BIOINFORMATICS ORIGINAL PAPER

Vol. 24 no. 2 2008, pages 151–157
doi:10.1093/bioinformatics/btm567

Genome analysis

MANTIS: a phylogenetic framework for multi-species genome comparisons

Athanasia C. Tzika[1,†], Raphaël Helaers[1,†], Yves Van de Peer[2] and Michel C. Milinkovitch[1,*]

[1]Laboratory of Evolutionary Genetics, Institute for Molecular Biology & Medicine, Université Libre de Bruxelles, 12 rue Jeener & Brachet, B6041 Gosselies and [2]Bioinformatics & Evolutionary Genomics, Department of Plant Systems Biology, Ghent University, VIB, Gent, Belgium

Received on September 22, 2007; revised and accepted on November 7, 2007

Advance Access publication November 19, 2007

Associate Editor: Martin Bishop

ABSTRACT

Motivation: Practitioners of comparative genomics face huge analytical challenges as whole genome sequences and functional/expression data accumulate. Furthermore, the field would greatly benefit from a better integration of this wealth of data with evolutionary concepts.

Results: Here, we present MANTIS, a relational database for the analysis of (i) gains and losses of genes on specific branches of the metazoan phylogeny, (ii) reconstructed genome content of ancestral species and (iii) over- or under-representation of functions/processes and tissue specificity of gained, duplicated and lost genes. MANTIS estimates the most likely positions of gene losses on the true phylogeny using a maximum-likelihood function. A user-friendly interface and an extensive query system allow to investigate questions pertaining to gene identity, phylogenetic mapping and function/expression parameters.

Availability: MANTIS is freely available at http://www.mantisdb.org and constitutes the missing link between multi-species genome comparisons and functional analyses.

Contact: mcmilink@ulb.ac.be

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

1 INTRODUCTION

Since the first fully sequenced genome of a free-living organism, *Haemophilus influenzae* (Fleischmann *et al.*, 1995), the list of available whole genome sequences is lengthening at an increasing pace, with over 500 completed and 2000 ongoing projects (Liolios *et al.*, 2006). This wealth of data has led to the appearance of a new biological field: comparative genomics. Algorithms and software have been developed for the alignment of whole genomes and for facilitating multi-genome comparisons, and public databases have been created for genome annotation, functional and expression data, as well as for linking taxonomic and relevant publications to complete and ongoing sequencing projects (Bray *et al.*, 2003; Brudno *et al.*, 2007; Curwen *et al.*, 2004; Gouret *et al.*, 2005; Liolios *et al.*, 2006; Odronitz *et al.*, 2007). Particularly effective is the ENSEMBL database (Hubbard *et al.*, 2007) that (i) provides a comprehensive and integrated source of metazoan sequence annotation generated through an automatic gene build pipeline, (ii) identifies orthologs and paralogs through the estimation of gene family phylogenetic trees and (iii) provides links to gene ontology terms (Ashburner *et al.*, 2000), as well as to eGenetics (Kelso *et al.*, 2003) and GNF (Su *et al.*, 2002) expression data. Popular databases complementary to ENSEMBL are PANTHER, relating molecular functions and biological processes to phylogenetically defined subfamilies of proteins (Mi *et al.*, 2007), and HMDEG, which classifies millions of human and mouse ESTs into tissue/organ categories (Pao *et al.*, 2006). Here, we present MANTIS (www.mantisdb.org), a java application system that builds a MySQL relational database integrating, in a phylogenetic framework, all ENSEMBL genes, corresponding molecular functions and biological processes, as well as expression data from multiple databases; makes extensive use of the ENSEMBL ortholog/paralog prediction pipeline for identifying gene duplication events; and infers the mapping of gene gains, duplications, and losses on the phylogenetic tree. Through a user-friendly interface, MANTIS allows the user to identify (i) gains and losses on specific branches of the tree, (ii) genome content of ancestral species, (iii) statistically over- or under-represented molecular functions and biological processes and (iv) tissue specificity of gained, duplicated and lost genes. Finally, the entire set of information available in MANTIS can be exploited further using an advanced system of queries by which gene identity, presence, mapping and function parameters can be combined using logical operators. In short, MANTIS allows the user to explore and interrogate animal genome content and associated functional data within a phylogenetic context.

*To whom correspondence should be addressed.

†The authors wish it to be known that, in their opinion, the first two authors should be regarded as joint First Authors.

2. SYSTEM AND METHODS

2.1 Characters mining

Up to version 38 of the ENSEMBL database, orthology prediction was based on best reciprocal BLAST hits and synteny extensions. Since version 39, ENSEMBL orthologs and paralogs are inferred with a much improved pipeline (www.ensembl.org/info/data/compara/homology_method.html) that includes the clustering of gene family members, multiple protein sequence alignment, protein tree inference and gene tree versus species tree reconciliation (to label duplication events on the internal nodes). The final product of this process is accessible via the '*protein_tree_member*/*node*/*tag*' tables of the *ENSEMBL-Compara* database. We generate two types of MANTIS characters: (i) one new character is created for each tree representing a protein family (i.e. these correspond to *de novo* gains, and not to duplication events) and (ii) one new character is created for each duplication event (i.e. each protein family may include one to many duplication events). Once all trees have been parsed, characters are also created for the genes that have no homologs (neither orthologs, nor paralogs), i.e. species-specific single genes that are not included in the ENSEMBL protein trees. Hence, strictly speaking, a character in MANTIS is not a single gene or protein, but the set of all homologous genes (i.e. a complete gene family) for a *de novo* gain, or the set of all orthologous genes for a duplication event. The only exceptions are species-specific single genes (i.e. those among species-specific genes that did not experience any duplication event). Two datasets, which serve as a basis for all MANTIS functionalities, are then built: the dataset '*with duplications*' combining all characters, and the dataset '*families only*', that excludes the characters corresponding to duplication events (by merging the characters within each protein tree).

