Journal of Computer Science 1 (3): 445-449, 2005
ISSN 1549-3636

OntoCell: An ontology of Cellular Biology

Lynda DIB and Mohamed T Laskri
University Badji Mokhtar Annaba, Algeria

Abstract: This article presents OntoCell, a cellular ontology that we constructed and edited under Protege2000. It regroups and unifies the main concepts and relations related to the cell's structure and behavior. OntoCell has been validated by experts in biology (by the UMRS INSERM 514). Moreover, it will be validated in the context of the development of a multi-agent system simulating the behavior of a cellular population.

Keywords*:* Ontology, cellular biology, cell, concepts, relations.

INTRODUCTION

This work is inscribed in project which aims is the multi-agent modelisation and simulation of cellular populations. The aim is to propose a multi-agent platform in order to facilitate the modeling of cellular populations and the observation of their behaviors. The observed collective behaviors will be then confronted by the biologists to the observations and to the quantitative data results obtained in video-microscopy. This step allows to refine the biology models existed notably in tumoral invasion.

The multi-agent simulation of cellular populations offer a very powerful tool to study the behavior of cells and their interactions. However, the necessary models for the development of this simulation require the knowledge in biology which are very difficult to identify and to structure.

The first reflex when we speak about structuration and representation of data and knowledge is to think about data bases and their usual models (entity-association model, relational model, oriented object model…) and to thesaurus. Ontologies come to enrich and emphasis this modelisation of the domain and to offer the powerful tools of research and manipulation.

The cellular biology, which is a part of experimental sciences domain, is characterized by a biologic language using an extremely rich and very difficult vocabulary to manipulate by the none biologists. In order to automate the treatment of biologic information, firstly, it must conceive a formal model. This model must unify the vocabulary, define clearly concepts and underlying relations and offer a research mechanisms and an easy manipulation to use. The ontologies are a very powerful formalisms of representation knowledge domains as complex and rich that the cellular biology.

The aim of this article is to present OntoCell, a cellular ontology that we constructed in order to represent knowledge of cellular biology. Hence, OntoCell represents the concepts of cellulars bases, their components their behaviors and their interactions. It regroups and unify most concepts and relations linked to the cell and the milieu in which it evolves. It's constructed and edited under Protege2000 [17]. OntoCell is very rich, it regroups about fifty concepts. Otherwise, it is very easy to enrich it basing on more detailed study of cells and by exploiting the Protege2000 tool.

This article is organized as follows : In the first section, we present an art state of ontologies for the cellular biology domain. In the second section, we describe the methodology followed by the OntoCell development which is summarized by the four steps : the requirements analysis, conceptualization, formalization and maintenance.

Ontologies in Biology: The last years, many ontologies appeared in the biology domain. They have a common objective: to facilitate the division and the exchange of knowledge.

The first category regroups *Gene Ontology* [2]. This ontology is very pragmatic. It is dedicated to the product's annotation of gene among a vocabulary reference. It objective is to palliate terminology heterogeneity problems by furnishing a common control vocabulary facilitating the exchange of information. The terms of GO were defined from "Oxford Dictionary Biology" [26] and from "Swissprot" [3]. GO [2, 24] is a system combining three independent ontologies describing biological processes, molecular functions and cellular components.

The conception of ontologies of this first category doesn't always respect the definition of the ontology's reference (T.Gruber defines the ontology as "descriptions structured and formal of concepts of domain and their interrelations"), but they answer to the division and to the exchange problematic of Gene knowledge. These first ontologies demonstrated their utility, the essential concepts having been defined, related to each other and serving to the annotation of the partners data bases. Actually, they were in restructuration in order to be formalized, it is the case of Gene Ontology [10].

