2012

# A Heuristic Ontological Model of Protein Complexes A Case Study Based on the E3 Ubiquitin Ligase Protein Complexes of Arabidopsis thaliana

Claudia DiNatale

# A Heuristic Ontological Model of Protein Complexes
# A Case Study Based on the E3 Ubiquitin Ligase Protein Complexes of *Arabidopsis thaliana*

**By Claudia DiNatale**

**A Thesis**
**Submitted to the**
**Faculty of Graduate Studies**
**through Biology in**
**Partial Fulfillment of the Requirements for**
**the Degree of Master of Science at the**
**University of Windsor**

**Windsor, Ontario, Canada**
**2012**

# Approval Page

A Heuristic Ontological Model of Protein Complexes
A Case Study Based on the E3 Ubiquitin Ligase Protein Complexes of *Arabidopsis thaliana*

By

Claudia DiNatale

APPROVED BY:

______________________________
Dr. Luis Rueda
Computer Science

______________________________
Dr. Michael Crawford
Biology Department

______________________________
Dr. Jan Ciborowski
Biology Department

______________________________
Dr. William Crosby, Advisor
Biology Department

______________________________
Dr. Andrew Hubberstey, Chair of Defence
Biology Department

Monday, September 17th 2012

## *Author Declaration*

I hereby certify that I am the sole author of this thesis and that no part of this thesis has been published or submitted for publication.

I certify that, to the best of my knowledge, my thesis does not infringe upon anyone's copyright nor violate any proprietary rights and that any ideas, techniques, quotations, or any other material from the work of other people included in my thesis, published or otherwise, are fully acknowledged in accordance with the standard referencing practices. Furthermore, to the extent that I have included copyrighted material that surpasses the bounds of fair dealing within the meaning of the Canada Copyright Act, I certify that I have obtained a written permission from the copyright owner(s) to include such material(s) in my thesis and have included copies of such copyright clearances to my appendix.

I declare that this is a true copy of my thesis, including any final revisions, as approved by my thesis committee and the Graduate Studies office, and that this thesis has not been submitted for a higher degree to any other University or Institution.

# ***Abstract***

Ontology (with a capital O) is the philosophical study of the nature of existence that was derived to define the relationships of entities that can be said to exist in nature. The concept of an ontology was later adopted by the biological sciences to formally represent knowledge within a biological domain in order to standardize the annotation of biological data, and further, enable more efficient and easier data collection, sharing, and reuse across biological and model organism databases. The Protein Ontology (PRO) is a specific biological ontology developed to represent the relationships between proteins and protein complexes. This thesis presents a revised PRO framework, modelled around *Arabidopsis thaliana* and associated SCF ubiquitin ligase complexes, with the aim to more adequately represent what is known about the process and dynamics of protein complex formation in order to better serve the broader scientific community.

## ***Dedication***

To my parents, Maria and Renato DiNatale for their endless support and many sacrifices that have provided me the opportunities I have today. To my brother, Walter DiNatale, who has always supported me and provided me guidance that has enabled me to reach my potential. Last, but not least, to my partner, my love, and best friend, Ronjon Paul Datta who has encouraged me to strive for excellence, pursue my dreams, and live life to the fullest. I love you.

## *Acknowledgements*

I would like to thank my supervisor Dr. William L. Crosby for providing me the opportunity to work in his lab and allowing me to engage in scientific research that has made me a better scientist. To my many lab-mates, past and present for their endless support, and for the new friendships we have made. A special thank you to my parents for their unconditional love and support. Lastly, to my partner, Ronjon Paul Datta, encouraging me every day to ‘keep-on-going’, and to believe in myself and in my work, I would not have been able to complete this thesis without him.

## *Table of Contents*

## *List of Tables*

## *List of Figures*

## *List of Appendices*

## *List of Abbreviations*

AI: Artificial Intelligence
APC: Anaphase Promoting Complex
ARF: Auxin Response Factors
ASK1: Arabidopsis Skp-1 like
ATP: Adenosine Triphosphate
BFO: Basic Formal Ontology
CDC53: Cell division control protein 53
COI1: Coronatine Insensitive 1
COP1: Constitutive Photomorphogenic 1
CRLs: Cullin-Ring Ligases
CRY1: Cryptochrome 1
CRY2: Cryptochrome 2
CSN: COP9 Signalosome
CUL1: Cullin 1
DAG: directed acyclic graph
DLs: Description Logics
DO: Disease Ontology
EHR: Electronic Health Record
EID1: Empfindlicher im Dunkelroten Licht 1 protein
EO: Plant Environmental Conditions Ontology
FMA: Foundational Model of Anatomy Ontology
FR: Far-red
GBROWSE: Generic Genome Browser
GFF3: Generic Feature Format
GI: Gigantea
GO: Gene Ontology
GRO: Gene Regulation Ontology
HRT1: High level expression reduces Ty3 transposition protein 1
IAA: Indole Acetic Acid
ICD: International Classification of Diseases
IHTSDO: International Health Terminology Standards Development of Organisation
JAs: Jasmonates
JAZ: Jasmonate ZIM-domain protein
Ka: Affinity constant
Kd: Dissociation constant
KEGG: Kyoto Encyclopedia of Genes and Genomes
MCM: Mini Chromosome Maintenance Proteins
MGED: Microarray Gene Expression Data
MGI: Mouse Genome Informatics
OBO: Open Biological and Biomedical Ontologies
OWL: Web Ontology Language

PFAM: Protein Families (database)
PHYA: Phytochrome A
PHYB: Phytochrome B
PHYC: Phytochrome C
PHYD: Phytochrome D
PHYE: Phytochrome E
PIR: Protein Information Resource
PO: Plant Ontology
PRO: Protein Ontology
PSI-MOD: Proteomics Standards Initiative-Modifications
R: Red
RACE-PRO: Rapid Annotation Interface for Protein Ontology
RBX1: Ring Box 1
RDF: Resource Description Framework
RDMS: related relational database management systems
RO: Relation Ontology
ROC1: Regulator of cullins 1
SCF: Skp1-Cul1-Fbox
SGD: Saccharomyces Genome Database
SKP1: S phase kinase associate protein 1
SLY1: Sleepy 1
SMAD: Small body size (S) Mothers against decapentaplegic (MAD)
SNOWMED CT: Systematized Nomenclature of Medicine Clinical Terms
SO: Sequence Ontology
TAIR: The Arabidopsis Information Resource
TGF-B: Tranforming growth factor beta
TIR1: Transport Inhibitor Response 1
TOC1: Timing of Cab Expression 1
UFO: Unusual Floral Organs
UMLS: Unified Medical Language System
UNIPROT KB: Universal Protein Resource Knowledgebase
W3C: Web Ontology Working Group of the WWW consortium
ZTL: Zeitlupe

## *Introduction*

### *I Formal Representation of Biological Complexity*

Biology, as the study of living things, was first organized into a structured hierarchy of groups of living systems by the Swedish botanist, Carolus Linnaeus. This hierarchical organization includes concepts such as ‘class’, ‘order’, ‘genus’, and ‘species’, from which the binomial naming convention of organisms evolved. For example, for the names *Homo sapiens* or *Arabidopsis thaliana,* the first term defines the genus and the second term the species. A taxonomical hierarchy of living systems classifies organisms into groups that range from the general to more specific groupings, for example, a 'kingdom' (animalia) would be more general than 'species' (*Homo sapiens*). Today, knowledge representation systems, including ontologies in the information sciences, are very much reminiscent of Linnaean classification and may be considered the modern day continuation of the Linnaean enterprise.

A biological hierarchy of groups of living systems further reduced to a micro level would include the cellular or molecular aspect of biology concerned with defining the interactions of systems within the cell including the regulation of DNA, RNA and proteins. There is a complex network of events that take place within the cell that are the consequences of cellular and environmental signals and that enable organisms to adapt, grow, develop, and reproduce. For example, DNA is transcribed into RNA which is then translated to protein, and has been coined the 'central dogma of molecular biology' by Francis Crick in 1958 (Crick, F. 1970). The basic building blocks (DNA) evidently require specific cues to turn the transcription of genes on or off to provide cues for growth and development. Proteins carry out many of their functions as multi-subunit complexes, be it minimally in a binary complex or with up to as many as 13 subunits as with the case of the Anaphase Promoting Complex (APC) - an important regulator of the cell cycle (Schreiber, A. et al., 2011). Examples of some complex biological functions mediated by higher-order quaternary protein complexes include proteins that act as chaperones to aid in the assembly of macromolecules and folding or unfolding of proteins, acting as transcription factors regulating gene transcription, and signalling other proteins or complexes including targeted protein degradation as with an E3 ubiquitin ligase (Hua and Vierstra, 2011).

With respect to plants as a model organism, a cellular plant network is increasingly complex compared to those of other higher eukaryotes being that plants are non-motile and must adapt to often a rapidly-changing environment. One example of this complexity is the requirement of photosynthesis for growth and survival. By comparison, humans may require sunlight for basic biochemical and physiological functions, for example the shuttling of calcium by vitamin D and in the production of the neurotransmitter serotonin, but sunlight is not required to the degree that it is essential for metabolism and survival. In response to environmental stresses, humans can manipulate their conditions and surroundings in order to adapt, whereas plants must rely on internal cellular signaling to induce biochemical and physiological responses. One particular example of this type of regulation, and is discussed in more detail throughout this work, is the exploitation of the 26S proteasome whereby targeted protein degradation is used to regulate the expression levels of proteins in the cell. Targeted protein degradation is linked to regulation of plant growth, development, and pathogen defense.

Taken together, the complexity of dynamic biological networks raises the question, “how can this information be described and captured in a way that is universally translated and understood? Further, how can the information be made easily accessible across all information mediums, including computational databases?” For example, a gene can be described as the “fundamental unit of heredity” or even as “a stretch of DNA”. In order to remove this ambiguity, biology-related terms and systems need to be formally defined. This thesis presents the case of adapting ontologies to formally represent biological knowledge as a tool for scientific research, with a focus on representing proteins and protein complexes as they exist in time and space.

## *II Ontology and its Relevance to Biological Research*

The term ontology originated in Philosophy, where Ontology is considered a fundamental branch of metaphysics and is concerned with the nature of existence. As early as the beginning of the 5th century B.C., ancient Greek philosophers began contemplating the nature of existence in an attempt to define that which exists. Approximately 150 years later, this metaphysical research was expanded by Aristotle whose contribution to this area can be found in the compilation entitled 'Metaphysics', which focused on answering fundamental questions pertaining to the nature of existence.

An ontology is a framework to represent relationships between objects or, in an Aristotelian sense, relationships of entities that can be said to exist in nature. The relationships derived between these objects lend a method to classify objects based upon what is similar or different, where the resultant organization then resembles a hierarchy. Similar types of hierarchical organization can be found in the biological sciences in the form of phylogenetic trees and taxonomies like that of the Linnaean classification of living systems; however, there are key differences between an ontology and a simple taxonomical hierarchy. In the main, a taxonomical hierarchy categorizes objects based on similarities and differences but does not attempt to capture meaning behind the classification as does an ontology. An ontology strives to define relationships between the objects in an attempt to model a more formal framework for the representation of reality.

The information sciences have historically adopted the concept of ontology as a way to create a more efficient and easier method for data collection, sharing, and reuse. An early pioneer of ontology engineering, Thomas Gruber recognized early-on that research and development in information systems was made difficult by an inability to share and reuse knowledge (Gruber,T.R., 1991). Thus, it was Gruber who presented the case for using ontologies in support of formal knowledge representation. In comparison to the Aristotelian usage of ontology to describe entities that exist in nature, what 'exists' in artificial intelligence (AI) systems, as noted by Gruber, are objects or elements that can be represented in some way, shape, or form (Gruber,T.R., 1993). Gruber's definition of ontology is among the most quoted in computer science and ontological literature and is as follows: “An ontology is an explicit specification of a conceptualization.” (Gruber, T.R., 1993).

The main objectives in developing an ontology are: identifying the domain of interest; indicating the scope of the domain with respect to granularity (i.e. single molecule verses whole populations), providing a language to represent time in the ontology as well as defining a vocabulary of terms in the form of 'classes' and 'relations' - all of which would have a some definition including defined restrictions upon how these objects or entities are represented. The language adopted to model the ontology would impose its own framework and restrictions on how these components can be formed or interpreted and, thus, the language choice is a very important part of the ontology modeling process. Ultimately, a well-crafted ontology requires that the text is human-readable but also easily interpreted by machines or computers, such that it can be universally understood, shared and reused by knowledge bases and researchers alike.

Interestingly, ontologies have been available in the medical field since the 17th century where lists were drawn up to describe the ways in which people died. For example, the formal term 'French pox' was used to describe all deaths associated with this specific illness (Bodenreider, O. and Stevens, R., 2006). Arising from these early efforts, a controlled vocabulary that classified diseases was published in the late 1880's as the "International Classification of Diseases" (ICD), and is still being used within hospitals today (Bodenreider, O. and Stevens, R., 2006; cf. http://www.who.int/classifications/icd/en/ ). Subsequently, the advent of more formal and expressive knowledge representation languages such as the description logics (DLs) enabled other medical ontologies to be developed like SNOMED Clinical Terms (SNOMED CT; Systematized Nomenclature of Medicine; Bodenreider and Stevens, 2006). SNOMED CT is used in more than fifty countries and provides the main clinical healthcare terminology for the electronic health record (EHR; see the 'About SNOMED CT' section of the International Health Terminology Standards Development of Organisation [IHTSDO]; http://www.ihtsdo.org/snomed-ct/snomed-ct0/ ; accessed July 24th 2012).

Not long after AI systems adopted ontologies to represent knowledge, the biological science community viewed this as an opportunity to represent biological data a means to facilitate access to and interpretation of biological data by the broader scientific community. This initiative arose at an opportune time since genomics-related technologies began to emerge and to produce large amounts of genomics-related data – principally through high-throughput DNA sequencing methods. From the mid- to late-90's several genomes were sequenced and annotated including yeast, fly, human, and *Arabidopsis thaliana.* These genomes continue to be updated on a regular basis while simultaneously new genomes are being sequenced and annotated. Indeed, the amount of genomics-related data continues to expand rapidly and requires a system for data collection, reuse, and sharing, as required with AI systems. The Gene Ontology (GO) was one of the very first biological ontologies developed by the GO consortium in a joint project of three model organism databases: Flybase (Consortium, T.F. 1999; cf. http://flybase.org ), Mouse Genome Informatics (MGI) (Blake, J.A., et al., 2000; cf. http://www.informatics.jax.org ), and the Saccharomyces Genome Database (SGD; Ball, C.A., et al., 2000; cf. http://www.yeastgenome.org ). The GO Consortium promoted GO as a way to integrate the genomic information contained within the three databases. The

Consortium planned to create GO as a 'tool for the unification of biology' (Ashburner, M. et al., 2000). This ambitious goal was generated in light of the finding that a significant fraction of genes in diverse organisms retain common biological functions, and are shared by all eukaryotes. However, the Consortium found that the system of nomenclature for genes and gene products was divergent, despite biologists recognizing the similarities across organisms (Ashburner, M. et al., 2000). The GO consortium created an ontology, as a controlled vocabulary to represent genes and gene products with respect to their biological process, molecular function, and cellular component. As a result, these three domains represent three independent sub-ontologies contained under the umbrella of GO. This goal to unify information was a key driving force to creating interoperable model organism databases, and subsequently aided scientists in making inferences of biological roles from shared genes and proteins across organisms – a major benefit ontologies provide for scientific research. The Gene Ontology website and database can be accessed at http://www.geneontology.org .

Subsequent to the development of GO, other biologically based ontologies began to emerge. One in particular, 'The Protein Ontology' (PRO), was created to provide a 'structured representation of protein forms and complexes' (Natale, D. et al., 2011). This ontology is the focus of this thesis. PRO characterizes proteins and protein complexes at a molecular level, species specific or not, and includes definitions for proteins, protein isoforms, co- or post- translationally modified proteins and their complexes, along with the associated subunit parts. To date the protein ontology has classified over 29000 classes (protein objects), and is therefore of similar scope and complexity in comparison to GO that contains 37778 classes (www.bioportal.org , accessed July 30th 2012; cf. Noy, N.F. et al., 2009). The Protein Ontology website and database can be accessed at http://pir.geogetown.edu/pro/pro.shtml .

Specifically, PRO models proteins and complexes as objects that endure through time. In other words, the objects remaining in their native protein or complex forms indefinitely regardless of the context in which they are described. At first glance, this seems to misrepresent the true nature of proteins and complexes in that protein complexes are made up of protein subunits in which the combination of subunits, and thus the function of the complex, is variable depending on the environmental conditions surrounding the cell or organism. For example, an Arabidopsis E3 ubiquitin ligase complex has the potential to form in different cellular compartments and with different combinations of subunits, depending on the stimulus (see section below, 'SCF Complexes as a Class of Model Protein Complexes in *Arabidopsis thaliana'* '; cf. Hua and Vierstra, 2011). Thus, the complex is defined as an E3 ubiquitin ligase, regardless of whether the subunits change in response to environmental conditions. Since the complex subunit composition is variable, the associated object class and definition does not coincide with a complex of an enduring type, but rather is better defined as a perduring or temporal type (see sections IV and 1.1 regarding enduring and perduring entities).

In this thesis, I present the case for expanding PRO to include a more robust framework that can represent protein forms and complexes as they exist across time and space. This expansion has been modeled upon *Arabidopsis thaliana* as a complex model organism within which protein complexes as objects can be more credibly defined in response to the dynamic changes in the environment that are typical of the adaptive strategy of plants.

### *III Knowledge Representation in the Information Sciences: Understanding "Classes" and "Relations"*

Knowledge representation emerged as a branch of artificial intelligence (AI) essentially to do what the name implies; represent knowledge pertaining to a specific area or domain using a computational system. In order to represent knowledge in the AI domain, the area of knowledge representation requires the foundations of logic and ontology in order to properly define, model, and represent the entities that are said to 'exist' within the domain that is to be defined. The formal logic (syntax and semantics) required to represent knowledge has its roots in first-order logic, however, it has been greatly enhanced to better represent knowledge in a more structured and formal way that can be easily understood (Baader, F. et al., 2004). The main class of languages used to represent knowledge - especially in an ontological format - are the description logics (DLs). DLs provide the means to describe notions of a domain by concept descriptions, hence the term 'description' in the title. DLs can describe concepts (unary predicates) and roles (binary predicates). As well, they provide an inference capability to deduce implicit knowledge from explicitly represented knowledge in the form of a 'subsumption' algorithm, that is, a framework to represent sub- and super- class relationships (Baader, F. et al., 2004). These attributes make DLs ideal for ontology languages and representation.

Ontologies can be modeled with varying degrees of complexity depending on the type of application required, and this is reflected by the type of language used. For example: a 'highly informal' ontology is expressed in a natural language, a 'semi-informal' is expressed in a more restricted and structured natural language with reduced ambiguity, a 'semi-formal' is expressed in an artificial yet formally defined language and finally a 'rigorously formal' ontology would contain formally defined terms with formal semantics, theorems, and proofs more closely resembling a DL language (Uschold and Gruninger, 1996).

Ontologies are widely used for different purposes including natural language processing, knowledge management, e-commerce, and the Semantic Web (Gomez-Perez, A. et al. 2004). The importance and value of knowledge representation is evident in the context of the Semantic Web where the main idea is to be able to represent information and data from different databases through the internet. More formally, the concept of the Semantic Web, derived by Tim Berners-Lee, is to provide a repository of information on the World Wide Web (WWW) that can be interpreted by machines and involves the inclusion of meta-data (data about the information found; Daconta, M.C. et al., 2003). Thus, a formal language is a critical component in providing the tools for a computer to interpret, process, and reason with information, along with providing the resources for a human to easily interpret it. Two main examples of widely used languages for the semantic web are the Resource Description Framework (RDF) and the Web Ontology Language (OWL). RDF was originally developed to describe various resources on the WWW, specifically web pages as the name implies. However, the language is not limited to information resources but can also model objects in the 'real-world' (Mika, P., 2007). Although useful, the RDF language was found by the Web Ontology Working Group of the WWW consortium (W3C) to be limited with respect to its expressive power in describing more elaborate hierarchical relationships (Antoniou, G. and van Harmelen, F., 2008). For example, the OWL language was created as an extension of the RDF language in order to provide a framework to describe disjoint relationships of classes and further to include properties like cardinality restrictions (Antoniou, G. and van Harmelen, F., 2008).

### *III.i Deriving the Logical Structure of Ontology in General*

Although a single standard methodology for developing an ontology does not exist, the methods that have been documented in the literature do have three features in common. First, a purpose and scope of the ontology should be identified. Second a language and logic must be chosen to construct the ontology keeping in mind issues of reuse and integration with existing ontologies. Finally, a method must be derived to 'capture' the ontology with respect to the class hierarchy and relationships between the classes (Uschold and Gruninger, 1996). This includes identifying key concepts that need to be modeled including creating unambiguous definitions of the concepts either as classes or relations and formalizing terms for the classes and relations that will suite the definitions. Moreover, when formalizing and defining the classes, there are three basic approaches to creating the hierarchy: top-down (where the classification moves from the more general to the more specific), bottom-up (where classification commences with more specific terms first) or a combination of both, in which there is no single defined correct method (Noy and McGuinness, 2001; Uschold and Gruninger, 1996). Uschold and Gruninger have suggested the combination method as superior, since it seems to balance out the level of detail required. This approach helps to ensure that terms are not so broadly defined as to become arbitrary and that the opposite situation is avoided where creating the ontology bottom-up results in terms that are too restrictive thereby making it difficult to exploit 'commonality' between concepts (Uschold and Gruninger, 1996).

There are two types of terms or entities that can be defined in an ontology: classes and instances, formally defined philosophically as universals and particulars, respectively (Smith, B., 2004; Smith, B. and Ceusters, W., 2010). The term 'class' is used to refer to some aspect of reality, which is also referred to as a 'concept', 'type', or 'kind' (Smith, B. et al., 2005). A universal is considered a more general or abstract thing, and particulars are specific instances of the universal. For example, a 'human' would be a universal, whereas a particular person would be considered an instantiation or `instance` of the human universal or class.

An ontology represents the relationships between classes or objects with relational terms called 'relations'. There are three basic forms of binary relations: “class:class,” “instance:class,” and “instance:instance” describing the relationship between two classes, between an instance and a class, or between  two instances, respectively (Smith, B. et al., 2005). The general format or syntax of representing relationships between classes is in the form: 'term A' - relation - 'term B'. For example, a `person`- is a -`human being`.

Relations that describe relationships between objects are directional in that they are applied and interpreted in one direction unless they are of a specific type, as in the case of a symmetrical or a cyclic relation. The most common and universally understood relation is 'is a', which is used in the context of describing ancestral relationships. For example, if a class 'A' is defined as having a relationship or relation 'is a' to another class 'B', this implies that A is a subclass or child of class B. This relational hierarchy is also referred to as a 'subtyping'  relation (see OBO-Edit 2 User's Guide, http://oboedit.org/docs/index.html ; accessed July 25th 2012) and can be found in other ontology languages. Examples include the OWL language, which refers to this type of relationship as 'subClassOf' (see http://www.w3.org/TR/owl-features ; accessed August 1, 2012). An `is a` relationship indicates a relationship such that a child term inherits all the properties of its predecessors while imposing new restrictions or properties on itself.

The outcome of such a network of terms and relations is described as a directed acyclic graph (DAG), where 'parent' and 'children' type nodes are connected by relational terms (edges in the graph). Each term in the hierarchy or DAG may be a child of more than one parent allowing for multiple inheritance of properties (Ashburner, M. et al., 2000). The general study of node and edge relationships is derived from `graph theory` or the study of graphs found in mathematics and computer science. For example, a graph structure resembles a hierarchical model, as with taxonomies, such graphs can also represent more complex networks (Daconta, M.C. et al., 2003).  The DAG is read in a bottom-up fashion where the top-most node or root are the more general classes that include universally accepted terms like 'entity' or 'object'. The leaf nodes that follow are more specific where each new sub-layer or sub-class in the graph becomes more restrictive in its definition and properties. As the name implies, a DAG is directed and acyclic in that the relationships (edges) connect the terms (nodes), and it is not possible for a term to be an ancestor of itself. In other words, when reading the graph, beginning at one vertex or node and following the directed relationships will never allow a sequence that leads back to the original vertex.

### *III.ii Deriving the Logical Structure of Ontology in Biology*

As stated previously, the general format for deriving an ontology is applicable to all knowledge domains including biology. With respect to the biological domain, the main criteria would be specifying the domain that is to be captured or, more specifically, defining the level of granularity and the scope of the ontology. Regarding the biomedical and biological domains, the derivation of an ontology can be facilitated by accessing the Open Biological and Biomedical Ontologies (OBO) Foundry framework (Smith, B. et al., 2007; cf. section 1.2 of thesis) since the OBO Foundry was generated to standardize biomedical and biological ontologies with respect to time and space according to the Basic Formal Ontology top level ontology structure (BFO; Bittner and Smith, 2004; cf. section 1.1 of thesis). The granularity and scope of an ontology must be considered with regards to modelling at a molecular level as for The Protein Ontology, or at a macro-level in support of anatomical entities such as those encapsulated by the Foundational Model of Anatomy Ontology (FMA; Rosse and Meejino, 2003; cf. http://sig.biostr.washington.edu/projects/fm ; accessed July 25th 2012). Subsequently, it is important to identify which bio-ontologies have already been developed to verify whether they can be re-cycled, expanded-upon or modified to accommodate the domain in question. This is easily achieved by querying the OBO Foundry database (http://obofoundry.org ; accessed July 25th 2012) or the BioPortal database (Noy, N.F. et al., 2009; cf. http://bioportal.bioontology.org ; accessed July 25th 2012) as a web portal for biological and biomedical ontologies. With respect to deriving the class hierarchy and defining the classes specifically, biological classes are defined on the premise of being used as general terms found in the biological literature such that the terms are universally understood by scientists; for example, terms like 'cell' or 'mitochondria' (Smith, B. et al., 2005). With respect to the syntax and semantics used, the most commonly used language in the biological domain is the OBO language that was in fact derived specifically for biologists due to its ease of use and simplicity (see section 2.2.1). With regard to deriving the relations used to describe the relationships between classes, the same theory of defining classes applies. This derivation of relations can also be simplified by referring to formally defined biologically related relations found in the Relation Ontology (RO; Smith, B. et al., 2005), as well as reviewing the relations used in the current biological ontologies and re-using relations where possible, or deriving new relations as the need arises. Finally, the entire process of deriving a biologically related ontology with respect to producing the hierarchy and DAG is simplified and semi-automated by the use of the ontology editor software such as OBO-Edit which was developed for biologists and specifically designed for deriving ontologies in the OBO language (Day-richter, J. et al., 2007; cf. http://oboedit.org ; accessed July 25th 2012; see section 2.2.2).

