TX Task:
Automatic Detection of Focus Organisms in Biomedical Publications

Thomas Kappeler, Kaarel Kaljurand, Fabio Rinaldi*
Institute of Computational Linguistics, University of Zurich
kappeler@bluewin.ch, kalju@cl.uzh.ch, rinaldi@cl.uzh.ch

Abstract

In biomedical information extraction (IE), a central problem is the disambiguation of ambiguous names for domain specific entities, such as proteins, genes, etc. One important dimension of ambiguity is the organism to which the entities belong: in order to disambiguate an ambiguous entity name (e.g. a protein), it is often necessary to identify the specific organism to which it refers.

In this paper we present an approach to the detection and disambiguation of the focus organism(s), i.e. the organism(s) which are the subject of the research described in scientific papers, which can then be used for the disambiguation of other entities.

The results are evaluated against a gold standard derived from IntAct annotations. The evaluation suggests that the results may already be useful within a curation environment and are certainly a baseline for more complex approaches.

1 Introduction

The task of identifying the organisms which are involved in research described in biomedical articles is extremely important for the field of biomedical information extraction (IE), both in itself and in connection with other tasks. In itself, because the concept of biological taxonomy is basic for every researcher: organisms and their taxonomic classification can be used very effectively in various contexts, for example to restrict searches, a classical information retrieval (IR) task. At the same time, any biomedical text mining system would be incomplete without the possibility to use organisms as concepts, e.g. in finding (statistical) associations, which can then be used to form hypotheses about causal relations.

The necessity of identifying organisms is even more evident as part of other important entity recognition tasks in biomedical information extraction (IE), e.g. identification and disambiguation of proteins mentioned in the literature. For example, within the PPI task (identification of protein-protein interactions) of Biocreative II (Krallinger et al., 2008), the identification of the focus organism was seen by many participants as an essential subtask in order to properly disambiguate protein names. Protein interactions are fundamental for most biological processes, therefore they are at the focus of a huge and fast growing number of biomedical papers. As these cannot all be read or even inspected by the researchers, databases such as IntAct (Kerrien et al., 2006) or MINT (Zanzoni et al., 2002) try to create a reliable catalogue of experimentally detected interactions by extracting them "manually" from the literature through the usage of human experts. This is known as "curation", a costly and time-consuming process, which could be speeded up much by efficient, robust and precise extraction tools.

One of the most important obstacles for efficient automatic identification of proteins is the extreme ambiguity of the commonly used protein names in the literature. The fragmentation of the biomedical scientific community into lots of extremely specialized sub-communities seems to be the main reason for this ambiguity. In most cases, the ambiguity is between homologous proteins of different species. Any human reader belonging to the sub-community concerned can, in general, disambiguate an ambiguous protein name like "goat" (which can refer to proteins found in four different organisms: human, rat, mouse and zebrafish), as the species is obvious to them from the context. However, this ambiguity

*Corresponding author

Proceedings of the Workshop on BioNLP, pages 80–88,
Boulder, Colorado, June 2009.

remains problematic for IE systems (and even for curators in some cases) and needs to be solved before more complex tasks, such as protein interaction detection, can be effectively tackled (Rinaldi et al., 2008).

Our goal is to be able to identify automatically the focus organisms, i.e. the organisms that are mentioned in the paper as the hosts of the experiments described, or as the sources of the entities involved. This information can then be used for tagging papers for more efficient organism-based information retrieval, or, more commonly, for the disambiguation of other entities mentioned in the same paper. Since organism recognition is normally performed with reference to a taxonomical organization (of Linnean origin) of all known organisms (in our case, the NCBI taxonomy) this task is often referred to as "TX task".

In the rest of this paper we describe in section 2 the resources used and the approach followed in order to extract and rank candidate organisms. In section 3 we present our results and propose a more fine grained interpretation of the task, which we again evaluate. Finally in section 4 we compare our approach to previous work and discuss its limitations.

2 Methods

Our approach can be described briefly as (1) find all explicit mentions of organisms either by their scientific or "common" names; (2) count these mentions and combine the resulting numbers with a simple use of statistics to arrive at a ranked list or a simple set of organisms which can be used, among other things, to disambiguate ambiguous protein names in the article under investigation.

