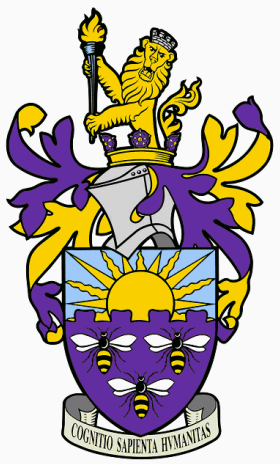
MANCHESTER 1824

The University of Manchester

Text mining molecular interactions and their context for studying disease

A thesis submitted to the University of Manchester
for the degree of Doctor of Philosophy (PhD)
in the Faculty of Life Sciences



2014

Daniel George Jamieson

Contents

LIST OF FIGURES 5

LIST OF TABLES 6

LIST OF ABBREVIATIONS 7

ABSTRACT 9

DECLARATION 10

COPYRIGHT 11

ACKNOWLEDGEMENTS 12

INTRODUCTION TO THE THESIS 13

GENERAL INTRODUCTION 16

1.1 MOTIVATION 16
1.2 HUMAN DISEASE 18
1.2.1 HUMAN PATHOGENS 21
1.2.2 PAIN AND ITS ASSOCIATED DISEASES 28
1.3 THE PUBLISHED LITERATURE: A WEALTH OF DATA 31
1.3.1 PUBMED 31
1.3.2 MANUALLY CURATING LITERATURE INTO STRUCTURED DATABASES 34
1.3.3 TEXT-MINING IN BIOMEDICAL LITERATURE 36
1.3.4 SEMI-AUTOMATIC APPROACHES TO DATA EXTRACTION AND CURATION 49
1.4 ANALYSING BIOLOGICAL DATA 51
1.4.1 SYSTEMS BIOLOGY 51
1.5 RESEARCH RATIONALE 54

TOWARDS SEMI-AUTOMATED CURATION: RECREATING THE HIV-1, HUMAN PROTEIN INTERACTION DATABASE 57

2.1 ABSTRACT 57
2.2 INTRODUCTION 58
2.3 METHODS 59
2.3.1 DATA 60
2.3.2 NAMED ENTITY RECOGNITION AND NORMALIZATION 61
2.3.3 EVENT EXTRACTION 62
2.3.4 EVENT EVALUATION 64
2.3.5 COMPARISON OF TM RESULTS TO HHPID INTERACTIONS 64
2.4 RESULTS 65
2.4.1 ACCURACY OF TEXT MINING TOOLS 65
2.4.2 COMPARISON OF HIV-1–HUMAN INTERACTIONS EXTRACTED BY TM AND THE HHPID 66
2.4.3 OTHER TYPES OF INTERACTIONS RETRIEVED BY TM 70
2.4.4 FULL-TEXT TM ANALYSIS 73
2.5 DISCUSSION 75
2.5.1 TM VERSUS MANUAL CURATION 75
2.6 CONCLUSIONS 79

2.7 ACKNOWLEDGEMENTS **80**

CATALOGUING THE BIOMEDICAL WORLD OF PAIN THROUGH SEMI-AUTOMATED CURATION **81**

3.1 ABSTRACT **81**
3.2 INTRODUCTION **82**
3.3 METHODS **84**
3.3.1 (I) BUILDING A TOPIC-SPECIFIC CORPUS 84
3.3.3 (II) DATA EXTRACTION 87
3.3.4 (III) AVAILABILITY AND VISUALISATION FOR MANUAL CURATION 92
3.4 RESULTS AND DISCUSSION **93**
3.4.1 (I) BUILDING A TOPIC-SPECIFIC CORPUS 93
3.4.2 (II) DATA EXTRACTION 97
3.4.3 (III) AVAILABILITY AND VISUALISATION FOR MANUAL CURATION 105
3.5 CONCLUSIONS **110**
3.6 ACKNOWLEDGMENTS **111**

THE PAIN INTERACTOME: CONNECTING PAIN SPECIFIC PROTEIN INTERACTIONS **112**

4.1 ABSTRACT **112**
4.2 INTRODUCTION **112**
4.3 METHODS **114**
4.3.1 DATA AVAILABILITY 114
4.3.2 THE CURATION PROCEDURE FOR PPIS 114
4.3.3 NETWORK ANALYSIS 115
4.3.4 GENE FUNCTIONAL ENRICHMENT 116
4.3.5 PAIN CATEGORY ASSIGNMENT 116
4.3.6 ANATOMICAL CATEGORIZATION 116
4.3.7 MICROARRAY ANALYSIS 117
4.4 RESULTS **117**
4.4.1 THE LITERATURE-DERIVED PAIN PPI NETWORK 117
4.4.2 COMPARATIVE ANALYSES BETWEEN ALTERNATIVE PAIN PROTEIN DATASETS 120
4.4.3 INSIGHTS INTO THE PATHOLOGY OF PAIN 122
4.5 DISCUSSION **130**
4.6 ACKNOWLEDGEMENTS **131**

EXPANDING THE HUMAN PATHOGEN INTERACTOME WITH TEXT MINING **132**

5.1 ABSTRACT **132**
5.2 INTRODUCTION **133**
5.3 METHODS **134**
5.3.1 IDENTIFYING HUMAN PATHOGENS 134
5.4.2 BUILDING A HUMAN PATHOGEN CORPUS 135
5.4.3 COLLECTING EXISTING TEXT-MINING DATA 136
5.4.4 ENHANCING TEXT-MINING DATA 136
5.4.5 INTEGRATING HOST-PATHOGEN PPI DATABASES 137
5.4.6 CURATING TEXT MINING RESULTS 138
5.2 RESULTS **139**
5.2.1 SOURCING TEXT MINING DATA 140
5.2.2 COMPARISONS OF TEXT-MINED AND PUBLIC DATABASE PPIS 143
5.2.3 CURATING TEXT MINING DATA 146
5.3 DISCUSSION **149**

5.4 ACKNOWLEDGMENTS **151**

DISCUSSION AND CONCLUSION **152**

6.1 BUILDING A DISEASE SPECIFIC CORPUS **152**
6.2 MOLECULAR INTERACTIONS AND THEIR CONTEXTS **154**
6.3 CURATING LARGE SCALE MOLECULAR INTERACTION DATASETS **156**
6.4 HARNESSING MIS FROM TM FOR FURTHER USE **158**
6.5 FUTURE DIRECTIONS **159**
6.6 CONCLUSION **161**

REFERENCES **163**

APPENDIX A

Jamieson, D.G., Gerner, M., Sarafraz, F., Nenadic, G. & Robertson, D.L. Towards semi-automated curation: using text mining to recreate the HIV-1, human protein interaction database. *Database,* bas023 (2012).

APPENDIX B

Jamieson, D.G., Roberts, P.M., Robertson, D.L., Sidders, B. & Nenadic, G. Cataloging the biomedical world of pain through semi-automated curation of molecular interactions. *Database*, bat033 (2013).

APPENDIX C

Jamieson, D.G., Moss, A., Kennedy, M., Jones, S., Nenadic, G., Robertson, D.L., Sidders, B. The pain interactome: Connecting pain-specific protein interactions. *Pain* in press (2014).

List of Figures

FIGURE 1.1 MECHANISM OF ACTION OF STATINS THROUGH THEIR INHIBITION OF HGM-COA REDUCTASE. 20
FIGURE 1.2 GLOBAL DISTRIBUTION OF INFECTIOUS DISEASE. 22
FIGURE 1.3 INNATE IMMUNITY VERSUS ADAPTIVE IMMUNITY. 23
FIGURE 1.4 CHARACTERISTICS AMONG VIRUSES. 24
FIGURE 1.5 PAIN CLASSIFICATIONS 30
FIGURE 1.6 TOTAL LITERATURE CITATIONS IN PUBMED SINCE 1990. 32
FIGURE 1.7 DEPENDENCY PARSE TREE. 41
FIGURE 1.8 THE BIOCONTEXT SYSTEM ARCHITECTURE. 48
FIGURE 1.9 NETWORK PROPERTIES 52
FIGURE 1.10 EXAMPLE PROTEIN INTERACTION NETWORKS. 53
FIGURE 2.1 SUMMARY OF THE METHODOLOGY. 60
FIGURE 2.2 METHODS OF EVENT EVALUATION. 63
FIGURE 2.3 EXAMPLES OF FALSELY REPORTED EVENT CHAINS. 76
FIGURE 2.4 TM INTERACTION INVOLVING TWO OR MORE PARTICIPANTS. 77
FIGURE 3.1. DIAGRAMMATIC REPRESENTATION OF METHODOLOGY. 84
FIGURE 3.2. PAIN DICTIONARY SUMMARY STATISTICS. 94
FIGURE 3.3. PAIN TERM MATCHES. 95
FIGURE 3.4 DOCUMENT PAIN RELEVANCY SCORES. 97
FIGURE 3.5 NUMBER OF NEGATED EVENT CHAINS. 99
FIGURE 3.6. EXAMPLE OF A TYPICAL MOLECULAR INTERACTION IN WIKI-PAIN.ORG. 106
FIGURE 4.1 THE PAIN INTERACTION NETWORK. 118
FIGURE 4.2 BIAS IN THE PAIN INTERACTION NETWORK. 119
FIGURE 4.3 DRUG TARGETS IN THE PAIN INTERACTION NETWORK. 123
FIGURE 4.4 PROTEIN REGULATION IN THE PAIN INTERACTION NETWORK. 125
FIGURE 4.5 PPIS SPECIFIC TO NEUROPATHIC PAIN. 126
FIGURE 4.6 PPIS SPECIFIC TO INFLAMMATORY PAIN. 127
FIGURE 5.1 TEXT-MINING METHODOLOGY TO RETRIEVING PATHOGEN RELATED PPIS. 140

List of tables

TABLE 1.1 EXAMPLE PROTEIN-PROTEIN INTERACTION DATABASES MANUALLY CURATED FROM THE LITERATURE. 35
TABLE 1.2 EXAMPLE BIOMEDICAL NER APPLICATIONS 42
TABLE 1.3 BIOMEDICAL EVENT TYPES USED IN THE BIONLP '09 SHARED TASK. 46
TABLE 1.4 SEMI-AUTOMATIC APPROACHES TO DATA CURATION 50
TABLE 2.1 EVENT EXTRACTION PERFORMANCE ON THE TEST-HIV GOLD STANDARD OF 50 ABSTRACTS AND TITLES 66
TABLE 2.2 THE NUMBER OF HIV-1-HUMAN INTERACTION MENTIONS EXTRACTED FROM 3,090 CITATIONS: A COMPARISON BETWEEN THE HHPID DATABASE AND THE TM RESULTS 67
TABLE 2.3 TOP 10 MOST FREQUENT PARTICIPANTS IN EVENTS AS PRESENTED IN THE HHPID AND AS EXTRACTED BY TM 67
TABLE 2.4 TOP 10 MOST FREQUENT HIV-1-HUMAN INTERACTIONS RETRIEVED THROUGH TM 69
TABLE 2.5 TOP 10 MOST FREQUENT BINDING PARTICIPANTS WITH THE HIV-1 TAT GENE 70
TABLE 2.6 TOP 10 MOST FREQUENTLY OCCURRING PARTICIPANTS WITHIN EVENT CHAINS IN THE TM RESULTS. 71
TABLE 2.7 TOP MOST FREQUENT INTERACTIONS RETRIEVED BY TM BUT NOT FOUND IN THE HHPID 72
TABLE 2.8 HIV-1 TAR AND LTR MOST FREQUENT INTERACTIONS EXTRACTED BY TM 73
TABLE 2.9 TOP 10 MOST FREQUENT INTERACTIONS RETRIEVED FROM 49 OA FULL-TEXT ARTICLES WITH TM 74
TABLE 3.1 TOP REPORTED PAIN TERMS IN P1. 96
TABLE 3.2 EVENT CHAINS FROM P1. 98
TABLE 3.3 EVENT TYPES INVOLVED IN EVENT CHAINS. 98
TABLE 3.4 TOP 10 ANATOMICAL REGIONS ASSOCIATED WITH EVENT CHAINS. 100
TABLE 3.5 OVERVIEW OF OVERALL PAIN RELEVANCY SCORES FOR UNIQUE EVENT CHAINS INVOLVING HUMAN, MOUSE OR RAT PROTEINS AND EXCLUDING SELF-INTERACTIONS. 100
TABLE 3.6 TOP DISEASES ASSOCIATED WITH DOCUMENTS CONTAINING EVENT DATA. 101
TABLE 3.7 EVALUATIONS OF TM SOFTWARE USED. 102
TABLE 3.8 PAIN GENES ENRICHMENT ANALYSIS. 104
TABLE 3.9 MANUAL CURATION EVALUATION. 107
TABLE 3.10 TOP 10 HOMOLOGUES APPEARING IN OUR MANUALLY CURATED DATA. 109
TABLE 5.1 TOP 20 MOST STUDIED HUMAN PATHOGENS. 142
TABLE 5.2 TOP 20 PATHOGENS ORDERED BY HP-PPIS FROM TEXT-MINING. 143
TABLE 5.3 PATHOGEN-RELATED PPIS FROM TEXT-MINING AND PUBLIC DATABASES 144
TABLE 5.4 PROTEINS TARGETED BY PATHOGENS 146
TABLE 5.5 RANKING METHODS FOR CURATING UNKNOWN HUMAN TARGETS FOR PATHOGENS. 149

List of abbreviations

AIDS acquired immunodeficiency syndrome

CRF conditional random field

DDIExtraction Drug-Drug Interaction Extraction

DNA deoxyribonucleic acid

DO Disease Ontology

DRG dorsal root ganglion

FP false positive

FN false negative

GH growth hormone

GHRH growth hormone releasing hormone

GO Gene Ontology

HBV hepatitis virus B

HCV hepatitis virus C

HHPID Human Protein Interaction Database

HIF-1 hypoxia inducible factor 1

HIV human immunodeficiency virus

HIV-1 human immunodeficiency virus 1

HMG-CoA 3-hydroxy-3-methylglutaryl coenzyme A

HP host-pathogen

HSV-1 herpes simplex virus 1

IAV influenza virus A

IR information retrieval

K thousand

LEP leptin

LTR long terminal repeat

M million

MCC Matthew's Correlation Coefficient

MeSH Medical Subject Heading

MI molecular interaction

ML machine learning

NCBI National Library of Medicine

NER named entity recognition

NLP natural language processing

NSAID non-steroidal anti-inflammatory drug

OA open access

PDB public database

PMC PubMed Central

PMID PubMed identifier

PNS peripheral nervous system

POS part-of-speech

PP pathogen-pathogen

PPI protein-protein interaction

RNA ribonucleic acid

TAR transactivation response element

TM text mining

TEES Turku Event Extraction System

TN true negative

TNT tibial nerve transection

TP true positive

UniProtKB UniProt Knowledgbase

UMLS Unified Medical Language System

Abstract

Molecular interactions enable us to understand the complexity of the human living system and how it can be exploited or malfunction to cause disease. The biomedical literature presents detailed knowledge of molecular functions and therefore represents a valuable reservoir of data for studying disease. However, extracting this data efficiently is difficult as it is spread over millions of publications in text that is not machine-readable. In this thesis we investigate how text mining can be used to automatically extract data for molecular interactions and their context relevant to disease. We focus on two globally relevant classes of diseases of which manifest from contrasting mechanisms: pain-related diseases and diseases caused by pathogenic organisms. Using HIV-1 as a case study, we first show that text mining can be used to partially recreate a large, manually curated database of HIV-1-human molecular interactions derived from the literature. We highlight both weaknesses in the quality of the data produced by the text-mining approach and strengths in it being able to extract this data rapidly, identifying instances missed in the manual curation and its potential as a support tool. We then expand on this approach by showing how an entirely new database of protein interactions relevant to pain can be created efficiently and accurately using text mining to generate the data and manual curation to validate the data quality. The following chapter then presents an analysis of 1,002 unique pain-related protein-protein interactions derived from this database, showing that it is of greater relevance to pain research than databases of pain interactions created from other common starting points. We highlight its value by, for example, identifying new drug repurposing opportunities and exploring differences in specific pain diseases using the contextual detail afforded by the text mining. Finally, we expand further on our approach to extracting molecular interactions from the literature, by showing how interactions between human proteins and pathogens can be curated across pathogenic organisms. We demonstrate how these techniques can be used to expand our knowledge of human pathogen related interaction data already stored in public databases, by identifying 42 new HIV-1-human molecular interactions, 108 new interactions between pathogen species and human proteins and 33 new human proteins that were found to interact with pathogens. Together, the results show that contexualised text mining, when supported by manual curation, can be used to extract molecular interactions for contrasting disease types in an efficient and accurate manner.

Declaration

The University of Manchester
PhD by Alternative Format Candidate Declaration

Candidate name: Daniel George Jamieson

Faculty: Faculty of Life Sciences

Thesis Title: Text mining molecular interactions and their context for studying disease

Declaration to be completed by the candidate:

I declare that no portion of the work referred to in this thesis has been submitted in support for an application for another degree or qualification of this or any other university or institute of learning.

Signed:  **Date**: September 27, 2014

Copyright

The author of this thesis (including any appendices and/or schedules to this thesis) owns any copyright in it (the "Copyright") and s/he has given The University of Manchester certain rights to use such Copyright, including for administrative purposes.

Copies of this thesis, either in full or in extracts and whether in hard or electronic copy, may be made only in accordance with the Copyright, Designs and Patents Act 1988 (as amended) and regulations issued under it or, where appropriate, in accordance with licensing agreements which the University has from time to time. This page must form part of any such copies made.

The ownership of certain Copyright, patents, designs, trade marks and other intellectual property (the "Intellectual Property") and any reproductions of copyright works in the thesis, for example graphs and tables ("Reproductions"), which may be described in this thesis, may not be owned by the author and may be owned by third parties. Such Intellectual Property and Reproductions cannot and must not be made available for use without the prior written permission of the owner(s) of the relevant Intellectual Property and/or Reproductions.

Further information on the conditions under which disclosure, publication and commercialisation of this thesis, the Copyright and any Intellectual Property and/or Reproductions described in it may take place is available in the University IP Policy (see http://documents.manchester.ac.uk/DocuInfo.aspx?DocID=487), in any relevant Thesis restriction declarations deposited in the University Library, The University Library's regulations (see http://www.manchester.ac.uk/library/aboutus/regulations) and in The University's policy on Presentation of Theses.

Acknowledgements

This PhD was funded by a BBSRC CASE studentship with Pfizer Worldwide Research & Development. There have been many people who have helped contribute to this thesis who have been duly acknowledged in the relevant chapters. Particular thanks must go to David Robertson, Goran Nenadic and Ben Sidders. Their dedication and support throughout this PhD was truly appreciated. I would also like to my advisor Jean-Marc Schwarz, the Robertson laboratory, members of Computational Evolutional Biology, the 'GN team', colleagues at Neusentis, friends, family and all those who helped my PhD come to fruition. I am forever grateful for all your time and efforts.

Introduction to the thesis

Permission was granted to present this thesis in "alternative" format. The thesis begins with a general introduction to the topics covered, followed by chapters 2-5 presented as journal style articles, while Chapter 6 provides a discussion of the full thesis. It was appropriate to present this thesis in alternative format as Chapters 2-4 have already been published in peer-reviewed journals and Chapter 5 is in preparation publication. Although other authors helped contribute to these (their contributions are detailed below), I, as the primary author had a major influence in all of the work undertaken.

Chapter 2

Jamieson, D.G., Gerner, M., Sarafraz, F., Nenadic, G. & Robertson, D.L. Towards semi-automated curation: using text mining to recreate the HIV-1, human protein interaction database. *Database,* bas023 (2012). See Appendix A.

DGJ was involved in all aspects of the work presented in this paper, including the writing of the paper. GM & SF provided event data. GN and DLR provided supervision throughout.

Chapter 3

Jamieson, D.G., Roberts, P.M., Robertson, D.L., Sidders, B. & Nenadic, G. Cataloging the biomedical world of pain through semi-automated curation of molecular interactions. *Database*, bat033 (2013). See Appendix B.

DGJ was involved in all aspects of the work presented in this paper, including the writing of the paper. PMR, BS, DLR & GN provided supervision throughout.

Chapter 4

Jamieson, D.G., Moss, A., Kennedy, M., Jones, S., Nenadic, G., Robertson, D.L., Sidders, B. The pain interactome: Connecting pain-specific protein interactions. *Pain* in press (2014). See Appendix C.

DGJ was involved in all aspects of the work presented in this paper, including the writing of the paper. AM provided the gene expression data. MK aided with the curation. SJ aided with the pain drug categorization. GN, DLR and BS provided supervision throughout.

Chapter 5

Jamieson, D.G., Oyeyemi, O.J., Sidders, B., Nenadic, G., Robertson, D.L. **Expanding the human pathogen interactome with text mining**. *In preparation* (2014)

DGJ was involved in all aspects of the work presented in this paper, including the writing of the paper. OJO aided with the curation. GN, BS and DLR provided supervision throughout.

In addition to the work presented in this thesis, various aspects of this PhD have been presented at conferences as posters and presentations. They are as follows:

- Jamieson, D.G., Dickerson, J., Gerner, M., Sarafraz, F., Nenadic, G. & Robertson, D.L. Automating host-pathogen interaction discovery: an HIV case study. ISMB, Vienna, Austria (2011) (poster).
- Jamieson, D.G., Robertson, D.L. & Nenadic, G. Task-specific Protein Tagging: an Experiment with BANNER on HIV-1/human interaction text. Languages in Biology and Medicine, Singapore (2011) (poster).
- Jamieson, D.G. Gerner, M., Sarafraz, F., Nenadic, G. & Robertson, D.L. Towards semi-automated curation: using text mining to recreate the HIV-1, human protein interaction database. Biocuration, Washington DC, USA (2012) (presentation).
- Jamieson, D.G., Robertson, D.L. & Nenadic, G. Contextualising and exploring human-pathogen molecular interactions through full-scale biomedical text mining. ISMB 2012, Long Beach CA, USA (2012) (poster).
- Jamieson, D.G., Roberts, PM., Robertson, D.L., Sidders, B. & Nenadic, G. Cataloguing the biomedical world of pain through semi-automated curation of molecular interactions. Biocuration, Cambridge, UK (2013) (poster and presentation).

- Jamieson, D.G., Robertson, DL, Sidders, B. & Nenadic, G. Exploring the biomedical world of neuropathic pain through molecular interactions. NeuPSig, Toronto, Canada & Pfizer Academic Forum, Cambridge, UK (2013) (poster).
- Jamieson, D.G., Robertson, DL. & Nenadic, G. Exploring the full spectrum of human-pathogen molecular events across the biomedical literature using text mining. ISMB, Berlin, Germany (2013) (poster).
- Jamieson, D.G., Moss, A., Kennedy, M., Jones, S., Nenadic, G., Robertson, D.L., Sidders, B. The pain interactome: Connecting pain-specific protein interactions. ISMB, Boston MA, USA (2014) (poster) & Pfizer Companion Talk Series, Cambridge MA, USA (2014) (presentation).
- Jamieson, D.G. Nenadic, G., Robertson, D.L. & Sidders, B.S. Ion channels and solute carriers: what do we already know? Pfizer Academic Forum 2014, Cambidge, UK (poster).

For consistency throughout this thesis we have changed the formatting of any published work, including any figure numbers and table numbers. Abbreviations appear as they did in the original published material. Original published articles are included in the appendices (A-C). Supplementary work is available online for published articles or otherwise available on the disk provided with this thesis.

Word count: 52194

CHAPTER 1

General Introduction

1.1 Motivation

Text-mining (TM) refers to the process of deriving knowledge from unstructured text. Much like data mining, it seeks to find interesting patterns from large bodies of data, except the key difference is that textual data is represented as natural language, designed and executed for the purpose of reading by humans, whereas in data mining facts are represented in a machine-readable form as databases. This distinction is decisive as both text and data mining are implemented through computational techniques, meaning that if TM is to reveal any knowledge from a textual body it must first process this into a more systematic form so that the barriers between human language and computational interpretation can be transcended.

Natural language processing (NLP) is often the term used to describe this process, and its origins can be traced back to 1950 in which Alan Turing proposed the 'Turing test' to gauge machine intelligence through natural language conversations[1]. While NLP's uses extend beyond processing raw text, for example in speech recognition or natural language generation, the vast majority of its tasks are directly relevant to TM. Using aspects such as named entity recognition (NER), part-of-speech (POS) tagging and parse trees, words and their individual meanings and relationships to each other can be labeled to decode computational meaning from otherwise abstract sequences of characters. It is only from here that data mining to reveal any knowledge of interest can then commence.

Because of the vast abundance of unstructured text used to communicate information in human society, TM has been applied across far-reaching disciplines, (e.g. in security, marketing and business intelligence), utilizing text

from a wide range of sources (e.g. online news outlets, social media, email, books etc.). This can facilitate knowledge acquisition for predicting complex phenomena, such as changes in stock prices[2, 3] or the spread of flu[4-6]. However, while common TM methodologies are applied across the board, the contrasting ways in which language is communicated across different fields often mean tailored TM approaches are required. For example, in social media (e.g. Twitter or Facebook) the use of slang and undocumented abbreviations are frequent and any TM systems implemented will be developed to account for this unique style. Reapplying these specific systems elsewhere, such as in government reports or historical publications, will thus be less effective as the more formal language styles are likely to be largely incompatible with social media.

Biomedical literature is no different with regards to the formality of language used, where the complex scientific writing style and use of unique concepts and terminology by trained researchers is designed to be read and understood by likeminded experts within their various sub-disciplines. This often presents a number of challenges in reading comprehension even for the intended scientific audience and for any successful biomedical TM system these must thus be overcome.

Publishing one's research is crucial for progress in biomedicine and presenting research from a study in the form of a journal style article for which other scientists can access represents something of an authentic end product to a scientific investigation[7]. This sharing of knowledge through publication is a fundamental cultural trait, so much so that career success (in science) hinges on the quality and quantity of these[8, 9]. As such, millions of articles have been published, many containing detailed descriptions of biological findings garnered through rigorous scientific study. Within this huge body of published knowledge there is thus useful data, which if unraveled and connected in a coherent and logical manner can be used to invoke new and compelling findings.

In this thesis we explore how TM can be used to exploit the published literature as a rich source of biomedical knowledge. We investigate what alternative methodologies exist to using TM for exploiting this knowledge and how these

can be merged to benefit each other. The range of data types stored within the biomedical literature is multifarious, so instead our chief focus will be on extracting and analyzing molecular interactions and how they have been contextually defined for studying disease. Human disease remains a critical aspect of biomedical research and comprehending its mechanisms so that it may be prevented, remediated and eradicated is often the major goal. The molecular interactions involved are vital in understanding these mechanisms for it is often at the molecular level that their pathologies become truly apparent[10, 11]. However, as we will see, this often gives rise to increasing levels of complexity and the ability to make biological sense of this data becomes more difficult when comparing diseases originating from multiple different pathogeneses. We thus restrict our efforts to exploring diseases and their molecular interactions between two disease classes: diseases caused by human pathogens and pain-related diseases.

The thesis introduction is structured to give an overview of human disease in general before exploring human-pathogen and pain related diseases more specifically. Following this we discuss how the published literature can be used to source knowledge related to these fields and in what ways TM can be of use. We then explore how any data acquired can be analysed and mined to draw interesting conclusions in the study of these diseases.

1.2 Human disease

Human disease can be defined as a specific illness or disorder characterized by a recognizable set of signs and symptoms attributable to heredity, infection, behaviour, ageing or environment[12]. Heritable diseases, such as cystic fibrosis, are passed on genetically from parents to kin; pathogenic organisms, like human immunodeficiency virus 1 (HIV-1) and *Escherichia coli*, cause infectious diseases; while environmental factors, e.g. smoking and poor diet, can also result in disease. The science of remediating, diagnosing and preventing disease, known as medicine, has existed for millennia, and has improved substantially in the last century. Thereupon, global life expectancy has increased on average 68.5%

between 1900 and 1990[13] and it is projected to continue rising at a steady rate for decades to come[14, 15].

Part of this rise in medical practice has come through a transformation in our understanding of human biology, from basic knowledge of the organs and tissues that make up the human body to being able to study the molecules that form these higher-level structures. Since 1953, when Watson and Crick first reached the conclusion that the DNA molecule is formed of a three dimensional double helix[16], we now have knowledge of the entire human genome consisting of an estimated 20-25 thousand (K) gene-coding proteins embedded in a genome of 4 billion bases[17, 18]. Each of these proteins then has their own role within various biological functions and together coordinate to make up complex living systems.

By understanding the full complexity of the human system we can then use this knowledge to decipher how they can be modified to correct underlying defects, such as genetic abnormalities, or prevent them from being manipulated by pathogenic organisms, both of which can manifest disease. In this way, pharmacologic treatment options are often designed to inhibit or activate biological entities such as proteins, DNA and RNA, known to play an implicit role in the pathogenesis of a disease[19, 20]. For example, statins are an effective treatment for atherosclerosis and cardiovascular disease through their inhibition of the HGM-CoA reductase (Figure 1.1). Inhibition of this enzyme means the mevalonate pathway cannot function and this leads to an overall decrease in low density lipoprotein deposits in blood vessels - the major cause of these diseases[21].

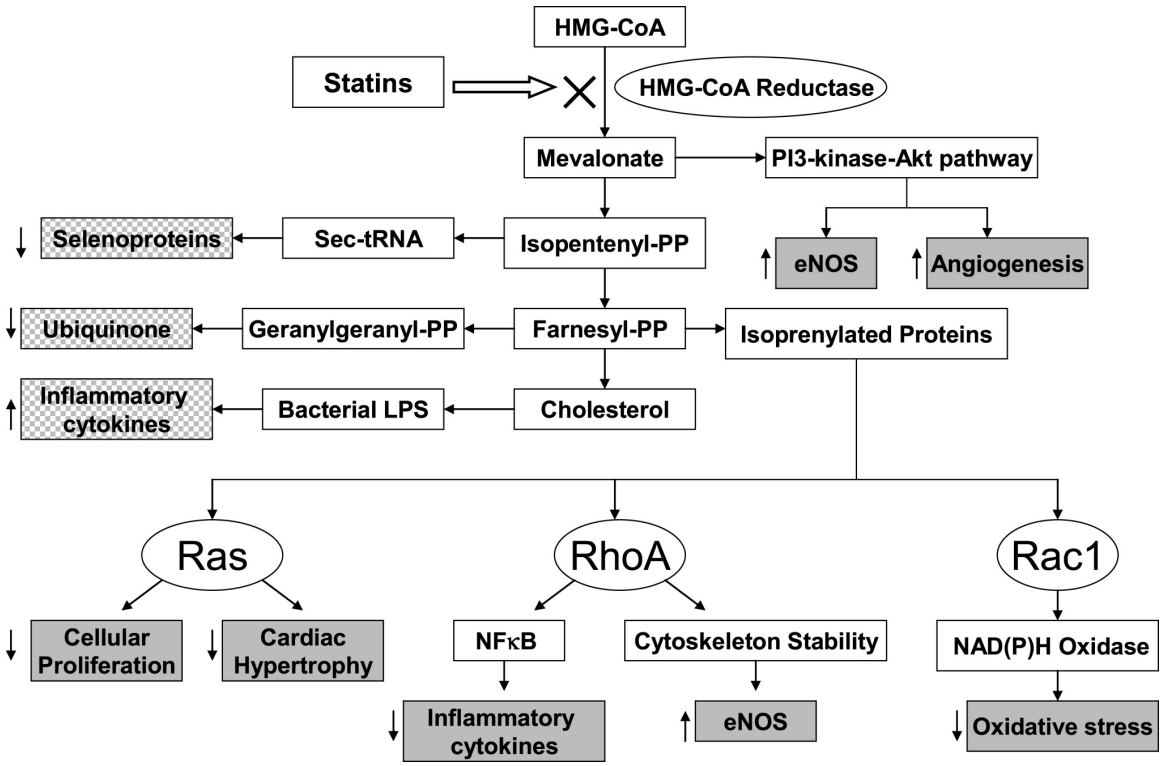


Figure 1.1 Mechanism of action of statins through their inhibition of HGM-CoA reductase. The inhibition of this singular coenzyme has multiple downstream affects. Boxes with grey backgrounds signify beneficial effects. Checkered boxes represent adverse effects. eNOS = endothelial nitric oxide synthase; HMG-CoA = 3-hydroxy-3-methylglutaryl coenzyme A; LPS = lipopolysaccharide; NAD(P)H = nicotinamide adenine dinucleotide phosphate; NFkB = nuclear factor kappa B; PI3 = phosphatidylinositol-3; PP = pyrophosphate; tRNA = transfer ribonucleic acid. Taken from Ramasubbu *et al*[22].

The full spectrum of human diseases is huge, ranging from the very rare, such as Dercum's disease (adiposis dolorosa)[23] or Fahr's disease/syndrome (familial idiopathic basal ganglia calcification)[24], to global pandemics like Malaria, HIV and tuberculosis[25]. The mechanisms behind different diseases are often contrasting, although various classes of diseases share common characteristics with each other. Cancer, for example, embodies over 100 related diseases and all are defined by their ability to cause abnormal cell growth with the potential to spread to outside regions of the body[26]. It is then within this shared feature that the individual pathogenesis of each disease can be studied for which unique medical solutions might be expanded across the full scope. For example, the upregulation of the hypoxia-inducible factor 1 (HIF-1) protein has been associated with increased mortality in cervical cancer, non-small-cell lung cancer, breast cancer, ovarian cancer etc., linked through its prominent role in the

transcription of genes involved in critical aspects of cancer biology. It thus represents a useful therapeutic target for multiple cancer-related diseases[27, 28].

This thesis seeks to investigate two broad disease classes of which understanding their mechanisms together might yield new insights into their overall biology and specific diseases within them. Firstly, we focus on human pathogens and their associated diseases. Human pathogens are all alike in that they produce infection in humans that can result in disease. To establish infection they or their products must interact with the human host at a molecular level and this separates them from other diseases whom the molecular mechanisms only involve human molecules. Secondly, we investigate pain and its related diseases. While pain is a natural part of the human defense strategy it can also persist in chronic disorders and understanding how these arise is important for the advance of therapeutics[29].

1.2.1 Human pathogens

A pathogen can be defined as a harmful agent causing infectious disease to its host. These include organisms from viruses, bacteria, fungi, helminthes, protozoa and prions, with around 1.4K individual species recognized as human pathogens[30-32]. Diseases arising from pathogens are common in humans, resulting in significant contributions to reduced mortality and morbidity across the globe (Figure 1.2).

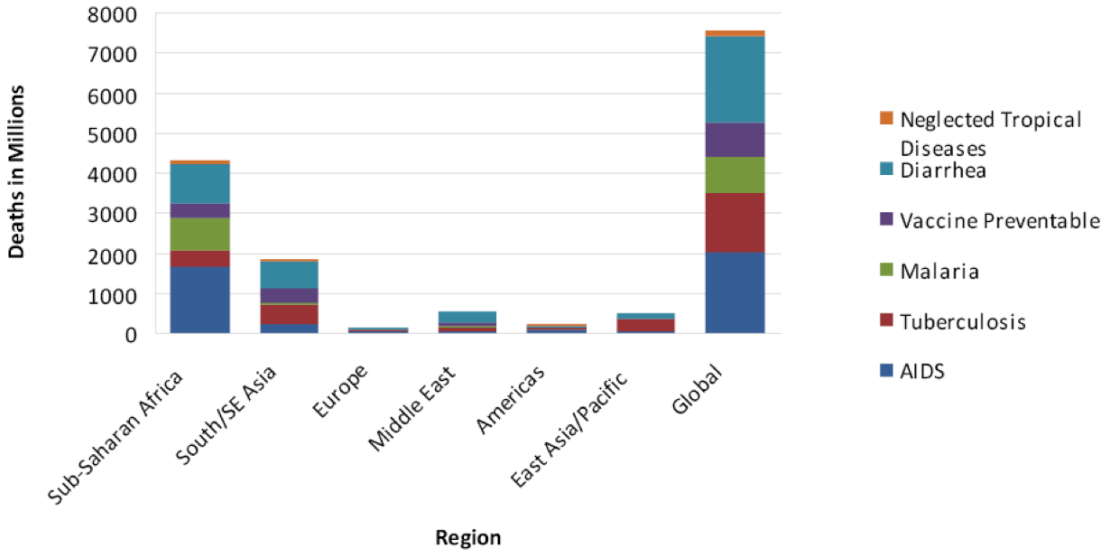


Figure 1.2 Global distribution of infectious disease.
Image sourced from http://www.smartglobalhealth.org/issues/entry/infectious-diseases

The role of the immune system in infection

In order for infection to ensue, a pathogenic organism (or their products) usually enters the cytoplasm of a host cell. It is here that they can survive, replicate and mediate their actions[33]. The strategies for achieving this vary across different pathogens, although they must all first overcome the body's natural defense mechanism, the immune system.

The human immune system can be divided into two largely interdependent arms: adaptive immunity and innate immunity (Figure 1.3). Innate immunity acts as a first line of defense, where host cells recognize the molecular signatures of foreign bodies through pattern recognition receptors, triggering an arsenal of defense mechanisms, e.g. additional immune cell recruitment and production of pro-inflammatory cytokines. Adaptive immunity is slower (typically three days to two weeks) and best characterized by B cell (B lymphocyte) and T cell (T lymphocyte) recognition of molecules unique to that infectious pathogen, termed antigens. This facilitates more specific responses to a particular pathogen, e.g. through the production of antibodies or cytotoxic T cells, and the formation of memory T and B cells provides protection against future pathogens carrying the same antigenic signatures[34].

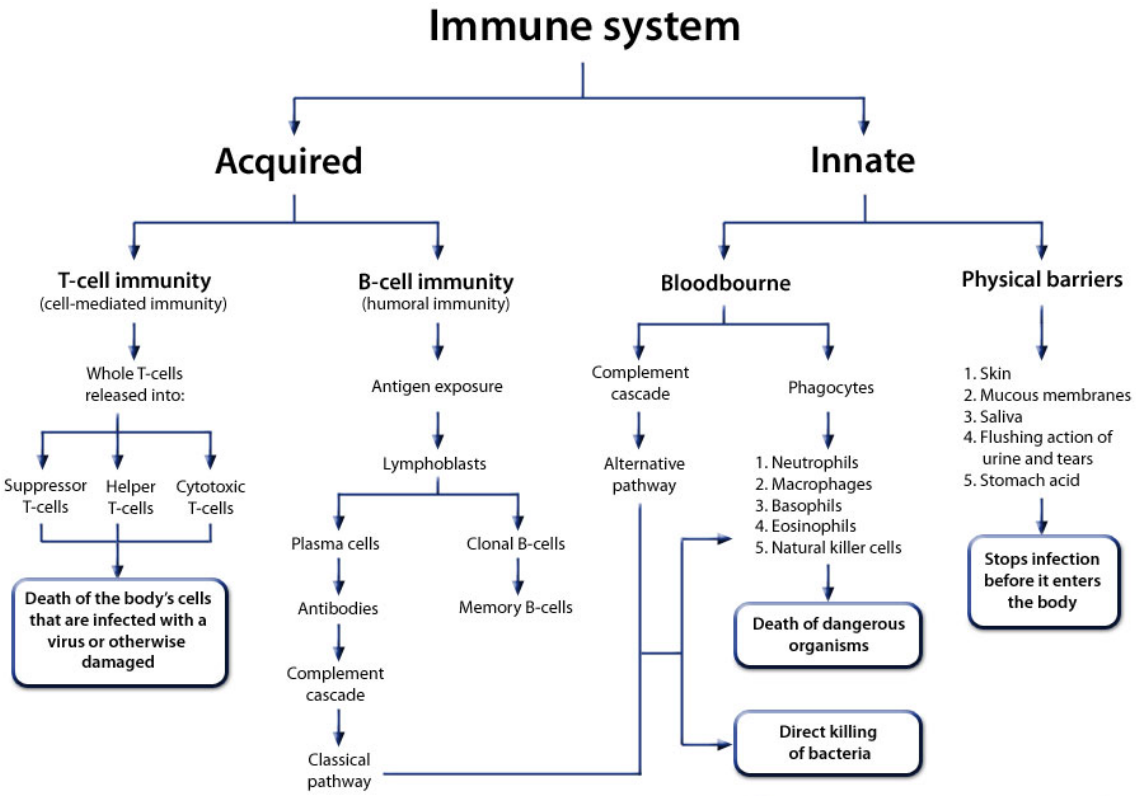


Figure 1.3 Innate immunity versus adaptive immunity.
Image sourced from http://www.myvmc.com/anatomy/human-immune-system/

Viruses

As pathogens, viruses are unique in that they are obligate intracellular parasites that cannot replicate independently[35]. There has been some debate as to whether they can be defined as living or otherwise[36-39], but regardless can be infectious, consist of complexes of biomolecules and share numerous biological characteristics with other microorganisms. Throughout history, pathogenic viruses have had huge negative impacts on human mortality and morbidity, e.g. through smallpox[40] or poliomyelitis[41], and continue to cause disease from a wide range of contrasting viral subtypes, e.g. single stranded RNA retroviridae[42] or double stranded DNA hepadnaviridae[43]. Figure 1.4 exhibits characteristics amongst viruses, with example species provided.

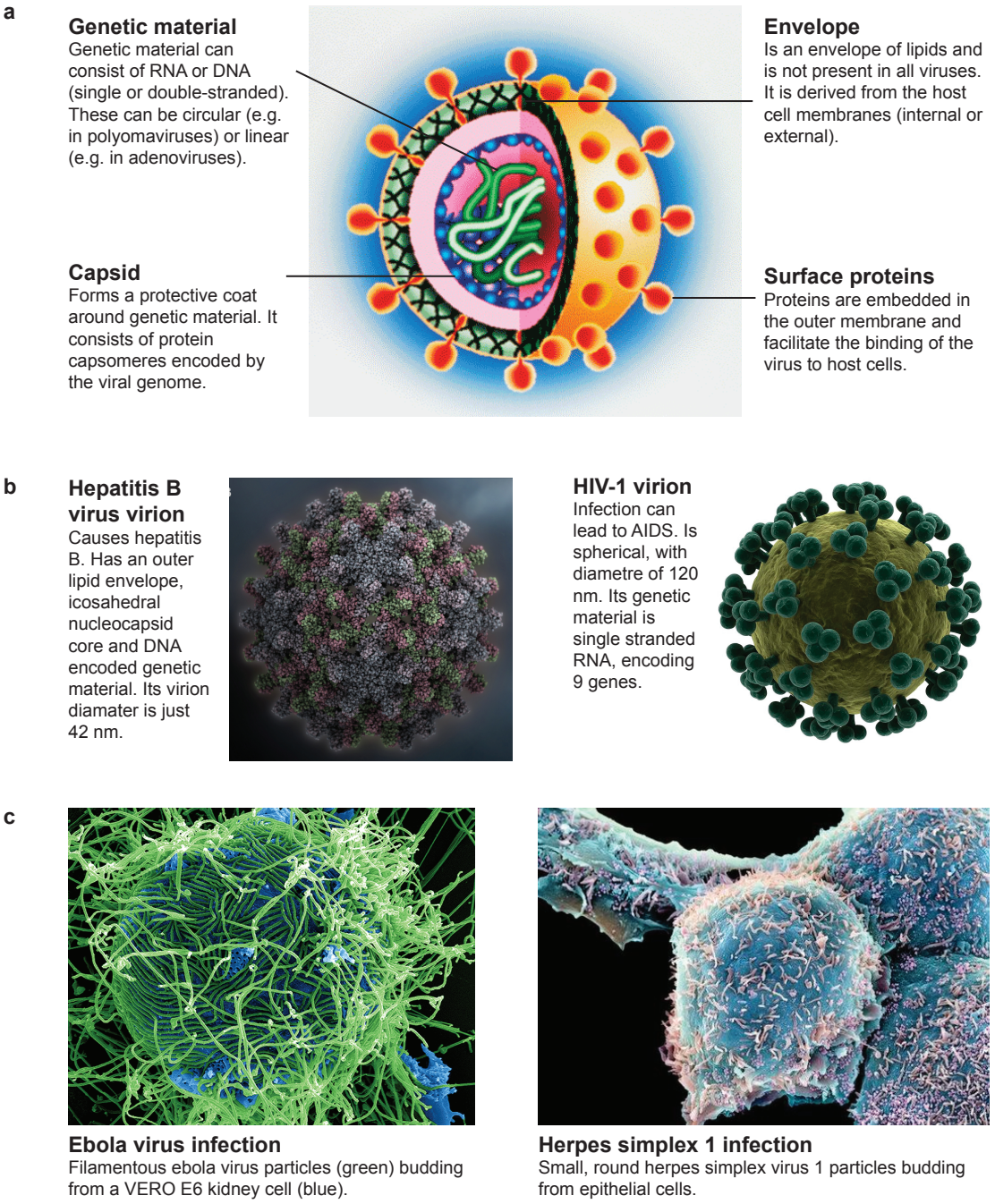


Figure 1.4 Characteristics among viruses.
a Structural characteristics common to all viral subtypes. Image adapted from http://www.microbiologyonline.org.uk/about-microbiology/introducing-microbes/viruses
b Examples of individual virion structures of pathogenic virues: Hepatitis B virus and HIV-1. Images sourced from http://wwwf.imperial.ac.uk/blog/studentblogs/bernadeta/2013/07/12/researching-hiv/ and http://www.virology.wisc.edu/virusworld/viruslist.php?virus=hpb
c Cells infected by viral pathogens: Ebola virus and Herpes simplex 1. Images sourced from http://cen.chempics.org/ and http://sciencephotolibrary.tumblr.com/post/32194695446/herpes-virus-coloured-scanning-electron

The typical viral reproductive cycle in humans follows a path of host cell attachment, cellular entry, uncoating, replication, virion assembly and virion release[44]. Attachment occurs through interactions of viral surface proteins and cellular receptors, for example HIV-1 gp120 interacts with CD4 and a coreceptor (e.g. CCR5)[45]. Cellular entry is then most commonly achieved through use of existing entry mechanisms, such as clathrin-mediated endocytosis, phagocytosis or macropinocytosis[33], or through fusion with the cellular membrane[46]. Once a virus has gained entry to the cytoplasm of a cell through one of these mechanisms, it then releases proteins from its capsid, in a process known as uncoating.

Prophylactic vaccines are key in ameliorating or preventing the effects of viral infection. These have eradicated smallpox and drastically reduced the incidence of polio across the globe in what have been described as some of the greatest achievements of modern medicine[47]. Antiviral vaccines work by exposing an attenuated or inactivated form of a virus to stimulate the adaptive immune system to produce antibodies that recognize specific antigens of that virus – ultimately providing immunization against future infection. However, due to the variation within each viral species these are often limited to particular viral strains and subsequently are not always a suitable preventative measure. For example, HIV is highly heterologous, where amino acid sequences of Env (the gene responsible for cellular binding) vary by as much as 20% within a clade and 35% between clades, and any immune response will have to be cross-reactive between these to be effective[48].

