Bioentities: a resource for entity recognition and relationship resolution in biomedical text mining

John A. Bachman[1,*], Benjamin M. Gyori[1,*] Peter K. Sorger[1]

1 Laboratory of Systems Pharmacology, Department of Systems Biology, Harvard Medical School, 200 Longwood Avenue, Boston, MA 02115

*** Authors contributed equally**

Abstract

Background: For automated reading of scientific publications to extract useful information about molecular mechanisms it is critical that genes, proteins and other entities be correctly associated with uniform identifiers, a process known as named entity linking or "grounding." Correct entity identification is essential for resolving relationships between mined information, curated interaction databases, and biological datasets. The accuracy of this process is largely dependent on the availability of resources allowing computers to link commonly-used abbreviations and synonyms found in literature to uniform identifiers.

Results: In a task involving automated reading of ∼215,000 articles using the REACH event extraction software we found that grounding was disproportionately inaccurate for multi-protein families (e.g., "AKT") and named complexes that involve multiple molecular entities (e.g. "NF-κB"). To address this problem we created Bioentities, a manually curated resource defining protein families and complexes as they are commonly referred to in text. In Bioentities the gene-level constituents of families and complexes are defined in a flexible format allowing for multi-level, hierarchical membership. To create Bioentities, text strings corresponding to entities were identified empirically from literature and linked manually to uniform identifiers; these identifiers were also mapped to equivalent entries in multiple related databases. Bioentities also includes curated prefix and suffix patterns that improve named entity recognition and event extraction. Evaluation on a distinct corpus of ∼54,000 articles showed that incorporating Bioentities into the set of databases used for grounding significantly increased grounding accuracy for families and complexes, from 15 to 71%. The hierarchical organization of entities also made it possible to integrate otherwise unconnected mechanistic information across families, subfamilies, and individual proteins.

Conclusion: Bioentities is an effective tool for improving named entity recognition, grounding, and relationship resolution in automated reading of biomedical text. The content in Bioentities is available in both tabular and Open Biomedical Ontology formats at `https://github.com/sorgerlab/bioentities` under the Creative Commons CC0 license and has been integrated into the TRIPS/DRUM and REACH reading systems.

Keywords: Text mining, protein families, grounding, named entity linking, named entity recognition, biocuration, event extraction, natural language processing

Background

A critical challenge in contemporary molecular biology is integrating detailed mechanistic information about specific genes and proteins with less precise genome-scale information about the networks in which these genes participate. Large networks of molecular

mechanisms powerful tools for interpreting large-scale data in the context of prior knowledge [1–4]. The creation of exchange formats such as BioPAX [5] has allowed disparate databases to be aggregated into uniform resources such as Pathway Commons [6], greatly expanding the availability of machine-readable information about biochemical mechanisms. However, a significant fraction of the information available in the literature has not been recorded in pathway databases. Text mining has the potential to address this and augment curated resources by automating the process of extracting molecular mechanisms from the literature, but machines are not yet able to extract mechanisms with a quality equal to that of human curators [7].

A key challenge in event extraction from biomedical text is correctly identifying named entities (e.g., genes, proteins, chemicals, cell lines, diseases, etc.) and linking these entities to corresponding identifiers in formal ontologies. These processes are known as *named entity recognition* (NER) and *named entity linking* or "grounding" respectively. Mechanisms involving entities without associated identifiers cannot be used by machines for downstream assembly and interpretation tasks, and systematic *misidentification* of entities clutters extracted networks with errors that skew subsequent analysis.

Grounding errors occur when the named entity text identified during text parsing cannot be mapped correctly to any relevant entity's official name or relevant synonym. Such errors have three principal causes. First, during the NER step the system may extract too little text for that entity (e.g., "MAP" instead of "MAP kinase"), or too much (e.g., "AKT1 protein" instead of "AKT1") for a successful match to an entry in an ontology. Second, the entity identified in the text might be missing from curated databases used for grounding. Finally, the matching algorithm used for grounding might not handle names with affixes, which are typically domain-specific (e.g., "FLAG-AKT1" or "pAKT1", indicating AKT1 fused to a FLAG epitope tag and phosphorylated AKT1, respectively).

Protein families (e.g., "RAS", "AKT") and multi-protein complexes (e.g., "NF-κB", "AP-1") present a recurrent challenge for NER and grounding because assertions about these entities are abundant in literature and the most comprehensive gene and protein databases (e.g., UniProt, HGNC) do not contain identifiers or lexical synonyms for them. Relevant identifiers can be found in protein family databases (InterPro, PFAM, NextProt) and curated interaction databases (Reactome, SIGNOR, OpenBEL) allowing complexes and families to be resolved into their constituent genes. However, such databases generally lack lexical synonyms corresponding to the many ways in which entities are referenced in text, limiting their value for literature mining; they only rarely contain mappings to equivalent entities across multiple resources, making it difficult to integrate diverse sources of information.

In this paper we describe Bioentities, a curated resource that improves grounding and relationship resolution for molecular entities commonly encountered in mining and curation of biomedical text. Bioentities contains a set of identifiers representing protein families and complexes along with multiple types of mappings: (i) links between text strings and Bioentities identifiers, (ii) between Bioentities identifiers and identifiers representing protein families and complexes in other resources, and (iii) between Bioentities families/complexes and their constituent members. Bioentities also contains a list of prefixes and suffixes frequently appended to protein names that impact NER and event extraction. The Bioentities resource consists of a set of comma-separated value (CSV) files listing entities and relations, along with Python scripts for checking consistency and identifying equivalent identifiers in other databases. Bioentities is hosted on GitHub at `https://github.com/sorgerlab/bioentities` and is made available under the Creative Commons CC0 license. It is also available in the Open Biomedical Ontology (OBO) format and can be accessed via the NCBO BioPortal [8] at `http://bioportal.bioontology.org/ontologies/BE`.

