UNIVERSITA' DEGLI STUDI DI PADOVA

Facoltà di Scienze MM. FF. NN.

Centro Ricerche Interdipartimentale Biotecnologie Innovative (CRIBI)

European Bioinformatics Institute (EBI, UK)

Scuola di Dottorato di Ricerca in Biochimica e Biotecnologie
Indirizzo in Biotecnologie
Ciclo XX

Contribution to OBO ontologies and application of structured vocabularies for data integration and biological reasoning

Direttore della Scuola
Ch.mo Prof. Lorenzo A. Pinna

Supervisore
Ch.mo Prof. Giorgio Valle

Co-supervisore
Dr. Jennifer Deegan (née Clark)

Dottorando
Erika Feltrin

A.A. 2007/2008

Abstract

As the amount of accessible biological data is growing exponentially, it is becoming harder and harder to extract the biological knowledge contained in thousands of databases. Biomedical scientists collect facts, often recording them in natural language, and then use their knowledge to make inferences about as yet uncharacterised observations. Therefore, to make the best use of biological databases and the knowledge they contain, different kinds of information from different sources must be integrated in ways that make sense to the scientific community. The Gene Ontology (GO) and other biomedical ontologies (OBO) are fundamental components in data integration and annotation.

This PhD project focuses on the improvement of some already existing resources, and the development of new methods that facilitate data integration and extraction, for genes, drugs and diseases, and their inter-relationships. The work consists of contributions to biological ontologies and definitions of cross-links between different semantic fields represented in several distinct databases. Significant changes in GO content and structure have been provided, resulting in the addition of hundreds of terms useful in the representation of muscle and nervous system biology. In addition, a resource has been developed to find preliminary correlations between genes, drugs and diseases. This resource integrates information from several very up-to-date sources, most of which are manually curated; and from a human disease ontology, the 'Disease Ontology'.

The revised ontologies will facilitate the interpretation of high-throughput experiments in the area of muscle biology and neurobiology, and more importantly, in the fields of neuromuscular and nervous system diseases. Furthermore, the developed ontology-based system will provide interoperability support for physicians and medical researchers in the interpretation of data from studies on human diseases.

Abstract

Negli ultimi anni, la quantità di dati a disposizione della comunità scietifica è cresciuta in maniera esponenziale, apportando indubbi benefici. D'altra parte, l'eterogeneità delle informazioni contenute nei database ne ha reso difficile l'estrazione e la successiva interpretazione. Spesso inoltre, i concetti e le definizioni presenti vengono descritte usando il linguaggio comune difficilmente interpretabile da un calcolatore. Per utilizzare nel migliore dei modi sia i database a nostra disposizione sia i dati in essi raccolti, diversi tipi di informazioni provenienti da varie fonti devono essere integrate, in modo da renderle disponibili alla comunità scientifica. La Gene Ontology (GO) e le altre ontologie biomediche (OBO), grazie all'uso di identificativi unici, forniscono le basi per l'interoperabilità tra i database, per l'annotazione e per l'analisi di dati provenienti da esperimenti su larga scala.

Durante questo progetto di dottorato, mi sono occupata di migliorare risorse già disponibili, e di definire e sviluppare nuovi metodi per facilitare l'estrazione e l'integrazione di dati riguardo a geni, farmaci, malattie, e le loro inter-relazioni. Questo lavoro ha apportato numerosi contributi a diverse ontologie biologiche. In particolare, sono stati forniti notevoli cambiamenti al contenuto e alla struttura della Gene Ontology, aggiungendo e ridefinendo centinaia di termini utili per esplicitare e concettualizzare la conoscenza relativa alle cellule nervose e muscolari in diverse condizioni e patologie. Successivamente, è stata sviluppata una risorsa per identificare correlazioni interessanti tra geni, farmaci e malattie, integrando informazioni provenienti da diversi database e associadole ai termini di un'ontologia di malattie. Questo strumento sarà utile nell'interpretazione di dati provenienti da esperimenti high-throughput sul tessuto muscolare o nervoso, e soprattutto, per lo studio di malattie neuro-muscolari e del sistema nervoso. Inoltre, questo sistema, basato sull'interoperabilità tra diverse risorse integrate, fornirà a medici e ricercatori un supporto nell'interpretazione dei dati provenienti da studi su malattie umane.

Acknowledgements

I would like to thank first Prof. Giorgio Valle and Dr. Fabrizio Caldara for giving me the opportunity to do a PhD studentship. I would also like to thank Dr. Matej Lexa and Dr. Elisabeth Ehler for taking the time to evaluate my thesis.

An enourmous thank to Jennifer Deegan who provided exquisite support and direction throughout the research process, who answered all my questions, who suggested interesting points, who corrected the manuscript, and finally who gave me practical help with language and technology.

I would like to acknowledge the help and support of the following people from EBI: Jane lomax, Emily Dimmer, Midori Harris, Evelyn Camon, Amelia Ireland and Rachael Huntley who accepted to have me as a visiting student in their office for six months, who made me feel part of a BIG project, and in whom I found not only colleagues but also friends.

I would also like to thank my colleagues who provided great insight and encouragement throughout my work. A particular thanks to Dr. Alessandro Albiero who gave me a huge amount of help in the developement of our database.

I would like to thank my husband, Federico for his support and his extremely high-level of patience throughout this process. Thanks to my parents and to my grand mother, who encouraged me during these years.

Contents

1 Introduction 1

1.1 Ontologies 1

1.1.1 Definition of ontology 2

1.1.2 Ontologies in biology 2

1.2 Open Biomedical Ontologies 6

1.2.1 OBO Foundry 7

1.3 The Gene Ontology project 9

1.3.1 GO structure 10

1.3.2 Updating the ontologies 11

1.3.3 GO software resources 13

1.4 Gene Ontology Annotation project 14

1.4.1 Electronic annotation 15

1.4.2 Manual annotation 16

1.5 Aims 17

2 Improvement of the neurobiology representation in Gene Ontology 19

2.1 Background 19

2.2 Representation of nervous system biology in GO 21

2.2.1 Method and results 21

2.2.2 A case: depression and mechanisms of regulation 24

2.2.3 Improved GO categories 25

2.2.4 The Neurobiology Content Meeting 28

2.3 Functional annotation of nervous system specific genes 29

2.3.1 List of annotated genes 30

2.3.2 Manual annotation method 30

2.3.3 An example: the annotation of proteins mGLUR2 and mGLUR3 32

2.3.4 An example: the annotation of proteins GRIK2, GRIK3 and GRIA1 . . 34

3 Muscle GO Content Meeting 37

3.1 Introduction 37

3.1.1 GO Content Meeting 38

3.2 Method 39

3.3 Results and discussion 41

3.3.1 Muscle contraction 42

3.3.2 Muscle plasticity 44

3.3.3 Sarcomere and its role in regulating the calcium ion dependent processes 47

3.3.4 Muscle development and regeneration 48

3.3.5 Muscle Biology Community Annotation 49

4 Representation of 'drugs' in GO 51

4.1 Introduction to the problem 51

4.2 Method 54

4.2.1 Set up of the 'response to drug' interest group 54

4.3 Evaluation of the available chemistry resources 54

4.4 ChEBI database and ontology 55

4.4.1 ChEBI structure 55

4.5 Implementation project 57

4.6 Results 58

4.6.1 Case of cycloheximide (GO:0046898; CHEBI:17076) 60

4.6.2 Case of morphine (GO:0043278; CHEBI:17303) 61

4.6.3 Case of methotrexate (GO:31427; CHEBI:null) 62

4.6.4 Case of fluoxetine (GO:null; CHEBI:44185) 63

4.6.5 Resulting 'response to drug' structure 64

5 Disease Ontology Annotation project 69

5.1 Introduction to the project . . . 69

5.2 Resources . . . 70

5.2.1 Disease Ontology . . . 70

5.2.2 Genetic Association Database . . . 73

5.2.3 OMIM . . . 74

5.2.4 DrugBank . . . 74

5.2.5 PharmGKB . . . 75

5.3 Method and Results . . . 76

5.3.1 Acquisition and integration of data . . . 76

5.3.2 Acquiring synonym terms for disease names . . . 78

5.3.3 Finding relationships between genes and diseases . . . 81

5.3.4 Compilation of the drug dictionary . . . 81

5.3.5 Characterising relations between drugs and diseases . . . 83

5.3.6 Finding relationships between drugs and target genes . . . 85

5.4 Possible applications . . . 85

5.5 Future perspective . . . 87

6 Tools 89

6.1 Ontology editor . . . 89

6.1.1 OBO-Edit, a tool for ontology development . . . 89

6.2 Ontology browsers . . . 92

6.2.1 Ontology Lookup Service . . . 92

6.2.2 AmiGO: GO Consortium browser . . . 93

6.2.3 QuickGO@EBI . . . 94

6.3 Further resources . . . 95

6.3.1 Information Hyperlinked Over Proteins . . . 97

6.3.2 GOPubMed . . . 97

7 Conclusion 99

A *Nervous system* GO new terms 103

B *Nervous system* gene annotations 113

C *Muscle* GO new terms 119

D List of abbreviations 127

Bibliography 139

Chapter 1

Introduction

This document is organised into 7 chapters. Following this introduction about ontologies and their role in molecular biology (with particular attention to Gene Ontology), Chapter 2 describes our contribution to GO content and structure in the area of neurobiology. Chapter 3 gives a detailed description of the GO Muscle Content Meeting that has been organised for the representation of muscle biology in Gene Ontology vocabularies. Chapter 4 is about the results obtained by redefinition of the 'response to drug' GO node. Chapter 5 offers an overview of the Disease Ontology Annotation method adopted in this project with suggestions for further developments. Chapter 6 lists the specific tools used in this work and finally, the conclusion of the project is given in Chapter 7.

1.1 Ontologies

The exponential growth of experimental data, owing to rapid biotechnological advances and to high-throughput technologies, as well as the advent of the World Wide Web as a new means for data exchange, made it more complicated and difficult to find the biological meaning hidden in the heterogeneous biological data available to the scientific community. Furthermore the huge amounts of information, that are now produced on a daily basis, require more sophisticated management solutions; and the availability of the Internet as a modern infrastructure for scientific exchange has created new demands with respect to data accessibility [1]. At the same time, in the era of genome-scale biology, the accumulation of biological data is accompanied by the widespread proliferation of biology-oriented databases [2].

Therefore, to make the best use of such databases and the knowledge they contain, different kind of information from different sources must be integrated in ways that make sense to biologists. With this respect, the integration of data from the existing databases has long been recognized as a fundamental component in the life science studies and several technologies and approaches to data integration have been pursued over the past decade [1]. A major component of the integration effort is the development and use of annotation standards such as ontologies.

1.1.1 Definition of ontology

The word 'ontology' comes from the Greek *ontos* (being) and *logos* (word) and its conceptual origin can be traced back to early philosophers which have been studying the theory of objects and their ties for centuries. In philosophy ontology is used to name the discipline that tries to describe reality. But the term 'ontology' is still controversial because different people have different ideas on the definition of an ontology. The first formal and explicit approach to ontologies in the technical (not philosophical) sense dates back to 1900, given by Husserl. Later in the 1980's, the ontologies entered the computer science field as a way to provide a simplified and well defined view of a specific area of interest or domain. There is a certain consensus in what an ontology is not: it is not a taxonomy (is not just a class-subclass hierarchy), a dictionary (ontology includes relationships between terms), nor a knowledge base that includes individual objects. According to Gruber, an ontology, is 'the specification of conceptualizations, used to help programs and humans share knowledge' [3].

Nowadays ontologies are more formalized conceptual models utilized in computer science, database integration, and artificial intelligence and they make available a common terminology, over a domain, necessary for communication between people and organizations. They provide the basis for interoperability between systems. They can be used for making the content in information sources explicit and serve as an index to a repository of information [4].

1.1.2 Ontologies in biology

Several decades ago, the main aim of the bioinformatics was to store, retrieve and analyse the data produced by biologists; such as nucleotide sequences and protein structures. At that time, the limited amount of data produced by biological researchers, required simple systems for their management, organisation and analysis.

However, the advent of the genome sequencing projects, high-throughput experiments, and other techniques produced a huge amount of data that needed to be analysed. Nowadays, bioinformatics systems have to deal with large amounts of complex information, unmanageable for a scientist without sophisticated knowledge of management and information processing tools [5]. Such data are growing at an exponential rate but the knowledge contained in them is not growing at the same pace. There are different reasons for this lack of productive knowledge and the most important one is that biological phenomena can be described in many different ways [6] and this complexity has not been tackled at a semantical level. That means that usually the biologists are left with a giant domain of information that they cannot access, analyse, or integrate in a sensible way [7]. The impossibility of drawing on information from the data available, adds additional pressure to implement standardised and compatible nomenclature in molecular biology.

The fundamental problem is that biomedical scientists collect facts, often recording them in natural language, and then use that knowledge to make inferences about yet uncharacterised observations. Because of this, the knowledge is highly heterogeneous. While it is easy to compare, for instance, nucleic acid or polypeptide sequences between bioinformatics resources, the knowledge component of these resources is very difficult to compare, both for humans and computers, because the knowledge is represented in a wide variety of lexical forms [8].

Often in biology, a word refers to two different concepts: for example, the concept of 'gametogenesis' means different processes in mammals or in plants and a user, querying a database for this concept, needs to deal with these terminological and conceptual incompatibilities. This situation makes it more complicated for a computer to process information because it would not able to reason over the data and capture the knowledge content. Thus, there is urgent need to find a strategy for the representation of biological knowledge in a formal way [9]. One way to do this is to represent the knowledge as ontologies: the resulting 'bio-ontologies', a relatively new area of bioinformatics [10].

An ontology is a 'controlled vocabulary' that provides a way to capture and represent the knowledge of a domain in a computer-comprehensive way. An ontology describes objects and the relations between them in a formal way, and has a grammar for using the vocabulary terms to express something meaningful within a specific domain of interest [11].

The labels used for the objects and the relationships in an ontological model can provide a language for a community to talk about the domain being modeled. By agreeing on a particular ontological representation, a common vocabulary can be used to describe and ultimately analyse data. Such sharing has obvious benefits because it helps humans to make inferences about a studied domain. The data, that are clues for enriching the knowledge about the domain, become much easier to handle as the same things are referred to in the same manner across the resources in which those data are stored. If different biological databases use the same ontologies to describe their data objects, the bio-ontologies can be used to link the databases and retrieve information from them. Ultimately, since ontologies give a well-defined semantics for the knowledge representation language, machine can make inferences about the facts expressed in that language [8].

Ontologies are designed for the domain and application that they are intended to support, however, it is forth pointing out that, for any ontology to be valuable, it has to be defined following specific rules and assertions. There are several fundamental criteria that an ontology must have to be considered complete and ready to be widely used [12]:

- complete: ontologies are designed to capture the maximum quantity of relevant concepts for the domain they represent;
- formality: ontologies are built using mathematical formalisms, making them readable by computer machine;
- understandable by humans: ontologies are built using natural language terms, making them accessible for scientists;
- general: ontologies aim to represent conceptual domains independently of any specific use or implementation [13].

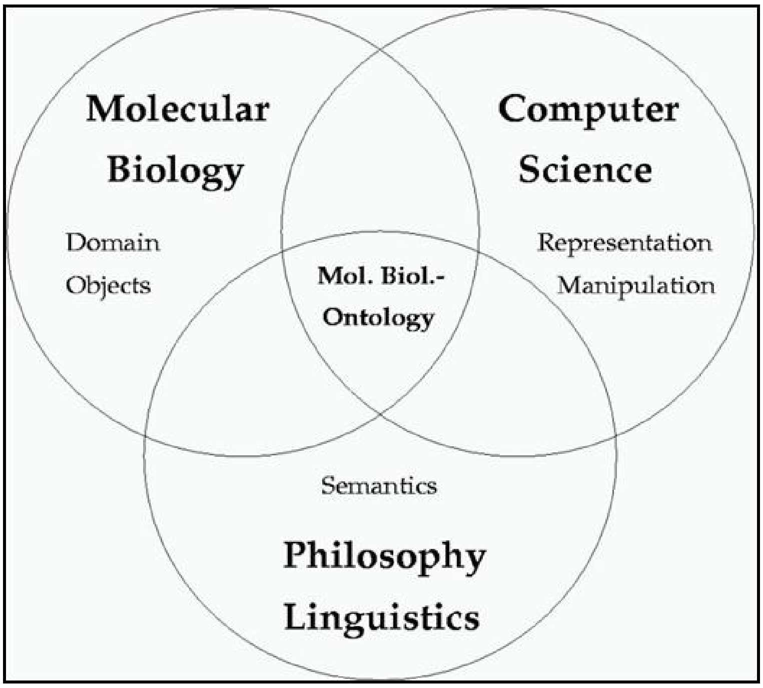


Figure 1.1: Interplay between ontologies, biology, computer science and philosophy. Molecular biologists discover facts that need to be organised and stored in databases. Computer scientists provide techniques for data representation and manipulation. Philosophers and linguists help in organising the meaning behind database labels [14].

Therefore, the development of an ontology requires the help of a computer scientist, providing techniques for data representation and manipulation, and from a philosopher and linguist, organising the meaning behind database labels. The interplay between ontologies, biology, computer science and philosophy is depicted in Figure 1.1.

Finally, it is worth pointing out that. An ontology, aiming to be public valuable, has to be widely accepted by the specific domain that it aims to summarize. The scientific community have to be deeply involved in the development of the ontology and have to ensure that only single ontologies for each area are placed in the public domain. The number and diversity of ontologies will grow in the following years as they have been demonstrated to be a useful tool not only for resource integration but also as tools of knowledge generation or prediction [15]. For example, it has been demonstrated that functional annotation of new sequences based on sequence similarity is not optimal [16] and semantic methods of functional annotation based on ontologies represent an improvement [17].

1.2 Open Biomedical Ontologies

As stated above, an ontology has to be widely disseminated and accepted among users of the field that it aims to summarize. In this respect, a strong community involvement is crucial to ensure that each specific domain is represented by only one ontology. This result is reached by the Open Biomedical Ontologies standards.

The Open Biomedical Ontologies (OBO)[1] is a collection of controlled vocabularies developed in 2001 for the ontological representation of several biological domains. The aim of this initiative, focused on object-level questions, is to represent in an exhaustive way the proteins, organisms, diseases or drug interactions that are of primary interest in biomedical research [18]. The main role of the OBO umbrella is to be an ontology resource. It is supported by the NIH Roadmap National Center for Biomedical Ontology (NCBO) through its BioPortal and it is continually kept up-to-date by ontology-based developers. There are currently over 60 live-science ontologies lodge in OBO, covering domains such as anatomy, development and phenotype, genomic and proteomic information and taxonomic information. All of them use a range of different attributes to describe the respective biological domain. There are many resources under the OBO umbrella, and most of these are shown in Figure 1.2, in which OBO have been roughly arranged along a spectrum of genotype to phenotype.

To be included in OBO, an ontology has to be developed following a set of principles that are used to give coherence to wider ontological efforts across the community:

- openness: ontologies must be available to all, without any constraint or license on their use and it is only asked that users acknowledge the original source. This encourages usage and community buy-in and effort;

- common representation: this is either the OBO format[2] or the Web Ontology Language (OWL)[3]. This provides common access via open tools and offers common semantics for knowledge representation;

[1] http://obofoundry.org

[2] http://www.geneontology.org/GO.format.shtml#oboflat

[3] http://www.w3.org/TR/owl-features/

- independence: lack of redundancy across separate ontologies encourages combinatorial re-use of ontologies and the interlinking of ontologies via relationships;
- identifiers: each term should have a semantic-free identifier, the first part of which refers to the originating ontology. This promotes easy management;
- natural language definitions: terms themselves are often ambiguous, even in the context of their ontology, and definition helps ensure appropriate interpretation. Thus, the terms in each ontology must have a proper textual definition explaining clearly the exact meaning of the concept within the context of a particular ontology.

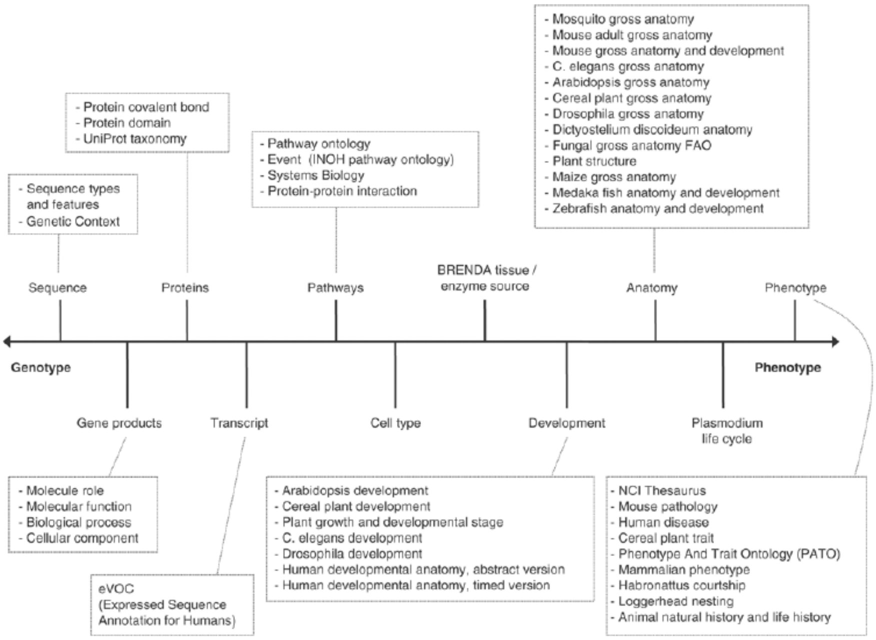

- Sequence types and features
- Genetic Context
- Protein covalent bond
- Protein domain
- UniProt taxonomy
- Pathway ontology
- Event (INOH pathway ontology)
- Systems Biology
- Protein-protein interaction
- Mosquito gross anatomy
- Mouse adult gross anatomy
- Mouse gross anatomy and development
- C. elegans gross anatomy
- Arabidopsis gross anatomy
- Cereal plant gross anatomy
- Drosophila gross anatomy
- Dictyostelium discoideum anatomy
- Fungal gross anatomy FAO
- Plant structure
- Maize gross anatomy
- Medaka fish anatomy and development
- Zebrafish anatomy and development
Sequence
Proteins
Pathways
BRENDA tissue / enzyme source
Anatomy
Phenotype
Genotype
Phenotype
Gene products
Transcript
Cell type
Development
Plasmodium life cycle
- Molecule role
- Molecular function
- Biological process
- Cellular component
eVOC (Expressed Sequence Annotation for Humans)
- Arabidopsis development
- Cereal plant development
- Plant growth and developmental stage
- C. elegans development
- Drosophila development
- Human developmental anatomy, abstract version
- Human developmental anatomy, timed version
- NCI Thesaurus
- Mouse pathology
- Human disease
- Cereal plant trait
- Phenotype And Trait Ontology (PATO)
- Mammalian phenotype
- Habronattus courtship
- Loggerhead nesting
- Animal natural history and life history

Figure 1.2: The OBO ontologies arranged on a spectrum of genotype to phenotype, according to their main domain [8].

1.2.1 OBO Foundry

The principles described above are necessary to ensure that the OBO ontologies remain a resource for the entire community. At the same time, the developers of a small set of OBO ontologies have initiated the OBO Foundry. The participants have established a set of principles in addition to the existing and well-defined original OBO rules. These further principles require that ontologies i) are a result of a collaboration among the other OBO members, ii) use a set of relationships defined in the OBO Relation Ontology (RO) [19], iii) provide procedure for identifying successive versions and iv) represent a clearly specified and delineated content to ensure additivity of annotations and to bring the benefits of modular development. Members can propose new principles using the OBO wiki page[4].

The OBO Foundry is also a valuable contribution to many other important projects carried out by non-Foundry communities and addresses several areas. For example to support the effort of ontology alignment essential for researches on different model organisms, the Foundry has created a Common Anatomy Reference Ontology (CARO) which provides guidelines for model organism communities. In addition, to support the integration of experimental data analysis, an Ontology for Biomedical Investigations (OBI) has been developed as a representation of designs, protocols, instrumentations, materials and processes in all areas of biological and biomedical investigation.

Figure 1.3 summarizes the scope and the state of each OBO Foundry ontology (as of April 2007). There are few mature ontologies like Gene Ontology and Cell type ontology that are continually improved and reorganized. Instead most of them either need to be reviewed like Disease Ontology and ChEBI or are quite new and need to be completed in content.

The OBO Foundry is an open community and, by joining the initiative, the authors of an ontology commit its maintenance, and ensure the improvement of the existing principles over time. Any research groups and individuals that are involved in biomedicine projects, might join the one or more Foundry mailing lists participating with suggestions and even comments on the discussions. Any developers of new ontologies are very welcome, they must interact with the existing members of the Foundry and ensure that the fundamental principles are respected.

The long-term goal is that the Foundry offers a resource, where data, that are produced by biomedical researches and available to the scientific community, are collected in a consistent and algorithmically traceable way. In this way it will be possible to solve some problems associated, for instance, with the differences between technical and biological language. OBO-ontologies have a role in supporting two approaches essential for the interpretation of genome-scale datasets: data integration, and comparative genomics. To date, OBO ontologies such

[4] http://obofoundry.org/wiki/index.php/OBO_Foundry_Principles

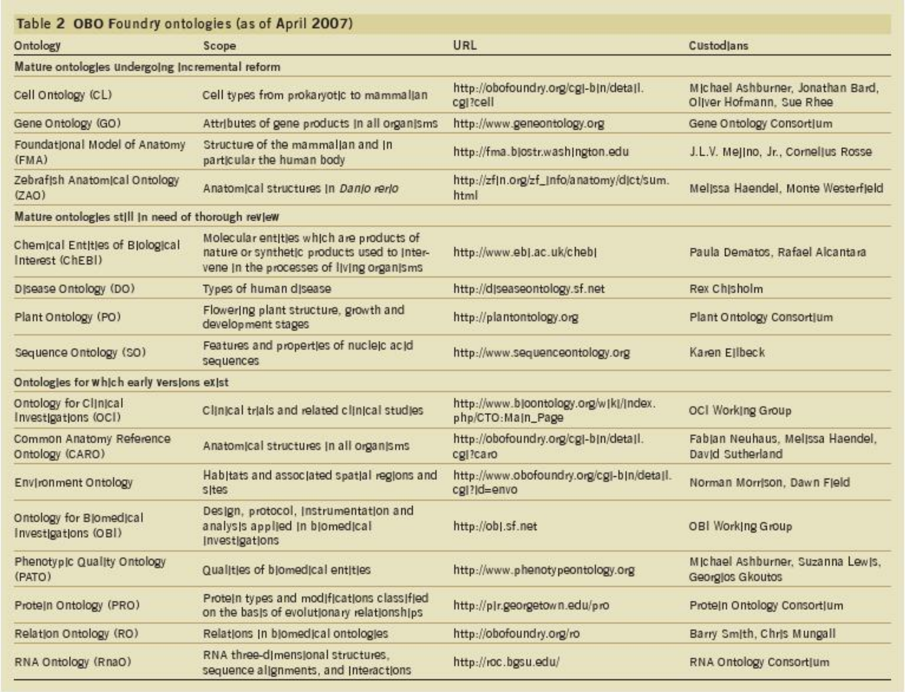


Table 2 OBO Foundry ontologies (as of April 2007)

Ontology	Scope	URL	Custodians
Mature ontologies undergoing incremental reform			
Cell Ontology (CL)	Cell types from prokaryotic to mammalian	http://obofoundry.org/cgi-bin/detail.cgi?cell	Michael Ashburner, Jonathan Bard, Oliver Hofmann, Sue Rhee
Gene Ontology (GO)	Attributes of gene products in all organisms	http://www.geneontology.org	Gene Ontology Consortium
Foundational Model of Anatomy (FMA)	Structure of the mammalian and in particular the human body	http://fma.biostr.washington.edu	J.L.V. Mejino, Jr., Cornelius Rosse
Zebrafish Anatomical Ontology (ZAO)	Anatomical structures in *Danio rerio*	http://zfin.org/zf_info/anatomy/dict/sum.html	Melissa Haendel, Monte Westerfield
Mature ontologies still in need of thorough review			
Chemical Entities of Biological Interest (ChEBI)	Molecular entities which are products of nature or synthetic products used to intervene in the processes of living organisms	http://www.ebi.ac.uk/chebi	Paula Dematos, Rafael Alcantara
Disease Ontology (DO)	Types of human disease	http://diseaseontology.sf.net	Rex Chisholm
Plant Ontology (PO)	Flowering plant structure, growth and development stages	http://plantontology.org	Plant Ontology Consortium
Sequence Ontology (SO)	Features and properties of nucleic acid sequences	http://www.sequenceontology.org	Karen Eilbeck
Ontologies for which early versions exist			
Ontology for Clinical Investigations (OCI)	Clinical trials and related clinical studies	http://www.bioontology.org/wiki/index.php/CTO:Main_Page	OCI Working Group
Common Anatomy Reference Ontology (CARO)	Anatomical structures in all organisms	http://obofoundry.org/cgi-bin/detail.cgi?caro	Fabian Neuhaus, Melissa Haendel, David Sutherland
Environment Ontology	Habitats and associated spatial regions and sites	http://www.obofoundry.org/cgi-bin/detail.cgi?id=envo	Norman Morrison, Dawn Field
Ontology for Biomedical Investigations (OBI)	Design, protocol, instrumentation and analysis applied in biomedical investigations	http://obi.sf.net	OBI Working Group
Phenotypic Quality Ontology (PATO)	Qualities of biomedical entities	http://www.phenotypeontology.org	Michael Ashburner, Suzanna Lewis, Georgios Gkoutos
Protein Ontology (PRO)	Protein types and modifications classified on the basis of evolutionary relationships	http://pir.georgetown.edu/pro	Protein Ontology Consortium
Relation Ontology (RO)	Relations in biomedical ontologies	http://obofoundry.org/ro	Barry Smith, Chris Mungall
RNA Ontology (RnaO)	RNA three-dimensional structures, sequence alignments, and interactions	http://roc.bgsu.edu/	RNA Ontology Consortium

Figure 1.3: Status of the OBO Foundry Ontologies as of April 2007 [18].

as Gene Ontology, have been used primarily in the community genome databases as structure controlled terminology and as data aggregators (for an example see [20]).

1.3 The Gene Ontology project

The Gene Ontology[5] (GO) project began in 1998 as a collaborative effort between three model organism databases: FlyBase (Drosophila), the Saccharomyces Genome Database (SGD) and the Mouse Genome Informatics (MGI) project [21]. Since then, many databases have joined the GO Consortium including several of the world's major repositories for plant, animal and microbial genomes. Nowadays, the GO is the most successful OBO ontology and it is used in several studies including expression profile analysis and proteomic studies to extract additional knowledge from the huge amount of data available.

The GO project starts from the consideration that a large fraction of the genes, derived by genomic sequencing and specifying the core of biological functions, are shared by all organisms.

[5] http://www.geneontology.org

At the moment, many robust methods are at hand for automated transferring of biological annotations from the experimentally tractable model organisms to the less tractable organisms based on gene and protein sequence similarity. The knowledge accumulated from one organism can be often transferred to other organism; but there are a wide range of hurdles to overcome. First, the current system of nomenclature for genes and their products is not followed correctly. Even when an underlying similarity between two genes can be appreciated, the experts are not very confident in using the right nomenclature. Secondly, the lack of the interoperability between genomic databases limits the use of the content of these databases. The Gene Ontology project was formed to help in the solution of these major barriers.

The GO project has three main goals: i) to develop and maintain a set of controlled and structure vocabularies, or ontologies [22, 23], for the description of genes and gene products; ii) to use these vocabularies to annotate genes and gene products in biological database from as many species as possible, iii) to provide a public resource allowing access to ontologies, to gene annotation files and to specific tools developed to utilize all GO data [24].

1.3.1 GO structure

The Gene Ontology Project provides three orthogonal vocabularies used for the description of genes and gene products: cellular component, biological process and molecular function ontologies. A gene product might be associated with or located in one or more cellular components; it is active in one or more biological processes, during which it performs one or more molecular functions. For example, the gene product cytochrome c can be described by the molecular function term oxidoreductase activity, the biological process terms oxidative phosphorylation and induction of cell death, and the cellular component terms mitochondrial matrix and mitochondrial inner membrane [9].

The building blocks of the Gene Ontology are the terms: a GO term represents a single biological concept and if one concept is known by many different names, the alternative names are added as synonyms of the same GO term. Each GO entity has a unique numerical identifier, and it is described by a textual definition, with references indicating the source of the definition [25].

GO is not a simple hierarchical tree but is implemented as a Directed Acyclic Graph (DAG) allowing multiple parent terms for each child term. The top of the DAG is populated by general terms (e.g. cell proliferation or binding activity) and moving down on the path, the terms become more specialized (e.g. skeletal muscle cell proliferation). The terms on the end of the path are called leaves and the terms in the path are usually called nodes. GO terms are linked by two relationships: *is_a* and *part_of* [26]. The *is_a* type refers to the situation when a child term is a type of the parent term (e.g. cellular component vocabulary, a mitochondrial membrane is a type of a membrane). The *part_of* type refers to when a child is a component of the parent (e.g. mitochondrial membrane is a component of a mitochondrion).

The implementation of three independent vocabularies structured as a DAG, with two kind of relationships, gives to GO great expressive capabilities. Compared to tree and logic language, Gene Ontology structure allows a better representation of the complex biological reality. A particular protein can be associated with more than one node within the three ontologies, reflecting the fact that it may function in several processes, contains domains that carry out different molecular functions, and participates in multiple alternative interactions with other proteins, organelles or locations in the cell (Figure 1.4).

It is very important to clarify that the development of the ontologies and the association of ontology terms with gene products (see Gene Annotation Section 1.4) are two independent operations that are carried on in parallel [27].

1.3.2 Updating the ontologies

The growth of GO has been spectacular in recent years because of its openness, community involvement, intuitive structure, and for other many reasons [28] and its success is best illustrated by how much GO is used[6].

The ontologies are dynamic, in the sense that they are progressively changed to reflect the current state of biological knowledge about genes and gene products, and they are continually updated to meet the needs of the user community. The improvement process involves both the structure and the content of the ontology as a whole. The GO editors are responsible for all changes to GO vocabularies; they usually collaborate directly with scientific curators in order to define the area to be revised [29].

[6] http://www.geneontology.org/cgi-bin/biblio.cgi

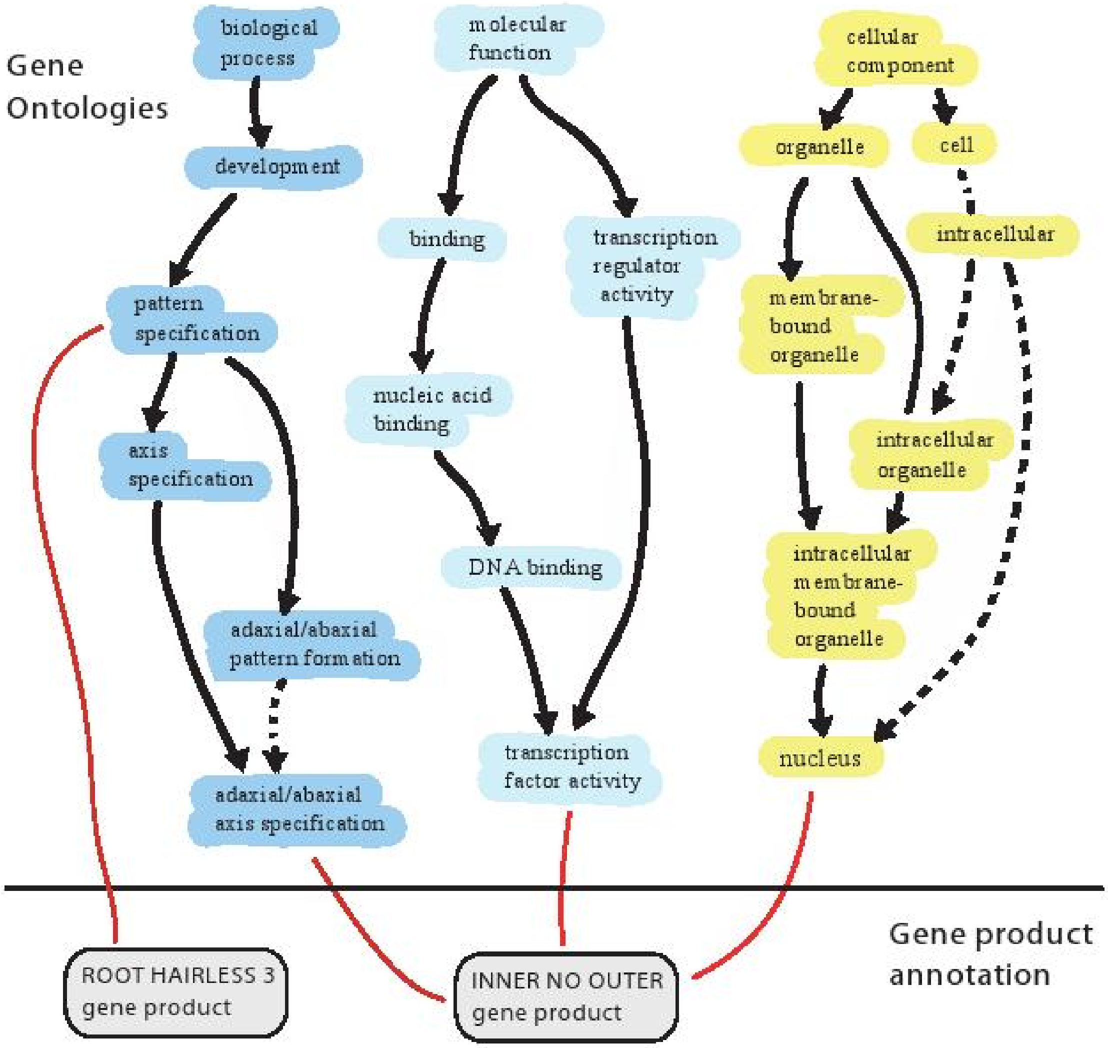


Figure 1.4: the three ontologies. The terms are linked by two relationships: *is_a* relationships (continuous black lines) and the *part_of* relationships (dashed black lines). A term may have multiple parent terms, as in the case of transcription factor activity. Separately and in parallel with the development of the GO, gene products are annotated (red lines) to the terms [9].

To add a new term in the GO, it is necessary to follow few editing rules decided within the GO Consortium: i) all paths must be true (True Path Rule[7]); ii) terms should not be species specific; iii) all terms must have a definition. Every GO term must obey the True Path Rule: if the child term describes the gene product, then all its parent terms must also apply to that gene product. Additions and changes to the ontologies are not exclusively proposed by GO consortium members but also by the broader community. If a curator or a user is interested to add a new term in the GO, he can use the curator request tracker on SourceForge[8], where it is possible to insert the request so that it will be considered by a GO curator. Suggesting new terms on the tracker system is the most immediate way to request for changes; but there are diverse approaches that curators can follow to add and improve Gene Ontology [9].

If a user is a scientific expert in a specific topic and he wants to collaborate on GO project, he might offer his expertise, and with the help of a GO curator, he can develop the Gene Ontology in that particular area. As an alternative, he might join the GO interest group that deals with the specific biological topics. The members of each group participate in discussions of areas that are likely to require extensive additions or revisions, or where proposed changes crop up frequently. If the topic of discussion is very large, there is the opportunity to organize a content-oriented meeting which brings together GO curators and community experts (for an example see [30]). This face-to-face meeting facilitates large-scale changes in specific areas of the ontology (for our Muscle Content Meeting see Chapter 3).

1.3.3 GO software resources

Vocabularies, annotations, databases and tools, and other GO web resources are freely available at the GO web page and can be used without payment, in accordance with its redistribution and citation policy [2].

The GO vocabularies are saved in several file formats: the older GO flat file format that is now deprecated, the newer OBO flat file format, the OBO-XML format, and others like RDF-XML and OWL. The OBO flat file format is the master and the most frequently updated format at the time of writing. In addition, the ontologies and the annotations have been implemented in a MySQL relational database that can be installed locally (Perl APIs are provided).

[7] http://www.geneontology.org/GO.usage.shtml#truePathRule

[8] https://sourceforge.net/tracker/?group_id=36855&atid=440764

To manipulate vocabularies and annotations, the GO Consortium makes available to the GO users two software tools: a Java-based editor named OBO-edit [31] and a Web-based browser named AmiGO[9]. The former is a stand alone program written in Java that supports the editing and developing of ontologies in GO or OBO format. An OBO-Edit working group has been created in order to test bugs and to help in the release of a stable version. The latter provides a web interface for searching and displaying the ontologies, term definitions and gene annotations made by all the databases contributing to GO annotation project. The software described above is developed within the consortium and frequently improved and expanded with advanced features. In addition many useful tools have been created outside of the consortium for use with the GO ontologies (see the GO Tool page[10]). These tools can be use for simply searching and browsing GO, for gene annotation, for gene expression and microarray analysis, and for other kind of applications.

