Trends in Developmental Biology
Vol. 12, 2019

Original Article

Revealing the clade of origin of Gene Ontology terms related to developmental processes

Carlos A. X. Gonçalves and J. Miguel Ortega*
Department of Biochemistry and Immunology, ICB, UFMG, Belo Horizonte, MG, Brazil.

ABSTRACT

Gene Ontology (GO) is a database comprised of terms describing biological information that can be associated to gene products, categorized in "biological process", "molecular function" and "cellular component". These GO terms have been attributed to genes of diverse organisms from unicellular to plants and animals. Many terms have been annotated to several different species and can be looked upon through a comparative biology view. This allows us to inspect their taxonomic distribution, from which we can infer the clade of origin of the processes and functions described by these Gene Ontology terms. We studied the taxonomic distribution and inferred a putative ancestral clade for GO terms related to development, focusing on children of the term "developmental process" GO:0032502. We observed that while some biological processes are ancient, there are processes which have originated more recently in evolutionary history, some being restricted to placental mammals, or plants with flowers. Our approach was able to reveal the approximate period in evolution when the processes have risen.

KEYWORDS: Gene Ontology, development, origin, organs.

1. INTRODUCTION

Ever since life originated on Earth, circa four billion years ago [1], it has spread out through evolution taking all sorts of forms, shapes, features and characteristics. If it took two of those four billion years for the first big rupture to appear, separating the domains of Archaea and Bacteria [2, 3], the more recent half of that time saw countless of those evolutionary ruptures shaping what eventually became the tree of life of organisms that populate the planet.

As genes appeared and evolved, they became the building blocks for complex biological systems that allowed organisms to develop new structures and conquer new metabolisms [4, 5]. As these systems were retained on descendant organisms, shared between them by the homology characterizing their common origin, they became enriched and further specified, as they acquired new genes down the trail of evolution. Complex processes such as these are often represented as biological pathways, which are commonly reunited on databases such as Kegg Pathway [6] and Panther [7].

As we attempt to study these pathways as they are constructed within their organisms at the present, it is hard to unravel when such complex processes became functional during evolution. To try and unravel such mysteries, we decided to turn to another type of data describing biological phenomena: ontological data.

Gene Ontology (GO) is a consortium created with the objective of cataloging all biological information associated with existing gene products in a massive database. In order to do that, GO makes use of three collections of terms (or ontologies), which aim to cover everything that is possible to

*Corresponding author: miguel@ufmg.br

describe about genes: the collections of "Biological Process", "Molecular Function" and "Cell Component" [8].

Biological Processes refer to complex biological systems such as those aforementioned, which involve the collaborative performance of multiple genes. They are used to describe metabolic pathways, regulatory mechanisms, events associated with the development of tissues or organs, etc. Molecular functions describe information directly associated with the biochemical role of genes. They mainly comprise enzymatic activities performed by genes and their ability to bind to ions or other molecules. Cellular components, in turn, are terms to indicate where gene products can be found in cells. They can refer to membranes, organelles, cytoplasmic spaces, nucleus or even specific intercellular regions (such as synapses).

GO contains both more generic terms and more specific terms. As a result, they are organized hierarchically in the database, in which some terms are taken as ancestors of others. A term can have multiple ancestral and descendant terms, which means that each ontology can be computationally represented in the form of a directed acyclic graph.

The association of a term from Gene Ontology to a gene product can be done by manual curation or electronically. Manual annotations require a reference in the literature that precisely documents the information to be annotated for the protein under study, while electronic annotations are usually performed by transferring annotations between similar proteins. Electronic annotations are often done for molecular function. New annotations are made without any discretion in relation to annotations previously made for the same protein, so that redundancies in the base are common.

The three ontologies of the GO database are comprised of terms that can be annotated to gene products. Since each gene product is specific to a given species, we can look upon the tree of life to search for the ancestral node from which all organisms containing any given GO term descend. In other words, we unraveled the origin of every biological process and molecular function by determining the lowest common ancestor (LCA) [9] clade, or ancestral clade on the tree of life that originated all of the organisms annotated to each term. In this work we sought to reveal the ancestral clade of GO terms related to developmental processes.

2. MATERIALS AND METHODS

Gene Ontology data describing ontological structure (terms and relationships established between them) and gene product annotation was downloaded from the Gene Ontology FTP repository located at: ftp://ftp.geneontology.org/pub/go/ [8]. NCBI Taxonomy data was downloaded from the FTP repository located at: ftp://ftp.ncbi.nlm.nih.gov/pub/taxonomy/ [10]. The taxonomic identifiers associated to all gene products annotated with each GO term were obtained from the annotation file and used to calculate the LCA clades based on the NCBI Taxonomy tree of life. Computational representations of the GO hierarchical ontologies and of the NCBI taxonomy were created through graphs using the networkX module of the Python3 language and in-house scripts when necessary. GO terms related to "development" were textually queried and had their origins and overall annotation distribution manually inspected.