Duplication events have the potential of generating shifts in gene functions and/or expression patterns and localizations. For example, compartmentation of specialized gene functions can be brought about by duplication of the protein coding sequence with its *cis*-regulatory non-coding modules (CRMs) followed by subfunctionalization (Lynch and Conery, 2000; Lynch and Force, 2000), i.e. the two gene copies specialize to perform complementary functions for example through evolution of the respective sets of CRMs (Force *et al*., 1999; Greer *et al*., 2000). Subfunctionalization would increase the probability of survival of duplicates, hence, would provide an extended time period during which duplicated genes can experience neofunctionalization, i.e. one copy acquiring a new function whereas the other retaining the ancestral function (Ohno, 1970) through coding sequence modifications (He and Zhang, 2005; Rastogi and Liberles, 2005). Finally, lineage-specific positive selection might also play a significant role in the retention of duplicates (Hurles, 2004; Kondrashov and Kondrashov, 2006; Shiu *et al*., 2006). One major asset of combining genome content, functional data and phylogenetic information lies in the possibility of analysing these shifts in gene functions and/or expression patterns and localizations. Hence, the definition of MANTIS characters originating from duplications requires the discrimination of the 'ancestral' versus 'derived' child branch after each duplication event. To this end, we use the following proxy (Fig. 1): for each of the two child sub-trees, the mean distance between the duplication node and all leaf nodes is calculated, and the sub-tree associated to the smallest value is regarded as 'ancestral'. This method is based on the assumption that 'ancestral' characters are constrained by selection, i.e. ancestral characters preserve their ancestral function (hence, less change is expected to occur), whereas 'new' (i.e. 'derived') characters are expected to experience some level of positive selection, i.e. of accelerated evolution (Zhang, 2003).

As (i) functional and expression data are associated to a single-specific ENSEMBL gene but (ii) a MANTIS character can be associated to several ENSEMBL genes (from different species, see above), we must decide what ENSEMBL gene (and associated expression data) each MANTIS character will be named after. To ensure assignment of the most likely functional/expression data, each MANTIS character is associated to the ENSEMBL gene ID (hereafter, called '*main gene*') of the species highest in a 'priority list' (see Supplementary Fig. 1): species are sorted partly according to their phylogenetic proximity with human then mouse, i.e. the two species with the largest amount of functional and expression data available. All results provided by MANTIS, concerning biological processes, molecular functions, and gene expression of any character, are based on the data available for the '*main gene*'. All non-priority species genes associated to a given character are considered as 'synonyms' to the corresponding MANTIS character (i.e. the gene of the first priority species) except when functional information is available, via the PANTHER database, for the non-priority species gene, as it is the case for *Mus musculus*, *Rattus novergicus* and *Drosophila melanogaster* genes.

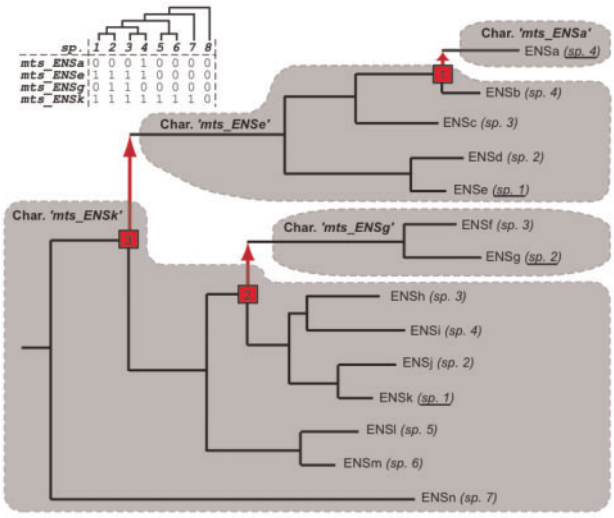

Fig. 1. Example of MANTIS character assignment for genes with imaginary ENSEMBL gene IDs ('ENSa' to 'ENSn'). Three duplication events (red boxes) occurred within this gene family: within species 4, in the ancestor of species 1–4, and in the ancestor of species 1–6. Species 1 and 4 lost a paralog after duplication 2 whereas species 5 and 6 lost a paralog after duplication 3. At each duplication event, MANTIS generates a new character and all species within the sister subtree with the largest mean branch length (from the duplication event to each tip node) is assigned to it. For example, the mean distance from duplication 2 to genes 'ENSf' and 'ENSg' is larger than the mean distance to genes 'ENSh', 'ENSi', 'ENSj' and 'ENSk'; hence, a new (derived) character ('*mts_ENSg*') is considered 'present' in species 2 and 3 whereas that character is considered 'absent' in species 1 and 4. The derived character is named after the ENSEMBL gene ID ('ENSg') of the first species in a priority list (here, species 2). Computation of mean distances disregards nested derived characters; e.g. to identify the ancestral and derived paralog lineages after duplication 3, MANTIS computes the mean distance from duplication 3 to, on one hand, 'ENSb', 'ENSc', 'ENSd' and 'ENSe' (and *not* 'ENSa') and, on the other hand, to 'ENSh', 'ENSi', 'ENSj', 'ENSk', ENSl' and 'ENSm' (and *not* 'ENSf', 'ENSg'). This example generates the presence/absence character matrix shown in upper left together with the true tree among the eight species. No member of the gene family is present in species 8 (hence, the '0' state for all characters in that species). 'mts_ENSk' is the most ancestral character and is present in species 1–7.