1.4 Gene Ontology Annotation project

Annotation data is primarily produced in species-specific databases resources, such as the Mouse Genome Informatics resources and FlyBase, and in multispecies resources such as Uniprot. The complete list of contributing database groups and the total numbers of annotations are listed on the GO web page[11].

Among such contributors, there is also the GOA group located at the European Bioinformatics Institute (EBI)[12]. The Gene Ontology Annotation (GOA)[13] project aims to provide high-quality GO annotations to proteins in the UniProtKnowledgebase (UniProtKB)[14] and International Protein Index (IPI)[15] and is a central dataset for other major multi-species databases such as Ensembl[16] and NCBI[17] [32].

GOA has been a member of the GO Consortium since 2001, and is responsible for the

[9] http://amigo.geneontology.org

[10] http://www.geneontology.org/GO.tools.shtml

[11] http://www.geneontology.org/GO.current.annotations.shtml

[12] http://ebi.ac.uk/

[13] http://www.ebi.ac.uk/GOA/

[14] http://www.ebi.ac.uk/uniprot/index.html

[15] http://www.ebi.ac.uk/IPI

[16] http://www.ensembl.org/

[17] http://www.ncbi.nlm.nih.gov/

integration and release of GO annotations to the human, chicken and cow proteomes. GOA is also committed to the comprehensive annotation of a set of disease-related proteins in human.

By annotating all characterized proteins with GO terms and facilitating the transfer of this knowledge to similar uncharacterized proteins, the Uniprot group will make a valuable contribution to biological and biotechnological research through a better understanding of all proteomes. High-quality GO annotations are generated through a combination of electronic and manual techniques, the latter of which employ of skilled biologists.

1.4.1 Electronic annotation

The electronic annotation approach is used to provide large-scale assignment of GO terms to SWISS-PROT and TrEMBL entries. This strategy is based on the use of existing properties of the entries including the presence of keywords and Enzyme Commission (EC) identifiers [27]. For example, UniProt protein description field may contain Enzyme Commission (EC) numbers. Using an existing mapping[18] of EC numbers to the GO molecular function ontology (ec2go) and a mapping of protein accession numbers to EC numbers, GOA can produce an association on protein accession numbers to GO[19]. For example, in the mapping file ec2go, the EC number EC:1.4.1.2 is mapped to GO:glutamate dehydrogenase activity (GO:0004352). Similarly, in the protein description field of the Uniprot protein P00367, the EC line provide the EC number EC:1.4.1.2. Therefore, using an automated method for parsing ec2go file, it is possible to annotate protein P00367 to GO term 'glutamate dehydrogenase activity' (GO:0004352). The GOA group also maintains a Swiss-Prot keyword to GO mapping (spkw2go) for similar reasons. This approach is relatively quick but produces low quality annotations derived without human validation. Therefore, the electronic methods are especially useful for the annotation of proteins that are not easily studied with experimental methods (e.g. human proteins). The quality of the electronic annotation has been assessed and it has been shown that the use of the mapping files for the generation of GO annotation precisely predicts the correct GO term 60% to 70% of the time [33].

[18] The word 'mapping' is used here to refer to the linking of various classification systems to GO terms.

[19] The word 'association' refers to a connection between a database object (which may represent a gene, transcript, or protein) and a GO term that describes the gene product.

1.4.2 Manual annotation

To provide more reliable and specific annotation, the GOA project also makes use of manual curation using information extracted from a variety of sources. However, manual annotation is time consuming and dependent on skilled biologists capable of extracting up-to-date information from the published scientific literature. Literature curation is an interpretive process. Although interactive text mining programs are being tested [34], these programs have very high error rates [35] and the interpretation of the literature by human biocurators remains the gold standard [36, 37, 38]. To minimize curator-to-curator variations and to promote consistency, GO terms are defined and curators undergo specialized training. Also the manual annotation process has been evaluated and the conclusion is that it will take far too long to complete the annotations of even just the most important organisms [39]. Several solutions have been proposed. One such solution is collaborative curation. There have been multiple calls to provide an incentive, such as a 'citable acknowledgement', for researchers to voluntarily contribute to public databases in general, and annotation of database contents in particular [40]. There have been efforts to produce open-source software for multi-user annotation of database contents [41, 42] and free text [43], as well as examples of successful community annotation projects (see Chapter 3 Section 3.3.5 for our Muscle Gene Annotation Community).

The manual annotation process is described in detail in Chapter 2.

1.5 Aims

My PhD project starts from the consideration that, in the post genomic era, researchers confront themselves with the phenomenon that, while the amount of accessible data is growing exponentially, it is becoming harder and harder to find the appropriate information. Within this scenario, there are two main obstacles that biologists have to overcome to extract knowledge from the data available. The first obstacle is that the introduction of the high-throughput techniques and the beginning of the big genome sequencing projects produce a vast amount of biological data that needs to be analysed. These heterogeneous data about genes, drugs, structures, images, diseases, proteins and so on are stored in literally thousands of databases. In addition the experiments are carried out in a wide range of species and using a variety of different experimental methods. As a result, it is very complicated to manage all the data and to try to find relationships between genes that are the cause of a certain disease, that are involved in specific biological processes and/or that are regulated by a therapeutic drug. The second obstacle is brought about by terminological and conceptual incompatibilities among databases, making the data inaccessible for computational analysis. However scientists wish to draw on information from these important data and several methods have already been developed for integrating data from different databases and for extracting useful and interesting relationships between the biological entities. Therefore, to make the best use of biological databases and the knowledge they contain, different kinds of information from different sources must be integrated in ways that make sense to the scientific community.

The aim of my PhD project is to explore the possibility of improving the existing resources and developing new methods in order to facilitate data integration about genes, drugs and diseases and their reciprocal relationships. For this purpose, I firstly decided to contribute to the development of ontologies, such as Gene Ontology and other OBO ontologies; secondly, but even more important, I developed a method to cross-link between different semantic fields represented in several distinct databases. These improved resources should result in more biological knowledge and will provide new opportunities for advanced data mining. Finally they will facilitate biological reasoning on diseases, genes, mechanisms and drugs. As a result, it will be possible to address more complicated questions like ‘how many genes are associated to a disease?’, ‘how many of these genes are also targets of a drug?’, and ‘what biological processes are these genes involved in?'. Once this additional knowledge is extracted, it will be possible to share it and make it available to the scientific community.

In collaboration with the GO Consortium, I have carried through an initiative to expand neurobiology and muscle biology representation in the GO ontologies; moreover, for inter-ontology consistency, I also contributed to the existing structure of other OBO ontologies. These revised ontologies should facilitate the interpretation of high-throughput experiments in the area of muscle and neurobiology and more important, in the fields of neuro-muscular and nervous system diseases. Finally, I used an OBO ontology, the Disease Ontology, together with other biological databases, for the integration of information about gene-drug-disease, collecting the data in a specifically designed relational database. Our ontology-based system should be able to provide interoperability support for physicians and medical researchers in the interpretation of data from studies on human diseases.

Chapter 2

Improvement of the neurobiology representation in Gene Ontology

The aim of this chapter is to describe our contribution in the improvement of neurobiology representation in the Gene Ontology vocabularies. Section 2.1 is a brief overview on the reasons for starting this work and the valuable collaboration with GO community. Then Sections 2.2 and 2.3 are respectively about our contribution to the GO with new terms and the annotation of genes relevant to nervous system diseases and involved in processes modulated by drug treatments. The list of new terms implemented in GO as well as the list of genes manually annotated can be find in Appendices A and B.

2.1 Background

This PhD project was in collaboration with Glaxo Smith Kline (GSK) in Verona (Italy), which is one of the several Centres of Excellence for drug discovery focused on neurodegenerative and neuropsychiatric disorders such as Alzheimer's disease, mood disorders, and in particular depression and schizophrenia.

With the fundamental contribution of GSK experts, the initial step was to integrate different kinds of information from several sources in order to better exploit with bioinformatics tools the knowledge about genes, drugs and disorders related to the nervous system (NS). At the beginning of my PhD project, the GO had a very shallow representation of processes pertaining to the nervous system. Such lack of defined terms was particularly serious and made impossible a description of nervous system specific events and functions. Therefore, in order to achieve one of the goals of this project, namely to create a high-quality Gene Ontology structure of nervous system terms, the first step was to contact the Gene Ontology Consortium and present our ideas.

This led to a collaboration and a visiting period of six months at EBI (European Bioinforamtics Institute) in Cambridge (UK), that gave me the opportunity to interact intensively with the Gene Ontology and GOA groups. Being in touch directly with the experts in editing ontologies and annotating genes made this experience very productive and useful for understanding the main activities around the GO and GOA initiatives. The principle advantage in visiting the GO community was that I was well-trained in ontology development and annotation. Several resources, both bioinformatics and literature, were made available, among them a variety of editing and annotation tools (e.g. OBO-Edit and protein2go). The first period was dedicated to learning the method used by a GO editor to find, define and finally add terms to GO, under the lead of GO editors: Jennifer Deegan (nee Clark) (for the process ontology), Amelia Ireland (for the function ontology), Jane Lomax and Midori Harris (for general problems about ontology editing). To become a GO editor, the first aim was to learn and correctly apply the rules at the base of GO editing (when a definition needs to be changed, when a term has to be obsoleted, when a relationship has to be fixed and many other rules).

Learning and applying the standard rules is essential for maintaining the structure and the syntax of the GO vocabularies. This process required an intense period of training. I learnt how to edit the Gene Ontology and how to check my edits using OBO-Edit. When adding terms, I applied the GO general conventions, including the True-path and the *is_a* complete rules. The *is_a* complete rule means that all non-root terms in the ontologies should have an is_a parent, and a complete path to the root of the ontology that traverses only is_a relations. Making changes to the Gene Ontology is not a straightforward process, and decisions on when a term can be inserted or modified are based on long discussions between GO curators and also GO users. This discussion process enables representation of the information in both an ontologically and biologically correct way. The SourceForge (SF) tracker system is an additional helpful resource for managing the discussions. It is an on-line discussion forum and it holds information about all of the work currently in progress in the GO consortium. Additionally, the SF system provides a GO SourceForge curator request tracker showing a list of the requests for changes to the ontologies that have been proposed by GO users. There are also lists about OBO-Edit bugs and annotation problems.

At the same time, tools and resources available to make manual gene annotations have been used; in particular, the proten2go tool, developed by GOA group at EBI, that is a web-based GO annotation tool that helps annotation of Swiss-Prot entries to GO terms. For the manual annotation of interesting genes, the teaching from the GOA annotators (Emily Dimmer, Evelyn Camon and Rachael Huntley) was essential.

Once I rejoined the group in Padua at the end of these few months, the collaboration with GO editors and curators continued, and in particular, the work has been carried out with the extremely valuable help of Jennifer Deegan (nee Clark), member of the GO editorial office and manager of the Annotation Outreach group. The benefits achieved with this collaboration are countless. First of all, the FREE access to the GO CVS repository that has been set up to house the GO data. CVS, or concurrent versions system, is a tool that allows multiple users to edit a file simultaneously. Secondly, the privileges to work on the live GO file. Then, the opportunity to join critical quality control working groups such as OBO-Edit working group and specific GO chats that utilise several software system (e.g. Webex, IRC, Skype, Google shared documents). Lastly, the possibility to keep a productive collaboration with GO groups based on frequent (once a week) conference call to discuss and agree upon concepts, to get problems resolved and to get GO changes completed and implemented.

Due to our contribution of annotations and significant collaboration on major GO content development, the Genomics Group[1] of Prof. Giorgio Valle at CRIBI (Padua) is now an associate member of GO Consortium.

2.2 Representation of nervous system biology in GO

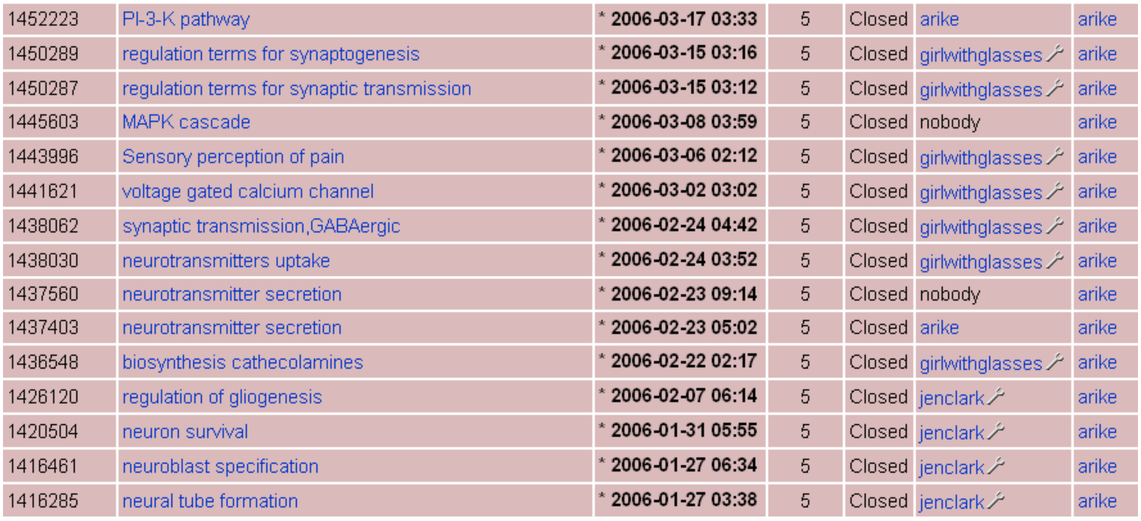
2.2.1 Method and results

To providing a high-quality representation of nervous system biology in the Gene Ontology, the working plan was roughly based on the re-definition of the existing GO nodes and the addition of new specific terms that were more informative and detailed than those available in

[1] http://genomics.cribi.unipd.it

the three GO. Classes of the biological process and molecular function ontologies more closely related to the depressive disorders and to the neurodegenerative diseases have been treated as mostly relevant. Such classes have been analysed in detail to pinpoint the nodes that were poorly described and that would benefit from an enrichment of new terms. Afterwards the improved GO structure was used for the manual annotation of genes associated to specific processes and functions with a reasonable degree of certainty.

In order to give a valuable and reliable contribution to GO content in the area of nervous system biology, several resources of information were exploited to create a solid basic knowledge on neurodegenerative and neuropsychiatric diseases, their neurobiology and the possible involved mechanisms. Many internet resources were available to accomplish this step.



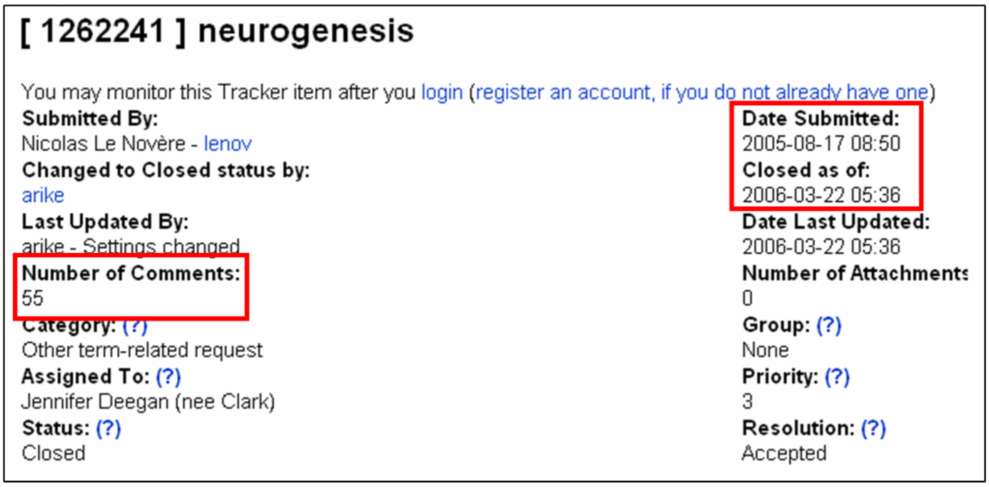
1452223	PI-3-K pathway	* 2006-03-17 03:33	5	Closed	arike	arike
1450289	regulation terms for synaptogenesis	* 2006-03-15 03:16	5	Closed	girlwithglasses	arike
1450287	regulation terms for synaptic transmission	* 2006-03-15 03:12	5	Closed	girlwithglasses	arike
1445603	MAPK cascade	* 2006-03-08 03:59	5	Closed	nobody	arike
1443996	Sensory perception of pain	* 2006-03-06 02:12	5	Closed	girlwithglasses	arike
1441621	voltage gated calcium channel	* 2006-03-02 03:02	5	Closed	girlwithglasses	arike
1438062	synaptic transmission,GABAergic	* 2006-02-24 04:42	5	Closed	girlwithglasses	arike
1438030	neurotransmitters uptake	* 2006-02-24 03:52	5	Closed	girlwithglasses	arike
1437560	neurotransmitter secretion	* 2006-02-23 09:14	5	Closed	nobody	arike
1437403	neurotransmitter secretion	* 2006-02-23 05:02	5	Closed	arike	arike
1436548	biosynthesis cathecolamines	* 2006-02-22 02:17	5	Closed	girlwithglasses	arike
1426120	regulation of gliogenesis	* 2006-02-07 06:14	5	Closed	jenclark	arike
1420504	neuron survival	* 2006-01-31 05:55	5	Closed	jenclark	arike
1416461	neuroblast specification	* 2006-01-27 06:34	5	Closed	jenclark	arike
1416285	neural tube formation	* 2006-01-27 03:38	5	Closed	jenclark	arike

Figure 2.1: List of some SF request items submitted by ‘arike’ (Erika Feltrin SF user name). There are requests for new terms such as ‘neurotransmitter secretion and uptake’, ‘synaptic transmission’ and ‘neuroblast specification’.

The information has been also collected from literature sources like books (e.g. Principle of Neural Science written by Eric R. Kandel) and several journals of neurobiology and neuroscience. However, the involvement of people at the GSK Biology department were absolutely necessary to properly take into consideration the complexity of both the processes and the cellular machinery involved in these disorders. They were also helpful to elucidate issues related to genes and signalling pathways involved in the root mechanisms of actions of certain drugs (e.g. antidepressants). The collected available information on mechanisms and pathways closely related to disorders, has been used to conceive new terms to be added to the Gene Ontology.

Choosing a certain NS disorder, the strategy was to gather as much information as possible about the fundamental mechanism and the affected biological processes. Then, GO was browsed seeking GO terms suitable for representing the retrieved information. If any appropriate term was not available, a request for a new GO term was added to the curator request SF tracker (Figure 2.1).

The SF proposal usually is taken up by a GO curator who discusses with other GO curators and users whether the proposed term should be accepted, and how it should be named and defined within the Gene Ontology. The discussion occurring on line, often continues for a long time until all participants agree with the suggested term. For example, the 'neurogenesis' item (SF:1262241) was submitted in August, 2005 and closed March, 2006 (Figure 2.2). The proposal required about 8 months, 36 pages of discussion correspondent to 6615 words before being accepted.



[1262241] neurogenesis

You may monitor this Tracker item after you login (register an account, if you do not already have one)

Submitted By:
Nicolas Le Novère - lenov
Changed to Closed status by:
arike
Last Updated By:
arike - Settings changed
Number of Comments:
55
Category: (?)
Other term-related request
Assigned To: (?)
Jennifer Deegan (nee Clark)
Status: (?)
Closed

Date Submitted:
2005-08-17 08:50
Closed as of:
2006-03-22 05:36
Date Last Updated:
2006-03-22 05:36
Number of Attachments
0
Group: (?)
None
Priority: (?)
3
Resolution: (?)
Accepted

Figure 2.2: 'neurogenesis' (SF1262241)[2] was submitted in August, 2005 and closed March, 2006. The proposal required about 8 months, 36 pages of discussion correspondent to 6615 words before being accepted.

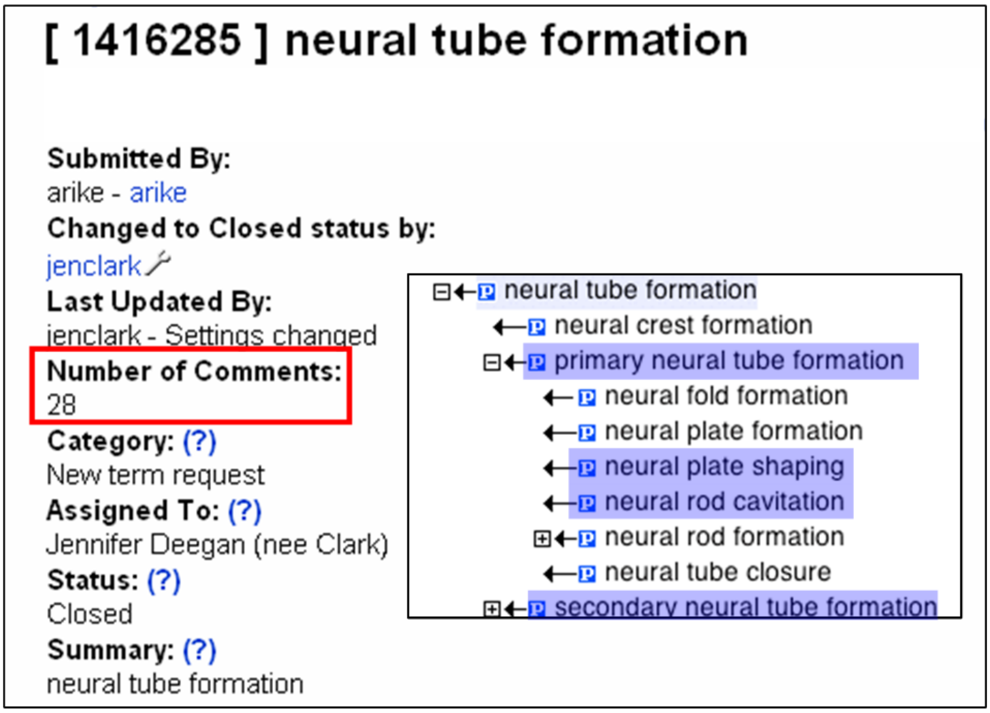
Once the requests are accepted, the curators add terms to the GO and then make them available for gene annotation (Figure 2.3).

[2] https://sourceforge.net/tracker/index.php?func=detail&aid=1262241&group_id=36855&atid=440764

2.2.2 A case: depression and mechanisms of regulation

The following sections describe an example of how the information on a certain disease (e.g. depression) has been collected and translated into GO terms.

There are several hypotheses on depression and antidepressant mode of action. Some of those are based largely on the dysregulation of the hypothalamic-pituitary-adrenal axis (HPA) and hippocampus and involve corticotropin-releasing hormone (CRH), glucocorticoids, brain-derived neurotrophic factors (BDNF) and CREB [44]. Others focus on the fact that depression is often described as a stress-related disorder and there are good evidences that episodes of depression often occur in response to stress to some trauma.



[1416285] neural tube formation

Submitted By:
arike - arike
Changed to Closed status by:
jenclark
Last Updated By:
jenclark - Settings changed
Number of Comments:
28
Category: (?)
New term request
Assigned To: (?)
Jennifer Deegan (nee Clark)
Status: (?)
Closed
Summary: (?)
neural tube formation

- ⊟ P neural tube formation
  - P neural crest formation
  - ⊟ P primary neural tube formation
    - P neural fold formation
    - P neural plate formation
    - P neural plate shaping
    - P neural rod cavitation
    - ⊞ P neural rod formation
    - P neural tube closure
  - ⊞ P secondary neural tube formation

Figure 2.3: 'neural tube formation' (SF1416285)[3]. The proposal suggested a new definition for the existing term 'neural tube formation' and some new child terms. After about 30 comments, the discussion resulted in the implementation of a new structure for the 'neural tube formation' node and in the addition of several new terms such as 'neural plate shaping' and 'neural rod cavitation'.

[3] https://sourceforge.net/tracker/index.php?func=detail&aid=1416285&group_id=36855&atid=440764

A prominent mechanism by which the brain reacts to acute and chronic stress is through the activation of the HPA axis. When activated by exposure to stressors, CRH is produced within the hypothalamus. In turn, CRH stimulates the anterior pituitary gland to release adrenocorticotropic hormone (ACTH) into the bloodstream. ACTH then stimulates the release of glucocorticoids (e.g. cortisol in human) from the adrenal cortex [45]. Circulating *glucocorticoids* interact with their receptors in various target organs such as the liver and muscle tissue, as well as the brain and the HPA axis itself. Here they are responsible for initiating feedback inhibition. Thus they exert profound effects on general metabolism and also affect several processes like *neurogenesis, survival of neurons, neuronal plasticity, neuronal cell proliferation and cell death* [46].

Other hypothesis suggests a role for *neurotrophic factors* at the basis of depression. They regulate *neuronal growth* and differentiation during development but are also known to be potent regulators of *plasticity* and *survival* of adult neurons and glia cells.

Many papers dealing with the neurotrophin hypothesis have shown that acute and chronic stress decreases levels of BDNF expression in several brain regions [47].

Starting with all this information, specific categories have been considered and then represented in GO in a comprehensive way.

2.2.3 Improved GO categories

Further interesting processes and mechanisms related to other disorders have been converted into specific GO terms, or have been redefined given the existing GO terms. This approach resulted in a great improvement of the representation of biological processes that contribute, and that are common, to different nervous system disorders. For a complete list of all new and modified GO terms see Appendices A. These terms related to specific topics (listed in Table 2.1) have been submitted to the Gene Ontology curators and implemented.

Apoptosis plays an important role during neuronal development [48]. Defects in the control machinery of this mechanism drive to modified cell death processes, and may also play an important role in the etiology of various pathologies of the nervous system such as Parkinson's disease and Alzheimer's disease. For covering this process, more specific GO terms have been provided (Figure 2.4).

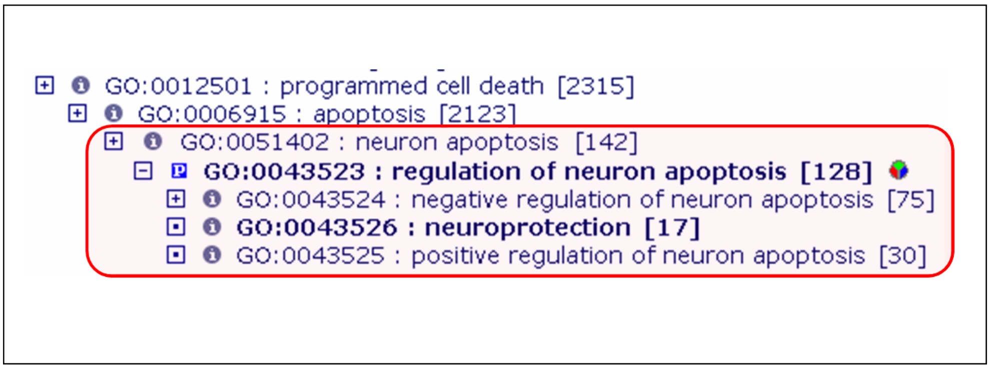
neuron apoptosis	steroid hormone (e.g. glucocorticoids)
neuroprotection mechanisms	lithium and its mechanism of action
synaptic plasticity and its modulation	stress and the response mechanism
neurotransmitters secretion, uptake and transmission	CRH, ACTH and their receptors
BDNF and other neurotrophic factors	cortisol and its secretion

Table 2.1: List of the topic, related to nervous system, covered in Gene Ontology

Neuronal apoptosis can be induced by different stimuli including deficiency of survival factors like **BDNF, NGF and neurotrophins 3 and 4/5** [49]. These same factors have been postulated to contribute to neuronal cell loss in human neurodegenerative disorders [50]. For this reasons, the representation of such factors in the GO has been improved (Figure 2.5).

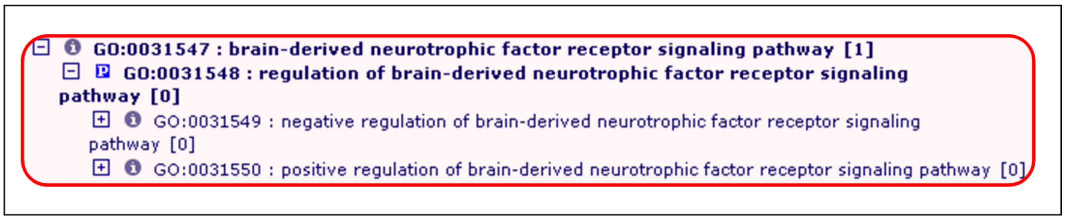
Recent studies have shown that lithium has an important role in regulating neuronal survival; it modulates **neurotransmitter levels** and readjusts the balance between **excitatory and inhibitory activities**. Lithium also modulates signals impacting on the cytoskeleton system relevant to the neuronal plasticity. Thus, new terms suitable to describe the action of lithium have been proposed (e.g. response to lithium ion (GO:0010226)).

Along with lithium, other neurotrophic factors like BDNF are necessary for the survival and function of neurons and for the maintenance and regulation of neuronal plasticity. **Plasticity** is the ability of the brain to rearrange the connections between its neurons. It is important in



⊞ GO:0012501 : programmed cell death [2315]
⊞ GO:0006915 : apoptosis [2123]
⊞ GO:0051402 : neuron apoptosis [142]
⊟ **GO:0043523 : regulation of neuron apoptosis [128]**
⊞ GO:0043524 : negative regulation of neuron apoptosis [75]
⊞ **GO:0043526 : neuroprotection [17]**
⊞ GO:0043525 : positive regulation of neuron apoptosis [30]

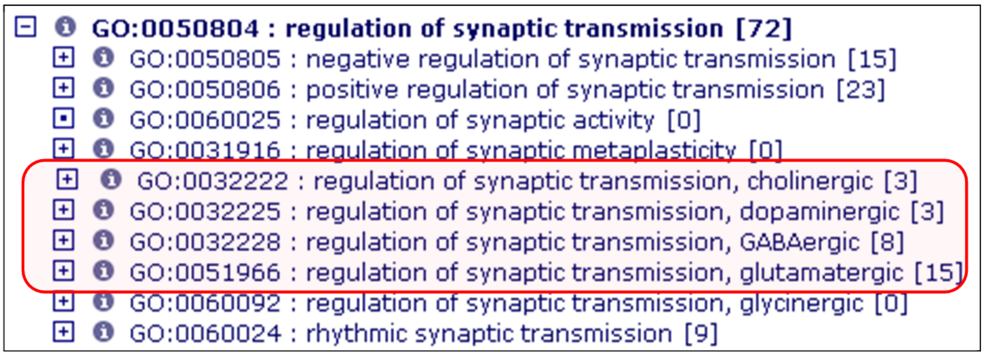
Figure 2.4: AmiGO screenshot of the new terms about ‘apoptosis’. The numbers in brackets show the number of genes annotated to each term.



⊟ GO:0031547 : brain-derived neurotrophic factor receptor signaling pathway [1]
⊟ GO:0031548 : regulation of brain-derived neurotrophic factor receptor signaling pathway [0]
⊞ GO:0031549 : negative regulation of brain-derived neurotrophic factor receptor signaling pathway [0]
⊞ GO:0031550 : positive regulation of brain-derived neurotrophic factor receptor signaling pathway [0]

Figure 2.5: The new ‘BDNF receptor signalling pathway’ GO category with regulation terms. The numbers in brackets show the number of genes annotated to each term.

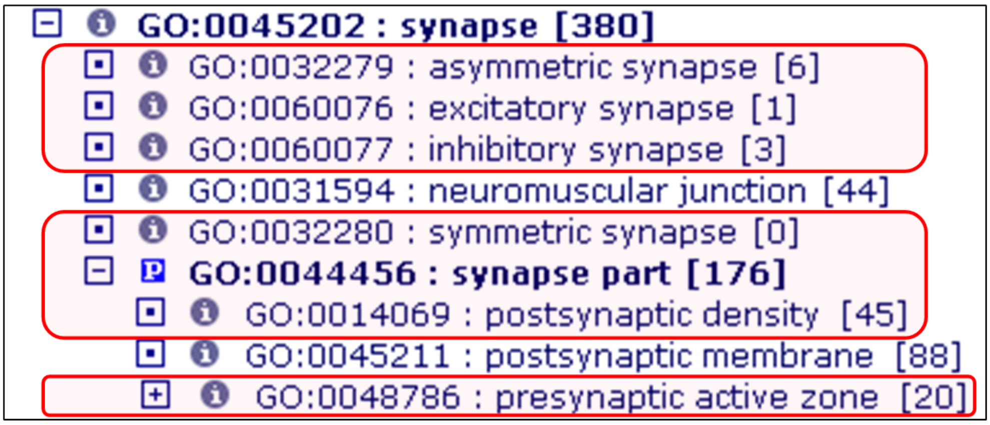
compensating brain damage by allowing the brain to create new networks of neurons. These local changes to brain structure depend on the environment and represent an adaptation to it. For this adaptation mechanism, it is necessary to select neurons, increase the number of connections, and release more neurotransmitters, among other actions. Neurotransmitters are essential chemical messengers that regulate brain, muscle, nerve and organ functions; and over 60 diseases and illnesses may be caused by, or associated with, neurotransmitter deficiency. On account of this, part of the new structure covered processes such as neurotransmitters re-uptake, release and secretion and, regulation of synaptic transmission (Figure 2.6).



⊟ GO:0050804 : regulation of synaptic transmission [72]
⊞ GO:0050805 : negative regulation of synaptic transmission [15]
⊞ GO:0050806 : positive regulation of synaptic transmission [23]
⊡ GO:0060025 : regulation of synaptic activity [0]
⊞ GO:0031916 : regulation of synaptic metaplasticity [0]
⊞ GO:0032222 : regulation of synaptic transmission, cholinergic [3]
⊞ GO:0032225 : regulation of synaptic transmission, dopaminergic [3]
⊞ GO:0032228 : regulation of synaptic transmission, GABAergic [8]
⊞ GO:0051966 : regulation of synaptic transmission, glutamatergic [15]
⊞ GO:0060092 : regulation of synaptic transmission, glycinergic [0]
⊞ GO:0060024 : rhythmic synaptic transmission [9]

Figure 2.6: New regulation terms for the ‘synaptic transmission’ process. The numbers in brackets show the number of genes annotated to each term.

Since the **synapse** has a crucial role in synaptic transmission and in interneuronal communication, the biological process and the cellular component ontologies were enriched with new terms. New biological process terms were added such as ‘regulation of synaptogenesis’, ‘regulation of excitatory postsynaptic membrane potential’ and ‘regulation of inhibitory postsynaptic membrane potential’, as well as additional cellular component concepts like ‘asymmetric and symmetric synapse' and 'postsynaptic density' (Figure 2.7).



⊟ GO:0045202 : synapse [380]
⊡ GO:0032279 : asymmetric synapse [6]
⊡ GO:0060076 : excitatory synapse [1]
⊡ GO:0060077 : inhibitory synapse [3]
⊡ GO:0031594 : neuromuscular junction [44]
⊡ GO:0032280 : symmetric synapse [0]
⊟ GO:0044456 : synapse part [176]
⊡ GO:0014069 : postsynaptic density [45]
⊡ GO:0045211 : postsynaptic membrane [88]
⊞ GO:0048786 : presynaptic active zone [20]

Figure 2.7: Improved representation of synapse in the cellular component ontology. The numbers in brackets show the number of genes annotated to each term.

Some studies have demonstrated that alteration of synaptic plasticity and regulation of neurogenesis might be caused by various stressors that act by decreasing expression of BDNF. Plasticity can also be modulated by glucocorticoids, excitatory aminoacids and NMDA receptors [51]. For this reason, the improvement of the representation of synaptic activities, neurogenesis and other mechanisms in the GO vocabularies, was considered of great importance. A total of 228 new terms have been implemented in the GO during these three-year project. In the Figure 2.8 such terms have been grouped in more general categories for a better representation.

2.2.4 The Neurobiology Content Meeting

Since GO people were very interested in improving the GO in the neurobiology area, they organised the First Neurobiology Content meeting[4].

In the Spring of 2006 curators decided to focus on better representation of the nervous system in GO. In particular, the interest was placed on three areas: forebrain development, hindbrain development and neural tube development. In June 2006, during a two-day meeting in Bar Harbor (Maine, US), the graphs were discussed among three curators and some experts

[4] For full description of GO Content Meetings see the next chapter.

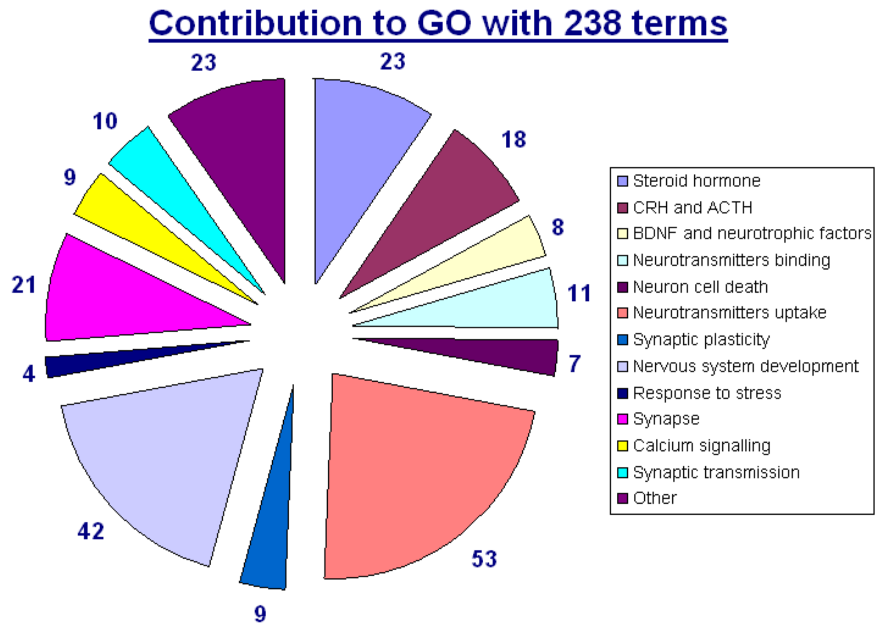


Figure 2.8: Representation of the 228 new terms grouped in general categories.

in NS development, neuroanatomy and ontology development (including myself). Changes to the ontologies were made directly as discussions proceeded and after the meeting further revisions to the graph were made based on the discussions. Then the final graph was committed to the live GO repository some weeks later after further community consultation.

Over 500 terms were added to the ontology dealing with nervous system development. The new NS terms provide a framework that is easily extensible for the addition of new terms, as they are required for literature-based curation.

2.3 Functional annotation of nervous system specific genes

The improvement of the Gene Ontology with terms relevant to nervous system (NS) and neurobiology took a few months. After that period, the functional annotation of interesting genes, obviously using the new GO terms, was started as a completion of our contribution to GOA project. The new GO terms gave new opportunities for capturing the most recent information about those genes. Starting from a list of genes (provided by GSK) that are either involved in NS diseases or able to regulate NS biological processes, their annotations in UniProt were evaluated and if necessary, improved and update to reflect the most recent studies. In the meanwhile, the contribution to GO vocabularies by adding new terms proceeded.

2.3.1 List of annotated genes

The manual annotation process involved the annotation of 78 genes from a total of 186. This list, populated by GSK experts, consisted of genes and proteins implicated in the different depression hypotheses (the neuroendocrine one (HPA axis dysregulation), the neuroimmune one, the neuroplasticity one and monoamine one). The genes are also interesting because they interact with corticotropin-releasing hormone (CRH) which is involved in the depression hypothesis based largely on dysregualtion of the HPA axis. In addition NS biological processes such as synaptic transmission and plasticity, neurotransmitter uptake and release might be regulated by the activity of these specific proteins. For example, neurotransmitters are key molecules in several NS mechanisms and the activity of their receptors might be modulated by drug treatments. Serotonin is the neurotransmitter targeted by the drug Prozac; the GABA neurotransmitter is affected by Valium; and most Alzheimer's drugs affect the neurotransmitter acetylcholine. The choice to start the annotation with the glutamate receptors came from consideration that these far outnumber the other types of neurotransmitter receptors in the human and mouse brain. The manual annotation covered the functional association of metabotropic and iontropic (kainate, NMDA and AMPA) glutamate receptors.

2.3.2 Manual annotation method

An annotation is basically a statement that a gene product has a particular molecular function, is involved in a particular biological process, is located within a certain cellular component as determined by a particular method and as described in a particular literature reference. As said in the introduction, there are two ways of creating annotations: by applying computational methods (electronic annotation) and by an annotator reading through scientific literature and manually creating each annotation (manual annotation).

