OXFORD

Briefings in Bioinformatics, 2016, 1–10

doi: 10.1093/bib/bbw036
Paper

Ontology-based annotations and semantic relations in large-scale (epi)genomics data

Eugenia Galeota and Mattia Pelizzola

Corresponding author. Mattia Pelizzola, Center for Genomic Science of IIT@SEMM, Fondazione Istituto Italiano di Tecnologia (IIT), Via Adamello 16, 20139 Milan, Italy. Tel.: 0039-0294375058; E-mail: mattia.pelizzola@iit.it

Abstract

Public repositories of large-scale biological data currently contain hundreds of thousands of experiments, including high-throughput sequencing and microarray data. The potential of using these resources to assemble data sets combining samples previously not associated is vastly unexplored. This requires the ability to associate samples with clear annotations and to relate experiments matched with different annotation terms. In this study, we illustrate the semantic annotation of Gene Expression Omnibus samples metadata using concepts from biomedical ontologies, focusing on the association of thousands of chromatin immunoprecipitation followed by high-throughput sequencing (ChIP-seq) samples with a given target, tissue and disease state. Next, we demonstrate the feasibility of quantitatively measuring the semantic similarity between different samples, with the aim of combining experiments associated with the same or similar semantic annotations, thus allowing the generation of large data sets without the need of additional experiments. We compared tools based on Unified Medical Language System with tools that use topic-specific ontologies, showing that the second approach outperforms the first both in the annotation process and in the computation of semantic similarity measures. Finally, we demonstrated the potential of this approach by identifying semantically homogeneous groups of ChIP-seq samples targeting the Myc transcription factor, and expanding this data set with semantically coherent epigenetic samples. The semantic information of these data sets proved to be coherent with the ChIP-seq signal and with the current knowledge about this transcription factor.

Key words: semantic annotation; semantic similarity; epigenetics; transcription factor; high-throughput sequencing; natural language processing

Introduction

High-throughput biological data sets available in large-scale public repositories constitute a valuable resource to investigate diseases and phenotypes in different organisms and biological conditions. Public repositories such as Gene Expression Omnibus (GEO) [1] and Array Express [2] include dozens of thousands of samples profiling transcriptional and epigenetic patterns and the binding of various regulatory proteins in entire genomes or a large fraction thereof. Integrative analysis of such data could shed light on the epigenetic and transcriptional patterns that lead to various diseases with respect to normal or healthy conditions [3].

Genes' transcriptional activity results from the synergistic action of multiple regulatory cues, including transcription factors (TFs) binding and epigenetic marks. Consequently, to either study the transcriptional patterns observed in a specific biological context of interest, or to advance our understanding in the intricate interplay between various epigenetic and regulatory marks, it is always highly desirable to acquire multi-layered comprehensive

Eugenia Galeota is a postdoc at the Center for Genomic Science of IIT@SEMM, Fondazione Istituto Italiano di Tecnologia. Her research interests include data mining and integrative analysis of heterogeneous biological data and epigenetics.

Mattia Pelizzola is the Head of the Computational Epigenomics unit at the Center for Genomic Science of IIT@SEMM (Fondazione Istituto Italiano di Tecnologia), which develops computational tools to decipher the contribution of epigenetic and regulatory factors in the establishment of transcriptional regulatory programs.

Submitted: 27 January 2016; **Received (in revised form):** 17 March 2016

data sets. Ideally, these would typically include the highest possible number of TFs binding maps and epigenetic marks in multiple tissues and disease states. Clearly, the cost of generating such data sets is often out of reach of individual research groups, and is something more commonly obtained through the constitution of large consortia [4]. Nevertheless, given a specific molecular target of interest, such as the binding of a specific TF in a tissue and disease state, it is often possible to find in public large-scale databases that experiments have been already performed for numerous regulatory factors or epigenetic marks profiled in the same or in a sufficiently similar tissue and disease condition. Thus, the process of expanding the initial data set with compatible publicly available data has a great potential to increase the value of the initial data.

Crucial for this expansion process are (i) the ability of assigning formal semantic labels to public samples (semantic annotation; for example, determining the targeted protein, the investigated disease state and tissue type), and (ii) the possibility of relating different samples through these labels (semantic similarity), to identify semantically coherent experiments that are suitable to be associated to the initial data set. Using ontology-based annotations is beneficial because it allows associating text to formal concepts, thus supporting interoperability, and guarantees automatic reasoning based on the ontologies underlying structure.

The association of samples to semantic information would be greatly facilitated if their metadata were relying on controlled vocabularies or specific biomedical ontologies [5]. Metadata of microarray experiments deposited in GEO and ArrayExpress were among the first to benefit from the creation of these standards [6–8]. Many ontologies, lexicons, databases and dictionaries emerged in various domains, including the Unified Medical Language System (UMLS) Metathesaurus [9] and the Open Biomedical Ontologies (OBO) [10]. Despite the efforts in data standardization, the heterogeneity and quality of natural language textual annotations, especially in the field of epigenetics and transcriptional regulation, still represents a bottleneck in the retrieval of data, in the programmatic access and organization of samples in semantically coherent groups.

While a number of large-scale projects and consortia resulted in high-quality well-annotated data sets [11, 12], this level of curation is typically lacking in individual samples, i.e. the vast majority of the available data. Different attempts have been made to facilitate access to resources that provide results of experiments and the corresponding raw data based on their associated metadata, such as EBI Linked Data (https://www.ebi.ac.uk/rdf/), LinkedCT (http://arxiv.org/abs/0908.0567) and Bio2RDF (http://dl.acm.org/citation.cfm?id=2878554). In particular, one of the main tools is the National Center for Biomedical Ontology (NCBO) Bioportal, which provides the ability to browse, search, visualize, annotate (using the NCBO annotator) and map text with concepts from more than 300 different ontologies [13, 14]. Similarly, the tool Gemma [15] is a database and analysis system capable of meta-analyses, such as differential gene expression or co-expression, using public gene expression data sets annotated both automatically and manually with specific ontologies. None of the two resources provide a comprehensive annotation of available high-throughput data sets, as they are limited to GEO data sets and gene expression samples, respectively.