If preventative measures are not provided for a virus (whether a vaccine is available or not), antiviral drugs can be practical treatments for alleviating symptoms of viral infection. The strategies for these are either designed to target viral proteins or host cellular receptors. Antiviral drugs targeting viral proteins are typically more specific (akin to vaccines), and as a result are vulnerable to becoming ineffective due to the development of virus-drug resistance[49]. For example, the DNA polymerase of human cytomegalovirus is targeted by the drugs ganciclovir, foscarnet and cidofivor, but resistance develops following mutations in this gene[50]. While, antiviral drugs afflicting host proteins have

increased likelihood of toxicity, but have a broader activity spectrum and are less susceptible to drug-virus resistance[49]. This enables treatment of highly variable viruses, such as Maraviroc, which prevents HIV-1 entry by binding non-competitively to the HIV-1 co-receptor CCR5[51].

Bacteria

The occurrence of bacteria (and viruses) in the human body as components of a complex microbial community is a natural occurrence, where their numbers vastly outweigh that of human host cells and are vital for aspects such as nutrition and even pathogen resistance[52, 53]. They can thus be seen as a double-edged sword when they act as pathogens themselves by causing prevalent diseases such as Bubonic plague (by *Yersinia pestis*)[54] or non-typhoid salmonellosis (by *Salmonella enterica*)[55, 56]. Bacterial infections can be conditional, e.g. in infections through open wounds[34], and like viruses, can infect as obligate intracellular parasites[57]. Bacteria, as prokaryotes, are unicellular, their DNA is usually presented as a singular circular chromosome, they may contain plasmids (for DNA transfer)[58] and reproduce mitotically through binary fission[59].

The strategies intra-cellular bacteria use to infect host cells are often as complex as those used by viruses and have developed numerous approaches to avoiding immune subversion. For example, *Mycobacterium tuberculosis* and *Legionella pneumophila* are engulfed by phagosomes of the host, but prevent their maturation into phagolysosomes (which would otherwise digest them) to maintain a pH neutral environment for survival[60].

Most bacterial infections are typically treated with antibiotics, such as penicillin, gentamicin or erythromycin. These are often naturally occurring compounds produced by bacteria or fungi, which inhibit the growth of often a broad spectrum of other competing bacteria. They are therefore effective treatments against many bacterial pathogens, though when used inappropriately can lead to bacterial antibiotic resistance. Resistant bacteria pose a major threat to future medicine[61], as the number of antibiotic treatments available is limited and alternative treatments are currently unavailable. There is therefore a huge

demand for new therapeutic treatments to provide other therapeutic options should our current treatments become ineffective.

Fungi

Pathogenic fungi, also termed mycoses, are a major burden to human health[62]. The majority of fungal infections occur in immunocompromised patients[63, 64], e.g. by *Candida* or *Cryptococcus* species[65]. For example, *Aspergillious sp.* has been responsible for 10-15% of all deaths among transplant recipients[66]. Mycoses are often classified by their primary sites of colonization in the tissues and organs they infect, e.g. superficial fungal infections (outermost layers of the stratum corneum of the skin), dermatophyte infections (skin, hair, and nails), subcutaneous mycoses (subcutaneous tissues) and systemic mycoses (primarily the respiratory tract)[67]. Fungi differ from bacteria and viruses in that they are eukaryotes. Most grow in cylindrical structures, known as hyphae, and can become visible at the macroscopic level. They can reproduce both asexually and sexually, and spores ejected for colonization of new habitats, with mycelial cells, are a common cause of allergies in humans (e.g. allergic asthma, allergic sinusitis, hypersensitivity pneumonitis etc.)[68]. Treatment for fungal infections usually involves the use of antifungal agents such as imidazoles, triazoles and allylamines[69].

Helminthes

Helminthes are large multicellular parasitic worms, of two major phyla: nematodes (roundworms) and platyhelminths (flatworms). Soil transmitted nematodes cause the human diseases ascariasis, trichuriasis and hookworm, and filarial nematodes are responsible for onchocerciasis, loiasis and dracunculiasis. Platyhelminth flukes result in schistosomiasis and food-borne trematodiases, and platyhelminth tapeworms cause cysticersosis[70]. Together, one or more of these are said to currently infect over 1 billion (B) people in developing regions of the world[71-73]. Unlike other pathogens, helminthes do not replicate within the human host and symptoms of infection are localized to a single, albeit large, invading entity and the toxins and eggs it produces[70, 74]. Infection often impacts on disability resulting in disease driven poverty traps, although helminths can

persist asymptomatically after invasion[73]. Preventing infection through health education or improved hygiene are common means of reducing helminth disease incidence and helminthic treatments such as praziquantel are also available[75, 76].

Protozoa

The majority of pathogenic protozoa are mobile unicellular eukaryotes. Protozoa cause significant human diseases, such as malaria, leishmaniasis and African and American trypanosomiasis. Their prevalence is confined mainly to poorer global regions, although they can present in immunocompromised patients or as emerging diseases[77-79]. Antiprotozoal agents are used to treat infection, for example furazolidone, pentamidine and metronidazole[80-82].

Prions

Prions comprise of a much smaller class of non-living pathogens, defined only 32 years ago in the prion hypothesis[83]. While other hypotheses for their existence prevail[84, 85], the 'protein only' hypothesis proposes that small infectious pathogens lacking in nucleic acid contain proteins that exist as conformational isoforms of their normal host proteins in the outer surface of neurons. These prion proteins are indistinguishable from their human protein counterparts in their amino acid sequences, but are structurally dissimilar[86]. In a prion infection, human proteins in neural regions are replaced by these and they propagate across neural structures causing irreparable damage[87, 88]. The nature of this causal agent is yet unclear, although this pathogenesis would fit the human pathogen that causes Creutzfeldt-Jacob disease[89], as well as pathogens afflicting other animals[90].

1.2.2 Pain and its associated diseases

Pain in its normal function either acts as a protective physiological early warning system to avoid noxious stimuli or as an adaptive system to discourage physical contact and movement of an injured body region[91] (Figure 1.5). Disease can arise when these types of pain become chronic (e.g. in inflammatory pain-related disorders), arise from abnormal function of the nervous system (i.e. in

neuropathic pain[92]) or when they are completely absent altogether (e.g. in congenital insensitivity to pain[93]). Chronic pain affects one third of Americans, up to 30% of Europeans and 14.5-33.9% of people in developing regions[94, 95], causing significant distress and impairment to its sufferers[96]. It is thus a significant global health issue and warrants appropriate action to mediate its impact on society[97].

Neuropathic pain

The International Association for the Study of Pain defines neuropathic pain as pain caused by a lesion or disease of the somatosentory nervous system[98]. Neuropathic pain presents centrally or peripherally depending on whether the central or peripheral nervous system has been damaged and both are maladaptive, conferring no evolutionary benefit like other nociceptive pain types[95]. Central neuropathic pain can result from spinal chord injury[99], multiple scleorsis[100] or after strokes[101] among many other predispositions[102]. Peripheral pain can be caused by traumatic injury[103], diabetes[104] and HIV[105] among others. The treatments available for neuropathic pain are diverse, and therapeutic options include the use of opioid antagonists, calcium channel ligands and trycyclic antidepressants. However, these are typically ineffective and unable to target the underlying mechanisms precisely, with only a 30% reduction in pain described as a clinically meaningful result[102]. At a molecular level, the research into neuropathic pain is extensive, although it is still not fully understood and further efforts are needed to help uncover these so that future pharmacologic design can exploit this knowledge.

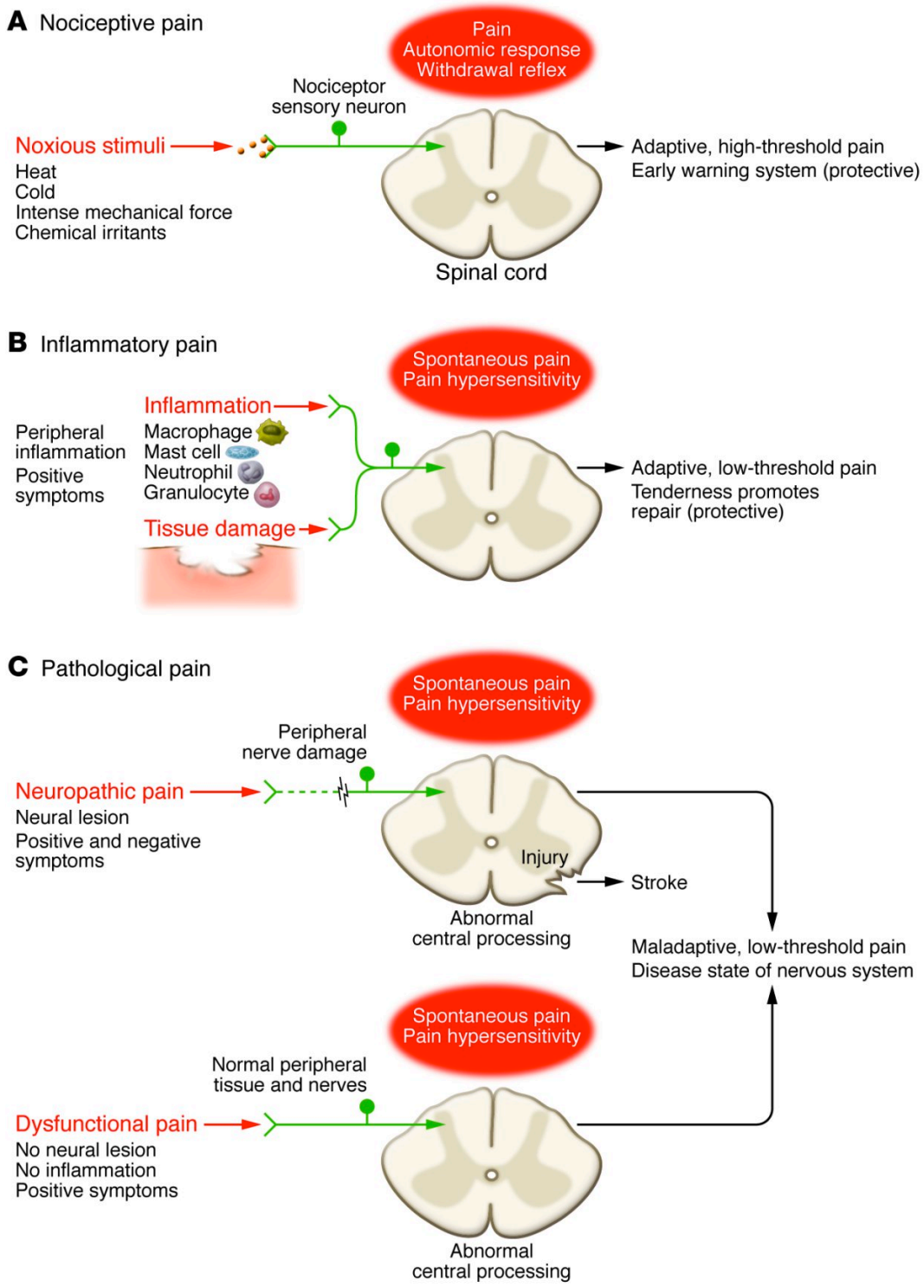


Figure 1.5 Pain classifications
(**a**) Nociceptive pain. This is designed to act as a physiological early warning system against noxious stimuli. (**b**) Inflammatory pain. Discourages physical contact of an injured body region. (**c**) Pathological pain, defined by two major pain types: neuropathic and dysfunctional pain, both characterized by a malfunction of the somatosensory nervous system. Image provided by http://pharmaceuticalintelligence.files.wordpress.com/2012/06/painclassfication.jpg

Inflammatory pain

Chronic inflammatory pain disorders result when inflammation persists over a long period of time, becoming exaggerated or inappropriate to the underlying tissue damage that caused it[95]. Notable causes include arthritic diseases (e.g. osteoarthritis and rheumatoid arthritis)[106], Crohn's disease[107] and injuries (leading to chronic disorders such as tendinitis)[108]. Each disorder is complex in its own right[109], and while chronic inflammatory pain can be treated with similar remedies to acute inflammatory pain, such as opioids and non steroidal anti-inflammatory drugs[95], these are not universally effective and pose potential significant aggravations to patient quality of life[107, 110]. As with neuropathic pain, new treatments are therefore in demand that reduce these side effects and again an improved understanding of the molecular mechanisms behind inflammatory pain will be beneficial for meeting this goal.

1.3 The published literature: a wealth of data

William Black in his 1782 publication[111] once pondered the value of medical publications raising concerns on the efficiency of handling and identification of knowledge from the 'immense pile of books' that existed through 'a revolution of ages and empires'[112]. These timeless words still underpin the modern day issues of a knowledge explosion in the world of biomedical publishing, where researchers are largely dependent on using literature as the main vehicle in which to publish their research findings and to access the work of others[113]. In this section we will explore these problems in more detail and what solutions have been proposed.

1.3.1 PubMed

In today's digital age, the majority of all peer-reviewed biomedical journal articles are available for researchers to access online. One of the largest and most widely recognized stores of these citations is that of the National Library of

Medicine's (NCBI) PubMed. Since its inception in 1996, PubMed now stores over 25 million (M) citations, including over 21M abstracts and titles from MEDLINE and over 3M open access full text articles linked in PubMed Central (PMC). Between 2 and 4K new citations are added each week, including the addition of publications from as early as the 18th century, across over 5.6K worldwide journals[114, 115]. Figure 1.6 shows the growth of citations in PubMed since 1990.

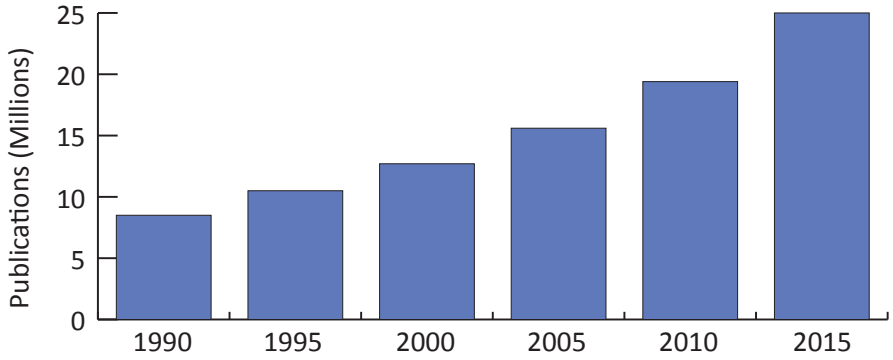


Figure 1.6 Total literature citations in PubMed since 1990.

The majority of citations in PubMed are presented with a title, an abstract and a link to a full-text source for that article. Other information, such as the contributing authors, their institutions, the journal and year published provide additional context. Furthermore, many articles are annotated with Medical Subject Headings (MeSH), which provide summary categories for indexing. Together, this data can be used for searching through citations to find relevant publications.

Retrieving relevant publications

As a result, PubMed and indeed other biomedical citation websites (e.g. Google Scholar and ScienceDirect) are often seen as excellent starting points in the biomedical information retrieval (IR) process[116, 117]. For example, a researcher can search for articles related to 'depression' on PubMed and will have instant access to 310K citations (Aug, 2014). However, as with this example, a number of problems arise when the number of relevant publications from a search exceeds the researcher's ability to read through the majority of relevant results.

Firstly, 'depression' is lexically ambiguous, where it can refer to a region of depression, the act of depressing, the state of being depressed or as perhaps intended, a condition of general emotional dejection and withdrawal. Biomedical terms are commonly like this[118], with 11.7% of phrases in biomedical abstracts determined to be ambiguous relative to the Unified Medical Language Thesaurus (UMLS)[119, 120]. Consequently, the precision of search results can be reduced by 'false positives' often caused by ambiguous terms[121, 122].

It is seldom the case that a search term is not synonymous with other terms[123], as is affirmed by the numerous biomedical thesauri (e.g., UMLS[120], National Cancer Institute Thesaurus[124], BioThesaurus[125] etc.) and ontologies (Disease Ontology[126], The Protein Ontology[127], Gene Ontology (GO)[128], etc.) that exist. Melancholia, unipolar disorder and depressive are some of the terms that can more directly refer to depression. While, indirect synonyms such as despondency, abjection and gloominess may also refer to the biomedical form of depression. Thus, to achieve the full scope of results available, the search mechanism must account for these and their own individual ambiguities.

However, even without perfect recall in IR, researchers like in our example are presented with insurmountable numbers of citations to review. Dogen *et al* have demonstrated that over a third of PubMed searches retrieve more than 100 citations[129]. With large numbers of citations it would be unnatural for any researcher to spend their time and effort scrutinizing each result with equal clarity. What is often a more convenient alternative is to adopt a limited selection process where only a handful of the total number of citations are reviewed. This is influenced primarily by the ordering of the results, where 80% of abstract views occur in the top 20 citations[129]. To compound this problem, Nourbakhsh *et a* [121] have demonstrated that users only found 67.6% of abstracts retrieved from the top 20 search results in PubMed to be useful.

To improve the precision and recall of biomedical IR there are, however, more advanced biomedical search systems that exist. For example, RefMED[130] allows users to rate the relevance of individual search results and then use this feedback to generate more personalized IR. In PolySearch[131], publications can be searched

using user-defined patterns of queries selected from 50 different classes. GoPubMed[132] and GO2PUB[133] group search results together by highly related Gene Ontology (GO) and MeSH concepts to allow users to select the groups of documents of interest.

While the existing publication search tools can be useful for researchers seeking background information on their topics of interest, it is often not possible to fulfill more complex requests. For example, a comprehensive list of protein-protein interactions (PPIs) for a given disease under specific contexts will be difficult to obtain using even the most advanced search systems. Furthermore, the researcher will most likely wish for this information to be represented in a structured format to save time locating and extracting the data from the text.

1.3.2 Manually curating literature into structured databases

To help address these issues in biomedical IR, the reorganization of the literature content into topic-specific structured databases has become a prominent feature in permitting researchers quick and targeted access to published data. Many of these focus primarily on storing the more valuable information from the text, such as PPIs (Table 1.1), often with added layers of context to aid the intended users. This allows a researcher to, for example, source an entire catalogue of protein interactions mediated by the phosphoprotein-binding domains[134], which would otherwise be impractical using standard search websites.

Name	Description
IntAct[135]	454,515 binary interactions (Aug 2014) are provided by literature curation or user-submissions from 11 different databases.
HHPID[136]	PPIs between HIV-1 and humans are selectively curated from the literature. Last updated June 2014.
BioGRID[137]	749,912 protein and genetic interactions from major model organism species curated from 43,149 publications (v3.2.115).
MPIDB[138]	24,295 microbial interactions curated from the literature or imported from external databases (2009-11-18 release).
DD database[139]	Comprehensive PPI data on death domain superfamily proteins.
MIPS[140]	Mammalian PPIs manually curated from the literature.
CORUM[141]	Manually annotated PPIs specific to mammalian organisms.
DIP[142]	Manually and computationally derived PPIs.
HPRD[143]	Over 36,500 unique PPIs from 25,000 proteins.
InnateDB[144]	23,779 interactions curated from 4,889 publications specific to the mammalian innate immune response.
VirHostNet[145]	Virus-host PPIs from public databases and curated data from the literature.

Table 1.1 Example protein-protein interaction databases manually curated from the literature.

The curation of databases from the literature is typically performed manually by highly specialized curators, who read through individual articles and input any relevant data into specific databases. The process is usually accurate in extracting and inputting the data into databases correctly and moreover the quality can be monitored by the inter-annotator agreement, e.g. through using Cohen's kappa coefficient[146]. This overall process often gives rise to high quality databases, however, the manual process is notoriously slow[147, 148]. For example, the HIV-1, Human Protein Interaction Database (HHPID) took 7 years to curate 2,589 unique PPIs from 3,200 articles[136, 149]. To stay up-to-date it has since required updating and this is of course a regular facet of database curation from the literature. Manual curation is flawed in this sense, as keeping pace with the existing and ongoing growth of primary literature often remains a challenge too costly and unfeasible even for those biomedical topics in highest demand.

Researcher-led curation

One solution to biomedical data curation is to adopt a system whereby researchers themselves manually input any key data from their research into

associated databases as a part of the publication process[150, 151]. To facilitate this, many topic-specific databases are available (e.g. many of those in Table 1.1), although the onus of submitting data from publications often lies with the publisher and is exerted only as an assurance of quality and not conformity[152]. Moreover, even if researchers were to strictly adhere to submitting all data from their work to their relevant databases it would still leave the existing body of literature untouched, requiring appropriate annotation.

In an extension to this paradigm, 'crowdsourcing' has been experimented with as a method of facilitating curation of the biomedical literature. Systems such as TagCurate, for disseminating biomedical annotations[153], or Amazon Mechanical Turk, for validating mutations related to specific genes[154], have mobilized communities of researchers to curate literature. However, while these have shown some success, they are often reliant on incentives for which are not easily provided and scaling these up to cover the entire published literature is unrealistic.

1.3.3 Text-mining in biomedical literature

As an alternative to manual curation, text-mining (TM) techniques have shown great promise in being able to extract and organize information from the literature into biological databases automatically. Despite its relatively short existence from the end of the 20th Century[155-158], the growth and progression of TM in biomedicine has been quite remarkable. However, it is by no means a 'complete' discipline and it has a number of challenges that compound its use. In this section we will explore the various ways in which TM has so far been instrumented to extract biomedical data from published articles, with particular reference to protein-protein interactions (PPIs). Some of these will have already been partially visited in relation to IR using article website search engines, although they will now be probed in greater detail.

Evaluating the performance of text-mining

Before we examine the various ways in which TM can be used to extract biomedical data from the published literature we first discuss how it is commonly evaluated. TM software often produces incorrectly retrieved results, known as false positives (FPs), while correctly retrieved results are termed true positives (TPs). Data can also be missed, called false negatives (FNs), and any remaining data not retrieved are designated true negatives (TNs). Measuring TPs, FPs, TNs, and FNs can be done manually by comparing the result set against the text from which they were extracted. However, it is more common to construct a corpus for which all of the correct mentions are annotated in the text beforehand, so that the results of the application can be automatically compared and validated. 'Gold standard' corpora are often described as those constructed manually, whose annotations are of the highest quality and are thus less likely to be erroneous. However, 'silver standard' corpora can also be constructed automatically to sidestep the time-consuming and costly process of manual annotation, by using high quality data often provided by a composite set of annotation services[159]. Once these figures have been calculated, it is then possible to evaluate the precision and recall of the results.

$$\text{Precision} = \frac{TP}{TP + FP}$$

$$\text{Recall} = \frac{TP}{TP + FN}$$

Precision represents the portion of retrieved instances that are correct, while recall represents the portion of all possible correct results that are retrieved. Given precision and recall are commonly judged together to reflect the overall performance of the text-mining software, an overall measure of the two is also often required. The F-score (F) is most commonly applied here, as the harmonic mean of the precision (p) and recall (r). Precision and recall can be weighted equally, or either can be favored by adjusting β (1 when balanced equally).

$$F = \frac{(\beta^2 + 1)pr}{\beta^2 p + r}$$

Other overall measures of precision and recall have also been used. The area (A) under the interpolated precision/recall curve (AUC iP/R) has been used to measure the quality of result ranking, which tends to favour recall. It is defined as,

$$A(f_{pr}) = \sum_{j=1}^{n} (p_{ij} * (r_j - r_{j-1})),$$

where p_i represents the interpolated precision and is defined as

$$p_i(r) = \max_{r' \geq r} p(r')$$

In this evaluation, provided its recall is high and the TPs are ranked correctly, a system can achieve a high score despite having a large number of FPs and thus lower precision. For result ranking this can be a more favourable evaluation in contrast to the traditional F-measure, where the overall number of TPs can be lower but still produce a high F-score if the precision is high[160]. The Threshold Average Precision (TAP-*k*) metric has also been used to measure ranking in a similar manner, but is less biased towards high recall systems[161, 162].

When TNs are known and should be taken into account (precision and recall do not use these), accuracy and Mathew's Correlation Coefficient (MCC)[163] are useful alternative measures of performance. Accuracy is the measure of the combined specificity and sensitivity of the results. Specificity (the true negative rate) can be used to measure the proportion of results correctly not retrieved, while sensitivity (the true positive rate) is the same as recall[160].

$$\text{Accuracy} = \frac{TP + TN}{TP + FP + FN + TN}$$

$$\text{Specificity} = \frac{TN}{TN + FP}$$

$$\text{Sensitivity} = \frac{TP}{TP + FN}$$

The MCC score is similar to accuracy, but it can be more reliable in that it is unbiased by sets where the two classes have varying sizes. A score of -1 to 1 is afforded, where 1 represents a perfect result set, 0 an average or random result set and -1 an inverse classification of results[160].

$$MCC = \sqrt{\frac{\chi^2}{n}}$$

$$MCC = \frac{TP * TN - FP * FN}{\sqrt{(TP + FP)(TP + FN)(TN + FP)(TN + FN)}}$$

Shared tasks

Shared tasks have emerged as a crucial driver for the development of TM systems to automatically extract key entities and their relations. In these, well-defined tasks are coordinated to allow participating teams to assess the performance of their systems against each other. For example, the BioCreative shared tasks have previously defined tasks for the extraction of genes/proteins, chemical entities and assisted curation among others[160, 164-167]. Other notable shared tasks include the BioNLP shared tasks[168-170], focusing on initiatives such as event extraction and concept recognition, and the Drug-Drug Interaction Extraction (DDIExtraction) shared tasks[171, 172], which sought systems for mining drug-drug interactions.

Teams are commonly given development and test corpora to fine-tune their systems for the task. They then submit their final systems and these are evaluated against a new corpus and their results are ranked against each other. This allows for more fair assessments of the quality of TM software for any given

task, as well as providing comparisons of the performance of different types of approaches. Typically the results from different tasks vary widely. A 'high' result for precision and recall can therefore be largely a relative term within each task, although those that approach comparable levels of quality to human curation are generally considered excellent[173].

Tokenization

Often the initial step in NLP involves breaking up the text into paragraphs, sentences, words, syllables and other meaningful elements, before other more sophisticated components are employed. This is often termed tokenization, and in TM it more frequently refers specifically to dividing text into words and sometimes sentences[174]. For other proceeding TM components this stage is imperative for matching entities and assigning semantic relations between words in textual zones.

Tokenization strategies vary across biomedicine, often depending on the goals of the software[175], and many challenges exist in order to achieve adequate results. For example, determining sentence splits in biomedicine is made difficult by the ambiguous use of periods, such as in abbreviations (Dr.), figure numbers (1.14) or in molecular nomenclature (Nav1.8). In word boundary determination the use of hyphens can complicate the process, where in some cases the words should be separated (e.g. text-mining) and in others they should be retained (e.g. down-regulate). However, in comparison to more complex text-mining tasks tokenization is typically considered more simplistic[176, 177], and generally the more rigorous systems in biomedicine produce high accuracy[178].

Part-of-speech tagging

Once the individual words have been derived from a block of text it is useful to determine their lexical categories or POS tags. Basic lexical items, such as adjectives, verbs or nouns are common POS examples; although more than 80 various classes have been described[179]. The process of POS tagging is complicated by words having multiple possible word classes, e.g. [example]. Thus, often to achieve accurate POS tags, POS taggers are optimally designed according to the

textual style. For biomedical text, example POS taggers include the GENIA tagger[180] and the MedPOST tagger[181], and are highly accurate (e.g. F score).

Syntactic Parsing

Syntactic parsing refers to the analysis of the syntactic structure of a sentence. At the sentence level, this can involve determining the relationships between phrases. For example, in the sentence 'DNA amplification of *p53* gene by PCR' PCR relates to DNA amplification and not the *p53* gene. However, resolving these syntactic structures are non-trivial through computational methods, mainly due to the variety of ambiguity involved[179].

Another method of parsing involves detecting the individual relationships between words, known as dependency parsing. In this form of parsing, dependency trees can be used to represent how each word relates to each other within a phrase (Figure 1.7).

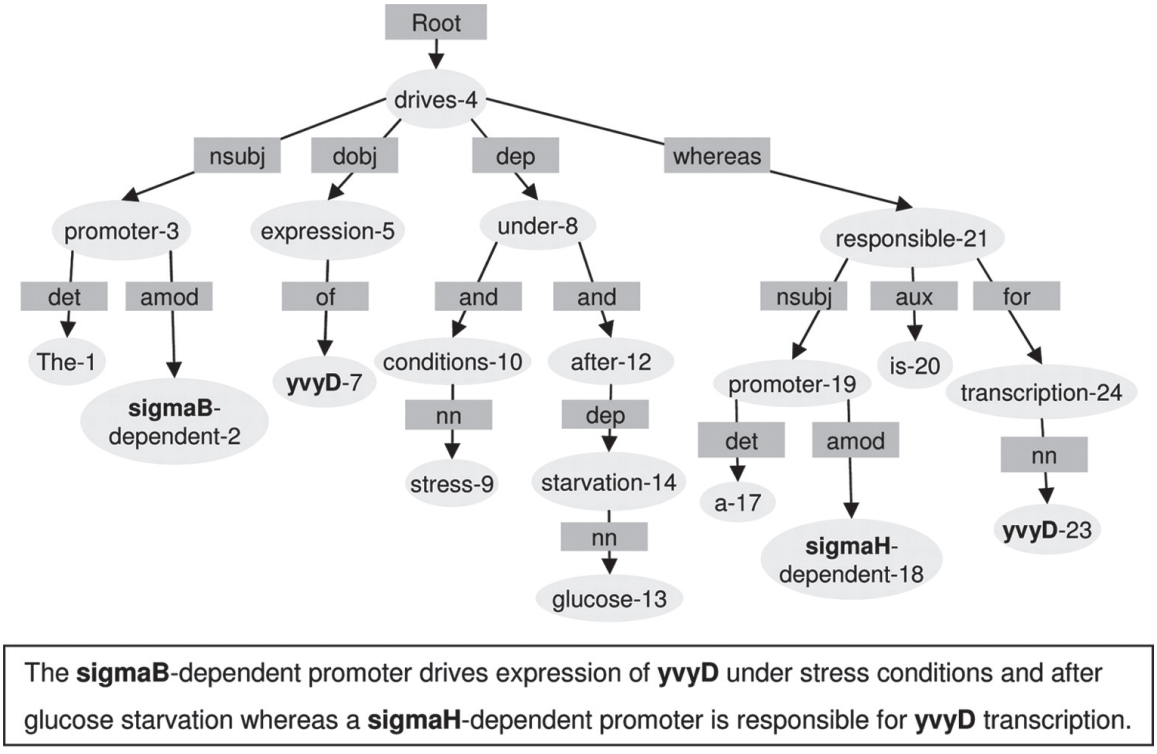


Figure 1.7 Dependency parse tree.
The tree was created using the Stanford Lexicalised Parser. Words are represented as ellipses provided with their positions. Dependency types are represented by rectangles except for the root of the sentence at the top of the tree. Image adapted from Fundel *et al* (2007)[182].

Named Entity Recognition

Named entity recognition (NER) concerns identifying mentions of specific semantic types in text[179]. In biomedicine, NER has been applied to a huge range of entity classes, broadly to aspects such as diseases, chemicals, genes and proteins, and to more specific groupings such as cell lines and cell types (Table 1.2). It is therefore a fundamental step in being able to extract and organize biomedical knowledge from the literature.

Entity type	Example systems	Description	Quality
Chemicals	ChemSpot[183] and OSCAR[184].	Entities can be trivial names, drugs, abbreviations, molecular formulas and International Union of Pure and Applied Chemistry denotations.	87.4% F-score in the BioCreative IV chemical entity recognition task[185].
Genes, proteins and RNA molecules.	BANNER[186], GenNorm[187], Abgene[188] and ABNER[189].	Protein, gene names and RNA molecules across all species can be extracted.	87.2% F-score in the BioCreative II gene name recognition task[190].
Species	LINNAEUS[191] and SR4GN[192]	Species name recognition, including official names, common names and others.	LINNAEUS shows 94% recall and 97% precision against their own corpus[191].
Mutations	MutationFinder[193], tmVar[194] and Open Mutation Miner[195].	Genetic variants (mutations) can be extracted. The possible range of these varies between systems.	tmVar shows 91.4% F-score against their corpus and 93.9% against another gold standard[194].
Pathways	PathNER[196]	Biological pathway mention recognition.	PathNER shows 84% F-score against GENIA derived gold standard corpus[196].
Diseases	DNorm (utilizing BANNER[186])[197]	Disease name recognition utilizing MeSH and OMIM disease terms.	DNorm shows 78.2% F-score against the NCBI disease corpus[197].
Cells	CellFinder[198, 199]	Dictionary based matching of cell lines and cell types.	Exact cell line matches showed 33-61% F-score; exact cell type matches showed 15-86% F-score[199].

Table 1.2 Example biomedical NER applications

The precision and recall of these systems has varied markedly across different entity types. This can be attributed partly to the linguistic characteristics of the desired entities as a consequence of the predefined naming standards and their adherence by the biomedical community[200]. For example, in gene/protein recognition, there are millions of different official names, synonyms and symbols in use[201], with new ones added all the time to databases like Entrez Gene[202] and UniProt knowledgebase (UnitProtKB)[203]. These are often ambiguous and can have meaning in several other concepts[204], such as the gene PCR (Gene ID: 952244) whose gene symbol is also commonly used to refer to 'polymerase chain reaction'. To compound this problem, researchers frequently invent their own protein terms in publications that do not follow the standard annotation guidelines, if any, and keeping track of these can be difficult[186]. Thus, NER in biomedicine is a complex task and it has led to the development of a number of different approaches, many of which are used in tandem.

In general, NER methods can be divided into three broad categories, 1) dictionary matching 2) rule-based and 3) machine-learning (ML) approaches. Dictionary-based approaches are where textual entities are matched to dictionary mentions. The successes of these are highly dependent on the plenitude of the dictionary and the ambiguity of the entities required. To achieve 100% precision and recall, every name for that entity class must be known and they must have only been mentioned in the context of that field in the form they appear in the dictionary. These are unlikely, but classes such as species names come close, with the system LINNAEUS demonstrating 94% recall and 97% precision for retrieving species name mentions[191]. The majority of NER tools, however, are not afforded the same privileges, with classes such as gene/protein recognition containing a much larger number of ambiguous terms and unaccounted dictionary entries[205].

One way to improve on dictionary-based matching is to combine these with rule-based approaches, which employ rules to linguistic inputs. For example, a specific rule might be to filter out any protein matches if they are immediately followed by 'disorder', i.e. 'the BMD *disorder*' to reduce false positives. Thus, in order to be successful, a huge number of rules often must be coded and this is

often difficult for the more ambiguous entity classes. Still, successful systems have been created such as Abgene, which has shown 66.7% recall and 85.7% precision for gene/protein tagging and has been described as one of the best of its type[206].

However, those NER systems that often perform best for matching difficult entity classes are those integrating ML approaches. ML approaches utilize training data and feature sets to perform classification of entities based on statistical confidence. Training data consists of a corpus of pre-annotated entities from a text. Features are then used to assign characteristics to these entities, generally including typographical aspects (e.g. ending with a number or not or having mixed cased characters), or features of the surrounding text (e.g. their part-of-speech (POS) tags)[179]. This data can then be used to create a model for which new text can be labeled for entities of that class based on its properties. If the features have been defined carefully, entities can then be matched predictively to a class according to the presence or absence of these features in text. BANNER, for example, utilizes an ML approach based on conditional random fields (CRFs), and performs particularly well for gene/protein tagging, with precision of 85.1% and recall of 79.1%[186].

As well as CRFs, many other discriminative models have also been applied to biomedical NER, such as support vector machines[207], maximum entropy[208] and semi-supervised learning models[209]. Furthermore, unsupervised methods that do not use training data have also demonstrated that they can be useful for tagging entity classes. For example, interaction terms have been identified using a pattern-clustering algorithm[210].

Entity normalization

Once an entity from an entity class has been retrieved it is then often necessary to normalize this to a specific identifier of that entity, e.g. linking a gene name to its Entrez Gene ID. This is an important step for further data analysis as it removes any remaining ambiguity between entities of the same class. For example, the gene symbol *ARP* can refer to several different genes within humans and also across several different species.

Normalisation can be divided into two overall stages, 1) entity mapping and 2) disambiguation. In the first step, candidate database identifiers are linked to the entity match. These are usually assigned according to the similarity of that entity name to a potential database identifier[211]. For example, if the gene name '*p53'* is denoted in the text, a system might map all gene records that have the term '*p53*' associated with it. The matching of the database terms to the entity match can be done exactly or with approximate matching, e.g. minimum edit distance[212], Jaro and Jaro-Winkler distance[213] or Dice coefficient[214]. Approximate matching is particularly useful for matching terms when the associated lexicon is likely to be incomplete or the variability of names is high[211].

Once associated database records with the entity have been identified, the relevance of these to the mention in the text must be decided. In many entity classes each mention should only be mapped to a single record, such as its unique CHEMBL ID[215] or disease ontology ID[126]. Thus, when there are multiple records, disambiguation must ensue. Hu *et al* demonstrate how contextual features of the surrounding text, such as GO annotations and species, can be used to assign the correct identifier[211]. Other gene/protein normalisation systems adopt similar approaches, e.g. GNAT[216] and GenNorm[187].

As with NER, the difficulty of gene normalization varies according to entity class. Again, gene normalization, particularly when it is applied across all species, proves to be more problematic. This is perhaps exemplified in the shift towards result ranking and the use of TAP-*k* as a means of testing the performance of gene normalization in the more recent BioCreative tasks (see 'evaluating the performance of text mining')[162]. In this regard, the expectation for achieving high enough precision and recall in gene normalisation automatically is low. For example, GenNorm one of the top performing gene normalization systems recently only demonstrated 38.1% precision and 26.9% recall against the BioCreative III corpus[217].

Relationships between entities

To fully comprehend the meaning of entities denoted in the text, it is necessary to deduce the relationships between these. On a basic level, the frequent co-

occurrence of entities in a body of text can often mean some form of relationship exists between them[218], e.g. a cell and protein constantly featured in the same sentences. However, clarifying what this relationship is can be more troublesome and often requires other entities (e.g. biomolecular events) and linguistic features (e.g. verbs). Approaches such as pattern-based, ML, statistical analyses and formal inference can then use this information to predict the true relationship between entities[218-220].

One large area of focus for the biomedical TM community has been the extraction of biomedical events that can be linked to gene mentions. The BioNLP '09 shared task defined nine event types using the GENIA ontology (Table 1.3). Non-regulatory events have at least one gene/protein theme; binding events may have one or more. Regulatory events are special in that they can have one gene/protein theme or another biomedical event. Causal gene/proteins are an additional optional property of regulatory events. Together, these can be used as 'single events' (e.g. gene expression of *p53*) or, when combined, can form 'event chains' (e.g. *Muc16* negative regulation of phosphorylation of *IL16*). Event chains are useful as these can be used to represent molecular interactions and, albeit less frequently, interactions involving more than two participants.

Type	Primary Args.	Second. Args.
Gene_expression	T(P)	
Transcription	T(P)	
Protein_catabolism	T(P)	
Phosphorylation	T(P)	Site
Localization	T(P)	AtLoc, ToLoc
Binding	T(P)+	Site+
Regulation	T(P/Ev), C(P/Ev)	Site, CSite
Positive_regulation	T(P/Ev), C(P/Ev)	Site, CSite
Negative_regulation	T(P/Ev), C(P/Ev)	Site, CSite

Table 1.3 Biomedical event types used in the BioNLP '09 shared task.
Taken from Kim *et al* (2011)[168].

The BioNLP shared tasks have attracted a myriad of competing teams and many of the systems developed are now available for use with varied perfomance[168].

For example, The Turku Event Extraction System (TEES), shows precision and recall of 56.3% and 46.2% respectively[221].

Negation and speculation

In linguistic terms, negation marks the absence of an entity or event[222], e.g. 'there was *no* up-regulation of CD4'. Speculation (also termed hedging[223]), marks the contemplation of an entity or event, e.g. 'we *hypothesize* the presence of T cells'. These linguistic traits are common in biomedical text, with 12.7% and 19.4% of sentences reported to contain negation and speculation respectively[224]. Therefore determining their presence is an important layer of context in addition to matching other entities and events.

Negation and speculation systems are often designed to detect cues, e.g. 'not', 'no', or 'without', and scope - the words of the sentence the cues refer to. As with other TM tasks, detecting these have yielded a variety of approaches. Rule based approaches have been used to detect negation of PPIs[225], while ML approaches have been implemented to detect negation and speculation more generally[222, 226].

Linking text mining components for large-scale data extraction

While it is useful to focus developments on individual TM tasks, such as those reviewed so far, to reach a complete understanding of a body of text often requires that many TM components must be linked together. For example the following sentence requires pre-processing (tokenization, POS-tagging etc.), various NER and entity normalization tools (i.e. protein, mutation, anatomy and disease), entity relationships (e.g. their events), speculation and negation detection:

'*We speculate the BDNF Val66Met mutant up-regulated the expression of NGF in the hippocampus resulting in no increased risk of neuropathic pain'*

To successfully perform all of these steps at once thus becomes more difficult as the chance of error increases substantially. Moreover, many of these steps are dependent on each other (Figure 1.7), i.e. pre-processing or the proteins linked to

the events – if one of these fails, the proceeding TM tasks then operate on erroneous data.

However, two prominent systems that have demonstrated the utility of multiple TM components are BioContext[227] and Evex DB[217] (as well as other frameworks, such as GATE[228] and UIMA[229]). Both of these have released full-scale TM databases containing the results of multiple of TM systems run collectively on the whole of Medline and open access PMC. BioContext, released in 2012, contains 11.4M biomolecular events for 290K unique genes and proteins with added context from anatomical associations and any negation or speculation involved; all provided by numerous different software. One particularly useful aspect of this system was the use of different protein recognition and event linking tools, which when combined offered increased recall (gene/protein NER 92%, 79-84% otherwise) and when in agreement increased precision (event extraction 66%, 46-50% otherwise) [227]. The more recently updated Evex DB is also centered on biomolecular events (>40M) for genes and proteins (>76M) and provides an expanded ontology of these[217].

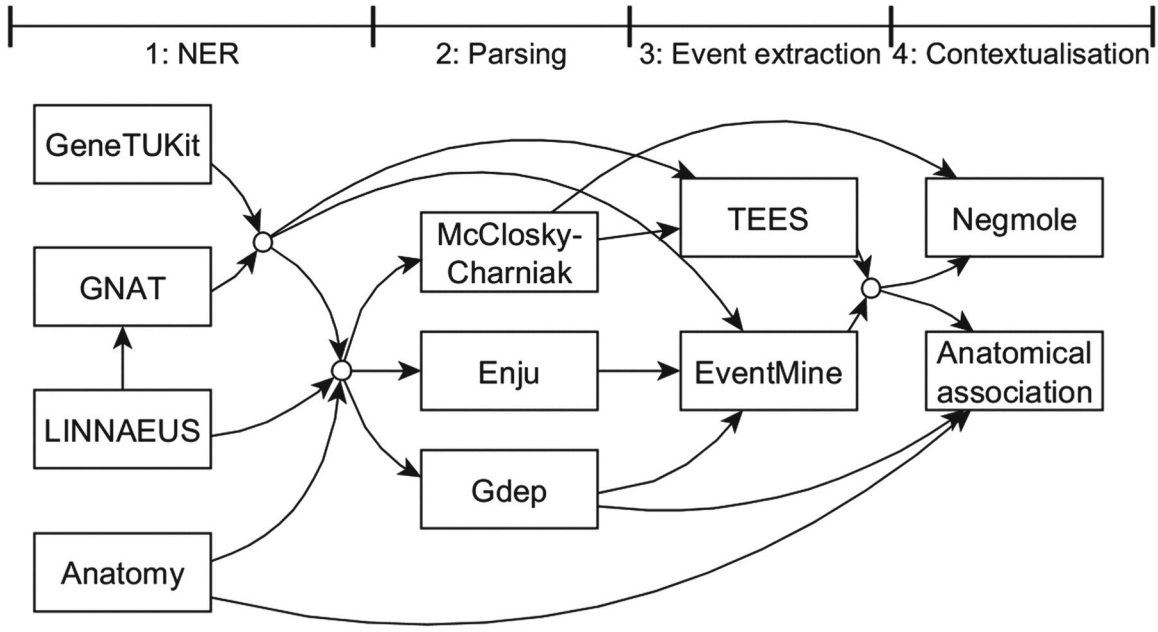


Figure 1.8 The BioContext system architecture.
Each box represents a different tool used in BioContext. Each tool furthermore has its own individual dependencies. Arrows represent the flow of data, while circles represent data merging and post-processing. See Gerner *et al* (2014) for further details (image provided from here also)[227].

Full text versus abstracts and titles

The way in which language is conveyed and stylized across a biomedical article varies across the title, abstract and the rest of the text. For instance, this can be through differences in sentence lengths, incidence and types of parenthesized text, and morphosyntactic and discourse features. These differences, not only present additional challenges for new TM software, but also, because many TM systems have traditionally been developed for use on abstracts and titles only, much of the existing state-of-the art is not optimally designed for full text use. For example, BANNER achieved 56.3% F score on abstracts and titles, whereas on full text this performance dropped 50.4%[230].

Part of the reason why TM systems were first developed for primary use on abstracts and titles was the lack of available full text to process. This has been a well-documented problem for text-miners[231], where most publishers prevent access to computationally processing full text articles even if a subscription has been obtained. All the more, many articles that are open access and thus freely accessible to view are also restricted from TM. However, in recent years PMC has made available a special open access subset of articles, which can be used for TM. Furthermore, in 2014 Elsevier began allowing access for TM to its collection of over 11 million published articles[232], following proposed changes in the European Commision's Text and Data Mining Report (2014)[233].

1.3.4 Semi-automatic approaches to data extraction and curation

So far we have seen how manual curation can be used to accurately extract data from the published literature, but it is highly time-consuming and often impractical for large-scale data extraction. On the other hand, TM can be used to rapidly extract data from articles, although, as we shall see later in this thesis, its precision and recall are likely to be too low to conduct rigorous biological analyses, particularly for more complex tasks. Therefore, efforts have been made

to develop approaches that sit somewhere in the middle of the two, taking advantage of the speed of one and the accuracy of another.

Often termed assisted curation, these approaches are built around guiding human curators to the desired data to simplify the curation process and prevent time wasted scanning irrelevant text (Table 1.4). Articles can be prioritized for curation using TM to calculate the likelihood of the text containing pertinent data for a task[234]. For example, basic article ranking for curating a murine database of PPIs might involve selecting all articles that contain murine-related terms (e.g. mouse, mice, *Mus musculus* etc.). However, more complex approaches have been implemented to rank articles based on the probability of containing a curatable interaction[235].

Name/task	Description
PCorral[236]	Interactive mining of PPIs for curation, utilising TM in IR and PPI extraction.
ODIN[237, 238]	Suggest entities in text derived from the OntoGene TM system in the ODIN interface.
CvManGO[239]	Comparisons between literature-based and computationally predicted Gene Ontology annotations to identify genes whose annotations are need of review.
CTD curation[235]	Documents ranked for curation based on gene, chemical and disease terms provided by TM.
PathText[240]	TM components were integrated with a pathway visualizer and annotation tools to aid curation of metabolic and signaling pathways.
MGI curation[241]	Documents triage and NER used to enhance curation of the MGI model organism database.

Table 1.4 Semi-automatic approaches to data curation

As well as providing document ranking, TM, when incorporated with data visualization platforms, can be used to suggest data mappings within the text. For example, the ODIN curation interface highlights relevant terms and any candidate relations are then specified on a separate panel. This allows a curator to validate whether the data has been extracted correctly, without having to map the data to its associated identifiers manually[237, 238].

