Orienteering Tools: Biomedical Research with Ontologies

Federico Boem †
federico.boem@gmail.com

ABSTRACT

Biomedical ontologies are considered a serious innovation for biomedical research and clinical practice. They promise to integrate information coming from different biological databases thus creating a common ground for the representation of knowledge in all the life sciences. Such a tool has potentially many implications for both basic biomedical research and clinical practice. Here I discuss how this tool has been generated and thought. Due to the analysis of some empirical cases I try to elaborate how biomedical ontologies constitute a novelty also from an epistemological point of view.

keywords: bio-ontologies, biomedical research, scientific practice.

Nel suo profondo vidi che s'interna,
Legato con amore in un volume,
ciò che per l'universo si squaderna:
sustanze e accidenti e lor costume
quasi conflati insieme, per tal modo
che ciò ch'i' dico è un semplice lume.

Dante, *Paradiso*, Canto XXXIII, vv. 85-90.

1. Introduction

If someone searched for the term 'ontology' on Google he/she could be surprised to realize that the first entries mainly refer to *applied ontology*. In this battle for notoriety, 'ontology' in a more traditional and philosophical sense is defended just by Wikipedia and a few of other websites. While

† University of Milan, Italy.

philosophical ontology was devoted to pure speculation, engineers and computer scientists revitalized such a notion in the light of its possible applications. Indeed, in modern computational jargon a *computational ontology* (or applied ontology) is a way to model and represent a domain of interest or a particular area of knowledge so that a computer can process it. As Gruber pointed out (Gruber 2009) "an ontology specifies a vocabulary with which to make assertions, which may be inputs or outputs of knowledge agents (such as a software program)". Here lies the difference. If *philosophical ontology* is pursued as a way to establish on pure speculative ground 'what there is' or the fundamental entities or things of the world, *applied ontology* is a subfield of informational research devoted to knowledge representation and data integration. To make a slogan from 'ontology' we came to 'ontologies'. Again, as Gruber writes "ontologies are typically specified in languages that allow abstraction away from data structures and implementation strategies; in practice, the languages of ontologies are closer in expressive power to first-order logic than languages used to model databases" (Gruber 2009). In other words, ontologies constitute a tool that allows comparison among data that were originally produced and stored in different manners. In addition, ontologies are conceived as the mode to translate a specific knowledge at a certain level of description to other levels. This is why ontologies are also said to be the "semantic level" of scientific modelling.

Biomedical research is one of the leading areas of inquiry for the implementation and application of these semantic instruments. *Bio-ontologies* (as they are called) are now proliferating in the management of many biological databases. Among them, the *Gene Ontology* (from now on GO), developed by the Gene Ontology Consortium, represents a promising project greatly employed by many different institutions and laboratories in all the life sciences. The semantic dimension of this enterprise is clear in its own mission. The aim of the Gene Ontology project is to provide a representation of the features of gene products across different species and databases through a controlled vocabulary of different "biological categories".

One may wonder whether such a tool could be relevant for clinical purposes. Indeed, GO features and applications seem to be closer to basic biological research than medical practice. However, in the perspective of a translational research agenda, a tool like go might be extremely useful. As a matter of fact, the implementation of computational ontology categorizations and classifications affect the way biological knowledge is (and will be)

represented thus influencing the way scientists and clinicians conceive and define physiological malfunctions and diseases. Because of that, computational ontologies constitute a promise potentially revolutionary for all biomedical practices, *from bench to bed*.

2. GO: an orienteering tool for biomedical research

Gene Ontology is probably the most famous ontological initiative developed for biological research. Gene Ontology aims to provide a standardized representation of gene products' features across different species and databases. GO actually covers three domain ontologies which are called *Cellular Component* (the parts of a cell of its extracellular environment), *Molecular Function* (the basic activities of a gene product) and *Biological Process* (the set of molecular events characterized by clear beginning and end).

GO *terms* describe gene product characteristics in a single, computationally controlled way, in order to provide a common format. Each GO term (fig.1) has a specific name which designates it and which can be a single word or an expression (*e.g.*: apoptotic process), a unique alphanumeric identifier (*e.g.*: GO:0006915), a definition (see the note)[1] with references, and the ontological dependence that indicates the domain to which it belongs to (*e.g.*: Biological Process).

fig.1 (taken by QuickGO, https://www.ebi.ac.uk/QuickGO/)

Each GO term has then a set of defined relationships (*e.g.*: *is_a*, *part_of*, or *positvely_regulates* etc.) towards one or more terms in the same domain, and

[1] "A programmed cell death process which begins when a cell receives an internal (*e.g.*: DNA damage) or external signal (*e.g.*: an extracellular death ligand), and proceeds through a series of biochemical events (signaling pathways) which typically lead to rounding-up of the cell, retraction of pseudopodes, reduction of cellular volume (pyknosis), chromatin condensation, nuclear fragmentation (karyorrhexis), plasma membrane blebbing and fragmentation of the cell into apoptotic bodies. The process ends when the cell has died. The process is divided into a signalling pathway phase, and an execution phase, which is triggered by the former"

sometimes in other domains. The GO terminology is designed to be species-neutral, in order to be exploitable from prokaryotes to eukaryotes and from single to multi-cellular organisms.

GO annotation is "the practice of capturing the activities and localization of a gene product with GO terms and it provides references and indicates what kind of evidence is available to support it" (GO website - http://www.geneontology.org/). Annotations are created on the basis of observations of the individual occurrences (*i.e.* the instances) of the type under examination. Hidden in this scientific and technical presentation, the philosopher may recognise the Aristotelian mark of such an endeavour. Indeed, while GO terms stand for types, GO annotations are singular evidences (obtained through experimental observations) that instantiate the term of relevance. Here lies the Aristotelian legacy. Knowledge, biological knowledge, belongs to *universals*. However it is possible to get to the universal through the *particular*. GO annotations display the gene product (*e.g.* PB1-F2 protein), the relevant GO terms involved (*e.g.* apoptotic process), the reference which provides ground for such an annotation (*e.g.* the Gene Ontology Database references), the type of scientific evidence that supports the annotation (*e.g.* Inferred from Electronic Annotation) and finally the author and the date of the annotation itself.

It is clear that the choice of the three domains is also motivated by reasons of convenience. In other words, since GO is meant to provide a semantic representation of knowledge in use for molecular biology, the conceptual framework adopted clearly refers to the way molecular biologists pursue their experimental work, display their information and conceive explanations. This illustrates why GO is built to present *terms* and *annotations* according to a mechanistic description of molecular events. Indeed, GO is a technical tool, not a metaphysical device. Its application reveals the reason behind the terminological choice. However, such a choice, given the scope and the hope for generality of GO, cannot be grounded just on logical consistency and empirical adequacy. Being a tool of knowledge-capture and representation, GO terms must satisfy the needs and the desiderata of the scientific community. Accordingly, the process of *curation* is the production of annotations on the basis of findings retrieved from experimental work. Thus, since the activity of curation requires a deep scrutiny of the relevant literature, it is important (no less than obvious) that curators possess a robust expertise in

the related field. Normally, annotations are created through a procedure that requires several steps.

