

# Me and my friends: gene mention normalization with background knowledge

Jörg Hakenberg<sup>1</sup>

hakenbergj@biotec.tu-dresden.de

Loic Royer<sup>1</sup>

loic.royer@biotec.tu-dresden.de

Conrad Plake<sup>2,1</sup>

cplake@transinsight.com

Hendrik Strobel<sup>1</sup>

hendriks@biotec.tu-dresden.de

Michael Schroeder<sup>1</sup>

ms@biotec.tu-dresden.de

<sup>1</sup> Biotechnological Centre, Technische Universität Dresden, Tatzberg 47–51, 01307 Dresden, Germany

<sup>2</sup> Transinsight GmbH, Tatzberg 47–51, 01307 Dresden, Germany

“Tell me who your friends are, and I will tell you who you are” – this proverb best illustrates our approach to the normalization of gene names. In this approach, we rely on background knowledge that describes various aspects of a gene: it is localized on a chromosomal band, it belongs to an operon structure, it is a member of a gene family, its products take part in biological processes, they fulfil molecular functions, they occur at dedicated cellular locations, mutations of the gene ultimately cause diseases, its proteins contain domains and form secondary, tertiary and quaternary structures. Whenever a gene (or one of its products) is discussed, some of these aspects –the gene’s friends, that is, semantically related information– will be mentioned as well. The paradigm we follow with this approach demands not only the presence of a gene’s name, but also of some of its friends.

We see every set of information available (see Methods section) for each gene as this gene’s description, or the context this gene typically “lives” in. Whenever we encounter ambiguities regarding proper identification of a gene, we assess each potential candidate gene by comparing its typical context against the predicted one (in this case, a PubMed abstract.) The descriptions for genes originate from various curated resources: EntrezGene provides organisms, summaries, chromosomal loci, Gene Ontology (GO) terms, and encoded proteins; UniProt provides functional descriptions, protein domains, interaction partners, keywords, and GO terms; more GO annotations are provided by GOA.

Consider the example of the oncogene p54 (reflected in Figure 1.) Having resolved the issue of potential organisms, there are still human genes from EntrezGene that share the same name. Indeed, they refer to completely different genes with disjoint annotations. Based on the name alone, this problem could not be solved. Only comparing each of the gene’s contexts to the text reveals that one of the potential candidates is a RNA helicase, and the text indeed mentions “RNA helicase.” The text also mentions the exact chromosomal location of the correct gene.

A gene encoding a putative **human RNA helicase, p54**, has been cloned and mapped to the band **q23.3 of chromosome 11**. The predicted amino acid sequence shares a striking homology (75% identical) with the female germline-specific RNA helicase ME31B gene of *Drosophila*. Unlike ME31B, however, the new gene expresses an abundant transcript in a large number of adult tissues and its 5' non-coding region was found split in a **t(11;14)(q23.3;q32.3) cell line** from a diffuse large **B-cell lymphoma**.

EntrezGene ID: <b>1656</b> P54; RCK; HLR2 Species: <b>H. sapiens</b> Chromosome: <b>11q23.3</b> GO: <b>RNA Helicase</b>	EntrezGene ID: 2289 P54; FKBP51; PPlase Species: <b>H. sapiens</b> Chromosome: 6p21.3-2 GO: isomerase activity	EntrezGene ID: 4841 P54; NMT55; NRB54 Species: <b>H. sapiens</b> Chromosome: Xq13.1 GO: RNA splicing	EntrezGene ID: 42828 S4; dRpt2; <b>p54</b> ; p56 Species: <b>D. melanogaster</b> Chromosome: 3R;95C13 GO: proteolysis
---	--	--	---

Figure 1: The given abstract as a whole points out one out of four genes: synonym, species, chromosomal location, biological process all fit best to the leftmost gene context (PubMed ID: 1579499).

## Methods

The system we propose for identification of gene names in texts consists of four major components. The basic step provides an initial recognition of candidate terms, which also assigns all potential EntrezGene IDs to each candidate. From there on, the next components deal with refining these candidate hits: removal of false positives and disambiguation of polysemous names. The second component finds text parts that never contain a gene name but might account for errors of the recognition step. The third component filters false positives by looking at term frequencies, and reduces the candidate IDs by comparing new to known texts (from the “noisy” training data.) The final component disambiguates remaining terms and identifiers using each gene’s typical context. On the BioCreative2 GN test set, our system achieves an F1-measure of 81% (highest recall: 87.5%, highest precision: 79%.) The highest recall we measured on the training data set was 92.7% (at <40% precision); this was achieved when not using the disambiguation.

Description of the submitted run	Precision	Recall	F1 (in %)	TP	FP	FN
NER with extended masterlist, FP+FN filter, disambiguation	78.9	83.3	81.0	654	175	131
NER with extended masterlist, FP filter, no disambiguation	49.6	87.5	63.3	687	699	98
NER with unextended masterlist, FP filter, disambiguation	70.7	72.5	71.6	569	236	216

### Named entity recognition

For the initial recognition of potential gene names and their EntrezGene identifiers, we extended the provided masterlist with additional synonyms found on the EntrezGene website, plus synonyms for the gene products. We then sorted each synonym into one of four categories:

- database identifiers (“KIAA0958”, “HGNC:17875”),
- abbreviations (“CD95L”, “Lin7c”),
- single- or multi-word terms (“tumor necrosis factor alpha”, “RAD51-interacting protein”), and
- spurious synonyms (“AA”, “ORF has no N-terminal ‘Met’,it may be non-functional”).

