An Ontology of Gene

Hiroshi Masuya[1,*] and Riichiro Mizoguchi[2]
[1] RIKEN BioResource Center, Tsukuba, Japan
2 Department of Knowledge Systems, ISIR, Osaka University, Ibaraki, Japan

ABSTRACT

The concept of a gene was established in the era of classical genetics and is now essential for life science for elucidating the molecular basis of the coding of genetic information necessary to realize the body of an organism and its biological functions. However, an ontology fully representing multiple aspects of a gene is still not available. In this study, we dissected the biological and ontological definitions of bearers of genetic information, including genes and alleles. Based on this analysis, we then proposed a basic way of modeling an ontology that represents the common definitions in classical and molecular genetics. This ontology is available at http://www.brc.riken.jp/lab/bpmp/ontology/ontology_gene.html.

1 INTRODUCTION

"Gene" is one of the most fundamental concepts composing the basis of modern biomedical science, established in the 1900s. Today, genetic information is known to be coded molecularly using a sequence of four types of polynucleotide bases, providing blueprints for the development of an organism's body and its biological functions.

Recently, gathering information using a computer has become more and more important as a component of research studies in life science. Ontology provides us with one of the most important means of processing varieties of data and representing knowledge models. Currently, several biomedical ontologies have been constructed with the aim of integrating a variety of information produced by different fields of biology. Therefore, development of a common model/ontology for genes is one of the key issues in bioinformatics studies.

However, the simple modeling of a gene and its allele into an ontology can raise some problems. For example, because a gene can be formed of two kinds of molecular entities, DNA or RNA, we cannot define a single class of gene under a hierarchical tree of molecular classifications.

Another problem with the classification arises with the hierarchy of "gene => allele => molecular instance". This hierarchy corresponds to a hierarchy of "*Gdf5* of mouse (gene) => *Gdf5*^{Rgsc451} (allele) => molecular instance of *Gdf5*^{Rgsc451}". As a result, we cannot define the gene or allele class in the ontology. In this case, each class of gene and allele seems to have another super-class, termed "gene" and "allele", respectively.

In this study, we dissected the role of a gene and built an ontology that represents a consistent data model of the basic concepts of genetics, including genes, alleles, and nucleic acid molecules.

2 DISSECTION OF THE MEANING OF GENE

In classical genetics, a gene is defined as a "particle-like thing to define a genetic trait". Genetic information conveyed by a gene was thought to have effects on specific body features or biological functions and to result in biological variation. To explain the levels of biological variation, a gene, or a factor that controls a specific trait in a species, and an allele, or a variant of a gene that produces phenotypic variations of a trait, must be clearly discriminated in genetics.

On the other hand, in molecular biology, genomic segments are classified according to their roles or features as genes, non-genics, promoters, exons, introns, and so on. A gene is defined as a segment that codes one or multiple products (functional polypeptide or RNA).

The role of "bearer of genetic information" is essential for a gene. It is notable that the definition of a gene was made prior to the discovery of its material basis. To put it extremely, any molecular entity that bears genetic information in a "unit of" functional role could be termed a gene.

The role of a gene can be classified into two categories. One is the transmission of genetic information from parents to their offspring. This is driven by a series of biological processes resulting in the self-replication of nucleic acid chains. Additionally, through the effects of mutation (the alteration of genetic information), the self-replication processes contribute to genetic variation. In this sense, the hierarchy of gene to allele corresponds to the levels of genetic variation: namely, inter- and intra-species variations.

Another role of genes is to encode the design plan of individual organisms, namely, the coding of gene products. This function is accomplished by a sequence of biological processes including gene transcription and the translation of mRNA to a polypeptide. This role is equivalent to the regional classification of genomic segments in molecular biology.

In summary, the classifications of genomic segments can be categorized into two dimensions. One is the direction to increase genetic diversification based on self-replication; the other is a region-based classification of genomic segments. These classifications are distinguished by their fundamental roles in providing different contents of genetic information through different biological processes. For instance, a gene, or a species-specific variation in a genome segment that

[*] To whom correspondence should be addressed: hmasuya@brc.riken.jp

codes a gene product, can be represented as the point of intersection of these two systems (Fig. 1).