The primary aim of GO annotation is to create annotations based on findings obtained from experiments on related organisms. However, information coming from different model organisms or by sources other than experiments (as sequence information in the genome browser) is also taken into account. Thus, the annotation file provides a way to discriminate the sources of annotation and to filter out what is not considered important by the researcher. As in a map, the single scientist can highlight this or that feature, remove or add elements, in order to orientate himself/herself in the topic.

The second step consists in linking the information captured by the annotation within the appropriate term. Some factors should be taken into account. Indeed, the kind of experiment itself shapes the nature of evidence that can be obtained and sets up the resolution and the quality of results. "For example, cell fractionation might localize molecules of a protein to the nucleus of a cell, but immunolocalization experiments might localize molecules of the same type of protein to the nucleolus of a cell. *As a result, the same gene may have annotations to different terms in the same ontology because annotations are based on different experiments*" (Hill *et al.* 2008, emphasis is mine). Last, but not least, annotation procedures are usually verified for their consistency. In doing so, both computational/logical tools and domain experts are involved. To further develop this aspect, it is possible to individuate distinct *epistemic moments* according to which annotations are created. First, information coming from scientific publications is captured, extracted and abstracted by annotators and then condensed into a unique *semantic designation*, according to the rules of term composition and the consistency of GO. Thus, even if most annotations are manually operated, the process is reviewed both by GO curators and by automatic reasoners. Such a product must finally face the judgment of the scientific community *i.e.* the experts of the field. Obviously, the process of annotation is not a static given. Both GO terms and annotations are in constant evolution and growth since they map the current state of the research. GO updates its content according to scientific debates and it is even able to display the disagreement among experts (*e.g.* the NOT annotation). For example, the vast part of terms and annotations pertaining to the range of phenomena which include the death of a cell are undergoing a revision due to the very latest scientific finding in the field (see for instance Kaczmarek, Vandenabeele, Krysko 2013; Christofferson and Yuan 2010). Gene

Ontology then, is not dictating, in a purely top down fashion, which terms are right or not for the research, but it is rather mapping the current use of *scientific vocabulary* trying to standardize it. However, such a feature shows why GO is also normative too. A tool like GO is a map of knowledge but it is also a way to standardize practices. Accordingly, GO presents a form of objectivity that, following Alberto Cambrosio's suggestions (Cambrosio, Keating, Schlich and Weisz 2006), can be called *regulatory objectivity*. This kind of objectivity "is based on the systematic recourse to the collective production of evidence. Unlike forms of objectivity that emerged in earlier eras, regulatory objectivity consistently results in the production of conventions, [...] most often arrived at through concerted programs of actions" (Cambrosio, Keating, Schlich and Weisz 2006, p 189). Indeed, both GO structure and its practical choices, heavily rely on collective concerted actions among different 'players' such as database curators, biologists, other researchers, computer scientists etc. A map at first glance, is a standardized representation of different elements under a shared, common framework. Standardization implies also agreed conventions both in the construction and in the interpretation of what is represented. In this sense a simple description might become a norm. It might be useful to use a metaphorical image to explain this epistemic passage from descriptive efforts to normative ones. In the field of western jurisprudence, it is possible to individuate two main legal systems that have been developed differently. These two systems, known as civil law and common law are primary distinguishable because of their different historical genesis, thus affecting the countries in which they are applied, and then because they embed different conceptions concerning the nature of jurisprudence itself and thus the nature of what a norm is. Civil law, preponderant in all European countries (with the exception of the UK) is rationalized in the framework of the ancient Roman law system, further developed by the code of Justinian and finally systematized by the Napoleonic code and the German BGB (Bürgerliches Gesetzbuch). Accordingly, civil law is based on the written codification of general norms and principles that constitute the primary source of law. Such systematic collections of principle, inform both citizens about the behaviour they should have and judges/magistrates on how interpret the law itself. Therefore, civil law establishes and explain, from above, principles, rights and duties and how the legal system works. Civil law founds and unifies jurisprudence by acting as a sort of a top down theoretical framework, thus determining what is consistent

with its principles and norms, and rejecting what is contrary to it. Common law is instead the system adopted by the UK and by most of the actual and former colonies/possessions of the British Empire. Common law has its raison d'être on precedents, praxis and routines of conduct rather than formal codifications of norms and principles. Common law is then systematizing and ordering the customary practices in a more general and coherent form. Thus common law founds and unifies the law not by dictating an overarching structure from above, but rather by conforming and standardizing the practice in a consistent way.

Bearing in mind this distinction, I would argue that the way GO performs its unification power, is closer, metaphorically speaking, to common law than to civil law. GO is unifying biological knowledge in a novel way that is different from theoretical unification but nevertheless practically useful and robust. Indeed, as the UK is a solid democracy without having a proper constitution, biology can be unified, in this sense, without having a general overarching theory.

The standardization created through GO affects the way information in databases and other electronic resources is presented. By expanding the experimental context, ontologies allow not just the use but especially the re-use of the represented knowledge. Thus a new lab, in the definition of its standards and terminology, would not start from scratch, following arbitrary criteria, but it would rather rely on a body of knowledge which is more and more organized and unified (I will come back to this point later in the study).

The structure of GO terms and relations among them is also displayed graphically (see an example in fig. 2)

This shows how GO is a kind of epistemic map very well. A map of knowledge. Indeed, each chart is highly interactive. Each term can be *opened* and further examined. *Parent* and *children* terms are thus shown along with related gene product annotations. All this information is literally mapped into a wider context. Therefore, it is possible to navigate GO through its terms and relations, check the gene products involved in certain phenomena and link them with other area of research out of the given experimental context. Moreover, GO is a live map. As already mentioned, the content of GO is not static but rather it tracks the changes and developments within the scientific community.

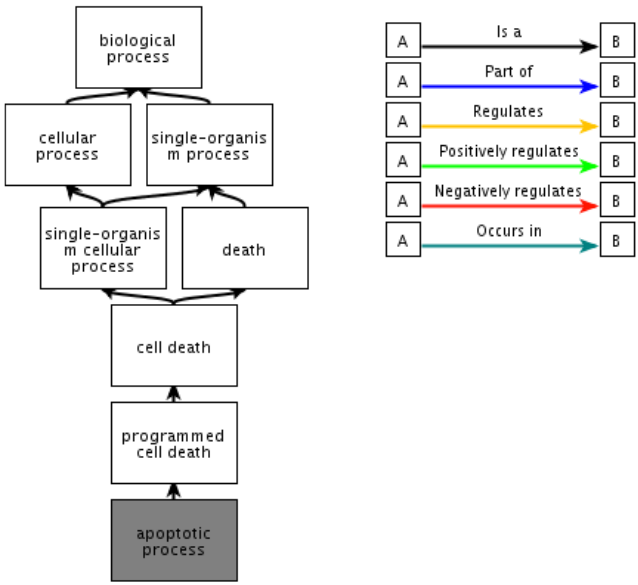

fig.2 (taken by QuickGO)

