Automatically Extracting Structured Information from Biomedical Text

Raymond J. Mooney

Department of Computer Science

Razvan Bunescu, Ruifang Ge, Rohit J. Kate, Yuk Wah Wong

Department of Computer Science

Edward M. Marcotte, Arun Ramani

Institute for Cellular and Molecular Biology

University of Texas at Austin
Extracting Information from Scientific Literature

• A great deal of scientific data is spread across a vast array of natural-language documents, rather than collected in structured computer databases.

• Information Extraction (IE) is a subarea of Natural Language Processing (NLP) that focuses on extracting structured data from text documents.

• IE systems can identify, extract, and collect scientific data spread across the primary literature and make them accessible to automated querying, analysis, and data mining.

• One type of “text mining”
Information Extraction

• Named Entity Extraction (NER)
  – Locate names of entities like people, places and organizations in text.
  – In biomedicine, named entities might include organisms, genes, proteins, diseases, drugs, etc.

• Relation Extraction
  – Locate stated relationships between previously identified entities, e.g. person works for organization.
  – In biomedicine, relations might be: protein interacts with protein, or drug treats disease.

• Both typically developed by training on a corpus of human annotated documents.
Biological Motivation

• Human Genome Project has produced huge amounts of genetic data.

• Next step is analyzing and interpreting this data.
PHASE: INTERPRETATION

I THINK I FOUND A CORNER PIECE.
Proteomics 101

• Genes code for proteins.
• Proteins are the basic components of biological machinery.
• Proteins accomplish their functions by interacting with other proteins.
• Knowledge of protein interactions is fundamental to understanding gene function.
• Chains of interactions compose large, complex gene networks.
Sample Gene Network
Knowledge in Biomedical Literature

• An ever increasing wealth of biological information is present in millions of published articles but retrieving it in structured form is difficult.
• Much of this literature is available through the NIH - NLM’s Medline (PubMed) repository.
• 11 million abstracts in electronic form are available through Medline.
• Excellent source of information on protein interactions.
Sample Medline Abstract

TI - Two potentially oncogenic cyclins, cyclin A and cyclin D1, share common properties of subunit configuration, tyrosine phosphorylation and physical association with the Rb protein

AB - Originally identified as a ‘mitotic cyclin’, cyclin A exhibits properties of growth factor sensitivity, susceptibility to viral subversion and association with a tumor-suppressor protein, properties which are indicative of an S-phase-promoting factor (SPF) as well as a candidate proto-oncogene.

Other recent studies have identified human cyclin D1 (PRAD1) as a putative G1 cyclin and candidate proto-oncogene.

However, the specific enzymatic activities and, hence, the precise biochemical mechanisms through which cyclins function to govern cell cycle progression remain unresolved.

In the present study we have investigated the coordinate interactions between these two potentially oncogenic cyclins, cyclin-dependent protein kinase subunits (cdks) and the Rb tumor-suppressor protein.

The distribution of cyclin D isoforms was modulated by serum factors in primary fetal rat lung epithelial cells.

Moreover, cyclin D1 was found to be phosphorylated on tyrosine residues in vivo and, like cyclin A, was readily phosphorylated by pp60c-src in vitro.

In synchronized human osteosarcoma cells, cyclin D1 is induced in early G1 and becomes associated with p9Ckshs1, a Cdk-binding subunit.

Immunoprecipitation experiments with human osteosarcoma cells and Ewing’s sarcoma cells demonstrated that cyclin D1 is associated with both p34cdc2 and p33cdk2, and that cyclin D1 immune complexes exhibit appreciable histone H1 kinase activity.

Immobilized, recombinant cyclins A and D1 were found to associate with cellular proteins in complexes that contain the p105Rb protein.

This study identifies several common aspects of cyclin biochemistry, including tyrosine phosphorylation and the potential to interact directly or indirectly with the Rb protein, that may ultimately relate membrane-mediated signaling events to the regulation of gene expression.
Sample Medline Abstract

TI - Two potentially oncogenic cyclins, cyclin A and cyclin D1, share common properties of subunit configuration, tyrosine phosphorylation and physical association with the Rb protein

AB - Originally identified as a ‘mitotic cyclin’, cyclin A exhibits properties of growth factor sensitivity, susceptibility to viral subversion and association with a tumor-suppressor protein, properties which are indicative of an S-phase-promoting factor (SPF) as well as a candidate proto-oncogene.