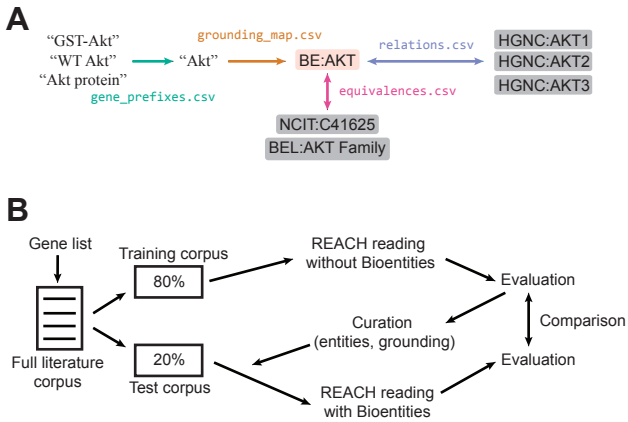

Figure 1. Bioentities links named entities to protein families and complexes and their constituents. **(A)** Structure of Bioentities content. The affixes in `gene_prefixes.csv` can be used to improve recognition of molecular entity names, which can be linked to database identifiers using the lexical synonyms in `grounding_map.csv`. Bioentities itself contains identifiers representing families and complexes which are mapped to corresponding identifiers in other databases in `equivalences.csv`. Hierarchical relationships among families, complexes, and genes are listed in `relations.csv`. **(B)** Workflow for curation and evaluation. A gene list was used to define a corpus of articles that was divided into two subsets, "training" and "test". The "training" corpus was processed with REACH and results were evaluated and used to guide curation. The "test" corpus was processed after incorporation of Bioentities and results were compared against the baseline from the training corpus.

Construction and Content

Development of Bioentities was motivated by an empirical analysis of grounding accuracy in events extracted by the REACH biomedical literature mining software [9]. As described in detail below, we found that grounding of protein families was disproportionately inaccurate and that a relatively small proportion of frequently occurring entities accounted for the bulk of grounding errors. An examination of existing resources highlighted the fragmented nature of data on protein families and complexes and its general lack of suitability for literature mining. We therefore set out to curate a "bridging" resource to link available information about families, complexes, and other frequently misgrounded entities across a diverse set of existing bioinformatics databases.

At the core of Bioentities is a set of identifiers representing protein families and complexes (Figure 1A). Bioentities represents the hierarchical relationships of these high-level entities to each other and to individual genes, along with corresponding synonyms in text and cross-references to other databases where available. Entities and mappings are recorded in a set of CSV files.

Selection of corpus for curation and evaluation

To empirically guide curation of entities and synonyms based on their appearance in literature we selected a corpus of articles relevant to the proteins, protein families, complexes, and molecular events frequently curated in existing databases (Figure 1B). Specifically, we compiled a list of 3,752 human genes from the signaling proteins included in Reactome [10] and the members of protein families and named complexes defined in OpenBEL resource files [11]. From this gene list a corpus of 269,489 papers was assembled

Table 1. Composition of article corpus by source.

Text Source Type	Number	%
MEDLINE Abstract	187,176	69.5
Elsevier XML	35,327	13.1
Pubmed Central Open Access Subset XML	32,113	11.9
Pubmed Central Author Manuscript XML	13,777	5.1
No content retrieved	267	0.1
Pubmed Central Open Access Subset text file	52	0.02
Total	269,489	100

by retrieving papers curated for each gene from the Entrez Gene database [12]. Article abstracts were obtained from MEDLINE; full texts were downloaded either from the Pubmed Central Open Access subset (in XML or text format), the Pubmed Central Author Manuscript Collection, or via the Elsevier text and data mining API (Table 1).

Event extraction from text using REACH and INDRA

The corpus of ∼270,000 papers was processed with the REACH event extraction software [9], yielding a set of sentences, named entities, and extracted relations (Figure 1B). REACH is built on widely-used approaches to syntactic parsing and named-entity recognition, suggesting that use of other event extraction tools would yield results similar to ours. In particular, REACH uses the Stanford CoreNLP parser [13] for syntactic parsing and draws information on biology-specific named entities from Uniprot, InterPro, PFAM, HMDB, ChEBI, Gene Ontology, MeSH, Cellosaurus, ATCC, and CellOntology. As a final step we used the INDRA software [14] to convert events extracted by REACH into INDRA Statements, a format suitable for analyzing and assembling sets of mechanisms into networks and executable models of various kinds.

Characterizing patterns of grounding errors

The set of entities and events extracted by REACH was used to characterize patterns of grounding errors and prioritize entities and their lexical synonyms for subsequent curation (Figure 1B). Prior to curation, the corpus of articles was divided into two sets: a "training" set and a "test" set consisting of 80% (215,360) and 20% (53,840) of articles, respectively. The "training" set of articles was processed with REACH before the addition of Bioentities to evaluate baseline grounding accuracy and guide curation. After curation was completed, the "test" set of articles was processed with a version of REACH incorporating Bioentities. The partitioning of articles was performed to ensure that estimates of grounding accuracy would not be biased towards the specific set of articles used for curation.

Definition of protein families and complexes and their constituents

Identifiers for protein families and complexes in Bioentities were created by drawing on two resources: 1) identifiers created *de novo* in Bioentities to correspond to named entities encountered in event extraction, and 2) identifiers drawn from the OpenBEL resource. In the first case, identifiers were prioritized by their frequency of occurrence among extracted events, with common entities such as "NF-kappaB", "Ras", "PI3-kinase", "Akt", etc., accounting for a significant fraction of grounding errors. In the case of OpenBEL, identifiers for protein families and complexes were drawn from the resource files `protein-families.xbel` and `named-complexes.xbel`, respectively. These files are accessible from the OpenBEL GitHub repository at