3. RESULTS

3.1. Development in Gene Ontology

The queried word "development" in Gene Ontology returns 2877 Biological Process terms, such as "eye development" (GO:0042460), "pollen development" (GO:0009564) and "heart development" (GO:0007511). It also returns 63 Molecular Function terms containing the word development; however, even though they are terms created by the consortium to exist in the ontology, they have not been so far attributed to any proteins, and therefore will not be analyzed in respect to their putative ancestral clade. The Biological Process terms attributed to the top ten more species are shown in Table 1.

Amongst these, the most attributed term is GO:0032502, "developmental process", which we elected to describe alongside its descendant terms. This term was also the one with most children terms bearing "development" in term name: 352 (numbers within parentheses in Table 1).

Table 1. Top ten terms bearing “development” in term name.

GO identifier	GO term name (#children development)	#Species
GO:0007275	multicellular organism development (19)	22990
GO:0007399	nervous system development (0)	3428
GO:0032502	developmental process (352)	3428
GO:0007517	muscle organ development (9)	594
GO:0007417	central nervous system development (0)	572
GO:0048666	neuron development (29)	517
GO:0031175	neuron projection development (8)	445
GO:0007507	heart development (0)	420
GO:0040034	regulation of development, heterochronic (7)	405
GO:0007286	spermatid development (0)	402
GO:0008544	epidermis development (2)	386

3.2. Developmental process, GO:0032502

When a protein is being annotated to a GO term, the curator will always attempt to use the most specific term possible in relation to the information they have as evidence for the annotation (usually the result of a scientific investigation described on a research article). However, often a protein will be annotated with several terms hierarchically related to each other, mostly because for the same protein different annotators had information which could be more or less specific. Whenever proteins of only one species receive annotation for a given GO term, annotation might have been conducted by a group of investigators specially interested in that organism. Therefore, after manual inspection of the database, we determined a threshold of 10 species to accept that the term underwent being subject of comparative biology, hence spreading throughout the taxonomic tree. Table 1 shows that GO:0032502 is largely used, being attributed to 3428 species. Table 2 shows the ontology children of GO:0032502. One term has been annotated to 2848 species. Six terms were attributed to proteins of less than ten species, and since these terms did not pass the chosen threshold for this analysis, they will not be considered. One of them, “development maturation” (GO:0021700), was attributed to only three species, which were *Homo sapiens*, *Mus musculus* and *Rattus novergicus*, three model organisms whose LCA is the ancestor of primates and rodents, the Euarchontoglires clade. Therefore, if this term was propagated to any wild or domesticated animal such as a bat or dog, the ancestral clade would move to Boreoeutheria; if further attributed to a chicken, it would go to the origin of the amniotes; and to a fish, it would be determined as the ancestral clade of Euteleostomi. Thus, the number of species must be reasonable to suggest that the term went into a comparative biology routine, in opposition to being associated only with physiological studies which tend to be more restrict as the example given. For instance, Figure 1 shows a common taxonomic tree for the 42 species which have had the term “developmental induction” (GO:0031128) from Table 2 annotated to them.

It can be noticed that the ancestral node Amniota is supported by the occurrence of the term in proteins from organisms of four classes: Mammalia, the most prevalent, Aves, with five organisms, Lepidosauria and the Class of Testudines, both with one species annotated. Therefore, the assignment to an ancient node such as Amniota does not depend on a too extensive occurrence along the taxonomic tree, even few approaches of comparative biology may support the asserting of an ancestral node. Remarkably, if very deeply annotated invertebrates such as *Drosophila melanogaster* would have the term

Table 2. Children of developmental process GO:0032502.

GO term	GO term name	Ancestral clade	#Species
GO:0009847	spore germination	Root	2848
GO:0043934	sporulation	Root	2700
GO:0007568	aging	cellular organisms	408
GO:0022611	dormancy process	cellular organisms	357
GO:0009653	anatomical structure morphogenesis	Root	249
GO:0048589	developmental growth	Eukaryota	190
GO:0048856	anatomical structure development	Eukaryota	114
GO:0009838	abscission	Eukaryota	95
GO:0003006	developmental process involved in reproduction	Eukaryota	87
GO:0048646	anatomical structure formation involved in morphogenesis	Euteleostomi	72
GO:0060033	anatomical structure regression	Bilateria	66
GO:0010014	meristem initiation	Magnoliopsida	56
GO:0044111	development involved in symbiotic interaction	Amniota	51
GO:0098727	maintenance of cell number	Eukaryota	51
GO:0007571	age-dependent general metabolic decline	Neopterygii	44
GO:0031128	developmental induction	Amniota	42
GO:0021700	developmental maturation	Euarchontoglires	3
GO:0048532	anatomical structure arrangement	Murinae	2
GO:0048869	cellular developmental process	Candida albicans SC5314	1
GO:0043696	dedifferentiation	Unknown	0
GO:0097737	acquisition of mycelium reproductive competence	Unknown	0
GO:0090644	age-related resistance	Unknown	0

