework (Demuth *et al.*, 2006). In addition to reaching a higher resolution of gene history by calculating and comparing genome composition of multiple ancestral genomes, we include two more species in the analysis and use more recently annotated gene sets. With that, we obtain results that concur qualitatively with Demuth *et al.* (2006) but differ quantitatively. Like them, we also report that mammalian gene family sizes have changed drastically during evolution, including numerous changes in gene number since the split between humans and chimps. Finally, we compare our results from five genome-wide analyses against those of SYNERGY, InParanoid and MultiParanoid.

2 METHODS

For a given set of species-specific sequences and a species tree, we describe an all-to-all similarity-clustering algorithm to reconstruct the ancestral genes of each ancestral node of the tree and pinpoint the timings of *GDLs*. In the first section, we describe a new similarity metric which is based on normalized, global optimal pair-wise alignments. In the second section, we explain how the orthologous groups are detected for ancestral nodes, which is essentially based on post-order traversal of the tree (from descendants to the root) while performing all-to-all similarity comparisons of each node's nearest descendants (*NDs*). In the Section 3, we explain how the tree is traversed back from root to descendants to find Dollo parsimonious timings of gene gains and losses. In the final sections, we present two new species topology-based trees that are generated as result, and summarize the implementation of the algorithm.

2.1 Normalized similarity scoring

Many ortholog detection algorithms use very fast scoring schemes. For example, all-against-all BLASTP (Blastall) searches offer high-speed detection of homologous sequence segments (Altschul *et al.*, 1997). However, BLAST often generates multiple, overlapping hits between a pair of genes. Any one of these overlapping hits may often represent inaccurate estimates of overall similarity and joining them together is complicated. In contrast to very rapid similarity scoring, dynamic programming algorithms—like Needleman–Wunsch (*NW*) optimal global alignments—offers a highly accurate scoring metric (Needleman and Wunsch, 1970), but is slow in comparison to BLAST. As a compromise between speed and accuracy, we first find the most similar gene pairs using BLAST, then generate more accurate similarity scores of those most similar genes by pairwise *NW* alignments assuming a BLOSUM62 scoring matrix (Henikoff and Henikoff, 1992). Gap extensions are minimally penalized to treat alignments of genes with different splice variants or with lost or recombined domains similarly to longer but evenly homologous alignments. Following standard *NW* alignments, similarity scores are normalized using:

$$NSS(i,j)=\frac{NWS(i,j)/GAL(i,j)}{Min(NWS(i,i),NWS(j,j))/Min(GAL(i,i),GAL(j,j))} \quad (1)$$

where $NSS(i,j)$ is the normalized similarity score between sequences i and j; NWS are NW alignment scores and GAL are gap-free alignment lengths where $GAL(i,i)$ and $GAL(j,j)$ are equal to length of the sequences i and j, respectively. If $NSS(i,j)$ returns a negative score, it is set to 0. With that, Equation (1) defines globally comparable similarity scores that range from 0 to 1 (1 being most similar). In the implementation, we multiply the final score by 1000 to have a more memory efficient, integer representation of the score.

2.2 Detecting ancestral genes using symmetrical best alignments

Orthologous genes are commonly identified between a pair of genomes as symmetrical best alignments that score above a minimum threshold (sym-bet) (Remm *et al.*, 2001). Sequences within the same genome that align better to a sym-bet gene than its ortholog are defined as 'in-paralogs' (Sonnhammer and Koonin, 2002). In this study, we employ these concepts and use the terms 'ancestral gene' and 'orthologous group' interchangeably to refer to a sym-bet and its in-paralogs. Therefore, once identified, an orthologous group is assumed to include the detectable descendants of a single, hypothetical ancestral gene of a given ancestral node (Fig. 1A).

In the first phase of the algorithm, the species tree is traversed in post-order (from descendants to root). For each ancestral node, sequences of its two NDs are compared in an all-to-all fashion using the NSS matrix to find sym-bets and their in-paralogs. In cases where the NDs are also ancestral nodes, similarity scores between their ancestral genes are calculated as the average similarity of members of one ancestral gene to those of the other. Genes that are neither a sym-bet nor an in-paralog of another sym-bet is defined as a 'singular' gene of the ancestral node. Those singular genes are considered as ambiguous ancestral genes at this stage where there is no evidence from the pairwise comparisons that the gene was present in that ancestral node. Through the second phase of the algorithm, earlier ancestral nodes are rigorously scanned for any such evidence, and the ambiguity is further resolved.