3. Analyses with GO and their Epistemic Meaning

Let us now consider one of the most common operations available with GO: the *enrichment analysis*. This type of investigation may allow scientists to map and evaluate possible scenarios given specific experimental conditions. For instance for "a set of genes that are up-regulated under certain conditions, an enrichment analysis will find which GO terms are over-represented (or under-represented) using annotations for that gene set" (http://geneontology.org). By doing this, researchers can characterize that set of genes under a common functional profile, revealing important features of the underlying biological phenomenon. The output of such an analysis is then an ordered list of GO terms, with the related p-value. For example, due to high-throughput analysis, it is now possible to compare the gene expression profiles of a healthy tissue with the cancerous one. By examining the semantic discrepancies resulting from the analysis it is possible to furnish indications about the differences on the hidden biological mechanisms. This kind of work can also be done

pursuing different research strategies. On the one hand it is possible to check which terms are significant in a particular set of genes or the other way round, that is to check if a biological phenomenon (such as apoptosis) is over-represented (or under-represented) in a particular set of genes. Several tools (such as DAVID, Panther, Ontologizer, Onto-Express) have been developed to perform this type of investigation deploying different statistical methods and different databases sources. This means that researchers should ideally perform different functional profiling adopting different tools before interpreting their experimental results.

Another kind of common analysis with GO consists in the prediction of putative gene function. "Typical approaches tend to be variations of the same theme: genes are grouped together on the basis of some criterion such as similar gene expression or through a protein–protein interaction network. Enrichment of GO terms is detected by methods such as those described above, and the uncharacterized genes are presumed to be involved in the same biological processes as the genes with which they are grouped" (Rhee, Wood, Dolinski and Draghici 2008). It is clear that such an operation pays close attention. By propagating a gene function just on the basis of annotations that are neither manually verified nor experimentally validated can lead to many false positives. On the other side "[g]ene functions can also be inferred from GO annotations without the need for a prior gene grouping, for instance, on the basis of a semantic analysis of the gene function association matrix. This type of analysis relies on capturing the implicit dependencies that might be present between genes" (*ibid*).[2]

By examining both the epistemic reasons for its implementation and the type of analyses provided by GO, I would argue that such a tool resembles some features of a *model* but nevertheless constitutes something new in the epistemological scenario. Not entirely a theory, more than a model (however structurally similar to it), my point is that GO efforts constitute a novel category within the epistemic repertoire. Indeed, my central claim is the knowledge provided by GO, should be seen as a more or less effective tool

[2] On this aspect some scholars (*i.e.* Noble 2008) have raised some critical points about the power and the limits of GO, by claiming that describing biological phenomena just in terms of their related, involved, gene products will miss the higher-level insight. In order to answer to this issue, many new projects in this field are devoted to the development of other ontologies, integrated with GO, but concerning different aspects and levels/granularities of biological phenomena.

through which we can discriminate, among an enormous amount of data, a *convenient* way of organizing those empirical results which were at the basis of the GO analysis. In this perspective GO is a very peculiar map. It is a unique, *sui generis* instrument. It is an *orienteering tool* for biological research. The view that GO is an *orienteering tool* means that it is an instrument through which scientists can *map* their data on a wider context and then, thanks to this, elaborate new experimental strategies. GO is truly a map for making the conceptual content of a particular experimental condition comparable across different research contexts. Such a map is essential not as a way to confirm experimental results but as a way to compare experimental results with the theoretical background (the so called 'big picture')

Therefore, it should be clear that GO does not provide any, *stricto sensu*, discovery, since a map cannot show what is not mapped. Precisely because GO is a map of what is already known, it is not, *per se*, a discovery tool. However, such a map can allow the interpreter to observe connections that are invisible without the map itself. This is exactly what GO can do. GO is not an instrument to make discoveries, but it rather creates the conditions to make discoveries.

4. Doing Science with GO: the Use of the Orienteering Tool

In this section I will analyze how GO constitutes a driving force for contemporary biomedical research. In doing so, along with the brief examination of some cases, I will focus on a particular example, a recent article, "*7q11.23 dosage-dependent dysregulation in human pluripotent stem cells affects transcriptional programs in disease-relevant lineages"* published on *Nature Genetics* (Adamo, Atashpaz, Germain *et al.* 2015). The article provides an excellent, paradigmatic, case of how a tool like GO changed the practice of current scientific research in the life sciences. Indeed, in the paper, the experimental strategy, the methods and the rationale are all embedded in a map thinking framework. However, there is more. Such an article really shows a peculiar and distinct, way of doing science. This way is not simply ascribable to a naïve dichotomy between hypothesis-driven vs. data driven. It is rather a complex combination of the two, in which the knowledge coming from databases, exploited by a tool like GO, drives the other components of scientific efforts by exchanging the epistemic primacy and priority of exploratory experiments with the navigation in the data sea.

The aim of the article is ambitious. Summing up, its purpose is to increase the reliability of *induced pluripotent stem cells* (iPSCs) as models for diseases. iPSCs are a type of cell, that has undergone molecular reprogramming, presenting, *bona fide*, the features of embryonic stem cells. iPSCs constitute a promising and trendy sector of biomedical research since they challenged the idea that specialised cells are inescapably committed to their fate. In 1960s John Gurdon (Gurdon, 1962) has shown how somatic, differentiated cells could be turned back into their embryonic state by transferring nuclei of epithelial cells into enucleated eggs of a frog. Recently, Shinya Yamanaka (Takahashi and Yamanaka, 2006) has demonstrated that, without any nuclear transfer, he could reproduce Gurdon's results by exposing differentiated cells to specific factors which eventually turned (*i.e.* reprogrammed) those committed cells back into their pluripotent state (*i.e.* iPSCs). Such a discovery granted the Nobel Prize in 2012 to both of them, and opened the door to the implementation of iPSCs in many areas of biomedical research. One application is precisely *disease modelling*. The potential of such an approach is that iPSCs should allow a better analysis of the complex picture of pathogenic drives in a developmental context via molecular approaches. In other words, iPSCs could shorten the gap between clinical and research contexts by permitting the track of the consequences of genetic alterations through the cell development thus providing hints of clinical relevance. In order to exploit such a potentiality, it is fundamental to provide an answer to, at least, two problems. On the one hand it is central to determine how much genetic alterations, in early developmental phases, are indicative over related pathological conditions and their molecular pathways (*i.e.* to observe the onset of the disease in a preclinical phase otherwise not detectable). On the other hand, it is crucial to establish how much iPSCs modelling is apt to identify these pathways. Moreover, this approach to disease modelling could, in theory, provide suggestions on relevant molecular mechanisms from the point of view of future therapeutic implementations.

