BIOINFORMATICS ORIGINAL PAPER

Vol. 27 no. 7 2011, pages 980–986
doi:10.1093/bioinformatics/btr043

Data and text mining

Advance Access publication February 16, 2011

Annotating genes and genomes with DNA sequences extracted from biomedical articles

Maximilian Haeussler*, Martin Gerner and Casey M. Bergman

Faculty of Life Sciences, University of Manchester, Manchester, M13 9PT, UK

Associate Editor: Alex Bateman

ABSTRACT

Motivation: Increasing rates of publication and DNA sequencing make the problem of finding relevant articles for a particular gene or genomic region more challenging than ever. Existing text-mining approaches focus on finding gene names or identifiers in English text. These are often not unique and do not identify the exact genomic location of a study.

Results: Here, we report the results of a novel text-mining approach that extracts DNA sequences from biomedical articles and automatically maps them to genomic databases. We find that $\sim 20\%$ of open access articles in PubMed central (PMC) have extractable DNA sequences that can be accurately mapped to the correct gene (91%) and genome (96%). We illustrate the utility of data extracted by text2genome from more than 150 000 PMC articles for the interpretation of ChIP-seq data and the design of quantitative reverse transcriptase (RT)-PCR experiments.

Conclusion: Our approach links articles to genes and organisms without relying on gene names or identifiers. It also produces genome annotation tracks of the biomedical literature, thereby allowing researchers to use the power of modern genome browsers to access and analyze publications in the context of genomic data.

Availability and implementation: Source code is available under a BSD license from http://sourceforge.net/projects/text2genome/ and results can be browsed and downloaded at http://text2genome.org.

Contact: maximilianh@gmail.com

Supplementary information: Supplementary data are available at *Bioinformatics* online.

Received on December 15. 2010; revised and accepted on January 21, 2011

1 INTRODUCTION

A common challenge encountered by many biomedical researchers is to obtain a summary of the relevant literature pertaining to a particular gene or genomic region. With nearly 2000 articles added to MEDLINE on a daily basis (http://www.nlm.nih.gov/bsd/index_stats_comp.html), it is increasingly difficult to keep up with the rapid pace of publication outside ones immediate domain of expertise. The problem of finding relevant articles for a particular locus is becoming more acute as researchers increasingly adopt high-throughput genomic technologies (microarrays, genome-wide association studies, high-throughput sequencing, etc.). These genome-wide approaches often generate low-level data on thousands of genes or genomic regions, the interpretation of which becomes much more valuable when integrated with previously published studies on individual loci.

The challenge of linking articles to genes is partially solved for a limited number of model organisms, where dedicated teams of curators scan the literature and link publications to gene records in individual model organism databases such as FlyBase (The FlyBase Consortium, 2003), or through federated multi-organism databases such as Entrez Gene (Maglott *et al.*, 2007). However, these collections are not comprehensive and for the majority of species, including human, efforts to curate gene–article associations remain incomplete. In principle, automatic linking of articles to genes could be achieved by developing text-mining tools that detect gene names or identifiers in abstracts or full-text articles. However, gene names are not consistently used and are often not unique and developing accurate methods to resolve and disambiguate gene names in text and link them to database identifiers remains an active area of research (Krallinger *et al.*, 2008).

Even with curated or automatically generated links between articles and genes, the exact genomic sequences referred to in an article currently can only be determined by human interpretation of the full text. Furthermore, specific questions such as ‘which transcript was cloned?’, ‘which exon was amplified?’ or ‘where in the genome is a particular mutation found?’ can take considerable time for an individual researcher to answer, often requiring labor-intensive manual interaction between the literature and genomic databases.

For publications where authors report DNA sequences directly, these problems could be solved if all published sequences were systematically sent to primary sequence databases such as GenBank (Benson *et al.*, 2010). However, in the post-genomic era fewer articles report primary DNA sequences directly and instead only report primers used for polymerase chain reaction (PCR)-based techniques that are designed from published genome sequences. Furthermore, in contrast to longer DNA sequences, journals generally do not require deposition of short primer sequences in databases, and the minimum sequence length required for a GenBank submission is 50 bp (http://www.ncbi.nlm.nih.gov/Sitemap/samplerecord.html). As such, many DNA sequences that could provide unique tags to link articles to specific genes and genomes remain locked in the biological literature.

The possibility that DNA and protein sequences can be extracted from biomedical text was first demonstrated by Wren *et al.* (2005) and subsequently by several other groups (Aerts *et al.*, 2008; Garcia-Remesal *et al.*, 2010a, b; Shtatland *et al.*, 2007). Aerts *et al.* (2008)

*To whom correspondence should be addressed.

extended this technique to show that DNA sequences extracted from biomedical text could be mapped to genome sequences to identify the location, organism and target gene mentioned in an article. The approach of Aerts *et al.* (2008) was inspired by, and tailored to, longer sequences typically found in publications on *cis*-regulatory regions. For this particular use case, genome mapping using the single best BLAST match on a small number of model organism genomes provided high-precision results. However, this basic approach is not suitable to the more ambitious application of mapping all sequences from articles to all genomes, since the short size of many sequences in articles (e.g. PCR primers) and the increasing size of genome databases requires more sophisticated mapping techniques. In addition, Aerts *et al.* (2008) did not provide software for users to run and extend or a database for users to download and browse results in an intuitive way.