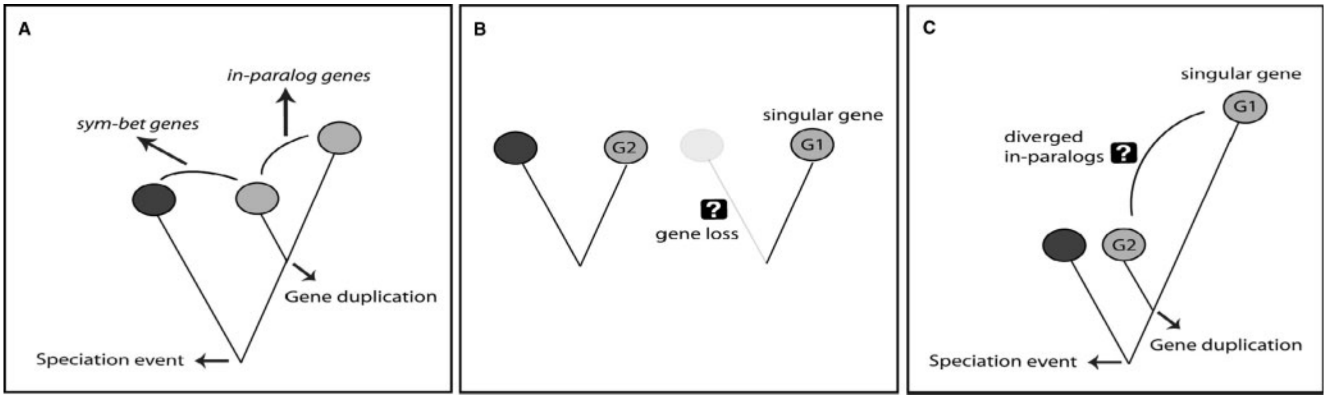

Fig. 1. Detecting ancestral genes. (**A**) Sym-bets are determined as symmetrical best aligning genes between two genomes, and in-paralogs as genes within the same genome that align better to a sym-bet gene than its ortholog. Together, a sym-bet and its in-paralogs represent an orthologous group whose members descended from a single ancestral gene locus present at the time of speciation. (**B**) Remaining (singular) genes might be lost from one of the genomes or (**C**) they might have diverged faster than their orthologs and are no longer recognizable as an in-paralog.

2.3 Detecting gene gains and losses using Dollo parsimony

In the second phase of the algorithm, the species tree is traversed in preorder (root to descendants) to better estimate the composition of the ancestral genomes and detect the timings of gene gains and losses. If a gene is ambiguously present (singular) in a given ancestral node and present (member of an orthologous group) in one of its *ND*s, then it was either present in this ancestral node and lost in the branch towards nearest descendant (*BTND*) missing the gene (Fig. 1B), or it was not present as an individual loci in this node and was duplicated in the other *BTND* and diverged quickly and lost its in-paralog signal (Fig. 1C). Comparing the corresponding genes of earlier ancestors may help clarify this ambiguity. Since the loss of an existing gene is more likely than a parallel gain of the same gene in another genome (ignoring horizontal gene transfers), a Dollo parsimony model (Farris, 1977) is fast and appropriate for inferring if and when a gene was gained or lost. For each singular, orthologous groups of all earlier ancestral nodes are searched to see if it was already inferred ‘present’ in any of those earlier ancestors. If such a ‘present’ orthologous group is found, and it does not include other paralogs from the *ND* missing the gene (i.e. it is not a larger group containing multiple paralogs), then the gene is inferred to be present in this ancestor and lost in the *BTND* missing the gene by the Dollo parsimony criterion (Fig. 2A). If none of the ancestors have such an orthologous group, then according to the Dollo criterion the gene is inferred ‘not present’ in this ancestor and is gained in the *BTND* containing the gene (Fig. 2B).

Gained genes are further categorized into one of the three groups of *in-paralogs*, *diverged in-paralogs* or *ambiguous gains*. *In-paralogs* are detected as genes that are inferred present in the ancestor as descendants of a single gene, but split into at least two genes in the descendant. They are essentially paralogous genes that were conserved better than the orthologs after the duplication event. *Diverged in-paralogs* are defined as singular genes of an ancestor, which are present in the descendant, and they have significant sequence similarity above a user defined statistical threshold to another present gene of the descendant. They are essentially genes that have likely duplicated in the given branch and diverged under relaxed selection. *Ambiguous gains* are defined as singular genes of the ancestor, which are inferred present in the descendant and do not have significant sequence similarity above the user defined threshold to other genes of the descendant. Such genes might have experienced very high rates of mutation that erased their history, horizontally transferred from another genome, lost in all earlier lineages or could just be dataset artifacts.