The authors of the article addressed then these issues by investigating two related genetic syndromes produced "by symmetrical copy number variations (CNVs) at 7q11.23[3] involving, respectively, the loss and gain of 26-28 genes: Williams-Beuren syndrome (WBS) and Williams-Beuren region duplication

[3] a genomic region on human chromosome 7

syndrome [...] that includes autistic spectrum disorder (7dupASD)" (Adamo, Atashpaz, Germain *et al.* 2015, p132). The main idea is that, thanks to iPSCs modelling, it would be possible to scrutinize such a biological symmetry (let us consider that genomic inverted alterations are mirrored by specular, behavioral phenotypes) starting from the 'origin' or the stem like state. Thus, due to a collaboration with clinicians, researchers were able to have samples from a cohort of patients resulting in 4 different genotypes: the WSB typical deletion, WSB atypical deletion (a shorter one, in terms of base pairs, and less frequent), the control case and the 7dupASD duplication. Skin fibroblasts have been reprogrammed via synthetic mRNA encoding different pluripotent factors, thus developing a total of 27 iPSC lines. Successively, the pluripotent state of these cells has been confirmed through transcriptomic analysis[4]. Such a step is a further confirmation of the standardizing power of databases for research. Indeed, the pluripotent state has been determined as such since the transcriptomic profile has been compared and matched with published datasets. RNA-seq (roughly, the sequence of the transcription) and Nanostring quantification (another methodology to assess gene expression) also confirmed that gene expression mirrored gene dosage and, again via *database consultation*, scientists were able to verify that two proteins, GTF2I and BAZ1B, are "encoded by genes associated with key traits of WBS and 7dupASD" (Adamo, Atashpaz, Germain *et al.* 2015, p133). In particular, GTF2I protein level correlates with gene dosage. Next, differential expression analysis between distinct genotypes has been conducted by RNA-seq profiling of iPSCs, then comparing the results against some control cell lines. This is again through the use of databases, which provided the reference context on which to give sense to experimental results. A pairwise comparison of the three genotypes (Williams-Beuren syndrome vs. Control Group, Williams-Beuren syndrome vs. 7dupASD and 7dupASD vs. Control Group) revealed 757 differentially expressed genes (DEGs). Finally, a GO term enrichment analysis of the union of DEGs has been performed.

At this stage Gene Ontology comes explicitly into play. However, I believe that GO rationale has driven much part of the experimental design and has highly influenced the earlier steps of such a research study. In other words, my argument is that GO is at the basis of the main heuristic strategy of the entire

[4] and also by IF (immunofluorescence) of pluripotent factors

study. In order to justify my claim, before discussing the use of GO and its results, I will go back to the previous phases of the experimental strategy in order to detect and unveil the role of GO.

Above all, the main strategy of the study rests on a well-designed, combined use of iPSCs and Gene Ontology. As already mentioned, iPSCs constitute an excellent surrogate of embryonic stem cells in terms of pluripotency and stem like features. Indeed, whereas the process of reprogramming had been conducted effectively, it would be virtually impossible to distinguish iPSCs from ESCs (embryonic stem cells). Pluripotency is defined as the capacity of a cell to differentiate itself into any other cell of an organism (see for instance the Oxford Dictionary of Biology, 6th Ed., 2008). At a molecular level, cell types are determined by peculiar gene network interactions. Being pluripotent means thus that a cell shows a particular molecular signature established by the modality of genes' activity. Indeed, since the genome of any cell of an organism is almost the same (there are some exceptions, but it is not fundamental to discuss this point here), the differences among cell types and states should be mainly attributed to the way genes are differentially expressed and regulated. Thus the pluripotent state (as any other cell state) is essentially related to epigenetics, or how different parts of the genome are alternatively transcribed, silenced and modulated.

In the case discussed here, the creation of iPSCs lines from patients affected by WBS and 7dupASD syndromes, can potentially allow scientists to obtain specific cell types (such as neurons) for further experiments. More than a joke, this could mean that it would be possible, in theory, to have a "brain in a dish" (see for instance Shen 2013). However, the authors of the article do not pursue that path (although it is possible that they will do in the future). Why is it so?

First, the production of specific cell differentiated lines is not straightforward. Both reprogramming and transdifferentiation (the artificial induction of a somatic cell to commit itself to another cell state) are not an easy task to perform. Due to technical difficulties some cell types are either almost impossible to obtain or the efficiency of the procedure is so low as to be useless (see for instance Hanna, Saha and Jaenisch 2010). Second, the materiality of somatic cell lines does not ground, *per se*, a better explanatory framework. Indeed, cell cultures do not constitute a reliable model in virtue of simple similarity. Moreover, given the complex nature of both syndromes, it would be very difficult to assess which neurons (among different types) will play a role,

and how they do so, in the disease. In addition, it would also be very troublesome to reproduce the material structure of relations of a brain, just through neuronal cultures. However, as explained in the previous chapter, this does not prevent a thing such as a cell from being a good model. Every scientist is aware that a bunch of cells does not accurately portray all the features the related tissue and organs. Still, as also shown in the previous chapter, this does not prevent a thing such as a cell from being a good model. But a good model for what?

The choice to focus on iPSCs is motivated precisely because they can provide a better model, compared to cultures of differentiated cells, for developmental conditions, given the implementation of certain type of analysis. It is also, again, a matter of style. Indeed, the types of evidence obtained through empirical experimentation are epistemically different from those coming from computational approaches. Certainly, the material production of distinctive cell types (*i.e.* neurons) is not mutually exclusive with bioinformatics work. On the contrary, they are complementary. By this I mean that computational approaches should not intended as a way to replace traditional experimental work. Different approaches can be used according to distinct interests and, in particular, according to the kind of results are searched. This is because different research strategies will prefer some types of evidence over other types. In our example, the production of specific cell lines, with no other indication, could have been potentially uninformative. On the contrary, the adoption of a tool like GO, allows scientists to globally map a sort of 'cell differentiation process *in silico*', thus suggesting what to look at in further experiments. This is possible because of the combination of the features of iPSCs and ontologies. As already said, within iPSCs there is all the potential of developing every cell of the organism. This means that iPSCs, given the genetic nature of the diseases taken into account, can contain, virtually and *ab origo*, all the relevant elements that could affect the molecular phenotype of interest (and hopefully suggesting therapeutic interventions in the clinical setting). On the other hand, GO, being an updated, global map of biological knowledge, allows comparing local findings with those ones coming from other experimental settings and to situate them into a wider picture. GO permits then to computationally explore the space of possible relations of different cell lineages through the comparison of given samples against all the relevant data stored in databases. Therefore, the possibility granted by GO, shapes the type of scientific strategy. A strategy that is: first, making a map.

Let us examine the reason why the construction of such a map is possible and probably, needed. First the kind of data. This type of science is heavily based on *omics*. Normally, transcriptional profiles show the global set of all RNA transcripts of a given genome (under specific conditions). However, their analysis affects cell populations rather than single cells. As a matter of fact, no cell behaves exactly as others, even if it is of the same type, in the same context. Uniqueness and intrinsic variation are indeed features of biological objects since 19th Century natural history. This means that, by performing transcriptomic analysis, scientists normally privilege the understanding of the average behavior rather than the detailed (hopefully mechanistic) description of single cell behavior and its fluctuations and relation with the other cells[5]. Transcriptional profiles are indeed general, cell-group behavioral maps. Certainly a map of this kind misses something. Tiny differences will be neglected and 'absorbed' by the background. This is not a problem. As a matter of fact, a map that is as detailed as the object it represents, is basically useless. Indeed, when the authors of the paper have identified 757 DEGs they did not care (for that moment) about *how* (i.e. which mechanism was responsible for it) these genes were differentially regulated (probably many genes are altered in different ways, one from another). Their concern was about *where* this distinct regulation happened. By that I mean that scientists have looked at the "number and distribution of DEGs across the comparison among the three genotypes" (Adamo, Atashpaz, Germain *et al.* 2015, p134, fig 2a) rather than looking at the mechanistic nature of such regulations.

