BIOINFORMATICS

Vol. 24 ECCB 2008, pages i126–i132
doi:10.1093/bioinformatics/btn299

Inter-species normalization of gene mentions with GNAT

Jörg Hakenberg[1,*], Conrad Plake[2,3], Robert Leaman[1], Michael Schroeder[2] and Graciela Gonzalez[4]

[1]Department of Computer Science and Engineering, Arizona State University, Tempe, AZ 85287, USA, [2]Biotechnological Centre, Technische Universität Dresden, Tatzberg 47–51, 01307 Dresden, [3]Transinsight GmbH, Tatzberg 47–51, 01307 Dresden, Germany and [4]Department of Biomedical Informatics, Arizona State University, Phoenix, AZ 85004, USA

ABSTRACT

Motivation: Text mining in the biomedical domain aims at helping researchers to access information contained in scientific publications in a faster, easier and more complete way. One step towards this aim is the recognition of named entities and their subsequent normalization to database identifiers. Normalization helps to link objects of potential interest, such as genes, to detailed information not contained in a publication; it is also key for integrating different knowledge sources. From an information retrieval perspective, normalization facilitates indexing and querying. Gene mention normalization (GN) is particularly challenging given the high ambiguity of gene names: they refer to orthologous or entirely different genes, are named after phenotypes and other biomedical terms, or they resemble common English words.

Results: We present the first publicly available system, GNAT, reported to handle inter-species GN. Our method uses extensive background knowledge on genes to resolve ambiguous names to EntrezGene identifiers. It performs comparably to single-species approaches proposed by us and others. On a benchmark set derived from BioCreative 1 and 2 data that contains genes from 13 species, GNAT achieves an F-measure of 81.4% (90.8% precision at 73.8% recall). For the single-species task, we report an F-measure of 85.4% on human genes.

Availability: A web-frontend is available at http://cbioc.eas.asu.edu/gnat/. GNAT will also be available within the BioCreative MetaService project, see http://bcms.bioinfo.cnio.es.

Contact: joerg.hakenberg@asu.edu

Supplementary information: The test data set, lexica, and links to external data are available at http://cbioc.eas.asu.edu/gnat/

1 INTRODUCTION

In biomedical text mining, researchers try to devise systems that extract relevant information from published literature in an automated way (Zweigenbaum *et al.*, 2007). Such information ranges from classifying texts for relevance to a certain topic, finding the named entities (objects of biomedical interest, such as names of genes or diseases), to associations between these entities (interactions between proteins, genes associated with diseases). Research in named entity recognition (NER) has focused mainly on detecting names of genes and proteins in texts, and extracting information on protein–protein interactions. A prerequisite for further analysis of extracted information by the user is the exact mapping of a gene name as found in a text to an entry in a database, for instance, EntrezGene or GenBank. Only then can the biomedical researcher access additional data on that gene (sequence, origin, SNPs, etc.), or a database curator cross-reference information from the article to a database entry. In the same way, being able to pinpoint the exact gene mentioned in a text and not just its name helps other tasks such as knowledge integration (hyperlinking between texts and different sources), information retrieval, text indexing and question answering.

The task for gene mention normalization (GN, also called gene symbol disambiguation, or more generally, named entity identification) is to find and identify individual genes mentioned in a text; this excludes, for example, references to gene families. In addition, we do not distinguish between genes and gene products (e.g. mRNAs and proteins). The main challenges for GN that need to be addressed are:

(1) ambiguous gene names, that is, names shared among different genes or products ('p21' as an example for a generic abbreviation);

(2) genes from the same family that have similar names, where perhaps the qualifier ('family member 3') cannot be determined;

(3) names that resemble gene names, but in the particular context refer to a disease ('neurofibromasosis 2'), a cell line ('CD8'), etc.;

(4) names that are also common English words ('taxi' or 'cab'), many of which originate from an observed phenotype ('white');

(5) the variety of names used for a single gene, some of which are known previously, but many of which are *ad hoc* names or spelling variations of known names.

Extending the task to inter-species gene mention normalization (ISGN) further complicates most of these challenges. For instance, the context for disambiguation will be more complicated to handle, with all the nuances for the particular species to take into account. Also, common English words are more often used in some species. On top of these issues, ISGN has to address the problem of (6) gene names that are shared among different species (orthologs).

As an introductory example, which also explains the basic idea of our methodology, consider the short text shown in Figure 1. It mentions a gene name ('P54'), and a variety of genes can be considered as potential candidates for this name. The mention of a species, or even a cell line ('RC–K8'), helps to narrow down the search to a particular species. Still, there are five human genes that

*To whom correspondence should be addressed.

The **P54** gene was previously isolated from the chromosome translocation breakpoint region on 11q23 of RC-K8 cells, with t(11;14)(q23;q32). It was found to encode a 472-483-amino-acid (aa) polypeptide belonging to an RNA helicase/translation initiation factor family.