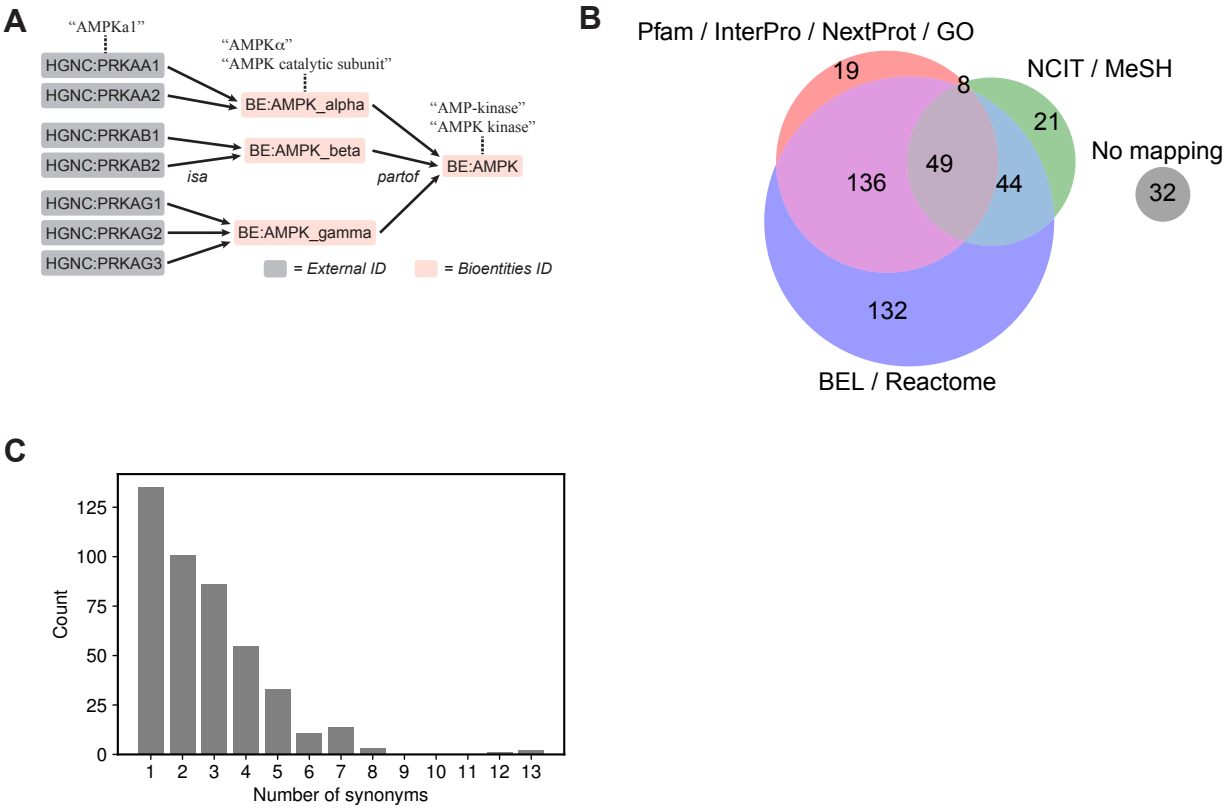

Figure 2. Bioentities links identifiers for families and complexes to members, other databases, and lexical synonyms. **(A)** Bioentities uses *isa* and *partof* predicates to represent the hierarchical relationships between specific genes, families and complexes. Lexical synonyms can be associated with entities at each level. **(B)** Mappings of Bioentities identifiers to outside databases. **(C)** Number of lexical synonyms curated for Bioentities identifiers in the grounding map.

`https://github.com/OpenBEL/openbel-framework-resources`. The full list of all Bioentities identifiers is contained in the text file `entities.csv`.

Members of protein families and complexes are enumerated in the file `relations.csv` using two types of relations: *isa* and *partof*, denoting membership in a family or a complex, respectively (Figure 1A). These relationships can be applied hierarchically to describe multi-level protein subfamily relationships or protein complexes that are hetero-oligomers of subunits belonging to distinct families (Figure 2A). For example, 5' AMP-activated protein kinase, or AMPK, is a heterotrimeric protein consisting of alpha, beta, and gamma subunits, the alpha and beta subunit families containing two isoforms each, and the gamma subunit family containing three. This hierarchical structure can be represented in Bioentities by using a combination of *isa* and *partof* relationships to link the identifiers for the subunit genes to Bioentities-specific identifiers for the subunit families and the full complex.

Information on protein family and complex membership was drawn from OpenBEL resource files, HGNC, Reactome, InterPro, and Wikipedia, and manually curated for consistency. Where there were discrepancies among sources about family membership we prioritized sets of members corresponding to common usage in literature and curation. For example, the InterPro entry corresponding to the Ras protein family (`IPR020849`) lists 145 human proteins as members, whereas common usage in literature and interaction databases

would recognize only KRAS, NRAS, and HRAS.

Mapping Bioentities identifiers to related resources

Entities defined in Bioentities are cross-referenced to corresponding identifiers in other databases and ontologies in the equivalences file (`equivalences.csv`; Figure 1A). Figure 2B shows the subsets of Bioentities identifiers containing mappings to different types of external databases: databases of interactions curated from literature (OpenBEL, Reactome), databases containing specific information about protein families and complexes (PFAM, InterPro, NextProt, and Gene Ontology), and general-purpose biomedical vocabularies (NCIT, MeSH). There are 32 unmapped entries for which no equivalent entry was found in external databases. The unmapped entries are implicitly defined in Bioentities by the specific genes that they contain as members.

Identifier mappings between Bioentities and Reactome and InterPro were obtained in a semi-automated fashion. The gene-level members of each Bioentities family and complex were used to query Reactome and InterPro for families and complexes containing these genes. Reactome and InterPro families with equivalent sets of members were incorporated into `equivalences.csv`. Python scripts for generating these mappings are available in the Bioentities GitHub repository at `import/reactome_mappings.py` and `import/interpro_mappings.py`.

Additional identifier mappings (to PFAM, NCIT, NextProt, GO and MeSH) were collected by entering Bioentities identifiers and lexicalization into the TRIPS/DRUM web service available at `http://trips.ihmc.us/parser/cgi/drum` [15]. The TRIPS/DRUM web service returned identifier mappings and their scores based on string matches to several databases, which were then manually curated.