1.4 Analysing biological data

Once a biological dataset has been acquired, through whatever means, it is then necessary to statistically analyse its properties, quality and to transform it into a form for which useful patterns and knowledge can be unearthed. For large-scale biological datasets these techniques are vital in assessing the viability of the data for making and supporting biological hypotheses in studying disease. This section will explore how these can be applied specifically to molecular interaction datasets, particularly through the use of biological networks.

1.4.1 Systems biology

The aim of systems biology is to model and discover emergent properties of biology as a system through holistic top-down approaches. A major component of this discipline is the study of interactions between components of the biological system, such as proteins, and how these contribute to the overall function of that biological system. These interactions can be represented as networks and, as such, they represent a perfect way for visualizing molecular interaction data extracted from the literature.

Network biology

In a broad sense, a network can refer to any collection of objects for which they are connected by links. The flexibility of this definition has allowed for the use of networks in a wide range of settings, for example, from its first use in Euler's topological maps of bridges in 1736[242], to social networks and the connectivity of friend groups. By allowing data to be represented in this way, it offers opportunities for seeking trends from the full dataset, for which they can be derived through the application of graph theory.

In graph theory, objects are termed *nodes* or *vertices* and their links are named *edges*. Two nodes connected by edges are called *neighbors.* The edges between neighbors can be *directed* or *undirected*, depending on whether one node points to another. Overall, the nodes and their edges together form a *graph*. Within a graph, we can measure individual properties of each node or look at the overall

structure. For example, *degree* refers to the total number of edges a node possesses, while we can calculate the average degree of a network through taking the mean of these[243]. *Hubs* are nodes within networks that have a high-degree relative to other nodes and networks containing a few of these among mainly low-degree proteins are said to be *scale-free* – termed so as the degree distribution is independent of scale. As well as scale-free formations, networks can also form *hierarchical* or *random* arrangements (Figure 1.8). We can also determine the *betweenness centrality* of a node, which measures the number of shortest paths from all nodes that pass through the node in question. Nodes with high betweenness centrality are often termed *bottlenecks* as they represent focal points for which information can be transferred through a network.

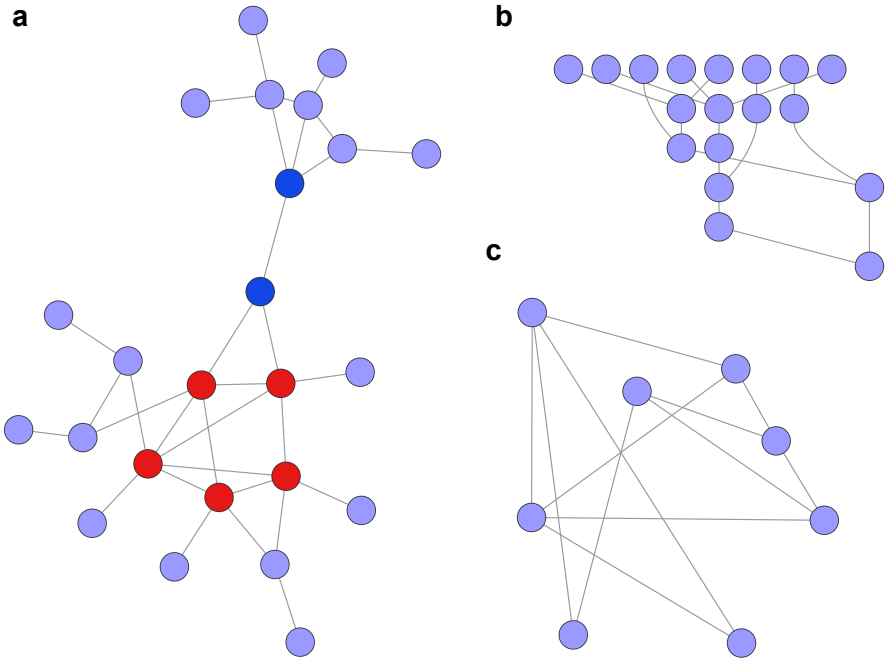


Figure 1.9 Network properties
a Scale free network. Red nodes represent hubs and dark blue nodes represent bottlenecks. **b** Hierarchical network. **c** Random network.

When applied to protein interaction networks, graph theory can be useful for understanding the role of each protein in the overall system. For example, proteins that form hubs and bottlenecks have been shown to be key to the function of biological systems[244]. Furthermore, when visualized with programs such as Cytoscape[245] or iGraph in R[246], networks can be useful ways to present

large-scale datasets in a more visually appealing and comprehensible form (Figure 1.9).

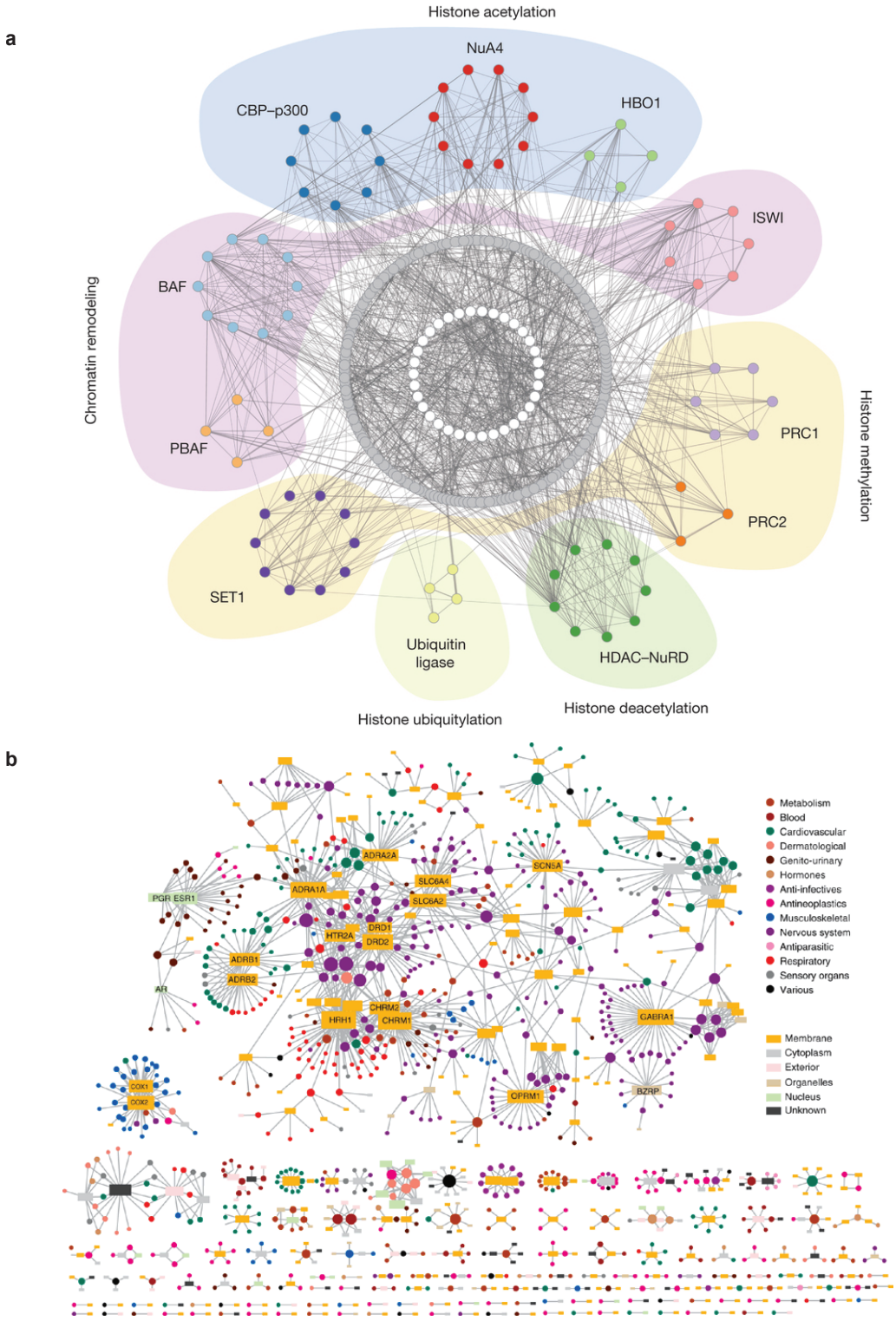


Figure 1.10 Example protein interaction networks.
a Protein-protien interaction network of human chromatin factors. From Huang et al (2013)[247].

b Drug-target network for drugs targeting human proteins. From Yildrim et al (2007)[248].

Using Gene Ontology terms

As well as using graph theory to provide analysis on the roles of proteins in biological systems, GO terms can be used to offer additional functional detail on collections of protein interactions. The GO provides an invaluable resource of annotations for genes on their biological processes, molecular functions and cellular components[128, 249]. These have been manually assigned from an ontology of over 40K biological concepts reported in experiments documented in over 100K published biomedical articles. Each term is linked to other databases, such as Entrez Gene, for which we can then use to automatically assign to proteins (provided the proteins have been mapped to a common gene ID). This added functional detail can then enable us to understand more clearly the role of each protein within a system or the properties of the entire system itself. For example, a group of proteins all with the GO annotation 'bone development' are likely to be related through this concept.

However, as an ontology, terms are represented hierarchically, and the numbers of genes each term is assigned to varies. For example, many genes are assigned to 'cytosol', whereas only a handful are assigned to 'cytosolic small ribosomal subunit'. It is thus often important to calculate the enrichment values of GO terms when analyzing their novelty in gene sets, using tools such as DAVID[250]. Furthermore, when aiming to represent more generic GO-gene associations, GO slim terms can be used - embodying a select group of top-level terms (e.g. 'cytoplasm' or 'cell-cell signaling').

1.5 Research rationale

As discussed above, diseases relating to human pathogens and pain have a huge negative impact on global society and make interesting and useful disease classes to study in greater molecular detail. We have demonstrated that the published literature contains a wealth of data for achieving this purpose, however accessing it is impractical without computational assistance. The best way of extracting and

understanding data from the published literature comes through the use of TM, although it has been shown to be an imperfect method. Our rationale for the research chapters undertaken in this thesis is therefore as follows:

We first aimed to investigate the strengths and weaknesses of TM using a specific disease as a case study. We decided to use HIV-1 and explored recreating the HHPID. This database was constructed from a large collection of published articles and was curated manually, thus acting as 'gold standard' for comparison with results produced by TM. The results from this study showed that TM could recreate a large proportion of this database, as well as interactions that were not present and potentially missed by manual curation. However, the TM methods also produced false positives and false negatives and it was therefore suggested that TM would be best served as a *support* solution to manual curation.

We next considered how an entirely new database of unique molecular interactions could be constructed for pain related diseases using TM. This presented new challenges, as firstly we had to identify publications for sourcing pain relevant interactions and secondly we had no way of knowing whether the data we had extracted was accurate. We therefore developed an approach to building a pain-specific corpus of literature and then expanded on our approach of grouping unique molecular interactions (proposed in Chapter 2) to create a novel way to curate TM data that was both highly efficient and accurate. The approach is detailed in Chapter 3. Furthermore, we showed how additional context could be added to molecular interactions to provide further detail on their roles within pain diseases.

Chapter 4 then presents an analysis of an expanded dataset of pain-related PPIs derived from the procedure outlined in Chapter 3. In this chapter, our aim was to show that we had not only created an accurate dataset of PPIs that was relevant to pain, but that it was also useful for investigating the mechanisms of pain and its diseases in line with other research conducted on experimental data. We achieved this by comparing TM-derived data against pain-related gene expression and manually curated gene lists, showing TM data to be more relevant to pain than these. Furthermore we highlighted the value of the contexts

we had added by exploring pain in different diseases and anatomical regions, identified new drug repurposing opportunities and offered detailed insight into the underlying mechanisms behind pain.

Now that we had demonstrated that we could create an entirely new database of molecular interactions between human proteins for disease research, our focus was to explore curating interactions where human proteins were the targets of multiple pathogen species. While there are differences in the pathogen proteins that interact with the human proteins it is ultimately their effect on the human proteins that leads to disease. We repeated our style of approach from Chapter 3 to deriving TM data, and improved it by showing how the TM data could be enhanced for five viruses. We then curated TM derived host pathogen interactions against those present in public databases in three tasks. Task one revisited extracting HIV-1-human protein interactions by applying our method of curation to see how we could extend the HIV-1 protein interaction database. Tasks two and three showed how our method of curation could be extended to uncover new interactions between human proteins and pathogen species and to identify human proteins that had not been identified as interacting with any pathogen in public databases. These findings are presented in Chapter 5.

Finally, in Chapter 6 we present a general discussion of the topics covered in Chapters 1-5. Here, we summarise our methodology and findings for extracting molecular interactions contextually for studying disease and outline any future directions, before concluding the thesis.

CHAPTER 2

Towards semi-automated curation: recreating the hiv-1, human protein interaction database

2.1 Abstract

Manual curation has long been used for extracting key information found within the primary literature for input into biological databases. The human immunodeficiency virus type 1 (HIV-1), human protein interaction database (HHPID), for example, contains 2589 manually extracted interactions, linked to 14 312 mentions in 3090 articles. The advancement of text-mining (TM) techniques has offered a possibility to rapidly retrieve such data from large volumes of text to a high degree of accuracy. Here, we present a recreation of the HHPID using the current state of the art in TM. To retrieve interactions, we performed gene/protein named entity recognition (NER) and applied two molecular event extraction tools on all abstracts and titles cited in the HHPID. Our best NER scores for precision, recall and F-score were 87.5%, 90.0% and 88.6%, respectively, while event extraction achieved 76.4%, 84.2% and 80.1%, respectively. We demonstrate that over 50% of the HHPID interactions can be recreated from abstracts and titles. Furthermore, from 49 available open-access full-text articles, we extracted a total of 237 unique HIV-1–human interactions, as opposed to 187 interactions recorded in the HHPID from the same articles. On average, we extracted 23 times more mentions of interactions and events from a full-text article than from an abstract and title, with a 6-fold increase in the number of unique interactions. We further demonstrated that more frequently occurring interactions extracted by TM are more likely to be true positives. Overall, the results demonstrate that TM was able to recover a large proportion of interactions, many of which were found within the HHPID, making TM a

useful assistant in the manual curation process. Finally, we also retrieved other types of interactions in the context of HIV-1 that are not currently present in the HHPID, thus, expanding the scope of this data set. All data is available at http://gnode1.mib.man.ac.uk/HIV1-text-mining.

2.2 Introduction

The human immunodeficiency virus type 1 (HIV-1), human protein interaction database (HHPID) is a manually curated database containing 2589 distinct HIV-1 to human protein interactions, linked to 14 312 mentions in 3090 Medline articles[136, 149]. Each of these documented interactions is potentially of value to researchers studying HIV-1, where improved treatment strategies are in urgent demand for a disease that reported 33.3 million confirmed positive cases in 2009, leading to 1.8 million acquired immune deficiency syndrome-related deaths a year[251]. As well as providing an instant resource to researchers seeking distinctive literature on specific HIV-1–human protein interactions, the HHPID has been used to construct detailed networks of the overall host–pathogen interactome[252] and has been vital in RNAi studies with HIV data[11, 253, 254].

The curation of the HHPID took over 7 years to complete and, ideally, it requires on-going updating. While an update based on manual curation is imminent, spanning from 2007 to 2011, future updates would benefit from some form of assisted curation effort. In the original design process of the HHPID, approximately 100 000 relevant HIV-1 documents were identified through PubMed queries, before further review and filtering reduced this number to 3200[254]. As of December 2011, a simple PubMed search for 'HIV' produces more than 233 000 results (including more than 64 000 new abstracts since 2007), highlighting the availability of a large body of potentially relevant literature for automated curation. Therefore, future updates to the HHPID will benefit from the ability to systematically process a much larger body of HIV-focused literature.

Text-mining (TM) techniques have emerged as a potential support solution to the knowledge extraction problem, helping to keep pace with the existing and

ongoing expansion of primary literature. TM systems are designed to convert text data into manageable information and knowledge[255]. Within TM, there exists a range of techniques used to identify, extract, analyse and visualize data stored within text[256]. A large degree of focus within the field has been placed on accurately and exhaustively extracting molecular interactions (MIs) from biomedical text, supported by collaborative events such as the BioCreative and BioNLP shared tasks[160, 169]. These have led to the overall advancement of biomedical TM, making large-scale data extraction an immediate possibility[257, 258]. However, the quality of TM data has historically been scrutinized in comparison to manual curation, where aspects such as gene name ambiguity[259] and conflicting event relationships[260], have impeded its overall accuracy.

Existing forms of assisted curation using TM approaches have benefitted the manual curation process by reducing the scale and complexity of information that curators have to process. For example, Wiegers *et al.*[235] have demonstrated potential in ranking documents according to chemical, gene/protein and disease identifiers in text to augment the efficiency of manual curation of the Comparative Toxicogenomics Database. Another example comes from Kemper *et al.*[240] who have integrated TM components with a pathway visualizer and annotation tools to aid curators in generating metabolic and signaling pathways more effectively.

In this article, we explore the reconstruction of the HHPID using a suite of tailored state-of-the-art TM tools. The results and analyses demonstrate that TM is able to recover a large proportion of interactions found within the HHPID with reasonable recall and precision, in addition to expanding the scope of the database by identifying interactions between other types of entities. These techniques have demonstrated that future curation of the HHPID and indeed other MI databases can be assisted by TM helping speed up the curation process.

2.3 Methods

Figure 2.1 summarizes our approach for recreating and evaluating the HHPID using text mining tools. The method has four main steps: (i) text retrieval (using

only citations from the HHPID), (ii) named-entity recognition (NER, finding mentions of molecules in text), (iii) molecular event extraction (finding any interactions that exist between entities) and (iv) various evaluations and comparisons of the results.

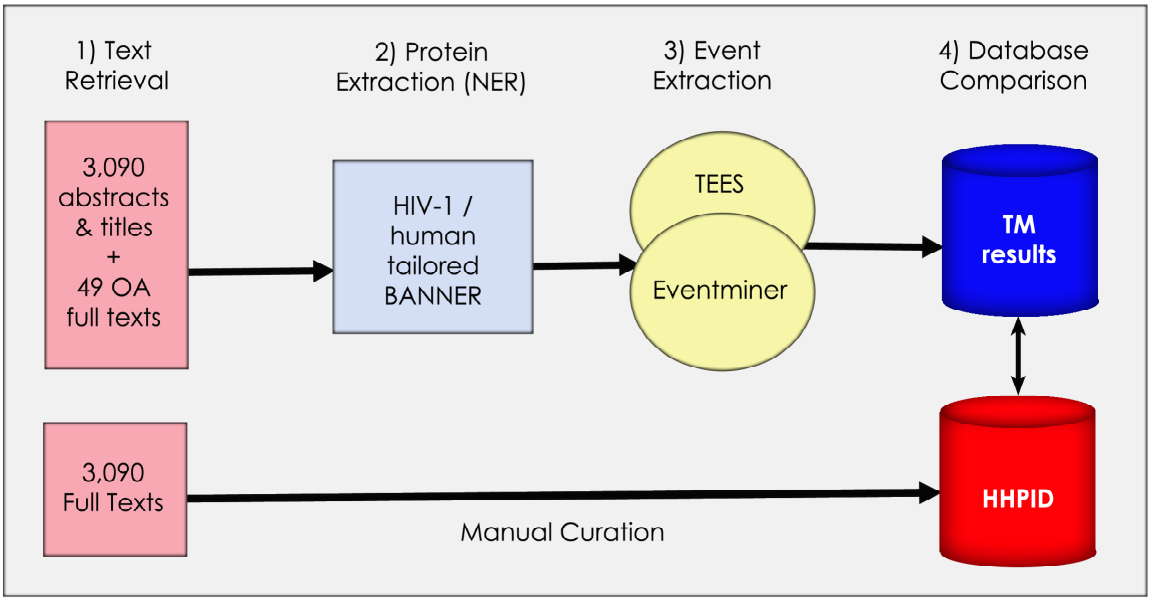


Figure 2.1 Summary of the methodology.
Our methodology is divided into four stages: (1) retrieval of all abstracts and titles, as well as 49 open-access full texts from the 3090 citations in the HHPID, (2) proteins were extracted using an HIV-1/human tailored version of BANNER, (3) events were extracted using two event extraction tools (TEES and Eventmine) and (4) a comparison of the results retrieved by TM was made with the manually curated HHPID.

2.3.1 Data

We limited our investigation to only those articles used in the HHPID to directly compare manual curation to TM. Of the 14 312 citations in the HHPID, we found 3090 of these to have unique PubMed identifiers (PMIDs). Only 49 articles (1.6%) were available through PubMed Central (PMC) as full-text open access (OA) articles. While it would be preferable to use full text for the entire set of 3090 citations, the limited availability of OA articles restricted our main analysis to using abstracts and titles. To illustrate the value of using full-text articles, we also performed a separate experiment using the 49 OA articles.

2.3.2 Named entity recognition and normalization

To extract proteins from the text, we used BANNER, which has been ranked as one of the top performing NER systems by the BioCreative shared task III[160, 186]. Since BANNER has been developed for use on NER across generic biomedical text, we decided to make adjustments to focus the tool on HIV-1 specific text so that we could enhance its overall performance[261]. To identify any specific BANNER performance weaknesses on the HIV-related literature, we first evaluated the performance on a corpus of 50 randomly selected abstracts and titles from the HHPID (referred to as 'Train-HIV'). We evaluated these abstracts using the same evaluation approach as used in NER evaluation in the BioCreative III shared task using precision, recall and F-score[160, 262].

The initial evaluation of BANNER revealed commonly occurring types of false positives such as protein regions (e.g. 'V3') or event mentions (e.g. 'superoxide release'), and false negatives such as hyphenated entities (e.g. 'tat-induced') or entities contained within brackets (e.g. '(SOD1)'). While false positives were difficult to distinguish computationally, we were able to reduce the number of false negatives by providing an additional training data set with HIV-1–human interaction-specific classes of false negatives annotated in text. Furthermore, we designed and implemented post-processing modules to work in unison with BANNER and reduce false negatives by applying dictionaries of HIV-1 and top occurring human-related gene names to match untagged proteins from the text. We then evaluated our modified version of BANNER on a new corpus of 50 randomly selected abstracts and titles from the HHPID (referred to as 'Test-HIV'). The abstracts and titles were marked for all proteins, genes and RNA molecules by DGJ.

In addition to recognition of gene names in text, we normalized our NER results to either HIV-1 or human genes using the Entrez Gene gene names, gene symbols and gene aliases[263]. While normalization has traditionally been made difficult by intra- and inter-species gene name ambiguity[201], HIV-1's small gene

set (nine genes) and the knowledge that each document was HIV-1 relevant, helped us to more confidently and accurately associate genes with HIV-1. Gene names that could not be normalized to an HIV-1 dictionary were, wherever possible, mapped to a human dictionary. If they were not matched to either an HIV-1 dictionary or a human dictionary, they were classified as 'other'.

2.3.3 Event extraction

We focus our investigation on specific types of events that represent interactions between proteins as defined by BioNLP'09[169, 264]. These interactions cover three types of protein metabolism (specifically, gene expression, transcription and protein catabolism), phosphorylation, localization, binding and regulatory events (regulation, positive regulation and negative regulation). Events are identified in text by using two event extraction tools, the Turku event extraction system (TEES)[265] and Eventmine[260]. The tools have been designed to conform to the BioNLP task. Events of gene expression, transcription, protein catabolism, phosphorylation and localization types are all required to act on a single gene or protein, called a theme. Binding events can have one or two gene/protein themes. Regulatory events differ in that their theme may be either a gene/protein or another event. While not required, a regulatory event can also have a gene/protein or another event as its cause. This allows for the possibility of 'event chains' involving multiple gene/proteins in multiple events. For example, the sentence "Tat increased the expression of NF kappa B" mentions an event chain that includes 'expression of NF kappa B' and positive regulation of that event by 'Tat' (Figure 2.2).

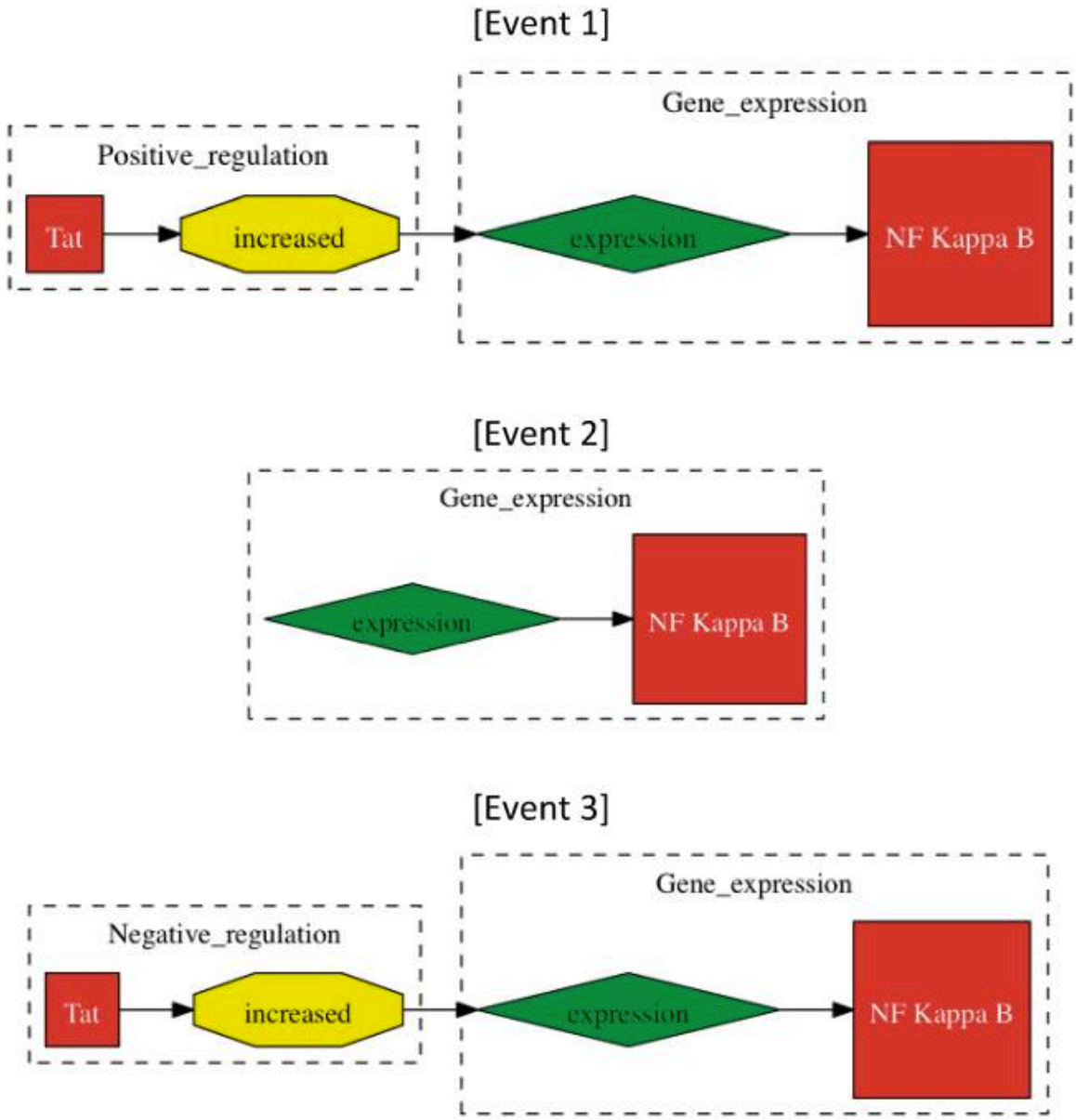


Figure 2.2 Methods of event evaluation.
The three events have been extracted from the sentence "Tat increased the expression of NF kappa B". In this sentence 'Tat' and 'NF kappa B' represent proteins and 'increased' and 'expression' represent events. In approximate evaluation, both events 1 and 2 would be counted as true positives, whereas only Event 1 would be considered a true positive in stringent event evaluation, as 'Tat positive regulation (increased)' is missing. Event 3 would be a false positive in both categories of evaluation, whereby 'increased' does not signify negative regulation.

We applied the two event extraction systems to 3,090 titles and abstracts and 49 full-text articles associated with HHPID, after these had been tagged by the HIV-1/human tailored version of BANNER. We considered molecular events identified by either of the systems (union) or by both systems (intersection).

2.3.4 Event Evaluation

Molecular interactions represented in the HHPID are characterized by 70 keywords that potentially indicate the type of interaction, many of which are potentially redundant (e.g. 'binds' and 'complexes with'). To enable us to compare the event extraction results with interactions from the HHPID, we mapped 51 out of the 70 HHPID interaction keywords to the nine event types (see Supplementary File S1). The remaining 19 interaction keywords (such as 'glycosylates') were designated as 'other' in the results.

To assess the performance of the event extraction systems, we used our Test-HIV corpus of 50 abstracts and titles. Rather than evaluating single events as is commonplace in the BioNLP shared tasks[169], we evaluated 'event chains' since these represent a more complete depiction of the full interaction and have been represented as such in the HHPID. Event chains were evaluated under two different sets of rules: (i) Stringent event evaluation required that any recorded event chain should be represented in its entirety, i.e. without any falsely reported information in order to be classified as a true positive. (ii) Approximate event evaluation differs in that each reported event chain should be represented without any falsely reported information, although it may still be classified as a true positive if some information is missing. This allows for event chains with missing themes or causes to still be classified as true positives provided the rest of the captured data is correct. Figure 2.2 provides some examples of event evaluation methods.

2.3.5 Comparison of TM results to HHPID interactions

In order to ensure that any comparisons made between the TM results and the HHPID were fair, we firstly limited our analysis to only citations from the HHPID and interactions between HIV-1 and human molecules. When comparing interactions from the HHPID against TM, we used the Entrez gene IDs as specified in the database and cross-referenced TM entities with Entrez Gene HIV-1 and human gene names, gene symbols and gene synonyms.

It was not possible to automatically evaluate all TM-extracted interactions against the HHPID due to incompatibility of the data format representations (e.g. unspecified triggers, textual positions and full text/abstract origin of interactions within the HHPID). Instead, a random sample of 50 abstracts and titles from the data set was chosen and interactions reported within the HHPID as originating from the set were compared against those extracted through TM. We only considered interactions from the HHPID that were present within the abstracts and titles and not the full text. In addition, interactions that could not be extracted by TM, since they did not conform to the nine event types, but were present in the HHPID (e.g. 'acetlyation' interactions), were ignored.

A separate analysis was performed on the 49 PMC full-text OA articles that were cited in the HHPID. Following a similar procedure as above, we compared interactions retrieved from full text by TM against those retrieved from the same subset in the HHPID and those retrieved from only abstracts and titles by TM.

2.4 Results

We report two types of results: the generic accuracy of text mining tools and accuracy specifically applied to the HHPID.

2.4.1 Accuracy of Text Mining Tools

The performance of the original version of BANNER[186] on our Test-HIV corpus showed precision, recall and F-score of 83.9%, 87.9% and 85.8%, respectively. When we used altered training data and combined BANNER with a post-processing module, our precision, recall and F-score were all improved to 87.5%, 90.0% and 88.6%, respectively, showing a marginal increase on the default BANNER configuration.

Table 2.1 shows the precision, recall and F-score for the event extraction tools. Results are provided for TEES and Eventmine individually, their union (i.e. both tools) and their intersection (i.e. when both tools are in agreement). Eventmine performed better than TEES in both stringent and approximate matching, with the highest precision, recall and F-score in approximate matching: 79.9%, 73.7%

and 76.7%, respectively. When the results of both tools are merged in a union of events, recall and F-score are both notably higher in the stringent and approximate evaluations compared to individual tools and the precision is greater in the stringent evaluation. Our analysis showed that this was due to full event chains now being completely represented. However, the precision of the union is slightly lower (−3.5%) in the approximate matching. The highest precision is achieved in the intersection of the two tools (87.4%), although recall (46.2%) and F-score (60.4%) are considerably lower. We therefore decided to use the union of the two tools for further investigation.

	Stringent evaluation			Approximate evaluation		
	Precision	**Recall**	**F Score**	**Precision**	**Recall**	**F Score**
TEES	0.373	0.524	0.436	0.726	0.682	0.703
Eventmine	0.460	0.622	0.529	0.799	0.737	0.767
Union	0.537	0.786	0.638	0.764	0.842	0.801
Intersection	0.663	0.392	0.493	0.874	0.462	0.604

Table 2.1 Event extraction performance on the Test-HIV gold standard of 50 abstracts and titles

2.4.2 Comparison of HIV-1–Human Interactions extracted by TM and the HHPID

Table 2.2 shows the total numbers of HIV-1-human molecular interactions for the HHPID and TM. We note that the TM results here are restricted to interactions between HIV-1 and human molecules only. The HHPID showed greater total numbers of interactions for all of the event types in comparison to TM. This is not surprising considering that the HHPID was derived from full text, whereas TM in this analysis was applied to abstracts and titles only. Table 2.3 further shows a comparison between the proteins involved in events ("participants") with the highest frequency in HIV-1-human interactions in the HHPID and TM. Here we observed eight out of ten of the same proteins shared between the two datasets.

Interaction Type	Total HHIPD interactions (abstracts, titles & full text)	Total TM interactions (abstracts and titles)
Binding	5,534	1,967
Protein catabolism	122	40
Positive regulation	3,517	329
Phosphorylation	223	33
Localisation	565	37
Transcription	N/A	31
Regulation	990	127
Gene expression	N/A	243
Negative regulation	1,935	124
Other	518	N/A
Total	13,404	2,931

Table 2.2 The number of HIV-1-human interaction mentions extracted from 3,090 citations: a comparison between the HHPID database and the TM results

HHPID	
Participant	**Total interactions**
Env gp160	4,863
Tat	4,247
CD4	1,188
Vif	1,005
Nef	980
Gag	867
Vpr	790
Gag-Pol	541
CXCR4	303
CCR5	285
Total interactions	13,404

TM	
Participant	**Total interactions**
Cd4	1,290
Tat	1,226
Gp120	1,161
Nef	531
Env	353
Vpr	230
Cxcr4	230
Cccr5	228
Rev	157
Vpu	65
Total interactions	2,931

Table 2.3 Top 10 most frequent participants in events as presented in the HHPID and as extracted by TM

To estimate how much of the HHPID we have replicated through TM, we compared interactions taken from abstracts and titles in the HHPID against HIV-

1–human TM interactions over a set of 50 randomly selected citations from the HHPID. We were able to match 22 TM interactions to interactions within the HHPID, while 20 interactions that were present in the abstracts and titles were either missed or not fully extracted by TM. Thus, we estimate TM has recreated over 50% of interactions derived from the 3090 abstracts and titles within the HHPID without considering any potential data from full text. The value of using full text in TM is explored later in our analysis.

When we only considered frequently occurring unique HIV-1–human interactions, our results for TM were particularly encouraging. Table 2.4 shows the frequency of the top ten most commonly occurring HIV-1–human interactions extracted by TM. With our analysis restricted to unique interactions, TM achieves a similar number of total interactions (2069) in comparison to the HHPID (2589). All of the top 10 interactions retrieved automatically from text were true positives; however, only 7/10 were present within the HHPID. For example, 'negative regulation of binding of gp120 to CD4' is not present within the HHPID due to there being no regulation of binding interactions recorded within it. The 'binding of gp120 to sCD4' is not distinguished within the HHPID as an interaction, as CD4 is only recorded as 'T-cell surface glycoprotein CD4 isoform 1 precursor' and neglects the 'soluble recombinant' prefix of the CD4 nomenclature from the interaction. Instead, this information is presented within a reference sentence for the interaction in the HHPID and is unable to be filtered in a standard database query.

TM Interaction	Frequency	True positive	Present in HHPID
Binding of Gp120 to CD4	207	Yes	Yes
Binding of Gp120 to CXCR4	32	Yes	Yes
Binding of Tat to Cyclin T1	30	Yes	Yes
Binding of Gp120 to CCR5	29	Yes	Yes
Negative regulation of binding of Gp120 to CD4	24	Yes	No
Binding of Vpu to CD4	19	Yes	No
Binding of Gp120 to sCD4	18	Yes	No
Binding of Nef to CD4	18	Yes	Yes
Vpu positive regulation of protein catabolism of CD4	15	Yes	Yes

Binding of Env to CD4	10	Yes	Yes
Total unique mentions	2,069	N/A	2,589

Table 2.4 Top 10 most frequent HIV-1-human interactions retrieved through TM

While these two instances of missing interactions from the HHPID can be accounted for by constraints in the way data in the HHPID is curated, there is no obvious reason as to why the 'binding of Vpu to CD4' is not present. We were able to confirm this interaction as a true positive from a number of references[266-268], all of which are present in the HHPID article set. We believe that—although binding of Vpu to CD4 has been documented as a direct interaction in a number of publications—the end result of this event is a down-regulation of CD4 and is documented in the HHPID as 'Vpu degrades CD4' and 'Vpu downregulates CD4'—an interaction also qualified in the TM data set by 'Vpu positive regulation of protein catabolism of CD4'. This discrepancy highlights issues for both the HHPID and TM. Here, it is evident that in the HHPID it is not completely clear from the interaction (when ignoring the reference sentence) that Vpu had bound to CD4 to cause its degradation. However, in TM, although both parts of the overall interaction (the binding and degradation) are represented in separate event chains, they cannot with the existing methodology be automatically linked together when spanning over one sentence. A combined TM and manual curation approach could help solve both of these problems, by using TM as a support to manual curation to provide additional descriptions for a candidate interaction.

Given the high number of binding events, we further analysed the most frequent interaction participants involved in this type of interaction. In Table 2.5, we compare the binding participants between the HHPID and TM for the HIV-1 Tat gene, as this gene was amongst the most frequent participants in both data sets. We observed similar numbers of total unique mentions of participants between the two data sets (388 for TM and 323 for the HHPID). 'Cyclin T1', 'p-tefb', 'tbp' and 'Cyct1' (a Cyclin T1 alias) were present in the top ten participants of both data sets. We observed 'Sp1' (11 mentions), 'Pkr ' (4 mentions) and 'Puralpha' (3 mentions) outside of the HHPID top ten, but within the top 10 in the TM results.

Tat Binding	HHPID
P-Tefb	57
Cyclin T1	52
TBP	22
CDK7	18
CCNH	17
ITGAV	16
ITGB3	16
CREBBP	15
GTF2H3	14
ERCC2	14
Total interactants	323

Tat Binding	TM
Tar	51
Cyclin T1	30
Tar RNA	26
p-tefb	18
Tbp	15
Sp1	13
Pkr	11
Pp1	11
Cyct1	9
Puralpha	9
Total interactants	388

Table 2.5 Top 10 most frequent binding participants with the HIV-1 Tat gene

2.4.3 Other Types of Interactions Retrieved by TM

As well as retrieving HIV-1–human molecular interactions, TM retrieved events and participants that were involved in other types of interactions or event chains. For example, in Table 2.5, the top occurring binding participant for Tat in the TM data set, Tar, was not present in the HHPID as this is an RNA molecule and the HHPID only contains protein–protein interactions.

Overall, TM retrieved 5674 events involving only a single HIV-1 protein, 7364 single human events, 437 HIV-1–HIV-1 interactions, 1265 human–human protein interactions and 243 interactions involving two or more participants (Table 2.6). Furthermore, we designated 8415 interactions as other, i.e. not involving an HIV-1 or human protein. We note that it is likely that this number is much lower, given that our normalization methods were not sufficient in categorizing all of the participants into their appropriate species.

Interactant	Number of interactions with			
	One participant	Two interactants	More than two interactants	Total interactions
Cd4	1,924	1,290	62	3,276
Tat	1,244	1,226	52	2,522
Gp120	1,468	1,161	60	2,689
Nef	914	531	18	1,463
Env	621	353	13	987
Vpr	301	230	6	537
Cxcr4	357	230	15	602
Cccr5	337	228	10	575
Rev	278	157	3	438
Vpu	184	65	5	254
HIV-1 protein	5,674	N/A	N/A	6,228
Human protein	7,364	N/A	N/A	7,464
HIV-1 – Human	N/A	2,931	N/A	2,931
HIV-1 – HIV-1	N/A	437	N/A	437
Human – Human	N/A	1,265	N/A	1,265
Other	5,560	2,855	N/A	8,415
Total Event Chains	18,598	7,488	243	26,329

Table 2.6 Top 10 most frequently occurring participants within event chains in the TM results. The table presents the number of interactions with one, two or more interactants.

Some of the most frequently occurring interactions that were not present in the HHPID, due to the restrictions in its scope are shown in Table 2.7. We noted that the majority of TM interactions that were false positives for HIV-1–HIV-1 and human–human MIs were each involving self-interactions, and as such can be filtered out easily. However, while these particular self-interactions represented false positives, we should take into account in future work that self-interactions may sometimes represent true positives as well[269].

Interaction	Interaction category	Frequency	True positive
Binding of Tat to Tar	HIV-1 – HIV-1	51	Yes
Binding of tat to tat	HIV-1 – HIV-1	21	No
Binding of gp120 to gp41	HIV-1 – HIV-1	9	Yes
Binding of gp120 to gp120	HIV-1 – HIV-1	8	No
Binding of Nef to Nef	HIV-1 – HIV-1	7	No
Binding of CD4 to CD4	Human - Human	22	No
Binding of CD4 to CXCR4	Human – Human	21	Yes
Binding of CD4 to CCR5	Human – Human	16	Yes
Binding of CCR5 to CCR5	Human – Human	5	No
Binding of CCR5 to CXCR4	Human – Human	5	No
Gp120 positive regulation of binding of CD4 to CD95	More than 2 interactants	2	Yes
HIV-1 Tat positive regulation of HIV-1 Tat positive regulation of protein catabolism of iKappab	More than 2 interactants	1	No
P73 negative regulation of binding of Tat to Cyclin T1	More than 2 interactants	1	Yes
Negative regulation of NF Kappa B/rel causes negative regulation of tat positive regulation of HIV-1 LTR	More than 2 interactants	1	Yes
Binding of CD4 to Okt4 antibody causes negative regulation of CD4 mobility	More than 2 interactants	1	Yes

Table 2.7 Top most frequent interactions retrieved by TM but not found in the HHPID

Table 2.7 also shows that the HIV-1 trans-activation response element (TAR) is involved in Tat binding. It is interesting that this interaction was not present in the HHPID. Although a fundamental molecule involved in HIV-1's biology[270], this TAR interaction is not included within the HHPID as it is an RNA molecule and the HHPID is limited to proteins only. This is also the case for the HIV-1 long-terminal repeat (LTR). To demonstrate the significance of TAR and LTR's involvement within HIV-1 interactions, Table 2.8 shows their most frequently occurring interactions retrieved through TM and whether they are supported by

the literature. Out of the 15 interactions involving LTR and TAR, only two were false positives.

Interaction	Frequency	True positive
Binding of Tat to TAR	51	Yes
Tat positive regulation of LTR	11	Yes
Binding of Cyclin T1 to TAR	7	No
Binding of RNA polymerase II to TAR	6	Yes
Negative regulation of binding of tat to TAR	6	Yes
Binding of CDK9 to TAR	3	No
Binding of TRP - 185 to TAR	3	Yes
Binding of Tat to Cyclin T1 positive regulation of binding of Tat to TAR	3	Yes
Tat positive regulation of transcription of LTR	3	Yes
Binding of Tat to Vpr positive regulation of LTR	2	Yes
Tat positive regulation of tat positive regulation of LTR	2	Yes
Tat regulation of transcription LTR	2	Yes
Binding of LTR to SP1	2	Yes
Vpr positive regulation of LTR	2	Yes
Ptb positive regulation of binding of RNA polymerase II to TAR	2	Yes

Table 2.8 HIV-1 TAR and LTR most frequent interactions extracted by TM

2.4.4 Full-text TM analysis

Table 2.9 shows most frequent interactions extracted from the 49 articles cited within the HHPID which were open access and available for text mining. We compared HIV-1–human interactions extracted from full text, abstracts and titles and those denoted within the HHPID for this set of articles. For the top 10 interactions retrieved through TM applied on full text, we could only account for four in the HHPID, despite all 10 being true positives, indicating that potentially 60% of top-ranked full-text TM interactions might be missing from the HHPID. In total, there were 237 unique HIV-1–human interactions extracted from the 49 articles. This is 27% more than what is in the HHPID from the same subset, suggesting a potential gap in the interaction references in the HHPID. Although TM will have almost certainly reported some false positives (and false negatives

for that matter) within these, the absence of 6 out of 10 true positive interactions found by full-text TM suggests that manual curation is not as exhaustive as we may have come to expect.

Interaction	Full text TM frequency	Abstracts & titles TM frequency	HHPID frequency	True Positive
Binding of Vif to APOBEC3G	27	0	No	TP
Binding of DC-SIGN to gp120	22	0	Yes	TP
Binding of Nef to ABCA1	20	1	No	TP
Binding of gp120 to CD4	17	0	Yes	TP
Nef Positive regulation of Rac	16	2	No	TP
Binding of Tat to CDK2	15	1	No	TP
Binding of DOCK2 to Nef	14	0	Yes	TP
Binding of Nef to ELMO1	14	0	Yes	TP
Vif Positive regulation of protein catabolism of APOBEC3G	13	0	No	TP
Binding of gp120 to CXCR4	11	0	0	TP
Total unique HIV-1-human interactions	237	39	187	N/A
Total HIV-1-human interaction mentions	4,342	40	N/A	N/A
Other mentions (single events, HIV-1-HIV-1 interactions, etc.)	6,995	441	N/A	N/A
Total mentions	11,337	481	N/A	N/A

Table 2.9 Top 10 most frequent interactions retrieved from 49 OA full-text articles with TM

A comparison of HIV-1–human interactions extracted from full-text to those extracted using only abstracts and titles revealed over a 6-fold increase in the number of unique interactions. Only three of the top 10 interactions from full-text TM were found in the abstracts and titles TM subset. Overall, TM on full text recorded an average of 231 interaction or single event mentions per article in contrast to just 10 in abstracts and titles, an increase of 23 times. These results provide a compelling justification for the use of full text as opposed to only abstracts and titles in TM.

2.5 Discussion

Our custom BANNER system was able to achieve precision, recall and F-score of 88%, 90% and 89%, respectively using a modified, specially tailored training data set and a post-processing module utilizing a dictionary with HIV-1 and top occurring human genes. Although only marginally better than the original system, these scores demonstrated TM to be capable of extracting genes and gene products from HIV text to a useful level. An error analysis shows that commonly occurring false positives were acronyms such as cell line names (e.g. HeLa) or strain names (e.g. HIV-1 subtype B).

For event evaluation, we chose to use a union of two event extraction tools, which—under our most strict method of evaluation—showed precision, recall and F-score of 54%, 79% and 64%, respectively. Our approximate form of event evaluation for our best system showed precision, recall and F score of 76%, 84% and 80%, respectively. These results indicate that a large proportion of false positives from our stringent evaluation were caused not through falsely reported information, but through incomplete event chains, such as missing interaction causes or binding partners. Here, there is potential to improve on the performance of event extraction through completing the event chains that have missing information. These findings support results from other event extraction studies and proposed solutions include tuning the confidence thresholds to improve recursive matching and to use confidence values of the predicted candidates as features in the proceeding modules[260]. However, generally the greatest challenge for event extraction tools comes from apprehending the various writing styles employed by different authors. False positive events were most persistently caused by complex grammatical sentences or just poor grammar, making it difficult for automated tools to ascertain their intended meaning. Figure 2.3 provides some examples of typical false positives.