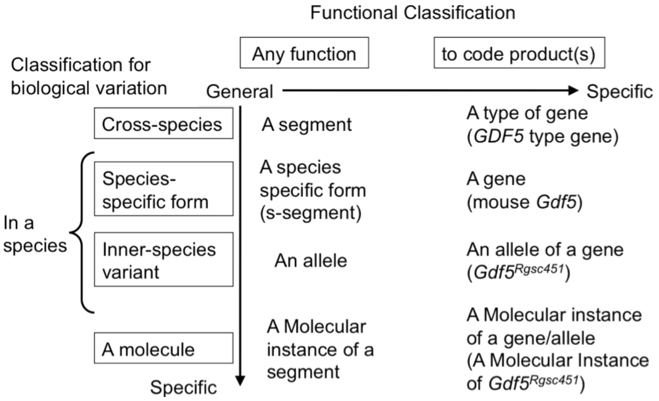

Functional Classification
Any function
to code product(s)
Classification for biological variation
General
Specific
Cross-species
A segment
A type of gene (*GDF5* type gene)
Species-specific form
A species specific form (s-segment)
A gene (mouse *Gdf5*)
In a species
Inner-species variant
An allele
An allele of a gene (*Gdf5*^{Rgsc451})
A molecule
A Molecular instance of a segment
A Molecular instance of a gene/allele (A Molecular Instance of *Gdf5*^{Rgsc451})
Specific

Fig. 1. Two-dimensional classification of genomic segment.

The above-mentioned problem, i.e., that the concepts of gene and allele seem to contain "another super-class" can be solved if they are regarded as top-level concepts for roles in a specific context.

3 DETAILED DISSECTION OF THE ROLES OF GENOMIC SEGMENTS

Next, we further dissected the detailed roles of genomic segments.

3.1 Similarities and differences between the two kinds of genetic information

The two kinds of genetic information carried by a gene are commonly represented by the specific form, or the functional structure of the polynucleotide molecule. This structure can be compared to an artificial information bearer in which symbol sequences (e.g., text) are conveyed on an information media (e.g., paper).

In the description of information using text, a "symbol" is the most fundamental unit. In the English language, the letter "G" in the alphabet represents an abstract meaning (informational object) of the symbol "G" with the form of a line image in the shape of "G". The coding of genetic information by a molecule has the same structure. The genetic information of "G" is representation of the abstract symbol of "G of the genetic code," the molecular entity of which corresponds to a guanine nucleotide. (Mizoguchi 2004)

Using a sequence of such molecular symbols, representing the polynucleotide groups, the genetic information for an organism can be conveyed. This is analogous to a paper document conveying a representation of a story or a specification of something using a symbol sequence.

On the other hand, the content represented by the two kinds of genetic information is clearly different. With self-replication, the content represented is the design plan for the nucleotide chain as itself. In the coding of a gene product, the content is a sequence of polypeptides or functional RNA as a design plan.

3.2 Detailed classification of role for self-replication

The roles of self-replication and diversification can be classified as follows:

(1) Level general to all organisms:

A molecular entity, such as polynucleotide group, is termed as a genomic segment and plays a role in bearing information for self-replication under the general context of the organism.

(2) Level reflecting the identity of a biological species:

The genomic segment plays a specified role in bearing species-specific genetic information under the context of the population of the species (or the gene pool). At this level, the genomic segment may be often termed as a genetic marker. In this paper, we call it an "s-segment".

(3) Level reflecting inter-species variation:

Under the context of species populations, another role exists: encoding inter-species variations. In this context, the s-segment is referred to as an allele.

3.3 Region-based classification of genomic segments

On the other hand, under the context of an organism, a genomic segment, which codes a polypeptide or a functional RNA, is generally termed as a "gene". This is not a species-specific gene, as described above. In this paper, we refer to this context as a "gene type" for discriminatory purposes.

4 ONTOLOGICAL MODEL FOR GENE

Based on these analyses, we worked out an ontological model to describe the concept of gene. We referred to Yet Another More-Advanced Top-level Ontology (YAMATO), which represents a basic framework of roles and representation (Mizoguchi 2004; 2009). For the modeling tool, we used the Hozo ontology editor, which enables a systematic description of role-related concepts and their contexts (Hozo).