In this article, we start by providing a brief description of the state-of-the art concepts and tools for recognizing and relating ontology concepts in biomedical texts. We specifically focus on publicly available samples for the binding of TFs and epigenetic marks resulting from chromatin immunoprecipitation followed by high-throughput sequencing (ChIP-seq) experiments. We illustrate and we evaluate the semantic annotation of GEO ChIP-seq samples associating them to an immunoprecipitated target, disease and tissue information. We proceed by demonstrating how the initial query can be expanded for the identification of semantically coherent data sets. Finally, we exemplify these concepts with ChIP-seq samples targeting the Myc TF, illustrating how it is possible to study the tissue-specific binding of this TF solely based on public data, and showing how these samples can be complemented with a set of semantically coherent epigenetic marks (while a TF was chosen as an example, our approach does not have to be intended to be tailored or limited to samples targeting these regulatory proteins).

Materials and methods

Accessing public repositories metadata using SRAdb

Genome-wide maps of DNA-binding regulatory proteins and epigenetic marks are typically obtained and released as ChIP-seq experiments. Each ChIP-seq sample in GEO is linked to the Sequence Read Archive (SRA), a public repository of high-throughput sequencing (HTS) raw data and their metadata. SRAdb [16] is an R/Bioconductor package collecting SRA metadata into an SQLite database and providing the user a way to quickly query samples metadata and eventually download the associated sequence files (in the Fastq or SRA format) for local reads-alignment and downstream analysis in the R environment.

Details of interest about the experiments and samples were obtained from 'experiment_attribute' and 'sample_attribute' metadata fields. The 'sample_attribute' field contains data about the considered physical sample, in form of free-text or as a pipe-separated sentence containing SRA metadata keywords corresponding to xml tags and the corresponding submitted information.

GEO samples are not always linked to unique SRA ids. Because the metadata contained in SRAdb closely match the GEO metadata, and because SRA ids are necessary for the retrieval of the raw data, we decide to report the number of samples as SRA ids, rather than GEO samples (for example, the 21 037 SRA ids, corresponding to human and mouse ChIP-seq samples, correspond to 20 913 unique GEO sample accessions).

Identification of the immunoprecipitated target

The *tm* R package was used to match the text to genes and histone marks identifiers. The procedure first splits the sentence in tokens, removes the stop words and punctuations and finally creates a document-term matrix containing matched histone modifications or official gene symbols. Finally, gene or protein identifiers were mapped to the corresponding gene symbol (contained in Bioconductor metadata packages for mouse and human), and histone marks were associated to reference lists obtained from the Histome database (following the standard nomenclature provided in [17]).

Semantic annotation: UMLS and biomedical ontologies

In this study, we considered, and compared, two main resources to perform the semantic annotation: UMLS and specific OBO ontologies. The UMLS is an ensemble of interconnected

controlled vocabularies in the biomedical sciences [9]. It is composed of several components. Among these, the Metathesaurus consists of different terms (concepts) from different controlled vocabularies or ontologies, defined as sources. Each concept is associated to a short description and can be identified by a unique concept identifier (CUI). Metathesaurus relationships can be divided in two main categories: relationships between concepts belonging to the same source (intra-source relationships) and relationships between concepts belonging to different sources (inter-source relationships). UMLS editors have manually generated the latter during the integration process of new sources into the Metathesaurus. Most of these relationships refer to synonymous concepts; others are introduced to connect isolated concepts (with few or no relationships in their source), thus explicitly creating a mapping between different sources. While extremely useful for the interoperation of heterogeneous data, creating links between concepts from different sources is a difficult task, and the heterogeneity of the sources represents the major bottleneck.

On the other hand, while ontologies belonging to the OBO foundry are also created by experts in the field and reflect human reasoning, and while they covered a more limited set of research fields, they have important advantages in terms of orthogonality, unambiguousity, scalability and topic specificity [10]. OBO ontologies considered within this study are the BRENDA Tissue Ontology (BTO) [18], which collects a hierarchy of tissues, cell types and cell lines for different organisms, and the Disease Ontology (DO) [19], which describes human diseases and phenotypes.

Semantic annotation: associating biomedical concepts with samples metadata using Metamap and Conceptmapper

In this study, two different concept mapping tools were considered and compared with semantically annotate GEO metadata: Metamap (based on UMLS), and Conceptmapper (based on OBO ontologies). Metamap [20] is a Natural Language Processing (NLP) tool developed with the aim of annotating biomedical text with UMLS Metathesaurus concepts. The large range of input, output and processing options makes Metamap suitable for different types of text. It provides different output formats for human or machine readability, the possibility to use different UMLS subsets, the use of two data models (relaxed and strict) to decide the type of filtering on the Metathesaurus content, a word sense disambiguation module, the identification of negations and gaps in the text. The annotation process applied in this study involves different steps: the syntactic parsing of the biomedical text into noun phrases with the Specialist Lexicon; the generation of lexical variants for each noun phrase with spelling variants, abbreviations acronyms, CUI candidate retrieval and evaluation; a final mapping construction where the best combinations of Metathesaurus (USABase 2014AA version, containing ~3M distinct CUIs) concepts is reported. The final mappings are a set of concept candidates that best describe the text given as input, scored based on centrality, variation, coverage, cohesiveness and involvement. The output was requested in the MetaMap Indexing format, where individual ranked concepts are returned. Concepts are scored based on a ranking function that characterizes the power (aboutness) of a given concept for a piece of text, which is the product of a frequency factor and a relevance factor. The relevance factor takes into account the depth of the concept in the tree hierarchy, the word length, the number of characters and the Metamap mapping score.

Conceptmapper [21] is a Unstructured Information Management Architecture-based tool to annotate text with concepts from a dictionary. BO and DO OBO ontologies were converted into Conceptmapper dictionaries using the components developed at Colorado Computational Pharmacology and available at http://sourceforge.net/projects/bionlp-uima/.

For both UMLS and Conceptmapper, the performance depends on the different configuration parameters and on the ontology used, and for this study we adopted settings previously identified as recommended [22]: Conceptmapper was configured to retrieve the longest match of contiguous tokens matching the text in the dictionary, using the PORTER stemming algorithm and considering all synonym types; for UMLS we used the term processing options (−z) to process each input sentence as a single phrase to identify more complex Metathesaurus terms, and the word sense disambiguation option (−y) to reduce as much as possible the ambiguity of concepts.