2.1 Resources Used

The first step for this approach was to choose a widely accepted taxonomy which not just includes unambiguous identifiers for all known organisms, but also provides a sufficiently large list of names for them. The taxonomy selected for this was the NCBI Taxonomy[1].

As most of these organism are unlikely to ever occur in biomedical literature, we decided to restrict our interest to the organisms for which a UniProt organism mnemonic identifier exists. UniProt (UniProt Consortium, 2007) is a database containing detailed information about known proteins, obtained by a process of curation of the biomedical literature. For every protein, a "mnemonic" identifier is defined (e.g. HBA_HUMAN for "Human Hemoglobin A") which is composed by a shorthand for the protein name and a simple unique identifier for the organism. Within the UniProt entry for the protein, the organism is also referred to by its NCBI identifier, allowing the construction of a mapping from the mnemonic identifiers for the organisms used by UniProt to their equivalent NCBI identifiers.

The set of organism that have a UniProt mnemonic identifier (11,444 organisms) probably covers the near totality of organisms that have been subject to research in molecular biology. In the NCBI taxonomy 31,733 names are defined for that subset of organisms. Although several classes of names are defined by NCBI, for the purpose of this work we distinguish only between "scientific names" and the other classes (pooled together as "common names").[2]

As an additional source of information, we used the IntAct database of protein interactions[3] for two different purposes:

- to derive statistical measures used later by the program, most importantly the frequency of each focus organism in papers curated by IntAct (using the IntAct annotations as the sources of the 'focus').
- to derive a gold standard against which our programs could be tested

IntAct provides an annotated set of protein interactions. Each interaction is enriched with detailed information about the two proteins involved

[1] Available as archive taxdmp.zip from ftp://ftp.ncbi.nih.gov/pub/taxonomy/. We worked with a version downloaded on July 10th 2008. The file nodes.dmp contains the taxonomy as a set of 443,299 nodes for the taxa and immediate-dominance-relations between them. The file names.dmp connects one or several names (619,325) of different nameclasses (such as "scientific" or "common") to each node. The nodes (taxa) are referred to by numeric identifiers.

[2] While there are no ambiguous "scientific names" in this taxonomy, there are several ambiguous "common names", but only very few of these occurred in our sample, e.g. "mink", "barley", "green monkey", and they are very rare.

[3] Version of May 2008, downloaded from http://www.ebi.ac.uk/intact/site/contents/downloads.jsf

(from which the reference organisms can be recovered), and with the identifier of the paper from which the interaction was originally derived in the curation process. This allows to build a gold standard by associating each paper to its focus organisms.

The sample used in our experiments is a set of 621 PubMed-indexed full text articles, dating from 1995 to 2007, for which IntAct annotations are available.[4]

2.2 First Experiments and Normalization

As an initial experiment, we performed a simple lexical lookup of the names of the 11,444 organisms under consideration. In previous applications of IE techniques for biomedical literature (Kappeler et al., 2008; Rinaldi et al., 2008) we found that simple techniques for the generation of variants of the known names significantly benefited the recall of the application. For example, multiword protein names can be subject to a number of minor variants, such as the introduction of hyphens or the separation of compound words, which make automatic recognition more challenging. In the case of organism names, although our initial expectations were similar, we found the benefit (in terms of additional recall) of such variants to be extremely limited, possibly because names of species are used more consistently than the names of proteins or genes.

Therefore it was possible to implement a simpler approach to recognition of organism names, based on lexical lookup against a database containing all names of interest, coupled with a simple normalization step which removes trivial orthographic differences (such as hyphens) between the key word in the database and the lookup word from the document (for details see (Kaljurand et al., 2009)). The inclusion of other biomedical NE's (such as protein names, method names, cell line names) in the database together with a strict implementation of the "longest match" principle leads to better precision by eliminating false positives caused by matching organism names with a fragment of a multiword term for another entity (such as the method "yeast two-hybrid").

As mentioned, the names provided by the NCBI taxonomy have been classified into "scientific names" or "common names". Using only "scientific names" appeared as an effective way to obtain better precision, but we soon discovered that precision of the common names suffered most by a few very bad names, such as "Li", which is a "common name" for LIV (Louping ill virus) in the taxonomy, but appears only (and very frequently) as Chinese surname in the texts. By eliminating about 25 of similar misleading "common names" the results of this class rose to the same level as the "scientific names", so there was no reason to exclude the whole class (as that would have harmed recall).