We ignored spurious synonyms in the remainder, as they never occur in text, but only in database fields. For each synonym class we applied specialized search strategies.

- Database identifiers were extracted using regular expressions, yielding immediate identification: “KIAA0958” could appear as “Kiaa0958”, yet it was unique and pointed to a single ID.
- Abbreviations got segmented around optical gaps: white spaces, punctuation, transitions between digits, lower case or upper case letters. We generated variations for each segment and re-combined them. Variations affected case changes, transformations between Latin/Arabic/Greek/English, and structural changes (“CD95 receptor”, “receptor of CD95.”) Starting with the known synonym “IFN-gamma”, the mentioning “Ifng” has to be recognized.
- Multi-word terms were tokenized and each token was evaluated for potential variations, comparable to the abbreviation class. This added possible spelling variants of each synonym. Some tokens were optional and not essential for recognition (“protein” at the end of a name), because they often are omitted in text.

Each synonym could correspond to several different genes and thus different identifiers. To remove obvious false positives, we used a filtering algorithm based on contextual rules. Each rule was a triplet consisting of three regular expressions, the first matching the context immediately before a potential gene name, the second matching the name itself, and the third matching the context right after the name. For example, an initial candidate name immediately followed by “cells” most likely referred to a cell line, and only implicitly to a gene/protein. “Mouse” before a name hinted to a mouse gene, but if the name was then followed by “homolog”, this rule did not apply immediately. We created these rules manually driven by examples from the training data.

The last step of this initial recognition merged consecutive candidate names (that shared one identifier) into one contiguous candidate. Such occurrences were most likely to refer to one and the same gene. Such tuples appeared, for instance, when abbreviations were introduced and a long form was followed by its abbreviation in brackets. We kept only such EntrezGene IDs that were assigned to all consecutive candidates; we kept the IDs of the long form when there were no IDs common to

all, dropping all other IDs. In addition, we expanded ranges (such as in “seven novel forkhead genes, freac-1 to freac-7”) to the full list of all names included therein.

### Disguising false positive sites

The second component of our system marked obvious irrelevant parts that often accounted for false positives. It removed the following types of phrases prior to NER: units like “497 amino acids”; cell types and descriptions (“CD34+”); DNA/RNA (“ACGGT”, “cDNA”); chromosomal locations (“chromosome 20 on band p13”, “21q22.1”); and abbreviations not related to genes/proteins (“Human granulocytic ehrlichiosis (HGE)”). This avoided some errors introduced by the first component, for instance, the detection of “p13” in the chromosomal location example. As another filter, we removed unspecific references (to protein families etc.) from the predicted candidate names. We noticed that in most cases, (even multi-word) names that consisted entirely of lower case letters could also be removed.

### Identification of candidates

After the initial recognition of gene names, we proceeded to identify each name. We passed the annotated texts through several filters to reduce the number of possible IDs for each gene name and to find the correct masterlist entry: We first searched for exact matches of candidate names in the masterlist. In case only one entry was found, we took this entry directly as the annotation. For ambiguous cases (multiple entries for the name), we compiled a set of representative texts for each entry from the noisy data and EntrezGene Summary. From 8243 ambiguous entries, 2954 had abstracts in the noisy training data (see GN task description), 3906 had an EntrezGene Summary, and 2074 had both. Every set of texts was transformed into a set of feature vectors with tf-idf feature weights. We then searched for the 100 abstract most similar to the current abstract (cosine-based distance.) From the set of IDs resulting from this comparison (each of the 100 representatives had one or more genes assigned), we selected the subset of IDs that had synonyms matching the candidate gene name. For matching, we used an approximative, character-based alignment. All IDs from this subset were taken into further consideration. To all remaining gene names we assigned a tf-idf score based on the current abstract and the overall text corpus. If a candidate name achieved a low tf-idf score, we dropped it as a likely false positive annotation. This step thus dealt with two types of errors introduced by the named entity recognition: it removed false annotations and it found genes initially missed.

### Disambiguation by candidate ranking

The fourth component disambiguated each polysemous name. We compared background knowledge available for each gene (gene context) with the current text and picked the gene which context best fitted the current text. We collected external knowledge from EntrezGene, UniProt, and GOA for each of the 30,000 genes (EntrezGene: summary, GO terms; UniProt: diseases, keywords, functions, GO terms; GOA: GO terms.) For EntrezGene and UniProt, we calculated the overlap of the text at hand with each annotation based on tokens. For calculating the similarity based on GO terms, we used GoPubMed to find GO terms in the current text [1]. For each potential tuple taken from the two sets (text & gene annotation), we computed a distance of the terms in the ontology tree (comparable to [2]). 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 LCA in the overall hierarchy. All five comparisons yielded likelihoods stating the similarity of the current text with the knowledge available on each gene. We combined the likelihoods into confidence measures, and picked the EntrezGene ID with the highest probability, if this was above a certain threshold.

## References

- [1] Doms, A. and Schroeder, M., GoPubMed: exploring PubMed with the Gene Ontology. *Nucleic Acids Res.*, 33:W783–786, 2005.

- [2] Schlicker, A., Domingues, F.S., Rahnenführer, J., and Lengauer, T., A new measure for functional similarity of gene products based on Gene Ontology, *BMC Bioinformatics*, 7:302, 2006.