Text Mining MEDLINE

A Simple Application to Mine MEDLINE

Raf Podowski
Life Sciences
Oracle Corporation
raf.podowski@oracle.com
Outline

- Loading
- Retrieval
- Indexing
- Searching
- Tokens

- Themes
- Thesauri
- Clustering
- Classification
- Demo
Loading

- Storage choice
  - XML DB ★
  - CLOB

- Loading method
  - SQL*Loader ★
  - INSERT
  - UPSERT ★

- Initial Loading with 12M Records
- Periodic Updates
Loading

```sql
create table medtab (
    pmid number primary key,
    text xmltype
);

Load DATA
    INFILE 'medline.dat'
    BADFILE 'medline.bad'
    DISCARDFILE 'medline.discard'
    INTO TABLE medtab
    REPLACE
    FIELDS TERMINATED BY '\t'
    (pmid, text char(1000000))

select pmid,
    extractValue(text, '/MedlineCitation/Article/Abstract/AbstractText') Abstract
from medtab
where PMID='15129431';
```
Loading

```
MERGE INTO table
USING table/view/subquery
ON ( condition )
WHEN MATCHED THEN update clause
WHEN NOT MATCHED THEN insert clause
```

SELECT * FROM table WHERE condition;
Estrogen receptor immunoreactivity in meningiomas. Comparison with the binding activity of estrogen, progesterone, and androgen receptors.

Estrogen receptor (ER) analysis was performed in 70 meningioma samples by means of two assays: an enzyme immunoassay that used monoclonal antibodies against human ER protein (estrophilin), and a sensitive radioligand binding assay that used iodine-125-labeled estradiol as the radioligand. Low levels of ER immunoreactivity were found in tumors from 51% of patients, whereas ER binding activity was demonstrated in 40% of the meningiomas examined. In eight (11%) of the tissue samples, multiple binding sites for estradiol were observed. The immunoreactive binding sites corresponded to those of the classic high-affinity ER. In ligand binding studies, however, measurement of classic ER was...
Retrieval

```sql
select extract(text,
    '/MedlineCitation/MeshHeadingList/MeshHeading/DescriptorName').getStringVal()
from medtab
where PMID='3298569';

<DescriptorName MajorTopicYN="N">Comparative Study</DescriptorName>
<DescriptorName MajorTopicYN="N">Female</DescriptorName>
<DescriptorName MajorTopicYN="N">Human</DescriptorName>
<DescriptorName MajorTopicYN="N">Immunoenzyme Techniques</DescriptorName>
<DescriptorName MajorTopicYN="N">Male</DescriptorName>
<DescriptorName MajorTopicYN="N">Meningeal Neoplasms</DescriptorName>
<DescriptorName MajorTopicYN="N">Middle Aged</DescriptorName>
<DescriptorName MajorTopicYN="N">Radioligand Assay</DescriptorName>
<DescriptorName MajorTopicYN="N">Receptors, Androgen</DescriptorName>
<DescriptorName MajorTopicYN="N">Receptors, Estrogen</DescriptorName>
<DescriptorName MajorTopicYN="N">Receptors, Progesterone</DescriptorName>
```
select extract(text,
    '/MedlineCitation/ChemicalList/Chemical/NameOfSubstance').getStringVal()
from medtab
where PMID='3298569';

<NameOfSubstance>Receptors, Androgen</NameOfSubstance>
<NameOfSubstance>Receptors, Estrogen</NameOfSubstance>
<NameOfSubstance>Receptors, Progesterone</NameOfSubstance>
Indexing

- **Filter**
  - File formats
- **Sectioner**
  - HTML
  - XML
- **Tokenizer**
  - Lexer - tokenize
  - Stoplists - mask
Indexing

• Index Types
  ♦ CONTEXT
    • Text retrieval
    • CONTAINS query operator
  ♦ CTXCAT
    • Item categories
    • CATSEARCH query operator
  ♦ CTXRULE
    • Classification rules
    • MATCHES query operator
Indexing

-- Enable theme indexing
exec ctx_ddl.create_preference('mylex','BASIC_LEXER');
exec ctx_ddl.set_attribute('mylex','MIXED_CASE','NO');
exec ctx_ddl.set_attribute('mylex','THEME_LANGUAGE','ENGLISH');
exec ctx_ddl.set_attribute('mylex','index_themes','YES');
exec ctx_ddl.set_attribute('mylex','index_text','YES');

-- Create XML sections
exec ctx_ddl.create_section_group('xmlgroup','auto_section_group');