annotated to any of its proteins, that would make the calculated clade of origin to be more ancient; thus, the existence of greatly annotated distant genomes grants that the origin is indeed more recent, supporting the approach.

3.3. Plant development

A child of “Developmental process” (Table 2) belongs to plant lineages, GO:0010014 “meristem initiation”. This term is annotated to 56 species from 19 orders: Alismatales, Amborellales, Apiales, Arecales, Asterales, Brassicales, Cucurbitales, Ericales, Fabales, Gentianales, Lamiales, Malpighiales, Malvales, Poales, Rosales, Sapindales, Solanales, Vitales, Zingiberales. Although the term has been attributed to just one organism from the order Vitales, *Vitis vinifera*, the support for the determination of clade of origin is cumulative; therefore it is possible to suppose that in plant development the origin of the Biological Process “meristem initiation” is restricted to the class Magnoliopsida or Dycotiledons. The term definition in GO is: “Initiation of a region of tissue in a plant that is composed of one or more undifferentiated cells capable of undergoing mitosis and differentiation, thereby effecting growth and development of a plant by giving rise to more meristem or specialized tissue”. Our analysis restricts this process to Dycotiledons.

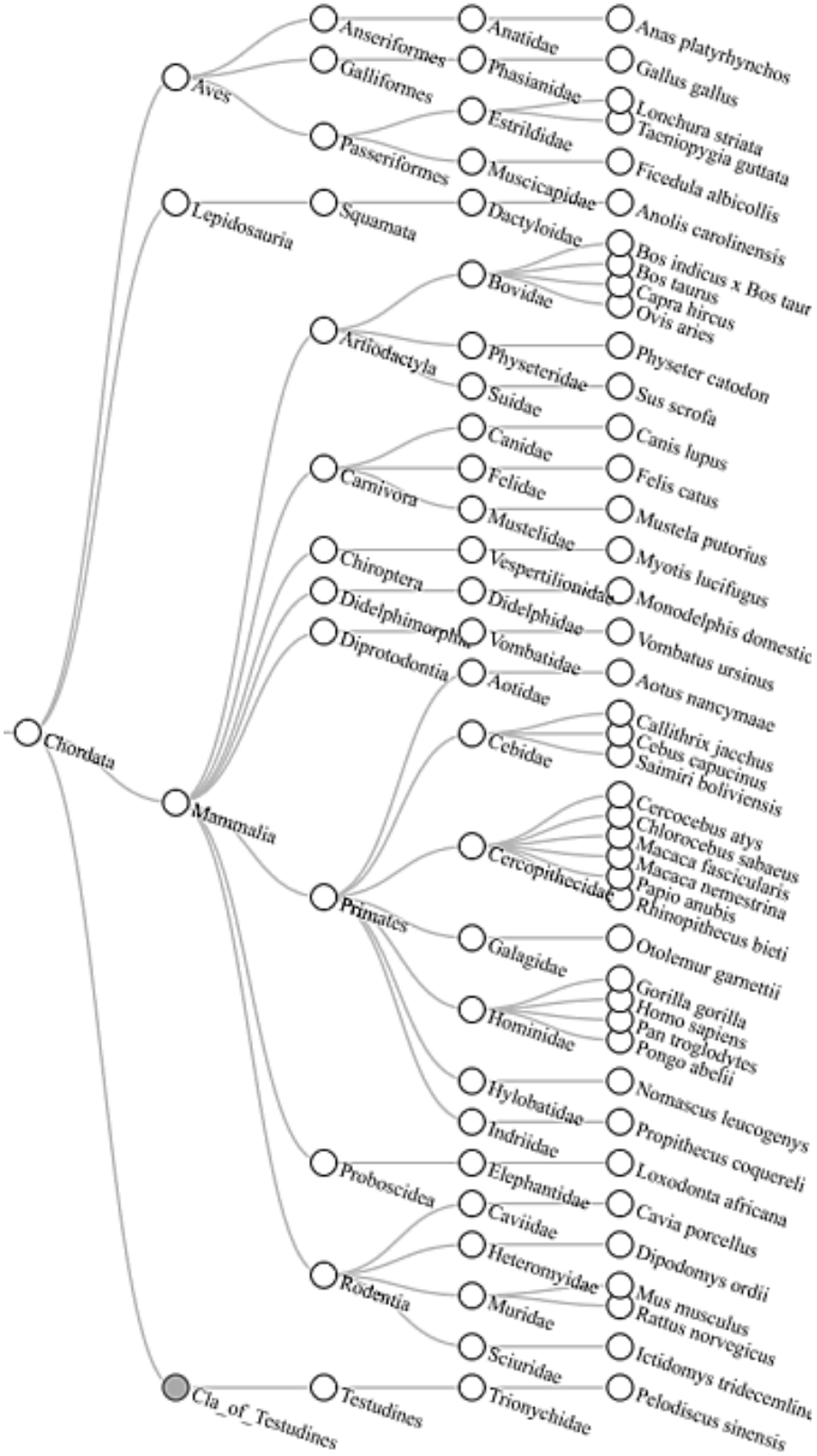

Figure 1. Taxonomic distribution of GO:0031128 developmental induction.

3.4. Bacterial, Fungus and Plant development

"Sporulation" (GO:0043934) is a child of "Developmental process" which maps to the root of the taxonomic tree due to annotation of metagenomics sequences, along with bacterial and fungi proteins. The process is shared between them but electronically annotated to metagenomics, which is considered to come from the root of tree. One child term of "Sporulation" is "Asexual sporulation" that has as child "Conidium formation", annotated to 28 Dicarya organisms, Ascomycota and Basidiomycota. Another child of "Sporulation" is GO:0030435 "Sporulation resulting in formation of a cellular spore", which maps to 7632 organisms, being mostly bacterial phyla such as Firmicutes, Bacteroidetes, Cyanobacteria but also to fungi. Therefore the origin of the process is as ancient as the clade cellular organisms, although there is a lag between bacteria and fungus. One child of "Sporulation" is GO:0034293 "Sexual sporulation", which is attributed to a single organism, and it is not recommended to look at origin in these cases, however it has as child GO:0048236 "Plant-type sporogenesis", mapped to Embryophyta. These observations turn "Sporulation" into a typical example of a very remote biological process.

3.5. Developmental processes of the Amniota