Semantic similarity measures

The task of assembling semantically homogeneous groups of experiments requires the definition of semantic similarity measures [23]. Two different approaches are commonly used in the biomedical domain: knowledge- and distributional-based methods. Knowledge-based methods use the structure of semantic networks as a knowledge source to build a graph of concepts connected by concept relationships and estimate concept similarities. These measures include the 'path finding', based on the shortest path between concepts, 'random walk' and 'intrinsic information content' that weights semantic relationships based on the specificity of the concepts in the semantic hierarchy. Distributional-based methods associate the knowledge about the structure of the source with the frequency of the concepts in a given text. The choice of the appropriate similarity measure is a difficult task and depends on the number and type of concepts that a knowledge source includes and the relationships it contains, but also on the text in which measures have to be computed.

In the case of GEO metadata, knowledge-based methods are expected to perform better than distributional-based ones, as the description of each sample is short and the frequency of each concept in the corpus is always low. In particular, within knowledge-based methods, those relying on 'intrinsic information content' provide good performance in terms of accuracy and are a good choice when a big corpus to compute concept frequencies is not available [24, 25]. The specific definitions of the knowledge-based semantic similarity metrics adopted in this study follow here.

The Path [26] similarity measure between two concepts (c_1 and c_2) is defined as the inverse of the shortest path length between them:

$$Path(c_1, c_2) = \frac{1}{shortest\_path(c_1, c_2)}$$

Leacock and Chodorow [27] propose a similar measure (LCH) scaling the path length by the maximum depth of the taxonomy d:

$$LCH(c_1, c_2) = 1 - \frac{\log(shortest\_path(c_1, c_2))}{\log(2d)}$$

Wu and Palmer [28] divide the depth of the most specific concept subsumer (lcs) of the two given concepts by the sum of the distances of the concepts to the subsuming concept:

$$WuPalmer(c_1, c_2) = \frac{depth(\text{lcs}(c_1, c_2))}{depth(c_1) + depth(c_2)}$$

Information content-based distributional measures consider the frequency of the occurrence of a given concept in a large text corpus. The corpus information content of a concept is simply the inverse of the log of the concept's frequency (which also accounts for the frequencies of its children). The Lin [29] semantic similarity measure falls within this class:

$$Lin(c_1, c_2) = \frac{2 \times IC(\text{lcs}(c_1, c_2))}{IC(c_1) + IC(c_2)}$$

This measure depends on the availability of a reference corpus, which is not available for our purpose. Eventually, we excluded the Lin similarity because in the GEO metadata the frequency of concepts is mostly associated with the frequency of a given type of experiments (which does not have any specific biological meaning) rather than reflecting the concepts specificity or importance. As an alternative, improving on the Lin measure and avoiding the dependency on the terms frequency, it is possible to use a version named Intrinsic Lin. This is based on the Intrinsic IC [30], obtained by replacing the IC with its intrinsic counterpart, defined as follows:

$$IC_{\text{intrinsic}}(c) = -\log\left(\frac{\frac{leaves(c)}{subsumers(c)} + 1}{\max_{leaves} + 1}\right)$$

where leaves(c) is the number of leaves that are descendant of c, subsumers(c) is the list of ancestors of concept c and max_{leaves} is the total number of leaves in the taxonomy.

Knowledge-based measures mentioned above can also be transformed so that they leverage on Intrinsic IC, as this has been shown to be advantageous [24]. This originates the Intrinsic Path, LCH versions and is achieved considering the following semantic distance instead of the classical shortest path distance:

$$dist(c_1, c_2) = IC(c_1) + IC(c_2) - 2 \times IC(\text{lcs}(c_1, c_2))$$

Results and discussion

We focus on the metadata of publicly available ChIP-seq samples to illustrate the potential of NLP tools such as Metamap and Conceptmapper in determining the semantic annotation and similarities between high-throughput biological samples. Specifically, we intended to demonstrate and evaluate the ability of these tools in recognizing the immunoprecipitated target, the tissue and the disease condition for ChIP-seq samples targeting TFs and epigenetic marks. Surprisingly, while several thousands of experiments are already available for these regulatory factors in public databases, and while studies involving TFs and epigenetic marks would greatly benefit from this abundant and comprehensive information, the semantic information associated to this rich data resource is rarely exploited.

Retrieval of ChIP-seq metadata and identification of the immunoprecipitated target

Metadata associated to the raw data of ChIP-seq experiments stored in SRA (and closely matching the corresponding GEO metadata) were obtained using the SRAdb Bioconductor package, by filtering for experiments associated to the 'ChIP-seq' library strategy (see the Methods section for additional details). We identified 21 037 SRA experiments for human (11 641 samples) and mouse (9396 samples).

Proper documentation of a ChIP-seq experiment implies the definition of the immunoprecipitated molecule. As an exception, ChIP-seq control samples are commonly obtained by repeating the experiment without the antibody (input samples) or using antibodies resulting in nonspecific binding (such as IgG). ChIP-seq samples metadata often contain a field identifying the targeted entity or the antibody used, if any. This field can be identified by different keywords like 'ChIP antibody' or 'Antibody target'. When these tags were available, the rest of the description was not used for the identification of the immunoprecipitated target. Overall, we were able to identify targets for 16 953 of 21 037 SRA ChIP-seq experiments (see the Methods section for additional details). Of these, 2493 were control samples (input) and 360 IgG (Figure 1A).

Identification of disease and tissue terms in ChIP-seq metadata

We configured Metamap to map sentences with CUIs in the 'disease syndrome' and 'neoplastic process' semantic classes. We evaluated both the possibility to run Metamap on the entire sentences describing the samples, or on parts of the sentence associated to specific tags informative on the disease condition: 'disease', 'disease state' and 'donor health status'. The latter option was available only for 2177 of 21 037 experiments. First, these tags are used to identify normal (healthy) samples. A sample was considered normal if the value associated to the tag in the attribute sentence contained one of the following: 'normal', 'healthy', 'presumed normal', 'None', 'NA', 'no ad present' (1896 experiments of 2177). Then, experiments not assigned to the normal state were annotated using Metamap. The annotation phase required 3h17m and returned 325 unique disease CUIs for 6248 ChIP-seq samples (Figure 1B). Metamap results were compared with those obtained with an alternative tool, Conceptmapper, using a dictionary of diseases based on the DO [19]. With Conceptmapper we were able to match 4831 samples to 124 unique disease ids in a few minutes (Supplementary Figure S1A).