GO annotators follow the guidelines set down by GO Consortium. The same method was applied for the annotation of our genes and when GO terms necessary for the annotation were not available within the GO vocabularies, SF requests for new term have been submitted.

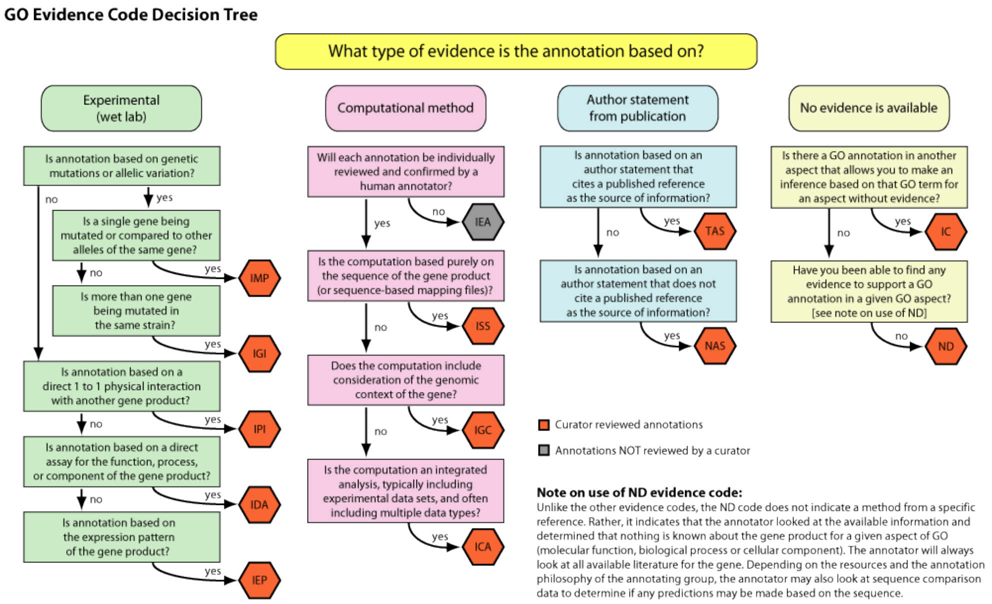
An understanding of how these annotations are made is crucial for their correct interpretation and utility. A GO annotation includes four essential pieces of information:

1. accession number of a gene product that the annotator is interested in;

2. ID number of the GO term or terms describing the gene product;

3. the evidence code[5] providing a broad categorization of the type of evidence on which the annotation is based;

4. the reference number of the paper or information source in which the evidence was found.


GO Evidence Code Decision Tree
What type of evidence is the annotation based on?
Experimental (wet lab)
Computational method
Author statement from publication
No evidence is available
Is annotation based on genetic mutations or allelic variation?
no
yes
Is a single gene being mutated or compared to other alleles of the same gene?
no
yes
IMP
Is more than one gene being mutated in the same strain?
yes
IGI
Is annotation based on a direct 1 to 1 physical interaction with another gene product?
no
yes
IPI
Is annotation based on a direct assay for the function, process, or component of the gene product?
no
yes
IDA
Is annotation based on the expression pattern of the gene product?
yes
IEP
Will each annotation be individually reviewed and confirmed by a human annotator?
yes
no
IEA
Is the computation based purely on the sequence of the gene product (or sequence-based mapping files)?
no
yes
ISS
Does the computation include consideration of the genomic context of the gene?
no
yes
IGC
Is the computation an integrated analysis, typically including experimental data sets, and often including multiple data types?
yes
ICA
Is annotation based on an author statement that cites a published reference as the source of information?
no
yes
TAS
Is annotation based on an author statement that does not cite a published reference as the source of information?
yes
NAS
Is there a GO annotation in another aspect that allows you to make an inference based on that GO term for an aspect without evidence?
no
yes
IC
Have you been able to find any evidence to support a GO annotation in a given GO aspect? [see note on use of ND]
no
ND
🟥 Curator reviewed annotations
▩ Annotations NOT reviewed by a curator

Note on use of ND evidence code:
Unlike the other evidence codes, the ND code does not indicate a method from a specific reference. Rather, it indicates that the annotator looked at the available information and determined that nothing is known about the gene product for a given aspect of GO (molecular function, biological process or cellular component). The annotator will always look at all available literature for the gene. Depending on the resources and the annotation philosophy of the annotating group, the annotator may also look at sequence comparison data to determine if any predictions may be made based on the sequence.

Figure 2.9: The Evidence Code Decision Tree[6] that has to be followed in the manual annotation process.

The EC types 'traceable author statement' (TAS) and 'non traceable author statement' (NAS) refer to remarks made by authors in the literature, and the code 'inferred by curator' (IC) is used where a curator has deduce the annotation from other available data. Several ECs, including IMP 'inferred from mutant phenotype' and IDA 'inferred from direct assay', refer to experimental data, and as such generally give the most detailed annotations, depending on the

[5] http://www.geneontology.org/GO.evidence.shtml
[6] http://www.geneontology.org/GO.evidence.tree.shtml

experimental technique (for example, annotation to 'inhibition of CREB transcription factor' versus less granular 'regulation of transcription'). Others ECs are 'inferred by electronic annotation' (IEA), 'reviewed computational annotation' (RCA) and 'inferred by sequence similarity' (ISS).

The flow chart in Figure 2.9 illustrates the steps that a curator typically takes when deciding what evidence code should be applied to an annotation.

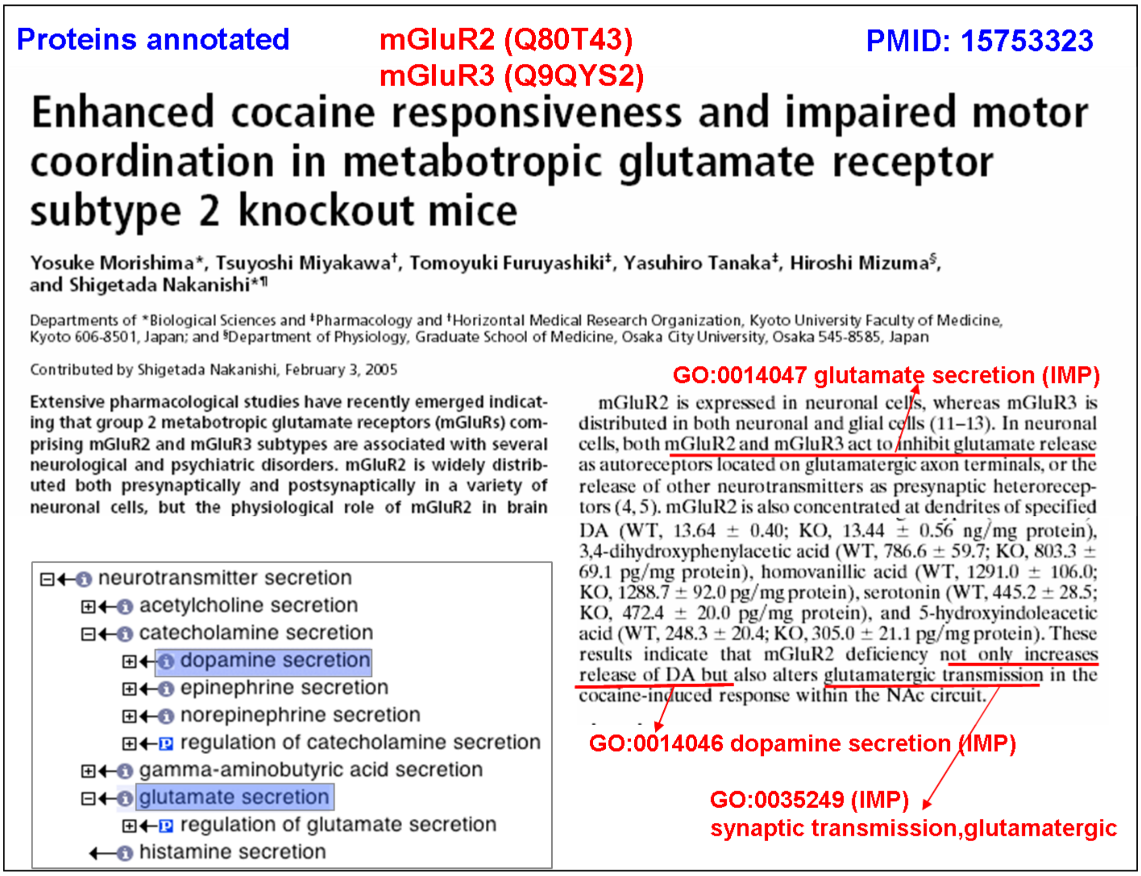
2.3.3 An example: the annotation of proteins mGLUR2 and mGLUR3

Figure 2.10 shows an example of how the annotator, reading the most recent scientific literature, converts the information into protein-GO term annotations. A paper reports experiments in a knockout mouse showing that metabotropic glutamate receptors type 2 and 3 (mGLUR2 and mGLUR3) can inhibit glutamate secretion. At the same time they can alter the dopamine secretion and the synaptic transmission mediated by glutamate neurotransmitter.

To make an annotation, the annotator has to find the four pieces of information. The accession number of the proteins can be easily found from the UniProt database (Q80T43 for mGluR2 and Q9QYS2 for mGluR3). Then, the annotator has to note the reference number for the paper (PMID:15753323) and the proper EC supporting the evidence (IMP, 'inferred from mutant phenotype'). Then, she must select the most appropriate GO term for biological processes in which these proteins take part (glutamate secretion (GO:0014047)) (Figure 2.10).

A single gene product could be annotated to multiple terms from all three ontologies to capture all of the relevant information. This is the case, mGluR2 can be annotated with the previous 'glutamate secretion' (GO:0014070) but also with 'synaptic transmission, glutamatergic' (GO:0035249).

The 'glutamate secretion' term and its descendants were created during the previous phase of GO editing and, as seen in the example above, these new nervous system specific terms were particularly helpful in the annotation of such proteins.



Proteins annotated mGluR2 (Q80T43) mGluR3 (Q9QYS2) PMID: 15753323

Enhanced cocaine responsiveness and impaired motor coordination in metabotropic glutamate receptor subtype 2 knockout mice

Yosuke Morishima[*], Tsuyoshi Miyakawa[†], Tomoyuki Furuyashiki[‡], Yasuhiro Tanaka[‡], Hiroshi Mizuma[§], and Shigetada Nakanishi[*¶]

Departments of [*]Biological Sciences and [†]Pharmacology and [‡]Horizontal Medical Research Organization, Kyoto University Faculty of Medicine, Kyoto 606-8501, Japan; and [§]Department of Physiology, Graduate School of Medicine, Osaka City University, Osaka 545-8585, Japan

Contributed by Shigetada Nakanishi, February 3, 2005

Extensive pharmacological studies have recently emerged indicating that group 2 metabotropic glutamate receptors (mGluRs) comprising mGluR2 and mGluR3 subtypes are associated with several neurological and psychiatric disorders. mGluR2 is widely distributed both presynaptically and postsynaptically in a variety of neuronal cells, but the physiological role of mGluR2 in brain

- ⊟ neurotransmitter secretion
  - ⊞ acetylcholine secretion
  - ⊟ catecholamine secretion
    - ⊞ dopamine secretion
    - ⊞ epinephrine secretion
    - ⊞ norepinephrine secretion
    - ⊞ regulation of catecholamine secretion
  - ⊞ gamma-aminobutyric acid secretion
  - ⊟ glutamate secretion
    - ⊞ regulation of glutamate secretion
  - histamine secretion

GO:0014047 glutamate secretion (IMP)

mGluR2 is expressed in neuronal cells, whereas mGluR3 is distributed in both neuronal and glial cells (11–13). In neuronal cells, both mGluR2 and mGluR3 act to inhibit glutamate release as autoreceptors located on glutamatergic axon terminals, or the release of other neurotransmitters as presynaptic heteroreceptors (4, 5). mGluR2 is also concentrated at dendrites of specified

DA (WT, 13.64 ± 0.40; KO, 13.44 ± 0.56 ng/mg protein), 3,4-dihydroxyphenylacetic acid (WT, 786.6 ± 59.7; KO, 803.3 ± 69.1 pg/mg protein), homovanillic acid (WT, 1291.0 ± 106.0; KO, 1288.7 ± 92.0 pg/mg protein), serotonin (WT, 445.2 ± 28.5; KO, 472.4 ± 20.0 pg/mg protein), and 5-hydroxyindoleacetic acid (WT, 248.3 ± 20.4; KO, 305.0 ± 21.1 pg/mg protein). These results indicate that mGluR2 deficiency not only increases release of DA but also alters glutamatergic transmission in the cocaine-induced response within the NAc circuit.

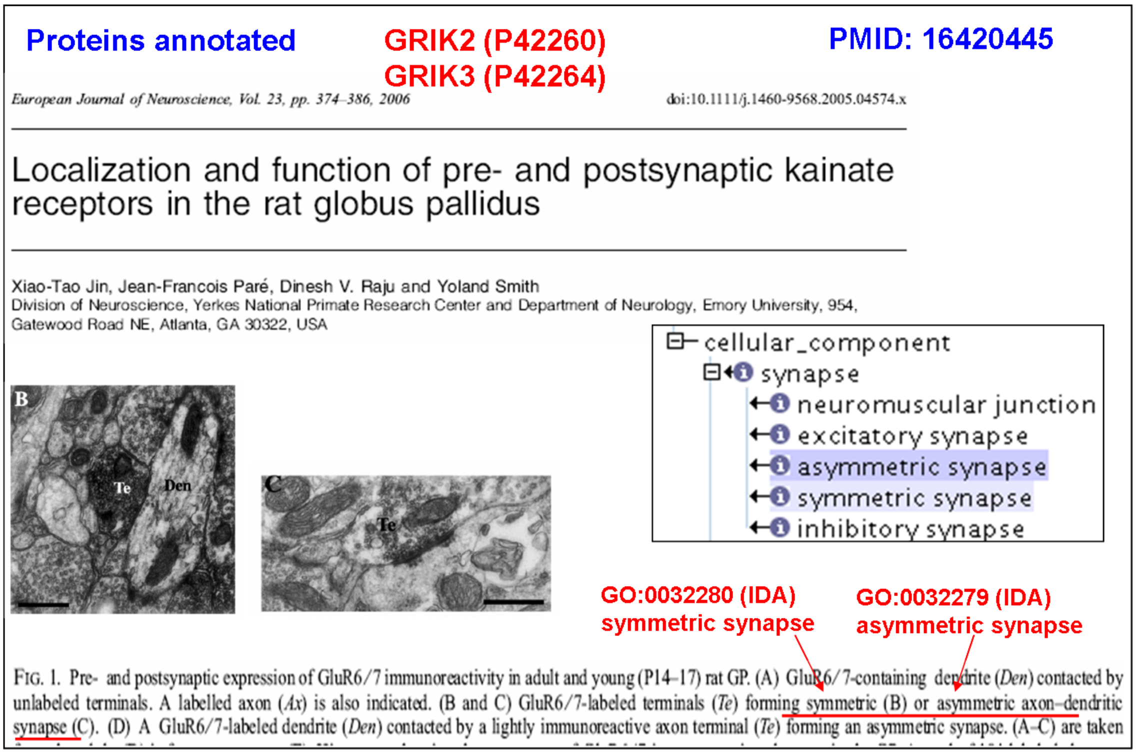
GO:0014046 dopamine secretion (IMP)

GO:0035249 (IMP) synaptic transmission,glutamatergic

Figure 2.10: Example of the annotation for two metabotropic glutamate receptors, mGLUR2 and mGLUR3. Using the protein names found in the paper, then the corresponding Uniprot sequence identifiers have been found: Q80T43 for mGLUR2, and Q9QYS2 for mGLUR3. This paper provides experimental evidence for biological process annotations for both mGLUR2 and mGLUR3 (words underlined in red). Based on this assay in knockout mice, mGLUR2 and mGLUR3 can be annotated with the GO biological process 'dopamine secretion' (GO:0014046) and 'glutamate secretion' (GO:0014047) and 'synaptic transmission, glutamatergic' (GO:0035249) using IMP 'inferred from mutant phenotype since the experiments was carried in mutant mice. On the left there is the GO tree showing the terms previously added that were useful in the annotation of such proteins.

2.3.4 An example: the annotation of proteins GRIK2, GRIK3 and GRIA1

Figure 2.11 and Figure 2.12 present further examples of gene annotations. Ionotropic glutamate receptors are divided into three classes: AMPA, NMDA and kainate (KA) receptors. GRIK2 and GRIK3 (also known as GluR6 and GluR7) belong to the KA receptor family, while GRIA1 (also GluR1) is an AMPA glutamate receptor. Papers discussing experiments to examine the localisation and the function of KA and AMPA receptors have been considered.



European Journal of Neuroscience, Vol. 23, pp. 374–386, 2006 doi:10.1111/j.1460-9568.2005.04574.x

Localization and function of pre- and postsynaptic kainate receptors in the rat globus pallidus

Xiao-Tao Jin, Jean-Francois Paré, Dinesh V. Raju and Yoland Smith
Division of Neuroscience, Yerkes National Primate Research Center and Department of Neurology, Emory University, 954, Gatewood Road NE, Atlanta, GA 30322, USA

FIG. 1. Pre- and postsynaptic expression of GluR6/7 immunoreactivity in adult and young (P14–17) rat GP. (A) GluR6/7-containing dendrite (*Den*) contacted by unlabeled terminals. A labelled axon (*Ax*) is also indicated. (B and C) GluR6/7-labeled terminals (*Te*) forming symmetric (B) or asymmetric axon–dendritic synapse (C). (D) A GluR6/7-labeled dendrite (*Den*) contacted by a lightly immunoreactive axon terminal (*Te*) forming an asymmetric synapse. (A–C) are taken

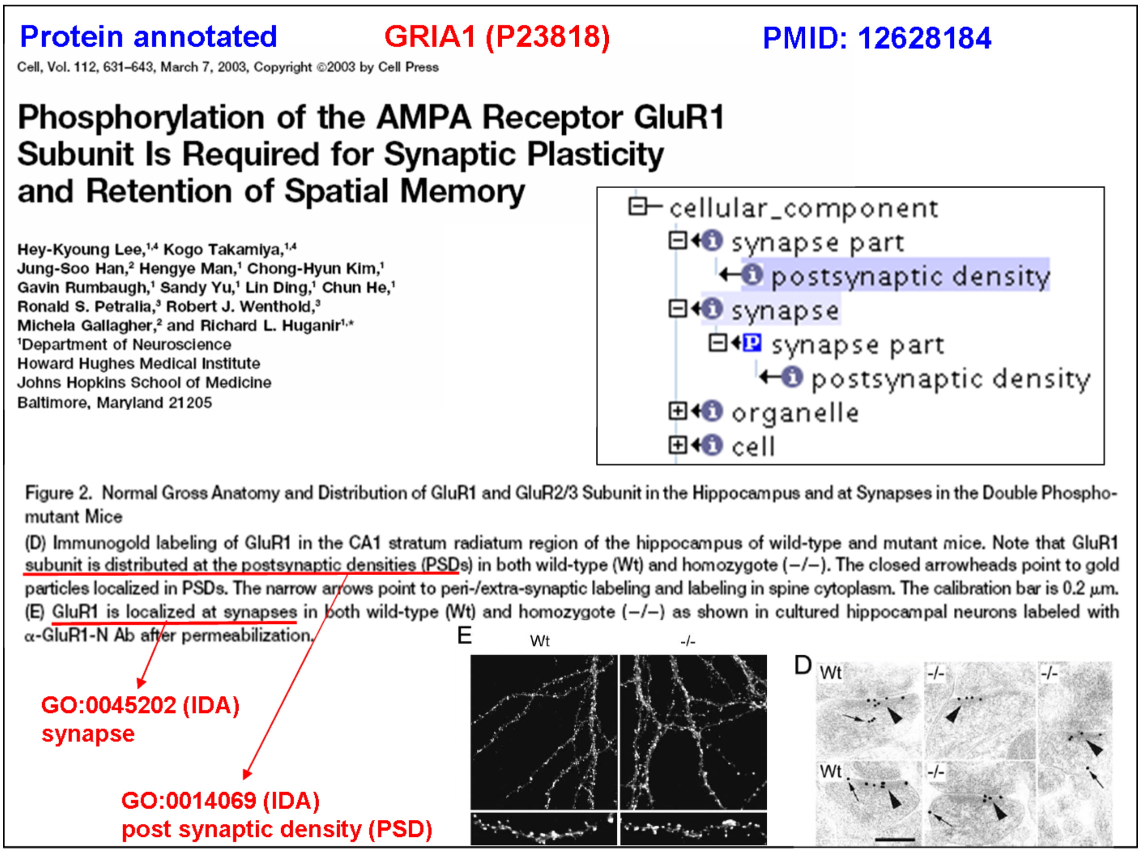
Figure 2.11: In this paper two proteins can be annotated: GRIK2 and GRIK3. Jin et al. provide experimental evidence for cellular component annotations for both GRIK2 and GRIK3. There is evidence that these glutamate receptors localise to symmetric and asymmetric synapses (words underlined in red). At the time of curation, terms for ‘symmetric and asymmetric synapse’ did not exist. Therefore a SourceForge item (SF:1481900)[7] was made to request that these new terms be added to GO. Thus, these proteins have been annotated to the new GO terms ‘asymmetric’ (GO:0032279) and ‘symmetric’ (GO:0032280) synapse (GO DAG graph on the right) using the evidence code IDA ‘inferred from direct assay’.

[7] https://sourceforge.net/tracker/index.php?func=detail&aid=1481900&group_id=36855&atid=440764

In the first paper (PMID:16420445) (Figure 2.11) immunoreactivity studies confirmed that both GRIK2 and GRIK3 receptors had a pre and postsynaptic localisation. In addition the presynaptic labelling of these two proteins was found also in terminals forming symmetric and asymmetric synapses. Thus, the receptors have been associated to the new GO terms 'asymmetric and symmetric synapse' using the IDA evidence code because the results were inferred from direct assays (immunocytochemistry).

Reading a paper, an annotator usually can annotate more than one protein with more than one GO term. For example, from one paper GRIK2 and GRIK3 have been annotated with the term 'regulation of excitatory postsynaptic membrane potential' (GO:0060079), because it has been demonstrated that these receptors can modulate the EPSC (excitatory postsynaptic current). In addition, KA receptor activation not only inhibited EPSC but also reduced the glutamatergic synaptic transmission. In this case, the term 'negative regulation of synaptic transmission, glutamatergic' (GO:0051967) has been applied for the annotation.

In the second paper (Figure 2.12) immunolabelling experiments showed that GRIA1 was distributed at the postsynaptic density (PSD) and localized at synapses in both wild type and mutant mice. Therefore, GO terms 'postsynaptic density' and 'synapse' have been used to annotate this protein. In this case, the correct evidence code to use was IDA 'inferred from direct assay', and not IMP 'inferred from mutant phenotype', since the results were obtained equally from the wild types and from the mutants. GRIA1 has been also annotated with 'receptor internalisation' (GO:0031623) (IMP) and 'long-term memory' (GO:0007616) (IMP).



Protein annotated GRIA1 (P23818) PMID: 12628184

Cell, Vol. 112, 631–643, March 7, 2003, Copyright ©2003 by Cell Press

Phosphorylation of the AMPA Receptor GluR1 Subunit Is Required for Synaptic Plasticity and Retention of Spatial Memory

Hey-Kyoung Lee,[1,4] Kogo Takamiya,[1,4]
Jung-Soo Han,[2] Hengye Man,[1] Chong-Hyun Kim,[1]
Gavin Rumbaugh,[1] Sandy Yu,[1] Lin Ding,[1] Chun He,[1]
Ronald S. Petralia,[3] Robert J. Wenthold,[3]
Michela Gallagher,[2] and Richard L. Huganir[1,*]
[1]Department of Neuroscience
Howard Hughes Medical Institute
Johns Hopkins School of Medicine
Baltimore, Maryland 21205

- cellular_component
  - synapse part
    - postsynaptic density
  - synapse
    - synapse part
      - postsynaptic density
  - organelle
  - cell

Figure 2. Normal Gross Anatomy and Distribution of GluR1 and GluR2/3 Subunit in the Hippocampus and at Synapses in the Double Phosphomutant Mice

(D) Immunogold labeling of GluR1 in the CA1 stratum radiatum region of the hippocampus of wild-type and mutant mice. Note that GluR1 subunit is distributed at the postsynaptic densities (PSDs) in both wild-type (Wt) and homozygote (−/−). The closed arrowheads point to gold particles localized in PSDs. The narrow arrows point to peri-/extra-synaptic labeling and labeling in spine cytoplasm. The calibration bar is 0.2 μm.

(E) GluR1 is localized at synapses in both wild-type (Wt) and homozygote (−/−) as shown in cultured hippocampal neurons labeled with α-GluR1-N Ab after permeabilization.

GO:0045202 (IDA)
synapse

GO:0014069 (IDA)
post synaptic density (PSD)

Figure 2.12: In this paper the protein GRIA1 (uniprot ID P23818) can be annotated to some GO terms. There is evidence that these AMPA receptors localize to postsynaptic density and synapse (words underlined in red). At the time of the annotation, GO term 'synapse' (GO:0045202) was already implemented in GO. Thus, this term has been used to annotate GRIA1 AMPA receptor. In addition, a new term 'postsynaptic density' (GO:0014069) has been added in GO and used for the annotation (GO DAG graph on the right). In this case, the correct evidence code to use was IDA 'inferred from direct assay', and not IMP 'inferred from mutant phenotype', since the results were obtained equally from the wild types and from the mutants.

Chapter 3

Muscle GO Content Meeting

This chapter presents our effort to improve the representation of muscle biology in the GO in order to support muscle biology research about muscular and neuromuscular diseases. The first paragraph briefly introduces the content-oriented meetings and the reasons for their organisation. Then, the second part of the chapter is about the method and the results of our Muscle GO Content Meeting and the resources made available for the annotation of genes involved in muscle processes. All new GO terms as well as modified GO terms are listed in Appendices B.

3.1 Introduction

The Gene Ontology has long included a number of terms describing processes, functions and cellular components related to muscle biology, and the GO system has already been used extensively for statistical data analysis for these studies. For example, an analysis of the global transcriptional changes that take place in skeletal muscle in relation to estrogen status [52] used GO annotations to define significant gene sets, and there has also been an expression profiling study of MyoD during myogenic differentiation to improve the understanding of skeletal muscle development [53].

However, particularly in the biological process ontology, the existing terms failed to cover the breadth of known muscular processes, and in many cases diverged from current usage and understanding in the field, as these terms were mainly created by non-muscle experts using older references. Thus, to fully support current muscle community research needs, especially with respect to the study of muscular diseases, a considerable revision of the terms within the GO was required.

For example in muscle biology there is some ambiguity in the use of the word *plasticity.* This word could be used to mean '*the ability of a muscle cell to change and adapt*', but is often used in the literature to mean the '*process of adaptation*'. That is to say, in one case the term refers to the potential inherent in a muscle cell, and in the other, the biological process itself. In addition to clarity for scientists, computer software must encode the relationship between plasticity and other important processes like atrophy, hypertrophy and hyperplasia. It is difficult and frustrating for a scientist to sort out the meaning of biological language when the same undefined word is used for two different concepts, and this task is impossible for a computer. This is especially important where high-throughput processing work by computation analysis is required.

Here is described an effort to improve the structure of muscle terms in the process and component ontologies. The work was carried out in a collaborative project that brought together the GO Consortium and the Genomics Research Group of CRIBI Biotechnology Center[1] at the University of Padua. The aim was to improve GO terms that would specifically support muscle biology research in areas relevant to the investigation of muscular and neuromuscular diseases.

3.1.1 GO Content Meeting

Since GO people are very interested in improving the GO in several areas of specific interest, they often organize content-oriented meeting called GO Content Meetings in which together GO members and a few domain experts usually discuss about GO categories concerning the focused domain. In 2004 the GO consortium organized the GO Immunology Content meeting held at the Institute for Genomic Research (TIGR), and in 2006, they organised the GO Neurobiology Content Meeting held at MGI at Bar Harbor (Maine, US). In both cases more than 500 new terms were added. I was actively involved in the CNS meeting.

Usually GO content meetings are short and involve small numbers of people, who are mostly

[1] http://genomics.cribi.unipd.it/

few GO curators and invited experts. Before the meeting, GO curators spend several months reading papers and make most of the terms in advance. Once the proposed structure has been discussed during the meeting, further refinements are made by email, and changes are made once consensus is reached. These content oriented meetings facilitate large-scale changes in specific areas of the ontology [29]. This approach has some advantages because it allows a lot of detailed work to be done on a very specific area and involves valuable external expertise. However, there are few problems that must be considered: the organisation is very expensive, the time to get terms into ontologies is long and finally terms added at these sessions may not have as much annotation as older terms but they are simply newer. But, on the other hand, this process is quite quick and cheap compared to the contribute of a single curator working alone.

The method adopted for the preparation of the GO preliminary draft is different from each meeting and dependent on the focused biological domain. For instance for the immunology meeting, a combined method of ‘top-down’ and ‘bottom-up’ approaches was used. In the top-down approach, authoritative textbooks and many current reviews of specific subject areas within immunology were consulted as a basis for providing a set of redefined and new high level terms to match their usage in the literature. The bottom-up approach involved identifing missing terms while doing annotation of gene products, and adding these terms to the GO biological process ontology; this is otherwise known as *annotation driven ontology development.* Most of these terms concerned lower-level processes of the immune system [30].

However, this Muscle Content Meeting was slightly different from the previous model because it introduced an alternative method.

3.2 Method

Following the example of the Immunology Content Meeting and of other GO ontology development meetings, the Muscle GO Content Meeting brought together experimental biologists and ontology developers to capture the details of specific research field. The work focused on the biology of skeletal and smooth muscle and was held at the University of Padua, Italy.

As the usual model, there was a preliminary phase during which we reviewed current knowledge of the specific chosen domain and prepared a draft of a revised GO graph structure.

Following this initial task, domain experts were invited to come together and reviewed the changes in a two day discussion and editing meeting.

The Muscle Content Meeting was similar and yet distinct from the previous ontology development workshops. The principle innovation in this meeting was that, I as a member of the muscle research community, was trained within the GO consortium for six months to have a good knowledge of ontology development and annotation. I then rejoined my research community and worked with them to make the initial round of edits to the subgraph of muscle biology. This process was supported by ontology experts in the GO Consortium via wiki, VoIP and screen sharing technologies to facilitate discussion of problems. Furthermore, wiki pages have been developed for meeting participants: one page[2] with links to all the information about the meeting such as the agenda and the implementation plan, and an other page[3] with information about helpful resources for browsing and editing the GO. A OBO-Edit tutorial written in Italian was also provided[4]. After this initial editing phase, ontology developers from the GO Consortium were invited to meet with the group of domain experts for further discussion and revision of the GO graph. The changes were then discussed over a period of two days, with live editing to make any further changes.

This novel community-based ontology update model was extremely beneficial as it allowed the ontology developer to access a wide range of domain experts, all locally available, and all with cutting edge knowledge of the field. This ontology development process, directed by a scientist working in a specific research field as the key ontology developer, was a new alternative to having ontology experts access knowledge through reviews with later checking of their understanding with the domain experts. In keeping with the usual model, the final ontology development meeting enabled the new and existing terms to be thoroughly reviewed by domain experts. By bringing together muscle and ontology experts, the participants were able to systematically improve the GO structure related to specific muscle biology areas. This system was an innovation for the GO consortium and provided an immensely valuable resource, quickly and cheaply. It also provided the all important buy-in of the gross roots biology community that is so essential for the proper functioning of an ontology-based information

[2] http://wiki.geneontology.org/index.php/Muscle_Development

[3] http://wiki.geneontology.org/index.php/Content_Meeting_Participants_Information

[4] http://www.geneontology.org/teaching_resources/tutorials/2007_08_Italian_o boedit_tutorial.pdf

system. This innovation is one of the main novel outcomes of this PhD project and is likely to be regularly repeated by GO Consortium.

For inter-ontology consistency, the new structure drew on, and cross-referenced with the existing structure of other ontologies. The Adult Mouse Anatomical Dictionary (AMA)[5] was used as a source of definitions and structures for muscle contraction anatomical terms. The Adult Mouse Anatomical Dictionary is an ontology providing a standardized nomenclature for anatomical terms in the postnatal mouse. It is structured as a directed acyclic graph, and is organized hierarchically both spatially and functionally [54]. In addition, the Cell Type ontology (CL)[6] was consulted for cell type definitions and the resulting terms were cross-referenced. The Cell Type ontology is an OBO ontology for prokaryotic, fungal, animal and plant cell types that are classified under several generic categories. It is organized as a directed acyclic graph and it is designed to be used in the context of model organism genome and other biological databases [55]. Where appropriate, new definitions were contributed to the Cell Type ontology. For example a new definition of the satellite cell type was created and introduced in both cell type and GO ontologies, and the resulting terms were cross-referenced.

The revised structure for the muscle biology subgraph was sent to the entire GO community for further discussion. At the end of October 2007, it was merged into the public GO file and is now available for all GO users.

3.3 Results and discussion

The muscle research community, in collaboration with the GO consortium, has completed an initiative to greatly expand muscle biology representation in the GO biological process and cellular component ontologies.

The main focus of the work was skeletal and smooth muscle, with specific consideration given to the processes of *muscle contraction*, *muscle plasticity*, *muscle development* and *regeneration*; and to the *sarcomere* and *membrane delimited compartments.* The aims were to update the existing structure to reflect current knowledge, and to resolve in an accommodating manner the ambiguity in the language of the community. This collaborative effort drew on the

[5] http://www.informatics.jax.org/

[6] https://lists.sourceforge.net/lists/listinfo/obo-celltype

expertise of an extensive community of muscle experts, and resulted in the addition of **159 new terms** and the improvement of **57 existing terms**.

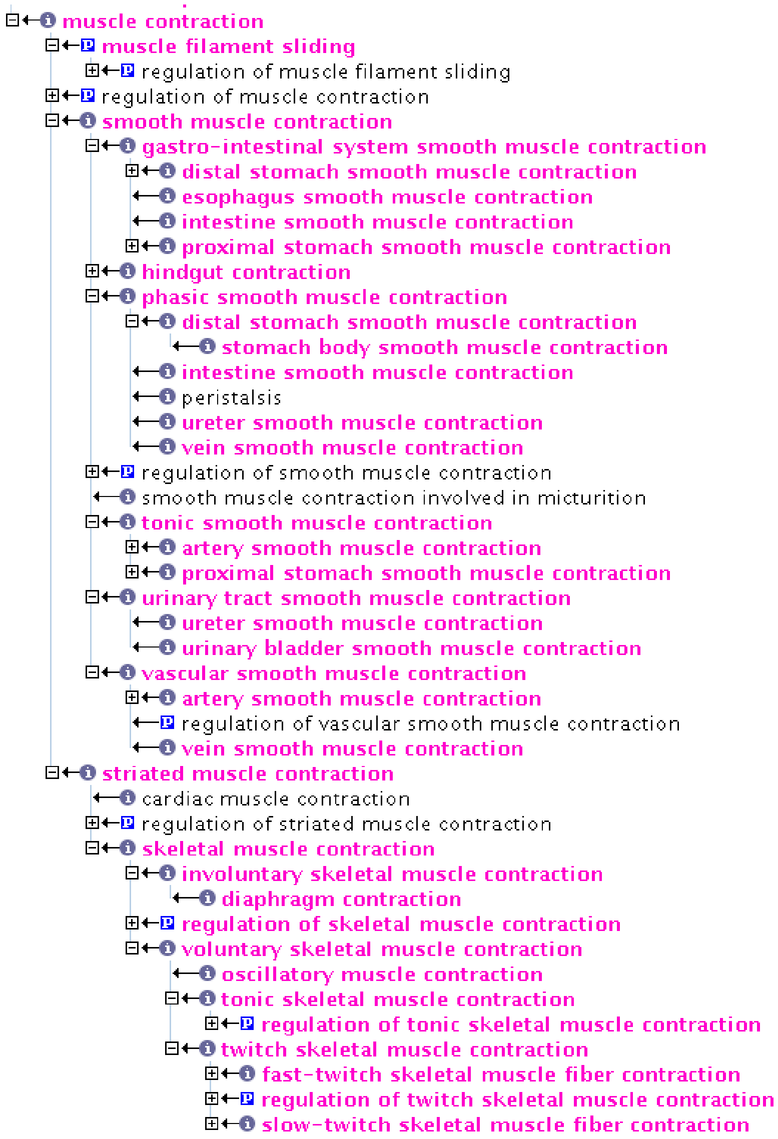
Muscles can be divided into striated and smooth types. Smooth muscle or ‘involuntary muscle’ is found within structures such as the oesophagus, stomach, intestines, bronchi, uterus and blood vessels; and unlike skeletal muscle, it is not under conscious control. Cardiac and skeletal muscles are ‘striated’ in that they contain sarcomeres and are packed into highly-regular arrangements of bundles. Skeletal muscle is further divided into two subtypes: slow-twitch and fast-twitch muscle, depending on their contractile capacity. The biology of these two muscle types is key in current research and so there is a need to correctly represent it as part of the biological process ontology. Importantly these terms were also cross-checked by a cardiovascular physiology community group[7], whose ontology development meeting took place at about same time, and which also touched on the voluntary/involuntary muscle processes.

3.3.1 Muscle contraction

Muscle contraction and muscle plasticity, both child terms of the high level term ‘muscle system process’, have been revised and the hierarchies expanded. Prior to the meeting, with the help of professor Carlo Reggiani, a muscle physiologist with a vast knowledge in muscle contraction, a revised structure on the biological process of muscle contraction was prepared. The new DAG structure contained GO terms about different kinds of muscle contraction (e.g. phasic and rhythmic) but also several regulation terms.

Since the meeting, the definition of the term ‘muscle contraction’, which previously existed in the GO, has been considerably improved and all descendents have been reorganized. The new structure represents several forms of muscle contraction and their inter-relationships with the various types of muscle. To reflect this, there is also a greatly expanded set of terms describing the different contractile capacity of muscle. Striated muscle contracts and relaxes in short, intense bursts, whereas smooth muscle sustains longer or even near-permanent contractions. This difference is captured by the creation of *is_a* children ‘phasic smooth muscle contraction’ and ‘tonic smooth muscle contraction’ under the parent term ‘smooth muscle contraction’. Since the process of smooth muscle contraction varies with the anatomical location of the

[7] http://wiki.geneontology.org/index.php/Cardiovascular

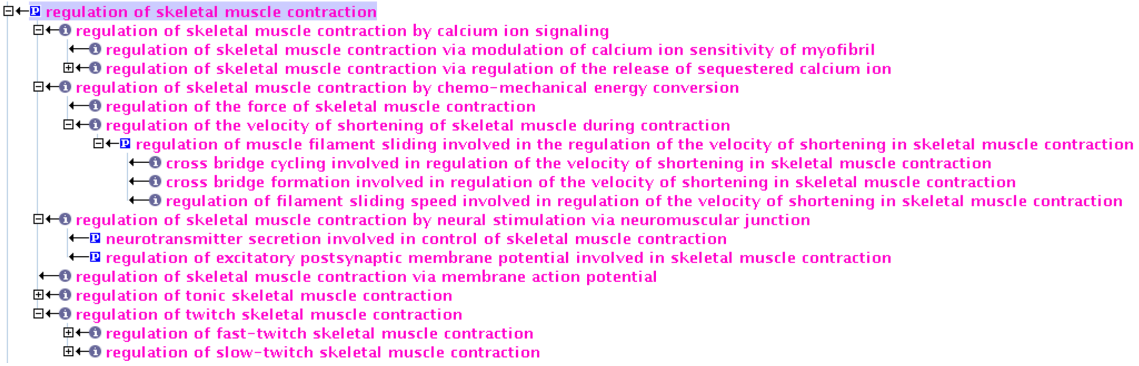


- ⊟ muscle contraction
  - ⊟ muscle filament sliding
    - ⊞ regulation of muscle filament sliding
  - ⊞ regulation of muscle contraction
  - ⊟ smooth muscle contraction
    - ⊟ gastro-intestinal system smooth muscle contraction
      - ⊞ distal stomach smooth muscle contraction
      - esophagus smooth muscle contraction
      - intestine smooth muscle contraction
      - ⊞ proximal stomach smooth muscle contraction
    - ⊞ hindgut contraction
    - ⊟ phasic smooth muscle contraction
      - ⊟ distal stomach smooth muscle contraction
        - stomach body smooth muscle contraction
      - intestine smooth muscle contraction
      - peristalsis
      - ureter smooth muscle contraction
      - vein smooth muscle contraction
    - ⊞ regulation of smooth muscle contraction
    - smooth muscle contraction involved in micturition
    - ⊟ tonic smooth muscle contraction
      - ⊞ artery smooth muscle contraction
      - ⊞ proximal stomach smooth muscle contraction
    - ⊟ urinary tract smooth muscle contraction
      - ureter smooth muscle contraction
      - urinary bladder smooth muscle contraction
    - ⊟ vascular smooth muscle contraction
      - ⊞ artery smooth muscle contraction
      - regulation of vascular smooth muscle contraction
      - vein smooth muscle contraction
  - ⊟ striated muscle contraction
    - cardiac muscle contraction
    - ⊞ regulation of striated muscle contraction
    - ⊟ skeletal muscle contraction
      - ⊟ involuntary skeletal muscle contraction
        - diaphragm contraction
      - ⊞ regulation of skeletal muscle contraction
      - ⊟ voluntary skeletal muscle contraction
        - oscillatory muscle contraction
        - ⊟ tonic skeletal muscle contraction
          - ⊞ regulation of tonic skeletal muscle contraction
        - ⊟ twitch skeletal muscle contraction
          - ⊞ fast-twitch skeletal muscle fiber contraction
          - ⊞ regulation of twitch skeletal muscle contraction
          - ⊞ slow-twitch skeletal muscle fiber contraction

Figure 3.1: New muscle contraction GO node (new terms and modified terms are pink-coloured). The new structure represents several forms of muscle contraction and their interrelationships with the various types of muscle. There are terms like ‘phasic smooth muscle contraction’ and ‘tonic smooth muscle contraction’ that help in the grouping of several kind of muscle contraction such as ‘vascular smooth muscle contraction’ or ‘gastro-intestinal system smooth muscle contraction’.

muscles, terms such as ‘vascular muscle contraction’ and ‘gastrointestinal muscle contraction’ were also created (Figure 3.1).