Since the bibliography might contain spurious mentions of other organisms, we automatically removed it from the main text. However, contrary to expectations, this did not lead to better results for this task (at least after the elimination of the misleading "common names" mentioned above), but was not reversed because of its effects on other tasks. An intuition from other tasks was to use the abstracts instead of the full text of the articles, because that would tend to exclude accidental mentions of organisms leading to false positives. But a main problem of this approach is that many abstracts do not yield any organism mentions. Whenever they do though, their precision is high. So there is a strong case for giving the mentions there a higher weight, but obviously the rest of the article plays an important role as well. We experimentally found that counting an "abstract mention" as equivalent to 25 "fulltext mentions" worked best.

2.3 Measures Improving Recall

An experiment using all names provided by NCBI and considering all mentions of those names in the fulltext version of each article led to a recall of 83%, leading us to conclude that either the taxonomy does not contain all names used, or some organisms are suggested to the human reader by the context and/or his anticipations. The first of these problems was adressed by adding some generated names to the termbase, the second by the use of a default.

Several possible ways of generating new names automatically from the names in the database were considered, but only two were applied successfully, as described below. One of them was the automatic generation of additional names from the nameclass

[4]The reason of this particular choice is that the same subset was used for experiments related to the automatic detection of experimental methods, also using IntAct annotations as a gold standard, described in (Kappeler et al., 2008).

"scientific name" (for organisms of species or subspecies level) by the process of replacing the first word (which would be the genus name in the classical Linnean binomial nomenclature) by its first letter and a dot. The resulting names, such as "E. coli", are widely used, but not included in the taxonomy. A seemingly large disadvantage of this approach is its potential for ambiguity: 338 of the resulting names refer to more than one organism. But the test on our sample showed that of these only 4 occurred at all, only 1 more than once: "C. elegans" (potentially referring to the organisms identified in UniProt as CAEEL, CENEL, CESEL and CUNEL) which always stood for CAEEL, i.e. "Caenorhabditis elegans". So excluding the other options for "C. elegans" eliminated the ambiguity (at least in our sample). We observed that this type of name is in frequent use only for few species and in this case the unabbreviated name is often used first, so the addition of this generated nameclass added little to recall.

The other type of name missing from the taxonomy is the use of the (Linnean) genus name for a very frequent species, e.g. "Arabidopsis" used for "Arabidopsis thaliana". Experiments showed that this type could not be reliably generated automatically from the "scientific names", as this nameclass includes many names which do not follow the rules of Linnean binomial nomenclature, mostly virus names such as "Human papillomavirus type me180" where the first word is generally not a genus name, but a host name. So the problem of (potentially huge) ambiguity in this type of names was not even researched, instead the names of this type for the most frequent organisms were generated manually and those which improved the results were included into the termbase (*Saccharomyces, Arabidopsis, Drosophila, Escherichia, Xenopus* and *Synechocystis*). The addition of this generated nameclass did not add much to recall for the same reason as for the first group: in most cases the unabbreviated name appears in the paper as well. Together both groups improved recall by about 3.4%.

As HUMAN is the most frequent organism in this context, it was obvious that a default HUMAN would take care of many cases where human readers disambiguate ambiguous protein names even without any explicit mentions of this species. As there are no cases (with the current termbase and sample) of articles with no organism mentions in the full text, we chose to have a default triggered by no findings in the abstract. Experiments showed that — contrary to intuition — a weight of the default proportional to the total number of mentions (just adding a percentage to HUMAN) would lead to worse results than an absolute value for the default.[5]

Table 1: Most frequent organisms in IntAct (derived from interactor proteins and host organisms)

ORG	freq
HUMAN	0.281
YEAST	0.272
MOUSE	0.091
ARATH	0.056
CERAE	0.037
RAT	0.033
DROME	0.028
SCHPO	0.023
ECOLX	0.020
ECOLI	0.013

2.4 Measures Improving Precision

The simple approach of considering every mention of each organism (after excluding the misleading common names, as described above), leads to a precision of only 27.6%, therefore the list of organism identifiers obtained in this way has to be considered as a "candidates list" from which a selection has to be made.

