Creation and evaluation of a dictionary-based tagger for virus species and proteins

Helen Victoria Cook
Rūdolfs Bērziņš
Cristina Leal Rodríguez
Novo Nordisk Foundation
Center for Protein Research,
Faculty of Health and Medical Sciences,
University of Copenhagen, Denmark
helen.cook@cpr.ku.dk

Juan Miguel Cejuela
TUM, Department of Informatics,
Bioinformatics & Computational Biology,
i12, Boltzmannstr. 3, 85748
Garching/Munich, Germany

Lars Juhl Jensen
Novo Nordisk Foundation
Center for Protein Research,
Faculty of Health and Medical Sciences,
University of Copenhagen, Denmark
lars.juhl.jensen@cpr.ku.dk

Abstract

Text mining automatically extracts information from the literature with the goal of making it available for further analysis, for example by incorporating it into biomedical databases. A key first step towards this goal is to identify and normalize the named entities, such as proteins and species, which are mentioned in text. Despite the large detrimental impact that viruses have on human and agricultural health, very little previous text-mining work has focused on identifying virus species and proteins in the literature. Here, we present an improved dictionary-based system for viral species and the first dictionary for viral proteins, which we benchmark on a new corpus of 300 manually annotated abstracts. We achieve 81.0% precision and 72.7% recall at the task of recognizing and normalizing viral species and 76.2% precision and 34.9% recall on viral proteins. These results are achieved despite the many challenges involved with the names of viral species and, especially, proteins. This work provides a foundation that can be used to extract more complicated relations about viruses from the literature.

1 Introduction

Viruses are major human and agricultural pathogens. Influenza A in the US alone costs billions of dollars each year in lost wages and medical expenses (Molinari et al., 2007). Worldwide, Influenza, Human papilloma virus and Hepatitis C virus are each responsible for at least a quarter of a million deaths each year (WHO, 2014). At the same time, viruses such as Zika virus are emerging as global health threats as the habitats of their vectors are expanding due to climate change (Mills et al., 2010; Fauci and Morens, 2016). Such arboviruses are previously neglected diseases, and as such vaccines and antiviral drugs are not available for them, posing a large health risk.

The impact of outbreaks in livestock can also be immense. The 2001 Foot and mouth disease virus outbreak in the UK cost an estimated 8 billion (Knight-Jones and Rushton, 2013) and still today much remains unknown about the virus, including the mechanisms for persistent infection (Paul et al., 2010), and the virus' interactions with the immune system that may aid cross serotype vaccine production (Paton and Taylor, 2011).

The study of viruses is aided by bioinformatics resources such as protein–protein interaction databases. Having a comprehensive picture of a virus protein's interaction partners is crucial to the understanding of the viral lifecycle and aids in the search for vaccines and antiviral drugs (Shah et al., 2015). However, manually creating and maintaining such resources is a cost, time and labour intensive endeavour (Attwood et al., 2015). Text mining provides a means to automatically identify relevant publications and the entities of interest that are mentioned in them quickly and at low cost. A first step towards building these resources for viruses is the identification of viral species and their proteins in text.

1.1 Background

Text mining for viruses presents several challenges over text mining for species of cellular organisms. Viruses are often known by many different names, either because a virus was identified in different countries and given different names (e.g. Bovine pestivirus and Bovine viral diarrhoea virus), or because the taxonomy has changed (e.g.

Proceedings of the BioNLP 2017 workshop, pages 91–98,
Vancouver, Canada, August 4, 2017.

polyomaviruses). Another source of synonyms is the use of the disease that the virus causes in place of the virus name.

Viral proteins are even more challenging to text mine, as they are often referred to by one-letter names such as E or M. Further, even if their names are longer, they can be written in many different orthographic variants e.g. U(S)11, Us11, or US11. Viral proteins may also have many synonyms related to the gene name, position on a segment, or may be referred to by their function e.g. "the polymerase". Some RNA viruses have polyproteins which complicates their analysis. Their viral mRNA codes for a single open reading frame that is translated to a polypeptide product, which is then post translationally cleaved into functional protein products. Bioinformatics databases such as UniProt (The UniProt Consortium, 2014) often give first class identifiers to the polyprotein but not to the cleavage products, thus complicating the process of referring to the functional protein product.