A second ontologies categories, such as *EcoCyc* [15, 16, 18] and *OMB* [20, 21, 22, 23], treat the key points in problematic related on the number and heterogeneity of biologic resource : definitions of the common vocabulary, consensual description of knowledge basis scheme proper to a theme, extraction information distributed on different sources, and definition of Web bioinformatics services. The EcoCyc *ontology* is described to cover the general knowledge. It is not adapted to the molecular biology because is too general. EcoCyc serves to visualize the biochemical reactions and the gene's disposition with the chromosomes. The treated concepts by this ontology are : the gene, the proteins, the small molecules and biochemical functions. It is presented to the biologists by using encyclopedic metaphor. It covers "E.coli genes", metabolism, regulation and traduction signal [5] . The *OMB ontology* (Ontology for the Molecular Biology) is elaborated by Schulze-Kremer, It models the molecular biology. In OMB the biological concepts are regrouped in three categories : the biological objects, the experimental procedures and the aspects in silico of biological molecular [19]. OMB is not rich (less deeper) because it treats less the biological concepts.

The ontologies of the last category combine two or several existing ontologies. For example, RiboWeb [6, 1] is a set of four ontologies allowing the stake in common of the Web resources dedicated to the ribosomes. This basis allows the comparison and the interpretation of data in order to establish new distributed models. Otherwise, *TAO* the ontology of the *Tambis* system [4, 5] has the objective to answer to the problematic raising the extraction of information from different sources. This ontology has a central role in the TAMBIS system. The interrogation of set of data extern basis is realized by using only one constructed request, by the user, via an interface, from defined concepts in Tao. Hence, the biologists don't need to learn the languages of the specific request, nor the scheme of all susceptible basis to be interrogated.

Since, these ontologies are well specific to their provided use (representation of biological knowledge for the study of the gene or of the protein), they treat the key points in the problematic linked to the number and the heterogeneity of the biological resources : definitions of the common vocabulary, consensual description of the scheme of the basis knowledge proper to a theme, extraction information dispatched on different sources, and definition of Web bioinformatics services, we propose, so, a new ontology OntoCell.

OntoCell is different to the other ontologies because it doesn't interest itself to the survey of the gene or proteins nor to give the detail of experiences employees to study the structure of the RNA/proteins complex, but it concerns the study of the behavior of cell's population. it is about the modeling of the cell, its components (its structure), its behavior in the environment, its adaptation and its change of state in one hand, and their interactions with other cells of its population on the other hand. This ontology is a hierarchical concepts, rich and full concerning the biological domain. These concepts were determined from different meetings of work and documents furnished by the biologists of the UMRS INSERM 514.

OntoCell :It exists a great number of methodologies for the development and the maintenance of the ontologies such as Tove [11, 12, 13], Methodology [8], One-To-knowledge [25] and KBSI IDEF5 [14]. A comparative study of these different methodologies has been realized by Fernández-Lopez, Gómez-Pérez and Juristo [9] who show that these methodologies adopt the same cycle of development:

- requirements analysis
- Conceptualization
- Formalization
- Maintenance.

OntoCell has been created while basing on the cycle of the development. The different steps are described in the following paragraphs.

Requirements Analysis : Beginning a development of an ontology, means to define its domain and its reach. OntoCell is conceived in the aim to represent the biological basis knowledge cellular. It represents the structures of the cells, their environment and their behaviors. Hence, the domain is clearly defined. One of the proposed methods to determine the reach of the ontology, consists to write a list of questions which the final knowledge's basis can answer, called questions of competence [12].

In the case of OntoCell, the objective is to modelize the cell (healthy or cancerous) in its environment. In fact, we have elaborated many questions of competences such as :

- What are the cell characteristics?
- What are the necessary resources to its survival?
- What are the characteristics of a cancerous cell?
- What are the differences between a healthy cell and a cancerous cell?
- What is the difference between a benign tumoral cell and a malignant cancerous cell?
- When the cancerous cell can migrate from its origin tissue?
- How cell passes from first phase of cancer to 4^{th} phase of cancer?
- When cancerous cells carry a metastasis?

From these questions and the different interactions with biologists, we have acquired the knowledge of basis in cellular biology and collected the most information and necessary documents to the conceptualization step.