There are three main components to biological ontologies with respect to their application. The first involves representing knowledge in a controlled way that can be interpreted by both machines and humans. A second component involves providing interoperability across model organism and biologically related databases. The third aspect is primarily a result of the first two goals, and is directed to providing the tools for annotation of biological data. Taken together, bio-ontologies provide scientists with efficient access to important biological information further providing the means to generate relevant biological questions while providing the ability to infer biological relationships across organisms.

## *IV Integrating a Spatio-Temporal Dimension in Biological Ontology Modelling*

Entities that exist in nature fall into two basic categories: 'enduring' or 'continuant' to describe those entities that persist in time and remain unchanged and generally described as objects. The terms 'perduring' or 'occurrent' are used to describe entities that unfold over time, have a temporal component and are often described as processes (Bittner, T. and Smith, B., 2003 ; Grenon, P. et al., 2004). Researchers have used examples such as a person's life, that cannot exist without the person. A more specific example pertaining to a use-case example is that of the ‘el Niňo’ phenomenon, that unfolds through time and is dependent on objects like wind, water, and the earth (Bittner, T. and Smith, B., 2003).

It has been argued that the adaptation of ontology to represent knowledge in AI systems is not being applied as an ontology in its purest sense with respect to modeling the nature of biological entities as they `exist` in time and space, but rather more of a controlled vocabulary in the form of simple hierarchies without formally describing relationships between objects. With regard to using ontologies to define biological systems, it is justifiable that incorporating an ontological framework should not be applied simply as a controlled vocabulary, but rather the framework must capture the reality of biological systems as they exist in time and space. Simply put, biological systems cannot be formally defined without encompassing the 4th dimension of time. For example, in the previous examples of perduring entities depending on enduring entities for their existence, a protein complex is dependent on many factors including the protein subunits that make up the complex, a cellular component, environmental stimuli, and other properties of the proteins that make for a viable interaction without which the complex could not form. The complex itself may also change shape or structure over time depending on the environmental and cellular conditions, thus lending further importance to describing natural objects as being of a perduring type. One example of this behaviour is with protein interaction competition arising from stronger or weaker affinity profiles (affinity being measured as an equilibrium constant value Ka, the inverse of the dissociation constant Kd; cf. Phizicky and Fields, 1995). Competition could conceivably result in the disruption of an interaction involving proteins bound in complex, leading subsequently to the formation of a new protein complex bearing the proteins that exhibit a higher interaction affinity for each another. Moreover, factors other than affinity can affect protein-protein interaction and stability including protein abundance, the presence of interaction co-factors, cellular conditions and compartmentation (Phizicky and  Fields, 1995). Another example is found in the case of SCF class of E3 ubiquitin ligase complexes where the canonical subunit composition of the complex includes a scaffold protein: CULLIN; an adapter protein: SKP; an E2 enzyme recruiting protein: RING-BOX; and an F-box subunit as the substrate recognition protein (Hotton and Callis, 2008; Hua and Vierstra, 2011). In general, the scaffold and adapter proteins are invariant, but it is the F-box subunit composition that changes depending on signalling molecules and to changing conditions of the cell. For example, the most commonly observed protein subunits part of an Arabidopsis SCF complex are RBX1, CUL1, and ASK1 (SKP1 homolog), in combination with various F-box proteins such as TIR1, COI1, UFO, ZTL, and SLY1 to name but a few (Hua and Vierstra, 2011). Thus, when taking the instance of a SCF complex as an example, one can imagine that the complex is in a dynamic state since the combination of subunits changes thus rendering the complex as definable as an occurrent or perduring entity (see sections *IV* and 1.1).

### *V SCF Complexes as a Class of Model Protein Complexes in Arabidopsis thaliana*

*Arabidopsis thaliana* is a small flowering plant and member of the Brassiceae (mustard family) that has been used as a premier model organism in plant research. Arabidopsis has many advantages as a model organism including its small genome size (~125 megabase pairs), short generation time and large seed-set (The Arabidopsis Genome Initiative, 2000). Arabidopsis was the first flowering plant to have its genome completely sequenced (The Arabidopsis Genome Initiative, 2000).

As sessile organisms that are transparent to abiotic changes in their environment, plants have evolved sophisticated mechanisms to adapt and survive under rapidly changing environmental conditions. This adaptive strategy includes the evolution of efficient physiological and biochemical methods to cope with abiotic and biotic stress. These mechanisms include the evolution of molecular pathways that protect the plant in favour of survival and reproductive success. One such molecular mechanism that is particularly emphasized in plants is the E3 ubiquitin ligase pathway that target proteins for degradation via the 26S proteasome. Protein complexes, like the SCF complex, participate in the E3 ubiquitin ligase pathway. The Arabidopsis genome has a large number of genes in each subunit-encoding gene family including approximately 700 F-box genes, 21 Arabidopsis SKP1-like (ASK) genes and 5 Cullin genes (Gagne, J. et al., 2002; Hua and Vierstra, 2011). In fact, the Arabidopsis genome contains an order of magnitude higher number of F-box proteins compared to humans, *Drosophila melanogaster* (fly), and *Saccharomyces cerevisae* (baker's yeast) with 31, 21, and 100 respectively (Calderon-villalobos, L. I. et al., 2010). Due to the large number of genes that encode known or predicted subunits in SCF complexes, a combinatorially diverse family of SCF complexes could in principle form, each with the potential for a unique functional role. In light of the potential for combinatorial complexity under varying conditions, Arabidopsis SCF complexes offer a model framework for the development of protein complex ontologies, and for this reason was chosen as the experimental foundation for the expansion of the current protein ontology described in this thesis.

In this work, I hypothesize that a revised protein ontology framework will serve to more adequately represent what is known about the process and dynamics of protein complex formation in comparison to current protein complex representation in the

Protein Ontology. Further, I predict that a revised ontology will provide added functionality for obtaining relevant protein and protein complex related information involving a structured query, while providing a mechanism to infer biological roles of shared proteins and complexes across organisms.

## *Chapter 1: Current Approaches to Ontological Modeling*

### *1.1 The Basic Formal Ontology (BFO)*

Ontologies have provided scientists an invaluable tool for discovering biological information across databases, and it is this interoperability of databases that is the key to enabling and facilitating their ongoing development. In order for ontologies to be orthogonal, there must be consistency in the framework used for representing the objects with respect to time and space. It is this issue that prompted the development of The Basic Formal Ontology (BFO) as a top-level ontology formulated to normalize domain ontologies for bio-medical and biological applications. The BFO website and information relevant to the framework can be accessed at: http://www.ifomis.org/bfo/.

As described by others, normalization requires identifying a specific set of 'trees' or hierarchies to form the back bone of the ontology which includes the representation of specific categorical distinctions, as well as a suitable set of binary relations (Bittner, T. and Smith, B., 2004). For example, making clear distinctions between separate domains of an ontology as in the case of the Gene Ontology that has three sub-ontologies. An additional example can be found in the case of this present work directed to modifying PRO, by separating the purely object based domain from the processual domain that will segregate protein objects from protein complexes respectively.

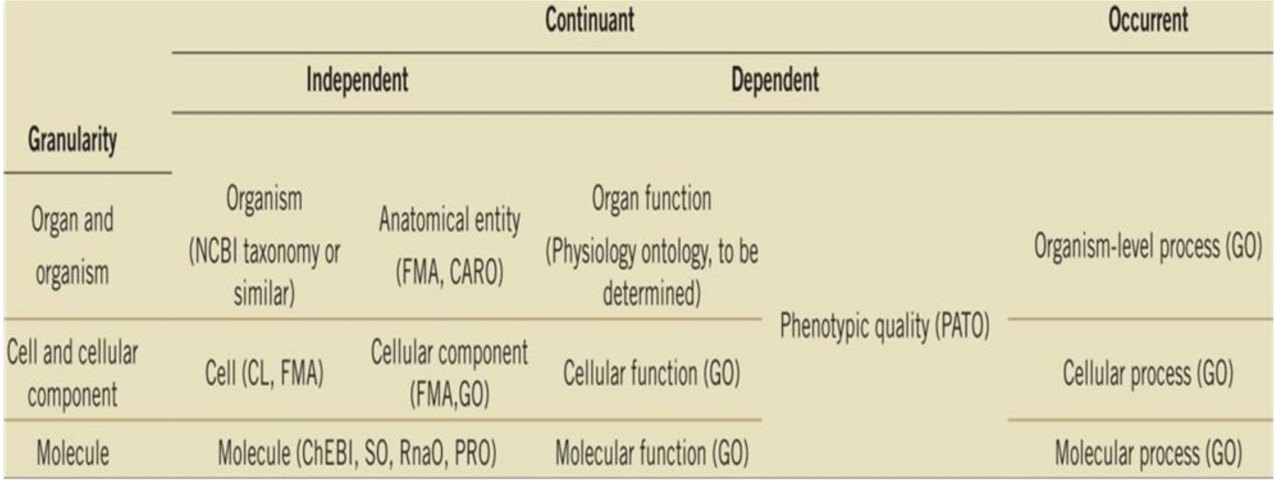
BFO provides a framework to organize bio-medical ontologies along two dimensions: representing different levels of granularity, from a single molecule to whole populations and, second, with respect to time for example persisting (enduring) or emerging (perduring) objects through time. Previous publications concede that a ‘good ontology' is able to capture this dichotomy of time – both 'synchronically' and 'diachronically', meaning as ‘reality’ exists at a given point in time and as it unfolds through time respectively (Grenon, P. et al., 2004). The methodology adopted by the BFO consortium to develop the framework is stated as 'realist' (that reality exists independent of the supposed representations), 'fallibilist' (accepting that the theories and classifications can be revised), 'perspectivalist' (there can be multiple, equally accepted, perspectives on reality) and 'adequatist' (opposing reductionism in that there is one ‘privileged’ perspective that all alternate representations are implied in) (Grenon, P. et al., 2004). By adopting this approach, the BFO framework maintains that it is able to capture reality or a “...genuine knowledge of the world” by using both common sense and a scientific approach (Grenon, P. et al., 2004).

In order to capture the dichotomy of reality as it pertains to time (objects verses processes), the BFO constructed two main types of ontologies - 'SNAP' and 'SPAN' - that provide the basis for the sub-ontologies of BFO under which they fall. The SNAP ontology represents 'continuant' entities that endure through time, where the name is derived from the concept of a 'snapshot' ontology. Examples of endurant or continuant entities provided in the BFO ontology include a heart, a symphony orchestra, and the color of a tomato. On the other hand, the SPAN ontology represents 'occurrent' or perdurant entities that  have temporal qualities, with the name derived from a description of objects that unfold through or 'span' time. Examples or span entities are the life of an organism, the formation of a synapse, and the course of a disease. The terms 'continuant' and 'occurrent' are derived from the philosophical literature relating to ontology and are defined as 'things' (objects; endurants) and 'occurrents' ( activities; events; perdurants) respectively (Smith, B. et al., 2005).

GRO is an example of a current biological ontology that maps to top-level terms of the BFO by having the root class of ‘material entity, and including both ‘continuant’ and ‘occurrent’ as root level classes of the Gene Regulation Ontology (GRO; Beisswanger, E. et al., 2008). In other cases, terms from the BFO are not directly mapped, rather snap or span concepts from the BFO are indirectly implied as in the case of the Gene Ontology. In this case the top-level classes or terms are reflective of the distinct sub-ontology, for example ‘biological process’, ‘cellular component’, and ‘molecular function’. As an OBO foundry ontology, GO is still defined within the spatio-temporal BFO framework as depicted in Figure 1.1 which outlines the scope of the OBO Foundry ontologies.

### *1.2 The Open Biomedical and Biological Ontology Foundry (OBO)*

As a result of the proliferation of biological data generated by high-throughput technologies, ontologies have been expanding in an effort to formally represent and model specific biological domains. It follows that the broader scientific community would take advantage of a powerful tool to aid in scientific discovery. However, the propagation of ontologies has also made the integration of biological data and databases more challenging. It is this barrier that the Open Biomedical Ontologies (OBO) was developed to address. The OBO Foundry was developed as a result of a strategy initiated in 2001 to support the integration of biological and biomedical data and databases



| | Continuant | | | | Occurrent |
|---|---|---|---|---|---|
| | Independent | | Dependent | | |
| Granularity | | | | | |
| Organ and organism | Organism<br>(NCBI taxonomy or similar) | Anatomical entity<br>(FMA, CARO) | Organ function<br>(Physiology ontology, to be determined) | Phenotypic quality (PATO) | Organism-level process (GO) |
| Cell and cellular component | Cell (CL, FMA) | Cellular component<br>(FMA,GO) | Cellular function (GO) | | Cellular process (GO) |
| Molecule | Molecule (ChEBI, SO, RnaO, PRO) | | Molecular function (GO) | | Molecular process (GO) |

Fig 1.1 The scope of the OBO Foundry ontologies represented by time along the x-axis and granularity (spatial scales) along the y-axis. ( Smith et al., 2007)

(Smith, B. et al., 2007). The goal of the OBO Foundry is to be an 'umbrella' for life-science ontologies that will allow database and ontology integration by standardizing the foundation upon which the ontologies are built. In order to do this, the OBO framework is modeled against the BFO upper level ontology classes with respect to time and space (granularity). There is direct evidence that the OBO Foundry (http://obofoundry.org) has succeeded in standardizing ontologies and one very good example is that of the consolidation of three cell-type ontologies into a single ontology 'cell-type ontology' (CL; Smith, B. et al., 2007; Bard, J. et al., 2005). The Gene Ontology and the Protein Ontology initiatives are among the eight founding ontologies (the GO sub-ontologies are actually counted as 3 independent ontologies in the OBO Foundry), while the remaining biological and biomedical ontologies remain under review.

OBO Foundry ontologies follow the standard 'graph-theory' structure consisting of terms (classes) connected by edges that represent 'relations', the relationships between the terms, as in 'is-a' (Smith, B. et al., 2007). The OBO consortium concedes that relations were initially used inconsistently and thus formed the 'OBO relation ontology' (RO; Smith, B. et al., 2005) to provide a controlled vocabulary of  relations for ontology engineers to use as a tool and as a method to remain consistent with other ontologies (Smith, B. et al., 2007). However, ontology developers are encouraged to derive new relations when relations are not available in the RO and new relations are required. This is evident in the case of the relation 'has-part'. This relation is not yet part of the RO but can be found as a relation in the PRO to define sub-unit protein parts of a macromolecular complex or in the case of annotating protein domains (see Figure 2.1).

The OBO Foundry is built upon an open-source framework such that ontology developers are encouraged to participate in its initiative and maintain activities within the ontological community to ensure interoperability. Thus, the OBO Foundry complies with the common format of ontology development in instructing ontology developers to survey the current database for domain specific ontologies, learn from current ontological formatting, and collaborate between groups wherever possible.

### ***1.3 The Gene Ontology (GO)***

Biological research and experimentation has been significantly impacted by the advent of novel methods of high-throughput genome sequencing. The use of high-throughput methods has facilitated and accelerated the process by which entire genomes can be sequenced and annotated. Molecular biologists now have entire genomic profiles readily available, thereby enabling the analysis of genomic information including the ability to identify the conservation of genes and proteins across different organisms.  One of the first discoveries of evolutionarily conserved genomic information was attainede by comparison of the first two eukaryote genomes sequenced;  *Saccharomyces cerevisiae* (budding yeast) completed in 1996 and *Caenorhabditis elegans* (nematode worm) completed in 1998 (Ashburner, M. et al., 2000). The comparison of these two genomes revealed evidence for a large number of orthologous genes, most of which were found to play  roles in fundamental biological processes such as DNA replication and cell division (Ashburner, M. et al., 2000).

The Gene Ontology Consortium (GO; http://www.geneontology.org) discovered early-on that the ability to decipher biological roles across organisms would be invaluable as a support to biological research. However, at the same time, the Consortium acknowledged the challenge of developing some form of computationally based framework to allow scientists to store and retrieve data. The Gene Ontology was developed under the OBO framework with the goal of being able to 'organize', 'describe', 'query' and 'visualize' biological knowledge that is constantly evolving (Ashburner, M. et al., 2000). Thus, the Consortium contrived an ontology that consists of three sub-ontologies consisting of 'biological process', 'molecular function' and 'cellular component'. GO defines biological process as a “...biological objective to which the gene or gene product contributes”, the molecular function is defined as “... the biochemical activity of a gene product” and cellular component is referred to as “… the place in the cell where a gene product is active” (Ashburner, M. et al., 2000).

GO maintains the standard hierarchical design of an ontology in the form of a directed acyclic graph (DAG; see section *III.i*) that includes the relations ‘is a’, ‘part of ‘ and ‘regulates’ to describe relationships between classes. For example, the ‘is a’ relation would describe common sub-typing relationships, while the ‘part of’ relation would describe ‘part-whole’ relationships, for example where mitochondria is ‘part of’ the cytoplasm that is ‘part of’ the cell. The ‘regulates’ relation includes two sub-relations; ‘positively regulates’ and ‘negatively regulates’. The Consortium defines ‘regulates’ relations as processual relationships where one process may directly affect the outcome of another process or quality. For example, a cell cycle checkpoint ‘regulates’ the cell cycle or in another scenario, some enzymatic reaction that may affect a quality like pH (see http://www.geneontology.org/GO.ontology.relations.shtml ).

GO was designed with the goal of providing a tool for annotating genes and gene products. For example, Figure 1.2 outlines implementation of the root class ‘molecular function’ and its sub-classes, as well as how various genes and gene products from three different model organisms ( yeast, fly, and mouse) are classified with respect to the function they perform. As with a DAG in which representative nodes can have multiple parent classes, a gene product in GO can be associated with more than one parent. This is represented in Figure 1.2 by the MCM family of proteins, that have been shown to bind chromatin and have ATP-dependent DNA helicase activity (Ashburner, M. et al., 2000). Since macromolecular complexes are found within cellular components, GO has included macromolecular complexes (this class also contains protein complexes as a sub-class) in the ontology as a sub-class of ‘cellular component’. Regarding the idea of ‘mapping’ and reusing terms across ontologies, this is evident in the case of PRO, which uses the GO terms ‘macromolecular complex’ and associated protein complex sub-class terms within the PRO ontology. The GO enterprise, including the ontology and database, has become a widely accepted research tool used in many aspects of biological research. This includes GO databases being available at various model organism databases, its adoption in microarray research and analysis (Ashburner, M. et al., 2000) and its inclusion as a standard annotation format used in biological schemas for construction of biologically

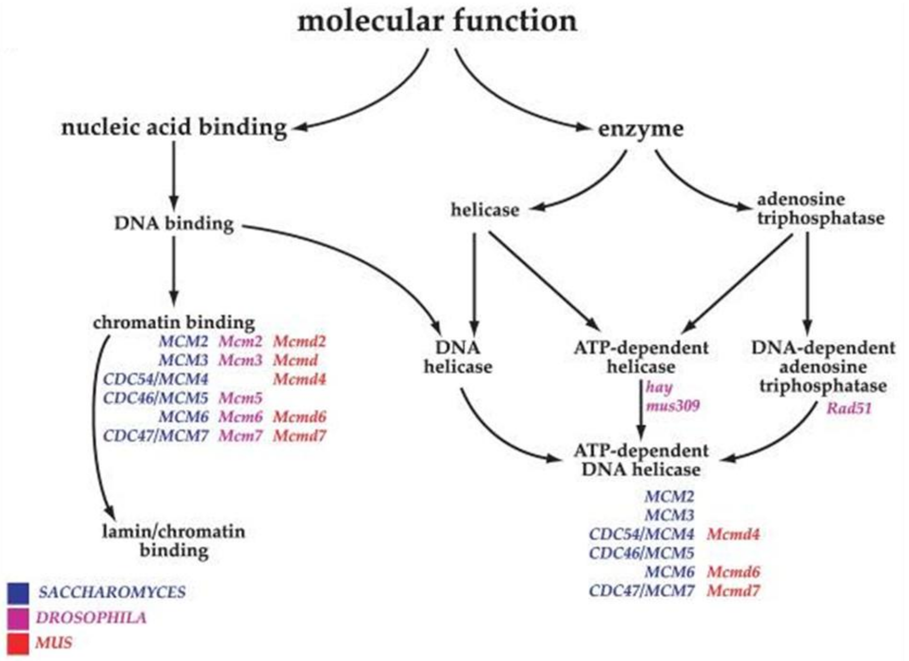


Fig 1.2 Snapshot of the GO hierarchy sub-ontology ‘molecular function’ ( Ashburner et al., 2000)

related relational database management systems (RDMS) like ‘CHADO’ (Mungall and Emmert, 2007).

### *1.4 The Protein Ontology (PRO)*

Following the development of the Gene Ontology several other biological ontologies emerged including the Protein Ontology (PRO). PRO focuses on the representation of protein forms and complexes and the relationships between these forms. The original framework of PRO included only protein forms, specifically those found in human, mouse, and *Escherichia coli*, and was later modified to also include macromolecular complexes (Natale, D. et al., 2011). Protein ‘forms’ specifically includes the representation of proteins, protein isoforms, co- and post-translationally modified forms, and with either generic or species-specific qualifiers (Bult, C. et al., 2011). PRO is comprised of three sub-ontologies that includes 'ProEvo' representing proteins based on evolutionary relatedness, 'ProForm' representing protein isoforms, mutation variants, or modified forms, and 'ProComp', which represents macromolecular complexes (Natale, D. et al., 2011).

The PRO model is composed of specific categories called 'levels of distinction' that provide further detail on the organization of PRO within the mentioned three sub-ontologies. For example, the ProEvo ontology includes two ‘levels’: family and gene where the 'family-level' category is comprised of protein products that are derived from a distinct gene family that share a common ancestor, whereas the 'gene-level' defines proteins more specifically as the products of a 'distinct gene (Natale, D. et al., 2011; Bult,

C. et al., 2011). For example, PRO defines a 'SMAD' protein or 'TGF-B receptor-regulated SMAD' at the ‘family-level’, whereas a 'SMAD2' or 'SMAD3' protein, that are the products of a more distinct *SMAD* gene, are characterized at the ‘gene-level’ (Natale, D. et al., 2011; Bult, C. et al., 2011). In addition the ProForm sub-ontology maintains two other layers of classification called 'sequence-level' and 'modification-level' that represent isoforms or splice variants and co- or post-translationally modified proteins, respectively. With respect to the ontology hierarchy, the family-level distinctions would be considered the 'parent-most' node and the modification-level would be the 'leaf-most' node. On the other hand, the ProComp sub-ontology contains 'complex', 'organism', 'subunit', and 'modification' level distinctions to represent complexes with respect to specific species and subunits (modified or not).

PRO is now regarded as one of the eight official OBO Foundry ontologies representing proteins at a molecular level of granularity, and as an 'independent continuant' with respect to time as outlined by the BFO. Specifically, the PRO framework is modeled to contain its ontology under the upper ontological class defined by the BFO as 'material entity', in which protein forms and complexes are considered 'objects'. With respect to the hierarchical ontological framework, PRO uses four main relations to describe the relationships between the protein objects: 'is a', 'derives from', 'has part' and 'only-in-taxon'. The former two relations 'is a' and 'derives from' have been formally characterized as OBO relations found in the relation ontology (RO; Smith, B. et al., 2005; cf. http://www.obofoundry.org/ro ) where the latter two relations were formulated by the PRO consortium.

PRO maintains that the protein information contained in the ontology is obtained by manual curation of scientific literature and by “large-scale processing” of resources, like the 'UniProt KB' database (Consortium T.U., 2010) that contains curated protein and pathway related data (Natale, D. et al., 2011). It is specifically the manual curation that the PRO consortium concedes is required to maintain a reliable, “...high-quality core data set”, but that will also limit the speed and coverage of further development (Natale, D. et al., 2011). PRO is continuing to further improve the available tools for curation, and in compliance with the foundations of the open-source ontology community has included a method for community curation and annotation called 'RACE-PRO' (Rapid Annotation Interface for Protein Ontology; see: http://pir.georgetown.edu/cgi-bin/pro/race_pro). RACE-PRO can be used by the broader scientific community (mainly domain 'specialists') to submit terms and protein related information that is subsequently reviewed by the PRO Consortium prior to inclusion in the ontology as protein-related data (Arighi, C.N., 2011; Natale, D. et al., 2011). This community annotation effort enables the continued development of PRO specifically with regards to protein related information comprised of mainly use-case scenarios that will further aid scientists with domain-specific research. One notable example is the case of the PRO expansion represented in this work which was the impetus for Arabidopsis being included as an organism sub-class and part of PRO as of December 2011 (PRO release 25).

### *1.5 Limitations of Current Approaches*

The coordinated effort to create and apply an upper-level ontology (BFO) for the OBO Foundry in order to standardize biological and biomedical ontologies has been and continues to be very beneficial to the ontological community. On the one hand, such efforts help to standardize all ontologies and force the community to consolidate overlapping ontologies. Secondly, it contributes positively to interoperability between model organism datasets. Third, it forces the ontological community to think critically about their ontologies and formulate representations according to what is required with respect to time and space (objects, processes, and granularity). Ontologies are consistently being expanded and modified in attempt to authentically represent that which 'exists', and these modifications are not exclusive of upper-level ontologies.