These challenges can be mitigated by using a dictionary approach to text mining. In such an approach, comprehensive dictionaries are created to contain all the alternative names that are likely to be referred to in a corpus. In this work, we have chosen to use a dictionary based method based on the success of this approach to identify bacteria species and biotopes (Cook et al., 2016). We have chosen to use curated databases (NCBI taxonomy and UniProt) to populate the dictionary instead of other approaches such as unsupervised methods to learn which items are named entities (Neelakantan and Collins, 2014), as the data available in these databases is high quality and openly available. Furthermore, starting with a resource dramatically reduces the difficulty of normalization of recognized entities.

Previous work in this field includes LINNAEUS (Gerner et al., 2010), a dictionary-based system that is also designed to recognize species in abstracts. The SPECIES tagger (Pafilis et al., 2013) is a newer and faster dictionary system that aims to identify names of any species in biomedical abstracts. SPECIES has achieved good performance when tagging names of viruses species in abstracts from virology journals. A more recent and specialized effort used the dictionary and template-based ANDSystem to text mine the HCV interactome (Saik et al., 2015).

Here, we improve on the SPECIES dictionary for all virus species, and additionally tag names of virus proteins for those proteins that have reference proteomes in UniProt. We have created viral species and protein dictionaries, and a gold-standard corpus that has been annotated by 4 human annotators.

2 Availability

The version of the dictionaries used in this publication are available at `http://figshare.com/articles/virus_entities_tsv/4721287` and the most recent version will be available at `http://download.jensenlab.org/`.
The V300 corpus and annotator guidelines is publicly available at `http://www.tagtog.net/-corpora`. The evaluation code is available at `http://github.com/bitmask/textmining-stats`.
The tagger software used for this work is available at `http://bitbucket.org/larsjuhljensen/tagger`.

3 Methods

3.1 Dictionary creation and tagging

Virus names were taken from NCBI Taxonomy (Sayers et al., 2009) and included all synonyms at all taxonomic levels under the viruses superkingdom. Disease names were taken from the Disease Ontology (Kibbe et al., 2015) and were manually mapped onto the correct virus taxid, giving an additional 387 names for 102 species that are human pathogens. This resulted in a total of 173,367 names for 150,885 virus tax levels.

Virus species name acronyms were taken from the ninth ICTV report on virus taxonomy (King et al., 2012) by text mining the document and extracting any text in parentheses that appears to be an acronym and that follows a match for a virus name. This way we found 778 acronyms, more than 500 of which were not found in the previous sources, for 662 virus species.

Virus protein names were taken from UniProt reference proteomes (The UniProt Consortium, 2014) as of Aug 31, 2015. Viruses that did not have complete proteomes were not included in the protein dictionary, although they are included in the species dictionary. Protein names and synonyms were taken from all fields in the UniProt record, including the protein name, short name,

gene, and chain entries if the protein is a polyprotein. Additionally, many variants of the protein names were generated following a set of rules to cover orthographic variation, such as "X protein" is expanded to "protein X" and "X". For a complete list of rules, refer to the code. This resulted in 16,580 proteins with 112,013 names from 397 virus species.

Stopwords were adapted from the text mining done for the text-mining channel of the STRING database (Szklarczyk et al., 2015). Additional stopwords were found by running the dictionary over all documents in PubMed and inspecting the 100 most frequent matches to determine if they should be stopworded. Although normally considered to be stopwords by the tagging system, specific one and two letter names from the dictionary were permitted to be matches to enable finding very short protein names.

Automated tagging used the dictionaries described above and the tagger text-mining system developed for the SPECIES resource (Pafilis et al., 2013).