Figure 1 illustrates the character assignment method: in the '*families only*' dataset, a single character (grouping all genes at the leaves of the tree) is defined and the most ancestral gene of the priority species is assigned as its name (here, '*mts_ENSk*'); whereas in the '*with duplications*' dataset, four characters are defined ('*mts_ENSk*', '*mts_ENSg*', '*mts_ENSe*' and '*mts_ENSa*').

2.2 Characters mapping

Gains and losses of characters are mapped by MANTIS on the 'true' species-tree, i.e. the best-supported tree on the basis of the available literature (see Supplementary Fig. 1 and Supplementary references). Mapping in MANTIS is performed as follows. First, a character presence/absence matrix for all species is built, and used for computing a distance matrix following a modified Jukes–Cantor model, adjusted for 2-states characters. Second, the distance matrix is used to compute the branch lengths of the MANTIS 'true' species-tree, using the least-squares approach under minimum evolution (Kidd and Sgaramella-Zonta, 1971; Rzhetsky and Nei, 1992, 1993). Third, the gain of a character is assigned to the corresponding branch in the species tree, i.e. to the branch leading to the most-recent-common ancestor of all species present either in the whole gene family tree (for *de novo* characters) or in the gene family subtree defined by the duplication event (for characters corresponding to duplication events). Species-specific characters are gained on the tip branch leading to the corresponding species. Fourth, using a recursive function, a maximum-likelihood approach is used to identify, for each character, the branch(es) to which gene loss(es) is (are) assigned. MANTIS computes the probability of losing the character at a specific branch as:

$$P(\text{loss}) = 1 - e^{-bl} \quad (1)$$

where bl is the length of the specific branch in the minimum-evolution tree, and the probability of not losing the character at this branch is:

$$P(\text{no\_loss}) = (1 - P(\text{loss}))P_{\text{left}}P_{\text{right}} \quad (2)$$

Where P_{left} and P_{right} are the maximum probability (of loss or no-loss) associated to the left and right child branches, respectively. P_{left} and P_{right} are recursively computed from tip to deep nodes such that only the most probable combination of losses and no-losses are considered (and computed, at worst, in polynomial time) rather than all possibilities. Figure 2 gives an example of character mapping performed by MANTIS.

Once the characters are mapped, two datasets are created: the '*all changes*' dataset includes all characters, independently of the number of times they were lost, whereas the '*single changes*' dataset only comprises, as '*gains*', the characters that were lost in none of the branches of the species tree, and, as '*losses*', the characters that were lost only in a single branch. Gains and losses can be visualized and analysed further through the '*character mapping*' view of MANTIS (Fig. 3).

2.3 Genome content

Once gains and losses have been mapped, MANTIS builds the genome content of each ancestral species (i.e. at each internal node) by (starting from the root) adding all gained and subtracting all lost characters along the branches leading to the node of interest. The result can be visualized and analysed further through the '*genome content*' view of MANTIS (Fig. 3). To allow for analysis of functional data associated to genes of interest, the '*genome content*' functionality keeps the assignment of '*main genes*' to characters, which explains the presence of gene names from different species even for characters at the leaves of the tree. MANTIS builds genome contents both for '*single*' and '*all changes*' datasets.

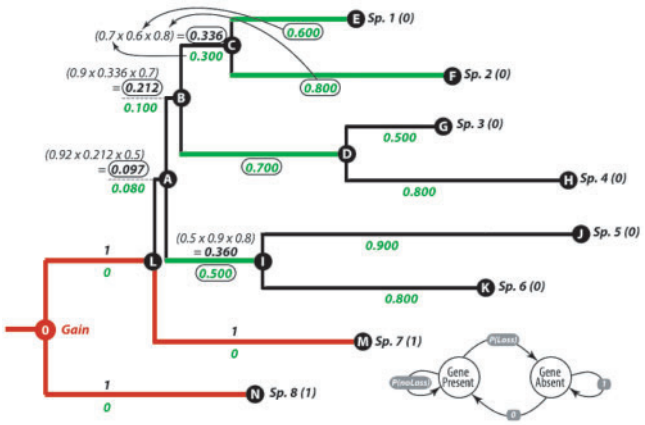

Fig. 2. Example of character mapping (i.e. assigning gains and losses of MANTIS characters on branches of the true tree) on an 8-species tree for one hypothetical character [presence (1) or absence (0) of the character is indicated after each species name]. Using all MANTIS characters, MANTIS estimates distances following a 2-states-adjusted Jukes–Cantor model and computes the branch lengths (least-squares approach under minimum evolution) of the MANTIS species tree. Gain of the character at hand is here simply assigned to the most recent common ancestor (node 'O') of all species exhibiting the specific character (here, taxa 'M' and 'N', red branches). Losses are assigned to branches (in green) using a maximum-likelihood automaton (lower right) with the probabilities of loss or no-loss being recursively computed from tip to deep nodes such that only the most probable combination of losses and no losses is computed. For example, the probabilities of loss on branches C-E, C-F and B-C are equal to 0.600, 0.800 and 0.300, respectively; the probability of no-loss on branch B-C is $[(1-P(\text{loss}))P_{\text{left}}P_{\text{right}}] = (1-0.3)*0.6*0.8 = 0.336$. Given that, for the branch B-C, $P(\text{no\_loss}) > P(\text{loss})$ (i.e. $0.336 > 0.300$), the loss is assigned on the branches C-E and C-F. See details in the text.