Conceptualization*:* In this step, we begin the real creation of OntoCell. It is to introduce the concepts and the different relations. Hence, we were based on classical approach of information systems analysis 'Merise' [7]. We started to define the dictionary of the data basing on the different discussions with biologists also the documents that they furnished to us. Mostly, the dictionary contains the properties of the cellular biology domain. Table 1 gives the examples of properties of the OntoCell dictionary data.

Table 1. Example of OntoCell data.

Property of OntoCell	Significance	Example of values
CatCell	Category of Cell	Cell of the crystalline lens
StateCell	State of Cell	Cancerous, Quiescent
LifeCell	Length of Cell's life	One day
NameDeplComp	Name of Displacement Components	Lamellipode
StateJonc	State Junction	Destroyed
CatMol	Category of Molecule	Occludine, Catenine
SizeMol	size of Molecule (fuzzy value)	Large Molecule
CatFact	Category of Factor	Creatinin, Flumazenil
FactType	Factor Type	Liposoluble Factor
CatMolAdh	Category of Adherence Molecule	Integrine, Cadherine
StateAdhMol	State of Adherence Molecule	Activate
CatDissMol	Category of Dissociation Molecule	Protease
StateDissMol	State of Dissociation Molecule	Activate
TypeActiDissMol	Type of activation of Dissociation Molecule	Activate with neutral pH

The second step consists to search the functional dependences between the different properties listed in the dictionary. Table 2 represents few functional dependences of data in Table 1.

The functional dependences (FD) permit to construct the theoretical access structure (TAS). This TAS schematizes the links (FD) that exist between all the properties. Fig 1 gives the TAS a part of the dictionary OntoCell (see Table 1).

The conceptual model of data is easily deduced from the TAS. This model offers a representation of data, easily comprehensible, describing the system with the help of entities and their relations.

Fig. 2 gives the relation entity model corresponding to TAS of Fig. 1.

Table 2: Example of Functional Dependences.

CatCell →StateCell
CatCell → NameDeplComp
CatCell →LifeCell
CatCell → CatFact
CatMol → SizeMol
CatMol → CatMolAdh
CatMol → CatMolDiss
CatMol → CatFact
CatAdhMol → StateAdhMol
CatDissMol → StateDissMol
CatMolDiss → TypeActi DissMol
CatFact → FactType
CatDissMol + CatCell → StateJonc

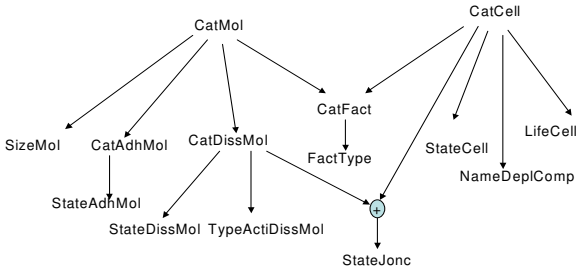


Fig. 1. TAS of Table 2.

An analysis of the different relations of relation-entities model allowed us to identify two categories of relations:

- General relations,
- Specific relations to cellular biology.

In what follows, we give two examples of relations: a general relation (isa) and specific relation (modified).

The relation *"isa"*

The isa relation conduct to the factorization of certain concepts in one generic concept and/or to specialization of generic concept (Fig.2). In the example of Fig 3, the Receptor is generic concept, but the Intracellular Receptor and Membranaire Receptor are a specific concepts. The is a relation represented by the arrows defines the hierarchical relation. Thus, Membranaire Receptor is a Receptor and Intracellular Receptor isa Receptor

Specific relations to cellular biology

A specific relation expresses the biologic dependence bond between concepts. Relations of dependences of the example given in Fig 4 are:

- Transforms "the Cell Transforms the Ligand",
- Secretes "the Cell secretes a Factor",
- Is in "the Cell is in the Position",
- Gives out "the Cell Gives out Components of displacement",
- Modifies properties "the Ligand Modifies properties of the Receptor",
- Ties to "the Ligand ties to a Receptor