[From PubMed–ID8543178]

Potential candidates for P54, with annotations for each, extracted from EntrezGene and UniProt:

Gene	DDX6	ETS1	FKBP5	NONO	SRFS11
Chromosome	11q23.3	11q23.3	6p21.3-p21.2	Xq13.1	1p31
Length (aa)	483	441	457	471	484
GO (examples)	RNA helicase activity	immune response	FK506 binding	DNA binding	mRNA processing

Fig. 1. The mention of 'P54' was found that potentially refers to a gene. The mention of a human lymphoma cell line, RC-K8, points towards the species human, so we can concentrate on finding the correct gene there. 'P54' is a name shared by five human genes (DDX6, ETS1, FKBP5, NONO and SFRS11). Only two of these genes can be found on band 23 of chromosome 11 (DDX6 and ETS1). Additionally, only the DDX6 protein has a length of 483aa. Looking at the Gene Ontology (GO) terms annotated for DDX6, we find that DDX6 has 'RNA helicase activity', which is also mentioned in the second sentence.

have the synonym 'P54' (perhaps as a slight spelling variation). We retrieve all information known on these five genes, and try to map each of these profiles to the current sentence and, if available, the text surrounding it. In the example, there is a reference to a chromosomal location, the length of a protein that is encoded by the gene and also a GO (Gene Ontology Consortium, 2008) term that is also annotated to one of the candidate genes. All these hints point towards the human DDX6 gene as the solution for the normalization of the gene name 'P54' in that particular sentence.

The problem of GN has been extensively studied in the BioCreative 1 and 2 challenges (Hirschman *et al.*, 2005; Morgan *et al.*, 2008). However, the task was restricted to normalizing gene names from a single species. Various approaches have been proposed, and performance ranged from 79% (mouse) to 92% (yeast) in *F*-measure; for human and fly, results were slightly better than for mouse (81 and 82%, respectively). In this article, we present GNAT, an attempt to solve the more general problem of finding and identifying every gene in a text, regardless of species.

The rest of the article is organized as follows. We first present the methodology of our approach and the evaluation. We then show quantitative and qualitative results. The article concludes with a discussion and related work, and conclusions drawn from our work.

2 SYSTEM AND METHODS

Our methodology is a multi-step procedure of refining an initial set of predictions until a final conclusion is reached. The first component in our approach handles the *recognition of gene mentions* in general. We use a dictionary-matching and machine learning to recognize such mentions. Dictionary-matching is able to *assign candidate identifiers* to each mention; the machine-learning component requires a subsequent step to look up candidate identifiers for predicted mentions. In the next step, we reconsider each predicted mention to *remove names* that resemble a gene name, but in the current context refer to, for instance, a disease, cell line or a gene family. After such *false positive mentions* were eliminated, we start narrowing down the set of candidate IDs per gene mention. First, we try to *find the correct species* (or list of species) and remove identifiers referring to different species. Then, we *disambiguate* each mention by comparing the current text with background information known for each candidate gene. Figure 2 shows an overview of the system's components. In the following, we describe each of them in more detail.

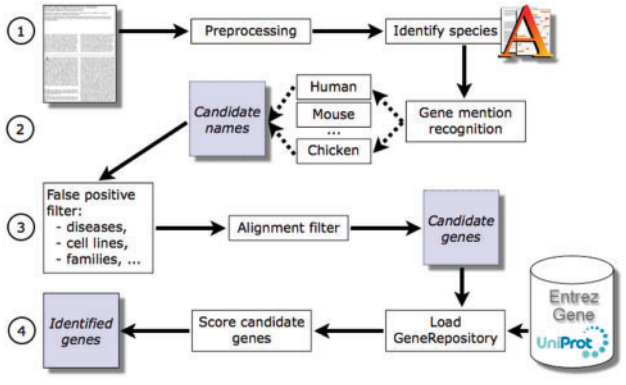

Fig. 2. Data processing in GNAT can be divided into four stages. (1) After pre-processing a text, we search for species using ALIBABA. (2) For each species that is referenced in the text, a dictionary is loaded and we annotate names of genes in the text. (3) Filters remove false positives, for example, names of gene families and diseases. (4) Remaining candidate genes are ranked using context profiles stored in a repository created from EntrezGene and UniProt annotations.

2.1 Named entity recognition for genes and species

The initial recognition of (potential) gene names was handled by two separate components. The first was based on a set of dictionaries (one for each species), the second used conditional random fields to learn a model on a training sample. For the dictionary-based component, we extracted all known names of each gene from EntrezGene and included names of the encoded protein(s), taken from UniProt/Swiss-Prot, where cross-references were available. We pre-processed each dictionary to filter out spurious names and names that are unlikely to lead to matches. Spurious names were names such as 'liver', 'fragment', or 'precursor'; examples for names that are unlikely to lead to matches are 'keratin, hair, acidic, 3A' and 'keratin 10 (epidermolytic hyperkeratosis; keratosis palmaris et plantaris)'. We removed single synonyms from each gene when they seemingly did not fit the others: considering the example of the human GPR12 gene, its known synonyms are GPR12, GPCR12, GPCR21, G protein-coupled receptor 12 and Probable G-protein coupled receptor 12. While four names refer to the family member 12, only one contains the number 21. Assuming that that name ended up in the database by mistake or is maintained for historical reasons, and that it will not be used, we removed it. Indeed, the name GPCR21 is only used twice in PubMed abstracts, both times referring to a murine gene. The two mentions occur in abstracts from 1993 and 1995, respectively, a possible extension

of this strategy thus could use validity checks including publication dates (of articles and genes/proteins). Tamames and Valencia (2006) showed an interesting analysis of gene name usage over time.