2.4 Biological processes and molecular functions

The biological processes and molecular functions data available in MANTIS are based on the PANTHER Classification System (Mi *et al.*, 2007). Using the *ENSEMBL-Compara* database (Stalker *et al.*, 2004), MANTIS maps *Homo sapiens*, *Mus musculus*, *Rattus norvegicus* and *D.melanogaster* ENSEMBL gene IDs to corresponding Entrez gene IDs used by PANTHER. The PANTHER categories and their hierarchy are maintained in MANTIS, but a '*No information*' category is added, corresponding to ENSEMBL genes with no functional information available in PANTHER. MANTIS is not limited to providing lists of Gene Ontology terms for each gene, but goes further by plotting the over- or under-representation of each category associated with gains or losses on each branch of the species tree. Statistical significance of category over- or under-representation is computed on the basis of the category distribution of the reference-species genes: e.g. a category C is over-represented in gains (or losses) when $k(C)$, the observed number of gained (or lost) genes of category C, is greater than $p(C)K$, the expected number of corresponding events [where $p(C)$ is the proportion of genes of category C in the reference species, and K is the total number of gains (or losses) on the branch considered]. Statistical significance is determined by the calculation of a P-value following the binomial statistics or classical approximations (see Supplementary Material).

On each branch, the representation of each functional category is displayed as a column of a histogram ['*Biological Processes*' and

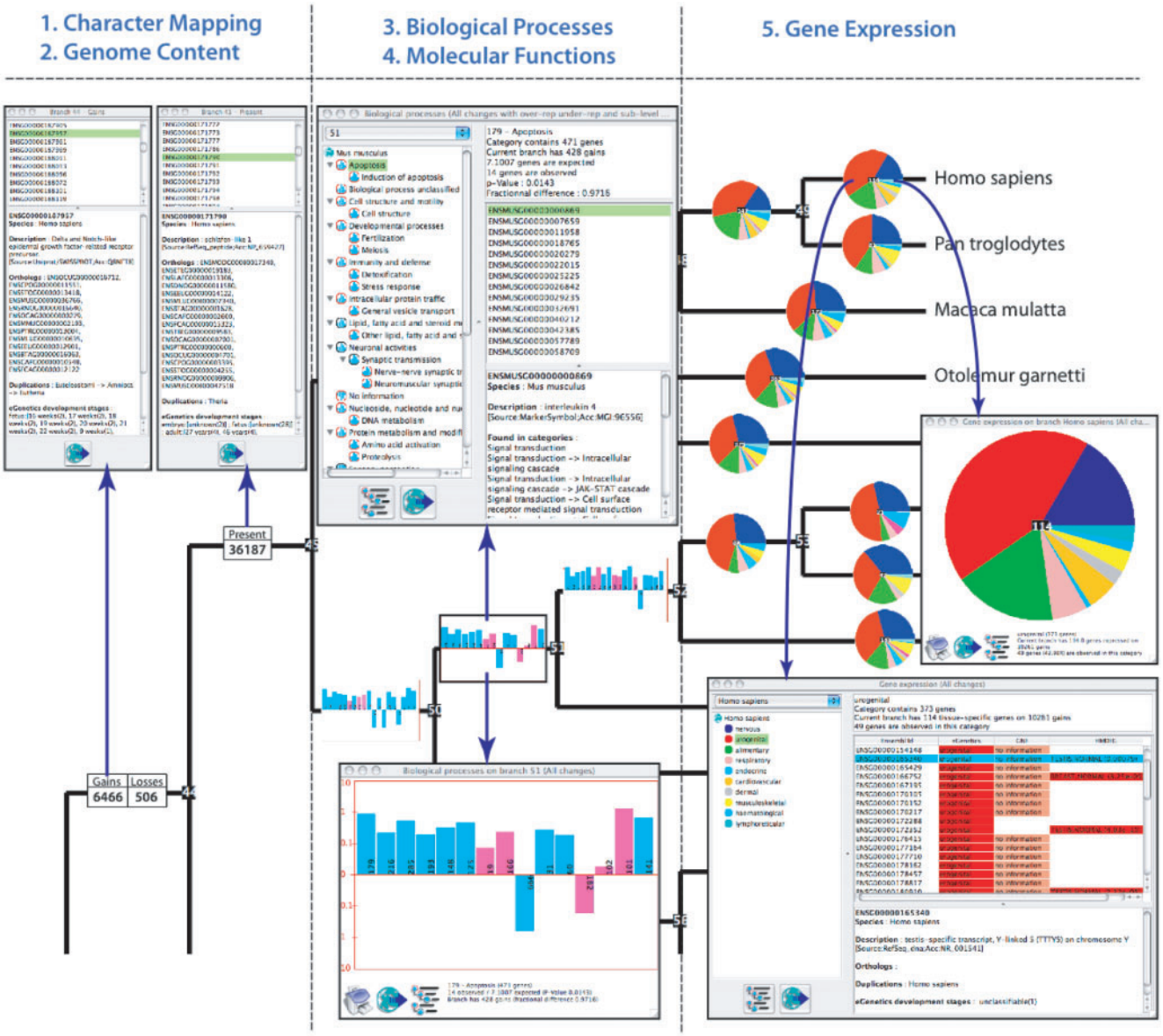

Fig. 3. Overview of the five main MANTiS displays. Numbers, graphs and pie charts on the branches of the phylogenetic tree give access to additional information. See text for details.