Muscle contraction is actively regulated by a series of events, and so the appropriate regulation terms have been added. These include several processes such as cross-bridge formation, cross-bridge cycling, and filament sliding, that are necessary for force generation during muscle contraction (Figure 3.2). Multiple molecular components such as sarcoplasmic proteins have a role in regulating the muscle contraction. For instance mutation in several Z-disc proteins in the sarcomere, which is important for the cross-linking of thin filaments and transmission of force generated by the myofilaments, have been shown to cause cardiomyopathies and/or muscular dystrophies [56].



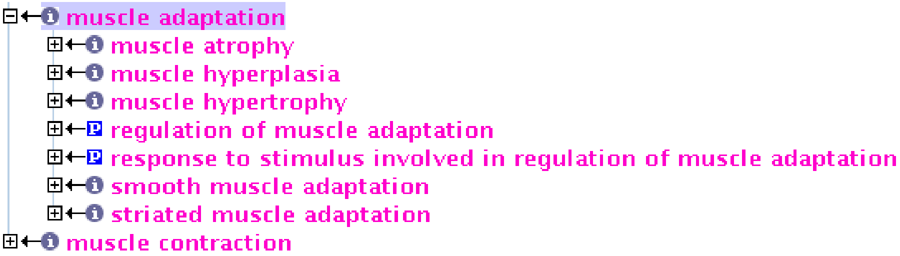
```
⊟ regulation of skeletal muscle contraction
  ⊟ regulation of skeletal muscle contraction by calcium ion signaling
      regulation of skeletal muscle contraction via modulation of calcium ion sensitivity of myofibril
    ⊞ regulation of skeletal muscle contraction via regulation of the release of sequestered calcium ion
  ⊟ regulation of skeletal muscle contraction by chemo-mechanical energy conversion
      regulation of the force of skeletal muscle contraction
    ⊟ regulation of the velocity of shortening of skeletal muscle during contraction
      ⊟ regulation of muscle filament sliding involved in the regulation of the velocity of shortening in skeletal muscle contraction
          cross bridge cycling involved in regulation of the velocity of shortening in skeletal muscle contraction
          cross bridge formation involved in regulation of the velocity of shortening in skeletal muscle contraction
          regulation of filament sliding speed involved in regulation of the velocity of shortening in skeletal muscle contraction
  ⊟ regulation of skeletal muscle contraction by neural stimulation via neuromuscular junction
      neurotransmitter secretion involved in control of skeletal muscle contraction
      regulation of excitatory postsynaptic membrane potential involved in skeletal muscle contraction
    regulation of skeletal muscle contraction via membrane action potential
  ⊞ regulation of tonic skeletal muscle contraction
  ⊟ regulation of twitch skeletal muscle contraction
    ⊞ regulation of fast-twitch skeletal muscle contraction
    ⊞ regulation of slow-twitch skeletal muscle contraction
```

Figure 3.2: New regulation terms for muscle contraction (new terms and modified terms are pink-coloured). Muscle contraction is actively regulated by a series of events, including processes such as cross-bridge formation, cross-bridge cycling, and filament sliding, that are necessary for force generation during muscle contraction. In order to cover these events, new regulation terms have been added such as ‘regulation of velocity of shortening of skeletal muscle during contraction’ and ‘regulation of muscle filament sliding involved in the regulation of the velocity of shortening’.

3.3.2 Muscle plasticity

The term ‘muscle plasticity’ has been renamed ‘muscle adaptation’, and the definition improved to resolve ambiguity in the meaning of the word plasticity. There are two different possible meanings of this word, such that plasticity could be either the quality of adaptability, or the process of adaptation. However the critical thing in ontology development is to be clear about which term represents which process, and to ensure that the language is unambiguous;

whilst still reflecting community usage. As the existing term was defined to describe the process of change, it was renamed ‘muscle adaptation’ (leaving ‘muscle plasticity’ as a synonym, to ensure that researchers can still find the term) and it now groups terms describing various forms of muscle adaptation (Figure 3.3).

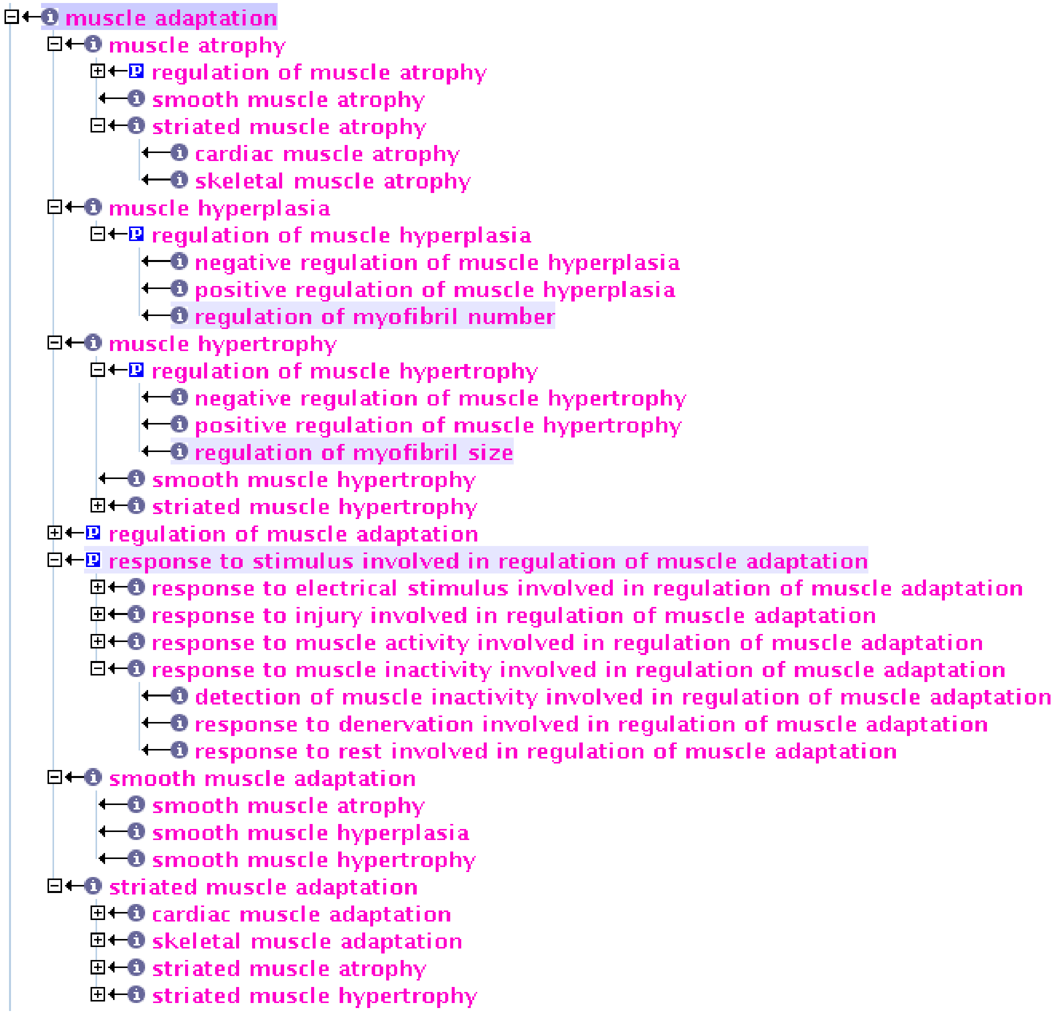


- muscle adaptation
  - muscle atrophy
  - muscle hyperplasia
  - muscle hypertrophy
  - regulation of muscle adaptation
  - response to stimulus involved in regulation of muscle adaptation
  - smooth muscle adaptation
  - striated muscle adaptation
- muscle contraction

Figure 3.3: New ‘muscle plasticity’ GO node (new terms and modified terms are pink-coloured). The term ‘muscle plasticity’ has been renamed ‘muscle adaptation’ and the whole GO category has been re-organised to cover other adaptive processes such as muscle atrophy, hypertrophy and hyperplasia.

There are many stimuli that bring about muscle adaptation. Musculo-skeletal adaptability studies include examination of a muscle’s response to joint immobilization, spinal cord injury, electrical stimulation, chronic stretch, exercise-induced injury and microgravity. These factors are now also accommodated in the ontology structure. Such adaptive events occur in muscle fibers and associated structures (motoneurons and capillaries); and they involve alterations in regulatory mechanisms, contractile properties, fiber-type compositions, and in metabolic capacities. These adaptive processes include atrophy, hypertrophy and hyperplasia. Terms covering these processes have been included as children of ‘muscle adaptation’ (Figure 3.4).

This set of terms will help in the annotation of gene products involved in the control of muscle fiber-type diversity, providing potential new targets for the treatment and prevention of different disorders ranging from metabolic to neuromuscular diseases, for example Type 2 diabetes and muscular dystrophy [57].

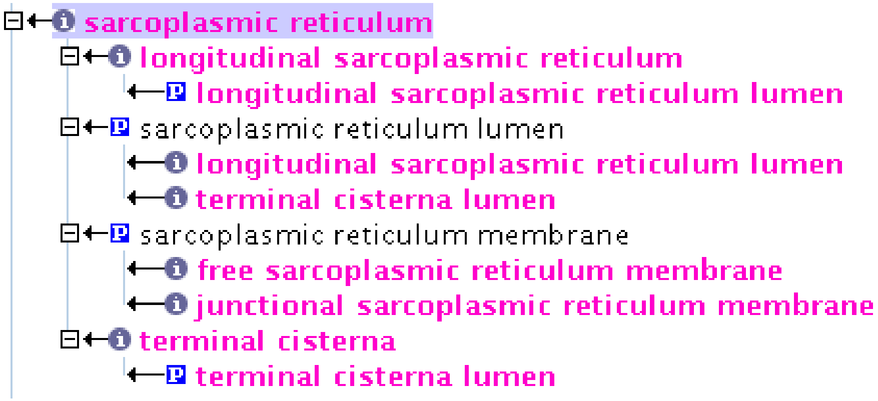


- muscle adaptation
  - muscle atrophy
    - regulation of muscle atrophy
    - smooth muscle atrophy
    - striated muscle atrophy
      - cardiac muscle atrophy
      - skeletal muscle atrophy
  - muscle hyperplasia
    - regulation of muscle hyperplasia
      - negative regulation of muscle hyperplasia
      - positive regulation of muscle hyperplasia
      - regulation of myofibril number
  - muscle hypertrophy
    - regulation of muscle hypertrophy
      - negative regulation of muscle hypertrophy
      - positive regulation of muscle hypertrophy
      - regulation of myofibril size
    - smooth muscle hypertrophy
    - striated muscle hypertrophy
  - regulation of muscle adaptation
  - response to stimulus involved in regulation of muscle adaptation
    - response to electrical stimulus involved in regulation of muscle adaptation
    - response to injury involved in regulation of muscle adaptation
    - response to muscle activity involved in regulation of muscle adaptation
    - response to muscle inactivity involved in regulation of muscle adaptation
      - detection of muscle inactivity involved in regulation of muscle adaptation
      - response to denervation involved in regulation of muscle adaptation
      - response to rest involved in regulation of muscle adaptation
  - smooth muscle adaptation
    - smooth muscle atrophy
    - smooth muscle hyperplasia
    - smooth muscle hypertrophy
  - striated muscle adaptation
    - cardiac muscle adaptation
    - skeletal muscle adaptation
    - striated muscle atrophy
    - striated muscle hypertrophy

Figure 3.4: Expanded ‘muscle adaptation’ GO node (new terms and modified terms are pink-coloured). There are many stimuli that bring about muscle adaptation and they involve alterations in regulatory mechanisms, contractile properties, fiber-type compositions, and in metabolic capacities. Terms covering these processes have been included as children of ‘muscle adaptation’; they include ‘regulation of myofibril type’, ‘regulation of myofibril number’ and ‘response to stimulus involved in regulation of muscle adaptation’.

3.3.3 Sarcomere and its role in regulating the calcium ion dependent processes

Muscle plasticity is closely linked with, and highly dependent on, the calcium handling system; as muscles use calcium ions as their main regulatory and signalling molecule. Therefore calcium ion-dependent processes control the properties of mechanisms of contraction and relaxation in different types of muscle fibres [58]. The sarcoplasmic reticulum (SR) is a subcompartment of the endoplasmic reticulum (ER) and it is molecularly specialized for calcium release, uptake and storage and for the contraction-relaxation cycle in skeletal muscle fibres [59]. Recognizing the importance of this, part of the work was focused on improving the existing terms describing 'sarcoplasmic reticulum' and its role in regulating the calcium ion dependent processes. Terms such 'regulation of skeletal muscle contraction by calcium ion signaling' and 'regulation of skeletal muscle contraction via modulation of calcium ion sensitivity of myofibril' were added as *part_of* children of muscle contraction. In addition, the sarcoplasmic reticulum compartment and its components are covered by a new hierarchy of terms including longitudinal sarcoplasmic reticulum, terminal cisternae, free sarcoplasmic reticulum membrane and the junctional sarcoplasmic reticulum membrane (Figure 3.5).



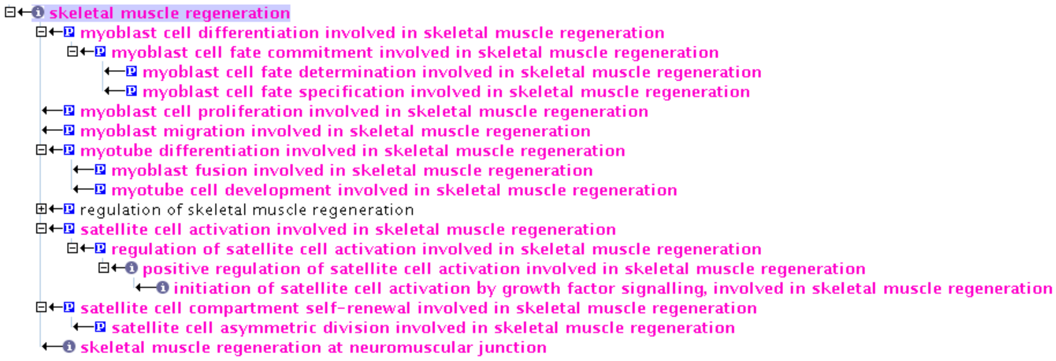
```
sarcoplasmic reticulum
    longitudinal sarcoplasmic reticulum
        longitudinal sarcoplasmic reticulum lumen
    sarcoplasmic reticulum lumen
        longitudinal sarcoplasmic reticulum lumen
        terminal cisterna lumen
    sarcoplasmic reticulum membrane
        free sarcoplasmic reticulum membrane
        junctional sarcoplasmic reticulum membrane
    terminal cisterna
        terminal cisterna lumen
```

Figure 3.5: The new hierarchy of terms covering the sarcoplasmic reticulum compartment and its components (new terms and modified terms are pink-coloured).

These new GO terms will allow additional muscle-specific gene associations and thus will aid our understanding of normal muscle processes and muscle pathological conditions such as dystrophinopathies, Brody's disease and malignant hyperthermia that have been shown to be due to alterations in calcium ion-dependent channel activities [58]. In addition, it is well established that calcium regulates other signalling pathways critical for myogenesis, muscle remodelling, and regeneration.

3.3.4 Muscle development and regeneration

Myofibers, the functional unit of skeletal muscle, are long cylindrical multinucleated cells that vary in their morphological, biochemical, and physiological properties. They are derived from myoblasts: cells committed to the skeletal muscle lineage. Upon fusion, myoblasts form myotubes, which are further remodeled into myofibers [60]. The 'skeletal muscle development' subtree is now covered by a new hierarchy of terms describing myoblast, myotube and myofiber development, and the mechanisms of regulation. Many terms have been added to cover the process of cell regeneration, and its regulation in skeletal muscle tissue (Figure 3.6).



```
⊟←i skeletal muscle regeneration
    ⊟←P myoblast cell differentiation involved in skeletal muscle regeneration
        ⊟←P myoblast cell fate commitment involved in skeletal muscle regeneration
            ←P myoblast cell fate determination involved in skeletal muscle regeneration
            ←P myoblast cell fate specification involved in skeletal muscle regeneration
    ←P myoblast cell proliferation involved in skeletal muscle regeneration
    ←P myoblast migration involved in skeletal muscle regeneration
    ⊟←P myotube differentiation involved in skeletal muscle regeneration
        ←P myoblast fusion involved in skeletal muscle regeneration
        ←P myotube cell development involved in skeletal muscle regeneration
    ⊞←P regulation of skeletal muscle regeneration
    ⊟←P satellite cell activation involved in skeletal muscle regeneration
        ⊟←P regulation of satellite cell activation involved in skeletal muscle regeneration
            ⊟←i positive regulation of satellite cell activation involved in skeletal muscle regeneration
                ←i initiation of satellite cell activation by growth factor signalling, involved in skeletal muscle regeneration
    ⊟←P satellite cell compartment self-renewal involved in skeletal muscle regeneration
        ←P satellite cell asymmetric division involved in skeletal muscle regeneration
    ←i skeletal muscle regeneration at neuromuscular junction
```

Figure 3.6: Improved representation of muscle regeneration process in the new GO hierarchy (new terms and modified terms are pink-coloured).

Satellite cell processes are considered particularly important, since their activation is involved in muscle regeneration. Satellite cell proliferation, differentiation and self-renewal are essential for proper myofiber turnover, an ongoing process that maintains proper muscle tissue viability [61]. Moreover, in adult skeletal muscle, the self-renewing capacity of satellite cells contributes to muscle growth and adaptation [62]. Skeletal muscle is capable of complete regeneration due to the presence of stem cells that reside in skeletal muscle and non-muscle stem cell populations. However, in severe myopathic diseases such as Duchenne Muscular Dystrophy, this regenerative capacity is exhausted [63].

All new muscle GO terms as well as all GO modified terms are listed in Appendix B.

3.3.5 Muscle Biology Community Annotation

This ontology development effort provides a valuable resource for annotation of gene products related to muscle biology. New terms supporting critical research areas are now available, and existing terms have been improved and reorganized to follow their usage in muscle literature. There are a number of important advantages to a research community in having their field accurately represented in the GO. The revised ontology structure facilitates the interpretation of high throughput experiments (e.g. gene expression microarrays) in the area of muscle science and muscular disease. Such studies yield a very large number of data points, making it a challenge to determine which genes specifically contribute to a disease phenotype [64].

However the use of GO ontologies and annotations greatly simplifies this analysis using GO-related statistical analysis tools[8]. Obviously, a critical component of such analysis is the comprehensive annotation of relevant gene products. To enable community annotation, a Muscle Biology Community Annotation Wiki[9] has been provided, which will complement the ontology development work carried out.

Using this page, researchers can provide data and annotation for gene products relevant to muscle biology research. The wiki contains working annotation pages for **172 genes** associated with muscle development and function. Users can review the Gene Ontology annotations for any gene of interest, and they can input information about any aspect of the biology of a gene from any species. Finally, creation of this tool will facilitate community participation in the Gene Ontology project and speed up the annotation process.

[8] http://www.geneontology.org/GO.tools.shtml#micro

[9] http://wiki.geneontology.org/index.php/Muscle_Biology

Chapter 4

Representation of ‘drugs’ in GO

This chapter provides a detailed description of the method used to re-organise and re-define the response to drug (GO:0042493) node. The chapter starts with a short description of the problem, followed by the explanation of the method applied for each case study. The last section is about the results obtained and the future work.

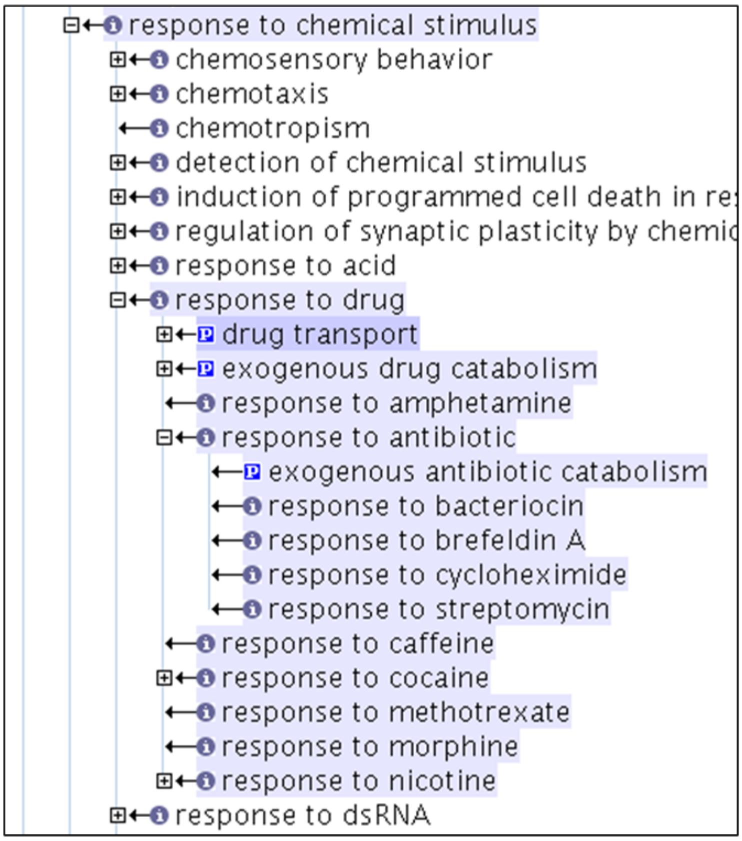
4.1 Introduction to the problem

Before introducing the following chapter, it is important to state again the scope of GO, and what it does and does not cover. The fundamental principle is that the GO vocabularies must describe the roles of a gene products in cells in their normal environment. The vocabularies do not attempt to represent every aspect of biology; therefore areas that are outside the scope of GO, and terms in these domains would not appear in the ontologies. An example is the case of processes, functions or components that are unique to mutants or diseases: e.g. oncogenesis is not a valid GO term because causing cancer is not the normal function of any gene[1].

At the beginning of PhD project, the GO process ontology included a number of terms grouped under ‘response to drug’ (GO:0042493). These GO terms included ‘response to caffeine’, ‘response to cocaine’ and ‘response to antibiotic’ and so on (Figure 4.1). The definition of ‘response to drug’ term was: ‘a change in state or activity of a cell or an organism (in terms of movement, secretion, enzyme production, gene expression, etc.) as a result of a drug

[1] http://www.geneontology.org/GO.doc.shtml#not

stimulus. A drug is a substance used in the diagnosis, treatment or prevention of a disease’. According to the definition above, a drug has to be consider with reference only to a disease; but this is not completely correct for ‘caffeine’. It is a plant alkaloid, found in numerous plant species, where it acts as a natural pesticide; but it can have many effects also on the mammals metabolism, including stimulating the central nervous system and in this case it has not to be seen as a drug.



```
⊟ response to chemical stimulus
    ⊞ chemosensory behavior
    ⊞ chemotaxis
      chemotropism
    ⊞ detection of chemical stimulus
    ⊞ induction of programmed cell death in re
    ⊞ regulation of synaptic plasticity by chemic
    ⊞ response to acid
    ⊟ response to drug
        ⊞ drug transport
        ⊞ exogenous drug catabolism
          response to amphetamine
        ⊟ response to antibiotic
              exogenous antibiotic catabolism
              response to bacteriocin
              response to brefeldin A
              response to cycloheximide
              response to streptomycin
          response to caffeine
        ⊞ response to cocaine
          response to methotrexate
          response to morphine
        ⊞ response to nicotine
    ⊞ response to dsRNA
```

Figure 4.1: ‘response to drug’ node before any changes in its structure and content. ‘response to drug’ term and the descendants refer to substances very different one to each other for their structures. However, they have only one characteristic in common: they might act as ‘drugs’.

In this context, there were some reservations about these terms since they are clearly very context dependent. Although agents, known as drugs, occur naturally, many of them are substances created by humans and in some cases do not bear any structural relationship to any naturally occurring substances. Moreover there are cases in which a drug to some species (e.g. humans) is not a drug to another species (e.g. bacteria, plants). Antibiotics are a perfect example of this ambiguity. In fact, an antibiotic might be consider a ‘drug’ when it is used as a chemotherapeutic agent in the treatment of a variety of infections in humans and it might be able to inhibit or abolish the growth of microorganisms, such as bacteria, fungi, or protozoans. However, the bacteria have responses to the antibiotics in their physiological environments where these agents would not act as a drug. In addition, it might be difficult to state that a gene product has evolved specifically to respond to a certain drug. Moreover, as already clarified, GO is used to annotate the ‘normal’ processes of gene products and obviously there were reservations whether responding to a drug represents a ‘normal’ process. Thus, the word ‘drug’ was considered too arbitrary and the classification of the ‘response to drug’ terms was often inconsistent and unreliable and was not enough appropriate for the GO purpose.

Nevertheless, many users of the GO find the terms under the ‘response to drug’ GO node useful and they would have appreciated an expanded hierarchy of this group of terms. Furthermore, these terms are currently of interest to pharmaceutical researchers for the annotation of gene products that respond to the administration of drugs. In a drug discovery process, following a gene expression experiment involving high-throughput microarrays, a biomedical scientist is confronted with a list of a few hundred to thousands genes, from which the researcher will need to extract useful information on the types of biological processes affected the experiments [65]. Therefore, a better reorganization of the content and structure of ‘response to drug’ node will help in the association of genes that are differentially express in response to a drug treatment and consequently, the set of new annotations will provide additional links between gene expression profiles and effects of a drug in regulating specific biological processes. In this case, for example, it has been shown that there are alterations in neuronal gene expression following single injection of a neurotropic drug (e.g morphine). In particular, this drug alters expression of two major groups of genes; those for proteins involved in mitochondrial respiration and those for cytoskeleton-related proteins [66].

4.2 Method

4.2.1 Set up of the ‘response to drug’ interest group

The Gene Ontology Consortium has created several interest groups organised by topics that roughly correspond to GO terms, usually high level term such as ‘transport’ or ‘RNA metabolism’. The aim is to facilitate discussions of areas within the ontologies that require extensive additions or revisions, or where proposed changes appear frequently. The groups are listed at GO Curator Interest Groups web page[2].

In order to solve problems relating to the misuse of the word ‘drug’, a ‘response to drug’ working group has been set up. I am responsible for this interest group and at the moment it has 9 members from several research groups such as MGI and EBI (all members are listed on the the Response to Drug Interest Group GO page[3]). Clearly anyone that is interested in this topic might join the group by writing a request to the GO interest group mailing list. Moreover, the very active interest groups have their own mailing lists and our group has its mailing list: response-to-drug@geneontology.org. A mail archive is also available to assure that discussions occurring on the mailing list are available at any time and can be consulted for reassessing decisions taken in the past. From the beginning, the group members have been handling discussions about what might be the best way to develop and improve the ‘response to drug’ node and what might be the possible content and structure changes to overcome the obstacle about ambiguity.

4.3 Evaluation of the available chemistry resources

Firstly, the main objective was to select a resource that could help in avoiding ambiguity at all costs and secondly, to find a method for the standardisation of language used for representing drugs in GO without breaking the ontological rules. Several chemistry resources have been considered and compared for choosing the best one that could provide a standard terminology to refer to the drugs of interest. The long-established chemical databases such

[2] http://www.geneontology.org/GO.interests.shtml
[3] http://www.geneontology.org/GO.interests.shtml#response

as Chemical Abstracts Service (CAS) Registry database[4] or Beilstein[5], as well as the more recent data repository PubChem[6], were excluded because they are a very large collection of data, characterised by high-redundancy (many entities are present more than one), by absence of ontological relationships and by ambiguity (the same ID may refer to more than one substance). The DRUG collection at the KEGG LIGAND database[7] is a new addition in the KEGG database (December 2005) providing chemical structures and additional information such as therapeutic categories and target molecules [67] but this resource is limited to the context of metabolic pathways and protein ligands. Thus the search was addressed to a chemical controlled vocabulary using a consistent and widely recognised terminology in combination with a unique identifier to refer to the molecule of interest [68]. A viable solution is ChEBI (Chemical Entities of Biological Interest) ontology that, containing both systematic and common names, provides a means for placing entities of interest into a wider chemical, biological or medical context.

4.4 ChEBI database and ontology

Chemical Entities of Biological Interest (ChEBI)[8] is a freely available high quality, thoroughly annotated controlled vocabulary, to promote the correct and consistent use of unambiguous biochemical terminology throughout the molecular biology databases [69]. Any chemical compound naturally occurring in living organisms can be called a biochemical compound and be classified according to structure, physico-chemical properties or biological function.

4.4.1 ChEBI structure

ChEBI is an OBO ontology of molecular entities focused on ‘small’ chemical compounds. The molecular entities in question are either natural products or synthetic products used to intervene in the processes of living organisms. The scope of ChEBI encompasses not only ‘biochemical compounds’ but also pharmaceuticals, agrochemicals, laboratory reagents and

[4] http://www.cas.org/EO/regsys.html

[5] http://beilstein.com/

[6] http://pubchem.ncbi.nlh.nih.gov/

[7] http://www.genome.jp/kegg/

[8] http://www.ebi.ac.uk/chebi/

subatomic particles. The ontology employs nomenclature and terminology recommended by international bodies such as International Union of Pure and Applied Chemistry (IUPAC)[9] and Nomenclature Committee of the International Union of Biochemistry and Molecular Biology (NCIUBMB)[10]. Data from a number of sources are incorporated, and merged and cross-referenced in ChEBI to eliminate redundancy, and at the same time data have been enriched by information from manual annotation.

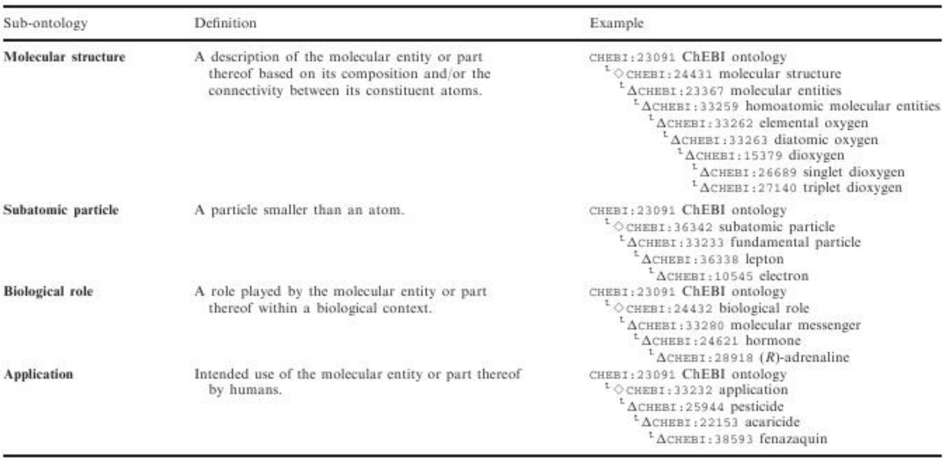
ChEBI ontology consists of four sub-ontologies (Figure 4.2):

- Molecular Structure, in which molecular entities or parts of them are classified according to structure;

- Biological Role, which classifies entities on the basis of their role within a biological context (e.g. antibiotic, coenzyme, hormone);

- Application, which classifies entities, where appropriate, on the basis of their intended use by humans (e.g. pesticide, drug, fuel);

- Subatomic Particle, which classifies particles smaller than atoms.



Sub-ontology	Definition	Example
Molecular structure	A description of the molecular entity or part thereof based on its composition and/or the connectivity between its constituent atoms.	CHEBI:23091 ChEBI ontology └◇CHEBI:24431 molecular structure └ΔCHEBI:23367 molecular entities └ΔCHEBI:33259 homoatomic molecular entities └ΔCHEBI:33262 elemental oxygen └ΔCHEBI:33263 diatomic oxygen └ΔCHEBI:15379 dioxygen └ΔCHEBI:26689 singlet dioxygen └ΔCHEBI:27140 triplet dioxygen
Subatomic particle	A particle smaller than an atom.	CHEBI:23091 ChEBI ontology └◇CHEBI:36342 subatomic particle └ΔCHEBI:33233 fundamental particle └ΔCHEBI:36338 lepton └ΔCHEBI:10545 electron
Biological role	A role played by the molecular entity or part thereof within a biological context.	CHEBI:23091 ChEBI ontology └◇CHEBI:24432 biological role └ΔCHEBI:33280 molecular messenger └ΔCHEBI:24621 hormone └ΔCHEBI:28918 (*R*)-adrenaline
Application	Intended use of the molecular entity or part thereof by humans.	CHEBI:23091 ChEBI ontology └◇CHEBI:33232 application └ΔCHEBI:25944 pesticide └ΔCHEBI:22153 acaricide └ΔCHEBI:38593 fenazaquin

Figure 4.2: The four ChEBI sub-ontologies [69].

[9] http://www.iupac.org/

[10] http://www.iubmb.org/

At the moment, there are 9 types of relationships used in ChEBI to link terms. Some are defined by the Relations Ontology (see Introduction 1.2.1) that are *is_a* and *part_of*; some others are unique and specifically required by ChEBI like *is_enantiomer_of* and *is_tautomer_of.* A general ChEBI entry consists of several data fields such as ChEBI identifier, ChEBI name, definition and synonyms:

- *ChEBI ID* is a unique and stable identifier for the entity, for example, carbohydrate acids (CHEBI:33720);
- *ChEBI Name* is the name for an entity recommended for use in biological databases and it may have been chosen to enhance clarity and avoid ambiguity;
- *Definition* is a short description of class name. For instance, the definition of oligosaccharides (CHEBI: 25679) is 'compounds in which monosaccharide units are joined by glycosidic linkages';
- *Synonyms* are alternative names such as systematic names, with an indication of their source.

4.5 Implementation project

Although compared to other chemistry databases, ChEBI database is small, the quality of its data is higher and it is able to provide a good standard for the representation of chemical substance and drugs in Gene Ontology vocabularies. Based on a long discussion on the GO mailing list, it was proposed that all the 'response to drug' terms had to be classified on rational basis and that a chemical classification rather then a pharmacological one was the best choice.

Promptly, a GO wiki page, providing the description of the project, was prepared[11] and it has been continuously maintained and updated with summaries about implementation work and all activities about 'response to drug' project. In addition a SourceForge tracker item was open to monitor the improvements of this activity: SF1494526[12] (see also related SF tracker items SF:1494548[13], SF:1658374[14]).

[11] http://gocwiki.geneontology.org/index.php/Response_to_drug

[12] http://sourceforge.net/tracker/index.php?func=detail&aid=1494526&group_id =36855&atid=440764

[13] http://sourceforge.net/tracker/index.php?func=detail&aid=1494548&group_i d=36855&atid=440764

[14] https://sourceforge.net/tracker/index.php?func=detail&aid=1658374&group_ id=36855&atid=440764

The project roughly consisted of obsoleting the ‘response to drug’ term (GO:0042493), in favour of specific ‘response to’ terms for specific chemical substances to be organized in parallel with the ChEBI molecular structure ontology. This strategy has a key role because it removes the burden of deciding what was or was not a ‘drug’ from the GO and it was outlined in six steps:

- **action 1**: move ‘response to drug’ child terms under the existing high level term ‘response to chemical stimulus’ (GO:0042221);
- **action 2**: use the ChEBI molecular structure ontology to better define the parent terms of ‘response to X’ term (where X is a chemical substance);
- **action 3**: request new terms to ChEBI curators, if there is not any available corresponding one in the ChEBI ontology;
- **action 4**: add synonyms useful for searching as required and also provide cross-references;
- **action 5**: obsolete ‘response to drug’ only when all its child terms have been removed;
- **action 6**: upon further discussion, all GO terms that specifically include the word ‘drug’ or ‘multi-drug’ in the term name would be obsolete and the children of such terms that refer to specific chemical substances would be re-homed appropriately.

For browsing and editing the vocabularies, the GO Consortium provides two tools: a browser called AmiGO and an editor OBO-Edit (see Chapter 6) that have been used in order to carry on the response-to-drug project. A range of ID numbers was set up specifically for this set of terms and the GO reference GOC:ef (Erika Feltrin) was included in the definition Dbxref field of each new or modified term to indicate which GO curator was responsible for all changes.

4.6 Results

Starting from the existing structure of ‘response to drug’ node (see Figure 4.1), the first thing was to browse ChEBI molecular structure sub-ontology and identify standard chemical terms that might be sources for the definition of new GO terms.

The ChEBI was cross-referenced and used as a source of definition for GO terms. Where appropriate, new definitions were contributed to the ChEBI ontology. For example a new entry of cycloheximide was created and introduced in ChEBI and re-defined in GO ontologies, and finally the resulting term was cross-referenced. Moreover, common names have been added as synonyms to help researchers in finding the terms.

In ChEBI, the ‘organic molecular compound’ (CHEBI:25700) was defined as a molecular entity that contain carbon. The possible correspondent GO term was ‘response to organic substance’ (GO:0010033), defined as ‘a change change in state or activity of a cell or an organism (in terms of movement, secretion, enzyme production, gene expression, etc.) as a result of an *organic substance* stimulus’. But exactly, what is the meaning of an *organic substance*? The answer to this question came from the ChEBI ‘organic molecular compound’ definition that therefore has been used to classify an organic substance in GO ontologies.

As a result, ‘response to organic substance’ (GO:0010033) gained the role of new parent term for some of the ‘response to drug’ child terms. Nevertheless, finding correspondences between the two ontologies, in order to move the child terms from ‘response to drug’ to the redefined ‘response to organic substance’, was difficult and at least three strategies were required:

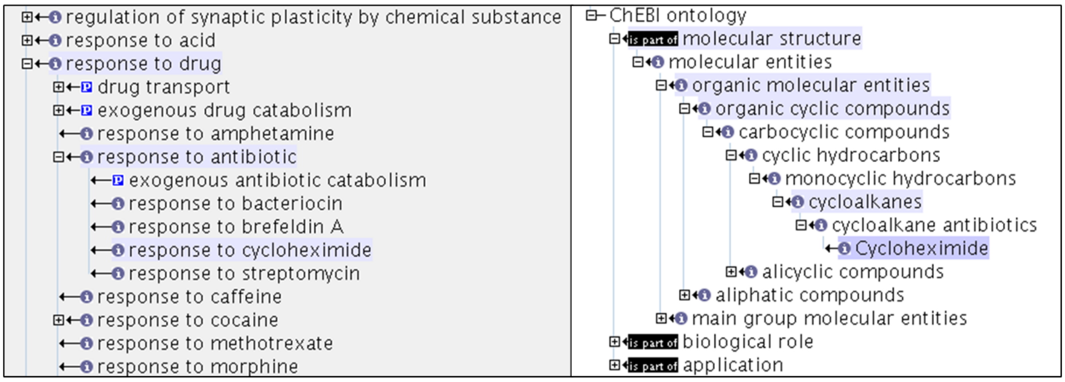
1. the X compound in the ‘response to X’ GO term was already present in the ChEBI molecular structure ontology (e.g. cycloheximide and morphine). The solution was to mirros part of ChEBI structure in the GO;

2. the X compound in the ‘response to X’ GO term was not included in the ChEBI molecular structure ontology but it was present in other ChEBI sub-ontologies (application, molecular role or unclassified) (e.g. methotrexate). The solution was first, forward a request to ChEBI curators to have the missing term under molecular structure ontology and once added, mirror it in the GO;

3. the X compound in the ‘response to X’ GO term was not included in the ChEBI ontology (e.g. citalopram) or was unclassified (e.g. fluoxetine (ChEBI:5118)). In this case the solution was first, forward a request to ChEBI curators to have the missing term and once added, mirror it in the GO.