The BFO has formulated its ontology on the basis of original philosophical ideas regarding that which ‘exists’ and thus presents a very good foundation for an upper-level ontology. However, the BFO too can be expanded or modified based upon what is required from the ontological community. One example has already presented itself from the ontological community emphasizing the need for new classes. In one case, arguments have presented that the current BFO does not exhaustively cover all possible types of ‘material entities’, and thus does not adequately capture the reality they were trying to model (Vogt, L. et al., 2012). These authors point out that most biomedical material entities are in fact a hybrid of objects and fiat parts (parts of objects), and thus suggest that the ontology be expanded to include new ‘material-entity’ categories to accommodate this reality.

With respect to proteins and protein complexes, the PRO has formulated its ontology to include proteins and complexes as objects under the 'material-entity' class of BFO, without a temporal component. One may argue that both proteins and protein complexes can be considered 'objects' as they persist in time, and this would seem accurate if the protein objects and complexes are being defined at the point from which they emerge as entities from complex biological processes. For example, if a protein is being defined at the point of emergence after translation of mRNA and not taking into account earlier intermediates in the production of the polypeptide. Another example involves defining a protein complex as an object that emerges as a result of a spontaneous combination of protein subunits (objects) interacting, while ignoring the incremental protein:protein subunit interactions leading up to formation of the fully mature complex. Pointing out that proteins and protein complexes are emergent entities requires that the objects themselves inherit a temporal quality. However, with respect to the BFO definitions of objects versus processes, where an object maintains its structure or identity through time whereas a process does not, and in keeping with the current framework of the PRO designed to simplify existing efforts at expanding PRO, a protein object is assumed to *not* inherit temporal qualities. On the other hand, it would be misleading to assume that a protein complex can be represented as simply an object lacking temporal qualities. Quaternary protein complexes incorporate subunit parts that can change depending on various environmental or developmental factors, rendering the protein complex dependent on other entities or processes. In the cellular environment specifically, protein complexes often form in response to specific cues in order to carry out provide biological function. There is an inherent process that must take place in order for a given complex to form, for example, in response to a signal from a hormone signal or stress. One specific example is that of an Arabidopsis SCF complex containing the F-box protein TIR1. The process leading to the formation of the $SCF^{(TIR1)}$ complex is not entirely understood, however a critical function of the complex is to regulate downstream gene activation processes in response to the phyto-hormone auxin that binds to the TIR1 subunit and activates the SCF complex of which it is a part.  Auxin ( indole-3-acetic acid), is an important phyto-hormone involved in and regulating many plant developmental processes including, but not limited to, embryogenesis, root and stem elongation, photo- and gravitropism (Calderon-villalobos, L.I.  et al., 2010).The molecular function of the $SCF^{TIR1}$ complex is to target IAA proteins for degradation via the 26S proteasome, thus liberating auxin response factors (ARFs) as transcription factors involved in the up-regulation of a set of downstream auxin-responsive genes (Calderon-villalobos, L.I. et al., 2010; Hua and Vierstra, 2011; Tan, X. et al., 2007)

There seems to be a limitation in the current BFO classes with respect to capturing the temporal qualities of a protein complex. Currently, the BFO occurrent class contains process-related subclasses that include:*processual entity, fiat process part, process, process aggregate, process boundry and processual context* (see; http://bioportal.bioontology.org/ontologies/40358?p=terms ). The class that more precisely defines the properties of an emergent protein complex is 'process boundary' defined as "A processual entity that is the *fiat* or *bona fide* instantaneous temporal process boundary" and includes the examples of: birth, death, and the formation of a synapse. This definition recognizes the initiation or final emergent form of a process with respect to the boundary element. Thus, it may be suitable to define the temporal qualities of a protein complex since a protein complex 'forms' or emerges over time like the forming of a synapse. Moreover, an ‘occurrent’ class is defined as “An entity that has temporal parts and that happens, unfolds or develops through time”, and a sub-class of occurrent 'processual entity' is defined as “An occurrent that exists in time by occurring or happening, has temporal parts and always involves and depends on some entity”. The properties of a protein complex fit within both of these classes, since on the one hand a protein complex develops through time and, secondly, since the formation of a protein complex depends upon the assembly of its constituent protein parts – often in response to complex cellular and/or environmental cues.

In light of these concepts, it seems fitting that the emergence of a protein complex can be defined as a process boundary. However, what is limiting from this definition is the representation of the complex as an emergent object of the process boundary itself. For example, the action of protein complex formation can be considered the process boundary, but the protein complex emerges having physical form as a result of the action, and so would be considered an object at this point. So, without removing the temporal elements from the process of conditional protein complex formation, it seems reasonable to create a subclass of process or process boundary that would encapsulate the idea of an object that has temporal properties and emerges as a consequence of the process or boundary - for example, a ‘processual object’. However, given the current state of the BFO and availability of defined classes, it is fitting to use the most general process term that will capture the concept while not imposing restrictions on the definition that may result in an inadequate and informal representation of the concept. Thus, it seems reasonable to use the class ‘processual entity’ as a super-class to defining a protein complex since it implies the idea of a processual object by the use of the term 'entity', and the definition of the term specifies that the processual entity has temporal parts and always depends on another entity as in the case of a protein complex formed by the interaction of two or more subunits.

The aforementioned extension to the BFO upper-level ontology presents an ideal foundation for the expansion of the PRO hierarchy to accommodate the known reality of protein complex formation. Currently, PRO does not adequately capture the reality of conditional protein complex formation defined by complex formation through the interaction of protein subunits in response to cellular, biochemical, and physiological cues. Moreover, there are specific cellular and protein properties such as protein abundance, protein affinity, interaction co-factors (e.g. ATP and Calcium) and cellular conditions (e.g. pH and ionic environment) that can influence protein-protein interactions and complex formation (Phizicky and  Fields, 1995). In addition, sub-cellular compartmentation can influence the formation of a complex, with the added complexity that some protein complexes shuttle between compartments in response to a post-translational modification state. An example includes that of Cyclin A- and E- Cdk complexes that are key regulators of DNA synthesis and mitosis, and have been shown to have this shuttling behavior between the nucleus and cytoplasm (Jackman, M. et al., 2002). In light of the model protein complexes used in this work, it has been speculated that ubiquitin-mediated proteolysis occurs in both the cytoplasm and nucleus, with implications for the formation and potential shuttling of SCF ubiquitin ligases whose biological function is to target proteins for degradation to the 26S proteasome (Calderon-villalobos, L.I. et al., 2010). By way of example, the COP1 protein and COP9 signalosome (CSN) are important regulators in plant light responses and development, where it has been shown that COP1 functions in both the nucleus and cytoplasm under regulation of the CSN complex which is involved in inducing the COP1 nuclear localization (Wang et al., 2009). In the case of the CSN complex, was shown to be not only associated with photomorphogenesis, but has also been shown to regulate cullin-ring ligases (CRLs), including SCF complexes by de-neddylating/rubylating the cullin backbone of the complex (Hua and Vierstra, 2011). The finding that protein complex formation and the corresponding object state is often dependent on diverse cellular factors and conditions further lends itself to the object being defined as a 'processual entity' having temporal qualities.

GO has been used as a case study to describe the application of ontologies in the life sciences, and in the process several limitations in the framework were revealed that could be avoided by following formal logic principles (Smith, B. et al., 2004). By pointing out these issues, the ontology development community was alerted to potential problems that can arise and how to avoid them in the future. One example involved a failure to maintain the formal integrity across the ontology, for instance what it is that a relation is deemed to capture by definition. In the cited case involving taste sensation, relations had not been used consistently to represent classes between the three sub-ontologies of GO leading to the suggestion that there is a missing link between the two terms 'taste' (a biological process term) and 'taste receptor activity' (a molecular function term; Smith, B. et al., 2004). It was further pointed out that by not connecting these two classes with the appropriate link or 'relation', a false return of data from a query pertaining to the linkage information was the result. Moreover, it was pointed out that, while GO is lacking in its formal definitions of relations, the literature is scant with respect to automated solutions to such problems that that must be otherwise addressed manually (Smith, B. et al., 2004). A critical point here is that, in order to ensure the validity and integrity of an ontology, it must be manually designed and curated. One solution to these common problems would be the implementation of ontology editor software like Protégé ( http://protege.stanford.edu ) or OBO-Edit (http://oboedit.org ) that are equipped with reasoners and a set of built-in relations (Smith, B. et al., 2004).

Another example of a biological ontology that presents similar limitations is the Microarray Gene Expression Data (MGED) ontology (http://mged.sourceforge.net/ontologies/index.php ). The MGED ontology was created as a resource for describing and annotating microarray based experiments (Whetzel, P.L. et al., 2006). The ontology is presented in OWL format and is not a very extensive ontology having only, as of May 2009, approximately 233 classes (see the BioPortal website, MGED Ontology for metrics). Others have pointed to several problems with the ontology including its structure, and similar issues with the formal definitions and uses of relations and classes (Soldatova and King, 2005).

Pointing out the various limitations of different ontologies reinforces the impression that ontology construction and maintenance is a complex process. The solutions for ontology engineering problems are not entirely straight forward, especially since there isn't one defined methodological approach to their development. A basic approach to preventing similar ontology modeling artifacts is to explicitly define the ontology with regards to the scope of the ontology, which includes formally defining the classes and relations used. A second requirement is to explicitly define the model and domain of the ontology using correct language or syntax in order that the ontology can be shared across databases. Thirdly, it is important to comply with any existing upper-level ontology to ensure a standard format is applied within a specific suite of ontologies and to further ensure interoperability across databases. Finally, the ontology should be built such that it can be expanded and modified so that the ontology remains reliable with respect to the domain it is modeling over time.

### *1.6 A Proposed Model for an Integrative Spatio-Temporal Framework to Represent Protein Forms and Complexes*

In light of the complexity surrounding protein complex assembly, this work presents an expanded and modified PRO model to capture this reality as it exists in nature. The new framework includes four additional layers. First, a temporal domain is presented in a direct and indirect manner, the use of directly- as a classifier for top-level classes that will differentiate an object from a process, and the use of indirectly-for characterizing signal transduction pathway components as class properties to define the emergence and assembly of macromolecular protein complexes. Secondly, a spatial domain is introduced at a class level with respect to defining protein object forms as they exist in cellular compartments. Thirdly, conditional protein complex formation is characterized as a response to environmental stimuli - in this case, in the form of light from the visible light spectrum, and is defined at a class level. Finally, conditional protein complex formation is defined as a consequence of the biochemical properties of affinity and abundance that are represented as a class property and as a relationship property (see Chapter 2 and 3 for more information regarding reasoning and development of the model).

# *Chapter 2. Methods and Reasoning*

## *2.1 Reasoning Through Model Construction*

### *2.1.1 Methodology*

A standardized methodology for ontology development is not currently available. However, there are commonalities between most methods and, with respect to the model expansion proposed in this work, the methodology followed de-emphasizes building a new ontology in favour of building upon the existing PRO framework by following a common development protocol. This approach complies with the principles of ontological development agreed-upon by the ontology community in that ontologies, wherever possible, should be considered for re-use or integration. In light of these principles, it seemed reasonable to refrain from construction of an entirely new ontology and to work toward an expansion of the existing PRO framework.

The PRO database (http://pir.georgetown.edu/pro/pro.shtml ; accessed Aug, 11 2012) and framework were analyzed to identify the reasoning and formatting behind the model and its development, while simultaneously assessing literature and references provided via the website and available in the public domain. PRO is a development effort of the Protein Information Resource (PIR) and the PIR website (http://pir.georgetown.edu ; accessed Aug, 11 2012) was accessed for resources that pertain to the PRO development and framework. The PRO framework is described in a figure provided on the PIR website as well as in the published literature (see Figure 2.1; cf. Natale, D. et al., 2011) and was key for understanding how PRO is organized (a

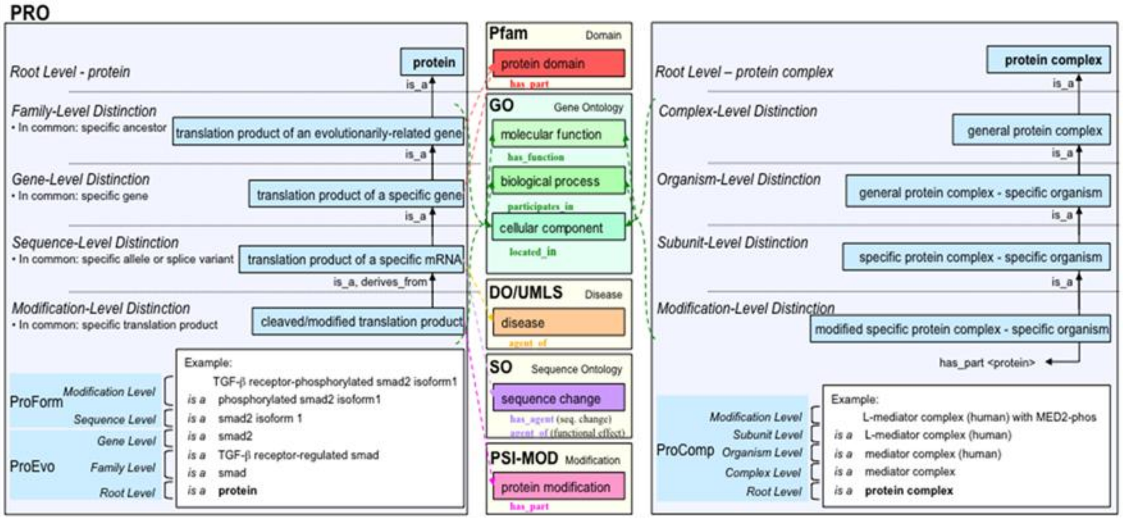

PRO
Root Level - protein
protein
is_a
Family-Level Distinction
• In common: specific ancestor
translation product of an evolutionarily-related gene
is_a
Gene-Level Distinction
• In common: specific gene
translation product of a specific gene
is_a
Sequence-Level Distinction
• In common: specific allele or splice variant
translation product of a specific mRNA
is_a, derives_from
Modification-Level Distinction
• In common: specific translation product
cleaved/modified translation product
Example:
TGF-β receptor-phosphorylated smad2 isoform1
ProForm Modification Level
is a phosphorylated smad2 isoform1
Sequence Level
is a smad2 isoform 1
Gene Level
is a smad2
ProEvo Family Level
is a TGF-β receptor-regulated smad
is a smad
Root Level
is a protein
Pfam Domain
protein domain
has_part
GO Gene Ontology
molecular function
has_function
biological process
participates_in
cellular component
located_in
DO/UMLS Disease
disease
agent_of
SO Sequence Ontology
sequence change
has_agent (seq. change)
agent_of (functional effect)
PSI-MOD Modification
protein modification
has_part
Root Level – protein complex
protein complex
is_a
Complex-Level Distinction
general protein complex
is_a
Organism-Level Distinction
general protein complex - specific organism
is_a
Subunit-Level Distinction
specific protein complex - specific organism
is_a
Modification-Level Distinction
modified specific protein complex - specific organism
has_part <protein>
Example:
L-mediator complex (human) with MED2-phos
Modification Level
Subunit Level
is a L-mediator complex (human)
ProComp Organism Level
is a mediator complex (human)
Complex Level
is a mediator complex
Root Level
is a protein complex

Fig 2.1 Framework of the Protein Ontology taken from the PIR website. The left panel includes the ProEvo and ProForm subontologies, the right panel contains the ProComp subontology representing protein complexes, and the middle panel contains the resources or databases used to define or annotate PRO terms (Natale et al., 2011)

detailed explanation of the figure can be found at the PIR website: http://pir.georgetown.edu/pro/documents/framework_Figure.pdf ; accessed Aug, 11 2012). More about the expansion of the PRO is discussed under section 2.3.2 "The Protein Ontology (PRO): Expanding the Ontology".

The OBO language was identified as the preferred language of the PRO development group and thus the expansion of PRO followed the procedures to format an ontology in the OBO language (see section 2.2.1). This procedure included locally implementing the OBO-Edit ontology editor software in order to upload PRO ontology files in obo format to view and modify the ontology (see section 2.2.2). The OBO Foundry and BioPortal were investigated for ontologies relating to the PRO and for ontologies that could be used as a reference, as well as for mapping of appropriate biological terms (see section 2.3).

The PRO ontology expansion was tested by first manually constructing a simulated version of the PRO in the local implementation of the OBO-Edit software that included the main concepts reflected in this thesis work, specifically Arabidopsis SCF complexes. Ontology searches in the local implementation of OBO-Edit were mirrored in the live PRO database search in order to test that the query results were consistent, in order to confirm the expansion of PRO could be performed using the OBO-Edit software (see section 4). The expanded PRO framework was further validated by querying the ontology using the OBO-Edit search and link panels (Section 4).

## *2.2 Program and Language Resources for the Revised Model*

### *2.2.1 OBO-Language: The Syntax Used*

The open biological and biomedical (OBO) language was developed in coordination with development of the OBO Foundry. Most OBO ontologies use the OBO file format and it is the favoured format by model-organism and other biologist communities (Smith, B. et al., 2007). This is primarily due to the language being more user friendly compared to a language like OWL, making it easier for biological domain experts to understand and operate the OBO-Edit ontology editor software. The OBO flat file format was modelled against a description logic (DL) language such as OWL. However, the language has been simplified to represent a subset of the concepts in the OWL DL language, while adding extensions for meta-data modelling and other modelling attributes that are not supported in the DL language (Day-Richter, J. 2006; Tirmizi, S. et al., 2009).

There are two main parts to an OBO formatted ontology with respect to how the language is expressed and formulated in an OBO text document (the file itself would have the extension or suffix '.obo'). The first part is the header found at the beginning of the OBO file that contains 'tag-value pairs' describing the ontology (this would include meta-data like the version of the software, the date created, the creator of the ontology etc.) The second part contains the main components that describe the domain knowledge in the form of classes and relationships (relations) referred-to as 'term' and 'typedef' respectively (Tirmizi, S. et al., 2009, Day-Richter, J., 2006). A stanza (in layman's terms, a group of lines) is used to define the concepts with respect to a term or typedef, and always begins with a unique id tag that represents the global identifier defining the object in the stanza, which will be the same in every file and ontology. There are two required tags in the stanza term that include the id and the name or term given to the object being referenced (Day-Richter, J. 2006). In addition, there are a number of optional tags, but those most often used include the definition of the term, comments, synonyms, and relationship descriptors that define relationship types between objects. Typedef stanzas may include almost all of the same tags as a term stanza with the addition of specific relationship (relation) type tags that include domain, range, as well as other relation descriptors that define the type of relation (e.g. if it is symmetric or transitive; Day-Richter, J., 2006).  By default, all OBO ontologies built in OBO-Edit come pre-formatted to include the basic, most commonly used relations or typedefs. As of 2006 these relations include ‘is a’ (the basic sub-classing relationship), ‘disjoint from’ (indicates two classes are disjoint or of separate domains), ‘inverse of’ (indicates one relationship type is the inverse of another), ‘union of’ (indicates a term is the union of others), ‘instance of’ (describes the a term being an instance of a class) and ‘intersection of’ (describes a term as being the intersection of several others; Day-Richter, J., 2006). However, recently there have been modifications and extensions to the relations used and included in the most recent version of OBO-Edit (2.2), but this information has yet to be defined or published. The aforementioned modifications to OBO-Edit regarding changes to relations included are the result of the need for ontologies to continue adapting to the evolving nature of science and proliferation of biological data over time. Figure 2.2 shows an example of a term type stanza taken from the latest PRO OBO file dated July, 23 2012. Deriving the term or typedef stanzas is a semi-automated process achieved with the OBO-Edit software (see the next section for more information).

```
[Term]
id: PR:000028458
name: SCF(TIR1) complex (Arabidopsis thaliana)
def: "An SCF(TIR1) complex consisting of TIR1, SKP1A(ASK1), rubylated CUL1, and one
of RBX1 proteins, whose components are encoded in the genome of Arabidopsis
thaliana." [PRO:CNA, PMID:21370976]
comment: Category=organism-complex.
is_a: PR:000028457 ! SCF(TIR1) complex
relationship: has_part PR:000027956 {cardinality="1"} ! protein TRANSPORT INHIBITOR
RESPONSE 1 (Arabidopsis thaliana)
relationship: has_part PR:000027960 {cardinality="1"} ! SKP1-like protein 1A
(Arabidopsis thaliana)
relationship: has_part PR:000028456 {cardinality="1"} ! RING-box protein 1
(Arabidopsis thaliana)
relationship: has_part PR:000027977 {cardinality="1"} ! cullin-1 rubylated
(Arabidopsis thaliana)
relationship: only_in_taxon NCBITaxon:3702 ! Arabidopsis thaliana
```

Fig 2.2 An example PRO term stanza in OBO format of the SCF(TIR1) complex expressed in *Arabidopsis thaliana* . The stanza includes the global identifier (id), the name as it is portrayed in the ontology, the definition of the term, the super-class shown by the is_a relation, and the relationships of subunit parts that are part of the complex having the relation has_part that includes the cardinality or number of each subunit part of the complex. Finally, the relationship of the complex being only found in Arabidopsis is specified by the relation only_in_taxon.

### *2.2.2 OBO-Edit Software: Creating the Revised Ontology*

The OBO-Edit software was developed by the GO consortium (the original was developed by John Day-Richter for biology domain experts as a standard ontology editor that would facilitate the development, editing, searching and browsing of ontologies; Day-Richter, J. et al., 2007). The software evolved from earlier ontology editor software called 'GO-Edit'. GO-Edit was originally created to edit the Gene Ontology. As various ontologies began to emerge, there was need for a more flexible software to accommodate more complex ontologies resulting in GO-Edit becoming 'DAG-Edit', where the acronym 'DAG' implied a directed acyclic graph (Day-Richter, J. et al., 2007; cf. The OBO-edit User's Guide. “OBO-Edit, The OBO Ontology Editor”; http://OBOedit.org ; accessed July 15th 2012). DAG-Edit evolved into OBO-Edit to meet the demands of the growing ontology community, especially with the advent of the OBO Foundry that developed the new OBO ontology format as the main language used for developing biological and biomedical ontologies.

Using the semi-automated process of ontology generation in OBO-Edit, a new ontology can be developed *ab initio* or from an OBO file of an already developed ontology that can be uploaded and modified or expanded as required. As it concerns this thesis work, PRO OBO files were uploaded and modified in OBO-Edit version 2.2. This process included generating new terms and relations with the OBO-Edit text editor. The OBO-Edit text editor provides the tools for generating a term stanza by manually inputting the term name, namespace (specifying sub-ontologies if required), definition, synonym and references. OBO-Edit was also used to examine PRO by viewing the hierarchy and searching for specific terms and relationships in the search panel. In order to validate the results, queries of the PRO in OBO-Edit were compared to equivalent searches at the PIR website in order to verify that the search mechanism returned the same results. One deviation in the search capacity of OBO-Edit compared to that of the PRO database was such that queries for terms linked via relations were not possible in the PRO database. The PRO Consortium has acknowledged this as a good use-case scenario and is discussing its implementation (C. Arighi, personal communication).  An example of this type of search is querying a term, for example a protein, and the search returning all relationships to either ‘parents’ or ‘children’ of the query protein. In effect, if the protein is a subunit of a complex, the search would be expected to return a result like: protein- has part- complex.

## *2.3 Ontology Framework Resources for the Revised Model*

### *2.3.1 The Basic Formal Ontology (BFO): Appropriating the Top-Level Categories*

Analysis of the PRO model and hierarchy contributed to the identification and understanding of top-level categories as defined by the BFO. The entire foundation of PRO is built under the top-level BFO class `material entity’, which is defined as being a continuant, and therefore does not impose temporality. This identification was the basis of the rationale for changing the PRO framework to include a temporal referent, specifically in order to accommodate the temporal aspects of protein complex assembly. Thus, in keeping with the current framework of PRO and its use of BFO top-level classes, and in order to comply with the ontology community standards of using BFO classes in biologically related ontologies, the BFO framework was investigated for terms that could be used to add a temporal dimension to PRO.

The BFO framework and hierarchy were accessed through BioPortal where all relationships and definitions of terms were available. BFO terms were examined by assessing the definitions and available examples. The two most apparently acceptable classes to include in PRO were the BFO upper-level classes `continuant` and `occurrent`, that when applied to PRO would immediately split the framework into two components with one including only objects , and the other including processual entities. In the case of PRO, this included transferring the entire class defined as 'macromolecular complex' to the new occurrent domain, while all other classes remained under the heading 'material entity' that now contains a new supra-class entitled 'continuant'. Currently PRO has the object referent `material entity` as a top-level class, however, it was fitting to include its supra-class `continuant` to make explicit that the ontology framework contained two separate domains that may not otherwise have been clear. More specifically, the terms and definitions of BFO in the `occurrent` spectrum were further analyzed to identify a more specific term that would more specifically define the concept of a protein complex having a temporal nature, compared to simply `occurrent`. From this work came the realization that the BFO term `processual entity` would be most suitable since it implies that the nature of a processual object has temporal parts and is always being dependent on some other entity (see section 3.1).

### *2.3.2 The Protein Ontology (PRO): Expanding the Ontology*

The PRO ontology was downloaded in OBO format from the PIR website (http://pir.georgetown.edu/pirwww/index.shtml ) and uploaded into a local implementation of OBO-Edit in order to assess the structure and format of PRO in its native form. New versions of PRO were regularly updated by the PRO developers and thus the PIR site was checked on a regular basis for these updates so that the most recent version of PRO was utilized for analysis. The last version of PRO downloaded and used for this research was release 27.0 dated April 23 2012. The ontology was searched both at the PIR website and within the local implementation of OBO-Edit for terms related to Arabidopsis SCF complexes and related protein subunits. It wasn't until the PRO ontology release 25, dated December 21 2011 that Arabidopsis and associated SCF complexes were included in the ontology (see the PIR website download section for the OBO file containing the PRO release). It was this omission in the PRO ontology that prompted communication with the PRO developers and the PIR consortium, specifically Cathy Wu (director of PIR) to initiate this development. Indeed, it was this thesis work that led to the addition of Arabidopsis and associated protein complexes (mainly $SCF^{(COI1)}$ and $SCF^{(TIR1)}$) to the PRO as a plant model organism. However, even with the addition of these complexes, it was apparent that the structure of PRO was limited with respect to describing the complicated nature of SCF complex assembly and function. This prompted further analysis of the PRO framework and related ontological literature in order to identify an appropriate framework to describe this complexity.