Curation of lexical synonyms for entities

Entities defined in the Bioentities resource are associated with lexical synonyms in the grounding map (`grounding_map.csv`; Figure 1A). These synonyms allow natural language processing tools to match named entities extracted from text to the protein families and complexes contained in the Bioentities hierarchy.

Lexical synonyms were curated in two ways. First, named entities extracted from the "training" articles by REACH were sorted by frequency, and named entities corresponding to Bioentities families and complexes were added to the grounding map. Entries were also added to the grounding map for frequently occurring but incorrectly grounded named entities of other types (e.g., proteins, chemicals, and biological processes). For less-frequently encountered families and complexes, synonyms were curated using a different approach: names and synonyms for the gene-level members of families and complexes were searched against the named entities extracted by REACH. Potential matches were identified by fuzzy string matching (Levenshtein distance [16]) using the Python `fuzzywuzzy` package and subsequently manually curated.

Of the 2,076 entries in the grounding map, 1,186 map to Bioentities identifiers; the remaining 890 mappings are to frequently occurring proteins, chemicals, and biological processes. The distribution of lexical synonyms across the set of Bioentities identifiers is shown in Figure 2C. The frequently-occurring entities `NFkappaB` and `ERK` have the most synonyms, with 13 and 9, respectively; many other less-frequently occurring entities have only a single synonym. Example synonyms for NF-κB include "NF-kB", "NFkappaB", and "NF-kappaB TFs"; synonyms for ERK include "ERK 1/2", "ERKs", and "Extracellular Signal Regulated Kinase".

Table 2. Gene/protein affix types.

Category	# of affixes	Example
Experimental context	63	`eGFP-{Gene name}`
Protein state	30	`phospho-{Gene name}`
Inhibition	22	`shRNA-{Gene name}`
Generic descriptor	12	`proto-oncogene {Gene name}`
Species	9	`mmu-{Gene name}`
mRNA grounding	1	`{Gene name} mRNA`

Table 3. Entity frequency and grounding accuracy for 300 entities, with and without Bioentities. Standard error was calculated using the formula $\sqrt{(k/n)(1-k/n)/n}$ where k is the number of samples in the given category and n is the total number of samples.

	No Bioentities				With Bioentities			
	#	Entity %	# Corr.	% Corr.	#	Entity %	# Corr.	% Corr.
Protein/gene	169	56.3	133	**78.7 ± 2.9**	172	57.3	154	**89.5 ± 2.9**
Family/complex	53	17.7	8	**15.1 ± 0.9**	52	17.3	37	**71.2 ± 1.9**
Small molecule	33	11.0	18	54.5 ± 1.4	26	8.7	14	53.8 ± 1.2
Biological process	28	9.3	24	85.7 ± 1.6	28	9.3	28	100.0 ± 1.7
Other/unknown	16	5.3	0	0.0 ± 0.0	21	7.0	0	0.0 ± 0.0
microRNA	1	0.3	0	0.0 ± 0.0	1	0.3	0	0.0 ± 0.0

Curation of gene/protein affixes

References to genes and proteins in the literature are often modified by affixes that describe modifications or other context. For example, "mmu-AKT1" and "pAKT1" refer to murine and phosphorylated AKT1, respectively. A list of 137 case-sensitive affixes was tabulated by alphabetically sorting a list of ∼80,000 named entities resulting from event extraction and manually identifying common affix patterns. These affixes were subsequently grouped into six semantic categories (Table 2). The largest category, "experimental context", contains affixes used to identify the precise variant of a gene used in an experiment; these often refer to protein tags or gene delivery methods. Two of the six categories affect event extraction as well as grounding: "protein state" affixes contain information on modification, location and mutation states, while "inhibition" affixes invert the apparent polarity of an extracted event. For example, a positive regulation event mediated by "BRAF siRNA" actually represents a negative regulation by BRAF itself. The full list of affixes can be found in the CSV file `gene_prefixes.csv` (Figure 1A).

Utility and Discussion

Protein families and complexes appear frequently in literature and are often incorrectly grounded

To evaluate baseline grounding performance (i.e., without Bioentities) we manually scored a random sample of 300 named entities generated by running REACH on the training corpus. Entities were categorized by type (protein/gene, family/complex, small molecule, biological process, microRNA, and other/unknown) and the database mappings identified by REACH were scored for correctness (Table 3). Where the entity text alone was insufficient to evaluate grounding accuracy, the sentence used for extraction was examined in the context of the original paper.

We found that references to protein families and complexes were second only to genes and proteins in the frequency of their occurrence in text, accounting for 17.7% of all extracted

entities (Table 3). Grounding accuracy was substantially lower for families and complexes relative to genes and proteins, with only 15.1% of families and complexes correctly grounded compared to 78.7% for individual proteins (Table 3). The 15% rate of correct grounding for families and complexes reflected accurate matches to identifiers in InterPro or PFAM.

Close inspection of attempts by REACH to ground frequently-occurring families and complexes in the absence of Bioentities revealed two types of errors. First, some named entities corresponding to families and complexes had no associated identifiers, indicating a failure to find a match in any database. This was true of the entity "Ras", as well as the most frequently occurring family-level entity, "NFκB". Second, many family-level entities were misgrounded due to perfect matches to obscure synonyms for other genes. For example, "ERK" was grounded to the human gene *EPHB2*, and "AKT" to the *Dictyostelium discoideum* gene *pkbA*, due to the fact that these strings were included in the lists of synonyms for these proteins in Uniprot (entries `P29323` and `P54644`, respectively).

In estimating the potential return on curation effort, we noted that a small fraction of frequently occurring entities accounted for a substantial fraction of all entities lacking any identifiers (Figure 3A, orange curve). For example, half of all ungrounded string occurrences in the training corpus were due to just the top 2.4% most frequently occurring strings (2,666 distinct strings). This suggested that curation focusing specifically on frequently occurring misgrounded entities could substantially improve overall grounding and reading performance.

