Universidad autónoma de Madrid

Facultad de Ciencias

Departmento 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

Facutad de Ciencias

Departmento 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 everybody 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 Phuritatkul, 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