4.6.1 Case of cycloheximide (GO:0046898; CHEBI:17076)

In the existing structure, ‘response to cycloheximide’ was a child of ‘response to antibiotic’ (GO:0046677) but according to the plan, this substance should not be classified following its role in the cell but on the base of its molecular structure. Then, looking at the ChEBI, it was clear that it was a child of cycloalkanes (Figure 4.3) and therefore, two new terms were created: ‘response to organic cyclic substance’ (GO:0014070) and ‘response to cycloalkane’ (GO:0014071). The implementation of part of the ChEBI structure allowed creation of a new DAG organisation in GO, moving the term ‘response to cycloheximide’ (GO:0046898) from ‘response to antibiotic’ to ‘response to cycloalkane’ (Figure 4.4).



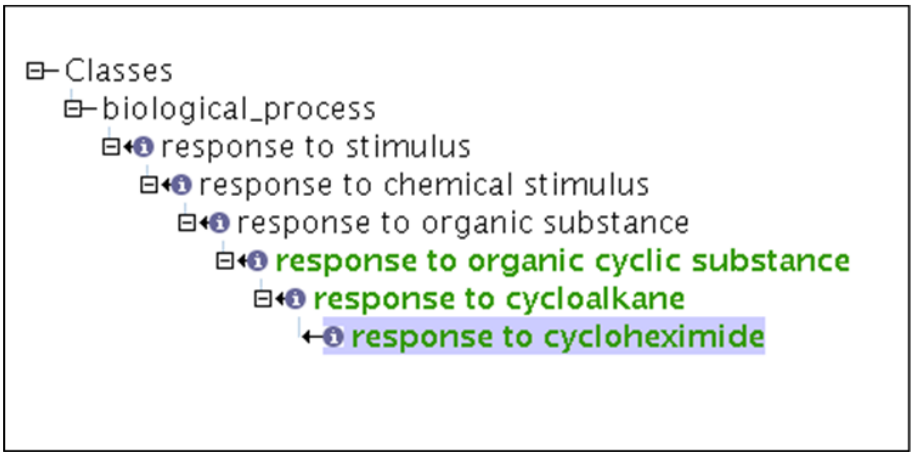
```
⊞ regulation of synaptic plasticity by chemical substance
⊞ response to acid
⊟ response to drug
    ⊞ drug transport
    ⊞ exogenous drug catabolism
      response to amphetamine
    ⊟ response to antibiotic
        exogenous antibiotic catabolism
        response to bacteriocin
        response to brefeldin A
        response to cycloheximide
        response to streptomycin
      response to caffeine
    ⊞ response to cocaine
      response to methotrexate
      response to morphine
```

```
⊟ ChEBI ontology
  ⊟ is part of molecular structure
    ⊟ molecular entities
      ⊟ organic molecular entities
        ⊟ organic cyclic compounds
          ⊟ carbocyclic compounds
            ⊟ cyclic hydrocarbons
              ⊟ monocyclic hydrocarbons
                ⊟ cycloalkanes
                  ⊟ cycloalkane antibiotics
                      Cycloheximide
            ⊞ alicyclic compounds
        ⊞ aliphatic compounds
      ⊞ main group molecular entities
  ⊞ is part of biological role
  ⊞ is part of application
```

Figure 4.3: Correspondences between GO (on the left) and ChEBI (on the right) DAG structure for the substance cycloheximide. Based on the ChEBI classification, the GO structure was reorganised following a chemical classification.

The same process could be applied for ‘response to bacteriocin’ (GO:0046678), ‘response to brefeldin A’ (GO:0031001) and ‘response to streptomycin’ (GO:0046679). For streptomycin it would be easy because it was already present in ChEBI ontology (CHEBI:17076); instead for the others new terms have to be requested to ChEBI curators and await implementation in ChEBI.

Unfortunately, a discussion about the ‘response to antibiotic’ node is still in progress because first, many GO curators and GO consortium groups are very interested in this topic and second, because about 150 proteins are annotated to this node. The re-definition or the change of position are processes that require much attention and especially community consensus. Therefore, it has been decided to move the response to antibiotic parent term, together with its children, out of the response to drug term to directly under the more general response to chemical stimulus term. This term will be further corrected by the Consortium at a later date.



- Classes
  - biological_process
    - response to stimulus
      - response to chemical stimulus
        - response to organic substance
          - response to organic cyclic substance
            - response to cycloalkane
              - response to cycloheximide

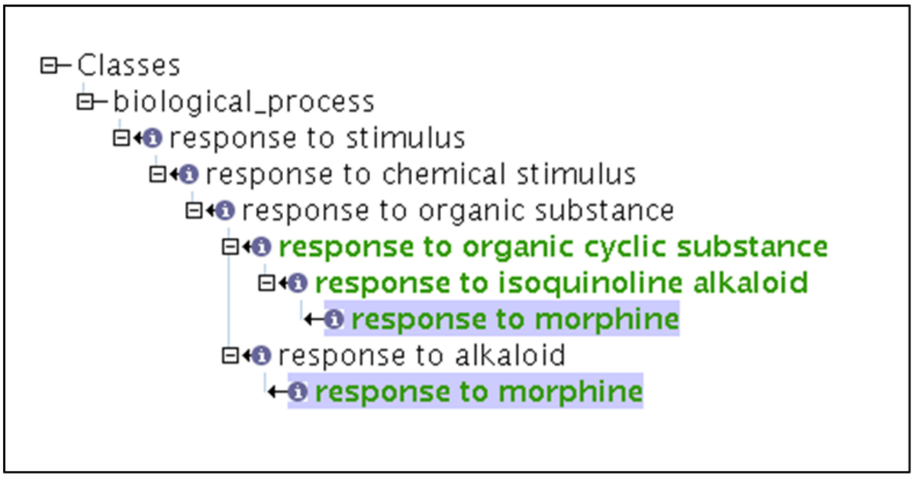
Figure 4.4: Resulting GO structure after changes of the ‘response to antibiotic’ and ‘response to organic cyclic substance’ nodes. The term ‘response to cycloheximide’ is now an *is_a* child of ‘response to cycloalkane’.

4.6.2 Case of morphine (GO:0043278; CHEBI:17303)

The term ‘response to morphine’ (GO:0043278) was an other example where the ChEBI represented a valuable help in the reorganisation of GO DAG, in fact this GO term had a correspondence to the ‘morphine’ ChEBI term (CHEBI:17303).

Morphine is derived from alkaloids and it is highly potent opiate analgesic drug. Like other opiates, morphine acts directly on the central nervous system and in particular at synapses to relieve pain. Interestingly, morphine has recently been found to be endogenously produced by humans, made by cells in the heart, pancreas and brain and it has also been isolated from a range of other mammals, as well as some invertebrates [70]. Clearly, a chemical classification of morphine in GO will improve the representation of biological processes in response to drug treatments.

Applying the same method of the previous ‘response to’ term, the term ‘response to morphine’ was moved and implemented as child term of ‘response to alkaloid’ (GO:0043279) in the ‘response to organic substance’ node. At the same time, a new GO term ‘response to isoquinoline alkaloid’ (GO:0014072) was created and added as child of ‘response to organic cyclic substance’ node. These changes were necessary to allow classification of morphine on its molecular structure and not on its application role (Figure 4.5).



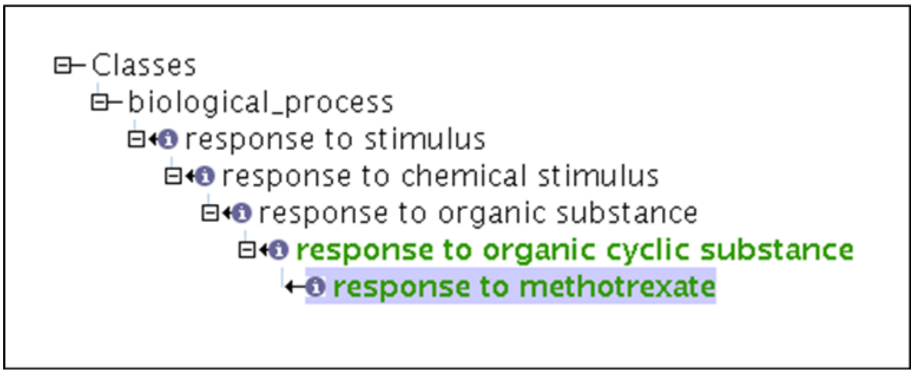
- Classes
  - biological_process
    - response to stimulus
      - response to chemical stimulus
        - response to organic substance
          - response to organic cyclic substance
            - response to isoquinoline alkaloid
              - response to morphine
          - response to alkaloid
            - response to morphine

Figure 4.5: This figure shows the final structure of the GO DAG after the implementation of changes in the case of ‘response to morphine’ (GO:0043278).

4.6.3 Case of methotrexate (GO:31427; CHEBI:null)

Methotrexate is an antimetabolite drug used in treatment of cancer, autoimmune diseases and severe skin conditions such as psoriasis, and rheumatoid arthritis by inhibiting the metabolism of folic acid. There is evidence that the combination of this immunosuppressive agent with standard or new therapies provides synergistic effects in patients affected by multiple sclerosis [71].

Differently from the previous examples, in this case there was an additional obstacle: ‘response to methotrexate’ was a child of ‘response to drug’ in GO but methotrexate was not classified in ChEBI ontology. In collaboration with ChEBI curators, there was created a new entry for methotrexate with the ChEBI:44185 in the molecular structure ontology and the new node was used as a model for modifying GO. The change gave a contribution to both ChEBI and GO vocabularies (Figure 4.6).



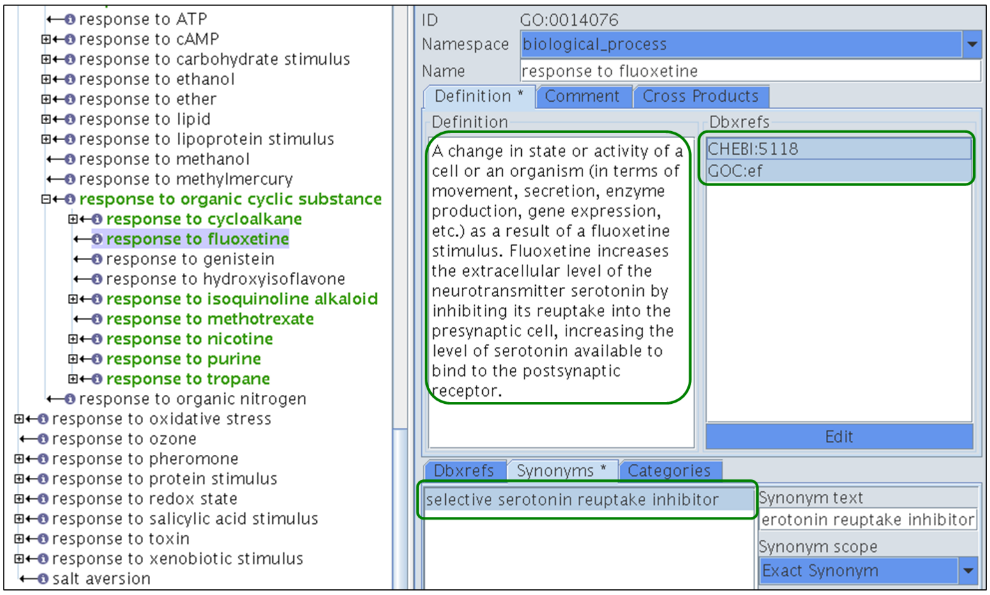
Classes
biological_process
response to stimulus
response to chemical stimulus
response to organic substance
response to organic cyclic substance
response to methotrexate

Figure 4.6: 'response to methotrexate' is now a child of 'response to organic cyclic compound' classified on a chemical basis.

4.6.4 Case of fluoxetine (GO:null; CHEBI:44185)

As previously anticipated, there are biological processes of interest to pharmaceutical researchers for the annotation of gene products that respond to the administration of specific drugs in the treatment of a certain disease. An example is fluoxetine, a widely used antidepressant of the selective serotonin reuptake inhibitors (SSRIs) class, that exerts its medicinal effects almost exclusively by blocking the serotonin uptake systems. It has a quite complex mechanism of action and affects a variety of membrane proteins, increasing levels of serotonin in many brain areas [72].

According to its pharmaceutical application, fluoxetine is an *is_a* child of 'central nervous system drug' (CHEBI:35470) and according to its structure, it is an *is_a* child of 'organic cyclic compound' entity (ChEBI:33832). But, the GO vocabularies were lacking a term 'response to fluoxetine' that would be helpful in representing the normal physiological response of an organism to a drug, for example, in terms of changes of gene expression levels or activity/inactivity of proteins. This term was added, (Figure 4.7) and in order to represent more response processes to other SSRI drugs, several entities have been described in ChEBI and have been crossed referenced in GO (Table 4.1).



response to ATP
response to cAMP
response to carbohydrate stimulus
response to ethanol
response to ether
response to lipid
response to lipoprotein stimulus
response to methanol
response to methylmercury
response to organic cyclic substance
response to cycloalkane
response to fluoxetine
response to genistein
response to hydroxyisoflavone
response to isoquinoline alkaloid
response to methotrexate
response to nicotine
response to purine
response to tropane
response to organic nitrogen
response to oxidative stress
response to ozone
response to pheromone
response to protein stimulus
response to redox state
response to salicylic acid stimulus
response to toxin
response to xenobiotic stimulus
salt aversion

ID GO:0014076
Namespace biological_process
Name response to fluoxetine

Definition: A change in state or activity of a cell or an organism (in terms of movement, secretion, enzyme production, gene expression, etc.) as a result of a fluoxetine stimulus. Fluoxetine increases the extracellular level of the neurotransmitter serotonin by inhibiting its reuptake into the presynaptic cell, increasing the level of serotonin available to bind to the postsynaptic receptor.

Dbxrefs: CHEBI:5118, GOC:ef

selective serotonin reuptake inhibitor

Synonym text: erotonin reuptake inhibitor
Synonym scope: Exact Synonym

Figure 4.7: A new GO term named 'response to fluoxetine' was created as an *is_a* child of 'response to organic cyclic substance' (on the left of the figure). The term now has a more detailed definition cross-referenced to ChEBI ontology (on the right of the figure) and a new synonym 'selective serotonin reuptake inhibitor'.

4.6.5 Resulting 'response to drug' structure

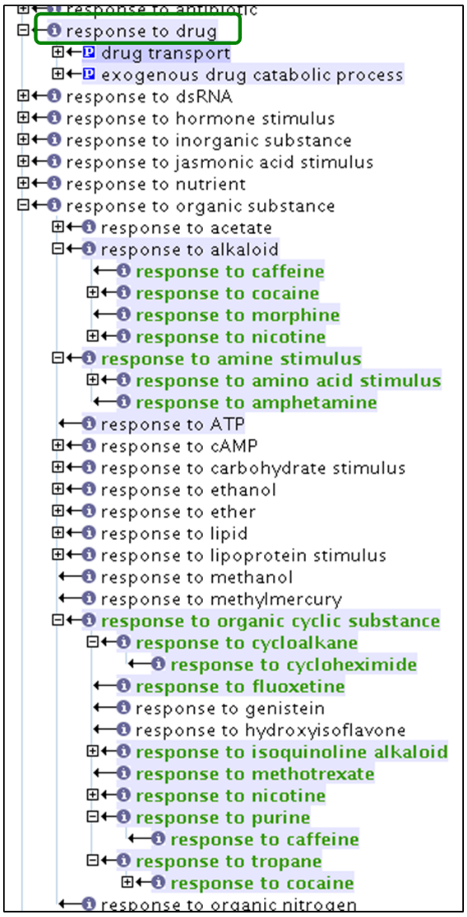
The cases described above are only examples of the method followed for carrying on the response-to-drug work. The final structure of the response to drug node is shown in Figure 4.8. The terms in green are new terms or terms that have been redefined. Not only the terms described in examples were changed but also more existing terms were rearranged such as 'response to caffeine' and 'response to amphetamine'. As shown in Figure 4.8, the response to drug node has currently only two *part_of* child terms: 'drug transport' and 'exogenous drug catabolic process'. As a general rule, the term can be obsolete only when it does not have any child terms. In this case, other work has to be done to complete the reorganisation of the structure and to made the term 'response to drug' obsolete.

In addition, requests for new terms are continually submitted to ChEBI curators via Source-Forge mailing list. These terms are of particular interest to pharmaceutical researchers because they are appropriate for the annotation of gene products that respond to the administration of drugs. The terms that are already implemented in ChEBI are listed in Table 4.1. Some of them will not be visible until next ChEBI release, which is scheduled for February, 2008.

Entity name	**CHEBI ID**
Sertraline	CHEBI:9123
paroxetine	CHEBI:7936
citalopram	CHEBI:3723
venlafaxine	CHEBI:9943
(S)-citalopram	CHEBI:36791
fluvoxamine	CHEBI:5138
bupropion	CHEBI:3219
amitriptyline	CHEBI:2666
duloxetine	CHEBI:36796
dothiepin	CHEBI:36798
clomipramine	CHEBI:47780
desipramine	CHEBI:47781
imipramine	CHEBI:47499
lofepramine	CHEBI:47782
nortriptyline	CHEBI:7640
protriptyline	CHEBI:8597
trimipramine)	CHEBI:9738
streptomycin	CHEBI:17076
brefeldin A	CHEBI:48080
bacteriocin	CHEBI:48081

Table 4.1: Terms that are already implemented in ChEBI. Some of them will not be visible until next ChEBI release which is scheduled for February, 2008.

Many terms with the word ‘drug’ in the name or synonym are currently present in the GO vocabularies (Table 4.2): 8 terms in the Molecular Function and 11 terms in the Biological



- response to antibiotic
- response to drug
  - drug transport
  - exogenous drug catabolic process
- response to dsRNA
- response to hormone stimulus
- response to inorganic substance
- response to jasmonic acid stimulus
- response to nutrient
- response to organic substance
  - response to acetate
  - response to alkaloid
    - **response to caffeine**
    - **response to cocaine**
    - **response to morphine**
    - **response to nicotine**
  - **response to amine stimulus**
    - **response to amino acid stimulus**
    - **response to amphetamine**
  - response to ATP
  - response to cAMP
  - response to carbohydrate stimulus
  - response to ethanol
  - response to ether
  - response to lipid
  - response to lipoprotein stimulus
  - response to methanol
  - response to methylmercury
  - **response to organic cyclic substance**
    - **response to cycloalkane**
      - **response to cycloheximide**
    - **response to fluoxetine**
    - response to genistein
    - response to hydroxyisoflavone
    - **response to isoquinoline alkaloid**
    - **response to methotrexate**
    - **response to nicotine**
    - **response to purine**
      - **response to caffeine**
    - **response to tropane**
      - **response to cocaine**
  - response to organic nitrogen

Figure 4.8: The current structure of the response to drug node in the GO biological process tree (GO revision 5.631 December 2007). The terms in green are new terms or terms that have been changed.

Function tree. Based on the similar reasoning about the ‘response to drug’ terms, all GO terms that specifically include the word ‘drug’ or ‘multidrug’ in the term name will be obsoleted, and the child terms of such terms that refer to specific chemical substances will be redefined appropriately. As this will affect lots of annotations, additional discussion of this issue will be necessary to find a better solution that will satisfy all GO consortium members.

It is very important to enphasize that the process for the reorganisation of an existing term and/or the introduction of a new term differ from term to term. The working group has to be aware, and in favour, of any changes that the GO curator is going to apply to GO structure

11 Biological Process terms	8 Molecular Function terms
drug catabolism	drug binding
drug export	drug transporter activity
drug metabolism	drug:hydrogen antiporter activity
drug transport	multidrug efflux pump activity
endogenous drug catabolism	multidrug endosomal transporter activity
multidrug transport	multidrug transporter activity
induction of synaptic plasticity by chemical substance	multidrug, alkane resistant pump activity
exogenous drug catabolism	xenobiotic-transporting ATPase activity
positive regulation of synaptic plasticity by chemical substance	
regulation of synaptic plasticity by chemical substance	
response to drug	

Table 4.2: List of the terms with the word 'drug' in the name that are currently present in the GO vocabularies. They are 8 terms in the Molecular Function and 11 terms in the Biological Function tree. These will be addressed in future work by the GO Consortium, to conform to the standards established as part of my PhD project.

and content. Therefore, discussions among GO curators and annotators who are not members of the working group, are extremely helpful because they ensure that all can have a chance to discuss both the case for change and its implications. Sometime the consensus is reached only after a face-to-face discussion at a GO consortium meeting. All these things might contribute to delay the work but assure that all changes have been done in respect of GO rules and purpose.

Chapter 5

Disease Ontology Annotation project

In this chapter is described the development of a resource for representing the relations between genes, drugs and diseases to help understand the mechanism of diseases. Firstly, the data sources are listed and described; then there is a section about how the data have been collected and organised in our database. Finally the last sections suggest a possible applications and further developments.

5.1 Introduction to the project

The first problem is that high-throughput functional genomic technologies have resulted in the rapid accumulation of genome-scale data sets. At the same time linkage analysis and association studies that identify disease-associated genes, generate increasingly large candidate gene sets that need to be analysed. However it remains difficult to identify the most likely relationship between the studied disease and genes. The etiology of most chronic diseases involves interactions between environmental factors and genes that modulate important biological processes [73]; however the molecular mechanisms underlying the correlation between chemicals and diseases are not well understood. Therefore there is a major, continuing need to integrate, aggregate and annotate data about genes, drugs and diseases. An other problem is that biologists, interested in different disease, are currently hampered by the differences in the technical language. For example, a physician might try to access information on gene products involved in 'Alzheimer's disease'. During the search, he would find that genes relevant for 'myokymia' are also involved in other diseases such as 'Morvan's chorea' without knowing that the latter is a correct synonym for the former. An other problem is the polysemy. Polysemy is the ambiguity of an individual word or phrase that can be used (in different contexts) to express two or more different meanings. To make the best use of biological databases, different kinds of information from different sources must be integrated in ways that make sense to scientific community. Ontologies are a valuable possibility for data integration [2].

Following the example of Gene Ontology Annotation (GOA) project, our goal is to classify and represent gene-drug, gene-disease or gene-drug-disease associations in a standardized way using ontologies. The idea is to associated genes both related to disorders and regulated by drug treatment, using the terms of the Disease Ontology (DO). The final result is the building of a knowledge base of genes, drugs and targets to help understand the mechanism of diseases.

5.2 Resources

Several sources such as ontologies and databases collecting data about genes, drugs and diseases have been evaluated and finally only 5 of them were chosen for developing our knowledge base. A dictionary for disease names has been developed from three of them, and at the same time a vocabulary for drug names has been populated using two of these sources.

5.2.1 Disease Ontology

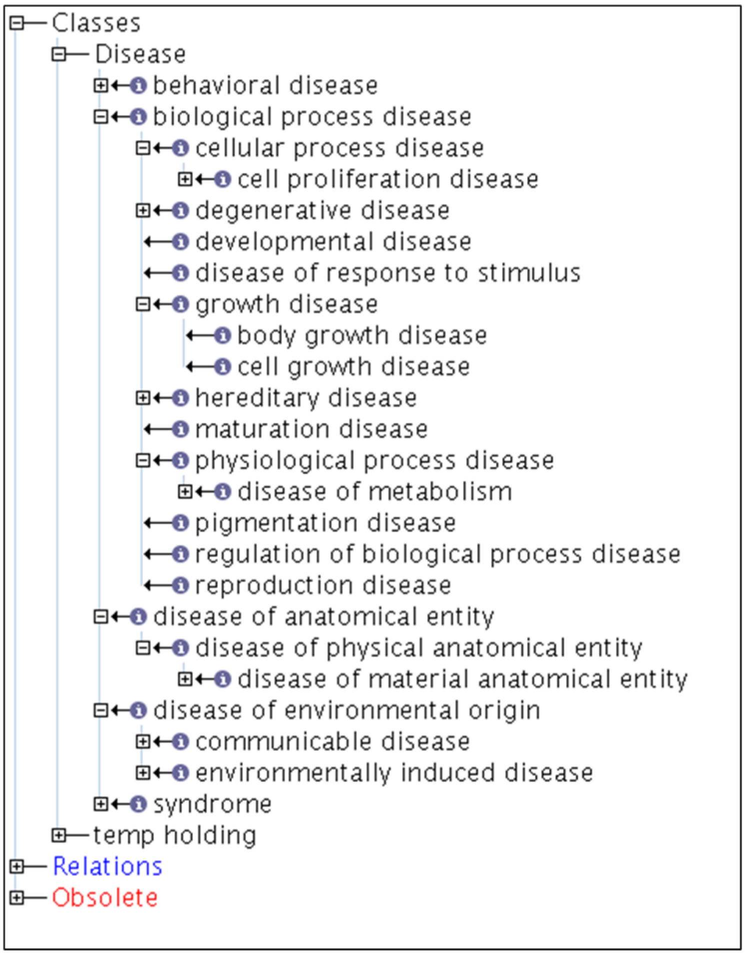
The Disease Ontology (DO)[1] is a controlled medical vocabulary, modelled on GO, developed at the Bioinformatics Core Facility in collaboration with the NuGene Project at the Center for Genetic Medicine (Chicago, US). It was designed to facilitate the mapping of diseases and associated conditions to particular medical codes such as ICD9CM[2], SNOMED[3] and others. The Disease Ontology is implemented as a directed acyclic graph (DAG) and utilizes the

[1] http://diseaseontology.sourceforge.net/#projects

[2] The International Classification of Diseases, Ninth Revision, Clinical Modification is the official system of for the classification of disease entries, diagnostic, and therapeutic procedures associated with hospital utilization in the US.

[3] Systematized Nomenclature of Medicine-Clinical Terms is a standardized vocabulary system that creates a common clinical language for medical databases. Current modules contain more that 357,000 concepts.

Unified Medical Language System (UMLS)[4] [74]. Using this standard, much of the process of updating the ontology can be handled by UMLS, freeing resources for clinicians to pursue more urgent tasks. In a manner similar to the GO curation process and open development, the ontology is continually extended and revised in order to broadly encompass diseases. The DO is available in OBO format and it can be readily edited and viewed using OBO-Edit. Figure 5.1 shows an OBO-Edit screenshot of the Disease Ontology version 3.



- Classes
  - Disease
    - behavioral disease
    - biological process disease
      - cellular process disease
        - cell proliferation disease
      - degenerative disease
      - developmental disease
      - disease of response to stimulus
      - growth disease
        - body growth disease
        - cell growth disease
      - hereditary disease
      - maturation disease
      - physiological process disease
        - disease of metabolism
      - pigmentation disease
      - regulation of biological process disease
      - reproduction disease
    - disease of anatomical entity
      - disease of physical anatomical entity
        - disease of material anatomical entity
    - disease of environmental origin
      - communicable disease
      - environmentally induced disease
    - syndrome
  - temp holding
- Relations
- Obsolete

Figure 5.1: Screenshot of DO using OBO-Edit version 1.1

[4] The UMLS contains a metathesaurus within medical concepts and a semantic network. It is intended to be used mainly by developers of systems in medical informatics and it provides facilities for natural language processing.

The previous version of Disease Ontology 2 (v2.1) was based almost entirely on ICD9CM with additional concepts that are useful for mapping common disease requests. For our project has been used the version 3 of Disease Ontology containing 12,448 concept nodes which is currently available for download from the SourceForge home page[5]. The choice of this ontology was based on the consideration that it has never been used for gene annotation, and no annotation file has been developed yet. Moreover, as already explain in the introduction, there are several advantages in using ontologies:

- through their semantic-free identifiers (unique IDs), they allow linkage to other resources that use them (e.g. GAD database);
- terms in natural language are often ambiguous even in the context of their ontology, however the hierarchical definition structure (DAG) ensures appropriate interpretation (Figure 5.2);
- ontologies are machine processable.

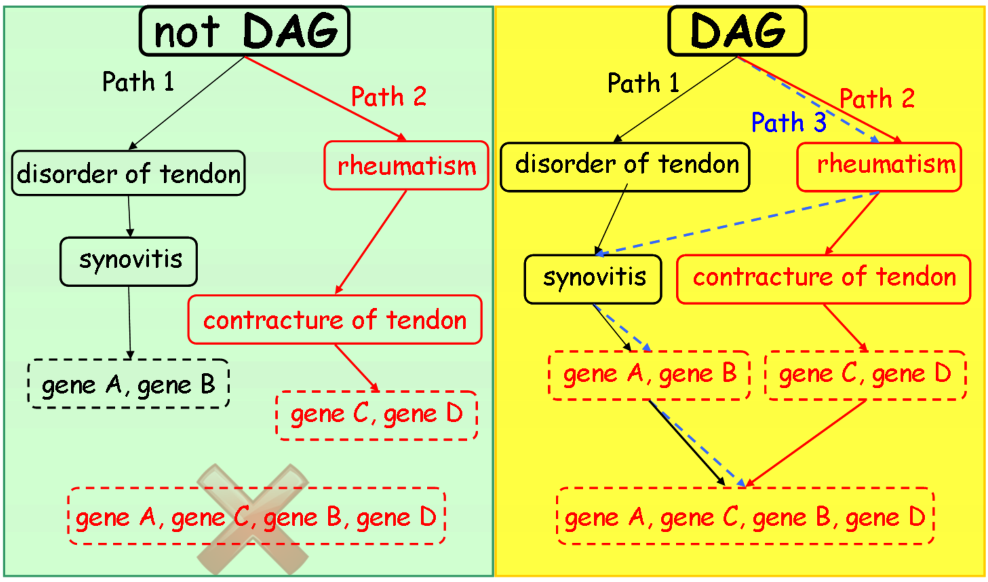


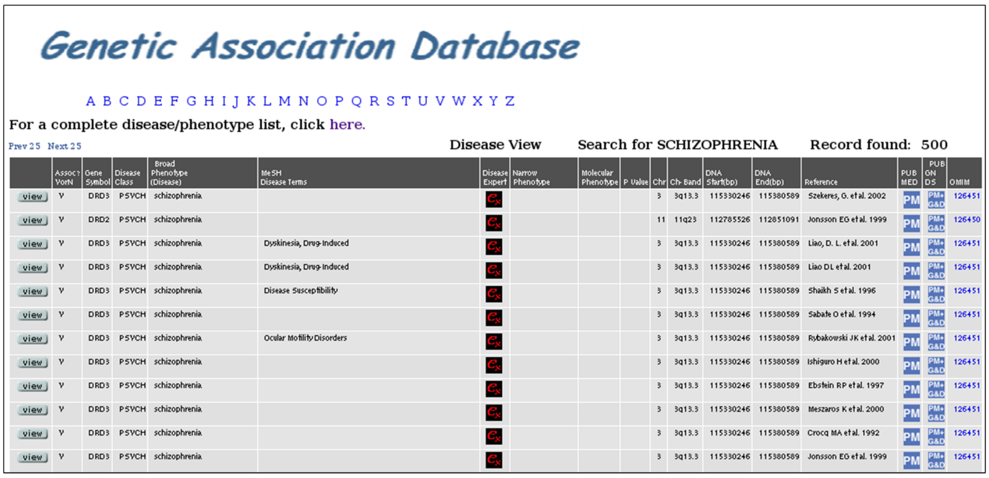
Figure 5.2: Genes annotated with the child concept 'synovitis' can be transitively annotated with parent terms 'disorder of tendon' and 'rheumatism' (on the right). This is not possible out of a DAG structure (on the left).

[5] http://sourceforge.net/project/showfiles.php?group_id79168&package_id=202115&release_id=508426

5.2.2 Genetic Association Database

The Genetic Association Database (GAD)[6] is a publicly available NIH based database of published gene-based genetic association studies which contains records of over 5,000 human genetic association studies. The database is gene centered and provides a standardized molecular nomenclature by including official HUGO gene symbol. Each record refers to a gene or a marker and is annotated with links to molecular databases (LocusLink, GeneCards) and references databases (PubMed, CDC) among others [75]. The goal of this database is to allow the user to rapidly identify medically relevant polymorphism from the large volume of polymorphism and mutational data.

In GAD, there are several data fields common to genetic association studies, such as disease, phenotypes, sample sizes and allele descriptions (Figure 5.3). Of particular interest are the fields about diseases. A top level 'disease class' is assigned followed by 'disease' from the original paper. Then, there is the 'Broad (or Narrow) Phenotype' disease class that is assigned if studies recognize clinical subphenotypes and finally there is the MeSH Disease Terms. A list of all disease/phenotype is available[7]. The OMIM gene field links each GAD official HUGO genes to OMIM genes.



Genetic Association Database

A B C D E F G H I J K L M N O P Q R S T U V W X Y Z

For a complete disease/phenotype list, click here.

Prev 25 Next 25

Disease View Search for SCHIZOPHRENIA Record found: 500

	Assoc? YorN	Gene Symbol	Disease Class	Broad Phenotype (Disease)	MeSH Disease Terms	Disease Expert	Narrow Phenotype	Molecular Phenotype	P Value	Chr	Ch-Band	DNA Start(bp)	DNA End(bp)	Reference	PUB MED	PUB GN DS	OMIM
view	Y	DRD3	PSYCH	schizophrenia		ex				3	3q13.3	115330246	115380589	Szekeres, G. et al. 2002	PM	PM+ G&D	126451
view	Y	DRD2	PSYCH	schizophrenia		ex				11	11q23	112785526	112851091	Jonsson EG et al. 1999	PM	PM+ G&D	126450
view	Y	DRD3	PSYCH	schizophrenia	Dyskinesia, Drug-Induced	ex				3	3q13.3	115330246	115380589	Liao, D. L. et al. 2001	PM	PM+ G&D	126451
view	Y	DRD3	PSYCH	schizophrenia	Dyskinesia, Drug-Induced	ex				3	3q13.3	115330246	115380589	Liao DL et al. 2001	PM	PM+ G&D	126451
view	Y	DRD3	PSYCH	schizophrenia	Disease Susceptibility	ex				3	3q13.3	115330246	115380589	Shaikh S et al. 1996	PM	PM+ G&D	126451
view	Y	DRD3	PSYCH	schizophrenia		ex				3	3q13.3	115330246	115380589	Sabate O et al. 1994	PM	PM+ G&D	126451
view	Y	DRD3	PSYCH	schizophrenia	Ocular Motility Disorders	ex				3	3q13.3	115330246	115380589	Rybakowski JK et al. 2001	PM	PM+ G&D	126451
view	Y	DRD3	PSYCH	schizophrenia		ex				3	3q13.3	115330246	115380589	Ishiguro H et al. 2000	PM	PM+ G&D	126451
view	Y	DRD3	PSYCH	schizophrenia		ex				3	3q13.3	115330246	115380589	Ebstein RP et al. 1997	PM	PM+ G&D	126451
view	Y	DRD3	PSYCH	schizophrenia		ex				3	3q13.3	115330246	115380589	Meszaros K et al. 2000	PM	PM+ G&D	126451
view	Y	DRD3	PSYCH	schizophrenia		ex				3	3q13.3	115330246	115380589	Crocq MA et al. 1992	PM	PM+ G&D	126451
view	Y	DRD3	PSYCH	schizophrenia		ex				3	3q13.3	115330246	115380589	Jonsson EG et al. 1999	PM	PM+ G&D	126451

Figure 5.3: A simple search of associations for the disease schizophrenia. Fields in this view include Official Gene Symbol, Disease Phenotype, Disease class, OMIM ID, MeSH Disease term.

[6] http://geneticassociationdb.nih.gov
[7] http://geneticassociationdb.nih.gov/diseaselist.html

This database was selected as external resource because it is based on manual curation and therefore provides an excellent baseline for constructing our database. A community of experts listed in ad-hoc list contributes to the GAD curation process. Anyone, specialized in either a specific disease, and/or a specific gene, or other related expertise, such as disease or gene specific data collections can easily enter the list.

5.2.3 OMIM

The Online Mendelian Inheritance in Man (OMIM)[8] is a comprehensive, authoritative and regularly updated knowledgebase of human genes and genetic disorders compiled to support human genetics research and education and the practice of clinical genetics [76]. OMIM data are organised in two different files: the 'gene map' and the 'morbid map' available from the FTP site[9]. The OMIM Gene Map is a single file, in tabular format, listing genes that are described in OMIM. Not all OMIM entries are included in the Gene Map, but only those for which a cytogenetic location has been published in the cited references. Each entry is a list of fields such as gene location, gene symbol, MIM number, disorders and reference. The OMIM Morbid Map is an alphabetical list of diseases used in the database and their corresponding cytogenetic locations.

5.2.4 DrugBank

The DrugBank[10] is a unique bioinformatics and cheminformatics resource that combines detailed drug (i.e. chemical, pharmacological and pharmaceutical) data with comprehensive drug target (i.e. sequence, structure, and pathway) information [77]. This includes physical property data, structure and image files, pharmacological and physiological data about thousands of drug products as well as extensive molecular biological information about their corresponding drug targets. Each DrugCard contains more than 80 data fields with half of the information being devoted to drug/chemical data and the other half devoted to drug target or protein data. Each entry is entered and prepared by one member of the curation team and then separately validated by a second member of the curation team. Human validation is a

[8] http://www.ncbi.nlm.nih.gov/omim/

[9] ftp://ftp.ncbi.nih.gov/repository/OMIM/

[10] http://redpoll.pharmacy.ualberta.ca/drugbank/index.html

warrantee of quality and completeness. In addition drug targets are confirmed using multiple sources (PuBMed, RxList, PharmGKB) as are all drug structures (KEGG, PubChem). Thanks to this manual curation, the data collected in this resource are high-quality and for this reason, these were employed for the construction of our database.

5.2.5 PharmGKB

The Pharmacogenetics and Pharmacogenomics Knowledge Base[11] (PharmGKB) is a public resource that contains genomic, phenotype and clinical information collected from ongoing research and from the literature [78]. It is devoted to cataloguing information about pharmacogenes such as those genes involved in modulating the response to drugs [79]. These genes are pharmacogenes because they are involved in the pharmacokinetics (PK) of a drug (how the drug is absorbed, distributed, metabolised and eliminated) or the pharmacodynamics (PD) of a drug (how the drug acts on its target and its mechanisms of action). The aim is to capture the relationships between drugs, diseases/phenotypes and genes including several other types of information such as literature annotations, primary data sets, PK and PD pathways, and expert-generated summaries of PK/PD relationships [80].

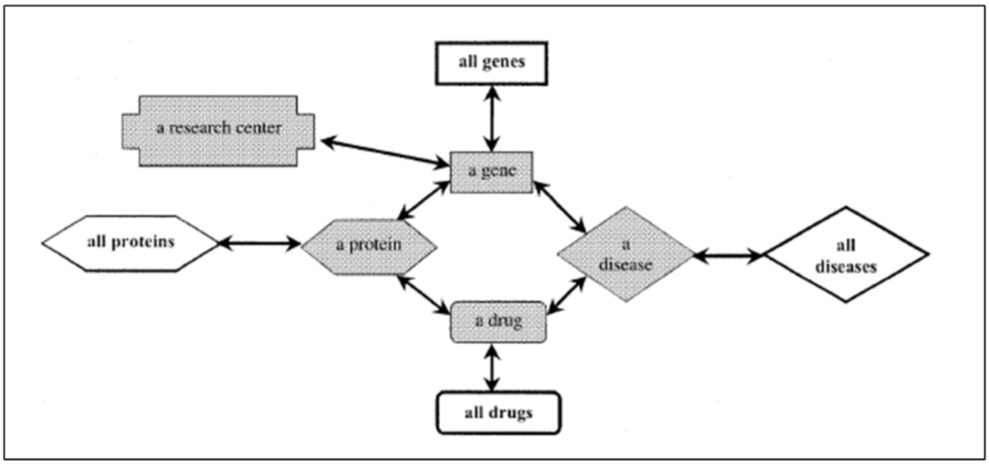


Figure 5.4: Some of the relationships among data objects in PharmGKB. Today, the PharmGKB has curated evidence for nearly 2,000 genes involved in drug response.

There are 519 drugs with associated phenotype data, genotype data or literature anno-

[11] http://www.pharmgkb.org/index.jsp

tations, 43 manually created drug-related pathway, 518 diseases with supporting information and 2,100 literature entries (data of November, 2007) (Figure 5.4). Moreover, the scientific community helps to curate the contents of the database by providing information about gene-drug, gene-disease or gene-drug-disease associations, as well as the type of evidence that is available for these associations and furthermore all submitted data are curated and validated by curators.

