Methods xxx (2014) xxx–xxx

Contents lists available at ScienceDirect

Methods

journal homepage: www.elsevier.com/locate/ymeth

How to learn about gene function: text-mining or ontologies?

Theodoros G. Soldatos [a,*], Nelson Perdigão [b], Nigel P. Brown [c], Kenneth S. Sabir [d], Seán I. O'Donoghue [d,e]

[a] MolecularHealth GmbH, Heidelberg, Germany
[b] Instituto Superior Técnico, Universidade de Lisboa, Portugal
[c] CEITEC, Masaryk University, Brno, Czech Republic
[d] Garvan Institute of Medical Research, Sydney, Australia
[e] CSIRO Computational Informatics, Sydney, Australia

ARTICLE INFO

Article history:
Available online xxxx

Keywords:
Functional annotation
Text mining
Keyword enhancement
GO term enrichment
Systems biology
Benchmarks

ABSTRACT

As the amount of genome information increases rapidly, there is a correspondingly greater need for methods that provide accurate and automated annotation of gene function. For example, many high-throughput technologies – e.g., next-generation sequencing – are being used today to generate lists of genes associated with specific conditions. However, their functional interpretation remains a challenge and many tools exist trying to characterize the function of gene-lists. Such systems rely typically in enrichment analysis and aim to give a quick insight into the underlying biology by presenting it in a form of a summary-report. While the load of annotation may be alleviated by such computational approaches, the main challenge in modern annotation remains to develop a systems form of analysis in which a pipeline can effectively analyze gene-lists quickly and identify aggregated annotations through computerized resources. In this article we survey some of the many such tools and methods that have been developed to automatically interpret the biological functions underlying gene-lists. We overview current functional annotation aspects from the perspective of their epistemology (i.e., the underlying theories used to organize information about gene function into a body of verified and documented knowledge) and find that most of the currently used functional annotation methods fall broadly into one of two categories: they are based either on 'known' formally-structured ontology annotations created by 'experts' (e.g., the GO terms used to describe the function of Entrez Gene entries), or – perhaps more adventurously – on annotations inferred from literature (e.g., many text-mining methods use computer-aided reasoning to acquire knowledge represented in natural languages). Overall however, deriving detailed and accurate insight from such gene lists remains a challenging task, and improved methods are called for. In particular, future methods need to (1) provide more holistic insight into the underlying molecular systems; (2) provide better follow-up experimental testing and treatment options, and (3) better manage gene lists derived from organisms that are not well-studied. We discuss some promising approaches that may help achieve these advances, especially the use of extended dictionaries of biomedical concepts and molecular mechanisms, as well as greater use of annotation benchmarks.

1. Introduction

The emergence of high-throughput technologies (such as expression microarrays and next-generation sequencing) generates large gene lists measured under a variety of conditions. While many computational tools have been developed to help biologists automatically gain insight into the biological processes underlying such lists, a great deal of them produce summary-reports in the form of (hypothesized) annotations. Such annotation tools are primarily designed to provide a quick first insight into the functional difference between two sets of genes (typically associated with a specific phenotype or condition), frequently by mining information from databases and the literature (as discussed in Section 2).

To help understand these processes, the current article engages in an epistemological trek through functional annotation, i.e., by investigating primary aspects (with regard to methods, validity and scope) and perspectives that refer to the systems employed to acquire, explore and transform (functional) information into an organized body of verified and documented knowledge. For example, a big part of the knowledge contained in gene resources

* Corresponding author.

http://dx.doi.org/10.1016/j.ymeth.2014.07.004
1046-2023/

(databases and literature) is written by and for human experts. To transform the representation of knowledge from natural language into other forms computer-aided reasoning is required; e.g., by using text mining (TM) methods to analyze all literature related to a given gene set (as discussed in Sections 2.2 and 3).

For TM-based methods, two key challenges are syntax and semantics: syntax refers to the definition of symbols that a representation method uses and the rules that these symbols may be combined with, whereas semantics refers to the definition of the concepts assigned to symbols and the combinations of symbols that the syntax allows [1]. Regarding syntax, TM in the biosciences is aided partly by the many efforts towards developing biomedical term ontologies that describe relationships between concepts via hierarchies (as discussed in Sections 2.2 and 3). However, a confounding factor is that biomedical ontologies are generally far from complete, since the phenomena they attempt to describe can be overwhelmingly complex and many times are only partly understood (Section 2.3). Besides syntax, TM has two main difficulties in coping with semantics, namely language ambiguity and context-dependent interpretation (Section 3.1.2).

In the design of functional annotation and TM tools (as with other systems for automated knowledge inference), two basic strategies are common: data-driven reasoning (also known as 'bottom-up') and goal-driven reasoning (also known as 'top-down') [1]. Data-driven reasoning starts from 'low level' data and gradually seeks to reach 'high level' observations – annotation of genes mostly corresponds to this (Sections 2.4 and 3), beginning from genes selected through experimental data and aiming to infer 'higher'-level information (e.g., about phenotype).

In contrast, goal-driven reasoning works in reverse, beginning from known high-level observations (e.g., a known phenotype) to find lower level, more detailed information (e.g., the set of genes or chemicals associated with that given phenotype). To achieve these goals directly via annotation requires the development of platforms that can successfully synthesize information in a holistic 'systems' perspective (as discussed in Section 2.4), an aspect that becomes especially important in an era of personal genomics (Section 5.1).

In Sections 2 and 3, we also frequently contrast the relative merits of the TM and ontology-based approaches; we expect that future methods will achieve improved annotation accuracy and coverage by combining the strengths of the ontology-based approaches with the increased sensitivity of keywords extracted directly from the literature (Sections 5.2 and 5.3).

While reviewing features that characterize but also synergistically effect the development of functional annotation tools is main objective of Sections 2 and 3, the somewhat overlooked aspect of systematically measuring performance of such methods is also examined in Section 4, as quantification of such results is often difficult. This in turn prevents objective evaluation through means of metrics such as precision and recall: precision describes 'how many from the extracted terms have been correct' whereas recall measures 'how many of the terms that should have been retrieved have indeed been identified' – i.e., the extent of 'result-success', in terms of coverage. To help bridge the gap between qualitative and quantitative evaluations of annotation assignments, Section 4 also investigates the features that a proper benchmark should facilitate.

2. Annotating gene function

While a large range of tools for functional analysis exists ([2,3] – at least 68 listed by [4] – Table 1) typically the underlying methods utilize sequence similarity, TM of database annotations or literature, and keyword hierarchies or ontologies. In this section, we discuss some of the relative merits of these strategies.

2.1. Functional annotation by homology

It is nowadays easily possible to obtain the complete genome sequence of an organism but determining the function of gene products remains highly non-trivial (Supplement, Part [B]) and often requires experimental validation [5]. The sequence of a gene is one of its primary properties, and many methods have been developed to predict gene function based on sequence similarity to genes of known function (e.g., GOtcha [5], OntoBlast [6], and GOblet [7]). Such assignments can work very well in the cases that (a) the two sequences are highly similar to each other and (b) when the reference product is functionally well studied so that it can support an adequate description of molecular behavior [5]. Nevertheless, both constraints are often not met for several reasons, including the following:

- lower sequence similarities may lead to multiple low-quality candidate reference products;
- various candidate products may have been assigned differing functions;
- some classes of sequences are similar in sequence but diverse in function; and
- vice versa, sequences with similar function may belong to different classes.

Such sequence-alignment methods are based on the evolutionary concept of homology (i.e., genes with common ancestry share function). Nevertheless, this concept does not always apply ([5,8]). For example, orthologs (i.e., similar sequences of different organisms originating from the same species) often, but not always, have the same function. Paralogs (i.e., duplicated copies that end up occupying different positions in the same genome) often have the same or similar function, but sometimes not, since each copy is free to mutate and acquire new functions or regulatory mechanisms independently of the other.

2.2. Functional annotation using text-mining

Some methods that utilize sequence similarity (e.g., GeneQuiz [9] and PEDANT [10]) try to extract directly, from databases or from literature, textual descriptions of gene product function. Such methods have to overcome the complications of TM and interpreting natural language [5]. For example, the same biological function may be described in different ways by different investigators. In addition, computational processing of annotations sometimes cannot determine whether different human curator assignments conflict with each other. A key reason underlying these complications is that two distinct terms may be used to describe the same function, and it can be difficult for TM to always recognize this situation. A partial solution for these problems has been the development of ontologies (such as Medical Subject Headings (MeSH) [11], or Gene Ontology (GO) [12]) and their use in systems like GOblet [7]. Ontologies often describe relationships between terms, and hence can help address the issues raised above.

2.3. Functional annotation using Gene Ontology

Many of the major gene databases provide, in each gene record, a description of the gene's function as a set of GO terms (e.g., Entrez Gene [13]; Supplement, Part [B]). The GO initiative was created to provide a unifying ontology to describe biological functions using terms that can be represented as a directed acyclic graph, where each node represents a clearly defined biological concept ([5,12]) – GO has since become widely adopted. Table 2 summarizes some of the features that make ontologies in general, and GO in particular, favorable for the development of gene annotation

Table 1
List of tools related to functional annotation (and text-mining; TM) mentioned in the main text of this article. The list includes named (and accessible via the web) annotation and TM entities contained in the main text of this article (and only those; providing an exhaustive list of such relevant resources is outside of the scope of this article); entities ordered by citation.