-- Index column 'text' of table 'medtab' for user 'hex'
create index medtab_idx on medtab(text)
indextype is ctxsys.context
parameters('lexer mylex filter ctxsys.null_filter section group xmlgroup');
```sql
COL Title FORMAT a60;
COL S FORMAT 999;
select score(1) s, pmid,
   extractValue(text, '/MedlineCitation/Article/ArticleTitle') Title
from medtab
where CONTAINS(text, 'aldose reductase WITHIN AbstractText', 1) > 0
ORDER BY score(1) DESC;

<table>
<thead>
<tr>
<th>S</th>
<th>PMID</th>
<th>TITLE</th>
</tr>
</thead>
<tbody>
<tr>
<td>12</td>
<td>9537432</td>
<td>New member of aldose reductase family proteins overexpressed in human hepatocellular carcinoma.</td>
</tr>
<tr>
<td>12</td>
<td>10322639</td>
<td>Developmental expression of urine concentration-associated genes and their altered expression in murine infantile-type polycystic kidney disease.</td>
</tr>
<tr>
<td>12</td>
<td>9565553</td>
<td>Identification and characterization of a novel human aldose reductase-like gene.</td>
</tr>
<tr>
<td>12</td>
<td>11261885</td>
<td>Overexpression of aldose reductase in liver cancers may contribute to drug resistance.</td>
</tr>
</tbody>
</table>
```
### Searching

```sql
select score(1) s, pmid, 
    extractValue(text, '/MedlineCitation/Article/ArticleTitle') Title 
from medtab 
where contains(text, 
    '<query><textquery>aldose reductase WITHIN AbstractText</textquery> 
    <score algorithm="COUNT"/></query>', 1) > 0;
```

<table>
<thead>
<tr>
<th>S</th>
<th>PMID</th>
<th>Title</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>9537432</td>
<td>New member of aldose reductase family proteins overexpressed in human hepatocellular carcinoma.</td>
</tr>
<tr>
<td>1</td>
<td>10322639</td>
<td>Developmental expression of urine concentration-associated genes and their altered expression in murine infantile-type polycystic kidney disease.</td>
</tr>
<tr>
<td>1</td>
<td>9565553</td>
<td>Identification and characterization of a novel human aldose reductase-like gene.</td>
</tr>
<tr>
<td>1</td>
<td>11261885</td>
<td>Overexpression of aldose reductase in liver cancers may contribute to drug resistance.</td>
</tr>
</tbody>
</table>
Basal cell carcinomas in mice overexpressing sonic hedgehog.

Mutations in the tumor suppressor gene PATCHED (PTC) are found in human patients with the basal cell nevus syndrome, a disease causing developmental defects and tumors, including basal cell carcinomas. Gene regulatory relationships defined in the fruit fly Drosophila suggest that overproduction of Sonic hedgehog (SHH), the ligand for PTC, will mimic loss of ptc function. It is shown here that transgenic mice overexpressing SHH in the skin develop many features of basal cell nevus syndrome, demonstrating that SHH is sufficient to induce basal cell carcinomas in mice. These data suggest that SHH may have a role in human tumorigenesis.
Basal cell carcinomas in mice overexpressing sonic hedgehog.

Mutations in the tumor suppressor gene PATCHED (PTC) are found in human patients with the basal cell nevus syndrome, a disease causing developmental defects and tumors, including basal cell carcinomas. Gene regulatory relationships defined in the fruit fly Drosophila suggest that overproduction of Sonic hedgehog (SHH), the ligand for PTC, will mimic loss of ptc function. It is shown here that transgenic mice overexpressing SHH in the skin develop many features of basal cell nevus syndrome, demonstrating that SHH is sufficient to induce basal cell carcinomas in mice. These data suggest that SHH may have a role in human tumorigenesis.
Thesauri

begin
 ctx_thes.create_thesaurus('genes', FALSE);
end;

begin
 ctx_thes.create_phrase('genes','LLID_231');
 ctx_thes.create_phrase('genes','LLID_367');
 ctx_thes.create_phrase('genes','LLID_374');
 ctx_thes.create_phrase('genes','AR');
end;

begin
 ctx_thes.create_phrase('genes','androgen receptor');
 ctx_thes.create_phrase('genes','dihydrotestosterone receptor');
 ctx_thes.create_phrase('genes','HGNC:644');
 ctx_thes.create_phrase('genes','AIS');
 ctx_thes.create_phrase('genes','DHTR');
 ctx_thes.create_phrase('genes','HUMARA');
 ctx_thes.create_phrase('genes','KD');
 ctx_thes.create_phrase('genes','NR3C4');
 ctx_thes.create_phrase('genes','SBMA');
 ctx_thes.create_phrase('genes','SMAX1');
 ctx_thes.create_phrase('genes','TFM');
end;
begin
    ctx_thes.create_relation('genes','gene','NT','LLID_231');
    ctx_thes.create_relation('genes','gene','NT','LLID_367');
    ctx_thes.create_relation('genes','gene','NT','LLID_374');
    ctx_thes.create_relation('genes','LLID_367','NT','AR');
    ctx_thes.create_relation('genes','LLID_367','NT','androgen receptor');
end;