With the Hozo ontology editor, concepts are fundamentally categorized into basic concepts, which can be defined context-independently, and role concepts to be played by an entity, which are dependent on a specific context. An entity playing a role concept is termed as a "role-holder". For example, a teacher, a person who plays the role of a teacher role in a school is a role-holder. The player is selected from the basic concepts or role-holders (Kozaki et al., 2007).

Definitions of key concepts in natural language are shown below and their schematic representation in Hozo is depicted in Figure 2

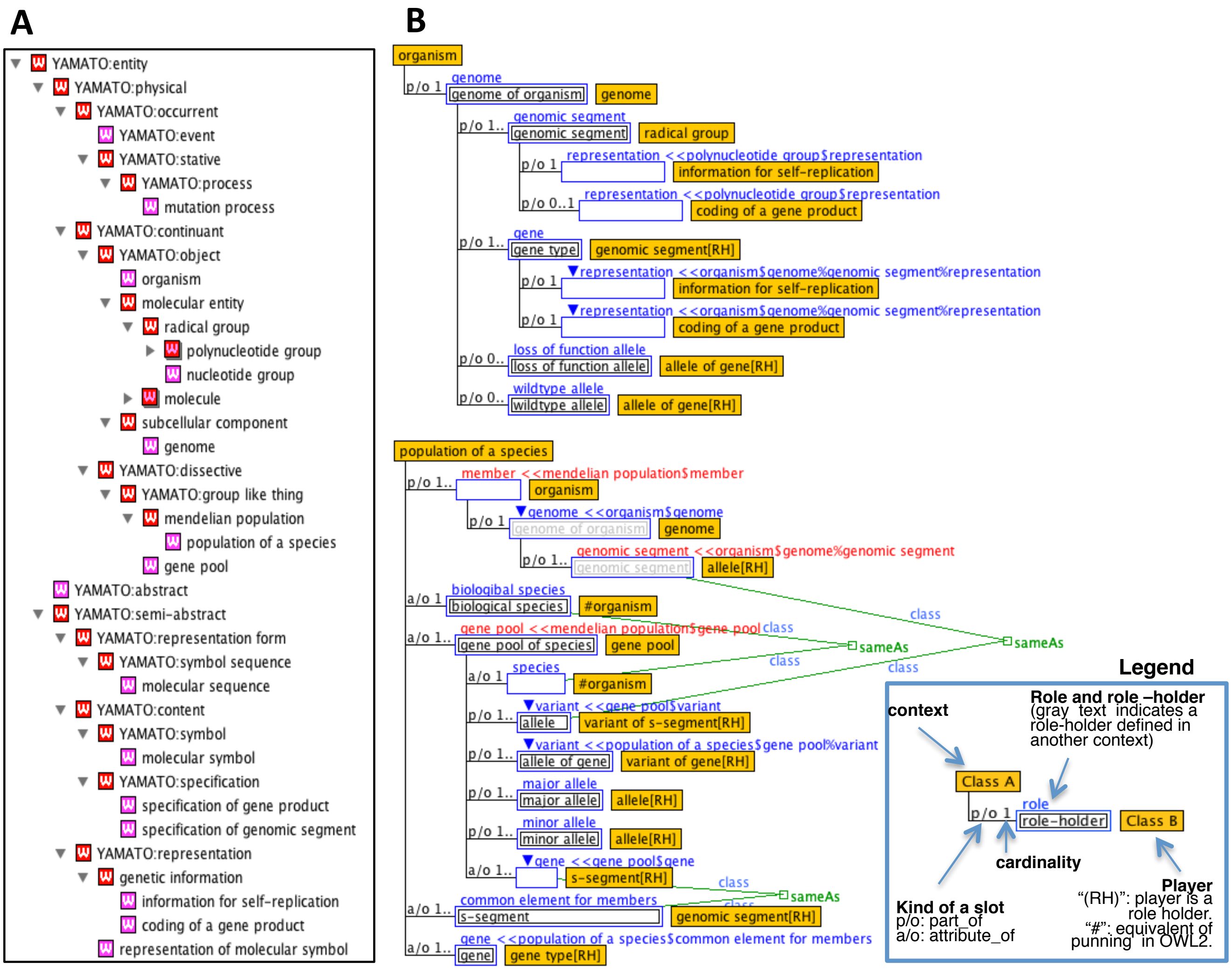


Fig. 2. Modeling of gene, allele and genetic information according to the Hozo ontology editor. A: Class tree of basic concepts. B: Composition of ***organism*** and ***population of a species***.