Similarly to what we did for the diseases, we repeated the annotation process for the identification of tissues and cell lines using both tools. To this purpose, we used the same Metamap options described above, selecting the 'bpoc', 'tisu' and 'cell' semantic types. The annotation of tissues with Metamap allowed us to match 14 328 of 21 037 samples to 580 different tissues and cell lines CUIs (Figure 1C). Results obtained by running Conceptmapper on the same set of sentences allowed us to associate 19 152 of 21 037 samples to 613 different concepts defined in the BTO (Supplementary Figure S1B).

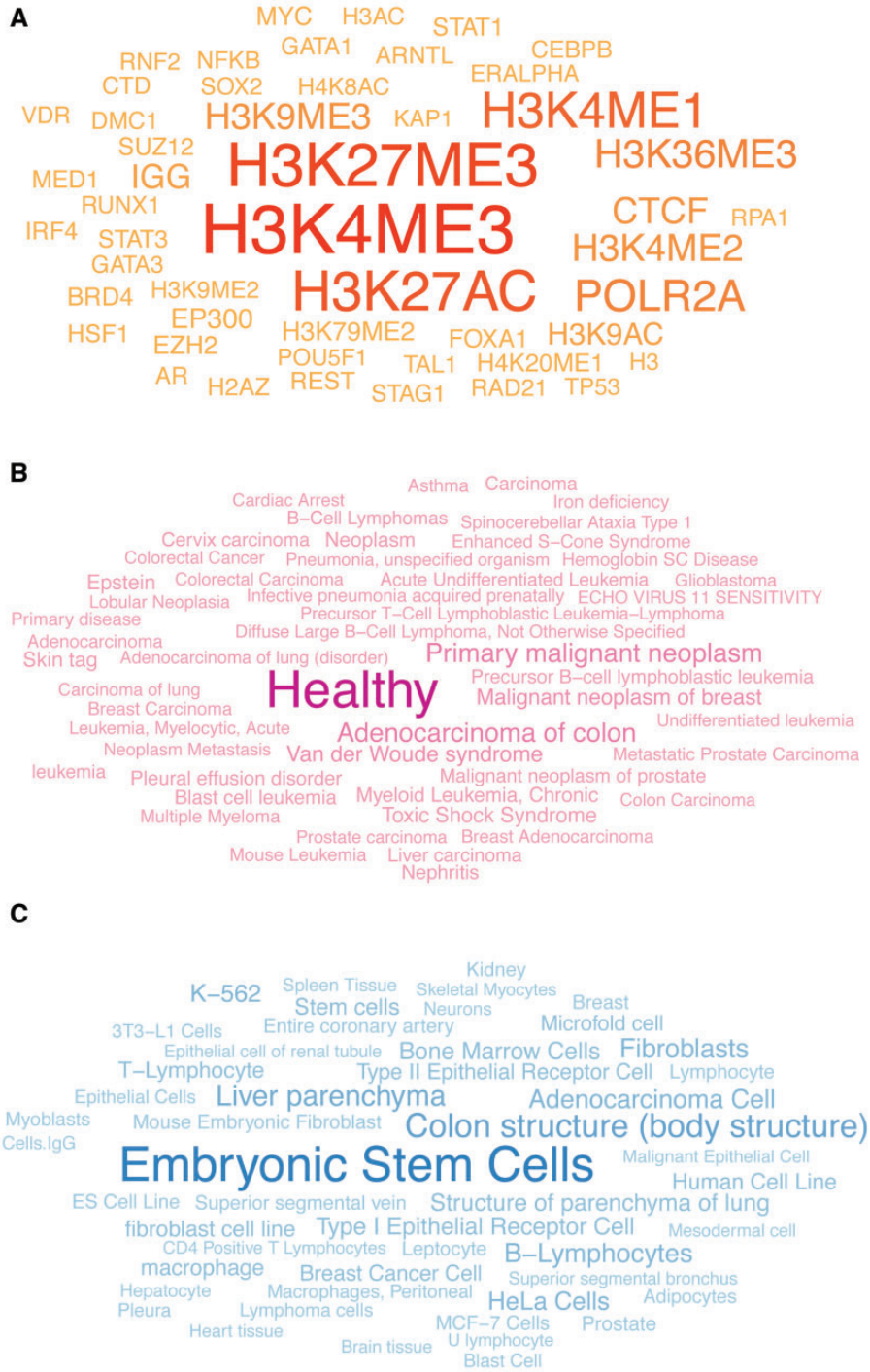

Figure 1. Word cloud with the most frequent targets, tissue and disease terms. (**A**) The 50 most frequent targets for the considered ChIP-seq experiments excluding the control samples; target size and color shade are proportional to the number samples. (**B**) As in (A) for the 50 most frequent disease terms identified from Metamap. (**C**) As in (B) for the tissue terms.

For both diseases and tissues, the top-ranking terms identified by Metamap and Conceptmapper qualitatively show a good agreement.

Evaluating the quality of semantic annotations

To evaluate the quality of the annotations, we randomly extracted 200 GEO samples from the set of 21 037 ChIP-seq experiments. For each sample, we manually compared the result obtained from the automatic annotators (Metamap or Conceptmapper) with the expected one, resulting from manual curation. We would like to stress that the aim of the manual curation is to evaluate the amount and quality of information that we could extract from the GEO metadata, without considering any additional information that could be, for example, retrieved from the corresponding scientific article. We adopted

Table 1. Evaluation of Metamap (MM) and Conceptmapper (CM) semantic annotations for tissue and disease terms

Annotations	TP	FP	TN	FN	Sensitivity	Specificity	Precision	Negative predictive value	F1
MM full tissues	267	257	2	82	0.76	0.01	0.67	0.05	0.61
MM keyw tissues	118	57	6	107	0.52	0.65	0.67	0.05	0.59
MM full diseases	106	112	113	1	0.99	0.5	0.49	0.99	0.65
MM keyword diseases	62	36	73	63	0.5	0.67	0.63	0.54	0.56
CM tissues	312	33	15	34	0.9	0.31	0.9	0.31	0.9
CM diseases	73	19	141	17	0.81	0.88	0.79	0.89	0.8

Note. For Metamap, the evaluation was done both providing as input the entire sentences (MM full) or only the part of it matching keywords specific for the topic of interest (MM keyw).

the following rules for a given annotation of a given sample: (i) a true-positive (TP) value was assigned if the concept corresponded to the manually curated annotation; (ii) a false-positive (FP) value was assigned to concepts not matching the manual annotations of samples (including concepts that were too generic, such as 'cell line' or 'disease', or poorly specific given a more specific concept available in the hierarchy); (iii) a true-negative (TN) value was given to those samples for which it was not possible to determine the annotation both manually and automatically (the concept is not contained in the metadata and is not available in the Metathesaurus, for UMLS, or ontology, for Conceptmapper); a false-negative (FN) value was assigned to manually recognized concepts that were missed by the automatic annotators. The performance of the annotation process is summarized in Table 1, which reports the values of sensitivity, specificity, precision, negative predictive value and F1 measure. The F1 score is the harmonic mean of precision and recall and can be used to summarize the accuracy of the annotation procedure.