Other recent studies have identified human cyclin D1 (PRAD1) as a putative G1 cyclin and candidate proto-oncogene. However, the specific enzymatic activities and, hence, the precise biochemical mechanisms through which cyclins function to govern cell cycle progression remain unresolved.

In the present study we have investigated the coordinate interactions between these two potentially oncogenic cyclins, cyclin-dependent protein kinase subunits (cdks) and the Rb tumor-suppressor protein.

The distribution of cyclin D isoforms was modulated by serum factors in primary fetal rat lung epithelial cells. Moreover, cyclin D1 was found to be phosphorylated on tyrosine residues in vivo and, like cyclin A, was readily phosphorylated by pp60c-src in vitro.

In synchronized human osteosarcoma cells, cyclin D1 is induced in early G1 and becomes associated with p9Ckshs1, a Cdk-binding subunit.

Immunoprecipitation experiments with human osteosarcoma cells and Ewing’s sarcoma cells demonstrated that cyclin D1 is associated with both p34cdc2 and p33cdk2, and that cyclin D1 immune complexes exhibit appreciable histone H1 kinase activity.

Immobilized, recombinant cyclins A and D1 were found to associate with cellular proteins in complexes that contain the p105Rb protein.

This study identifies several common aspects of cyclin biochemistry, including tyrosine phosphorylation and the potential to interact directly or indirectly with the Rb protein, that may ultimately relate membrane-mediated signaling events to the regulation of gene expression.
Two potentially oncogenic cyclins, **cyclin A** and **cyclin D1**, share common properties of subunit configuration, tyrosine phosphorylation and physical association with the **Rb** protein.

Originally identified as a ‘mitotic cyclin’, **cyclin A** exhibits properties of growth factor sensitivity, susceptibility to viral subversion and association with a tumor-suppressor protein, properties which are indicative of an **S-phase-promoting factor** (**SPF**) as well as a candidate proto-oncogene.

Other recent studies have identified human **cyclin D1** (**PRAD1**) as a putative G1 cyclin and candidate proto-oncogene.

However, the specific enzymatic activities and, hence, the precise biochemical mechanisms through which cyclins function to govern cell cycle progression remain unresolved.

In the present study we have investigated the coordinate interactions between these two potentially oncogenic cyclins, cyclin-dependent protein kinase subunits (**cdks**) and the **Rb** tumor-suppressor protein.

The distribution of **cyclin D** isoforms was modulated by serum factors in primary fetal rat lung epithelial cells.

Moreover, **cyclin D1** was found to be phosphorylated on tyrosine residues in vivo and, like **cyclin A**, was readily phosphorylated by **pp60c-src** in vitro.

In synchronized human osteosarcoma cells, **cyclin D1** is induced in early G1 and becomes associated with **p9Ckshs1**, a Cdk-binding subunit.

Immunoprecipitation experiments with human osteosarcoma cells and Ewing’s sarcoma cells demonstrated that **cyclin D1** is associated with both **p34cdc2** and **p33cdk2**, and that **cyclin D1** immune complexes exhibit appreciable histone H1 kinase activity.

Immobilized, recombinant cyclins A and D1 were found to associate with cellular proteins in complexes that contain the **p105Rb** protein.

This study identifies several common aspects of cyclin biochemistry, including tyrosine phosphorylation and the potential to interact directly or indirectly with the **Rb** protein, that may ultimately relate membrane-mediated signaling events to the regulation of gene expression.
Large-Scale Text Mining for Protein Interactions

- Applied trained extractors to 753,459 Medline abstracts that reference “human”.
- Verified on independent test data that extractors are about as accurate as human annotators.
- Automatically mined a large number of protein interactions from this scientific text.
  - Extracted 6,580 interactions between 3,737 human proteins
- Integrated extracted data with existing databases to construct the world’s largest database of human protein interactions.
  - 31,609 interactions between 7,748 human proteins
Conclusions

• Information Extraction technology can automatically assemble large scientific databases by just “reading” the literature.

• Extractors just need to be trained on a fairly large set of documents (hundreds) annotated by a human expert in order to learn to identify the requisite entities and relations.