5.3 Method and Results

The approach for creating our tool combines automated and manually curated strategies to address two distinct task: i) extracting gene, disease and drug data from the selected resources and ii) characterising relationships using several strategies. The proposed method can be divided in 4 phases:

- phase 1: acquisition and integration of data from the external resources;
- phase 2: compilation of a vocabulary for disease names and synonyms and a dictionary for drug names and synonyms;
- phase 3: association of diseases, genes and drugs to DO terms based on automated and manually curated approaches;
- phase 4: design and implementation of a MySQL database.

There have been several problems in the development of our resource. For this reason, tools have been designed case by case to parse data and pull out only relevant information, then the information was re-organized to be easily accessible to the query tools and to allow easy maintenance and update.

5.3.1 Acquisition and integration of data

The first effort was to retrieve relevant data about genes, drugs and diseases and their relationships from each external database and resources. The problem was that databases differ in terms of information content and data format, thus different approaches were adopted to produce files with accesible format to better exploit the resources available.

After downloading the newest revision of the **Disease Ontology 3** (revision 21)[12], the DO text file was parsed to extract all disease names and synonyms with DO identifiers, leaving out the ‘temp holding’ and the ‘obsolete’ terms. As already said, the terms in the DO are structure as a DAG, where a parent term can have more than one child term and in turn a child term can have more than one parent term. A Perl script to navigate the terms in the DO has been developed and it allows to draw inferences from selected terms, going down throught descendent or up through ancestor of a given term, and taking account of multiple paths.

The **PharmGKB** provides access to a selected subset of data via a SOAP interface and documentation. The sample client code is freely available and the client programs are downloadable from the home page[13]. Some Perl scripts have been combined in order to extract different kinds of data from the PharmGKB knowledgebase. In particular the *specialSearch.pl* script was run with option 6 to obtain all diseases with supporting information. The result was parsed and given as input file for the *disease.pl* script obtaining information about related genes and drugs for each disease. Finally *drugs.pl* and *genes.pl* scripts were used to retrieve information about each drug and gene. In addition when available, the drug chemical structures were collected and then implemented in our database.

To download the complete database, **GAD** requires filling in a form and then sends out a tab-delimited text files. Several attributes are used to described each GAD entry. Afterwards filters have been applied to select only relevant fields like Broad Phenotype, Disease Class, MeSH Disease Term, Gene, Gene Name, OMIM ID.

Furthermore the **OMIM** morbid map was used to extract information on disorders and genes involved in the disorders with the correspondent OMIM ID.

Finally, since **DrugBank** is a freely available resource, a full set of DrugBank Approved DrugCards in a single flat file was downloaded[14] and used as a source for drug names and synonyms, and gene target symbols.

For each databases, a list of all diseases was prepared and used for the compilation of the disease dictionary. Then, association data between genes and diseases were extracted from each resource and later used for the gene annotation process.

[12] http://sourceforge.net/project/showfiles.php?group_id$\bar{7}$9168&package_id=202115&release_id=508426

[13] http://www.pharmgkb.org/home/projects/webservices/index.jsp

[14] http://redpoll.pharmacy.ualberta.ca/drugbank/cgi-bin/download.cgi

5.3.2 Acquiring synonym terms for disease names

One problem that had to be tackled was the presence of synonyms, because in the case of diseases, there often exist different names for the same diseases. Also genes associated to the same disease are often annotated with different synonyms. Therefore, in order to solve this problem, the external resources, GAD, PharmGKB and OMIM, were used to provide an additional set of disease synonyms.

The strategy used to compare associated DO terms to disease names is based on a combination between an automated association process (comparison algorithm) and a manual curation. The former produces relatively low-quality associations derived without human validation, instead the manual curation produces high-quality, specific annotations but it is very time consuming.

Once all data were collected and the formats were re-organised to be more suitable for the analysis, the list of disorders included in the DO was used to align the ontology to other external resources. Each DO concept was mapped to each external database first by running a Perl script designed expressly for a term to term comparison. In order to perform the comparison process in all databases, the script was adjusted to be applied to different file format inputs. Automated processing was focused on the principle of reducing as much the false negatives as possible accepting meanwhile limited stringency on the false positives. This initial approach allowed maximising the identification of possible synonyms from the beginning devoting accuracy to the manual step. Being in the context of standard definitions and not of the natural language, intended in its widest accepted meaning, no sophisticated learning algorithms were necessary to make comparisons. The automated comparison method was based on the application of simple rules to score the level of identity between sentences, also taking into consideration some semantic content of composing words when possible. Similar definitions were considered synonyms if at the first instance they responded to the following condition: $I >= \text{int}(K/2)$ where $K=T-N$ (I=Identities, T=total number of words, N= words not relevant). Conjunctions, generic medical words and order of terms were considered either irrelevant for the identification of synonyms of diseases or negatively correlated to the level of identity to be calculated. Main limitations of this comparison approach were the impossibility to spot synonyms when definitions contained different words with the same meaning and also when completely different definitions of the same disease existed (e.g. depressive disorder and major depression).

When a DO term has been mapped successfully to a disease name present in one of the source databases, all its synonyms were extracted. The next step corresponded to the accurate curation of the results, addressing to reduce the number of false positives (Figure 5.5).

The database entries matching to a DO term were considered synonyms of the given disease and linked to the corresponding DO term. A main vocabulary was created, where almost all the diseases described in the external databases are associated to at least one DO term with a unique identifier.

The highest number of exact matches was found between the DO and PharmGKB database. A total of 2,633 exact matches between these two resources were obtained, e.g. osteoporosis (DOID: 11476 and GKB:PA445190), and rheumatoid arthritis (DOID:7148 and GKB:PA443434).

	A	B	C	D	E
	Associations	Total matches (column C+D)	Identity matches	Manual curated matches	Unmatched
PharmGKB (3,998)	186,894	2,866	2,633	233	1,132
GAD (5,635)	3,901	2,700	424	2,276	2,935
OMIM (4,121)	27,016	2,084	184	1,900	2,037

Table 5.1: Results of the comparison between DO and the three resources (the number in brackets correspond to total number of terms for each databases). Column A: total number of associations generated by the script used for the comparison. Column B: sum of totals in column C and D. Column C: total number of identities between the DO name and the name or synonyms in the other database. Column D: total number of matches found after the manual curation.

Table 5.1 recapitulates the number of matches between DO and external databases. Column A represents the total number of associations generated by the script used for the comparison. The associations, including also the false positives, were redundant and required a manual curation process. For instance, in the case of PharmGKB, the script found 186,894 possible

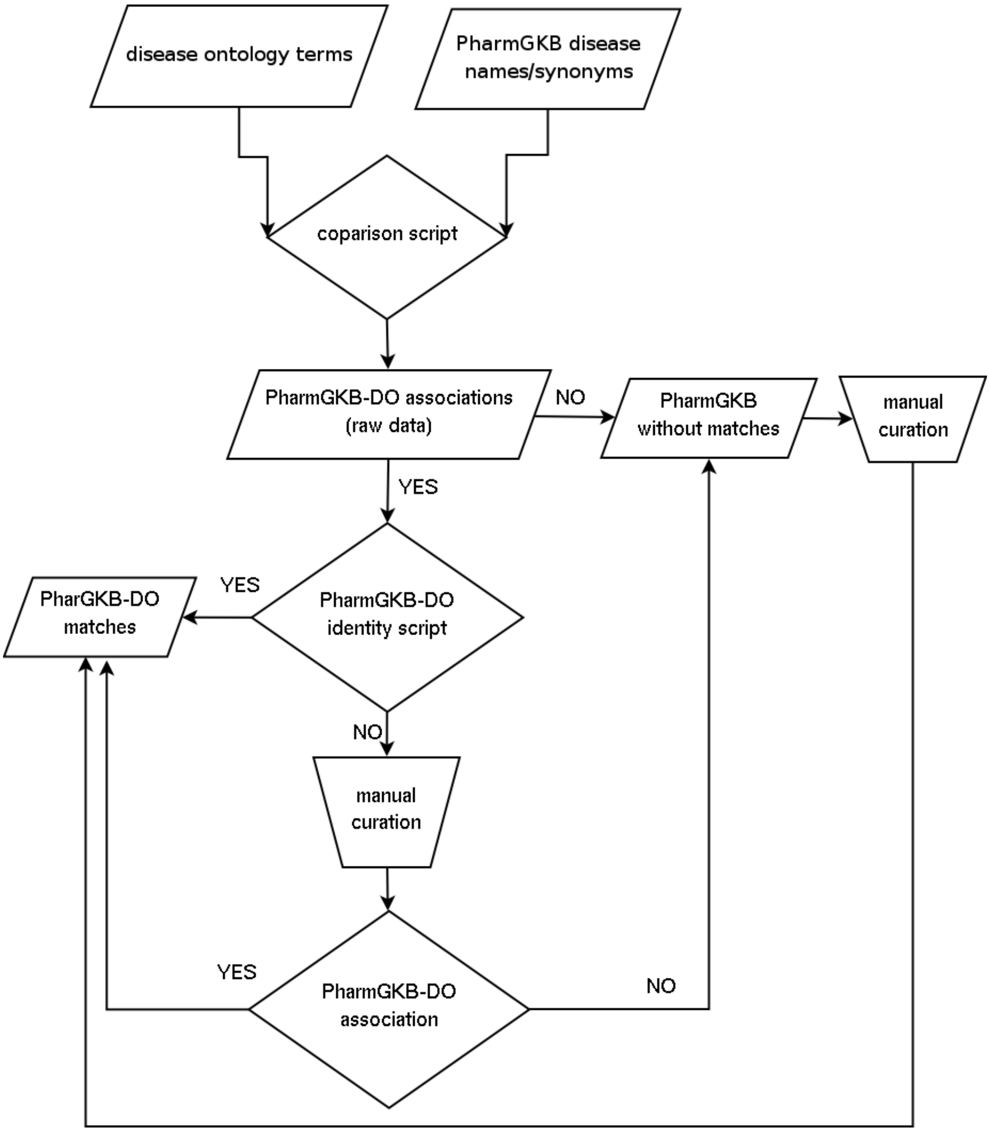


Figure 5.5: Overview of the method. The input files of the method corresponds to two lists of disease names and synonyms: one from the Disease Ontology and one among GAD, OMIM and PharmGKB. In this example, the DO dictionary is augmented by synonyms provided by PharmGKB. After the initial filtering using comparison alghorithm, a manual curation have been applied to correrct the result and find additional associations.

redundant results that corresponded to 2,976 non-redundant associations. After the manual curation of this large set of data, 2,866 were correctly matched by the script (column B). The total matches are derived from the addition of the matches in column C and D. Column C shows the identities between the DO name and the name or synonyms in the other database; column D shows the matches found after the manual curation. The highest overlap (71,68%) is found between the DO and PharmGKB database.

The DO terms presented in the high-level of the ontology hierarchy are common to all databases e.g. osteoporosis (DOID:11476; PharmGKB:PA445190; OMIM:166710). The low-level DO terms, which refer to more specific disease classes, have been found in at least one external database.

5.3.3 Finding relationships between genes and diseases

Gene information was retrieved and downloaded from the NCBI FTP site[15]. The file with all human gene-based information was parsed to collect data from specific fields and the extracted information was implemented in a MySQL database (Figure 5.9). Gene annotation data were obtained from the GOA gene association files[16] containing the GO assignment for the proteins of the non-redundant human proteome set.

Disease names and synonyms of the general disease vocabulary were used for searching in the gene association file obtained from the databases. All matches between the two files were collected in a gene annotation file where genes are associated with one or more DO disease with a unique DO ID (Figure 5.6).

5.3.4 Compilation of the drug dictionary

The list of drugs used in our database was compiled from two resources, DrugBank and PharmGKB. DrugBank was the first database used for selecting relevant information for the compilation of the drug dictionary. This resource is based on the 'active principle' or active ingredient set of drugs; although some drugs are currently in the queue of 'to be added' drugs (e.g. nimesulide).

[15] ftp://ftp.ncbi.nlm.nih.gov/gene/DATA/gene_info.gz

[16] ftp://ftp.ebi.ac.uk/pub/databases/GO/goa/HUMAN/gene_association.goa_human.gz

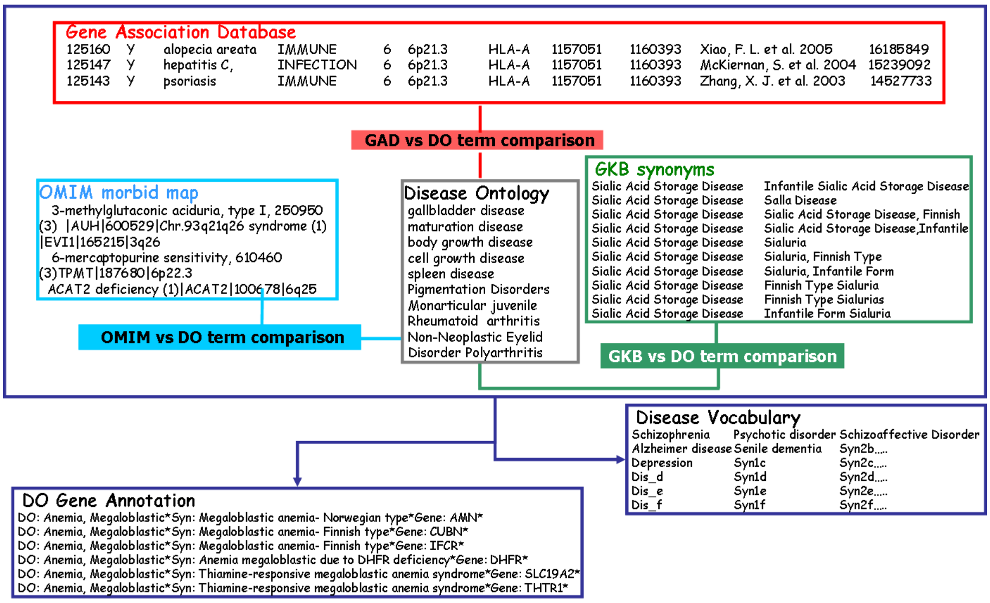

Gene Association Database
125160 Y alopecia areata IMMUNE 6 6p21.3 HLA-A 1157051 1160393 Xiao, F. L. et al. 2005 16185849
125147 Y hepatitis C, INFECTION 6 6p21.3 HLA-A 1157051 1160393 McKiernan, S. et al. 2004 15239092
125143 Y psoriasis IMMUNE 6 6p21.3 HLA-A 1157051 1160393 Zhang, X. J. et al. 2003 14527733
GAD vs DO term comparison
OMIM morbid map
3-methylglutaconic aciduria, type I, 250950
(3) |AUH|600529|Chr.93q21q26 syndrome (1)
|EVI1|165215|3q26
6-mercaptopurine sensitivity, 610460
(3)TPMT|187680|6p22.3
ACAT2 deficiency (1)|ACAT2|100678|6q25
OMIM vs DO term comparison
Disease Ontology
gallbladder disease
maturation disease
body growth disease
cell growth disease
spleen disease
Pigmentation Disorders
Monarticular juvenile
Rheumatoid arthritis
Non-Neoplastic Eyelid
Disorder Polyarthritis
GKB synonyms
Sialic Acid Storage Disease Infantile Sialic Acid Storage Disease
Sialic Acid Storage Disease Salla Disease
Sialic Acid Storage Disease Sialic Acid Storage Disease, Finnish
Sialic Acid Storage Disease Sialic Acid Storage Disease,Infantile
Sialic Acid Storage Disease Sialuria
Sialic Acid Storage Disease Sialuria, Finnish Type
Sialic Acid Storage Disease Sialuria, Infantile Form
Sialic Acid Storage Disease Finnish Type Sialuria
Sialic Acid Storage Disease Finnish Type Sialurias
Sialic Acid Storage Disease Infantile Form Sialuria
GKB vs DO term comparison
Disease Vocabulary
Schizophrenia Psychotic disorder Schizoaffective Disorder
Alzheimer disease Senile dementia Syn2b.....
Depression Syn1c Syn2c.....
Dis_d Syn1d Syn2d.....
Dis_e Syn1e Syn2e.....
Dis_f Syn1f Syn2f.....
DO Gene Annotation
DO: Anemia, Megaloblastic*Syn: Megaloblastic anemia- Norwegian type*Gene: AMN*
DO: Anemia, Megaloblastic*Syn: Megaloblastic anemia- Finnish type*Gene: CUBN*
DO: Anemia, Megaloblastic*Syn: Megaloblastic anemia- Finnish type*Gene: IFCR*
DO: Anemia, Megaloblastic*Syn: Anemia megaloblastic due to DHFR deficiency*Gene: DHFR*
DO: Anemia, Megaloblastic*Syn: Thiamine-responsive megaloblastic anemia syndrome*Gene: SLC19A2*
DO: Anemia, Megaloblastic*Syn: Thiamine-responsive megaloblastic anemia syndrome*Gene: THTR1*

Figure 5.6: The general disease vocabulary has been populated using GAD, OMIM and PharmGKB data. This vocabulary was used for searching in the gene association file obtained from the databases. All matches between the two files were collected in a gene annotation file where genes are associated with one or more DO diseases with a unique DO ID.

Relevant data from several fields were extracted from each DrugCard entry, among them:

- generic name: standard name of drug as provided by drug manufacturer;
- brand name and synonym: alternate names of the drug, brand names from different manufacturers;
- brand name mixtures: brand names and composition of mixtures that include the drug described in the DrugCard file;
- indication: description or common names of diseases that the drug is used to treat;
- drug target(s) name: name of the protein or macromolecule (or other small molecule) that the drug is supposed to act upon. Some drugs act on multiple targets, so these fields may be repeated several times, reflecting the number of drug targets that a specific drug may have;

- drug target(s) gene name: gene name of drug target;
- drug target(s) synonyms: alternate names (protein names, abbreviations, etc.) of the drug target;
- other fields such as ChEBI ID, CAS Registry Number and PharmGKB ID.

Starting from the drug generic name list, Perl scripts were run on PharmGKB database to augment the list with other drug names or synonyms in order to populate a general dictionary in which each generic drug name is associated to all possible synonyms and to mixtures that include the drug. A mixture might be associated to more than one drug e.g. Ana-Kit is a mixture composed of Chlorpheniramine (APRD00001) and Epinephrine (APRD00450).

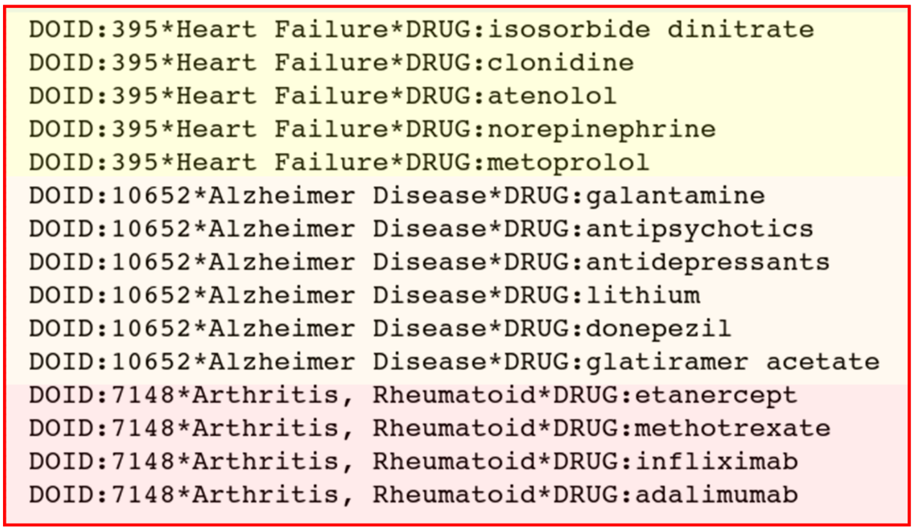
A total of 1,349 drug active principle names and 24,303 synonyms have been identified, each of them have an average of 19 associated synonyms (Table 5.2).

Active principle (total)	Synonym (average)	DrugBank	PharmGKB	In common	Mixture
1,349	19	216	149	984	253

Table 5.2: A total of 1,349 drug active principle names with an everage of 19 synonyms have been identified. 216 drugs are from DrugBank, 149 from PharmGKB and 984 are in common. The total number of synonym is 24,303 and of mixture is 253.

5.3.5 Characterising relations between drugs and diseases

The next step was to extract and characterise the relationships between drugs and diseases. The *Indication* field in DrugBank provides a description of the possible uses of the drug for the treatment of specific disorders. Unfortunately this definition does not contain a standard name but it is derived by natural language. Therefore finding the association between drug-disease has been more complicated because of the large number of possible false positives. In order to address this problem, Perl scripts were developed and performed to be as predictive as possible; their output files have been manually curated, increasing the level of accuracy of the relationships. As a result, 888 drugs from DrugBank have been associated to 801 DO diseases (Figure 5.7).



```
DOID:395*Heart Failure*DRUG:isosorbide dinitrate
DOID:395*Heart Failure*DRUG:clonidine
DOID:395*Heart Failure*DRUG:atenolol
DOID:395*Heart Failure*DRUG:norepinephrine
DOID:395*Heart Failure*DRUG:metoprolol
DOID:10652*Alzheimer Disease*DRUG:galantamine
DOID:10652*Alzheimer Disease*DRUG:antipsychotics
DOID:10652*Alzheimer Disease*DRUG:antidepressants
DOID:10652*Alzheimer Disease*DRUG:lithium
DOID:10652*Alzheimer Disease*DRUG:donepezil
DOID:10652*Alzheimer Disease*DRUG:glatiramer acetate
DOID:7148*Arthritis, Rheumatoid*DRUG:etanercept
DOID:7148*Arthritis, Rheumatoid*DRUG:methotrexate
DOID:7148*Arthritis, Rheumatoid*DRUG:infliximab
DOID:7148*Arthritis, Rheumatoid*DRUG:adalimumab
```

Figure 5.7: Example of associations between DO diseases and drugs.

A drug can be used for the treatment of several diseases. For example Chlorpheniramine (DrugBankID: APRD00001), used for the treatment of rhinitis, urticaria, allergy, common cold, asthma and hay fever, has been associated to six DO entries (Table 5.3). Chlorpheniramine has been associated to 'hypersensitivity' because the disease name 'allergy' is a synonym of the DO term name 'hypersensitivity'.

Drug	**Indication**	**DO Association**
APRD00001	For the treatment of rhinitis, urticaria, allergy, common cold, asthma and hay fever.	rhinitis DOID:4483
		asthma DOID:2841
		urticaria DOID:1555
		common cold DOID:10459
		hypersensitivity DOID:1205 (synonym: allergy)
		hay fever DOID:14030

Table 5.3: Chlorpheniramine (DrugBankID:APRD00001), used for the treatment of rhinitis, urticaria, allergy, common cold, asthma and hay fever, has been associated to six DO entries.

Thus, using our general disease vocabulary, it was possible to find the main DO term name. At the same time, the same diseases might be treated using different drugs (Table 5.4).

Drug	**Indication**	**Association**
Divalproex (APRD00066)	For treatment and management of seizure disorders, mania, and prophylactic treatment of migraine headache.	migraine DOID:6364
Rizatriptan (APRD00008)	For treatment of acute migraine attacks.	migraine DOID:6364

Table 5.4: Divalproex and Rizatriptan are both used for the treatment of migraine DOID:6364.

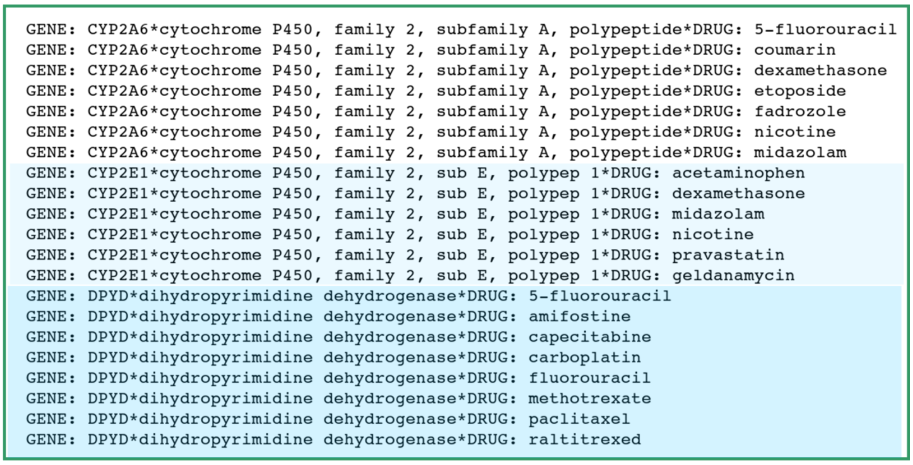
5.3.6 Finding relationships between drugs and target genes

A gene or a protein might be involved in a disease because it is a key molecule involved in a particular metabolic or signalling pathway that is specific to a disease condition or pathology. But a protein might also be considered a key molecule in the treatment of a disease where the inhibition of this protein by a drug might stop the functioning of the pathway in the diseased state. Drug Bank database also provides information about target genes. These data were used to characterise relationships between drugs and the respective target genes. The result is a file where each drug is linked with specific targets that are genes listed in the GENE table of the MySQL database containing information about UniprotID, alternative gene name and so on (Figure 5.8).

All data about genes, drugs and diseases have been implemented in a MySQL database (Figure 5.9).

5.4 Possible applications

Considering the large number of diseases, it is impossible for a single researcher to get an overview of the genes already studied. Suppose a researcher is performing a microarray study for a specific disorders, e.g. multiple sclerosis. Analysis of differential gene expression might reveal a list of several candidate genes. Some of them are known to the researcher and well-



```
GENE: CYP2A6*cytochrome P450, family 2, subfamily A, polypeptide*DRUG: 5-fluorouracil
GENE: CYP2A6*cytochrome P450, family 2, subfamily A, polypeptide*DRUG: coumarin
GENE: CYP2A6*cytochrome P450, family 2, subfamily A, polypeptide*DRUG: dexamethasone
GENE: CYP2A6*cytochrome P450, family 2, subfamily A, polypeptide*DRUG: etoposide
GENE: CYP2A6*cytochrome P450, family 2, subfamily A, polypeptide*DRUG: fadrozole
GENE: CYP2A6*cytochrome P450, family 2, subfamily A, polypeptide*DRUG: nicotine
GENE: CYP2A6*cytochrome P450, family 2, subfamily A, polypeptide*DRUG: midazolam
GENE: CYP2E1*cytochrome P450, family 2, sub E, polypep 1*DRUG: acetaminophen
GENE: CYP2E1*cytochrome P450, family 2, sub E, polypep 1*DRUG: dexamethasone
GENE: CYP2E1*cytochrome P450, family 2, sub E, polypep 1*DRUG: midazolam
GENE: CYP2E1*cytochrome P450, family 2, sub E, polypep 1*DRUG: nicotine
GENE: CYP2E1*cytochrome P450, family 2, sub E, polypep 1*DRUG: pravastatin
GENE: CYP2E1*cytochrome P450, family 2, sub E, polypep 1*DRUG: geldanamycin
GENE: DPYD*dihydropyrimidine dehydrogenase*DRUG: 5-fluorouracil
GENE: DPYD*dihydropyrimidine dehydrogenase*DRUG: amifostine
GENE: DPYD*dihydropyrimidine dehydrogenase*DRUG: capecitabine
GENE: DPYD*dihydropyrimidine dehydrogenase*DRUG: carboplatin
GENE: DPYD*dihydropyrimidine dehydrogenase*DRUG: fluorouracil
GENE: DPYD*dihydropyrimidine dehydrogenase*DRUG: methotrexate
GENE: DPYD*dihydropyrimidine dehydrogenase*DRUG: paclitaxel
GENE: DPYD*dihydropyrimidine dehydrogenase*DRUG: raltitrexed
```

Figure 5.8: Example of associations between genes and drugs.

discussed in the literature. Others are new in this context. For instance, a search for the official gene name PTPRC in context of multiple sclerosis retrieves 82 PubMed abstracts containing also different synonyms of PTPRC such as CD45 and Ly5. Finding useful information about genes, drugs and/or diseases in which the researcher is interested, is a time consuming and tedious job. Our database will be potentially useful in analysis of microarray results because it will reduce the number of genes to analyse, and save a literature search.

The aim of our tool is to provide the researcher with a quick overview of potential links between genes, drugs and diseases. The Disease Ontology successfully connects data of diseases to genes and drugs from different databases. It is a source of information about genes, diseases and drugs developed using several very up to date resources, some of them also manually curated and with detailed descriptions. As soon as the web interface is available, the user will be able to browse the list of disease concepts or query the database to search disease terms or genes of interest. Intersections of gene sets for different disease concepts which are of particular interest, will also be feasible. For instance, it is widely recognised that persons suffering one type of mood disorder have an increased susceptibility to other mood disorders. With this approach it would be possible to find an answer to several questions such as how many genes are associated to depression, and among them how many are also target genes for a therapeutic drug used in the treatment of a mood disorders. Moreover finding gene-drug relationships can form the basis of more detailed pharmacogenetic experimental investigations.

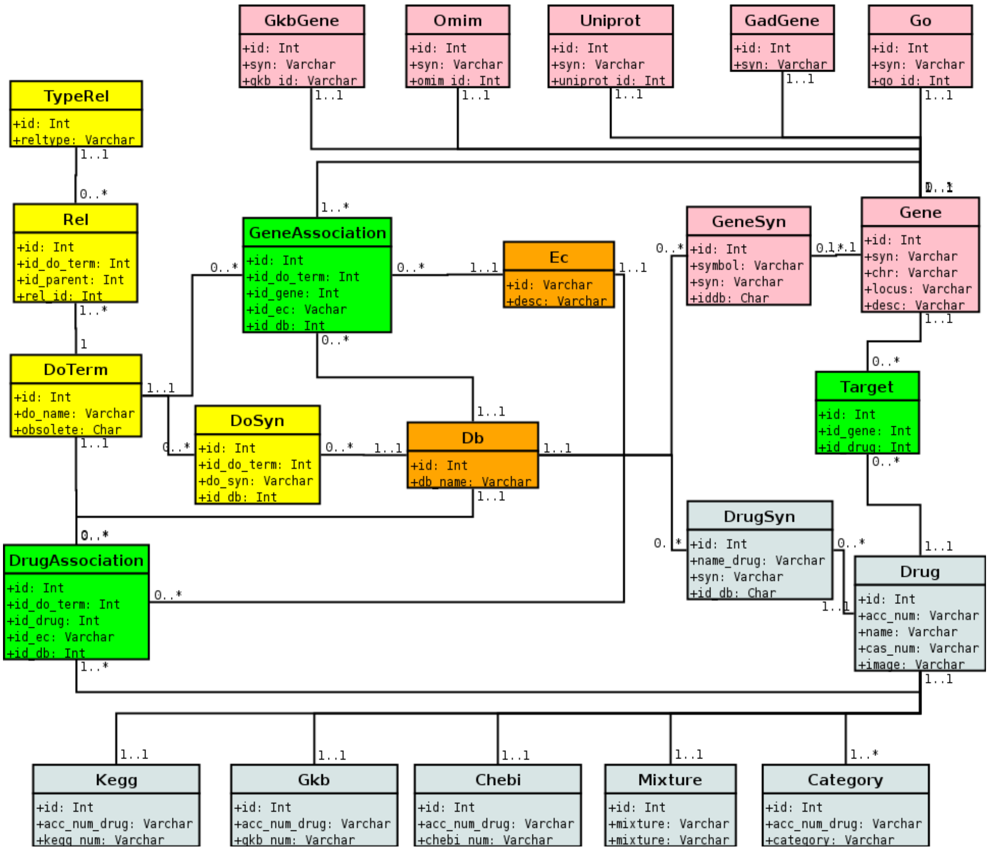


Figure 5.9: Schema of the MySQL database.

5.5 Future perspective

A Web interface will be developed that will include user registration and comment forms, and basic and advanced query options to access data for genes, drugs, disorders and their relationships. BioMart is a possible solution for the web interface.

BioMart[17] is a query-oriented data management system developed jointly by the European Bioinformatics Institute (EBI) and Cold Spring Harbor Laboratory (CSHL). BioMart simplifies the task of creation and maintenance of advanced query interfaces backed by a relational database and it is particularly suited for providing the 'data mining' like searches of complex descriptive (e.g. biological) data. It can work with existing data repositories by converting them to a required BioMart format, as well as newly created databases.

[17] http://www.biomart.org/

At the completion of the interface development phase, the DO annotation database will be made accessible to collaborators and participating members of the scientific community to evaluate its functionality and to test the system. It will be necessary to work on case studies to validate the system and to improve it on the basis of the feedback from the community. The annotation process will continue increasing the number of DO annotations collected in the database. At the same time the vocabulary will be improved with additional synonyms, and supplementary information such as data on pathways and reaction will be integrated.

Moreover, it would be necessary to develop a new strategy to keep this resources up to date even as knowledge of genes, drugs and diseases is accumulating and changing; and in addition to design a data curation plan helpful for speeding up the process of manual curation that it is currently very time consuming.

Chapter 6

Tools

In this chapter are listed the main bioinformatics resources and tools that have been used for this project. They are an ontology editor (paragraph 6.1), several ontology browsers (paragraph 6.2) and ontology-based literature searching tools (paragraph 6.3).

Some tools (e.g. OBO-Edit and AmiGO) are developed within the Gene Ontology Consortium. They are continuously improved and expanded and the users can receive help in using them by emailing the GO helpdesk. Many additional tools have been created outside the consortium for use with the GO and other ontologies as well. They can be useful for searching and browsing ontologies (e.g. OLS), for gene annotation (e.g protein2go) and for gene expression microarray analysis and a variety of other applications.

6.1 Ontology editor

6.1.1 OBO-Edit, a tool for ontology development

OBO-Edit (OE) [31] is a well-known representative for the GO tools and it has been used for editing the Gene Ontology in several phases of our project (e.g. during the implementation of muscle specific terms and in the re-organisation of response to drug node). Now it is being used to browse, query and edit Gene Ontology and any other vocabulary that has an OBO file format. OE is an open source, platform-independent ontology editor written in Java, maintained by the Gene Ontology Consortium. It is developed by John Day-Richter, as part of the Berkeley Bioinformatics and Ontologies Project and it is available via SourceForge web site[1].

It features an easy to use editing interface and powerful search capabilities. In the ontology editor, the ontology is displayed as normalized tree view. Each line in the tree view represents a relationship between terms, with an arrow indicating the direction of the relationship (Figure 6.1[A] on the right). The Graph Viewer plugin offers a more traditional graphical view of the GO graph (Figure 6.1[A] on the left). Finally, the DAG viewer plugin displays every path from a selected term to the ontology root. Since a term may have any number of relationships to other terms, a term may appear several times in the tree view (Figure 6.1[A] on the right).

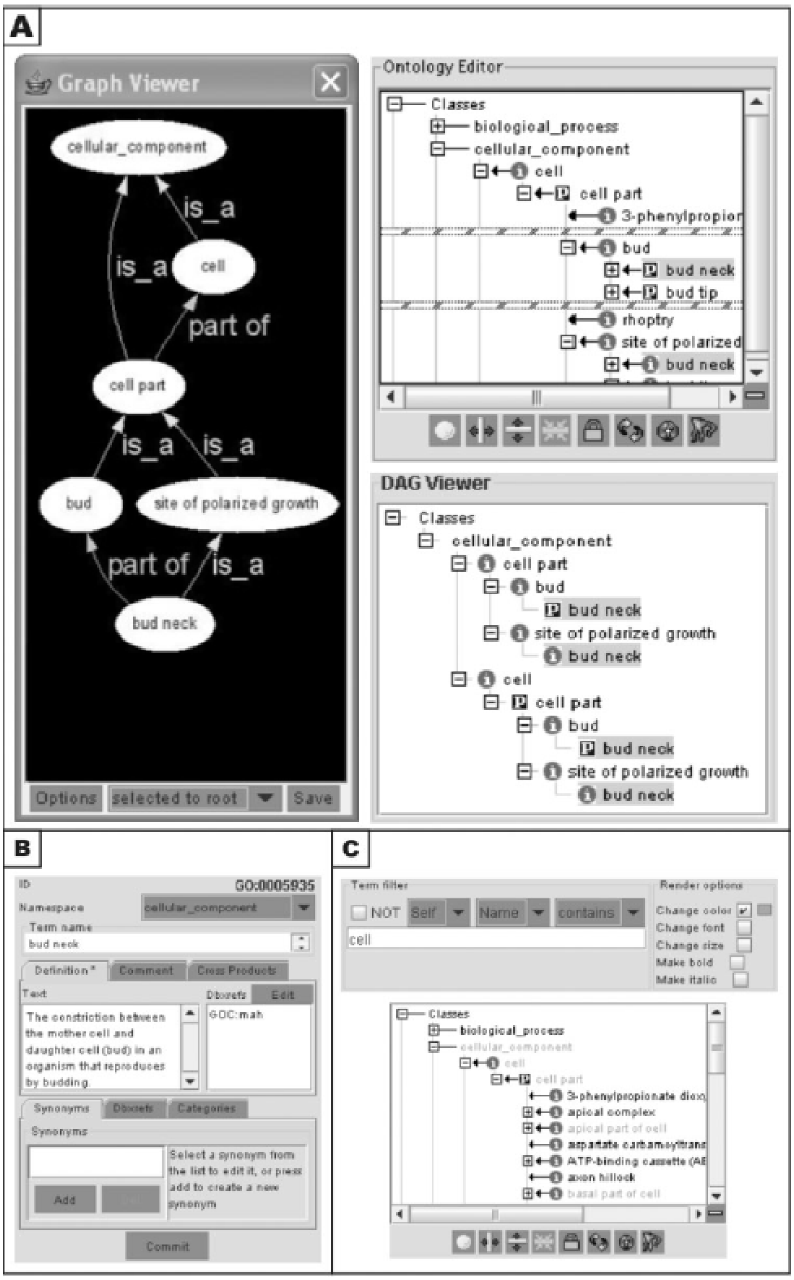
Structural ontology edits can be made from the ontology editor view or using the Edit Menu. Textual additions, such as the term name and definition, can be edited using the Term Editor Component (Figure6.1[B]). OBO-Edit provides also a filter interface to apply basic and also compound searches that return a list of all the terms that match a given filter. Using filters, it can be possible to alter the display and highlight matching terms by changing fonts, colours and style (Figure 6.1[C]). For example, all the GO terms implemented during the Muscle Content Meeting were pink coloured to differ from the other terms.

The most recent version of OBO-Edit offers sophisticated filtering, editing, reasoning and error checking capabilities. Built-in checks help maintain the correct ontology structure. To use OBO-Edit for this project, it was necessary to set up parameters, IDs, filters suitable for our purposes. First of all, GO community assigned a GO ID number range for each activity: one for nervous system-muscle editing and one for ‘response to drug’ work. When a new term is created, OBO-Edit determines what id generation rule to use and such rule contains information about what new ids should look like. The ID rules was specified using the ID Manager Plugin and set up like below:

- Response to drug ID profile: from GO:0014000 to GO:0014300;
- Muscle ID profile: from GO:0014301 to GO:0014999.

A diverse group of biologists contributes to improving OBO-Edit, ensuring that it is versatile, configurable and user friendly. Its development is overseen by the OBO-Edit Working Group (I am a member of this group), a team of OBO-Edit users from a dozen sites around

[1] https://sourceforge.net/project/showfiles.php?group_id=36855&package_id=192411



ID GO:0005935

Namespace cellular_component

Term name

bud neck

Text

The constriction between the mother cell and daughter cell (bud) in an organism that reproduces by budding.

Dbxrefs

GOC:mah

Synonyms

Select a synonym from the list to edit it, or press add to create a new synonym

□ NOT Self Name contains

cell

Change color ☑

Figure 6.1: Key components of the OBO-Edit interface. (A) Visualization. The Ontology Editor, DAG Viewer and Graph Viewer provide three different views of an ontology term. The Ontology Editor also serves as the editing interface for adding or removing terms and relationships. (B) Term Editor. The Term Editor panel allows users to edit the term name, definition, synonyms and other textual metadata. (C) Filtering. Criteria specified in the filter interface can be used to highlighte matching terms using a renderer [31].

the world. The Working Group is responsible for directing future OBO-Edit development, bug testing, and the generation of documentation. This working group meets every week on Thursday to discuss about new OE features, bugs and other OE related topics using Webex technology. In addition, an OBO-Edit wiki page[2] is provided as a central repository for OBO-Edit developer info, announcements, and technical information. For the Muscle Meeting participants, an Italian OBO-Edit tutorial[3] was provided by me.

6.2 Ontology browsers

6.2.1 Ontology Lookup Service

The number of biomedical ontologies and controlled vocabularies developed by each scientific community is growing very fast. Unfortunately, each group tends to create its own locally available ontologies and to develop its own online browser to query these ontologies (among them, the afterwritten AmiGO and QuickGO browsers).