3.2 Corpus creation and gold standard creation

300 abstracts were selected randomly by filtering abstracts mentioned in reviewed UniProt entries for viral proteins for top virology journals as determined by impact factor. Documents were divided among four annotators such that each pair of annotators shared 10 documents, implying that 20% of the documents were annotated by two annotators. These overlapping documents were used to calculate inter-annotator agreement (IAA), and the annotators were blind to which documents were in this set throughout the project.

Annotation guidelines were agreed upon following the annotation of 10 documents in a pilot set, which were not used in the evaluation of IAA or to assess the performance of the tagger. All abstracts were manually annotated using tagtog (Cejuela et al., 2014), an online system for text mining. Species names were normalized to NCBI taxonomic identifiers. Protein names were normalized to UniProt entry names, unless they were the cleavage product of a polyprotein, in which case they were normalized to their chain name.

3.3 Evaluation

The IAA among the human annotators was determined separately for viral species and proteins by determining the number of annotations that overlap and contain the same normalization. Boundaries of annotations were considered to match if the annotations overlapped.

Species normalizations were considered to match if one was a parent of the other and if both were at or below species level, or if both were below species level and had a common parent. For example, both of the following pairs were considered matches: "Influenza A" and "Influenza A H1N1", and "Influenza A H1N1" and "Influenza A H7N9". This allowed for an annotation to not be penalized if the strain was annotated instead of the species, or if two different strains of the same species were annotated. Protein normalizations were considered to match if they were within 90% identity according to BLAST (Zhang et al., 2000).

IAA was measured by F-score, however since we allow boundaries to overlap, this measure may not be symmetric. If one annotator has annotated "long form (short form)" as one annotation, and another annotator has annotated it as two annotations, then this will count as one true positive when comparing the first annotator to the second, but as two true positives when comparing the second annotator to the first. To avoid this asymmetry, we counted all the true positives, false negatives and false positives across both annotators.

The guidelines specify that if a span refers to multiple entities, then it should be normalized to each of them. Each normalization was treated as contributing separately to the number of true or false positives. A special case was established for Adenovirus, which is a large genus containing very many species of viruses that have a highly conserved set of proteins. Adenovirus proteins are often referred to in general in the literature, without specifying a specific species. Manual annotation of Adenovirus proteins required that only one representative protein from one species be tagged, thus effectively treating this genus as a single species.

The recall and precision of the tagger was calculated against the consensus of the human annotations. The consensus was determined as follows. If only one annotator annotated the document, their annotations were taken as the gold standard. The annotations were similarly accepted as the gold standard if two annotators agreed on position and normalization. However, if there was a disagreement, then a third annotator was asked to

resolve it. For positions that overlapped, the union of the spans was used as the consensus.

The precision and recall were calculated in three different ways. The first method required that the boundaries and normalizations of the consensus and tagger annotations match. The second method, "boundaries only", required only the boundaries of the annotations to match. The last method, "document level normalization", compared the lists of unique normalizations found in the document, regardless of position and number of occurrences.

4 Results and Discussion

4.1 Corpus and Inter-annotator agreement

The corpus consisted of 300 documents with 1,826 species and 2,540 protein annotations. There was overall good agreement between annotators for both species and proteins. The mean IAA F-score for species was 87.3%, and considering boundaries only was 90.0%. For proteins, the mean IAA F-score was 76.5%, which rose to 86.9% when considering boundaries only. Detailed results are shown in figure 1.

There was substantial agreement between annotators regarding the location of species and protein annotations, and there was also good agreement on the normalization of species. However, there was less agreement among protein normalizations than those for species. 26% of these disagreements involve one annotator normalizing a protein name to a UniProt entry, and the second annotator reporting the normalization as unknown. An additional 20% of the disagreement is due to an annotator normalizing a span to multiple entities and another annotator normalizing it to fewer entities. Such cases, in which an abstract discusses a protein in one virus and compares it to a closely related protein, can be ambiguous and refer to the protein without being completely clear about which species is being referred to.