(i) ***Genomic segment*** $=_{\text{def}}$ A role holder played by a ***molecular entity*** or ***polynucleotide group***, under the context of ***organism***.

(ii) ***Gene type*** $=_{\text{def}}$ A role holder played by a ***genomic segment***, which has both of the two kinds of genetic information, i.e., ***information for self-replication*** and ***coding of gene product***, and inherited ***genomic segment*** under the ***organism*** context.

(iii) ***S-segment*** $=_{\text{def}}$ A role holder played by a ***genomic segment*** under the context of a ***population of a species***, a ***Mendelian population***.

(iv) ***Gene*** $=_{\text{def}}$ A role holder played by a ***gene type*** under the context of a ***population of a species***, which is defined by the specialization of the ***s-segment***.

(v) ***Mutation*** $=_{\text{def}}$ A role holder played by a ***genomic segment***, which is the terminal state in the process under the context of the ***mutation process***.

(vi) ***Variant of s-segment*** $=_{\text{def}}$ A role holder played by a ***s-segment*** under the context of a ***mutation process in gene pool.***

(vii) ***Allele*** $=_{\text{def}}$ A role holder played by a ***variant of*** *an* ***s-segment*** under the context of a ***gene pool of*** *a* ***population of*** *a* ***species.***

(viii) ***Genetic information entity*** $=_{\text{def}}$ A ***YAMATO: representation***, composed of a ***molecular sequence*** of a representation form and ***YAMATO:specification*** as a content.

(ix) ***Molecular sequence*** $=_{\text{def}}$ A ***YAMATO:symbol sequence*** composed of a ***molecular symbol***.

(x) ***Molecular symbol*** $=_{\text{def}}$ A ***YAMATO:symbol***, which is a ***YAMATO:content***.

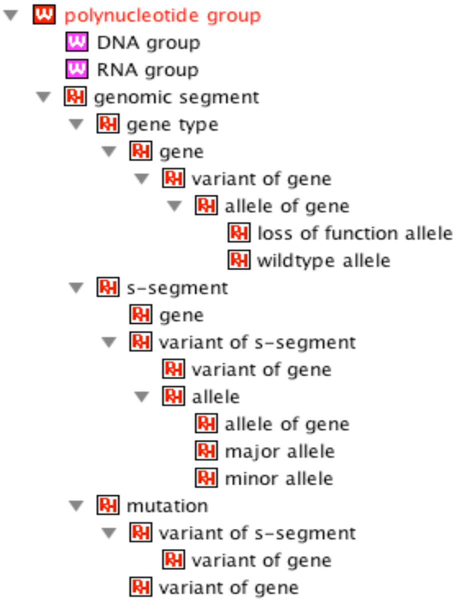
(xi) ***Representation of molecular symbol*** $=_{\text{def}}$ A ***YAMATO:representation***, which is composed of **a**

radical group as a form and a ***molecular symbol*** as a content.

(xii) ***Major allele*** =$_{\mathrm{def}}$ A role holder played by an ***allele*** under the context of the ***population of a species***.

*(xiii)****Loss of function allele*** =$_{\mathrm{def}}$ A role holder played by a ***gene allele*** under the context of an ***organism***.

Hozo's reasoner dynamically generates a classification of role-holders and their players by referring to inter-relationships among slots and class restrictions (Fig. 3). Because the classification of *genomic segment* is dependent on multiple contexts, the inferred hierarchy represents multiple inheritances of role-holders. Here, the hierarchy demonstrates the practical orders of classification as "*polynucleotide group => genomic segment => s-segment => allele => major allele*", "*s-segment => gene => gene allele => loss of function allele*" with the help of roles.