Table 2 shows two biological processes distributed amongst 51 and 42 species, namely GO:0044111 "development involved in symbiotic interaction" and the aforementioned GO:0031128 "developmental induction", respectively. Their taxonomic distribution comprises Mammalia, Aves and the classes of Testudines and Lepidosauria. "Developmental induction" has two children which are also Amniota-originated process: GO:0072034 "Renal vesicle induction" (58 species) and GO:0060235 "Lens induction in camera-type eye" (63 species); however it has another child with broader distribution. That may occur because "developmental induction" groups more information than renal vesicle and lens in eye. It is GO:0001759 "Organ induction" (216 species), with putative origin in the clade Gnathostomata, an ancestor clade reuniting all jawed animals, from sharks to humans. A child of "organ induction" with more than ten species assigned to it is GO:0060492 "lung induction" (Amniota, 65 species). Thus, several processes are probably inexistent before the origin of the Amniota organisms.

3.6. Anatomical structure formation involved in morphogenesis, GO:0048646

GO:0048646 is a child of "developmental process" (Table 2) and its children are shown in Table 3. Proteins from diverse organisms have been annotated with this term, e.g. from the classes: Mammalia, Aves, the classes of Lepidosauria (snakes and lizards) and Testudines (turtles), which together would bring the ancestral clade to Amniota, but with occurrences detected in Amphibia and with participation of fish brought the origin of the process to the clade of Euteleostomi. Therefore, anatomical structure formation involved in morphogenesis may be understood as a biological process that did not exist since the remote origin of life, but rather was originated in the ocean, with the appearance of ancient fish.

However, the parent term might be connected with children terms that originated in more ancient clades. Particularly the terms "syncytium formation" and "cellularization", are distributed along the Eukaryota clade, because they are attributed to plant and animal proteins related to the parent issue of "anatomical structure formation involved in morphogenesis". Moreover, children of GO:0048646 also comprise terms of ancient clades such as Metazoa, Bilateria and Chordata, and a clade in the lineage of plants, Magnoliopsida, as shown in Table 3.

Describing the origin of biological processes in the lineage of man, we notice the term "tube formation" first appearing in the clade Metazoa. Later in evolution, "germinal center formation", "somitogenesis", "retina layer formation", "Spemann organizer formation", "formation of anatomical boundary", "formation of primary germ layer", "central nervous system formation", "eggshell formation and micropyle formation" originated in the clade Bilateria.

