Text Mining for Health Care and Medicine

Sophia Ananiadou
Director
National Centre for Text Mining
www.nactem.ac.uk
The Need for Text Mining

- MEDLINE
  - 2005: ~14M
  - 2009: ~18M
- Overwhelming information in textual, unstructured format
- Full papers, reports, grey literature (notes, lab records, discharge summaries)

"About a quarter of late stage failures we surveyed could have been eliminated 2 years earlier by making all internal information in documents more widely available."

Top 5 Pharma senior VP
The problem with information overload

- Humans cannot easily:
  - Keep up-to-date with all relevant literature
  - Find relevant and precise information
  - Synthesize information from many diverse sources
  - Exploit the mass of information to generate hypotheses
  - Discover new knowledge
What can Text Mining do for you?

- Helps with information overload and overlook
- Discovers unsuspected links from the huge amount of literature and supports medical research
- Integrates knowledge from many sources
- Enhances clinical decision support systems
- Supports translational medicine
- Reduces costs and errors in handling information
From Text to Knowledge

Unstructured Text
(implicit knowledge)

Structured content
(explicit knowledge)
National Centre for Text Mining

- 1st national text mining centre in the world
  www.nactem.ac.uk
- **Location**: Manchester Interdisciplinary Biocentre (MIB) www.mib.ac.uk
- **Remit**: Provision of text mining services to support UK research
- **Funded by**: the JISC, BBSRC, EPSRC
- **Domains**: Biology, Medicine, Social Sciences
How Text Mining can Support Medical Research

- Mines textual data (literature, admission notes, reports, summaries)
- Adds meaning to data → semantic metadata
- Yields precise knowledge nuggets from the sea of information → Information Extraction
- Supports not just medical research but also:
  - Clinicians in their care
Applications

- Gene-disease associations
- Disease-disease associations
  - Inferring relationships
  - Showing the evidence from literature
- Toxicity prediction
  - Discovery of promising drug targets for clinical trials
- Facilitates translational research into causes and treatment of diseases
- Clinical trials (collaborative project BRC-NaCTeM)
Text mining supports hypothesis generation

- Data driven methods complementing human hypothesis generation
- Rapid mining of candidate hypotheses from text, validated against experimental data
  - Migraine and magnesium deficiency
  - Indomethacin and Alzheimer’s disease
  - Using thalidomide for treating a series of diseases such as acute pancreatitis and chronic hepatitis C
  - *Curcuma longa* and retinal diseases
Translational Medicine

- Biomedical text mining customised for different users and needs
- Linking with “omics” data
- Supporting clinician and researcher in decision making through literature
Text Mining Systems @ NaCTeM

- **TerMine**
  - Automatic extraction of important concepts from text

- **AcroMine**
  - Acronym disambiguation and lookup

- **FACTA**
  - Association mining from MEDLINE, direct and indirect

- **ASSERT**
  - Aids systematic reviews, summarisation

- **KLEIO**
  - Searching using semantic types and several facets

- **MEDIE**
  - Searching using relations and semantic templates
TerMine
TerMine (C-value) analysis

Homologous desensitization of beta2-adrenergic receptors has been shown to be mediated by phosphorylation of the agonist-stimulated receptor by G-protein-coupled receptor kinase 2 (GRK2) followed by binding of beta-arrin2 to the phosphorylated receptor. Binding of beta-arrestin2 to the receptor is a prerequisite for subsequent receptor desensitization, internalization via clathrin-coated pits, and the initiation of alternative signaling pathways. In this study we have investigated the interactions between receptors and beta-arrestin2 in living cells using fluorescence resonance energy transfer. We show that (a) the initial kinetics of beta-arrestin2 binding to the receptor is limited by the kinetics of GRK2-mediated receptor phosphorylation, (b) repeated stimulation leads to the accumulation of GRK2-phosphorylated receptor, which can bind beta-arrestin2 very rapidly, and (c) the interaction of beta-arrestin2 with the receptor depends on the activation of the receptor by agonist because agonist withdrawal leads to swift dissociation of the receptor-beta-arrestin2 complex. This fast agonist-controlled association and dissociation of beta-arrin2 from prephosphorylated receptors should permit rapid control of receptor sensitivity in repeatedly stimulated cells such as neurons.