Ontological literature and associated databases, mainly the OBO Foundry and BioPortal, were reviewed for ontologies that pertained to proteins, protein complexes, protein and gene regulation, and plants in order to narrow down the ontologies that could be used for reference and/or mapping. This research returned very limited resources for the PRO expansion, but did lead to the identification of Gramene as a useful resource.

Gramene is a curated open source repository and information resource for comparative grass and rice genomics data (Jaiswal, P. et al., 2002; Ware, D. et al., 2002; cf. http://www.gramene.org ; accessed Aug, 11 2012). Gramene contains plant and plant environment based ontologies, specifically the Plant Environmental Conditions Ontology (EO) and the Plant Ontology (PO), that were referenced for purposes of modelling the expansion of PRO to incorporate plant abiotic stress (light) and pathway referents respectively (see section 2.3.3).

The PRO expansion involved reiteratively constructing the ontology in various class-relation formats in order to formally define conditional protein complex assembly in the most concise way without completely disrupting the current PRO framework. In addition to including pathway referents and a radiation regimen to define conditional protein complex assembly, other concepts included in the expansion were the attraction of protein subunits (affinity) and cellular compartmentation. Subsequently, and in order to represent the structure and function of an SCF ubiquitin ligase, defining the negative regulation of protein abundance is also included (see Chapter 3).

The attraction or association of protein subunits is generally represented in biological and biochemical literature by an affinity ‘Ka’ or dissociation constant ‘Kd’. The biochemical property of protein attraction is more commonly represented by the constant ‘Kd’ which is the inverse of ‘Ka’, and is represented in the units of molar concentration (cf. Nelson and Cox, 2004). This common representation of protein affinity by Kd prompted the reasoning for its inclusion in the PRO expansion. In order to characterize affinity in the ontology it was most reasonable to include affinity in two formats:  as a property of protein complex formation being represented at a class level by the constant Kd and, secondly- represented as a directional relation ‘has affinity for’ in order to follow the emergence of a protein complex (see section 3. 4).

The PRO does not currently include cellular compartmention in the ontology directly; however, it is defined with respect to the functional annotation of PRO terms. The PRO annotates cellular compartmention by mapping to the GO database and using the relation ‘located in’ (Natale, D. et al., 2011; Arighi, C.N., 2011). In order to formally define compartment-dependent complex formation and/or persistence, and to take full advantage of querying complexes with respect to localization (which cannot currently be done in the PRO database) it was fitting to include cellular compartment as part of the term definition (see section 3.3).

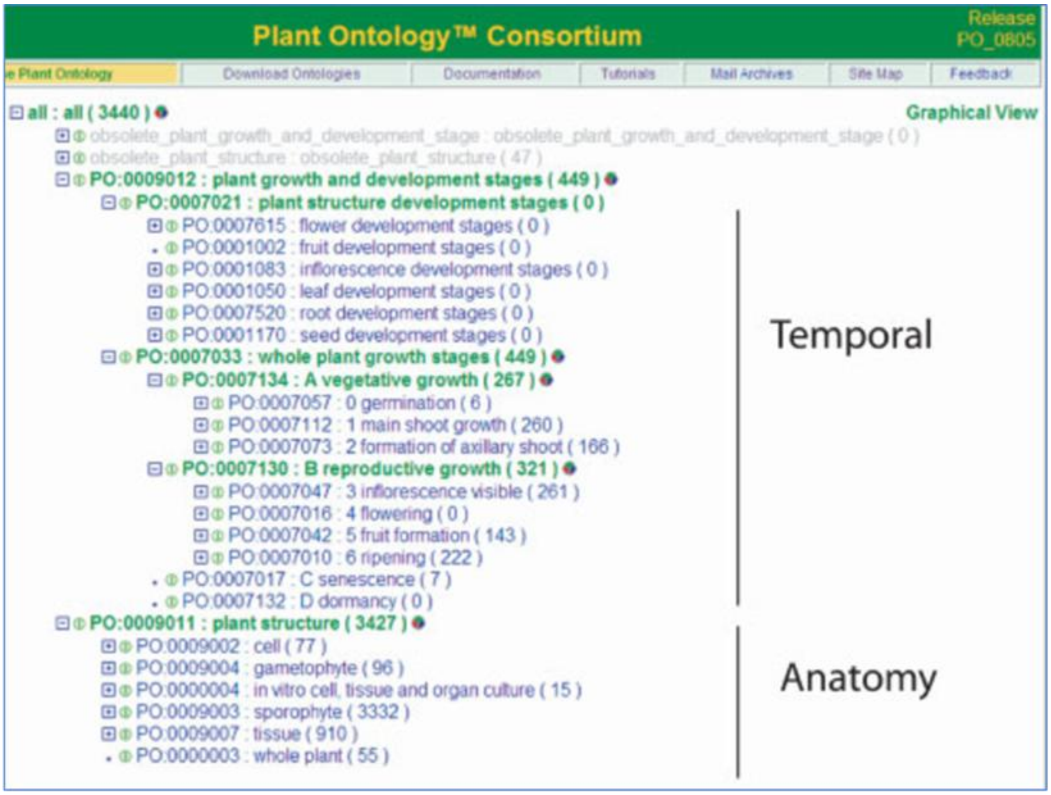
Finally, in order to accommodate what is known about SCF ubiquitin ligase structure-function complexity, and to be consistent with the addition of pathway referents in the PRO expansion, it was appropriate to include the representation of the negative regulation of protein abundance by the SCF complex. Since the function of the SCF complex is to target proteins for degradation to the 26S proteasome and therefore negatively regulate target abundance, it seemed reasonable to construct a new relation in the ontology to represent this aspect of SCF function and regulation. This new relation was defined as ‘negatively regulates abundance’ in order to precisely capture the relationship of an SCF complex targeting a protein for degradation to the 26S proteasome (see section 3.2).

### *2.3.3 Gramene*

#### *2.3.3.i The Plant Ontology (PO)*

The Plant Ontology was developed by the Plant Ontology Consortium (POC) in collaboration with several plant databases, including Gramene, to develop a controlled vocabulary to describe plant anatomy, growth, and developmental stages (Jaiswal, P. et al., 2006). The PO is coded in the OBO format, and describes plant growth and developmental stages in both an anatomical and temporal manner. It was this aspect of temporal representation that prompted further investigation with regards to the methods of the temporal modelling that could be mirrored in the PRO expansion.

The PO uses a method to include temporality into the ontology by specifying alpha and numerical values as prefixes to specific categorical terms or classes that are hierarchically represented in sequence based on the order that the events occur. For example, under the parent class 'flower development stage', there exist five main subclasses numbered 1 through 5 respectively: *flower meristem visible stage, flower meristem notched stage, flower organ development stage, anthesis stage,* and *post anthesis stage*, in which the numbers provide specificity to the order in which the processes occur (this hierarchical information was accessed from a downloaded PO OBO file released on July 3rd 2012 that was viewed in OBO-Edit v 2.2). PO uses this format throughout the ontology where temporality is required. An example of the temporal vs. anatomical format can be seen in Figure 2.3, taken from a 2006 PO publication (Jaiswal, P. et al., 2006). The prefixed values provide a method for tracking the stages of development or growth, although the values are not necessarily inherited as part of the controlled vocabulary itself. For example, if the terms were used for annotation, flower



```
⊟ all : all ( 3440 )
    ⊞ obsolete_plant_growth_and_development_stage : obsolete_plant_growth_and_development_stage ( 0 )
    ⊞ obsolete_plant_structure : obsolete_plant_structure ( 47 )
    ⊟ PO:0009012 : plant growth and development stages ( 449 )
        ⊟ PO:0007021 : plant structure development stages ( 0 )
            ⊞ PO:0007615 : flower development stages ( 0 )
            • PO:0001002 : fruit development stages ( 0 )
            ⊞ PO:0001083 : inflorescence development stages ( 0 )
            ⊞ PO:0001050 : leaf development stages ( 0 )
            ⊞ PO:0007520 : root development stages ( 0 )
            ⊞ PO:0001170 : seed development stages ( 0 )
        ⊟ PO:0007033 : whole plant growth stages ( 449 )
            ⊟ PO:0007134 : A vegetative growth ( 267 )
                ⊞ PO:0007057 : 0 germination ( 6 )
                ⊞ PO:0007112 : 1 main shoot growth ( 260 )
                ⊞ PO:0007073 : 2 formation of axillary shoot ( 166 )
            ⊟ PO:0007130 : B reproductive growth ( 321 )
                ⊞ PO:0007047 : 3 inflorescence visible ( 261 )
                ⊞ PO:0007016 : 4 flowering ( 0 )
                ⊞ PO:0007042 : 5 fruit formation ( 143 )
                ⊞ PO:0007010 : 6 ripening ( 222 )
            • PO:0007017 : C senescence ( 7 )
            • PO:0007132 : D dormancy ( 0 )
    ⊟ PO:0009011 : plant structure ( 3427 )
        ⊞ PO:0009002 : cell ( 77 )
        ⊞ PO:0009004 : gametophyte ( 96 )
        ⊞ PO:0000004 : in vitro cell, tissue and organ culture ( 15 )
        ⊞ PO:0009003 : sporophyte ( 3332 )
        ⊞ PO:0009007 : tissue ( 910 )
        • PO:0000003 : whole plant ( 55 )
```

Fig 2.3 Snapshot of the PO hierarchy representing the anatomical verses temporal domains (Jaiswal et al., 2006)

meristem' would not be referred to as '1-flower meristem', but simply ‘flower meristem’.

In this work, I present this type of temporal organization in a manner analogous to characterizing signal transduction pathways but applied to protein complex assembly and function. For example, classes or terms include numerical values as prefixes for both proteins and complexes to define the particular instance of the object, a suffix value that includes information pertaining to the total number of steps in a particular pathway the object is involved in, and a number defining the specific step at which the object is taking place in the pathway. The suffix value is also appended by the letter ‘P’ as an abbreviation of the word ‘pathway’ and an arbitrary number that reflects the particular pathway instance, for example ‘P1’ allows one to track all events pertaining to the particular instance of the pathway ‘P1’, and the total number of pathways represented in the ontology. If the first instance ‘protein-A’ is taking part in a particular pathway ‘P1’ that consists of 7 total steps, and this ‘protein-A’ is specifically interacting with a ‘protein-complex ABC’ to form a complex at step number 4, then ‘protein-A’ would be defined as: 1-protein-A-(P1:4/7), the complex would also contain the suffix ‘P1’. The reasoning behind this format is that all objects taking part in a particular pathway, for example ‘P1’, would include the suffix ‘P1’ which can then be tracked and queried in a database. Defining the particular instance of each object is important to be able to differentiate the numerous functions and pathways in which a particular object, protein or protein complex participates. As previously stated, the protein or complex in question would not be annotated with the prefix or suffix values, but simply as ‘protein-A’ or ‘protein-complex-ABC’ (see section 3.2).

### *2.3.3.ii The Environment Ontology (EO)*

The Environment Ontology is part of the Gramene development and was developed to describe different types of treatments or environments that a plant is exposed to (EO is also referred to as 'Plant Environmental Conditions'; see: http://bioportal.bioontology.org/ontologies/45260?p=terms ; or from the Gramene site: http://www.gramene.org/db/ontology/search?id=EO:0007359 ; accessed Aug, 11 2012). This ontology was specifically surveyed for its use of terms describing plant abiotic stress in order to keep the naming convention similar in the PRO expansion. This allowed the modeling to remain consistent with the ontology community with regards to re-using terms where possible. The EO specifically was the source of terms in the hierarchy that pertains solely to plants; however, the terms were not borrowed or mapped in full for the PO expansion, but rather improvised in accordance with defining protein complexes that respond to light (radiation) abiotic stress. Figure 2.4 depicts the group of radiation related terms from the EO that were used to develop the terms in the PRO expansion. The improvised terms were specifically related to light regimens that plants have been shown to respond to and that had been documented in literature. For example, plant responses to light verses dark conditions, or in the case of particular visible light spectrum the wavelengths included were red (R), far-red (FR), and blue-light (see section 3.3).

## *2.4 Assessing the Logical Coherence of an Ontology*

A 'reasoner' is a type of software built into ontology editor software that

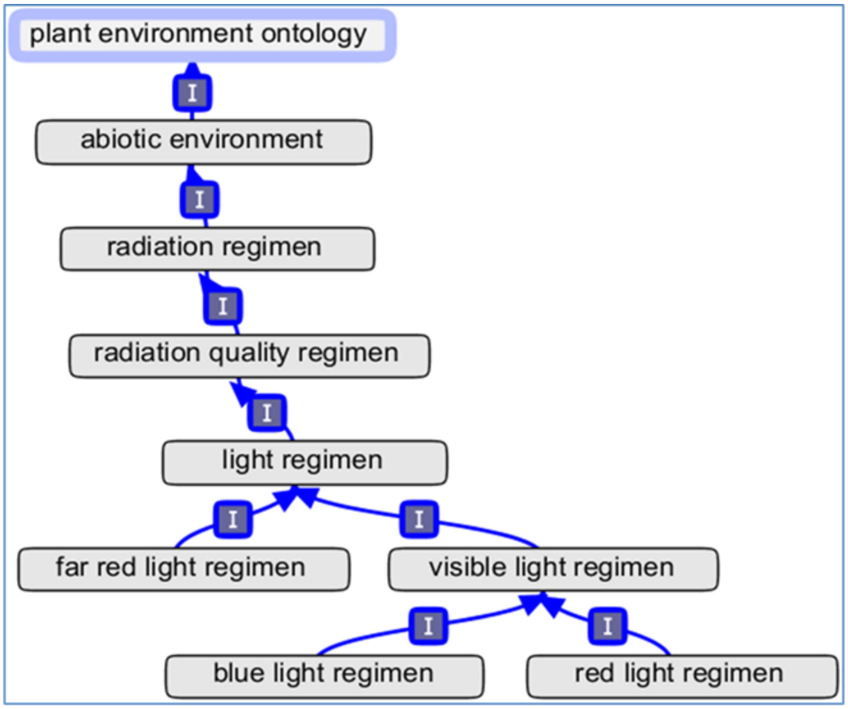


Fig 2.4 Snapshot of the EO derived in OBO-Edit depicting the light radiation related terms used in the expansion and modification of the PRO

automatically checks the consistency of the hierarchy or ontology. Reasoners come in a variety of forms and can differ in their inference procedures, type of reasoning, expressiveness and implementation language (Mishra, R.B. and Kumar, S., 2010). A reasoner can infer links or relationships between objects that are otherwise not explicit, and can also use the relationship types between objects or classes to infer more information about the ontology. The OBO-Edit reasoner is a ‘rule-based reasoner’ that uses a set of rules for making inferences about an ontology. These rules include: 1-'transitivity', 2-'simple genus/differentia implications',  and 3-'cross product definition resolution' (see OBO-Edit2 User's Guide, “The OBO-Edit Reasoner”, http://oboedit.org/docs/index.html ; accessed Aug, 11 2012). The transitivity rule is such that the reasoner infers all relationships in an ontology that are implied by transitivity, but not explicitly stated, through transitive relations like, 'is a', and 'part of'. For example, if there is a link: 'A-is a-B', and 'B-part of- C', the reasoner will infer 'A-part of-C' through transitivity. A cross product in an ontology is generally defined as a relationship between two terms in different ontologies, however OBO-Edit defines a cross-product as an 'intersection' of two or more terms. Rules 2 and 3 involve the reasoner making inferences across cross-product relationships.

OBO-Edit also includes a 'verification manager' that can perform additional ontology checks including but not limited to cycle, disjointedness, and redundancy checks. For example, a cycle check would report if there is a non-explicitly defined illegal cycle causing a term to be an ancestor of itself. The disjointedness check functionality serves to verify that no term has two 'is a' ancestors that are disjoint super-classes, whereas a redundancy check ensures that a term name in an ontology is unique (see OBO-Edit 2 User's Guide, “Introduction to Verification”, http://oboedit.org/docs/index.html ; accessed Aug, 11 2012). The verification manager immediately reports errors to the end-user enabling efficient identification and repair of the problems.

As previously stated, the OBO-Edit search and link panels were also used to check terms and relationships in the PRO expansion by imposing specific search queries related to Arabidopsis and associated SCF complexes. The search panel operations are reminiscent of a PRO database search that enables testing the ontology for logical coherence and typical search queries otherwise performed in the live PRO database. These searches specifically involved the expanded and modified concepts and terms presented here. OBO-Edit searches are explored in more detail in Chapter 4.

### *2.5 The PRO Consortium: Collaborative Networking*

This thesis work was conducted in consultation with the international PRO/PIR development group led by Dr. Cathy Wu (University of Delaware; cf. http://bioinformatics.udel.edu/People/Cathy_Wu/ ; accessed Aug, 11 2012). As described, the discovery that Arabidopsis was not defined in the PRO as a model organism led to collaborative efforts with the PRO Consortium leading to a framework that now includes Arabidopsis and associated SCF complexes. The Crosby lab group became invited members of the PRO/PIR initiative as Arabidopsis consultants, in which our contributions have already been incorporated in the PRO. This included the addition of Arabidopsis as a model organism, and to date, the inclusion of specific SCF complexes, $SCF^{(TIR1)}$ and $SCF^{(COI1)}$. All ontology developments are supervised by Dr.

Barry Smith (University of Buffalo, NY) as the director of the BFO and editor of the OBO, in order to ensure logical coherence and adhesion to generally acceptance ontology development standards (http://www.philosophy.buffalo.edu/people/faculty/smith/ ; accessed Aug, 11 2012.). Part of this 'collaborative networking' involved taking part in tele-conferences with the PRO Consortium that provided an understanding of typical ontology operation and development procedures. The on-going community and collaborative efforts within the broader ontological community is an integral part to ontology development in maintaining formal, reliable, and orthogonal ontologies.

# *Chapter 3. Revisions to the Current Protein and Protein Complex Ontological Model*

## *3.1 Adding the Temporal Component*

The PRO hierarchy has been split into two sub-ontologies to encapsulate two domains; a continuant domain representing only objects *not* inheriting a temporal component, and an occurrent domain including processual entities that include macromolecular complexes (see Figure 2.1). As previously indicated, the current PRO already contains three sub-ontologies, thus the two protein based sub-ontologies "ProEvo" and "ProForm" were maintained under the continuant domain, while the "ProComp" subontology was maintained under the occurrent domain.

### *3.1.1 Defining Continuant and its Referent*

The continuant domain includes 'material entity' and all existing sub-classes of 'material entity' currently maintained in the PRO, except for 'macromolecular complex' that is maintained under the occurrent domain. These PRO parent sub-classes of 'material entity' include: *fiat object part and object,* where the 'object' sub-class contains *amino acid chain and protein, molecular entity* and *organism.* The continuant class is defined as indicated by the BFO and is as follows: "An entity that exists in full at any time in which it exists at all, persists through time while maintaining its identity and has no temporal parts". The synonym of continuant is 'endurant', where a continuant is defined as being disjoint from the occurrent class. Figure 3.1 represents snapshots of the current PRO model (restricted to the specific classes under review in this thesis) in two separate forms

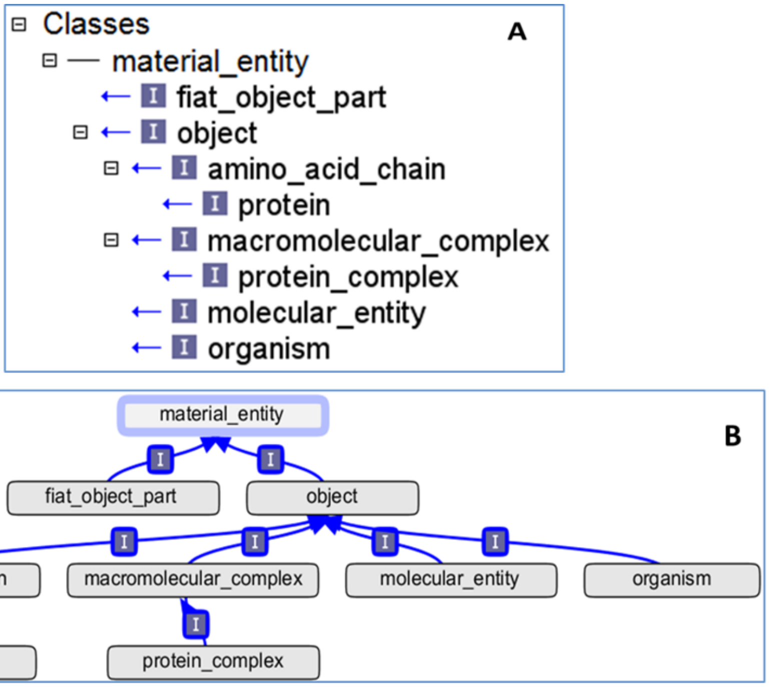


Fig 3.1 A snapshot of the PRO hierarchy representing the main top-level classes under review in this thesis in (a) hierarchical format and (b) as a directed acyclic graph produced with OBO-Edit. Note: this does not represent the entire PRO hierarchy

generated in OBO-Edit. Figure 3.1 panel A displays the tree format (as a hierarchy) and Figure 3.1 panel B displays the directed acyclic graph (DAG). The snapshots represent the main PRO classes under investigation in this work and include: ‘material entity’, ‘object’, ‘protein’, ‘macromolecular complex’, ‘protein complex’, ‘molecular entity’ and ‘organism’. The current PRO hierarchy can be downloaded in full from the PRO website in OBO format

### *3.1.2 Defining Occurrent and its Referent*

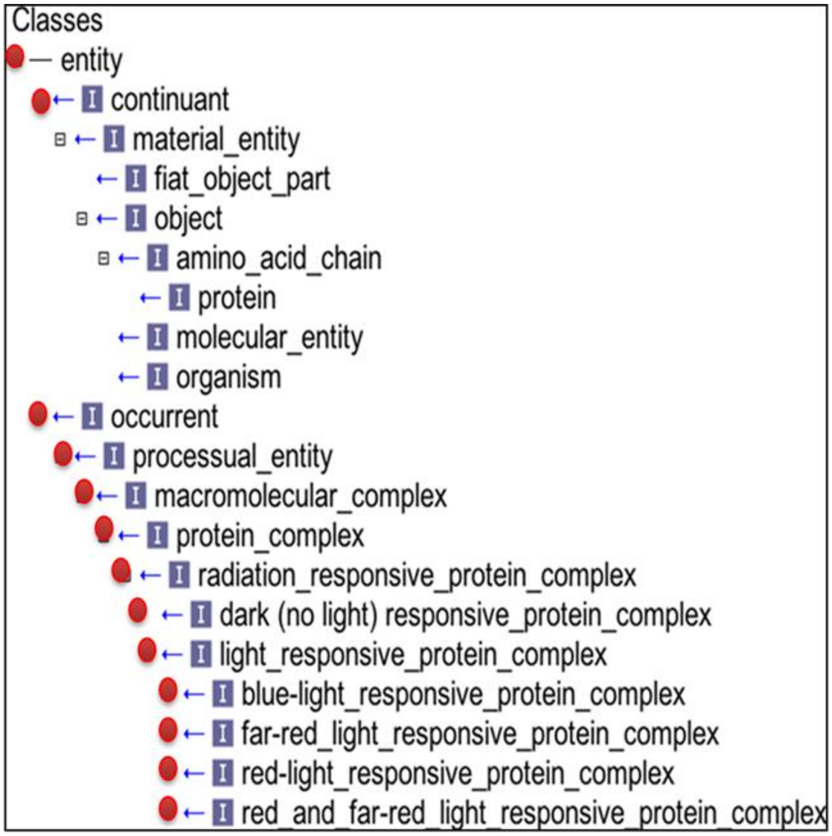
The occurrent entities include the current ‘macromolecular complex’ PRO class that contains *protein complex* as a sub-class. The occurrent class is defined by the BFO as follows: “An entity that has temporal parts and that happens, unfolds or develops through time”. The synonym of occurrent is ‘perdurant’ where occurrent entity is defined as being disjoint from a continuant term. The sub-class of continuant and super-class of macromolecular complex includes ‘processual entity’ and is defined as indicated by the BFO as follows: “An occurrent that exists in time by occurring or happening, has temporal parts and always involves and depends on some entity”.

The current PRO uses the existing macromolecular complex term from the GO, the global identifier is GO:0032991, and thus the definition remains as defined by the GO: “A stable assembly of two or more macromolecules, i.e. proteins, nucleic acids, carbohydrates or lipids, in which the constituent parts function together”. Since the term macromolecular complex resides as a sub-class of a processual entity, the term macromolecular complex inherits the properties of this sub-class through transitivity of the relation ‘is a’, and thus the definition does not need to be altered. Subsequently, the term protein complex is also mapped to the GO, the global identifier is GO:0043234, and is defined as: “Any macromolecular complex composed of two or more polypeptide subunits, which may or may not be identical. Protein complexes may have other associated non-protein prosthetic groups, such as nucleotides, metal ions or carbohydrate groups”.

New sub-classes of ‘protein complex’ include radiation-responsive protein complexes. The main sub-class of protein complex is the class ‘radiation responsive protein complex’ that will define particular protein complexes that respond to a radiation regimen as either in visible light spectrum or no light at all. These radiation-based classes are discussed in more detail under in section 3. 3. Figure 3.2 represents a snapshot of the modified and expanded PRO framework derived in this thesis work in an ontology tree format where new or modified classes are indicated by a red dot. The DAG format is not shown due to space restrictions. Table 3.1 lists the current PRO relations and the newly derived relations as part of this thesis work. See appendix A for new and old term definitions included in the PRO expansion and appendix B for new relation definitions.

