The chordate proteome history database.

Anthony Levasseur, Julien Paganini, Jacques Dainat, Julie D Thompson, Olivier Poch, Pierre Pontarotti, Philippe Gouret

Evolutionary Bioinformatics

ORIGINAL RESEARCH

The Chordate Proteome History Database

Anthony Levasseur[1,2,*], Julien Paganini[3,*], Jacques Dainat[3], Julie D. Thompson[4], Olivier Poch[4], Pierre Pontarotti[3] and Philippe Gouret[3]

[1]INRA, UMR1163 Biotechnologie des Champignons Filamenteux, Aix Marseille Université, ESIL Polytech, 163 avenue de Luminy, CP 925, 13288 Marseille Cedex 09, France. [2]Aix Marseille Université, UMR1163 BCF, 163 avenue de Luminy, CP 925, 13288 Marseille Cedex 09, France. [3]UMR7353, Evolutionary Biology and Modeling, Aix Marseille Université, 3 place Victor-Hugo, 13331 Marseille, France. [4]Département de Biologie Structurale et Génomique, IGBMC (Institut de Génétique et de Biologie Moléculaire et Cellulaire), CNRS/INSERM/Université de Strasbourg, Illkirch, France.
*These authors contributed equally to this article. Corresponding author email: anthony.levasseur@univ-amu.fr

Abstract: The chordate proteome history database (http://ioda.univ-provence.fr) comprises some 20,000 evolutionary analyses of proteins from chordate species. Our main objective was to characterize and study the evolutionary histories of the chordate proteome, and in particular to detect genomic events and automatic functional searches. Firstly, phylogenetic analyses based on high quality multiple sequence alignments and a robust phylogenetic pipeline were performed for the whole protein and for each individual domain. Novel approaches were developed to identify orthologs/paralogs, and predict gene duplication/gain/loss events and the occurrence of new protein architectures (domain gains, losses and shuffling). These important genetic events were localized on the phylogenetic trees and on the genomic sequence. Secondly, the phylogenetic trees were enhanced by the creation of phylogroups, whereby groups of orthologous sequences created using OrthoMCL were corrected based on the phylogenetic trees; gene family size and gene gain/loss in a given lineage could be deduced from the phylogroups. For each ortholog group obtained from the phylogenetic or the phylogroup analysis, functional information and expression data can be retrieved. Database searches can be performed easily using biological objects: protein identifier, keyword or domain, but can also be based on events, eg, domain exchange events can be retrieved. To our knowledge, this is the first database that links group clustering, phylogeny and automatic functional searches along with the detection of important events occurring during genome evolution, such as the appearance of a new domain architecture.

Keywords: phylogenetic reconstruction, ortholog groups, protein architecture, functional inference, family size, genome evolution

Evolutionary Bioinformatics 2012:8 437–447

doi: 10.4137/EBO.S9186

This article is available from http://www.la-press.com.

Introduction

The genetic information encoded in the genome sequence contains the blueprint for an organism's potential development, physiology and activity. This information can only be fully comprehended in the light of the evolutionary events acting on the genome (duplication, gains and gene losses, nucleotide substitutions, genome recombination), reflected in changes in the sequence, structure and function of the gene products (nucleic acids and proteins) and ultimately in the organism's biological complexity.

The recent availability of the complete genome sequences of a large number of model organisms means we can now begin to unravel the mechanisms involved in the evolution of the genomes and their implications for the study of biological systems. At the same time, theoretical advances in biological information representation and management have revolutionized the way experimental information is collected, stored and exploited. Ontologies, such as Gene Ontology (GO) or Sequence Ontology (SO),[1] provide a formal representation of the data for automatic, high-throughput data parsing by computers. These ontologies are being exploited in new information management systems to allow large-scale data mining, pattern discovery and knowledge inference.

The vast number and complexity of the events shaping genomes means that a complete understanding of evolution at the genomic level is not currently feasible. At the lowest level, point mutations affect individual nucleotides. At a higher level, large chromosomal segments undergo duplication, lateral transfer, inversion, transposition, deletion and insertion. Ultimately, whole genomes are involved in processes of hybridization, polyploidization and endosymbiosis, often leading to rapid speciation (for a review see[2]).

Several databases dedicated to homologous gene families from vertebrates and microbial organisms have recently been developed for use in comparative genomics projects (for example,[3,4]). At present, the genomic context of a specific gene can be easily displayed using different user-friendly databases[5,6] and the evolutionary dynamics of gene clustering can be accurately inspected. In our study, the main objective is the characterization and study of the evolutionary histories of the chordate proteome, in particular the detection of genomic events and automatic functional searches. We make use of formal descriptions of biological data, together with recent developments concerning automated reliable protein sequence alignment and accurate phylogenetic reconstruction. These approaches have been combined in a multi-agent, expert system for the construction of evolutionary histories to facilitate the automatic definition of the important genetic events shaping a given protein. Here we present the computational strategies that we have developed and the first steps towards our final goal, in the form of a novel database: the chordate proteome history database. This database provides phylogenies for the chordate proteomes, reconstructed using a gene-based approach in which the same high quality phylogenetic pipeline is applied to each individual gene in a given genome. Genomic events, at the gene level or at the protein domain level, were detected automatically and localised on the gene phylogenies and on the genomic sequence, wherever possible. We focused on the orthologous relationships between sequences from 14 species: *Homo sapiens, Pan troglodyte, Pongo abelii, Macaca mulata, Canis lupus familiaris, Mus musculus, Rattus norvegicus, Monodelphis domestica, Gallus gallus, Xenopus tropicalis, Tetraodon nigriviridis, Oryzias lentipes, Amphioxus, Ciona intestinalis* and used these relationships for functional transfer wherever possible. We note that *Amphioxus* is used only in phylogroup analyses. The orthologous associations were obtained by clustering the protein sequences using OrthoMCL,[7] followed by correction based on a detailed phylogenetic analysis. All multiple alignments, phylogenetic trees, and tree-based functional predictions and genomic events affecting protein domain architecture can be easily accessed *via* a web-based user interface.

