New Biotechnology • Volume 30, Number 3 • March 2013

ELSEVIER

The Functional Genomics Network in the evolution of biological text mining over the past decade

Christian Blaschke[1,2], and Alfonso Valencia[1,2]

[1] Spanish National Cancer Research Centre, C/Melchor Fernández Almagro, 3, E-28029 Madrid, Spain
[2] The Spanish Institute of Bioinformatics, Spain[3]

Different programs of The European Science Foundation (ESF) have contributed significantly to connect researchers in Europe and beyond through several initiatives. This support was particularly relevant for the development of the areas related with extracting information from papers (text-mining) because it supported the field in its early phases long before it was recognized by the community.
We review the historical development of text mining research and how it was introduced in bioinformatics. Specific applications in (functional) genomics are described like it's integration in genome annotation pipelines and the support to the analysis of high-throughput genomics experimental data, and we highlight the activities of evaluation of methods and benchmarking for which the ESF programme support was instrumental.

Introduction

The 'Frontiers of Functional Genomics' Programme (2006–2011) and its predecessor the 'Integrated Approaches for Functional Genomics' (2001–2006) of The European Science Foundation (www.functionalgenomics.org.uk) have contributed significantly to connect researchers in Europe and beyond through workshops, training courses, conferences and travelling grants.

Bioinformatics was conceived as the central hub connecting the various areas of functional genomics in which the programme developed activities (see the central figure of the programme below). Of the areas of bioinformatics from sequence analysis to systems biology that were supported by the programme, those related with extraction of information from papers (text-mining) were particularly relevant. The support of the programme was particularly relevant because it supported the field in its early phases, where it was particularly vulnerable and was possibly more difficult to recognize by other organizations.

In the following, now that biological text mining is a well-established area of bioinformatics, we look back to review the historical development of the field. We put particular emphasis in the activities of evaluation of methods and benchmarking for which the programme support was instrumental (Fig. 1).

Computational methods and the (massive) integration of all existing information have become fundamental aspects of contemporary biology and biomedicine. But a considerable fraction of the existing data consist of publications written for human consumption that are not directly usable for data processing. The growth of databases such as PubMed (now approaching 22 million records), together with the increasing demand for efficient tools that can extract biologically relevant information from the published literature has resulted in the development of a range of text-mining and information-extraction tools specifically for the biological domain.

In the area of genomics text mining methods have contributed, among other things, to support the annotation of newly sequenced genomes and to predict functions for those genes that still lack any significant annotation. Furthermore it has become an important tool for database curators to increase their efficiency, support the analysis of massive data sets and enhance our knowledge of disease processes.

In this short review we will give an overview of the beginnings of text mining in its application to biology and its specific applications in the field of functional genomics. We will close with a

Corresponding author: Blaschke, C. (cblaschke@cnio.es)
[3] http://www.inab.org/.