We expanded each name to a regular expression that covered likely spelling variations, using heuristics based on observations on training data (Hakenberg, 2007). To do so, we first sorted each name into either of three categories: database identifier ('KIAA0958'), abbreviation ('Ifn-g'), and compound name ('tumor necrosis factor alpha'). Variations of database identifiers allowed only for different patterns of capitalization ('Kiaa0958'). We segmented each abbreviation and each token of a compound name at the following positions: changes in capitalization, changes from letters to numbers or vice versa, symbols. For example, 'CD95R' and 'Hfn–3beta' would get segmented in the following ways:

- 'CD95R' → 'CD', '95' and 'R'.
- 'Hfn–3beta' → 'Hfn', '3' and 'beta'.

A set of heuristics then transformed every segment into a regular expression; these were combined into one expression for the whole name, allowing for different gaps between segments (whitespace, dash, no gap). For the given examples, this would lead to

- CD95R → (CD|Cd|cd)[–]?95[–]?(R|r|)(ceptor)?
- Hfn-3beta → (HFN|Hfn|hfn|HfN)[–]?(3|III|iii)[–]?(B|b|)(eta)?

Depending on the origin of the gene, i.e. its species, an abbreviation can also be preceded by the first letter of the species' name, for instance, 'hCHST6' when the original name of a human gene was 'CHST6'. For each species-specific dictionary, we combined all regular expression into a finite state automaton, using the BRICS library.[1] Matching the regular expressions then became linear in the length of the text to match against. In addition, each end state in the automaton stored the corresponding identifiers of all names that potentially end at this state. Thus, each match in a text immediately provided us with the set of candidate IDs. A detailed description of the dictionary-based method can be found in Hakenberg *et al.* (2008).

Recognition of species was handled by AliBaba[2] (Plake *et al.*, 2006), using all names of species from NCBI Taxonomy data. We constructed regular expressions from each name, using heuristic rules suited for names of species (e.g. dealing with Latin/Greek suffixes). Thus, the name 'Drosophila melanogaster' would result in an expression like

- [Dd]([.]|rosophilae?)[-_]*[Mm]el([.]|anogaster)?

to also match 'D. melanogaster' and 'D. mel.' Often, the species is not mentioned literally in a text; thus, we included lists of cell lines commonly used in laboratories (see Supplementary Material). A mention of 'HEK–293T' or 'C3H–10T1' would thus trigger the annotation of human or mouse, respectively.

In addition to the dictionary-based NER system, we also employed BANNER,[3] an open-source, portable NER system for biomedical text (Leaman and Gonzalez, 2008). BANNER uses the MALLET[4] implementation of second-order CRFs, with a feature set consisting primarily of orthography, morphology and shallow syntax. BANNER was extended with an abbreviation resolution (Schwartz and Hearst, 2003) and a dictionary-lookup feature. For the latter, we collected names from EntrezGene, UniProt, HUGO and mentions from the BioCreative 2 GN training set. The dictionary consisted of all *single tokens* of each full gene name collected; all tokens which occurred more frequently outside a gene name than as part of a gene name were pruned (for instance, the token 'gene'). Starting with 905 914 gene names, we ended up with 344 074 different tokens (after pruning 1675 tokens). We implemented the lookup as a binary feature, stating whether or not a token in a sentence matched (case-insensitive, normalized symbols). Mentions found using BANNER were matched to their corresponding gene by TF*IDF string similarity, using the SecondString software package.[5] The n genes with the highest TF*IDF score, where n is parameterizable to tune towards precision or recall, were then designated the gene candidates for the mention. We combined dictionary-based NER with BANNER by taking the union of all predictions. When predicted mentions were identical or overlapped, we kept only the mention and candidate IDs assigned by the dictionary. The reason for this was that we trained BANNER on BioCreative 2 GM data, where the task was slightly different (it included, for instance, gene families or species' names at the beginning of a gene name).

Another important step was the pre-processing of abstracts, to find locally resolved abbreviations and conjunctions. Baumgartner *et al.* (2007) note that 8% of all gene names in the BioCreative 2 data were part of some form of conjunctions or ranges. For instance, the text 'freac–1 to freac–7' includes references to five genes that are not mentioned literally. Conjunctions of compound names proved more problematic to handle, and we dealt only with some variations. For instance 'IL-7 and IL-15 receptors' should be expanded to 'IL–7 receptor and IL–15 receptor'. Otherwise, 'IL-7' potentially is recognized as a gene name on its own and mapped to a wrong ID.

2.2 Validation of gene mentions

In this step, we reconsidered each of the previously recognized gene names to check whether this name indeed referred to a gene (or protein). Two main reasons necessitate this step:

(1) Gene names are not only ambiguous within the class genes. They are also similar or even identical to names of diseases ('Neurofibromatosis 2', 'myotonic dystrophy'), cells or cell lines ('CD34', 'TRPV6–expressing cells'), common English words ('white', 'hairy'), etc. Ambiguity is especially high for abbreviations ('CRF', 'PCR').

(2) We are interested in references to specific genes, not entire gene families or classes of functionally related genes ('vesickle-trafficking protein', 'G-protein-coupled receptors'), which are often found by the initial named entity recognition.