Materials and Methods

General features

The chordate proteome history database is deployed *via* a web application named Interface for Ontological Data Analysis (I.O.D.A), developed with the Google Web Toolkit technology, which uses Java and Javascript/AJAX languages. Each results menu on the left-hand side of the site can be annotated by a registered user *via* a wiki system on the right-hand side of the site. All users can read these wiki pages when they browse

the database. I.O.D.A is currently fully functional on the browsers Firefox and Google Chrome (download available on the I.O.D.A. homepage). For Macintosh users, I.O.D.A works correctly with MacOS 10.6.7 or higher and Java 1.6.0_24 or higher.

Data model

As its name suggests, I.O.D.A does not rely on a relational database model, but on a more accurate and flexible model structured by an ontology. The ontology used in the chordate database focuses on the specific evolutionary concepts manipulated in the laboratory. More specifically, we use an approach based on mathematical first-order logic named *Description Logic* (http://dl.kr.org/). The W3C-standardized OWL language (http://www.w3.org/TR/owl-ref/) is an XML representation of DL that we use to define our model. The database itself is formed by RDF triples persisting on an underlying relational database server (PostgreSQL: http://www.postgresql.org). The server is not accessed directly, but *via* a JAVA (http://www.java.com) API named Jena (http://jena.sourceforge.net/), which provides access to classes, instances and relationships. Wherever possible, we adopt the Relational Ontology terminology,[8] designed to standardize relationships in biological ontologies. The ontological database model scheme is described in.[9]

Phylogeny construction and event detection

All the phylogenetic trees present in the database were built automatically using the software platform FIGENIX[10,11] driven by the DAGOBAH expert system.[9] DAGOBAH is a multi-agent system in which specific agents have been developed for genetic event detection and verification. The phylogenetic trees were automatically analysed by a Java API: PhyloPattern.[12]

Identification of vertebrate homologs and construction of a multiple sequence alignment

The 19,837 human proteins defined by the Human Protein Initiative (http://expasy.org/sprot/hpi/) were used in this analysis. For each protein, database searches of the Swissprot and Ensembl databases[13] were performed using the BlastP program. Multiple alignments of complete sequences (MACS) were then constructed using the MAFFT program,[14] containing up to 500 full-length protein sequences. The quality of the MACS was then validated using the NorMD objective function, and unrelated sequences were excluded using the LEON program.[15]

Once a high quality MACS was obtained, the next step was to extract structural/functional information related to the protein family from the public databases. This was done using the in-house BIRD data retrieval system, and covered a wide range of information, from taxonomic data and functional descriptions (protein definition, EC number, GO, pFAM, Interpro) to sequence features, such as structural domains and active site residues. The retrieved data was integrated in the multiple alignment, together with a number of *ab initio* calculations (disordered regions, low-complexity segments and transmembrane helices), using the MACSIMS Information Management System.[16]

Construction of an accurate phylogenetic tree

Based on the main FIGENIX phylogeny pipeline, a new phylogeny pipeline was specifically developed to initiate phylogenetic studies from MACSIMS alignment files. In this pipeline, the alignment was intelligently cut to detect alignment areas associated with specific protein domains and repeats. For each domain, a phylogenetic tree was built and used for the study of domain architecture events.

In addition, a gene-level phylogeny was produced. All alignment areas associated with the domains in the protein query (the one that initiated the alignment) were concatenated and the resulting alignment was used for tree building. The gene phylogenies were used to study gene losses/gains and horizontal gene transfers and to compile duplication events and orthology and paralogy relationships.

Functional data

From all the homolog pages in I.O.D.A, the user can search functional data from: GO,[1] KEGG,[17] ArrayExpress,[18] String,[19] and QuickGO.[20] To do this, I.O.D.A converts Ensembl references to Uniprot references, which are all indexed in these databases.[21] To extract the functional data, these references are then sent to the web services associated with each of these databases. I.O.D.A presents the functional data, either directly on the web pages or through a link to these sites.

New protein domain architecture events, localization on the chordate species tree and verification at genome level

An apomorphic protein can be formed by any of five kinds of events detected by a dedicated DAGOBAH agent:

- *Gain*: one or more domains are gained at the beginning or end of the ancestral protein,
- *Loss*: one or more domains are lost at the beginning or end of the ancestral protein,
- *Insertion*: one or more domains are inserted between two domains of the ancestral protein,
- *Deletion*: one or more domains are deleted between two domains of the ancestral protein,
- *Shuffling*: one or more domains are exchanged at the beginning or end of the ancestral protein with a pendant protein.

The general strategy for domain event detection involved a nine-step process driven by the DAGOBAH multi-agents system:

1. Domain-annotated protein alignments built from a query protein are used to outsource phylogeny tree construction (domain trees and protein trees) to the FIGENIX pipeline.
2. The Mirkin parsimony algorithm[22] is used on each tree produced to infer ancestral domain architectures on internal nodes. Unfortunately, no efficient algorithm is currently available to infer the order of ancestral domain architectures.
3. The query's domain architecture is divided into a list of consecutive domain pairs. We note that two artificial domains (without any associated phylogenetic tree) are added at the tips, in order to study events occurring at the beginning and end of the protein. For example, for a protein with three domains A, B and C, our process studies each of the four pairs *[A-before, A], [A, B], [B, C],* and *[C, C-after]*. For each pair, the phylogenetic trees produced at step 1 are used in the first steps of event detection (steps 4–6).
4. Ideally, a phylogenetic pattern consistent with the event should be found on each domain tree of a pair, which strengthens the event hypothesis. Nevertheless, events found only on one tree of the domain pair are considered as valid, but weaker, candidates. Two patterns are applied on the domain pair trees with our API: PhyloPattern,[12] one for deletion events and one for other events (Fig. 1). A pattern is a triple, ie, a well-supported ancestral node with two children: a plesiomorphic node corresponding to a domain architecture close to the ancestral one, an apomorphic node corresponding to the derived domain architecture.