2.5.1 TM versus manual curation

We have successfully managed to recreate a large proportion of the interactions denoted within the HHPID using the current state of the art in TM. We have shown that TM tools are at least capable of precisely replicating over 50% of the interactions denoted within the HHPID from an evaluation sample of 50 abstracts and titles. Considering the manual curation of the HHPID took 7 years to perform, our tools have proven to be markedly more efficient by replicating a large percentage of this data automatically in a matter of hours.

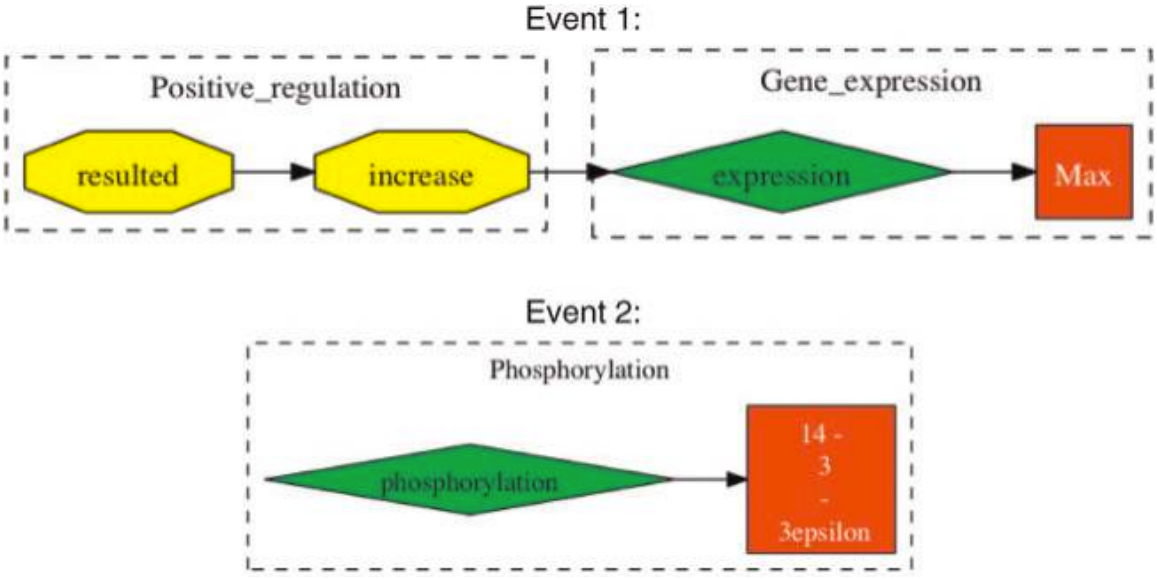


Figure 2.3 Examples of falsely reported event chains.
Events are extracted from the sentence "In parallel to the modulation of cell growth, gp 120 at low concentrations resulted in an increase in the expression of c-Myc, Max, and 14–3-3epsilon proteins and phosphorylation of ATP-dependent tyrosine kinases (Akt) at Ser (473)". Taken from Ref. (20). Event 1 shows an example of an incomplete event chain, where gp120 is missing as the cause for positive regulation. In Event 2, there is falsely reported information in that 14-3-3epsilon is expressed and not phosphorylated.

Across the full list of citations within the HHPID, we have retrieved 2069 total unique HIV-1–human interaction mentions in comparison to 2589 unique HHPID interactions. Although some of these TM interactions probably represent false positives, this result is still extremely encouraging considering that curators of the HHPID had access to interactions from full text as well as abstracts and titles. From these HIV-1–human interactions, we found 7 of the top 10 binding interactants between Tat retrieved by TM to be present in the HHPID. Thus, we

feel that those interactions recovered using TM represent a strong demonstration of how manual curation could be supported by sophisticated TM.

A top participant detected by TM for Tat binding that was not present in the HHPID was the HIV-1 TAR element. We found that the HHPID does not have any mentions of the HIV-1 TAR or any other RNA interactions involving HIV. It was not an objective of the HHPID to document these kinds of interactions, although, they are a potentially valuable resource to researchers studying HIV-1. To determine the role of TAR and another HIV-1 RNA molecule, the LTR, we highlighted interactions involving only these molecules (Table 2.8). Across the 15 interactions that we examined, only two were false positives and thus, we feel TM have the potential to identify valuable information from HIV-specific text on HIV-1 interactions that are not currently present in the HHPID. Given the other types of interactions that could be extracted (interactions between HIV-1 molecules, interactions between human molecules, interactions between two or more participants, etc.), TM tools could facilitate a semi-automated approach to the expansion of the scope of the HHPID database.

From five interactions involving more than two participants that we examined (Table 2.7), we were able to find four true positives. The true positives for interactions involving more than two participants are especially beneficial as that they provide a more complete illustration of interactions in contrast to the HHPID. Figure 2.4 shows an example.

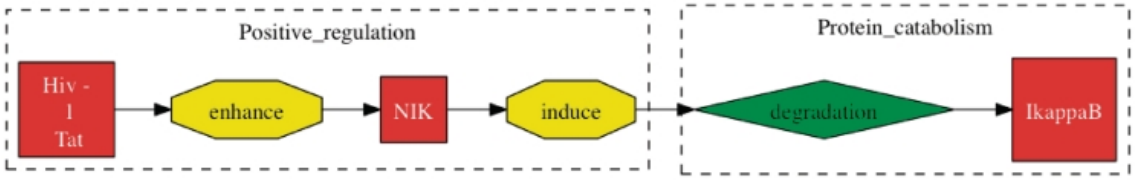


Figure 2.4 TM interaction involving two or more participants.
This event was extracted from the sentence "HIV-1 Tat can substantially enhance the capacity of NIK to induce IkappaB degradation"[271]. Here, we can see that the full interaction is identified by TM, across multiple participants and events. The HHPID documents this same interaction as 'Tat enhances mitogen-activated protein kinase kinase kinase 14', which is clearly a misrepresentation of the actual full interaction.

To consider the potential of full-text TM, we investigated the available OA articles cited in the HHPID. Interactions extracted from this subset highlighted that 6 of the top 10 interactions retrieved by full-text TM were missing, with 27% fewer unique interactions compared to the HHPID. These particular full-text articles referred to large numbers of gene and gene product mentions, contributing to some 11 337 interaction mentions as deduced by TM. While inaccuracies of TM cannot be ignored, these results do perhaps draw attention to limitations of manual curation, especially when dealing with more interaction-saturated literature, e.g. in high-throughput studies which are likely to contain more interaction mentions. However, it should be noted that curators from the HHPID may have chosen to only document the most important interactions denoted within these papers, accounting for the lower numbers of interactions.

In our subset of OA full-text articles, a comparison of TM using only abstracts and titles of the same articles exposed a significantly lower frequency of interaction mentions. On average, there were only 10 interaction mentions in abstracts and titles in contrast to 231 in full text. When only unique HIV-1–human interaction mentions were considered, full text still showed a 6-fold increase in data, with seven of the top 10 full-text TM interactions not present in the abstracts and titles TM data set. Although it has already been demonstrated that full text contains more information[272], only a small number of more than 233 000 HIV-related articles are accessible through PMC OA, thus, limiting the full potential of full-text TM to provide a large-scale systematic approach to information extraction from the entire literature.

One major weakness in our approach was the lack of an advanced normalization system able to fully categorize all of our retrieved participants into either HIV-1 or human species types. The dictionary-based methods we used can potentially be improved by using more sophisticated normalization systems such as GNAT[216, 273] or GeneTUKit[273, 274], capable of normalizing generic protein matches to their Entrez Gene IDs. Better normalization of participants will enable us to more precisely identify the interactions that TM has retrieved. However, we will be careful to ensure that useful context in descriptive prefixes and suffixes of molecules, e.g. ‘mutant’, are not lost while normalizing, as this information can

potentially be useful to researchers in understanding what was originally documented.

2.6 Conclusions

In this article, we explored the potential of a TM-driven approach to curation of the HHPID. The results and analyses demonstrate that TM is able to recover a large proportion of interactions found within the HHPID with a reasonable recall/precision ratio, in addition to potentially expanding the scope of the database by identifying interactions between other types of entities. In principle, TM methods are more likely to retrieve true positives that are more frequently recorded in the literature. With such a large body of citations available for HIV, we believe that in the future we will be able to apply confidence to interactions based on how frequently they were recorded, and thus provide better support to the curation process.

Our analysis of full-text TM has revealed a convincing support for its usefulness, compared to solitary abstracts and titles. With such a dramatic difference in the frequencies of interaction mentions, we believe that in our future work we will be able to retrieve huge numbers of interactions if we have access to all full-text articles. A potential problem in full-text analysis in comparison to using only abstracts and titles will be to identify the 'value' and 'novelty' of an interaction, where aspects such as defining interactions as 'referenced' or 'recorded' will present new TM challenges. However, we believe neglecting such huge amounts of potentially valuable data would vastly hinder any future efforts to curate a more complete HIV-1–human protein interaction database.

Overall, although it is unlikely that TM will ever be able to replicate the accuracy that manual curation can achieve in MI extraction, its main strength is in the speed at which it can generate data that can be used to, amongst other aspects, support the curation process. Our results have shown that TM can retrieve reasonably accurate results for MI extraction and therefore a TM-assisted manual curation approach could be most beneficial, in particular for the more frequent

interactions that can be checked first via references to the text. In the future, we intend to apply the current techniques with any improvements to the full list of HIV-1 citations in Medline and PMC, and make our results available to researchers online. The corpora generated are available on request.

2.7 Acknowledgements

The authors would like to thank Ben Sidders from Pfizer, UK for helpful comments and feedback; and Jonathan Dickerson and Jamie MacPherson for providing help throughout the investigation. We would also like to thank Roger Ptak and William Fu for feedback and comments on the curation of the HHPID.

CHAPTER 3

Cataloguing the biomedical world of pain through semi-automated curation

3.1 Abstract

The vast collection of biomedical literature and its continued expansion has presented a number of challenges to researchers who require structured findings to stay abreast of and analyze molecular mechanisms relevant to their domain of interest. By structuring literature content into topic-specific, machine-readable databases, the aggregate data from multiple articles can be used to infer trends that can be compared and contrasted to similar findings from topic-independent resources. Our study presents a generalized procedure for semi-automatically creating a custom topic-specific molecular interaction database through the use of text mining to assist manual curation. We apply the procedure to capture molecular events that underlie 'pain', a complex phenomenon with a large societal burden and unmet medical need. We describe how existing text mining solutions are used to build a pain-specific corpus, extract molecular events from it, add context to the extracted events, and assess their relevance. The pain-specific corpus contains 765,692 documents from Medline and PubMed Central, from which we extracted 356,499 unique, normalized molecular events, with 261,438 single protein events and 93,271 molecular interactions supplied by BioContext. Event chains are annotated with negation, speculation, anatomy, Gene Ontology terms, mutations, pain and disease relevance, which collectively provide detailed insight into how that event chain is associated with pain. The extracted relations are visualized in a wiki platform (wiki-pain.org) that enables efficient manual curation and exploration of the molecular mechanisms that underlie pain. Curation of 1,500 grouped event chains ranked by pain relevance revealed 613 accurately extracted unique molecular interactions that in the future can be used to study the underlying mechanisms involved in pain. Our approach

demonstrates that combining existing text mining tools with domain-specific terms and wiki-based visualization can facilitate rapid curation of molecular interactions to create a custom database.

3.2 Introduction

One of the largest and most widely used resources of online biomedical literature is the National Library of Medicine's PubMed[275]. PubMed now searches over 23 million biomedical records and with other biomedical literature search engines (e.g. Google Scholar, Web of Science and Scopus) is a typical starting point in biomedical knowledge acquisition and information retrieval (IR)[116, 117]. For example, a researcher searching for 'pain' on PubMed will retrieve 521,141 citations (March 6th 2013). This highlights the key problem that arises when the number of relevant unstructured documents from a topical search exceeds the limits of a researcher's ability to read all (or many) of them. An alternative is to use manually curated resources. Topic-specific curated databases often arise because of unmet needs from existing resources, leading to curation of data not captured by more general sources. They often contain added context that aids the intended users[134, 136, 139, 276]. Extracting, normalizing and cataloging relevant concepts and facts from free text by dedicated curators make it possible to deal with otherwise unwieldy amounts of information. Accordingly, topic-specific databases that house these findings are rapidly accumulating at an increasing rate[277]. Creation of topic-specific databases is well documented[278-280], and there are recurrent themes in the processes used to build high-quality resources. Document triage can be as simple as keyword searches[281-283], but many of these sources have matured enough to shift to sophisticated document classification algorithms[282, 284].

In parallel, there is increasing focus on building tools to help defray the high cost of manual curation[276]. There are few databases that are up-to-date with all available relevant information; funding for manual curation is the limiting factor, rather than finding papers to curate. Assisted curation, e.g. through the process of applying text-mining (TM) tools to highlight curatable events, has been repeatedly shown to increase efficiency and reduce curatorial errors[285].

In addition to using TM tools to highlight facts within a paper, they can also be used to highlight common facts across papers. We recently reported the recreation of a database of human-HIV-1 protein interactions[286] wherein we proposed a method to group identical interactions mentioned in multiple papers. To increase coverage of unique interactions, it is then only a matter of manually curating selected examples from each group of potentially equivalent interaction mentions. In this system, only one instance of a grouped text mined interaction is required to confirm it as a true positive, enabling rapid validation of molecular interactions derived from TM. Such an approach would acknowledge unique interactions as the primary target of knowledge capture rather than individual mentions, as these are often a very valuable feature used by researchers in inferring trends from the overall interactome[11].

In this study we explore whether TM tools can be used to create a full-scale disease-specific molecular interaction database from start to finish. Chronic neuropathic pain is an important public health problem which approximately 5-8% of the European population suffers with[287]. Current treatment regimens are not universally adequate with only 30-50% of patients reporting an appreciable reduction in pain and improvement in their quality of life using the currently available analgesic drugs such as the gabapentinoids, opioids and selective serotonin reuptake inhibitors such as Carbamazepine[288]. In addition, the use of these drugs is often limited by unwanted side effects. There is therefore a significant need for new therapeutics, which requires a better understanding of the mechanisms that mediate chronic pain so that new therapeutic mechanisms might be uncovered. However, there are no existing extensively curated pain-specific molecular interaction databases to facilitate this.

To build a comprehensive pain-related molecular interaction database, we created a pain-specific corpus of biomedical documents using all abstracts and titles from PubMed and full text from open access PubMed Central (PMC). From this pain-relevant corpus, we extracted all molecular interactions using the existing BioContext database constructed from the state of the art in TM. We used existing contexts from this database and added further contexts, such as pain and disease relevancy to interactions, to increase their value to researchers.

Finally, we made available the interaction data retrieved to allow manual curation of the grouped results, with the ultimate aim of creating a highly accurate, pain-relevant molecular interaction database.

3.3 Methods

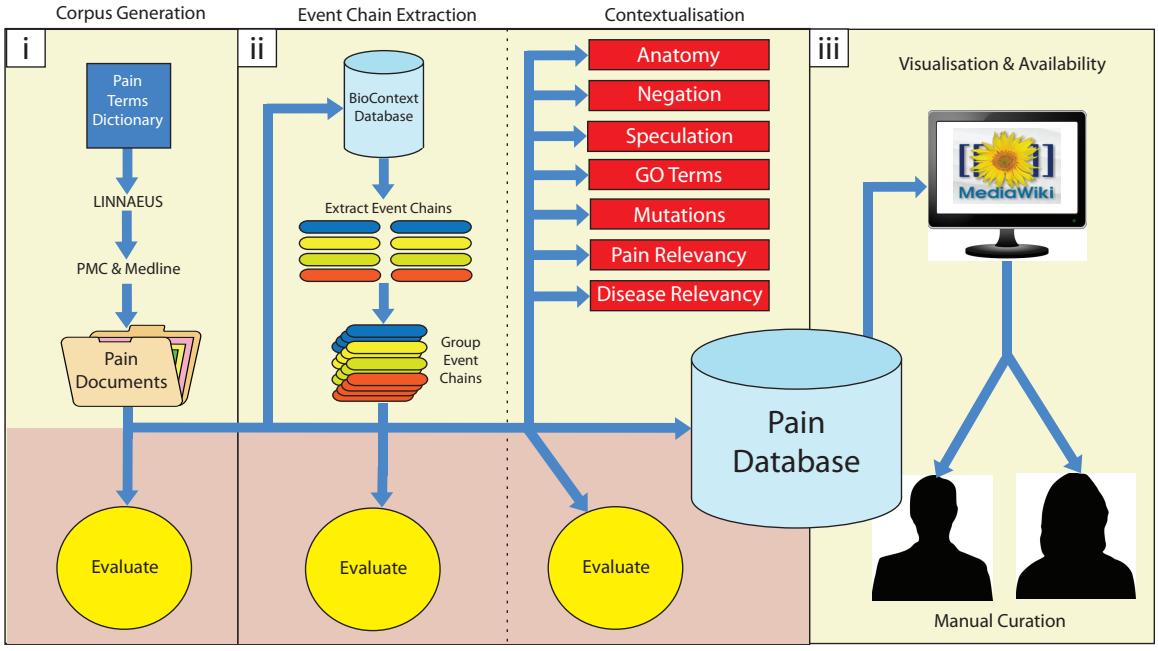


Figure 3.1. Diagrammatic representation of methodology.
Our methodology is divided into three main parts: (i) building a topic-specific corpus and evaluation of document scoring; (ii) data extraction (extracting molecular interactions and adding contexts) and their associated evaluations (iii) visualization and availability for manual curation of results. Each of these is described in detail within the methods section.

3.3.1 (i) Building a topic-specific corpus

Dictionary generation and document retrieval

The first step in generating a full-scale biomedical corpus of documents relevant to pain was to create a pain terms dictionary that could be used to match pain-associated biomedical text. As a basis for the pain terms dictionary we added terms from an online glossary[289], various pain review articles[92, 290, 291] and an in-house term set. Case-sensitive synonyms used in the literature supplemented long forms of pain terms. Ambiguous terms were excluded (e.g. acronyms and

terms commonly used in other fields), as these have been shown to increase false positive results in IR[292, 293].

Dictionary development was iterative, with two rounds of: dictionary term review; document retrieval; manual assessment of retrieved documents for absent or ambiguous terms; and dictionary modification. After an initial review of retrieved pain documents, we enhanced the pain terms dictionary to improve this procedure. Firstly, we added terms to the dictionary that we flagged as false negatives from the initial corpus evaluation. Secondly, we developed a support tool able to rank strings of tokens based on the proportion of stop words they contained and their size (in number of words). Using this tool we took the text from the top ranked 10,000 pain articles to create a list of potential phrases that might be associated with pain. We then manually went through the top terms in this list, adding 33 extra terms to our dictionary. The final dictionary contained 583 terms and 3,144 synonyms.

Each term in the dictionary was assigned one of 12 pain-related categories (e.g. pain type, disorder, pain drug, anatomy, condition, etc.; see supplementary file 1 for details) to provide more contextual data later in our analysis. Furthermore, a specificity assignment was given to each term to reflect whether the term is specifically relevant to the biomedical research field of pain or if it is a more general term that could apply to other research areas, but still has a prominent relevance to pain research. For example, the term 'neuropathic pain' was categorized as a pain type and classified as 'pain-specific'. On the other hand, the brain region *locus caeruleus* is not a term synonymous with pain, but it is relevant to pain as an anatomical region involved in the sensation; these are called "pain-relevant". In general, terms were classified as pain-specific (and assigned a weight of 2) if they were a type of pain disorder, a drug or surgical procedure used to treat pain, a gene with genetic association to pain, or a target of a pain drug. Pain-relevant terms (weight of 1) tended to be anatomically or physiologically relevant concepts. The terms and synonyms, including their categorization and pain specificity scores, were inspected by a biologist with pain expertise.

In order to match pain-specific terms from our dictionary to biomedical text we used LINNAEUS[191], a named entity recognition tool able to match terms from a pre-defined dictionary to text. We note that only pain-specific terms were used for document retrieval. We implemented LINNAEUS' in-built post-processing feature to resolve ambiguity in the results (i.e. terms that corresponded to more than one pain term identifier) and allow the capture of abbreviations associated with terms in the dictionary. We applied this to all abstracts, titles and MeSH terms in Medline (May 2012 release) and to full text in open access PMC (2011 release) that were classified as review or research articles. From herein we refer to our final pain corpus as P1.

Document relevance scoring

To quantify the relevance of each retrieved document in the corpus, a document relevance scoring scheme was developed that makes use of both pain-specific and pain-relevant terms, as well as the position of each term's mention in the document (i.e. title, abstract, MeSH or body). Each pain term matched in a document in P1 was given an individual score based on its textual position (2 if appearing in the title; 1 if appearing in the abstract and in associated MeSH description of the document; 0.25 otherwise) and the pain-specificity of the term (2 if pain-specific; 1 if pain-relevant). These individual scores are then used to determine an overall document relevancy score to pain by summing up the score of all pain terms:

$$document\ pain\ relevance = \sum_{i=1}^{n} t_i p_i$$

where t_i is a term's pain-specificity weight, p_i is a term's position weight and n is the number of pain terms in the document. We can similarly calculate pain category relevancy scores (by summing up the score of all pain terms mentioned for each category) and individual pain term relevancy scores (by using all mentions of a given pain term) in each document.

Evaluations

To evaluate the effectiveness of our document-scoring scheme we selected all documents from P1 containing the MeSH term 'Pain' and then compared the distribution of document scores for those that had 'Pain' as a major MeSH term and those that had 'Pain' as a minor MeSH term. We also evaluated individual pain terms matched within 50 documents that had been retrieved in P1. To ensure that we evaluated documents across our pain document scoring range, we randomly selected 10 that scored between 1-3 in pain relevancy, ten between 3 and 10, ten between 10 and 25, ten between 25 and 50 and ten with a score of 50 or greater.

3.3.3 (ii) Data Extraction

Extracting molecular interactions

In order to retrieve the molecular interactions from P1, we used the BioContext database[227]. The BioContext database was created from a pipeline of state of the art biomedical TM tools applied to the whole of Medline (May 2011 release) and OA PMC (May 2011 release). Each record in the BioContext database is organized into an event chain originating from a single sentence. Every event chain has a minimum of one and a maximum of three events that were extracted by a union of two event extraction tools[260, 265].

Events are categorized into nine types as defined by the GENIA ontology[169, 294], covering protein metabolism (protein catabolism, gene expression and transcription), phosphorylation, localization, binding and regulatory events (positive regulation, negative regulation and regulation). Metabolic events, phosphorylation and localisation have a single gene, protein or RNA molecule(s) as their theme (subject), whereas binding events have one or more gene(s), protein(s) or RNA molecule(s) as their theme. Regulatory events are special in that their theme may be a gene, protein, RNA molecule or another event. They are also unique in that they may have a gene, protein, RNA molecule or another event as their cause. Event chains can thus be formed involving multiple molecules and events. For example, "CCK-induced expression of fos" would create an event chain of "*CCK* **Positive Regulation** (induced) of **Gene Expression**

of *Fos*". A summary of the events and examples of the event chains that can be formed is provided in supplementary file 2.

The genes, transcripts and proteins that form the themes and causes of each event were extracted using GNAT[216, 273, 295] and GeneTUKit[274]. Where possible, each mention is then normalized to a species using LINNAEUS[191] and further normalized to an Entrez Gene ID[296] and finally a homologene ID[297].

We took all event chains from BioContext that were extracted from documents present in P1. We then grouped event chains together that contained the same sequence of proteins and events. For example, mentions of the event chain "*Ros1* **Positive Regulation** of *NFKB1*" extracted from multiple sentences and documents were grouped into a single record. Entrez Gene IDs were used to group proteins instead of gene symbols to prevent erroneous grouping caused by naming ambiguity.

To group event chains involving a binding event with two molecules we had to resolve instances where the order of the proteins varied across analogous event chains. For example, one event chain may be directed as, "Binding of *CD44* and *MMP9*" whereas another may vary as such, "**Binding** of *MMP9* and *CD44*". Since the order of proteins in binding events does not infer any functional characteristic of the data (binding of CD44 and MMP9 is the same), classing these as separate unique event chains when grouping would be erroneous. Thus, we rearranged binding proteins numerically using Entrez Gene IDs when proteins were normalized or alphabetically otherwise.

During the grouping of each event chain we recorded the total frequency of that event chain and the number of documents that each event chain was reported in. We also stored the number of molecules involved in each event chain. This enabled us to define molecular interactions as those event chains containing two proteins, genes or RNA molecules. Those containing only a single molecule are referred to as single events. TM confidence scores provided by BioContext for each grouped event chain were determined by taking the highest confidence score from the associated event chains used in the grouping.

Molecular interaction extraction evaluation

The individual tools used in BioContext to create the event chains used in this study have already been extensively evaluated[227]. We used the results from the final manual curation step (see below) for direct evaluation of grouped molecular interactions.

Pain relevant interactions extracted for this study should be enriched for proteins previously linked to pain. Therefore, we also undertook an enrichment analysis, comparing event chains retrieved from P1 with a set of interactions derived from a random set of documents for the presence of known pain associated proteins. The genes/proteins used as a gold standard pain set were taken from the Pain Genes DB[298]. This set contained 297 unique, manually curated genes. We measured how many unique and total mentions of genes were present in our event chains (both single events and molecular interactions). The generic set of event chains was formed from the same number of randomly selected Medline and PMC documents as P1, but which were not present in P1. Event chains from this random document set (referred to as R1) were then extracted from the BioContext database and grouped using the same procedure as used in constructing the event chains from P1. Unique and total mentions of pain genes present in R1 event chains were then determined. Fisher's exact test was used to statistically evaluate whether P1 was enriched for pain genes in event chains in contrast to R1.

Adding context to molecular interactions

As well as the species context for proteins, BioContext also contains anatomy, negation and speculation context for each event chain. Anatomical mentions in the text (such as "peripheral nerve" or "spinal cord") and cell-line mentions used as proxies for anatomical locations were extracted using GETM[299]. These anatomy mentions were, where possible, mapped to events to provide detail on the anatomical location of an event.

Negation and speculation detection was provided for each event in BioContext using a modified version of Negmole[300]. Instances of negation (e.g., "Lep did not

bind to Obsty1") and speculation (e.g., "Lep maybe binds to Obsty1") are extracted and annotated on the resulting event chain (i.e., "[Negative] **Binding** of *Lep* and *Obsty1*" or "[Speculative] **Binding** of *Lep* and *Obsty1*").

We additionally provide four other contextual features: associated GO terms and mutations, and pain and disease relevance scores.

GO terms[128] and their overarching GO Slim terms[301, 302] were added to normalized proteins where feasible to provide more functional information on proteins involved in each event chain. This was achieved using the publicly available Gene2Go mapping of Entrez Gene IDs to GO IDs available on the National Center for Biotechnology Information FTP service[303].

Point mutation context was added to proteins in event chains by using MutationFinder to match and normalise mutation instances in the text [193]. MutationFinder was run only on sentences that were the source of each event chain in our pain set. However, since MutationFinder is unable to link mutations to any associated protein mentions in the text, we designed and implemented our own system to do this. We formulated a number of priority-ranked regular expressions to match commonly occurring textual patterns, e.g., "<protein> - <mutation>" or "<mutation> for the <protein>". Our system also allowed individual proteins to match multiple mutations, e.g. "mutations <mutation A>, <mutation B> and <mutation C> for <protein>". The regular expressions used are provided at wiki-pain.org/downloads.

We designed a novel method to calculate the relevance of each pain term to an event chain in a document (note that this is distinct from the document relevance method described above). The score ranges from 5 to 100 and reflects the likelihood that a pain term is relevant to a given event chain. The algorithm uses the document sections in which the pain term and the event chain are mentioned (i.e. title, abstract, MeSH and body), whether they co-occur in a sentence, and where appropriate the distance between the two and the order that each is presented. For example, a pain term mentioned in the same sentence as an event chain receives a score of between 75-100. Pain terms matched in different sections to a given event chain are given lower relevancy scores. We were then able to

produce an overall relevancy score to pain for an event chain using individual relevance scores of each pain term above 50 to that event chain and weighting by pain-term specificity. A more detailed description of the scoring calculation with examples is provided in supplementary file 3.

The final context added to our event chains is disease relevancy. Pain, although often considered a disease in itself, is commonly related to symptoms of a whole host of other diseases. To allow researchers to explore these trends in relation to interactions, we matched disease terms from an in-house disease lexicon (containing 4,861 terms with 205,373 case-sensitive synonyms) to P1 using LINNAEUS[191]. We then adopted the same method used in the pain relevancy scoring to calculate the relevancy of each event chain to each disease term match and from these the overall disease relevancy of each grouped event chain (without the term weighting).

Context evaluations

We did not repeat the existing evaluations performed in BioContext[227] for anatomy, negation and speculation contexts. Similarly, mutation detection and normalization had also been previously evaluated for MutationFinder[193]. However, to evaluate the mutation to protein linking method we selected 100 event chains that matched at least one mutation in the original sentence used to extract the data. As well as noting true positives, false positives and false negatives we marked true negatives defined as those mutation mentions correctly left unlinked to a protein in an event chain.

To assess the event chain relevancy scoring system to individual pain terms we randomly selected 100 linked event chains and pain terms that scored above 50 and another 100 that scored below 50. A true positive was given if the term bared some notable relevance to the event chain in question, whether a direct or indirect association.

Our disease relevancy evaluation first assessed the disease term matching performed by LINNAEUS in 50 randomly selected documents that had matched at least one disease term. As above for the pain relevancy evaluation, we selected

100 linked event chains and disease terms that scored over 50 and another 100 that scored below 50 for disease term to event chain relevance evaluation.

3.3.4 (iii) Availability and visualisation for manual curation

To visualize and make our data available to researchers, the MediaWiki (version 1.19) framework was used, as this platform has been successfully utilized in other database representations[304]. The primary use of this system (available at wiki-pain.org) is to support curation of pain-related molecular interactions by providing an infrastructure for assessing data proposed by TM as described above. We built wiki-pain.org using the MediaWiki API to automatically upload pages constructed from our databases[305].

As a pilot, we performed manual curation on the top 1,500 grouped molecular interactions (ordered by overall pain relevancy scores) involving human, mouse or rat proteins and excluding self-interactions, marking each as either a true positive or false positive. The task was spread across three curators, 500 assigned to DJ, 500 to BS and a further 500 to 3 biologists.

Traditional evaluations of events and their protein constituents have focused on selecting a set of articles and scanning the text for requisitioned data and comparing this against the data retrieved[306]. As grouped interactions can be formed from a number of different documents, to fully evaluate even a small number of these using a traditional evaluation would require masses of documents to be assessed. Thus, we chose to evaluate grouped event chains by selecting individual mentions of an event chain ordered by TM confidence and their associated sentences (and documents if needed for further verification) and used these to determine whether an overall grouped event chain was a true positive or a false positive. We required only one correct individual event chain of a group to determine it as an overall true positive. While this form of evaluation requires much less time spent reading each full article, we recognise that as a result we do not measure the frequency of false negative instances.

We evaluated each individual event chain using the stringent form of evaluation as described previously[286]. This evaluation requires the full event chain including

all of its participants to have been extracted and normalized accurately to their correct species and Entrez Gene ID in order to be classed as a true positive.

To assess the quality of our manual curation, we determined the inter- and intra-annotator agreement by one curator blindly re-curating 50 randomly selected molecular interactions previously curated by that curator (intra) and 50 randomly selected molecular interactions previously curated by other curators (inter). Furthermore, to assess how many individual mentions a curator needed to curate in order to determine a grouped molecular interaction as a true positive, we sampled 100 random true positive grouped interactions that contained at least 5 mentions of that interaction from our curated data. We then assessed the proportion of individual mentions that were correct in each grouped molecular interaction.

3.4 Results and Discussion

3.4.1 (i) Building a topic-specific corpus

Pain terms dictionary

Figure 3.2 displays the final counts of pain-specific and pain-relevant terms and synonyms for the 12 categories of pain terms. In total there were 583 terms (235 pain-specific and 348 pain-relevant) and 3,144 case-sensitive synonyms (1,506 pain-specific and 1,638 pain-relevant). We note that there are high proportions of pain-specific 'disorder' and 'pain type' pain terms. This is perhaps due to the fact that the field of pain is more succinctly encapsulated in these categories. We note, that while in this study the pain terms dictionary created was sufficient for building an accurate corpus of pain documents, future recreations of our approach in other biomedical fields may be better suited to using existing ontologies and controlled vocabularies (such as, for example, SNOWMED CT[307]).

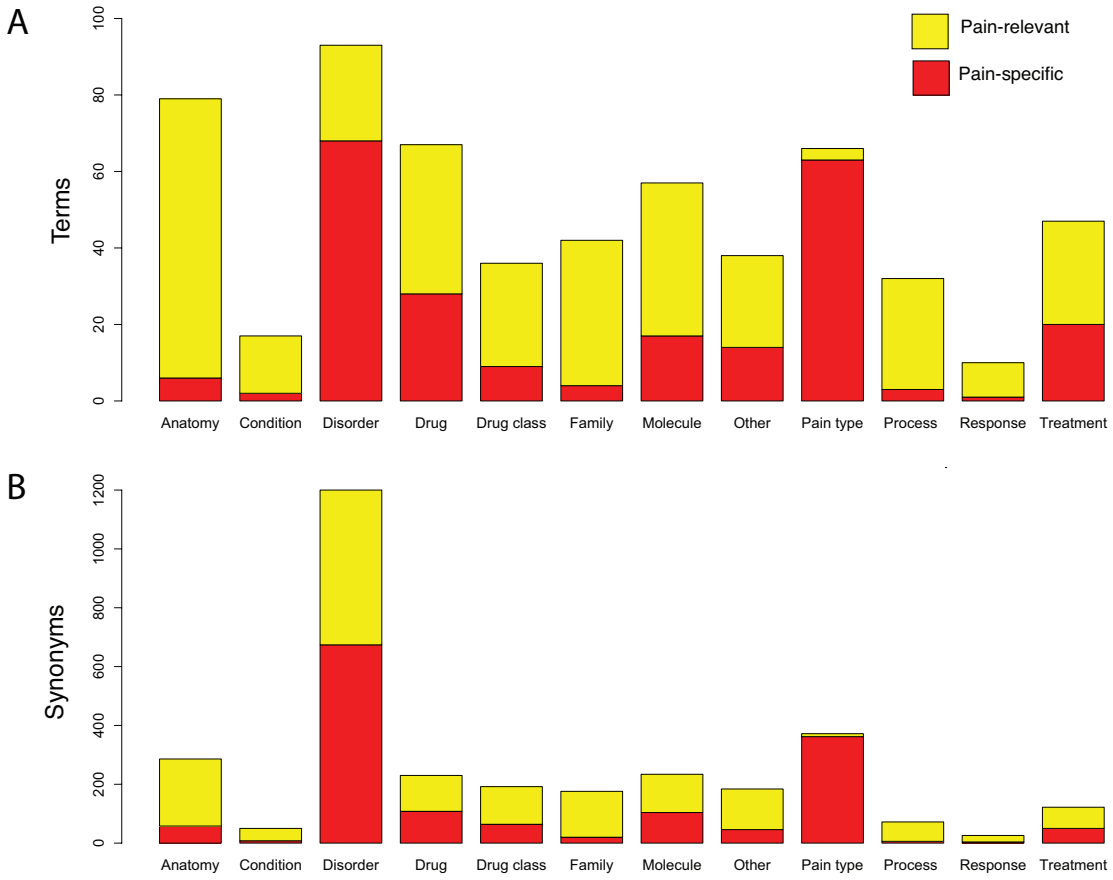


Figure 3.2. Pain dictionary summary statistics.
(A) represents the numbers of pain-specific and pain-relevant terms in the pain dictionary for each category of pain term. (B) shows the numbers of pain-specific and pain-relevant synonyms in the dictionary for each category of pain term.

Document Retrieval

The total number of matches in different document sections of pain-specific and pain-relevant terms for each pain term category is shown in Figure 3.3. There were matches of pain-specific and pain-relevant terms in all of the 12 categories with a large proportion coming from disorder terms. Altogether there were 4,645,861 pain term matches, 2,548,287 pain-specific and 2,097,574 pain-relevant. Matches of pain-specific and pain-relevant terms were made across each type of document section in P1 with a large proportion being made in the abstracts. However, while this distribution of terms across different textual sections is representative of our corpus, we would expect that the proportion of terms found in the body of a document would be far greater had we had excess to full

text not available in our Medline dataset. For instance, if we exclude Medline documents from our sectional analysis, 91% of matches are found in the body.

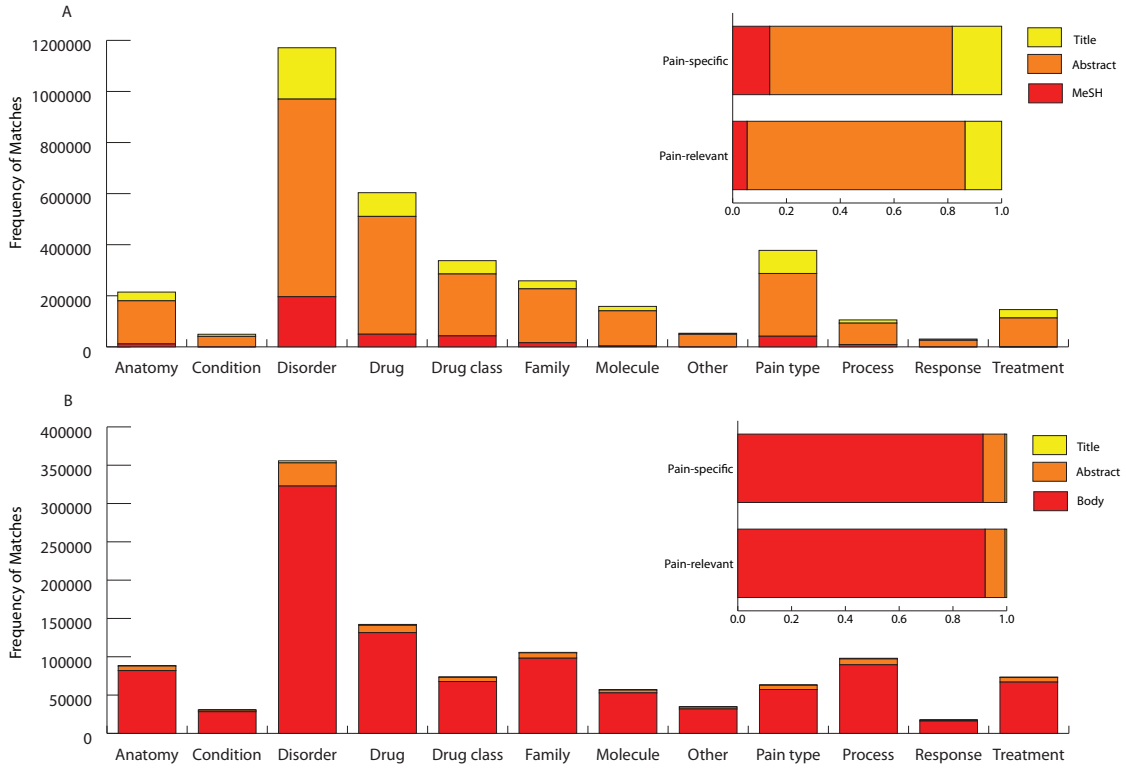


Figure 3.3. Pain term matches.
Pain term matches from Medline (A) and open access PMC documents (B) in each type of document section across the 12 pain term categories are displayed. The overall percentage of pain-specific and pain-relevant terms from Medline and open access PMC documents are shown for each type of document section. "Body" represents full-text outside of abstracts and titles. MeSH refers to textual document tags used by PubMed articles in indexing.

Table 3.1 displays the top 10 reported pain terms in P1, ordered by the number of documents that they were reported in. Nine out of ten terms were pain-specific and they accounted for roughly 25% of all matches. From our pain-specific matches there were 765,692 documents (732,826 Medline and 32,866 PMC) that matched at least one term. Of the 32,866 PMC open access documents that were part of P1, these composed roughly 17% of the entire PMC open access corpus in comparison to 7% of Medline from 732,826 documents. It is likely that this disparity was caused by a greater availability of text accessible for matching terms from our pain dictionary in full text documents. This perhaps indicates that many documents that are pain relevant in Medline have been missed, as we have not had access to terms located in associated full text.

Pain Term	Category	Pain Specific	Frequency	Documents
Pain	Disorder	Yes	627,644	247,312
Anaesthesia	Pain type	Yes	190,376	115,614
Analgesic	Drug class	Yes	112,703	61,223
Headache	Disorder	Yes	118,956	50,249
Brain haemorrhage	Disorder	No	85,702	45,214
Opioid	Drug class	Yes	77,921	33,486
Morphine	Drug	Yes	119,985	33,337
Analgesia	Pain type	Yes	64,777	31,982
Palliative	Treatment	Yes	51,401	27,536
Abdominal pains	Pain type	Yes	33,916	25,062

Table 3.1 Top reported pain terms in P1.
Pain term refers to the individual pain term and all it's synonyms. Pain terms are pain-specific (yes) or pain-relevant (no). Pain term categories are defined in supplementary file 1. 'Frequency' refers to the total number of times that that term was mentioned. 'Documents' refers to the number of documents that that term was mentioned in.

The overall pain document relevancy scores are summarized in Figure 3.4. The analysis of this scoring scheme showed that documents with the MeSH term 'Pain' as a major term scored significantly higher than those that had 'Pain' as a minor MeSH term when using a Wilcoxon/Krustai-Wallis test ($Z=-49.326$ and $p<0.001$). Further information is provided in supplementary file 4. This initial evaluation shows that as well as being able to retrieve pain documents, we can also differentiate between these in terms of their overall relevance to pain using our scoring system. As well as this overall pain relevancy score, the pain category and individual pain term scores allow for exploration of specific aspects of pain. Indeed, our evaluation of the pain terms present in 50 reported pain documents showed 100% precision and 89.6% recall (see Table 3.7), highlighting that we have been able to extract individual pain concepts with high accuracy.

However, we note from Figure 3.4 that documents where full text was used scored higher than articles with only abstracts and titles available, highlighting a potential issue in our scoring method when using documents of varying textual lengths. At present, our method partially addresses this by scoring terms matched in the body of an article with 0.25, in comparison to terms scored with 1 in the abstract and 2 in the title. However, in future corpus generation the section weights could be adjusted to produce a score that does not bias full text articles into being scored higher.

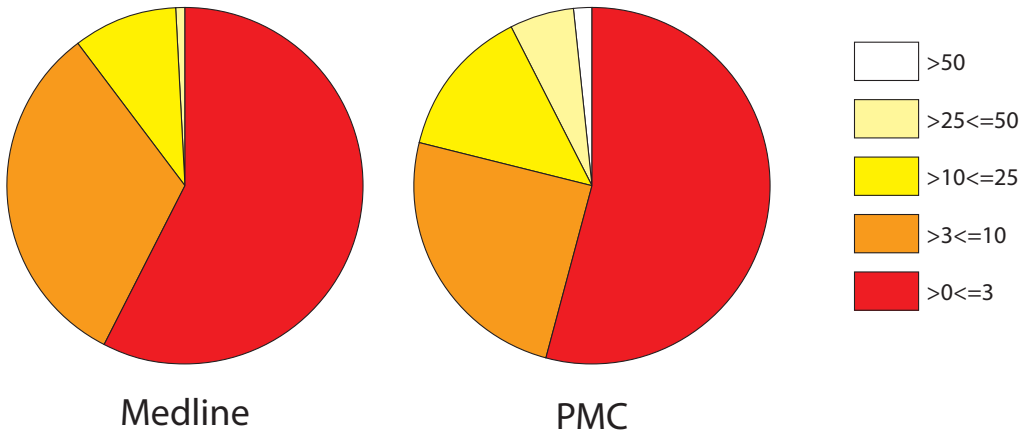


Figure 3.4 Document pain relevancy scores.
Pie charts represent the overall pain scores for Medline (abstracts and titles) and open access PMC documents. Pain relevancy ranges between 0-3, 3-10, 10-15, 25-50 and >50.

3.4.2 (ii) Data extraction

Event chains

In total there were 1,578,654 event chains from the BioContext database present in P1. After grouping these event chains, there were 356,499 unique event chains, with 261,438 single events, 93,271 containing two participants (i.e. molecular interactions) and 1,790 involving more than two participants. Table 3.2 shows the frequencies of single events, molecular interactions and interactions with more than two participants involving proteins normalized to humans, mice, rats and other species. Human, mouse and rat proteins incorporated 44% of unique single events and 37% of unique molecular interactions with the other proteins in event chains being normalized to 1,230 different species. As humans, mice and rats are the standard animal species studied in pain molecular research, these results show that there are large amounts of useful data available for curating a pain relevant molecular interaction database.

Table 3.3 shows the number of grouped event chains involving events of protein metabolism, binding, localization, phosphorylation and regulation. We found large numbers of regulatory and binding events involved in all types of event chains and high numbers of gene expression events in single events.

Involving only	Single Events	Molecular Interactions	More than 2 Participants	Total
Human Proteins	45,731	14,568	262	60,561
Mice Proteins	41,671	12,956	230	54,857
Rat Proteins	26,736	7,369	132	34,237
Other Proteins	147,300	58,378	1166	206,844
Total	261,438	93,271	1790	356,499

Table 3.2 Event chains from P1.
Event chains are shown for those involving only human, mice, rat and other proteins as their cause and/or theme. Event chains are divided into single events, molecular interactions (i.e. those containing two participants) and event chains with more than two participants. Total numbers of events chains by number of participants and by proteins involved are displayed.

Event Type	Single Events	Molecular Interactions	More than 2 Participants
Binding	33,358	37,291 (37,315)	897 (919)
Gene expression	78,255	12,223 (12,482)	95
Transcription	12,158	1,238	10
Localisation	27,329	5,355 (5,368)	50
Phosphorylation	7,360	1,782 (1,784)	37
Protein catabolism	5,296	467	6
Positive regulation	69,846 (75,064)	32,222 (35,740)	1,174 (1,650)
Negative regulation	52,754 (54,729)	13,698 (14,870)	541 (624)
Regulation	41,137 (42,422)	19,271 (19,783)	468 (551)

Table 3.3 Event types involved in event chains.
Non-redundant frequencies of single events, molecular interactions (i.e. those containing two participants) and event chains containing more than two participants are displayed for each of the 9 categories of events used by the event extractors. The numbers in brackets represent the total number of occurrences of that event type where some events have duplicate (redundant) event types, e.g. '**positive regulation** of **positive regulation** of *protein A*'.