We handled the first set of problematic cases by considering the immediate context (words to the left and right) of each gene name. Several dedicated filters checked for references to cells and cultures ('CD34+ cells' and 'EPC culture', where 'CD34' and 'EPC' were initially recognized as gene names). We compared the annotation of an abbreviation to the one of its long form; often, an ambiguous abbreviation was marked as a gene name, but the long form clearly referred to something else ['vitelliform macular dystrophy (VMD2)']. Some of these cases could be resolved globally (on the entire abstract), for instance, abbreviations, so that every occurrence of an abbreviation was filtered out if there was evidence for a false positive somewhere. Other cases were resolved locally (single occurrence). From a training sample (BioCreative 1 and 2 data), we counted the frequencies of tokens in a window of size two around true positive and false positive gene names. From that, we calculated likelihoods and log-likelihood ratios to estimate whether a new name was more likely a false positive.

The second problem, unspecific mentions of genes that refer, for instance, to gene families, could be solved using the name itself. We compiled a list of tokens that, in any combination, always refer to an unspecific name (not a gene name at all, only a gene family, etc.; see Supplementary Material). A similar list for single words can be obtained from Morgan *et al.* (2008). We also check names for identity to amino acids, species, tissues, and diseases.

We also decided on a heuristic to filter names that are ambiguous with common English words and biomedical terms. Gene names such as 'abnormal', 'embryonic gonad', and 'alpha-like' were filtered out when no second, unambiguous synonym ('abn', 'EGON', 'dALS') occurred in the

[1] BRICS — `http://www.brics.dk/automaton/`
[2] AliBaba — `http://alibaba.informatik.hu-berlin.de`
[3] BANNER — `http://banner.sourceforge.net`
[4] MALLET — `http://mallet.cs.umass.edu`
[5] SecondString — `http://secondstring.sourceforge.net`

same abstract. An example for such a co-occurrence of an ambiguous mention with a specific one can be found in PubMed abstract 2555153:

'… we termed this gene 'embryonic gonad' (egon),'

but such mentions do not necessarily have to be adjacent.

As our initial dictionary-based NER allowed for inexact matches, and the CRF-based NER found names that do not occur literally in the dictionary, we computed the similarity of each predicted name with all synonyms of all candidate genes assigned to it. Using alignment, this allowed us to introduce a threshold for names that were too far from any of the known synonyms.

2.3 Correlating gene mentions with species

The example in Figure 3 shows some of the typical evidence we may exploit to map genes to species. We relied on a multi-stage procedure with descending reliability to assign species to genes. A gene and a species could occur in the

(1) same compound noun: 'murine Eif4g1',

(2) same phrase, including enumerations: 'rat and murine Eif4g1' or

(3) same sentence.

If they did not occur in a single sentence, we checked whether the

(4) previous sentence mentions a species,

(5) title of abstracts mentions a species,

(6) first sentence of abstract mentions a species,

(7) a species occurs anywhere in the abstract or

(8) a species was annotated as MeSH term.

If all steps failed, we checked the abstract for in general mentions of kingdoms, classes, etc. Whenever a general mention ('mammals', 'fly') occurred that could not be resolved within the abstract or MeSH terms (additional mention of a species), we used parameterizable definitions for mapping, e.g. 'mammals' to human, or 'fly' to *D. melanogaster*.

2.4 Gene mention disambiguation

GNAT predicts the identifier (in our case, an EntrezGene ID) for each gene name recognized in a text. In cases where multiple candidate IDs for one name exist, the system has to pick the ID for the gene that is most likely referred to in the text. We draw hints on which gene is potentially discussed from terminology appearing in the same text. Each gene, as it is annotated in databases for various aspects, has a set of terminology (single-term annotations, complex descriptions, etc.) that is specific to it. Such terminology refers to chromosomal locations, species, molecular functions of the genes' products, known mutations and single nucleotide polymorphisms, etc., which all help to identify a gene if they are mentioned together in a text. All in all, we collected the following information on each gene and the protein(s) the gene encodes (where available): summaries, GeneRIFs, chromosomal location (all from EntrezGene), diseases, functions, tissues, keywords, protein length and mass, mutations, domains (from UniProt), interaction partners and GO terms (from both sources).

For each candidate ID that was assigned to a gene mention and thus to a text, we tried to find all such information in the text and picked the ID with the highest likelihood. Complex descriptions (summaries, descriptions of diseases and functions) were compared to a text using the cosine similarity, after filtering ca. 200 stop words (most frequent English words plus additional words observed in training data that have no impact on disambiguation, like 'gene' or 'suggest').

To calculate the similarity based on GO terms, we searched for GO terms in the current abstract and compared them to the set of GO terms assigned to each gene candidate. For each potential tuple taken from the two sets (text and gene annotation), we calculated a distance of the terms in the ontology tree. These distances yielded a similarity measure for two terms, even if they did not belong to the same sub-branch or were immediate parents/children of each other. The distance took into account the shortest path via the lowest common ancestors, as well as the depth of this lowest common ancestor in the overall hierarchy [comparable to Resnik (1999); Schlicker *et al.* (2006)]. Distances were calculated within each of the three branches of GO, but not across. The distances for the closest terms from each set then defined a similarity between the gene and the text.

Each comparison of a gene's annotations to a text resulted in a set of scores. For each type of annotation (e.g. diseases), we set the highest score that was obtained by any of the candidate genes to one, normalizing all other scores with the resulting factor. The final score for each gene was then the sum of all its single scores. Thus, using the 13 types of annotations shown before, the highest score obtainable was 13, although we never found all 13 annotations for a single gene in one PubMed abstract; the largest number of annotations we found in one text was eight.