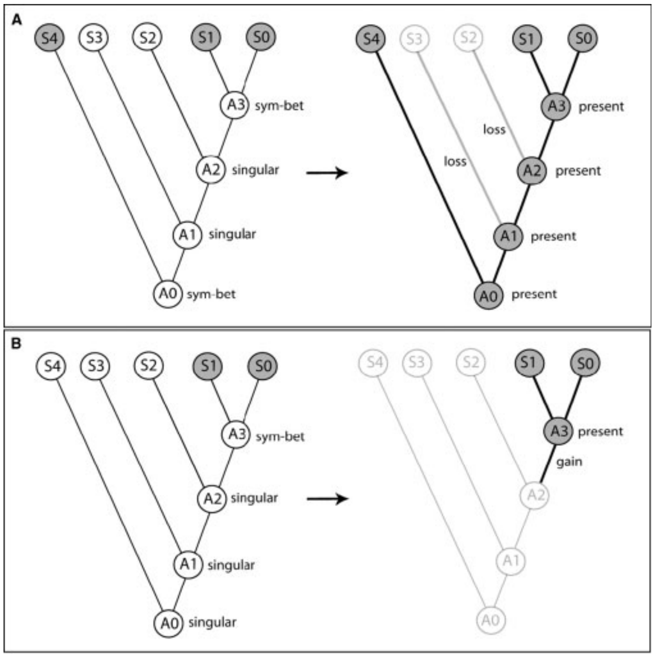

Fig. 2. Identifying timings of gene gains and losses based on Dollo parsimony model. (**A**) Assuming a hypothetical case where a sym-bet detected at A0 including all members of a sym-bet at A3, and that sym-bet from A0 not including any genes from the species S2 and S3, Dollo parsimony infers two gene losses at lineages descending from A2 to S2 and A1 to S3. (**B**) As another hypothetical case, if a sym-bet detected at A3 is a singular in all earlier nodes or is an in-paralog of an earlier orthologous group that includes other sym-bets from A3, a gene gain event is inferred for lineages descending from A2 to A3.

2.4 Average ortholog divergence and gene expansion trees

As a summary of results, two species topology-based trees are generated. A first tree, called the average ortholog divergence (*AOD*) tree, displays the average distance of orthologs of ancestral genes as the length of each branch of the species tree. A second tree, called the ‘gene expansion’ tree, displays branch lengths as absolute number of genes

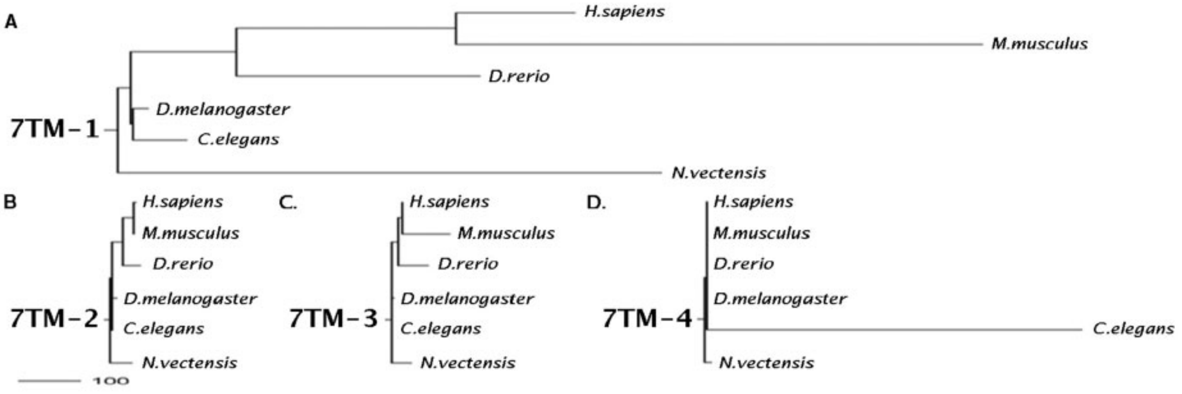

Fig. 3. Expansion of GPCR-coding genes. Gene expansion trees were generated using EvolMAP with all-to-all *NSS* matrix for genes containing selected GPCR domains of (**A**) *7tm_1* (Opsin-rhodopsin family) (**B**) *7tm_2* (Secretin family) (**C**) *7tm_3* (metabotropic glutamate receptor family) and (**D**) *7tm_4* (chemoreceptor family). For each analysis, all-to-all *NSS* score matrices were used to find ancestral genes with the minimum ortholog threshold set as 250. Scale bar corresponds to hundred gene gains.