Here, we address the question of whether annotation of all genomes with sequences from the available open access (OA) biomedical literature is a realizable and practical goal. We show that the automated extraction and mapping of DNA sequences from more than 150 000 OA full-text articles in PubMed Central (PMC) is indeed possible and present a software implementation to achieve this aim called 'text2genome'. We map extracted sequences to 224 genomes and provide easily searchable results in the form of both a web interface and genome annotation tracks for the Ensembl (Hubbard *et al.*, 2009) and UCSC genome browsers (Rhead *et al.*, 2010). We demonstrate that we can associate articles with relevant genes and genomes by evaluating text2genome results on the subset of articles that also have GenBank records. Finally, we provide example use cases to demonstrate potential applications of our approach, by intersecting text2genome mappings with ChIP-seq data, and by querying for articles that report quantitative reverse transcriptase (RT)-PCR experiments for a given genomic locus. Our work provides a unique and timely resource for interpreting both biomedical literature and genomic data, and will help aid discovery across many domains of the life sciences.

2 SYSTEM AND METHODS

2.1 Data sources

PMC-OA (Roberts, 2001) full-text articles were downloaded in June 2010. PMCID–PMID associations were obtained from the PMC FTP server (ftp://ftp.ncbi.nlm.nih.gov/pub/pmc/PMC-ids.csv.gz). When available, we used XML files, or the ASCII text version of the PMC-OA article based on optical character recognition (OCR) of the original pdf. For PMC-OA articles where neither XML nor text files were available, pdftotext (http://www.foolabs.com/xpdf/) was used to convert the PDF file to ASCII text. Additionally, we processed text (or converted text) from supplementary files of the following document types: HTML, CSV, TXT, XML, DOC, XLS, PPT and PDF.

Starting with a complete download of GenBank version 176, we kept only non-high-throughput divisions that are relevant to this study: bct, inv, mam, pln, pri, rod, vrt. The resulting dataset of 7.09 million records were parsed with BioPython (Cock *et al.*, 2009) into relational database tables. Sequences > 1Mb were eliminated to remove large sequences from high-throughput studies that were deposited in the incorrect division of GenBank (e.g. chromosome sequences of *Drosophila melanogaster*). Entrez Gene data were downloaded in October 2009.

RepeatMasked genomes and gene transcripts were obtained from Ensembl Release 56 (Hubbard *et al.*, 2009) and EnsemblGenomes Release 3 (Kersey *et al.*, 2010). The taxa represented in these genomes include 68 animals, 134 bacteria, 10 fungi, 8 plants and and 4 protists, totalling 224 organisms with a NCBI taxon ID.

2.2 Text and sequence processing algorithm

We developed a simple procedure to detect nucleotide sequences in articles that accounts for the presence of OCR errors and the fact that sequences can be separated by spaces and line breaks (See Supplementary File 1 for details). In brief, we first removed all non-letter mark-up characters from a text, then concatenated words that (i) contained exclusively nucleotide letters or (ii) contained a certain percentage of nucleotide letters (a, c, t, g and u) above a length cut-off of 19. FASTA sequences extracted from PMC-OA were then searched with BLAT (Kent, 2002) in genes and genomes from Ensembl/EnsemblGenomes with a minimum number of 19 identical base pairs. Articles with exceedingly long (> 1Mb) or many (> 100) sequences were removed from further processing to increase the precision of our approach, since some supplementary files contain genome-wide sequence data (e.g. microarray probes).

The resulting BLAT matches from genomes and transcripts were then filtered to obtain the best matching species, genes and genomic regions. For each extracted sequence, only the highest scoring hits were retained. Hits to common plasmids and sequencing vectors were removed (using data from NCBI Univec). In order to disambiguate sequences matching several different organisms equally well, we extracted all mentions of organism names from the articles using default settings of LINNAEUS (Gerner *et al.*, 2010). If the full text contained organism names detected by LINNAEUS, only BLAT matches for these genomes were kept. If no organism mentions were found, the matches were limited to human and major model organisms. If there was no best match among these genomes, all remaining matches were retained. The best genome was determined as the one with the highest number of matching sequences at the gene or genomic level. To account for conserved sequences that may hit highly similar genomes (e.g. chimpanzee and human), the best genome for species that had the same number of best hits was decided by ranking genomes based on the species with the higher number of publications in Entrez Gene.

Hits on the best genome were fused into 'chains' if they were located closer than 50 kb; hits on transcripts from the best genome were chained if they matched the same gene. When a sequence was a member of several chains (e.g. caused by matches to segmental duplications), the hit was retained only for the chain with the maximum number of other matching sequences. Genes were predicted to be hit only if they matched at least two text-extracted sequences. If two genes passed this threshold and were hit by exactly the same sequences, only the gene with the largest number of publications in Entrez Gene was retained.

In general, our filtering steps are designed to achieve high precision, which can result in no prediction for either genes or genomic features. For instance, genomic features but no genes are predicted if sequence map to non-coding regulatory DNA (promoters and enhancers). Conversely, genes but no genomic features are predicted if sequences are designed to span exon–exon boundaries such as morpholinos or primers for quantitative RT-PCR.