'*Molecular Functions*' views of MANTIS (Fig. 3)], where the y axis is the fractional difference (on a logarithmic scale):

$$f_{\text{diff}} = \left(\frac{k(C) - p(C)K}{p(C)K}\right) \tag{3}$$

2.5 Gene expression

Three sources of gene expression data are used in MANTIS: *eGenetics*, *GNF* and *HMDEG*, the two former available from the *EnsemblMart* retrieval tool (http://www.ensembl.org/biomart/martview). The *eGenetics* database uses Expressed Sequence Tags (ESTs) annotated with eVOC ontology terms (Kelso *et al.*, 2003) by SANBI (South African National Bioinformatics Institute) and mapped to ENSEMBL transcript predictions. The *GNF* (Genomics Institute of the Novartis Research Foundation) database includes expression data based on the Affymetrix HG-U95A microarray, annotated with eVOC ontology terms by SANBI, and Affymetrix probe sequences mapped to ENSEMBL transcript predictions (Su *et al.*, 2002). *HMDEG* [Human and Mouse Differentially Expressed Genes; (Pao *et al.*, 2006)] is a database that classifies more than 8 million human and mouse ESTs into tissue/organ categories to which one can apply statistical tests for differential expression.

MANTIS applies a filter to *eGenetics* and *GNF* expression data for determining the tissue-specific genes. To be declared '*tissue-specific*', a gene has to be expressed only in one higher-level anatomical system, regardless of the expression site sub-categories (i.e. a gene expressed both in '*Anatomical System* → *nervous* → *peripheral nervous system* → *visual apparatus* → *retina*' and in '*Anatomical System* → *nervous* → *central nervous system* → *brain*' is considered as '*nervous-system*

specific'). Furthermore, the '*unclassifiable*' category is ignored when it is associated with a defined anatomical system. Genes found in more than one anatomical systems are not considered '*tissue specific*'. MANTIS then integrates tissue-specificity information from each expression database into categories, representing the following eVOC anatomical systems: '*nervous*', '*urogenital*', '*alimentary*', '*respiratory*', '*endocrine*', '*cardiovascular*', '*dermal*', '*embryo*', '*musculoskeletal*', '*haematological*', '*lymphoreticular*' and '*unclassifiable*'. Furthermore, MANTIS assigns a '*no information*' category to genes for which no expression data is available. For each gene assigned as '*tissue-specific*' on the basis of the *eGenetics* and *GNF* data, MANTIS also displays the tissue with highest specificity of expression (i.e. having the lowest *P*-value) as inferred by *HMDEG*, and its correspondence with the 13 eVOC anatomical system categories (version 2.8).

In the '*Gene Expression*' view, MANTIS displays the anatomical system assignments of gains or losses using interactive pie charts for rapid browsing (Fig. 3).

2.6 Queries

A powerful functionality of MANTIS is the possibility to build elaborate queries concerning gene identity, mapping and function parameters (biological processes, molecular functions and gene expression). The queries generated through a user-friendly interface (see Supplementary Fig. 2) are then automatically translated into complex SQL queries. Each MANTIS query is composed of one action performed on one or several '*statement*(*s*)' linked by logical operators. In each statement, four criteria can be considered in combination or isolation: a subset of user-defined genes, the type of events mapped (gene presence or gene gains and/or losses, with '*all*' or '*single*' changes), a subset of branches, and specific functions (biological processes, molecular functions or gene expression). The complementary of any of the chosen '*gene*', '*branch*' and '*function*' criteria can be selected within a statement or in reference to all previous statements. Any ENSEMBL ID (gene, transcript or protein), Unigene ID, or Entrez ID provided by the user is converted to its '*main gene*' synonym (i.e. the MANTIS character, see above) with the possibility to extend the entry to all genes from the same multigene family. MANTIS also provides a graphical interface for easy selection of branches and/or functions.

Statements are executed following priorities and logical operators explained in the Supplementary Material. Once all criteria within all statements have been set, one action is chosen. For example, the user can request the '*List of Genes*', or the '*List of Branches*' or the '*List of Functions*' that meet all statement criteria. When the '*Restrict Tree*' action is selected, not only are the relevant genes listed, but a new MANTIS dataset is generated (all gains, losses, biological processes/molecular functions histograms and gene expression pie charts are recomputed) such that the species tree can be browsed, with only the genes of the query result, under the five MANTIS views (Character Mapping, Genome Content, Biological Processes, Molecular Functions and Gene Expression). The user can then easily switch between the full and restricted datasets. Additional parameters for fine-tuning the queries are explained in the Supplementary Material.

Finally, part or all of the data in the query result can be used as input information for a new query (see Supplementary Fig. 2), allowing the user to perform successive queries of any complexity.