Candidates can be of course ranked according to number of mentions in each article. A ranking based on the mention counts, taking into account the correction factor of 25 for mentions in the abstract (as described in section 2.2), was still far from optimal, so we multiplied the mentions with the relative frequencies of the organisms in a micro-averaged frequency table (table 1) computed over all of IntAct (not just our sample, to avoid overfitting) and smoothed roughly by attributing 1% of the probability mass to all unseen organisms (over 11,000). This ranking did far better than expected and after nor-

[5] A tentative explanation: In a small paper, the effect of accidental mentions of "wrong" organisms is much larger than in big papers (where the important organisms are mentioned again and again). This detrimental effect may be counterbalanced by a relatively stronger default.

malizing the whole list to 1, a minimal threshold for the score could be set up to maximize the f-score by improving precision at the cost of recall. The actual value of the threshold (currently 0.04) is of course arbitrary, depending on what measure one wants to maximize.

Another problem to be tackled is that different papers will have different numbers of focus organisms, ranging from one (in about 70% of the cases), to several hundreds (in a few very infrequent cases). It could be assumed that being able to correctly guess the number of focus organisms would lead to improvement in the TX task, as we could pick only as many candidate organisms (in their ranking order) as the expected number for the paper. However, an experiment using the gold standard as an oracle to predict the number of organisms to be returned as a result, instead of using a threshold in the ranking, did not perform much better (recall was about 1.7% higher), so we decided not to spend any energy on exploring ways to predict the number of organisms as the effect would be minimal, even with perfect prediction.

Further experiments, such as giving different weights to mentions of names of different name-classes, did not lead to better results. Including information about the precision or recall of the names encountered in our test set (or the organisms predicted by them) in the formula for the weights[6] did not lead to better results either.

3 Evaluation and analysis of results

So finally the program in its current form considers all organism mentions, as delivered by the termbase search, eliminates the problematic common names, counts the mentions for each organism in fulltext and abstracts, multiplies the latter by 25 and adds them to the fulltext mentions. In case of no abstract mentions, a default of 28 fulltext mentions is added to HUMAN (equivalent to about one abstract mention).

The result for each organism is multiplied by the relative frequency of the organism in IntAct and divided by the sum of the results over all organisms to normalize the sum of the values to 1 (100%). All organisms under the threshold of 0.04 (or 4%) are then eliminated from the list.

[6] An idea suggested by its successful use in the detection of experimental methods in (Kappeler et al., 2008) and (Rinaldi et al., 2008).

Table 2: Most frequent false positives for the best results with our sample

ORG	freq
HUMAN	121
YEAST	104
MOUSE	68
ECOLX	18
DROME	13
ARATH	11
RAT	9

Table 3: Most frequent false negatives for the best results with our sample

ORG	freq
CERAE	73
MOUSE	59
RAT	40
YEAST	21
BOVIN	14
ECOLI	13
ECOLX	13

Our best results (max. f-score) for the task of finding all organisms in the gold standard combining organisms of interacting proteins and host organisms are: precision: 0.742; recall: 0.738; f-score: 0.740.

An analysis of the most frequent false positives is reported in table 2. The ranking is more or less identical with the frequency table (table 1), which is what we would expect. Manual inspection of some of the papers causing these false positives gave the following results:

- Some names of experimental methods containing organism names (which could avoid false positives if recognized as methods) were not yet included in the termbase.
- Some organisms (or their proteins respectively) are discussed in the paper, but not as results of the authors own experiments, so they do not appear in the gold standard. Obviously the curators consider only the novel findings reported in the paper, and all background information is ignored.

Table 4: Most frequent organisms in IntAct (derived from interactor proteins only)

ORG	freq
HUMAN	0.380
MOUSE	0.123
YEAST	0.108
ARATH	0.080
RAT	0.047
DROME	0.040
SCHPO	0.032
ECOLI	0.019
BOVIN	0.016
CAEEL	0.014

- While in some cases the annotators seem to decide that an organism is just used as part of the method and does not merit an inclusion, in other cases the annotators do not seem to treat the problem the same way.

An analysis of the most frequent false negatives is reported in table 3. The ranking is certainly not identical with the frequency table (table 1), which was unexpected. Manual inspection of some of the papers causing these false negatives gave the following results:

- Some common names such as "mice", and adjectives such as "murine", were absent from the taxonomy (while "transgenic mice" e.g. was present).
- There are probably more hints to recognize ECOLI (Escherichia coli K12) than just the presence of the string "K12" (or "K-12"). Our program tends to attribute all mentions of "Escherichia coli" without this string to ECOLX, generating false negatives for ECOLI and false positives for ECOLX.
- The extremely high false negative rate for CERAE (Chlorocebus aethiops, also known as Cercopithecus aethiops) is a consequence of its very different frequencies as source of interactor proteins and as a host organism.