2.3 Data analysis

GenBank records were used to generate a benchmark set of links between articles and species or genes. The PubMed document ID and organism of a submission were parsed directly from GenBank records using the chronologically first article associated with the record (i.e. the last or second-to-last 'REFERENCE' entry). GenBank accession numbers in full-text articles were identified using the following set of regular expressions: (([A-Z]{1}[0–9]{5})|([A–Z]{2}[0–9]{6})|([A–Z]{4}[0–9]{8,9})|([A–Z]{5}[0–9]{7}))(\[0–9]{1,3}). Precision was defined as the number of species–article or gene–article predictions by text2genome that matched at least one species–article or gene–article association defined by the GenBank record, as a proportion of the total number of text2genome predictions. Recall was defined as the number of species–article or gene–article predictions by text2genome that matched the species–article or gene–article associations defined by the GenBank record, as a proportion of the total number of associations defined in the sample of GenBank records.

To obtain the most likely Ensembl gene identifier for each GenBank record, BLAT was used to map each GenBank sequence to Ensembl/EnsemblGenomes transcripts, keeping only the best matching gene ID. As non-high-throughput divisions still can contain submissions with several thousand sequences, we filtered this set to retain sequences from small-scale analyses only and therefore removed articles that submitted more than 100 sequences. The resulting table contains articles identified by their PubMed ID and the genomes and predicted genes that were submitted to GenBank, represented by their NCBI Taxonomy ID and the Ensembl Gene ID (Supplementary File 2).

ChIP-seq data from Visel *et al.* (2009) was obtained from NCBI GEO Accession GSM348064. p300-bound peaks were mapped to the most current mouse genome assembly (mm9) and loaded into the UCSC genome browser (Rhead *et al.*, 2010) as a custom track in combination with the text2genome mm9 genome annotation track. ChIP-seq data was intersected with the full-text extracted sequences using the ‘overlap’ function of the UCSC table browser.

We retrieved articles describing RT-PCR-related experiments by using NCBI Entrez Programming Utilities to query PubMed abstracts or PMC full-text articles with the following query: ‘qpcr’ OR ‘q-pcr’ OR ‘qrt-pcr’ OR ‘quantitative pcr’ OR ‘quantitative poly*’ or ‘quantitative polymerase’ OR ‘quantitative realtime pcr’ OR ‘reverse transcription polymerase’ OR ‘reverse transcription pcr’ OR ‘rtpcr’ OR ‘rt-pcr’ OR ‘rt-qpcr’ OR ‘rtq-pcr’. A list of common RT-PCR control loci was obtained from (Vandesompele *et al.*, 2002). From this list, only the prefixes were used to account for different gene names in non-mammalian model organisms. Genes were thereby counted as RT-PCR control genes if they start with one of the following prefixes: ACT, B2M, GAPD, HMBS, HPRT1, RPL13, SDHA, TBP, UBC and YWHAZ.

2.4 Implementation

All extracted sequences, BLAT matches and genome–gene associations generated by text2genome are stored in a MySQL database. Custom Python CGI scripts render data into HTML pages and act as a light-weight Distributed Annotation System (DAS; Dowell *et al.*, 2001) server, making it possible to overlay the matches onto the Ensembl Genome Browser and provide metadata including links to the corresponding articles. An additional script exports the same data in Browser Extensible Data (BED) format, allowing visualization and filtering of chained BLAT matches on the UCSC genome browser. Extracted sequences, predicted genes, browser tracks and additional metadata, such as gene names recognized in full-text articles by GNAT (Hakenberg *et al.*, 2008), can be searched, downloaded and browsed at http://www.text2genome.org. Source code for the text extraction, mapping and display are available as a set of Python 2.4 scripts that can be downloaded from http://sourceforge.net/projects/text2genome/.

3 RESULTS

3.1 Full-text articles contain a wealth of DNA sequences that are not in GenBank

We extracted DNA sequences from 153 513 full-text research articles and their associated supplementary files in the OA subset of PMC (downloaded June 2010; later referred to as PMC-OA) using a procedure similar to the one that we presented previously in Aerts *et al.* (2008). Briefly, text was first stripped of XML tags and non-letter characters, after which all words with an [A,C,T,G,U] content greater than a threshold value were concatenated and output if the resulting sequence exceeded a length cutoff (see Section 2 and Supplementary File 1 for details). Using this algorithm, we obtained 350 888 nucleotide-like strings with an average length of 115.81 bp and a total size of 40.6 Mb (Supplementary File 3). The mode of the length distribution of individual sequences is 20 bp, and sequences more frequently occur in multiples of 2, suggesting that many sequences are PCR primers (Supplementary File 4).

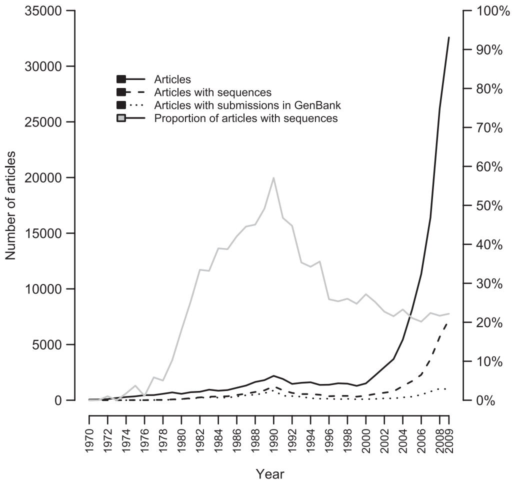

Fig. 1. Trends in sequence extraction from full-text articles. Lines show the growth of the PMC-OA subset from 1970 through 2009, together with the number and proportion of articles that contain DNA sequences in their full text or supplementary files, and the number of articles that could be linked to non-high-throughput GenBank submissions.