Name	Feature (major reason for mention)	URL (status)	References[a]; cited also in.
GOtcha	Use of sequence similarity searches	http://www.compbio.dundee.ac.uk/gotcha/gotcha.php	[5]; Sections 2.1, 2.2, 2.3, and 4.1
GOblet	Use of sequence similarity searches	http://goblet.molgen.mpg.de/cgi-bin/goblet2008/goblet.cgi	[7]; Sections 2.1 and 2.2
PEDANT[c]	Use of sequence similarity searches	http://pedant.gsf.de	[10]; Section 2.2
GOrilla[c]	Analysis and visualization	http://cbl-gorilla.cs.technion.ac.il	[14]; Section 2.3
GS2[c]	Comparison and similarity	http://bioserver.cs.rice.edu/gs2	[15]; Sections 2.3 and 2.4
OAT[c]	Analysis and visualization (interpretation; browsing)	http://bioinfo.ifm.liu.se/services/oat	[16]; Section 2.3
TXTGate	Literature; within set analysis.	http://tomcat.esat.kuleuven.be/txtgate	[20]; Table 4
PubMatrix	Literature; comparison of term lists (e.g., gene names)	http://pubmatrix.grc.nia.nih.gov	[21]; Table 4
CoPub	Literature; expanded dictionary.	http://services.nbic.nl/copub5	[47] ([22]); Section 2.4, 3 (Table 4)
Martini	Literature; two-set comparison.	http://martini.embl.de	[23]; Sections 2.4, 3, 4.1, 4.2, 4.3, 5.1 and 5.3; Fig. 3; Tables 4, 5, 7 and 8
FunSpec[c]	Enrichment; yeast (one input list)	http://funspec.med.utoronto.ca	[24]; Table 4
VAMPIRE[c]	Specific data focus (microarray data)	http://genome.ucsd.edu/microarray	[26]; Table 4
High-Throughput GoMiner	Specific data focus (microarray data)	http://discover.nci.nih.gov/gominer	[27]; Table 4
Onto-Express (of Onto-Tools)	Integrated platforms or software suits	http://vortex.cs.wayne.edu/projects.htm	[28] of [29]; Table 4
DAVID[c]	Integrated platforms or software suits	http://david.abcc.ncifcrf.gov	[30]; Section 2.4; Table 4
BABELOMICS	Integrated platforms or software suits	http://babelomics.org	[32]; Table 4
CLICK and EXPANDER[c]	Analysis and visualization	http://acgt.cs.tau.ac.il/expander	[33]; Table 4
BiNGO[c]	Analysis and visualization	http://www.psb.ugent.be/cbd/papers/BiNGO	[34]; Table 4
NetAffx Gene Ontology Mining Tool	Analysis and visualization	http://Affymetrix.com/analysis	[35]; Table 4
FatiGO[c]	Enrichment; allows two sets as input	http://babelomics.bioinfo.cipf.es/functional.html	[37]; Sections 2.4 and 5.1; Table 4
CoCiter	Comparison and similarity	http://www.picb.ac.cn/hanlab/cociter/	[38]; Sections 2.4, 3, 3.1.1, 4.1 and 4.2; Table 4
GSFS[c] toolkit	Comparison and similarity	http://bioinfo.hrbmu.edu.cn/GSFS	[39]; Sections 2.4, 4.1, 4.2; Table 4
GOstat	Enrichment; allows two sets as input	http://gostat.wehi.edu.au	[40]; Section 2.4; Table 4
GSS[c]	Tools for gene set analysis	http://bio.ccs.miami.edu/cgi-bin/GSS/AnalyzeGeneSets.cgi	[41]; Table 4
ProfCom[c]	Enrichment; allows two sets as input	http://webclu.bio.wzw.tum.de/profcom	[43]; Section 2.4; Table 4
PANTHER[c]	Pathway analysis	http://pantherdb.org/tools/compareToRefListForm.jsp	[45]; Sections 2.4 and 5.1
Reactome	Pathway analysis	http://reactome.org	[46]; Sections 2.4 and 5.1
Marmite[c]	Literature-based enrichment analysis	http://babelomics.bioinfo.cipf.es/functional.html	[48]; Sections 2.4 and 3
eTBLAST	TM: IR[b,d]	http://etest.vbi.vt.edu/etblast3	[56]; Section 3.1.1
Caipirini	TM: IR[b,d]	http://caipirini.org	[57]; Section 3.1.1
MedlineRanker	TM: IR[b,d]	http://cbdm.mdc-berlin.de/~medlineranker/cms/medline-ranker	[58]; Section 3.1.1
MScanner	TM: IR[b,d]	http://mscanner.stanford.edu	[59]; Section 3.1.1
Génie	TM: IR[b,d]	http://cbdm.mdc-berlin.de/~medlineranker/cms/genie	[61]; Section 3.1.1
Peer2ref	Suggesting authors, finding experts.	http://peer2ref.ogic.ca	[62]; Section 3.1.1
Jane	Suggesting authors, finding experts.	http://biosemantics.org/jane	[63]; Section 3.1.1
Reflect	TM: ER (highlights terms)	http://Reflect.embl.de	[66]; Section 3.1.2
Alkemio	TM: IR[b,d]	http://cbdm.mdc-berlin.de/~medlineranker/cms/alkemio	[73]; Section 5.1
Metab2MeSH	Compound annotation with MeSH	http://Metab2mesh.ncibi.org	[74]; Section 5.1

[a] Reference: citations as in the main text of the article.
[b] Abbreviations: *GO:* (Gene Ontology), *TM* (Text-Mining), *IR* (Information Retrieval), *ER* (Entity Recognition).
[c] Declared software abbreviations (see also Supplement, Part [A]).
[d] Characteristics and details of TM-IR tools (see also Supplement, Part [A]).

and enrichment tools (such as GOrilla [14]), gene set similarity applications (such as GS2 [15]) and gene set browsing and interpretation systems (such as OAT [16]); a comprehensive list of GO tools is maintained by GO itself ([17,18]). As discussed also in Section 4.1 with examples, these ontology- and term-specific characteristics can also provide metrics for direct and objective comparison, independent of the arbitrary cut-off values that different methods may apply. On the other hand, some GO features impose qualitative difficulties (Table 3).

In spite of the potential limitations in GO and its application to gene annotation, it remains today a widely used standard. In a GO-based analysis, usually, all annotated GO terms and all other GO terms that are associated with them (i.e., lower or higher in the hierarchy) are found. Then, the significantly over-represented terms for a gene set are considered to be those that describe the gene set. To identify over-represented terms for one gene set, usually, the number of appearances of each GO term inside the group of interest is counted, and compared to that of a group of reference genes. A number of statistical tests are available, but usually Fisher's Exact Test is performed to judge whether the observed difference is significant or not, and in the end of the analysis a *p*-value score for each GO term is calculated that indicates the likelihood that the observed counts occur by chance ([2,4]); Supplement, Part [C]. The most significantly overrepresented GO terms are then

Table 2
Advantageous features of GO. GO has a clear advantage over more generic term hierarchies or dictionaries.

Feature	Description
Continuously developed	GO is continuously being improved and updated
Clear biological meaning	Since GO is developed exclusively by the biological community, all terms are given precise biologically relevant definitions
Inter-term relationships	In GO, each term is assigned a unique identifier and relationships with other terms are clearly defined. This specification of term interrelationship can help greatly in TM
Multiple levels of abstraction	In the GO hierarchy, genes are annotated at various levels of abstraction. For example, 'induction of apoptosis by hormones' is a type of 'induction of apoptosis', which in turn is a part of 'apoptosis'. 'Apoptosis' represents a higher level of abstraction, whereas 'induction of apoptosis by hormones' represents a lower level of abstraction. Thus, gene function can be described at varying levels of abstraction, depending on the needs of the user or the requirements of a particular application
Computational analysis	The tree-like organization of GO makes it appropriate for using automated similarity measures that are applicable to quantitative comparisons

Table 3
Qualitative challenges of GO. The quality and completeness of annotations made using GO is far from perfect for various reasons.

Feature	Description
Continuously developed	While continuous development of GO in general leads to improved accuracy, it also means that annotations made previously with GO can become out of date
Incompleteness of the ontology	Ideally, GO would contain a complete description of all gene functions, organized within ontological hierarchies. Unfortunately, the current hierarchy in GO is incomplete, although it is constantly being improved
Annotation coverage	Many genes currently have few or no GO annotations
Quality of annotations	The quality of annotations made using GO may be restricted either due to limitations of specific human annotators, or due to limitations in accuracy or precision of annotations inferred by computational methods using GO
Annotation consistency	Genes are not always consistently annotated at the highest level of detail possible. For example, one gene may be annotated simply as involved in 'apoptosis', while another may be annotated as involved in 'induction of apoptosis by hormones', when in fact both relate to the same function. This can lead to ambiguous or redundant annotations

assumed to describe functional properties shared by the input gene set.

2.4. Tools for annotation: a combined variety of features

Altogether, there is a large variety of features that can characterize annotation tools and their use (Table 4). Partly this is due to the fact that biological data integration efforts have so far failed to efficiently and successfully make available a central data resource with information for all genes and to standardize the used data formats. However, this has not prevented the development of more tools; most recent annotation techniques rely on a combination of the available resources (Table 4). Special emphasis can be paid to three recent directions discussed next.

2.4.1. The scope of semantics

The larger variety of concepts represented by the underlying dictionaries, the more 'systems applications' can be derived. Annotation systems have gradually expanded their scope from searching for gene names, biological processes and molecular functions to diseases and chemical compounds (e.g., Martini [23,44]) as well as interactions and pathways (e.g., PANTHER [45] or Reactome [46] tools). This '*systems innuendo*' denotes a more goal driven approach – an underlying desire to infer higher-level associations directly from gene lists. For example, FatiGO [37] and CoPub [47] rely primarily on GO, but also additional information such as pathways, and by comparison, Marmite [48] and Martini [23] incorporate further term categories such as 'diseases' and 'chemicals' – a large variety of dictionaries and hierarchies exists that can satisfy this broader scope, from GO, KEGG and MeSH to ATC [49] and MedDRA [50]. PANTHER [45] on the other hand provides another permutation of approaches by incorporating homology, pathway and ontology analysis comprised of a subset of GO terms (GO slim) complemented by their own ontology.