The Ontology Lookup Service (OLS)[4] integrates publicly available ontologies in OBO format into a single database [81]. All modified ontologies are kept up to date daily and available to be browsed.

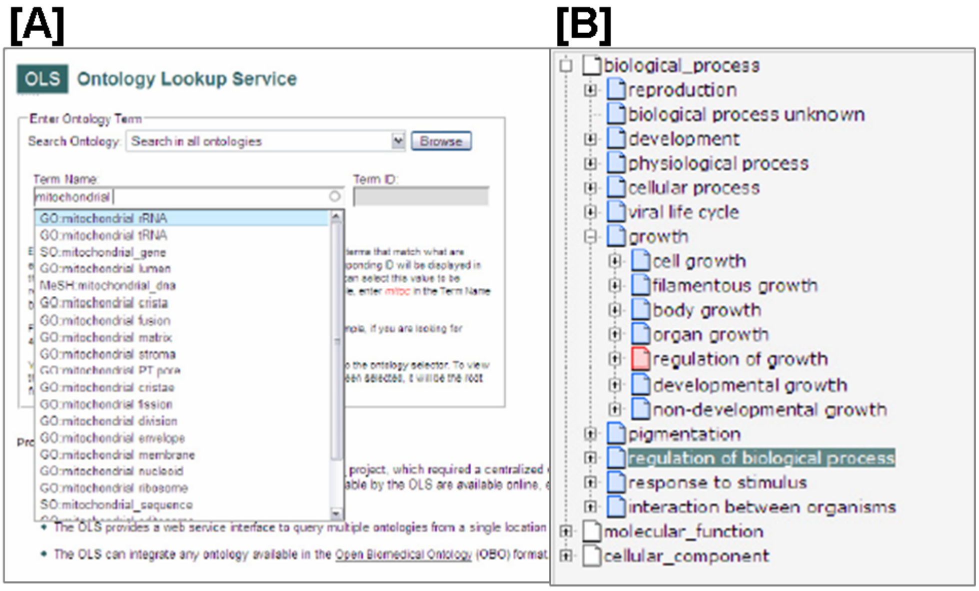
It is possible to query a selected ontology or all of them to obtain information on a single term; the query is performed on the preferred term name as well as on any synonyms. The search process is made easier by an auto-completion mechanism: a collection of suggested terms that match what is entered are displayed in a drop-down menu. If you select one from the pull-down list, its corresponding ID will be displayed in the form (Figure 6.2[A]).

Users can browse the selected ontology, as well as a subset of the ontology, using a dynamically generated tree structure very similar to the AmiGO and QuickGO ones (Figure 6.2[B]).

[2] wiki.geneontology.org/index.php/OBO-Edit

[3] http://www.geneontology.org/GO.teaching.resources.shtml#tut

[4] http://www.ebi.ac.uk/ontology-lookup/



- The OLS provides a web service interface to query multiple ontologies from a single location
- The OLS can integrate any ontology available in the Open Biomedical Ontology (OBO) format

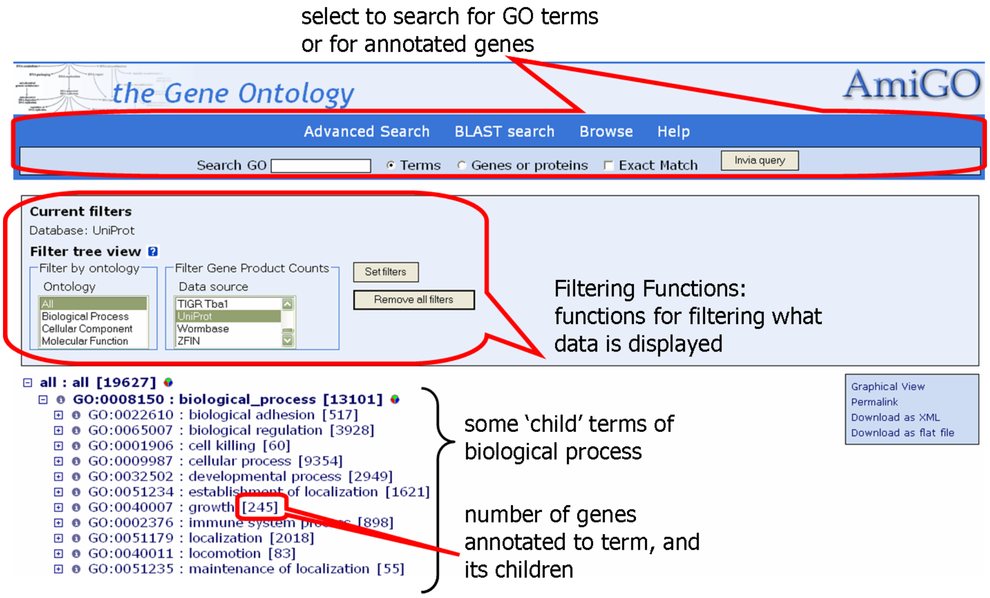
Figure 6.2: (A) The search process is made easier by an autocompletion mechanism. (B) Users can browse the selected ontology, as well as a subset of the ontology, using a dynamically generated tree structure.

6.2.2 AmiGO: GO Consortium browser

AmiGO[5] is a web resource developed within the GO Consortium [2, 82] and it has been used for searching and browsing the Gene Ontology terms and gene products annotations.

AmiGO displays the GO terms as a tree: the branches of the tree can be expanded by clicking the + icons next to the terms. AmiGO also allows searching for a particular GO term using the search box. The number at the end of each term is the number of genes annotated to that GO term and its child terms. AmiGO allows also searching by gene product, and displays all of the GO terms with which a gene is associated from all species unless filters are applied (Figure 6.3). Recently it has been enhanced by the addition of new navigation and search options, an improved display of search results, and a simplified user interface [29]. It is also possible to perform a sequence identity BLAST search and view the GO term associations for the genes or proteins returned.

[5] http://amigo.geneontology.org



select to search for GO terms
or for annotated genes

Filtering Functions:
functions for filtering what
data is displayed

```
⊟ all : all [19627]
  ⊟ GO:0008150 : biological_process [13101]
    ⊞ GO:0022610 : biological adhesion [517]
    ⊞ GO:0065007 : biological regulation [3928]
    ⊞ GO:0001906 : cell killing [60]
    ⊞ GO:0009987 : cellular process [9354]
    ⊞ GO:0032502 : developmental process [2949]
    ⊞ GO:0051234 : establishment of localization [1621]
    ⊞ GO:0040007 : growth [245]
    ⊞ GO:0002376 : immune system process [898]
    ⊞ GO:0051179 : localization [2018]
    ⊞ GO:0040011 : locomotion [83]
    ⊞ GO:0051235 : maintenance of localization [55]
```

some ‘child’ terms of
biological process

number of genes
annotated to term, and
its children

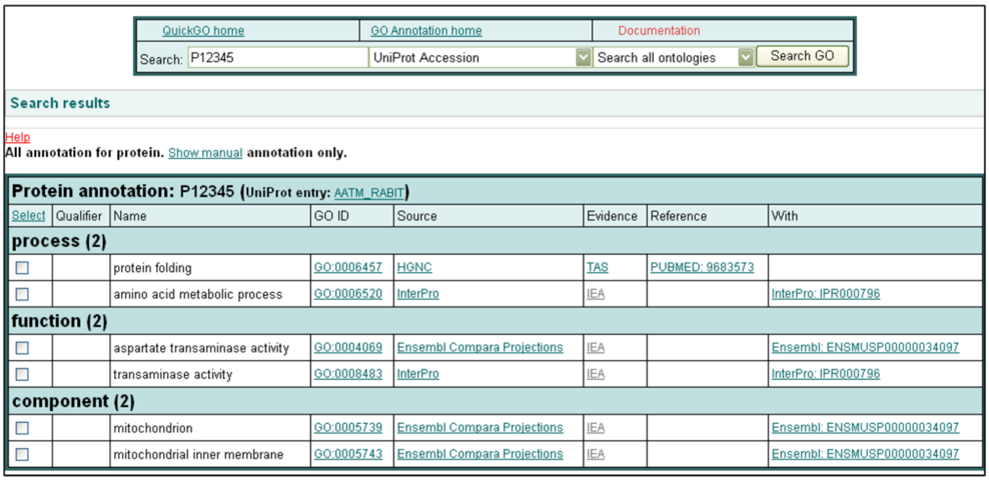
Figure 6.3: AmiGO search interface with the filtering function box and the treeview expanded showing also the number of gene annotated to GO:0040007.

6.2.3 QuickGO@EBI

QuickGO[6] is a ‘fast’ web-based browser developed at EBI to allow users to search and browse GO data and associated links to other data sets. In addition, QuickGO accesses the manually curated annotations and mappings of Swiss-Prot keywords, InterPro entries and the EC classification schemes to GO terms, as well as electronically and manually curated associations of GO terms to Swiss-Prot/TrEMBL entries (GOA).

It provides various search facilities: selecting the required search field, it is possible to query by GO ID, GO term name or synonym, Uniprot accession, InterPro ID, GO definition, Comments and so on. Querying by protein accession number shows all terms mapped to that Swiss-Prot entry and the source of each term association (Figure 6.4). QuickGO was extremely helpful during the process of annotation of genes involved in neuropsychiatric and neurodegenerative disorders. With QuickGO it is possible to display only manually assigned GO terms and check the annotations one by one.

[6] http://www.ebi.ac.uk/ego/



All annotation for protein. Show manual annotation only.

Protein annotation: P12345 (UniProt entry: AATM_RABIT)

Select	Qualifier	Name	GO ID	Source	Evidence	Reference	With
process (2)							
□		protein folding	GO:0006457	HGNC	TAS	PUBMED: 9683573	
□		amino acid metabolic process	GO:0006520	InterPro	IEA		InterPro: IPR000796
function (2)							
□		aspartate transaminase activity	GO:0004069	Ensembl Compara Projections	IEA		Ensembl: ENSMUSP00000034097
□		transaminase activity	GO:0008483	InterPro	IEA		InterPro: IPR000796
component (2)							
□		mitochondrion	GO:0005739	Ensembl Compara Projections	IEA		Ensembl: ENSMUSP00000034097
□		mitochondrial inner membrane	GO:0005743	Ensembl Compara Projections	IEA		Ensembl: ENSMUSP00000034097

Figure 6.4: Sample of the QuickGO display page showing all GO terms that have been mapped electronically and manually to Swiss-Prot entry P12345.

6.3 Further resources

Currently, to access information, it is a common practice to use a search engine, such as GoogleTM and to follow hyperlinks rather than reach for a reference book. This is part of the approach that has been followed for building our knowledge on gene-drug-disease and their possible representation within ontologies. During the development of this project, several sources have been exploited. In addition to tools developed to browse and retrieve data about GO and gene annotations, other resources were helpful to create and enrich knowledge on the studied biological domain, among them several dictionaries, encyclopaedias and ontologies such as Encyclopedia of Life Science[7], Dictionary of Anxiety and Panic Disorders[8] (ASAP), Medical Subject Headings[9] (MeSH), National Library of Medicine[10] and International classification of Diseases[11] (ICD9). Two sources, GOPubMed and iHOP, were very valuable for many activities.

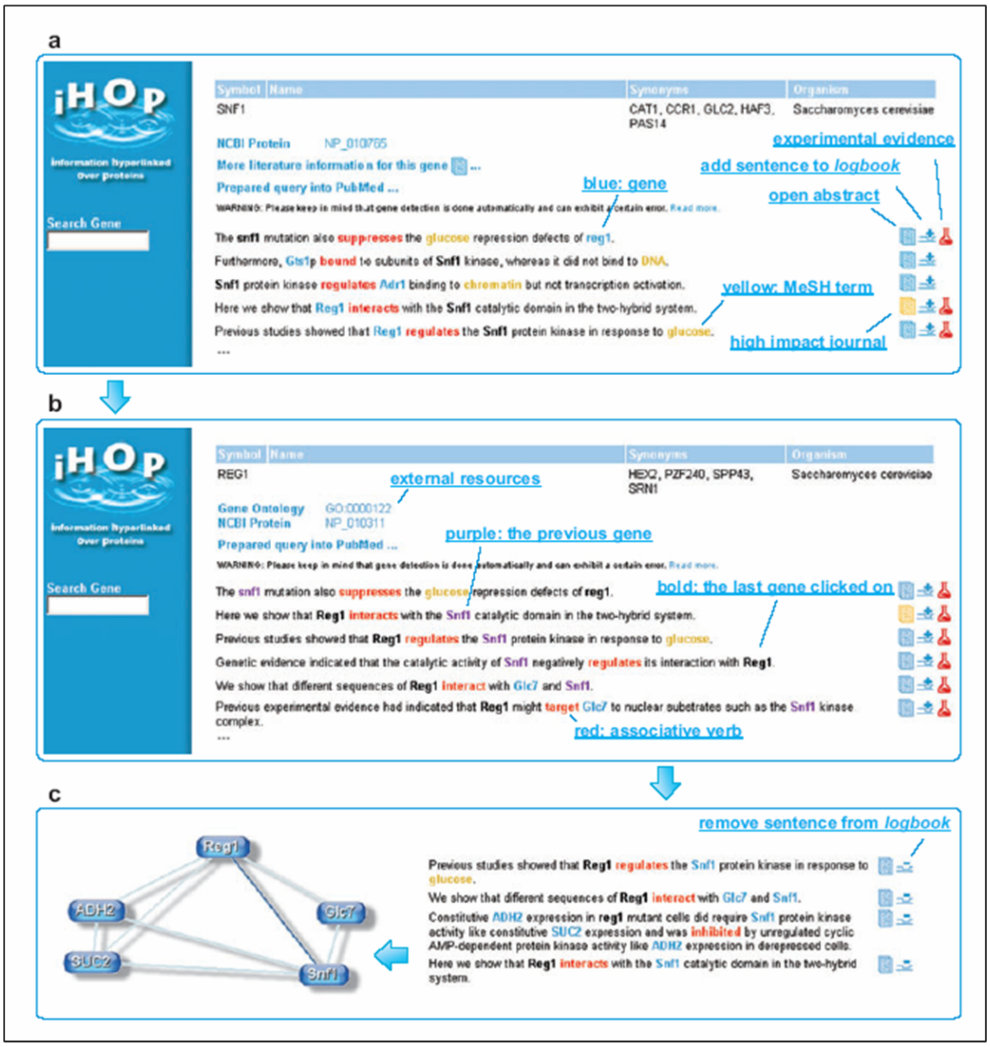
[7] http://www.els.net

[8] http://anxiety-panic.com/dictionary/en-main.htm

[9] http://www.nlm.nih.gov/mesh

[10] http://wwwcf.nlm.nih.gov

[11] http://icd9cm.chrisendres.com


a
iHOP
Information hyperlinked over proteins
Search Gene
experimental evidence
add sentence to *logbook*
blue: gene
open abstract
yellow: MeSH term
high impact journal
b
iHOP
Information hyperlinked over proteins
Search Gene
external resources
purple: the previous gene
bold: the last gene clicked on
red: associative verb
c
remove sentence from *logbook*
Reg1
ADH2
Glc7
SUC2
Snf1

Symbol	Name	Synonyms	Organism
SNF1		CAT1, CCR1, GLC2, HAF3, PAS14	Saccharomyces cerevisiae

NCBI Protein NP_010765

More literature information for this gene ...

Prepared query into PubMed ...

WARNING: Please keep in mind that gene detection is done automatically and can exhibit a certain error. Read more.

The **snf1** mutation also suppresses the glucose repression defects of reg1.

Furthermore, Gts1p bound to subunits of **Snf1** kinase, whereas it did not bind to DNA.

Snf1 protein kinase regulates Adr1 binding to chromatin but not transcription activation.

Here we show that Reg1 interacts with the **Snf1** catalytic domain in the two-hybrid system.

Previous studies showed that Reg1 regulates the **Snf1** protein kinase in response to glucose.

...

Symbol	Name	Synonyms	Organism
REG1		HEX2, PZF240, SPP43, SRN1	Saccharomyces cerevisiae

Gene Ontology GO:0000122
NCBI Protein NP_010311

Prepared query into PubMed ...

WARNING: Please keep in mind that gene detection is done automatically and can exhibit a certain error. Read more.

The snf1 mutation also suppresses the glucose repression defects of **reg1**.

Here we show that **Reg1** interacts with the Snf1 catalytic domain in the two-hybrid system.

Previous studies showed that **Reg1** regulates the Snf1 protein kinase in response to glucose.

Genetic evidence indicated that the catalytic activity of Snf1 negatively regulates its interaction with **Reg1**.

We show that different sequences of **Reg1** interact with Glc7 and Snf1.

Previous experimental evidence had indicated that **Reg1** might target Glc7 to nuclear substrates such as the Snf1 kinase complex.

...

Previous studies showed that **Reg1** regulates the Snf1 protein kinase in response to glucose.

We show that different sequences of **Reg1** interact with Glc7 and Snf1.

Constitutive ADH2 expression in **reg1** mutant cells did require Snf1 protein kinase activity like constitutive SUC2 expression and was inhibited by unregulated cyclic AMP-dependent protein kinase activity like ADH2 expression in derepressed cells.

Here we show that **Reg1** interacts with the Snf1 catalytic domain in the two-hybrid system.

Figure 6.5: (A) Example of a search for a gene or a protein of interest in iHOP. Information on such a gene and its interactions is provided in the form of sentences extracted directly from their source abstracts. Sentences that include proteins whose interaction is experimentally supported are highlighted and ranked higher. (B) All sentences associating A to B are ranked first when the users arrives at gene B from gene A. (C) Interesting sentences can be collected and dynamically represented as a graph.

6.3.1 Information Hyperlinked Over Proteins

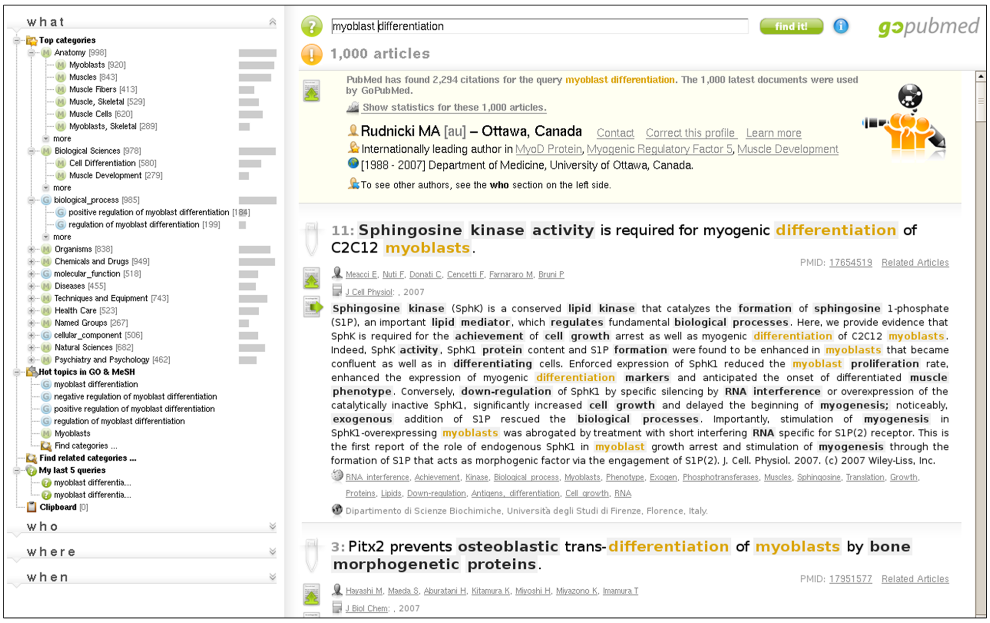
Information Hyperlinked Over Proteins (iHOP)[12] is a network of genes and proteins extended through the scientific literature, touching on phenotypes, pathologies and gene functions [83]. By using genes and proteins as hyperlinks between sentences and abstracts, the information in PubMed is converted into one navigable resource. iHOP is an online service that provides a gene-guided network as a natural way of accessing millions of PubMed abstracts bringing all the advantages of the internet to scientific literature research (Figure 6.5).

6.3.2 GOPubMed

GOPubMed[13] was very helpful in the gene annotation process but also in the improvement of GO contents. It allowed identification of new GO terms to be submitted and creation of new GO annotations. GOPubMed is an ontology-based literature search engine. It submits keywords to PubMed, extracts GO terms from the retrieved abstracts, and presents the resulting ontology for browsing (Figure 6.6). The ontology is the minimal subset of GO, which comprises all the GO terms found in the documents [38].

[12] http://www.ihop-net.org/UniPub/iHOP/

[13] http://www.gopubmed.org



myoblast differentiation

1,000 articles

PubMed has found 2,294 citations for the query myoblast differentiation. The 1,000 latest documents were used by GoPubMed.

Show statistics for these 1,000 articles.

Rudnicki MA [au] – Ottawa, Canada Contact Correct this profile Learn more

Internationally leading author in MyoD Protein, Myogenic Regulatory Factor 5, Muscle Development

[1988 - 2007] Department of Medicine, University of Ottawa, Canada.

To see other authors, see the **who** section on the left side.

11: **Sphingosine kinase activity** is required for myogenic differentiation of C2C12 myoblasts.

PMID: 17654519 Related Articles

Meacci E, Nuti F, Donati C, Cencetti F, Farnararo M, Bruni P

J Cell Physiol: , 2007

Sphingosine kinase (SphK) is a conserved **lipid kinase** that catalyzes the **formation** of **sphingosine** 1-phosphate (S1P), an important **lipid mediator**, which **regulates** fundamental **biological processes**. Here, we provide evidence that SphK is required for the **achievement** of **cell growth** arrest as well as myogenic differentiation of C2C12 myoblasts. Indeed, SphK **activity**, SphK1 **protein** content and S1P **formation** were found to be enhanced in myoblasts that became confluent as well as in **differentiating** cells. Enforced expression of SphK1 reduced the myoblast **proliferation** rate, enhanced the expression of myogenic differentiation **markers** and anticipated the onset of differentiated **muscle phenotype**. Conversely, **down-regulation** of SphK1 by specific silencing by **RNA interference** or overexpression of the catalytically inactive SphK1, significantly increased **cell growth** and delayed the beginning of **myogenesis;** noticeably, **exogenous** addition of S1P rescued the **biological processes**. Importantly, stimulation of **myogenesis** in SphK1-overexpressing myoblasts was abrogated by treatment with short interfering **RNA** specific for S1P(2) receptor. This is the first report of the role of endogenous SphK1 in myoblast growth arrest and stimulation of **myogenesis** through the formation of S1P that acts as morphogenic factor via the engagement of S1P(2). J. Cell. Physiol. 2007. (c) 2007 Wiley-Liss, Inc.

RNA interference, Achievement, Kinase, Biological process, Myoblasts, Phenotype, Exogen, Phosphotransferases, Muscles, Sphingosine, Translation, Growth, Proteins, Lipids, Down-regulation, Antigens, differentiation, Cell growth, RNA

Dipartimento di Scienze Biochimiche, Università degli Studi di Firenze, Florence, Italy.

3: Pitx2 prevents **osteoblastic** trans-differentiation of myoblasts by **bone morphogenetic proteins**.

PMID: 17951577 Related Articles

Hayashi M, Maeda S, Aburatani H, Kitamura K, Miyoshi H, Miyazono K, Imamura T

J Biol Chem: , 2007

Figure 6.6: User interface of GOPubMed displaying the results for the query 'myoblast differentiation'. On the left, the part of the GO relevant to the query is shown, and on the right the abstracts of the selected GO term. After searching GOPubMed, the tool shows how many citations have been found for the query. The 1,000 latest documents are displayed by GOPubMed. The keywords and their synonyms are highlighted in orange.

Chapter 7

Conclusion

The introduction of high-throughput functional genomics techniques has resulted in an exponential increase in the accumulation of biological data. Currently, thousands of bioinformatics databases are available, containing heterogeneous data about genes, drugs and disorders. Therefore there is a major need to integrate and annotate data to make knowledge about diseases and disease-associated genes available to the scientific community. A fundamental component in the integration effort is the development and application of standardized ontologies for annotation. Over the past years, a lot of biomedical ontologies have been developed, including the Gene Ontology (GO), and these have been made available through the open biomedical ontology (OBO) project.

During my PhD studentship, I improved several of the ontologies that are used primarily as structured controlled terminologies and as data aggregators in the study of genes, drugs and diseases. In addition, I have developed a new method for data integration and data extraction of gene-drug-disease relations.

My work included significant changes in GO content and structure, resulting in the addition of hundreds of terms useful in the representation of muscle and nervous system biology in the three vocabularies. Although the improvement of the Gene Ontology in specific biological domains is essential, its utility has been increased by cross domain links to other OBO ontologies such as ChEBI, Cell Ontology and Disease Ontology.

These ontology development efforts provide valuable resources for annotation of genes related to the relevant biological domains, and the revised structure facilitates the interpretation of high throughput experiments (e.g. gene expression microarrays) in the area of neuromuscular and neurodegenerative disorders. Gene annotation wiki pages have been provided to help speed up the annotation process and especially to facilitate community participation in the Gene Ontology project

Part of the significance of this work is the achievement of bringing together ontology developers and community scientists, enabling the latest scientific knowledge and language to be captured in a rigorous computable form. Decisions on when a term can be inserted or modified are based on long discussions between GO curators and biological experts, in order to represent the information in both an ontologically and biologically correct way. To facilitate this, I organised a GO content meeting, bringing together biology and ontology experts, in a location that was very much dominated by bench scientists. As the work was carried out in a bench science institute, rather than in a distant bioinformatics centre, we were able to systematically improve the GO whilst achieving the all-important buy-in of the grass roots biology community, that is so essential for the proper functioning of an ontology-based information system. This work made a very significant contribution to the representation of muscle biology in the gene ontology structure.

Considering the large number of diseases we seek to effectively treat, systems to handle drug data are also a high-priority requirement. For this reason, I contributed to the reorganisation of the structure of GO 'response to drug' node mirroring the structure of a chemical controlled vocabulary, the ChEBI ontology. Following the ChEBI ontology, almost all 'response to drug' terms have been classified on a rational basis applying a chemical classification rather than a pharmacological one. The new structure will help in the annotation of genes that are differentially expressed in response to a drug treatment, and consequently, the set of new annotation data will provide additional links between gene expression profiles and effects of drugs in regulating specific biological processes.

Finally, I developed a resource to find preliminary correlations between genes, drugs and diseases. My project integrates information from several very up-to-date sources, most of them manually curated. One of these is, a human disease ontology, the Disease Ontology. I constructed a general dictionary of disease and drug names, and by dictionary matching and manual curation, I extracted gene-drug-disease relation candidates. As soon as the web interface is available, users will be able to browse the list of disease concepts or query the database to search for disease names and interesting genes, and to find intersections of gene sets.

An important factor for further improvement is feedback and the collaboration with experimental groups that are benefited by this tool. Thus, the Disease Ontology annotation database will be made accessible to collaborators and participating members of the scientific community to evaluate its functionality, to test the system, and to come back with constructive feedback.

The collaboration with the GO consortium has been a valuable experience and will continue beyond the lifespan of this project. Due to our contribution of annotations and significant collaboration on major GO content development, our group at CRIBI (Padua) is now an associate member of GO Consortium.

Appendix A

Nervous system GO new terms

This section shows the terms that have been modified or added in the GO vocabularies divided in categories. These terms cover several topics related to nervous system biology.

BDNF and neurotrophic factors

brain derived neurotrophic factor receptor signaling pathway (GO:0031546)
negative regulation of brain derived neurotrophic factor receptor activity (GO:0031552)
negative regulation of brain derived neurotrophic factor signaling pathway (GO:0031549)
positive regulation of brain derived neurotrophic factor receptor activity (GO:0031553)
positive regulation of brain derived neurotrophic factor receptor signaling pathway (GO:0031550)
regulation of brain derived neurotrophic factor receptor activity (GO:0031551)
regulation of brain derived neurotrophic factor receptor signaling pathway (GO:0031548)

Calcium signalling

negative regulation of calcium ion transport via voltage gated calcium channel (GO:0051927)
positive regulation of calcium ion transport (GO:0051928)
positive regulation of calcium ion transport via voltage gated calcium channel (GO:0051929)
regulation of calcium ion transport (GO:0051924)
regulation of calcium ion transport via voltage-gated calcium channel (GO:0051925)

regulation of calcium ion transport via voltage-gated calcium channel (GO:0051925)
regulation of sensory perception (GO:0051931)
regulation of sensory perception of pain (GO:0051930)

CRH and ACTH

corticotropin-releasing hormone binding (GO:0051424)
corticotropin-releasing hormone receptor 1 binding (GO:0051430)
corticotropin-releasing hormone receptor 2 binding (GO:0051431)
corticotropin-releasing hormone receptor activity (GO:0043404)
corticotropin-releasing hormone receptor binding (GO:0051429)
corticotropin-releasing hormone secretion (GO:0043396)
gonadotropin-releasing hormone binding (GO:0051448)
gonadotropin-releasing hormone receptor binding (GO:0031530)
hormone receptor binding (GO:0051427)
negative regulation of adrenocorticotropin hormone secretion (GO:0051460)
negative regulation of corticotropin-releasing hormone secretion (GO:0051465)
polypeptide hormone receptor binding (GO:0051428)
positive regulation of adrenocorticotropin hormone secretion (GO:0051461)
regulation of adrenocorticotropin hormone secretion (GO:0051459)
regulation of corticotropin-releasing hormone secretion (GO:0043397)
thyrotropin-releasing hormone binding (GO:0051449)
thyrotropin-releasing hormone receptor binding (GO:0031531)

Nervous system development

axon regeneration in the peripheral nervous system (GO:0014012)
glial cell proliferation (GO:0014009)
medullary cord formation (GO:0014021)
mesenchymal cell development (GO:0014031)
mesenchymal cell fate commitment (GO:0014030)

negative regulation of gliogenesis (GO:0014014)

negative regulation of nervous system development (GO:0051961)

negative regulation of neuron maturation (GO:0014043)

negative regulation of Schwann cell differentiation (GO:0014039)

neural crest cell development (GO:0014032)

neural crest cell differentiation (GO:0014033)

neural crest cell fate commitment (GO:0014034)

neural crest cell fate determination (GO:0014035)

neural crest cell fate specification (GO:0014036)

neural crest cell migration (GO:0001755)

neural crest formation (GO:0014029)

neural keel formation (GO:0014025)

neural plate shaping (GO:0014022)

neural rod cavitation (GO:0014024)

neural rod formation (GO:0014023)

neuroblast development (GO:0014019)

neuroblast differentiation (GO:0014016)

neuroblast fate commitment (GO:0014017)

neuroblast proliferation (GO:0007405)

notochord formation (GO:0014028)

positive regulation of gliogenesis (GO:0014015)

positive regulation of nervous system development (GO:0051962)

positive regulation of neuron maturation (GO:0014042)

positive regulation of Schwann cell differentiation (GO:0014040)

primary neural tube formation (GO:0014020)

regulation of gliogenesis (GO:0014013)

regulation of nervous system development (GO:0051960)

regulation of neuron maturation (GO:0014041)

regulation of Schwann cell differentiation (GO:0014038)

Schwann cell development (GO:0014044)

Schwann cell differentiation (GO:0014037)

Schwann cell proliferation (GO:0014010)

Schwann cell proliferation during axon regeneration (GO:0014011)

secondary neural tube formation (GO:0014021)

Neuron cell death

induction of programmed cell death in response to chemical substance (GO:0031558)

negative regulation of neuron apoptosis (GO:0043524)

neuroprotection (GO:0043526)

positive regulation of neuron apoptosis (GO:0043525)

Programmed cell death, neurons (GO:0051402)

regulation of neuron apoptosis (GO:0043523)

Neurotransmitters binding

BH domain binding (GO:0051400)

BH1 domain binding (GO:0051432)

BH2 domain binding (GO:0051433)

BH3 domain binding (GO:0051434)

BH4 domain binding (GO:0051435)

epinephrine binding (GO:0051379)

histamine binding (GO:0051381)

kainate selective glutamate receptor complex (GO:0032983)

norepinephrine binding (GO:0051380)

serotonin binding (GO:0051378)

Neurotransmitters uptake

amino acid uptake during transmission of nerve impulse, (GO:0051933)

catecholamine transport (GO:0051937)

catecholamine uptake during transmission of nerve impulse (GO:0051934) epinephrine uptake (GO:0051625)

gamma-aminobutyric acid import (GO:0051939)

gamma-aminobutyric acid uptake during transmission of nerve impulse (GO:0051936)

glutamate uptake during transmission of nerve impulse (GO:0051935)

histamine uptake (GO:0051615)

inhibition of acetilcholine uptake (GO:0051634)

inhibition of epinephrine uptake (GO:0051629)

inhibition of histamine uptake (GO:0051619)

inhibition of norepinephrine uptake (GO:0051624)

inhibition of serotonin uptake (GO:0051614)

L-glutamate import (GO:0051938)

response to peptide hormone stimulus (GO:0043434)

neuropeptide binding (GO:0042923)

opioid receptor binding (GO:0031628)

negative regulation of acetilcholine uptake (GO:0051632)

negative regulation of amine transport (GO:0051953)

negative regulation of amino acid transport (GO:0051956)

negative regulation of amino acid uptake during transmission of nerve impulse (GO:0051942)

negative regulation of catecholamine uptake during transmission of nerve impulse (GO:0051945)

negative regulation of epinephrine uptake (GO:0051627)

negative regulation of gamma-aminobutyric acid uptake during transmission of nerve impulse (GO:0051949)

negative regulation of glutamate uptake during transmission of nerve impulse (GO:0051948)

negative regulation of histamine uptake (GO:0051617)

negative regulation of neurotransmitter uptake (GO:0051581)

negative regulation of norepinephrine uptake (GO:0051622)

negative regulation of serotonin uptake (GO:0051612)

norepinephrine uptake (GO:0051620)

positive regulation of acetilcholine uptake (GO:0051633)

positive regulation of amine transport (GO:0051954)

positive regulation of amino acid transport (GO:0051957) positive regulation of amino acid uptake during transmission of nerve impulse (GO:0051943)

positive regulation of catecholamine uptake during transmission of nerve impulse (GO:0051944)

positive regulation of epinephrine uptake (GO:0051628)

positive regulation of gamma-aminobutyric acid uptake during transmission of nerve impulse (GO:0051950)

positive regulation of glutamate uptake during transmission of nerve impulse (GO:0051951)

positive regulation of histamine uptake (GO:0051618)

positive regulation of neurotransmitter uptake (GO:0051582)

positive regulation of norepinephrine uptake (GO:0051623)

positive regulation of serotonin uptake (GO:0051613)

regulation of acetilcholine uptake (GO:0051631)

regulation of amine transport (GO:0051952)

regulation of amino acid transport (GO:0051955)

regulation of amino acid uptake during transmission of nerve impulse (GO:0051941)

regulation of catecholamine uptake during transmission of nerve impulse (GO:0051940)

regulation of epinephrine uptake (GO:0051626)

regulation of gamma-aminobutyric acid uptake during transmission of nerve impulse (GO:0051947)

regulation of glutamate uptake during transmission of nerve impulse (GO:0051946)

regulation of histamine uptake (GO:0051616)

regulation of neurotransmitter uptake (GO:0051580)

regulation of norepinephrine uptake (GO:0051621)

regulation of serotonin uptake (GO:0051611)

serotonin uptake (GO:0051610)

Response to stress

general adaptation syndrome (GO:0051866)

general adaptation syndrome, behavioural process (GO:0051867)

response to long exposure to lithium ion (GO:0043460)

response to short exposure to lithium ion (GO:0043459)

Steroid hormone

corticosteroid receptor signaling pathway (GO:0031958)

cortisol receptor activity (GO:0031963)

cortisol receptor binding (GO:0031961)

cortisol secretion (GO:0043400)

detection of glucocorticoid stimulus (GO:0051468)

detection of steroid hormone stimulus (GO:0051467)

glucocorticoid mediating signaling (GO:0043402)

mineralocorticoid receptor activity (GO:0017082)

mineralocorticoid receptor binding (GO:0031962)

mineralocorticoid receptor signaling pathway (GO:0031959)

negative regulation of cortisol secretion (GO:0051463)

positive regulation of cortisol secretion (GO:0051464)

regulation of cortisol secretion (GO:0051462)

response to corticosteroid stimulus (GO:0031960)

response to corticosterone stimulus (GO:0051412)

response to cortisol stimulus (GO:0051414)

response to cortisone stimulus (GO:0051413)

response to glucocorticoid stimulus (GO:0051384)

response to mineralocorticoid stimulus (GO:0051385)

response to steroid hormone stimulus (GO:0048545)

steroid hormone mediating signaling (GO:0043401)

Synapse

cytoskeletal matrix organization at active zone (GO:0048789)

excitatory synapse (GO:0060076)

inhibitory synapse (GO:0060077)

membrane hyperpolarization (GO:0060081)

negative regulation of synaptic vesicle fusion to presynaptic membrane (GO:0031631) negative regulation of synaptogenesis (GO:0051964)

positive regulation of synaptogenesis (GO:0051965)

post synaptic density (GO:0014069)

presynaptic active zone (GO:0048786)

presynaptic cytoskeletal matrix assembled at active zones (GO:0048788)

regulation of excitatory postsynaptic membrane potential (GO:0060079)

regulation of inhibitory postsynaptic membrane potential (GO:0060080)

regulation of postsynaptic membrane potential (GO:0060078)

regulation of resting potential (GO:0060075)

regulation of synaptic vesicle fusion to presynaptic membrane (GO:0031630)

regulation of synaptogenesis (GO:0051963)

symmetric synapse (GO:0032280)

synaptic maturation (GO:0060075)

synaptic vesicle fusion to presynaptic membrane (GO:0031629)

Synaptic plasticity

negative regulation of synapse structural plasticity (GO:0051826)

negative regulation of synaptic metaplasticity (GO:0031917)

negative regulation of synaptic plasticity (GO:0031914)

positive regulation of synapse structural plasticity (GO:0051835)

positive regulation of synaptic metaplasticity (GO:0031918)

positive regulation of synaptic plasticity (GO:0031915)

regulation of neuronal synaptic plasticity in response to neurotrophin (GO:0031637)

regulation of synapse structural plasticity (GO:0051823)

regulation of synaptic metaplasticity (GO:0031916)

regulation of synaptic plasticity by drug (GO:0051913)

Synaptic transmission

negative regulation of synaptic transmission, dopaminergic (GO:0032227)

negative regulation of synaptic transmission, GABAergic (GO:0032229)

positive regulation of synaptic plasticity by chemical substance (GO:0051914)

positive regulation of synaptic transmission, cholinergic (GO:0032224)

positive regulation of synaptic transmission, dopaminergic (GO:0032226)

positive regulation of synaptic transmission, GABAergic (GO:0032230)

regulation of synaptic transmission, cholinergic (GO:0032222)

regulation of synaptic transmission, dopaminergic (GO:0032225)

regulation of synaptic transmission, GABAergic (GO:0032228)

synaptic transmission, GABAergic (GO:0051932)

Appendix B

Nervous system gene annotations

This section shows the 21 metabotropic glutamate receptor proteins and the 57 ionotropic glutamate receptor proteins manual annotated (15 KA, 15 AMPA and 27 NMDA receptors)

Metabotropic glutamate receptor proteins

MGR1_HUMAN/Q13255
Metabotropic glutamate receptor 1 (mGluR1)
MGR1_MOUSE/P97772
Metabotropic glutamate receptor 1 (mGluR1)
MGR1_RAT/P23385
Metabotropic glutamate receptor 1 precursor (mGluR1)
MGR2_HUMAN/Q14416
Metabotropic glutamate receptor 2 precursor (mGluR2)
MGR2_RAT/P31421
Metabotropic glutamate receptor 2 precursor (mGluR2)
MGR3_HUMAN/Q14832
Metabotropic glutamate receptor 3 precursor (mGluR3)
MGR3_MOUSE/Q9QYS
Metabotropic glutamate receptor 3 precursor (mGluR3)
MGR3_RAT/P31422

Metabotropic glutamate receptor 3 precursor (mGluR3)

MGR4_HUMAN/Q14833

Metabotropic glutamate receptor 4 precursor (mGluR4)

MGR4_RAT/P31423

Metabotropic glutamate receptor 4 precursor (mGluR4)

MGR4_MOUSE/Q68EF4

Metabotropic glutamate receptor 4 precursor (mGluR4)

MGR5_HUMAN/P41594

Metabotropic glutamate receptor 5 precursor (mGluR5)

MGR5_RAT/P31424

Metabotropic glutamate receptor 5 precursor (mGluR5)

MGR6_HUMAN/O15303

Metabotropic glutamate receptor 6 precursor (mGluR6)