In total, 22.6% (34 828/153 513) of all research articles in PMC-OA contain sequence-like strings. The proportion of articles containing sequences in PMC-OA reached a peak of ∼45% in the mid-1990s and has subsequently levelled at just > 20% (Fig. 1). Over 33% (119 281/350 888) of sequences and the majority of nucleotides (64%, 26.0/40.6 Mb) were extracted from supplementary files. Only 9.7% (3381/34 828) of these articles contain sequences exclusively in their supplementary files. The majority of extracted strings are likely to be *bona fide* DNA sequences, since out of 3443 articles published before 1960 [i.e. prior to the advent of nucleic acid sequencing (Holley *et al.*, 1965)], only one article contains a sequence-like string (which was caused by an OCR error). To further validate our method, we manually inspected a randomly selected 1% subset of the nucleotide strings and found that all were valid DNA sequences, implying that most extracted strings represent true DNA sequences.

To compare the proportion of articles with DNA sequences to the proportion of articles accompanied by a GenBank submission, we estimated the number of PMC-OA articles that have a non-high-throughput GenBank record. Overall, we found that 6.7% (10 378/153 513) of PMC-OA articles are linkable to a GenBank submission (Supplementary File 5), and that this number of articles has remained relatively low over time (Fig. 1). As expected, we can extract nucleotides from the full text of the majority of PMC-OA articles with a GenBank submission (76.5%, 7937/10 378). Surprisingly, 77.2% (26 891/34 828) of articles with extractable nucleotides in their full text are not linkable to a GenBank submission. This result implies that the majority of sequences extracted from full-text articles have not been submitted to any nucleotide data bank.

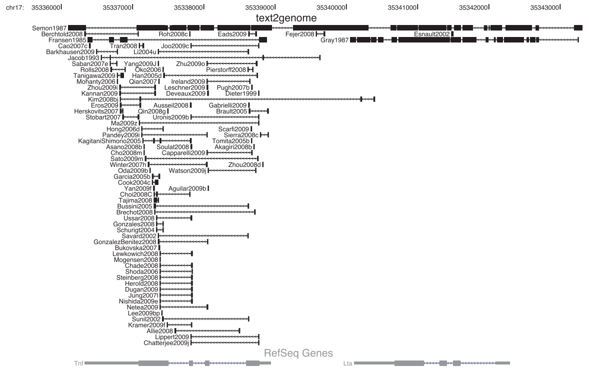

Fig. 2. Example of text2genome mappings for the mouse Tnf region. Exons for RefSeq gene models of Tnf and lymphotoxin A (Lta) are shown as grey rectangles below, and chained BLAT hits from text2genome mappings are shown as black rectangles above. Note that the majority of mapped papers contain pairs of sequences that are consistent with being PCR primers. The two larger mapped sequences come from original publications reporting Tnf and Lta primary sequences (Gray *et al.*, 1987; Semon *et al.*, 1987).

3.2 Sequences link articles to species, genes and genomic locations with high precision

DNA sequences extracted from text do not themselves contain information about the species or gene from which they were obtained, but instead must be mapped to other annotated sequences in order to propagate this meta-information to the articles in which they were found. Therefore, we searched all sequences extracted from text against all genome and transcript sequences in the Ensembl and EnsemblGenomes databases and resolved the best matching species, gene and genomic region (see Section 2 and Supplementary File 1 for details). An example of a genome browser view of the text2genome mappings for a region of the mouse genome containing the tumor necrosis factor (Tnf) gene is shown in Figure 2. Overall, 79.3% of articles with sequences (27 632/34 828) lead to a single best species prediction that is based on hits to a best matching genomic region (99.3%, 27 452/27 632) and/or a best matching gene (40.0%, 9935/27 632). Roughly one-third of all article–species associations are based on both a best matching genomic region and a best matching gene hit (35.3%, 9755/27 632), indicating that many sequences extracted from articles map to intronic and intergenic regions. In total, these articles generate 247 007 unique associations between articles and genomic regions (Supplementary File 6) and 23 388 unique gene–article associations (Supplementary File 7). The ∼21% of articles with sequences that do not lead to a best genome prediction at all arise from sequences in repeated regions, from cloning/sequencing vectors, or species not currently represented in Ensembl or EnsemblGenomes.

To evaluate the accuracy of text2genome species and gene mappings, we used as a reference the set of articles where the original authors submitted sequences to GenBank. We chose not to use data from Entrez Gene as a gold standard since it is part of our pipeline, and since Entrez Gene may curate species or genes that are mentioned in an article but for which no sequences are reported. The text2genome-inferred gene or species was considered to be correct if it matched any of the Genbank-derived information for an article. We attempted to filter out articles from this evaluation set that reported either (i) high-throughput sequencing results (e.g. expressed sequence tag projects) by excluding articles with more than 100 submitted sequences, or (ii) genome-scale sequence contigs, by limiting the length of the GenBank sequence to 1 Mb. This resulted in a dataset with the species and the best matching gene for 4800 articles based on GenBank submissions (Supplementary File 2).