2.4.2. The two-set comparison

The systems described earlier typically refer to statistical measures for comparisons of two groups, such as a treatment group and a control group that in the case of functional annotation are gene sets. The choice of the 'control' or 'reference' list is an important consideration when identifying statistically significant terms and different tools approach this aspect in different ways: CoCiter [38] and Marmite [48] require that users explicitly upload both lists of interest, whereas CoPub [47] is based on fixed reference sets. Also FatiGO [37] and ProfCom [43] use a predefined reference set, unless the user specifies otherwise, whereas PANTHER [45] (v 9.0) allows for multiple lists to be compared. Usually in functional annotation, the set of all genes in a genome is used as the reference, but this may be an inappropriate choice when the selected list of input genes is derived from a condition the mechanism of which may possibly involve only a very specific class of genes – i.e., a part of the genome (e.g., see [2,42] for microarrays). This is because ideally term significance should be measured against a gene set that belongs to a related pool ([2,3,42]). Although in many cases this is not easy (e.g., a second list is not always available), to avoid contradicting this rule it is in principle best when users explicitly define the reference set (e.g., genes known to be involved in same/similar pathway or tested under the same experimental condition). Explicitly specifying the background also allows addressing more interesting questions, since the characteristic annotations derived after the comparison (whether searching for similarities or for differences between the two sets) are done '*with respect to*' an informed reference that also represents an important topic of interest (e.g., genes measured under a certain other condition).

2.4.3. Similarity instead of difference

In contrast to the above, some methods for gene set comparison search for similarities or *associations* between two input gene sets. Some of these methods analyze the gene annotation overlap between gene sets (e.g., GS2 [15], DAVID [30] and GOstat [40]), but because the degree of overlap between two gene sets can influence the analysis of functional similarity, newer methods examine association based on categories identified to be significant for each set (e.g., GSFS [39]) or via additional features, such as

Table 4
Ten key characteristics of annotation tools. Some of the principle differences include focus, methodology and data source as there is considerable difference in the databases that tools use as their primary source for deriving annotations.

Feature	Description
Information resource	The used data sources are primarily PubMed [19] for literature based tools like [20–22], and/or GO for ontology based methods (like [6]). In most cases, a variety of other databases (e.g., KEGG) and keyword hierarchies are co-integrated
Dictionary integration	Most tools create their own dictionaries, rely on external keyword hierarchies and ontologies, or integrate previously existing resources into their own vocabulary
Scope of semantics	Most tools for functional annotation utilize keywords that represent molecular components, and biological processes or mechanisms (e.g., gene names, GO/KEGG), whereas others expand by associating also compounds or diseases (e.g., [23]). The scope is largely represented by the choice of underlying dictionary
Numbers of species	Some tools (such as [24], useful only for Yeast) are restricted to a very small number of species
Specific data focus	Some systems build on top of previously published tools and use gene annotation as a method to analyze specialized data sets (e.g., [25], [26] and [27] are useful for the interpretation of microarray experiments)
Analysis and visualization beyond annotation	Some tools are integrated in larger projects, pipelines or suite of tools that can perform also other types of analysis; examples include [28] (part of [29]), [30] (a set of bioinformatics resources some of which perform gene annotation and classification), [31] (a package of web-based tools for gene annotation), and [32] (suite of tools, some of which can perform functional annotation). Some (like [33–35]), incorporate further functionalities for analysis (e.g., to cluster or classify the input genes) and also place emphasis towards visualization expertise that can support and extend the annotation results (e.g., by linking the output to numerous other databases, or by using enhanced interactivity techniques, such as graph and network analysis). Gene/Protein set and pathway analysis applications [36] are similar in scope that come however with their own downsides, e.g., gene expression is often specific to cell type and changes over time; in addition, pathway databases can be incomplete (i.e., not comprehensively covering all genes known to be involved in a process)
Number of gene lists	Many tools focus on describing a single input set of genes, whereas others can take as input two gene sets. Allowing the user to define a second input gene set (a specific reference set), helps characterize functional differences between exactly these two sets (two-set comparison) and answer more detailed questions
Comparison of gene-lists	Most tools for two-set comparison focus on inferring functional differences between two gene sets (e.g., [37] and [23]) instead of similarity (e.g., [38] or [39])
Within vs overall analysis	Some tools look *within* the set of genes, either by focusing on the annotation of each gene individually ('single gene analysis', e.g., [40]), or by interactively dividing a single gene set into functionally related sub-clusters (e.g., [20] and [41]). Most *within* methods that 'group and profile' (i.e., identify subsets of genes with related function from the input set) aim to do gene set analysis or to make the resulting group of enhanced keywords or GO terms more interpretable by processing further the results (e.g., by clustering or displaying GO terms associated with individual genes), whereas by contrast others look at the *overall* gene set attempting to derive a functional description for the entire input set (e.g., [37] or [23])
Statistical method	While a variety of statistical tests and distributions exist (e.g., Fisher's exact test, hypergeometric, binomial) to model the annotation task ([42,2]), most tools rely on enrichment analysis (Supplement, Part [C]) searching for over- and under-represented associations, or both (e.g., [41]). Other advances include more complex models or heuristics (e.g. [43]) and taking advantage of co-citation and other (semi-) structured literature features (e.g., [38])

literature co-citation (e.g., CoCiter [38]) or protein network interactions and pathway overlays (e.g., [51,45]).

3. Functional annotation using literature

A key feature of gene records in the major databases is the list of literature associated with the gene, typically provided as PubMed identifiers [19]. The literature associated with a gene provides a potentially very rich source of information about gene function, and in some cases can contain information that partly overcomes problems mentioned above arising from insufficient annotation with GO terms (Fig. 1; Supplement, Part [B]), or incompleteness of the GO ontology itself (Section 2.3, Tables 2 and 3). To extract gene information from literature, a range of TM-based annotation tools have been developed (e.g., Martini [23], CoCiter [38] and Marmite [48]). However, TM is not mutually exclusive or necessary a competing method to GO-based, or other ontology-based, approaches (Table 5): in fact some TM-based methods are built using GO as their reference dictionary (e.g., [47]). In some use scenarios, GO-based approaches may give better performance than purely TM-based approaches, and vice versa. For example, TM is better than GO terms for well-studied organisms, but not for most organisms (e.g., in SwissProt), due to lack of literature; however, this limitation of TM could potentially be overcome by using sequence homology to transfer functional annotation (Fig. 1; Supplement, Part [B]).

Putting aside implementation details such as storage, indexing, calculation of relevance, or the algorithms incorporated, an annotation implementation using literature typically consists of four components (Fig. 2):

- Defining the topic of interest: mostly the input can be one or two sets described in accessions, names of genes/proteins, other query terms, or abstract identifiers.
- Retrieval of literature: usually, any type of input is translated in a list of PubMed identifiers (e.g., GeneRIFs from Entrez Gene records).
- Keyword extraction: the next step is to convert each retrieved result into a list of keywords. Some tools allow term type selection where a user can optionally select if certain types of terms should be considered, or not (e.g., as in Martini [23]).
- Statistical analysis: as discussed in Sections 1, 2.3. and 2.4., functional annotation approaches are mostly data-driven as commonly tools start from gene sets and 'go up' searching for significantly over-represented terms. These methods can be applied to keywords found in the literature associated with each gene (*keyword enhancement*) or GO terms associated with each gene in the input set (*GO term enrichment*); Supplement, Part [C].

While defining the topic of interest is a step that precedes the retrieval of literature, both steps are closely related and their aspects are commonly discussed together under the data mining term 'information retrieval'. Next to information retrieval, keyword extraction is part of another key TM step, the so called 'entity recognition' task. Each of these TM components poses a variety of challenges discussed in Section 3.1., below.

3.1. Text mining considerations

Several methods have been proposed for finding literature related to a gene, a gene set, or a field of specific biomedical

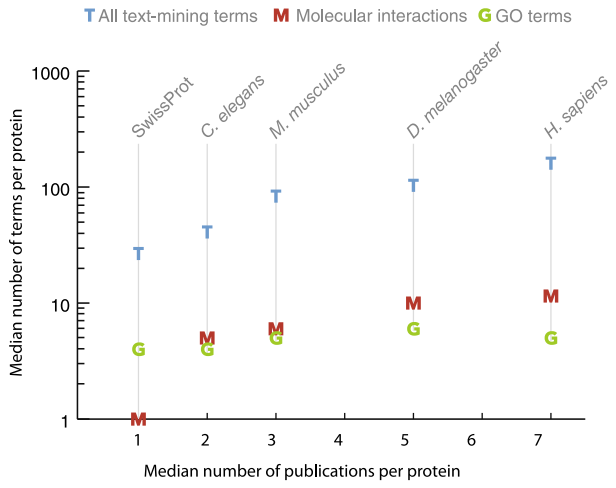

Fig. 1. Comparison of functional annotation depth in SwissProt via GO terms, publications, and text-mining (TM). For well-studied organisms (e.g., human), a much richer description of protein function can be obtained via TM of related publications compared to relying only on GO terms. Conversely, GO-based functional annotation may be best when studying proteins from most organisms, since the median number of publication per protein is one, compared with four GO terms. These data were derived from the December 2013 version of SwissProt; the median number of GO terms and publications per protein were calculated for the indicated organisms, as well as for all SwissProt proteins. The number of biomolecular interactions and TM terms per protein were estimated for each category as follows: (1) we found all proteins for which both the number of publications and GO terms were exactly equal to the median values; (2) from these, we selected 20 proteins randomly and collected all related abstracts into a single text file that was then sent to the Reflect service to tag either biomolecules (chemicals and proteins) or Wikipedia terms; (3) the median number for these 20 proteins are reported, with all TM terms equal to the sum of biomolecules plus Wikipedia terms. The TM measures derived here are an approximate estimate of the number of terms that can be automatically extracted from publications to describe protein function – the precise values will vary greatly depending on the method used. The Reflect method used here has two disadvantages: firstly, it overestimates to number of terms per abstract, due to using Wikipedia entries as a dictionary; secondly, it underestimates the number of terms, as it considers only abstracts, not full-text documents.

interest (e.g., [53–55] list such tools). Yet, none of the TM components of literature-based annotation (describing a topic interest, identifying matching documents, or extracting keywords) are straightforward. For example, retrieving relevant literature can be hindered by the large size of literature related to a given specific topic in biosciences, by the heterogeneity of the related studies and by the lack of a clearly predefined categorization of abstracts.