   The pattern associated with a deletion event candidate is an ancestral node with the two domains of the pair and other domains (denoted DL) located between them, an apomorphic child node with the two domains of the pair whose subtree contains the query sequence, and a plesiomorphic child node with the two domains of the pair whose representative sequence contains the DL domain list.

   The pattern associated with other event candidates is an ancestral node with one of the two domains of the pair, ie, the one for which the tree receives the pattern, an apomorphic child node with the two domains of the pair, and a plesiomorphic child node with one of the two domains of the pair, ie, the same one as in the ancestral node.

   We note that when we refer to a node's domains, we mean the inferred architecture for an internal node or the known architecture for a leaf.
5. The choice of apomorphic and plesiomorphic representative sequences is very important for the

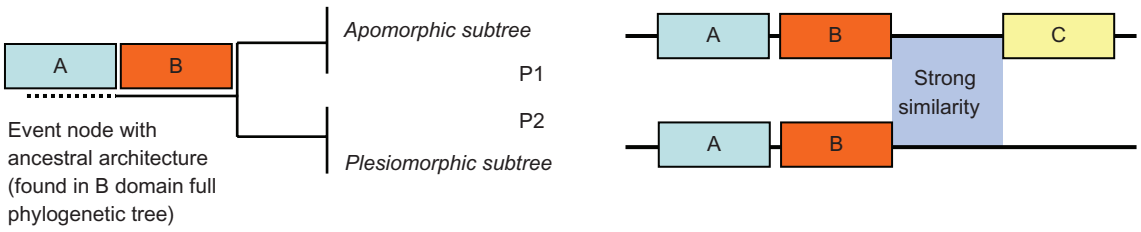

Figure 1. A virtual example of an event leading to a novel domain architecture.
Notes: Here a gain event is confirmed because the genome sequences between domains B and C on the apomorphic sequence and after domain B on the plesiomorphic sequence are strongly conserved. P1 and P2 indicate the two protein sequences chosen as representatives of the apomorphic and plesiomorphic subtrees.

subsequent steps. It will influence the reliability of event type determination (step 6), and also the reliability of event verification at the genomic level (step 8). Thus the process chooses the least remotely derived sequences, ie, the sequences with the domain architecture closest to the ancestral one and with the shortest branch to the ancestral node. These sequences are assumed to have accumulated fewer mutations and recombinations than the others. For the choice of apomorphic sequence, there is an exception to this last rule: when an event candidate's pattern is relevant for the two trees of the domain pair, we choose a sequence that belongs to the two apomorphic subtrees found in the two trees of the domain pair (for deletion candidates, we choose the query). When the criteria are not sufficient to choose between plesiomorphic sequences, the sequence closest to the apomorphic sequence species in the species tree is chosen.

6. The next step is to determine precisely the type of the event and all the domains involved in the transformation. This is done by computing the difference between the representative plesiomorphic and apomorphic sequence domain architectures. Sometimes several event types are similar. We will see at step 7 how we validate the event type. Table 1 summarizes all the different cases.
7. To produce "definitive" conclusions, the process confronts each individual event candidate (produced by the study of all the domain pairs in the query sequence architecture) with all the others, through an expert system, applying logical rules. We cannot give details of all the specific rules here, but their aim is to group some individual events or to remove some ambiguity, whenever possible. Non-grouped events are identically conserved. As an example of grouped event candidates, given an apomorphic domain architecture A B C D, the process could identify two insertion candidates by studying the A, B pair and the C, D pair, but they are probably linked to a single event, the insertion of B and C between A and D.

   In addition, we can see that the "shuffling/insertion" ambiguity in Table 1 could also be removed if, for example, the plesiomorphic architecture was A D when the process studied the A domain tree. In this case, the shuffling hypothesis is eliminated.
8. When event candidates are confirmed, the next step is to try to verify them at the genome level, by trying to find an alignment break position between two DNA segments, one associated with the representative apomorphic protein and the other with the representative plesiomorphic protein. DNA segments are extracted between concerned domains using Ensembl online access, and a Blast (tblastn) search is then performed on the DNA regions associated with the proteins, using the domain's amino acid segments as a query. Overlapping of Blast high similarity pairs is managed to extract the most significant area.

   BlastZ[23] is then used to align the two segments and to detect the alignment break position that should be the recombination point. More details of this process can be found in.[9]

   We note that for many events we find no such position, because the divergence date between the apomorphic and plesiomorphic species is often too far distant, and many other accumulated events have since masked the recombination event. When this information is found, it is supplied to the I.O.D.A user in an "Expert comment" field.

Table 1. In the studied query sequence, the domain pair A B or a pair with a virtual tip A A-after is shown in bold.

Event type	Plesiomorphic architecture part	Apomorphic architecture part	Event description
Gain	A A-after	**A B X**	Gain of domains B and X after domain A
Loss	A X	**A after**	Loss of domains X after domain A
Insertion	A X	**A B Y X**	Insertion of domains B and Y between A and X
Deletion	A X B	**A B**	Deletion of domains X between A and B
Shuffling	A X	**A B Y**	Replacement of domains X by domains B and Y
Shuffling/ insertion	A X	**A B Y X**	Replacement of domains X by domains B, Y and X or Insertion of domains B and Y between domains A and X

Notes: X and Y indicate lists of other domains. The event candidate is detected on the phylogenetic tree of domain A. When the tree of domain B is studied, a symmetric case is obtained.