Table 1. Evaluation of text2genome (t2g) and GNAT species and gene predictions against associated GenBank submissions

Set	Cutoff	*N*	TP	FP	FN	Precision	Recall
t2g species	1	1248	1201	47	47	0.96	0.96
	5	1334	1279	55	173	0.96	0.88
	10	1338	1283	55	197	0.96	0.87
	100	1338	1283	55	197	0.96	0.87
t2g genes	1	890	814	76	76	0.91	0.91
	5	1647	1223	424	457	0.74	0.73
	10	1813	1278	535	592	0.70	0.68
	100	2017	1325	692	895	0.66	0.60
GNAT species	1	518	373	145	145	0.72	0.72
	5	1793	867	926	301	0.48	0.74
	10	1809	875	934	324	0.48	0.73
	100	1809	875	934	324	0.48	0.73
GNAT genes	1	143	73	70	70	0.51	0.51
	5	1489	261	1228	393	0.18	0.40
	10	3682	429	3253	724	0.12	0.37
	100	7591	627	6964	1092	0.08	0.36

Cutoff refers to the number of predictions allowed for text2genome or GNAT predictions and the GenBank evaluation set. *N* refers to the number of predicted species–article or gene–article associations for each method and for the GenBank evaluation set. TP, FP, and FN refer to true positives, false positives and false negatives, respectively. Precision is defined as TP/(TP + FP) and recall is defined as TP/(TP + FN).

As with GenBank submissions, the number of predicted species and genes can vary for text2genome predictions. By limiting the number of text2genome species or gene predictions for a given article, we observed that all performance measures with the exception of species precision decrease with increasing numbers of predictions per article (Table 1). The most easily interpretable understanding of the true performance of text2genome can be obtained when documents are limited to those with one predicted species/gene and one reference species/gene. In this case, each false positive prediction creates an associated false negative and precision equals recall, and the accuracy of species prediction is 96%, while the accuracy of gene prediction is 91%. When we allow greater than one prediction per article, recall becomes lower than precision, reflecting the fact that not all sequences in a GenBank submission are reported in the full text, confirming the intended role of nucleotide databases as repositories that complement the main publication. At all cutoffs, species predictions are better in terms of precision and recall than gene predictions.

We analyzed the distribution of species and genes in our dataset to provide an overview of the taxonomic and genomic data extracted. As expected, we found that sequences in full-text articles most frequently map to the human and mouse genomes, as well as other organisms used in basic or agricultural genetics (Fig. 3). Trends in species identified using sequences in full text are largely consistent with the relative proportion of species in PMC-OA articles with GenBank records and the relative proportion of species mentions

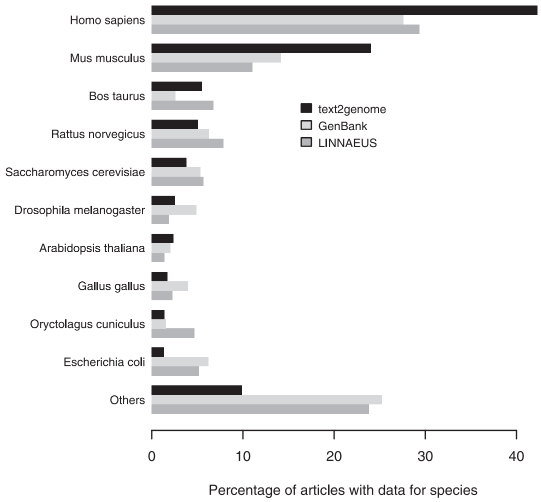

Fig. 3. Top 10 species identified by text2genome. Shown are the proportions of articles matching sequences extracted from the full text (text2genome), with sequence submissions in GenBank, or with mentions of the species name in the full text (LINNAEUS) for species with a sequenced genome in Ensembl/EnsemblGenomes.

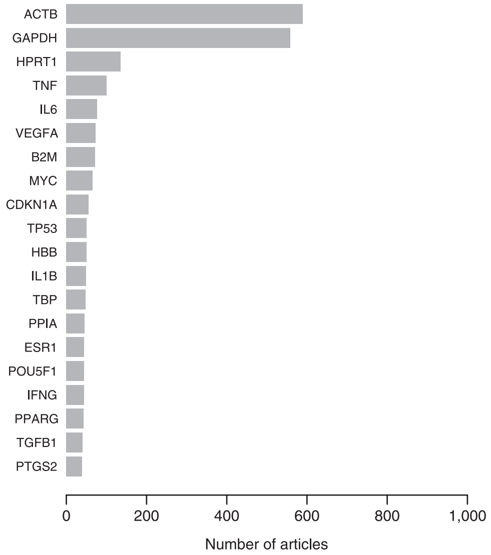

Fig. 4. Top 20 genes identified by text2genome. Shown are the number of articles with text2genome hits in PMC-OA for the human and mouse genome.

that can be extracted by LINNAEUS (Gerner *et al.*, 2010). Only 2 of the top 10 species mentioned in the entire set of MEDLINE abstracts (Gerner *et al.*, 2010) are missing from the top 10 text2genome best species list: HIV and dog.

