

Universidad autónoma de Madrid
Facultad de Ciencias
Departamento de Biología Molecular

Data mining from Scientific Literature

TESIS DOCTORAL

Parantu K Shah

European Molecular Biology Laboratory, Heidelberg, Germany
Max-Delbrück Centrum für Molekulare Medizin, Berlin, Germany
Universidad Autónoma de Madrid, Madrid, Spain

Universidad Autónoma de Madrid
Facultad de Ciencias
Departamento de Biología Molecular

Data mining from Scientific Literature

Memoria presentada para optar al grado de Doctor en Biología
Molecular por:

Parantu K Shah

Director: Dr. Peer Bork

Tutor: Dr. Alfonso Valencia Herrera

Acknowledgements

I spent four productive and fun-filled years at EMBL growing academically and personally. I would like to thank every body who has been part of this experience.

- First and foremost, Dr. Peer Bork for giving me the opportunity to work under his guidance. Peer provided scientific advice, ideas and resources so that the work presented here could be completed. But above all, he provided me with ample freedom to work in still developing field of natural language processing (NLP) of biomedical texts and encouraging me to *push the limits* of the current knowledge.
- Dr. Miguel Andrade and Dr. Carolina Perez-Iratxeta in helping me getting started in NLP in biology.
- Dr. Nigel Collier at the NII in Japan for teaching me basics of NLP and for providing me with an opportunity to work under his guidance and interact with his group composed of theoretical and computational linguists.
- Dr. Rob Russell for his useful comments as a member of my advisory committee and giving me opportunity to work under his guidance in the area of Structural Bioinformatics. Other members of my advisory committee, Dr. Christos Ouzounis and Prof. Alfonso Valencia.
- Dr. Lars Juhl Jensen for enlightening me towards the power of statistics and data analysis, and helping me deal with scientific writing and reviewers' reports, the other side of scientific world.
- Dr. Francesca Ciccarelli and Dr. Tobias Doerks for being wonderful colleagues to share the office with. Dr. Monica Campillos for helping me out with the submission procedure. Also, present and past members of Bork group and members of EMBL Biocomputing program for making EMBL a place full of fun and lots of science.
- Andrés Gaytan and Dr. Monica Campillos for helping me with the Spanish version of the summary of thesis work.
- My friends Carolene Lemerle, Maria-Vittoria Verga-falzacappa, Magdalena Kraus, Gabriela Zuliani, Sumati Mattoo, Sonal Patel, Francesca Diella, Peter Winn, Marina Chekulaeva, Barbara Diventura, Fabiana Perocchi, Sandra Esteras, Lodovica Borghese, Rune Linding, Tuangthong

Wattarujeekrit, Tony Mullen, Luree Schneider, Anirban Bhaduri, Elham Andaroodi, Jumpot Phuritakul, Pedro Beltrao, and others for touching my life in a positive way and enriching it immensely (names are in no particular order, some may be invariably missing).

- My parents and other members of my family for their unconditional love and support. For teaching me to fight for all things I deserve, believe in myself and keep a balanced head.
- My brother for teaching me to work hard and concentrate my energies to the positive direction in life. This thesis is one of the examples of things that can be achieved with a positive outlook.

Index

List of Publications	I
Summary	II
Summary in Spanish (Resumen en Español)	III
Abbreviations	XIV
List of figures	XV
List of Tables	XVII
I. Introduction	1
1.1. - Prologue-Introduction	1
1.2. - Automated handling of text	2
1.2.1. - Logical view of documents	3
1.3. - An overview of IE methods	4
1.3.1. - Named entity extraction	4
1.3.2. - Relationship extraction	4
1.3.3. - Hypothesis generation	6
1.3.4. - Integration frameworks	6
1.3.5. - Ontologies in biology	7
1.4. - Event extraction	9
1.4.1. - Events in molecular biology	9
1.4.2. - Template based extraction of relationships and events	9
1.4.3. - Spray alterations and the problem of syntactic patterns in IE	9
1.4.4. - Need for Semantic Relationships in molecular event extraction	10
1.5. - Predicate Argument Structures	12
1.5.1. - Resources for PAS	13
1.5.2. - Introduction to PropBank	14
1.6. - Classification using inductive machine learning	15
1.7. - Generation of Alternative transcripts	16
1.7.1. - Alternative promoters	18
1.7.2. - Alternative Splicing	19
1.7.3. - Alternative polyadenylation	19
II. Objectives	20