However, the largest part this disagreement comes from instances in which annotators have normalized to different proteins that are different enough to not pass the 90% identity BLAST criterion. Manual inspection of these proteins indicate that the majority are correct, but that fast viral evolution has caused the protein sequences of similar isolates to diverge. The set of documents randomly chosen to calculate IAA was unlucky to contain a few documents containing proteins that

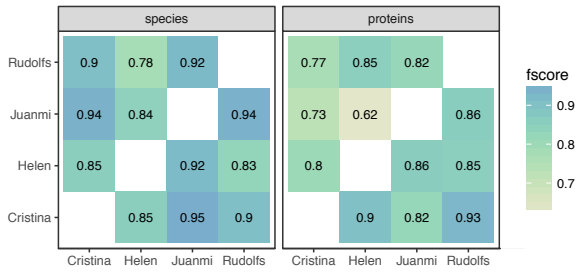

Figure 1: Inter-annotator agreement for viral proteins and species. Above the diagonal both normalization and boundaries are required to be correct, below the diagonal only identification of boundaries are required to be correct.

are quite divergent, but this is not representative of the whole corpus. This can be seen by dropping the BLAST identity criterion to 50%, which then accounts for 29% of the difference between annotators, but increases the tagger precision and recall by only 1%.

4.2 Tagger performance for species

The automatic tagger achieved 81.5% precision and 73.3% recall for the combined task of recognizing and normalizing viral species. When requiring only the boundaries to be correct, i.e. recognition but not normalization, the precision and recall were 93.1% and 79.8% respectively. At the document level, the normalization precision was 74.9% and the recall was 85.4%. Results are summarized in table 1. Combined, this shows that if the tagger identifies a viral species, it is very likely that a viral species is mentioned at the reported position, and it is also likely that the tagger has normalized it correctly. Also, the tagger correctly identifies most of the species that are mentioned in a document.

In 43% of the cases of incorrect species normalization, the tagger has identified both the correct species normalization and additional normalizations with the same abbreviation. For example, the tagger normalized SV40 to "Simian virus 40", which is correct, but also to "Polyomavirus sp." under unclassified Polyomaviridae because both taxa have SV40 as an abbreviation in the NCBI taxonomy. The abbreviation SV40 will thus count as both a true positive and a false positive with respect to normalization. If instead such partially correct normalizations were counted only as true positives, the precision would rise from 81.5% to 85.8%.

The tagger does not attempt to correctly identify all referenced entities in sentence constructs such as "HSV types 1 and 2" although such normalizations are obvious to human annotators. More ambiguously, papers that discuss Influenza proteins or Adenovirus proteins, without specifying the species (such as Influenza A, or Adenovirus type 1) are not clear about what exactly is being referred to.

In an additional 32% of the cases of incorrect species normalization, an annotator identified the virus as unclassified in which case it and the taxa identified by the tagger joined the taxonomic tree above the species level, and so was not considered to be a match by the matching code. If the match is relaxed to genus level, then the precision will rise from 81.5% to 85.0% and to 86.3% if accepting also partially correct normalizations as described above.

Despite efforts to be comprehensive, some abbreviations are missing from the virus dictionary, for example the abbreviations Ad2 and Ad5 for Adenovirus type 2 and 5 respectively were not included in the dictionary. The tagger does contain logic to identify and expand acronyms on the fly, but has very strict matching criteria to prevent false positives (Pafilis et al., 2013). Further, synonyms that are not present in NCBI taxonomy will not be identified. For example "Blackberry yellow vein disease" was not identified as as synonym for "Blackberry yellow vein virus" and so was not found by the tagger. This could be improved with more comprehensive synonym generation.

The tagger will tag all instances of entries in its dictionary, even in contexts that are not appropriate. The annotation guidelines state that viruses that are used as vectors should not be tagged, since the scientific work they are mentioned in is not primarily about the virus. However, this is a matter of opinion and the opposite case could also be argued. Regardless, the tagger cannot distinguish the context in which viruses are mentioned, and will blindly tag all occurrences of the virus name.