3.1.1. Information retrieval

To ensure that abstracts most related to the input query are not buried way down in the returned list of results, often evaluation and ranking by relevance is incorporated. To improve text retrieval in this way, a wide variety of tools have been implemented, most of which rely on the content-based philosophy: *i.e.*, that abstracts sharing more similar annotations or words with the input are likely to be more related (e.g., eTBLAST [56], Caipirini [57], MedlineRanker [58], MScanner [59]); see Supplement, Part [A]. Others use structured features of literature to improve inference or prioritization, such as co-citations or co-occurrence (e.g., [60]) for a variety of tasks - from functional annotation (e.g., CoCiter [38]), to ranking genes instead of literature (e.g., Genie [61]), or even helping editors find reviewers (e.g., Peer2ref [62] and Jane [63]).

To ease the retrieval, sometimes curators associate documents with categories derived from hierarchically organized ontologies, such as GO or the MeSH thesaurus; sometimes such ontologies can be used also as the underlying dictionary from which keywords can be derived (Section 2.4; Table 4). This allows abstracts to be organized into categories, thus defining sets of papers related to specific topics. Retrieval may be confounded a range of factors: e.g., when documents refer to several topics simultaneously, or belong to heterogeneous types (*e.g.*, reviews, laboratory notes, clinical records, patents, etc.), or come in different formats (e.g., PDF, XML, etc.) or different languages. Most annotation tools discussed here consider primarily PubMed, for which these issues are usually manageable.

Queries constructed by keywords can also be difficult and may still not capture all relevant abstracts. In many cases the user needs to take into account non-standard nomenclature for very specific biological fields, and non-expert users can have difficulty in providing all the relevant keywords. Also, different terms with similar meaning when used in a query can give different results. To help users, some systems automatically expand the query by adding synonyms and alternative expressions that can be interchangeably assigned (e.g., 'tagging' instead of 'annotation'). Eventually, these keywords are matched against candidate documents – usually, those that contain the query terms in their text are returned as the result.

3.1.2. Entity recognition

Identifying the keywords of a text can be tedious; the two major tasks involved are first finding the terms and then assigning meaning to the terms (as discussed in Section 1). A wide range of methods have been developed for identifying terms in a text, the most elementary of which rely on simple matching of text patterns (e.g., word strings). Semantics are usually disambiguated with the use of synonyms or cross-references to clearly defined database records – in that respect, dictionary-based methods have a crucial advantage over those based only on syntactical features or patterns, since dictionaries help not only to recognize names but also to use synonyms and accession numbers for linking to summaries reported in mapped external records (e.g., for identifying which gene a term refers to). Finally, hybrid approaches combine dictionary matching with rule-based and statistical methods to reduce the number of false positives.

This is especially important for gene, protein, and drug names, which constitute a special challenge as often they are comprised of multiple words (e.g., 'brentuximab vedotin'), or are referred to by abbreviations (e.g., 'FOLFIRI therapy'), often in combination with alphanumeric symbols (e.g., 'Sti-571' instead of 'imatinib'). Furthermore, distribution of gene names and detection accuracy may vary among texts with different length or within different sections of a full text article [64]. Other arduous aspects of biomedical entity recognition include: name ambiguity; unclear synonym relations; typographical errors; misspellings; orthographic and language variants; domain-specific terminology and styles (e.g., non-standard nomenclature for genomic variants); ambiguous semantics, part-of-speech and grammatical relationships; and finally, incompleteness in databases, hierarchical vocabularies and ontologies. A partial solution to some of these issues can be the use of open dictionaries (e.g., [65] or [66]) that are frequently updated, edited, and corrected by the scientific community; however, the most reliable approach to tagging of biologically relevant terms remains manual curation from trained experts. However, this is not feasible for all biomedical literature corpora.

4. Functional annotation assessment

While significant effort that has gone into benchmarking the performance of pure TM tasks applied to life science literature (especially the BioCreAtIvE initiative [67]), annotation tasks have not undergone the same scrutiny. Typically, function assignment methods are assessed against incompletely annotated datasets.

Table 5
TM vs GO: summary of relative merits. The reliability of GO-based approaches has different strengths compared with the increased sensitivity of keywords extracted directly from the literature.

Challenge	Description
Quality of terms	The performance of methods that rely on specific keyword lists depends on the quality of the terms incorporated in these ontologies, hierarchies, or dictionaries. Moreover, reliance on a specific ontology or dictionary may lead to over-representation of certain functional aspects. By contrast, TM techniques are typically not restricted only to the biological knowledge incorporated in a specific ontology – however, TM has the disadvantage that it relies on keyword-based dictionaries that may not describe biological functions as coherently and clearly as ontologies. For example, a keyword-based dictionary may instead contain many 'noise' terms, with limited biomedical meaning or interest
Coverage	Compared to ontology-based methods, approaches that rely on literature can sometimes retrieve more information related to the function of genes, since ontologies often contain terms related to a specific functional aspect, whereas literature is not as focused in scope (e.g., [23,44,52]). For example, GO does not contain drug names, whereas an article may in addition refer to clinical aspects of the expression of a gene, or to the structure of its products, and so on
Biological relevance	Significant keywords extracted from literature may sometimes not describe a gene's function, whereas GO-terms have in comparison a clear biological meaning. However, the number of GO-terms associated with a gene is often not as large as the number of keywords extracted from prose
Method complexity	Although TM can potentially extract a wealth of functional information for a gene set from biomedical literature, the complexity of natural language often limits such methods. For example, the quality of the underlying keyword dictionary may make the analysis prone to biased results, or may introduce a high number of artificial associations
Custom level of abstraction	In the hierarchical structure of GO, genes are annotated at various levels of abstraction – while this organization facilitates direct quantitative metrics for comparing functional annotation tools more objectively than when using keywords, it can also lead to ambiguous, incomplete, or redundant annotations
Terminology	Several systems rely on dictionaries compiled from a combination of different hierarchies and ontologies, thus helping enhance results and compatibility with other tools. Nevertheless, ontologies and dictionaries do not always correspond exactly with the personal terminology used by an author of a scientific article
Annotation quality	Not all genes have been annotated with GO terms, annotations may not be always up-to-date, and annotation quality depends on a curator's expertise or on a method's accuracy
Implementation	TM, as well as ontology-based systems and their combinations, can become complex and hard to maintain, especially with respect to coordinating the updates of the underlying resources
Pre-computation	A drawback of literature-based methods is that unless pre-processing is carried out, the large volumes of information make it difficult to develop online, interactive applications, such as web-based tools, services, or desktop applications. The use of GO-terms in this perspective can have advantages

Substantially more research efforts are needed to develop useful, generic benchmarks for functional annotations – particularly, since it is clear that this will be a difficult goal to achieve.

4.1. Limitations to overcome

One key issue is lack of annotation completeness, since the current state of knowledge for any specific biological process or function is generally incomplete. For example, even for human and well-studied model organisms, only a fraction of all genes currently has GO annotations (see also Supplement, Part [B]). This fact imposes severe restrictions on the ability to objectively assess the performance of functional prediction methods, and to compare tools. Thus it is typically the case that, when assessing functional prediction methods, only an estimated, lower bound on accuracy can be used to describe performance [5].

Comparison of annotation tools is also complicated due to the different database or dictionary resources upon which they are built. In addition, the presentation of results from GO-based methods may be very different compared to that of keyword-based methods. Finally, each system can retrieve only information related to the source databases or dictionaries they are built upon. Thus, different systems may characterize different aspects of the functions associated with a gene set.

Moreover, GO- and literature-based methods can pose computational challenges, and incompatibility issues arise when comparing them. For example, different GO-based methods may use a set of terms from its hierarchy that is less specific than that of another method. In that respect, compared with literature-based methods, GO shows computational advantages (Table 2) as it can facilitate quantitative metrics for direct comparison (e.g., counting common nodes present among different annotations, or measuring distances within the hierarchy). While calculating such metrics can be simple (e.g., counting of common ancestors in the hierarchy, or using graph-theory to compute the length of the shortest path between terms), more advanced semantic similarity measures have been developed that rely not only on the organization of the hierarchy (e.g., measuring distance from the root of the graph), but may also utilize probabilities to weight nodes and edges based on content-information extracted from the literature or the GO-topology itself (e.g., [68,69]).