III. Methods	21
3.1- Analysis of full-text articles for comparison of information in different sections	21
3.1.1 Text Corpus for the analysis of full-text articles	21
3.1.2 Derivation of Associations between the words of a section	21
3.1.3 Selection of Keywords	21
3.1.4 Classification of Words in Subjects	22
3.2. -Definitions of precision, recall and F-measure	22
3.3. – Predicate argument structure analysis for written texts in molecular biology	23
3.3.1. - Selection of Verbs for PAS analysis	23
3.3.2. - Selection of Example Sentences for PAS analysis	23
3.3.3. - Use of parsers reduces manual work	23
3.4. - Semi-automated generation of the database of transcript diversity	24
3.4.1. - Description of transcript diversity in abstracts	24
3.4.2. - Definition of Sentence classification task for inductive learning	26
3.4.3. - Training corpus and pre-processing for sentence classification	26
3.4.4. - Set for benchmarking of recall for SVM classifier	28
3.4.5. - Mapping of sentence classification results to Sequence databases	28
3.4.6. - Quantifying the gain in gene annotation	28
3.4.7. - Merging multiple syntactic patterns to define semantic categories	29
3.4.8. - Rules for extracting semantic categories	30
3.4.9. - Benchmarking of the tagging performance	30
3.4.10. - Associating TD-generating mechanisms with organ systems	30
IV. Results	31
4.1. - Analysis of full text articles with keywords	31
4.1.1. - Performance at keyword detection	31
4.1.2. - Keyword Selection by section	32
4.1.3. - Sections display heterogeneous information	33
4.1.4. - Qualitative analysis of subjects per section	35
4.1.5. - Analysis of distribution of gene names	37
4.2. – PASBio: Towards event extraction from biomedical texts	38
4.2.1. - Mapping from surface structures to PAS	39
4.2.2. - Defining predicate-argument structures for molecular biology	40
4.2.3. - Guidelines for defining PAS	41
4.2.4. - Examples of defined PAS	43

4.2.5. - Complexities in Biology Texts	51
4.3. - Extraction of information about transcript diversity from MEDLINE	52
4.3.1. - Overall strategy and generation of the database	52
4.3.2. – Experiments on sentence classification	55
4.3.3. - Analysis of extracted sentences	62
4.3.4. - Semantic role labeling	62
4.4. – Data mining of LSAT	63
4.4.1. - Proposing new annotations in sequence databases	63
4.4.2. - Quantification of the different mechanisms that lead to transcript diversity	63
4.4.3. - Identifying tissue specific differences in the extent of alternative splicing	65
4.4.4. - Assigning function to the transcripts generated by computational analysis	66
V. Discussion	68
5.1. - Analysis of full-text articles for IE	69
5.1.1. - Choice of the data-set	69
5.1.2. - The distribution of information is heterogeneous	69
5.1.3. - Introduction and Discussion are also information rich	70
5.1.4. - Context matters	70
5.1.5. - Related work on analysis of full-text articles	70
5.2. Exploitation of sentence semantics for accurate event extraction	71
5.2.1. - Specialization of domains affects various text processing tools	71
5.2.2. - PASBio: a database of predicate argument structures for molecular biology.	72
5.2.3. – Utilization of PASBio	72
5.2.4. - Related work on Information Extraction from biomedical texts	74
5.3. – Generating event-specific database with a two-step procedure	74
5.3.1.- Description of LSAT	74
5.3.2. - Retrieving event describing sentences using text categorization methods	76
5.3.3. - Rule-based tagging for IE would help database curation	77
5.3.4. - Rule based versus semantic role labeling using machine learning	78
5.3.5. - Related work on relationship/event extraction	78
5.4. - Analysis and integration of text-mining data to present knowledge	79
5.4.1. – Automated MeSH term assignments to Abstracts	79
5.4.2. - Function annotation using text-mining	79
5.4.3. - Transcript diversity generating mechanisms, synergy and preference	79
VI. Conclusions	81

VII. Supplementary material	83
Appendix A	83
Appendix B	85
Appendix C	91
VIII. References	92