Table 1. Number of abstracts that mention a particular species (or refer to a kingdom/class in general), and number of genes per species contained in the test set

Species	Abs.	Genes	Species	Abs.	Genes
Mus musculus	45	81	*Schizosaccharomyces pombe*	4	7
Homo sapiens	43	75	*Xenopus laevis*	3	2
Saccharomyces cerevisiae	30	79	*Oryctolagus cuniculus*	2	2
Drosophila melanogaster	28	60	*Escherichia coli*	3	2
Rattus norvegicus	8	8	*Arabidopsis thaliana*	2	2
Eukaryotes	3	–	Mammals	1	–
Metazoans	2	–	Plants	1	–

Shown are only species mentioned in two or more abstracts (see text for others). Overall, there are 320 genes mentioned in 100 abstracts.

2.5 Data sets and evaluation

EntrezGene currently stores information on 3 446 768 genes for which a species is known (January 2008). We mapped genes to proteins in UniProt using the `gene2accession` file provided by EntrezGene and found 242 279 such mappings. `gene_info` contains information such as names of genes, synonyms, and taxonomy ID.[6]

We used five different data sets for the evaluation of our system; four of them were the test sets from the BioCreative 1 and 2 challenges (Hirschman *et al.*, 2005; Morgan *et al.*, 2008). BioCreative 1, task 1B, provided defined training and test sets for GN of fruit fly, yeast and murine genes (three separate sets). BioCreative 2, GN task, provided normalization data for human genes, also split into training and test data. All these data sets consist of a range of PubMed abstracts, where identifiers of occurring genes were annotated on the abstract level. That means, whenever a gene occurred multiple times in an abstract, perhaps with different names (for instance, an abbreviation and its long form), it is sufficient to recognize, identify, and report a single instance. Note that for each of these four benchmarks, only genes from one particular species were annotated. Genes from other species, although occurring in the same abstract(s), were not annotated in either benchmark.

To get to the fifth, inter-species benchmark, we took 25 abstracts from each of the four aforementioned BioCreative sets. These 100 abstracts were re-annotated (existing annotations copied) for all other species with the help of PubMed's Links facility and by two human annotators. In these abstracts, we found references to 34 different species, including indirect mentions ('patients', 'COS cells'). Table 1 shows the number of abstracts and gene mentions for each species. On average, there were 1.93 species mentioned

[6]EntrezGene data — `ftp://ftp.ncbi.nih.gov/gene/`

(1) *Human* corneal GlcNac 6–O–sulfotransferase and *mouse* intestinal GlcNac 6–O–sulfotransferase both produce keratan sulfate. (11278593)
(2) We have determined the complete cDNA coding sequences of both the *human* and the *mouse* isoforms of Ksp–cadherin. (9721215)
(3) Here we describe new *human* and *murine* semaphorin homologues. The respective genes were cloned and sequenced, and they were termed *H*–Sema–L and *M*–Sema–L. (9721204)
(4) prk mRNA expression is also detected at a low level in the megakaryocytic *cell line Dami*, *MO7e*, and three brain glioma cell lines. (8702627)

	Gene name	Species	EntrezGene	Evidence
(1)	corneal GlcNac 6–O–sulfotransferase	human	4166	same compound noun
(1)	intestinal GlcNac 6–O–sulfotransferase	mouse	56773	same compound noun
(2)	Ksp-cadherin	human	1014	same phrase
(2)	Ksp-cadherin	mouse	12556	same phrase
(3)	H-Sema-L	human	8482	preceding sentence and implicitly in name (H–)
(3)	M-Sema-L	mouse	20361	preceding sentence and implicitly in name (M–)
(4)	prk	human	1263	reference to cell lines (Dami/MO7e)

Fig. 3. Examples of references to species (printed in italics) as they typically appear in abstracts; PubMed IDs given in brackets.

per abstract. Not counting the four species from which the test set was derived (human, mouse, yeast, fly), there were 0.49 additional species per abstract. The annotations consist of 320 genes; in addition to the numbers shown in Table 1, the species *Caenorhabditis elegans, Trichoderma reesei, Tetra thermophila* all had one gene mention in the test set.

We could not find an ID for five gene mentions, but decided to keep these abstracts and annotations in the test set; some of the gene's products could be found in UniProt or model organism-specific databases (such as SGD), but without any reference to EntrezGene. We mapped some genes to a species that was not mentioned explicitly in the abstract, but only in the full-text. For instance, 'AIF' in PubMed 11259394, which refers to human and mouse, but the murine gene is not mentioned literally in the abstract. For some species mentioned in the abstract, EntrezGene contained that gene only for a strain of that species. An example for that is 'Nrap' in PubMed ID 2541176, where it is discussed as a gene from *Schizosaccharomyces pombe*, but EntrezGene contained this gene only for some specific strains, e.g. *S. pombe strain 972h–*.

We ran all experiments on a dual quad-core Linux server with 32GB RAM. During our experiments, we loaded only dictionaries for the 25 species that were likely to be needed (common model organisms) permanently into memory (roughly 10 GB). Software was written in Java and interacts with a MySQL database. Without any offline pre-processing of abstracts, our method is able to handle one abstract in roughly 10 seconds.