The problem with CERAE suggests that it might be necessary to consider separately organisms in their roles as sources of the interactor proteins and as hosts for the experiments. CERAE is only frequent as a host organism, but in this role it does not appear in the papers by any of the organism names given by the taxonomy (such as "Chlorocebus aethiops", "Cercopithecus aethiops", "African green monkey", "grivet", "savanah monkey" or "vervet monkey"). The reason is that often only the names of cell lines (e.g. "Vero") derived from the organism appear in the paper.[7] To a lesser degree, this is true as well for papers where YEAST appears in this role.

A first step to deal with this problem consisted in creating different frequency tables for organisms as source of interactor proteins and as hosts of the experiment (tables 4 and 5). As these frequency tables are very different from each other and from the combined one (table 1) and as the combined task of identifying "protein organisms" and "host organisms" seems to be artificial in any case, we decided to split the problem accordingly: (a) identify organisms from which interacting proteins are derived; (b) identify host organisms. The results for each of these new tasks are not yet as good as the result for the combined task we described above, but as the information we are looking for now is more specific, this was to be expected.

3.1 Identification of "Interactor Organisms"

In order to obtain a solution for this more specific task, we just kept the formula as for the original task, but replaced the frequency table for "interactor and host organisms" (table 1) by a new one for "interactors only" (table 4). At the same time we raised the threshold to 18%: as the new freqency tables tended to nearly eliminate several typical host organisms, the remaining candidates for "interactor organisms" profited by this, so the threshold had to be raised to maximize f-score. The rest of the parameters remained identical.

Obviously, a new gold standard for "interactors only" had to be derived from IntAct. Our best results for this new task are: precision: 0.697; recall: 0.693; f-score: 0.695.

3.2 Identification of "Host Organisms"

For this alternative task we also had to improve the input, not just the formula, as we noticed that of-

[7] The Vero lineage is a very popular cell line isolated from kidney epithelial cells extracted from an African green monkey ("Cercopithecus aethiops").

Table 5: Most frequent organisms in IntAct (host organisms only), freq* is computed excluding "in vitro"

ORG	freq	freq*
"in vitro"	0.363	-
YEAST	0.262	0.412
HUMAN	0.167	0.264
CERAE	0.036	0.057
MOUSE	0.035	0.055
ARATH	0.021	0.034
DROME	0.021	0.034
SCHPO	0.020	0.031
ECOLX	0.017	0.027
RAT	0.010	0.015

ten species which were given as hosts by IntAct were not mentioned by any of their names (most importantly CERAE). So we decided to include another category of biological named entities in our termbase, namely cell line names. These were derived from one of the largest collections of cell lines information: the Cell Lines Knowledge Base (CLKB, (Sarntivijai et al., 2008)). However, a few cell line names which are type-ambiguous with other types of NE's in our termbase (normally proteins) had to be ignored to avoid conflicts. Another new input to the formula was the mention of "in vitro", contained in our termbase as a method, but used by the IntAct annotators as annotation for the "host organism".

The following adaptations to the ranking formula were necessary. The frequency table for "interactor and host organisms" (table 1) was replaced by a new one for "hosts only", including "in vitro" (table 5). At the same time the default had to be changed to "in vitro" and was given a nearly identical weight of 30 fulltext mentions (instead of 28), the threshold remained at 4% and the abstract mentions were given a weight of 35 fulltext mentions. The new cell line mentions were given a weight of 3 fulltext mentions for their respective organisms. Of course, a new gold standard for "interactors only" was derived from IntAct also in this case. Our best results yet for this new task are: precision: 0.689; recall: 0.737; f-score: 0.712.

4 Related Work and Discussion

The task of organism recognition is only recently starting to emerge as an independent subtask in biomedical IE. For example, the latest BioCreative competitive evaluation of text mining system for biology[8] included a task of protein-protein interaction detection (Krallinger et al., 2008). Although organism recognition was not officially evaluated, many participants found that it was an indispensable step in order to perform accurate protein recognition and disambiguation. As a consequence, the BioCreative meta-server (Leitner et al., 2008), offers organism recognition as one of its services (called "TX task").