- W polynucleotide group
  - W DNA group
  - W RNA group
  - RH genomic segment
    - RH gene type
      - RH gene
        - RH variant of gene
          - RH allele of gene
            - RH loss of function allele
            - RH wildtype allele
    - RH s-segment
      - RH gene
      - RH variant of s-segment
        - RH variant of gene
        - RH allele
          - RH allele of gene
          - RH major allele
          - RH minor allele
    - RH mutation
      - RH variant of s-segment
        - RH variant of gene
      - RH variant of gene

Fig. 3. An partial "is_a" hierarchy of classes (represented as "W") and role-holders (RH) generated by the Hozo-reasoner. One role-holder appears multiple times in the hierarchy because of multiple inheritance.

5 DISCUSSION

5.1 Major outcomes of this study

The meaning of the term "gene" seems to have changed historically. However, biologists have no difficulty thinking in Mendelian terms when applying traditional genetic techniques and are able to switch seamlessly to molecular conceptions of the gene (Griffiths et al. 2007). Therefore, the core concept of "gene" should be considered as being consistent between classic and molecular genetics. Although further refinements of the presently reported theory are needed, the present study may provide a foundation for the construction of a semantic data model for the concept of gene applicable to broad fields of life science including genetics, molecular biology, and population genetics.

First, we defined the body of information carried by a gene using YAMATO, which is a top-level ontology based on the traditional and commonplace Aristotelian ontology, such as the Basic Formal Ontology (BFO) and the Descriptive Ontology for Linguistic and Cognitive Engineering (DOLCE). The ontology of representation in YAMATO shows the definition of the general structure for variously styled "content-bearing informational entities", or ***representations***, composed of a ***representation form*** and a ***content***. A physical thing conveying a ***representation*** (i.e., books or electric document files) is defined as a ***representing thing***. The ontology of representation provides a detailed theory for the classification and instantiation of these abstract and physical entities (Mizoguchi, 2004).

A ***genetic information entity***, which is transmitted from a parent to its progeny (or from a genomic segment to a replicated segment) was modeled as a specific combination of a symbol-sequence pattern of nucleotides and informational content to specify a one-dimensional structure of itself or a non-self molecule (gene product). This model revealed the separation of "information for self-replication", which all genomic segments have, and "coding of gene products", which only genes have. The modeling of a ***genomic segment*** analogous to a ***representing thing*** enables multiple copies of a specific gene to share the same genetic information.

Another achievement is the modeling of a multiplex classification of genomic segments that play different roles in different contexts. By dissecting the biological role of ***genetic information entities*** and their contexts (i.e., organisms or biological populations), we demonstrated the systematic organization of concepts derived from genomic segment without multiple inheritance. This achievement was obtained thanks to the Hozo tool, in which the role theory of YAMATO is well embodied as operations to edit ontologies (Kozaki et. al. 2007, Mizoguchi 2009, Hozo).

5.2 Related work

Various broad concepts of "gene" have been proposed among existing biomedical ontologies. For example, the Sequence Ontology (SO) classifies a gene as a "*region*", which is a "*sequence feature*" composed of bases and a polypeptide region composed of amino acids (Eilbeck et al., 2005). Other ontologies classify a gene as a biological macromolecule (Foundational Model of Anatomy Ontology: FMA), a genetic observation type (Health Level Seven: HL7), or a genetic interval (Ontology of Genetic Interval: OGI). Although each of these classifications represents an aspect of a gene, they do not encompass the meaning generally accepted. To solve this problem, several efforts have been undertaken by the OBO Foundry and related groups.

Hoehndorf et al**.** proposed a system of axioms for SO's top-level categories based on three primitive terms **(***Molecular sequence*, *Syntactic sequence***,** and *Abstract sequence***)** to specify the meaning of sequence-related terms used in the biological domain (Hoehndorf et al**.,** 2009). As they mentioned, because the axiom system is compatible with multiple top-level ontologies, these primitive terms are equivalent to a molecular entity as a ***representing thing***, an electric physical entity as a ***representing thing*****,** and a symbol-sequence pattern as a ***representation form*** in YAMATO. They provided a detailed mereology of sequences, which is not provided in the present study. However, they did not deal with ***genetic information entity*** as ***representation***, which has been addressed here. Consequently, the system cannot solve the problem of the multiplex classification of genomic segments.

