UNIVERSIDADE DE LISBOA
FACULDADE DE CIÊNCIAS
DEPARTAMENTO DE INFORMÁTICA

Thesis' Title

Your Name

DISSERTAÇÃO
MESTRADO EM BIOINFORMÁTICA E BIOLOGIA COMPUTACIONAL
ESPECIALIDADE EM BIOINFORMÁTICA

Dissertação orientada por:
Prof. Doutor Your Supervisor
Dr.

YYYY

Resumo

Palavras Chave: palavras chave

Abstract

Keywords: keywords

Resumo Alargado

Desde os anos 60 que se tem observado uma explosão de informação biomédica proveniente sobretudo dos esforços de sequenciação dos genomas de diversos organismos. Neste contexto promoveu-se o desenvolvimento e o aperfeiçoamento de técnicas que permitiram o avanço e o estabelecimento de disciplinas da pós-genómica, como a proteómica e metabolómica, o que gerou ainda mais informação. Assim surgiu um grande interesse pela Bioinformática como a disciplina que procura armazenar, gerir a analisar todos estes dados de interesse biológico.

Acknowledgements

Contents

List of Figures

List of Tables

Chapter 1

Chapter name

1.1 Motivation

In recent years, there has been an exponential increase in the research of biological domain, much due to the extensive use of high-throughput technologies such as yeast-two hybrid-based methods, DNA expression arrays and mass spectrometry. These methods have been producing large and heterogeneous collections of data, including proteomic and genomic sequence data, expression profiles, and protein structure coordinates that represent an important fraction of existing biological information (**?**). High-throughput technologies generate massive amounts of data that in turn require efficient information retrieval before any analysis is attempted. Since the 19th century, natural language is the main vehicle through which humans transmit and exchange the facts discovered in biological research, by means of scientific publications, patents, or reports (**?**). The most comprehensive repository for such articles is the MEDLINE database (**?**). MEDLINE is a bibliographic database of life sciences and biomedical information that currently contains over 19 million records that can be accessed through PubMed. Figure 1.1 shows the exploding number of articles available from MEDLINE since its creation in 1965, showing the exponential growth of published biomedical documents.
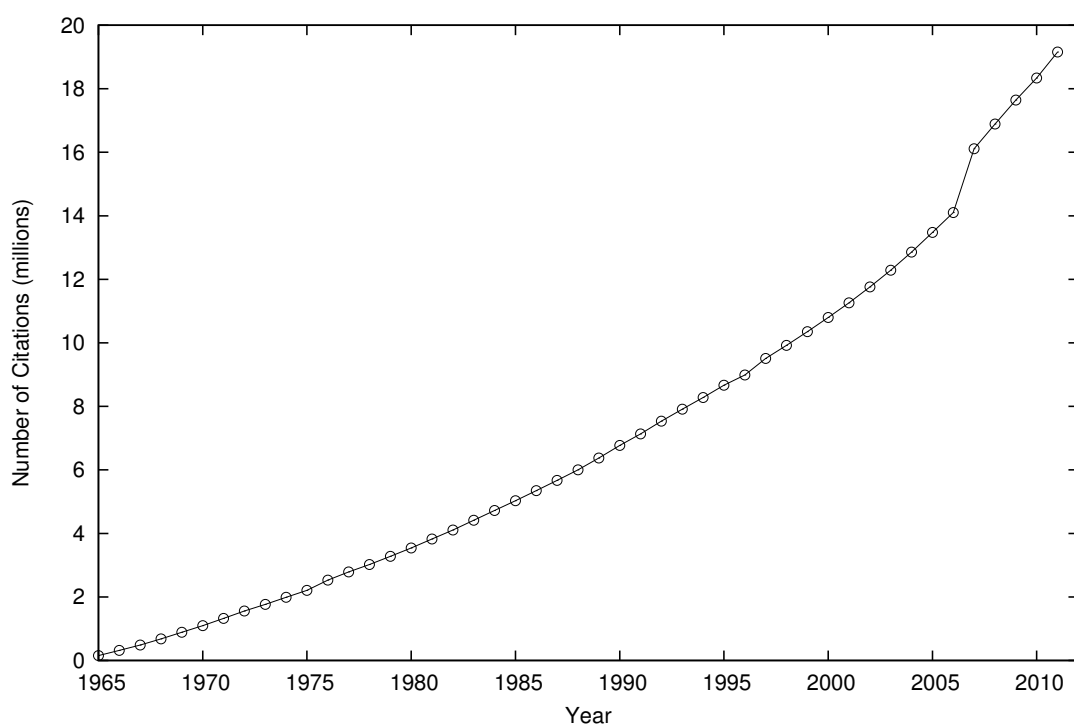
Figure 1.1: Number of citations present in MEDLINE since its beginning in 1950. Data from official statistics available at `www.nlm.nih.gov/bsd/index_stats_comp.html`.

1.2 Objectives

Hypothesis: The results achieved by current chemical entity identification systems can be improved by exploiting a Machine Learning approach and Semantic Similarity techniques using dictionaries as domain knowledge.

1.3 Contributions

Thus, the following specific contributions can be enumerated as follows:

Contribuition1:

1.4 Overview

The overview of this document is as follows.

1.5 Methods

1.6 Results

1.7 Discussion

1.8 Conclusions

References