In "ideal" studies, we identify two close recombination points on the common apomorphic sequence found on the two domain trees that show the event. If the same position is identified based on two different plesiomorphic sequences, then the event hypothesis is very strongly supported. However, these cases are very infrequent in the database.

9. We introduced this final step to detect, *a posteriori*, the most obvious artefacts. An artefact probability is supplied to the I.O.D.A user for all events. Our process detects two kind of artefacts:
   - Wrong propagation of domain architecture in the MACSIMS alignments (used to initiate our studies). The artefact detection agent re-predicts the apomorphic and plesiomorphic sequence domain architectures from the Pfam database to verify them.
   - Sequencing, assembling or gene prediction errors in the genomes used in this study. This agent is able to detect frequent artefacts resulting from the use of a gene isoform as an apomorphic one, although the plesiomorphic variant still exists, or reciprocally as a plesiomorphic one when the apomorphic variant exists.

Phylogroups and gene loss/gain study

The OrthoMCL algorithm was used to create groups of orthologous sequences from the same set of species as used for the phylogeny reconstruction. Phylogroups were created by clustering the groups using ortholog information obtained by the phylogenetic analyses. The "gene loss/gain" module is based on the phylogroup analysis. Gene gain and loss events were identified using the PARS algorithm.[22] As gene transfer in chordates is unlikely, a gene gain was assumed to occur only once. An event that occurred more than once was thus assumed to signal an artefact. The PARS algorithm minimizes the gain and loss events. For example, when an ortholog is frequently absent on a given tree, the algorithm predicts several gains. These cases should be considered as putative artefacts, possibly due to problems with the sequencing/assembly process.

Rules for event validation

If orthologs are recorded absent only on the leaves (except for the well-annotated genome: human and mouse), the expert system (a DAGOBAH agent) will not confirm the loss, which might be due to an annotation artefact or unfinished sequences. If the orthologs are recorded absent higher up the tree, or if all the orthologs are also absent in daughter branches, then DAGOBAH will valid the loss events.

External access

To facilitate access to all the data contained in the chordate database, I.O.D.A entries can be easily linked to and from external pages using the URL: http://ioda.univ-provence.fr/IodaSite/Site.jsp?id=XXXXX, where XXXXX can be replaced by any reference or keyword searchable on the I.O.D.A site (eg, P35125, which is a Uniprot reference). In this way, other databases focused on specific themes can include additional evolutionary information in their data.

Results and Discussion

The data in the chordate proteome history database are divided into two subprojects. The first subproject includes phylogenies, new architecture and duplication events. The second one is dedicated to chordate phylogroups analysis.

Phylogenetic data

As we were interested in the evolution of the human proteome, the scope of the phylogenetic analyses was limited to the chordate, focusing exclusively on well-annotated genome species. The phylogenetic analysis was assumed to be robust for small families, as all the homologous sequences should be present, forming reliable ortholog and paralog groups. Based on the phylogenetic tree, genetic events that affect different protein characteristics were investigated, including orthology/paralogy and domain architecture.

Functional data for the different orthologous groups were collected from the GO,[1] KEGG,[17] ArrayExpress,[18] STRING[19] databases, and links are provided to the Ensembl,[21] Uniprot and Pfam[24] databases and the NCBI taxonomy.[25]

Phylogroups

The phylogroup analysis is used as a filter and provides information about the size of the gene families, about potential gene loss in a given lineage, and finally about the appearance or gain of a novel gene family. Phylogroups are in fact OrthoMCL[7] ortholog groups that we overclusterized using orthology relationships

offered by the automatic analysis of phylogenetic trees produced. OrthoMCL clustering can lead to artefact groups made up of fast-evolving orthologs. We correct this artefact by clustering the groups using ortholog information obtained by the phylogenetic analyses. Thus several OrthoMCL groups can be integrated in a single group, denoted "phylogroup". The phylogroups can then be used to perform functional analyses as described for phylogenetic analyses.

In addition, the phylogroups are exploited in the evolutionary analyses for the detection of events such as gene loss and gene appearance. Gene appearance can be the result of various scenarios,[26] eg, (i) pseudo-appearance due to duplication followed by high rates of mutation, (ii) gene rearrangement leading to different domain architectures in the orthologs, (iii) horizontal gene transfer (only a few examples in the chordates) or (iv) de novo genes.

Our phylogroup approach and the associated gene loss and gain results offer a number of advantages over other published ortholog databases that use clustering: (i) the ortholog group is corrected by the phylogeny and (ii) we include expert rules to give greater confidence to the ortholog loss/gain events.

Database access and web interface features

Browsing and querying

The chordate proteome history database is publicly accessible at http://ioda.univ-provence.fr. The database is organized in two interconnected projects: (i) domain events and phylogenies and (ii) chordate phylogroups and gene loss/gain. The two subprojects are linked: the corresponding phylogroup can be accessed from a gene's phylogeny study page, and conversely, the domain events and phylogeny studies can be accessed from the phylogroup page.

The database can be browsed using the "search" window by entering various queries, eg, (i) the human protein name, using Ensembl or Uniprot identifiers, (ii) the Ensembl identifier for nonhuman species, (iii) key words, (iv) one or more domain names, (v) partial domain names or (vii) a combination of these key words. We note that numbered information and user guidelines are provided in wiki pages.

Phylogenies and domain event searches in the phylogeny subproject

The phylogenetic reconstructions for each gene are available and can be retrieved directly from queries. The phylogeny subproject can be searched for events leading to new domain architectures, ie, caused by the loss or gain of a domain or domain shuffling. Figure 2 shows an example of a search using the UniProtKB/Swiss-Prot accession number P35125 as a query. By clicking on the search window (entry page), two results pages are available; phylogenetic study and events studies (see Fig. 2).