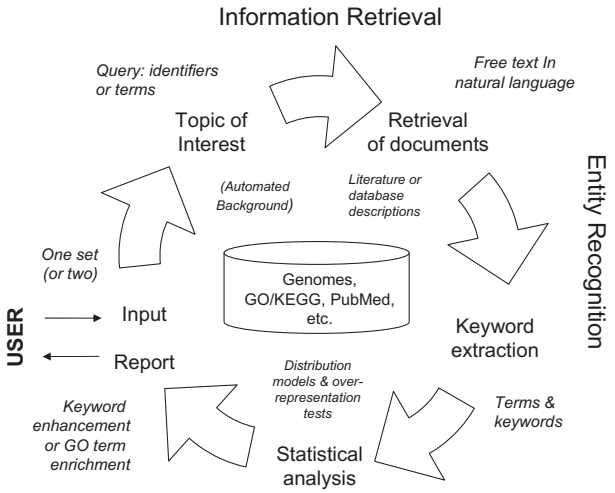

Fig. 2. Major components of literature-aided functional annotation tools. The user may specify as input the topic(s) of interest, typically genes. When a second set is not specified, the systems automatically assign a reference set to compare with. The input (usually expressed in terms of words or identifiers) is then translated into a list of matching (categories of) documents. In turn, their context is represented via the underlying dictionary terms contained in the text of the matched documents and significant keywords are compiled in a report that summarizes the findings of the annotation.

Still, comparing GO terms with literature keywords remains difficult and may need human involvement ([23,52]). Sometimes, categorical evaluations regarding the results, the computational performance, or the features of each annotation method are used

Table 6
Key reasons for unreliable judgements without benchmarking. Comparing keyword enhancement methods with each other is not straightforward because each relies on keywords from different sources and using directly the terms for assessing performance poses several analytical challenges.

Qualitative	
Semantics of terms	In many cases, terms isolated from their context can have unspecific and ambiguous meaning. Clear and potential synonyms may help overcome context-dependent interpretation in these cases
Term redundancy	Different terms may refer to the same or similar processes – redundancy can have an impact on a tool's evaluation since it is possible that certain functional concepts may be over or under-represented. Some such cases can be managed by mapping two different terms to a single concept, when both terms refer to the same external database record; but for many redundant terms, no such cross-references are available. Alternatively, automated mapping of identified keywords onto a standard ontological structure may help in describing results within a common functional space
Term-relationships	Direct counting may be misleading because many terms are semantically related with each other (particularly within hierarchies). In comparison to free text keywords, quantifying such relationships with GO is more straightforward
Quantitative	
True positives	Identifying true positives can be difficult: (a) prior knowledge whether a term can be considered as correct, or not, is necessary; (b) in general, it is not always the case that such knowledge is available; and (c) this knowledge may include subjective evaluations
False positives	Comparisons based only on false positives can be unfair as some systems may give no results – *i.e.*, no terms
Precision	When measured on the actual number of terms extracted from each method does not depict appropriately how successful the method has been, because one method may have retrieved more terms than another one. However, few successful terms may describe better the biological processes
Recall	Generally cannot be computed, as the number of false negative terms is often unbounded

to assess quality instead. Other times, evaluations may be based upon a previously published data set with annotations (e.g., [38] or [39]), but the current lack of a standard methodology for quantitative assessment of such methods leaves some comparison studies open to the criticism of subjectivity (Table 6).

4.2. Features of a gold standard

As discussed above, direct comparison of annotation tools is not credible as the function space over which the evaluation is performed (e.g., GO terms or other dictionary compilations) is different for each program (Table 6). While the lack of high-quality 'gold standards' is a generic issue [70], to the knowledge of this and other previous work (e.g., [23], [38] or [39]), there is no accepted data set upon which a fair comparison among different keyword enhancement systems can take place. The issue has been dealt with multiple times by developing ways that produce mappings between different systems of keywords helpful to serve as common platform to compare and re-annotate heterogeneous gene lists (e.g., [71] or [72]).

Another proposed solution for encoding annotations in a structured format suggests that the extracted from each method terms should be projected to those of an accepted benchmark that describes the minimum performance of a method with respect to a specified data set [23]. This would allow comparisons and data exchange be based on the same reference and be described in a common '*level-play field*' by unifying terms regarding 'quality' and 'definition', and by describing results without redundancies. To achieve that, the definition of a benchmark should consider the following aspects:

- *Benchmark-terms*. The benchmark should provide a definition of the minimum requirements (terms expected to be extracted by any tool), described in a common space of terms in order to assess upon. These benchmark-terms should be selected carefully so that they have specific and unambiguous meaning, and so that they are all exclusively related to the task at hand. Based on these non-redundant concepts true and false positives can be counted and terms without specific biological meaning or relevance for the examined data set can be discarded.
- *Coverage of content*. The construction of an annotation benchmark should be primarily guided by its purpose to be used as a common reference. For this it should not be restricted to the completeness of results that only a subset of tools could achieve. Although some terms may not have chance be mapped in this way, the benchmark should indicate a minimum performance estimate allowing both for quantification of results and comparison of methods.

While attention to such considerations should be paid for the creation of a Benchmark, Martini [23] proposed using a *Benchmark Table* and a methodology that allowed comparing different tools using a specific *Cell Cycle* dataset of human genes (Fig. 3). In specific, this example offers the mentioned benefits (level-playing-comparisons, defining min-performance, counting non-redundant matches) and the use of its Benchmark Table also facilitates further analytical computations, as one can systematically describe the results with quantitative measures (like precision and recall) and effectively compare.

The usual approach without benchmarking is annotating whether the identified keywords are indeed correct or relevant (Fig. 3). However, comparing performance across tools using measurements that rely only on the results of each (such as number of terms) – independent from each other – are not fair to compute and in many cases the result-terms should be 'normalized' (Table 6). Instead, with a benchmark, one can describe for each system individually how many from the expected result-terms (defined in the benchmark) have been identified, and thus quantitatively assess the performance (Fig. 3). Table 7 summarizes the features of the benchmark mechanism and how it helps quantify the performance of each method without bias, allowing calculate both precision and recall.

The certain approach followed in Martini [23] makes it thus a quite good example as it satisfies the criteria mentioned above, although differences between hierarchical, strictly defined terms, and non-hierarchical, context-free keywords with respect to their mapping onto the benchmark-terms were not explored in detail. Moreover, the organization of Martini's Cell Cycle Benchmark Table [23] is one only such example of how the benchmark-terms can be used to define a 'score-card' in this way.

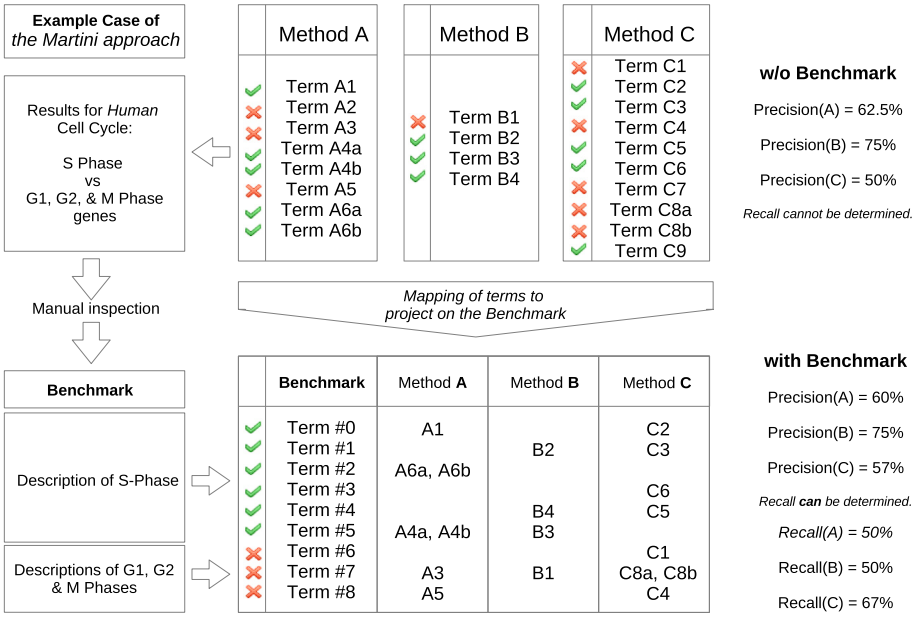

Fig. 3. Measuring performance with and without benchmark. Assessment without benchmark (upper middle and right part) can be misleading. On the contrary, benchmarking (lower middle and right part) can be more descriptive: notably the recall measure is also essential for a fair comparison as it shows 'how much of the expected knowledge has been identified' by each system, next to the precision that shows 'how much of the retrieved knowledge has been expected (correct)'. While defining a set of true negatives is difficult (since they can be uncountable), the Martini [23] approach (left part) overcomes this problem when testing for one cell cycle phase by using the matches to descriptions of the other phases as false positives.

Table 7
Benchmarking for objective comparison. Benchmarking can be critical for the fair comparison of different systems in a uniform way; measures for reliable quantification of performance can be facilitated this way.