(Wang and Matthews, 2008) is perhaps the most comprehensive study to date dealing with species disambiguation for term disambiguation. They combine a rule-based species disambiguation approach with a maximum entropy classifier based on contextual features of the term to be disambiguated. They evaluate in detail the contribution of both approaches over two separate corpora. While previous work has shown the benefits of using species information for term disambiguation (Alex et al., 2008; Rinaldi et al., 2008), this is perhaps the first study which also provides a separate evaluation of species disambiguation in itself. Since their purpose is to use the organism mentions to disambiguate entities, they evaluate how far their system can identify the organisms associated with each entity mention in the document. They report a level of accuracy that reaches 74.24% on one of their test corpora.

Since our results are for whole articles, not single entity mentions, they are not directly comparable. The advantage of our approach resides in its simplicity, since it does not require a specifically designed training set, being based only on publicly available standard databases. This reduces not only the cost compared to building own resources, but also ensures that their quality is monitored.

In this paper we have not discussed how our results can be used in the disambiguation of entities. As long as only one organism is selected as the focus of a given research publication, this is a rather trivial task. However, as mentioned already in section 2.4, it is often the case that multiple organisms are considered within the same publication. In that

[8] http://www.biocreative.org/

case, organism mentions would need to be 'localized' within the article in order to serve for disambiguation purposes, as done in (Wang and Matthews, 2008). Our own approach to this problem is presented and discussed in (Kaljurand et al., 2009).

One important limitation of our approach is its reliance on explicit mentions of organisms by their names as stored in the termbase (or minor variants thereof). Using all the names available to us (including cell lines) and their variants we could so far achieve only a maximal value of 88% recall, which means that 12% of the organisms are not referred to by any name in our resources. This may be due to either missing names in the termbase (the organisms are mentioned, but by different names) or because they are identified by human readers through other contextual hints which may consist of any sort of information,[9] and may presuppose massive amounts of background knowledge. The first problem might be adressed by adding other sources of names to our termbase. The second problem might be adressed by using a machine learning approach, which however brings with it a whole set of new problems, such as selection and representation of the features relevant for training, as well as the fact that a sufficiently large training corpus needs to be available.

Another limitation of our approach is the fact that its development and testing rests on its application to the identification of either organisms of protein interactors or host organisms. The original formulation of the goal that motivated this work was "to identify automatically the organisms forming part of the subject matter of scientific papers". This leaves open the question of the application of the results, and is deliberately vague in the wording "part of the subject matter", which includes but is not confined to the cases mentioned above. This formulation was motivated by a desire to keep the task as generic as possible, so that the resulting application could not only be used as a module for the protein disambiguation task, but also for other tasks of NE disambiguation with respect to organisms, as well as for organism identification as an independent task. Additionally, the ranked list of candidate organisms delivered by our program could also be presented to human users, who might want to use them in novel ways, for example in an assisted curation environment.

However, the gold standard by which we test our results is tailored to its application as a protein disambiguation module, just as the frequency tables we use. Even apart from this, the appropriateness of the gold standard is partly questionable, as it does not only prefer organisms involved in protein interactions to those that are not, but also "new" knowledge to "old" knowledge, etc. Our approach, based on "correcting" simple counts of organism mentions using frequency tables, can only be successful as long as there is a gold standard for the specific application that is being pursued. We can derive from IntAct useful gold standards for organisms from which protein interactors are derived or host organisms, but we have no gold standard for "organism identification" as an independent task.

5 Conclusion

In this paper we discussed an approach to the problem of "organism identification" as an independent task, based only on standard resources. While the initial results were interesting, the experimental setup led us to identify more specific aspects of the problem, and in particular to distinguish organisms mentioned in their roles as sources of the interacting proteins and as hosts of the experiments. We have shown that a clear identification of the different functional roles played by organism mentions can lead to more accurate results.

Although a fully automated disambiguation process based on organism mentions is not within immediate reach, the results described in this paper appear already potentially useful for protein name disambiguation in a curation environment. Another possible application would be in biomedical curation-based databases, for the semi-automatic tagging of publications with their focus organisms.