declare
    synonyms varchar2(2000);
begin
    synonyms := ctx_thes.nt('LLID_367',1,'genes');
    dbms_output.put_line('Thesaurus: genes');
    dbms_output.put_line('The synonym expansion for LLID_367 is: '||synonyms);
end;

Thesaurus: genes
The synonym expansion for LLID_367 is: {LLID_367}|{ANDROGEN RECEPTOR}|{AR}|
{DIHYDROTESTOSTERONE RECEPTOR}|{HGNC:644}|{AIS}|{DHT}|{HUMARA}|{KD}|{NR3C4}|
{SBMA}|{SMAX1}|{TFM}
Document Clustering

• Unsupervised Classification
  ♦ No training needed
  ♦ Good for initial overview of a group of documents
  ♦ Identifies shares attributes

• CTX_CLS.CLUSTERING
  ♦ KMEAN
    • requires setting the number of clusters
  ♦ TEXTK
    • experimental hierarchical clustering
Document Clustering

- Prepare database objects
  - Collection table
  - Clusters table
  - Document results table
- Set clustering preferences
- Populate and index collection table
- Run clustering
Clustering

- **Collection table**

```sql
create table collection (id number primary key, text clob);
```

- **Clusters table**

```sql
create table clusters (clusterid NUMBER,
descript varchar2(4000),
label varchar2(200),
sze NUMBER,
quality_score NUMBER,
parent NUMBER);
```

- **Document results table**

```sql
create table restab (docid NUMBER,
clusterid NUMBER,
score NUMBER);
```
Clustering

- Index collection table

```sql
create index collectionx on collection(text)
  indextype is ctxsys.context
  parameters('STOPLIST CTXSYS.DEFAULT_STOPLIST');
```

- Examine token TF and DF

```sql
Declare
    x clob := null;
Begin
    ctx_report.index_stats('collectionx', x);
    insert into collection_stats values (x);
    commit;
    commit;
    dbms_lob.freetemporary(x);
    end;

Select * from collection_stats;
```
Clustering

- **CTX_REPORT.INDEX_STATS**

  Query: liver cancer
  No.docs: 9, No.tokens: 702

<table>
<thead>
<tr>
<th>Token</th>
<th>TF</th>
<th>DF</th>
<th>TFIDF</th>
</tr>
</thead>
<tbody>
<tr>
<td>DNA</td>
<td>22</td>
<td>2</td>
<td>11.00</td>
</tr>
<tr>
<td>TISSUE</td>
<td>30</td>
<td>3</td>
<td>10.00</td>
</tr>
<tr>
<td>CELLS</td>
<td>37</td>
<td>5</td>
<td>7.40</td>
</tr>
<tr>
<td>TUMOR</td>
<td>29</td>
<td>4</td>
<td>7.25</td>
</tr>
<tr>
<td>AR</td>
<td>40</td>
<td>6</td>
<td>6.67</td>
</tr>
<tr>
<td>ENZYME</td>
<td>12</td>
<td>2</td>
<td>6.00</td>
</tr>
<tr>
<td><strong>LIVER</strong></td>
<td>50</td>
<td><strong>9</strong></td>
<td>5.56</td>
</tr>
<tr>
<td>PROTEIN</td>
<td>11</td>
<td>2</td>
<td>5.50</td>
</tr>
<tr>
<td>DIETS</td>
<td>11</td>
<td>2</td>
<td>5.50</td>
</tr>
</tbody>
</table>

- **Add stopwords to collection index**

  ALTER INDEX collectionx REBUILD PARAMETERS ('ADD STOPWORD 1');
  ALTER INDEX collectionx REBUILD PARAMETERS ('ADD STOPWORD CANCER');
  ALTER INDEX collectionx REBUILD PARAMETERS ('ADD STOPWORD LIVER');
Clustering

♦ Set clustering preferences

BEGIN
  ctx_ddl.drop_preference('my_cluster');
  ctx_ddl.create_preference('my_cluster','KMEAN_CLUSTERING');
  ctx_ddl.set_attribute('my_cluster','CLUSTER_NUM',5);
  ctx_ddl.set_attribute('my_cluster','MAX_FEATURES',200);
  ctx_ddl.set_attribute('my_cluster','MAX_DOCTERMS',20);
END;