In total there were 37,628 grouped event chains that were reported negatively at least once. 24,142 of these potentially represented contradictions with some mentions of a grouped event chain being reported negatively and others positively. Of those event chains that were reported more than once there were only 25% (369/1457) that were reported entirely negatively (Figure 3.5). A total of 31,275 (26,268 single events and 4,909 molecular interactions) grouped event chains were reported speculatively at least once. Of those event chains that had been reported more than once, 277/1207 molecular interactions and 382/20,931 single events were reported entirely speculatively

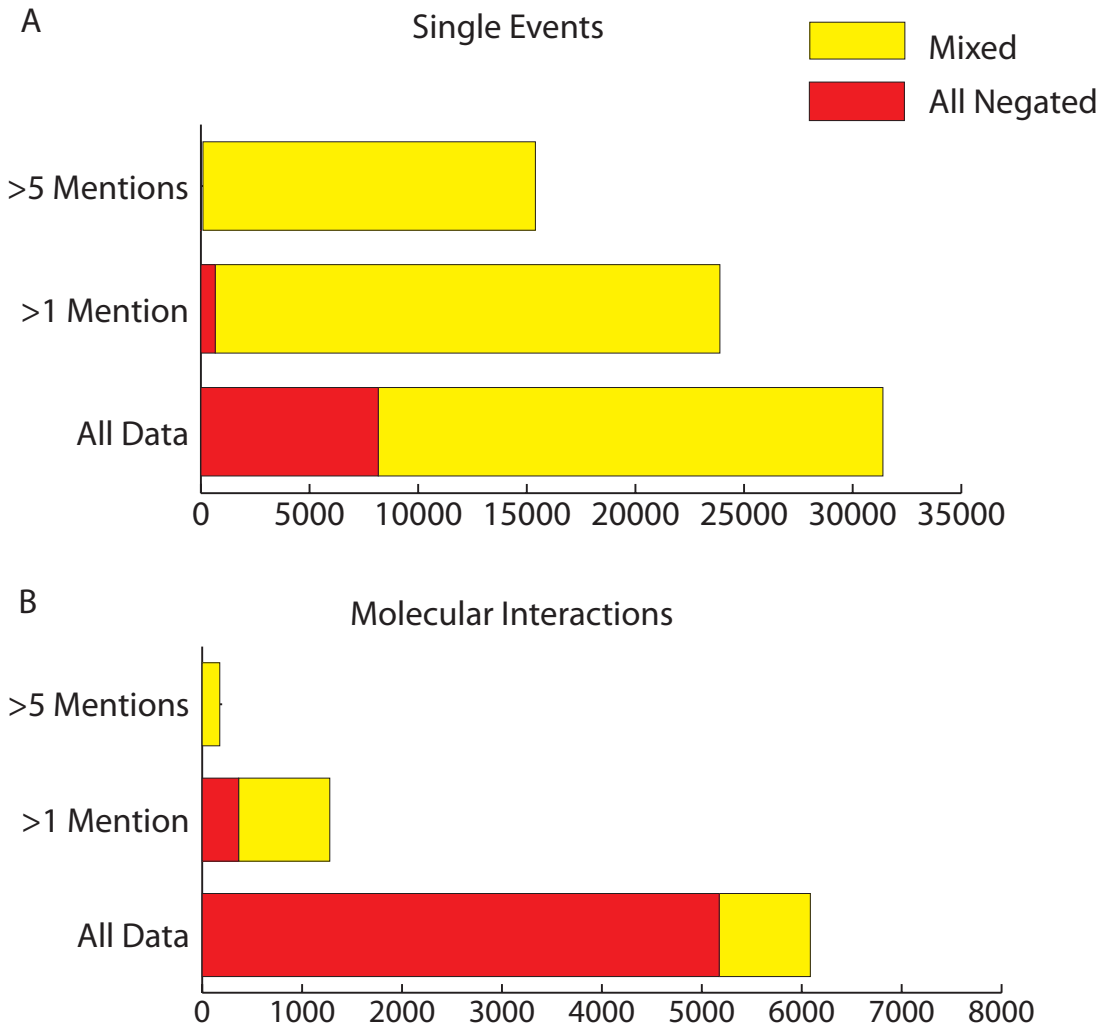


Figure 3.5 Number of negated event chains.
"Mixed" refers to event chains that have been mentioned both negatively and positively. "All negated" refers to the number of event chains that are only mentioned negatively. Proportions of mixed and negated data are shown for all molecular interactions and single events that have been mentioned more than once or more than five times.

From the 356,499 grouped events chain, 172,294 were mapped to at least one anatomical region. Table 3.4 exhibits the top 10 anatomical regions (of 2,774 total) associated with event chains retrieved from P1; these made up about 27% of all anatomical mentions in our pain dataset. We note high numbers of immune anatomical structures, which is not unexpected with pain-related data [290].

Name	Frequency
Neurons	37,666
Plasma	36,969
Brain	31,775
Blood	19,291
T cells	16,092
Liver	15,650
Spinal Cord	14,453
Macrophage	13,409
Neuronal	12,368
Nerve	11,355
Total	761,990

Table 3.4 Top 10 anatomical regions associated with event chains.
Anatomy terms are extracted using GETM.

From sentences used to extract event chains in P1 we were able to map 2,997 mutations to proteins involved in single events and 721 mutations to proteins involved in molecular interactions.

Table 3.5 provides an overview of the overall pain relevancy scores calculated for each unique event chain in our dataset involving human, mouse or rat proteins (the most commonly studied animal models in pain research) and excluding self-interactions (e.g. "**Binding** of *Tprv1* and *Tprv1*"). The mean overall pain relevancy score for these was 0.33, with a median of 0.15 and standard deviation of 0.64. There were 25,593 medium pain ranked (between 0 and 1 in overall pain-relevancy) and 2,646 highly relevant (greater than 1 in overall pain relevancy) unique pain molecular interactions.

Pain Relevancy Score	Single Events	Molecular Interactions	More than 2 participants	Total
Low (0)	22,623	9,240	191	32,054
Medium (>0,<=1)	62,640	25,593	520	88,753
High (>1)	28,875	2,646	42	31,563

Table 3.5 Overview of overall pain relevancy scores for unique event chains involving human, mouse or rat proteins and excluding self-interactions.
We show the frequency of unique single events, molecular interactions (i.e. two participants) and event chains with more than two participants with a low (0), medium (>0,<=1) or high (>1) overall pain relevancy score.

In total we matched 6,792,990 disease terms in 618,487 documents from P1, allowing 3,041,109 disease terms to be mapped to 1,402,560 event chains. Table 3.6 displays the top diseases associated with P1 documents containing event chains. While generic classes of disease terms, such as 'disease', 'injury' and

'inflammation', featured in the top 10, there were also high numbers of 'diabetes', 'pain', 'depression', 'cancer' and 'HIV' associated event chains. We note that these have a large neuropathic pain component.

Disease Name	Disease term mentions
Disease	135,367
Pain	122,233
Cancer	117,041
Inflammation	101,059
Injury	59,237
Infection	57,481
Diabetes Mellitus	50,705
Stress	41,056
Depression	39,762
AIDs or HIV infection	30,872
Total	3,041,109

Table 3.6 Top diseases associated with documents containing event data.
Here we report the total number of disease term mentions in documents that contain at least one event chain.

Data extraction evaluations

Table 3.7 displays the results for all of the new evaluations of methods used in this study.

Our mutation-to-protein linker (of co-occurring mentions in sentences) extension for MutationFinder showed precision of 97.3% and recall of 72% to give an F score of 82.7%. The mutation-to-protein linker also showed a 99.1% true negative rate to give an accuracy of 90.6%. Improvements to recall can be facilitated by extending our library of regular expressions. At present our tool is only able to normalize proteins to mutations that are both denoted in the same sentence; however, in our analysis we noted a large number of proteins associated with mutations that were defined outside of the sentence. This limitation, as well as the accuracy involved in extracting the original event chain and the mutation mention itself, is important to consider when using such data.

Tool	Data	True Positives	False Positives	True Negatives	False Negatives	True Negative Rate	Precision	Recall	Accuracy	F Score
Pain Terms (LINNAEUS)	50 Documents	3,803	0	N/A	443	N/A	100	89.6	N/A	94.5
Mutation to Protein Linker	100 Event Chains	36	1	109	14	99.1	97.3	72	90.60	82.7
Pain Relevancy (>50 confidence)	100 Event Chains	78	22	N/A	N/A	N/A	78 (92 expected)	N/A	N/A	N/A
Pain Relevancy (<=50 confidence)	100 Event Chains	39	61	N/A	N/A	N/A	39 (20 expected)	N/A	N/A	N/A
Disease Terms (LINNAEUS)	25 Documents	345	16	N/A	15	N/A	95.6	95.8	N/A	95.7
Disease Relevancy (>50 Confidence)	100 Event Chains	84	16	N/A	N/A	N/A	84 (88 expected)	N/A	N/A	N/A
Disease Relevancy (<=50 Confidence)	100 Event Chains	30	70	N/A	N/A	N/A	30 (13 expected)	N/A	N/A	N/A

Table 3.7 Evaluations of TM software used.
For each tool evaluated we display a summary of the data used in the evaluation (either documents or event chains), and the frequencies of true positives, false positives, false negatives and true negatives for each tool wherever possible. From the true positives, false positives, false negatives and true negatives we calculated the true negative rate, precision, recall, accuracy and F score of each tool where applicable. In pain and disease relevancy we also note the expected precision calculated from the average relevancy score of each term in the respective evaluation.

In the evaluation of pain terms relevant to event chains with scores above 50, we judged 78/100 as relevant. These results are lower than the predicted 92/100 taken from the average relevancy score across the 100 event chains evaluated. We noted that 'molecule' and 'family' category pain terms were more likely to be irrelevant to an event chain when mentioned outside of the sentence the event chain was denoted in. By contrast, the evaluation of pain terms relevant to event chains with scores below 50 showed that 39/100 relevant pain term-event chain pairs, whereas the expected value was 20/100. The higher than expected number was mainly caused by 'disorder' pain terms that, although mentioned in distant sentences to the event chain, were still perceivably relevant.

Judging from 25 documents, our disease term matching showed a precision, recall and F score of 96%. Our evaluation of the linking of these terms to event chains in which relevancy scores were above 50 showed 84/100 relevant disease term-event chain pairs. The average predicated relevance score across each linked disease term-event chain pair was 88, indicating that our high relevance predictions were fairly accurate. However, in the evaluation of relevancy of disease terms-event chain pairs with scores below 50 we found 30/100 disease terms to be relevant compared to the 13/100 predicted. As with our low pain relevancy evaluation findings, we found that disease terms could still be relevant to an event chain even if they were mentioned in paragraphs and sentences far away from the event chain in the text. These issues for both pain and disease relevancy can be resolved by adjusting each approach to more closely reflect the likelihood of actual disease or pain relevancy.

Because we are using event chains directly from BioContext, we expect that event extraction precision and recall will be consistent with previously reported ones [227, 286]. Indeed, benchmarking against a small manually curated gold standard of five full text documents reported similar precision, recall and F-score of 35%, 58% and 44%, respectively. A detailed analysis is available in supplementary file 5. A comparison of TM data against existing generic manually curated databases is difficult as there are no extensive pain-focused resources that can be used directly. Instead, we have explored the intersection between our TM results and iRefIndex, a large generic molecular interaction database containing interactions

from numerous species sourced from various individual manually curated databases [308]. As expected, the overlap is not significant (only 21 interactions) given the difference in the criteria used to extract and represent the data between datasets. We have provided this analysis in supplementary file 6.

To determine whether genes known to be related to pain were enriched in the P1 extracted events, an enrichment analysis was performed (Table 3.8). In total 280/297 genes in the Pain Genes DB were mentioned in at least one of our event chains. These genes were mentioned in 4.54% of event chains in P1, which was more than double the 2.14% found in R1. Fisher's exact test confirmed P1 to be enriched for these genes with an odds ratio of 2.18 and a highly significant P value (<2.2e-16), suggesting that the overall dataset of molecular events recovered from our corpus are relevant to pain.

Corpus	Event chains mentioning a pain gene	Event chains not mentioning a pain gene	Total event chains	% of event chains with a pain gene
P1	71,685	1,506,969	1,578,654	4.54
R1	47,998	2,196,618	2,244,616	2.14

Table 3.8 Pain genes enrichment analysis.
P1 represents the pain corpus and R1 represents the randomly generated generic corpus. We show frequencies of event chains mentioning a gene from the Pain Gene DB for each corpus and event chains not mentioning a gene from the Pain Gene DB. We also display total event chains for each corpus and the percentage of event chains that contain genes from the Pain Gene DB. Fisher's exact test showed significant enrichment of pain genes within P1, having an odds ratio of 2.177008 with a p value <2.2e-16.

Of the 17 genes from the Pain Genes DB that were not mentioned in event chains from P1, 15 were mentioned in BioContext event chains extracted from Medline and PMC documents there were not in P1. To determine why these genes were found outside of P1, we selected five random articles that mentioned one of these genes in an event chain for each of the 15 genes (75 articles in total). Of the 75 articles, they were all pain-irrelevant, with a small number mentioning pain-relevant terms (e.g. GABA). Four of the genes did not have a correctly reported mention in the articles sampled, with the majority of the errors coming from erroneous gene name normalization (see 'Manually curated data').

3.4.3 (iii) Availability and visualisation for manual curation

Data availability

We uploaded the data retrieved from our investigation onto wiki-pain.org. At the core of the wiki are the 'INT' pages (Figure 3.6) used to display each grouped event chain relevant to pain. Within each page summary, contextual data is displayed at the top, providing a concentric view of that event chain's relevance to pain. Beneath the summary information are the sentences where the data was extracted from highlighting entities extracted using a color-coded key. Each sentence then has its own summary providing links back to its original source among other useful contexts that can be used for further investigation.

The INT pages are named using INT IDs to enable linking across the site. Most INT links stem from summary tables created to help guide users to the most relevant information. For example, the entry page on the wiki contains summary tables of all interactions and single events in the database ordered by their overall relevance to pain. Other summary tables can be found on gene pages, journal pages, event type pages, disease term pages etc. linking interactions specific to page type e.g., on the G:60628 (CXCR4) page only event chains mentioning this gene are displayed.

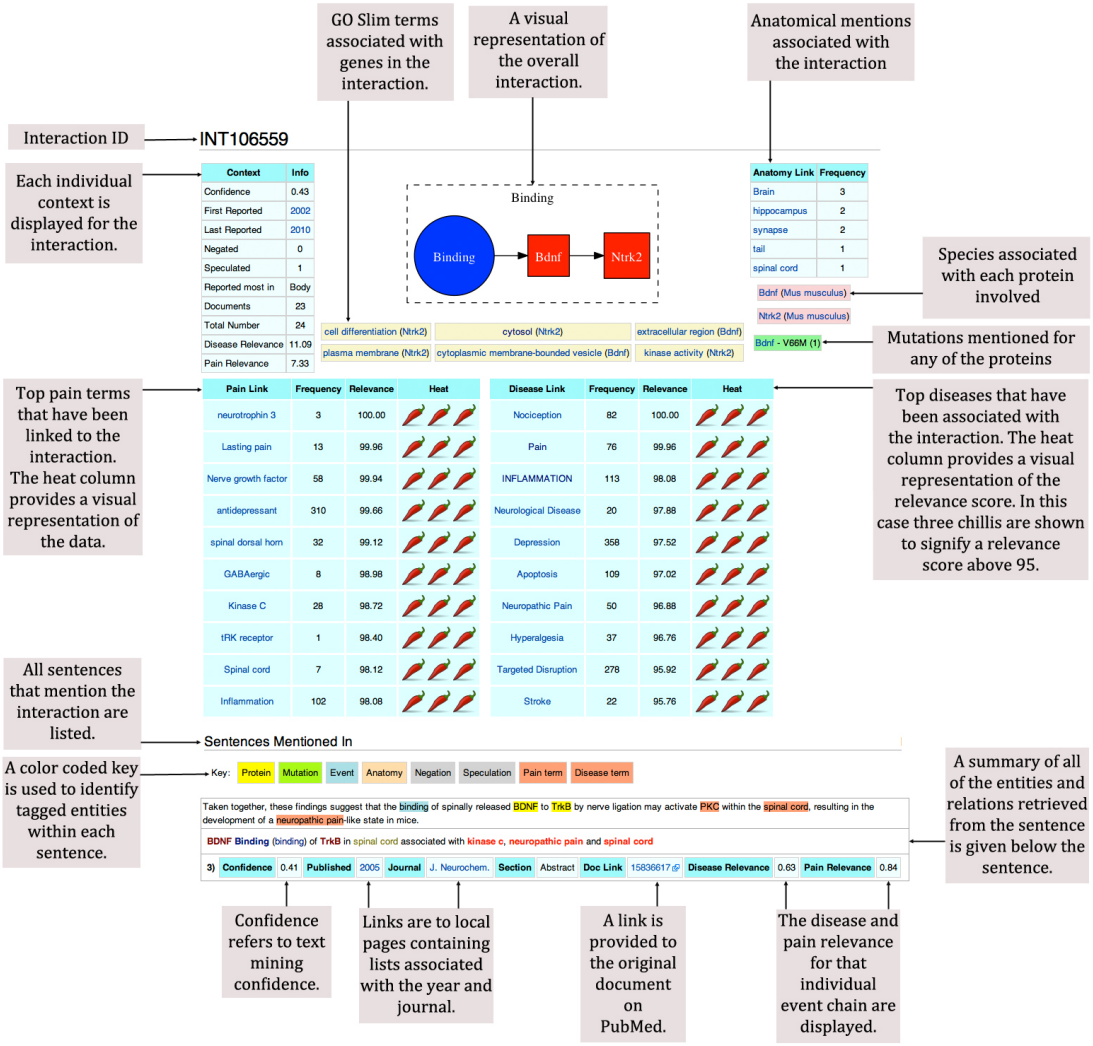


Figure 3.6. Example of a typical molecular interaction in wiki-pain.org.
We have removed the page borders that are typical of a Mediawiki interface and annotated each region of the page that we have designed and is novel. All 'INT' pages on wiki-pain.org follow the same framework including single events and event chains containing more than two participants. The specific page shown can be viewed by searching 'INT106559' on wiki-pain.org.

Manually curated data

The manual curation of the top 1,500 grouped molecular interactions showed 613 true positives and 887 false positives. This means that grouped molecular interactions have a precision of 40.87% before they have been curated. However, if we set a cut-off of 50% for the TM confidence (coming from BioContext), our precision more than doubled to 84.17% (117 true positives and 22 false positives). We also found that unique interactions mentioned in more than one document

were more likely to be a true positive, with precision of 59.71% (252 true positives and 170 false positives) in comparison to 33.48% (361 true positives and 717 false positives) mentioned in only one document. Therefore, for supporting curation, it makes sense to prioritize using high confidence interactions only.

Overall, the 613 true positives included 487 different genes, with 161 human proteins, 170 mouse proteins and 156 rat proteins. These genes could be grouped into 351 homologues (by using their homologene IDs), indicating a variety of different types of proteins in the curated data and not simply those proteins synonymous between species. Table 3.9 shows the top 10 homologues ordered by frequency of unique molecular interactions that each is involved in. We also found 90/276 homologues and 61/297 of the previously identified pain relevant genes from the Pain Genes DB in our manually curated dataset. These results indicate that we have identified additional 261 homologous sets of genes that could potentially be associated with pain, including 426 specific genes.

Analysis	TPs before	TPs after	FPs before	FPs after	Agreed	Disagreed	P(A)	P(E)	K
Intra	18	12	32	38	42	8	0.84	57.3	0.427
Inter	27	22	23	28	45	5	0.9	49.5	0.802
Overall	45	34	55	66	87	13	0.87	51.6	0.731

Table 3.9 Manual curation evaluation.
We evaluate the quality of our manual curation using an intra analysis (data quality is evaluated by the same curator), an inter analysis (data quality originally curated by a different curator is evaluated) and these two are combined to show an overall evaluation of our manual curation. We present the number of true positives (TPs) and false positives (FPs) in the original curation (before) and the new curation results (after). Results that were the same were marked as 'Agreed' and those that were different, 'Disagreed'. The absolute agreement, P(A), was calculated from the proportion of agreement (agreed/disagreed). Cohen's Kappa coefficient (K) was calculated from the proportion of agreement, corrected for expected agreement by chance (P(E)), such that $K = ((P(A) - P(E))/(1-P(E)))$.

Of the false positives, we noted commonly occurring causes such as incorrect protein normalization to Entrez Gene IDs and event mismatches. We also noted a large number of false positives caused by abbreviations tagged as proteins that were in fact other types of entities (e.g. "long term potentiation (LTP)" that was erroneously normalized to the "LTP gene"). This problem can be resolved by better integration of biomedical entity tagging tools to filter out instances of data by pre- or post-processing that which had been previously defined as another entity type.

To determine how many of the true positive molecular interactions were present in existing manually curated databases, we checked protein pairs from our data against MiMi [309], a large online database incorporating multiple data sources (BIND [310], HPRD [143], IntAct [311] etc.)), through the MiMi API. In total we retrieved 59 protein pairs in MiMi from 505 present in our dataset, indicating that the majority of our true positive (curated) data has yet to have been incorporated into the existing curated databases.

In the assessment of the proportion of individual mentions of a grouped event chain that were correct we removed 12 grouped interactions from the analysis that had previously been reported as true positives and after review were determined to be false positives. From the remaining 88 grouped interactions we found a total of 335 correctly identified individual mentions against 105 incorrectly identified mentions, highlighting on average 3 true positives in the top five mentions of a grouped interaction. These results show that for grouped interactions a high proportion of the top five individual mentions are correct and therefore curators do not need to spend added time curating each and every individual mention when the overall grouped molecular interaction is a true positive.

Having manually curated 4% of all extracted interactions, we sought to infer what proportion of the uncurated interactions were likely to be true positives. The TM confidence score for each interaction (deduced by BioContext) separates the true from false positives relatively well. The true and false positive interactions have a mean confidence score of 0.3 and 0.1 respectively and are significantly different ($p<0.0001$). We therefore fit a generalized linear model following a binomial distribution with a logit link function to the confidence scores from the curated data, so that we can assign a probability of being correct to the remaining 36,732 grouped interactions. We found that interactions with a TM confidence score above 28% were likely to be true positives. Using this measure, we can predict that 5,816 of the remaining interactions are more likely to be true positives than false positives (see supplementary file 7 for further details on these calculations). For this study, it took on average one working day for a curator to curate 250 molecular interactions. Therefore, we can assume that

it would take one curator a further 23 days to review the remaining predicted true positive data (those with a TM score above 28%).

Manual curation quality

Table 3.10 shows the review of our manual curation quality. The intra agreement rate was 0.84, while the inter agreement rate was 0.9 to give an overall agreement rate of 0.87. Cohen's Kappa coefficient [146] showed a moderate intra agreement rate of 0.43, a substantial inter annotator agreement rate of 0.80 and a substantial overall agreement rate of 0.73. Upon review of the curation results that were in disagreement, 7 of the 8 new curation results in the intra analysis were correct. Four were caused by incorrect normalization to protein IDs and one by incorrect protein tagging and it is likely that these were identified in the second attempt due to increased experience in curating pain related proteins. A further two were attributed to event mismatches. In the inter analysis, 5/5 of the new curation results were correct and the original curation errors were again due to erroneous protein normalization and also more complex interactions that were perhaps more difficult to curate correctly. While this assessment of our manual curation quality showed that our curated results were of a high standard, they also show that it is likely that some of the curated data that has not been reviewed is likely to be incorrect. Therefore, in order to be sure that the final curated results used in subsequent analyses are entirely accurate, it is important to perform multiple curations.

Homologue ID	Symbol	Frequency
1876	NGF	53
37368	OPRM1	50
723	POMC	45
12920	TRPV1	44
88337	CALCB	40
4528	PENK	39
502	IL6	27
599	CRH	22
496	TNF	19
4537	PNOC	16

Table 3.10 Top 10 homologues appearing in our manually curated data.
These have been ranked by frequency of unique molecular interactions that each homologue is involved in in our manually curated data. Homologue ID refers to the ID used by NCBI homologene database (http://www.ncbi.nlm.nih.gov/homologene).

3.5 Conclusions

In this study we have demonstrated that a pain-specific contextual molecular interaction database can be created using TM to rapidly generate content and support manual curation to confirm its accuracy. The whole process of building the pain-relevant corpus, extracting and contextualizing the interactions and curating the data took just two months, which is in contrast to a typical fully manual procedure that may take years. We have used the existing state-of-the-art in TM methods to generate the core data used in our curation (e.g. corpus generation using LINNAEUS and event chains and context taken from BioContext). Therefore, the approach used in this study is not limited to the pain domain and would potentially suit many other biomedical fields that consider molecular interactions a focal point of the research. For example, the approach could be repeated for another topic by applying a relevant dictionary to generate a corpus in the same way as for pain and using this as a basis for data extraction and curation. To facilitate such instantiations of our approach in other fields we have therefore provided a full list of methods used in this study on wiki-pain.org/downloads.

As well as the existing TM methods and data used in this study we have also proposed a i) new method for scoring documents for their relevance to pain and any individual concepts; ii) new methods for determining the relevance of an event chain to pain or disease terms and iii) a novel sentence based mutation-protein linking extension to MutationFinder. Furthermore, wiki-pain.org is the first extensive pain-specific molecular interaction database that researchers can use to explore context specific pain data extracted from the literature.

In the future, we wish to continue curating the grouped molecular interactions for pain and to expand this curation process to each individual context to ensure that all of our data displayed is accurate. We then plan to investigate more closely the biological implications of the correctly distinguished data. For example, it would be interesting to compare and contrast the most connected and frequently occurring proteins between different pain-related disorders and anatomical regions. Furthermore, our procedure has been carefully designed so

that additional context can be built into our database and adding aspects such as chemical interactions will be considered.

3.6 Acknowledgments

The authors would like to thank Louise Riley (Neusentis), Victoria Albrow (Neusentis) and Mike Kennedy (Neusentis) for their help with the curation, and Rie Suzuki (Neusentis) for helping assess the quality of the pain terms dictionary. We would also like to thank Alex Gutteridge (Neusentis) for useful comments and feedback throughout the investigation and Panagiotis Papastamoulis (University of Manchester) and Xiaowei Jiang (University of Manchester) for assistance with the pain relevancy equations and statistics.

CHAPTER 4

The pain interactome: Connecting pain specific protein interactions

4.1 Abstract

Understanding the molecular mechanisms associated with disease is a central goal of modern medical research. As such, many thousands of experiments have been published that detail individual molecular events that contribute to a disease. Here we use a semi-automated text mining approach to accurately and exhaustively curate the primary literature for chronic pain states. In so doing, we create a comprehensive network of 1,002 contextualised protein-protein interactions (PPIs) specifically associated with pain. The PPIs form a highly interconnected and coherent structure, and the resulting network provides an alternative to those derived from connecting genes associated with pain using interactions that have not been shown to occur in a painful state. We exploit the contextual data associated with our interactions to analyse sub-networks specific to inflammatory and neuropathic pain, and to various anatomical regions. Here, we identify potential targets for further study and several drug-repurposing opportunities. Finally, the network provides a framework for the interpretation of new data within the field of pain.

4.2 Introduction

Acute pain has evolved as a key physiological alert system for avoiding noxious stimuli and protecting damaged regions of the body by discouraging physical contact and movement[91]. This form of pain is crucial; in its absence, e.g. in those with congenital insensitivity to pain, we are more prone to damaging or non-protective behaviors that can hinder our quality of life[312, 313]. Conversely, persistent or chronic pain can be similarly debilitating with those affected typically suffering psychological disturbance and significant activity

restrictions[314]. The incidence of chronic pain is widespread across the global population, with estimates in the adult general population of 12.7-29.9% in developed and 14.5-33.9% in developing nations[94]. Pharmacological therapeutics such as opioids, non-steroidal anti-inflammatory drugs (NSAIDs), such as COX-2 inhibitors, are often prescribed as the standard treatment regimens for chronic pain sufferers[96, 315], but while these and a huge range of other treatment options are available, their efficacy often proves at best modest and their use is limited by unwanted side effects[316, 317]. There is therefore an urgent need to better understand the molecular systems that mediate chronic pain and to use this knowledge to develop improved therapeutics.

Pain researchers have published hundreds of thousands of articles, many of which detail knowledge of the molecular interactions involved in pain. However, digesting and utilizing this knowledge is impractical without the use of text mining. In our previous work[318], we used state of the art computational methods to retrieve molecular interactions associated with pain from the primary literature: the whole of Medline and open access articles in PubMed Central (PMC). These data are catalogued at wiki-pain.org, which contains 93,271 molecular interactions derived from 765,692 pain-related articles. Each interaction is annotated with detailed contextual information such as anatomy, associated point mutations and disease relevance. However, as fully-automated text-mining results can be error prone[218], we implemented a novel strategy to curate mentions of protein-protein interactions (PPIs) grouped from multiple publications to create the first pain-specific dataset of interactions. Through ongoing curation, this dataset now contains over 1,000 unique contextualised PPIs[318].

Here we explore the relevance and accuracy of our pain related PPIs using network and functional enrichment analyses and gene bias assessment methods. To emphasize the quality and effectiveness of this approach of sourcing interaction data we provide comparisons with pain related interaction networks derived from gene expression data, a manually curated list of pain genes and the known targets of pain drugs. Our results demonstrate that a semi-automated text-mined interaction network allows us to interpret the sum knowledge of the

biomedical domain of pain in an integrated manner, providing a more complete portrait than is possible from other common means of inferring disease networks. The network has immediate utility to researchers in the field as a framework for the interpretation of new findings and high-throughput 'omic datasets. Importantly, this approach has a broad applicability to other diseases or syndromes to which a combination of text-mining and network biology might be applied.

4.3 Methods

4.3.1 Data availability

The data generated in this study is available in supplementary tables 1-24.

4.3.2 The curation procedure for PPIs

In a previous study detailing our text-mining methodology, we curated over 1,500 unique PPIs involving mouse, rat and/or human proteins ranked by their overall relevance to pain[318]. Raw interactions were extracted from text automatically and displayed on wiki-pain.org to be verified by an expert. For a PPI to be considered accurate, the proteins, including underlying species and associated Entrez Gene IDs, and interactions had to have been extracted accurately in at least one instance when all mentions of that interaction were grouped together. We continued curating interactions in this study following these guidelines. We focused on those interactions that had a text mining confidence score above a threshold (28%) that was empirically determined to be a good indicator of true-positive interactions[318]. We grouped orthologous proteins from rats, mice and humans (using NCBI Homologene IDs) and simplified the interactions to either positive regulation, negative regulation, regulation or binding to remove superfluous data.

PPIs for the neuropathic and inflammatory pain tasks were curated in the same way as with general pain associated interactions, with the addition of one more condition: each interaction had to have a specific association with the relevant

pain disorder, e.g. 'activiation of c-Jun in DRGs induces VIP and NPY upregulation and contributes to the pathogenesis of neuropathic pain'[319]. Those interactions that were selected for curation had neuropathic or inflammatory pain relevance above 90%[318] and, again, a text mining confidence threshold of 28%. The involvement was noted as being either part of the mechanism of that disorder or having an inhibitory effect on it. We note that interactions were curated from literature published over decades and so any changes in the formal definition of these indications used may not have been accounted for.

4.3.3 Network analysis

Networks were analyzed using iGraph for R[246] and visualized using Cytoscape 3.0[245].

Enrichment analysis of proteins was performed using Fisher's exact test to determine proteins that had a statistically significant number of interactions in the sub-graph under study compared to the relevant main graph. This follows similar implementations of Fisher's exact test as described in Poirel *et al*[320] and Wuchty[321]. Enrichment was determined by calculating the number of interactions each protein features in the sub-graph (a) and those that it does not (c), as well as the number of interactions each protein features in, in a comparison main-graph (b) and those that it does not (d). The probability a protein is enriched is then determined using the hypergeometric distribution, such that

$$p = \frac{\frac{a+b}{a} \times \frac{c+d}{c}}{\frac{a+b+c+d}{a+c}}$$

The hypergeometric distribution assumes that proteins appearing in interactions in the main graph and sub-graph are equally likely and thus if p is below 0.05 we can reject this null hypothesis. iRefIndex (version 06062013) was used as a source of generic PPIs in order to construct a comparison main graph representative of the human interactome. iRefIndex is a large generic molecular interaction database containing interactions that have been sourced from numerous

manually curated databases[308]. Using only human proteins from this database, the network contains 14,818 nodes and 167,413 edges, with an average degree of 22.6.

4.3.4 Gene functional enrichment

To determine enriched GO terms, we used the DAVID functional annotation tool to assign genes with their affiliate terms and to order them by enrichment[250].

4.3.5 Pain category assignment

In order to determine which drugs are used to treat pain associated indications we manually assigned pain categories to all pharmacologically treatable indications. Indications were assigned to one of the four categories: i) 'Pain specific' are indications that are specifically associated with pain, (e.g., neuropathic pain and headaches); ii) 'Typically painful' are indications that are typically painful, where pain is consistently presented as a symptom of the disorder (e.g., endometriosis and arthritis); iii) 'Can be painful' are indications that can be painful, but can also manifest in a pain-free state (e.g. certain cancers and diabetes); and iv) 'Typically non-painful' are indications that are typically not associated with pain (e.g. alopecia and wrinkling skin), including mental illnesses (e.g. schizophrenia and depression). Protein targets of drugs were sourced from an in-house database and were then assigned a pain category using the most pain related indication.

4.3.6 Anatomical categorization

To build pain networks specific to the brain, spinal chord, peripheral nervous system (PNS) and immune system, all interactions that had at least one mapping to an anatomical term derived from wiki-pain.org data were used. Anatomical terms were then mapped into one or more of the four anatomical regions or other (see Supplementary file 24 for mappings). Each network was then built for the four anatomical regions according to interactions that had an associated anatomical term.

4.3.7 Microarray analysis

We performed the tibial nerve transection (TNT) surgery[322] on adult female rats (n = 8) alongside sham controls (n = 8). Rats were confirmed for tactile allodynia in response to mechanical pressure and both dorsal root ganglion (DRG) and spinal cord were harvested at 7 days post surgery. Gene expression analysis was performed using the Affymetrix Rat 230 2.0 chip. After QC, data were RMA normalized and limma was used to identify differentially expressed genes versus sham, which were considered significant if their FDR corrected p value was <0.05 and their fold change was >1.5. These experiments were approved and monitored by the local ethics committee. The data from this experiment are available in the ArrayExpress database (www.ebi.ac.uk/arrayexpress) under the accession number E-MTAB-2260.

4.4 Results

4.4.1 The literature-derived pain PPI network

Using a semi-automated text-mining procedure[318] we identified 1,002 unique PPIs associated with pain, involving 611 different proteins (Figure 4.1a and Supplementary table 1; see Methods). In total, there are 124 interactions classed as negative regulation, 403 as positive regulation, 180 as regulation (either positive or negative) and 295 as binding. When connected as a network, the PPIs form a highly interconnected and coherent structure with the largest component containing 481 (79%) of the 611 proteins (Figure 4.1b). The network has an average degree of 2.8, a clustering coefficient of 0.07 and a power law fits the node degree distribution with 0.993 correlation indicating it is scale-free, consistent with other molecular interaction networks[323]. The proteins in the network show a statistically significant enrichment for pain associated Gene Ontology (GO) biological processes (e.g., response to wounding and inflammatory response), cellular components (e.g., neuron projection and postsynaptic membrane) and molecular functions (e.g., ion channel activity and neurotrophin binding) (Supplementary table 2).

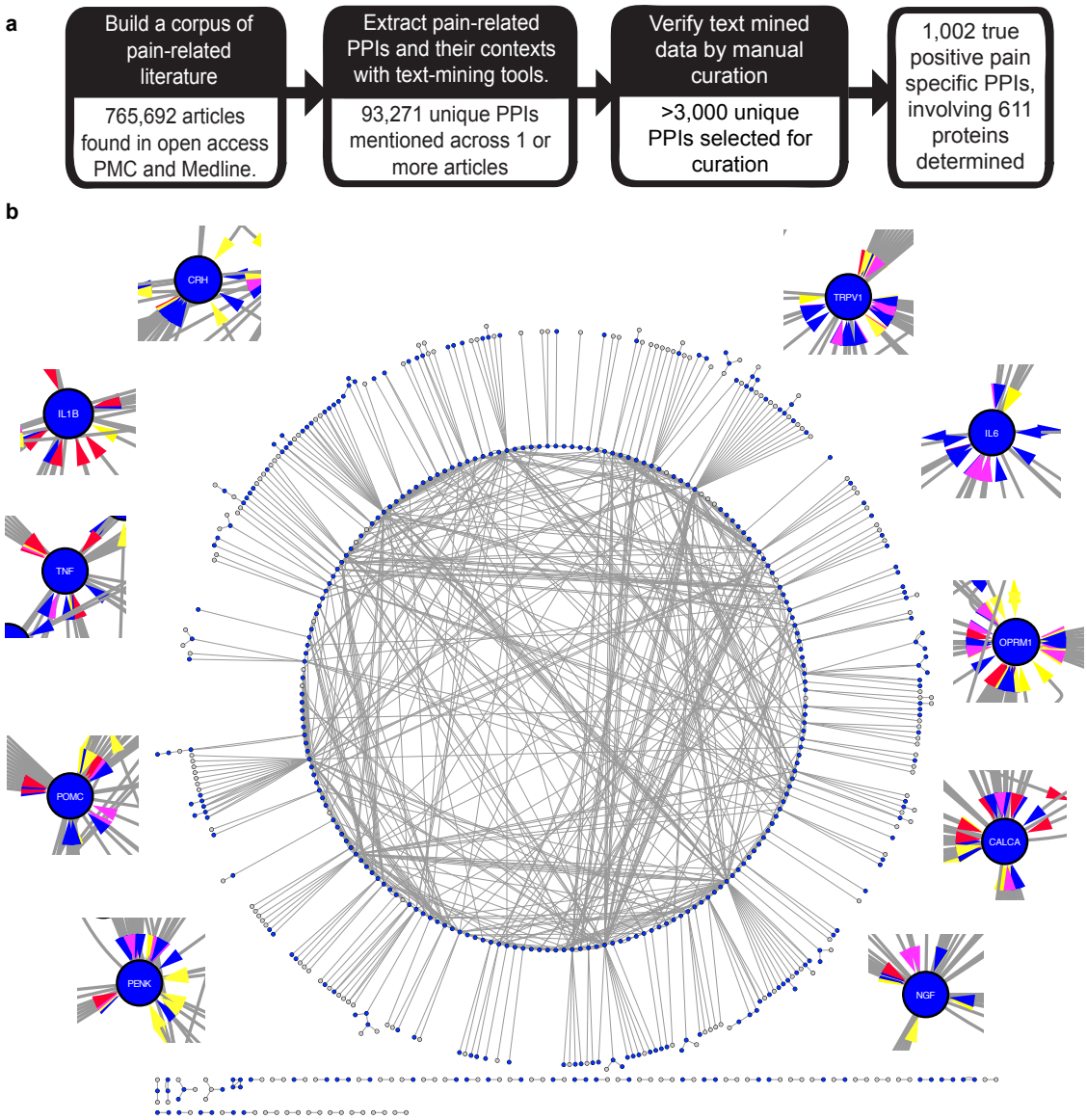


Figure 4.1 The pain interaction network.
a, Workflow for creating a pain specific PPI network. **b**, The PPI network for all pain associated proteins derived from the curated data. Proteins enriched against iRefIndex ($p<0.05$) are highlighted in blue (3887 total, see Supplementary table 13). Inserts show the top 10 enriched proteins. Colored arrows refer to interaction type: blue corresponds to positive regulation, red to negative regulation, turquoise to regulation and yellow to binding (these edges are bi-directional).

Given that our interaction data is derived from the primary literature, there is potential for ascertainment bias in our network[324]. For example, we will only have data for proteins that have been studied in a pain context and the most central nodes to our network could be biased by the fact that they have been studied for the longest time. As expected (see [324]), there is a positive linear trend

between the degree of a protein and the number of publications describing its interactions (Supplementary Figs 1, 2 and Supplementary table 3; Figure 4.2a), but there is no significant increase in the average number of documents per interaction observed as degree increases (rho = 0.08, Figure 4.2b). However, there is an inverse correlation between the date of a node's first publication in our dataset and its degree (rho = -0.4, Figure 4.2c and Supplementary table 4).

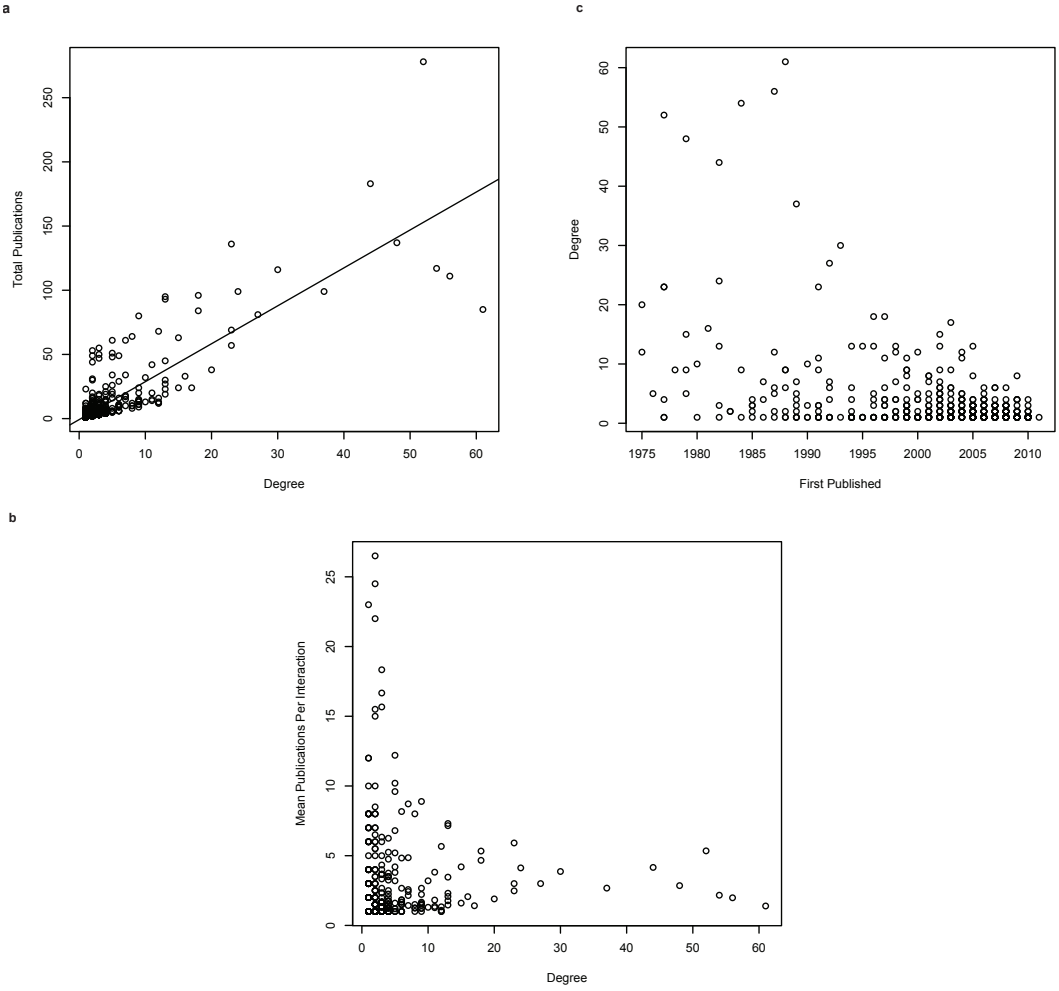


Figure 4.2 Bias in the pain interaction network.
a Correlation between the number of publications and degree for nodes in our network showing a linear trend (rho = 0.83), **b** The average number of publications per interaction for a pain protein remains flat (rho = 0.08) suggesting most interactions are reported individually (Supplementary Figs 1 and 2; Supplementary table 3). **c** There is an inverse relationship between the date of first publication on a protein's interactions and the protein's degree (rho = -0.4) (Supplementary table 4).

The first interactions in our network were published as early as 1975, with 25% of interactions published before the year 2000. Those published before 2000 include

the 17 highest degree proteins; supporting the assertion that degree correlates with length of study and knowledge of a protein's perceived importance. We therefore need to be aware of the fact that – within literature derived networks – the longer and more thoroughly a protein has been studied, the more interactions it is likely reported to have.

4.4.2 Comparative analyses between alternative pain protein datasets

To investigate the scope and relevance of our text-mined network to pain we compared it to networks derived from two other commonly used sources of disease-associated gene datasets, using generic interaction data from iRefIndex to determine known interactions between the proteins in these datasets (see Methods). It would be preferable to provide comparisons with datasets whose interactions are derived entirely from pain-specific experiments, but there are no such datasets currently available. Firstly, we generated gene expression data from dorsal root ganglion (DRG) and spinal cord in the rat TNT model of neuropathic pain (see Methods) to derive a set of pain associated differentially expressed genes. Secondly, we utilized a list of pain associated proteins from the Pain Genes DB[298] that have been manually curated from the literature. We reasoned that the gene expression data would not be prone to the same biases as literature-associated data (i.e., data derived from small-scale experiments), whereas the Pain Genes DB list is curated from the literature but is not dependent on text mining.

From the gene expression experiment, we find 237 genes to be differentially regulated across both DRG and spinal cord tissue; also, there are 399 genes in the Pain Genes DB dataset (Supplementary tables 5 and 6). These are considerably fewer than the 611 proteins in the text-mining derived dataset. Using the generic interactions from iRefIndex to connect proteins in these datasets, it was only possible to make 67 connections between 63 proteins in the gene expression data (Supplementary Fig. 3a and Supplementary Table 7). Therefore, we expanded this dataset to include first order neighbors with high-betweeness, stipulating that they have interactions with at least two of the input genes, so acting as bridges to connect the network. As a consequence, 192 (81%) of the differentially

expressed genes are included in the network. The resulting networks from Pain Genes DB and gene expression contain, respectively, 272 and 901 nodes, 510 and 12,318 edges, average degrees of 3.75 and 27.34, clustering coefficients of 0.115 and 0.264, and power law correlations of 0.959 and 0.645 (Supplementary Figs. 3b and 4; Supplementary tables 8 and 9). These networks include relatively few of the proteins from our text-mined network; 125 (20%) and 137 (22%) respectively. We note in particular the high average degree of 27.34 in the gene expression network, while the Pain Genes DB and text-mined networks both have similar ratios of 3.75 and 2.8 respectively.

These data indicate that our text-mined network has similar properties to a network derived from manual curation. The network derived from gene expression data has a far higher average degree and so presumably contains many more non-specific/non-relevant interactions despite the constraints we placed on introducing new nodes. Indeed, both the gene expression data (without first order neighbors) and the Pain Genes DB curated data show similar pain relevant enriched GO terms to the text-mined proteins (Supplementary tables 10 and 11). However, when analyzing only those bridge proteins that were added to the gene expression network there is much lower enrichment of pain related GO terms in comparison to the original gene expression gene list (Supplementary table 12), which would suggest that there is considerable noise introduced into this network.

We next cross-referenced nodes in all three networks (text-mining derived, gene expression and Pain Genes DB) with known therapeutic targets of FDA approved drugs (see Methods), taking this as an additional measure of relevance to pain. In the text-mined network we find 181 targets for existing therapeutics, with 88 targets for anesthetics and pain specific indications (e.g., migraines, neuropathic pain, abdominal pain etc.) and 51 targets for typically painful indications (e.g., arthritis, endometriosis etc.) (Figure 4.3a and Supplementary table 13). Examples of pain specific targets in the text-mined network include OPRM1, the target of analgesics such as morphine [325] and key pro-inflammatory cytokines such as TNF that is targeted by numerous drugs for Rheumatoid Arthritis[326]. The Pain Genes DB and gene expression networks have fewer

therapeutic targets than the text-mined network, with 100 and 132 respectively (Supplementary tables 14 and 15). We also note that the gene expression network contains a much lower proportion of pain specific targets (36%) in comparison to the Pain Genes DB (66%) and the text-mined network (49%); see Figure 4.3b.