4.3 Tagger performance for proteins

For combined recognition and normalization of viral proteins, the precision and recall of the tagger were 76.2% and 34.9% respectively. Observing boundaries only, the precision and recall rose to 87.4% and 40.0% respectively. At the document level, the normalization precision was 76.2% and the recall was 38.1%. Results are summarized in table 2.

	Precision	Recall
Normalisation	81.5%	73.3%
Boundaries only	93.1%	79.8%
Doc level normalisation	74.9%	85.4%
Partially correct norm	85.8%	73.3%
Match at genus level	85.0%	73.5%
Previous two criteria	86.3%	73.5%

Table 1: Summary of species precision and recall for different evaluation criteria: Normalization and recognition, recognition of boundaries only, normalization at the document level, treating entities that have been normalized to multiple entities as correct if one of the normalizations is correct, relaxing the matching criterion to the genus level, and finally allowing both of the previous two criteria.

Since viral protein names are so short and not unique to one species, the tagger will only tag protein names for species that have already been identified. This means that the theoretical upper bound for tagging proteins is equivalent to the species document level normalization recall (85.4%) assuming that all the proteins are present in the dictionary. However, the dictionary only contains protein names for species that are contained in reviewed UniProt proteomes, a total of 348 species and 88.1% of the proteins mentioned in the corpus. This gives a maximum possible recall of 75.2% for proteins. Conversely, since the tagger detects proteins only after the species has been detected, the normalization of the viral proteins that are found is quite accurate.

Considering only annotation of the proteins in the dictionary, the precision was 86.0% and the recall 35.5%. Recall does not change significantly from considering all proteins because there are 10 times more false negatives due to not locating the protein compared to false negatives due to incorrectly normalizing the protein. At the document level, the normalization precision of proteins that were present in the dictionary is 77.1% and the recall is 50.7%.

Viral proteins are very hard for the tagger to identify due to the diversity of names that are used to refer to them. For example, the tagger has missed 97% of names in which the protein is referred to by its molecular weight (e.g. "the 33K

protein"). Including these synonyms would increase the recall by 4 percentage points. Similarly, the tagger has tagged only 10% of the cases in which the viral protein is referred to by its function (e.g. "the helicase"). Including these synonyms would increase the recall by 6 percentage points. As observed for species, the tagger does not recognize novel abbreviations, such as "sGP" for the Ebola virus nonstructural small glycoprotein, and such constructs are used quite frequently in the literature. Better on-the-fly acronym identification in the tagger may help increase this recall rate.

Another source of error is the ambiguity of terms used in the text to refer to parts of the virus that are also names of proteins such as "capsid". Although the frequently-named capsid protein is the main constituent of the viral capsid, references in the text to "capsid" are often ambiguous as to whether they refer to the protein or to the assembled virus part. The annotation guidelines state that such terms should only be tagged if they refer to the protein and should not be tagged if they refer to part of the virus, but these cases are often difficult to distinguish in practice.

The tagger identifies false positives at a much lower rate than false negatives. Since very short protein names are present in the dictionary, it is much more likely for these names to appear in places that are not in the context of a protein. For example, Coronavirus infectious bronchitis virus has a spike protein abbreviated S, however discussion of the polyprotein cleavage site before a serine residue will be false positively tagged as serine is also abbreviated S.

Normalization of protein names to multiple entities can also be incorrect in instances where an abstract discusses both a specific protein in one species, and the same protein in many species. The tagger will tag all instances of the protein name with all species and will not be able to distinguish the instances that refer only to the protein in a specific species, whereas human annotators are more easily able to distinguish these cases.

4.4 Results in other corpora

Compared to the S800 virus corpus (Pafilis et al., 2013), the improved dictionary finds over 100 more mentions, including new abbreviations, but does not tag more general terms such as "infectious virus" and "avian viruses" which refer to more than one species. Measured against the S800 gold standard for only virus annotations in the virus subset of the corpus, the improved tagger has a precision and recall of 63.3% and 57.0% respectively, compared to the initial results from SPECIES of 63.2% and 53.0% respectively.