The phylogenetic study results page includes the phylogenetic trees, the ortholog list with the functional

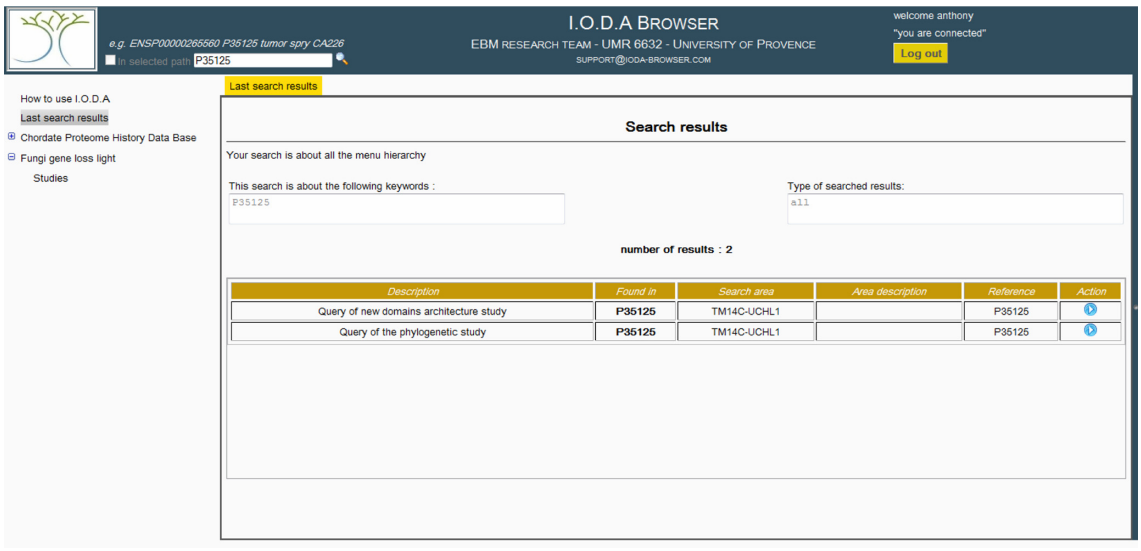

Figure 2. The chordate proteome history database entry page.
Notes: The entry page of a query protein (P35125) includes links to two available results: (i) phylogeny study and (ii) events study.

links, the paralog list and the list of homologs if the phylogenetic analysis results in some weakly supported nodes.

The events study results page includes links to each possible type of domain architecture evolution, ie, domain shuffling, domain insertion or deletion inside the sequence and domain loss or gain at the N- or C-terminus. For example, in the case of P35125, a domain shuffling event was detected (Fig. 3). By clicking on the "Shuffling events" tab and selecting a specific shuffling event, the user gains access to two information pages: "from Tree" and "Event pattern" (Fig. 4). The "from Tree" tab shows the phylogenetic tree used to deduce the event, together with the domain organization of the leaf sequences. In addition, the branch on which the event is assumed to occur is identified. The "Event pattern" tab provides more details about the domain organization of the apomorphic (derived) and the plesiomorphic (similar to the ancestral) representative sequences.

Phylogroup subproject

By clicking on the "Chordate phylogroups and gene loss/gain" and "Studies" menus, the study box shows the ortholog distribution on the different species under investigation. The "Group statistics" menu gives the user an overall view of the group distribution, the sequence number and the number of events, while the "Groups" menu gives the list of all ortholog groups. The tree box shows the species tree where the gene appears and when it is lost. The ortholog box provides the list of all the orthologs, and the functions of the ortholog sequences can be easily retrieved by clicking on the functional request button (see below: *Gulo* gene, for example).

A case study: the example of *Gulo* gene analysis from phylogroup data and phylogenies

The *Gulo* gene encodes an enzyme known to be involved in the pathway of vitamin C biosynthesis. This gene has been lost in primates,[27,28] resulting in the inability of primates to produce vitamin C. Any GULO protein found in our selected species could be used; for example, the user can type the mouse protein reference: ENSMUSP00000060912 in the search toolbar. Several results are available in the phylogroup or phylogeny analyses. Firstly, the phylogroup