In total, text2genome identified 12 655 genes in 9935 articles, the majority of which (6907, 70%) are from the human and mouse genomes. For sequences mapping to human and mouse, the most frequently hit genes are ACTB and GAPDH (Fig. 4), which are well-known control loci for RT-PCR experiments. Other genes with sequences frequently reported in the literature are from heavily investigated loci involved in immunity (TNF, IL6, IFNG) or cancer (P53, MYC). Detected genes are spread over a substantial proportion of human and mouse genomes. For example, articles linked to the human genome cover 21.1% of the 19 814 human and 13.4% of the 20 192 mouse gene models that are listed in Ensembl 56 and Entrez Genes.

Finally, we compared results of gene–article associations from text2genome with results of the gene name identification and normalization software GNAT (Hakenberg *et al.*, 2008; Supplementary File 8). GNAT detected at least one gene name in 73.2% (112 445/153 513) of PMC-OA articles, a rate approximately 3 times higher than articles with DNA sequences and more than 10 times higher than articles with genes found by text2genome. Likewise, genes identified by GNAT cover approximately three times more of the human (64.9%) and mouse (31.4%) genomes than text2genome. On our GenBank-derived benchmark, GNAT detected 7591 gene names in the full text of 1072 articles. For 50.9% of these articles (546/1072), one of these gene names corresponded to the gene mapped to by a sequence submitted to Genbank. Using the same evaluation criteria as for text2genome, we found that all GNAT performance measures except species recall also decrease with increasing numbers of predictions (Table 1). In all cases, text2genome outperforms GNAT for species or gene prediction in terms of precision and recall. For documents with only one prediction, the accuracy of GNAT predictions is 72% for species and 51% for genes. For documents with greater than one prediction, GNAT recall is higher than precision, consistent with the higher rate of predictions per article by GNAT relative to text2genome. When predictions for both text2genome and GNAT are directly compared on all of PMC-OA, 50.4% (5192/10 294) of human and 27.2% (1534/5629) mouse gene–article associations inferred by text2genome could be corroborated by GNAT predictions.

3.3 Sequences from articles accelerate the interpretation and design of genomics experiments

To demonstrate the utility of text2genome for research in genetics and genomics, we highlight possible use cases in the following two examples. In addition to these examples, the dataset of sequences extracted from articles by text2genome should also be useful in many other contexts, not least for annotators of various biological databases (Aerts *et al.*, 2008; Wren *et al.*, 2005).

3.3.1 Example 1: interpreting ChIP-seq data High-throughput ChIP-seq experiments providing information on the binding of transcription factors to thousands of loci can only be properly interpreted when calibrated against positive control data. By intersecting ChIP-seq regions with genomic regions annotated by text2genome, one can automatically find articles that have previously studied ChIP-seq regions. For example, Visel *et al.* (2009) conducted ChIP-Seq against p300, a common cofactor in many transcriptional complexes, with the aim of predicting enhancers in the forebrain, midbrain and hindbrain of mouse embryos. We intersected the 5119 p300-bound fragments in this dataset with mouse text2genome genomic regions, and found a region upstream of gene Lmo1 that is bound by p300 in the forebrain and overlaps previously reported primer sequences from Fulp *et al.* (2008). These authors have shown using ChIP-PCR that Lmo1 is expressed in the mouse forebrain and that its upstream region is bound by the transcription factor ARX in neuroblastoma cell lines. The ChIP-PCR fragment covers part of the interval published by Fulp *et al.* (2008) and confirms that this p300 bound region is actively bound by a transcription factor in mouse neuronal cells. To enable streamlined automation of this type of analysis using the UCSC Table Browser (Karolchik *et al.*, 2004), we provide text2genome data as BED tracks for selected assemblies in the UCSC genome database (Fig. 2).

3.3.2 Example 2: finding quantitative RT-PCR primers Biologists using quantitative RT-PCR (Gibson *et al.*, 1996) to measure transcript levels need to select control genes and find validated primers and cycling conditions before conducting their experiments. As many of the sequences in our database map to genes that are commonly used in RT-PCR experiments (Fig. 4), text2genome-extracted sequences offer a potentially rich source of validated RT-PCR primers. To evaluate this possibility and index articles for their potential utility in RT-PCR, we scanned all PMC-OA articles for keywords related to quantitative RT-PCR (see Section 2 for details). In PMC-OA, 3410 articles have RT-PCR-related keywords in their abstract, 18 912 have RT-PCR keywords in their full text or supplementary files and 1129 articles have text2genome predictions for genes that are commonly used in RT-PCR experiments. The vast majority of text2genome predictions that hit RT-PCR-related genes (81.6%, 922/1129) also have RT-PCR keywords in their full text, demonstrating that text2genome gene predictions can be useful when searching the literature during the design of an RT-PCR experiment. In contrast, only 21% of text2genome predictions that hit RT-PCR related genes (248/1129) have RT-PCR keywords in their abstracts, indicating that information about putative RT-PCR control genes cannot be obtained readily by searching abstracts alone. To aid in the selection of primers for RT-PCR, the text2genome web site offers a function to limit gene searches to articles that contain RT-PCR keywords in their full text. In this mode, the database currently shows sequences for 9045 genes from 5694 articles from 81 species. Ninety-eight percent of RT-PCR-related sequences are assigned to the human, mouse and rat genomes.