	Precision	Recall
Normalisation	77.5%	35.5%
Boundaries only	87.4%	40.0%
Doc level normalisation	77.1%	50.7%
Theoretical max recall	-	75.2%

Table 2: Summary of results for protein detection for different evaluation criteria: normalization and recognition, recognition of boundaries only and document level normalization. The theoretical maximum recall based on requiring the species to be recognized and present in the dictionary is also listed.

Running the tagger over all of Medline finds over 53 million mentions of 8063 viral species in more than 1.5M articles. Of these, we have protein level detail for 348 species, and find over 10M mentions of 4668 unique proteins. The most commonly mentioned species is HIV-1, making up over 3% of species mentions.

5 Conclusions and Perspectives

As the biomedical literature continues to grow at an exponential rate (Lu, 2011), automated tools, such as text mining, are necessary to enable extracting information from the literature in a timely and efficient manner. Text mining is a means to automatically extract information from the literature without requiring manual curation of a large number of documents. It can be used successfully to extract virus species and proteins from abstracts that pertain to viruses with good precision and also, in the case of species, good recall. There is still much room to improve the recall of proteins due to the abundance of alternative names that are used to refer to them. Further, the tagger does not recognize disjoint entities, and since there has recently been progress in this field (Tang et al., 2013), this could also be an area for future improvement of the tagger.

These results can be used in future work to extract co-occurrences of virus and host proteins, which could imply an interaction between these proteins. Integrating virus-host protein-protein in-

teractions into the larger host interaction network may provide insight into viral mechanisms of disease. Work done specifically on EBV, HPV, and Hepatitis C virus (Gulbahce et al., 2012; Mosca et al., 2014) revealed that host proteins local to viral targets form network modules that are related to the diseases caused by these viruses. With the virus-agnostic tools presented here, such work can be scaled up to easily enable investigation of all viruses for which there is sufficient data.

The work presented here could also be used as a foundation to identify viruses that are understudied compared to their impact, and may reveal future directions that are promising to study. The interrelationship of proteins and diseases has been explored recently using text mining to assess both the strength of an interaction between a protein and a disease, and also the scarceness of publications about a given protein target (Cannon et al., 2017). This gives researchers an overview of understudied proteins that could be relevant for disease etiology. A similar approach could be taken to reveal new directions in virus research.

Acknowledgements

The authors would like to thank Jorge Campos for his work on the interface of tagtog to support this project.

References

Teresa Attwood, Bora Agit, and Lynda Ellis. 2015. Longevity of Biological Databases. *EMBnet.journal* 21(0).

Daniel C Cannon, Jeremy J Yang, Stephen L Mathias, Oleg Ursu, Subramani Mani, Anna Waller, Stephan C Schurer, Lars Juhl Jensen, Larry A Sklar, Cristian G Bologa, and Tudor I Oprea. 2017. TIN-X: Target Importance and Novelty Explorer. *Bioinformatics* pages 1–3.

Juan Miguel Cejuela, Peter McQuilton, Laura Ponting, S. J. Marygold, Raymund Stefancsik, Gillian H. Millburn, and Burkhard Rost. 2014. Tagtog: Interactive and text-mining-assisted annotation of gene mentions in PLOS full-text articles. *Database* 2014:1–8. https://doi.org/10.1093/database/bau033.

Helen Cook, Evangelos Pafilis, and Lars Jensen. 2016. A dictionary- and rule-based system for identification of bacteria and habitats in text. In *Proceedings of the 4th BioNLP Shared Task Workshop*. pages 50–55. http://www.aclweb.org/anthology/W/W16/W16-30.pdf#page=60.

Anthony S Fauci and David M Morens. 2016. Zika Virus in the Americas Yet Another Arbovirus Threat. *The New England journal of medicine* 374:601–604. https://doi.org/10.1056/NEJMp1600297.