Some processes putatively appear only by the epoch of the Chordata, "neural crest formation", "heart valve formation", "floor plate formation" and "anterior neural plate formation".

Table 3. Children of anatomical structure formation involved in morpho-genesis GO:0048646.

GO term	GO term name	Ancestral clade	#Species
GO:0048645	animal organ formation	Amniota	68
GO:0021819	layer formation in cerebral cortex	Amniota	68
GO:0021502	neural fold elevation formation	Amniota	60
GO:0001842	neural fold formation	Amniota	60
GO:0021697	cerebellar cortex formation	Amniota	59
GO:0061198	fungiform papilla formation	Amniota	55
GO:0072033	renal vesicle formation	Amniota	54
GO:0003207	cardiac chamber formation	Amniota	53
GO:0021694	cerebellar Purkinje cell layer formation	Amniota	47
GO:0072003	kidney rudiment formation	Amniota	45
GO:0060661	submandibular salivary gland formation	Amniota	44
GO:0021688	cerebellar molecular layer formation	Amniota	42
GO:1905327	tracheoesophageal septum formation	Amniota	38
GO:0002467	germinal center formation	Bilateria	216
GO:0001756	Somitogenesis	Bilateria	136
GO:0010842	retina layer formation	Bilateria	129
GO:0060061	Spemann organizer formation	Bilateria	124
GO:0048859	formation of anatomical boundary	Bilateria	102
GO:0001704	formation of primary germ layer	Bilateria	62
GO:0021556	central nervous system formation	Bilateria	60
GO:0030703	eggshell formation	Bilateria	23
GO:0046844	micropyle formation	Bilateria	12
GO:0014029	neural crest formation	Chordata	116
GO:0003188	heart valve formation	Chordata	96
GO:0021508	floor plate formation	Chordata	95
GO:0090017	anterior neural plate formation	Chordata	38
GO:0007378	amnioserosa formation	Drosophila	12
GO:0046843	dorsal appendage formation	Drosophila	12
GO:0007293	germarium-derived egg chamber formation	Drosophila	12
GO:0035202	tracheal pit formation in open tracheal system	Drosophila	12
GO:0007370	ventral furrow formation	Drosophila	12
GO:0007376	cephalic furrow formation	Drosophila	10
GO:0007349	Cellularization	Eukaryota	106
GO:0006949	syncytium formation	Eukaryota	99

Table 3 continued..

GO term	GO term name	Ancestral clade	#Species
GO:0060174	limb bud formation	Euteleostomi	116
GO:0014028	notochord formation	Euteleostomi	116
GO:0003272	endocardial cushion formation	Euteleostomi	107
GO:0060214	endocardium formation	Euteleostomi	102
GO:0097187	Dentinogenesis	Euteleostomi	100
GO:0021588	cerebellum formation	Euteleostomi	95
GO:0021905	forebrain-midbrain boundary formation	Euteleostomi	91
GO:0021509	roof plate formation	Euteleostomi	85
GO:0030220	platelet formation	Euteleostomi	67
GO:0090009	primitive streak formation	Euteleostomi	65
GO:0097186	Amelogenesis	Euteleostomi	59
GO:0003408	optic cup formation involved in camera-type eye development	Euteleostomi	52
GO:0032475	otolith formation	Euteleostomi	51
GO:0003315	heart rudiment formation	Euteleostomi	46
GO:0003403	optic vesicle formation	Euteleostomi	40
GO:0021594	rhombomere formation	Euteleostomi	40
GO:0035992	tendon formation	Euteleostomi	30
GO:1905393	plant organ formation	Magnoliopsida	35
GO:0010618	aerenchyma formation	Magnoliopsida	24
GO:0035148	tube formation	Metazoa	137
GO:0120077	angiogenic sprout fusion	Neopterygii	24
GO:0061195	taste bud formation	Osteoglossocephalai	45
GO:0021576	hindbrain formation	Osteoglossocephalai	36
GO:0030435	sporulation resulting in formation of a cellular spore	root	7632
GO:0001525	Angiogenesis	root	1576
GO:0060900	embryonic camera-type eye formation	Tetrapoda	58
GO:0035802	adrenal cortex formation	Tetrapoda	45
GO:0001825	blastocyst formation	Theria	58
GO:0060592	mammary gland formation	Theria	45
GO:0060615	mammary gland bud formation	Theria	36
GO:0001946	Lymphangiogenesis	Vertebrata	115
GO:0021501	prechordal plate formation	Vertebrata	94
GO:0021547	midbrain-hindbrain boundary initiation	Vertebrata	86
GO:0060788	ectodermal placode formation	Vertebrata	51

After some more steps in evolution, in the ancestor of man and lampreys, the Vertebrata clade, the “lymphangiogenesis”, the “prechordal plate formation”, the “midbrain-hindbrain boundary initiation” and the “ectodermal placode formation” originated.