gained along each branch. In this case, absolute gene gain is defined as total number of gene gains minus losses. *AOD* for each ancestral node *A* of the tree is derived by the following formula:

$$AOD(A) = 1 - \left(\frac{\sum_{\text{for each sym}-bet(i,\,j)\text{ of }A} NSS(i,j)}{\text{number of sym-bets }(A)} \right) \tag{2}$$

For species nodes, *AOD* is defined as 0. From this equation, lengths for each branch (A→D) of the *AOD* tree are calculated using the equation:

$$AOD(A \to D) = \frac{AOD(A) - AOD(D)}{2} \tag{3}$$

Number of present genes (*NPG*) of ancestral nodes is used to find the branch lengths of the gene expansion tree where branch lengths are derived by the following formula (with negative branch lengths set to 0):

$$GE(A \to D) = NPG(D) - NPG(A) \tag{4}$$

2.5 Implementation and data collection

EvolMAP is implemented using Java 1.6.0 in a package called EvolMAP. It will work under most operating systems running a Java Virtual Machine. Source codes and an executable (JAR) file are available. The JAR file may be used from command line with a simple options file and a graphical user interface (GUI) is also provided. Basic input to EvolMAP is a binary newick format species tree and fasta files containing sequences from the corresponding species. Once an analysis is complete, inferred ancestral genes (and their descendants) of each ancestral node are reported as results. Other results include numbers of *in-paralogs*, *diverged in-paralogs*, *ambiguous gains* and gene losses for each branch in table format, and *AOD*, gene expansion, gene gain and loss trees in newick format. An HTTP query server is also provided to search for ancestral genes and their descendants within the completed analyses. For more information on implementation, see user's manual at EvolMAP web page.

Protein peptide sequences of *Homo sapiens* (human), *Pan troglodytes* (chimp), *Macaca mulatta* (monkey), *Mus musculus* (mouse), *Rattus norvegius* (rat), *Canis familiaris* (dog), *Monodelphis domestica* (opossum), *Danio rerio* (zebrafish), *Caenorhabditis elegans* (worm) and *Drosophila melanogaster* (fly) were downloaded from Ensembl FTP site (release 46) (ftp://ftp.emsembl.org). *Nematostella vectensis* (sea anemone) peptide sequences were downloaded from JGI FTP site (ftp://ftp.jgi-psf.org/pub/JGI_data/). For Ensembl genes, splice variants were removed so that only the largest isoform of a gene was kept. For GPCR analyses, genomes were searched using HMMER 2.3.2 (http://hmmer.wustl.edu/) and PFAM curated GA thresholds to identify peptides containing PFAM 7tm_1 (PF00001), 7tm_2 (PF00002), 7tm_3 (PF00003) and 7tm_4 (PF01461) domains (Bateman *et al*., 2002). For comparisons to previous methods, ortholog datasets were downloaded from the web sites of the SYNERGY (http://www.broad.mit.edu/regev/orthogroups/index.html), Inparanoid v.2 (http://inparanoid.cgb.ki.se/download/) and Multiparanoid (http://www.sbc.su.se/~andale/multiparanoid/html/index.html).

3 RESULTS

3.1 GPCR domain expansions in the animal kingdom

Genes containing any of the four GPCR domain families of *7tm_1 to 4* were collected from six animal species and analyzed separately using EvolMAP. Gene expansion trees were generated for all four families and displayed proportionally to allow comparison of their expansion patterns (Fig. 3). For the *7tm_1* analysis, there were 24 and 34 ancestral *7tm_1*-containing genes estimated present in the eumetazoan and bilaterian ancestors, respectively. Both sea anemone and vertebrate lineages have experienced massive expansions of this family independently, while the investigated protostome genomes (fly and worm) did not (Fig. 3A). The analyses of *7tm_2* and *7tm_3* containing genes displayed slower and otherwise different profiles of expansion (Fig. 3B and C). The *7tm_4* domain, which has very high copy number in the *C.elegans* genome, is not highly represented in other genomes, including the sea anemone. A phylogenetic tree with branch lengths proportional to gene numbers containing this domain shows a massive lineage-specific expansion on the branch leading to *C.elegans* (Fig. 3D).