3 THE MANTIS VIEWS

Most of MANTIS functionalities are performed in a phylogenetic framework, i.e. using the 'true' phylogenetic tree among the fully sequenced metazoan genomes accessible in ENSEMBL. Each MANTIS view ('*Characters mapping*', '*Genome content*', '*Biological processes*', '*Molecular functions*' and '*Gene expression*') displays different types of information on the branches of the tree. A zoom function allows the user to focus on specific portions of the tree. Each specific histogram and pie chart can be opened in a separate interactive window for printing, exporting and inspection of detailed information.

An overview of the five MANTIS displays is shown in Figure 3. The '*Character Mapping*' view (Fig. 3, left panel) shows the number of inferred gains and losses at each branch of the tree. Double clicking on the corresponding number generates a full list of the genes that have been gained or lost at the chosen branch; selecting a gene in the list prompts the display of associated information: gene description, known orthologs, number of duplication events (and their localizations in the tree) in the gene family until this specific gene gain or loss occurred and the developmental stages (according to *eGenetics* and *GNF* databases) at which expression of the gene has been detected (independently of tissue-specificity). Similarly, the '*Genome content*' view (Fig. 3, left panel) displays the number of genes present in each ancestral genome (i.e. internal node of the tree), and gives access to the same detailed information about each gene as in the '*Character Mapping*' view.

The '*Biological Processes*' and '*Molecular Functions*' views (Fig. 3, central panel) are very similar in form as they both provide, at each branch of the tree, a histogram that displays, respectively, the biological processes and molecular functions that are significantly over- or under-represented in the set of genes gained or lost at the corresponding branch. The *y* axis of the histogram represents the fractional difference [Equation (3)] on a logarithmic scale. By default, the first level categories of gene ontology terms are shown, if they are themselves significantly (P-value < 0.05) over- or under-represented (blue and yellow columns for gains and losses, respectively) or if they contain significantly over- or under-represented sub-categories (mauve columns). The type of displayed information can be however modified by considering or ignoring lower-level or non-significantly under/over-represented categories. Finally, each histogram is associated to a '*category browser*' that displays the classification hierarchy of the represented categories and provides detailed information on selected categories and the genes they contain. The user can switch among *H.sapiens*, *M.musculus*, *R.norvegicus* and *D.melanogaster* as the source species for '*Biological Processes*' and '*Molecular Functions*' data: MANTIS changes the display accordingly and re-calculates the representation (and significance) of each category. By definition, computation is possible only for branches that exhibit gains or losses of genes present in the reference species. For example, if *H.sapiens* is used as reference, human genes gained in any branch leading to *Homo* or lost in any other branch will be used whereas it is non-sensical to ask for molecular function data obtained in *H.sapiens* for genes gained in, e.g. teleost fishes.

Finally, the '*Gene Expression*' view (Fig. 3, right panel) shows, at each branch of the tree, a pie chart with the number of genes specifically expressed in each anatomical system (the user can switch between *eGenetics*, *EST* and *GNF* microarray data). As MANTIS only provides human expression data, all pie charts display tissue-specific genes that have either been gained at branches leading to *H.sapiens* or lost at any of the other branches in the tree. Each pie chart gives access to a gene

expression '*category browser*' displaying the list of genes and corresponding expression data from *eGenetics*, *GNF* and *HMDEG* with colour coding for easy comparisons of anatomical systems among the three databases. Selecting a gene provides similar detailed information as in other views: gene description, known orthologs, number of duplication events (and their localizations in the tree) in the gene family, and developmental stages at which expression of the gene has been detected.

4 A CASE STUDY

The functionalities of MANTIS make this software system very flexible: e.g. MANTIS allows, (i) the easy identification of genes present in any (set of) species and absent from any other (set of) species, (ii) to list the genes gained in any (set of) branch(es) that were subsequently lost in any (set of) more recent lineage(s) and (iii) detect and investigate the tempo and mode of duplication within gene families. Good examples of the latter are the well-known gamma-crystallin and bitter-taste receptor families (that greatly expanded within, respectively, various vertebrate and mostly mammalian lineages) or the enigmatic ENSF00000000209 opossum-specific family that includes 46 paralogs, none of which has been studied so far. We use below some examples centred on questions pertaining to the marsupial lineage.

A high-quality draft sequence of the grey short-tailed opossum (*Monodelphis domestica*) genome was recently published, with an estimated 18 648 protein-coding genes, and compared to available eutherian genomes (Mikkelsen *et al*., 2007): 15 320 (82.1%) and 2704 (14.5%) of Monodelphis genes matched human genes with or without, respectively, clear orthology, leaving a small number of genes (624 = 3.3%) as opossum-specific. MANTIS allows a comprehensive analysis of gains and losses in a multi-species phylogenetic framework. Using version 44 of the ENSEMBL database (including 21 359 opossum genes, of which 6263 are genes newly identified by the ENSEMBL pipeline), MANTIS identifies 5304 (24.8%) opossum-specific gains (= the number of gains in the 'mapping view' of the opossum lineage). Switching to the 'families only' dataset allows MANTIS to identify that 2971 of the 5304 genes gained in the opossum lineage do not have homologs in other species (i.e. did not originate from genes existing before the metatherian–eutherian split); inspection of the exported table listing these gains easily identifies that 2881 of them are *de novo* gains, whereas the remaining 90 genes originated through duplication events within the opossum lineage (i.e. they are members of gene families specific to Monodelphis). Conversely, 2333 (i.e. 5304–2971) opposum-specific gains originated from duplications (in the opossum lineage) of genes that existed before the metatherian–eutherian split. The much larger number of strictly opossum-specific genes (or families) identified by MANTIS than by Mikkelsen *et al*. (2007) (2971 and 624, respectively) could be explained by the improved ENSEMBL gene prediction pipeline due to the comparison to multiple genome sequences (although it is possible that a significant portion of these projected genes are false positives or pseudogenes).