Thus, these finding are suitable precisely to build the kind of map in question. If iPSCs 'contain' the entire horizon of developmental possibility (in terms of different cell types) and their transcriptional behavior suggests the directions of such a development, GO is then a map to navigate this computational horizon. And GO allows such a 'virtual tour', not by virtue of a direct and experimental examination of specific cell type lines, but rather through the fact that this tool is capable of computationally disclosing the information that is biologically enclosed in pluripotency. Indeed, as the authors themselves comment "[s]trikingly, Gene Ontology (GO) analysis of the union of DEGs showed significant enrichments for biological processes of

[5] Some researchers have argued the necessity to improve single cell analysis study. This is perfectly fine and compatible with what I said, as it will also respond to different epistemic desiderata. See again for instance Hanna, Saha and Jaenisch, 2010.

obvious relevance to the hallmark phenotypes and target organ systems of the two conditions" (Adamo, Atashpaz, Germain *et al.* 2015, p134). This shows very well why GO is an orienteering tool. GO is able to situate the information coming from the experimental work into the most updated map of current biological knowledge, thus highlighting connections and relations that are practically invisible to any single researcher or group. The power of GO is therefore to unveil existing, hidden links of biological knowledge. If the map of species and organisms provided by taxonomists is capable of suggesting possible indications on the relationships among those species, then the map provided by GO shows the capacity to do something similar for biological processes at molecular level. GO thus revealed that "[t]he top-ranking categories were related on one hand to cell adhesion, migration and motility, which appear especially relevant in light of the wide range of connective tissue alterations that characterize WBS, and on the other hand to the nervous system, providing a molecular context for the defining neurodevelopmental features of the two conditions. Additionally, further enrichments were related to remarkably specific features of the two diseases, including (i) cellular calcium ion homeostasis, a category of potential relevance across disease areas but that acquires particular salience given the high prevalence of hypercalcemia in WBS; (ii) inner ear morphogenesis, consistent with the hyperacusis and sensorineural hearing loss in WBS, as well as with the balance and sensory processing alterations found in ASD; (iii) a number of categories relevant for the craniofacial phenotypes, as represented by several categories, such as skeletal muscle organ development, migration and neural crest cell differentiation; (iv) blood vessel development and cardiovascular system development, reflecting the wide range of cardiovascular problems in WBS; and (v) kidney epithelium development, in line with the highly prevalent kidney abnormalities in WBS" (Adamo, Atashpaz, Germain *et al.* 2015, p134).

The possibility of such an approach suggested also a further step. If the GO analysis on iPSCs provided such a global map, the researchers, in order to prove whether transcriptional dysregulation would be amplified during development, derived also three lineages of cell types precursors: *PAX6-positive telencephalic neural progenitor cells* (NPCs, responsible for radial glia cells formation which, in turn, form cerebral cortex); *neural crest stem cells* (NSCSs, involved in the formation of craniofacial structures; and *mesenchymal stem cells* (MSCs, which are progenitors of osteocytes, chondrocytes and other cell types relevant for both syndromes). All these three lineages are crucially

significant for the pathological conditions under examination. The GO analysis can be seen here as the creation of three sub-maps of the previous one, against which they should be compared and judged. Indeed, the researchers "evaluated, for each of the three differentiated lineages under study, the proportion of DEGs showing conservation of the GO categories that were found to be enriched in iPSCs. Upon differentiation, iPSC DEGs were preferentially retained by category in a lineage-appropriate manner such that, for each target lineage, the proportion of conserved iPSC DEGs was much greater in categories relevant to that lineage (such as axonogenesis and axon guidance in the neural lineage, synapse-related categories in NCSCs that originate the peripheral nervous system and smooth muscle–related categories in MSCs)" (Adamo, Atashpaz, Germain *et al.* 2015, p138).

By looking at the conclusion of the study, it is quite clear that the main result of the research is the production of a specific kind of map. In particular, GO perfectly served the purpose of exploiting the potential of iPSCs. First, GO was an indispensable tool in order to manage the intrinsic *variability* of iPSCs as model for diseases, given that such a variability occurs across both individuals and lines derived by the same individual. Indeed, in order to obtain a reliable, and as much as global, picture, variability has had to be taken into account and produced, by the creation of the greatest cohort of iPSCs lines for any relevant condition. Next, of course, all this information should have been processed via high-throughput approaches. As in a complex climate forecast model where scientists need to take into account and compare different kinds of data such as geographical details, temperature differences, winds' directions and intensity, geological factors etc. and to display all of them on a common representation format, here the different transcriptional behaviours of distinct genetic conditions, the developmental issues and the pathological considerations were all consistently represented and managed by GO analysis. Indeed, such an approach perfectly exploits the potentiality embedded in iPSCs by predicting, already in the pluripotent state, which pathways will be affected given the specificity of the conditions under investigation. Moreover, the creation of a such a map, is indeed an orienteering tool by which scientists navigated the developmental trajectories thus showing how such a dysregulation "selectively amplified in a lineage-specific manner, with disease-relevant pathways preferentially and progressively more affected in differentiated lineages matching specific disease domains" (Adamo, Atashpaz, Germain *et al.* 2015, p139).

Once such a complex, multi-map has been built, it is also possible to better locate and address single factors (such as that particular protein) into the wider context of the disease development, thus suggesting further possible steps and experimental approaches. Indeed, as the relevance of specific gene products is globally assessed by GO analysis, then it would be possible to better focus on them (also with more traditional, mechanistic approaches). As the authors themselves argue "[n]otably, our analysis of symmetrically dysregulated targets also uncovered the following genes as prime candidates for mediating the molecular pathogenesis of defining aspects of the two conditions: (i) *PDLIM1*, which has been associated with ADHD, neurite outgrowth, cardiovascular defects and hyperacusis; (ii) *MYH14*, which is involved in hearing impairment; and (iii) *BEND4*, encoding a transcription factor harboring the BEN domain that distinguishes a recently characterized family of neural repressors and that was sensitive to both *GTF2I* dosage and its LSD1-mediated repressive activity, a finding that also resonates with the inversely correlated pattern of *GTF2I* and *BEND4* expression in the human brain" (Adamo, Atashpaz, Germain *et al.* 2015, p140). Such a scientific contribution does not certainly exhaust all the possibilities of map generation in this context. On the contrary it promotes the implementation of new maps and it suggests possible directions for more traditional, mechanistic experiments in order to investigate the single elements displayed on the generated map. As argued before, such efforts will be better addressed given the standardization created by GO. Hence, in order to promote and enhance such a common frame, researchers have also designed a web platform, named *WikiWilliams-7qGeneBase* to make data available to the research community working on these syndromes. Such a database will be open to external contributions given the adherence to shared format principles. In the end, by granting an original kind of scientific results, aimed at disentangling some crucial aspects of complex syndromes, and by contributing to the implementation of regulatory standardisation procedures in data display and management, such a study provides a clear example of a new way to conceptually and experimentally address the practice of epigenetic studies and transcriptional analysis.