In general, Conceptmapper clearly outperformed Metamap. Noteworthy, Metamap performance was higher when considering the entire sentences, for both tissue and disease terms. This indicated that, whenever specific keywords were available, the quality of the contained information did not meet the expected quality and did not justify discarding the full sentence. The lower performance of Metamap is mainly owing to a higher number of CUIs incorrectly assigned to a sample compared with Conceptmapper. This is often attributable to poorly informative CUIs, and ambiguous or conflicting CUIs associated to the same sample. These issues have lower effect on Conceptmapper annotations because OBO ontologies are less complex and better organized.

With the aim of improving Metamap annotations, we tried different solutions: (i) filtering the annotations removing those having mapping scores under a threshold; (ii) removing generic annotations based on the hierarchical structure of UMLS Metathesaurus; (iii) reducing the number of the assigned CUIs to a given sample based on their mapping score. Neither one of these filters nor their combination was able to significantly improve the performance in terms of the measured quality metrics (data not shown).

Semantic similarity

The availability of samples semantic annotations allows an effortless identification of samples matching a particular disease (or tissue). Importantly, it also allows identifying samples associated to semantically close conditions. This could be useful to (i) overcome the lack of samples matching a condition, relying on samples matching a similar tissue or disease state; (ii) quantify the semantic relatedness of samples associated to different

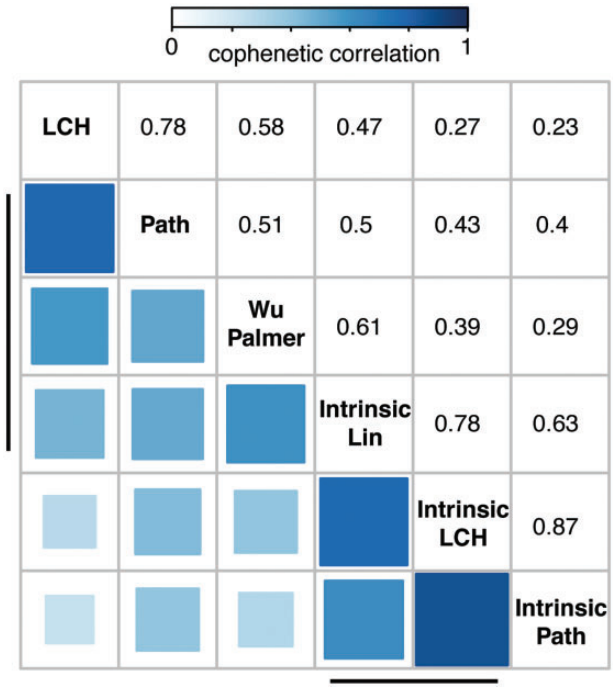

Figure 2. Matrix of pair-wise cophenetic correlations between the CUIs tissue dendrograms obtained with the indicated semantic similarity measures. The two groups of semantic similarity metrics discussed in the main text are highlighted.

biological conditions; (iii) identify (or stratify) samples associated to terms being children of a common parent-term (more generic). Importantly, the degree of similarity can be quantitatively determined.

We used YTEX [31] to compute the pairwise semantic similarities between UMLS CUIs annotating the test samples. The clustering of tissue CUIs using different similarity metrics is exemplified in Supplementary Figure S2 using the Intrinsic Lin metric. On one hand we could not compare the results obtained with the different measures to any gold standard. On the other hand, we reasoned that semantic similarity measures are more robust if there is agreement between the groups created by each of them. To this end, we computed the correlation between the resulting semantic similarity dendrograms (Figure 2) based on the cophenetic correlation (the correlation between two cophenetic distance matrices, obtained considering the height at which two close components are combined into a single cluster). We could identify two groups, the first referring to the intrinsic information content measures (intrinsic Lin, LCH and Path) showing the highest pairwise correlation (average 0.76 cophenetic correlation), and the second pointing to the LCH,

Path and WuPalmer measures (average 0.62 cophenetic correlation).

The correlation of dendrograms obtained using Slib [32] based on BTO and Conceptmapper shows higher coherence among the measures (average 0.86 cophenetic correlation, Supplementary Figure S3). Eventually, we decided to adopt Intrinsic Lin for the rest of the study, as this is the measure that best correlates with the other five measures both for Metamap and Conceptmapper.

Noteworthy, the analyses presented so far relate tissue (or disease) terms to each other directly. As ultimately one would like to determine the semantic similarity between samples, based on their metadata annotation, and as a sample can be assigned to multiple annotation terms, we decided to adopt the best match average criteria [33], which averages the similarity of the best matching concepts in two different samples.

UMLS-specific issues and possible solutions

In the case of tissue UMLS-based semantic similarities, concepts belonging to the same semantic type are often closer to each other compared with the expected similar concepts belonging to other types. For example, the concept 'MCF7-cells' (human breast cancer cell line) always appears closer to the concept 'K-562' (myelogenous leukemia cell), than to the concept 'Breast cancer cell'. Indeed, both 'MCF7-cells' and 'K-562' map to the 'cell' semantic type, while 'Breast cancer cell' belongs to the 'bpoc' semantic type. This would be problematic in a query for samples matching a disease like breast cancer, in which one would expect samples associated to both the terms 'Breast cancer cell' and 'MCF7-cells', with the exclusion of samples matching to 'K-562' cells. To overcome this issue, we considered the definitions of terms belonging to the 'cell' semantic type, taken from the corresponding Metathesaurus or, in its absence, using the short UMLS definition provided for each term. We used Metamap to associate these definitions to 'bpoc' CUIs, trying to enforce a link between the 'cell' and the 'bpoc' semantic types. This only partially resolved this issue, as we were not able to match the majority of 'cell' concepts to any 'bpoc' concept.