Martin Gerner, Goran Nenadic, and Casey M Bergman. 2010. LINNAEUS: a species name identification system for biomedical literature. *BMC bioinformatics* 11(1):85. https://doi.org/10.1186/1471-2105-11-85.

Natali Gulbahce, Han Yan, Amélie Dricot, Megha Padi, Danielle Byrdsong, Rachel Franchi, Deok-Sun Lee, Orit Rozenblatt-Rosen, Jessica C. Mar, Michael A. Calderwood, Amy Baldwin, Bo Zhao, Balaji Santhanam, Pascal Braun, Nicolas Simonis, Kyung-Won Huh, Karin Hellner, Miranda Grace, Alyce Chen, Renee Rubio, Jarrod A. Marto, Nicholas A. Christakis, Elliott Kieff, Frederick P. Roth, Jennifer Roecklein-Canfield, James A. DeCaprio, Michael E. Cusick, John Quackenbush, David E. Hill, Karl Münger, Marc Vidal, and Albert-László Barabási. 2012. Viral Perturbations of Host Networks Reflect Disease Etiology. *PLoS Computational Biology* 8(6):e1002531. https://doi.org/10.1371/journal.pcbi.1002531.

Warren A. Kibbe, Cesar Arze, Victor Felix, Elvira Mitraka, Evan Bolton, Gang Fu, Christopher J. Mungall, Janos X. Binder, James Malone, Drashtti Vasant, Helen Parkinson, and Lynn M. Schriml. 2015. Disease Ontology 2015 update: An expanded and updated database of Human diseases for linking biomedical knowledge through disease data. *Nucleic Acids Research* 43(D1):D1071–D1078. https://doi.org/10.1093/nar/gku1011.

Andrew M.Q. King, Michael J. Adams, Eric B. Carstens, and Elliot J. Lefkowitz, editors. 2012. *Virus Taxonomy: Classification and Nomenclature of Viruses: Ninth Report of the International Committee on Taxonomy of Viruses*. Elsevier Academic Press. https://doi.org/10.1016/B978-0-12-384684-6.X0001-8.

T. J D Knight-Jones and J. Rushton. 2013. The economic impacts of foot and mouth disease - What are they, how big are they and where do they occur? *Preventive Veterinary Medicine* 112(3-4):162–173. https://doi.org/10.1016/j.prevetmed.2013.07.013.

Zhiyong Lu. 2011. PubMed and beyond: A survey of web tools for searching biomedical literature. *Database* 2011:1–13. https://doi.org/10.1093/database/baq036.

James N. Mills, Kenneth L. Gage, and Ali S. Khan. 2010. Potential influence of climate change on vector-borne and zoonotic diseases: A review and proposed research plan. *Environmental Health Perspectives* 118(11):1507–1514. https://doi.org/10.1289/ehp.0901389.

Noelle Angelique M. Molinari, Ismael R. Ortega-Sanchez, Mark L. Messonnier, William W. Thompson, Pascale M. Wortley, Eric Weintraub, and Carolyn B. Bridges. 2007. The annual impact of seasonal influenza in the US: Measuring disease burden and costs. *Vaccine* 25(27):5086–5096. https://doi.org/10.1016/j.vaccine.2007.03.046.

Ettore Mosca, Roberta Alfieri, and Luciano Milanesi. 2014. Diffusion of Information throughout the Host Interactome Reveals Gene Expression Variations in Network Proximity to Target Proteins of Hepatitis C Virus. *PLoS ONE* 9(12):e113660. https://doi.org/10.1371/journal.pone.0113660.

Arvind Neelakantan and Michael Collins. 2014. Learning Dictionaries for Named Entity Recognition using Minimal Supervision. *Proceedings of the 14th Conference of the European Chapter of the Association for Computational Linguistics* pages 452–461. http://www.aclweb.org/anthology/E14-1048.