3.2 Evolution of the mammalian protein-coding genes

To scrutinize the gene complexity of ancestral mammalian genomes and identify the timings of *GDLs* during their descent, we analyzed seven mammalian proteomes using EvolMAP (Fig. 4). Dog and opossum genomes were used as outgroups

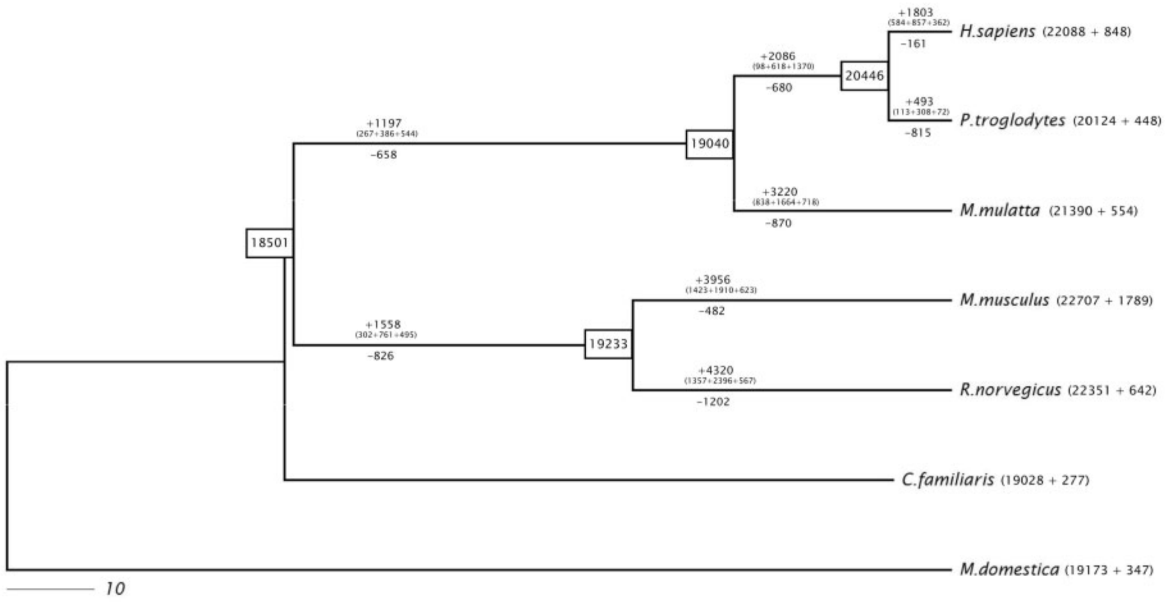

Fig. 4. Gene duplication and loss history of the mammalian proteome. EvolMAP was run with BLAST first to retrieve 300 top-scoring gene pairs for which *NSS* scores were generated where minimum ortholog threshold was set at 250. Scale bar corresponds to *AOD* (average ortholog divergence). Numbers within boxes show estimated number of present genes for that ancestral genome. Gene gains ('+') and losses ('−') are printed for each branch. Number of *in-paralogs, diverged in-paralogs and ambiguous gains* are printed, respectively below each gene gain with a '+' separator. Number of genes from the species is shown within parenthesis next to the species name where the second number following the '+' symbol shows the number of genes omitted from the study. Those genes were omitted since they did not score a BLAST hit (above minimum threshold) to other genes at all.

to support the values for the inner clades. Gene expansion, gene gain and gene loss trees are displayed in Supplementary Figure 1. Number of modern descendants of the inferred ancestral genes is summarized in Supplementary Figure 2. On average, gene gains were detected to be 3.3 times more frequent than losses over all branches suggesting a steady expansion of genome size. An exception was the lineage leading to chimp after their split with human where the number of losses was more than the gains. *In-paralogs*, *diverged in-paralogs* and *ambiguous gains* constituted 27, 48 and 25% of total gene gains, respectively. This suggests that most gained genes (75%) have recognizable sources of duplication and yet most of them (48%) diverged more than the orthologs, undetectable as proper *in-paralogs*. Ambiguously gained genes represent perhaps the most interesting gene gains as they have significantly diverged from their original sources and may be important in species differences. Although some may be annotation errors, many are likely not. For example, the 544 and 1370 genes estimated to be ambiguously gained in the primate and human–chimp ancestral lineages, respectively, are detectable symmetrical best hits within their respective clades, but are missing any significant homolog in the outgroup genomes. Those are unlikely to be due to annotation errors since it would require the same annotation error in multiple genomes. In any event, it is expected that the numbers presented in this study will improve as the genomes are better annotated and more species added to the analysis.

3.3 Comparisons to previous methods

EvolMAP was first compared against SYNERGY on a previously analyzed set of nine yeast genomes from Wapinski *et al.* (2007). *AOD* tree was generated and *GDLs* were mapped onto its branches (Supplementary Fig. 3). The orthologous assignments were observed to be in general agreement (avg. 91%) (Table 1). Furthermore, the numbers of genes estimated for seven out of nine ancestors were on average only 3% different. But for two remaining ancestors, EvolMAP predicted on average 15% fewer genes.