♦ Cluster collection

BEGIN
  ctx_cls.clustering('collectionx','id','restab','clusters','my_cluster');
END;
Document Classification

- Supervised Classification
  - Needs training
  - Can be applied to any document
- Rule-based
  - Manual rule creation
- Decision Trees
  - Automatic rule creation (editable)
- SVM
  - Automatic rule creation (opaque)
SVM Classification

- Training Documents
  - Create and populate training document table
  - Generate CONTEXT index on documents
- Categories
  - Assign documents to categories
- Set classifier preferences
  - MAX_FEATURES
- Train Classifier
  - Create Rules Table
  - Train
  - Generate a CTXRULE index on the rules table
- Classify New Documents
SVM Classification

- Create and populate training documents table

  ```sql
  create table svmtrain (  
    docid number primary key,  
    text clob);
  ```

- Create training document index

  ```sql
  create index svmtrainx on svmtrain(text)  
  indextype is ctxsys.context  
  parameters('STOPLIST CTXSYS.DEFAULT_STOPLIST');
  ```
SVM Classification

- Create and populate the category table

```sql
create table svmcats (
    docid number,
    cat_id number,
    catname varchar2(250));
```

- Create the rules table

```sql
create table svmtab (
    cat_id number,
    type number(3) NOT NULL,
    rule blob);
```
SVM Classification

- Set SVM classifier preferences
  ```java
  begin
  ctx_dml.set_attribute('mysvm','MAX_FEATURES','100');
  end;
  ```

- Train SVM classifier
  ```java
  begin
  ctx_cls.train('svmtrainx',
  'docid',
  'svmcats',
  'docid',
  'cat_id',
  'svmtab',
  'mysvm');
  end;
  ```
SVM Classification

- Create rules index
  ```sql
  create index svmx on svmtab(rule)
  indextype is ctxsys.ctxrule
  parameters ('filter svmfilter classifier mysvm');
  ```

- Classify unknown documents
  ```sql
  select cat_id, match_score(1) SCORE
  from svmtab
  where matches(rule, (select extractValue(text, '/MedlineCitation/Article/ArticleTitle')||' '||extractValue(text, /MedlineCitation/Article/Abstract/AbstractText')
          from medtab
          where pmid='7587903'), 1) > 50;
  ```
DEMO
Oracle Text Application

Load MEDLINE documents → Index with Oracle Text → Create SVM category → SVM Classification → Retrieve Document → Document Clustering → Co-occurrence matrix → Stopwords → Cluster Visualization

Interactive UI
Search

Oracle Text Medline Query System

There are 4318 MEDLINE records in the database.

Query: angiogenesis

Search Fields: 
- Abstract Text
- Article Title
- MeSH Term
- Name of Substance
- Whole Record

Search Type: 
- Context
- Theme

Search Button
Clear Button
Dups, start over

Browse Thesauri
- NCI Thesaurus
- Transcription Factors
Search

Oracle Text Medline Query System

There are 4318 MEDLINE records in the database.

Query

angiogenesis

Search Fields

- Abstract Text
- Article Title
- MeSH Term
- Name of Substance
- Whole Record

Search Type

- Context
- Theme

Search
Clear
Dops, start over

Searching for angiogenesis context in Article Title

SELECT score(l) as S, pmid,
extractValue(text, '/MedlineCitation/article/articleTitle') title
from medtab
where CONTAINS
(text, '(angiogenesis) WITHIN ArticleTitle', 1)>=0
ORDER BY score(l) DESC