5. The Epistemic Side

By looking at this kind of science, one may ask what is different compared to more traditional molecular studies. Interestingly, from a methodological point

of view, the order of epistemic steps in the discovery strategy has been inverted. Indeed, while the traditional strategy would have privileged genetic manipulation in order to detect phenotypic variations, here researchers started from the map of known phenotype (a database) and, through a tool capable of integrating such information with other maps, they were able to detect genes and pathways of interest, filtering then possible candidates for further, more classical, experiments. As a matter of fact, such way of doing changes the meaning and the role of experiments themselves in the current practice of science.

Moreover, by considering this type of scientific effort, it should be now obvious how much it embeds a different *way of doing* (see Pickstone 2000) rather than a change in the theoretical paradigm. Indeed, the molecular tenets are still there. The molecular stance, which drove biomedical research since the 1970s has been certainly modified, definitely extended and revised here and there, but its guiding principle are still valid. This is why I would argue that these new approaches pertain more to the epistemic and methodological side than to the theoretical dimensions of scientific paradigm. They concern how scientific evidences are produced, and how the methods to produce them can be considered reliable and scientific. If traditional molecular biologists were like old fishermen, carefully selecting the bait, the fishing pole, and focused on specific varieties of fish, the new generation of biologists seem to adopt a sort of bottom trawling, trying to collect as much information as possible. Accidental or not relevant elements such as crabs, prawns, rocks and old shoes (a metaphor for the biological noise) does not constitute a problem, given that the intellectual efforts of scientific practice will shift towards the theoretical principles and the practical constraints of collection design. In the next section I will precisely address the peculiarity of working in science with ontology from an epistemological point view.

6. Doing Science with Ontologies: Epistemic Categories

Molecular biologists look for mechanisms (see for instance Craver and Darden 2013). Moreover, molecular biology notoriously seems to lack a theoretical unification which is present in other scientific areas (such as physics). Molecular biology has then been described more as a set of techniques or, better, experimental cultures (see also Morange 2000, 2006, Rheinberger 1997). Thus one may also argue that what really makes molecular biology what

it is, can be found in the adherence of molecular biologists to a certain way of doing. It is again, a way of doing.

What is then this way of doing? Practice of science is sometimes more fluid than theoretical reflection. This is because practices may slightly vary (diversity here is a virtue) while theory tends to fill discrepancies. In order to describe a way of doing it is not possible to establish precise necessary and sufficient conditions. Rather, I will try to characterize some notions or hallmarks that clearly circumscribe the practice of molecular biology.

First, *experimental systems*. The Nobel Prize François Jacob writes that "[i]n analyzing a problem, the biologist is constrained to focus on a fragment of reality, on a piece of the universe which he arbitrarily isolates to define certain of its parameters. In biology, any study thus begins with the choice of a 'system'" (Jacob 1988, p 234). Experimental systems delimit the purpose, the boundaries and constraints of scientists' research efforts. These systems are constituted by the range of techniques adopted, the types of material instruments and resources, and, of course, the model organism on which the research will be conducted. Experimental systems are then those portions of reality, epistemically and practically demarcated, in which molecular biologists try to make discoveries. However science is not just discovery. Scientists do not just want to number phenomena. They also want to explain them. By looking at the practice of research, scientific models can be seen as one of the main tools of explanation in science. Thus, in molecular biology, experiments and models are inextricably connected. Indeed, "in molecular biology many experiments serve the purpose of developing and shaping hypotheses – about working models" (Boem and Ratti *forthcoming*). As nicely argued by William Bechtel and Robert Richardson, in order to make the complexity of biological phenomena (that are experimentally addressed) tractable, biologists use models to *decompose* the system into functional or structural elements and then try to *localize* to which structures belong certain functions and vice versa (see Bechtel and Richardson 2010).

The second notion is what Rheinberger (1997) calls *conjuncture*. Accordingly, a conjuncture is the potential intrinsic to the experimental process that can lead scientists to something that was not initially estimated. Following Rheinberger, the discovery of *transfer RNA* is a good example of this aspect. While protein synthesis was originally an area of pure biochemical investigation, the discovery of such a new molecule made it a central research field in molecular biology. Indeed, the fact that tRNA is a biochemical

intermediary between DNA and proteins, fostered the idea that it could be also an intermediary in genetic information transfer, thus establishing new paths of scientific inquiry.

Third, there is *hybridization*. Such a process occurs when parts of different experimental systems are combined in unforeseen ways. This can reveal unexpected, promising features. "The history of molecular biology is replete with hybridization events. The fusion, e.g., of François Jacob's bacterial conjugation and phage replication system with Jacques Monod's system of induced enzyme synthesis led to the emergence of another novel RNA entity, messenger RNA, and to a pathbreaking model of genetic regulation" (Rheinberger 1997, p s250).

Fourth, *bifurcation*. Briefly, a bifurcation is constituted by a new experimental system stemming out from another one (as when an *in vitro* technique is translated *in vivo*). Sometimes different systems present some degree of sharing, other times they become fully disconnected.

All these elements contribute to creating what Rheinberger calls *experimental culture*. As he points out, the adhesion of biologists to such a culture is not determined just by a theoretical commitment (which often is a set of guiding principles imported from other scientific disciplines such as chemistry and physics) but more on material tools and practical behaviors. It is how things are done that best individuate the nature of molecular biology. The seductive metaphor adopted by Rheinberger is that biological research looks then like a net of interconnected experimental systems, deploying different strategies, employing distinct approaches and materials. Namely, the *patchwork view of research*.

I would like to argue that the rise of ontologies may, somehow, challenge this picture. However, more than dismantling it, it is broadening it. A tool like GO does not have the purpose (not even the potentiality) to make traditional molecular biology obsolete. It rather has the power to change the meaning that experiments, experimental systems and other categories have for contemporary research. If heuristic strategy of molecular biology is *decomposing* complexity and *localizing* its building elements, now ontologies open the possibility to *re-compose* complexity thus adding a new, or at least an additional, layer of what scientific understanding is.

However, such a change should not be intended as a *paradigm shift* since ontologies are not shaking the main theoretical tenets of contemporary biomedicine. The point here is to examine what is the peculiarity of doing

science with ontologies from an epistemological perspective that takes into account the elements discussed in the previous paragraphs.