Then, we compared EvolMAP against InParanoid and MultiParanoid. For the InParanoid comparison, we observed almost complete overlap for the closely related species comparison of human and chimp (avg. 99%). The fly and worm comparison yielded somewhat less overlap (avg. 89%) and the biggest differences were observed for the comparison of human and *Ciona* (avg. 81%). Comparison to Multiparanoid analysis of four species from Alexeyenko *et al.* (2006) display similar differences with the comparison to InParanoid on human and *Ciona* dataset (avg. 81%). EvolMAP results tried against different cut-off settings of MultiParanoid (0, 25, 50, 75 and 100) gave the most ortholog overlap with the cutoff set at 25.

4 DISCUSSION

4.1 Advantages and limitations of EvolMAP and comparisons to other methods

The first step of EvolMAP's algorithm is based on consecutive all-to-all *NSS* comparisons of two genomes to find evolutionary relationships of orthology and paralogy. The comparison of EvolMAP and InParanoid, a previous method that generates all-to-all comparisons of a pair of genomes based on BLAST scores, shows that even though closely related species

Table 1. Orthologous gene family comparison with other methods

	Matching groups		Matching genes	
SYNERGY (9 yeasts)	9224/9224 (100%)	9008/9062 (99%)	42587/48654 (88%)	45401/48768 (93%)
InParanoid (HS–PT)	19306/19603 (98%)	19343/19594 (99%)	39101/39361 (99%)	39066/39396 (99%)
InParanoid (DM–CE)	3984/4333 (92%)	3985/4265 (93%)	9305/9823 (95%)	9254/11139 (83%)
InParanoid (HS–CI)	4373/5590 (78%)	4339/4512 (96%)	10397/12395 (83%)	10345/13140 (78%)
MultiParanoid (HS/CI versus DM/CE)	5431/6281 (87%)	5463/5650 (97%)	23,194/26,625 (87%)	23,302/31,253 (75%)

First column for each comparison displays the comparison of the program from left versus EvolMAP and second column displays EvolMAP versus the other program. First, for each orthologous group from one result set, the corresponding groups from the other result set were identified (*Matching groups*). Out of these corresponding groups, the one with most overlapping members with the original group was picked for comparison. For each gene of a *matching group*, we counted whether it is also contained in the corresponding group (*Matching genes*). Abbreviations used are HS, *Homo sapiens*; PT, Pan troglodytes; CI, *Ciona intestinalis*, DM, *Drosophila melanogaster*; CE, *Caenorhabditis elegans*. For all analyses, EvolMAP was run with BLAST first to retrieve 250 top-scoring genes for which *NSS* scores were generated, and orthologous groups were identified with the minimum ortholog threshold set as 300.

comparisons, like that of human and chimp, are in very high agreement (99%), the agreement drops for evolutionarily more distant comparisons, such as that of human and *Ciona* (81%). The differences are likely due to different scoring schemes used by the two programs. As explained in Section 2.1, we reason that *NSS* is more reliable than BLAST scores for predicting orthology between evolutionarily more diverged sequences.

Species tree-based clustering of the pairwise all-to-all comparisons enables reconstruction of ancestral genes of multiple hypothetical ancestors. In contrast to MultiParanoid, which clusters results from separate InParanoid analyses to find orthologous groups of a single ancestral genome, we estimate contents of multiple ancestral genomes. When multiple ancestral genomes are available, comparing ancestral gene families with the descendant gene families can help detect the timings of *GDLs*. We applied the Dollo parsimony model to detect those events rapidly.

SYNERGY is a gene phylogeny-based algorithm developed for the same purpose of ancestral gene content estimation. In that method, both gene and synteny similarity are utilized. Synteny similarity, which corresponds to the similarity of order of genes on a chromosome, might be advantageous for finding orthologs of closely related species. Even though EvolMAP does not consider synteny, orthologous assignments of the two programs for the nine yeast genomes were highly overlapping (91% same). A further challenge for synteny-based ortholog detection is the near absence of synteny for evolutionarily more distant species comparisons, such as that of a protostome and a deuterostome.

