New Biotechnology • Volume 30, Number 3 • March 2013

ELSEVIER

The emergence of Semantic Systems Biology

Erick Antezana, Vladimir Mironov and Martin Kuiper

Semantic Systems Biology Group, Department of Biology, Norwegian University of Science and Technology, Realfagbygget, Høgskoleringen 5, 7941 Trondheim, Norway

Over the past decade the biological sciences have been widely embracing Systems Biology and its various data integration approaches to discover new knowledge. Molecular Systems Biology aims to develop hypotheses based on integrated, or modelled data. These hypotheses can be subsequently used to design new experiments for testing, leading to an improved understanding of the biology; a more accurate model of the biological system and therefore an improved ability to develop hypotheses. During the same period the biosciences have also eagerly taken up the emerging Semantic Web as evidenced by the dedicated exploitation of Semantic Web technologies for data integration and sharing in the Life Sciences. We describe how these two approaches merged in Semantic Systems Biology: a data integration and analysis approach complementary to model-based Systems Biology. Semantic Systems Biology augments the integration and sharing of knowledge, and opens new avenues for computational support in quality checking and automated reasoning, and to develop new, testable hypotheses.

The emergence of Systems Biology

At the start of this century, the time appeared ripe for a call to adopt systems biology approaches to further our understanding of biological form and function. The Human Genome Project [1,2] had just produced the first of a series of drafts of the complete human DNA sequence, and knowledge about the genetic potential of genomes had provided a tremendous boost to the development and use of high throughput genome wide functional genomics data production approaches. The ability to produce genome scale data for virtually every molecule class of an organism provided one of the pillars of Systems Biology: the analysis of biological systems or subsystems through a large number of systems perturbations, each perturbation characterised by molecular snapshots as a proxy for the inner workings of a system [3,4].

As originally proposed by Ludwig von Bertalanffy [5], the properties of a biological system cannot be described in terms of its isolated elements. This notion has now pervaded the biomedical research domain, explaining the increasing popularity of Systems Biology: *an analysis approach that focuses on studying the whole rather than its parts*. A cornerstone of Systems Biology is a computational or mathematic model [6] that captures the dynamics of such interactions, allowing simulations of the system's behaviour over time, and the discovery of emergent properties of that system: properties not resulting from individual components but from their interaction.

The adoption of Systems Biology was further signified by the initiation in the year 2000 of research institutes devoted to Systems Biology, for example, The Institute for Systems Biology in Seattle [https://www.systemsbiology.org/],co-founded by Leroy Hood; and the Systems Biology Institute in Tokyo [http://www.sbi.jp/], founded by Hiroaki Kitano. Both Hood and Kitano were early adopters and developers of systems biology approaches, and specified the changes it would bring to biological research. Hood [3] put special emphasis on the systematic perturbation of systems, the production of molecular snapshot data for each perturbation, and the integration of these data into a (mathematical) model. The model mimics the biological system wiring and should be both descriptive and predictive. Kitano [4] focused more on the iterative nature of the approach, with consecutive cycles of data production, analysis and testing of results against a

Corresponding author: Kuiper, M. (martin.kuiper@bio.ntnu.no)

1871-6784/$ - see front matter http://dx.doi.org/10.1016/j.nbt.2012.11.008

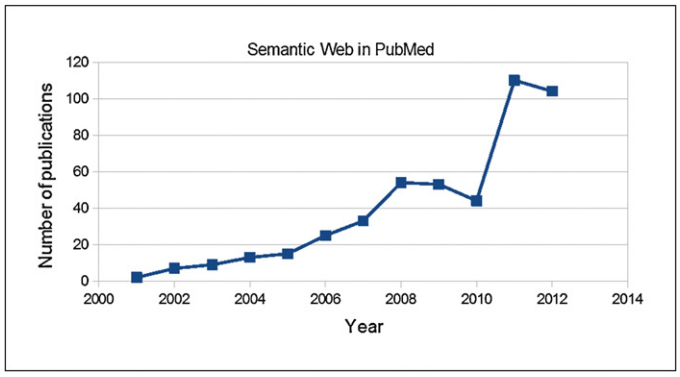

FIG. 1

Publications per year retrieved from PubMed, searching title and abstract for the occurrence of 'Systems Biology' (survey date: 23 August 2012).

mathematic model, in which each cycle would allow better hypothesis formulation and experimental design.