3 RESULTS

Table 2 shows the performance of our system on different benchmark sets. On the overall problem of ISGN, our method achieved an *F*-measure of 81.4% (90.8% precision at 73.8% recall). *F*-measures for single species vary between 75.3% (fly) and 89.6% (yeast), as evaluated on the respective BioCreative 1/2 test sets. The recall for the initial named entity recognition was 90.6% at 7.7% precision; this excluded any form of disambiguation. Filtering of the initial dictionaries (unlikely/spurious names) increased the *F*-measure by about 0.8%.

We initially evaluated the CRF-based recognition of named entities with BANNER on the BioCreative 2 GM test data; BANNER achieves an *F*-measure of 86.30% (89.20% precision at 83.59% recall), which is an increase of 2.0% precision and 0.8% recall (1.4% *F*-measure) to the version published before. The increase is due to adding a dictionary lookup (single tokens) as a feature to BANNER. Note that this constitutes a tuning of BANNER towards recognizing protein names; BANNER itself is sought to remain task-independent. The BioCreative 2 gene mention task data

Table 2. Performance of the overall system and for single species

Benchmark set	P	R	F	TP	FP	FN
Overall	90.8	73.8	81.4	236	24	84
Human, test set	90.1	81.1	85.4	637	70	148
Mouse, test set	91.6	72.6	81.0	355	36	149
Yeast, test set	94.9	84.8	89.6	520	28	93
Fly, test set	82.1	69.5	75.3	298	65	131

Species-specific benchmarks refer to BioCreative 1/2 data. Precision, recall, and F1-measure in percentage. Number of true positives, false positives and false negatives.

was not an exact fit for inter-species gene normalization since, while it does include all genes mentioned in the abstract, it also includes gene families and other mentions which cannot be uniquely identified (Wilbur *et al*., 2007). We thus devised two experiments to gain insight into the performance of BANNER on the GN and ISGN tasks, respectively. On the human GN data (BioCreative 2) and our inter-species data set, we calculated precision and recall over ranked lists of predicted candidate identifiers. These graphs are shown in Figure 4. The initial recall of BANNER reached 90% when the top two ranked IDs were submitted. However, the overlap with the dictionary-based tagger was not large, and a combination of both techniques improved the overall performance by only 1% in *F*-measure. The figure also points out the difference in ambiguity ratios for the inter- and single-species tasks. On the full lexicon from Entrez Gene, we calculated an average of 1.83 genes per name (case-insensitive, no symbols). The average number of species per name was 1.4. If we add names of the referenced proteins, we get an average of 2.13 genes and 2.23 species per name.

We analyzed the performance of our method to recognize names and references to species. We found six false positively recognized names: Hippocampus, Bia, Glycine, Human echovirus 1 (synonym: E-1), Pan and Cis, in all cases not referring to a species; these were due to spelling variations that allowed for a match with, for instance, '*cis*-regulatory element', 'bias' and 'E1'. Driven by observations on a training sample, our method tried to resolve some ambiguous names (cancer, codon, axis, thymus, bear, etc.), but did not include these six new names. For the task of recognizing species' names, there were no false negatives. For the task of finding the exact species (taxon ID), which is necessary for correct GN, there was one false

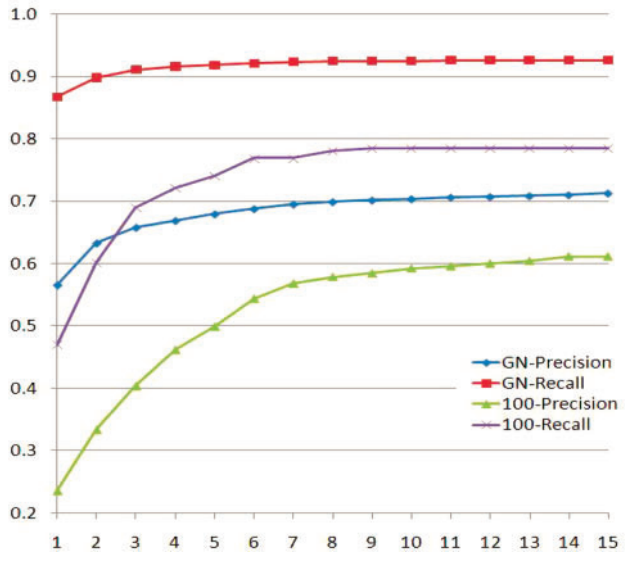

Fig. 4. Predictions of named entity recognition by BANNER and TD*IDF string similarity to assign candidate genes; evaluated on human GN test data and our inter-species set (100). A prediction at rank 3 (horizontal axis) would include the predictions from before, assuming a perfect resolution to species and disambiguation of genes.

negative; it directly led to false negatives in the ISGN performance. It was due to *M. musculus* not mentioned in either the abstract or MeSH annotations ('AIF' in PubMed 11259394). In this case, the information could have been picked up from the full text of the article.

3.1 Error analysis

An analysis of all errors (false positive and false negative predictions) that occurred on the test set is shown in Table 3. Our method failed to report 84 gene identifiers and predicted 24 false positive identifiers. We assigned each error to a category that best explains the cause of this mistake, to better understand the shortcomings of our approach and assess which problems to address first in the future. Most errors resulted from too restrictive filtering of names (stopwords, names that look like diseases) and by their context (cell lines, token frequencies). Ten missed identifiers resulted from missing rules to generate spelling variants ('Humly9' for human 'LY9'), and 16 from *ad hoc* creations of names with no similar enough entry in any database used. Twenty-four percent of the errors (12 false negatives and 14 false positives; some implicating the other) were due to an erroneous mapping of genes to species.