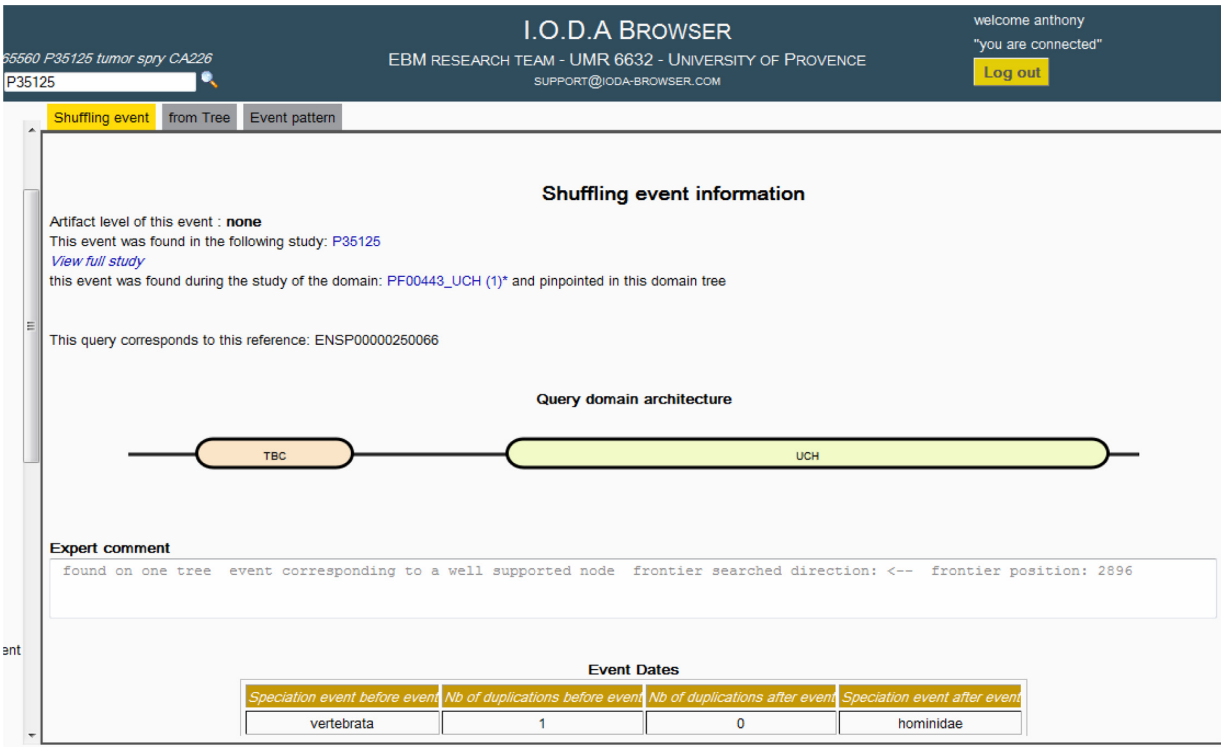

Figure 3. Domain structure organization.
Notes: The events study results page provides links to domain architecture evolution, eg, domain shuffling, domain insertion or deletion, domain loss or gain. In this example (P35125), a shuffling domain exchange was detected.

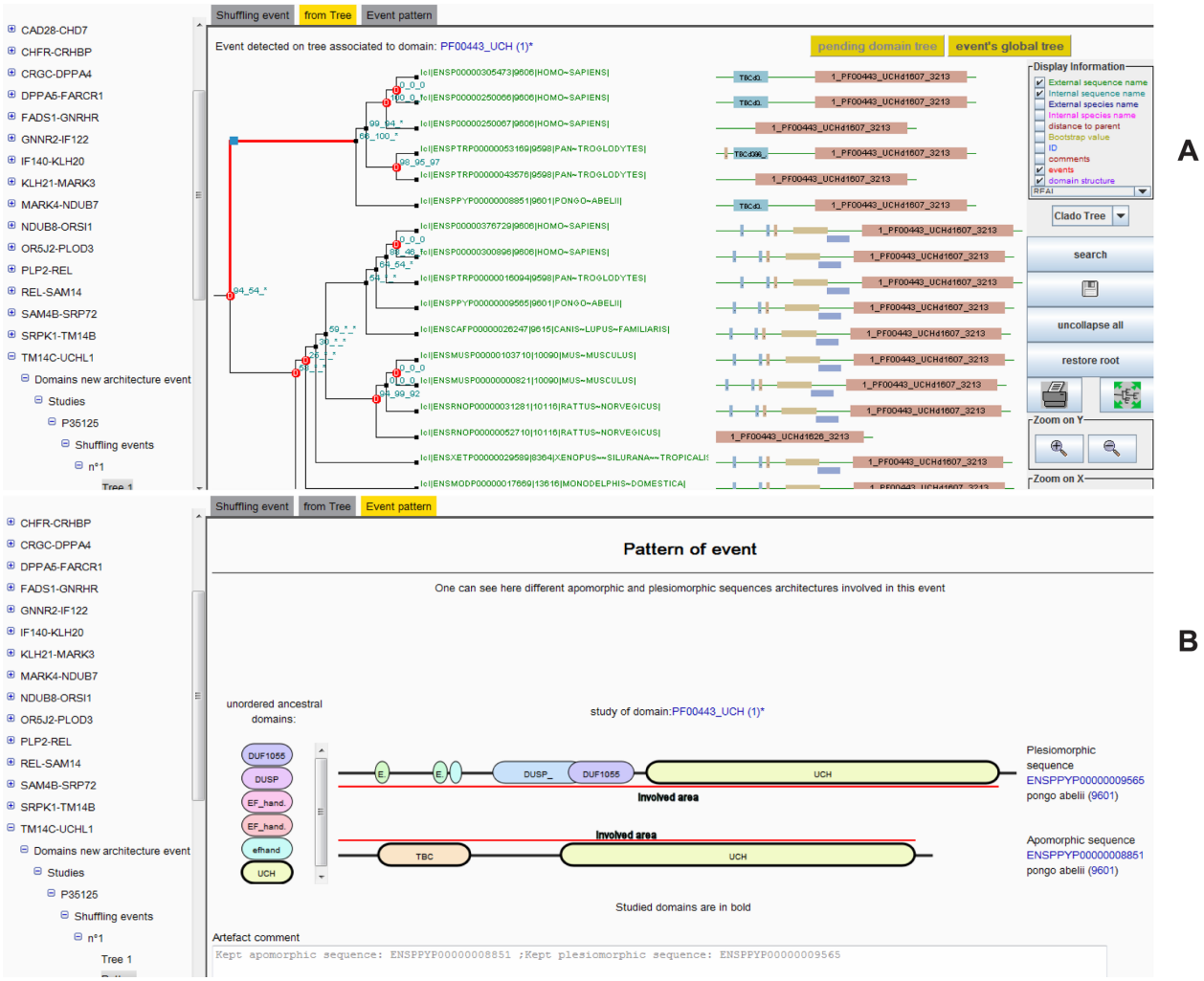

Figure 4. Tree and event pattern pages.
Notes: (**A**) The "from Tree" tab depicts the topology of the tree on the left-hand side and the domain organization for the leaves on the right-hand side. Gene duplications (red circles) and any detected domain architecture events (blue rectangles) are localized on the tree. Bootstrap values for each node are shown as a triplet corresponding to the three algorithms used to construct the tree. (**B**) The "Event pattern" tab shows the domain organization of the apomorphic (derived) and the plesiomorphic (similar to the ancestral) representative sequences.

results (OG_113469) indicate that the protein is found in 8 out of 14 species. In the *tree* tab, the loss events associated with this phylogroup are depicted (Fig. 5). This orthologous group existed before the last chordata ancestor, and subsequently two loss events in primates and actinopterygii ancestors occurred. The gene loss in teleosts has been observed previously[29] and this result agrees with the loss inferred in actynopterygii. These two loss events explain the six missing species and agree with the results already published. Secondly, the user can browse the phylogeny analysis in which ENSMUSP00000060912 is present (ie, Q15392: *All trees* tab) and examine the phylogeny based on the Q15392 entire protein by clicking on *Protein's best tree*. According to the phylogenetic tree, Q15392 is paralogous to Gulo. The Gulo ortholog group (paralogous to the Q15392 ortholog group) is found in this phylogenetic analysis and confirms that the gene is missing in both primates and actinopterygii.