There is a significant relationship between the enrichment (see Methods) of a node in the text-mined network and the likelihood of it being a drug target for a painful indication (Chi squared test for trends in proportions p=0.002, see Figure 4.3b), which is not the case for either the Pain Genes DB (p=0.05) or gene expression networks (p=0.9). We see strong enrichment for targets of drugs currently in development for pain indications e.g. NGF (Tanezumab)[327] and genes that have been earmarked as potential therapeutic options, e.g. BDNF[328, 329]. Moreover, the highly enriched IL6 and SST are currently targets for other indications (diabetes and prostate tumors) and thus their associated drugs may represent promising re-purposing opportunities to treat more typically painful and pain specific indications.

4.4.3 Insights into the pathology of pain

We next explored the molecular biology of pain apparent from our network. There are a number of proteins in the pain network with a high degree, indicating the importance of these nodes to the structure of the network[330]. As this pain network is a sub-graph of the much larger human interactome, we confirmed this connectivity by controlling for proteins that are highly connected in general and thus more likely to appear highly connected in our network. To do this, we developed a method to identify proteins with a significant enrichment of their known interactions within our pain network. We again used iRefIndex as a source of generic interactions to facilitate this[308], (see Methods).

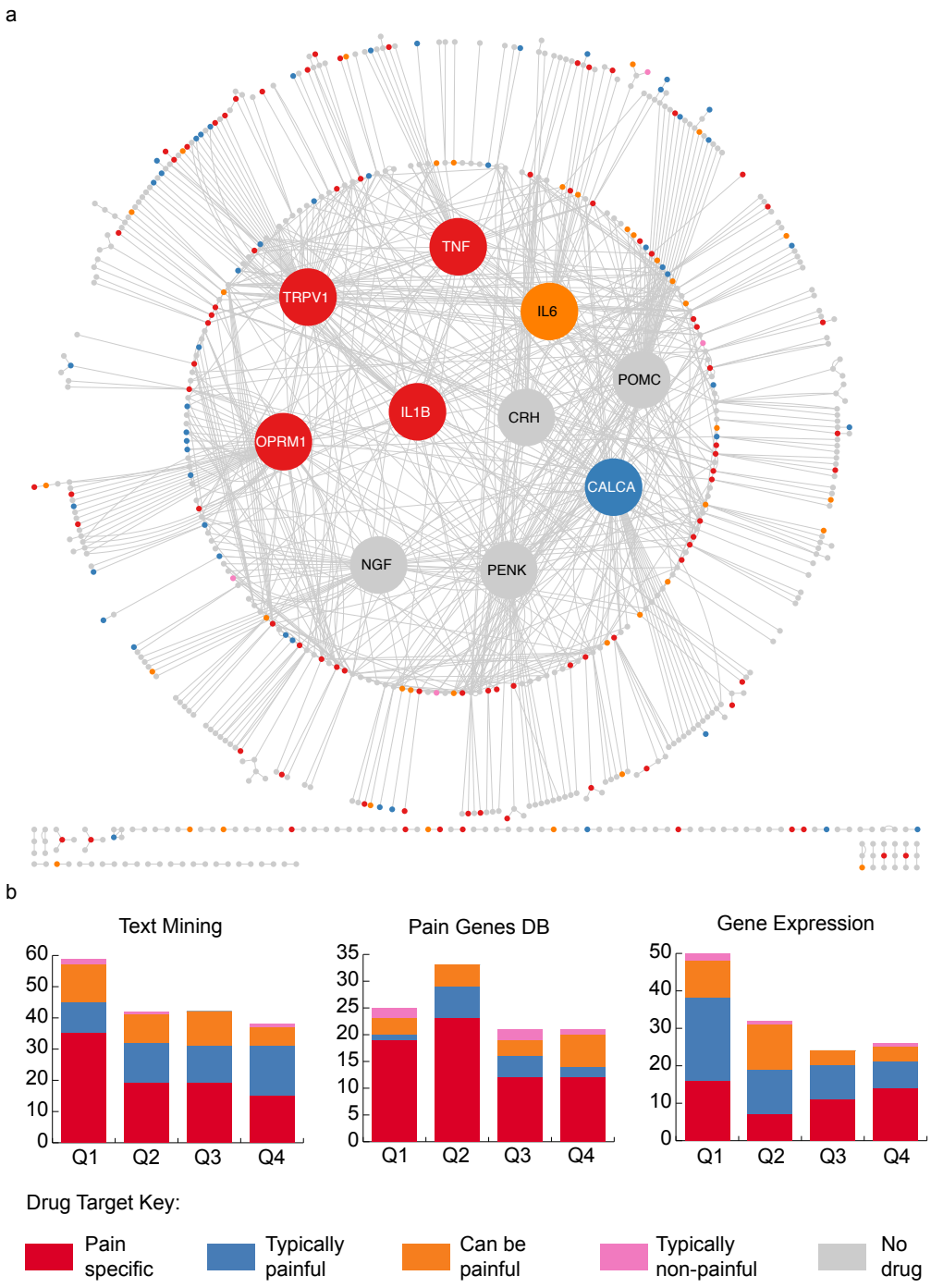
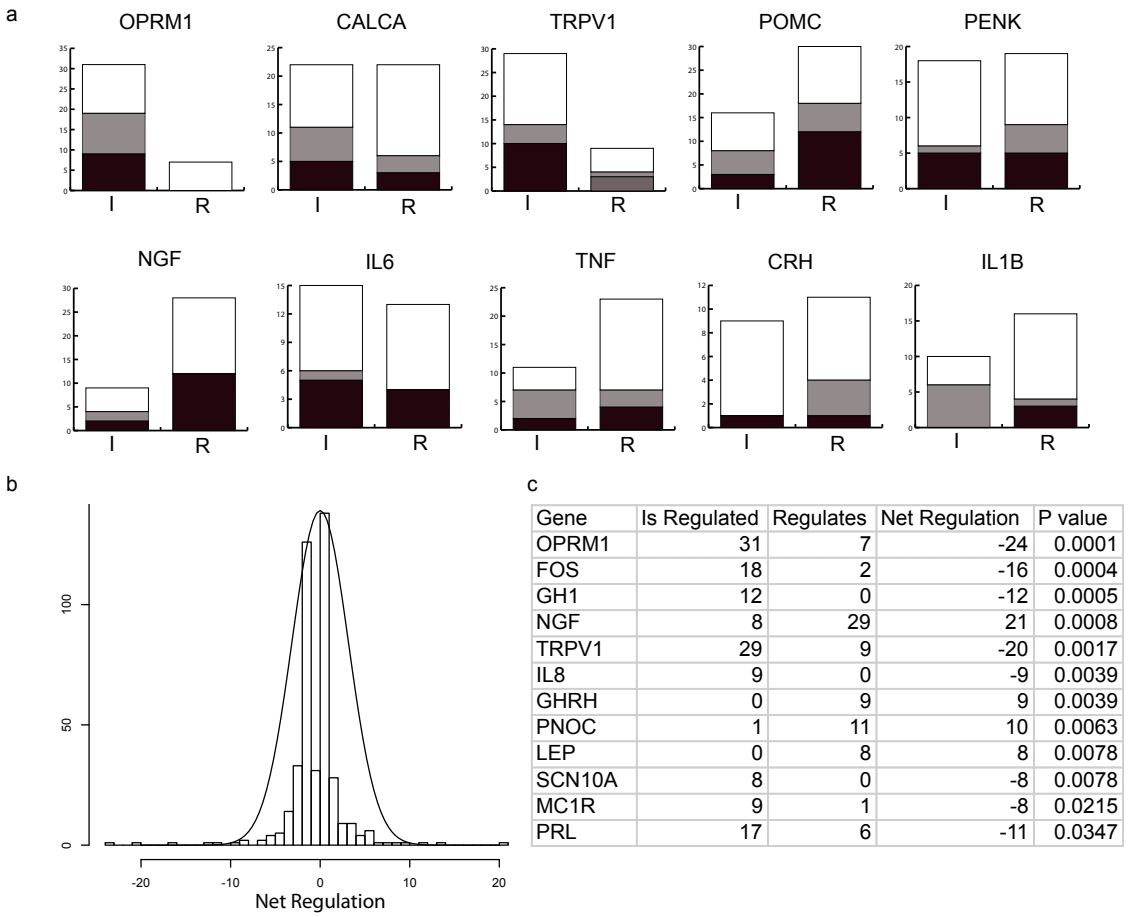


Figure 4.3 Drug targets in the pain interaction network.
a Drug targets are colour coded by the contribution of pain to their primary indication (see methods) as indicated in the key. The ten most enriched nodes are enlarged and moved into the centre for clarity. **b** Drug target profiles of each pain network. Proteins from each dataset are ranked by their enrichment p value and binned into quartiles (Q1-4). The numbers of associated drugs that target proteins in each quartile are then indicated. There is a significant relationship between the enrichment of a node in the text-mined network and the likelihood of it being a drug target for a pain specific indication (Chi squared test for trends in proportions p=0.002). However, neither the Pain Genes DB network, nor the gene expression data show the same significant trend (p = 0.05 & 0.9, respectively).

Of the most enriched proteins in the network (Figure 4.1b, 4.4a and Supplementary table 13), many are key to the pathology of pain, for example OPRM1[331], TRPV1[332, 333] and NGF[334, 335]. We find that enriched nodes have multiple regulatory roles, both up and down regulating numerous proteins (Figure 4a, b and Supplementary table 16). There are 8 enriched proteins that are more significantly regulated by others, e.g. OPRM1, TRPV1 and FOS, and 4 proteins that more significantly regulate others: NGF, GHRH, PNOC and LEP (Figure 4c). This is consistent with the known roles of NGF and nociceptin (PNOC) as mediators of pain signaling[336, 337]. Interestingly, growth hormone (GH), but not growth hormone releasing hormone (GHRH) has been associated with the chronically painful condition fibromyalgia[338]. Further, recent evidence has suggested a role for leptin (LEP) in the modulation of pain[339, 340]. Our data suggest that both GHRH and leptin might play a more prominent regulatory role in pain than has hitherto been appreciated.

To demonstrate the utility of our contextualised interaction dataset, we chose to investigate inflammatory and neuropathic pain, two fundamental aetiologies that manifest chronic pain states[290, 341]. In the construction of these sub-networks, we have also curated the overall effect of the interaction on the outcome of the pain type, i.e. an inhibitory or positive effect. There are 144 interactions associated with neuropathic pain in our dataset, with 122 found to be contributory to its pathology, 17 inhibitory and 5 denoted as both (Figure 4.5; Supplementary table 17). In comparison, 181 interactions are related to inflammatory pain, including 154 contributory interactions, 22 inhibitory interactions and 5 that have been documented as both (Figure 4.6; Supplementary table 18). 61% of the proteins from inflammatory pain form a coherent core graph, while there are two distinct sub graphs in the neuropathic pain network that together account for 73% of the proteins.



c

Gene	Is Regulated	Regulates	Net Regulation	P value
OPRM1	31	7	-24	0.0001
FOS	18	2	-16	0.0004
GH1	12	0	-12	0.0005
NGF	8	29	21	0.0008
TRPV1	29	9	-20	0.0017
IL8	9	0	-9	0.0039
GHRH	0	9	9	0.0039
PNOC	1	11	10	0.0063
LEP	0	8	8	0.0078
SCN10A	8	0	-8	0.0078
MC1R	9	1	-8	0.0215
PRL	17	6	-11	0.0347

Figure 4.4 Protein regulation in the pain interaction network.
a The top 10 most enriched genes in the pain network are shown with their regulation profiles broken down by incoming (is regulated, "I") and outgoing (regulates others, "R") interactions. Black denotes positive regulation, grey denotes negative regulation and white denotes other types of interaction. Undirected binding interactions are excluded. **b** The distribution of net regulation for all proteins in the pain network shows a normal distribution with long tails. This indicates that only a few proteins act as master regulators. **c** These master regulators were determined using the exact binomial test (Supplementary table 16). The proteins that are significantly more regulated than they are regulators and vice versa are shown; NGF is the most significant net regulator.

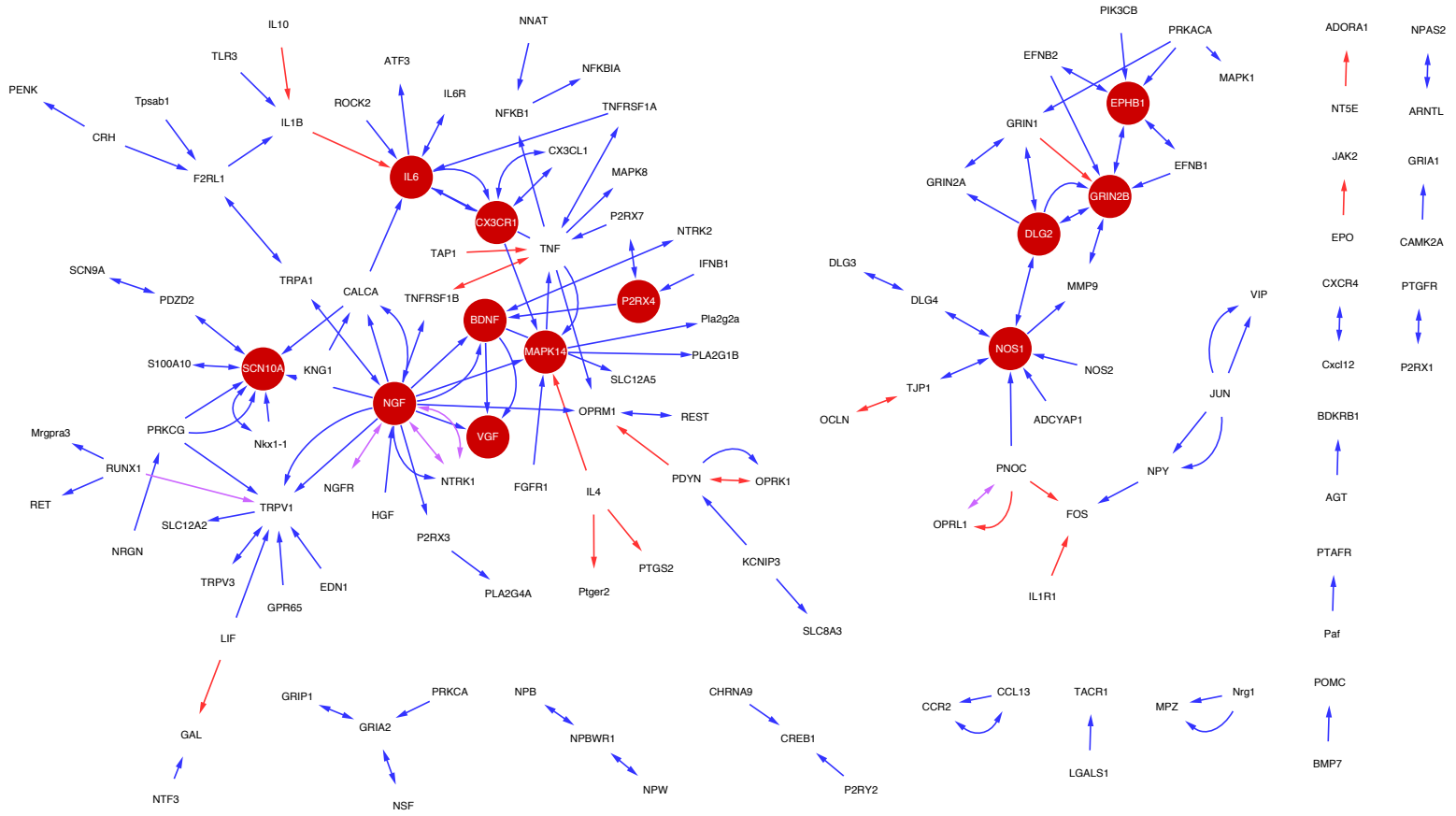


Figure 4.5 PPIs specific to neuropathic pain.
a Neuropathic pain specific sub-network. Blue edges are those interactions that have been curated as increased in a neuropathic pain state, red edges decreased and pink edges are those that have been denoted as both. Dark red nodes are those that are enriched against the general pain network (Supplementary table 17).

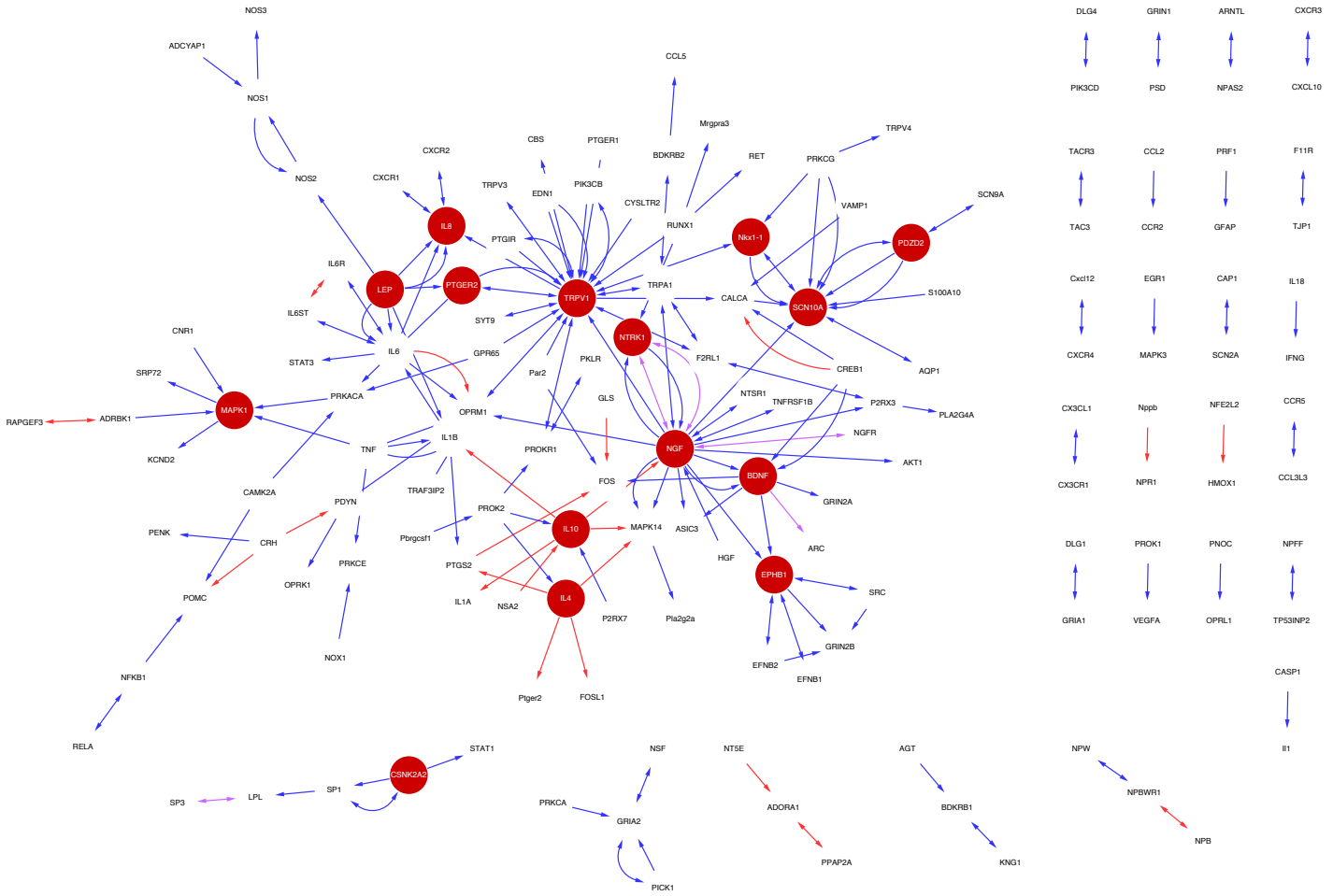


Figure 4.6 PPIs specific to inflammatory pain.
a Inflammatory pain specific sub-network. Blue edges are those interactions that have been curated as increased in a neuropathic pain state, red edges decreased and pink edges are those that have been denoted as both. Dark red nodes are those that are enriched against the general pain network (Supplementary table 18).

The neuropathic and inflammatory pain networks contain 127 and 157 proteins respectively, with 80 featuring in both datasets. This large overlap in proteins demonstrates that both forms of pain have underlying core pathology even if the initiation of pain is distinct. This is highlighted by the biological processes that the 80 proteins are highly enriched for, e.g. response to wounding ($p=2.88x10^{-14}$) and sensory perception of pain ($p=6.41x10^{-16}$) (Supplementary table 19). However, the proteins unique to the inflammatory and neuropathic pain datasets also reveal the more subtle differences between these disorders. For example, proteins unique to inflammatory pain show a much higher enrichment of inflammatory associated biological processes in comparison to proteins unique to neuropathic pain, such as inflammatory response ($p=5.16x10^{-12}$ vs. $p=3.54x10^{-4}$) and defense response ($p=1.30x10^{-10}$ vs. $p=6.03x10^{-05}$) (Supplementary table 19). The most enriched biological processes unique to the neuropathic pain dataset include regulation of membrane potential ($p=4.41x10^{-5}$) and regulation of action potential ($p=4.66x10^{-5}$). Moreover, there are 12 proteins in the intersection between the neuropathic pain dataset and the rat TNT gene expression dataset in comparison to just 3 from the intersection of the inflammatory pain dataset and the gene expression data. Odds ratio test confirms that proteins in the neuropathic pain network are more likely to feature in the gene expression dataset compared to proteins from the inflammatory pain network (odds ratio=5.36,z=2.55,p=0.01). Given that the gene expression dataset was derived from a neuropathic pain model, this would suggest that our neuropathic pain curated data is indeed more relevant to neuropathic than inflammatory pain.

To identify the key molecules within the two pain types, we repeated our method to reveal enriched proteins against the human interactome and, in addition, against the main pain network. This revealed 116 and 135 proteins where the majority of their interactions were present in the neuropathic and inflammatory pain sub-networks, respectively, compared to the generic human interactome (Supplementary tables 20 and 21). There were 12 and 15 proteins enriched in each sub-network compared to our main pain network. Of these, only NGF, SCN10A (NaV1.8), BDNF and EPHB1 (ephrin receptor) feature in both the neuropathic and inflammatory pain datasets. The neurotrophic NGF

and BDNF[329] as well as ephrin[342] play a key role in neuronal growth and axonal guidance and have been linked to multiple pain aetiologies. There are nine genes that encode alpha subunits of voltage gated sodium channels, many of which have been linked to multiple types of pain[343]. It is therefore surprising that only NaV1.8 is identified here. This raises an interesting perspective on our data, which does not seek to identify general gene-disease functional associations but rather to uncover which proteins are highly interacting within a diseased state compared to a normal state. Based on this measure, whilst other ion channels are important to pain, NaV1.8 appears to be the only sodium channel whose interactome spans multiple pain aetiologies.

There are 8 enriched proteins that are specific to the neuropathic pain network and 11 to the inflammatory pain network. Of the proteins specific for neuropathic pain, GRIN2B and NOS1 are already targeted for pain specific indications. MAPK14 and IL6 are targets for other indications and so might represent drug re-purposing opportunities for neuropathic pain specific disorders. In addition, DLG2, CX3CR1, P2RX4 and VGF appear to be promising leads for the specific investigation of neuropathic pain[344-346]. Similarly, IL10, PTGER2 and IL4 are existing targets for inflammatory pain associated disorders (e.g. rheumatoid arthritis), while TRPV1 is targeted by analgesics (e.g. Propofol[347] and capsaicin[348]). LEP, Nkx1-1, PDZD2, NTRK1, IL8, MAPK1 and CSNK2A2 would also appear to be specifically important to inflammatory pain. NTRK1 (TrkA) is the receptor for NGF, which we previously saw to be enriched in both types of pain. That NGF's receptor is only enriched in the inflammatory pain dataset (although it is present in the neuropathic pain network) emphasizes the need to apply caution when interpreting such data as complete. The curated data, while extensive, is not complete and indeed the body of published work itself does not detail the full pain interactome.

Finally, to illustrate further the possibilities associated with our data, we repeated the same style of analysis but this time creating networks for different anatomical regions. From the 1,002 PPIs, we used the anatomy context in wiki-pain.org to determine 607 interactions that could be mapped to at least one or more of the following pain relevant anatomical associations: brain, spinal cord,

PNS, immune system and other (Supplementary table 22). We determined 245, 204, 162, and 92 interactions associated with the brain, spinal cord, PNS and immune system, with 211, 190, 152 and 106 proteins in each respectively (Supplementary Figs. 5-8). We used our enrichment analysis to identify proteins more highly connected in each of the anatomical regions compared to the general pain network (Supplementary table 23). We find NGF and BDNF to be key to the network in multiple anatomical locations, being enriched in the brain, spinal cord and PNS networks. PENK, OPRL1, GHRH were specifically enriched in the brain, FOS in the spinal cord, while CALCA, TRPV1, RUNX1, RUNX3, NTRK2, TNFRSF1A and GDNF were only enriched in the PNS networks. There are also 20 proteins enriched in immune related anatomical regions, e.g., CCL5 and IL8. These data allow us to explore the anatomical interplay that contributes to the development of pain, in particular the interplay between the peripheral and central nervous systems. In addition, this also aids drug development by informing the necessary central or peripheral distribution of a drug candidate.

4.5 DISCUSSION

We have shown that our large semi-automated text-mining derived network is relevant to pain and forms a more complete representation of the molecular mechanisms underlying the disease than is possible using other common starting points. We identify several drug re-purposing opportunities and use our enrichment method to identify novel mediators of pain. In particular, we show that NaV1.8 is a key ion channel for both neuropathic and inflammatory pain. Further, as we are able to extract specific context with each interaction, we can create and explore networks specific to individual pain indications or anatomical regions. Recent studies have undertaken meta-analyses of gene expression data from pain models[349, 350] or have described resources that enable the network visualisation of known pain genes by incorporating PPIs from non-diseased contexts[350]. Our method, using disease specific interactions identified from the pain relevant literature, offers a considerable advance in specificity and relevance.

Text mining has long been heralded as the practical solution to efficiently retrieving data denoted in the ever expanding body of published biomedical literature[218], but poor precision and recall has restricted its wider use in delivering reliable data[256]. Instead, the majority of data derived from free text that is subsequently used in biological analyses is identified and extracted by manual curation, a process that is costly, time-consuming and often unable to offer more exhaustive coverage[286]. As a method of extracting and characterizing key proteins and interactions that are denoted in the literature, our study offers a strong case for a semi-automated approach that uses text mining to rapidly generate the data and manual curation of the results to achieve high precision. While the protein interaction data we have retrieved and curated in this study is not complete, the datasets have proven sufficiently broad, accurate and relevant enough to make compelling biological findings.

The results in this study represent the most extensive summary of all the published research conducted on pain-associated proteins. The power of such an approach comes from integrating the data at the network level, which allows novel hypotheses to be drawn in the context of the global picture. Further, the network can be used as a framework to provide context to the interpretation of datasets generated by researchers within the field. This is increasingly recognized as a successful approach to the study of disease biology[351]. It is foreseeable therefore that a similar approach to data retrieval and analysis could be applied to a huge range of biomedical disorders under various different contexts in order to provide networks and targets for further study.

4.6 Acknowledgements

We dedicate this paper to the memories of our colleagues Dr Phoebe M Roberts and Michael Kennedy. Phoebe contributed to the text mining of the pain data and sadly died on December 8th 2013. Co-author Michael passed away on February 7th 2014.

CHAPTER 5

Expanding the human pathogen interactome with text mining

5.1 Abstract

Disease, through infection by pathogens, has a huge negative impact on human morbidity and mortality worldwide. While human pathogens originate from a wide range of different taxonomic groupings, for example viruses, bacteria, fungi etc., they commonly share similarities in the human proteins they target, which ultimately leads to disease. There is therefore increasing focus on discerning these interactions between pathogen and human proteins and many of these are already stored in public databases. However, the literature contains many more interactions that are unaccounted and strewn over millions of publications. In this study we thus used text-mining to capture these, using enhanced data from the text-mining databases BioContext and Evex DB to find over 25.5 thousand new unique host-pathogen protein interactions involving 223 pathogens. As text-mining can produce false positives, we curated this data in three tasks to demonstrate how large amounts of data can be rapidly validated to identify new interactions between human proteins and pathogens. Although the text-mining data demonstrated particularly low precision in the curation, we were still able to find 42 new HIV-1-human protein interactions, 108 new interactions between human proteins and pathogen species and 33 human proteins that had not been known targets of any pathogen species before. These results highlight the value of the literature as a reservoir of uncatalgued host-pathogen interaction data and the method of curation we offer provides a way of rapidly discerning this data in an accurate and efficient manner.

5.2 Introduction

Human pathogens that cause infectious disease originate from a range of microorganisms, including viruses, bacteria, fungi, helminthes, protozoa and prions. Together, they accounted for roughly 19% of deaths[352] and 35% of the disability-adjusted life years lost[353] worldwide between 1990 and 2010. They range from global pandemics, e.g. human immunodeficiency virus (HIV)/ acquired immunodeficiency syndrome (AIDS), malaria and tuberculosis, to isolated outbreaks, such as with haemorrhagic fever caused by the ebola virus[354].

For many pathogenic diseases, the route to infection begins with invasion of the host cell. Thus, uncovering the specific protein-protein interactions (PPIs) that are involved in this process has been imperative for the advance of modern therapeutics. For example, viral gp41 and human coreceptor CCR5 are targets for FDA-approved enfuvirtide and maraviroc respectively, used to treat HIV-1[355]. Many pathogens have been studied in this way, and their PPIs have been documented widely across the literature and public databases[145, 356-359]. However, this knowledge often remains disconnected, preventing biomedical researchers from being able to answer more fundamental questions on the nature of pathogenicity.

Recent studies have begun to address this issue. For example, Dyer *et al* integrated seven publically available databases of 190 pathogen strains, producing 10,477 human-pathogen PPIs for analysis. Smith *et al* were able to identify possible repurposed drug targets though large-scale analysis of microarray datasets between multiple respiratory viruses[360]. Pichlmair *et al* experimentally determined new common signaling pathways and cellular processes between 30 viral species[361]. However, no study has yet fully utilized the largest source of all human pathogen PPIs available, the published literature.

The key issues with exploiting knowledge published in the literature stem from the inability to accurately and efficiently extract and convert it into a machine-readable structure. Existing databases of host-pathogen interactions are predominantly manually curated[145, 356, 358, 359], a slow and costly process[318]. While

text mining methods have been specifically developed for automating this entire process[362], the results need to be precise for further biological analysis, and given that they are often inaccurate, its use has been limited mainly to article prioritization so far[363]. However, novel semi-automated approaches to extracting PPIs in HIV-1[286] and pain[318] have shown that generating accurate data rapidly can be achieved by text-mining when it is organized and grouped into a format that allows manual curation to operate as a quality control step.

In this study our aim was to establish whether a semi-automated approach to extracting PPIs from the literature could be used to expand our knowledge of PPIs across all human pathogens in existing public databases. After identifying 1,419 pathogens known to infect humans, we found 13.6K host-pathogen PPIs already present in public databases. We used text-mining data to determine 26.5K HP-PPIs, of which 25.5K were not already present in public databases. From the newly identified HP-PPIs, we adopted three different approaches to validating the quality of this data for further analysis. First, we curated all PPIs involving both HIV-1 and human proteins finding 42 new interactions not in the already extensively curated HIV-1, human protein interaction database (HHPID). Second, we expanded the existing knowledge of the range pathogens known to interact with commonly targeted human proteins by a further 108 interactions between human proteins and pathogen species. Third, we curated PPIs involving human proteins that had not previously been associated with any pathogen, finding 33 new human proteins that interact with pathogens.

5.3 Methods

5.3.1 Identifying human pathogens

A systematic literature survey in 2001 listed 1,415 pathogens capable of causing disease in humans under natural transmission conditions[30]. Updates to this in 2005[32] and 2007[31] showed only 1,407 and 1,399 species respectively, changes due to certain species being reclassified. However, only 177 and 87 pathogen species were listed in each. Using these three sources of pathogen names, we mapped each pathogen to 1,224 unique NCBI taxonomic identifiers where possible,

leaving a further 195 pathogens that could not be mapped (1,419 pathogens in total). Taxonomic groupings into bacteria, viruses, helminthes, protozoa, prions and fungi were retained, with the addition of genus to NCBI mapped pathogens.

5.4.2 Building a human pathogen corpus

With the species names for the 1,419 pathogens identified, we added 3,810 NCBI taxonomic DB pathogen mapped synonyms, as well as 25,730 sub-species names and synonyms for all of these. We also linked diseases associated with pathogens by mapping 371 diseases to 180 pathogens listed as causal agents in the Disease Ontology[126]. For example, HIV-1 is the causal agent of AIDS. Linked diseases like this contributed a further 1,070 terms associated with the pathogens. Regular expressions were then used to add any case-sensitive variants to the terms (e.g. uppercase and lowercase), as well as additional species terms (e.g. *S Cerevisiae*) for non-viral pathogens and terms not originating from common names or acronyms. In total we determined 38,284 unique pathogen-associated terms with 161,753 case-sensitive variations of these (supplementary table 5).

Using this pathogen term list, we matched these to all abstracts and titles in Medline (2014 version), including MeSH terms, and open access PMC full-text (2014 version) using LINNAEUS[191]. We also used LINNAEUS to capture any abbreviations and implemented a separate post-processing system to resolve ambiguity between pathogen terms associated with more than one species.

Additionally, we matched general pathogen related terms, e.g. pathogenic, infection, outbreak etc., to enable us to distinguish more effectively between pathogens studied as pathogens and pathogens studied in another context. For example, *S. Cerevisiae* is a well studied fungus, however it is only rarely pathogenic[364, 365] and is an organism associated more prominently with other areas of research, such as an experimental organism for revealing eukaryotic gene function[366]. We considered specific pathogens that co-occurred with general pathogen terms as more likely to be studying the pathogenicity of a pathogen.

5.4.3 Collecting existing text-mining data

BioContext[227], released in 2012, and Evex DB[367], updated in 2013[217], are two databases containing the results of text-mining software employed on the whole of Medline and PMC open access full-text. These contain biomedical events for proteins that when combined in chains can be used to infer PPIs[286, 318, 368]. Gene, protein and RNA molecules were matched in both Evex DB and BioContext by BANNER[186]. Protein normalisation to NCBI Homologene IDs, Species IDs and Entrez Gene IDs was then performed by GenNorm[187] and Evex in Evex DB and GNAT[216] and GeneTUKit[274] in BioContext. Events for these entities were generated with TEES (version 2)[217] in Evex DB and TEES (version 1)[265] and EventMine[260] in BioContext.

We extracted all data from each database that had been derived from documents in our pathogen corpus. For merging of the two databases we then converted Evex DB into event chains, as represented in the BioContext format, and organised the events into a single table, containing >15.7M event chains. As with a previous study[318], we then removed superfluous events and grouped the individual event chains into unique event chains. This produced >4.4M unique event chains, of which >1.6M involved two participants and could be considered PPIs.

5.4.4 Enhancing text-mining data

While the text-mining data from BioContext and Evex DB was derived from the current 'state of the art', the software had been designed for generic usage and not for specifically extracting pathogen related PPIs. Furthermore, many of the proteins involved in the event chains were not mapped to Entrez Gene IDs making it difficult to determine their relevance to human pathogens. We therefore sought to improve the pathogen protein normalization aspect of the text-mining data, focusing specifically on HIV-1, hepatitis B virus (HBV), influenza A virus (IAV), hepatitis C virus (HCV) and herpes simplex virus 1 (HSV-1), whose genomes are small and thus more practical to facilitate

improvements. Moreover, these viruses all have a significant burden on human mortality and morbidity and thus represent priorities for further research into their molecular mechanims.

To enhance PPI extraction for the five viruses we expanded the dictionary of terms associated with each pathogen protein on Entrez Gene, using publications[369, 370] and various websites (e.g. Wikipedia) to create more thorough representations of their potential mentions in the literature. New gene records, not represented on Entrez Gene at the species level, were created for IAV and HCV to ensure that all of their proteins were catalogued appropriately, as well as removing any redundant records. We then re-normalised entities matched in Evex DB and BioContext, implementing a novel normalization tool with the expanded pathogen term sets. The tool utilised species mentions from the surrounding text to predict whether mentions from each pathogen term set had been correctly assigned. This enabled us to map an additional 153,998 event chains to unique proteins for HIV-1 (85 874), HBV (41 361), HCV (9 993), IAV (11 558) and HSV-1 (5 469). We then measured the precison of the pathogen protein normalization (and not other aspects of the text-mining data) by randomly selecting 100 event chains containing the newly normalized pathogen proteins for each of the 5 viruses. HIV-1, HBV, HCV, IAV and HSV-1 showed 93%, 97%, 92%, 99% and 99% precision respectively (96% overall). The majority of false positives were caused by incorrect species association with the proteins.

Finally, using the modified 1.6M event chains that could be considered PPIs we filtered any remaining event chains containing protein mentions that were not normalized to an Entrez Gene ID. The remaining PPIs were then classified as host-pathogen (HP) PPIs or pathogen-pathogen (PP) PPIs depending on the proteins involved. HP-PPIs were then further reduced to only those containing host proteins that were human or had a conserved human homologue.

5.4.5 Integrating Host-Pathogen PPI Databases

We extracted all PPIs involving human pathogens from the Pathogen Interaction Gateway[371], the human-HIV-1 protein interaction database[136], VirHostNet[145],

MIMI [309] and BioGRID[137]. Proteins were all mapped to Entrez Gene IDs (using ID conversions if they did not have an Entrez Gene ID already) and Homologene IDs were added where possible. We note that some UniprotKb IDs did not have an Entrez Gene ID mapping and could therefore not be used in this study. As with the text-mining data, proteins that were not human or did not have a human orthologue were removed. Duplicate PPIs that contained the same interactants were filtered out.

To determine whether these PPIs were present in the text mining data we cross-referenced the Entrez Gene IDs of the interactants in each PPI between each dataset. To provide this comparison we only used PPIs from the text mining dataset that contained two Entrez Gene normalized participants with at least one protein normalized to a human pathogen.

5.4.6 Curating text mining results

We curated text mining data for three different tasks/use cases:

1) To identify new protein interactions between human proteins and HIV-1 proteins
2) To identify new interactions between human proteins and pathogen species
3) To identify new interactions between human proteins and pathogens in general

In task one, the goal was to decipher whether an HIV-1 protein had been documented to have an interaction of any kind with a human protein that had not already been catalogued in the public databases. To do this we visualised all individual event chains and any useful relevant data (e.g. sentences, publication link, gene names etc.) corresponding to a unique interaction between an HIV-1 protein and human protein onto a single web page. This mimicked a previous approach for curating unique PPIs related to pain diseases[286]. A correct HIV-1-human PPI had to have each protein name mapped to their correct Entrez gene ID and have a clear direct or indirect interaction of whatever kind. We also

marked any interactions that had been documented negatively or speculatively. We required only one instance of a unique PPI to have been annotated correctly for the overall PPI to be considered correct.

In task two we seeked to curate interactions between human proteins, known to interact with at least one pathogen in public databases, and other pathogen species that did not have a catalogued interaction with that protein. PPIs were curated in the same way as task one, except that in this task a correct instance could be between a human protein and any protein of a specified pathogen species. The human protein-pathogen species interactions we selected for curation were those that had a text-mining confidence greater than 0.3. In a previous study we had shown this level to be a good indicator of correctly annotated PPIs[286]. We also filtered any interactions between HIV-1 as these had already been curated in task one.

Task three was similar to task one in that we were seeking interactions between specific human proteins. However, in task three we aimed to identify human proteins targeted by any pathogen, where the human proteins had no interactions with pathogens present in public databases. Therefore, in this task a correctly identified interaction was between a human protein and any protein from any pathogen species. As with task two, we filtered out any interactions involving HIV-1 proteins and selected interactions for curation that had a text-mining confidence above 0.3. We also selected interactions by ranking human proteins by the number of their pathogen associated Gene Ontology (GO) terms. Pathogen-associated GO terms were determined by selecting all human proteins from public databases that have been shown to interact with pathogens. Using DAVID[250], we then determined enriched GO terms, which we considered to be pathogen-associated GO terms. The enrichment p value was then used to provide multiple thresholds for which each GO term was used to rank human proteins. For example, in rank one we used all GO terms with a p value <0.1. Ranks two and three used p value of $p < 1E5$ and $p < 1E10$ respectively.

5.2 RESULTS

5.2.1 Sourcing Text Mining Data

Figure 5.1 summarises the metholodology used to retrieve pathogen related PPIs with text-mining, as well as the data retrieved with each step.

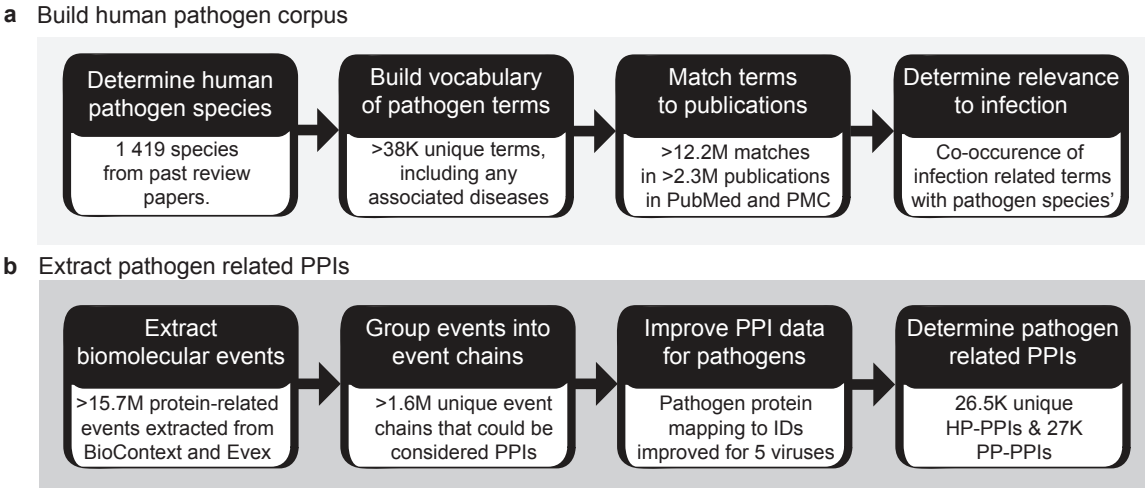


Figure 5.1 Text-mining methodology to retrieving pathogen related PPIs.
Our text-mining methodology consisted of **a** building a human pathogen corpus and **b** extracting pathogen related PPIs associated with this corpus.

We first identified 1,419 species that are considered pathogenic to humans, finding 2.3M associated publications with 1,362 species of these mentioned. Table 5.1 shows the top 20 pathogens ranked by total number of associated publications, as well as their first recorded publication. After filtering out likely non-pathogen specific studies, we find HIV-1, *E coli*, IVA, *M tuberculosis* and *S aureus* the most studied pathogens.

We next used the existing text-mining databases, BioContext[227] and Evex DB[217, 367], as sources of biomolecular events from the pathogen related publications for which we could infer pathogen related PPIs. Furthermore, to improve the quality of these databases we had modified the mapping of pathogen proteins to their Entrez Gene IDs for HIV-1, HBV, HCV, IAV and HSV-1 (see Methods). In total, we found 26,497 unique HP-PPIs and 27,056 unique PP-PPIs across 223 pathogens (table 2). The enhancements to the PPI data for HIV-1, HBV, HCV, IAV and HSV-1 yielded an additional 2,964 HP-PPIs and 76 PP-PPIs. For example, HIV-1 had 114 unique HP-PPIs in the original BioContext and Evex DB databases, but after the enhancements now has unique 2,193 HP-PPIs.

Pathogen	Genus	Class	First Publication	Total Publications	Pathogen Specific Publications
Human immunodeficiency virus 1	Lentivirus	VIRUS	1983	328,694	191,278
Escherichia coli	Escherichia	BACTERIA	1896	352,188	79,859
Influenza A virus	Influenzavirus A	VIRUS	1890	83,939	55,392
Mycobacterium tuberculosis	Mycobacterium	BACTERIA	1867	178,542	55,125
Staphylococcus aureus	Staphylococcus	BACTERIA	1896	92,422	48,004
Hepatitis B virus	Orthohepadnavirus	VIRUS	1946	79,298	41,379
Hepatitis C virus	Hepacivirus	VIRUS	1968	57,656	39,282
Saccharomyces cerevisiae	Saccharomyces	FUNGI	1929	208,482	30,036
Pseudomonas aeruginosa	Pseudomonas	BACTERIA	1894	54,372	26,560
Human Herpesvirus 1	Simplexvirus	VIRUS	1925	43,515	26,113
Helicobacter pylori	Helicobacter	BACTERIA	1986	38,928	24,685
Human Herpesvirus 4	Lymphocryptovirus	BACTERIA	1900	32,907	19,894
Plasmodium falciparum	Plasmodium	PROTOZOA	1916	35,492	17,956
Streptococcus pneumoniae	Streptococcus	BACTERIA	1948	29,881	17,264
Human papillomavirus	N/A	VIRUS	1950	28,406	16,462
Candida albicans	Candida	FUNGI	1948	32,985	16,225
Toxoplasma gondii	Toxoplasma	PROTOZOA	1940	23,377	12,569
Chlamydia trachomatis	Chlamydia	BACTERIA	1930	19,148	11,511
Haemophilus influenzae	Haemophilus	BACTERIA	1926	21,212	11,427
Measles virus	Morbillivirus	VIRUS	1880	24,800	11,148

Table 5.1 Top 20 most studied human pathogens.
Pathogens are ordered by their number of associated pathogenic publications. Pathogenic publications are those that mention a specific human pathogen in combination with a pathogenic related term, e.g. 'infection'. Total numbers of publications associated with each pathogen are also shown, as well as the first pathogen publication found.

Name	HP-PPIs	PP-PPIs	Total PPIs
Saccharomyces cerevisiae	9170	21348	30518
Escherichia coli	5574	2345	7919
Human Immunodeficiency Virus 1	2193 (114)	45 (39)	2238 (153)
Staphylococcus aureus	548	169	717
Mycobacterium tuberculosis	389	281	670
Bacillus subtilis	308	338	646
Pseudomonas aeruginosa	309	284	593
Candida albicans	298	191	489
Plasmodium falciparum	330	122	452
Salmonella choleraesuis	355	72	427
Influenza A virus	305 (36)	35 (6)	340 (42)
Hepatitis B virus	321 (0)	10 (0)	331 (0)
Helicobacter pylori	248	62	310
Streptococcus pyogenes	252	44	296
Human Herpesvirus 4	230	65	295
Enterococcus faecium	260	21	281
Yersinia enterocolitica	158	112	270
Hepatitis C virus	247 (0)	20 (0)	267 (0)
Human Herpesvirus 5	177	78	255
Neisseria gonorrhoeae	247	4	251
Human Herpesvirus 1	58 (56)	125 (44)	183 (100)

Table 5.2 Top 20 pathogens ordered by HP-PPIs from text-mining.
Numbers in brackets refer to counts before modifying BioContext and Evex DB. HSV-1, which is outside of the top 20, is also included. See Supplementary Table 1 for the full list of 220 pathogens.