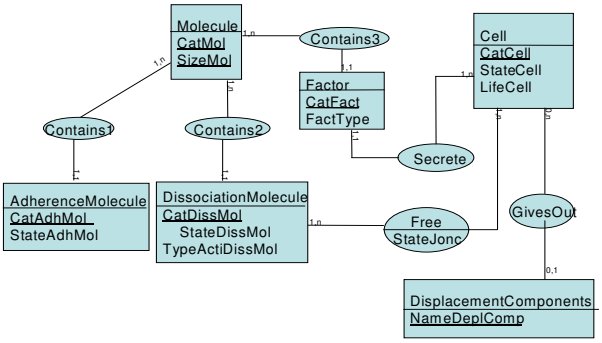
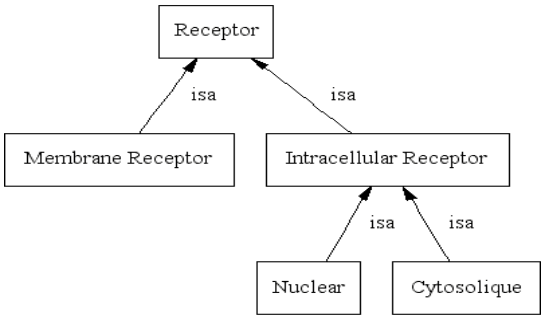


Fig. 2. The Conceptual Model of the TAS.



.

Fig. 3: Example Extracted from OntoCell representing a Hierarchical relation.

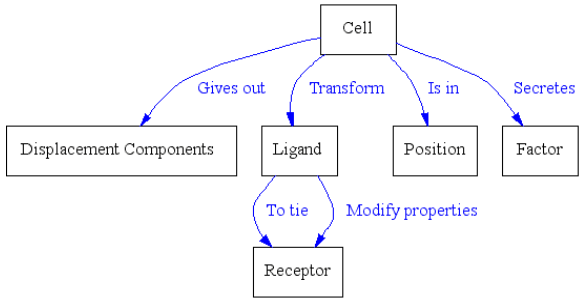


Fig. 4. Example extracted from OntoCell representing the relations specific to cellular biology.

Formalization: DAML (DRAPAAgent Markup Language) (http://www.daml.org/) is a language of an ontology description based on XML. It allows the division of the semantic. DAML can be associated to OIL (Ontology Inference Layer) (http://www.ontoknowledge.org/oil/) which is another language of description and inference on the ontologies and which takes support on the logics of description.
DAML+OIL is language chosen for the formalization of OntoCell. This formalization is realized by using the ontology's editor 'Protege-2000' [17]. This editor allow to construct an ontology for a given domain, to define the input forms of data, and to reach data with the help of these forms under the form of instances of this ontology.
Maintenance: The step of conception validation allowed to construct a core of OntoCell (Fig 5). This core can be easily putted at point, when the changes were suggested by the evaluation (with experts of the domain) or by the introduction of new objectives. This step allows to enrich and to correct the OntoCell ontology.
Since the cell lives in environment, we enriched the first core by adding the presence of molecule (Fig 6).

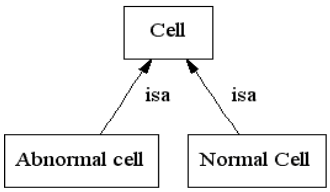


Fig. 5. OntoCell describing the cell. Here we have few examples of OntoCell enrichment.