### *3.2 Pathway Attributes: Analyzing Occurrent and Continuant Outcomes*

As previously stated, proteins and protein complexes may participate in numerous functional pathways. In the case of Arabidopsis SCF complexes, these complexes may form in response to various growth and development signals in order to carry out their

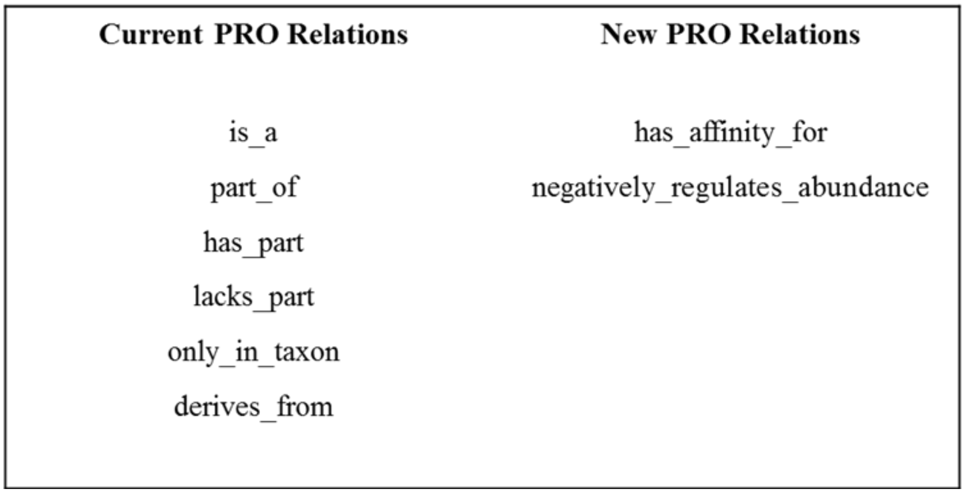


Classes

- entity
    - continuant
        - material_entity
            - fiat_object_part
            - object
                - amino_acid_chain
                    - protein
                - molecular_entity
                - organism
    - occurrent
        - processual_entity
            - macromolecular_complex
                - protein_complex
                    - radiation_responsive_protein_complex
                        - dark (no light) responsive_protein_complex
                        - light_responsive_protein_complex
                            - blue-light_responsive_protein_complex
                            - far-red_light_responsive_protein_complex
                            - red-light_responsive_protein_complex
                            - red_and_far-red_light_responsive_protein_complex

Fig 3.2 A hierarchical format of the expanded and modified PRO model produced in OBO-Edit. New or modified classes are indicated by a red dot.

Table 3.1 Relationship terms (relations) in the current PRO model and new relations derived for this thesis work as part of the PRO expansion and modification.



| **Current PRO Relations** | **New PRO Relations** |
|---|---|
| is_a | has_affinity_for |
| part_of | negatively_regulates_abundance |
| has_part | |
| lacks_part | |
| only_in_taxon | |
| derives_from | |

primary function for the targeting of select proteins for degradation. For example, in response to the presence of the plant growth regulatory compound (phyto-hormone) auxin, the SCF$^{TIR1}$ complex acts to target IAA proteins for degradation. While the order of subunit assembly of the SCF$^{TIR1}$ complex is not clear, it is known that four canonical subunits of the complex must associate in some temporal order, resulting in formation of the functional complex. These particular auxin signalling events can be characterized as a pathway that commences with the availability of molecular auxin as a functional effector, the triggering of the formation of the SCF$^{TIR1}$ complex, the binding of auxin to TIR1, and finally, the targeted degradation of  IAA proteins (Calderon-villalobos, L.I. et al., 2010; Tan, X. et al., 2007). In order to represent the events leading up to the formation and function of Arabidopsis SCF complexes, pathway properties were added to the defined instances of the protein complex and protein subunit parts. This addition to the instance definition of a protein or protein complex allows one to query a protein or related protein complex and generate an output of data that describes all events involved in the formation and function of a protein complex. These pathway attributes are not limited to describing SCF complexes but can be applied generally to the formation of any protein complex or to any signalling pathway.

### *3.2.1 Defining Proteins and Protein Complexes*

In order to define a protein or protein complex instance as part of a particular pathway the Formula 3.1 can be applied as shown below.

Formula 3.1:

[(instance no.)(term name)('P' {pathway instance no.}: step no. /total no. steps)]

- 'instance no.'= the value related to the instance of the particular entity
- 'term name'= the name of the entity in question (protein or protein complex)
- 'pathway instance no.' = an arbitrary number corresponding to a particular pathway 'P' instance
- 'step no.'= a number corresponding to the particular step an entity is involved in as part of pathway 'P' that has a defined number of steps
- 'total no. steps' = the total number of steps that are part of the particular pathway 'P'

Taking the instance of 'protein-A' as an example; this protein has been defined as the first instance of 'protein-A' that takes part in step 1 of a pathway 'P', and is deemed pathway instance 5 that is composed of 5 steps. Thus, this protein instance would be defined as: “1- 'protein-A' (P5:1/5)”.

Taking the instance of a 'protein complex AB' as a second example; this protein has been defined as the second instance of the protein complex that forms as part of step 2 in a pathway (pathway instance 4), involving 6 steps. This protein instance would be defined as: “2-'protein complex AB' (P4:2/6)”.

Although the pathway instance is currently arbitrary in that it has no definition, it allows one to track the particular pathway instance and all related protein forms that take part in the pathway. In the future, this could be modified to include desired pathway names.

### *3.2.2 The Emergent Constitution of Protein Complexes*

The emergence of a protein complex can be followed in the ontology with respect to pathway attributes indicated by the pathway instance, as well as by the inclusion of the directional relation ‘has affinity for’ (see section 3.4). All protein forms (proteins and protein complexes) that take part in a particular pathway ‘P’ will be defined with the pathway instance appended to the term or class name, and the relation ‘has affinity for’ will indicate the protein-protein interactions as well as protein-protein complex interaction.

By way of example, I use the hypothetical scenario of a pathway ‘P1’ where a ‘protein complex ABC’ targets ‘protein X’ for degradation, and the protein subunits ‘A’, ‘B’, and ‘C’ associate in sequential order. Step one of pathway ‘P1’ would be deemed the association of the first two subunits, where both protein terms would be defined as being step 1 and related by the relation ‘has affinity for’. The protein subunits would be defined in the continuant domain as having affinity for one another, however, the emergent protein complex formed ‘protein complex AB’ would be defined in the occurrent domain and would be defined as step 2 of pathway ‘P1’. An example of the P1 pathway and associated protein forms is shown in Figure 3.3.

### *3.2.3 The Emergent Functions of Protein Complexes*

The emergent function of a protein complex is currently defined in PRO by ‘mapping’ to the GO database molecular function sub-ontology, in which PRO terms are annotated with the relation ‘has function’. In addition, there are specific types of protein complexes, such as an SCF ubiquitin ligase complex, where PRO has ‘mapped’ the term

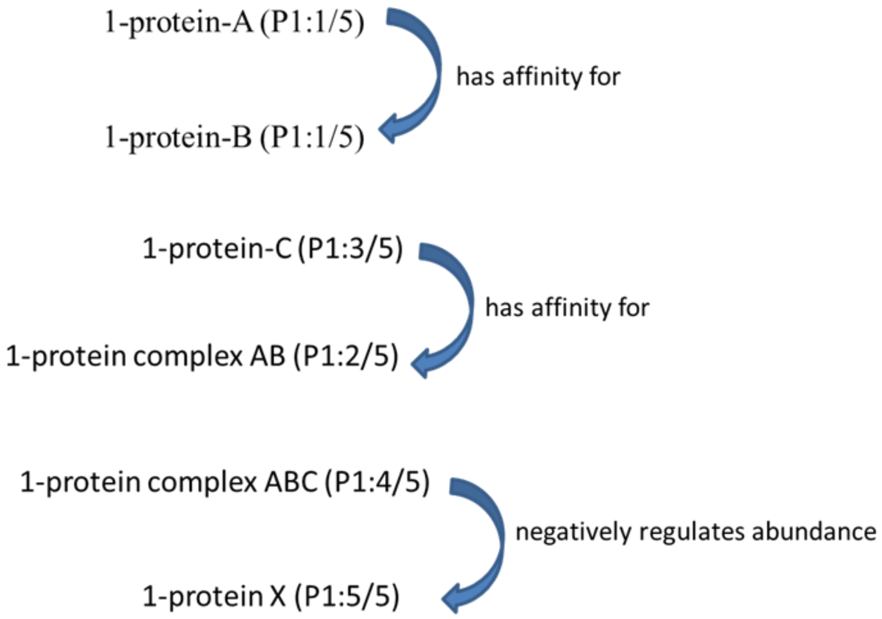


Fig 3.3 Example of the hypothetical pathway 'P1' depicting the formation of the 'protein complex ABC' and the function of the protein complex through the relation 'negatively regulates abundance' that is representative of targeted protein degradation.

to GO where the GO term has been formally defined to include a function in its definition. For example, the class "SCF ubiquitin ligase complex" (GO:0019005) is defined as: " A ubiquitin ligase complex in which a cullin from the Cul1 subfamily and a RING domain protein form the catalytic core; substrate specificity is conferred by a Skp1 adaptor and an F-box protein. SCF complexes are involved in targeting proteins for degradation by the proteasome. The best characterized complexes are those from yeast and mammals (with core subunits named Cdc53/Cul1, Rbx1/Hrt1/Roc1)". In this case, any sub-class of this particular class would inherit the properties of this class and thus, it would be assumed that the function of a SCF ubiquitin ligase is to target proteins for degradation. What is lacking in the PRO definition of a specific SCF ubiquitin ligase (for example SCF$^{TIR1}$) is the specificity of the targeted substrates. For example, the PRO definition of a SCF$^{TIR1}$ (PR:000028457) is: "An SCF ubiquitin ligase complex consisting of TIR1, SKP1A(ASK1), rubylated CUL1, and RBX1". In future, this should be modified to include that this particular complex targets IAA proteins for degradation. In the meantime, the addition of the relation 'negatively regulates abundance' will add a layer of structure-function not currently present as discussed in the next section.

#### *3.2.3.i Defining Protein Abundance Regulation*

The function of a SCF ubiquitin ligase complex is to target proteins for degradation to the 26S proteasome. In order to capture the negative regulation of protein abundance by an SCF complex, a new directional relation has been devised and defined as 'negatively regulates abundance'. This relation will provide a layer of specificity regarding the function of a SCF complex as well as contribute to pathway attributes as described earlier. This relation, in most cases, is defined as part of the last step of the particular pathway instance involving SCF complexes and targeted degradation. Specifically, the particular protein targeted for degradation will be defined as involving the final step in the pathway. The relation ‘negatively regulates abundance’ is not transitive and is formally defined as: “A relationship between an occurrent and continuant entity, in which the occurrent entity acts upon the continuant entity in a manner such that the relative abundance of the continuant object is negatively affected and results in a reduced abundance of the continuant object” (see also appendix B).

## *3.3 Conceptualizing Environmental Influences: Effects on Protein Complex Assembly*

### *3.3.1 Light: an Abiotic Stress*

Plant growth and development are in part controlled via signals perceived by the light environment, making it important to include radiation-responsive protein complex-related classes in the ontology. In particular, the control of growth and development is mediated by a network of photoreceptors that include the phytochrome and cryptochrome families (Casal, J.J. et al, 2003).*Arabidopsis* contains five phytochromes (PHYA, PHYB, PHYC, PHYD ,and PHYE) that perceive mainly red (R) and far-red (FR) light wavelengths (600 to 750nM), plus two cryptochrome photoreceptors (CRY1 and CRY2) that primarily perceive blue-light wavelengths (Casal, J.J. et al., 2003; Devlin and Kay, 2000). These photoreceptors perceive light signals and transduce signals to downstream targets that include proteins and protein complexes including SCF complexes as regulatory targets. For example, jasmonates (JAs) are a group of plant hormones that regulate diverse physiological processes including wound (herbivory) defense, growth, development, and senescence (Robson, F. et al., 2010). A link has been described between

JA and phytochrome A (phyA) signalling where wound and shade responses, mediated by phytochromes, are responsible for monitoring the change in the ratio of R to FR light (Robson, F. et al., 2010). The F-box protein COI1 has been identified as a key player in JA signal transduction where COI1 behaves as a receptor for JA as part of the SCF$^{(COI1)}$ complex in Arabidopsis. The Arabidopsis SCF$^{(COI1)}$ complex targets JAZ transcriptional repressors for degradation, resulting in the up-regulation of JA genes involved in defense, secondary metabolism, hormone biosynthesis, and JA synthesis as part of a feedback loop (Robson, F. et al., 2010).

Another example pertains to an Arabidopsis SCF complex containing the F-box subunit 'Zeitlupe' (ZTL) that responds to darkness and plays an important role in the free-running periodicity of the circadian clock in plants (Han, L. et al., 2004; Nakamichi, N., 2011). The ZTL protein exists bound in complex to another protein 'Gigantea' (GI) in the presence of blue-light, but dissociates from GI in response to darkness. This dissociation frees ZTL to join an SCF complex that in turn targets the protein 'Timing of Cab Expression 1' (TOC1) for degradation (Nakamichi, N.,  2011). The light regimen included in the ontology with respect to protein complex assembly can be applied to other organisms, but currently is beyond the scope of this thesis. Nevertheless, this expansion offers an exciting starting point from which further expansion of the ontology can take place as the need arises.

A further example is that of an F-box protein EID1 that has been shown to function as a negative regulator in phyA signalling, resulting in the regulation of photomorphogenesis in seedlings, rosette leaf development and flowering (Marrocco, K. et al., 2006). Research has implicated EID1 as a member of the SCF complex targeting phyA signalling transducers for degradation, mainly because it is an F-box protein and has been shown to associate with ASK1 and ASK2 proteins (Dieterle, M. 2001; Marrocco, K.  et al., 2006).

In order to include classes that define conditional protein complex assembly in response to particular light regimens, sub-classes of ‘protein complex’ were added in the occurrent domain. The main sub-class of ‘protein complex’ is ‘radiation responsive protein complex’ that is defined by combining the given EO definition for ‘radiation regimen’ (the global identifier is EO:0007151) in combination with the definition for a protein complex as outlined by the PRO and is as follows: “Any macromolecular complex that is composed of two or more polypeptide subunits, which may or may not be identical, that assembles and carries out some function in response to an exposure to a radiation type, intensity, or quantity”. The class radiation responsive protein complex has two main sub-classes: ‘dark (no light) responsive protein complex’ and ‘light responsive protein complex’ that define protein complexes responding to a dark or light radiation regimen respectively. The terms are defined with the same procedure as for the ‘radiation responsive protein complex’. The term ‘light responsive protein complex’ is defined as: “Any macromolecular complex that is composed of two or more polypeptide subunits, which may or may not be identical, that assembles and carries out some function in response to an exposure to day light as the light of the sun”. Whereas, ‘dark (no light) responsive protein complex’ is defined as: “Any macromolecular complex that is composed of two or more polypeptide subunits, which may or may not be identical, that assembles and carries out some function in response to an exposure to darkness (no light)”.

The light responsive protein complex class has four sub-classes representing four common visible light spectrum wavelengths to which plants have been shown to respond and that includes: ‘blue-light responsive protein complex’, ‘far-red light responsive protein complex’, ‘red-light responsive protein complex’, and ‘red and far-red light responsive protein complex’. Each class is defined in the same manner as described above. The term ‘blue-light responsive protein complex’ is defined as: “Any macromolecular complex that is composed of two or more polypeptide subunits, which may or may not be identical, that assembles and carries out some function in response to an exposure to blue light in the wavelength range of 455-492 nM”. The term ‘far-red light responsive protein complex’ is defined as: “Any macromolecular complex that is composed of two or more polypeptide subunits, which may or may not be identical, that assembles and carries out some function in response to an exposure to far-red light in the wavelength range of 700-800 nM”. The term ‘red-light responsive protein complex’ is defined as: “Any macromolecular complex that is composed of two or more polypeptide subunits, which may or may not be identical, that assembles and carries out some function in response to an exposure to red-light in the wavelength range of 622-780 nM”. Finally, the term ‘ red and far-red light responsive protein complex is defined as: “Any macromolecular complex that is composed of two or more polypeptide subunits, which may or may not be identical, that assembles and carries out some function in response to an exposure to red or far-red light in the wavelength range of 622-800 nM”.

### *3.3.2 Cellular Compartments*

In order to define the formation and function of a protein complex that is dependent upon location to a particular cellular compartment, the term name is defined at a class level by including the particular cellular compartment in question. The particular term is derived in a similar fashion as formula 3.1, but would include cellular compartment information as follows in formula 3.2 shown below.

Formula 3.2:

[(instance no.)(cellular compartment)(term name)('P' {pathway instance no.}: step no. /total no. steps)]

For example, taking the previous example derived from formula 3.1 regarding the emergence of the protein complex “2-'protein complex AB' (P4:2/6)”, and including nuclear cellular compartment information, the term would now be defined as “2-nuclear-'protein complex AB' (P4:2/6)”.

## *3.4 Conceptualizing Affinity Relationships: Potential Attractions Between Proteins and Protein Complexes*

In biology and biochemistry, affinity relationships that describe the stability of protein-ligand interactions can be formally defined through the dissociation constant ‘Kd’. The Kd constant is defined as follows:

$$Kd = \frac{[Pf][Lf]}{[PL]}$$

where ‘Pf’ and ‘Lf’ represent the concentration of free protein and ligand respectively, and PL represents bound protein-ligand (Phizicky and Fields, 1995; Nelson and Cox, 2004). Kd represents the molar concentration of ligand where half of the available binding sites are occupied by the interaction partner or protein. A lower value of Kd represents a more stable protein-ligand interaction since it implies that a lower concentration of ligand is required for half of the binding sites to be occupied. To date, the strongest known interaction documented is that of biotin and avidin (egg white) with a Kd value of about $10^{-15}$M (Nelson and Cox, 2004).

As previously pointed out, protein interaction relationships are defined via the relation ‘has affinity for’. This relation can define both a protein-protein interaction as well as a protein-protein complex relationship. With respect to the relation defining an affinity relationship defined by the Kd constant, this relation assumes the conditions for the interaction of proteins has been met. The relation ‘has affinity for’ is not transitive and is formally defined as follows: “A relationship between two entities either two continuant entities or a continuant and occurrent entity, in which an affinity relationship has been defined between the two entities through experimental evidence, and may or may not be represented by an experimentally defined Kd value” (see also appendix B). The reason for defining the relation in this manner is because, in some cases, experimental results in literature may state affinity relationships but may or may not include a quantitative Kd value.

Moreover, if a Kd value has been experimentally derived for a particular protein complex and is available from the literature, then the protein complex instance can be defined to include the Kd value to properly define the emergence of the particular instance of the protein complex. The term name is derived in the same manner as the previous derivations defined by Formula 3.1 and 3.2, however would include the Kd as follows in Formula 3.3 shown below.

Formula 3.3:

[(instance no.)(term name)(Kd:[M])('P' {pathway instance no.}: step no. /total no. steps)]

The Kd is expressed a as molar concentration ‘M’ and will represent the most recent protein interaction taking place. For example, if the complex is binary then the Kd would imply the value for the binary interaction. However, if the complex contains three subunits, the Kd value would represent the last subunit interaction with the binary complex represented. Taking a ‘protein complex AB’ for example, involving the interaction of a third protein ‘C’ with the complex, the Kd can be defined as ‘Kd (ab-c)’ where the last joining subunit to the protein complex would be appended by a hyphen. As previously stated, when formally annotating a protein complex, the Kd value will not be included. It should be stated that there are several reasons behind including Kd as a property at a class level. First, since currently there is only one available slot for defining cardinal values in OBO-Edit, where this slot is currently being used to define the number of subunits part of a protein complex. Second, it would make the Kd value easily accessible to an end user at a class level rather than being part of the definition of the term.

An example term derivation would be as follows: taking the previous example derived from formula 3.2 concerned with the emergence of a nuclear protein complex “2-nuclear-'protein complex AB' (P4:2/6)”, if Kd information was available for the interaction of protein subunits A and B, and using a hypothetical Kd value of $10^{-8}$M, then the term would be defined as “2-nuclear-'protein complex AB' (Kd:$10^{-8}$M) (P4:2/6)”. When annotating this particular protein interaction and subsequent complex, it would be understood that this was a formally derived term with evidence of the complex forming with a specific affinity and resident in the nucleus. It is apparent that this would be the second particular instance of the protein complex by the instance value ‘2’, so that more information about the complex in a different context could be investigated. Furthermore, pathway-related information would be made available so that one could follow, leading to an awareness of yet more information relating to the complex.  Again, this protein complex would only be annotated as ‘nuclear protein-complex AB’ while not containing any property information in the title.

### *3.5 Analyzing the Ontological Graph: Node distribution and Metrics*

Differentiating classes from instances or individuals in an OBO-Edit generated ontology is currently not possible. The OBO-Edit author, John Day-Richter validates this assertion in the OBO-Edit user-guide (see “An Introduction to OBO Ontologies”; http://oboedit.org/docs/index.html , accessed Aug, 18 2012). In this case, an instance of a particular class is still represented as a class instead of an individual of the class. This is evident by reviewing particular sub-classes represented in OBO-Edit that are in fact an instantiation of a major parent class, for example, taking a SCF ubiquitin ligase complex class as a use-case example.  An instantiation of this class would in theory be a specific SCF ubiquitin ligase complex with specific subunits, for example, an Arabidopsis SCF$^{(TIR1)}$. It is not clearly represented in OBO that the Arabidopsis SCF$^{(TIR1)}$ is in fact an instance of SCF ubiquitin ligase, but rather, it is still represented as a sub-class. Thus, when analyzing the ontological graph with respect to the PRO expansion, instances will still be represented as classes.

One aspect of the PRO expansion as part of this thesis work includes representing instances of classes by using a numerical value as a prefix to the term. This will aid in differentiating a class from an instance despite the lack of formal representation. Moreover, with the addition of temporality, cellular compartment, affinity, and pathway information to define protein forms, marking instances is critical to discern the many instantiations or forms that a particular protein complex can have. For example, a hypothetical ‘protein complex AB’ may be represented as being nuclear and with the same Kd value representing its affinity in two different cases, however, it may possibly be defined as participating in two different pathways, and therefore it should be represented by two different instances. This could be realized in the case of a Arabidopsis $SCF^{(TIR1)}$ complex that targets IAA proteins for degradation in response to auxin. Since auxin is released in response to many growth and developmental cues, the $SCF^{(TIR1)}$ complex effectively would also respond to auxin in these different cases. Currently, the PRO expansion is not equipped to formally define the particular pathway instance outside of the arbitrary pathway characterization (‘P’; cf. section 3.2); however, this is something that could be expanded upon in the future.

Consequently, current metrics of the PRO representing the number of classes and specific complex type classes, as of the latest PRO release version 28, are as indicated in Table 3.2. These metrics were taken from both BioPortal under the respective PRO ontology search as well as the PRO website downloads section (see the release note that pertains to the latest PRO obo file, ‘pro_release_note.txt’). Without taking into account potential protein instance additions and organism specific classes, the current PRO expansion will have contributed 11 major parent classes or nodes, in which, 4 nodes would be top-level classes and 7 would contribute to the ‘complex-category’. As well, the PRO expansion has contributed two new relations to represent affinity and abundance relationships. It is not possible to calculate the number of potential instances in each category; however, regarding the inclusion of pathway, affinity, abundance, and compartment properties, the hierarchy, when fully annotated, would be significantly expanded with respect to defining both proteins and protein-complexes.

Table 3.2 PRO metrics pertaining to the current PRO and the PRO after its expansion as part of this thesis work. The number of total classes and the number of specific complex category classes are represented.

| Current and Expanded PRO Metrics | | |
|---|---|---|
| | **Current PRO** | **Expanded PRO** |
| **Number of Classes** | 29005 | 29016 |
| **Number of Instances** | 0 | n/a |
| **Complex Category Classes** | 126 | 133 |

## *Chapter 4. Validating the Model*

### *4.1 OBO-Edit Search and Link Search Panel*

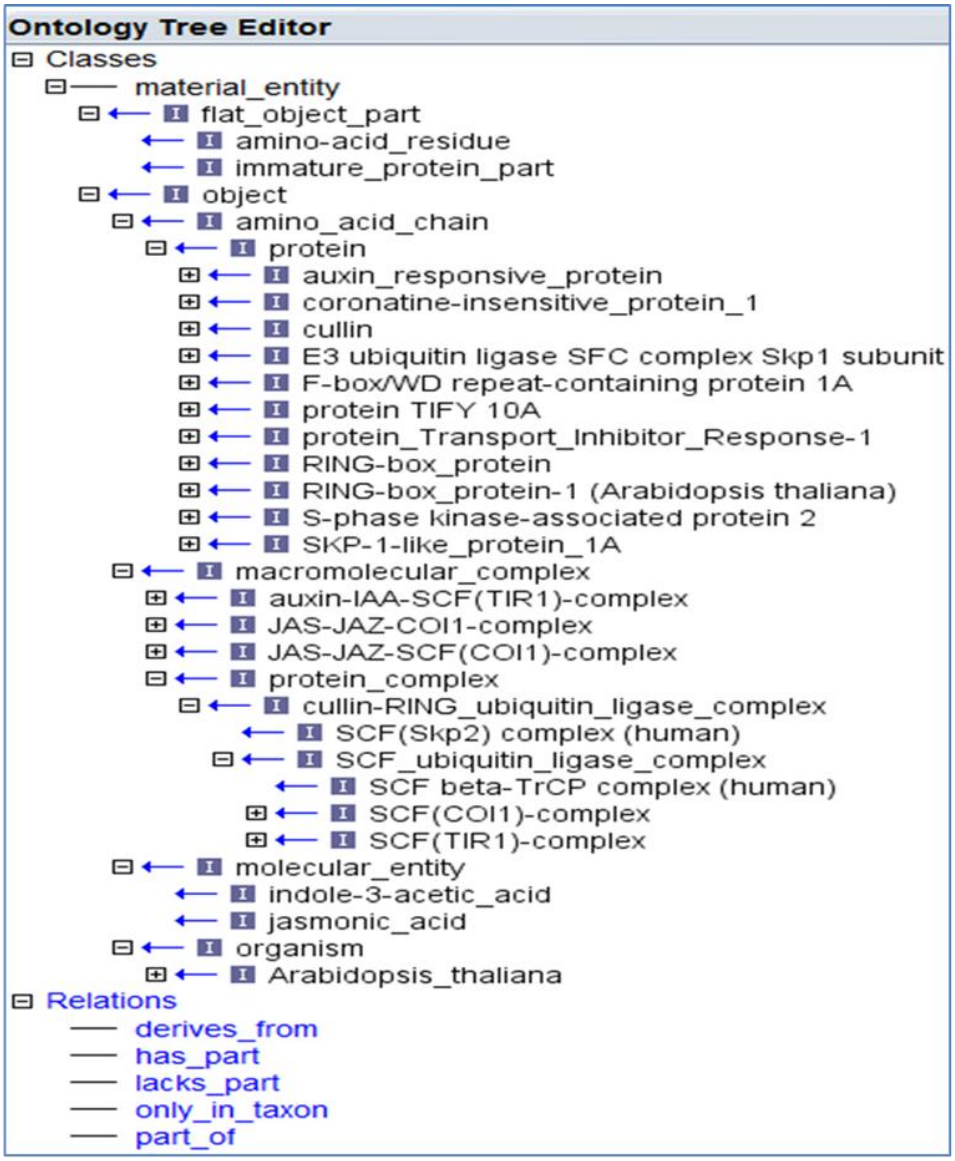
OBO-Edit has a search and link search panel that allows one to search an ontology for terms and relationships. The OBO-Edit search functions were used in this thesis work to validate the PRO expansion in a number of ways; for example, to test the ontological model for coherency; in verifying relationship links; and, in simulating PRO database queries. The OBO-Edit search function offers a number of drop-down menus that allows an end-user to generate general to very specific searches including, for example, searching by name, ID, definition, or any text field. Moreover, a user can generate compound searches using the ‘matches all’ or ‘matches any’ function that imposes the standard Boolean search types. The OBO-Edit link search panel adds another layer of search functionality in that it enables one to search by relationship (relation) criteria (between terms) rather than by only the terms. As previously indicated, the OBO-Edit search panels are very similar to the PRO database search function except that the later OBO-Edit search type regarding relationship links cannot currently be executed in the PRO database.