4 DISCUSSION AND CONCLUSION

We present here the first approach reported to handle ISGN, called GNAT. We use extensive background knowledge on genes and their products to resolve ambiguous names. A variety of data on genes, like their chromosomal location, molecular function, implications in diseases, help distinguish between genes sharing similar names. Assigning the correct species to a gene proved comparatively simple, when both gene and species are mentioned in the abstract of a publication. We consider MeSH annotations to find species when none are discussed in the abstract.

Table 3. Categories of false negatives (top) and false positives (bottom), with number of occurrences

Category	Number
Too restrictive filtering by context	24
No similar synonym known	16
Too dissimilar synonym	10
Too restrictive filtering of stopwords	6
No species/wrong species found in abstract	4
Unspecific species found in abstract	5
No assignment to species	3
Too restrictive filtering of names	2
Too many IDs left after disambiguation	3
Missed conjunction	1
Abbreviation resolution failed	1
Miscellaneous cases	10
Additional/wrong species assigned to gene	14
Disambiguation assigned wrong ID	4
System found a too general mention	1
Closer synonym for wrong gene	3
Name does not refer to a gene	2

Extensions to a previously reported system include covering all genes in EntrezGene, independent of species; adding machine learning to improve named entity recognition; mapping of genes to species based on evidence from the abstract; and a test set for ISGN.

We evaluate the performance of GNAT on a data set derived from the BioCreative 1 and 2 challenges. At an *F*-measure of 81.4%, the systems performs comparable to other systems that have been proposed for single-species normalization (see Related work section). Our methods yields a precision of 90.8% at a recall of 73.8%. On human genes alone, we achieve an *F*-measure of 85.4%, outperforming systems presented at BioCreative 2 by 4.3%.

Next steps include work on the initial named entity recognition. Our system misses 9.4% of the genes names entirely; almost 20% get lost during assignment of species, filtering of false positives and disambiguation. We attribute many of the initial misses to characteristics in naming conventions our method for generating dictionaries does not cover so far, since it was initially designed to handle human gene names.

Another way of relating genes to species is to use information available from parsing sentences, for instance, links derived from LinkGrammar (Grinberg *et al.*, 1995). We developed a database schema and query language, PTQL, to store and query such links. This will allow us to post queries like 'Does the gene named XYZ somewhere occur in the same noun phrase as the species ABC'? or, more to the point, 'Does the gene identified by the ID 123 somewhere occur together with the species 456 in the {same compound noun | same noun phrase | ... | title of the article | ... }'? to efficiently analyze the current abstract for any such occurrence (in descending order of reliability, see list in Section 2.3).

A web interface and Supplementary Material are available (see below); the tool will also be available as a web service for the BioCreative MetaServer (BCMS). BCMS combines efforts presented at the 2nd BioCreative Workshop, which deal with recognition and normalization of genes/proteins and species, as well as predictions of protein–protein interactions, into a single framework (Leitner *et al.*, 2008). We restrict both online

processing tools to genes from eleven different species (see Supplementary Material), because of hardware limitations (the full set of genes currently would contain ca. 3.5 million genes from 4700 species/strains). We are currently studying methods for efficient (time and memory) ways for named entity recognition to include this full set of genes; one way would be to rely on the machine learning-based NER only, which gives slightly inferior results, but does not require to load large dictionaries into memory.

We are currently extending the test set to at least 300 abstracts, which would be about the same size as the original BioCreative benchmarks. In addition, we want to annotate full-text articles for UniProt IDs, as information on the exact species (for example, on strains) is not always included in the abstract. Better evidence for assigning genes to species often is present in the full text. Data published with BioCreative 2 (IPS task) can be taken as a starting point for annotating all proteins and all species.

4.1 Related work

There are some systems which use components similar to ours. Fundel and Zimmer (2007) use a large dictionary (consisting of human gene names taken from EntrezGene, SwissProt and HUGO) and combine dictionary matching with the ProMiner tool (Fluck *et al*., 2007; Hanisch *et al*., 2005). Some filtering removes, for instance, gene names from other organisms, names of cell lines that resemble gene names, diseases named after a gene. Normalization is achieved by computing the cosine similarity of the given abstract (represented by all noun phrases) and known synonyms of each gene candidate. The system achieves an *F*-measure of 80.4% on the BioCreative 2 GN test set for human genes. ProMiner (Fluck *et al*., 2007) relies on a large dictionary that is augmented with a manually curated list and contains automatically generated spelling variants of gene names ('IL1' for 'IL-1', etc.) Approximate string matching searches for the dictionary entries in text. Disambiguation in ProMiner is based on finding another known synonym of the ambiguous gene in the same text. ProMiner achieves an *F*-measure of 79.9% on the BioCreative 2 GN test data. Both approaches rely on solving the problem of disambiguation using a dictionary of gene names and known or automatically added synonyms and concepts.