Found 19 matching records
<table>
<thead>
<tr>
<th>PMID</th>
<th>Score Contents</th>
</tr>
</thead>
<tbody>
<tr>
<td>14566013</td>
<td>8 Perlecan and tumor angiogenesis.</td>
</tr>
<tr>
<td>14585356</td>
<td>8 Disruption of Id1 reveals major differences in angiogenesis between</td>
</tr>
<tr>
<td></td>
<td>transplanted and autochthonous tumors.</td>
</tr>
<tr>
<td>12816951</td>
<td>8 Honokiol, a small molecular weight natural product, inhibits angiogenesis</td>
</tr>
<tr>
<td></td>
<td>in vitro and tumor growth in vivo.</td>
</tr>
<tr>
<td>12462194</td>
<td>8 Nitric oxide synthase-cyclooxygenase interactions are involved in tumor</td>
</tr>
<tr>
<td></td>
<td>cell angiogenesis and migration.</td>
</tr>
<tr>
<td>12454298</td>
<td>8 Complex interactions between the laminin alpha 4 subunit and integrins</td>
</tr>
<tr>
<td></td>
<td>regulate endothelial cell behavior in vitro and angiogenesis in vivo.</td>
</tr>
<tr>
<td>11803470</td>
<td>8 Low plasma levels of matrix metalloproteinase 9 permit increased tumor</td>
</tr>
<tr>
<td></td>
<td>angiogenesis.</td>
</tr>
<tr>
<td>11728927</td>
<td>8 Prevention of fracture healing in rats by an inhibitor of angiogenesis.</td>
</tr>
<tr>
<td>11684441</td>
<td>8 The many faces of metalloproteases: cell growth, invasion, angiogenesis,</td>
</tr>
<tr>
<td></td>
<td>and metastasis.</td>
</tr>
<tr>
<td>10188911</td>
<td>8 Increased transcriptional activity of prostate-specific antigen in the</td>
</tr>
<tr>
<td></td>
<td>presence of TNP-470, an angiogenesis inhibitor.</td>
</tr>
<tr>
<td>11147680</td>
<td>8 Regulation of angiogenesis and tumorigenesis by signal transduction</td>
</tr>
<tr>
<td></td>
<td>cascades: lessons from benign and malignant endothelial tumors.</td>
</tr>
<tr>
<td>10077584</td>
<td>8 Troponin I is present in human cartilage and inhibits angiogenesis.</td>
</tr>
<tr>
<td>10611308</td>
<td>8 Thrombospondin-2 a potent endogenous inhibitor of tumor growth and</td>
</tr>
<tr>
<td></td>
<td>angiogenesis.</td>
</tr>
</tbody>
</table>
Perlecan is a major heparan sulfate proteoglycan (HSPG) of basement membranes (BM)s and connective tissues. The core protein of perlecan is divided into five domains based on sequence homology to other known proteins. Commonly, the N-terminal domain I of mammalian perlecan is substituted with three HS chains that can bind a number of matrix molecules, cytokines, and growth factors. Perlecan is essential for metazoan life, as shown by genetic manipulations of nematodes, insects, and mice. There are also known human mutations that can be lethal. In vertebrates, new functions of perlecan emerged with the acquisition of a closed vascular system and skeletal connective tissues. Many of perlecan’s functions may be related to the binding and presentation of growth factors to high-affinity tyrosine kinase (TK) receptors. Data are accumulating, as discussed here, that similar growth factor-mediated processes may have unwanted promoting effects on tumor cell proliferation and tumor angiogenesis. Understanding of these attributes at the molecular level may offer opportunities for therapeutic intervention.
Perlecanc is a connective tissue protein and has a sequence homology to mammalian perlecan molecules, cytogenetic manipulations that lead to the acquisition of a variety of functions may provide insights into tyrosine kinases and tumor factor-mediated and tumoral opportunities for growth and angiogenesis.
SVM classification

Initialize SVM  View/Edit SVM Categories  Add SVM Category

Category Name: angiogenesis  SVM Classification

Oracle Text Medline Query System

SVM Classification Setup

- Refreshing SVM objects...
- Removing old categories...
- Removing old training documents...
- Clearing SVM preferences...
- Removing old rules...

Done! Please close window.

Oracle®

LIFE SCIENCES USERS GROUP
SVM classification

Oracle Text Search - Mozilla Firefox

SVM classification

Initialize SVM  View/Edit SVM Categories  Add SVM Category

Category Name: angiogenesis

SVM Classification

Oracle Text Medline Query System

SVM Classification Category Setup

Category Name: angiogenesis

Getting new category id...
... new category id is 1...
Populating category table...
Populating training document table...

Done! Please close window.
Searching for **growth factor context** in Article Title

Found 181 matching records

**SVM classification**

<table>
<thead>
<tr>
<th>PMID</th>
<th>Score</th>
<th>Contents</th>
</tr>
</thead>
<tbody>
<tr>
<td>7522958</td>
<td>Fetch</td>
<td>15 Androgen receptor activation in prostatic tumor cell lines by insulin-like <strong>growth factor-I</strong>, keratinocyte <strong>growth factor</strong>, and epidermal <strong>growth factor</strong>.</td>
</tr>
<tr>
<td>7515785</td>
<td>Fetch</td>
<td>15 Epidermal <strong>growth factor</strong> suppresses insulin-like <strong>growth factor</strong> binding protein 3 levels in human papillomavirus type 16-immortalized cervical epithelial cells and thereby potentiates the effects of insulin-like <strong>growth factor</strong> 1.</td>
</tr>
<tr>
<td>14969747</td>
<td>Fetch</td>
<td>10 Transforming <strong>growth factor</strong> beta receptor family ligands inhibit hepatocyte <strong>growth factor</strong> synthesis and secretion from astrocytoma cells.</td>
</tr>
<tr>
<td>12446587</td>
<td>Fetch</td>
<td>10 Heparin-binding epidermal <strong>growth factor</strong>-like <strong>growth factor</strong> stimulates androgen-independent prostate tumor growth and antagonizes androgen receptor function.</td>
</tr>
</tbody>
</table>
**SVM Classification**

Removing old rules...