Use of Bioentities in text mining significantly improves grounding for protein families and complexes

Following the manual curation of Bioentities identifiers and associated synonyms and the integration of Bioentities into REACH and INDRA, we performed a second evaluation on a random sample of 300 named entities drawn from the results of processing the test corpus (Table 3). The frequency of entity types was comparable between the training and test samples, with proteins/genes and families/complexes accounting for roughly three-quarters of all entities. Improvements in grounding were substantial for both classes, with grounding accuracy for families and complexes rising from 15% to 71% (Figure 3B; Table 3). Grounding accuracy for proteins and genes increased from 79% to 90%, an improvement attributable to the curation of synonyms for frequently occurring proteins.

An analysis of the distribution of the remaining ungrounded entities showed that Bioentities addressed a substantial proportion of the most frequently occurring grounding failures (Figure 3A, green curve). To examine the impact of grounding improvements at the level of extracted events, we calculated the proportion of events consisting either of any or all ungrounded entities, and found that both metrics improved with the use of Bioentities (Figure 3C). It is important to note that these measures, which deal only with entities that were *ungrounded*, represent an underestimate of the overall improvement in grounding because they do not account for cases in which entities were grounded to the *wrong* identifier in the absence of Bioentities.

To determine whether improvements in grounding were driven by a small subset of frequently-occurring entities in Bioentities or were more broadly distributed across families and complexes, we counted the occurrences of mappings to each Bioentities identifier in events extracted from the test corpus. We found that the 15 most frequently-referenced Bioentities identifiers accounted for 50% of all Bioentities groundings (blue bars in Figure 3D); the top five are shown in Table 4. While the distribution is skewed towards these frequently-occurring entities, we also found that 363 of the 441 Bioentities identifiers were mapped to text at least once, suggesting that the great majority of identifiers and lexical synonyms in Bioentities are useful for improving grounding (Figure 3D).

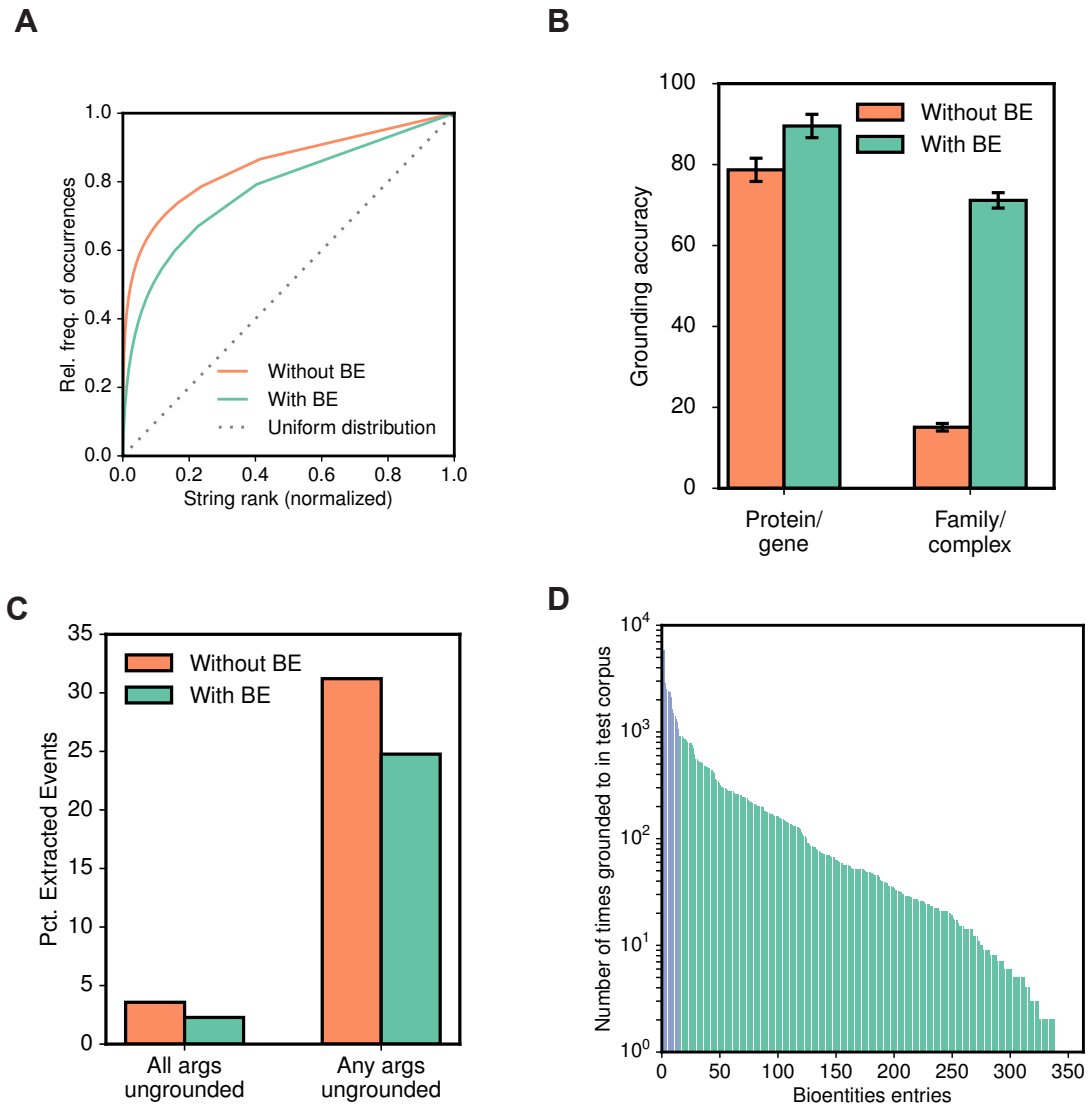

Figure 3. Bioentities improves grounding accuracy. **(A)** Cumulative occurrences of ungrounded entities by frequency of the entity text. Deviation from the gray line, representing a uniform frequency distribution, indicates the extent to which a small number of frequently occurring entities account for a disproportionate share of missed groundings. **(B)** Improvements in grounding accuracy for proteins/genes and families/complexes, with and without the use of Bioentities. **(C)** Reduction in the proportion of extracted events containing ungrounded entities, with and without Bioentities. **(D)** Number of groundings to Bioentities identifiers in the test corpus. The 15 most frequent identifiers account for 50% of all groundings and are shown in blue.