Xu *et al*. (2007) use information from publications that were cross-linked from gene entries in EntrezGene. For each gene, the system has a set of PubMed abstracts that discuss this particular gene. In these abstracts, the system searches for UMLS terms and words that are not common English; it also uses MeSH terms assigned to each abstract. All these data were taken to describe each gene in a profile. The disambiguation task then was to find the profile among the candidate genes that best matches the current text (for instance, identical MeSH terms) using cosine similarity. Xu *et al*. showed that this approach yields a precision of 92.2% on a subset of BioCreative 2 GN test data (124 abstracts out of 262). For their experiments, Xu *et al*. assumed a perfect recognition of named entities (no missed mentions, no non-gene names). Farkas (2008) studied the use of information on authors for the task of gene mention disambiguation. They built the inverse co-authorship graph (nodes: abstracts; edges: shared author) potentially using all PubMed citations. Many of these citations occurred as references for annotations in various databases, including EntrezGene, and thus could be mapped directly to one or more genes. For every test instance, the method obtained the citation that could be reached using the shortest path and contained the same gene name. The annotation for the gene under consideration was then taken as prediction for the test case. For path lengths of at most one and two, the method yielded 95–100% precision and dropped significantly with longer paths; coverage was relatively low (<50%) at a path length of one. The test scenario was the same as in Xu *et al*.

ACKNOWLEDGEMENT

We thank the anonymous reviewers for some valuable suggestions.

Funding: This work was partially supported by NSF CISE grant 041200, and by the Science Foundation Arizona 2007 Competitive Advantage Award. We kindly acknowledge partial support by the European Commission 6th Framework Programme, project IST-2006-027269.

Conflict of Interest: none declared.

REFERENCES

Baumgartner,W. *et al.* (2007) An integrated approach to concept recognition in biomedical text. In *Proceedings of Second BioCreative Workshop*, Fundación CNIO Carlos III, Madrid, Spain, pp. 257–271.

Farkas,R. (2008) The strength of co-authorship in gene name disambiguation. *BMC Bioinformatics*, **9**, 69.

Fluck,J. *et al.* (2007) ProMiner: recognition of human gene and protein names using regularly updated dictionaries. In *Proceedings of Second BioCreative Challenge Evaluation Workshop*, Fundación CNIO Carlos III, Madrid, Spain, pp. 149–151.

Fundel,K. and Zimmer,R. (2007) Human gene normalization by an integrated approach including abbreviation resolution and disambiguation. In *Proceedings of Second BioCreative Challenge Evaluation Workshop*, Fundación CNIO Carlos III, Madrid, Spain, pp. 153–155.

Gene Ontology Consortium. (2008) The Gene Ontology project in 2008. *Nucleic Acids Res.*, **36**, D440–D444.

Grinberg,D. *et al.* (1995) A Robust Parsing Algorithm for Link Grammars. In *Proceedings of International Workshop on Parsing Technologies*, ACL, Prague, Czech Republic, pp. 111–125.

Hakenberg,J. (2007) What's in a gene name? Automated refinement of gene name dictionaries. In *Proceedings of BioNLP at ACL 2007*, ACL, Prague, Czech Republic, pp. 153–160.

Hakenberg,J. *et al.* (2008) Gene mention normalization and interaction extraction with context models and sentence motifs. *Genome Biol.*, **9**, S14.

Hanisch,D. *et al.* (2005) ProMiner: rule-based protein and gene entity recognition. *BMC Bioinformatics*, **6 (Suppl 1)**, S14.

Hirschman,L. *et al.* (2005) Overview of BioCreAtIvE task 1B: normalized gene lists. *BMC Bioinformatics*, **6 (Suppl 1)**, S11.

Leaman,B. and Gonzalez,G. (2008) BANNER: an executable survey of advances in biomedical named entity recognition. *Pac. Symp. Biocomput.*, **13**, 652–663.

Leitner,F. *et al.* (2008) Introducing meta-services for biomedical information extraction. *Genome Biol.*, **9** (S2), S6.

Morgan,A. *et al.* (2008) Overview of BioCreative II Gene Normalization. *Genome Biol.*, **9** (S2), S3.

Plake,C. *et al.* (2006) AliBaba: PubMed as a graph. *Bioinformatics*, **22**, 2444–2445.

Resnik,P. (1999) Semantic similarity in a taxonomy: an information based measure and its application to problems of ambiguity in natural language. *J. Artif. Intell. Res.*, **11**, 95–130.

Schlicker,A. *et al.* (2006) A new measure for functional similarity of gene products based on Gene Ontology. *BMC Bioinformatics*, **7**, 302.

Schwartz,A.S. and Hearst,M.A. (2003) A simple algorithm for identifying abbreviation definitions in biomedical text. In *Proceedings of Pacific Symposium on Biocomput*, Hawaii, USA, pp. 451–462.

Tamames,J. and Valencia,A. (2006) The success (or not) of HUGO nomenclature. *Genome Biol.*, **7**, 402.

Xu,H. *et al.* (2007) Combining multiple evidence for gene symbol disambiguation. In *Proceedings of BioNLP at ACL 2007*, ACL, Prague, Czech Republic, pp. 41–48.

Wilbur,J. *et al.* (2007) BioCreative 2. Gene mention task. In *Proceedings of Second BioCreative Challenge Evaluation Workshop*, Fundación CNIO Carlos III, Madrid, Spain, pp. 7–16.

Zweigenbaum,P. *et al.* (2007) Frontiers of biomedical text mining: current progress. *Brief Bioinformatics*, **8**, 358–375.