Beside general investigation of the tempo of gains and losses in a (set of) specific branch(es), MANTIS also allows to ask gene-driven questions. For example, in the initial report of the opossum genome (Mikkelsen *et al*., 2007), it has been underlined that only eight of the opossum genes with no homolog in human (ENSMODG00000000115, Malate synthase; ENSMODG00000001080, Inosine/uridine-preferring nucleoside hydrolase; ENSMODG00000018409, CPD-Photolyase; ENSMODG00000017963, Fucolectin precursor 1, weak homology; ENSMODG00000024867, Fucolectin precursor 2, weak homology; ENSMODG00000008308, Fatty acid synthase; ENSMODG00000002691, Unknown function; ENSMODG00000021523) have strong evidence of being functional. Restricting the dataset to these eight genes (with or without their family members) with the use of a MANTIS query (see Methods section) allows easily extending investigation of gains and losses to the whole phylogenetic tree. For example, it allows to immediately visualize that, in fact, only one of these eight genes is strictly *Monodelphis*-specific and was generated, within the opossum lineage, from the duplication of a much older paralog (that originated in tetrapods), whereas the seven others are ancient genes (gained in the ancestor of either Metazoa or Bilateria or vertebrates or tetrapods) that were simply lost in one of the branches leading to human (i.e. in the ancestor of Eutheria and Euarchontoglires, for 6 and 1 genes, respectively). Such important information on the mode and tempo of gene gains and losses is much difficult to obtain without the MANTIS phylogenetic framework.

MANTIS allows easy access not only to genes presence/absence information, but also to biological processes or molecular functions data. For example, MANTIS indicates that gains of gene families in the 'Immunity and defense' category are over-represented both in the branch leading to mammals (31 gains observed/11.54 expected; *f*diff = 1.68; *P*-value < 1.5 *E*−6) and in the branch leading to [opossum + eutherians] (35 gains observed/16.35 expected; *f*diff = 1.14; *P*-value = 4.1 *E*−5), whereas they are not significantly over-represented in the branch leading to eutherians (45 gains observed/39.02 expected; *f*diff = 0.15; *P*-value = 0.19). Furthermore, only 4 '*Immunity and defense*' subcategories are assigned to gene gains in the eutherian branch whereas 9 and 8 subcategories are assigned in the mammalian and the [opossum + eutherians] branches, respectively. This result is consistent with the recent and somewhat surprising demonstration that significant increase in complexity of the mammalian immune system occurred prior to the divergence between the marsupial and eutherian lineage (Belov *et al*., 2007).

5 CONCLUSIONS AND AVAILABILITY

MANTIS (www.mantisdb.org) is a user-friendly application system with a relational database integrating genes from metazoan genomes, corresponding molecular functions and biological processes, as well as expression data. MANTIS provides phylogenetic-driven (focusing on any set of branches), gene-driven (focusing on any set of genes), function/process-driven and expression-driven functionalities. Among others, it allows identifying the phylogenetic position of gains and losses, the genome content of ancestral species, the over- or

under-representation of molecular functions and biological processes, and tissue specificity of gained, duplicated and lost genes. The functionalities of the software system make it very flexible: e.g. MANTIS allows, (i) the easy identification of genes present in any (set of) species and absent from any other (set of) species, (ii) to list the genes gained in any (set of) branch(es) that were subsequently lost in any (set of) more recent lineage(s) and (iii) detect and investigate the tempo and mode of duplication within gene families. 'Mantis' (μάντης) is the Greek word for a 'seer', i.e. someone able to give prophesies in response to questions.

MANTIS will follow ENSEMBL updates (i.e. every 2 months), as well as updates of the other databases that are incorporated, such as PANTHER and HMDEG. However, given that regular updates can be a burden in specific cases (e.g. researchers might wish to combine analyses performed more than 2 months apart), MANTIS gives access to the six latest versions such that the oldest and newest available versions are 1 year apart. By definition, the precision and breadth of the questions that can be tackled through the use of MANTIS will improve through the addition of new metazoan genomes as well as enhanced and extended functional and expression data in the years to come. As the positions of some lineages in the eukaryote phylogenetic tree are subject to discussions, it might become necessary, in future releases of MANTIS incorporating additional species, to provide the user with the possibility to choose among a small set of alternative and competing species trees. Furthermore, we plan to extend the MANTIS database beyond the ENSEMBL orthology pipeline by incorporating high-quality gene trees from other sources, such as the Human Phylome project (Huerta-Cepas *et al*., 2007), that include a larger number of species.

Stand-alone versions of the interface (for Windows, Mac OSX and Linux) as well as an applet version are freely available at http://www.mantisdb.org. The interface performs connections to a MySQL database. The character presence/absence matrix and the table of gains/losses generated by MANTIS are available as flat files at the MANTIS web site.