New protein domain architecture: the example of shuffling events

A number of shuffling domain exchanges could be evidenced by using the database (as described above in the case of P35125). To summarize, 1943 shuffling domains were reported in the current version of the database. These 1943 shuffling events exclude all putative artefacts and could be assigned as relevant shuffling events with no ambiguity. In the field of new

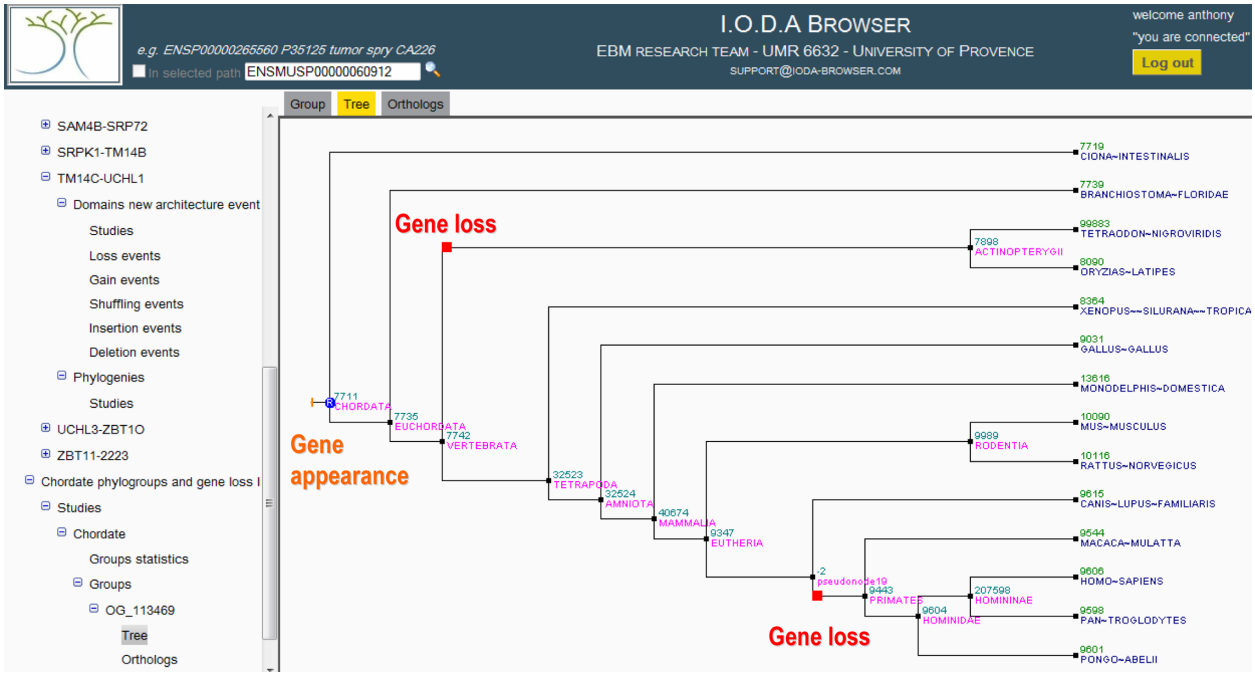

Figure 5. Detection of gene gain and loss in phylogroups.
Note: Example of *Gulo* gene analysis (ENSMUSP00000060912), the gene appearance and loss are directly depicted in the phylogenetic tree.

gene origination, these shuffling events are of prime importance for users, as the creation of new proteins/function could be carried out by bringing different domains together.[2]

Conclusions and Perspectives

In summary, the chordate proteome history database combines ortholog clustering, phylogeny and automatic functional link searches with automatic detection of important genomic events at the gene or protein domain levels. We are focusing on new enhancements for the medium-term including: (i) detection of other evolutionary events to achieve a more overall view of the genomic changes (eg, pseudogenization), (ii) introduction of other chordate genomes thanks to the current growing number of genomes sequenced and improved quality (structural and functional annotation) of the present genomes and (iii) development of other databases focusing on different kingdoms (eg, fungi) under the I.O.D.A. umbrella.

Author Contributions

Conceived and designed the experiments: AL, JP, PP, PG. Analysed the data: AL, JP, JD, JDT, OP, PP, PG. Wrote the first draft of the manuscript: AL, JDT, PP, PG. Contributed to the writing of the manuscript: AL, JDT, PP, PG. All authors reviewed and approved of the final manuscript.

Competing Interests

Author(s) disclose no potential conflicts of interest.

Funding

This research was supported by the ANR EvolHHuPro (ANR-07-BLAN-0054-01).

Disclosures and Ethics

As a requirement of publication author(s) have provided to the publisher signed confirmation of compliance with legal and ethical obligations including but not limited to the following: authorship and contributorship, conflicts of interest, privacy and confidentiality and (where applicable) protection of human and animal research subjects. The authors have read and confirmed their agreement with the ICMJE authorship and conflict of interest criteria. The authors have also confirmed that this article is unique and not under consideration or published in any other publication, and that they have permission from rights holders to reproduce any copyrighted material. Any disclosures are made in this section. The external blind peer reviewers report no conflicts of interest.