Feature	Description
Non-automated definition	Best approach when creating a benchmark is that the expected terms to be found significant for the examined data set be identified and listed manually. Extracting terms in an automated way from various resources would not comply with the principles that feature a benchmark: content should be selected so as to facilitate the opportunity that any system assessed against it has chance to identify any of the benchmarked concepts
Semantics of terms	The terms of the compiled list of benchmark-terms should be carefully selected so that they have specific meaning and so that they are all exclusively related to the topic at hand (e.g., the Cell Cycle Phases in [23]) – for a benchmark to be as appropriate as possible, the mapping should either consider the contexts in which a keyword has been extracted from, or be unambiguous (*i.e.*, rely on terms that have a clearly defined meaning, independent of context)
Term-redundancy	Term-redundancy is removed by projecting the keywords of each method on the terms of the Benchmark Table. This minimizes redundant representation and brings the comparison over a 'level playing field'
Common functional space.	Mapping of the identified keywords onto the same ontological structure, *i.e.*, the Benchmark Table, assists in both an objective but also fair comparison. This happens by bringing the quantitative measures to rely on a uniform background: multiple terms from the extracted results may map to the same term from the Benchmark Table. This makes the comparison rely on a same 'unit' – the non-uniform results of different systems are projected to the shared, common and closed space of the terms listed in the Benchmark
Analytical benefits	Key benefit is that individual precision and recall for each candidate method can be calculated with respect to the requirements defined by the Benchmark Table. This can facilitate an objective comparison among tools, unaffected from the individual advantages and disadvantages of each method
Precision	With the projection to the benchmark table, the terms of the different methods are brought to a common functional space: e.g., precision should not be defined as the number of true positives divided by the total numbers of terms retrieved (since non-mapped terms have to be discarded and multiple true positives may map to the same term); more 'properly', precision calculation can be equivalently defined instead by the ratio between the number of true positives and the sum of the true and false positives, this time as defined over the uniform and concrete space of the benchmark-terms
Recall	Benchmarking allows facilitating the recall measure as well, i.e., ratio between the number of true positive terms and the number of terms expected to be extracted

4.3. Downsides and further implications

Even when a benchmark is available, it is required that oversights that can result in imbalanced performance estimates are avoided. For example, when the compared tools are applied with their default settings (e.g., statistical configuration or selection of concept types to be used) some results may not favor equivalence of the analysis. Other difficulties, such as context implications and unambiguous mapping of terms to the benchmark, also support the view that such a methodology should be nurtured further.

While the benchmark-terms are clearly defined in meaning, the keywords extracted by the various tools remain mostly puzzling: whether it is appropriate to map a keyword to a given benchmark-term is not always straightforward, just as in previous times some of the decisions about what terms are correct and informative, or not, may appear somewhat arbitrary due to context-dependent interpretation and the different nature of terms extracted per method. For example, in GO there are terms that belong to different levels of abstraction but belong to the same process while simultaneously different free text terms may be represented by

Table 8
Human intervention for better benchmarking. The help of experts may be involved in benchmark tasks in order to be able to assess the accuracy of a method appropriately: from creating content, to its organization and the detailed handling of terms (selection, interpreting meaning, mapping and projection).

Aspect of benchmark	Discussion
Topic and content	A benchmark data set must be acceptable by the broader life sciences community and also be able to overcome the complexities discussed. For example, the proposed data set by [23] met key criteria: - Human cell cycle is both sufficiently well studied and of interest to a large spectrum of life scientists - The human cell cycle gene set is well characterized either with GO terms or with extended literature - Nevertheless, more data sets have to be proposed in order to satisfy a broader range of scope efficiently. For that, community based expertise and knowledge should be incorporated
Size and completeness	Lack of broad coverage by neglecting to compensate for possible spread of results may prevent performing a level play comparison as not all methods may end up having an equal opportunity to achieve maximal recall. Whilst most tools to be compared are assumed to be able to identify benchmark-terms, it may be the case that none identifies all of them. Also, it is difficult to grasp how many keywords are missed when there is no way of knowing what the annotations are that have been missed. Human experts can help identify these cases and constrain such pitfalls or flaws
Context interpretation and mapping of terms	Curation may have to take place for the projection of the terms of each method to the common level playing field that is defined by the benchmark. This is because in general keyword enhancement strips extracted terms from their context and a given keyword that may appear significant (and be mapped to a benchmark-term), may when context is included not be significant at all. In comparison the strength of GO is that the meaning of its terms is very strictly defined so that mappings to benchmark-terms are not in doubt. Failure to account for such idiosyncrasies could lead to overstating the performance of a tool
Objectivity via criticism	The main benefit from manually examining results relies on a human's expertise. However, when only one annotator is available an assessment may be subjective no matter how explicit some criteria may be. Reasons for that include that these criteria may eventually be relatively unclear, as their interpretation can vary considerably from one annotator to another, e.g., one may be 'optimist', 'pessimist' or neutral. This observation underscores the importance of using several independent annotators
Up-to-datedness	A benchmark cannot remain static as the knowledge over a certain field expands. Human experts should adapt content regularly

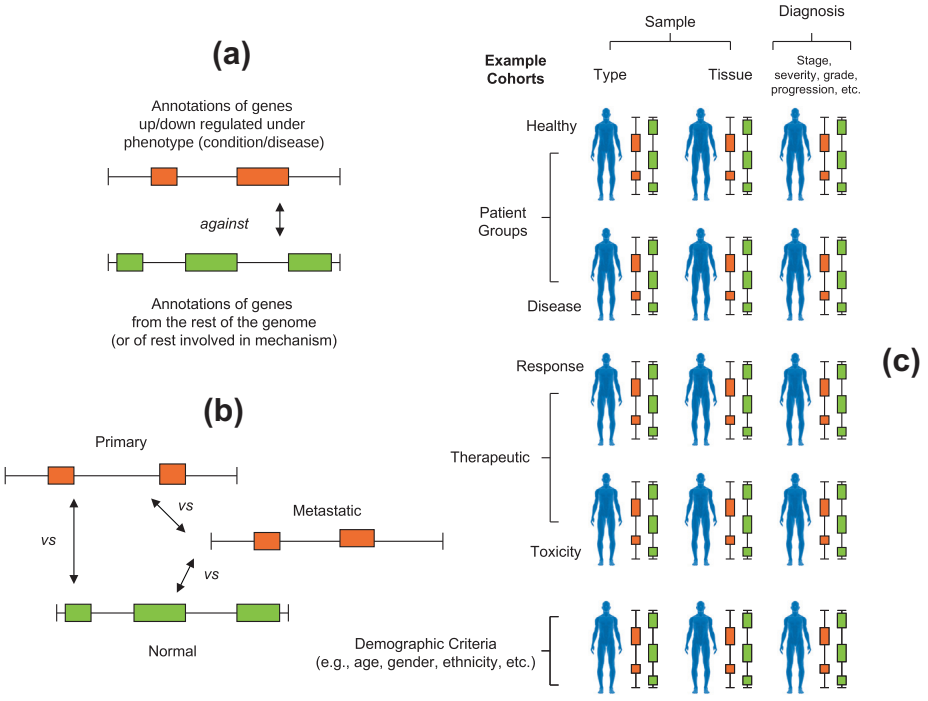

Fig. 4. Personalized approaches file clinical annotation within the rest of the modern genomics scene. Functional annotation this time embodies not (a) a single gene-set vs a 'representative' human genome comparison at a time, but (b) multiple ones that may be considered for each patient individually or(c) across larger population cohorts – e.g., gene groups observed in different tumour variations, gene profiles with somatic vs germline mutations, or driver vs passenger mutation carrier genes can be further compared – beyond the residue level – via enrichment association studies.

the same GO term. In practice, such an effect is much more acute for keywords extracted from the literature since the clearly defined GO terms can possibly be mapped 'more unambiguously' onto a benchmark, whereas for other keywords (especially when stripped off their context) there may be doubt.

While automatically mapping terms to the benchmark can be an ambiguous process due to the difficulty coping with the semantic interpretation of terms, yet another aspect of reducing the number of terms to the canonical 'benchmark set' is that it is a much smaller number of terms. Thus, it is possible that the reported counts used to compute precision and recall are biased. For these reasons, Martini [23] proposes that the mapping of the terms (extracted by a method) to those of the benchmark should be performed by an expert. Although, there are considerations with respect to objectivity of human-based annotations, curation by experts can be considered as one from the best ways to perform the mapping to the benchmark-terms (Table 8).

5. Discussion

Automated functional annotation methods are important since function prediction can be used to give a focused direction towards verification experiments. While gene annotation is mostly used to

refer to the process of consolidating various sources of information regarding gene function into a concise record for each gene, this work focused also on the closely related and equally important problem in modern biology, namely the analysis of groups of genes produced from high-throughput experiments (e.g., microarrays, etc.). The importance of addressing these problems (gene annotation and functional inference of gene sets) is indicated by the large variety of methodologies proposed to address them (Table 1).

5.1. Modern and future dimensions: clinical (meta-) annotation

Although the functional annotation task has relied on a plethora of methods and data sources, yet it is still in an actively improving and growing state today, as similar techniques can be used not only for inferring function (in the form of terms or biological pathways and interactions, like from Martini [23], FatiGO [37], PANTHER [45] or Reactome [46] analysis) but also to annotate phenotypes (medical conditions, diseases, side effects) or even compounds (e.g., Alkemio [73] or Metab2MeSH [74]). This is not only attributable to the modern biomedical research relying increasingly on the combination of disciplines previously considered as distinct (medicine, chemistry, and biology), but also to the rise of the personalized '-omics' era. As personalized medicine promises using genomic profiles of individuals to assign the right treatment, the focus shifts from the functional impact of single variants (e.g., as discussed in [75]) to elements from the annotation paradigm again (Fig. 4): expectations suggest that we need to start asking not only what is in a genome but also for such composite reports that can quickly raise some clinically oriented awareness regarding observed gene clusters.

Though there are still several scientific and technical limitations that make the direct implementation of such strategies unfeasible, modern treatment decision support systems for precision medicine (based already today on genomic analyses of patients; e.g., [76]) often compare tumors with both control tissue and the available database information to provide an output that consists of prioritized lists of genes or treatment regimens – in turn, these may be analyzed in follow-up experiments by researchers or in clinical studies by physicians to test their actual role in a person's cancer. Translating such findings in 'holistic' knowledge and distilled 'reductionist' views [70] is not yet straightforward and requires expert domain rules. Thus, broader annotation questions rise and with the advent of personal genomics identifying the systems that are perturbed (and their entities) by gene-sets of specific aberration profiles becomes not only highly of interest but soon also the norm. Integral part among these developments, modern functional annotation (even at the 'simplified' gene level) remains a key player with central role in the advances towards this direction.