Bioentities resolves hierarchical relationships in extracted events

A key feature of Bioentities is that it allows for relationship resolution not only "horizontally" (between different databases) but also "vertically" (between genes, families, complexes, and any intermediate sets involving these elements). Lexical synonyms can be defined at all levels in the Bioentities hierarchy (Figure 2A). The combination of a hierarchical representation with a mapping of entities to text at each level allows

Table 4. Bioentities entries most frequently grounded to in test corpus, with the absolute number of times grounded to in the test corpus and the percentage normalized to all Bioentities groundings.

	#	% BE
ERK	6,301	7.6
AKT	5,839	7.1
NFkappaB	5,768	7.0
TGFB	2,877	3.5
PI3K	2,486	3.0

information about biological interactions to be correctly organized and cross-referenced.

For example, the Bioentities family `PLC`, representing the family of phospholipase C enzymes, contains both individual genes (e.g., *PLCE1*) and Bioentities subfamilies (e.g., `PLCG`, a sub-family consisting of the genes *PLCG1* and *PLCG2*) as members (Figure 4A). In results from the test corpus we found descriptions of meaningful mechanisms associated with all three levels of this hierarchy—family, subfamily, and genes (Figure 4A). Moreover, relevant events were extracted for 12 of the 15 entities in the phospholipase C entity hierarchy, demonstrating the diversity of available mechanistic information and the importance of relationship resolution.

To characterize the relevance of multi-level relationship resolution more broadly, we counted the number of times a named entity identified in the test corpus was mapped to a Bioentities identifier at three or more hierarchical levels: the gene level (lowest), the top-level family or complex (highest), and any intermediate level (Figure 4B). We found that the identifier level that was used most frequently for mapping differed among protein families and complexes. For example, for `AMPK`, the majority of references in the literature were to the top-level AMPK complex, with a relatively small fraction of references to constituent genes or intermediates. On the other hand, most mappings to the family representing Phospholipase C (`PLC` in Bioentities) were to constituent genes such as *PLCG1*, *PLCD1*, etc. Finally, for the family of Activins (hetero- and homo-dimers of the transforming growth factor beta family, `Activin` in Bioentities), most references were to specific dimer subtypes—Activin A, Activin AB and Activin B—which are at an intermediate level in the Bioentities hierarchy.

Limitations

Although Bioentities provides substantial coverage of protein families and complexes, their lexical synonyms, and mappings to identifiers in other resources, it is not exhaustive. As a resource primarily for literature mining and biocuration, it focuses on the subset of proteins that are described in the literature as being members of families and complexes. While it demonstrably improves grounding performance, further empirically-guided curation of the grounding map will reduce errors still further, and with additional work mappings to other ontologies could become more comprehensive.

Bioentities does not directly address the problem of *ambiguity:* that is, selecting among multiple alternative groundings for the same entity. For example, "MEK" can refer to the family of MAPK/ERK Kinases or to the solvent methyl ethyl ketone. Resolving such ambiguities requires an examination of the named entity in the broader context of the sentence or article. However, the use of Bioentities does increase the likelihood that relevant groundings to protein families will not be missed, and can therefore be considered alongside possible alternative groundings in any ambiguity resolution procedure.