The beta2-adrenergic receptor (beta2-AR) belongs to the group of G-protein-coupled receptors and is present on skeletal and cardiac muscle cells and on lymphocytes. The gene encoding beta2-AR (ADRB2) displays a significant degree of polymorphism in the human population and the distributions of single-nucleotide polymorphisms (SNPs) at amino acid positions 46 and 37 in patients with essential hypertension and in the autoimmune disease myasthenia gravis. An involvement of ADRB2 has also been suggested in human rheumatoid arthritis (RA) and its animal model. We describe here an increased prevalence of the alleles Arg46 and Gln27 and a lower prevalence of Gln46 and Thr27 in patients with RA. Patients having the genetic combination C/Thr46, GlnThr27 had higher levels of...
1. Beta2-adrenergic receptor: 68.25
2. Blood pressure: 16.75
3. Beta2-adrenergic receptor gene: 11.5797
4. Odd ratio: 10
5. Protein kinase: 9.599999
7. Adrenergic receptor: 9.142857
8. Gly16 allele: 8
9. A549 cell: 8
10. Body mass index: 7.924812
11. Cystic fibrosis patient: 7.924812
12. Cystic fibrosis: 7.428571
13. Metabolic syndrome: 7
14. Confidence interval: 7
15. Bioluminescence resonance energy transfer: 6.8
17. Gene polymorphism: 6.4
18. Diastolic blood pressure: 6.32085

The table shows the ranked terms with their corresponding scores. The term 'Beta2-adrenergic receptor' has the highest score of 68.25, followed by 'Blood pressure' with a score of 16.75.

In addition to the table, there is a text box that appears to be discussing beta-arrestin binding and phosphorylation. The text mentions that the binding of beta-arrestins to the phosphorylated receptor is mediated by phosphorylation. The binding of beta-arrestins to the phosphorylated receptor is limited by the kinetics of dephosphorylation, which can influence the accumulation of GRK2-phosphorylated receptor, which can affect the activity of the receptor.
FACTA

- FACTA: Finds Associated Concepts with Text Analysis
  - What diseases are related to a particular chemical?
- FACTA+: finds indirect associations between concepts
- Quick and interactive, classifies documents based on associations, provides indirect associations
Retrieving related concepts

What kind of proteins are related to diabetes?

Insulin, albumin, ...

Diabetes is ...
... when \textbf{insulin} is ...
... lower \textbf{albumin} level

216,000 documents relevant to diabetes

MEDLINE (18 million abstracts)
ASSERT – aiding systematic reviews

- Goes beyond your average search engine
- Concept detection and highlighting to improve navigation
- Features include: document clustering, summarisation, user document management (MySystematicReview)
Assisting Systematic Reviews

- Searching
  - Query expansion, document clustering

- Screening
  - Document classification, sectioning

- Synthesising

Summary of: After Myocardial Infarction
Documents Processed: 11

- We used extended Cox hazards analysis to identify independent correlates of posttransplantation de novo CHF (adjusted hazard ratio 1.96, 95% confidence interval 0.50, P = 0.027).
- 361 patients in the pioglitazone group and 362 in the placebo group reached this endpoint (9.8 vs. 10.9, P = 0.68).
- Higher levels of myocardial G-isoprostane (G-Iso PROstaglandin (Ga)) and oxidized glutathione (GSSG), as well as greater upregulation of superoxide dismutase (SOD) and catalase (CAT) protein expression paralleled by increases in enzymatic activity, was observed in the diabetic MT animals, indicating higher oxidative stress.
- Twenty minutes after treatment with sublingual nitroglycerin and short-acting oral antiplatelet agent, blood pressure had dropped from 201.91 mm Hg to 158.84 mm Hg. Followed by abrupt onset of weakness in lower limbs, urinary retention and sensory loss in bilateral T4-L1 levels and the left lower limb at two hours after treatment.
- Advanced age and longer time on HD are factors related to LEA in non-diabetics.
- No association was found between the SPECT result and systolic function and left ventricular hypertrophy, however.

- INTERPRETATION: Plugging sodium reduces the composite of all-cause mortality, non-fatal myocardial infarction, and stroke in patients with type 2 diabetes who have a high risk of macrovascular events.
- Antagonism of aldosterone receptors with spironolactone benefits patients with severe heart failure, and spironolactone benefits those after myocardial infarction who have left ventricular dysfunction.
- Risk factors for de novo CHF included older recipient age, female sex, unemployed status at transplantation, pretransplantation comorbidities (diabetes mellitus, myocardial infarction, angina, cardiac arrhythmia, and peripheral vascular disease), transplant from older donors, donor cardiovascular death, and delayed graft function.
- These clinical manifestations and laboratory findings suggested catastrophic antiphospholipid antibody syndrome.

[Cluster results for 'diabetes']

- Diabetes Mellitus Type 2 (21) | Exp | Summ
- Diabetes Mellitus Type 1 (9) | Exp | Summ
- Type 2 Diabetes (20) | Exp | Summ
- For Diabetic Retinopathy (12) | Exp | Summ
- Metabolic Syndrome Group Patients (11) | Exp | Summ
- For Renal (13) | Exp | Summ
- For Blood Pressure (13) | Exp | Summ
- After Myocardial Infarction (11) | Exp | Summ
- For Creatinine (6) | Exp | Summ
- Management For (9) | Exp | Summ