5.2. Clash of annotations: democracy or aristocracy?

Finally, the epistemological perspective taken in this work with regard to the functional annotation of genes was not only restricted to comparing the different knowledge acquisition systems used, but also considered some conflicting beliefs and opinions, especially regarding the topic of where the information comes from – with the main two at least theories revolving around core GO- and TM-based approaches. While major conflicting points cover most qualities of each (substance, forms, resources, capabilities, value and limits) the contrast lies on a dominant impression that GO refers mostly to an 'oligarchy by experts' whereas TM represents a 'vox populi' approach for annotation. The former argue that curated GO annotations are better than automated TM inferences, and emphasize that the latter may coincidentally identify terms (which just arise by chance and are noisy and uninformative) or can sometimes be based on literature that is only general (and not specific to a gene or function). Arguments against a strictly GO-based annotation strategy (curated or inferred) include that it remains both limited in extent and restricted in scope, justifying claims from scientists such as that "sometimes feels as if the annotation stopped half-way through". Indeed, both observations apply as TM approaches can potentially give better recall (potentially retrieving all relevant literature descriptions of a protein's function), whereas GO terms are likely to be more precise (delivering generally fewer but more reliable annotations). While we have contrasted the differences between these two approaches in this work, they can also be perceived as complementary rather than competing, and indeed we believe that the most accurate and useful methods would ultimately combine the strengths of both approaches.

5.3. Concluding remarks

Although performing an exhaustive overview of produced genomic data to gain understanding is frequently not possible, current analytical methods for functional annotation and careful integration continue to improve our ability to extract knowledge and generate new hypotheses. Inspired by these developments this article highlights that:

- Recent tools rely on synthesized integration towards a 'systems' perspective. Modern computational methods for functional annotation seek to aid understanding about the synergistic biological functions that genes perform together. An emerging trend to do this is by elucidating large-scale systems and effects, revealing connections to system components, especially molecular actors such as proteins, other genes, and chemicals. Doing this in a systemic way requires an extensive vocabulary of terms that represent a broad ensemble of clearly defined biomedical concepts. To improve performance of functional annotation in this respect many tools incorporate keywords that rely on a plethora of dictionaries derived from multiple resources (e.g., GO, pathways, etc.) enabling thus keywords to belong in a broad range of semantic categories (Section 2.4).
- TM plays a central role in the task of functional annotation. Many methods mine the text of database annotations and of the literature to extract annotations (Section 2.2). In comparison, mining of literature can provide more sensitive results and has better potential capturing specific keywords that describe gene function because it is more information rich in content (Section 3; Table 5). Finally, data suggest more than one ways forward (Supplement, Part [B]) and annotation improvements can be accomplished when the mining is combined with homology based expansion and sequence similarity ranking.
- Main methods combine mining of literature and the use of ontologies. Studying gene properties using information extracted directly from the biomedical literature, or from the GO, still poses drawbacks (Table 5). While compiling a structured vocabulary from unstructured or semi-structured resources remains a challenge today, typically, many tools derive their dictionary primarily from GO, while others rely on literature itself in order to extract keywords (Section 3). Yet, combining the powers of GO-based approaches with the increased sensitivity of keywords extracted directly from the literature has been shown to be a viable solution that can be well performing (e.g., [23]).
- Two set comparison can help addressing more interesting questions. Commonly, for the functional annotation of gene lists, keyword enhancement is applied, often for genes from only a limited number of organisms (see Table 4). To improve functional annotation performance, using two sets for input allows asking questions that are more specific by enabling users to

control the reference set (see Section 2.4), and comes with further data management advantages (e.g., handling smaller volumes of data).

- Existence of proper benchmarks can help objectively assess functional annotation performance. In spite of its downsides, it is emphasized that using benchmarks like that of [23] presents analytical opportunities for objective (non-biased and quantifiable) performance comparison of results over a common functional space that is otherwise non-applicable (Section 4.2). However, the data set of Martini's experience [23] is one only such example and very specific. What is required is a set of reliable benchmarks tailored specifically for both qualitative and quantitative evaluation, ideally spanning a wide range of functions and organisms.

Acknowledgement

We gratefully acknowledge the help of Christian Stolte in finalising figures for this work.

Appendix A. Supplementary data

Supplementary data associated with this article can be found, in the online version, at http://dx.doi.org/10.1016/j.ymeth.2014.07.004.