1871-6784/$ - see front matter http://dx.doi.org/10.1016/j.nbt.2012.11.020

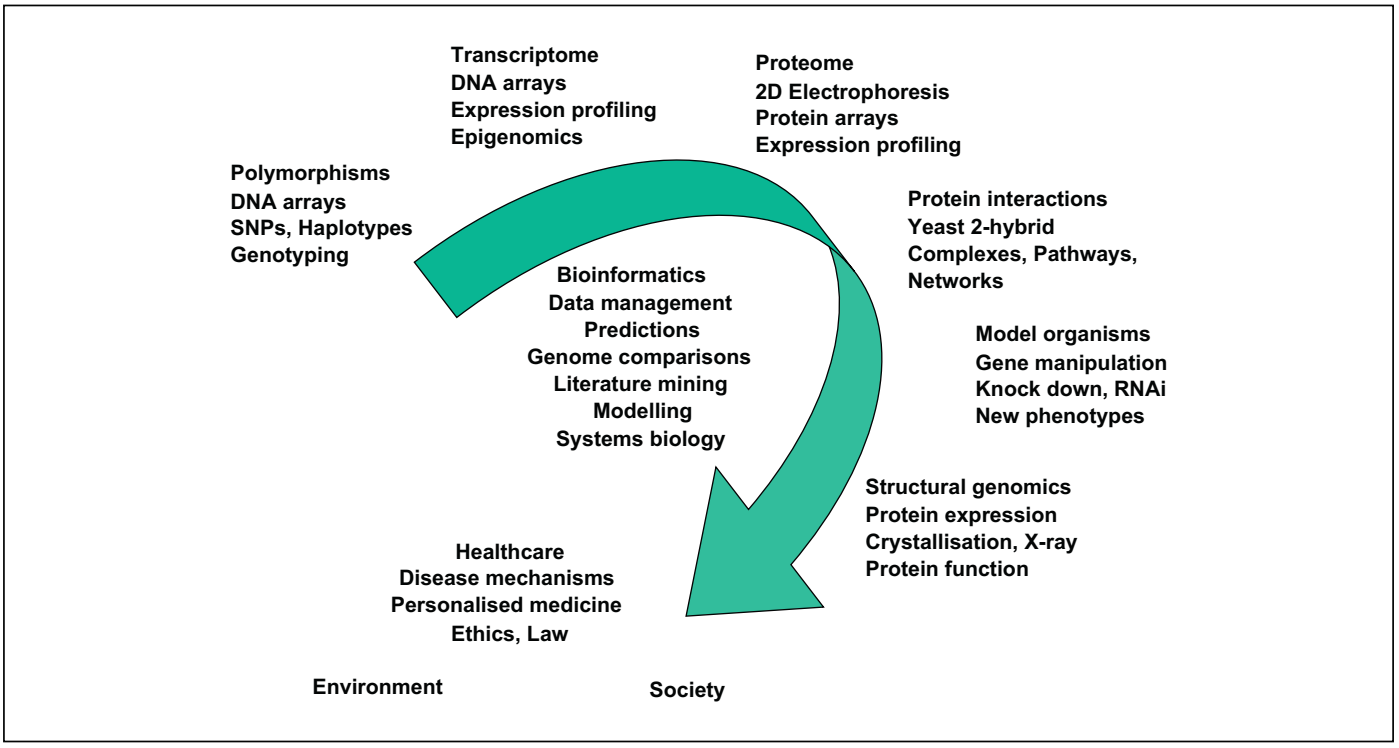

FIGURE 1

Bioinformatics as the central hub connecting various areas of functional genomics in which the ESF programme developed activities.

discussion of initiatives to assess the performance of text mining systems that have been instrumental in organizing the field and directing the development of new systems that respond to important information needs by biologists.

The beginnings of biomedical text mining

One of the predecessors of text mining applied in this area can be seen in the work of Swanson who manually searched the MEDLINE index to find connections between pathologies and their possible causes or potential treatments. Some of the well known examples are the case of dietary fish oils that lead to certain blood and vascular changes that might benefit patients with Raynaud's syndrome, and that magnesium deficiency might be a causal factor in migraine headache (see for example [1,2]). Later this search strategy was implemented in the Arrowsmith system [3] and has been followed by many others over the years (e.g. [4]).

The late 90s probably marked the beginning of the intensive research around text mining in the biomedical area. In fact, many of the now established lines of research like term and named entity recognition (NER), information extraction (IE), support to database annotation and use of text mining in protein functional annotation were initiated at that time. Wilbur and Yang applied information retrieval methodology for indexing molecular biology texts to deal more effectively with the large specialized vocabulary present in MEDLINE documents and to improve information processing [5–7] to describe a system that uses keyword statistics for automatically annotating protein function that was shown to be of similar quality to the ones contained in sequence data bases at that time. Ohta et al. [8] applied information retrieval, information extraction and automatic dictionary construction in a system that was used for constructing the Transcription Factor DataBase (TFDB).

Extracting relevant terms and their classification into predefined categories such as genes is an important prerequisite for many information extraction tasks. Collier et al. [9] explored the use of statistical methods and shallow parsing for the identification and classification of terms in biological abstracts from MEDLINE, whereas Fukuda et al. [10] applied a rule-based system to identify protein names, and Proux et al. [11] used machine learning, rules and dictionary look-up to detect gene symbols and names.

Information extraction then intends to generate structured information (for example specific relationships between the previously extracted entities) from unstructured sources (i.e. text written in natural language). Craven and Kumlien [12] showed that machine learning could be used to extract facts from biomedical text and explored possibilities to automatically construct training corpora from database entries. Blaschke et al. [13] developed a rule-based system to extract protein–protein interactions that showed good agreement with real-world interaction networks. By contrast, the system described by Sekimizu et al. [14] used shallow parsing for identifying the interaction between genes and gene products, and the method of Rindflesch et al. [15] demonstrated that classical natural language processing techniques could be applied successfully to extract binding events from the literature.

For an overview of the state of text mining in this area in the early 2000s see Blaschke et al. [16] and for a more in-depth review of the history and areas or text mining and the applications in genomics see Krallinger et al., 2005 and 2008 [17,18].