When we compared ancestral genes detected by SYNERGY and EvolMAP for a reconstructed yeast species tree, we observed only 3% average difference for seven of nine ancestral nodes (Supplementary Fig. 3). However, for the remaining two nodes EvolMAP predicted 15% fewer genes on average. Those nodes were the ancestral nodes immediately following the branch where the proposed yeast whole genome duplication (WGD) occurred. In fact, in the SYNERGY analysis, the branch of the WGD was assigned a priori a much higher duplication rate than other branches (0.5 for WGD-branch versus 0.05 for all other branches). That assignment led to differences between the two results, since EvolMAP finds the branches where the *GDLs* occurred based on the most Dollo parsimonious scenario. Fewer stabilized genes predicted by EvolMAP for those branches may be due to massive concurrent gene losses at after WGD lineages, and retention of different paralogs in lineages that branched shortly after the WGD. We suggest that using the option of SYNERGY to set different rates of duplications and losses for different branches may generate 'circular' results supporting the original hypothesis, whether or not it is fully supported by gene/synteny similarity data.

4.2 Analyses of specific gene families and complete proteomes using EvolMAP

Restricting EvolMAP analyses on selected gene families is informative for discerning clade-specific expansions of certain genes. In this study, we focused on the evolutionary history of genes containing GPCR domain families of *7tm_1* to *4*. These domains are not found in plant or yeast genomes by HMMER searches. Further, no *7tm_1* or *7tm_4* domains and only one *7tm_2* and *7tm_3* domains are found in the genome of the choanoflagellate *Monosiga brevicollis*. This suggests that two of these domains originated in metazoan ancestors and all of them expanded only in metazoans. For such animal-specific expansions, the recently sequenced sea anemone genome lies in a highly informative phylogenetic position (Putnam *et al.*, 2007). Of the four analyzed GPCR domain families, the Opsin/Rhodopsin (*7tm_1*) domain containing family is the largest animal-specific GPCR in size (Wistrand *et al.*, 2006). HMMER identified 916 genes containing that domain in the sea anemone. This number surprisingly exceeds that of worm (133) and fly (64), and is comparable to that of human (713) and zebrafish (755), all of which are dwarfed by that of mouse (1426). However, EvolMAP estimated only 24 genes containing *7tm_1* domains that were present in the eumetazoan ancestor suggesting that both vertebrate and the sea anemone genomes experienced massive parallel expansions of this family, whereas the two investigated protostome lineages did not undergo such a large expansion.

Using EvolMAP, we studied the history of *GDLs* of seven mammalian genomes. Our results concur with previous studies that used CAFE (Demuth *et al.*, 2006) to demonstrate numerous *GDLs* during evolution of five mammalian genomes (our study added newly available monkey and opossum genomes). Yet, our results differ in detail from Demuth *et al.*

(2006) and indicate that *GDLs* may be even more common than previously suspected. The biggest difference between results from EvolMAP and CAFE was the number of gene gains estimated along the chimp-specific lineage (26 with CAFE versus 493 with EvolMAP). We do not attribute this difference to changes in annotation with our use of a newer dataset (Ensembl v46) because analyzing the same dataset used by Demuth *et al.* (2006) (five genomes, Ensembl v41) with EvolMAP also yields much higher numbers of gene gains along the chimp lineage (694).

Instead, we suggest that the differences between EvolMAP and CAFE may be caused by highly conservative estimates made by CAFE in cases where genes radiated rapidly and were differentially lost. More specifically, after a rapid gene radiation, different subsets of paralogs may be retained in descendant species (Supplementary Fig. 4). In such cases, CAFE infers those paralogs to share a direct ancestry because the method only counts the number of genes in closely related gene families without reference to the relative similarity between them. To test for such occurrences, we performed homology tests between genes inferred by EvolMAP to be gained along the human and chimp lineages. We found 16% of human-specific and 61% of chimp-specific sequences to be significantly similar to a gene inferred as gained along the other species' lineage (i.e. they would be in the same orthologous group in CAFE analysis). Therefore, we were able to attribute some of the differences between the two results to rapid gene duplications. The results of Demuth *et al.* (2006) were conservative with respect to their provocative claim of a genomic 'revolving door', where genes are gained and lost routinely. Our results indicate that *GDLs* occur at an even faster rate than previously estimated.

But perhaps the most interesting genes categorized by EvolMAP are *ambiguous gains*, which are highly novel sequences that are clearly lineage specific. Here there is no concern for being an artifact of rapid gene radiations. Although these genes uninterestingly may be due to annotation errors, those that are not may contribute unique functions to the biology of their species. Such genes may originate by mechanisms such as exon/domain shuffling, bursts of positive selection or horizontal gene transfers.

ACKNOWLEDGEMENTS

We like to thank Tsunglin Liu, Firat Kart and Kawther Abu-Elneel for comments and discussions. Thanks also to I. Wapinski for help with SYNERGY data and analyses. Support for this project derived from the W. M. Keck Foundation and a gift from Harvey Karp.