Later, in the Euteleostomi clade several biological processes may have originated: “tendon formation”, “optic vesicle formation”, “rhombomere formation”, “heart rudiment formation”, “otolith formation”, “optic cup formation involved in camera-type eye development”, “amelogenesis”, “primitive streak formation”, “platelet formation”, “roof plate formation”, “forebrain-midbrain boundary formation” and “cerebellum formation”.

With the appearance of amphibia (whose ancestor clade shared with human is Tetrapoda), some processes appeared: “embryonic camera-type eye formation” and “adrenal cortex formation”.

Moreover, many processes might have appeared or have at least been studied in Amniota, our ancestor with chickens and turtles: “tracheoesophageal septum formation”, “cerebellar molecular layer formation”, “submandibular salivary gland formation”, “kidney rudiment formation”, “cerebellar Purkinje cell layer formation”, “cardiac chamber formation”, “renal vesicle formation”, “fungiform papilla formation”, “cerebellar cortex formation”, “neural fold elevation formation”, “neural fold formation”, “animal organ formation” and “layer formation in cerebral cortex”.

And with the appearance of Mammalia, in the Theria clade, which comprises the placentals and marsupials, recent processes arose: “mammary gland bud formation”, “mammary gland formation” and “blastocyst formation”.

3.7. Grandchild of developmental process (GO:0032502), Animal organ morphogenesis, GO:0009887

Another child of “developmental process” is “anatomical structure morphogenesis” GO:0009653, which has a lonely child, “animal organ morphogenesis” GO:0009887 that provides interesting information about the putative origin of biological processes occuring in animal structures: the organs (Table 4).

The remotest clade shared with man is Bilateria, in which all GO terms have been attributed to more than ten species. Bilateria groups our ancestor with *Drosophila* and all other arthropods, as well as several other phyla of animals. The terms mapped to Bilateria are: “oviduct morphogenesis”, “post-embryonic animal organ morphogenesis”, “uterus morphogenesis”, “trachea morphogenesis”, “gonad morphogenesis”, “muscle organ morphogenesis”, “brain morphogenesis” and “heart morphogenesis”. Therefore these organs are understood as the most ancient origin in this set. Some biological processes arise only in the ancestor of Chordata: “swim bladder morphogenesis”, “odontogenesis”, “cartilage morphogenesis”, “bone morphogenesis”. Moreover by the origin of Euteleostomi, our ancestor with fishes, there appears the usage of terms probably related to the origin of the organs: “lung morphogenesis”, “pancreas morphogenesis”, “embryonic organ morphogenesis” and “skin morphogenesis”. Furthermore, the term “gland morphogenesis” originated in Tetrapoda and deserves the analysis of its children, presented in Table 5.

“Salivary gland morphogenesis” is the remotest process in this set, found in Bilateria. “Mammary gland morphogenesis” as expected is the most recent, occurring in Theria, that groups placentals and marsupials. Clupeocephala is a clade in the fish lineage, and ocasionally neurohypophisis has been deeply studied in zebra fish and propagated to closely-related organisms, thus producing this bias.

Furthermore, in Table 3 we see the origin of a broad term named “embryonic organ morphogenesis” (GO:0048562) in the clade Euteleostomi, our ancestor with fish, which also deserves inspection, shown in Table 6. We observe the putative origin of terms related to eye mapping to Bilateria - including “invertebrate eye”, “embryonic skeletal system”, “notochord”, “embryonic digestive tract” and “genitalia”, “otic vesicle” and “ear”. The most recent are restricted to Euteleostomi clade.

Also in Table 3 we notice terms of more recent origin, appearing only in Amniota, our ancestor with chickens, turtles and lizards; these biological processes are related to modern organs: “cerebral cortex”, “neural fold”, “papilla”, “kidney”, “cardiac chamber”, “salivary gland” and “tracheoesophageal septum”.

Table 4. Children of animal organ morphogenesis GO:0009887.