MGR6_MOUSE/Q8CFQ7

Metabotropic glutamate receptor 6 (mGluR6)

MGR6_RAT/P35349

Metabotropic glutamate receptor 6 precursor (mGluR6)

MGR7_HUMAN/Q14831

Metabotropic glutamate receptor 7 precursor (mGluR7)

MGR7_MOUSE/Q68ED2

Metabotropic glutamate receptor 7 precursor (mGluR7) MGR7_RAT/P35400

Metabotropic glutamate receptor 7 precursor (mGluR7)

MGR8_HUMAN/O00222

Metabotropic glutamate receptor 8 precursor (mGluR8)

MGR8_MOUSE/P47743

Metabotropic glutamate receptor 8 (mGluR8)

Ionotropic glutamate receptor proteins AMPA receptors

GRIA4_HUMAN/P48058 Glutamate receptor ionotropic, AMPA 4 (GluR4)

GRIA4_MOUSE/Q9Z2W8

Glutamate receptor ionotropic, AMPA 4 (GluR4)

GRIA4_RAT/P19493

Glutamate receptor ionotropic, AMPA 4

Q2NKM6_HUMAN/Q2NKM6

Glutamate receptor, ionotropic, AMPA 1

Q7TNB5_MOUSE/Q7TNB5

Glutamate receptor, ionotropic, AMPA1 (Alpha 1)

GRIA1_HUMAN/P42261

Glutamate receptor ionotropic, AMPA 1 (GluR1) (GluRA)

GRIA1_MOUSE/P23818

Glutamate receptor ionotropic, AMPA 1 (GluR1)

GRIA1_RAT/P19490

Glutamate receptor ionotropic, AMPA 1 (GluR1) (GluRA)

Q59F93_HUMAN/Q59F93

Glutamate receptor, ionotropic, AMPA 2 variant

GRIA2_HUMAN/P42262

Glutamate receptor ionotropic, AMPA 2 (GluR2)

GRIA2_MOUSE/P23819

Glutamate receptor ionotropic, AMPA 2 (GluR2)

GRIA2_RAT/P19491

Glutamate receptor ionotropic, AMPA 2 (GluR2)

GRIA3_HUMAN/P42263

Glutamate receptor ionotropic, AMPA 3 (GluR3)

GRIA3_MOUSE/Q9Z2W9

Glutamate receptor ionotropic, AMPA 3 (GluR3)

GRIA3_RAT/P19492

Glutamate receptor ionotropic, AMPA 3 (GluR3)

Ionotropic glutamate receptor proteins - KA receptors

GRIK1_HUMAN/P39086

Glutamate receptor, ionotropic kainate 1 precursor (GluR5)

GRIK1_MOUSE/Q60934

Glutamate receptor, ionotropic kainate 1 precursor (GluR5)

GRIK1_RAT/P22756

Glutamate receptor, ionotropic kainate 1 precursor (GluR5)

GRIK2_HUMAN/Q13002

Glutamate receptor, ionotropic kainate 2 precursor (GluR6)

GRIK2_MOUSE/P39087

Glutamate receptor, ionotropic kainate 2 precursor (GluR6)

GRIK2_RAT/P42260

Glutamate receptor, ionotropic kainate 2 precursor (GluR6)

Q5T646_HUMAN/Q5T646

Glutamate receptor, ionotropic, kainate 3 (GluR7)

GRIK3_HUMAN/Q13003

Glutamate receptor, ionotropic kainate 3 precursor (GluR7)

GRIK3_RAT/P42264

Glutamate receptor, ionotropic kainate 3 precursor (GluR7)

GRIK4_HUMAN/Q16099

Glutamate receptor, ionotropic kainate 4 precursor (KA1)

GRIK4_MOUSE/Q8BMF5

Glutamate receptor, ionotropic kainate 4 precursor

GRIK4_RAT/Q01812

Glutamate receptor, ionotropic kainate 4 precursor (KA1)

GRIK5_HUMAN/Q16478

Glutamate receptor, ionotropic kainate 5 precursor (KA2)

GRIK5_MOUSE/Q61626

Glutamate receptor, ionotropic kainate 5 precursor (KA2)

GRIK5_RAT/Q63273

Glutamate receptor, ionotropic kainate 5 precursor (KA2)

Ionotropic glutamate receptor proteins - NMDA receptors

NMD3A_RAT/Q9R1M7

Glutamate receptor subunit 3A precursor (NMDA receptor subtype 3A) (GRIN3A)

Q5VTR3_HUMAN/Q5VTR3

Glutamate receptor, ionotropic, N-methyl-D-aspartate 3A (GRIN3A)

NMD3B_RAT/Q8VHN2

NMDA receptor subunit 3B precursor (NMDA receptor subtype 3B) (GRIN3B)

NMD3B_MOUSE/Q91ZU9

NMDA receptor subunit 3B precursor (NMDAR subunit NR3B) (GRIN3B)

Q5VSF3_HUMAN/Q5VSF3

Glutamate receptor, ionotropic, N-methyl D-aspartate 1 (GRIN1)

Q5VSF4_HUMAN/Q5VSF4

Glutamate receptor, ionotropic, N-methyl D-aspartate 1 (GRIN1)

Q5VSF5_HUMAN/Q5VSF5

Glutamate receptor, ionotropic, N-methyl D-aspartate 1 (GRIN1)

Q5VSF6_HUMAN/Q5VSF6

Glutamate receptor, ionotropic, N-methyl D-aspartate 1 (GRIN1)

Q5VSF7_HUMAN/Q5VSF7

Glutamate receptor, ionotropic, N-methyl D-aspartate 1 (GRIN1)

Q5VSF8_HUMAN/Q5VSF8

Glutamate receptor, ionotropic, N-methyl D-aspartate 1 (GRIN1)

Q5VSF9_HUMAN/Q5VSF9

Glutamate receptor, ionotropic, N-methyl D-aspartate 1 (GRIN1)

Q6P6I6_MOUSE/Q6P6I6

Glutamate receptor, ionotropic, N-methyl D-aspartate-like 1A (GRIN1)

NMDZ1_MOUSE/P35438

NMDA receptor subunit zeta 1 precursor (GRIN1)

Q62683_RAT/Q62683

NMDAR1 glutamate receptor subunit (GRIN1)

NMDZ1_RAT/P35439

NMDA receptor subunit zeta 1 precursor (GRIN1)

Q00959_ RAT/Q00959

NMDA receptor subunit NR2A (GRIN2A)

NMDE1_MOUSE/P35436

NMDA receptor subunit epsilon 1 precursor (GRIN2A)

Q00960_ RAT/Q00960

NMDA receptor subunit NR2B (GRIN2B)

NMDE2_MOUSE/Q01097

NMDA receptor subunit epsilon 2 precursor (GRIN2B)

NMDE2_HUMAN/Q13224

NMDA receptor subunit epsilon 2 precursor (GRIN2B)

NMDE3_RAT/Q00961

NMDA receptor subunit epsilon 3 precursor (GRIN2C)

Q61984_MOUSE/Q61984

NMDA receptor subunit NR2C (GRIN2C)

NMDE3_MOUSE/Q01098

NMDA receptor subunit epsilon 3 precursor (GRIN2C)

NMDE3_HUMAN/Q14957

NMDA receptor subunit epsilon 3 precursor (GRIN2C)

NMDE4_MOUSE/Q03391

NMDA receptor subunit epsilon 4 precursor (GRIN2D)

NMDE4_RAT/Q62645

NMDA receptor subunit epsilon 4 precursor (GRIN2D)

NMDE4_HUMAN/O15399

NMDA receptor subunit epsilon 4 precursor (GRIN2D)

Appendix C

Muscle GO new terms

This section shows the terms added or modified during the muscle content meeting.

Muscle cellular component

C zone (GO:0014705)

costamere (GO:0043034)

desmosome (GO:0030057)

fascia adherens (GO:0005916)

free sarcoplasmic reticulum membrane (GO:0014702)

gap junction (GO:0005921)

H zone (GO:0031673)

intercalated disc (GO:0014704)

junctional sarcoplasmic reticulum membrane (GO:0014701)

longitudinal sarcoplasmic reticulum (GO:0014801)

longitudinal sarcoplasmic reticulum lumen (GO:0014803)

muscle tendon junction (GO:0005927)

release of sequestered calcium ion by sarcoplasmic reticulum into cytosol (GO:0014808)

sarcoplasmic reticulum (GO:0016529)

spectrin-associated cytoskeleton (GO:0014731)

striated muscle thick filament (GO:0005863)

T-tubule (GO:0030315)

terminal cisterna (GO:0014802)
terminal cisterna lumen (GO:0014804)
Z disc (GO:0030018)

Muscle contraction

arteriole smooth muscle contraction (GO:0014830)
cross bridge cycling involved in regulation of the velocity of shortening in skeletal muscle contraction (GO:0014868)
cross bridge formation involved in regulation of the velocity of shortening in skeletal muscle contraction (GO:0014871)
detection of activity (GO:0014865)
detection of electrical stimulus involved in regulation of muscle adaptation (GO:0014879)
detection of inactivity (GO:0014863)
detection of injury involved in regulation of muscle adaptation (GO:0014885)
detection of muscle activity (GO:0014864)
detection of muscle activity involved in regulation of muscle adaptation (GO:0014875)
detection of muscle inactivity (GO:0014869)
detection of muscle inactivity involved in regulation of muscle adaptation (GO:0014884)
detection of wounding (GO:0014822)
diaphragm contraction (GO:0002086)
distal stomach smooth muscle contraction (GO:0014828)
esophagus smooth muscle contraction (GO:0014846)
fast-twitch skeletal muscle fiber contraction (GO:0031443)
gastro-intestinal system smooth muscle contraction (GO:0014831)
hindgut contraction (GO:0043133)
intestine smooth muscle contraction (GO:0014827)
involuntary skeletal muscle contraction (GO:0003011)
muscle contraction (GO:0006936)
muscle filament sliding (GO:0030049)
muscle thick filament assembly (GO:0030241)

muscle thin filament assembly (GO:0030240)

negative regulation of fast-twitch skeletal muscle contraction (GO:0031447)

negative regulation of slow-twitch skeletal muscle contraction (GO:0031450)

negative regulation of tonic skeletal muscle contraction (GO:0014748)

neurotransmitter secretion involved in control of skeletal muscle contraction (GO:0014860)

oscillatory muscle contraction (GO:0014703)

phasic smooth muscle contraction (GO:0014821)

positive regulation of fast-twitch skeletal muscle contraction (GO:0031448)

positive regulation of skeletal muscle contraction via regulation of the release of sequestered calcium ion (GO:0014810)

positive regulation of slow-twitch skeletal muscle contraction (GO:0031451)

positive regulation of tonic skeletal muscle contraction (GO:0014747)

proximal stomach smooth muscle contraction (GO:0014847)

regulation of excitatory postsynaptic membrane potential involved in skeletal muscle contraction (GO:0014853)

regulation of fast-twitch skeletal muscle contraction (GO:0031446)

regulation of filament sliding speed involved in regulation of the velocity of shortening in skeletal muscle contraction (GO:0014915)

regulation of muscle filament sliding involved in the regulation of the velocity of shortening in skeletal muscle contraction (GO:0014880)

regulation of skeletal muscle contraction (GO:0014819)

regulation of skeletal muscle contraction by calcium ion signaling (GO:0014722)

regulation of skeletal muscle contraction by chemo-mechanical energy conversion (GO:0014862)

regulation of skeletal muscle contraction by neural stimulation via neuromuscular junction (GO:0014852)

regulation of skeletal muscle contraction via modulation of calcium ion sensitivity of myofibril (GO:0014723)

regulation of skeletal muscle contraction via regulation of the release of sequestered calcium ion (GO:0014809)

regulation of slow-twitch skeletal muscle contraction (GO:0031449)

regulation of the force of skeletal muscle contraction (GO:0014728) regulation of the velocity of shortening of skeletal muscle during contraction (GO:0014729)

regulation of tonic skeletal muscle contraction (GO:0014746)

regulation of twitch skeletal muscle contraction (GO:0014724)

skeletal muscle contraction (GO:0003009)

slow-twitch skeletal muscle fiber contraction (GO:0031444)

smooth muscle contraction (GO:0006939)

stomach body smooth muscle contraction (GO:0014845)

stomach fundus smooth muscle contraction (GO:0014825)

striated muscle contraction (GO:0006941)

tonic skeletal muscle contraction (GO:0014720)

tonic smooth muscle contraction (GO:0014820)

twitch skeletal muscle contraction (GO:0014721)

ureter smooth muscle contraction (GO:0014849)

urinary bladder smooth muscle contraction (GO:0014832)

urinary tract smooth muscle contraction (GO:0014848)

vascular smooth muscle contraction (GO:0014829)

vein smooth muscle contraction (GO:0014826)

ventricular cardiac muscle cell development (GO:0055015)

voluntary skeletal muscle contraction (GO:0003010)

Muscle development

branchiomeric skeletal muscle development (GO:0014707)

extraocular skeletal muscle development (GO:0002074)

muscle fiber development (GO:0048747)

myoblast cell division (GO:0014872)

myoblast cell fate commitment (GO:0048625)

myoblast cell fate commitment in head (GO:0014714)

myoblast cell fate commitment in trunk (GO:0014715)

myoblast cell fate specification (GO:0048626)

myoblast development (GO:0048627)

myoblast differentiation (GO:0045445)

myoblast fusion (GO:0007520)

myoblast maturation (GO:0048628)

myoblast migration (GO:0051451)

myoblast proliferation (GO:0051450)

myotube cell development (GO:0014904)

myotube differentiation (GO:0014902)

negative regulation of branchiomeric skeletal muscle development (GO:0014713)

negative regulation of extraocular skeletal muscle development (GO:0014726)

negative regulation of myoblast differentiation (GO:0045662)

negative regulation of skeletal muscle cell proliferation (GO:0014859)

negative regulation of skeletal muscle contraction via regulation of the release of sequestered calcium ion (GO:0014811)

negative regulation of skeletal muscle fiber development (GO:0048744)

negative regulation of somitomeric trunk muscle development (GO:0014710)

plasma membrane fusion (GO:0045026)

positive regulation of branchiomeric skeletal muscle development (GO:0014712)

positive regulation of extraocular skeletal muscle development (GO:0014727)

positive regulation of myoblast differentiation (GO:0045663)

positive regulation of skeletal muscle cell proliferation (GO:0014858)

positive regulation of skeletal muscle fiber development (GO:0048743)

positive regulation of somitomeric trunk muscle development (GO:0014709)

regulation of branchiomeric skeletal muscle development (GO:0014711)

regulation of extraocular skeletal muscle development (GO:0014725)

regulation of myoblast differentiation (GO:0045661)

regulation of satellite cell proliferation (GO:0014842)

regulation of skeletal muscle cell proliferation (GO:0014857)

regulation of skeletal muscle fiber development (GO:0048742)

regulation of somitogenesis (GO:0014807)

regulation of somitomeric trunk muscle development (GO:0014708)

satellite cell asymmetric division (GO:0014833) satellite cell commitment (GO:0014813)

satellite cell differentiation (GO:0014816)

satellite cell fate specification (GO:0014817)

satellite cell proliferation (GO:0014841)

skeletal muscle cell proliferation (GO:0014856)

skeletal muscle development (GO:0007519)

skeletal muscle fiber development (GO:0048741)

skeletal myofibril assembly (GO:0014866)

smooth muscle fiber development (GO:0048746)

somatic muscle development (GO:0007525)

striated muscle cell proliferation (GO:0014855)

striated muscle development (GO:0014706)

striated muscle fiber development (GO:0048740)

syncytium formation by plasma membrane fusion (GO:0000768)

Muscle plasticity

cardiac muscle adaptation (GO:0014887)

cardiac muscle atrophy (GO:0014899)

cardiac muscle hypertrophy (GO:0014898)

muscle adaptation (GO:0043500)

muscle atrophy (GO:0014889)

muscle cell migration (GO:0014812)

muscle hyperplasia (GO:0014900)

muscle hypertrophy (GO:0014896)

negative regulation of muscle adaptation (GO:0014745)

negative regulation of muscle atrophy (GO:0014736)

negative regulation of muscle hyperplasia (GO:0014740)

negative regulation of muscle hypertrophy (GO:0014741)

positive regulation of muscle adaptation (GO:0014744)

positive regulation of muscle atrophy (GO:0014737)

positive regulation of muscle hyperplasia (GO:0014739) positive regulation of muscle hypertrophy (GO:0014742)

regulation of muscle adaptation (GO:0043502)

regulation of muscle atrophy (GO:0014735)

regulation of muscle hyperplasia (GO:0014738)

regulation of muscle hypertrophy (GO:0014743)

regulation of myofibril number (GO:0014882)

regulation of myofibril size (GO:0014881)

regulation of skeletal muscle adaptation (GO:0014733)

response to activity (GO:0014823)

response to denervation involved in regulation of muscle adaptation (GO:0014894)

response to electrical stimulus involved in regulation of muscle adaptation (GO:0014878)

response to inactivity (GO:0014854)

response to injury involved in regulation of muscle adaptation (GO:0014876)

response to muscle activity (GO:0014850)

response to muscle activity involved in regulation of muscle adaptation (GO:0014873)

response to muscle inactivity (GO:0014870)

response to muscle inactivity involved in regulation of muscle adaptation (GO:0014877)

response to rest involved in regulation of muscle adaptation (GO:0014893)

response to stimulus involved in regulation of muscle adaptation (GO:0014874)

skeletal muscle adaptation (GO:0043501)

skeletal muscle atrophy (GO:0014732)

skeletal muscle fiber adaptation (GO:0043503)

skeletal muscle hypertrophy (GO:0014734)

smooth muscle adaptation (GO:0014805)

smooth muscle hyperplasia (GO:0014806)

smooth muscle hypertrophy (GO:0014895)

striated muscle adaptation (GO:0014888)

striated muscle atrophy (GO:0014891)

striated muscle hypertrophy (GO:0014897)

transition between fast and slow fiber (GO:0014883)

transition between slow and fast fiber (GO:0014886)

Muscle regeneration

axon regeneration at neuromuscular junction (GO:0014814)

growth factor dependent regulation of satellite cell proliferation (GO:0014843)

initiation of satellite cell activation by growth factor signalling, involved in skeletal muscle regeneration (GO:0014815)

multicellular organismal movement (GO:0050879)

myoblast cell differentiation involved in skeletal muscle regeneration (GO:0014835)

myoblast cell fate commitment involved in skeletal muscle regeneration (GO:0014836)

myoblast cell fate specification involved in skeletal muscle regeneration (GO:0014838)

myoblast cell proliferation involved in skeletal muscle regeneration (GO:0014844)

myoblast fusion involved in skeletal muscle regeneration (GO:0014905)

myoblast maturation involved in muscle regeneration (GO:0014914)

myoblast migration involved in skeletal muscle regeneration (GO:0014839)

myotube cell development involved in skeletal muscle regeneration (GO:0014906)

myotube differentiation involved in skeletal muscle regeneration (GO:0014908)

negative regulation of smooth muscle cell migration (GO:0014912)

positive regulation of satellite cell activation involved in skeletal muscle regeneration (GO:0014718)

positive regulation of smooth muscle cell migration (GO:0014911)

regulation of satellite cell activation involved in skeletal muscle regeneration (GO:0014717)

regulation of smooth muscle cell migration (GO:0014910)

satellite cell activation (GO:0014719)

satellite cell activation involved in skeletal muscle regeneration (GO:0014901)

satellite cell asymmetric division involved in skeletal muscle regeneration (GO:0014716)

satellite cell compartment self-renewal involved in skeletal muscle regeneration (GO:0014834)

skeletal muscle regeneration (GO:0043403)

skeletal muscle regeneration at neuromuscular junction (GO:0014730)

smooth muscle cell migration (GO:0014909)

Appendix D

List of abbreviations

GOA Gene Ontology Annotation.

GO Gene Ontology.

DAG Directed Acyclic Graph

DO Disease Ontology

UMLS Unified Medical Language System

OBO Open Biomedical Ontologies

OE OBO-Edit

GSK Glaxo Smith Kline

GAD Genetic Association Database

OMIM Online Mendelian Inheritance in Man

PharmGKB Pharmacogenetics and Pharmacogenomics Knowledge Base

PK Pharmacokinetics

PD Pharmacodynamics

OLS Ontology Lookup Service

MeSH Medical Subject Headings

ASAP Dictionary of Anxiety and Panic Disorders

iHOP Information Hyperlinked Over Proteins

OWL Web Ontology Language

RO Relation Ontology

CARO Common Anatomy Reference Ontology

OBI Ontology for Biomedical Investigations

MGI Mouse Genome Informatics

SGD Saccharomyces Genome Database

EBI European Bioinformatics Institute

UniProtKB UniProtKnowledgebase

IPI International Protein Index

EC Enzyme Commission

DE UniProt description lines

NCBO National Center for Biomedical Ontology

ICD9CM The International Classification of Diseases, Ninth Revision, Clinical Modification

SNOMED Systematized Nomenclature of Medicine-Clinical Terms

NS Nervous System

SF SourceForge

CRH Corticotropin-releasing hormone

HPA Hypothalamic-Pituitary-Adrenal axis

BDNF Brain-derived neurotrophic factor

ACTH adrenocorticotropic hormone

PSD Postsynaptic density

AMA Adult Mouse Anatomical Dictionary

CL Cell Type ontology

SR Sarcoplasmic Reticulum

ER Endoplasmic Reticulum

ChEBI Chemical Entities of Biological Interest

CAS Chemical Abstract Service Registry Database

IUPAC International Union of Pure and Applied Chemistry

Bibliography

[1] S. Philippi and J. Kohler. Addressing the problems with life-science databases for traditional uses and systems biology. *Nat Rev Genet*, 7(6):482–8, 2006.

[2] M. A. Harris, J. Clark, A. Ireland, J. Lomax, M. Ashburner, R. Foulger, K. Eilbeck, S. Lewis, B. Marshall, C. Mungall, J. Richter, G. M. Rubin, J. A. Blake, C. Bult, M. Dolan, H. Drabkin, J. T. Eppig, D. P. Hill, L. Ni, M. Ringwald, R. Balakrishnan, J. M. Cherry, K. R. Christie, M. C. Costanzo, S. S. Dwight, S. Engel, D. G. Fisk, J. E. Hirschman, E. L. Hong, R. S. Nash, A. Sethuraman, C. L. Theesfeld, D. Botstein, K. Dolinski, B. Feierbach, T. Berardini, S. Mundodi, S. Y. Rhee, R. Apweiler, D. Barrell, E. Camon, E. Dimmer, V. Lee, R. Chisholm, P. Gaudet, W. Kibbe, R. Kishore, E. M. Schwarz, P. Sternberg, M. Gwinn, L. Hannick, J. Wortman, M. Berriman, V. Wood, N. de la Cruz, P. Tonellato, P. Jaiswal, T. Seigfried, and R. White. The Gene Ontology (GO) database and informatics resource. *Nucleic Acids Res*, 32(Database issue):D258–61, 2004.

[3] T. R. Gruber. Towards principles for the design of ontologies used for knowledge sharing. In N. Guarino and R. Poli, editors, *Formal Ontology in Conceptual Analysis and Knowledge Representation*, Deventer, The Netherlands, 1993. Kluwer Academic Publishers.

[4] P. Lambrix, M. Habbouche, and M. Perez. Evaluation of ontology development tools for bioinformatics. *Bioinformatics*, 19(12):1564–71, 2003.

[5] R. Mack and M. Hehenberger. Text-based knowledge discovery: search and mining of life-sciences documents. *Drug Discov Today*, 7(11 Suppl):S89–98, 2002.

[6] M. Harris and H Parkinson. Standards and Ontologies for Functional Genomics: Towards Unified Ontologies for Biology and Biomedicine. *Comparative and Functional Genomics*, 4(1):116–120, 2003. doi:10.1002/cfg.249.

[7] M. Deng, Z. Tu, F. Sun, and T. Chen. Mapping Gene Ontology to proteins based on protein-protein interaction data. *Bioinformatics*, 20(6):895–902, 2004.

[8] O. Bodenreider and R. Stevens. Bio-ontologies: current trends and future directions. *Brief Bioinform*, 7(3):256–74, 2006.

[9] J. I. Clark, C. Brooksbank, and J. Lomax. It's all GO for plant scientists. *Plant Physiol*, 138(3):1268–79, 2005.

[10] J. B. Bard and S. Y. Rhee. Ontologies in biology: design, applications and future challenges. *Nat Rev Genet*, 5(3):213–22, 2004.

[11] T. R. Gruber. A Translation Approach to Portable Ontology Specifications. *Knowledge Acquisition*, 5(2):199–220, 1993.

[12] B. Andersen. What is an ontology? *Ontology Works (http://www.ontologyworks.com)*, 2001.

[13] O. Bodenreider, J. A. Mitchell, and A. T. McCray. Biomedical ontologies. *Pac Symp Biocomput*, pages 76–8, 2005.

[14] S. Schulze-Kremer. Ontologies for molecular biology and bioinformatics. *In Silico Biol*, 2(3):179–93, 2002.

[15] J. S. Caldwell. Ontology recapitulates physiology. *Chem Biol*, 10(9):784–6, 2003.

[16] D. Devos and A. Valencia. Intrinsic errors in genome annotation. *Trends Genet*, 17(8):429–31, 2001.

[17] C. Blaschke and A. Valencia. Automatic ontology construction from the literature. *Genome Inform*, 13:201–13, 2002.

[18] B. Smith, M. Ashburner, C. Rosse, J. Bard, W. Bug, W. Ceusters, L. J. Goldberg, K. Eilbeck, A. Ireland, C. J. Mungall, N. Leontis, P. Rocca-Serra, A. Ruttenberg, S. A. Sansone, R. H. Scheuermann, N. Shah, P. L. Whetzel, and S. Lewis. The OBO Foundry: coordinated evolution of ontologies to support biomedical data integration. *Nat Biotechnol*, 25(11):1251–1255, 2007.

[19] B. Smith, W. Ceusters, B. Klagges, J. Kohler, A. Kumar, J. Lomax, C. Mungall, F. Neuhaus, A. L. Rector, and C. Rosse. Relations in biomedical ontologies. *Genome Biol*, 6(5):R46, 2005.

[20] J. A. Blake and C. J. Bult. Beyond the data deluge: data integration and bio-ontologies. *J Biomed Inform*, 39(3):314–20, 2006.

[21] M. Ashburner, C. A. Ball, J. A. Blake, D. Botstein, H. Butler, J. M. Cherry, A. P. Davis, K. Dolinski, S. S. Dwight, J. T. Eppig, M. A. Harris, D. P. Hill, L. Issel-Tarver, A. Kasarskis, S. Lewis, J. C. Matese, J. E. Richardson, M. Ringwald, G. M. Rubin, and G. Sherlock. Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nat Genet*, 25(1):25–9, 2000.

[22] D. M. Jones and R. C. Paton. Toward principles for the representation of hierarchical knowledge in formal ontologies. *Data & Knowledge Engineering*, 31(2):99–113, 1999.

[23] R. Stevens, C. A. Goble, and S. Bechhofer. Ontology-based knowledge representation for bioinformatics. *Brief Bioinform*, 1(4):398–414, 2000.

[24] J. Blake and M. Harris. *The Gene Ontology Project: Structured vocabularies for molecular biology and their application to genome and expression analysis.* Baxevanis, A.D. Davison, D.B. Page, R. Stormo, G. Stein, L.Current Protocols in Bioinformatics, Wiley & Sons, New York., 2003.

[25] M. Harris, J. Lomax, A. Ireland, and J. I. Clark. *The Gene Ontology project.* John Wiley & Sons, Ltd. Encyclopedia of Genetics, Genomics, Proteomics and Bioinformatics. Part 4. Bioinformatics 4.7. Structuring and Integrating Data, 2005.

[26] The Gene Ontology Consortium. Creating the gene ontology resource: design and implementation. *Genome Res*, 11(8):1425–33, 2001.

[27] E. Camon, M. Magrane, D. Barrell, V. Lee, E. Dimmer, J. Maslen, D. Binns, N. Harte, R. Lopez, and R. Apweiler. The Gene Ontology Annotation (GOA) Database: sharing knowledge in Uniprot with Gene Ontology. *Nucleic Acids Res*, 32(Database issue):D262–6, 2004.

[28] M. Bada, R. Stevens, C. Goble, Y. Gil, M. Ashburner, J.A. Blake, J.M. Cherry, M. Harris, and S. Lewis. A short study on the success of the gene ontology. *Journal of Web Semantics*, 1(2), 2004.

[29] The Gene Ontology Consortium. The Gene Ontology project in 2008. *Nucleic Acids Res*, 2007.

[30] A. D. Diehl, J. A. Lee, R. H. Scheuermann, and J. A. Blake. Ontology development for biological systems: immunology. *Bioinformatics*, 23(7):913–5, 2007.

[31] J. Day-Richter, M. A. Harris, M. Haendel, and S. Lewis. OBO-Edit–an ontology editor for biologists. *Bioinformatics*, 23(16):2198–200, 2007.

[32] E. Camon, M. Magrane, D. Barrell, D. Binns, W. Fleischmann, P. Kersey, N. Mulder, T. Oinn, J. Maslen, A. Cox, and R. Apweiler. The Gene Ontology Annotation (GOA) project: implementation of GO in SWISS-PROT, TrEMBL, and InterPro. *Genome Res*, 13(4):662–72, 2003.

[33] E. B. Camon, D. G. Barrell, E. C. Dimmer, V. Lee, M. Magrane, J. Maslen, D. Binns, and R. Apweiler. An evaluation of GO annotation retrieval for BioCreAtIvE and GOA. *BMC Bioinformatics*, 6 Suppl 1:S17, 2005.

[34] Z. Z. Hu, I. Mani, V. Hermoso, H. Liu, and C. H. Wu. iProLINK: an integrated protein resource for literature mining. *Comput Biol Chem*, 28(5-6):409–16, 2004.

[35] S. Pyysalo, F. Ginter, T. Pahikkala, J. Boberg, J. Jarvinen, and T. Salakoski. Evaluation of two dependency parsers on biomedical corpus targeted at protein-protein interactions. *Int J Med Inform*, 75(6):430–42, 2006.

[36] T. Z. Berardini, S. Mundodi, L. Reiser, E. Huala, M. Garcia-Hernandez, P. Zhang, L. A. Mueller, J. Yoon, A. Doyle, G. Lander, N. Moseyko, D. Yoo, I. Xu, B. Zoeckler, M. Mon-

toya, N. Miller, D. Weems, and S. Y. Rhee. Functional annotation of the Arabidopsis genome using controlled vocabularies. *Plant Physiol*, 135(2):745–55, 2004.

[37] F. M. Couto, M. J. Silva, V. Lee, E. Dimmer, E. Camon, R. Apweiler, H. Kirsch, and D. Rebholz-Schuhmann. GOAnnotator: linking protein GO annotations to evidence text. *J Biomed Discov Collab*, 1:19, 2006.

[38] A. Doms and M. Schroeder. GoPubMed: exploring PubMed with the Gene Ontology. *Nucleic Acids Res*, 33(Web Server issue):W783–6, 2005.

[39] W. A. Baumgartner, Jr, K. B. Cohen, L. M. Fox, G. Acquaah-Mensah, and L. Hunter. Manual curation is not sufficient for annotation of genomic databases. *Bioinformatics*, 23(13):i41–8, 2007.

[40] M. R. Seringhaus and M. B. Gerstein. Publishing perishing? Towards tomorrow's information architecture. *BMC Bioinformatics*, 8:17, 2007.

[41] S. D. Schlueter, M. D. Wilkerson, Q. Dong, and V. Brendel. xGDB: open-source computational infrastructure for the integrated evaluation and analysis of genome features. *Genome Biol*, 7(11):R111, 2006.

[42] M. D. Wilkerson, S. D. Schlueter, and V. Brendel. yrGATE: a web-based gene-structure annotation tool for the identification and dissemination of eukaryotic genes. *Genome Biol*, 7(7):R58, 2006.

[43] C Baral. Collaborative curation of data from bio-medical texts and abstracts and its integration. *In Proceedings of the Second International Workshop on Data Integration in the Life Sciences.*, 2005.

[44] E. J. Nestler, M. Barrot, R. J. DiLeone, A. J. Eisch, S. J. Gold, and L. M. Monteggia. Neurobiology of depression. *Neuron*, 34(1):13–25, 2002.

[45] F. Holsboer. Stress, hypercortisolism and corticosteroid receptors in depression: implications for therapy. *J Affect Disord*, 62(1-2):77–91, 2001.

[46] R. M. Sapolsky. Glucocorticoids and hippocampal atrophy in neuropsychiatric disorders. *Arch Gen Psychiatry*, 57(10):925–35, 2000.

[47] A. A. Russo-Neustadt and M. J. Chen. Brain-derived neurotrophic factor and antidepressant activity. *Curr Pharm Des*, 11(12):1495–510, 2005.

[48] D. Nijhawan, N. Honarpour, and X. Wang. Apoptosis in neural development and disease. *Annu Rev Neurosci*, 23:73–87, 2000.

[49] M. Hetman, K. Kanning, J. E. Cavanaugh, and Z. Xia. Neuroprotection by brain-derived neurotrophic factor is mediated by extracellular signal-regulated kinase and phosphatidylinositol 3-kinase. *J Biol Chem*, 274(32):22569–80, 1999.

[50] P. Desjardins and S. Ledoux. The role of apoptosis in neurodegenerative diseases. *Metab Brain Dis*, 13(2):79–96, 1998.

[51] B. S. McEwen. Stress and hippocampal plasticity. *Annu Rev Neurosci*, 22:105–22, 1999.

[52] E. Pollanen, P. H. Ronkainen, H. Suominen, T. Takala, S. Koskinen, J. Puolakka, S. Sipila, and V. Kovanen. Muscular Transcriptome in Postmenopausal Women With or Without Hormone Replacement. *Rejuvenation Res*, 2007.

[53] C. Bean, M. Salamon, A. Raffaello, S. Campanaro, A. Pallavicini, and G. Lanfranchi. The Ankrd2, Cdkn1c and calcyclin genes are under the control of MyoD during myogenic differentiation. *J Mol Biol*, 349(2):349–66, 2005.

[54] T. F. Hayamizu, M. Mangan, J. P. Corradi, J. A. Kadin, and M. Ringwald. The Adult Mouse Anatomical Dictionary: a tool for annotating and integrating data. *Genome Biol*, 6(3):R29, 2005.

[55] J. Bard, S. Y. Rhee, and M. Ashburner. An ontology for cell types. *Genome Biol*, 6(2):R21, 2005.

[56] D. Frank, C. Kuhn, H. A. Katus, and N. Frey. The sarcomeric Z-disc: a nodal point in signalling and disease. *J Mol Med*, 84(6):446–68, 2006.

[57] S. Schiaffino, M. Sandri, and M. Murgia. Activity-dependent signaling pathways controlling muscle diversity and plasticity. *Physiology (Bethesda)*, 22:269–78, 2007.

[58] M. W. Berchtold, H. Brinkmeier, and M. Muntener. Calcium ion in skeletal muscle: its crucial role for muscle function, plasticity, and disease. *Physiol Rev*, 80(3):1215–65, 2000.

[59] A. Nori, G. Valle, E. Bortoloso, F. Turcato, and P. Volpe. Calsequestrin targeting to sarcoplasmic reticulum of skeletal muscle fibers. *Am J Physiol Cell Physiol*, 291(2):C245–53, 2006.

[60] R. Bassel-Duby and E. N. Olson. Role of calcineurin in striated muscle: development, adaptation, and disease. *Biochem Biophys Res Commun*, 311(4):1133–41, 2003.

[61] A. Scime and M. A. Rudnicki. Anabolic potential and regulation of the skeletal muscle satellite cell populations. *Curr Opin Clin Nutr Metab Care*, 9(3):214–9, 2006.

[62] J. E. Anderson. The satellite cell as a companion in skeletal muscle plasticity: currency, conveyance, clue, connector and colander. *J Exp Biol*, 209(Pt 12):2276–92, 2006.

[63] X. Shi and D. J. Garry. Muscle stem cells in development, regeneration, and disease. *Genes Dev*, 20(13):1692–708, 2006.

[64] J. A. Timmons, O. Larsson, E. Jansson, H. Fischer, T. Gustafsson, P. L. Greenhaff, J. Ridden, J. Rachman, M. Peyrard-Janvid, C. Wahlestedt, and C. J. Sundberg. Human muscle gene expression responses to endurance training provide a novel perspective on Duchenne muscular dystrophy. *FASEB J*, 19(7):750–60, 2005.

[65] C. A. Joslyn, S. M. Mniszewski, A. Fulmer, and G. Heaton. The gene ontology categorizer. *Bioinformatics*, 20 Suppl 1:i169–77, 2004.

[66] A. V. Loguinov, L. M. Anderson, G. J. Crosby, and R. Y. Yukhananov. Gene expression following acute morphine administration. *Physiol Genomics*, 6(3):169–81, 2001.

[67] M. Kanehisa, S. Goto, M. Hattori, K. F. Aoki-Kinoshita, M. Itoh, S. Kawashima, T. Katayama, M. Araki, and M. Hirakawa. From genomics to chemical genomics: new developments in KEGG. *Nucleic Acids Res*, 34(Database issue):D354–7, 2006.

[68] K. Degtyarenko, M. Ennis, and J. S. Garavelli. "Good annotation practice" for chemical data in biology. *In Silico Biol*, 7(2 Suppl):S45–56, 2007.

[69] K. Degtyarenko, P. D. Matos, M. Ennis, J. Hastings, M. Zbinden, A. McNaught, R. Alcantara, M. Darsow, M. Guedj, and M. Ashburner. ChEBI: a database and ontology for chemical entities of biological interest. *Nucleic Acids Res*, 2007.

[70] C. Boettcher, M. Fellermeier, C. Boettcher, B. Drager, and M. H. Zenk. How human neuroblastoma cells make morphine. *Proc Natl Acad Sci U S A*, 102(24):8495–500, 2005.

[71] W. H. Stuart. Combination therapy for the treatment of multiple sclerosis: challenges and opportunities. *Curr Med Res Opin*, 2007.

[72] Y. Takahashi, K. Washiyama, T. Kobayashi, and S. Hayashi. Gene expression in the brain from fluoxetine-injected mouse using DNA microarray. *Ann N Y Acad Sci*, 1074:42–51, 2006.

[73] K. Olden and S. Wilson. Environmental health and genomics: visions and implications. *Nat Rev Genet*, 1(2):149–53, 2000.

[74] O. Bodenreider. The Unified Medical Language System (UMLS): integrating biomedical terminology. *Nucleic Acids Res*, 32(Database issue):D267–70, 2004.

[75] K. G. Becker, K. C. Barnes, T. J. Bright, and S. A. Wang. The genetic association database. *Nat Genet*, 36(5):431–2, 2004.

[76] A. Hamosh, A. F. Scott, J. S. Amberger, C. A. Bocchini, and V. A. McKusick. Online Mendelian Inheritance in Man (OMIM), a knowledgebase of human genes and genetic disorders. *Nucleic Acids Res*, 33(Database issue):D514–7, 2005.

[77] D. S. Wishart, C. Knox, A. C. Guo, S. Shrivastava, M. Hassanali, P. Stothard, Z. Chang, and J. Woolsey. DrugBank: a comprehensive resource for in silico drug discovery and exploration. *Nucleic Acids Res*, 34(Database issue):D668–72, 2006.

[78] M. Hewett, D. E. Oliver, D. L. Rubin, K. L. Easton, J. M. Stuart, R. B. Altman, and T. E. Klein. PharmGKB: the Pharmacogenetics Knowledge Base. *Nucleic Acids Res*, 30(1):163–5, 2002.

[79] R. B. Altman. PharmGKB: a logical home for knowledge relating genotype to drug response phenotype. *Nat Genet*, 39(4):426, 2007.

[80] T. Hernandez-Boussard, M. Whirl-Carrillo, J. M. Hebert, L. Gong, R. Owen, M. Gong, W. Gor, F. Liu, C. Truong, R. Whaley, M. Woon, T. Zhou, R. B. Altman, and T. E. Klein. The pharmacogenetics and pharmacogenomics knowledge base: accentuating the knowledge. *Nucleic Acids Res*, 2007.

[81] R. G. Cote, P. Jones, R. Apweiler, and H. Hermjakob. The Ontology Lookup Service, a lightweight cross-platform tool for controlled vocabulary queries. *BMC Bioinformatics*, 7:97, 2006.

[82] The Gene Ontology Consortium. The Gene Ontology (GO) project in 2006. *Nucleic Acids Res*, 34(Database issue):D322–6, 2006.

[83] R. Hoffmann and A. Valencia. Implementing the iHOP concept for navigation of biomedical literature. *Bioinformatics*, 21 Suppl 2:ii252–8, 2005.