### *4.2 Simulated PRO Database Queries with OBO-Edit Searches*

In order to test the validity of using OBO-Edit for the PRO expansion, a simulated PRO ontology and database was generated in a local implementation of OBO-Edit. 65 terms and definitions were manually generated that included specific classes pertaining to SCF ubiquitin ligases, as well as all top-level categories. All PRO relations were also generated. Figure 4.1 displays a snapshot of the simulated PRO ontology created in OBO-Edit. SCF ubiquitin ligase-related searches were generated in OBO-Edit as well as in the live PRO database and query outputs were compared. Figure 4.2 and 4.3 show the results of two separate queries that display matching outputs. The first query shown in Figure 4.2 and involving ‘SCF and Arabidopsis’ as search terms, implements an ‘AND’ Boolean search, while the second query shown in Figure 4.3 is a single query pertaining to the $SCF^{COI1}$ complex. Similarly, Figure 4.4 shows the comparison of the definition for $SCF^{COI1}$ that was derived in OBO-Edit (panel A) and compared to that of the PRO database in panel B.

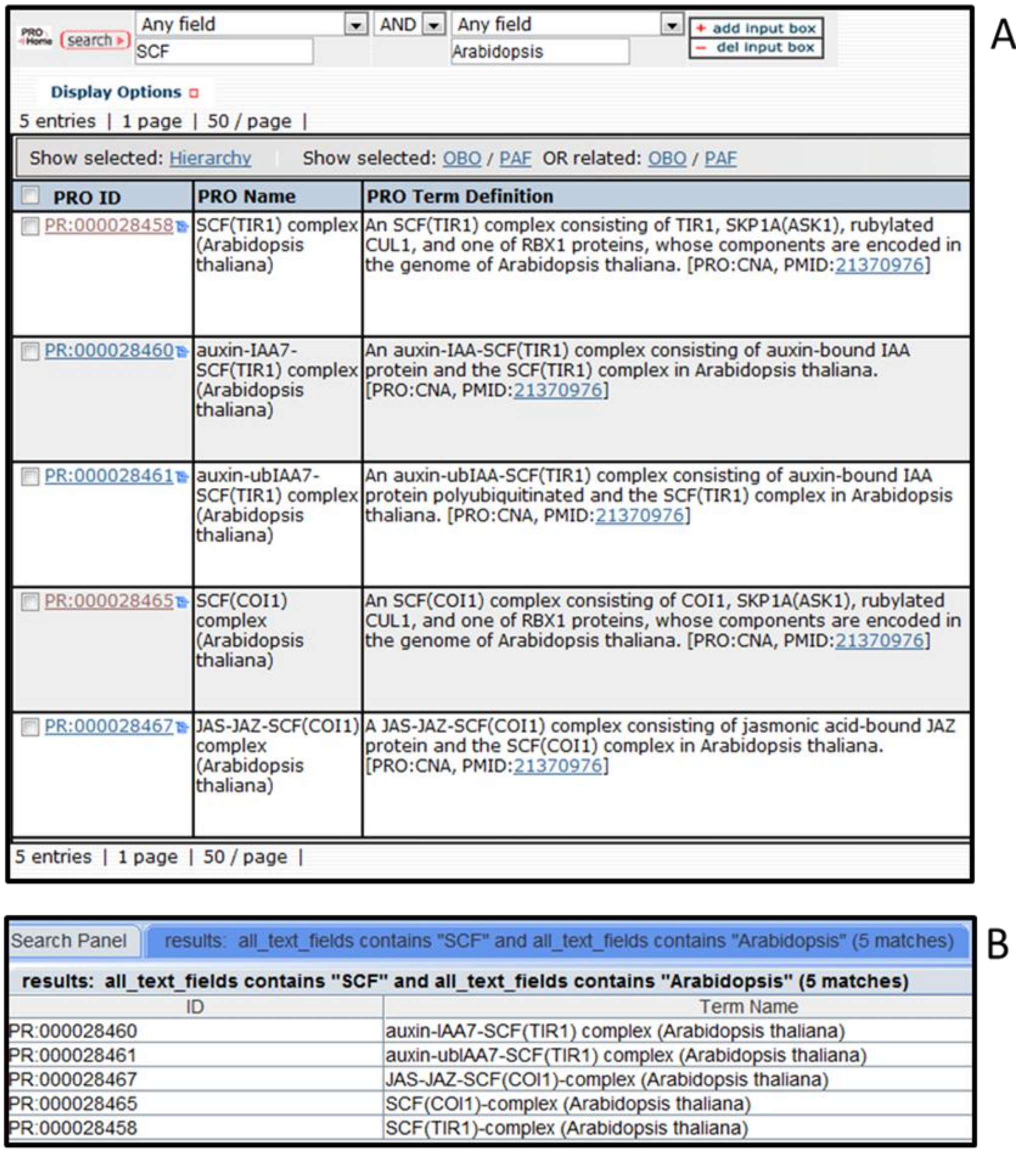
### *4.3 New Model Queries*

With respect to new classes and concepts added to the PRO as part of its expansion through this thesis work, query options now include (but are not limited to) pathway attributes, light responsive protein complexes, cellular compartment specific protein complexes and relation-dependent queries in the link search panel. Figure 4.5 illustrates the OBO-Edit search result of a query pertaining to a hypothetical pathway named ‘P1’ and displays all objects that pertain to P1 (note that this is a hypothetical pathway, where no cellular compartment or Kd is defined, and involves a hypothetical organism instance ‘organism-A’). The output of this query provides the end-user information pertaining to the sequential formation of the protein complex part of the pathway ‘protein complex ABC’ involving three protein subunits and a small molecule effector. The pathway properties of each object part of ‘P1’ would subsequently alert an



```
Ontology Tree Editor
⊟ Classes
  ⊟ material_entity
    ⊟ flat_object_part
        amino-acid_residue
        immature_protein_part
    ⊟ object
      ⊟ amino_acid_chain
        ⊟ protein
          ⊞ auxin_responsive_protein
          ⊞ coronatine-insensitive_protein_1
          ⊞ cullin
          ⊞ E3 ubiquitin ligase SFC complex Skp1 subunit
          ⊞ F-box/WD repeat-containing protein 1A
          ⊞ protein TIFY 10A
          ⊞ protein_Transport_Inhibitor_Response-1
          ⊞ RING-box_protein
          ⊞ RING-box_protein-1 (Arabidopsis thaliana)
          ⊞ S-phase kinase-associated protein 2
          ⊞ SKP-1-like_protein_1A
      ⊟ macromolecular_complex
        ⊞ auxin-IAA-SCF(TIR1)-complex
        ⊞ JAS-JAZ-COI1-complex
        ⊞ JAS-JAZ-SCF(COI1)-complex
        ⊟ protein_complex
          ⊟ cullin-RING_ubiquitin_ligase_complex
              SCF(Skp2) complex (human)
            ⊟ SCF_ubiquitin_ligase_complex
                SCF beta-TrCP complex (human)
              ⊞ SCF(COI1)-complex
              ⊞ SCF(TIR1)-complex
      ⊟ molecular_entity
          indole-3-acetic_acid
          jasmonic_acid
      ⊟ organism
        ⊞ Arabidopsis_thaliana
⊟ Relations
    derives_from
    has_part
    lacks_part
    only_in_taxon
    part_of
```

Fig 4.1 Snapshot of the simulated PRO ontology created in OBO-Edit.



Any field: SCF AND Any field: Arabidopsis

5 entries | 1 page | 50 / page |

Show selected: Hierarchy Show selected: OBO / PAF OR related: OBO / PAF

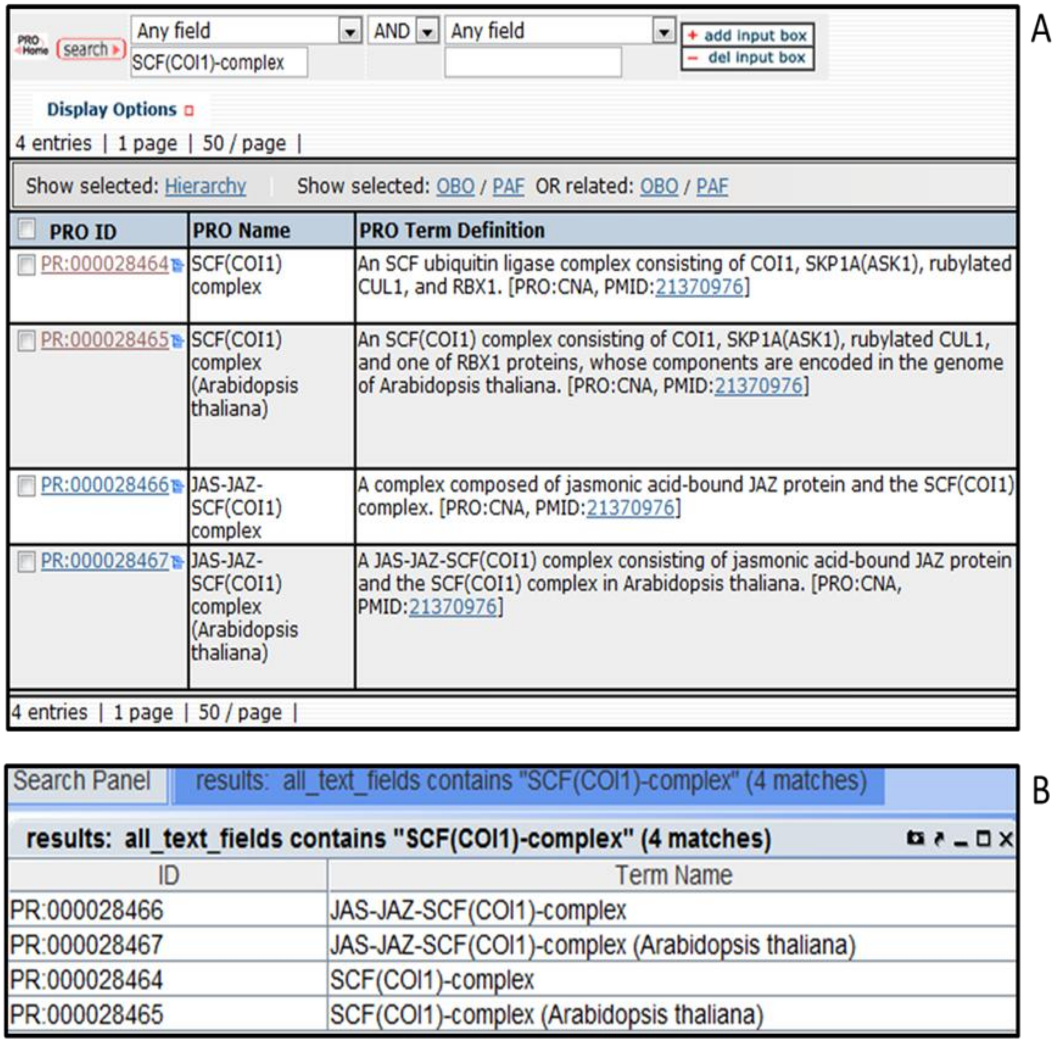
| PRO ID | PRO Name | PRO Term Definition |
|---|---|---|
| PR:000028458 | SCF(TIR1) complex (Arabidopsis thaliana) | An SCF(TIR1) complex consisting of TIR1, SKP1A(ASK1), rubylated CUL1, and one of RBX1 proteins, whose components are encoded in the genome of Arabidopsis thaliana. [PRO:CNA, PMID:21370976] |
| PR:000028460 | auxin-IAA7-SCF(TIR1) complex (Arabidopsis thaliana) | An auxin-IAA-SCF(TIR1) complex consisting of auxin-bound IAA protein and the SCF(TIR1) complex in Arabidopsis thaliana. [PRO:CNA, PMID:21370976] |
| PR:000028461 | auxin-ubIAA7-SCF(TIR1) complex (Arabidopsis thaliana) | An auxin-ubIAA-SCF(TIR1) complex consisting of auxin-bound IAA protein polyubiquitinated and the SCF(TIR1) complex in Arabidopsis thaliana. [PRO:CNA, PMID:21370976] |
| PR:000028465 | SCF(COI1) complex (Arabidopsis thaliana) | An SCF(COI1) complex consisting of COI1, SKP1A(ASK1), rubylated CUL1, and one of RBX1 proteins, whose components are encoded in the genome of Arabidopsis thaliana. [PRO:CNA, PMID:21370976] |
| PR:000028467 | JAS-JAZ-SCF(COI1) complex (Arabidopsis thaliana) | A JAS-JAZ-SCF(COI1) complex consisting of jasmonic acid-bound JAZ protein and the SCF(COI1) complex in Arabidopsis thaliana. [PRO:CNA, PMID:21370976] |

5 entries | 1 page | 50 / page |

results: all_text_fields contains "SCF" and all_text_fields contains "Arabidopsis" (5 matches)

| ID | Term Name |
|---|---|
| PR:000028460 | auxin-IAA7-SCF(TIR1) complex (Arabidopsis thaliana) |
| PR:000028461 | auxin-ubIAA7-SCF(TIR1) complex (Arabidopsis thaliana) |
| PR:000028467 | JAS-JAZ-SCF(COI1)-complex (Arabidopsis thaliana) |
| PR:000028465 | SCF(COI1)-complex (Arabidopsis thaliana) |
| PR:000028458 | SCF(TIR1)-complex (Arabidopsis thaliana) |

Fig 4.2 Snapshots of the matching data output of the query 'SCF and 'Arabidopsis' in (A) the live PRO database and (B) the simulated PRO database generated in
OBO-Edit



Any field: SCF(COI1)-complex

4 entries | 1 page | 50 / page |

Show selected: Hierarchy   Show selected: OBO / PAF OR related: OBO / PAF

| □ PRO ID | PRO Name | PRO Term Definition |
|---|---|---|
| □ PR:000028464 | SCF(COI1) complex | An SCF ubiquitin ligase complex consisting of COI1, SKP1A(ASK1), rubylated CUL1, and RBX1. [PRO:CNA, PMID:21370976] |
| □ PR:000028465 | SCF(COI1) complex (Arabidopsis thaliana) | An SCF(COI1) complex consisting of COI1, SKP1A(ASK1), rubylated CUL1, and one of RBX1 proteins, whose components are encoded in the genome of Arabidopsis thaliana. [PRO:CNA, PMID:21370976] |
| □ PR:000028466 | JAS-JAZ-SCF(COI1) complex | A complex composed of jasmonic acid-bound JAZ protein and the SCF(COI1) complex. [PRO:CNA, PMID:21370976] |
| □ PR:000028467 | JAS-JAZ-SCF(COI1) complex (Arabidopsis thaliana) | A JAS-JAZ-SCF(COI1) complex consisting of jasmonic acid-bound JAZ protein and the SCF(COI1) complex in Arabidopsis thaliana. [PRO:CNA, PMID:21370976] |

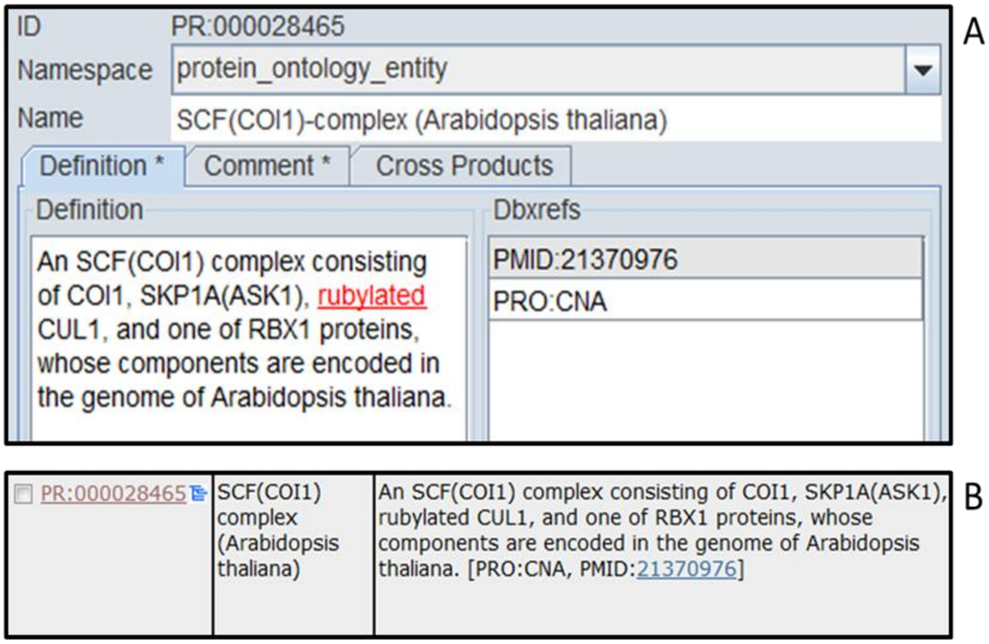
4 entries | 1 page | 50 / page |

results: all_text_fields contains "SCF(COI1)-complex" (4 matches)

results: all_text_fields contains "SCF(COI1)-complex" (4 matches)

| ID | Term Name |
|---|---|
| PR:000028466 | JAS-JAZ-SCF(COI1)-complex |
| PR:000028467 | JAS-JAZ-SCF(COI1)-complex (Arabidopsis thaliana) |
| PR:000028464 | SCF(COI1)-complex |
| PR:000028465 | SCF(COI1)-complex (Arabidopsis thaliana) |

Fig 4.3 Snapshots of the matching data output of the query ‘SCF(COI1)-complex’ in (A) the live PRO database and (B) the simulated PRO database generated in OBO-Edit



ID PR:000028465

Namespace protein_ontology_entity

Name SCF(COI1)-complex (Arabidopsis thaliana)

Definition

An SCF(COI1) complex consisting of COI1, SKP1A(ASK1), rubylated CUL1, and one of RBX1 proteins, whose components are encoded in the genome of Arabidopsis thaliana.

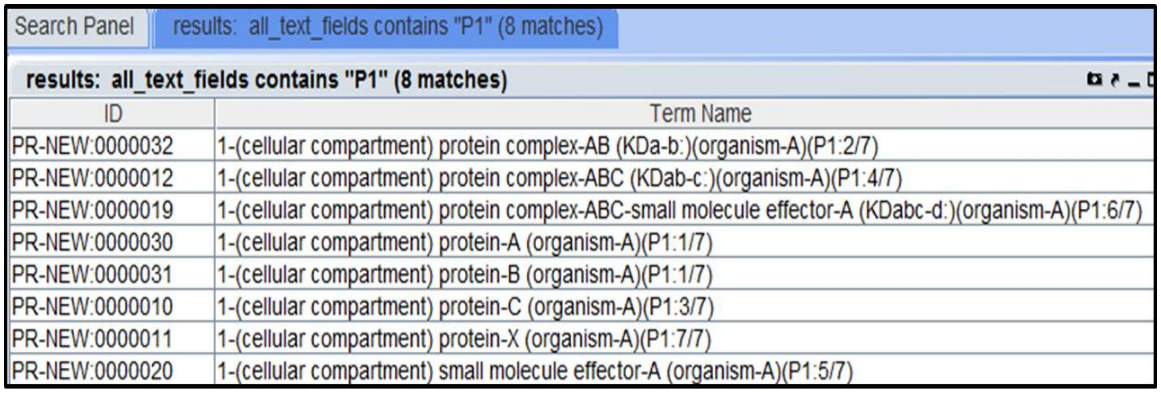
Dbxrefs

PMID:21370976

PRO:CNA

| | | |
|---|---|---|
| ☐ PR:000028465 | SCF(COI1) complex (Arabidopsis thaliana) | An SCF(COI1) complex consisting of COI1, SKP1A(ASK1), rubylated CUL1, and one of RBX1 proteins, whose components are encoded in the genome of Arabidopsis thaliana. [PRO:CNA, PMID:21370976] |

Fig 4.4 Snapshots of the matching data and definition output of the query ‘SCF(COI1)-complex (Arabidopsis thaliana)’ in (A) the simulated PRO database generated in OBO-Edit and (B) the live PRO database.



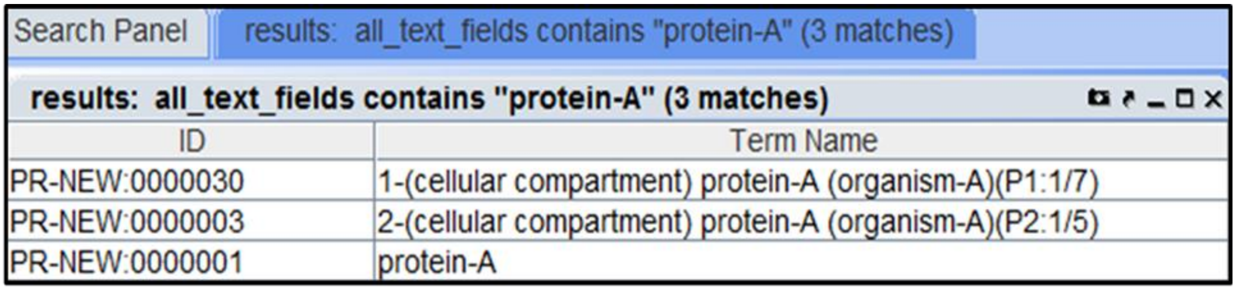
results: all_text_fields contains "P1" (8 matches)

| ID | Term Name |
|---|---|
| PR-NEW:0000032 | 1-(cellular compartment) protein complex-AB (KDa-b:)(organism-A)(P1:2/7) |
| PR-NEW:0000012 | 1-(cellular compartment) protein complex-ABC (KDab-c:)(organism-A)(P1:4/7) |
| PR-NEW:0000019 | 1-(cellular compartment) protein complex-ABC-small molecule effector-A (KDabc-d:)(organism-A)(P1:6/7) |
| PR-NEW:0000030 | 1-(cellular compartment) protein-A (organism-A)(P1:1/7) |
| PR-NEW:0000031 | 1-(cellular compartment) protein-B (organism-A)(P1:1/7) |
| PR-NEW:0000010 | 1-(cellular compartment) protein-C (organism-A)(P1:3/7) |
| PR-NEW:0000011 | 1-(cellular compartment) protein-X (organism-A)(P1:7/7) |
| PR-NEW:0000020 | 1-(cellular compartment) small molecule effector-A (organism-A)(P1:5/7) |

Fig 4.5 Snapshot of the output produced from the query 'P1' generated in the OBO-Edit search panel.

end-user to the number of steps involved in the formation and function of the protein complex. Each object that is part of the general ‘P1’ query output can further be searched to obtain additional information. For example, searching ‘protein-A’ specifically would retrieve information pertaining to the involvement of this hypothetical protein in other pathways and complexes, thus informing an end-user about the protein and related protein complexes (see Figure 4.6).