We concluded that both the quality of annotations and the robustness of semantic similarity metrics are higher when using individual ontologies (as in the case of BTO, or DO, using Conceptmapper) rather than UMLS annotations.

A case study: the Myc TF

As a case study to illustrate the application of semantic annotations and similarities, we retrieved 77 ChIP-seq samples targeting the Myc TF in human, and their associated metadata. ConceptMapper was used for the semantic annotation of tissue and disease terms, based on the BTO and DO, respectively. Resulting annotations were clustered based on the Intrinsic Lin semantic similarity (Supplementary Figure S4).

It could be expected that the semantic similarity between the samples reflects the similarity in the ChIP-seq signal, i.e. the Myc binding sites on the genome. To verify this, we selected within the Myc data set a subset of 22 samples from eight studies, representing nine distinct tissues and five disease states. The Intrinsic Lin semantic similarity was computed for each pair of samples, and the corresponding hierarchical clustering revealed five main groups (Figure 3A). For the same samples, we aligned the corresponding HTS reads on the human genome and we identified the Myc enriched regions (peaks). Figure 3B reports the proportion of shared peaks between pairs of samples. Samples are ordered based on the association to the same or similar tissue (tissue class; based on the clustering in Figure 3A).

The average peaks overlap for samples within the same tissue class is generally higher compared with samples belonging to different tissues and diseases. In the case of Hela cells for example, three samples from two independent studies show an average 56% of overlap (Figure 3A and B). As further example, six samples from three different studies in 'mammary gland' and 'breast cancer cell' tissues have 42% overlap. In the latter example, the samples could also be stratified, looking at the peaks overlap only, in two main groups that nicely correspond to the disease state ('adenocarcinoma' and 'breast adenocarcinoma') versus unknown (on closer inspection this could be classified as healthy samples). When the inter-tissue overlap is high, it often matches a high semantic similarity between the tissues. Compare, for example, the 45% overlap between 'K-562 cells' and 'B-lymphoma'/'B-lymphocyte' in Figure 3B, with the high similarity between these tissue terms in Figure 3A. These examples illustrate how, based on the semantic closeness of tissue terms, we could combine samples from different laboratories and independent studies, obtaining additional and coherent information without the need of new experiments.

Myc peaks typically are associated with epigenetic marks and regulatory factors characterizing open chromatin and transcriptionally active regions, such as H3K4me3, H3K27ac and Pol2 [34]. Starting from a collection of samples for Myc in a 'B-lymphoma' cell line (as tissue) and 'B-cell lymphoma' (as disease) from a specific study (GSE30726), we illustrate the concept of semantic expansion: ChIP-seq samples are collected targeting Myc and additional factors and marks associated to similar tissues and disease states.

First, all the ChIP-seq samples directed to Myc, H3K4me3, H3K27ac and Pol2 having the same tissue and disease condition of GSE30726 were identified: the block in the upper-left side of the heatmap in Figure 3C collects all these 'seed' samples associated to the 'B-lymphoma' tissue and 'B-cell lymphoma' disease state. Relaxing the tissue semantic similarity to values $\geq$0.9 we could expand this initial data set with additional experiments targeting Myc, H3K4me3, H3K27ac or Pol2, which were annotated with the 'B-lymphocyte', 'BCB-L1 cell' and 'Lymphoma cell' tissue terms. The heatmap in Figure 3C illustrates that most of the additional Myc peaks co-occur with Pol2, H3K4me3 and H3K27ac peaks (54%, 57% and 52%, respectively), pointing to active promoter or enhancer regions. As expected, Pol2 positive regions are in turn often associated with H3K4me3 and H3K27ac peaks (64% and 63%, respectively). As it has been described [34], while Myc often binds in presence of Pol2 (52% of Myc peaks are Pol2 positive) the opposite does not necessarily hold: not all the Pol2 positive regions are Myc bound (only 16% of them are Myc positive). Similarly, also for H3K4me3 and H3K27ac the fraction of peaks that are Myc positive (9% and 7%, respectively) and that are associated Pol2 (34% and 28%, respectively) is particularly reduced. These are expected to be mostly promoters of low-expressed genes or active enhancers that are not Myc bound. Noteworthy, within the Pol2 ChIP-seq samples, there are four samples annotated as 'B-lymphocytes' with 'Unknown' disease, whose overlap with the other Pol2 samples is markedly reduced. These samples were manually verified to refer to healthy samples, while they could not be identified as healthy by the annotator. This highlights the importance of considering both the tissue and disease state when collecting