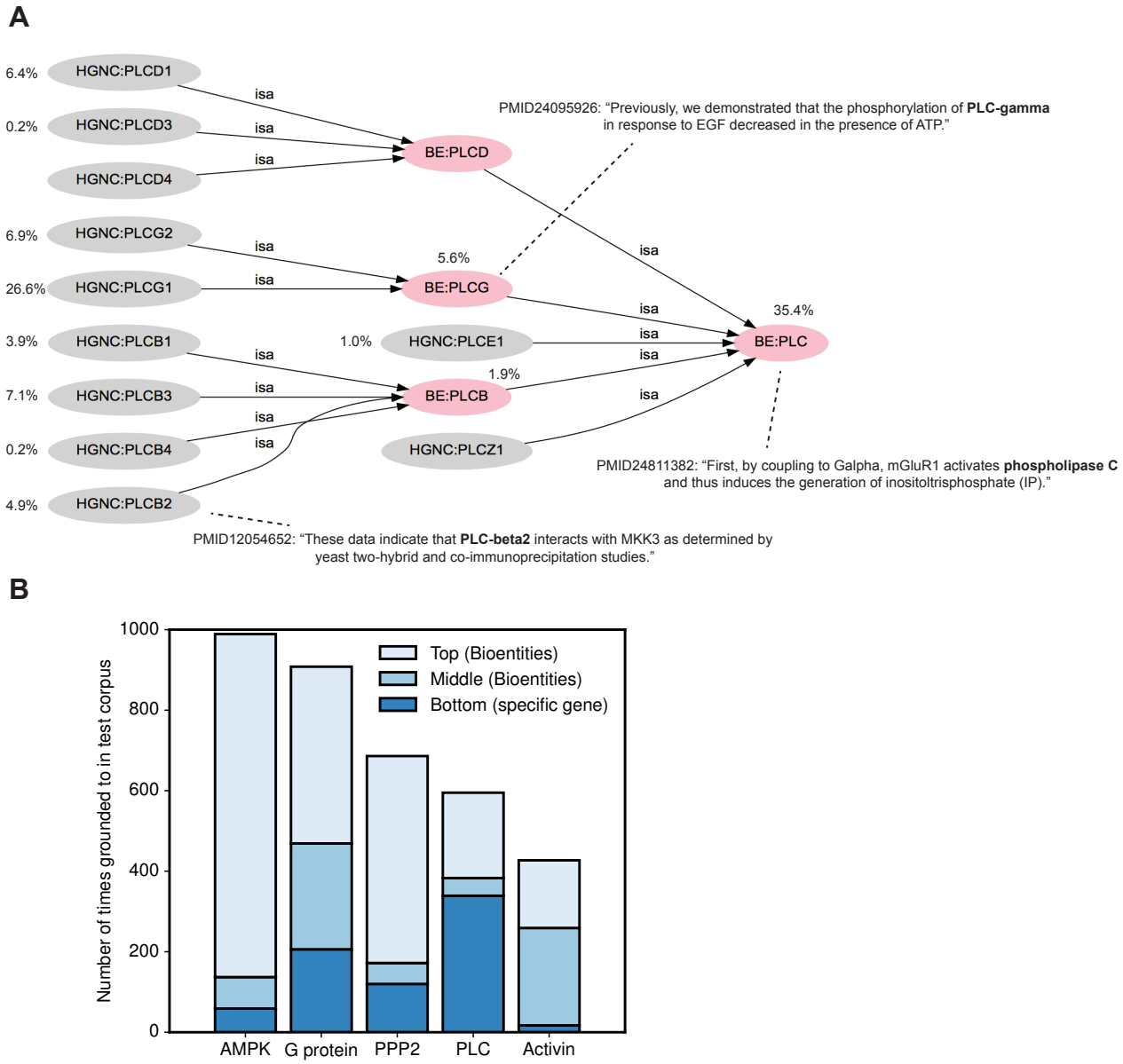

Figure 4. Bioentities facilitates hierarchical resolution of extracted information. **(A)** Hierarchical organization of the phospholipase C protein family (Bioentities identifier `PLC`) along with the proportion of occurrences of each member in the test corpus and examples of sentences yielding information at the different levels. Pink nodes indicate Bioentities families; gray nodes indicate genes. **(B)** Proportion of groundings in the test corpus to gene-level, intermediate-level, or top-level entities for five multi-level families/complexes in Bioentities.

Accessibility and Extensibility

We chose CSV files as the primary representation format for Bioentities to maximize accessibility and extensibility. CSV files can be opened and edited in any spreadsheet program or text editor, allowing biologists with no background in literature mining to assist

in the curation of the grounding map or mappings to other resources. Because the files are hosted on GitHub, other users can easily fork and make use-case specific extensions or other contributions that can be merged back into the main repository.

In addition to the CSV files, Bioentities includes an Open Biomedical Ontologies (OBO) [17] export feature to facilitate integration into OBO-based workflows. Bioentities relations and mappings have been integrated into the TRIPS/DRUM reading system [15] via OBO-exported content.

Conclusions

In this paper we describe the challenge that protein families and multi-protein complexes pose for machine reading of the biomedical literature, and we introduce a new resource, Bioentities, that addresses these challenges and substantially improves grounding. Bioentities fills a gap in existing bioinformatics resources, linking information about families and complexes in protein and pathway databases to a set of lexical synonyms that occur with high frequency in the scientific literature. Empirical evaluation shows that the hierarchical organization of Bioentities enables the integration of mechanistic information about gene families, complexes, and their individual subunits. This is important because information about biochemical mechanisms is often reported in terms of classes of entities whereas large-scale profiling experiments yield data about individual genes and proteins. Bioentities therefore supports the broader goal of making text mining a key contributor to the process of obtaining biological insight from high throughput -omic data by drawing on relevant mechanistic knowledge. We speculate that similar resources for resolving hierarchical relationships among entities could be useful in other domains of machine reading and natural language processing.

Declarations

Ethics approval and consent to participate. Not applicable.

Consent for publication. Not applicable.

Availability of data and material The datasets generated and analyzed during the current study, as well as the source code used to generate results is available in the repository `https://github.com/sorgerlab/bioentities_paper`. Bioentities is available at `https://github.com/sorgerlab/bioentities` under the Creative Commons CC0 license.

Competing interests. The authors declare that they have no competing interests.

Funding. This work has been supported by DARPA grants W911NF-14-1-0397 and W911NF-15-1-0544.

Authors' contributions. JAB and BMG conceived the software. JAB, BMG and PKS wrote the paper.

Acknowledgments. The authors would like to thank Tom Hicks, Mihai Surdeanu and Lucian Galescu for useful discussions and assistance with REACH and TRIPS, and Petar Todorov, Bobby Sheehan, Lily Chylek, Jeremy Muhlich and Isabel Latorre for their contributions to entity curation.

References

1. Babur, O., Gönen, M., Aksoy, B.A., Schultz, N., Ciriello, G., Sander, C., Demir, E.: Systematic identification of cancer driving signaling pathways based on mutual exclusivity of genomic alterations. Genome Biology **16**, 45 (2015). doi:10.1186/s13059-015-0612-6

2. García-Campos, M.A., Espinal-Enríquez, J., Hernández-Lemus, E.: Pathway Analysis: State of the Art. Frontiers in Physiology **6**, 383 (2015). doi:10.3389/fphys.2015.00383

3. Korkut, A., Wang, W., Demir, E., Aksoy, B.A., Jing, X., Molinelli, E.J., Babur, O., Bemis, D.L., Sumer, S.O., Solit, D.B., *et al.*: Perturbation biology nominates upstream–downstream drug combinations in RAF inhibitor resistant melanoma cells. Elife **4**, 04640 (2015)

4. Campbell, J., Ryan, C.J., Brough, R., Bajrami, I., Pemberton, H.N., Chong, I.Y., Costa-Cabral, S., Frankum, J., Gulati, A., Holme, H., Miller, R., Postel-Vinay, S., Rafiq, R., Wei, W., Williamson, C.T., Quigley, D.A., Tym, J., Al-Lazikani, B., Fenton, T., Natrajan, R., Strauss, S.J., Ashworth, A., Lord, C.J.: Large-Scale Profiling of Kinase Dependencies in Cancer Cell Lines. Cell Reports **14**(10), 2490–2501 (2016). doi:10.1016/j.celrep.2016.02.023. Accessed 2017-10-08

