GeneView
Gene-Centric Ranking of Biomedical Text

Philippe Thomas
Humboldt Universität zu Berlin
Article retrieval and ranking

• Research results are primarily published in scientific journals

• Curation efforts can not keep up with the fast increase of literature [1]

• Searching for biological concepts like genes is difficult
  - Synonyms, Homologous genes, Other concepts, ...

• Queries lead often to thousands of documents
  - Appropriate article ranking is important
Workflow

Full text articles → Named entity recognition

- Genes [2]
- Mutations [3]
- Drugs
- Diseases

Section detection → IMRAD

Inverted index

Aspirin and Stroke prevention
- Aspirin and codeine in two

<table>
<thead>
<tr>
<th>Section</th>
<th>Boost</th>
</tr>
</thead>
<tbody>
<tr>
<td>Title</td>
<td>3.0</td>
</tr>
<tr>
<td>Abstract</td>
<td>2.0</td>
</tr>
<tr>
<td>Introduction</td>
<td>1.0</td>
</tr>
<tr>
<td>Result</td>
<td>2.0</td>
</tr>
</tbody>
</table>

Ranker using section specific weights
Task 1: Retrieval

- **Given:** Gene of interest
- **Task:** Search for articles relevant for this gene
- **Ranking:**
  - Requires gene name recognition and normalization
  - Lucene ranking as baseline
  - Section specific boosting
  - Query expansion using GO terms
Evaluation of Ranking

- Gene2PubMed as gold standard
- Top 20 results for 10 randomly selected genes
  - True positive if a hit is contained in gene2pubmed
- Search for best ranker configuration
  - Section boosts
  - Query expansion
Results of Ranking

- Precision of 75.5%
- Section specific boosting increases precision 3.5%
- Incorporation of captions lead to a small decrease
- Current method for query expansion using GO has no positive effect

<table>
<thead>
<tr>
<th>Section</th>
<th>Boost</th>
</tr>
</thead>
<tbody>
<tr>
<td>Title</td>
<td>3.0</td>
</tr>
<tr>
<td>Abstract</td>
<td>2.0</td>
</tr>
<tr>
<td>Introduction</td>
<td>1.0</td>
</tr>
<tr>
<td>Result</td>
<td>2.0</td>
</tr>
</tbody>
</table>
## Gene Centric Ranking of Biomedical Text Retrieval - Interface

### Query Filter Options

- Genes
- SNPs
- Genes from list
- Personal Genes

### Top 5 genes associated with your query (Entrez): 999, 20613, 7431, 7291, 2335

<table>
<thead>
<tr>
<th>Options</th>
<th>Entry</th>
</tr>
</thead>
</table>
| 1. Alterations of E-cadherin and β-catenin in gastric cancer  
Chen Huiping  
BMC Cancer - 2001 – (PMID: 11747475) | 9 2 |
| 2. Sialyl Lewis x expression in canine malignant mammary tumours: correlation with clinicopathological features and E-Cadherin expression  
Salomé S Pinho  
BMC Cancer - 2007 – (PMID: 17617904) | 5 0 |
| 3. Molecular Plasticity of E-Cadherin and Sialyl Lewis X Expression, in Two Comparative Models of Mammary Tumorigenesis  
Salomé S. Pinho  
- 2009 – (PMID: 19675678) | 7 0 |
| 4. BMP-6 promotes E-cadherin expression through repressing 5EF1 in breast cancer cells  
Shuang Yang  
BMC Cancer - 2007 – (PMID: 17997862) | 15 1 |
| 5. Heterogeneity in the modification and involvement of chromatin components of the CpG island of the silenced human CDH1 gene in cancer cells  
Shiro Kozuma  
Nucleic Acids Research - - (PMID: 12409468) | 25 0 |
| 6. A novel diffuse gastric cancer susceptibility variant in E-cadherin (CDH1) intron 2: A case control study in an Italian population  
Sorosh Naser  
BMC Cancer - 2008 – (PMID: 18482459) | 6 1 |
| 7. Expression analysis of E-cadherin, Slug and GSK3β in invasive ductal carcinoma of breast  
Chandra P Prasad  
BMC Cancer - 2009 – (PMID: 19751508) | 15 0 |
| 8. Absence of germline mono-allelic promoter hypermethylation of the CDH1 gene in gastric cancer patients  
Hidetaka Yamada  
Molecular Cancer - 2009 – (PMID: 19671136) | 10 3 |
Evaluation of query associated genes