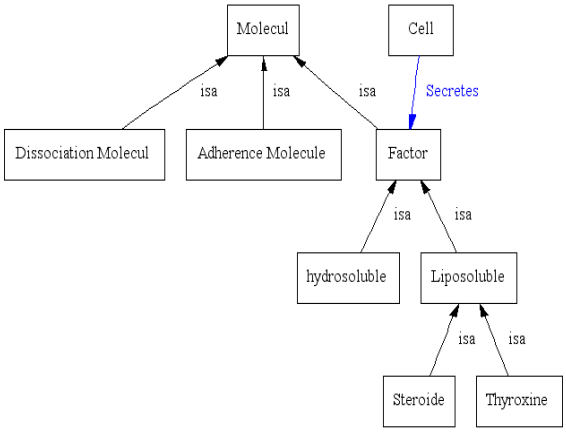


Fig. 6. OntoCell enriched in its environment.

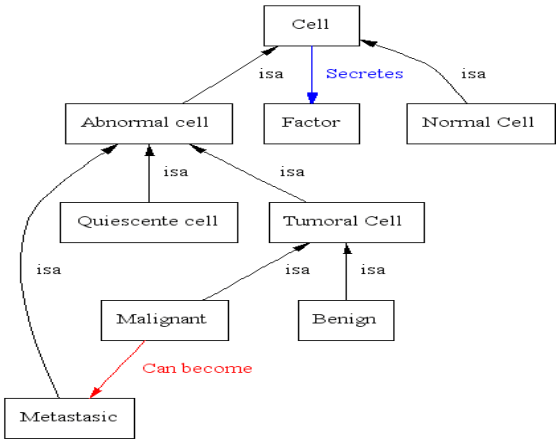


Fig 7: OntoCell enriched according to a more precise objectives.

An important problem for the biologists, in different domain of study and specially that of tumoral invasion and that of cellular migration. The OntoCell has been enriched in order to incorporate this particular aspect (Fig 7).
The enrichment of OntoCell in this example is illustrated by a description more detailed of Abnormal Cell: Abnormal cell is either a Quiescent Cell or Tumoral Cell and Tumoral Cell is Malignant Tumoral Cell or Benign Tumoral Cell.
Actually, OntoCell contains fifty concepts of cellular biology domain. It can be enriched according to future objectives such that the communication between different types of cellular populations, the union of two different cellular populations,…etc.

CONCLUSION

In this article, we presented the OntoCell ontology that we constructed. It allows to regroup the most of the concepts used in the domain of cellular biology. The objectives is to offer an ontology as complete as possible, coherent and easy to manipulate can be exploited in the development of system of prevision simulation of the cellular behavior.

We described the methodology of OntoCell development which is passed by the requirements analysis, the conceptualization, the formalization, and the maintenance. The explicit and formal representation of OntoCell given by the ontological language DAML+oil is important in order to establish and to exploit the system of previewed simulation. This formal representation is edited by the *Protege-2000* tool .

Actually, OntoCell contains fifty concepts. We judge that those accepted are the most important for the cell and its behavior in its milieu of life.

It has been validated by experts of the biological domain (by the UMRS INSERM 514), nevertheless it can be enriched by other concepts specific to biologic and specific objectives waited in function of the needs of our project.