First, ontologies seem to extend the notion of experimental system. By the implementation of procedures that allow *packaging* and *un-packaging* data, database seem to allow data to travel (see Leonelli 2010) across different research contexts and experimental systems. In other words, data do not just serve the purposes for which they have been created. They can also be *re-used*. This is certainly true in everyday practice of research. However, it is necessary to specify the epistemic nature of such a travel. According to Emanuele Ratti (2015), such a re-use should be intended as a way scientists can pursue in order to establish the presence of common features among different experimental systems. Indeed, following Ratti, data do not simply make a journey across several contexts. The fact that GO provides indications about the type of evidence supporting a given claim, shows that data are not simply packed, unpacked and re-used neglecting their original experimental context. On the contrary, by creating a *map* that unifies the vocabulary of experimental procedures and resources, ontologies are able to make distinctions across research contexts emerge. Indeed, ontologies are enhancing comparison power and not smoothening diversities. This is because they allow data comparison rather than data homogenization. With the use of ontologies, the feature of locality of experimental systems is diminished. The "piece of universe" (recalling Jacob's words) isolated by the scientist is not fully confined any longer. On the contrary, it is now always possible to situate the space of experimental manoeuvres into a wider context. In this sense the implementation of ontological work changes also the nature of conjunctures. While in traditional experimental contexts conjunctures have an intrinsic, unforeseen potential for further discoveries which, nevertheless, cannot be disclosed from the beginning, the map provided by a tool like GO makes this epistemic horizon explorable (consider the case of Adamo, Atashpaz, Germain *et al.* 2015) described in the previous section, at least in its directions. Moreover, ontologies modify also hybridization and bifurcation. By standardizing the way knowledge is represented, ontologies can either enhance the connections between different experimental contexts or dissolve them. Indeed, the idea of a global map for biological knowledge could mean the end of different epistemic cultures interweaving and contrasting one with another, towards the establishment of a more uniform epistemic scenario. However, again, due to the peculiar form of unification provided by ontologies, I suggest

that, rather than suppressing intrinsic and distinctive features of different experimental cultures, ontologies are favoring the appreciation of differences under a common view and not the dissolution of them. As translational dictionaries, ontologies are not conflating different idioms neither reducing one language into another. They are rather creating a way to grasp the meaning of a sentence (*i.e.* an experimental system) expressed in a given language into another one.

Moreover, *map thinking*, embedded in the application of bio-ontologies, produces a distinctive signature in the way scientific research is thought and perceived, also by scientists themselves. This is because the capacity of ontologies to represent, in a human understandable fashion, the patterns emerging from databases, sets a new frame into which understanding the peculiarity of a prominent part of contemporary research. Indeed, ontologies offer a fruitful perspective in order to analyze two important ways of thinking of biological sciences, the *comparative* style and the *exemplary* style (see Bruno Strasser and Soraya de Chadarevian 2011), and their epistemic relationship. My claim is that such a distinction is fundamental to understanding the peculiarity of many current approaches in doing science.

7. Styles and Ontologies

The two styles embed to different strategies of scientific generalisations of particular findings. While, for instance in comparative anatomy or taxonomy, the generality of a scientific claim is grounded on the *comparison* among many different samples, the discovery of the so-called molecular basis of living things by new biology, promoted the idea that, as famously stated by Monod, "anything found to be true of *E.coli* must also be true of elephants" (1961). This perspective means that, since the 'code of life' has the same structure for all living beings, the universality of certain finding at the molecular level can be generalised through the assumption that the model organism, taken as the exemplary case, serves as a reliable proxy for the phenomenon under investigation. However, these two ways of thinking should not be conceived as characteriszng the disciplinary and epistemic boundaries between natural history and molecular biology. On the contrary, such a distinction has been proposed by Bruno Strasser and Soraya de Chadarevian (2011) to analyze different components of scientific practices within molecular biology. In their study, Strasser and Chadarevian point out that the historical reconstruction

that has depicted the rise of molecular biology as simply the triumph of experimentalism over observations and collection methods employed by natural history, is partially erroneous. Indeed, Strasser and Chadarevian have shown that many great achievements of molecular biology, such as the study of protein structure and function or even the 'cracking' of the genetic code, were made possible also because of comparative strategies (think, for instance, about the collections of mutations gathered and classified by Morgan). Molecular biology flourished because of the combination, sometimes even the proficuous contrast, between different *styles of reasoning* (see also Crombie 1994 and Hacking 1985, 1994, 2004, 2012 [6]). Very often these styles were anchored to specific phases of scientific progress. This means that the *exemplary* and the *comparative* style do not represent a way of thinking peculiar to this or that research program. Rather, these styles were often combined.

The early history of genetics provides a good example of this fact. Indeed, by examining the rise of modern genetics it is possible to detect when and how the generalization about certain phenomena has been differentially justified by appealing to this or that style. Let us briefly focus on the case of one of the most famous model organism: the fruit-fly *Drosophila melanogaster*. In 1910 Thomas Hunt Morgan "discovered" the first mutant *white eyes* and in 1926, due to his study on those flies, he published his famous *Theory of the Gene*. Here lies the exemplary style. The theory of Morgan was not only about fruit-flies. The gene became the fundamental unit of biological explanation (see for instance Griffiths and Stotz 2006, 2013). Every living thing has genes as any material object is composed by atoms. Because of that (and its use in the laboratory work), *Drosophila* has become a symbol of biological research for many experimental biologists. From an experimental point of view, *Drosophila* became really a standard laboratory instrument like a microscope or chemical compounds. However, although fruit-flies were clearly a key component of an experimental work, the way of thinking of Morgan rested also on a very detailed classificatory strategy. Moreover, the capacity of inferring as universals those findings obtained through the fruit-flies was based on the great number of samples and specimens produced and compared. Morgan adopted a first system (called neo-Mendelian) of classifying genetic factors "into organ group

[6] Hacking's view, although stemming from Crombie, is not entirely reducible to Crombie's. However, for our purposes here it is not important, at the moment, to highlight such distinctions.

systems - eye color, wing shape, body color, thorax pattern" (Kohler 1994, p56) which helped him to identify "how many genetic factors were involved in the formation of each morphological feature" (*ibid.*). This system was helpful to understand the developmental processes and relationships between different strains. Another classification system Morgan adopted was rather "structural and spatial". The aim of this classificatory approach was useful instead to help scientists to locate physically genetic factors, forming a sort of *genetic map*. Observing, collecting, comparing, were not replaced by the rise of experimental practice, instead they coexisted along with experiments. It is important to notice that not only the practice of classification has been fundamental to complete the genetic study of *Drosophila* but also that different systems of classification provide different answers to questions which often are seen as typically experimental. Moreover, the following failure of the neo-Mendelian system of classification due to the vastness of new mutants, on the one hand forced scientists to elaborate new classificatory systems and on the other hand helped geneticists to understand the limits of Mendelian genetics.

In this case, I would say that different ways of knowing have "mixed" with each other. In other words, again, a problem of classification involves directly the practice and the theory of experimental science. However, if we consider the question the other way round, we see how the experimental work affects the strategy of classification. Indeed "drosophilists were the first to encounter the limits of Mendelian system because they were only ones whose breeding experiments *were big enough to produce new mutants*" (Kohler 1994, p60, emphasis is mine). So choosing a specific tool (*Drosophila*) was the fundamental condition to understand the limits of Mendelian approach. Because of that "Mendelians who worked with mice or fowl had no such experience, because new mutants appeared infrequently if at all in their experiments"(*ibid.*).