It is interesting that more than ten years after the concept of Systems Biology became a scientific research approach in its own right, the discipline is still defined in many different ways. Be that as it may, it is the result that counts, and in addition to a quick pervasion of the term 'Systems Biology' in the life science literature searchable through PubMed [7], see Fig. 1, perhaps the most useful result of this quest for Systems Biology was the fact that it put the hypothesis back at a central position in large-scale data production efforts. Another significant result is the development of models of biological subsystems (e.g. the Biomodels database [8,9]; Reactome [10,11]) that allow an understanding of the dynamics of the systems operations. These models are the result of a careful analysis and integration of many different types of data and knowledge.

Data integration and knowledge management

Systems Biology involves integration of huge volumes of heterogeneous data, produced by a global research community that developed many diverse repositories. Just to name a few to illustrate the diversity: ArrayExpress for gene expression data [12]), UniProtKB [13], for protein sequence and annotation data, GenBank [14], for gene sequence data and dbSNP [15] for sequence variation data. The knowledge gleaned from this data requires proper care and management to be useful. Both aspects can only be successfully secured with the use of rather sophisticated information technologies.

There are two typical approaches to data integration: *centralised* and *distributed*. In the former, the schemas of the individual databases are translated into a single unifying schema, and the data are deposited into a single database (warehouse), for example, the ArrayExpress Archive [16]. The second approach (database federation) leaves all the data in the original sources and relies on an agreed protocol to query the data (e.g. the BioMart system [17]). Neither of the two approaches is perfect and each has specific limitations. For example, warehouses are difficult to keep up-to-date and may be a suboptimal solution whereas querying federated databases can be rather inefficient in terms of performance.

Biological databases in use ten years ago, as well as in the two preceding decades, almost universally were based on the relational data model. This technology, even though well established, has several limitations when it comes to global data integration, such as the use of local identifiers (in the relational model, entities have no independent identification or existence: the unique name assumption (UNA) [18] is premised that different names always refer to different entities in the domain), idiosyncratic schemas (a single schema is necessary to define the scope and interpretation of the domain), closed world assumption (i.e., nothing is assumed, something is true only if you say it is [19]), and issues related to complexity and scalability handling [20].

Knowledge management (for a review see [21]) is the process of systematically capturing, retaining and reusing information for imparting an understanding of how a system works, and subsequently to convey this information meaningfully to other systems. In order to process knowledge computationally, it must be formally represented. Formal knowledge representation languages are means to ensure a shared understanding and an unambiguous exchange between systems (interoperability). Such interoperability is ensured via two components: syntax (symbols plus rules) and semantics (the meaning of things). Finally, knowledge must be properly conceptualised so that both humans and computers have a shared comprehension of the domain of discourse. This is normally achieved through the use of ontologies: *computer-interpretable specifications that are used by an agent, application, or other information resource to declare what terms it uses, and what the terms mean*. Formal ontologies are built using a logical framework (e.g. description logics, a branch of mathematics that allows computational reasoning) whereas in non-formal ontologies the intended meaning is described non-rigorously (e.g. natural language).

The life science research community has very early, in the use of high throughput functional genomics technologies, realised the importance of proper structuring and governance of the data. This can be illustrated by the remarkable amount of efforts undertaken to extend XML with semantics required to support particular research areas in the form of dedicated mark-up languages (e.g. for gene expression data [22,23], protein molecular interactions [24] and Systems Biology model descriptions [25]).

A new paradigm in biological knowledge management was set with the development of the Gene Ontology [26], which works towards a general, unified description of the function of genes along three axes: descriptions of biological processes, molecular functions and cellular locations. Limited as these axes may seem, the work has had a tremendous impact on the analysis of the many different genome scale data types, and new analysis approaches based on ontological descriptions continue to be developed. However, initially GO was developed non-formally and eventually the need to formalise it became obvious [27–29].

Semantic Web

The Internet, or Web, as it existed ten years ago was a web of documents linked by hyperlinks (Web 1.0). Hyperlinks lack any formal semantics and are thus inscrutable for computers; only humans can discern the intended meaning. The inventor of the Web Tim Berners-Lee proposed the Semantic Web as a layer on top of Web 1.0 that would turn it into a web of data meaningful to computers [30,31]. The Semantic Web (SW) is founded on several different technology layers. The bottom layer is formed by the Resource Description Framework (RDF [32]). RDF is a knowledge representation formalism for describing resources on the Web. Every single resource receives a global identifier, called Universal Resource Identifier (URI). The URI is a generalisation of the