REFERENCES

1. R. Altman, M. Bada, X.J. Chai, M. Whirl Carillo, R.O. Chen, and N.F. Abernethy. RiboWeb: An Ontology-Based System for Collaborative Molecular Biology. IEEE Intelligent Systems, 14(5):68-76, 1999.
2. M. Ashburner, A. Catherine. Go consortium. Gene ontology: tool for the unification of biology. Volume 25 no. 1 pp 25-29, 2000. http://genome-www.stanford.edu/GO/.
3. A. Bairoch, R. Apweiler. The SWISSPROT protein sequence database and its supplement TREMBL. Nucleic Acid Research, 28, 45-48,2000.
4. P.G. Baker, A. Brass, S. Bechhofer, C. Goble, N. Paton, and R. Stevens. TAMBIS: Transparent Access to Multiple Bioinformatics Information Sources. An Overview. In Proceedings of the Sixth International Conference on Intelligent Systems for Molecular Biology, pages 25-34. AAAI Press, June 28-July 1, 1998 1998.
5. P.G. Baker, C.A. Goble, S. Bechhofer, N.W. Paton, R. Stevens, and A Brass. An Ontology for Bioinformatics Applications. Bioinformatics, 15(6):510-520, 1999.
6. R.O. Chen, R. Felciano, and R.B. Altman. RiboWeb: Linking Structural Computations to a Knowledge Base of Published Experimental Data. In Proceedings of the 5th International Conference on Intelligent Systems for Molecular Biology, pages 84-87. AAAI Press, 1997.
7 D. Donisi the main thing of Merise. Eyrolles edition 1998.
8 M. Fernández-Lopez, A. Gómez-Pérez, N. Juristo. METHONTOLOGY: From Ontological Art Towards Ontological Engineering. Symposium on Ontological Engineering of AAAI. Stanford (California). March 1997.
9 M. Fernández-López. Overview of Methodologies For Building Ontologies. IJCAI'99.
10 Gene Ontology Consortium 2004. Te Gene Ontology (GO) database and informatics ressource. Nuleic acids Res. 1,32 Database issue : D258-61.
11 M. Gruninger, M. S. Fox. The design and Evaluation of Ontologies for Entreprise Engineering. In Workshop on Implemented Ontologies, European Conference on Artificial Intelligence 1994, Amsterdam, NL.
12 M. Gruninger, M. S. The Role of Competency Questions in Entreprise Engineering. IFIP WG5.7 Workshop on Benchmarking – Theory and Practice, Trondheim, Norway.
13 M. Gruninger, M. S. Fox. Methodology for the Design and Evaluation of Ontologies. In: Proceedings of the Workshop on Basic Ontological Issues in Knowledge Sharing, IJCAI-95, Montreal.
14 KBSI 1994. The IDEF5 Ontology Description Capture Method Overview. In : KBSI Report, Texas,1994.
15 P. Karp, S. Paley. Integrated Access to Metabolic and Genomic Data. Journal of Computational Biology, 3(1):191-212, 1996.
16 P. Karp, M. Riley, S. Paley, A. Pellegrini-Toole, M. Krummenacker. EcoCyc: Electronic Encyclopedia of E. coli Genes and Metabolism. Nucleic Acids Research, 27(1):55-58, 1999.
17 The Protege Project. http://protege.stanford.edu.
18 M. Riley. Functions of the gene products of Escherichia coli. Microbiological Reviews, 57:862-952, 1993.
19 S. Saha. Go consortium. Creating the gene ontology resource: design and implementation. Genome Research, 2001 Aug;11(8):1425-33.
20 S. Schulze-Kremer. Adding Semantics to Genome Databases: Towards an Ontology for Molecular Biology. In Proceedings of the Fifth International Conference for Intelligent Systems for Molecular Biology Conference, pages 272-275. AAAI Press, Palo Alto, 1997
21 Schulze-Kremer S. Integrating and Exploiting Large-Scale, Heterogeneous and Autonomous Databases with an Ontology for Molecular Biology. In R. Hofestaedt and H. Lim, editors, Molecular Bioinformatics, Sequence Analysis - The Human Genome Project, pages 43-56. Shaker Verlag, Aachen, 1997.
22 Schulze.-Kremer S. Ontologies for Molecular Biology. Pac.Symp.Biocomp., pp 695-706.
23 S. Schulze-Kremer. Ontologies for Molecular Biology. In Proceedings of the Third Pacific Symposium on Biocomputing, pages 693-704. AAAI Press, 1998.
24 A. D. Smith. Oxford Dictionnary of Biochemistry and Molecular Biology. Oxford University Press,1997.
25 Y. Sure, A Tool supported Methodology fo Ontology-based Knowledge Management. Dans :The Ontology and Modeling of Real Estate Transactions in European Juristictins E. Stubkaer (ed), International Land Management Series, Ashgate,2002.
26 M. Ushold, M. Gruninger. Ontolgies : Principles Methods and Applications. Knowledge Engineering Review, 11 (2), 93-155, 1996.