5.2.2 Comparisons of text-mined and public database PPIs

Using publically available PPI databases we determined 13,331 HP-PPIs and 259,213 PP-PPIs (236,373 from *S Cerevisiae*). HIV-1 has considerably more unique HP-PPIs recorded than other pathogens, with 6,774. *Y pestis* next has 2,762 unique HP-PPIs, followed by *S Cerevisiae*, *F tularensis* and *B anthracis*, with 1033, 961 and 747 respectively. We then integrated this data with the PPIs derived from text mining (Table 5.3 and Supplementary Table 1).

Name	TM HP-PPIs	PDB HP-PPIs	TM PP-PPIs	PDB PP-PPIs	TM PPIs	PDB PPIs	TM in PDB HP-PPIs	TM in PDB PP-PPIs	TM in all PDB PPIs	New TM HP-PPIs	New TM PP-PPIs	New TM PPIs	All HP-PPIs	All PP-PPis	All PPIs
Saccharomyces cerevisiae	9170	1033	21348	236373	30518	237406	58	9875	9933	9112	11473	20585	10145	247846	257991
Human Immunodeficiency Virus 1	2193	6475	45	16	2238	6491	880	16	896	1313	29	1342	7788	45	7833
Escherichia coli	5574	35	2345	5612	7919	5647	1	43	44	5573	2302	7875	5608	7914	13522
Yersinia pestis	149	2762	68	0	217	2762	0	0	0	149	68	217	2911	68	2979
Francisella tularensis	38	961	14	0	52	961	0	0	0	38	14	52	999	14	1013
Bacillus anthracis	29	746	7	3	36	749	0	0	0	29	7	36	775	10	785
Human Herpesvirus 4	230	425	65	24	295	449	8	3	11	222	62	284	647	86	733
Influenza A virus	304	294	35	44	339	338	0	0	0	304	35	339	598	79	677
Staphylococcus aureus	548	2	169	1	717	3	0	0	0	548	169	717	550	170	720
Mycobacterium tuberculosis	389	0	281	373	670	373	0	6	6	389	275	664	389	648	1037
Total (across all pathogens)	26496	13331	27056	259213	53552	272544	961	9995	10956	25535	17061	42596	38866	276274	315140

Table 5.3 Pathogen-related PPIs from text-mining and public databases
Only the top 10 pathogens are shown (see Supplementary Table 2 for full list). Columns are provided for text-mining (TM) and public databases (PDBs) and their host-pathogen (HP) and pathogen-pathogen (PP) protein-protein interactions (PPIs). We also show those HP-PPIs that have been found by TM not in PDBs and also the total PPIs between TM and PDBs.

The text-mining data contains 961 unique HP-PPIs from the public DB data, with 880 of these originating from HIV-1. This shows that the text-mining data only represents 7.21% of the HP-PPIs present in publically available DBs. This is likely due to a combination of poor accuracy in the text-mining data, where HP-PPIs have either been missed or incorrectly annotated, and that many of the PPIs in the public databases may not have been individually described in the text used. Furthermore, much of the full text from the published literature was not openly available to TM and if provided would have most likely improved the percentage of public DB PPIs retreived. Conversely, a further 25 535 unique HP-PPIs were present in the text-mining data and not the publically available databases. These included 9,112, 5,573, 1,332, 548 and 389 unique HP-PPIs from *S cerevisiae*, *E coli*, HIV-1, *S aureus* and *M tuberculosis* respectively, among 223 pathogens with new data.

Table 5.4 shows the top human proteins targeted by pathogens (see Supplementary Table 2 for full list), with the public DB data, text-mining data and their union and intersect all detailed separately. There are 9 035 human proteins that interact with pathogen proteins, with 5 635 from public DB data and 6 283 from text-mining data. From the text-mining data, 2 883 human proteins feature in public DB HP-PPIs, although many are between human proteins and pathogen species not present in the public DB data. For example, NFKB1 has been shown to interact with 6 different pathogens in public DB data, whereas a further 37 are observed in the text-mining data. Moreover, in the text-mining data there are 3,364 human proteins that interact with pathogens that do not have any recorded interactors in the public data.

Protein Name	TM pathogens that interact	Unique TM Pathogens	PDB pathogens that interact	Unique PDB Pathogens	All Unique pathogens that interact
NFKB1	40	37	6	3	43
TNF	42	40	2	0	42
GOPC	39	37	2	0	39
TTF2	35	34	1	0	35
MAPK1	35	34	1	0	35
IL8	34	33	1	0	34
SOLH	33	33	0	0	33
ACTR6	30	30	0	0	30
HIVEP1	27	26	2	1	28
TLR2	27	26	1	0	27
TLR4	26	25	2	1	27
FN1	26	24	3	1	27
IFNG	26	25	1	0	26
IL2	26	25	1	0	26
CAT	24	23	3	2	26
CD4	26	25	1	0	26
IFNA1	25	24	1	0	25
MAPK14	25	22	3	0	25
MUC7	24	23	1	0	24

Table 5.4 Proteins targeted by pathogens
The top human proteins that interact with pathogen species are shown. We show data separately from text-mining (TM) and public databases (PDB), highlighting unique pathogen species interactions for each and the total number of unique pathogen interactions across both datasets.

5.2.3 Curating Text Mining Data

In past studies[318, 368] our approach to curating PPIs from text-mining data has been to organize all individual events for a unique PPI from numerous publications onto a single web page, for which the data can then be validated

together to determine whether that unique PPI has been correctly extracted. This approach was designed primarily for validating interactions between human (or animal model) proteins in a particular class of disease. To see if this approach could be extended more broadly to curating interactions between human proteins and various pathogens we focused on three tasks/use cases, outlined here and in further detail in the Methods above.

1) Extending the HHPID

A large proportion of the HP-PPIs not present in public databases are from HIV-1 (1,313). We would perhaps expect this number to be much lower, if accurate, as the HIV-1 human protein interaction database (HHPID) has already been extensively curated from the literature, including access to full texts that are not openly accessible to text-mining. Therefore, it is of interest to determine if the HHPID is truly representative of the PPIs denoted in the literature and if useful data may have been missed.

We curated each HIV-1 associated PPI, assessing whether an HIV-1 protein had been documented to have an interaction of any kind with a human protein. Overall, we were only able to find 42 new HP-PPIs related to HIV-1, with an additional 12 reported negatively and 5 reported speculatively. Many of these new interactions were documented in the 1990s and these publications were unlikely to have been selected for curation in the HHPID.

However, while the new PPIs we have uncovered for HIV-1 with humans are no doubt useful, the remaining 95.5% of the HP-PPIs retrieved by TM not in the HHPID were incorrectly annotated. If we combine these figures with the 880 HP-PPIs from text-mining that were present in the HHPID, the overall precision for extracting HIV-1 HP-PPIs is 42.8% (939 are correct and 1 280 are incorrect). During the curation, we noted common causes of error such as names mapped to human Entrez Gene IDs that were not proteins or events that had been incorrectly assigned between proteins. For example, terms such as 'Fig' or 'HIV' itself were assigned to human protein IDs. These types of matches shared similar features with gene names and could potentially be resolved by a post-processing system.

2) Identifying additional pathogens that interact with commonly targeted human proteins

The intersection between HP-PPIs from text-mining and public databases revealed 2,670 human proteins that have interactions with various pathogens catalogued in public databases, but also with other pathogen species only identified by text-mining. To determine whether these pathogens had been correctly indentified as additional interactors, we selected high confidence HP-PPIs for 302 interactions between human proteins and new pathogen species. A pathogen was considered to interact with a human protein if at least one of its proteins had been documented to interact with it. This approach enabled us to find 108 new interactions between pathogens and commonly targeted human proteins (35.8% precision). 51 of these were direct physical interactions, while the remainder were indirect interactions.

3) Identifying new human protein targets for pathogens

To validate interactions between the 3,364 human proteins that had no reported interactions with any pathogens in the public databases, we selected human proteins for curation in two different ways. Firstly, we selected human proteins that had at least one high confidence text-mining interaction between it and a pathogen protein from any pathogen species. Secondly, we asked if enriched GO terms from human proteins that have been shown to interact with pathogens in public databases are a useful indicator of correctly identified human proteins interacting with pathogens from text-mining not in public databases. To test this we ranked candidate human proteins by their number of corresponding enriched GO terms from other known human protein pathogen targets, setting various thresholds of enrichment p values for the inclusion of each GO term (see Methods).

Table 5.5 shows the results for curating newly recognised human proteins targeted by pathogens in each of these ranking methods. In total these methods were able identify 33 new human proteins targeted by pathogens (21 direct

interactions and 12 indirect interactions). Human proteins ranked by number of enriched GO terms all had particularly low precision with no more than 12% of human proteins correctly identified as pathogen targets in each ranking. Those interactions between human proteins and pathogens from high confidence text-mining data showed slightly better precision at 17.2% and filtering proteins from these that had no enriched pathogen-associated GO terms had no significant effect on this score. We therefore concluded that the most effective way of identifying human proteins from text-mining for curation was by using text-mining confidence alone.

Protein ranking	Human Proteins curated	True positives
GO enriched ($p < 0.1$)	50	5 (10%)
GO enriched ($p < 1E5$)	50	6 (12%)
GO enriched ($p < 1E10$)	50	4 (8%)
Text-mining confidence (>0.3)	163	28 (17.2%)
GO enriched ($p <0.1$) and text-mining confidence (>0.3)	144	24 (16.7%)
GO enriched ($p <1E5$) and text-mining confidence (>0.3)	128	21 (16.4%)
GO enriched ($p <1E10$) and text-mining confidence (>0.3)	110	18 (18%)

Table 5.5 Ranking methods for curating unknown human targets for pathogens.
A true positive is a correctly annotated interaction between a human protein and any pathogen protein.

5.3 Discussion

In this study we set to out to determine how text-mining data can be utilized to expand our knowledge of interactions between pathogens and human proteins in public databases. It is clear that from the 1,419 pathogen species known to infect humans the PPIs relevant to these are only well represented for a few species (e.g. HIV-1 and *Y pestis*). Thus, to study how pathogenesis varies at a molecular level across all these pathogens many more PPIs need to be determined before accurate comparisons can be made.

While the published literature might not be the solution to uncovering all this knowledge, we have shown that it contains many PPIs that are not present in

public databases, finding 25.5K new HP-PPIs for 223 pathogens in 2.3M pathogen-relevant publications from text-mining data. However, we demonstrated that this data is not useful unless validated, as it contains significant numbers of incorrectly annotated PPIs. For example, the curation task producing the best precision equaled only 35.8% despite it involving high-confidence text-mining data and proteins already known to be targeted by pathogens.

The fact that the precision of the text-mining data was so low is perhaps surprising, given that in previous studies we had shown that PPIs derived from BioContext, used again in this study with Evex DB, showed precision of 84.2% when using high-confidence data[318]. However, one notable difference between this study and previous ones was that in this study we were seeking PPIs between two different species. This could make it more difficult to map proteins to their correct gene IDs, increasing the number of candidate IDs, particularly when the protein names are conserved across multiple species. Moreover, we demonstrated how the protein to ID mapping for pathogen proteins could be improved substantially when using more tailored approaches to normalisation. This had enabled us to uncover nearly 3K new HP-PPIs for five common viruses and it is foreseeable that tailored approaches to protein normalization like these could be as effective in other pathogens, particularly for those with smaller genomes, and thus with less protein names to decipher.

Conversely, while the text-mining data contained large numbers of incorrectly annotated interactions, our curation methods showed that these can be filtered out in an efficient manner. We showed that even for a host-pathogen PPI database that has been extensively curated from the literature, new PPIs can be found, that we can expand the range of pathogens known to interact with human protein targets and that new human protein targets for pathogens can be unearthed. To uncover more likely correctly identified new human protein targets for pathogens from text-mining we experimented with using GO terms to rank proteins more closely aligned with other known human protein targets, although text-mining confidence still proved a more useful ranking method.

In future work it is therefore viable to continue curating the new PPI data identified from text mining to further expand on the existing knowledge stored in public databases. To facilitate this, the text-mining data produced in this study is readily available for this task; however, in parallel, we will investigate further how text-mining data can be specifically improved for retrieving pathogen related PPIs. This could be through improving the quality of existing text-mining databases using tailored approaches to protein normalisation, as we have shown is useful in this study, or through creating entirely new datasets from text-mining, implementing tailored approaches to matching proteins, such as that used in Jamieson *et al* (2012)[286].

5.4 Acknowledgments

The authors would like to thank Craig Lawless, David Talavera and Ryan Ames from the University of Manchester for assistance with aspects of the pathogen protein interaction analysis conducted in this study.

CHAPTER 6

Discussion and Conclusion

In this thesis we have investigated how molecular interactions, derived from TM, can be utilized for more rigorous biological analyses in studying disease. We limited our investigation to studying two contrasting, but critically important classes of disease: pain-related interactions and host-pathogen interactions. In order to be able to analyse molecular interactions for these diseases we have followed a set TM approach of first identifying relevant literature, next extracting molecular interactions and their contexts, and finally validating this data through efficient large-scale curation. In each step we have introduced new methods, improved existing software and made available useful data for which other complementary research might be conducted. We will now discuss the merits and failings of these and what might be done to improve these in future research.

6.1 Building a disease specific corpus

If one is familiar with biomedical TM research, and this thesis is no different, a publication or presentation of such findings will very commonly begin by outlining the unprecedented number and growth of publications in biomedicine. This cliché is necessary to underline the messages that yes, these are insurmountable without computational assistance and no, this state of affairs is unlikely to change in the near future. For extracting molecular interactions specific to pain and pathogens, our solution to this has been to process publications limited to their fields by first building lists of relevant terms and then assessing their relevance to a publication.

Building lists of relevant terms was relatively straightforward. Indeed, we have been able to take advantage of the efforts of other researchers in utilizing species names for pathogens or have, for example, consulted with pain biologists to aid

with the creation of a new controlled vocabulary of 583 pain-related terms (see Chapter 3). We have then been able to supplement these efforts by using basic computational techniques, such as regular expressions for expanding these to cover a wider range of denotations, or more complex methods, such as ranking n-grams to identify new candidate terms.

Once we had created sufficiently broad lists of terms these were then matched directly to text in published articles (albeit through the utility of LINNAEUS[191]). Pathogen species terms are relatively unambiguous and are unlikely to be matched in error, whereas for pain terms we ensured only unambiguous terms were used to determine pain-relevant publications by assigning pain specificity scores to each term. This approach enabled us to create a corpus of 766K documents for pain, which we had empirically proven to be pain-relevant, while for pathogens we were able to associate 2.3M publications. These corpora were then vital starting points for deriving molecular interactions and other context specific to these fields.

However, for this approach to building corpora to be successful in other biomedical fields, like all NER, it is largely dependent on the ambiguity and breadth of the terms needed to assign relevant documents. For example, discerning relevant documents to a large family of genes might require a more advanced entity mapping solution similar to those for matching more generic entity types. Although for specific classes of disease, as for pain and pathogens, these methods are well suited to corpus building.

As well as being able to match publications for the overall fields of pain and pathogens, in Chapter 3 we experimented with ranking publications for their overall relevance to pain and the individual terms (and their categories) matched. We have shown that by utilizing the pain specificity scores we had given to each term, combined with where they were matched in a document (title, abstract etc.) that we could empirically predict how relevant a publication was to pain, a pain category or pain term. While these rankings were not needed for subsequent TM data or curation in this thesis, they may well prove useful in other areas where document triage within a corpus is more necessary.

6.2 Molecular interactions and their contexts

Rather than use software designed more specifically for extracting molecular interactions (e.g. He (2009)[372]) we had instead opted to use chains of biomolecular events, whom when containing two molecular participants resembled MIs. This choice was motivated by a need to understand more closely how individual proteins, genes and RNA molecules had been documented to interact, thereby affording us greater detail when coming to understand their biology in the complete system. Furthermore, the BioNLP shared tasks had provided a backdrop of high quality software available for achieving this goal.

Our first challenge in utilizing chains of biomolecular events had been how to evaluate them. Prior to our study presented in Chapter two, biomolecular events had only been evaluated as single entities and while useful for assessing the quality of each individual element of an overall chain an approach that evaluated the full linked chains was necessary. The 'stringent' and 'approximate' methods of evaluation we developed solved this issue and enabled accurate assessment of how well suited biomolecular events were for extracting molecular interactions.

In Chapter two, we first experimented with using these by linking them to molecule mentions in HIV-1 text. The mentions were derived from using a customized version of BANNER, with modified training data and post-processing, we had developed to improve precision and recall for matching HIV-1 genes, proteins and RNA molecules. However, while these did provide improvements to the performance of BANNER for this task, in Chapter 6 where we revisited matching HIV-1 proteins, it was more convenient to utilize existing TM data from BioContext and Evex DB[217, 367] that also employed more sophisticated normalization – particularly key for disambiguating human protein mentions.

By reusing these existing datasets (and also in Chapters three and four with BioContext), we were able to make use of data from previous TM research that had taken considerable efforts to produce. Indeed, it is perhaps disappointing that this practice is not more common. Since the publication of BioContext in

2012, with the exception of two publications from this thesis, it has been cited 10 times with only one of these studies making use of the data for biological analysis (see Wu *et al* (2013)[196]). For a database containing contextual biomedical events for the whole of Medline and PMC this should represent a goldmine for any biologist seeking to study the molecular mechanisms behind disease or to provide a ready dataset for integrating with new data types.

The data types we had enriched BioContext within this thesis were mutations linked to proteins and pain and disease relevance. Associating pain and disease relevance were two important innovations as they allowed MIs to be ranked for curation, providing a prediction of their relevance to a particular field. These predictions were fairly accurate in confirming an association, although no attempts were made to automatically qualify what these might be, as other initiatives have tried to[170]. However, the major advantage of our method was that these associations were made at the document level, enabling us to capture any associations between very distant candidates – a method more fitting for our purpose.

Can TM produce accurate MI data for immediate large-scale biological analysis? We believe in the majority of cases the answer to this is no, although it does depend on the type of MIs and the level of detail that is required. From a standalone perspective a MI, with its interaction type, proteins normalized to the correct species and gene IDs, and any negation or speculation indicated, will most commonly be incomplete or incorrectly annotated by TM. Our evaluations and others from BioContext and Evex DB have supported this assertion.

Conversely, from a biological perspective it is not always necessary to ensure absolute precision and recall for MI extraction. Our aims throughout this thesis had been to extract *unique* molecular interactions for a particular disease-type. This meant that we could use multiple instances of a unique MI to qualify its existence, relying on only one of these to be correct. Then, using the TM confidence for each instance it is possible to predict automatically how likely a molecular interaction is to be correct. In Chapter three we showed that by doing this for human, mice and rat MIs relevant to pain we could find 85% of these to

have been correctly extracted. Furthermore, achieving high recall of MIs becomes easier as MIs mentioned multiple times in different denotations increases the chances of TM capturing this knowledge correctly.

On the other hand, for this approach to be successful it is reliant on the TM data displaying a reasonable level of precision and recall to begin with. To this end, we found that the quality of generic TM data varied widely between MIs sourced for pain and those for pathogens. Pathogens PPIs from BioContext and Evex DB did not display high levels of precision comparable to pain and it is likely that many interactions were missed or incorrectly annotated. We had only recreated a tiny fraction (7.21%) of the pathogen interactions documented in public databases from this data and given the text available we would have expected this to be much higher.

However, in Chapter 5 we did show that by using a tailored approach to normalizing pathogen proteins in five viruses that we could dramatically improve the coverage of generic TM data for sourcing pathogen-related PPIs. Moreover, we suspect that if this approach was combined with the type of approach used to match HIV-1 proteins in Chapter 2 that we could increase this coverage even further. The issue in extracting host-pathogen interactions, we believe, is not in the matching and normalizing of proteins from species with small genomes (e.g. viruses like HIV-1), but for matching and normalising proteins from larger genomes (e.g. human proteins) and correctly ascertaining the events that determine the interactions between them. These are both difficult aspects to improve on as normalizing proteins becomes more problematic when dealing with larger numbers of proteins names (through the ambiguity of the terms and the breadth of terms required) and biomolecular events have to account for often complexly defined relationships.

6.3 Curating large scale molecular interaction datasets

To bridge the gap between producing MIs automatically and being able to examine them with confidence to infer insights into biological function and disease biology, we have designed and developed a new style of curating data

that, rather than reviewing data from individual publications, validated unique interactions mentioned across the entire literature. This approach to curation was extremely efficient, as once a single instance of a unique MI had been confirmed as correct, other instances of it were no longer needed for review. Up to 250 unique MIs a day could be curated in this way by a single person and when compared against manual curation tasks this finally gave TM a key role in providing the speed of data acquisition it had always promised and the accuracy it had heretofore lacked. While this approach to validating MIs did not ensure every single reported instance of a unique MI was correct, it had brought TM in line with the research goals of a major pharmaceutical company by providing data that could be used to conduct meaningful biological analysis.

We showed that this style of curation was not just useful for curating disease-related interactions between proteins, but also between host proteins and pathogen species and pathogens in general. Moreover, work outside of this thesis has been initiated in curating the expression of proteins in specific diseases for biomarker discovery and also for gaining insights into the overall roles of particular proteins across different diseases[373]. Provided each entity and its relations have been well defined by TM, there are few limits to how many data types stored in the literature might be suitable for this style of curation.

To visualize the data from TM for each curation task we had used an off-the-shelf solution in the Mediawiki framework and its APIs, to create static wiki pages. The most extensive wiki we had built in this thesis was wiki-pain.org, which served both as a platform for researchers to view summary information on PPIs and as a curation interface to provide all the necessary information to validate a PPI. While we believe this served both of these purposes well, a system that might encourage any visiting researchers into aiding with the curation might be of use (akin to other crowd sourcing systems). Indeed, this was a project that was recently initiated in a master's thesis at this University, but will require further work for it to reach its potential.

6.4 Harnessing MIs from TM for further use

There are three general methods for sourcing MIs in disease research: newly derived experimental data, those from former studies stored in public databases (e.g. iRefIndex, MiMi, IntAct etc.) and databases curated from the peer-reviewed literature. Each of these has their own strengths and weaknesses. Experimental data is often limited by the cost and time conducting the experiments, but allows flexibility in the kinds of data and level of detail that can be obtained. Public databases have the advantage of containing relatively large amounts of readily available data, but these often lack context for understanding the role of MIs in more detail. While databases curated from the literature can contain this context, but take large efforts to curate and are often expensive to access as a result (e.g. products from companies like Ingenuity and Biobase).

Now that we could curtail the problems with large-scale curation by TM we have shown that it is possible to build a new database of PPIs representative of an entire disease, as we did for pain (Chapters 3 and 4). The PPI database we built for pain had not only shown that it was highly relevant to this topic, but also that it was more representative than ones crafted from gene expression or manually curated gene lists, both connected by generic interactions from public databases. We suspect that this is due to the fact that interactions sourced from TM had come predominantly from experiments detailing interactions that occurred in pain states, whereas the generic interactions used to connect the pain gene lists may not have occurred in these.

Could we have obtained a list of PPIs only relevant to pain experimentally? This would be possible, though as we have mentioned it would have taken significant time and effort even for a pharmaceutical institution like Pfizer whom this project had been conducted in collaboration with. Here we see a key value proposition for TM. In being able to efficiently reconstruct existing knowledge contextually as we have, this style of approach has the ability to provide feasible solutions to otherwise unrealistic goals.

The value of the pain interactome we had constructed was obvious. By overlaying all existing drug target data onto enriched proteins in our network and applying pain categories to their relatedness to treat pain indications we had found a way of identifying new drug repurposing opportunities and targets for which novel drugs might be developed. Since these have been identified, further work has been conducted within Pfizer on their potential for further developments. We believe that many other diseases like those in pain could be studied in this way to deliver new therapeutic opportunities.

6.5 Future directions

Immediately, it is possible to continue curating interactions from the pain and pathogen studies. Further analysis of the current pain interactome is also possible, particularly through using the contexts, i.e. any of the pain-disease associations, anatomy etc., to provide further insights into pain diseases. This is likewise possible with the pathogens data, although we believe further improvements to the quality of the TM data and the addition of useful contexts (e.g. cell-types, disease stage etc.) would be advisable before any efforts to conduct a pan-pathogen style analysis are made.

Similar projects to those in this thesis have also been penned. For example, we have begun building a database of ion channel and solute carrier interactions, as these proteins are involved in many disorders for which new pharmacologic treatments are in need. Moreover, it is hoped our approach to curating unique knowledge will enable us to curate many other data types (e.g. for chemicals, diseases etc.) in a similar manner to how we have curated molecular interactions in this thesis.

A major vision of TM is to one-day offer complete confidence in the knowledge that the data outputted is as complete a representation of that that was described in the literature as possible – representing a true depiction of the scientific consensus and dissent. We believe that there are two major barriers that must be overcome for this aim to be met.

Firstly, TM must be able to demonstrate that it is capable of extracting this information, accounting for the unique writing styles and idiosyncrasies of scientists that form a global community of a huge range of sub-disciplines. Most TM approaches attempt to solve this by simplifying the ways in which this knowledge is conveyed. For example, core data concepts such as proteins, chemicals and diseases are removed from their ambiguous denotations and mapped to single IDs, while more descriptive terms, such as those used to signify negation or speculation are reduced to binary concepts. While these techniques are adequate in summarizing often complexly described scientific findings, they also leave behind vital context that the reader intended to convey.

How will this be resolved in the future? Initiatives such as the open biological expression language (BEL) project are attempting to by providing a framework and computable language for curating knowledge that aims to allow causal and correlative relationships and their context to be captured as they were originally described in the scientific literature. This community-wide effort between industry and academia holds promise, although it is in its infancy and it remains to be seen how well it will be developed to cover the more obscure scientific findings and whether it can represent quantitative data as well as qualitative findings.

However, even if languages like BEL are capable of representing all scientific findings in the literature it is unlikely that TM will ever be able to perfectly match all text to it. While advances are continually made to the quality of TM software and data outputted, scientific language is constantly evolving, and with the persistent introduction of new concepts and data types it is only likely to get more difficult. It may well be then that the best solution to ensuring that all future published data is computable is to fundamentally alter the culture in which knowledge is published in science. For example rather than allowing researchers to submit manuscripts written in their own linguistic styles, they would have to conform to a more standardized computable format. This could either replace the existing publication system or sit alongside it, allowing journal articles to still be produced for human consumption. Although, for these changes to occur they would most likely be gradual, would rely on the publishers to

enforce them (and these often have their own agendas) and would still not address the huge existing body of literature. To this end, we expect TM will always be necessary for conveying knowledge computationally.

At present, we believe a second major barrier to knowledge extraction from literature is having full access to all published literature in full text. This has been a persistent problem for biomedical text-miners over the years, where only abstracts and a small subset of open access articles have been available for use. It is clear that key biological data and knowledge is stored in full text and if access is granted to manually view that text then the same privileges should be extended to text-miners for automatically extracting information. However, with the recent change in UK law[233], it is hoped that publishers will now comply with providing such access, and it would be a welcome amendment in other countries too.

Greater access to full text is, however, likely to create new challenges. The processing demands are likely to dramatically increase and this will mean text-miners will have to take greater care in developing TM systems able to cope with these[227]. Furthermore, TM applied on full text is often more erroneous compared with TM applied on abstracts and titles and the novelty of the data retrieved is likely to vary more where work is referenced[230, 374]. Although these challenges would be welcome in exchange for more potential data and aspects such as applying TM confidence scores could help filter out data likely to be incorrect.

6.6 Conclusion

This thesis has shown that TM, when used with our approach to manual curation, can efficiently build accurate contextual databases of molecular interactions that are extremely useful for studying diseases. While the approach to data extraction does not ensure that all data published in the literature is captured, it is practical for assessing large-scale datasets, widely applicable across many disease domains and allows biologists to analyse TM data with confidence to make new and compelling findings. Using this approach we have contributed a new database of pain-specific molecular interactions, of which we

have confirmed 1,002 and provided detailed analyses and hypotheses that other researchers can now investigate further. In pathogen molecular interaction extraction, we found that TM software performed best for extracting molecular interactions when it had been tailored to the individual task and species, and we therefore encourage future studies to follow suit rather than designing software to extract general data more diffusely. By following on from our initial work in expanding the pathogen interactome it is possible that we will be able to continue to complement knowledge of the existing pathogen interactome with data provided by TM, where TM acts as both a support tool for databases already extensively curated (e.g. the HHPID) and a way of rapidly building new databases of pathogen interactions that are not already curated.

References

1. TURING, A.M. COMPUTING MACHINERY AND INTELLIGENCE. *Mind* **LIX**, 433-460 (1950).
2. Nikfarjam, A., Emadzadeh, E. & Muthaiyah, S. in Computer and Automation Engineering (ICCAE), 2010 The 2nd International Conference on, Vol. 4 256-2602010).
3. Thanh, H.T.P. & Meesad, P. Stock Market Trend Prediction Based on Text Mining of Corporate Web and Time Series Data. *Journal of Advanced Computational Intelligence and Intelligent Informatics* **18**, 22-31 (2014).
4. Corley, C.D., Cook, D.J., Mikler, A.R. & Singh, K.P. Text and structural data mining of influenza mentions in Web and social media. *Int J Environ Res Public Health* **7**, 596-615 (2010).
5. Olson, D.R., Konty, K.J., Paladini, M., Viboud, C. & Simonsen, L. Reassessing Google Flu Trends data for detection of seasonal and pandemic influenza: a comparative epidemiological study at three geographic scales. *PLoS computational biology* **9**, e1003256 (2013).
6. Ginsberg, J. et al. Detecting influenza epidemics using search engine query data. *Nature* **457**, 1012-1014 (2009).
7. Fraser, J. How to publish in biomedicine : 500 tips for success. (Radcliffe Medical, Abingdon; 1997).
8. Publish or perish. *Nature* **467**, 252 (2010).
9. Rawat, S. & Meena, S. Publish or perish: Where are we heading? *J Res Med Sci* **19**, 87-89 (2014).
10. Brooks, S.A., Lomax-Browne, H.J., Carter, T.M., Kinch, C.E. & Hall, D.M. Molecular interactions in cancer cell metastasis. *Acta histochemica* **112**, 3-25 (2010).
11. MacPherson, J.I., Dickerson, J.E., Pinney, J.W. & Robertson, D.L. Patterns of HIV-1 protein interaction identify perturbed host-cellular subsystems. *PLoS computational biology* **6**, e1000863 (2010).
12. in Mosby's Medical Dictionary, 8th editionhttp://medical-dictionary.thefreedictionary.com/disease; 2009).
13. Kinsella, K.G. Changes in life expectancy 1900-1990. *Am J Clin Nutr* **55**, 1196S-1202S (1992).
14. Oeppen, J. & Vaupel, J.W. Demography. Broken limits to life expectancy. *Science* **296**, 1029-1031 (2002).
15. Raftery, A.E., Chunn, J.L., Gerland, P. & Sevcikova, H. Bayesian probabilistic projections of life expectancy for all countries. *Demography* **50**, 777-801 (2013).
16. Portin, P. The birth and development of the DNA theory of inheritance: sixty years since the discovery of the structure of DNA. *J Genet* **93**, 293-302 (2014).
17. Finishing the euchromatic sequence of the human genome. *Nature* **431**, 931-945 (2004).

18. Pennisi, E. Genomics. ENCODE project writes eulogy for junk DNA. *Science* **337**, 1159, 1161 (2012).
19. Imming, P., Sinning, C. & Meyer, A. Drugs, their targets and the nature and number of drug targets. *Nat Rev Drug Discov* **5**, 821-834 (2006).
20. Hughes, J.P., Rees, S., Kalindjian, S.B. & Philpott, K.L. Principles of early drug discovery. *Br J Pharmacol* **162**, 1239-1249 (2011).
21. Stancu, C. & Sima, A. Statins: mechanism of action and effects. *J Cell Mol Med* **5**, 378-387 (2001).
22. Ramasubbu, K., Estep, J., White, D.L., Deswal, A. & Mann, D.L. Experimental and clinical basis for the use of statins in patients with ischemic and nonischemic cardiomyopathy. *J Am Coll Cardiol* **51**, 415-426 (2008).
23. Hansson, E., Svensson, H. & Brorson, H. Review of Dercum's disease and proposal of diagnostic criteria, diagnostic methods, classification and management. *Orphanet J Rare Dis* **7**, 23 (2012).
24. Kobari, M., Nogawa, S., Sugimoto, Y. & Fukuuchi, Y. Familial idiopathic brain calcification with autosomal dominant inheritance. *Neurology* **48**, 645-649 (1997).
25. Murray, C.J. et al. Global, regional, and national incidence and mortality for HIV, tuberculosis, and malaria during 1990-2013: a systematic analysis for the Global Burden of Disease Study 2013. *Lancet* (2014).
26. in Defining Cancer, Vol. 2014 (National Cancer Institute, http://www.cancer.gov/cancertopics/cancerlibrary/what-is-cancer; 2014).
27. Semenza, G.L. Targeting HIF-1 for cancer therapy. *Nat Rev Cancer* **3**, 721-732 (2003).
28. Wang, R., Zhou, S. & Li, S. Cancer therapeutic agents targeting hypoxia-inducible factor-1. *Curr Med Chem* **18**, 3168-3189 (2011).
29. Panerai, A.E. Pain emotion and homeostasis. *Neurol Sci* **32 Suppl 1**, S27-29 (2011).
30. Taylor, L.H., Latham, S.M. & Woolhouse, M.E. Risk factors for human disease emergence. *Philos Trans R Soc Lond B Biol Sci* **356**, 983-989 (2001).
31. Woolhouse, M. & Gaunt, E. Ecological origins of novel human pathogens. *Crit Rev Microbiol* **33**, 231-242 (2007).
32. Woolhouse, M.E. & Gowtage-Sequeria, S. Host range and emerging and reemerging pathogens. *Emerg Infect Dis* **11**, 1842-1847 (2005).
33. Gruenberg, J. & van der Goot, F.G. Mechanisms of pathogen entry through the endosomal compartments. *Nat Rev Mol Cell Biol* **7**, 495-504 (2006).
34. Hancock, R.E., Nijnik, A. & Philpott, D.J. Modulating immunity as a therapy for bacterial infections. *Nat Rev Microbiol* **10**, 243-254 (2012).
35. Raoult, D. & Forterre, P. Redefining viruses: lessons from Mimivirus. *Nat Rev Microbiol* **6**, 315-319 (2008).
36. Burnet, M. General discussion of virus nomenclature. *Ann N Y Acad Sci* **56**, 621-622 (1953).
37. Burnet, M. Virus classification and nomenclature. *Ann N Y Acad Sci* **56**, 383-390 (1953).
38. Lwoff, A. The concept of virus. *J Gen Microbiol* **17**, 239-253 (1957).
39. La Scola, B. et al. A giant virus in amoebae. *Science* **299**, 2033 (2003).

40. Behbehani, A.M. The smallpox story: life and death of an old disease. *Microbiol Rev* **47**, 455-509 (1983).
41. Lacey, B.W. The natural history of poliomyelitis. *Lancet* **1**, 849-859 (1949).
42. Kurth, R. & Bannert, N. Retroviruses : molecular biology, genomics and pathogenesis. (Caister Academic, [Wymondham]; 2010).
43. Ganem, D. & Prince, A.M. Hepatitis B virus infection--natural history and clinical consequences. *N Engl J Med* **350**, 1118-1129 (2004).
44. Poehlmann, S.e.o.c. & Simmons, G.e.o.c. Viral entry into host cells.
45. Cormier, E.G. et al. Specific interaction of CCR5 amino-terminal domain peptides containing sulfotyrosines with HIV-1 envelope glycoprotein gp120. *Proceedings of the National Academy of Sciences of the United States of America* **97**, 5762-5767 (2000).
46. Schelhaas, M. Come in and take your coat off - how host cells provide endocytosis for virus entry. *Cell Microbiol* **12**, 1378-1388 (2010).
47. Huang, D.B., Wu, J.J. & Tyring, S.K. A review of licensed viral vaccines, some of their safety concerns, and the advances in the development of investigational viral vaccines. *The Journal of infection* **49**, 179-209 (2004).
48. Barouch, D.H. Challenges in the development of an HIV-1 vaccine. *Nature* **455**, 613-619 (2008).
49. De Clercq, E. Strategies in the design of antiviral drugs. *Nature reviews. Drug discovery* **1**, 13-25 (2002).
50. Lurain, N.S. & Chou, S. Antiviral drug resistance of human cytomegalovirus. *Clinical microbiology reviews* **23**, 689-712 (2010).
51. Tan, Q. et al. Structure of the CCR5 chemokine receptor-HIV entry inhibitor maraviroc complex. *Science* **341**, 1387-1390 (2013).
52. Costello, E.K. et al. Bacterial community variation in human body habitats across space and time. *Science* **326**, 1694-1697 (2009).
53. Dethlefsen, L., McFall-Ngai, M. & Relman, D.A. An ecological and evolutionary perspective on human-microbe mutualism and disease. *Nature* **449**, 811-818 (2007).
54. Parkhill, J. et al. Genome sequence of Yersinia pestis, the causative agent of plague. *Nature* **413**, 523-527 (2001).
55. McClelland, M. et al. Complete genome sequence of Salmonella enterica serovar Typhimurium LT2. *Nature* **413**, 852-856 (2001).
56. Knodler, L.A. et al. Dissemination of invasive Salmonella via bacterial-induced extrusion of mucosal epithelia. *Proceedings of the National Academy of Sciences of the United States of America* **107**, 17733-17738 (2010).
57. Young, D., Hussell, T. & Dougan, G. Chronic bacterial infections: living with unwanted guests. *Nat Immunol* **3**, 1026-1032 (2002).
58. Lipps, G. Plasmids : current research and future trends. (Caister Academic, Wymondham; 2008).
59. Koch, A.L. Control of the bacterial cell cycle by cytoplasmic growth. *Crit Rev Microbiol* **28**, 61-77 (2002).
60. Briken, V. Molecular mechanisms of host-pathogen interactions and their potential for the discovery of new drug targets. *Curr Drug Targets* **9**, 150-157 (2008).

61. The evolving threat of antimicrobial resistance : options for action. (World Health Organization, Geneva; 2012).
62. Wilson, L.S. et al. The direct cost and incidence of systemic fungal infections. *Value Health* **5**, 26-34 (2002).
63. Moran, G.P., Coleman, D.C. & Sullivan, D.J. Comparative genomics and the evolution of pathogenicity in human pathogenic fungi. *Eukaryot Cell* **10**, 34-42 (2011).
64. Van Thiel, D.H., George, M. & Moore, C.M. Fungal Infections: Their Diagnosis and Treatment in Transplant Recipients. *International Journal of Hepatology* **2012**, 19 (2012).
65. Brown, G.D. et al. Hidden killers: human fungal infections. *Sci Transl Med* **4**, 165rv113 (2012).
66. Paterson, D.L. & Singh, N. Invasive aspergillosis in transplant recipients. *Medicine (Baltimore)* **78**, 123-138 (1999).
67. Kobayashi, G.S. Disease of Mechanisms of Fungi. (1996).
68. Simon-Nobbe, B., Denk, U., Poll, V., Rid, R. & Breitenbach, M. The spectrum of fungal allergy. *Int Arch Allergy Immunol* **145**, 58-86 (2008).
69. Richardson, M.D. & Warnock, D.W. Fungal infection : diagnosis and management, Edn. 3rd ed. (Blackwell, Oxford; 2003).
70. Hotez, P.J. et al. Helminth infections: the great neglected tropical diseases. *J Clin Invest* **118**, 1311-1321 (2008).
71. Hotez, P.J. et al. Control of neglected tropical diseases. *N Engl J Med* **357**, 1018-1027 (2007).
72. Brooker, S. Estimating the global distribution and disease burden of intestinal nematode infections: adding up the numbers--a review. *Int J Parasitol* **40**, 1137-1144 (2010).
73. Hotez, P.J. et al. Incorporating a rapid-impact package for neglected tropical diseases with programs for HIV/AIDS, tuberculosis, and malaria. *PLoS Med* **3**, e102 (2006).
74. Allen, J.E. & Maizels, R.M. Diversity and dialogue in immunity to helminths. *Nat Rev Immunol* **11**, 375-388 (2011).
75. Mascarini-Serra, L. Prevention of Soil-transmitted Helminth Infection. *Journal of global infectious diseases* **3**, 175-182 (2011).
76. Panic, G., Duthaler, U., Speich, B. & Keiser, J. Repurposing drugs for the treatment and control of helminth infections. *International Journal for Parasitology: Drugs and Drug Resistance* **4**, 185-200 (2014).
77. Zucca, M. & Savoia, D. Current developments in the therapy of protozoan infections. *Open Med Chem J* **5**, 4-10 (2011).
78. Pink, R., Hudson, A., Mouries, M.A. & Bendig, M. Opportunities and challenges in antiparasitic drug discovery. *Nat Rev Drug Discov* **4**, 727-740 (2005).
79. Reguera, R.M., Diaz-Gonzalez, R., Perez-Pertejo, Y. & Balana-Fouce, R. Characterizing the bi-subunit type IB DNA topoisomerase of Leishmania parasites; a novel scenario for drug intervention in trypanosomatids. *Curr Drug Targets* **9**, 966-978 (2008).
80. Gupta, Y.K., Gupta, M., Aneja, S. & Kohli, K. Current drug therapy of protozoal diarrhoea. *Indian journal of pediatrics* **71**, 55-58 (2004).

81. Bray, P.G., Barrett, M.P., Ward, S.A. & de Koning, H.P. Pentamidine uptake and resistance in pathogenic protozoa: past, present and future. *Trends in parasitology* **19**, 232-239 (2003).
82. Petri, W.A., Jr. Therapy of intestinal protozoa. *Trends in parasitology* **19**, 523-526 (2003).
83. Prusiner, S.B. Novel proteinaceous infectious particles cause scrapie. *Science* **216**, 136-144 (1982).
84. Farquhar, C.F., Somerville, R.A. & Bruce, M.E. Straining the prion hypothesis. *Nature* **391**, 345-346 (1998).
85. Chesebro, B. BSE and prions: uncertainties about the agent. *Science* **279**, 42-43 (1998).
86. Prusiner, S.B. Molecular biology of prion diseases. *Science* **252**, 1515-1522 (1991).
87. Araujo, A.Q.-C. Prionic diseases. *Arquivos de Neuro-Psiquiatria* **71**, 731-737 (2013).
88. Somerville, R.A. TSE agent strains and PrP: reconciling structure and function. *Trends Biochem Sci* **27**, 606-612 (2002).
89. Dirix, P., Vanhoenacker, F.M., Staumont, G. & Gille, M. Creutzfeldt-Jacob disease. *JBR-BTR* **89**, 128-129 (2006).
90. Krejciova, Z. et al. Genotype-Dependent Molecular Evolution of Sheep BSE Prions In Vitro Affects Their Zoonotic Potential. *The Journal of biological chemistry* (2014).
91. Woolf, C.J. What is this thing called pain? *J Clin Invest* **120**, 3742-3744 (2010).
92. Costigan, M., Scholz, J. & Woolf, C.J. Neuropathic pain: a maladaptive response of the nervous system to damage. *Annu Rev Neurosci* **32**, 1-32 (2009).
93. Theodorou, S.D., Klimentopoulou, A.E. & Papalouka, E. Congenital insensitivity to pain with anhidrosis. Report of a case and review of the literature. *Acta Orthop Belg* **66**, 137-145 (2000).
94. Elzahaf, R.A., Tashani, O.A., Unsworth, B.A. & Johnson, M.I. The prevalence of chronic pain with an analysis of countries with a Human Development Index less than 0.9: a systematic review without meta-analysis. *Curr Med Res Opin* **28**, 1221-1229 (2012).
95. Cohen, S.P. & Mao, J. Neuropathic pain: mechanisms and their clinical implications. *Bmj* **348**, f7656 (2014).
96. Breivik, H., Collett, B., Ventafridda, V., Cohen, R. & Gallacher, D. Survey of chronic pain in Europe: prevalence, impact on daily life, and treatment. *Eur J Pain* **10**, 287-333 (2006).
97. Lynch, M.E. & Watson, C.P. The pharmacotherapy of chronic pain: a review. *Pain research & management : the journal of the Canadian Pain Society = journal de la societe canadienne pour le traitement de la douleur* **11**, 11-38 (2006).
98. Merskey, H. & Bogduk, N. Classification of chronic pain : descriptions of chronic pain syndromes and definitions of pain terms, Edn. 2nd ed. (IASP Press, Seattle; 2012).
99. Gwak, Y.S. & Hulsebosch, C.E. GABA and central neuropathic pain following spinal cord injury. *Neuropharmacology* **60**, 799-808 (2011).

100. Foley, P.L. et al. Prevalence and natural history of pain in adults with multiple sclerosis: systematic review and meta-analysis. *Pain* **154**, 632-642 (2013).
101. Andersen, G., Vestergaard, K., Ingeman-Nielsen, M. & Jensen, T.S. Incidence of central post-stroke pain. *Pain* **61**, 187-193 (1995).
102. Baron, R., Binder, A. & Wasner, G. Neuropathic pain: diagnosis, pathophysiological mechanisms, and treatment. *The Lancet. Neurology* **9**, 807-819 (2010).
103. Kouyoumdjian, J.A. Peripheral nerve injuries: a retrospective survey of 456 cases. *Muscle & nerve* **34**, 785-788 (2006).
104. Callaghan, B.C., Cheng, H.T., Stables, C.L., Smith, A.L. & Feldman, E.L. Diabetic neuropathy: clinical manifestations and current treatments. *The Lancet. Neurology* **11**, 521-534 (2012).
105. Mangus, L.M. et al. Unraveling the pathogenesis of HIV peripheral neuropathy: insights from a simian immunodeficiency virus macaque model. *ILAR journal / National Research Council, Institute of Laboratory Animal Resources* **54**, 296-303 (2014).
106. Lee, Y.C., Nassikas, N.J. & Clauw, D.J. The role of the central nervous system in the generation and maintenance of chronic pain in rheumatoid arthritis, osteoarthritis and fibromyalgia. *Arthritis research & therapy* **13**, 211 (2011).
107. Sanford, D., Thornley, P., Teriaky, A., Chande, N. & Gregor, J. Opioid use is associated with decreased quality of life in patients with Crohn's disease. *Saudi journal of gastroenterology : official journal of the Saudi Gastroenterology Association* **20**, 182-187 (2014).
108. Alfredson, H. The chronic painful Achilles and patellar tendon: research on basic biology and treatment. *Scandinavian journal of medicine & science in sports* **15**, 252-259 (2005).
109. Khan, K.M., Cook, J.L., Kannus, P., Maffulli, N. & Bonar, S.F. Time to abandon the "tendinitis" myth. *Bmj* **324**, 626-627 (2002).
110. Lynch, M.E. & Campbell, F. Cannabinoids for treatment of chronic non-cancer pain; a systematic review of randomized trials. *Br J Clin Pharmacol* **72**, 735-744 (2011).
111. Black, W.M.D. An Historical Sketch of Medicine and Surgery, from their origin to the present time, etc. [With a table.]. (J. Johnson, London; 1782).
112. Beasley, S.W. The value of medical publications: 'to read them would...burden the memory to no useful purpose'. *Aust N Z J Surg* **70**, 870-874 (2000).
113. Welch, S.J. Selecting the right journal for your submission. *J Thorac Dis* **4**, 336-338 (2012).
114. NLM, Vol. 2014 (NIH, http://www.nlm.nih.gov/pubs/factsheets/medline.html; 2013).
115. NLM, Vol. 2014 (NIH, http://www.nlm.nih.gov/pubs/factsheets/dif_med_pub.html; 2013).
116. Falagas, M.E., Pitsouni, E.I., Malietzis, G.A. & Pappas, G. Comparison of PubMed, Scopus, Web of Science, and Google Scholar: strengths and weaknesses. *The FASEB journal : official publication of the Federation of American Societies for Experimental Biology* **22**, 338-342 (2008).