Creating training document index...

```sql
create index svmtrainx on svmtrain(text)
indextype is ctxsys.context
parameters('STOPLIST_CTXSYS_DEFAULT_STOPLIST')
```

Creating SVM classifier preferences...

```sql
begin
ctx_dll.set_attribute('mysvm','MAX_FEATURES','100');
end;
```

Training SVM classifier...

```sql
begin
ctx_cls.train('svmtrainx','docid','svmcats','docid','cst_id','svmtab','mysvm');
end;
```

Creating rules index...

```sql
create index svmx on svmtab(rule)
indextype is ctxsys.ctxrule
parameters ('filter svmfilter classifier mysvm')
```

Classifying documents...
### SVM

**Category** | **PMID** | **Score**
--- | --- | ---
angiogenesis | 11775025 | 79
| 12823209 | 62
| 9562977 | 57

metastasis | 11250937 | 87
| 11830552 | 81
| 2556408 | 68
| 1696279 | 68
| 1406654 | 65
| 11401606 | 63
| 8682172 | 59
| 1946376 | 59
| 9096655 | 57
| 8180021 | 57
| 2825016 | 56
| 10232609 | 55
| 10192430 | 55
| 1642153 | 55
| 9766531 | 54
| 8288913 | 53
| 1577731 | 52
| 9228090 | 52

apoptosis inhibition | 7784269 | 89
| 3322936 | 88
| 1425449 | 87
| 10362109 | 86
| 2144462 | 86
| 7784268 | 89
| 9322936 | 88
| 1425449 | 87
| 10362109 | 86
| 8194469 | 86
| 7606741 | 85
| 10638964 | 84
| 12646579 | 84
| 15138586 | 84
| 10568211 | 83
| 14662770 | 83
| 2036957 | 83
| 769570 | 83
| 7532576 | 83
| 1330491 | 82
| 11248024 | 81
| 11402326 | 81
| 8636339 | 81
| 15033751 | 81
| 11707452 | 80
| 1419897 | 80
| 8621250 | 80
| 9452071 | 80
| 10706107 | 79
| 11069386 | 79
| 8427046 | 79
| 12212679 | 79
## Clustering

### Document Clustering

- **KMEAN Cluster**
  - Cluster Size: 10
  - Max Distinct Terms per Doc: 30

- **TEXTK Cluster**
  - Maximum Splits per Node: 3
  - Hierarchy Depth: 1
  - Minimum Leaf Cluster Similarity Score: 0.2

### PMID | Score | Contents
--- | --- | ---
7522859 | Fetch | 15 Androgen receptor activation in prostatic tumor cell lines by insulin-like growth factor-I, keratinocyte growth factor, and epidermal growth factor.
7515765 | Fetch | 15 Epidermal growth factor suppresses insulin-like growth factor binding protein 3 levels in human papillomavirus type 16-immortalized cervical epithelial cells and thereby potentiates the effects of insulin-like growth factor 1.
14969747 | Fetch | 10 Transforming growth factor beta receptor family ligands inhibit hepatocyte growth factor synthesis and secretion from astrocytoma cells.
12446587 | Fetch | 10 Heparin-binding epidermal growth factor-like growth factor stimulates androgen-independent prostate tumor growth and antagonizes androgen receptor function.
11556655 | Fetch | 10 Expression of androgen receptor, epidermal growth factor receptor, and transforming growth factor alpha in salivary duct carcinoma.
Clustering

Oracle Text Medline Query System

Document Clustering

Populating collection table...

Creating clustering index...

cREATE INDEX collectionx ON collection(text)
indextype is ctksys.context
parameters({'STOPLIST CTKSYS.DEFAULT_STOPLIST'}

Clustering collection...
Gathering collection stats...