Conflict of Interest: none declared.

REFERENCES

Alexeyenko,A. *et al.* (2006) Automatic clustering of orthologs and inparalogs shared by multiple proteomes. *Bioinformatics*, **22**, e9–e15.

Altschul,S.F. *et al.* (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res.*, **25**, 3389–3402.

Bateman,A. *et al.* (2002) The Pfam protein families database. *Nucleic Acids Res.*, **30**, 276–280.

Chiu,J.C. *et al.* (2006) OrthologID: automation of genome-scale ortholog identification within a parsimony framework. *Bioinformatics*, **22**, 699–707.

De Bie,T. *et al.* (2006) CAFE: a computational tool for the study of gene family evolution. *Bioinformatics*, **22**, 1269–1271.

Deluca,T.F. *et al.* (2006) Roundup: a multi-genome repository of orthologs and evolutionary distances. *Bioinformatics*, **22**, 2044–2046.

Demuth,J.P. *et al.* (2006) The evolution of Mammalian gene families. *PLoS ONE*, **1**, e85.

Durand,D. *et al.* (2006) A hybrid micro-macroevolutionary approach to gene tree reconstruction. *J. Comput. Biol.*, **13**, 320–335.

Farris,J.S. (1977) Phylogenetic analysis under Dollo's law. *Syst. Zool.*, **26**, 77–88.

Fitch,W.M. (1970) Distinguishing homologous from analogous proteins. *Syst. Zool.*, **19**, 99–113.

Henikoff,S. and Henikoff,J.G. (1992) Amino acid substitution matrices from protein blocks. *Proc. Natl Acad. Sci. USA*, **89**, 10915–10919.

Ingram,V.M. (1961) Gene evolution and the haemoglobins. *Nature*, **189**, 704–708.

Koonin,E.V. (2005) Orthologs, paralogs, and evolutionary genomics. *Annu. Rev. Genet.*, **39**, 309–338.

Li,H. *et al.* (2006) TreeFam: a curated database of phylogenetic trees of animal gene families. *Nucleic Acids Res.*, **34**, D572–D580.

Li,L. *et al.* (2003) OrthoMCL: identification of ortholog groups for eukaryotic genomes. *Genome Res.*, **13**, 2178–2189.

Needleman,S.B. and Wunsch,C.D. (1970) A general method applicable to the search for similarities in the amino acid sequence of two proteins. *J. Mol. Biol.*, **48**, 443–453.

O'Brien,K.P. *et al.* (2005) Inparanoid: a comprehensive database of eukaryotic orthologs. *Nucleic Acids Res.*, **33**, D476–D480.

Ohno,S. (1970) *Evolution by Gene Duplication*. Springer-Verlag, Berlin.

Plachetzki,D.C. and Oakley,T.H. (2007) Key transitions during the evolution of animal phototransduction: novelty, 'tree-thinking,' co-option, and co-duplication. *Integrative and Comparative Biology*, **47**, 759–769.

Putnam,N.H. *et al.* (2007) Sea anemone genome reveals ancestral eumetazoan gene repertoire and genomic organization. *Science*, **317**, 86–94.

Remm,M. *et al.* (2001) Automatic clustering of orthologs and in-paralogs from pairwise species comparisons. *J. Mol. Biol.*, **314**, 1041–1052.

Sonnhammer,E.L. and Koonin,E.V. (2002) Orthology, paralogy and proposed classification for paralog subtypes. *Trends Genet.*, **18**, 619–620.

Storm,C.E. and Sonnhammer,E.L. (2002) Automated ortholog inference from phylogenetic trees and calculation of orthology reliability. *Bioinformatics*, **18**, 92–99.

Tatusov,R.L. *et al.* (2003) The COG database: an updated version includes eukaryotes. *BMC Bioinformatics*, **4**, 41.

Wall,D.P. *et al.* (2003) Detecting putative orthologs. *Bioinformatics*, **19**, 1710–1711.

Wapinski,I. *et al.* (2007) Automatic genome-wide reconstruction of phylogenetic gene trees. *Bioinformatics*, **23**, i549–i558.

Wistrand,M. *et al.* (2006) A general model of G protein-coupled receptor sequences and its application to detect remote homologs. *Protein Sci.*, **15**, 509–521.

Zmasek,C.M. and Eddy,S.R. (2002) RIO: analyzing proteomes by automated phylogenomics using resampled inference of orthologs. *BMC Bioinformatics*, **3**, 14.