Nevertheless, despite this methodological blurriness in which distinct approaches hybridize one into another, the epistemic primacy of experimentalism has definitely prevailed within molecular studies, maybe not entirely in the practice, but certainly in the way the results of biology were publicly disclosed and justified (within and without the scientific community). For instance, an article like the first one examined (Adamo, Atashpaz, Germain *et al.* 2015) would probably not have been published fifteen years ago. This is not because such a study relies on a different theoretical framework, but rather because it employs a diverse working approach. The *map thinking* shapes the

entire rationale of the article allowing to count as evidence what, in the past, was just noise or it could have been considered not relevant.

The epistemological point here is not just on the adoption of this or that methodology, but rather on the order and hierarchy of distinct ways of thinking. In other words, both opponents in this debate (see, for instance, the controversy on *Nature* 2010 between Robert Weinberg and Todd Golub) do not claim that one scientific practice should entirely replace the other, but they rather state which way of thinking should come first (epistemically, chronologically or economically). Therefore, the rise of ontologies within bioinformatics and their impact on the design of research, should not be understood as a shift from the experimental practice to the advent of a sort of '*in silico* age' of the life sciences. Even if some projects can be certainly pursued purely in a computational fashion, biologists will keep doing experiments. It is not the practice of experimentation that is changing. Rather, it is the transformation of the epistemic role of experiments within research. Thus, such an innovation indicates a difference in the general practice of science. It is something that concerns the *way of doing* science.

8. Conclusion

To sum up, in this essay, I provided several examples of current research actually driven by the application of bio-ontologies. I examined different areas of biomedical sciences, by showing how ontologies are not only applied within computational studies, but they also start to be adopted for approaching more traditional problems (such as gene function prediction), offering different and unusual perspectives. Then I proposed an analysis of how bio-ontologies change the practice of science, not just implementing the *comparative* style over the *exemplary* one, but by modifying the hierarchy of methods and evidences of research.

ACKNOWLEDGMENTS

I would like to thank Pierre-Luc Germain, Giulia Barbagiovanni, Ilaria Galasso, Maria Damjanovic and Emanuelle Ratti for the comments and suggestions on earlier version of this paper.

REFERENCES

Adamo, A., Atashpaz, S., Germain, P.-L., Zanella, M., D'Agostino, G., Albertin, V., ... Testa, G. (2015). 7q11.23 dosage-dependent dysregulation in human pluripotent stem cells affects transcriptional programs in disease-relevant lineages. *Nature Genetics*, *47*(2), 132–41. http://doi.org/10.1038/ng.3169

Bechtel, W., & Richardson, R. C. (2010). *Discovering Complexity: Decomposition and Localization as Strategies in Scientific Research*. MIT Press.

Christofferson, D. E., & Yuan, J. (2010). Necroptosis as an alternative form of programmed cell death. *Current Opinion in Cell Biology*, *22*(2), 263–8. http://doi.org/10.1016/j.ceb.2009.12.003

Craver, C. F., & Darden, L. (2013). *In Search of Mechanisms: Discoveries across the Life Sciences*. University of Chicago Press.

Crombie, A. C. (1994). *Styles of Scientific Thinking in the European Tradition: The History of Argument and Explanation Especially in the Mathematical and Biomedical Sciences and Arts, Volume 2*.

Golub, T. (2010). Counterpoint: Data first. *Nature*, *464*(7289), 679. http://doi.org/10.1038/464679a

Griffiths, P. E., & Stotz, K. (2006). Genes in the postgenomic era. *Theoretical Medicine and Bioethics*, *27*(6), 499–521. http://doi.org/10.1007/s11017-006-9020-y

Griffiths, P., & Stotz, K. (2013). *Genetics and Philosophy: An Introduction*. Cambridge University Press.

Gruber, T. R. (2009). What is an ontology? In L. Liu & T. Özsu (Eds.), *Encyclopedia of Database Systems*. Springer-Verlag.

Gurdon, J. B. (1962). The Developmental Capacity of Nuclei taken from Intestinal Epithelium Cells of Feeding Tadpoles. *J Embryol Exp Morphol*, *10*(4), 622–640. Retrieved from http://dev.biologists.org/content/10/4/622.abstract

Hacking, I. (1985). Styles of Scientific Reasoning. In J. Rajchman & C. West (Eds.), *Postanalytic Philosophy*.

Hacking, I. (1994). Styles of scientific thinking or reasoning: A new analytical tool for historians and philosophers of the sciences. In K. Gavroglu, J. Christianidis, & E. Nicolaidis (Eds.), *Trends in the historiography of science*. Kluwer Academic Publishers.

Hacking, I. (2004). *Historical Ontology*. Harvard University Press.

Hacking, I. (2012). "Language, Truth and Reason" 30years later. *Studies in History and Philosophy of Science Part A*, *43*(4), 599–609. http://doi.org/10.1016/j.shpsa.2012.07.002

Hanna, J. H., Saha, K., & Jaenisch, R. (2010). Pluripotency and cellular reprogramming: facts, hypotheses, unresolved issues. *Cell*, *143*(4), 508–25. http://doi.org/10.1016/j.cell.2010.10.008

Kaczmarek, A., Vandenabeele, P., & Krysko, D. V. (2013). Necroptosis: the release of damage-associated molecular patterns and its physiological relevance. *Immunity*, *38*(2), 209–23. http://doi.org/10.1016/j.immuni.2013.02.003

Kohler, R. E. (1994). *Lords of the Fly: Drosophila Genetics and the Experimental Life*. University of Chicago Press.

Morange, M. (2006). Post-genomics, between reduction and emergence. *Synthese*, *151*(3), 355–360. http://doi.org/10.1007/s11229-006-9029-9

Morange, M., & Cobb, M. (2000). *A History of Molecular Biology*. Harvard University Press.

Noble, D. (2008). Claude Bernard, the first systems biologist, and the future of physiology. *Experimental Physiology*, *93*(1), 16–26. http://doi.org/10.1113/expphysiol.2007.038695

Pickstone, J. V. (2001). *Ways of Knowing: A New History of Science, Technology, and Medicine*. University of Chicago Press.

Ratti, E. (2015). Big Data Biology: Between Eliminative Inferences and Exploratory Experiments. *Philosophy of Science*, *82*(2), 198–218. Retrieved from http://philpapers.org/rec/RATBDB

Rhee, S. Y., Wood, V., Dolinski, K., & Draghici, S. (2008). Use and misuse of the gene ontology annotations. *Nature Reviews. Genetics*, *9*(7), 509–515. http://doi.org/10.1038/nrg2363

Rheinberger, H.-J. (1997). *Toward a History of Epistemic Things: Synthesizing Proteins in the Test Tube*. Stanford University Press.

Shen, H. (2013). Stem cells mimic human brain. *Nature*. http://doi.org/10.1038/nature.2013.13617

Strasser, B. J., & Chadarevian, S. de. (2011). The comparative and the exemplary: revisiting the early history of molecular biology. Retrieved from http://philpapers.org/rec/STRTCA-5

Takahashi, K., & Yamanaka, S. (2006). Induction of pluripotent stem cells from mouse embryonic and adult fibroblast cultures by defined factors. *Cell*, *126*(4), 663–76. http://doi.org/10.1016/j.cell.2006.07.024

Weinberg, R. (2010). Point: Hypotheses first. *Nature*, *464*(7289), 678. http://doi.org/10.1038/464678a