SUMMARY
- No. Docs = 161
- No. tokens = 4039

<table>
<thead>
<tr>
<th>Token</th>
<th>TF</th>
<th>DF</th>
<th>TFIDF</th>
</tr>
</thead>
<tbody>
<tr>
<td>IGF</td>
<td>260</td>
<td>27</td>
<td>9.63</td>
</tr>
<tr>
<td>EOF</td>
<td>444</td>
<td>53</td>
<td>8.38</td>
</tr>
<tr>
<td>TOF</td>
<td>425</td>
<td>52</td>
<td>8.17</td>
</tr>
<tr>
<td>BETA</td>
<td>337</td>
<td>42</td>
<td>8.02</td>
</tr>
<tr>
<td>ALPHA</td>
<td>344</td>
<td>45</td>
<td>7.64</td>
</tr>
<tr>
<td>ANDROGEN</td>
<td>198</td>
<td>27</td>
<td>7.33</td>
</tr>
<tr>
<td>AR</td>
<td>439</td>
<td>60</td>
<td>7.32</td>
</tr>
<tr>
<td>CELLS</td>
<td>882</td>
<td>127</td>
<td>6.94</td>
</tr>
</tbody>
</table>
## Clustering

<table>
<thead>
<tr>
<th>Cluster</th>
<th>Quality</th>
<th>Size</th>
<th>Cluster Tokens</th>
</tr>
</thead>
<tbody>
<tr>
<td>TGFbeta1</td>
<td>0.80</td>
<td>13</td>
<td>TGFbeta1, Stromal, Endothelial, Melanocytes, Vivo, Metalloproteinase, Signal, MMP, Two, Induction, Independent, System, Tumour, Transfected, VEGF, Antibody, Transduction, Phosphorylation, Studies, Melanoma, Squamous, Addition, Ovarian, HB, Activated, ACD, Cultures, Enhanced, Heparin, Significant, ID, Secretion, Ligand, Pathway, Observed, Culture, Regulation, Tyrosine, H, Mast, Mediated, Fold, T, Matrix, Vitro, Presence, Level, Line, 10, Reduced, Response, Activation, Type, Breast, Inhibition, Transformed, Stimulated, Between, Effects, C, Kinase, R, Carcinoma, Found, Epithelial, Both, Expressed, EGFR, Receptors, Effect, Lines, Prostate, Proliferation, Tumors, Activity, May, Androgen, Induced, Tumor, Binding, Normal, Cancer, MRNA, Alpha, Protein, Ability, Epidermal, TGF, These, EGF, AR, Receptor, Human, Expression, Cell, Cells, Biological, Factor, Growth, Vascular, ACT, Melanomas, Retinoic, SHowed, EXogenous, Inhibitor, Inhibitors, While, Induces, Three, Stimulates, Stimulation, Secreted, Suppression, 8, Differential, Mechanisms, Used, Inhibited, Soluble, 7, Migration</td>
</tr>
<tr>
<td>Plasma</td>
<td>0.50</td>
<td>13</td>
<td>Plasma, Transformation, 8, Age, Detected, Platelet, Melanocytes, Vivo, Formation, Development, System, Non, Control, During, Metastatic, Years, Time, ML, Phosphorylation, Duct, Ovarian, NG, Conditioned, Reacton, Elevated, N, Cultures, EGF, Ras, Production, Pathway, Patients, Culture, Associated, Tyrosine, T, Progression, MCF, Vitro, Significantly, Reduced, Breast, Inhibition, Transformed, Role, Skin, Signaling, Effects, Mice, Kinase, R, Related, Gene, Found, Treatment, Epithelial, Dependent, MM, Derived, II, EGFR, Factors, I, Effect, Tumors, Activity, May, Increased, Bone, Induced, Tumor, Beta, Levels, Cancer, Alpha, Protein, Transforming, Ability, Epidermal, TGF, These, EGF, Receptor, Human, Expression, Cell, Cells, Factor, Growth, Bearing, Various, Kidney, IL, Inhibitor, EGFR, Against, PTBH, Positive, Exposure, 23, Calcium, 9, Renal, Model, Antigen, Used, Inhibited, 7</td>
</tr>
<tr>
<td>Sequence</td>
<td>0.30</td>
<td>12</td>
<td>Sequence, Hepatocyte, Early, Platelet, Endothelial, Epidermis, Melanocytes, Metalloproteinase, Using, Signal, Development, MMP, Induction, HGF, Years, Transduction, MAB, Small, Differentiation, Cellular, Genes, Several, Staining, Melanoma, Squamous, Types, Malignant, Elevated, Activated, Fibroblast, Specific, BFGF, Production, Ligand, Pathway, Observed, Patients, Regulation, Tyrosine, Mast, Carcinomas, Mediated, Vitro, Presence, Results, 10, Response, Activation, Type, Inhibition, Transformed, POle, Scattered, Keratinocytes, Skin, Signaling, Between, Effects, C, Kinase, Carcinoma, Gene, Found, Epithelial, Both, Dependent, Expressed, Derived, Receptors, Factors, Proliferation, Demonstrated, Tumors, Activity, May, Increased, Induced, Tumor, Benign, Normal, Like, Levels, Cancer, MRNA, Alpha, Protein, Transforming, Epidermal, TGF, EGF, Receptor, Human, Expression, Cell, Cells, Factor, Growth, Melanomas, SF, Messenger, Thyminidine, ER, Through, Mutations, Basal, RNA, Stimulates, Shown, Altered, S, Differential, Collagen, Soluble, Form, Migration</td>
</tr>
<tr>
<td>Cases</td>
<td>0.59</td>
<td>23</td>
<td>Cases, Androgens, Peptides, Cripto, Whether, Epidermis, Vivo, Survival, Colon, Stage, Prostatic, Lung, Two, Ligands, Independent, 100, Non, Low, Similar, Control, During, DHT, Endogenous, MRNA, Antibody, Mammary, MAB, Veterinary, Small, Differentiation, Apepy, Central, Duct, Addition, Tissue, Early, Ad Malignant, No, Calculated</td>
</tr>
</tbody>
</table>