References

[1] I. Vlahavas, P. Kefalas, N. Bassiliades, I. Refanidis, F. Kokkoras, H. Sakellariou, Artificial Intelligence, first ed., Gartaganis Publications, 2002.
[2] P. Khatri, S. Drăghici, Bioinformatics 21 (18) (2005) 3587–3595.
[3] T.N. Villavicencio-Diaz, A. Rodriguez-Ulloa, O. Guirola-Cruz, Y. Perez-Riverol, Curr. Top. Med. Chem. 14 (3) (2014) 435–449.
[4] D.W. Huang, B.T. Sherman, R.A. Lempicki, Nucleic Acids Res. 37 (1) (2009) 1–13.
[5] D.M.A. Martin, M. Berriman, G.J. Barton, BMC Bioinformatics 5 (2004) 178.
[6] G. Zehetner, Nucleic Acids Res. 31 (13) (2003) 3799–3803.
[7] S. Hennig, D. Groth, H. Lehrach, Nucleic Acids Res. 31 (13) (2003) 3712–3715.
[8] J.A. Gerlt, P.C. Babbitt, Can sequence determine function? Genome Biol. 1(5) (2000) REVIEWS0005.
[9] M.A. Andrade, N.P. Brown, C. Leroy, S. Hoersch, A. de Daruvar, C. Reich, A. Franchini, J. Tamames, A. Valencia, C. Ouzounis, C. Sander, Bioinformatics 15 (5) (1999) 391–412.
[10] M.L. Riley, T. Schmidt, C. Wagner, H.-W. Mewes, D. Frishman, The PEDANT genome database in 2005, Nucleic Acids Res. 33(Database issue) (2005) D308–D310.
[11] Medical Subject Headings – (MeSH Home Page). [Online]. Available: http://www.nlm.nih.gov/mesh/,2010 (accessed 19.09.10).
[12] M. Ashburner, C.A. Ball, J.A. Blake, D. Botstein, H. Butler, J.M. Cherry, A.P. Davis, K. Dolinski, S.S. Dwight, J.T. Eppig, M.A. Harris, D.P. Hill, L. Issel-Tarver, A. Kasarskis, S. Lewis, J.C. Matese, J.E. Richardson, M. Ringwald, G.M. Rubin, G. Sherlock, Nat. Genet. 25 (1) (2000) 25–29.
[13] Entrez Gene. [Online]. Available: http://www.ncbi.nlm.nih.gov/gene, 2010 (accessed 12.10.10).
[14] E. Eden, R. Navon, I. Steinfeld, D. Lipson, Z. Yakhini, BMC Bioinformatics 10 (2009) 48.
[15] T. Ruths, D. Ruths, L. Nakhleh, Bioinformatics 25 (9) (2009) 1178–1184.
[16] A. Bresell, B. Servenius, B. Persson, Appl. Bioinformatics 5 (4) (2006) 225–236.
[17] Gene Ontology Tools. [Online]. Available: http://www.geneontology.org/GO.tools.shtml, 2014 (accessed 23.03.14).
[18] Gene Ontology Tools – NeuroLex. [Online]. Available: http://neurolex.org/wiki/Category:Resource:Gene_Ontology_Tools, 2014 (accessed 23.03.14).
[19] PubMed. [Online]. Available: http://www.ncbi.nlm.nih.gov/pubmed, 2010 (accessed 19.09.10).
[20] P. Glenisson, B. Coessens, S. Van Vooren, J. Mathys, Y. Moreau, B. De Moor, Genome Biol. 5 (6) (2004) R43.
[21] K.G. Becker, D.A. Hosack, G. Dennis, R.A. Lempicki, T.J. Bright, C. Cheadle, J. Engel, BMC Bioinformatics 4 (2003) 61.
[22] B.T.F. Alako, A. Veldhoven, S. van Baal, R. Jelier, S. Verhoeven, T. Rullmann, J. Polman, G. Jenster, BMC Bioinformatics 6 (2005) 51.
[23] T.G. Soldatos, S.I. O'Donoghue, V.P. Satagopam, L.J. Jensen, N.P. Brown, A. Barbosa-Silva, R. Schneider, Nucleic Acids Res. 38 (1) (2010) 26–38.
[24] M.D. Robinson, J. Grigull, N. Mohammad, T.R. Hughes, BMC Bioinformatics 3 (2002) 35.
[25] G. Lu, T.V. Nguyen, Y. Xia, M. Fromm, AffyMiner: mining differentially expressed genes and biological knowledge in GeneChip microarray data, BMC Bioinformatics 7(Suppl. 4) (2006) S26.
[26] A. Hsiao, T. Ideker, J.M. Olefsky, S. Subramaniam, VAMPIRE microarray suite: a web-based platform for the interpretation of gene expression data, Nucleic Acids Res. 33(Web Server issue) (2005) W627–632.
[27] B.R. Zeeberg, H. Qin, S. Narasimhan, M. Sunshine, H. Cao, D.W. Kane, M. Reimers, R.M. Stephens, D. Bryant, S.K. Burt, E. Elnekave, D.M. Hari, T.A. Wynn, C. Cunningham-Rundles, D.M. Stewart, D. Nelson, J.N. Weinstein, BMC Bioinformatics 6 (2005) 168.
[28] P. Khatri, S. Draghici, G.C. Ostermeier, S.A. Krawetz, Genomics 79 (2) (2002) 266–270.
[29] P. Khatri, C. Voichita, K. Kattan, N. Ansari, A. Khatri, C. Georgescu, A.L. Tarca, S. Draghici, Onto-Tools: new additions and improvements in 2006, Nucleic Acids Res. 35(Web Server issue) (2007) W206–211.
[30] G. Dennis, B.T. Sherman, D.A. Hosack, J. Yang, W. Gao, H.C. Lane, R.A. Lempicki, Genome Biol. 4 (5) (2003) P3.
[31] V. Beisvag, F.K.R. Jünge, H. Bergum, L. Jølsum, S. Lydersen, C.-C. Günther, H. Ramampiaro, M. Langaas, A.K. Sandvik, A. Laegreid, BMC Bioinformatics 7 (2006) 470.
[32] F. Al-Shahrour, P. Minguez, J.M. Vaquerizas, L. Conde, J. Dopazo, BABELOMICS: a suite of web tools for functional annotation and analysis of groups of genes in high-throughput experiments, Nucleic Acids Res. 33(Web Server issue) (2005) W460–464.
[33] R. Sharan, A. Maron-Katz, R. Shamir, Bioinformatics 19 (14) (2003) 1787–1799.
[34] S. Maere, K. Heymans, M. Kuiper, Bioinformatics 21 (16) (2005) 3448–3449.
[35] J. Cheng, S. Sun, A. Tracy, E. Hubbell, J. Morris, V. Valmeekam, A. Kimbrough, M.S. Cline, G. Liu, R. Shigeta, D. Kulp, M.A. Siani-Rose, Bioinformatics 20 (9) (2004) 1462–1463.
[36] P. Khatri, M. Sirota, A.J. Butte, PLoS Comput. Biol. 8 (2) (2012) e1002375.
[37] F. Al-Shahrour, R. Díaz-Uriarte, J. Dopazo, Bioinformatics 20 (4) (2004) 578–580.
[38] N. Qiao, Y. Huang, H. Naveed, C.D. Green, J.-D.J. Han, PLoS One 8 (9) (2013) e74074.
[39] S. Lv, Y. Li, Q. Wang, S. Ning, T. Huang, P. Wang, J. Sun, Y. Zheng, W. Liu, J. Ai, X. Li, J. R. Soc. Interf. 9 (70) (2012) 1063–1072.
[40] T. Beissbarth, T.P. Speed, Bioinformatics 20 (9) (2004) 1464–1465.
[41] M. Gosink, S. Khuri, C. Valdes, Z. Jiang, N.F. Tsinoremas, Adv. Bioinformatics 2011 (2011) 271563.
[42] I. Coulibaly, G.P. Page, Int. J. Plant Genomics 2008 (2008) 893941.
[43] A.V. Antonov, T. Schmidt, Y. Wang, H.W. Mewes, ProfCom: a web tool for profiling the complex functionality of gene groups identified from high-throughput data, Nucleic Acids Res. 36(Web Server issue) (2008) W347–351.
[44] R. Jelier, J.J. Goeman, K.M. Hettne, M.J. Schuemie, J.T. den Dunnen, P.A.C. 't Hoen, Literature-aided interpretation of gene expression data with the weighted global test, Brief. Bioinformatics 12(5) (2011) 518–529.
[45] H. Mi, A. Muruganujan, P.D. Thomas, Nucleic Acids Res. 41 (D1) (2013) D377–D386.
[46] D. Croft, A.F. Mundo, R. Haw, M. Milacic, J. Weiser, G. Wu, M. Caudy, P. Garapati, M. Gillespie, M.R. Kamdar, B. Jassal, S. Jupe, L. Matthews, B. May, S. Palatnik, K. Rothfels, V. Shamovsky, H. Song, M. Williams, E. Birney, H. Hermjakob, L. Stein, P. D'Eustachio, The Reactome pathway knowledgebase, Nucleic Acids Res. 42(Database issue) (2014) D472–D477.
[47] R. Frijters, B. Heupers, P. van Beek, M. Bouwhuis, R. van Schaik, J. de Vlieg, J. Polman, W. Alkema, CoPub: a literature-based keyword enrichment tool for microarray data analysis, Nucleic Acids Res. 36(Web Server issue) (2008) W406–W410.
[48] P. Minguez, F. Al-Shahrour, D. Montaner, J. Dopazo, Bioinformatics 23 (22) (2007) 3098–3099.
[49] WHO | The Anatomical Therapeutic Chemical Classification System with Defined Daily Doses (ATC/DDD). [Online]. Available: http://www.who.int/classifications/atcddd/en/, 2010 (accessed 23.03.10).
[50] MedDRA. [Online]. Available: http://www.meddramsso.com/, 2010 (accessed 23.09.10).
[51] Q. Wang, J. Sun, M. Zhou, H. Yang, Y. Li, X. Li, S. Lv, X. Li, Y. Li, Bioinformatics 27 (11) (2011) 1521–1528.
[52] R. Küffner, K. Fundel, R. Zimmer, Expert knowledge without the expert: integrated analysis of gene expression and literature to derive active functional contexts, Bioinformatics 21(Suppl. 2) (2005) ii259–ii267.
[53] A.M. Cohen, W.R. Hersh, Brief. Bioinformatics 6 (1) (2005) 57–71.
[54] L.J. Jensen, J. Saric, P. Bork, Nat. Rev. Genet. 7 (2) (2006) 119–129.
[55] Z. Lu, PubMed and beyond: a survey of web tools for searching biomedical literature, Database (Oxford), vol. 2011.
[56] J. Lewis, S. Ossowski, J. Hicks, M. Errami, H.R. Garner, Bioinformatics 22 (18) (2006) 2298–2304.
[57] T.G. Soldatos, S.I. O'Donoghue, V.P. Satagopam, A. Barbosa-Silva, G.A. Pavlopoulos, A.C. Wanderley-Nogueira, N.M. Soares-Cavalcanti, R. Schneider, BioData Min. 5 (1) (2012) 1.
[58] J.-F. Fontaine, A. Barbosa-Silva, M. Schaefer, M.R. Huska, E.M. Muro, M.A. Andrade-Navarro, MedlineRanker: flexible ranking of biomedical literature, Nucleic Acids Res. 37(Web Server issue) (2009) W141–W146.
[59] G.L. Poulter, D.L. Rubin, R.B. Altman, C. Seoighe, BMC Bioinformatics 9 (2008) 108.
[60] F.M. Ortuño, I. Rojas, M.A. Andrade-Navarro, J.-F. Fontaine, BMC Bioinformatics 14 (2013) 113.
[61] J.-F. Fontaine, F. Priller, A. Barbosa-Silva, M.A. Andrade-Navarro, Génie: literature-based gene prioritization at multi genomic scale, Nucleic Acids Res. 39(Web Server issue) (2011) W455–W461.
[62] M.A. Andrade-Navarro, G.A. Palidwor, C. Perez-Iratxeta, BioData Min. 5 (1) (2012) 14.

[63] M.J. Schuemie, J.A. Kors, Bioinformatics 24 (5) (2008) 727–728.
[64] P.K. Shah, C. Perez-Iratxeta, P. Bork, M.A. Andrade, BMC Bioinformatics 4 (2003) 20.
[65] Wikipedia. [Online]. Available: http://www.wikipedia.org/, 2010 (accessed 23.09.10).
[66] Reflect – Highlighting Proteins, and Small Molecule Names. [Online]. Available: http://reflect.embl.de/, 2010 (accessed 23.09.10).
[67] ∗BioCreAtIvE homepage∗. [Online]. Available: http://biocreative.sourceforge.net/, 2014 (accessed 23.03.14).
[68] P.W. Lord, R.D. Stevens, A. Brass, C.A. Goble, Bioinformatics 19 (10) (2003) 1275–1283.
[69] F.M. Couto, M.J. Silva, P.M. Coutinho, Data Knowl. Eng. 61 (1) (2007) 137–152.
[70] J.C. Fuller, P. Khoueiry, H. Dinkel, K. Forslund, A. Stamatakis, J. Barry, A. Budd, T.G. Soldatos, K. Linssen, A.M. Rajput, EMBO Rep. 14 (4) (2013) 302–304.
[71] I. Iliopoulos, S. Tsoka, M.A. Andrade, A.J. Enright, M. Carroll, P. Poullet, V. Promponas, T. Liakopoulos, G. Palaios, C. Pasquier, S. Hamodrakas, J. Tamames, A.T. Yagnik, A. Tramontano, D. Devos, C. Blaschke, A. Valencia, D. Brett, D. Martin, C. Leroy, I. Rigoutsos, C. Sander, C.A. Ouzounis, Bioinformatics 19 (6) (2003) 717–726.
[72] A.J. Pérez, C. Perez-Iratxeta, P. Bork, G. Thode, M.A. Andrade, Bioinformatics 20 (13) (2004) 2084–2091.
[73] J.A. Gijón-Correas, M.A. Andrade-Navarro, J.F. Fontaine, Alkemio: association of chemicals with biomedical topics by text and data mining, Nucleic Acids Res. (2014).
[74] M.A. Sartor, A. Ade, Z. Wright, D. States, G.S. Omenn, B. Athey, A. Karnovsky, Bioinformatics 28 (10) (2012) 1408–1410.
[75] the International Cancer Genome Consortium Mutation Pathways and Consequences Subgroup of the Bioinformatics Analyses Working Group, Computational approaches to identify functional genetic variants in cancer genomes, Nat. Meth. 10(8) (2013) 723–729.
[76] MolecularHealth – Putting the Person in Personalized Medicine. [Online]. Available: http://www.molecularhealth.com/, 2014 (accessed 15.02.14).