Language Models for Genomics Information Retrieval

UIUC at TREC 2007 Genomics Track

Yue Lu, Jing Jiang, Xu Ling, Xin He, ChengXiang Zhai
University of Illinois at Urbana-Champaign
Goal of Participation

• Apply language models to genomics retrieval
• Extend standard language models for
  – gene synonym expansion
  – conjunctive query interpretation
• Experiment with relevance feedback
**Basic Framework**

1. **Pre-processing**
   - Medline articles
   - Candidate passages

2. **Retrieval**
   - Conjunction Query
   - Synonym Expansion

3. **Passage Post-processing**
   - Final output

**Our Focus**

**Query**
- Query \( Q \)

**Basic Framework**

- **Pre-processing**
  - Medline articles
  - Candidate passages

- **Retrieval**
  - Conjunctive Query
  - Synonym Expansion

- **Passage Post-processing**
  - Final output

**Our Focus**
Gene Synonym Expansion

“What [MOLECULAR FUNCTIONS] is LITAF involved in?”

Synonyms:
- lps-induced tn factor
- tbx 1

How to distinguish good synonyms from bad ones?
How to assign weights?
Overlap-Based Synonym Weighting

Q = “What [MOLECULAR FUNCTIONS] is LITAF involved in?”

Q1 = “… lps-induced tn factor…”

Qi = “… tbx 1…”

New Score for p

\[
s(p) = \text{MAX}\{ w_1 \cdot s(p;Q_1) \}
\]

\[
s(p;Q) \land \text{MAX}\{ w_i \cdot s(p;Q_i) \}
\]
Conjunctive Query Interpretation

“What [MOLECULAR FUNCTIONS] is LITAF involved in?”

p1 = “LITAF …involve … LITAF… involved … LITAF …”  
   Missing “Molecular Function”

p2 = “… LITAF … involve … molecular function …”  
   Match all query terms
KL-Divergence Retrieval Model

\[ D(\theta_Q \| \theta_D) = \sum_{w \in V} p(w | \theta_Q) \log \frac{p(w | \theta_Q)}{p(w | \theta_D)} \]

Query

\( Q \)

molecular 0.25
functions 0.25
LITAF 0.25
involved 0.25

Query LM

Document LM

Passage

\( p \)

the 0.120
for 0.085
involve 0.068
LITAF 0.052
function 0.034
molecular 0.034
… …

Background

\( B \)

the 0.210
a 0.181
for 0.085
function 0.034
involve 0.028

Dirichlet Smoothing
Conjunctive Scoring in LM
Experiments

Query Q

Medline articles

Candidate passages

Retrieval

Synonym Expansion
Conjunctive Query

Relevance Feedback

ranked passages

pseudo feedback
## Gene Synonym Expansion

<table>
<thead>
<tr>
<th>Method</th>
<th>DocMAP</th>
<th>Psg2MAP</th>
</tr>
</thead>
<tbody>
<tr>
<td>No expansion</td>
<td>Baseline1</td>
<td>0.1777</td>
</tr>
<tr>
<td>Gene Synonym Expansion</td>
<td>UIUCsyn</td>
<td>0.1926</td>
</tr>
</tbody>
</table>
## Gene Synonym Expansion

<table>
<thead>
<tr>
<th>Method</th>
<th>DocMAP</th>
<th>Psg2MAP</th>
</tr>
</thead>
<tbody>
<tr>
<td>No expansion</td>
<td>0.1777</td>
<td>0.0391</td>
</tr>
<tr>
<td>Gene Synonym Expansion</td>
<td>0.1926</td>
<td>0.0392</td>
</tr>
<tr>
<td>Improvement over Baseline1</td>
<td>+8.38%</td>
<td></td>
</tr>
</tbody>
</table>

**Improvement over Baseline1:** +8.38%
## Gene Synonym Expansion

<table>
<thead>
<tr>
<th>Method</th>
<th>DocMAP</th>
<th>Psg2MAP</th>
</tr>
</thead>
<tbody>
<tr>
<td>No expansion</td>
<td>0.1777</td>
<td>0.0391</td>
</tr>
<tr>
<td>Gene Synonym Expansion</td>
<td>0.1926</td>
<td>0.0392</td>
</tr>
<tr>
<td>Improvement over Baseline1</td>
<td>≈0</td>
<td></td>
</tr>
</tbody>
</table>

Improvement over Baseline 1: 0.0391 vs 0.0392
UIUCsyn improves DocMAP on many topics
UIUCsyn decreases DocMAP on a few topics
UIUCsyn improves Psg2MAP on some topics
UIUCsyn decreases Psg2MAP on some topics
Conjunctive Query Interpretation

<table>
<thead>
<tr>
<th>Method</th>
<th>DocMAP</th>
<th>Psg2MAP</th>
</tr>
</thead>
<tbody>
<tr>
<td>Std KL-Div.+fb</td>
<td>0.1918</td>
<td>0.0422</td>
</tr>
<tr>
<td>Official run</td>
<td>0.1495</td>
<td>0.0296</td>
</tr>
<tr>
<td>Strict Conj. Boolean</td>
<td>UIUCconj2</td>
<td>0.1688</td>
</tr>
<tr>
<td>Partly discount IDF</td>
<td>UIUCconj3</td>
<td>0.1932</td>
</tr>
<tr>
<td>Partly discount TF</td>
<td>UIUCconj4</td>
<td>0.1931</td>
</tr>
</tbody>
</table>

The best performing method is **Partly discount TF** with a DocMAP of 0.1931 and a Psg2MAP of 0.0423.
KL-Divergence Retrieval Model

Query

```
θ_Q
```

Document LM

```
θ_D
```

```
D(θ_Q || θ_D)
```

```
= \sum_{w \in V} p(w | θ_Q) \log \frac{p(w | θ_Q)}{p(w | θ_D)}
```

Passage

```
the 0.120
for 0.085
involve 0.068
LITAF 0.052
function 0.034
molecular 0.034
… ...
```

Background

```
\oplus
```

```
B
```

```
μ
```

Dirichlet Smoothing

```

the 0.210
a 0.181
for 0.085
function 0.034
involve 0.028
```
μ - Dirichlet Smoothing

Performance of our best official run

DocMAP

Psg2MAP

Std KL-Div. + FB

Std KL-Div.
Conjunctive Scoring over Optimum $\mu$

Optimum: $\mu=1000$

<table>
<thead>
<tr>
<th>Method</th>
<th>DocMAP</th>
<th>Psg2MAP</th>
</tr>
</thead>
<tbody>
<tr>
<td>Std KL-Div.+fb</td>
<td>0.2598</td>
<td>0.0570</td>
</tr>
<tr>
<td>Baseline2</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Partly discount IDF</td>
<td></td>
<td></td>
</tr>
<tr>
<td>UIUCconj3</td>
<td>0.2660</td>
<td>0.0680</td>
</tr>
<tr>
<td>Improvement</td>
<td>+2.4%</td>
<td>19.3%</td>
</tr>
</tbody>
</table>
Both feedback methods improve DocMAP, but NOT necessarily Psg2MAP
Conclusions and Future Work

• Standard KL-Div. retrieval method are effective but also sensitive to Dirichlet smoothing $\mu$

• Conjunctive scoring improves performance based on optimum $\mu$

• Synonym expansion and User relevance feedback tend to improve DocMAP but not Psg2MAP

• Future work
  – Automatically set optimum Dirichlet smoothing
  – More aggressive synonym expansion
Questions?