GO term	GO term name	Ancestral clade	#Species
GO:0060434	bronchus morphogenesis	Amniota	57
GO:0035112	genitalia morphogenesis	Amniota	58
GO:0072197	ureter morphogenesis	Amniota	64
GO:0060993	kidney morphogenesis	Amniota	68
GO:0035848	oviduct morphogenesis	Bilateria	10
GO:0048563	post-embryonic animal organ morphogenesis	Bilateria	56
GO:0061038	uterus morphogenesis	Bilateria	73
GO:0060439	trachea morphogenesis	Bilateria	78
GO:0035262	gonad morphogenesis	Bilateria	123
GO:0048644	muscle organ morphogenesis	Bilateria	131
GO:0048854	brain morphogenesis	Bilateria	145
GO:0003007	heart morphogenesis	Bilateria	157
GO:0048795	swim bladder morphogenesis	Chordata	48
GO:0042476	Odontogenesis	Chordata	117
GO:0060536	cartilage morphogenesis	Chordata	118
GO:0060349	bone morphogenesis	Chordata	126
GO:0008407	chaeta morphogenesis	Diptera	13
GO:0048734	proboscis morphogenesis	Drosophila	10
GO:0035211	spermathecum morphogenesis	Ecdysozoa	13
GO:0060425	lung morphogenesis	Euteleostomi	72
GO:0061113	pancreas morphogenesis	Euteleostomi	82
GO:0048562	embryonic organ morphogenesis	Euteleostomi	106
GO:0043589	skin morphogenesis	Euteleostomi	126
GO:0022612	gland morphogenesis	Tetrapoda	70
GO:0048705	skeletal system morphogenesis	Vertebrata	116

The scenario presented here illustrates how developmental biology appears progressively following the appearance of novel clades during evolution.

4. DISCUSSION

A complete inference of clades in which GO biological processes might have originated is a subject of research in progress that aims to build a database which will be updated periodically. It is expected that each update will further complete the annotations for already existent species and also add new species to the database, which will turn the inference of the taxonomic distribution of the terms more accurate.

A previous (unpublished) version of this work used as input a GO annotation file from 2016 which was less than half of the current size; the comparison of results from both versions indicates that the usage of a significant number of species

Table 5. Children of gland morphogenesis, GO:0022612.

GO term	GO term name	LCA	#Species
GO:0061682	seminal vesicle morphogenesis	Amniota	27
GO:0048850	hypophysis morphogenesis	Amniota	44
GO:0007435	salivary gland morphogenesis	Bilateria	101
GO:0048848	neurohypophysis morphogenesis	Clupeocephala	41
GO:0072576	liver morphogenesis	Euteleostomi	100
GO:1905905	pharyngeal gland morphogenesis	Rhabditida	10
GO:0060512	prostate gland morphogenesis	Tetrapoda	64
GO:0060443	mammary gland morphogenesis	Theria	50

Table 6. Children of embryonic organ morphogenesis (GO:0048562).

GO term	GO term name	Ancestral clade	#Species
GO:0048557	embryonic digestive tract morphogenesis	Bilateria	137
GO:0030538	embryonic genitalia morphogenesis	Bilateria	68
GO:0048048	embryonic eye morphogenesis	Bilateria	208
GO:0071600	otic vesicle morphogenesis	Chordata	99
GO:0042471	ear morphogenesis	Eumetazoa	253
GO:0048704	embryonic skeletal system morphogenesis	Euteleostomi	77
GO:0048570	notochord morphogenesis	Euteleostomi	108

results into a stable determination of the origin of processes, and therefore the perspective is that the continuous use of GO in comparative biology will contribute for accuracy and stability in determination of origin. Accordingly, terms that were already largely annotated had shown the same ancestral clade as in this updated analysis.

This work exemplifies with development and developmental processes the first attempt of using a sequence-independent approach to depict the origin of biological features. Our group has developed an approach to build biological pathways and study the epoch of origin of the genes [11], inspired by a previous work of our group that focused on the preimplantation embryo development [12]. In those studies, gene origin was based on orthologue clustering of amino acid sequences, which allowed us to determine their taxonomic distribution and subsequently the ancestral clade from which descend all organisms containing the gene. Here we decided not to rely on sequences, since the existence of a biological process, after our manual analysis of pathways, usually requires the cooperation of both remote and recent sequences, which makes it difficult to point out a clade in which the pathway or the process became functional. The same can be said regarding the origin of complex organs, as they evolve through a series of incremental (and often indirect) steps, which happen over the course of several clades and thus requiring the integration and exaptation of features and structures existing since different epochs within the organism [13]. On the other hand, by relying only on the ontology and its attributions to proteins, it is expected that some processes that have been studied in model organisms will show a certain lag in time until its actual taxonomic distribution

is achieved. For instance, we suspected that many terms associated with Amniota might indeed be placed more remotely after the comparative biology may be extended to fish biology, what would bring the ancestral clade to Euteleostomi. However, an updatable database that is in our objectives might stimulate comparative biologists to query literature on remote vertebrates to see if the process is indeed shared with them. Whatsoever, the existence of largely studied invertebrates grants somehow that an attribution of ancestral clade to Euteleostomi is accurate; otherwise the term should have been annotated to proteins from largely studied organisms, such as *Drosophila*.