Acknowledgements

This research is partially funded by the Swiss National Science Foundation (grant 100014-118396/1). Additional support is provided by Novartis Pharma AG, NITAS, Text Mining Services, CH-4002, Basel, Switzerland. We thank the anonymous reviewers for their insightful comments.

[9] A trivial example would be a publication in a journal which specializes in research on a single organism.

References

[Alex et al.2008] Beatrice Alex, Claire Grover, Barry Haddow, Mijail Kabadjov, Ewan Klein, Michael Matthews, Richard Tobin, and Xinglong Wang. 2008. Automating curation using a natural language processing pipeline. *Genome Biology*, 9(Suppl 2):S10.

[Kaljurand et al.2009] Kaarel Kaljurand, Fabio Rinaldi, Thomas Kappeler, and Gerold Schneider. 2009. Using existing biomedical resources to detect and ground terms in biomedical literature. In *12th Conference on Artificial Intelligence in Medicine (AIME'09)*, Verona, Italy, 18–22 July.

[Kappeler et al.2008] Thomas Kappeler, Simon Clematide, Kaarel Kaljurand, Gerold Schneider, and Fabio Rinaldi. 2008. Towards automatic detection of experimental methods from biomedical literature. In *Third International Symposium on Semantic Mining in Biomedicine (SMBM)*.

[Kerrien et al.2006] S. Kerrien, Y. Alam-Faruque, B. Aranda, I. Bancarz, A. Bridge, C. Derow, E. Dimmer, M. Feuermann, A. Friedrichsen, R. Huntley, C. Kohler, J. Khadake, C. Leroy, A. Liban, C. Lieftink, L. Montecchi-Palazzi, S. Orchard, J. Risse, K. Robbe, B. Roechert, D. Thorneycroft, Y. Zhang, R. Apweiler, and H. Hermjakob. 2006. IntAct — Open Source Resource for Molecular Interaction Data. *Nucleic Acids Research*.

[Krallinger et al.2008] Martin Krallinger, Florian Leitner, Carlos Rodriguez-Penagos, and Alfonso Valencia. 2008. Overview of the protein-protein interaction annotation extraction task of BioCreative II. *Genome Biology*, 9(Suppl 2):S4.

[Leitner et al.2008] Florian Leitner, Martin Krallinger, Carlos Rodriguez-Penagos, Jörg Hakenberg, Conrad Plake, Cheng-Ju Kuo, Chun-Nan Hsu, Richard Tzong-Han Tsai, Hsi-Chuan Hung, William W. Lau, Calvin A. Johnson, Rune Saetre, Kazuhiro Yoshida, Yan Hua Chen, Sun Kim, Soo-Yong Shin, Byoung-Tak Zhang, William A. Baumgartner, Lawrence Hunter, Barry Haddow, Michael Matthews, Xinglong Wang, Patrick Ruch, Frédéric Ehrler, Arzucan Özgür, Günes Erkan, Dragomir R. Radev, Michael Krauthammer, ThaiBinh Luong, Robert Hoffmann, Chris Sander, and Alfonso Valencia. 2008. Introducing meta-services for biomedical information extraction. *Genome Biology*, 9(Suppl 2):S6.

[Rinaldi et al.2008] Fabio Rinaldi, Thomas Kappeler, Kaarel Kaljurand, Gerold Schneider, Manfred Klenner, Simon Clematide, Michael Hess, Jean-Marc von Allmen, Pierre Parisot, Martin Romacker, and Therese Vachon. 2008. OntoGene in BioCreative II. *Genome Biology*, 9(Suppl 2):S13.

[Sarntivijai et al.2008] Sirarat Sarntivijai, Alexander S. Ade, Brian D. Athey, and David J. States. 2008. A bioinformatics analysis of the cell line nomenclature. *Bioinformatics*, 24(23):2760–2766.

[UniProt Consortium2007] UniProt Consortium. 2007. The universal protein resource (UniProt). *Nucleic Acids Research*, 35:D193–7.

[Wang and Matthews2008] Xinglong Wang and Michael Matthews. 2008. Distinguishing the species of biomedical named entities for term identification. *BMC Bioinformatics*, 9(Suppl 11):S6.

[Zanzoni et al.2002] A. Zanzoni, L. Montecchi-Palazzi, M. Quondam, G. Ausiello, M. Helmer-Citterich, and G. Cesareni. 2002. MINT: a Molecular INTeraction database. *FEBS Letters*, 513(1):135–140.