117. Henzinger, M. & Lawrence, S. Extracting knowledge from the World Wide Web. *Proceedings of the National Academy of Sciences of the United States of America* **101 Suppl 1**, 5186-5191 (2004).
118. Stevenson, M. & Guo, Y. Disambiguation of ambiguous biomedical terms using examples generated from the UMLS Metathesaurus. *Journal of biomedical informatics* **43**, 762-773 (2010).
119. Weeber, M., Mork, J.G. & Aronson, A.R. Developing a test collection for biomedical word sense disambiguation. *Proc AMIA Symp*, 746-750 (2001).
120. Bodenreider, O. The Unified Medical Language System (UMLS): integrating biomedical terminology. *Nucleic acids research* **32**, D267-270 (2004).
121. Nourbakhsh, E., Nugent, R., Wang, H., Cevik, C. & Nugent, K. Medical literature searches: a comparison of PubMed and Google Scholar. *Health Info Libr J* **29**, 214-222 (2012).
122. Anders, M.E. & Evans, D.P. Comparison of PubMed and Google Scholar literature searches. *Respir Care* **55**, 578-583 (2010).
123. McCrae, J. & Collier, N. Synonym set extraction from the biomedical literature by lexical pattern discovery. *BMC bioinformatics* **9**, 159 (2008).
124. Golbeck, J. et al. The National Cancer Institute's Thésaurus and Ontology. *http://www.mindswap.org/papers/WebSemantics-NCI.pdf* (2004).
125. Liu, H., Hu, Z.Z., Zhang, J. & Wu, C. BioThesaurus: a web-based thesaurus of protein and gene names. *Bioinformatics* **22**, 103-105 (2006).
126. Schriml, L.M. et al. Disease Ontology: a backbone for disease semantic integration. *Nucleic acids research* **40**, D940-946 (2012).
127. Natale, D.A. et al. The Protein Ontology: a structured representation of protein forms and complexes. *Nucleic acids research* **39**, D539-545 (2011).
128. Ashburner, M. et al. Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nat Genet* **25**, 25-29 (2000).
129. Islamaj Dogan, R., Murray, G.C., Neveol, A. & Lu, Z. Understanding PubMed user search behavior through log analysis. *Database (Oxford)* **2009**, bap018 (2009).
130. Yu, H. et al. Enabling multi-level relevance feedback on PubMed by integrating rank learning into DBMS. *BMC bioinformatics* **11 Suppl 2**, S6 (2010).
131. Cheng, D. et al. PolySearch: a web-based text mining system for extracting relationships between human diseases, genes, mutations, drugs and metabolites. *Nucleic acids research* **36**, W399-405 (2008).
132. Doms, A. & Schroeder, M. GoPubMed: exploring PubMed with the Gene Ontology. *Nucleic acids research* **33**, W783-786 (2005).
133. Bettembourg, C., Diot, C., Burgun, A. & Dameron, O. GO2PUB: Querying PubMed with semantic expansion of gene ontology terms. *J Biomed Semantics* **3**, 7 (2012).
134. Gong, W. et al. PepCyber:P~PEP: a database of human protein protein interactions mediated by phosphoprotein-binding domains. *Nucleic acids research* **36**, D679-683 (2008).

135. Orchard, S. et al. The MIntAct project--IntAct as a common curation platform for 11 molecular interaction databases. *Nucleic acids research* **42**, D358-363 (2014).
136. Fu, W. et al. Human immunodeficiency virus type 1, human protein interaction database at NCBI. *Nucleic acids research* **37**, D417-422 (2009).
137. Chatr-Aryamontri, A. et al. The BioGRID interaction database: 2013 update. *Nucleic acids research* **41**, D816-823 (2013).
138. Goll, J. et al. MPIDB: the microbial protein interaction database. *Bioinformatics* **24**, 1743-1744 (2008).
139. Kwon, D. et al. A comprehensive manually curated protein-protein interaction database for the Death Domain superfamily. *Nucleic acids research* **40**, D331-336 (2012).
140. Pagel, P. et al. The MIPS mammalian protein-protein interaction database. *Bioinformatics* **21**, 832-834 (2005).
141. Ruepp, A. et al. CORUM: the comprehensive resource of mammalian protein complexes--2009. *Nucleic acids research* **38**, D497-501 (2010).
142. Salwinski, L. et al. The Database of Interacting Proteins: 2004 update. *Nucleic acids research* **32**, D449-451 (2004).
143. Keshava Prasad, T.S. et al. Human Protein Reference Database--2009 update. *Nucleic acids research* **37**, D767-772 (2009).
144. Breuer, K. et al. InnateDB: systems biology of innate immunity and beyond--recent updates and continuing curation. *Nucleic acids research* **41**, D1228-1233 (2013).
145. Navratil, V. et al. VirHostNet: a knowledge base for the management and the analysis of proteome-wide virus-host interaction networks. *Nucleic acids research* **37**, D661-668 (2009).
146. Ananiadou, S. & McNaught, J. Text Mining for Biology and Biomedicine. (Boston, MA: Artech House; 2006).
147. Baumgartner, W.A., Jr., Cohen, K.B., Fox, L.M., Acquaah-Mensah, G. & Hunter, L. Manual curation is not sufficient for annotation of genomic databases. *Bioinformatics* **23**, i41-48 (2007).
148. Seringhaus, M.R. & Gerstein, M.B. Publishing perishing? Towards tomorrow's information architecture. *BMC bioinformatics* **8**, 17 (2007).
149. Ptak, R.G. et al. Cataloguing the HIV type 1 human protein interaction network. *AIDS research and human retroviruses* **24**, 1497-1502 (2008).
150. Martinez-Uribe, L. & Macdonald, S. in Research and Advanced Technology for Digital Libraries, Vol. 5714. (eds. M. Agosti, J. Borbinha, S. Kapidakis, C. Papatheodorou & G. Tsakonas) 309-314 (Springer Berlin Heidelberg, 2009).
151. Ceol, A., Chatr-Aryamontri, A., Licata, L. & Cesareni, G. Linking entries in protein interaction database to structured text: the FEBS Letters experiment. *FEBS Lett* **582**, 1171-1177 (2008).
152. Tikk, D., Thomas, P., Palaga, P., Hakenberg, J. & Leser, U. A comprehensive benchmark of kernel methods to extract protein-protein interactions from literature. *PLoS computational biology* **6**, e1000837 (2010).
153. Sateli, B., Luong, S. & Witte, R. TagCurate: crowdsourcing the verification of biomedical annotations to mobile users. *2013* **19** (2013).

154. Burger, J. et al. in Data Integration in the Life Sciences, Vol. 7348. (eds. O. Bodenreider & B. Rance) 83-91 (Springer Berlin Heidelberg, 2012).
155. Hearst, M.A. Untangling Text Data Mining. *ACL '99 Proceeding of the 37th annual meeting of the Association for Computational Linguistics on the Computational Linguistics*, 3-10 (1999).
156. Kostoff, R.N. & DeMarco, R.A. Extracting information from the literature by text mining. *Analytical chemistry* **73**, 370A-378A (2001).
157. Tanabe, L. et al. MedMiner: an Internet text-mining tool for biomedical information, with application to gene expression profiling. *Biotechniques* **27**, 1210-1214, 1216-1217 (1999).
158. Fukuda, K.I., Tsunoda, T., Tamura, A. & Takagi, T. Toward information extraction: identifying protein names from biological papers. . *In Pac Symp Biocomput* **707**, 707-718 (1998).
159. Rebholz-Schuhmann, D. et al. CALBC silver standard corpus. *Journal of bioinformatics and computational biology* **8**, 163-179 (2010).
160. Leitner, F. et al. An Overview of BioCreative II.5. *IEEE/ACM Trans Comput Biol Bioinform* **7**, 385-399 (2010).
161. Carroll, H.D., Kann, M.G., Sheetlin, S.L. & Spouge, J.L. Threshold Average Precision (TAP-k): a measure of retrieval designed for bioinformatics. *Bioinformatics* **26**, 1708-1713 (2010).
162. Lu, Z. et al. The gene normalization task in BioCreative III. *BMC bioinformatics* **12 Suppl 8**, S2 (2011).
163. Matthews, B.W. Comparison of the predicted and observed secondary structure of T4 phage lysozyme. *Biochim Biophys Acta* **405**, 442-451 (1975).
164. Hirschman, L., Yeh, A., Blaschke, C. & Valencia, A. Overview of BioCreAtIvE: critical assessment of information extraction for biology. *BMC bioinformatics* **6 Suppl 1**, S1 (2005).
165. Morgan, A.A. et al. Overview of BioCreative II gene normalization. *Genome Biol* **9 Suppl 2**, S3 (2008).
166. Arighi, C.N. et al. Overview of the BioCreative III Workshop. *BMC bioinformatics* **12 Suppl 8**, S1 (2011).
167. Arighi, C.N. et al. BioCreative-IV virtual issue. *Database (Oxford)* **2014** (2014).
168. Kim, J.D. et al. Overview of BioNLP Shared Task 2011. *Proceedings of BioNLP Shared Task 2011 Workshop* **Portland, Oregon, USA**, 1-6 (2011).
169. Kim, J.D., Ohta, T., Pyysalo, S., Kano, Y. & Tsujii, J. Overview of BioNLP'09 Shared Task on Event Extraction. *Proceedings of the Workshop on BioNLP: Shared Task*, 1-9 (2009).
170. Nedellec, C. et al. in Proceedings of the BioNLP Shared Task 2013 Workshop 1-7Sofia, Bulgaria; 2013).
171. Segura-Bedmar, I., Martinez, P. & Sanchez-Cisneros, D. in Proceedings of the 1st Challenge Task on Drug-Drug Interaction Extraction 2011 1-9Huelva, Spain; 2011).
172. Segura-Bedmar, I., Martinez, P. & Herrero Zazo, M. in Joint Conference on Lexical and Computational Semantics2013).
173. Arighi, C.N. et al. BioCreative III interactive task: an overview. *BMC bioinformatics* **12 Suppl 8**, S4 (2011).

174. Feldman, R. & Sanger, J. The text mining handbook : advanced approaches in analyzing unstructured data. (Cambridge University Press, Cambridge; 2007).
175. Jiang, J. & Zhai, C. An empirical study of tokenization strategies for biomedical information retrieval. *Information Retrieval* **10**, 341-363 (2007).
176. Tomanek, K., Wermter, J. & Hahn, U. Sentence and Token Splitting Based on Conditional Random Fields. *PACLING 2007 - Proceedings of the 10th Conference of the Pacific Association for Computational Linguistics, Melbourne, Australia, September 19-21, 2007*, 49-57 (2007).
177. Hassler, M. & Fliedle, G. Text preparation through extended tokenization. (WIT Press/Computational Mechanics Publications, 2006).
178. Barrett, N. & Weber-Jahnke, J. Building a biomedical tokenizer using the token lattice design pattern and the adapted Viterbi algorithm. *BMC bioinformatics* **12 Suppl 3**, S1 (2011).
179. Cohen, K.B. & Hunter, L.E. Chapter 16: text mining for translational bioinformatics. *PLoS computational biology* **9**, e1003044 (2013).
180. Tsuruoka, Y. et al. in Advances in Informatics, Vol. 3746. (eds. P. Bozanis & E. Houstis) 382-392 (Springer Berlin Heidelberg, 2005).
181. Smith, L., Rindflesch, T. & Wilbur, W.J. MedPost: a part-of-speech tagger for bioMedical text. *Bioinformatics* **20**, 2320-2321 (2004).
182. Fundel, K., Kuffner, R. & Zimmer, R. RelEx--relation extraction using dependency parse trees. *Bioinformatics* **23**, 365-371 (2007).
183. Rocktaschel, T., Weidlich, M. & Leser, U. ChemSpot: a hybrid system for chemical named entity recognition. *Bioinformatics* **28**, 1633-1640 (2012).
184. Jessop, D.M., Adams, S.E., Willighagen, E.L., Hawizy, L. & Murray-Rust, P. OSCAR4: a flexible architecture for chemical text-mining. *J Cheminform* **3**, 41 (2011).
185. Krallinger, M. et al. in BioCreative IV - CHEMDNER track, Vol. 2 (Proceedings of the fourth BioCreative challenge evaluation workshop, 2013).
186. Leaman, R. & Gonzalez, G. BANNER: an executable survey of advances in biomedical named entity recognition. *Pacific Symposium on Biocomputing. Pacific Symposium on Biocomputing*, 652-663 (2008).
187. Wei, C.H. & Kao, H.Y. Cross-species gene normalization by species inference. *BMC bioinformatics* **12 Suppl 8**, S5 (2011).
188. Kinoshita, S., Cohen, K.B., Ogren, P.V. & Hunter, L. BioCreAtIvE task1A: entity identification with a stochastic tagger. *BMC bioinformatics* **6 Suppl 1**, S4 (2005).
189. Settles, B. ABNER: an open source tool for automatically tagging genes, proteins and other entity names in text. *Bioinformatics* **21**, 3191-3192 (2005).
190. Smith, L. et al. Overview of BioCreative II gene mention recognition. *Genome Biol* **9 Suppl 2**, S2 (2008).
191. Gerner, M., Nenadic, G. & Bergman, C.M. LINNAEUS: a species name identification system for biomedical literature. *BMC bioinformatics* **11**, 85 (2010).

192. Wei, C.H., Kao, H.Y. & Lu, Z. SR4GN: a species recognition software tool for gene normalization. *PLoS One* **7**, e38460 (2012).
193. Caporaso, J.G., Baumgartner, W.A., Jr., Randolph, D.A., Cohen, K.B. & Hunter, L. MutationFinder: a high-performance system for extracting point mutation mentions from text. *Bioinformatics* **23**, 1862-1865 (2007).
194. Wei, C.H., Harris, B.R., Kao, H.Y. & Lu, Z. tmVar: a text mining approach for extracting sequence variants in biomedical literature. *Bioinformatics* **29**, 1433-1439 (2013).
195. Naderi, N. & Witte, R. Automated extraction and semantic analysis of mutation impacts from the biomedical literature. *BMC Genomics* **13 Suppl 4**, S10 (2012).
196. Wu, C., Schwartz, J.M. & Nenadic, G. PathNER: a tool for systematic identification of biological pathway mentions in the literature. *BMC systems biology* **7 Suppl 3**, S2 (2013).
197. Leaman, R., Islamaj Dogan, R. & Lu, Z. DNorm: disease name normalization with pairwise learning to rank. *Bioinformatics* **29**, 2909-2917 (2013).
198. Stachelscheid, H. et al. CellFinder: a cell data repository. *Nucleic acids research* **42**, D950-958 (2014).
199. Neves, M. et al. Preliminary evaluation of the CellFinder literature curation pipeline for gene expression in kidney cells and anatomical parts. *Database (Oxford)* **2013**, bat020 (2013).
200. Campos, D., Matos, S. & Oliveira, J.L. Biomedical Named Entity Recognition: A Survey of Machine-Learning Tools, Theory and Applications for Advanced Text Mining. (InTech, http://www.intechopen.com/books/theory-and-applications-for-advanced-text-mining/biomedical-named-entity-recognition-a-survey-of-machine-learning-tools; 2012).
201. Fundel, K. & Zimmer, R. Gene and protein nomenclature in public databases. *BMC bioinformatics* **7**, 372 (2006).
202. Maglott, D., Ostell, J., Pruitt, K.D. & Tatusova, T. Entrez Gene: gene-centered information at NCBI. *Nucleic acids research* **33**, D54-58 (2005).
203. Boutet, E., Lieberherr, D., Tognolli, M., Schneider, M. & Bairoch, A. UniProtKB/Swiss-Prot. *Methods Mol Biol* **406**, 89-112 (2007).
204. Leser, U. & Hakenberg, J. What makes a gene name? Named entity recognition in the biomedical literature. *Brief Bioinform* **6**, 357-369 (2005).
205. Tsuruoka, Y. & Tsujii, J. Improving the performance of dictionary-based approaches in protein name recognition. *Journal of biomedical informatics* **37**, 461-470 (2004).
206. Cohen, A.M. & Hersh, W.R. A survey of current work in biomedical text mining. *Brief Bioinform* **6**, 57-71 (2005).
207. Isozaki, H. & Kazawa, H. in Proceedings of the 19th international conference on Computational linguistics - Volume 1 1-7 (Association for Computational Linguistics, Taipei, Taiwan; 2002).
208. Saha, S.K., Sarkar, S. & Mitra, P. Feature selection techniques for maximum entropy based biomedical named entity recognition. *Journal of biomedical informatics* **42**, 905-911 (2009).

209. Li, Y., Lin, H. & Yang, Z. Incorporating rich background knowledge for gene named entity classification and recognition. *BMC bioinformatics* **10**, 223 (2009).
210. Quan, C., Wang, M. & Ren, F. An unsupervised text mining method for relation extraction from biomedical literature. *PLoS One* **9**, e102039 (2014).
211. Hu, Y., Li, Y., Lin, H., Yang, Z. & Cheng, L. Integrating various resources for gene name normalization. *PLoS One* **7**, e43558 (2012).
212. Lau, W. & Johnson, C. in Proceedings of the Second BioCreative Challenge Evaluation Workshop 165-168Madrid, Spain; 2007).
213. Grover, C., Haddow, B., Klein, E., Matthews, M. & Neilsen, L. in Proceedings of the Second BioCreative Challenge Evaluation Workshop 272-286Madrid, Spain; 2007).
214. Granchev, K., Crammer, K., Pereira, F., Mann, G. & Bellare, K. in Proceedings of the Second BioCreative Challenge Evaluation Workshop 119-124Madrid, Spain; 2007).
215. Gaulton, A. et al. ChEMBL: a large-scale bioactivity database for drug discovery. *Nucleic acids research* **40**, D1100-1107 (2012).
216. Solt, I. et al. Gene mention normalization in full texts using GNAT and LINNAEUS. *In Proceedings of the BioCreative III Workshop Bethesda, USA.* (2010).
217. Van Landeghem, S. et al. Large-scale event extraction from literature with multi-level gene normalization. *PLoS One* **8**, e55814 (2013).
218. Rebholz-Schuhmann, D., Oellrich, A. & Hoehndorf, R. Text-mining solutions for biomedical research: enabling integrative biology. *Nat Rev Genet* **13**, 829-839 (2012).
219. Hearst, M.A. in Proceedings of the 14th Conference in Computational Linguistics, Vol. 2 539-5451992).
220. Leach, S.M. et al. Biomedical discovery acceleration, with applications to craniofacial development. *PLoS computational biology* **5**, e1000215 (2009).
221. Bjorne, J. & Salakoski, T. in BioNLP Shared Task Workshop 16-252013).
222. Agarwal, S. & Yu, H. Biomedical negation scope detection with conditional random fields. *J Am Med Inform Assoc* **17**, 696-701 (2010).
223. Medlock, B. Exploring hedge identification in biomedical literature. *Journal of biomedical informatics* **41**, 636-654 (2008).
224. Vincze, V., Szarvas, G., Farkas, R., Mora, G. & Csirik, J. The BioScope corpus: biomedical texts annotated for uncertainty, negation and their scopes. *BMC bioinformatics* **9 Suppl 11**, S9 (2008).
225. Sanchez-Graillet, O. & Poesio, M. Negation of protein-protein interactions: analysis and extraction. *Bioinformatics* **23**, i424-432 (2007).
226. Cruz Díaz, N.P., Maña López, M.J., Vázquez, J.M. & Álvarez, V.P. A machine-learning approach to negation and speculation detection in clinical texts. *Journal of the American Society for Information Science and Technology* **63**, 1398-1410 (2012).
227. Gerner, M., Sarafraz, F., Bergman, C.M. & Nenadic, G. BioContext: an integrated text mining system for large-scale extraction and

contextualization of biomolecular events. *Bioinformatics* **28**, 2154-2161 (2012).
228. Cunningham, H., Tablan, V., Roberts, A. & Bontcheva, K. Getting more out of biomedical documents with GATE's full lifecycle open source text analytics. *PLoS computational biology* **9**, e1002854 (2013).
229. Kano, Y. et al. U-Compare: share and compare text mining tools with UIMA. *Bioinformatics* **25**, 1997-1998 (2009).
230. Cohen, K.B., Johnson, H.L., Verspoor, K., Roeder, C. & Hunter, L.E. The structural and content aspects of abstracts versus bodies of full text journal articles are different. *BMC bioinformatics* **11**, 492 (2010).
231. Van Noorden, R. Trouble at the text mine. *Nature* **483**, 134-135 (2012).
232. Van Noorden, R. Elsevier opens its papers to text-mining. *Nature* **506**, 17 (2014).
233. Hargreaves, I., Guibault, L. & Valcke, P. (European Commision, European Commision; 2014).
234. Van Auken, K., Jaffery, J., Chan, J., Muller, H.M. & Sternberg, P.W. Semi-automated curation of protein subcellular localization: a text mining-based approach to Gene Ontology (GO) Cellular Component curation. *BMC bioinformatics* **10**, 228 (2009).
235. Wiegers, T.C., Davis, A.P., Cohen, K.B., Hirschman, L. & Mattingly, C.J. Text mining and manual curation of chemical-gene-disease networks for the comparative toxicogenomics database (CTD). *BMC bioinformatics* **10**, 326 (2009).
236. Li, C., Jimeno-Yepes, A., Arregui, M., Kirsch, H. & Rebholz-Schuhmann, D. PCorral--interactive mining of protein interactions from MEDLINE. *Database (Oxford)* **2013**, bat030 (2013).
237. Clematide, S. & Rinaldi, F. Ranking relations between diseases, drugs and genes for a curation task. *J Biomed Semantics* **3 Suppl 3**, S5 (2012).
238. Rinaldi, F. et al. Using ODIN for a PharmGKB revalidation experiment. *Database (Oxford)* **2012**, bas021 (2012).
239. Park, J., Costanzo, M.C., Balakrishnan, R., Cherry, J.M. & Hong, E.L. CvManGO, a method for leveraging computational predictions to improve literature-based Gene Ontology annotations. *Database (Oxford)* **2012**, bas001 (2012).
240. Kemper, B. et al. PathText: a text mining integrator for biological pathway visualizations. *Bioinformatics* **26**, i374-381 (2010).
241. Dowell, K.G., McAndrews-Hill, M.S., Hill, D.P., Drabkin, H.J. & Blake, J.A. Integrating text mining into the MGI biocuration workflow. *Database (Oxford)* **2009**, bap019 (2009).
242. Euler, L. Solutio problematis ad geometriam situs pertinentis. *Comment. Academiae Sci. I. Petropolitanae* **8**, 128-140 (1736).
243. Easley, D. & Kleinberg, J. Networks, crowds, and markets : reasoning about a highly connected world. (Cambridge University Press, Cambridge; 2010).
244. Yu, H., Kim, P.M., Sprecher, E., Trifonov, V. & Gerstein, M. The importance of bottlenecks in protein networks: correlation with gene essentiality and expression dynamics. *PLoS computational biology* **3**, e59 (2007).

245. Shannon, P. et al. Cytoscape: a software environment for integrated models of biomolecular interaction networks. *Genome Res* **13**, 2498-2504 (2003).
246. Csardi, G. & Nepusz, T. The igraph software package for complex network research. *InterJournal* **Complex Systems**, 1695 (2006).
247. Huang, H.T. et al. A network of epigenetic regulators guides developmental haematopoiesis in vivo. *Nat Cell Biol* **15**, 1516-1525 (2013).
248. Yildirim, M.A., Goh, K.I., Cusick, M.E., Barabasi, A.L. & Vidal, M. Drug-target network. *Nat Biotechnol* **25**, 1119-1126 (2007).
249. Blake, J.A. & Harris, M.A. The Gene Ontology (GO) project: structured vocabularies for molecular biology and their application to genome and expression analysis. *Curr Protoc Bioinformatics* **Chapter 7**, Unit 7 2 (2008).
250. Huang da, W., Sherman, B.T. & Lempicki, R.A. Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources. *Nat Protoc* **4**, 44-57 (2009).
251. (UNAIDS), J.U.N.P.o.H.A. Global Report: UNAIDS report on the global AIDS epidemic 2010. *WHO Library Cataloguing-in-Publication Data* (2010).
252. Dickerson, J.E., Pinney, J.W. & Robertson, D.L. The biological context of HIV-1 host interactions reveals subtle insights into a system hijack. *BMC systems biology* **4**, 80 (2010).
253. Bushman, F.D. et al. Host cell factors in HIV replication: meta-analysis of genome-wide studies. *PLoS Pathog* **5**, e1000437 (2009).
254. Brass, A.L. et al. Identification of host proteins required for HIV infection through a functional genomic screen. *Science* **319**, 921-926 (2008).
255. Krallinger, M., Erhardt, R.A. & Valencia, A. Text-mining approaches in molecular biology and biomedicine. *Drug discovery today* **10**, 439-445 (2005).
256. Zweigenbaum, P., Demner-Fushman, D., Yu, H. & Cohen, K.B. Frontiers of biomedical text mining: current progress. *Brief Bioinform* **8**, 358-375 (2007).
257. Zaremba, S. et al. Text-mining of PubMed abstracts by natural language processing to create a public knowledge base on molecular mechanisms of bacterial enteropathogens. *BMC bioinformatics* **10**, 177 (2009).
258. Bjorne, J., Ginter, F., Pyysalo, S., Tsujii, J. & Salakoski, T. Complex event extraction at PubMed scale. *Bioinformatics* **26**, i382-390 (2010).
259. Mani, I. et al. Protein name tagging guidelines: lessons learned. *Comp Funct Genomics* **6**, 72-76 (2005).
260. Miwa, M., Saetre, R., Kim, J.D. & Tsujii, J. Event extraction with complex event classification using rich features. *Journal of bioinformatics and computational biology* **8**, 131-146 (2010).
261. Jamieson, D.G., Robertson, D.L. & Nenadic, G. Task-specific Protein Tagging: an Experiment with BANNER on HIV-1/human interaction text. *LBM 2011 : Fourth International Symposium on Languages in Biology and Medicine* (2011).
262. Tanabe, L., Xie, N., Thom, L.H., Matten, W. & Wilbur, W.J. GENETAG: a tagged corpus for gene/protein named entity recognition. *BMC bioinformatics* **6 Suppl 1**, S3 (2005).
263. NCBI Entrez Gene. *http://www.ncbi.nlm.nih.gov/gene* (2011).

264. Kim, J.D., Ohta, T., Tateisi, Y. & Tsujii, J. GENIA corpus--semantically annotated corpus for bio-textmining. *Bioinformatics* **19 Suppl 1**, i180-182 (2003).
265. Björne, J. et al. Extracting complex biological events with rich graph-based feature sets. *In Proceedings of the Workshop on BioNLP Shared Task Boulder, Colorado*, 10-18 (2009).
266. Buonocore, L., Turi, T.G., Crise, B. & Rose, J.K. Stimulation of heterologous protein degradation by the Vpu protein of HIV-1 requires the transmembrane and cytoplasmic domains of CD4. *Virology* **204**, 482-486 (1994).
267. Bour, S., Schubert, U. & Strebel, K. The human immunodeficiency virus type 1 Vpu protein specifically binds to the cytoplasmic domain of CD4: implications for the mechanism of degradation. *Journal of virology* **69**, 1510-1520 (1995).
268. Margottin, F. et al. Interaction between the cytoplasmic domains of HIV-1 Vpu and CD4: role of Vpu residues involved in CD4 interaction and in vitro CD4 degradation. *Virology* **223**, 381-386 (1996).
269. Ispolatov, I., Yuryev, A., Mazo, I. & Maslov, S. Binding properties and evolution of homodimers in protein-protein interaction networks. *Nucleic acids research* **33**, 3629-3635 (2005).
270. Bannwarth, S. & Gatignol, A. HIV-1 TAR RNA: the target of molecular interactions between the virus and its host. *Current HIV research* **3**, 61-71 (2005).
271. Li, X., Josef, J. & Marasco, W.A. Hiv-1 Tat can substantially enhance the capacity of NIK to induce IkappaB degradation. *Biochemical and biophysical research communications* **286**, 587-594 (2001).
272. Blake, C. Beyond genes, proteins, and abstracts: Identifying scientific claims from full-text biomedical articles. *Journal of biomedical informatics* **43**, 173-189 (2010).
273. Hakenberg, J. et al. The GNAT library for local and remote gene mention normalization. *Bioinformatics* **27**, 2769-2771 (2011).
274. Huang, M., Liu, J. & Zhu, X. GeneTUKit: a software for document-level gene normalization. *Bioinformatics* **27**, 1032-1033 (2011).
275. Hunter, L. & Cohen, K.B. Biomedical language processing: what's beyond PubMed? *Mol Cell* **21**, 589-594 (2006).
276. Barshir, R. et al. The TissueNet database of human tissue protein-protein interactions. *Nucleic acids research* **41**, D841-844 (2013).
277. Fernandez-Suarez, X.M. & Galperin, M.Y. The 2013 Nucleic Acids Research Database Issue and the online molecular biology database collection. *Nucleic acids research* **41**, D1-7 (2013).
278. Gao, F., Luo, H. & Zhang, C.T. DoriC 5.0: an updated database of oriC regions in both bacterial and archaeal genomes. *Nucleic acids research* **41**, D90-93 (2013).
279. Frenkel-Morgenstern, M. et al. ChiTaRS: a database of human, mouse and fruit fly chimeric transcripts and RNA-sequencing data. *Nucleic acids research* **41**, D142-151 (2013).

280. Pachkov, M., Balwierz, P.J., Arnold, P., Ozonov, E. & van Nimwegen, E. SwissRegulon, a database of genome-wide annotations of regulatory sites: recent updates. *Nucleic acids research* **41**, D214-220 (2013).
281. Vita, R. et al. The immune epitope database 2.0. *Nucleic acids research* **38**, D854-862 (2010).
282. Davis, A.P. et al. The Comparative Toxicogenomics Database: update 2013. *Nucleic acids research* **41**, D1104-1114 (2013).
283. Eppig, J.T., Blake, J.A., Bult, C.J., Kadin, J.A. & Richardson, J.E. The Mouse Genome Database (MGD): comprehensive resource for genetics and genomics of the laboratory mouse. *Nucleic acids research* **40**, D881-886 (2012).
284. McDonagh, E.M., Whirl-Carrillo, M., Garten, Y., Altman, R.B. & Klein, T.E. From pharmacogenomic knowledge acquisition to clinical applications: the PharmGKB as a clinical pharmacogenomic biomarker resource. *Biomark Med* **5**, 795-806 (2011).
285. Arighi, C.N. et al. An overview of the BioCreative 2012 Workshop Track III: interactive text mining task. *Database (Oxford)* **2013**, bas056 (2013).
286. Jamieson, D.G., Gerner, M., Sarafraz, F., Nenadic, G. & Robertson, D.L. Towards semi-automated curation: using text mining to recreate the HIV-1, human protein interaction database. *Database (Oxford)* **2012**, bas023 (2012).
287. Harifi, G. et al. Prevalence of chronic pain with neuropathic characteristics in the moroccan general population: a national survey. *Pain Med* **14**, 287-292 (2013).
288. Dworkin, R.H. et al. Pharmacologic management of neuropathic pain: evidence-based recommendations. *Pain* **132**, 237-251 (2007).
289. Leavitt, S.B. & Kersta-Wilson, S., Vol. 2012http://pain-topics.org/glossary/; 2010).
290. Calvo, M., Dawes, J.M. & Bennett, D.L. The role of the immune system in the generation of neuropathic pain. *Lancet Neurol* **11**, 629-642 (2012).
291. Woolf, C.J. & Mannion, R.J. Neuropathic pain: aetiology, symptoms, mechanisms, and management. *Lancet* **353**, 1959-1964 (1999).
292. Kim, Y., Hurdle, J. & Meystre, S.M. Using UMLS lexical resources to disambiguate abbreviations in clinical text. *AMIA ... Annual Symposium proceedings / AMIA Symposium. AMIA Symposium* **2011**, 715-722 (2011).
293. Liu, H., Aronson, A.R. & Friedman, C. A study of abbreviations in MEDLINE abstracts. *Proc AMIA Symp*, 464-468 (2002).
294. Ohta, T., Tateisi, Y. & Kim, J. The GENIA Corpus: an Annotated Research Abstract Corpus in Molecular Biology Domain. *Proceedings of the second international conference on Human Language Technology Research* **Morgan Kaufmann Publishers Inc.** (2002).
295. Hakenberg, J., Plake, C., Leaman, R., Schroeder, M. & Gonzalez, G. Inter-species normalization of gene mentions with GNAT. *Bioinformatics* **24**, i126-132 (2008).
296. Maglott, D., Ostell, J., Pruitt, K.D. & Tatusova, T. Entrez Gene: gene-centered information at NCBI. *Nucleic acids research* **35**, D26-31 (2007).

297. Sayers, E.W. et al. Database resources of the National Center for Biotechnology Information. *Nucleic acids research* **40**, D13-25 (2012).
298. Lacroix-Fralish, M.L., Ledoux, J.B. & Mogil, J.S. The Pain Genes Database: An interactive web browser of pain-related transgenic knockout studies. *Pain* **131**, 3 e1-4 (2007).
299. Gerner, M., Nenadic, G. & Bergman, C.M. An exploration of mining gene expression mentions and their anatomical locations from biomedical text. *Proceedings of the BioNLP Workshop* **Uppsala, Sweden** (2010).
300. Sarafraz, F. & Nenadic, G. Using SVMs with the Command Relation Features to Identify Negated Events in Biomedical Literature. *The Workshop on Negation and Speculation in Natural Language Processing* **Uppsala**, Sweden (2010).
301. , Vol. 2012http://www.geneontology.org/GO.slims.shtml; 2012).
302. Davis, M.J., Sehgal, M.S. & Ragan, M.A. Automatic, context-specific generation of Gene Ontology slims. *BMC bioinformatics* **11**, 498 (2010).
303. , Vol. 2012 (National Center for Biotechnology Information, http://www.ncbi.nlm.nih.gov/Ftp/; 2012).
304. Wang, J. et al. Factorbook.org: a Wiki-based database for transcription factor-binding data generated by the ENCODE consortium. *Nucleic acids research* **41**, D171-176 (2013).
305. MediaWiki, Vol. 2012 (MediaWiki, http://www.mediawiki.org/wiki/API:Main_page; 2012).
306. Hersh, W. Evaluation of biomedical text-mining systems: lessons learned from information retrieval. *Brief Bioinform* **6**, 344-356 (2005).
307. Cornet, R. & de Keizer, N. Forty years of SNOMED: a literature review. *BMC Med Inform Decis Mak* **8 Suppl 1**, S2 (2008).
308. Razick, S., Magklaras, G. & Donaldson, I.M. iRefIndex: a consolidated protein interaction database with provenance. *BMC bioinformatics* **9**, 405 (2008).
309. Tarcea, V.G. et al. Michigan molecular interactions r2: from interacting proteins to pathways. *Nucleic acids research* **37**, D642-646 (2009).
310. Bader, G.D., Betel, D. & Hogue, C.W. BIND: the Biomolecular Interaction Network Database. *Nucleic acids research* **31**, 248-250 (2003).
311. Kerrien, S. et al. The IntAct molecular interaction database in 2012. *Nucleic acids research* **40** (2011).
312. Simpson, A., Graham, M.E. & Williams, J. Chronic plantar ulcer secondary to congenital indifference to pain. *J Wound Care* **20**, 540, 542 (2011).
313. Rotthier, A., Baets, J., Timmerman, V. & Janssens, K. Mechanisms of disease in hereditary sensory and autonomic neuropathies. *Nat Rev Neurol* **8**, 73-85 (2012).
314. Gureje, O., Von Korff, M., Simon, G.E. & Gater, R. Persistent pain and well-being: a World Health Organization Study in Primary Care. *JAMA* **280**, 147-151 (1998).
315. Kroenke, K., Krebs, E.E. & Bair, M.J. Pharmacotherapy of chronic pain: a synthesis of recommendations from systematic reviews. *Gen Hosp Psychiatry* **31**, 206-219 (2009).
316. Finnerup, N.B., Sindrup, S.H. & Jensen, T.S. The evidence for pharmacological treatment of neuropathic pain. *Pain* **150**, 573-581 (2010).

317. Turk, D.C., Wilson, H.D. & Cahana, A. Treatment of chronic non-cancer pain. *Lancet* **377**, 2226-2235 (2011).
318. Jamieson, D.G., Roberts, P.M., Robertson, D.L., Sidders, B. & Nenadic, G. Cataloging the biomedical world of pain through semi-automated curation of molecular interactions. *Database (Oxford)* **2013**, bat033 (2013).
319. Son, S.J. et al. Activation of transcription factor c-jun in dorsal root ganglia induces VIP and NPY upregulation and contributes to the pathogenesis of neuropathic pain. *Exp Neurol* **204**, 467-472 (2007).
320. Poirel, C.L., Owens, C.C., 3rd & Murali, T.M. Network-based functional enrichment. *BMC bioinformatics* **12 Suppl 13**, S14 (2011).
321. Wuchty, S. Controllability in protein interaction networks. *Proceedings of the National Academy of Sciences of the United States of America* **111**, 7156-7160 (2014).
322. Hofmann, H.A., De Vry, J., Siegling, A., Spreyer, P. & Denzer, D. Pharmacological sensitivity and gene expression analysis of the tibial nerve injury model of neuropathic pain. *Eur J Pharmacol* **470**, 17-25 (2003).
323. Barabasi, A.L. & Oltvai, Z.N. Network biology: understanding the cell's functional organization. *Nat Rev Genet* **5**, 101-113 (2004).
324. Pfeiffer, T. & Hoffmann, R. Temporal patterns of genes in scientific publications. *Proceedings of the National Academy of Sciences of the United States of America* **104**, 12052-12056 (2007).
325. Crist, R.C. & Berrettini, W.H. Pharmacogenetics of OPRM1. *Pharmacol Biochem Behav* (2013).
326. Miyasaka, N. [Etanercept for therapy of rheumatoid arthritis]. *Nihon Rinsho* **63 Suppl 1**, 521-525 (2005).
327. Cattaneo, A. Tanezumab, a recombinant humanized mAb against nerve growth factor for the treatment of acute and chronic pain. *Curr Opin Mol Ther* **12**, 94-106 (2010).
328. Coull, J.A. et al. BDNF from microglia causes the shift in neuronal anion gradient underlying neuropathic pain. *Nature* **438**, 1017-1021 (2005).
329. Siniscalco, D., Giordano, C., Rossi, F., Maione, S. & de Novellis, V. Role of neurotrophins in neuropathic pain. *Curr Neuropharmacol* **9**, 523-529 (2011).
330. Jeong, H., Mason, S.P., Barabasi, A.L. & Oltvai, Z.N. Lethality and centrality in protein networks. *Nature* **411**, 41-42 (2001).
331. Holliday, K.L. et al. Do genetic predictors of pain sensitivity associate with persistent widespread pain? *Mol Pain* **5**, 56 (2009).
332. Gunthorpe, M.J. & Chizh, B.A. Clinical development of TRPV1 antagonists: targeting a pivotal point in the pain pathway. *Drug Discov Today* **14**, 56-67 (2009).
333. Pomonis, J.D. et al. N-(4-Tertiarybutylphenyl)-4-(3-cholorphyridin-2-yl)tetrahydropyrazine -1(2H)-carbox-amide (BCTC), a novel, orally effective vanilloid receptor 1 antagonist with analgesic properties: II. in vivo characterization in rat models of inflammatory and neuropathic pain. *J Pharmacol Exp Ther* **306**, 387-393 (2003).
334. Hefti, F.F. et al. Novel class of pain drugs based on antagonism of NGF. *Trends Pharmacol Sci* **27**, 85-91 (2006).

335. Holmes, D. Anti-NGF painkillers back on track? *Nat Rev Drug Discov* **11**, 337-338 (2012).
336. Pezet, S. & McMahon, S.B. Neurotrophins: mediators and modulators of pain. *Annu Rev Neurosci* **29**, 507-538 (2006).
337. Mika, J., Obara, I. & Przewlocka, B. The role of nociceptin and dynorphin in chronic pain: implications of neuro-glial interaction. *Neuropeptides* **45**, 247-261 (2011).
338. Leal-Cerro, A. et al. The growth hormone (GH)-releasing hormone-GH-insulin-like growth factor-1 axis in patients with fibromyalgia syndrome. *J Clin Endocrinol Metab* **84**, 3378-3381 (1999).
339. Lim, G., Wang, S., Zhang, Y., Tian, Y. & Mao, J. Spinal leptin contributes to the pathogenesis of neuropathic pain in rodents. *J Clin Invest* **119**, 295-304 (2009).
340. Li, X. et al. Intrathecal leptin inhibits expression of the P2X2/3 receptors and alleviates neuropathic pain induced by chronic constriction sciatic nerve injury. *Mol Pain* **9**, 65 (2013).
341. Basbaum, A.I., Bautista, D.M., Scherrer, G. & Julius, D. Cellular and molecular mechanisms of pain. *Cell* **139**, 267-284 (2009).
342. Cibert-Goton, V. et al. Involvement of EphB1 receptors signalling in models of inflammatory and neuropathic pain. *PLoS One* **8**, e53673 (2013).
343. Liu, M. & Wood, J.N. The roles of sodium channels in nociception: implications for mechanisms of neuropathic pain. *Pain Med* **12 Suppl 3**, S93-99 (2011).
344. Ferrini, F. et al. Morphine hyperalgesia gated through microglia-mediated disruption of neuronal Cl(-) homeostasis. *Nat Neurosci* **16**, 183-192 (2013).
345. D'Haese, J.G., Friess, H. & Ceyhan, G.O. Therapeutic potential of the chemokine-receptor duo fractalkine/CX3CR1: an update. *Expert Opin Ther Targets* **16**, 613-618 (2012).
346. Moss, A. et al. Origins, actions and dynamic expression patterns of the neuropeptide VGF in rat peripheral and central sensory neurones following peripheral nerve injury. *Mol Pain* **4**, 62 (2008).
347. Fischer, M.J. et al. The general anesthetic propofol excites nociceptors by activating TRPV1 and TRPA1 rather than GABAA receptors. *The Journal of biological chemistry* **285**, 34781-34792 (2010).
348. Knotkova, H., Pappagallo, M. & Szallasi, A. Capsaicin (TRPV1 Agonist) therapy for pain relief: farewell or revival? *Clin J Pain* **24**, 142-154 (2008).
349. LaCroix-Fralish, M.L., Austin, J.S., Zheng, F.Y., Levitin, D.J. & Mogil, J.S. Patterns of pain: meta-analysis of microarray studies of pain. *Pain* **152**, 1888-1898 (2011).
350. Perkins, J.R. et al. PainNetworks: A web-based resource for the visualisation of pain-related genes in the context of their network associations. *Pain* (2013).
351. Jacunski, A. & Tatonetti, N.P. Connecting the dots: applications of network medicine in pharmacology and disease. *Clin Pharmacol Ther* **94**, 659-669 (2013).

352. Lozano, R. et al. Global and regional mortality from 235 causes of death for 20 age groups in 1990 and 2010: a systematic analysis for the Global Burden of Disease Study 2010. *Lancet* **380**, 2095-2128 (2012).
353. Murray, C.J. et al. Disability-adjusted life years (DALYs) for 291 diseases and injuries in 21 regions, 1990-2010: a systematic analysis for the Global Burden of Disease Study 2010. *Lancet* **380**, 2197-2223 (2012).
354. Nkoghe, D., Kone, M.L., Yada, A. & Leroy, E. A limited outbreak of Ebola haemorrhagic fever in Etoumbi, Republic of Congo, 2005. *Trans R Soc Trop Med Hyg* **105**, 466-472 (2011).
355. Arts, E.J. & Hazuda, D.J. HIV-1 antiretroviral drug therapy. *Cold Spring Harb Perspect Med* **2**, a007161 (2012).
356. Chatr-aryamontri, A. et al. VirusMINT: a viral protein interaction database. *Nucleic acids research* **37**, D669-673 (2009).
357. Kumar, R. & Nanduri, B. HPIDB--a unified resource for host-pathogen interactions. *BMC bioinformatics* **11 Suppl 6**, S16 (2010).
358. Wattam, A.R. et al. PATRIC, the bacterial bioinformatics database and analysis resource. *Nucleic acids research* **42**, D581-591 (2014).
359. Winnenburg, R. et al. PHI-base update: additions to the pathogen host interaction database. *Nucleic acids research* **36**, D572-576 (2008).
360. Smith, S.B., Dampier, W., Tozeren, A., Brown, J.R. & Magid-Slav, M. Identification of common biological pathways and drug targets across multiple respiratory viruses based on human host gene expression analysis. *PLoS One* **7**, e33174 (2012).
361. Pichlmair, A. et al. Viral immune modulators perturb the human molecular network by common and unique strategies. *Nature* **487**, 486-490 (2012).
362. Thieu, T., Joshi, S., Warren, S. & Korkin, D. Literature mining of host-pathogen interactions: comparing feature-based supervised learning and language-based approaches. *Bioinformatics* **28**, 867-875 (2012).
363. Zhou, H., Jin, J. & Wong, L. Progress in computational studies of host-pathogen interactions. *Journal of bioinformatics and computational biology* **11**, 1230001 (2013).
364. Eng, R.H., Drehmel, R., Smith, S.M. & Goldstein, E.J. Saccharomyces cerevisiae infections in man. *Sabouraudia* **22**, 403-407 (1984).
365. Miceli, M.H., Diaz, J.A. & Lee, S.A. Emerging opportunistic yeast infections. *Lancet Infect Dis* **11**, 142-151 (2011).
366. Winzeler, E.A. et al. Functional characterization of the S. cerevisiae genome by gene deletion and parallel analysis. *Science* **285**, 901-906 (1999).
367. Van Landeghem, S. EVEX: A PubMed-Scale Resource for Homology-Based Generalization of Text Mining Predictions. *Proceedings of BioNLP Workshop* **Association for Computational Linguistics** (2011).
368. Jamieson, D.G. et al. The pain interactome: Connecting pain-specific protein interactions. *Pain* (2014).
369. Tan, S.-L. Hepatitis C viruses : genomes and molecular biology. (Horizon Bioscience, Wymondham, Norfolk, U.K.; 2006).

370. Trifonov, V., Racaniello, V. & Rabadan, R. The Contribution of the PB1-F2 Protein to the Fitness of Influenza A Viruses and its Recent Evolution in the 2009 Influenza A (H1N1) Pandemic Virus. *PLoS Curr* **1**, RRN1006 (2009).
371. Driscoll, T., Dyer, M.D., Murali, T.M. & Sobral, B.W. PIG--the pathogen interaction gateway. *Nucleic acids research* **37**, D647-650 (2009).
372. He, M., Wang, Y. & Li, W. PPI finder: a mining tool for human protein-protein interactions. *PLoS One* **4**, e4554 (2009).
373. Lawrence, K.M. et al. Urocortin – a peptide for all ills? *int j bioch & cell biol (in review)* (2014).
374. Verspoor, K.M. et al. A corpus of full-text journal articles is a robust evaluation tool for revealing differences in performance of biomedical natural language processing tools. *BMC bioinformatics* **13**, 207 (2012).