ACKNOWLEDGEMENTS

We thank Mike Steel (Department of Mathematics and Statistics, University of Canterbury, NZ) and Raffaele Pesenti (Department of Applied Mathematics, University of Venice, Italy) for the validation of our maximum-likelihood recursive function for the mapping of gene losses. This work was supported by grants from the '*Communauté Français de Belgique*' (ARC 1164/20022770), the '*National Fund for Scientific Research Belgium* (*FNRS*)' and the *'Université Libre de Bruxelles* (*ULB*)'. A.C.T. is PhD candidate at the '*Fonds pour la formation à la Recherche dans l'Industrie et dans l'Agriculture* (*FRIA*)', Belgium.

Conflict of Interest: none declared.

REFERENCES

Ashburner,M. *et al*. (2000) Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nat. Genet*., **25**, 25–29.

Belov,K. *et al*. (2007) Characterization of the opossum immune genome provides insights into the evolution of the mammalian immune system. *Genome Res*., **17**, 982–991.

Bray,N. *et al*. (2003) AVID: a global alignment program. *Genome Res*., **13**, 97–102.

Brudno,M. *et al*. (2007) Multiple whole genome alignments and novel biomedical applications at the VISTA portal. *Nucleic Acids Res*., **35**, W669–W674.

Curwen,V. *et al*. (2004) The Ensembl automatic gene annotation *system*. *Genome Res*., **14**, 942–950.

Fleischmann,R.D. *et al*. (1995) Whole-genome random sequencing and assembly of Haemophilus influenzae Rd. *Science*, **269**, 496–512.

Force,A. *et al*. (1999) Preservation of duplicate genes by complementary, degenerative mutations. *Genetics*, **151**, 1531–1545.

Gouret,P. *et al*. (2005) FIGENIX: intelligent automation of genomic annotation: expertise integration in a new software platform. *BMC Bioinformatics*, **6**, 198.

Greer,J.M. *et al*. (2000) Maintenance of functional equivalence during paralogous Hox gene evolu-tion. *Nature*, **403**, 661–665.

He,X. and Zhang,J. (2005) Rapid subfunctionalization accompanied by prolonged and substantial neofunctionalization in duplicate gene evolution. *Genetics*, **169**, 1157–1164.

Hubbard,T.J. *et al*. (2007) Ensembl 2007. *Nucleic Acids Res*., **35**, D610–D617.

Huerta-Cepas,J. *et al*. (2007) The human phylome. *Genome Biol*., **8**, R109.

Hurles,M. (2004) Gene duplication: the genomic trade in spare parts. *PLoS Biol*., **2**, E206.

Kelso,J. *et al*. (2003) eVOC: a controlled vocabulary for unifying gene expression data. *Genome Res*., **13**, 1222–1230.

Kidd,K.K. and Sgaramella-Zonta,L.A. (1971) Phylogenetic analysis: concepts and methods. *Am. J. Hum. Genet*., **23**, 235–252.

Kondrashov,F.A. and Kondrashov,A.S. (2006) Role of selection in fixation of gene duplications. *J. Theor. Biol*., **239**, 141–151.

Liolios,K. *et al*. (2006) The Genomes On Line Database (GOLD) v.2: a monitor of ge-nome projects worldwide. *Nucleic Acids Res*., **34**, D332–D334.

Lynch,M. and Conery,J.S. (2000) The evolutionary fate and consequences of duplicate genes. *Science*, **290**, 1151–1155.

Lynch,M. and Force,A. (2000) The probability of duplicate gene preservation by subfunctionalization. *Genetics*, **154**, 459–473.

Mi,H. *et al*. (2007) PANTHER version 6: protein sequence and function evolution data with expanded representation of biological pathways. *Nucleic Acids Res*., **35**, D247–D252.

Mikkelsen,T.S. *et al*. (2007) Genome of the marsupial Monodelphis domestica reveals innovation in non-coding sequences. *Nature*, **447**, 167–177.

Odronitz,F. *et al*. (2007) diArk – a resource for eukaryotic genome research. *BMC Genomics*, **8**, 103.

Ohno,S. (1970) *Evolution by Gene Duplication*. Springler Verlag, Heidelberg.

Pao,S.Y. *et al*. (2006) In silico identification and comparative analysis of differentially expressed genes in human and mouse tissues. *BMC Genomics*, **7**, 86.

Rastogi,S. and Liberles,D.A. (2005) Subfunctionalization of duplicated genes as a transition state to neofunctionalization. *BMC Evol. Biol*., **5**, 28.

Rzhetsky,A. and Nei,M. (1992) Statistical properties of the ordinary least-squares, generalized least-squares, and minimum-evolution methods of phylogenetic inference. *J. Mol. Evol*., **35**, 367–375.

Rzhetsky,A. and Nei,M. (1993) Theoretical foundation of the minimum-evolution method of phylogenetic inference. *Mol. Biol. Evol*., **10**, 1073–1095.

Shiu,S.H. *et al*. (2006) Role of positive selection in the retention of duplicate genes in mammalian genomes. *Proc. Natl Acad. Sci. USA*, **103**, 2232–2236.

Stalker,J. *et al*. (2004) The Ensembl Web site: mechanics of a genome browser. *Genome Res*., **14**, 951–955.

Su,A.I. *et al*. (2002) Large-scale analysis of the human and mouse transcriptomes. *Proc. Natl Acad. Sci. USA*, **99**, 4465–4470.

Zhang,J. (2003) Evolution by gene duplication: an update. *Trends Ecol. Evol*., **18**, 292–298.