- Single sided $\chi^2$ test
- ~1 second per query and entity type
- Evaluated the top 5 genes for the following specific queries:
  - Alzheimer (5/5)
  - Colorectal cancer (4/5)
  - Diabetes mellitus (1/5)
  - Parkinson (5/5)
Task 2: Indexing

- **Given:** Full text article of interest
- **Contribution:**
  - Full text visualization
  - Entity mark up and enrichment with link outs
  - User adjustable ranking of genes found in full text
Alterations of E-cadherin and β-catenin in gastric cancer

Chen Huiping
BMC Cancer – 2001

Abstract

Background The E-cadherin-catenin complex plays a crucial role in epithelial cell-cell adhesion and in the maintenance of tissue architecture. Perturbation in the expression or function of this complex results in loss of intercellular adhesion, with possible consequent cell transformation and tumour progression. Methods We studied the alterations of E-cadherin and β-catenin in a set of 50 primary gastric tumours by using loss of heterozygosity (LOH) analysis, gene mutation screening, detection of aberrant transcripts and immunohistochemistry (IHC). Results A high frequency (75%) of LOH was detected at 16q22.1 containing E-cadherin locus. Three cases (6%) showed the identical missense mutation, A592T. This mutation is not likely to contribute strongly to the carcinogenesis of gastric cancer, because a low frequency (1.6%) of this mutation was also found in 187 normal individuals. We also detected a low frequency (0.36%, 0%) of this mutation in 280 breast tumours and 444 other tumours, including colon and rectum, lung, endometrium, ovary, tests, kidney, thyroid carcinomas and sarcomas, respectively. We also analyzed the aberrant E-cadherin mRNAs in the gastric tumours and found that 7 tumours (14%) had aberrant mRNAs in addition to the normal mRNA. These aberrant mRNAs may produce abnormal E-cadherin molecules, resulting in weak cell-cell adhesion and invasive behaviour of carcinoma cells. Reduced expression of E-cadherin and β-catenin was identified at the frequency of 42% and 28%, respectively. Specially, 11 tumours (22%) exhibited positive cytoplasmic staining for β-catenin IHC. An association was found between reduced expression of E-cadherin and β-catenin. Moreover, an association was detected between reduced expression of E-cadherin and diffuse histotype. Conclusion Our results support the hypothesis that alterations of E-cadherin and β-catenin play a role in the initiation and progression of gastric cancer.

Background

E-cadherin (120 kDa; chromosome 16q) is a classical cadherin and forms the key functional component of adherence junctions between epithelial cells [1]. It is bound via a series of undercoat proteins, the catenins (α, β and γ) to the actin cytoskeleton [2]. This linkage between transmembraneous cadherins and actin
Interface - Indexing

Gene Centric Ranking of Biomedical Text

Article specific gene ranking

Entity specific link outs

BioCreative III,
Conclusion

- GeneView is developed as part of the ColoNet project
- Gene-centric article ranking and visualization
- Evaluation using gene2pubmed
- Detection of query associated entities
Thank you for your attention!

Questions?


## Section normalization

<table>
<thead>
<tr>
<th>Normalized Section</th>
<th>Number</th>
</tr>
</thead>
<tbody>
<tr>
<td>Title</td>
<td>17,780</td>
</tr>
<tr>
<td>Abstract</td>
<td>17,780</td>
</tr>
<tr>
<td>Results</td>
<td>13,879</td>
</tr>
<tr>
<td>Methods</td>
<td>13,650</td>
</tr>
<tr>
<td>Introduction</td>
<td>13,140</td>
</tr>
<tr>
<td>Discussion</td>
<td>12,255</td>
</tr>
<tr>
<td>Other</td>
<td>12,228</td>
</tr>
<tr>
<td>Conclusion</td>
<td>5,063</td>
</tr>
<tr>
<td>Supplement</td>
<td>3,836</td>
</tr>
<tr>
<td>Not normalized</td>
<td>2,503</td>
</tr>
<tr>
<td>abbreviations</td>
<td>2,088</td>
</tr>
</tbody>
</table>
NER results

<table>
<thead>
<tr>
<th>Entity Type</th>
<th>Articles</th>
<th>Entities</th>
</tr>
</thead>
<tbody>
<tr>
<td>Genes</td>
<td>16,013</td>
<td>1,294,875</td>
</tr>
<tr>
<td>Drugs</td>
<td>15,102</td>
<td>535,707</td>
</tr>
<tr>
<td>Diseases</td>
<td>12,166</td>
<td>940,225</td>
</tr>
<tr>
<td>SNPs</td>
<td>4,846</td>
<td>91,410</td>
</tr>
</tbody>
</table>