5. Demir, E., Cary, M.P., Paley, S., Fukuda, K., Lemer, C., Vastrik, I., Wu, G., D'Eustachio, P., Schaefer, C., Luciano, J., Schacherer, F., Martinez-Flores, I., Hu, Z., Jimenez-Jacinto, V., Joshi-Tope, G., Kandasamy, K., Lopez-Fuentes, A.C., Mi, H., Pichler, E., Rodchenkov, I., Splendiani, A., Tkachev, S., Zucker, J., Gopinath, G., Rajasimha, H., Ramakrishnan, R., Shah, I., Syed, M., Anwar, N., Babur, O., Blinov, M., Brauner, E., Corwin, D., Donaldson, S., Gibbons, F., Goldberg, R., Hornbeck, P., Luna, A., Murray-Rust, P., Neumann, E., Ruebenacker, O., Samwald, M., van Iersel, M., Wimalaratne, S., Allen, K., Braun, B., Whirl-Carrillo, M., Cheung, K.-H., Dahlquist, K., Finney, A., Gillespie, M., Glass, E., Gong, L., Haw, R., Honig, M., Hubaut, O., Kane, D., Krupa, S., Kutmon, M., Leonard, J., Marks, D., Merberg, D., Petri, V., Pico, A., Ravenscroft, D., Ren, L., Shah, N., Sunshine, M., Tang, R., Whaley, R., Letovksy, S., Buetow, K.H., Rzhetsky, A., Schachter, V., Sobral, B.S., Dogrusoz, U., McWeeney, S., Aladjem, M., Birney, E., Collado-Vides, J., Goto, S., Hucka, M., Le Novère, N., Maltsev, N., Pandey, A., Thomas, P., Wingender, E., Karp, P.D., Sander, C., Bader, G.D.: The BioPAX community standard for pathway data sharing. Nature Biotechnology **28**(9), 935–942 (2010). doi:10.1038/nbt.1666. Accessed 2017-10-05

6. Cerami, E.G., Gross, B.E., Demir, E., Rodchenkov, I., Babur, O., Anwar, N., Schultz, N., Bader, G.D., Sander, C.: Pathway Commons, a web resource for biological pathway data. Nucleic Acids Research **39**(Database issue), 685–690 (2011). doi:10.1093/nar/gkq1039

7. Ananiadou, S., Thompson, P., Nawaz, R., McNaught, J., Kell, D.B.: Event-based text mining for biology and functional genomics. Briefings in Functional Genomics **14**(3), 213–230 (2015). doi:10.1093/bfgp/elu015

8. Whetzel, P.L., Noy, N.F., Shah, N.H., Alexander, P.R., Nyulas, C., Tudorache, T., Musen, M.A.: BioPortal: enhanced functionality via new Web services from the National Center for Biomedical Ontology to access and use ontologies in software applications. Nucleic acids research **39**(Web Server issue), 541–545 (2011)

9. Valenzuela-Escárcega, M.A., Hahn-Powell, G., Hicks, T., Surdeanu, M.: A domain-independent rule-based framework for event extraction. In: ACL-IJCNLP

2015 - 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing, Proceedings of System Demonstrations, pp. 127–132. Association for Computational Linguistics (ACL), Beijing, China (2015)

10. Fabregat, A., Sidiropoulos, K., Garapati, P., Gillespie, M., Hausmann, K., Haw, R., Jassal, B., Jupe, S., Korninger, F., McKay, S., Matthews, L., May, B., Milacic, M., Rothfels, K., Shamovsky, V., Webber, M., Weiser, J., Williams, M., Wu, G., Stein, L., Hermjakob, H., D'Eustachio, P.: The Reactome pathway Knowledgebase. Nucleic Acids Research **44**(D1), 481–487 (2016). doi:10.1093/nar/gkv1351

11. The OpenBEL Consortium: OpenBEL Framework Resources (2017). `https://github.com/OpenBEL/openbel-framework-resources` Accessed 2017-10-01

12. Maglott, D., Ostell, J., Pruitt, K.D., Tatusova, T.: Entrez Gene: gene-centered information at NCBI. Nucleic Acids Research **33**(Database Issue), 54–58 (2005). doi:10.1093/nar/gki031. Accessed 2017-10-05

13. Manning, C.D., Surdeanu, M., Bauer, J., Finkel, J.R., Bethard, S., McClosky, D.: The stanford corenlp natural language processing toolkit. In: ACL (System Demonstrations), pp. 55–60 (2014)

14. Gyori, B.M., Bachman, J.A., Subramanian, K., Muhlich, J.L., Galescu, L., Sorger, P.K.: From word models to executable models of signaling networks using automated assembly. bioRxiv, 119834 (2017). doi:10.1101/119834. Accessed 2017-10-05

15. Allen, J., de Beaumont, W., Galescu, L., Teng, C.M.: Complex Event Extraction using DRUM. ACL-IJCNLP 2015, 1 (2015)

16. Navarro, G.: A Guided Tour to Approximate String Matching. ACM Comput. Surv. **33**(1), 31–88 (2001). doi:10.1145/375360.375365. Accessed 2017-10-05

17. Smith, B., Ashburner, M., Rosse, C., Bard, J., Bug, W., Ceusters, W., Goldberg, L.J., Eilbeck, K., Ireland, A., Mungall, C.J., Leontis, N., Rocca-Serra, P., Ruttenberg, A., Sansone, S.-A., Scheuermann, R.H., Shah, N., Whetzel, P.L., Lewis, S.: The OBO Foundry: coordinated evolution of ontologies to support biomedical data integration. Nature biotechnology **25**(11), 1251 (2007). doi:10.1038/nbt1346. Accessed 2017-10-05