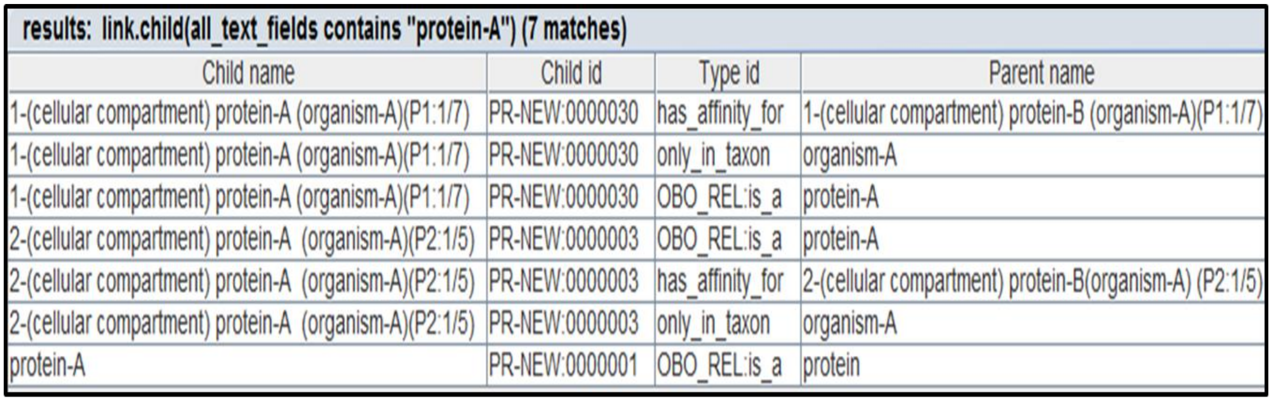
In order to acquire relationship information between objects, one can use the link search panel. Although this type of query is not currently available in the PRO database, it is under review for inclusion as a potentially useful search function for an end-user. For example, one could assess the output of a general query like the ‘P1’ query stated above, and search a specific object from the output like ‘protein-A’ to gather further information regarding its relationship between other proteins and protein-complexes (see Figure 4.7). The particular example shown in Figure 4.7 displays relationships involving the new relation ‘has affinity for’ together with the general usage of the ‘is a’ and ‘only in taxon’ relations. Another example involving the newly developed radiation dependent protein complex classes is shown in Figure 4.8. Here, a compound link-search using the terms ‘dark (no light)’ and ‘protein complex’. The output from this search includes the protein subunits that are part of the dark (no light) responsive protein complex via the relation ‘has part’, together with affinity relationships as well as information about the function of the complex regulating the abundance of ‘protein X’ via the relation ‘negatively regulates abundance’.



results: all_text_fields contains "protein-A" (3 matches)

| ID | Term Name |
|---|---|
| PR-NEW:0000030 | 1-(cellular compartment) protein-A (organism-A)(P1:1/7) |
| PR-NEW:0000003 | 2-(cellular compartment) protein-A (organism-A)(P2:1/5) |
| PR-NEW:0000001 | protein-A |

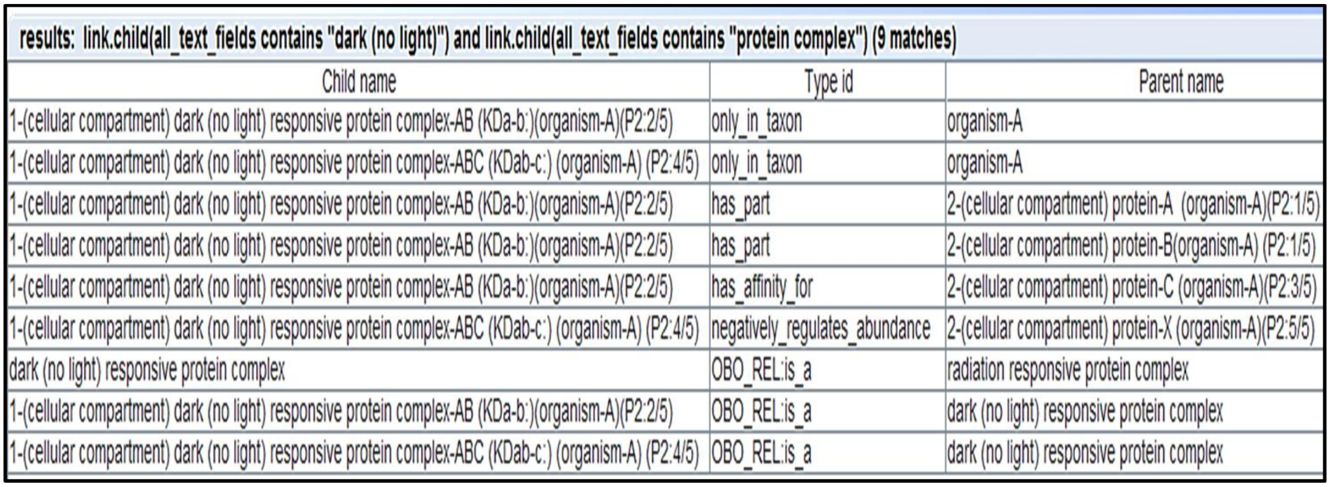
Fig 4.6 Snapshot of the output produced from the query 'protein-A' generated in the OBO-Edit search panel



**results: link.child(all_text_fields contains "protein-A") (7 matches)**

| Child name | Child id | Type id | Parent name |
|---|---|---|---|
| 1-(cellular compartment) protein-A (organism-A)(P1:1/7) | PR-NEW:0000030 | has_affinity_for | 1-(cellular compartment) protein-B (organism-A)(P1:1/7) |
| 1-(cellular compartment) protein-A (organism-A)(P1:1/7) | PR-NEW:0000030 | only_in_taxon | organism-A |
| 1-(cellular compartment) protein-A (organism-A)(P1:1/7) | PR-NEW:0000030 | OBO_REL:is_a | protein-A |
| 2-(cellular compartment) protein-A (organism-A)(P2:1/5) | PR-NEW:0000003 | OBO_REL:is_a | protein-A |
| 2-(cellular compartment) protein-A (organism-A)(P2:1/5) | PR-NEW:0000003 | has_affinity_for | 2-(cellular compartment) protein-B(organism-A) (P2:1/5) |
| 2-(cellular compartment) protein-A (organism-A)(P2:1/5) | PR-NEW:0000003 | only_in_taxon | organism-A |
| protein-A | PR-NEW:0000001 | OBO_REL:is_a | protein |

Fig 4.7 Snapshot of the output produced from the query 'protein-A' generated in the OBO-Edit link search panel. The type id column displays the links or relationships between the child and parent terms.



results: link.child(all_text_fields contains "dark (no light)") and link.child(all_text_fields contains "protein complex") (9 matches)

| Child name | Type id | Parent name |
|---|---|---|
| 1-(cellular compartment) dark (no light) responsive protein complex-AB (KDa-b:)(organism-A)(P2:2/5) | only_in_taxon | organism-A |
| 1-(cellular compartment) dark (no light) responsive protein complex-ABC (KDab-c:) (organism-A) (P2:4/5) | only_in_taxon | organism-A |
| 1-(cellular compartment) dark (no light) responsive protein complex-AB (KDa-b:)(organism-A)(P2:2/5) | has_part | 2-(cellular compartment) protein-A (organism-A)(P2:1/5) |
| 1-(cellular compartment) dark (no light) responsive protein complex-AB (KDa-b:)(organism-A)(P2:2/5) | has_part | 2-(cellular compartment) protein-B(organism-A) (P2:1/5) |
| 1-(cellular compartment) dark (no light) responsive protein complex-AB (KDa-b:)(organism-A)(P2:2/5) | has_affinity_for | 2-(cellular compartment) protein-C (organism-A)(P2:3/5) |
| 1-(cellular compartment) dark (no light) responsive protein complex-ABC (KDab-c:) (organism-A) (P2:4/5) | negatively_regulates_abundance | 2-(cellular compartment) protein-X (organism-A)(P2:5/5) |
| dark (no light) responsive protein complex | OBO_REL:is_a | radiation responsive protein complex |
| 1-(cellular compartment) dark (no light) responsive protein complex-AB (KDa-b:)(organism-A)(P2:2/5) | OBO_REL:is_a | dark (no light) responsive protein complex |
| 1-(cellular compartment) dark (no light) responsive protein complex-ABC (KDab-c:) (organism-A) (P2:4/5) | OBO_REL:is_a | dark (no light) responsive protein complex |

Fig 4.8 Snapshot of the output produced from the query ‘dark (no light)’ and ‘protein complex’ generated in the OBO-Edit link search panel. The type id column displays the links or relationships between the child and parent terms.

### *4.3.1 Use-case Scenarios: Arabidopsis SCF Complex Queries*

The above examples can be used to better represent what is known about the structure and function of an Arabidopsis SCF ubiquitin ligase. In this section, I will further demonstrate the search capabilities by using real Arabidopsis SCF ubiquitin ligase. Taking for example the case of Arabidopsis SCF$^{TIR1}$ that includes four main subunits: ASK1, RBX1, CUL1, and TIR1; while the order of subunit assembly is not understood, it is known that the small-molecule effector (phytohormone) auxin binds TIR1 as part of the complex and activates the complex for the targed ubiquitinylation and degradation of IAA proteins via the 26S proteasome. Since the order of subunit assembly to the complex is not understood, this example will demonstrate the complex as a whole, and include only the affinity relationship of auxin for TIR. In this case, the interaction of auxin for the SCF$^{TIR1}$ complex is informed by a scatchard analysis that determined the Kd of TIR1 for auxin to be 84nM (8.4x10$^{-8}$M)  although the experimental data was not shown (Dharmasiri and Estelle, 2005) . It is known from our work (M. Dezfulian, personal communication) and that of others that the SCF$^{(TIR1)}$ complex functions in the nucleus (Calderon-villalobos, L.I., et al., 2010), which defines the sub-cellular compartment for the illustration of this example. Taking each subunit of the complex as being the only instance, and since the order of subunit assembly is not understood, each subunit would be equivalently defined as step one of the pathway. However, as experimental evidence becomes available, this could be readily changed to reflect the actual order that the subunits assemble. Figure 4.9 displays the manner in which each term is defined as well as the order in which the process takes place. Not included in the image (but that would still apply) is the fact that the relation ‘has part’ would define the

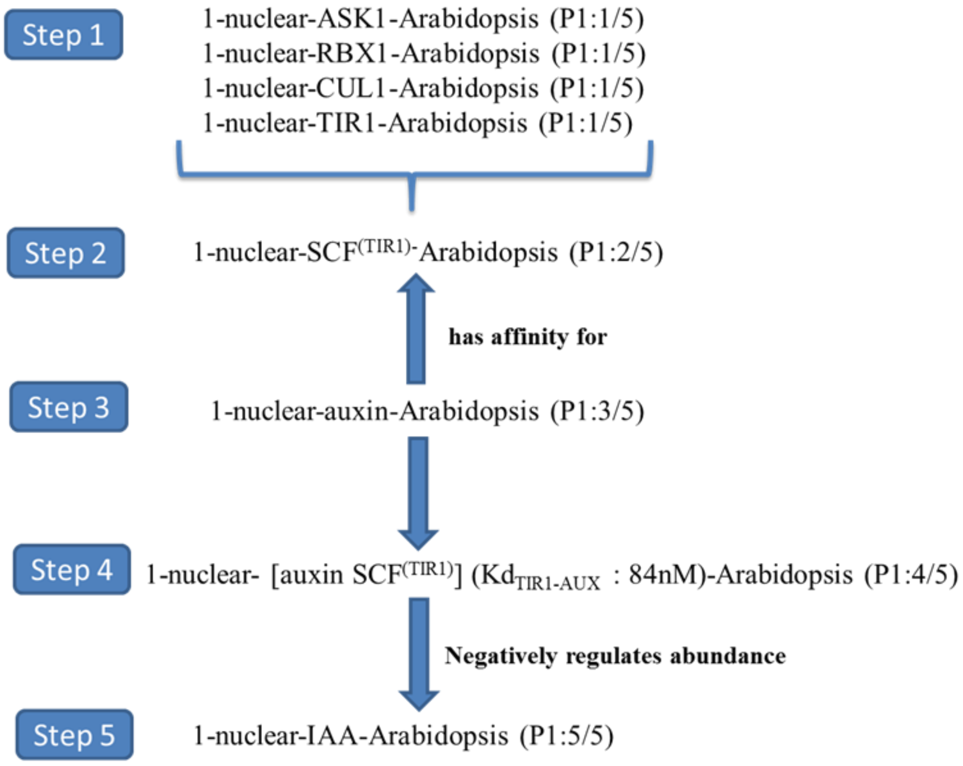

Step 1
1-nuclear-ASK1-Arabidopsis (P1:1/5)
1-nuclear-RBX1-Arabidopsis (P1:1/5)
1-nuclear-CUL1-Arabidopsis (P1:1/5)
1-nuclear-TIR1-Arabidopsis (P1:1/5)
Step 2
1-nuclear-SCF$^{(TIR1)-}$Arabidopsis (P1:2/5)
has affinity for
Step 3
1-nuclear-auxin-Arabidopsis (P1:3/5)
Step 4
1-nuclear- [auxin SCF$^{(TIR1)}$] (Kd$_{TIR1-AUX}$ : 84nM)-Arabidopsis (P1:4/5)
Negatively regulates abundance
Step 5
1-nuclear-IAA-Arabidopsis (P1:5/5)

Fig 4.9 A flow diagram displaying the manner in which the structure and function of an Arabidopsis SCF$^{(TIR1)}$ would be defined with respect to the PRO expansion.

relationship of each protein subunit as being part of the larger complex, as well as the inclusion of the relation ‘only in taxon’ to demonstrate that these are Arabidopsis-specific proteins and complexes. Figure 4.9 shows the initiation of the assembly of all four main subunits that takes place as the first step, and each subunit is defined as such by the pathway attribute: ‘P1:1/5’. The next step, ‘P1:2/5’ involves the actual formation of the complex. The third step (P1:3/5) demonstrates the affinity of auxin for the subunit TIR1 as part of the complex by the relation ‘has affinity for’. The fourth step (P1:4/5) shows the actual association of auxin with the complex with an experimentally defined Kd of 84nM and, finally, the last step (P1:5/5) indicates the targeting of IAA proteins for degradation by the relation ‘negatively regulates abundance’. As in the previous search or link search examples, these terms can be queried in a similar manner, except that each generic protein form defined in the above examples would be replaced by an actual protein subunit (e.g. ASK1).

Another Arabidopsis-specific example involving a variable light regimen is found with the $SCF^{ZTL}$ protein complex, that has been demonstrated to respond to darkness and target TOC1 proteins for ubiquitinylation and subsequent degradation. Figure 4.10 demonstrates the same procedure for defining the structure and function of this complex. However, unlike the example involving $SCF^{TIR1}$, there is no small molecular effector involved and no experimental evidence regarding the cellular compartment in which the complex functions, although it has been shown that TOC1 is a nuclear protein (Strayer, C. et al., 2000). In the face of this paucity of evidence, the TOC1 protein will be defined as residing in the nucleus. The order of subunit assembly for the $SCF^{ZTL}$ complex is similarly not understood. Thus, the pathway has been arbitrarily defined as ‘P2’ to

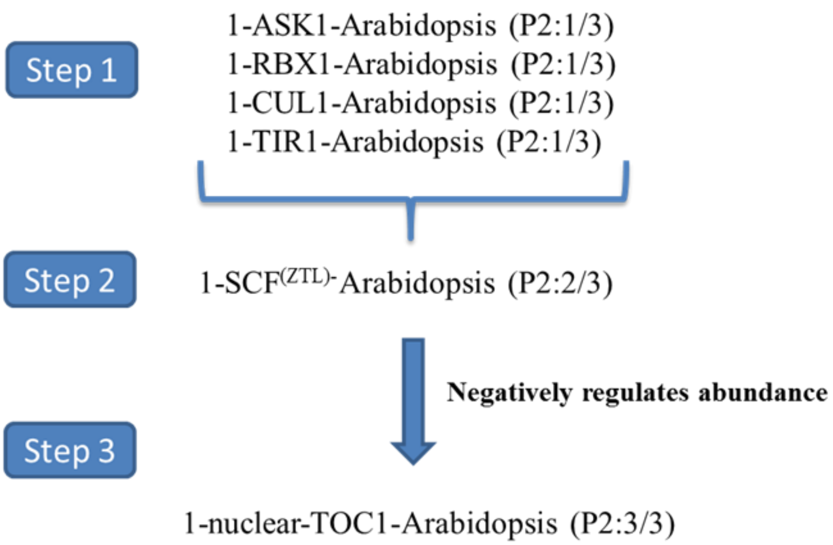

Step 1
1-ASK1-Arabidopsis (P2:1/3)
1-RBX1-Arabidopsis (P2:1/3)
1-CUL1-Arabidopsis (P2:1/3)
1-TIR1-Arabidopsis (P2:1/3)
Step 2
1-SCF$^{(ZTL)-}$Arabidopsis (P2:2/3)
Negatively regulates abundance
Step 3
1-nuclear-TOC1-Arabidopsis (P2:3/3)

Fig 4.10 A flow diagram displaying the manner in which the structure and function of an Arabidopsis SCF$^{(ZTL)}$ would be defined with respect to the PRO expansion.

differentiate it from the previous example. The SCF$^{ZTL}$ complex is defined a sub-class of the dark (no light) responsive protein complex class, and it is therefore implied through transitivity that the complex functions in response to darkness.

Since it is not possible to re-create the entire PRO database and test the functionality of the expanded PRO framework therein, testing and validating the expanded PRO model was undertaken using two approaches. First, a simulated PRO database was created using OBO-Edit and 65 PRO terms were manually created that related to SCF complexes. The terms were then queried within OBO-Edit and within the PRO database where the output of information (corresponding to individual identical queries) was evaluated in order to ensure consistency across the information sets employed. Matching outputs would functionally confirm that OBO-Edit was being used in the same manner as with the PRO development. Secondly, the newly expanded framework was tested for coherency by using the 'search' and 'link-search' panels within OBO-Edit that enable one to search class- and relation-based data, respectively. Specific terms and relations were queried and the output was analyzed to ensure that the correct information was being retrieved. For example, querying 'P1' generated an output that contained all pathway (P1) related components that was manually generated in OBO-Edit. In addition, the class and relation searches confirmed that a researcher can retrieve novel information from an ontology that he or she may not have previously known, thus aiding to the retrieval of related scientific information.

## *Chapter 5. Results and Applications*

### *5.1 Benefits of the New Model*

There are a number of benefits to the new PRO model presented in this thesis. For one, protein forms (proteins and complexes) are now more formally defined with respect to time and space, thus contributing to the reliable discovery of structure and function information that pertains to proteins and protein complexes. The current format of the PRO is geared more towards being a simple controlled vocabulary and not an ontology in its purest sense of fully capturing the underlying biological complexity of, specifically, the conditional formation and function of protein complexes. The current PRO is to be regarded as an important protein information resource and research tool for the biological community, and the PRO consortium has conceded the fact of its development purely as an object-based ontology that does not include temporality. Nevertheless, any forward-looking evolution of biological ontologies will require adding temporality as an important component.

The main idea of adding temporality to the current PRO allows the inclusion of more specific protein and protein complex properties to the current term definitions and relations. This added specificity provides enhanced functionality to scientific researchers in that the information they are obtaining is more reliable, while at the same time providing enriched search capabilities and enhanced information discovery. For example, the new model includes concepts not previously defined in the PRO directly such as interaction affinity, cellular compartmentation, light dependent protein complex formation and function and pathway properties - all of which can be searched within the database with the respective search type. Including affinity relationships in the PRO expansion - both as relations and within class definitions - enables a researcher to make inferences about the proteins involved in protein complex formation, including the sequence of subunit assembly as well as in understanding the conditions under which an interaction will take place. The same thinking applies to including cellular compartmentation and light-dependent protein complex formation. In the same manner, pathway information is extremely useful in understanding how a macromolecular or protein complex forms and functions under certain conditions, although the current PRO does not directly provide this information. The ability to search an enriched set of protein and protein complex properties included in the PRO expansion enables efficient discovery of protein complex-related information, while at the same time providing added mechanisms to make inferences about protein forms within and across species. In the absence of the suggested added functionality within the PRO, the reality and complexity of protein complex formation is not represented, thus relegating the end-user to identify and investigate the specifics via alternative database or literature-based resources.

### *5.2 Application of the New Ontological Model for Experimental Research and Design*

Bio-ontologies are continuing to be constructed, expanded and modified with the goal of providing a framework for annotation of genomic data as well as providing repositories of biological information. Bio-ontologies are still a relatively new concept so that, arguably, they are not being utilized to their full potential as an aid to biomedical research. As ontologies evolve and become more main-stream with regards to biological research, they will prove to be increasingly useful with regards to experimental research and design. For example, instead of needing to father a plethora of papers published on a particular topic, a scientist can simply look up a particular domain-based ontology to retrieve relevant information, thus significantly decreasing the search time and effort otherwise required. Equipped with an expanded ontology, a researcher could search a particular protein subunit within PRO and retrieve important information about the subunit alone and as part of a complex. This includes retrieving information regarding protein orthologues, affinity relationships of the protein (including actual Kd values where available), information pertaining to specific pathways or complexes in which the protein participates as well as whether the protein forms a complex in particular cellular compartments. If the Kd values are available for a particular binary interaction, this will provide a researcher biochemical information that can be utilized in the laboratory; for example, the required steady-state abundance (expressed as concentrations) of two subunits in order for an interaction to take place. Having this type of data readily available may lead a researcher to specific conclusions or inferences about a protein or complex that may accelerate the experimentation or contribute to the design of an experiment. Since ontologies are integrative and interoperable, additional information can be conveniently obtained across databases through mapping and database links. All of these features contribute to effective and efficient research design and implementation.

### *5.3 Interoperability With Existing Ontologies and Databases*

As indicated, PRO 'maps' to various ontologies and databases in order to borrow existing terminology to structure the ontology as well as to annotate protein ontology terms that are part of the ontology. There are five main databases that the PRO uses to annotate terms: 'Pfam'(Bateman et al., 2002), 'GO', the 'Disease Ontology' (DO; http://diseaseontology.sourceforge.nget ; cf. Natale, D., et al 2007) together with 'The Unified Medical Language System' (UMLS; Bodenreider, O., 2004), the Sequence Ontology (SO; cf. Eilbeck, K., et al. 2005) and 'PSI-MOD' (Montecchi-Palazzi L, et. al, 2008). Regarding the Gene Ontology, PRO annotates terms pertaining to the three GO domains by using the relations 'has function', 'participates in', and ‘located in'. PRO also uses the Pfam database to annotate protein domains via the relation 'has part' (not to be confused with the PRO relation 'has part' to link protein subunits to macromolecular complexes), the DO/UMLS databases to annotate disease related terms via the relation 'agent of', the SO database to annotate sequence related changes in proteins via the relations 'has agent' and 'agent of' and, finally, the PSI-MOD database to annotate protein modifications also with the relation 'has part'.

Moreover, when the PRO defines a term within the ontology, it is generally based upon some literature-derived or other direct experimental evidence so that PRO references specific databases within the definition of the term, for example UniProt Kb (Consortium, T.U., 2010) and Reactome (Croft, D. et al., 2011) are used to validate the definition. These databases are used in different ways, where the UniProtKb is specifically referenced for information pertaining to individual proteins while Reactome is used to reference pathway and interaction related information.

The expanded PRO would continue to map and annotate terms in the same manner as before, with some potentially new mapping definitions and new databases to be added. For example, since PRO expansion involves the inclusion of pathway attributes, future work could include mapping to the Reactome database for related pathway information, and vice versa where Reactome could be mapped to the new pathway information contained in the PRO expansion. An additional pathway-based database that could be mapped is the 'KEGG Database' (Kyoto Encyclopedia of Genes and Genomes; cf. Kanehisa, M. and Goto, S.. 2000; http://www.genome.jp/kegg/kegg1.html ), since this database is enriched for information relating to plant metabolism  in comparison to Reactome. Another potential mapping database would be Interactome (Cusick, M. et al. 2005), which is currently being used as a link for protein interaction data by the PIR database specifically (the PIR contains a protein information database that is distinct from the PRO). Since Interactome provides information relating protein interactions based on experimental evidence, this could be used as tool for mapping protein interaction and affinity data. As already pointed out, the PRO annotates cellular component-related terms by mapping to the GO database. Since the PRO expansion includes defining terms with respect to cellular compartmentation, this annotation feature may not be required in the future. These suggested examples constitute suggestions by which the current and future PRO would better integrate and interoperate with existing ontologies and databases.

Regarding model organism databases such as TAIR and Flybase, PRO is not currently being adopted in the model organism database infrastructure to define proteins and protein complexes. However,  like the common implementation of GO in these model organism databases to link genomic related information, the PRO can be just as easily adapted to include data pertaining to proteins and their attendant complexes. In the same manner that the PRO links to GO related terms, and that genomic information within the model organism databases link out to the GO databases, these databases could also include the PRO database and subsequently link to protein and protein complex-related terms. This would all be facilitated through the use of global identifiers (term IDs) that define a particular term. Consequently, ontologies and databases could be orthogonally linked via the use of common ontology languages, syntax, and IDs - ultimately providing a framework for making inferences about genomic information. Should the broader ontological community elect to comply with constructing ontologies in an interoperable way then, like the GO database, all model organism databases and related scientific and ontological databases would be functionally linked as a more efficient tool for scientific discovery.

## *Discussion and Future Perspectives*

### *VI Expanding the Current Protein Ontology*

The theoretical and analytical work undertaken in this thesis presents the case for expanding the current Protein Ontology framework to include an enriched representation of proteins and protein complexes. The PRO expansion proposed includes adding a temporal domain that in itself lays a foundation for including specific properties that are necessary to define conditional protein complex formation and function. These properties include the addition of affinity attributes, environmental stimulus response and cellular compartmentation, followed by the addition of pathway attributes defining the instances involved in complex formation and function. Including these additional aspects in the current PRO framework enables a more adequate and realistic picture of protein complex formation to be represented. Importantly, the proposed expansion enables the broader scientific community to acquire more dense and reliable protein and protein-complex related information as an assist to scientific research and discovery. Furthermore, this added functionality would allow scientists to infer protein and protein complex relationships within and across species, with a correspondingly enhanced contribution to generating biological hypothesis worthy of further investigation.

### *VII Future Expansion of the Protein Ontology*

The proposed expansion of the Protein Ontology described here revolves around

Arabidopsis and its associated SCF complexes as a model system, while maintaining the applicability of the proposed expansion to all organisms. Indeed, the proposed PRO expansion could provide a foundation for including more formally represented proteins and protein complexes with respect to time and space. As such, the proposed expansion is not limited to the ideas represented in this thesis, but rather encourages future expansion and/or modification by the wider ontology research community. The choice of Arabidopsis as a model system is faithful to the idea that plants exhibit complex biological networks in relation to other higher eukaryotes. It seems fitting to develop ontology around the most complex network example so that more simplistic models can readily fit within the more complex and inclusive framework. Moreover, using Arabidopsis SCF complexes as use-case scenarios was both intuitive and illustrative, since these complexes are found within many eukaryotic systems. By far, Arabidopsis SCF protein subunit expression, assembly and function is on par in terms of complexity in comparison to other eukaryotic systems. The initiative to add specific light regimen-based classes to the PRO to define conditional protein complex assembly coincided with those unique examples in Arabidopsis where SCF complex assembly has been shown to be light-dependent.