References

1. Harris MA, Clark J, Ireland A, et al. The Gene Ontology (GO) database and informatics resource. *Nucleic Acids Res*. 2009;32:D258–61.
2. Levasseur A, Pontarotti P. The role of duplications in the evolution of genomes highlights the need for evolutionary-based approaches in comparative genomics. *Biol Direct*. 2011;6:11.
3. Huerta-Cepas J, Capella-Gutierrez S, Pryszcz LP, et al. PhylomeDB v3.0. an expanding repository of genome-wide collections of trees, alignments and phylogeny-based orthology and paralogy predictions. *Nucleic Acids Res*. 2011;39:D556–60.
4. Penel S, Arigon AM, Dufayard JF, et al. Databases of homologous gene families for comparative genomics. *BMC Bioinformatics*. 2009;16:10.
5. Lopez MD, Samuelsson T. eGOB: eukaryotic Gene Order Browser. *Bioinformatics*. 2011;27:1150–1.
6. Wang D, Zhang Y, Fan Z, Liu G, Yu J. LGCbase: A comprehensive database for lineage-based co-regulated genes. *Evol Bioinform Online*. 2012;8:39–46.
7. Li L, Stoeckert CJ Jr, Roos DS. OrthoMCL. Identification of ortholog groups for eukaryotic genomes. *Genome Res.* 2003;13:2178–89.
8. Smith B, Ceusters W, Klagges B, et al. Relations in biomedical ontologies. *Genome Biol*. 2005;6:R46.
9. Gouret P, Paganini J, Dainat J, et al. Integration of evolutionary biology concepts for functional annotation and automation of complex research in evolution: the multi-agent software system DAGOBAH, evolutionary biology—concepts, biodiversity, macroevolution and genome evolution, Chap. 5, 2011 Springer; In press.
10. Gouret P, Danchin EGJ, Gilles A, Vitiello V, Balandraud N, Pontarotti P. FIGENIX: Intelligent automation of genomic annotation: expertise integration in a new software platform. *BMC Bioinformatics*. 2005;6:198.
11. Paganini J, Gouret P. Reliable phylogenetic trees building: a new web interface for FIGENIX. *Evolutionary Bioinformatics*. 2012;In press.
12. Gouret P, Thompson JD, Pontarotti P. PhyloPattern. regular expressions to identify complex patterns in phylogenetic trees. *BMC Bioinformatics*. 2009; 10:298.
13. Flicek P, Aken BL, Ballester B, et al. Ensembl's 10th year. *Nucleic Acids Research*. 2010;38:D557–62.
14. Katoh K, Misawa K, Kuma K, et al. MAFFT: a novel method for rapid multiple sequence alignment based on fast Fourier transform. *Nucleic Acids Res*. 2002;30:3059–66.
15. Thompson JD, Prigent V, Poch O. LEON: multiple aLignment Evaluation Of Neighbours. *Nucleic Acids Res*. 2004;32:1298–307.
16. Thompson JD, Muller A, Waterhouse A, et al. MACSIMS: multiple alignment of complete sequences information management system. *BMC Bioinformatics*. 2006;7:318.
17. Kanehisa M, Goto S. KEGG: kyoto encyclopaedia of genes and genomes. *Nucleic Acids Res*. 2000;28:27–30.
18. Parkinson H, Sarkans U, Kolesnikov N, et al. ArrayExpress update—an archive of microarray and high-throughput sequencing-based functional genomics experiments. *Nucleic Acids Res*. 2011;39:D1002–4.
19. Szklarczyk D, Franceschini A, Kuhn M, et al. The STRING database in 2011: functional interaction networks of proteins, globally integrated and scored. *Nucleic Acids Res*. 2011;39:D561–8.
20. Binns D, Dimmer E, Huntley R, Barrell D, O'Donovan C, Apweiler R. QuickGO: a web-based tool for Gene Ontology searching. *Bioinformatics*. 2009;25:3045–6.
21. Hubbard TJ, Aken BL, Ayling S, et al. Ensembl. *Nucleic Acids Res*. 2009;37:D690–7.
22. Mirkin BG, Fenner TI, Galperin MY, Koonin EV. Algorithms for computing parsimonious evolutionary scenarios for genome evolution, the last universal common ancestor and dominance of horizontal gene transfer in the evolution of prokaryotes. *BMC Evol Biol*. 2003;3:2.
23. Altschul SF, Madden TL, Schaffer A, et al. Gapped BLAST and PSI-BLAST—a new generation of protein database search programs. *Nucleic Acids Res*. 1997;25:3389–402.
24. Finn RD, Mistry J, Tate J, et al. The Pfam protein families database. *Nucleic Acids Res*. 2010;Database Issue 38:D211–22.
25. Sayers EW, Barrett T, Benson DA, et al. Database resources of the National Center for Biotechnology Information. *Nucleic Acids Res*. 2009;37:D5–15.
26. Long M, Betran E, Thornton K, Wang W. The origin of new genes: glimpses from the young and old. *Nat Rev Genet*. 2003;4:865–75.
27. Nishikimi M, Fukuyama R, Minoshima S, Shimizu N, Yagi K. Cloning and chromosomal mapping of the human nonfunctional gene for L-gulono-gamma-lactone oxidase, the enzyme for L-ascorbic acid biosynthesis missing in man. *The Journal of biological chemistry*. 1994;269: 13685–8.
28. Ohta Y, Nishikimi M. Random nucleotid substitution in primate nonfunctionnal gene for L-gulono-gammalactone oxidase, the missing enzyme L-ascorbic acid biosynthesis. *Biochimica et Biophysica Acta* (BBA). 1999;1472:408–4141.
29. Maeland A, Waagbø R. Examination of the qualitative ability of some cold water marine teleost to synthesis ascorbic acid. *Comparative Biochemistry and Physiology*. 1998;121:249–55.