**ORACLE**

**LIFE SCIENCES USERS GROUP**
Clustering
Clustering
Clustering
Clustering

Cluster 17 :: TGFbeta1 (cluster quality: 0.60; document score: 0.73)

TGFbeta1, Stromal, Endothelial, Melanocytes, Vivo, Metalloproteinase, Signal, MMP, Two, Induction, Independent, System, Tumour, Transfected, Yeast, Antibody, Transduction, Phosphorylation, Studies, Melanoma, Squamous, Addition, Ovarian, HB, Activated, Acid, Cultures, Enhanced, Heparin, Significant, Dic, Secretion, Ligand, Pathway, Observed, Culture, Regulation, Tyrosine, H, Mast, Mediated, Fold, T, Matrix, Vitro, Presence, Level, Line, 10, Reduced, Response, Activation, Type, Breast, Inhibition, Transformed, Stimulated, Between, Effects, D, Kinase, R, Carcinoma, Found, Epithelial, Both, Expressed, EGFR, Receptors, Effect, Lines, Prostate, Proliferation, Tumors, Activity, May, Androgen, Induced, Tumor, Binding, Normal, Cancer, mRNA, Alpha, Protein, Ability, Epidermal, TGF, These, EGF, AR, Receptor, Human, Expression, Cell, Cells, Biological, Factor, Growth, Vascular, ACT, Melanomas, Retinoic, Showed, Exogenous, Inhibitor, Inhibitors, While, Induces, Three, Stimulates, Stimulation, Secreted, Suppression, 9, Differential, Mechanisms, Used, Inhibited, Soluble, 7, Migration

Themes: receptors (53), DECORIN (32), EGF (29), KINASE (22), inducement (21), proteins (20), epidermis (20), United States (2), ligands (15), growth factors (15)

9988678 Decorin is a biological ligand for the epidermal growth factor receptor.

Ectopic expression of decorin induces profound cytostatic effects in transformed cells with diverse histogenetic backgrounds. The mechanism of action has only recently begun to be elucidated. Exogenous decorin activates the epidermal growth factor (EGF) receptor, thereby triggering a signaling cascade that leads to phosphorylation of mitogen-activated protein (MAP) kinase, induction of p21, and growth suppression. In this study we demonstrate a direct interaction of decorin with the EGF receptor. Binding of decorin induces dimerization of the EGF receptor and rapid and sustained phosphorylation of MAP kinase in squamous carcinoma cells. In a cell-free system, decorin induces autophosphorylation of purified EGF receptor by activating the receptor tyrosine kinase and can also act as a substrate for the EGF receptor kinase itself. Using radioligand binding assays we show that both immobilized and soluble decorin bind to the EGF receptor ectodomain or to purified EGF receptor. The binding is mediated by the protein core and has relatively low affinity (Kd approximately 87 nM). Thus, decorin should be considered as a novel biological ligand for the EGF receptor, an interaction that could regulate cell growth during remodeling and cancer growth.
Clustering Medline Abstracts
Comparing Document Collections

[Diagram of document processing workflow]
Analysis by Scatter Plot
InforSense Data Model Facilitates Cross-domain Data Analysis

Data Mining

Microarray Analysis

Text Selection

Metabonomic Analysis

Spectrum Data Mining

Text Mining

Domain Maps

Domain Maps

Chemical/sequence Data Model