Regarding the addition of light dependent protein complex assembly and function, specific light regimens were included based upon common light wavelength response found in Arabidopsis as a typical dicot plant model , although the light regimen classes described can be further expanded to include other light regimens as the need arises. One future expansion could involve the addition of the length of exposure time to specific light regimens, for example in the case of plants that entrain their circadian rhythm in response to specific light-dark cycles. One outcome of this work could include a method to define circadian rhythms within the ontology, which would be beneficial not just for plants but for other organisms that also rely on circadian rhythms including Drosophila and humans.

Other future additions to the PRO could include a method to define abundance levels of proteins in a way that contributes to defining the reality of protein quaternary complex formation. Currently, the inclusion of affinity properties assumes that the abundance of a protein has been met in order to meet the affinity requirements, although the inclusion of a qualitative value representing protein abundance would also be of value. The current OBO-Edit model only allows one definable cardinality value which is currently being used in the PRO to define the number of protein subunits part of a complex. In the future, the OBO-Edit model could be modified to include other cardinal values that would more specifically represent and define protein abundance and affinity values.

Future contributions to PRO expansion may include accurately defining the pathway attributes with respect to a biologically defined pathway or process term, rather than an arbitrary suffix and number (i.e. 'P1'). Such a formal definition could also lead to the proper mapping of the process or pathway term to an appropriate database like Reactome, KEGG or GO.

### *VIII Ontologies and Databases: A Modern Day Textbook*

As with other aspects of scientific research and experimentation, it is reasonable to foresee an expanded application of biological ontologies and databases as rich information resources akin to modern-day textbooks. Adopting bio-ontologies in this manner would be similar to the adoption of e-books, where volumes of information would be easily accessible and searchable through a computational device but with the added benefit that specific queries into an ontology could obviate the need to locate and peruse entire passages of a text. In future, one can imagine biomedical classrooms filled with computational devices instead of the typical textbook, notebook and pencils, where researching biologically relevant data would be implemented via bio-ontological databases. For example, a student may need to research the process by which a SCF ubiquitin ligase forms and functions; instead of reaching for a molecular biology textbook, the student would type a query pertaining to SCF ubiquitin ligases into an integrated an inter-connected bio-ontological database to retrieve the relevant information. This is just one example of the endless uses of bio-ontologies as a reference resource and assist to research. As bio-ontologies continue to be modified and expanded to represent progressively more complex biological reality, they would be more likely to be applied as an innovative reference and discovery tool.

## *IX Food for Thought: The Future of Ontological Research and Modelling*

### *IX.i Ontologies in Relational Database Schemas: A Future for PRO?*

The concept of using bio-ontologies in relational database schemas (RDSs) is not commonly documented, but it is also a growing theme in genomic related research. The current mechanism for using ontologies in RDSs focuses primarily on automating biological annotation and managing biological data. One relatively popular schema currently being used to annotate genomic data is called 'CHADO' (Mungall, C.J., et al, 2007). This schema was specifically created to help manage biological knowledge in biological databases, and is generally included in relational databases management systems like PostgreSQL (see http://www.postgreql.org for more information regarding this architecture). The schema involves implementing specific ontologies like GO, PO, and the 'Sequence Ontology' (SO; cf. Mungall C.J. et al, 2005) to annotate genomic information. Annotation software like 'MAKER' (Cantarel B., et al., 2008) also implement ontological formats such as SO to link genomic information. The GFF3 (generic feature format) annotation file-type generated by programs like MAKER can be uploaded into a biological database containing a CHADO-like schema that interprets the information via the ontologically formatted genomic information. For example, the SO terms utilized in a MAKER-generated GFF3 files include genomics-related terms like gene, mRNA, and exon that are connected by parent-child relationships. In this example, an exon would be a 'child' of mRNA that would in turn be a 'child' of a gene. Such an ontological format facilitates the manner in which genomic information is linked together. Further, in order to visualize genomic information in a biological database via a web interface, the information must be interpreted by a program such as 'GBROWSE' (Stein, L.D., et al., 2002), which displays genomic information graphically. This type of graphical representation is incorporated in many model organism databases such as TAIR (http://www.Arabidopsis.org ). The end result is that an end-user not only receives genomic information, but the genomic information is represented in a graphical format that facilitates the interpretation of the data.

It is the structure of the genomic data organized in an ontological format through the use of developed ontologies that enables software engines like MAKER, CHADO and GBROWSE to interpret, annotate, and represent the genomic data. Currently, programs like MAKER and CHADO implement only a handful of ontologies However, this area could be a potential fruitful area of application for an ontology such as the PRO directed to enabling and facilitating the annotation of specific protein and protein complex types.

### *IX.ii Live Ontologies*

In light of the many present and future benefits bio-ontologies provide to the broader scientific community, it is likely that bio-ontologies will continue to proliferate and increase in complexity. As previously mentioned, this can be both extremely beneficial and at the same time troublesome, although the issues to be addressed are not so daunting as to impede the formal construction and maintenance of bio-medical ontologies. If the broader ontological community complies with the conventions set out in developing bio-ontologies, the nature of bio-medical scientific research would greatly benefit. One can imagine bio-ontologies providing not just controlled vocabularies and a method to annotate biological data, but also evolving into more intelligent 'machines' able to interpret and make inferences about data contained within individual ontologies and across multiple model organism databases. By way of example, one could query an ontological database for the sub-cellular compartment in which a particular protein interaction was known to take place, where the output would be definitive for the compartmentation status, and include all the properties and reasoning for the returned conclusion. This level of functionality can only be realized if ontologies are modified and expanded so as to be as formal and specific as possible - in other words, to fully capture the reality of biological knowledge and complexity.

### *IX.iii Biological Systems Engineering*

Taken together, bio-ontologies provide a system for organizing and formally defining biological information from the general to the more specific. In light of this, if ontologies are developed and maintained in a formal and complex manner, they could evolve to be trusted and genuine sources of biological information. Taking the example of the expanding area of biological engineering, ontologies could be regarded as a source of pertinent information required to build biological systems. For example, in the case of a protein complex that forms dependent upon very specific conditions, PRO could be accessed for this particular information. A researcher working to develop a genetically engineered eukaryote that exploits a 26S proteasome system to degrade proteins could refer to the PRO to identify the potential subunits, conditions, and protein interaction affinities required for a specific E3 ubiquitin ligase to assemble and function. These are just some hypothetical examples of many instances that an ontology could be helpful for the timely, accurate and automated retrieval of biologically relevant information.

The future of biological ontology development and deployment appears to be very promising. By remaining faithful and inclusive in the representation of biological complexity, future ontology development is poised to have a major impact on the future of scientific research.

***References***

Antoniou, G. and van Harmelen, F. (2008). A Semantic Web Primer. Cambridge, MA and London, England: The MIT Press.

Arighi, C. N. (2011). Chapter 6 A Tutorial on Protein Ontology Resources for Proteomic Studies. Methods Mol Biol *694*, 77–90.

Ashburner, M., Ball, C., Blake, J., Botstein, D., Butler, H., Cherry, J., Davis, A., Dolinski, K., Dwight, S., Eppig, J., et al. (2000). Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. Nature genetics *25*, 25–34.

Badder, F., Horrocks, I., and Sattler, U. (2004). Description Logics. In *Staab, S. and Studer,R., Handbook on Ontologies*, pp. 3-28: Berlin: Springer-Verlag.

Ball, C. A., Dolinski, K., Dwight, S. S., Harris, M. A., Issel-tarver, L., Kasarskis, A., Scafe, C. R., Sherlock, G., Binkley, G., Jin, H., et al. (2000). Integrating functional genomic information into the Saccharomyces Genome Database. *28*, 77–80.

Bard, J., Rhee, S. Y., and Ashburner, M. (2005). An ontology for cell types. Genome Biology *6*, R21.

Bateman,A ., Birney, E., Cerruti, L., Durbin, R., Etwiller, L., Eddy, S.R., Jones, S., Howe, K.L., Marshall, M., Sonnhammer, E.L.L. (2002). The Pfam Protein Families Database. Nucleic Acids Research. *30*, 276-280.

Beisswanger, E., Lee, V., Kim, J.J., Rebholz-Schuhmann, D., Splendiani, A., Dameron, O., Schulz, S., Hahn, U. (2008). Gene Regulation Ontology (GRO): design principles and use cases. Stud Health Technol Inform. *136*, 9-14.

Bittner, T., and Smith, B. (2003). Formal ontologies for space and time. Technical report, IFOMIS-University of Leipzig. Retrieved from http://www.ifomis.org/bfo/publications, August 1, 2012.

Bittner, T., and Smith, B. (2004). Normalizing Medical Ontologies using Basic Formal Ontology. Proceedings of GMDS Innsbruck, 26-30 September 2004. Niebull:Videel OHG, 199-201. Retrieved from http://www.ifomis.org/bfo/publications. August 1, 2012.

Blake, J. A., Eppig, J. T., Richardson, J. E., and Davisson, M. T. (2000). The Mouse Genome Database (MGD): expanding genetic and genomic resources for the laboratory mouse. *28*, 108–111.

Bodenreider, O., and Stevens, R. (2006). Bio-ontologies: current trends and future directions. Briefings in bioinformatics *7*, 256–274.

Bodenreider, O. (2004). The Unified Medical Language System (UMLS): integrating biomedical terminology. Nucleic Acids Res. 32 (database issue), D267–D270.

Bult, C. J., Drabkin, H. J., Evsikov, A., Natale, D., Arighi, C., Eustachio, P. D., Smith, B., Blake, J. A., and Wu, C. (2011). The Representation of Protein Complexes in the Protein Ontology ( PRO ). BMC bioinformatics *12*, 371.

Calderon-villalobos, L. I., Tan, X., Zheng, N., & Estelle, M. (2010). Auxin Perception-Structural Insights. Cold Spring Harb Perspect Biol. *2,* a005546.

Cantarel, B., Korf, I., Robb, S., Parra, G., Ross, E., Moore, B., Holt, C., Alvarado, A.S., Yandell, M. (2008) MAKER: An easy-to-use annotation pipeline designed for emerging model organism genomes. Genome Res *18*:188–196.

Casal, J. J., Luccioni, L. G., Oliverio, K. a., & Boccalandro, H. E. H. E. (2003). Light, phytochrome signalling and photomorphogenesis in Arabidopsis. Photochemical & Photobiological Sciences, *2*(6), 625.

Consortium, T. F. (1999). The FlyBase Database of the Drosophila Genome Projects and community literature. *27*, 85–88.

Consortium, T. U. (2010). The Universal Protein Resource (UniProt) in 2010. Nucleic acids research *38*, D142–8.

Crick, F. (1970). Central Dogma of Molecular Biology. Nature. *227,* 561-563.

Croft D, O'Kelly G, Wu G, Haw R, Gillespie M, Matthews L, Caudy M, Garapati P, Gopinath G, Jassal B, et al. (2011). Reactome: a database of reactions, pathways and biological processes. Nucleic Acids Res. *39*:D691-D697

Cusick, N. Klitgord, M. Vidal, D.E. Hill. (2005). Interactome: gateway into systems biology. Hum. Mol. Genet. *14*, R171–R181

Day-richter, J., Harris, M. A., Haendel, M., Obo-edit, T. G. O., Group, W., and Lewis, S. (2007). Databases and ontologies OBO-Edit — an ontology editor for biologists. Bioinformatics *23,* 2198–2200.

Day-Richter, J. (2006). The OBO Flat File Format Specification, version 1.2. The Gene Ontology. Retrieved July, 15, 2012, from www.geneontology.org

Daconta, M.C., Obrst, L.J., and Smith, K.T. (2003). The Semantic Web. Indianapolis, Indiana: Wiley Publishing.

Devlin, P. F., and Kay, S. A. (2000). Cryptochromes are required for phytochrome signaling to the circadian clock but not for rhythmicity. The Plant cell, *12*(12), 2499–2510.

Dieterle, M., Zhou, Y. C., Schäfer, E., Funk, M., & Kretsch, T. (2001). EID1, an F-box protein involved in phytochrome A-specific light signaling. Genes & Development, *15*(8), 939.

Eilbeck K., Lewis S., Mungall C., Yandell M., Stein L., Durbin R., Ashburner M. (2005). The sequence ontology: a tool for unification of genome annotations. Genome Biol. *26*:R44

Gagne, J. M., Downes, B. P., Shiu, S.-H., Durski, A. M., & Vierstra, R. D. (2002). The F-box subunit of the SCF E3 complex is encoded by a diverse superfamily of genes in Arabidopsis. Proceedings of the National Academy of Sciences of the United States of America, *99*(17), 11519–24.

Gomez-Perez, A., Fernandez-Lopez, M., and Corcho, O. (2004). Ontological Engineering. London: Springer-Verlag.

Grenon, P., Smith, B., and Goldberg, L. (2004). Biodynamic Ontology : Applying BFO in the Biomedical Domain. In *Ontologies in Medicine*. pp. 20-38: Amsterdam: IOS Press.

Gruber, T. R. (1991). The Role of Common Ontology in Achieving Sharable, Reusable Knowledge Bases. In J.A Allen, R. Fikes, & E. Sandewall, (Eds.) *Principles of Knowledge Representation and Reasoning: Proceedings of the Second International Conference*, pp. 601-602: Cambridge, MA, Morgan Kaufmann.

Gruber, T. R. (1993). Toward Principles for the Design of Ontologies Used for Knowledge Sharing. Int.J. Human-Computer Studies. *43,* 907-928

Han, L., Mason, M., Risseeuw, E. P., Crosby, W. L., Somers, D. E., & Group, G. E. (2004). Formation of an SCF ZTL complex is required for proper regulation of circadian timing, 1, 291-301.

Hotton, S. K., and Callis, J. (2008). Regulation of cullin RING ligases. Annual review of plant biology *59,* 467–489.

Hua, Z., and Vierstra, R. D. (2011). The cullin-RING ubiquitin-protein ligases. Annual review of plant biology *62,* 299–334.

Jackman, M., Kubota, Y., Elzen, N. D., and Hagting, A. (2002). Cyclin A- and Cyclin E-Cdk Complexes Shuttle between the Nucleus and the Cytoplasm. 13, 1030–1045.

Jaiswal, P., Avraham, S., Ilic, K., Kellogg, E. A., Mccouch, S., Stevens, P., Vincent, L., et al. (2006). Plant Ontology ( PO ): a controlled vocabulary of plant. Comparative and Functional Genomics, 388-397.

Jaiswal, P., Ware, D., Ni, J., Chang, K., Zhao, W., Schmidt, S., Pan, X., et al. (2002). Gramene: development and integration of trait and gene ontologies for rice. Comparative and functional genomics, 3(2), 132-6

Kanehisa, M. and Goto, S. (2000). KEGG: Kyoto Encyclopedia of Genes and Genomes. Nucleic Acids Res. *28,* 27-30.

Marrocco, K., Zhou, Y., Bury, E., Dieterle, M., Funk, M., Genschik, P., Krenz, M., et al. (2006). Functional analysis of EID1, an F-box protein involved in phytochrome A-dependent light signal transduction. The Plant journal : for cell and molecular biology, *45*(3), 423–38.

Mika, P. (2007). Social Networks and the Semantic Web. In *Semantic Web and Beyond*. Volume 5, Part II. pp 65-92: New York, NY. Springer US

Mishra, R. B., and Kumar, S. (2010). Semantic web reasoners and languages. Artificial Intelligence Review, 35(4), 339–368.

Montecchi-Palazzi L, Beavis R, Binz PA, Chalkley RJ, Cottrell J, Creasy D, Shofstahl J, Seymour SL, Garavelli JS. (2008). The PSI-MOD community standard for representation of protein modification data. Nat. Biotechnol. *26*:864-866.

Mungall, C.J. and Emmert, D.B. (2007). A Chado case study: an ontology-based modular schema for representing genome-associated biological information. Bioinformatics. *23*, i337-i346.

Nakamichi, N. (2011). Molecular mechanisms underlying the Arabidopsis circadian clock. Plant & cell physiology, *52(10)*, 1709-18.

Natale, D. a, Arighi, C. N., Barker, W. C., Blake, J. a, Bult, C. J., Caudy, M., Drabkin, H. J., D'Eustachio, P., Evsikov, A. V., Huang, H., et al. (2011). The Protein Ontology: a structured representation of protein forms and complexes. Nucleic acids research. *39*, D539–45.

Natale, D., Arighi, C.N., Barker, W.C, Blake, J., Chang,T., Hu, Z., Liu, H., Smith. B., and Wu, C.H. (2007). Framework for a Protein Ontology. BMC Bioinformatics 2007, *8(Suppl 9)*:S1, 1-12.

Nelson, L.D. and COX, M.M. (2004). Lehninger Principles of Biochemistry 4th edition. W.H.Freeman. Pgs, 160-161.

Noy, N. F., and McGuinness, D. L. (2001). Ontology Development 101 : A Guide to Creating Your First Ontology. Technical Report SMI-2001-0880, Stanford Medical Informatics.

Noy,N.F., Shah,N.G., Whetzel, P.L, Dai, B., Dorf, M., Griffith, N., Jonquet, C., Rubin, D.L., Storey, M.A., Chute, C.G., and Musen, M.A. (2009). BioPortal: ontologies and integrated data resources at the click of a mouse. Nucleic Acids Research (web server issue). 37, W170-W173

Phizicky, E. M., and Fields, S. (1995). Protein-protein interactions: methods for detection and analysis. Microbiological reviews. *59,* 94–123.

Robson, F., Okamoto, H., Patrick, E., Harris, S.-R., Wasternack, C., Brearley, C., & Turner, J. G. (2010). Jasmonate and phytochrome A signaling in Arabidopsis wound and shade responses are integrated through JAZ1 stability. The Plant cell, *22*(4), 1143–60.

Rosse, C., & Mejino, J. L. V. (2003). A reference ontology for biomedical informatics: the Foundational Model of Anatomy. Journal of biomedical informatics, *36*(6), 478–500.

Schreiber, A., Stengel, F., Zhang, Z., Enchev, R. I., Kong, E. H., Morris, E. P., Robinson, C. V., da Fonseca, P. C. a, and Barford, D. (2011). Structural basis for the subunit assembly of the anaphase-promoting complex. Nature *470*, 227–232.

Smith, B. (2004). Beyond Concepts : Ontology as Reality Representation. In *Formal Ontology and Information Systems*. pp. 73-84: Amsterdam: IOS Press.

Smith, B., Ashburner, M., Rosse, C., Bard, J., Bug, W., Ceusters, W., Goldberg, L. J., Eilbeck, K., Ireland, A., Mungall, C. J., et al. (2007). The OBO Foundry : coordinated evolution of ontologies to support biomedical data integration. Nature Biotechnology *25*, 1251–1255.

Smith, B., and Ceusters, W. (2010). Ontological realism: A methodology for coordinated evolution of scientific ontologies. Applied ontology *5*, 139–188.

Smith, B., Ceusters, W., Klagges, B., Köhler, J., Kumar, A., Lomax, J., Mungall, C., Neuhaus, F., Rector, A. L., and Rosse, C. (2005). Relations in biomedical ontologies. Genome biology *6*, R46.

Smith, B., Köhler, J., and Kumar, A. (2004). On the Application of Formal Principles to Life Science Data : A Case Study in the Gene Ontology. In Data Integration in the Life Sciences (DILS). pp 79-94: Proceedings of the First international Workshop, DILS. Berlin and Heidelberg: Springer.

Soldatova, L. N., and King, R. D. (2005). Are the current ontologies in biology good ontologies? Nature biotechnology *23*, 1095–1098.

Stein,L.D., Mungall,C., Shu,S., Caudy,M., Mangone,M., Day,A., Nickerson,E., Stajich,J.E., Harris,T.W., Arva,A. et al. (2002). The generic genome browser: a building block for a model organism system database. Genome Res. *12*, 1599–1610.

The Arabidopsis Genome Initiative. (2000). Analysis of the genome sequence of the flowering plant *Arabidopsis thaliana*. Nature *408*, 796–815.

Tan, X., Calderon-Villalobos, L.I., Sharon, M., Zheng, C., Robinson, C.V., Estelle, M., Zheng, N. (2007). Mechanism of auxin perception by the TIR1 ubiquitin ligase. Nature. *446*, 640-645.

Tirmizi, S. H., Aitken, S., Moreira, D. A., & Mungall, C. (2009). OBO & OWL : Roundtrip Ontology Transformations. Semantic Web Applications and Tools for Life Sciences. Aachen, Germany: CEUR Workshops Proceedings.

Uschold, M., and Gruninger, M. (1996). Ontologies : Principles, Methods and Applications. Knowledge Engineering Review *11*, 93-136.

Vogt, L., Grobe, P., Quast, B., and Bartolomaeus, T. (2012). Accommodating ontologies to biological reality--top-level categories of cumulative-constitutively organized material entities. PloS one *7*, 1–19.

Wang, X., Li, W., Piqueras, R., Cao, K., Deng, X. W., and Wei, N. (2009). Regulation of COP1 nuclear localization by the COP9 signalosome via direct interaction with CSN1. The Plant Journal. *58*, 655–667.

Ware, D., Jaiswal, P., Ni, J., Pan, X., Chang, K., Clark, K., Teytelman, L., et al. (2002). Gramene: a resource for comparative grass genomics. Nucleic acids research, 30(1), 103-5.

Whetzel, P. L., Parkinson, H., Causton, H. C., Fan, L., Fostel, J., Fragoso, G., Game, L., Heiskanen, M., Morrison, N., Rocca-Serra, P., et al. (2006). The MGED Ontology: a resource for semantics-based description of microarray experiments. Bioinformatics (Oxford, England) *22*, 866–873.

# *Appendices*

## *Appendix A Table of terms and definitions included in the PRO expansion*

| Term | Super-class<br>(is a) | Definition |
|---|---|---|
| continuant | entity | (BFO) An entity that exists in full at any time in which it exists at all, persists through time while maintaining its identity and has no temporal parts |
| occurrent | entity | (BFO) An entity that has temporal parts and that happens, unfolds or develops through time |
| processual entity | occurrent | (BFO) An occurrent that exists in time by occurring or happening, has temporal parts and always involves and depends on some entity |
| material entity | continuant | (BFO) An independent continuant that is spatially extended whose identity is independent of that of other entities and can be maintained through time |
| object | material entity | (BFO) A material that is spatially extended, maximally self-connected and self-contained (the parts of a substance are not separated from each other by spatial gaps) and possesses an internal unity. The identity of substantial object entities is independent of that of other entities and can be maintained through time. |
| protein | object | An amino acid chain that is produced de novo by ribosome-mediated translation of a genetically-encoded mRNA |
| macromolecular complex | processual entity | A stable assembly of two or more macromolecules, i.e. proteins, nucleic acids, carbohydrates or lipids, in which the constituent parts function together |
| protein complex | macromolecular complex | Any macromolecular complex composed of two or more polypeptide subunits, which may or may not be identical. Protein complexes may have other associated non-protein prosthetic groups, such as nucleotides, metal ions or carbohydrate groups |
| radiation responsive protein complex | protein complex | Any macromolecular complex that is composed of two or more polypeptide subunits, which may or may not be identical, that assembles and carries out some function in response to an exposure to a radiation type, intensity, or quantity. |
| dark (no light) responsive protein complex | radiation responsive protein complex | Any macromolecular complex that is composed of two or more polypeptide subunits, which may or may not be identical, that assembles and carries out some function in response to an exposure to darkness (no light) |
| light responsive protein complex | radiation responsive protein complex | Any macromolecular complex that is composed of two or more polypeptide subunits, which may or may not be identical, that assembles and carries out some function in response to an exposure to day light as the light of the sun |

| | | |
|---|---|---|
| blue-light responsive protein complex | light responsive protein complex | Any macromolecular complex that is composed of two or more polypeptide subunits, which may or may not be identical, that assembles and carries out some function in response to an exposure to blue light in the wavelength range of 455-492 nM |
| far-red light responsive protein complex | light responsive protein complex | Any macromolecular complex that is composed of two or more polypeptide subunits, which may or may not be identical, that assembles and carries out some function in response to an exposure to far-red light in the wavelength range of 700-800 nM |
| red light responsive protein complex | light responsive protein complex | Any macromolecular complex that is composed of two or more polypeptide subunits, which may or may not be identical, that assembles and carries out some function in response to an exposure to red-light in the wavelength range of 622-780 nM |
| red and far-red light responsive protein complex | light responsive protein complex | Any macromolecular complex that is composed of two or more polypeptide subunits, which may or may not be identical, that assembles and carries out some function in response to an exposure to red or far-red light in the wavelength range of 622-800 nM |

## *Appendix B* *Table of new relations and definitions*

| Relation | Properties | Definition |
|---|---|---|
| has affinity for | is not transitive | A relationship between two entities, either two continuant entities or a continuant and occurrent entity, in which an affinity relationship has been defined between the two entities through experimental evidence, and may or may not be represented by an experimentally defined Kd value.<br>Example: ‘protein A’ - has affinity for - ‘protein B’ |
| negatively regulates abundance | is not transitive | A relationship between an occurrent and continuant entity, in which the occurrent entity acts upon the continuant entity in a manner such that the relative abundance of the continuant object is negatively affected and results in a reduced abundance of the continuant object.<br>Example:<br>‘protein complex ABC’ - negatively_regulates_abundance - ‘protein X’ |

# *Vita Auctoris*

Claudia Lucia DiNatale was born in 1981 in Windsor, Ontario. She graduated from Catholic Central High School in 2000. From there she went to the University of Windsor where she obtained a B.Sc honours with Thesis in Biology in 2010. She is currently a candidate for the Master`s degree in Biology at the University of Windsor and hopes to graduate in Fall 2012.