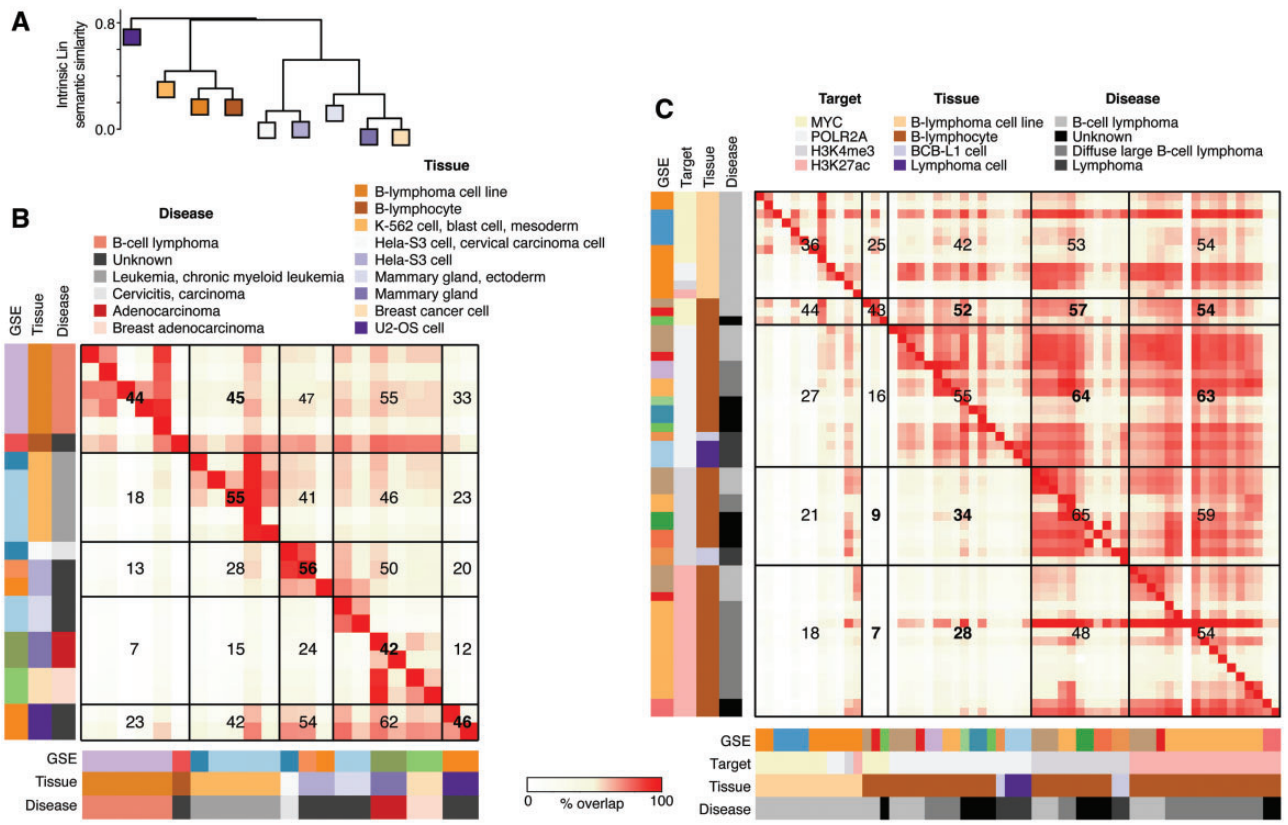

Figure 3. Agreement between semantic similarity and ChIP-seq signal. (**A**) Hierarchical clustering of tissue annotations for selected ChIP-seq samples targeting the Myc TF, based on the Intrinsic Lin semantic similarity. (**B**) Heatmap showing the percentage of peaks shared by the samples on the rows with the samples on each column. The colorbars on the left of the heatmap denote samples having identical GEO Series id (GSE), tissue or Disease id, according to the legend. (**C**) Percentage of overlap of peaks of ChIP-seq samples targeting Myc, H3K4me3, H3K27ac and Pol2. A first set of samples (upper left block) was identified for these marks in B-lymphoma cell line (tissue) and B-cell lymphoma (disease). Additional samples from different studies were added by relaxing the semantic similarity threshold to include samples for the same marks that are associated to similar tissues and diseases. For each block, the average overlap is reported, excluding same-to-same overlaps, and values in bold are discussed in the text.

samples from GEO. Eventually, this analysis illustrates how we were able to recapitulate the expected biology in the interplay between Myc, Pol2 and epigenetic marks [34], by using ontological reasoning to integrate previously unrelated ChIP-seq samples matching various targets and associated to similar tissue and disease conditions.

Conclusions

In this work, we have explored the efficacy of available text-mining tools to allow users to retrieve and relate samples of interest from public repositories of HTS data such as GEO. Specifically, we intended to test the feasibility of identifying ChIP-seq samples directed to a given target and associated to specific diseases and tissues. We adopted two popular concept-mapping tools, Metamap and Conceptmapper, to annotate GEO samples metadata using particular categories of UMLS concepts or ontology terms. We also explored the possibility of relating concepts to each other, as this would allow to group collected samples in semantically homogeneous clusters or to expand user queries to retrieve samples with the desired level of semantic similarity.

The large amount of UMLS relationships and concepts makes UMLS both a rich source of information and an extremely complex resource in terms of usability and accessibility. Many sources of error and ambiguity are introduced during the integration of new sources (ontologies and vocabularies). Consequently, determining the path from a given concept to another is affected by issues such as hierarchical cycles [35–37], redundant relationships, conflicting relationships, multiple mutual exclusive relationships between two concepts or orphaned components [38]. In addition, different semantic types are often not adequately interconnected: in our case, we could not associate cell lines to the tissues they refer to. Ultimately, we showed that tools as Conceptmapper, which can be used with topic-specific OBO ontologies, outperformed Metamap both in the annotation process and in the computation of semantic similarity measures.

To demonstrate the usefulness of these resources and tools in a practical application, we considered ChIP-seq samples targeting the Myc TF. We showed how it is possible to retrieve samples with precise tissue and disease annotations, and how one could complement the available data with additional samples having a controlled degree of semantic similarity. In conclusion, we demonstrated how it is possible to leverage on ontological reasoning for assembling large data sets including samples previously not associated. This illustrates the potential of public repositories (including, while not limited to GEO) in allowing scientists both to confirm their findings and to assemble larger data sets possibly leading to new or refined discoveries. Future work might be dedicated to the development of user-friendly software applying these concepts, with emphasis

on the field of epigenetics and transcriptional regulation, and possibly incorporating a database of semantically annotated GEO metadata.

Supplementary data

Supplementary data are available online at http://bib.oxfordjournals.org/.

Key Points

- The potential of using public repositories for assembling data sets including high-throughput biological data previously not associated is remarkably unexplored.
- Semantic annotation of the metadata of public data sets with concepts from biomedical ontologies standardizes their representation and is instrumental for the retrieval of samples for a given condition of interest.
- The semantic similarity between different samples can be quantitatively determined leading to possibly large groups of semantically coherent samples.
- Comparison of tools based on UMLS with tools that use topic-specific OBO ontologies showed that the latter outperforms the former both in the annotation process and in the computation of semantic similarity measures.
- The potential of this approach was illustrated selecting semantically homogeneous groups of ChIP-seq samples targeting the Myc TF and expanding the data set with semantically similar epigenetic samples; importantly, the semantic information proved to be coherent with the ChIP-seq signal and the current knowledge about this TF.