4 DISCUSSION

In an age of rapidly increasing amounts of DNA sequence data and published literature, finding peer-reviewed experimental results for a sequence of interest is more time consuming than ever. Here, we show that DNA sequences in full-text articles provide a rich source of 'unique identifiers' that can be automatically extracted and mapped to genomic data in order to link articles to species, genes and genomic regions. We confirm recent findings that a substantial number of OA articles in PMC contain extractable DNA sequences (Garcia-Remesal *et al.*, 2010b), and provide the first quantitative estimate of the proportion of PMC-OA articles with DNA sequences (∼22%), the majority of which we show are short sequences that are not found in GenBank.

Our study is also unique in that it presents the first attempt to apply sequence extraction techniques at a large scale to all types of both full text and supplementary data files, and in fact may be the first systematic application of text mining to supplementary files in any domain. Our observation that the majority of nucleotides in the PMC-OA corpus were extracted from supplementary files underscores the increasing reliance of authors to deposit important information contained in these files (Weiss, 2010), as well as the importance of using these resources for biological data mining and requiring ancillary research data to be persistently stored together with the main publication (Anderson *et al.*, 2006). Future work will be necessary to determine if the quality of data from full text differs in any way from that obtained in supplementary files.

We find that 96% of species–article and 91% of gene–article associations predicted using text2genome match those based on GenBank submissions from articles discussing a single species or gene. When compared with a state-of-the-art text-mining method that attempts to associate articles to species or genes by named entity recognition, text2genome exhibits much higher performance than GNAT for species (72%) or gene (51%) prediction. Thus, if researchers are looking for genes specifically investigated at the molecular level in an article, our results indicate that DNA sequences in text provide a richer source of information than gene names. It is important to point out that our evaluation of these systems is benchmarked against genes from associated GenBank sequence submissions spanning a wide range or organisms. Since many more genes are mentioned in the literature than are actually studied experimentally and since GNAT only recognizes genes for a limited set of species, the performance of GNAT on our GenBank evaluation set may be reduced relative to benchmarks performed on gene names (Hakenberg *et al.*, 2008).

For both text2genome and GNAT, system performance is related to the number of predictions made per paper. The effects of multiple predictions are greater for genes relative to species for both systems, and influence precision and recall differentially for text2genome and GNAT. The difficulty that both systems have for gene prediction in documents that discuss many genes is consistent with the fact that human annotators do not always agree when asked to curate genes in articles [69–91% depending on the dataset (Colosimo *et al.*, 2005; Morgan *et al.*, 2008)]. Despite these differences, there is a substantial degree of overlap between GNAT and text2genome gene–article mappings for some species such as human, suggesting that future full-text mining systems could fruitfully integrate sequence extraction together with named entity recognition to predict gene–article associations (Aerts *et al.*, 2008).

In addition to providing bidirectional links between articles and genes or species, text2genome allows accessing the biomedical literature using the powerful tools of modern genome browsers. In this manner, text2genome joins a limited number of other hybrid text-mining/genome bioinformatics systems that provide mechanisms to interpret the biomedical literature via genome browsers, such as PosMed (Yoshida *et al.*, 2009) and LitTrack (http://littrack.chop.edu/cgi-bin/hgTracks). However, since PosMed and LitTrack rely on gene name recognition methods and therefore can only map articles to the gene level, genomic coordinates must be inferred indirectly by these systems, whether they are appropriate or not. By mapping at the DNA sequence level itself, text2genome can directly identify the exact set of nucleotides in a genome sequence that is analyzed in a study. This distinguishing feature of our system is critical for researchers studying non-genic sequences such as *cis*-regulatory regions or miRNA binding sites. Database curators in these and other areas could use our system to aid in the prioritization and extraction of experimental data from papers.

Only ∼1% of all MEDLINE articles are available at the moment for full-text mining in the OA section of PMC. If we were to mine the full text and supplementary files of all 16.5 million articles in

MEDLINE from 1970 to the present, we would expect to harvest sequences from ~3 million articles using the text2genome approach. As we work toward this goal, we hope that the results presented here encourage other free-access and subscription-model publishers to permit the extraction and mapping of DNA sequences within their articles, to the mutual benefit of researchers, database curators and publishers alike.

ACKNOWLEDGEMENTS

We thank Guy Cochrane, Micaela Gallozzi, Nick Gresham, Nicole Rusk, Bruno Strasser, Jonathan Wren and members of the Bergman group for contributing ideas and assistance during the project. We thank Jörg Hakenberg, Jean-Stephane Joly, Sam Griffiths-Jones, Adriaan Klinkenberg, Goran Nenadic, Pedro Olivares-Chauvet and five anonymous referees for helpful comments on the manuscript. This work was performed in part using computational facilities at the Vital-IT (http://www.vital-it.ch) Center for high-performance computing of the Swiss Institute of Bioinformatics.