Evangelos Pafilis, Sune P. Frankild, Lucia Fanini, Sarah Faulwetter, Christina Pavloudi, Aikaterini Vasileiadou, Christos Arvanitidis, and Lars Juhl Jensen. 2013. The SPECIES and ORGANISMS Resources for Fast and Accurate Identification of Taxonomic Names in Text. *PLoS ONE* 8(6):2–7. https://doi.org/10.1371/journal.pone.0065390.

David J Paton and Geraldine Taylor. 2011. Developing vaccines against foot-and-mouth disease and some other exotic viral diseases of livestock. *Philosophical transactions of the Royal Society of London. Series B, Biological sciences* 366(1579):2774–81. https://doi.org/10.1098/rstb.2011.0107.

William E Paul, Michael Mcheyzer-williams, and David Barthold, Stephen. Bowen, R. Hedrick, Ronald. Knowles, Donald. Lairmore, Michael. Parrish, Colin. Saif, Linda. Swayne. 2010. Fenner'S Veterinary Virology. *Elsevier* 5th editio(August):43–51. https://doi.org/10.1016/B978-0-12-375158-4.X0001-6.

Olga V. Saik, Timofey V. Ivanisenko, Pavel S. Demenkov, and Vladimir A. Ivanisenko. 2015. Interactome of the hepatitis C virus: Literature mining with ANDSystem. *Virus Research* 218:40–48. https://doi.org/10.1016/j.virusres.2015.12.003.

Eric W Sayers, Tanya Barrett, Dennis A Benson, Stephen H Bryant, Kathi Canese, Vyacheslav Chetvernin, Deanna M Church, Michael DiCuccio, Ron Edgar, Scott Federhen, Michael Feolo, Lewis Y Geer, Wolfgang Helmberg, Yuri Kapustin, David Landsman, David J Lipman, Thomas L Madden, Donna R Maglott, Vadim Miller, Ilene Mizrachi, James Ostell, Kim D Pruitt, Gregory D Schuler, Edwin Sequeira, Stephen T Sherry, Martin Shumway, Karl Sirotkin, Alexandre Souvorov, Grigory Starchenko, Tatiana A Tatusova, Lukas Wagner, Eugene Yaschenko, and Jian Ye. 2009. Database resources of the National Center for Biotechnology Information. *Nucleic Acids Research* 37:D5–15. https://doi.org/10.1093/nar/gkn741.

Priya S. Shah, Jason A. Wojcechowskyj, Manon Eckhardt, and Nevan J. Krogan. 2015. Comparative mapping of host-pathogen protein-protein interactions. *Current Opinion in Microbiology* 27:62–68. https://doi.org/10.1016/j.mib.2015.07.008.

Damian Szklarczyk, Andrea Franceschini, Stefan Wyder, Kristoffer Forslund, Davide Heller, Jaime Huerta-Cepas, Milan Simonovic, Alexander Roth, Alberto Santos, Kalliopi P. Tsafou, Michael Kuhn, Peer Bork, Lars J. Jensen, and Christian Von Mering. 2015. STRING v10: Protein-protein interaction networks, integrated over the tree of life. *Nucleic Acids Research* 43(D1):D447–D452. https://doi.org/10.1093/nar/gku1003.

Buzhou Tang, Yonghui Wu, Min Jiang, Joshua C. Denny, and Hua Xu. 2013. Recognizing and Encoding Discorder Concepts in Clinical Text using Machine Learning and Vector Space. *Proceedings of the ShARe/CLEF Evaluation Lab* http://www.clef-initiative.eu/documents/71612/d596ae25-c4b3-4a9a-be4a-648a77712aaf.

The UniProt Consortium. 2014. UniProt: a hub for protein information. *Nucleic Acids Research* 43(D1):D204–212. https://doi.org/10.1093/nar/gku989.

WHO. 2014. WHO Fact Sheets: Influenza, HCV, HPV. http://www.who.int/mediacentre/factsheets/.

Zheng Zhang, Scott Schwartz, Lukas Wagner, and Webb Miller. 2000. A Greedy Algorithm for Aligning DNA Sequences. *Journal of Computational Biology* 7(12):203–214. https://doi.org/10.1089/10665270050081478.