References

1. Edgar R, Domrachev M, Lash AE. Gene expression omnibus: NCBI gene expression and hybridization array data repository. *Nucleic Acids Res* 2002;**30**:207–10.
2. Rustici G, Kolesnikov N, Brandizi M, *et al*. ArrayExpress update–trends in database growth and links to data analysis tools. *Nucleic Acids Res* 2013;**41**:D987–90.
3. Butte AJ, Kohane IS. Creation and implications of a phenome-genome network. *Nat Biotechnol* 2006;**24**:55–62.
4. ENCODE Project Consortium. The ENCODE (ENCyclopedia Of DNA Elements) project. *Science* 2004;**306**:636–40.
5. Hoehndorf R, Schofield PN, Gkoutos GV. The role of ontologies in biological and biomedical research: a functional perspective. *Brief Bioinform* 2015;**16**:1069–80.
6. Brazma A, Hingamp P, Quackenbush J, *et al*. Minimum information about a microarray experiment (MIAME)-toward standards for microarray data. *Nat Genet* 2001;**29**:365–71.
7. Taylor CF, Field D, Sansone S-A, *et al*. Promoting coherent minimum reporting guidelines for biological and biomedical investigations: the MIBBI project. *Nat Biotechnol* 2008;**26**: 889–96.
8. Whetzel PL, Parkinson H, Causton HC, *et al*. The MGED Ontology: a resource for semantics-based description of microarray experiments. *Bioinformatics* 2006;**22**:866–73.
9. Bodenreider O. The Unified Medical Language System (UMLS): integrating biomedical terminology. *Nucleic Acids Res* 2004;**32**:D267–70.
10. Smith B, Ashburner M, Rosse C, *et al*. The OBO Foundry: coordinated evolution of ontologies to support biomedical data integration. *Nat Biotechnol* 2007;**25**:1251–5.
11. , Weinstein JN, Collisson EA. Network TCGAR, *et al*. The cancer genome atlas pan-cancer analysis project. *Nat Genet* 2013;**45**:1113–20.
12. Bernstein BE, Stamatoyannopoulos JA, Costello JF, *et al*. The NIH Roadmap Epigenomics Mapping Consortium. *Nat Biotechnol* 2010;**28**:1045–8.
13. Whetzel PL, NCBO Team. NCBO technology: powering semantically aware applications. *J Biomed Semantics* 2013;**4(Suppl 1)**: S8.
14. Jonquet C, Shah NH, Musen MA. The open biomedical annotator. *Summit Translat Bioinforma* 2009;**2009**:56–60.
15. Zoubarev A, Hamer KM, Keshav KD, *et al*. Gemma: a resource for the reuse, sharing and meta-analysis of expression profiling data. *Bioinformatics* 2012;**28**:2272–3.
16. Zhu Y, Stephens RM, Meltzer PS, *et al*. SRAdb: query and use public next-generation sequencing data from within R. *BMC Bioinformatics* 2013;**14**: 1
17. Turner BM. Reading signals on the nucleosome with a new nomenclature for modified histones. *Nat Struct Mol Biol* 2005;**12**:110–12.
18. Gremse M, Chang A, Schomburg I, *et al*. The BRENDA Tissue Ontology (BTO): the first all-integrating ontology of all organisms for enzyme sources. *Nucleic Acids Res* 2011; **39**:D507–13.
19. Schriml LM, Arze C, Nadendla S, *et al*. Disease ontology: a backbone for disease semantic integration. *Nucleic Acids Res* 2011;**40**:D940–6.
20. Aronson AR, Lang F-M. An overview of MetaMap: historical perspective and recent advances. *J Am Med Inform Assoc* 2010;**17**:229–36.
21. Tanenblatt MA, Coden A, Sominsky IL. The ConceptMapper approach to named entity recognition. *LREC* 2010; 546–51.
22. Funk C, Baumgartner W, Garcia B, *et al*. Large-scale biomedical concept recognition: an evaluation of current automatic annotators and their parameters. *BMC Bioinformatics* 2014;**15**: 1.
23. Pesquita C, Faria D, Falcão AO, *et al*. Semantic similarity in biomedical ontologies. *Plos Comput Biol* 2009;**5**:e1000443.
24. Pirró G. A semantic similarity metric combining features and intrinsic information content. *Data Knowl Eng* 2009;**68**: 1289–308.
25. Seco N, Veale T, Hayes J. An intrinsic information content metric for semantic similarity in WordNet. *ECAI* 2004.
26. Pedersen T, Pakhomov SVS, Patwardhan S, *et al*. Measures of semantic similarity and relatedness in the biomedical domain. *J Biomed Inform* 2007;**40**:288–99.
27. Leacock C, Chodorow M. Combining local context and WordNet similarity for word sense identification. *WordNet* 1998;**49**:265–83.
28. Wu Z, Palmer M. Verbs semantics and lexical selection. *Proceedings of the 32nd annual meeting on Association for Computational Linguistics*, 1994, pp. 133–38.
29. Lin D. An information-theoretic definition of similarity. *ICML* 1998; **98**; 296–304
30. Seddiqui MH, Aono M. Metric of intrinsic information content for measuring semantic similarity in an ontology. *Proceedings of the Seventh Asia-Pacific Conference on Conceptual Modelling*, 2010; pp. 89–96

31. Garla V, Re Lo V, Dorey-Stein Z, *et al*. The Yale cTAKES extensions for document classification: architecture and application. *J Am Med Inform Assoc* 2011;**18**:614–20.
32. Harispe S, Ranwez S, Janaqi S, *et al*. The semantic measures library and toolkit: fast computation of semantic similarity and relatedness using biomedical ontologies. *Bioinformatics* 2014;**30**:740–2.
33. Schlicker A, Domingues FS, Rahnenführer J, *et al*. A new measure for functional similarity of gene products based on Gene Ontology. *BMC Bioinformatics* 2006;**7**:302.
34. Sabò A, Kress TR, Pelizzola M, *et al*. Selective transcriptional regulation by Myc in cellular growth control and lymphomagenesis. *Nature* 2014;**511**:488–92.
35. Bodenreider O. Circular hierarchical relationships in the UMLS: etiology, diagnosis, treatment, complications and prevention. *Proceedings of the AMIA Symposium*, 2001, pp. 57–61.
36. Mougin F, Bodenreider O. Approaches to eliminating cycles in the UMLS Metathesaurus: naïve vs. formal. *AMIA Annu Symp Proc* 2005;**2005**:550–4.
37. Halper M, Morrey CP, Chen Y, *et al*. Auditing hierarchical cycles to locate other inconsistencies in the UMLS. *AMIA Annu Symp Proc* 2011;**2011**:529–36.
38. Gu H, Elhanan G, Halper M, *et al*. Questionable relationship triples in the UMLS. *2012 IEEE-EMBS International Conference on Biomedical and Health Informatics (BHI)*, 2012, pp. 713–16.