- Mark as Viewed | Diabetes - Treatments | Commit changes
- Diabetes and the kidney, 16194585
- Diabetes is epidemic, 1621717
- Diabetes basics: Understanding insulin, 16220591
- 7 tips for controlling with diabetes, 16220599
- Treatment of gestational diabetes mellitus, 16221791
- Treatment of gestational diabetes mellitus, 16224824
- Treatment of diabetes mellitus in the elderly, 16246409
- H-tech advancements in diabetes management, 16187642
- Single-donor islet transplantation for diabetes, 16189359
- Diabetes and your marriage, Making things work, 16202596
- The Diabetes Prevention Program and the metabolic syndrome, 1620471
- The Diabetes Prevention Program and the metabolic syndrome, 1620473
- Stepping up care for diabetic foot ulcers, 16205268
- [Contribution of apoptosis to pathogenesis of type 1 diabetes], 16209940
- Type 2 diabetes with alcohol abuse, 16245400
Semantic search

- Specialised biomedical named entities, e.g. protein, gene...
- Linked to external knowledge sources
- Allows typed searching
- Normalises terms to include acronyms, synonyms and variants
- Faceted browsing

KLEIO SERVICE
Select listed entities to add them to query and narrow down the abstract list.

List of retrieved documents is updated with the new queries.
PubMedID: 6506428

Title: Posttranslational modification of human T-cell growth factor.

Abstract:
Amino-terminal sequence analysis of human T-cell growth factor indicated that the amino acid in position 3 of the polypeptide chain was Examination of the N-terminal octapeptide using the amino acid analyzer and mass spectrometry demonstrated that position 3 was a thi linked to N-acetyl-D-galactosamine. This site of glycosylation is of practical significance since it appears to play a role in the selectivity antibody for the factor.

Legend:

Legend:

Author(s): Popp BJ, Kutny RM, Panico M, Morris H, DeGrado WF, Chowdhry V
Mesh Heading(s): Amino Acid Sequence, Cell Line, Glycoproteins, Glycoproteins -- genetics, Hexosamines, Hexosamines -- analysis, Hu Interleukin-2 -- genetics, Peptide Fragments, Peptide Fragments -- analysis, Protein Processing, Post-Translational, Spectrum Analysis, T-Lymphocytes, Trypsin

Acronym(s): [No acronyms discovered]

Named Entities:
NE form: glycosylation
NE type: NATURAL PHENOM
CUI Number: C0047982,C0376322

NE form: polypeptide
NE type: METABOLITE
ID Number: KEGG:C00403[trDB]

NE form: N-acetyl-D-galactosamine
NE type: METABOLITE
MEDIE

- Semantic information extraction
  - extracts nuggets of knowledge from MEDLINE
  - interactive semantic retrieval from sentences
  - segments documents into sections, conclusion, methodology, etc
What does p53 activate?
p53 also activates the transcription of Mdm2, …

the growth inhibitory effects of Triphala is mediated by the activation of ERK and p53 …
Perform advanced search
In conclusion, our data also suggests that...
Our data also suggests that the growth inhibitory effects of Trichloroacetic acid (TCA) are mediated by the activation of ERK and p53, and shows potential for the treatment and prevention of human pancreatic cancer.

4. **p53 family in development**
   - Name: Santore, Kwame M. Salamano, Cho Lin, pp. 919-31, Volume 125, Issue 11-12, Mechanism of development, YYYY [PMID:18893440]
   - Imbalance of p53 protein family may contribute to a significant proportion of congenital developmental abnormalities in humans.

5. **Gambogic acid mediates apoptosis as a p53 inducer through down-regulation of mdm2 in wild-type p53-expressing cancer cells**
   - Gu, Hongyan, Xingfeng Wang, Shuyun Ren, Jia Wang, Jie Zhao, Fang Li Ren, Rong Mu, Yong Yang, Qi Qi, Wei Liu, Na Lu, Ha Ling, Qingdong You, Qingdong Guo, pp. 3295-305, Volume 7, Issue 10, Molecular cancer therapeutics, 2008 [PMID:18981293]
   - It is Click a gene name to show links to external databases

6. **Modulation of the DNA-damage response to HZE particles by shielding**
   - Interestingly, activation of the tumor suppressor p53 in FeR irradiated cells is uniquely biphasic and culminates in the induction of high levels of TP53 (Homo sapiens), p16, and H2AX.

7. **Combined effects of the p21 (Cip1/Waf1) and p16 (INK4a) genes on the risk of HPV-16-associated oral cancer in never-smokers**
   - These findings suggest that the interaction between p21 and p16 significantly increases the risk of HPV16-associated oral cancer, especially among never-smokers.

8. **Differential effect of camptothecin treatment on topoisomerase II alpha expression in ML-1 and HL-60 leukemia cell lines**
   - These results demonstrated that induction of p53 by camptothecin treatment can lead to a decreased level of TOP2 alpha and should be considered in design of combination therapy.

9. **Intrinsically unstructured domains of Arf and Hdm2 form bimolecular oligomeric structures in vitro and in vivo**
Conclusion

- Text Mining is an enabling technology for knowledge discovery, an integral part of medical informatics.

- Helps you to see the forest from the trees...