Recently, the SO was revised to define the mereological, spatial, and temporal aspects of a biological sequence (Mungall et al., 2010). As an extension of the SO, they defined a new ontology, the Sequent Ontology: Molecules (SOM), an ontology for genomic segments to create a partially isomorphic hierarchy to the SO. However, the class hierarchy of SO itself cannot solve the above-mentioned problem for genomic segments. For example, the relationship, "allele variant_of gene", does not contribute to the inheritance of a gene's attributes in an allele.

Thus, we think our effort is complementary to sequence-oriented approaches with SO, providing a biological role-oriented solution for the problem of the multiplex classification of genomic segments in genetics.

5.3 Toward further cooperation with domain ontologies

Transferring these merits into the OBO Foundry ontologies (Smith et al., 2009) is one of the major future issues for this study. The current version of our ontology is compatible with several domain ontologies. For example, the top level of molecular entity is fully compatible with the Chemical Entities of Biological Interest (ChEBI). However, open issues to build interoperability with the SO and Information Artifact Ontology (IAO) remain to be resolved.

In biological databases, genes and alleles are often instantiated at the level of an individual gene (e.g., *Gdf5* and *Gdf5*^{Rgsc451} allele). This issue is not well addressed in this paper. From an ontological viewpoint, the following possibilities can be suggested:

(a) They are instantiated from the view of a meta-model, as a rough equivalent of the "name" of class generated by the punning operation in OWL2.

(b) They are instances of the specifications of genes and alleles. They are described in Fig. 2A as "***specification of genomic segment***", which is a content of ***information for self-replication***.

We prefer (b) to (a). The ***specification***, as well as the "plan"**,** which is a specification of a sequence of actions mentioned elsewhere (Schulz et al**.,** 2011), is an instance of "realizable entity"**,** which is needed for the existence of designed entities, including genes. In the self-replication process, a genomic segment acts as a design plan for itself for its existence. Therefore, the ***specification of a genomic segment*** is an essential attribute of a ***genomic segment***.

In the SO, a gene is classified as a *biological region* defined as a "generically dependent continuant." This means that the SO also represents individual genes as instances of abstract things. To establish inter-operability with the SO and the IAO, further consideration of ***specification*** and its realization of designed entities is needed.

ACKNOWLEDGEMENTS

This work was supported by the Management Expenses Grant for RIKEN BioResource Center, MEXT.

REFERENCES

Eilbeck K., Lewis S., Mungall C.J., Yandell M., Stein L., Durbin R., Ashburner M. (2005) The Sequence Ontology: A tool for the unification of genome annotations. Genome Biology 6:R44

Griffiths, Paul E. and Karola Stotz (2007) "Gene", in Michael Ruse and David Hull (eds.), Cambridge Companion to Philosophy of Biology, Cambridge: Cambridge University Press, 85-102.

Hoehndorf R, Kelso J, Herre H. (2009) The ontology of biological sequences. BMC Bioinformatics. 18;10:377.

Hozo: http://www.hozo.jp/

Kozaki, K., Sunagawa, E., Kitamura, Y., Mizoguchi, R. (2007) Role Representation Model Using OWL and SWRL, Proc. of 2nd Workshop on Roles and Relationships in Object Oriented Programming, Multiagent Systems, and Ontologies, 39-46.

Mungall, C. J. Batchelor C. Eilbeck K. (2010) Evolution of the Sequence Ontology terms and relationships J Biomed Inform. 44(1), 87-93

Mizoguchi, R. (2004) Tutorial on ontological engineering - Part 3: Advanced course of ontological engineering, New Generation Computing, OhmSha & Springer, 22, No.2, pp.198--220.

Mizoguchi R. (2009). Yet Another Top-level Ontology: YATO, Proceedings of the 2nd Interdisciplinary Ontology Meeting, 2, 91—101.

YAMATO: http://www.ei.sanken.osaka-u.ac.jp/hozo/onto_library /upperOnto.htm

Schulz S, Brochhausen M, Hoehndorf R. (2011).Higgs Bosons, Mars Missions, and Unicorn Delusions:How to Deal with Terms of Dubious Reference in Scientific Ontologies, Proceedings of the 2nd International Conference on Biomedical Ontology, 183-189

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