Electronic propagation of terms can occasionally push a process or function to a more ancestral clade where it should not occur. However, Gene Ontology has in place a system for taxon constraints on which certain terms can be forbidden to be annotated within or outside specific taxons; although its penetrance on the database of terms is still shallow, there have been attempts [14] to programmatically enhance it by integration with other databases. As the distribution of taxon constraints on the database becomes broader, the determination of the ancestral clade for each process and function will also continue to improve.

5. CONCLUSION

Throughout the evolution of organisms, a series of complex biological phenomena have arisen to diversify life and allow for the establishment and regulation of new structures, forms, mechanisms of obtaining energy and so on. In time, what was once just a simple cellular organism on a primitive ocean, developed through all sorts of different manners into a multitude of forms of life that inhabit and have inhabited our shared planet. Today, when we look upon to these complex systems created by the appearing sequences over the time, we cannot establish when they became fully functional, since they often require the participation of products whose sequences appeared separated, sometimes, by billions of years. Thus, by adopting a sequence-independent strategy, we are able to attempt to answer the question “when did that particular phenomena started to occur?” by looking to the current organisms recognized as sharing that feature. Although Gene Ontology is, and always will be, lacking on information, due to the sheer amount of work needed to fully capture all biological data possible on a single database of terms, our work has shown that it can provide a general sense of time for when has a certain development novelty had arisen during evolution, thus providing some clue for the investigator. Lastly, since it is a growing, massive and continuous effort, GO will only get more expansive and better constructed over time, thus making this sort of analysis more accurate and useful with each new version.

DATABASES

This work was conducted with GO Annotation File version 2.1 generated on 2019-12-17 and NCBI Taxonomy data downloaded on 2020-01-06.

ACKNOWLEDGEMENTS

We would like to thank Dr. Karen R. Christie, Mouse Genome Informatics, from the GO Consortium for criticism and helpful discussions. The authors also thank Dr. Tetsu Sakamoto, from UFRN, for providing Taxallnomy webtool, used to assist the production of Figure 1.

FUNDING

This work was supported by Capes Computational Biology networks BSC and fGEF, and FAPEMIG.

CONFLICT OF INTEREST STATEMENT

The authors declare no conflict of interest regarding this work.

REFERENCES

1. Orgel, L. E. 1998, Trends in Biochemical Sciences, 23(12), 491-495.
2. De Duve, C. 2007, Nature Reviews. Genetics, 8(5), 395-403.
3. Woese, C. R., Kandler, O. and Wheelis, M. L. 1990, Proceedings of the National Academy of Sciences of the United States of America, 87(12), 4576-4579.
4. David, J. R. 2001, An Acad. Bras. Cienc., 73(3), 385-395.
5. Fani, R. and Fondi, M. 2009, Phys. Life Rev., 6(1), 23-52.

6. Kanehisa, M and Goto, S. 2000, Nucleic Acids Research, 28(1), 27-30.
7. Mi, H., Muruganujan, A. and Thomas, P. D. 2013, Nucleic Acids Research, 41 (Database issue), D377-86.
8. Ashburner, M., Ball, C. A., Blake, J. A., Botstein, D., Butler, H., Cherry, J. M., Davis, A. P., Dolinski, K., Dwight, S. S., Eppig, J. T., Harris, M. A., Hill, D. P., Issel-Tarver, L., Kasarskis, A., Lewis, S., Matese, J. C., Richardson, J. E., Ringwald, M., Rubin, G. M. and Sherlock, G. 2000, Nat. Genet., 25(1), 25-29.
9. Aho, A. V., Hopcroft, J. E. and Ullman, J. D. 1973, Proceedings of the fifth annual ACM symposium on Theory of computing - STOC '73. Anais...New York, New York, USA: ACM Press.
10. Federhen, S. 2012, Nucleic Acids Research, 40, Database issue, D136-143.
11. Trindade, D., Orsine, L. A., Barbosa-Silva, A., Donnard, E. R. and Ortega, J. M. 2015, Methods, 74, 16-35.
12. Donnard, E., Barbosa-Silva, A., Guedes, R. L., Fernandes, G. R., Velloso, H., Kohn, M. J., Andrade-Navarro, M. A. and Ortega, J. M. 2011, BMC Genomics, 12, S3.
13. Gregory, T. R. 2008, Evo. Edu. Outreach, 1, 358-389.
14. Tang, H., Mungall, C., Mi, H. and Thomas, P. 2018, arXiv:1802.06004.