Funding: European Science Foundation (Frontiers in Functional Genomics travel award to M.H.); the Biotechnology and Biological Sciences Research Council (CASE studentship to M.G., ISIS travel award 2076 to C.M.B., grant BB/G000093/1 to C.M.B., BB/E012868/1 to C.M.B.); the European Commission (grant HEALTH-F4-2008-223210 to C.M.B.).

Conflict of Interest: none declared.

REFERENCES

Aerts,S. *et al.* (2008) Text-mining assisted regulatory annotation. *Genome Biol.*, **9**, R31.

Anderson,N.R. *et al.* (2006) On the persistence of supplementary resources in biomedical publications. *BMC Bioinformatics*, **7**, 260.

Benson,D.A. *et al.* (2010) GenBank. *Nucleic Acids Res.*, **38**, D46–D51.

Cock,P.J. *et al.* (2009) Biopython: freely available Python tools for computational molecular biology and bioinformatics. *Bioinformatics*, **25**, 1422–1423.

Colosimo,M.E. *et al.* (2005) Data preparation and interannotator agreement: BioCreAtIvE task 1B. *BMC Bioinformatics*, **6** (Suppl. 1), S12.

Dowell,R.D. *et al.* (2001) The distributed annotation system. *BMC Bioinformatics*, **2**, 7.

The FlyBase Consortium (2003) The FlyBase database of the Drosophila genome projects and community literature. *Nucleic Acids Res.*, **31**, 172–175.

Fulp,C.T. *et al.* (2008) Identification of Arx transcriptional targets in the developing basal forebrain. *Hum. Mol. Genet.*, **17**, 3740–3760.

Garcia-Remesal,M. *et al.* (2010a) A method for automatically extracting infectious disease-related primers and probes from the literature. *BMC Bioinformatics*, **11**, 410.

Garcia-Remesal,M. *et al.* (2010b) PubDNA Finder: a web database linking full-text articles to sequences of nucleic acids. *Bioinformatics*, **26**, 2801–2802.

Gerner,M. *et al.* (2010) LINNAEUS: a species name identification system for biomedical literature. *BMC Bioinformatics*, **11**, 85.

Gibson,U.E. *et al.* (1996) A novel method for real time quantitative RT-PCR. *Genome Res.*, **6**, 995–1001.

Gray,P.W. *et al.* (1987) The murine tumor necrosis factor-beta (lymphotoxin) gene sequence. *Nucleic Acids Res.*, **15**, 3937.

Hakenberg,J. *et al.* (2008) Inter-species normalization of gene mentions with GNAT. *Bioinformatics*, **24**, i126–i132.

Holley,R.W. *et al.* (1965) Structure of a ribonucleic acid. *Science*, **147**, 1462–1465.

Hubbard,T.J. *et al.* (2009) Ensembl 2009. *Nucleic Acids Res.*, **37**, D690–D697.

Karolchik,D. *et al.* (2004) The UCSC Table Browser data retrieval tool. *Nucleic Acids Res.*, **32**, D493–D496.

Kent,W.J. (2002) BLAT–the BLAST-like alignment tool. *Genome Res.*, **12**, 656–664.

Kersey,P.J. *et al.* (2010) Ensembl genomes: extending Ensembl across the taxonomic space. *Nucleic Acids Res.*, **38**, D563–D569.

Krallinger,M. *et al.* (2008) Linking genes to literature: text mining, information extraction, and retrieval applications for biology. *Genome Biol.*, **9** (Suppl. 2), S8.

Maglott,D. *et al.* (2007) Entrez Gene: gene-centered information at NCBI. *Nucleic Acids Res.*, **35**, D26–D31.

Morgan,A.A. *et al.* (2008) Overview of BioCreative II gene normalization. *Genome Biol.*, **9** (Suppl. 2), S3.

Rhead,B. *et al.* (2010) The UCSC genome browser database: update 2010. *Nucleic Acids Res.*, **38**, D613–D619.

Roberts,R.J. (2001) PubMed central: the GenBank of the published literature. *Proc. Natl Acad. Sci. USA*, **98**, 381–382.

Semon,D. *et al.* (1987) Nucleotide sequence of the murine TNF locus, including the TNF-alpha (tumor necrosis factor) and TNF-beta (lymphotoxin) genes. *Nucleic Acids Res.*, **15**, 9083–9084.

Shtatland,T. *et al.* (2007) PepBank - a database of peptides based on sequence text mining and public peptide data sources. *BMC Bioinformatics*, **8**, 280.

Vandesompele,J. *et al.* (2002) Accurate normalization of real-time quantitative RT-PCR data by geometric averaging of multiple internal control genes. *Genome Biol.*, **3**, RESEARCH0034.

Visel,A. *et al.* (2009) ChIP-seq accurately predicts tissue-specific activity of enhancers. *Nature*, **457**, 854–858.

Weiss,M. S. *et al.* (2010) Citations in supplementary material. *Acta Cryst.*, **D66**, 1269–1270.

Wren,J.D. *et al.* (2005) Markov model recognition and classification of DNA/protein sequences within large text databases. *Bioinformatics*, **21**, 4046–4053.

Yoshida,Y. *et al.* (2009) PosMed (Positional Medline): prioritizing genes with an artificial neural network comprising medical documents to accelerate positional cloning. *Nucleic Acids Res.*, **37**, W147–W152.