Mining the BioMedical Literature

An Introduction

Hagit Shatkay

© Hagit Shatkay, 2002/3, All Rights Reserved
Literature can be used to *explain* and *predict*:

**Pathways:**

Correlation in expression level of genes/proteins:

http://www.ana.ed.ac.uk/rnusse/wntwindow.html

Eisen et al. PNAS, 95;25, 1998

---

**gene ↔ symptom ↔ disease relationships**

TNFRSF1B

Insulin Resistance

Type 2 diabetes

© Hagit Shatkay, 2002/3, All Rights Reserved
Overview

- Text, sources and methods
  - NLP
  - Information Extraction
  - Information Retrieval

- Applications in Bio-Medical Literature
- Functional relations among genes through IR
- Conclusion
Text Sources

- Full text documents (e.g. Elsevier, Nature)
- Text annotations (e.g. Swiss-prot, GeneCards)
- Internal documents, patent information
- Ontologies (e.g. GO, UMLS, MeSH)
Disciplines Handling Text

- Natural Language Processing (NLP)
- Information Extraction
- Information Retrieval
• All aspects of automated natural-language communication: Processing and understanding spoken, handwritten and printed language.

• Stages in natural language communication:
  Speaker: \textit{Intention, Generation, Synthesis}
  Hearer: \textit{Perception, Analysis, Disambiguation, Incorporation}

• Stages relevant to mining on-line text:
  1. \textit{Analysis}: Part of speech tagging, Parsing, Semantic Interpretation.
  2. \textit{Disambiguation}.

[Russell\&Norvig95] Ch. 22,23, and references therein.
[Charniak93, Allen95, Mann\&Schutze99]
Information Extraction

Automated identification of certain kinds of facts from text. Typically, populating a relational table from the identified facts.

Example:

<table>
<thead>
<tr>
<th>Kinase</th>
<th>Protein</th>
</tr>
</thead>
<tbody>
<tr>
<td>ABC-kinase</td>
<td>X</td>
</tr>
</tbody>
</table>

[Cowie&Lehnert96, Cardie97]

© Hagit Shatkay, 2002/3, All Rights Reserved
Information Extraction

Automated identification of certain kinds of facts from text. Typically, populating a relational table from the identified facts.

Example:

<table>
<thead>
<tr>
<th>Kinase</th>
<th>Protein</th>
</tr>
</thead>
<tbody>
<tr>
<td>ABC-kinase</td>
<td>X</td>
</tr>
</tbody>
</table>

“…BES1 is phosphorylated and appears to be destabilized by the glycogen synthase kinase-3 (GSK-3) BIN2…”

[Cowie&Lehnert96, Cardie97]
Information Extraction

Automated identification of certain kinds of facts from text. Typically, populating a relational table from the identified facts.

Example:

<table>
<thead>
<tr>
<th>Kinase</th>
<th>Protein</th>
</tr>
</thead>
<tbody>
<tr>
<td>ABC-kinase</td>
<td>X</td>
</tr>
</tbody>
</table>

“…BES1 is phosphorylated and appears to be destabilized by the glycogen synthase kinase-3 (GSK-3) BIN2…”

[Cowie&Lehnert96, Cardie97]
Information Extraction

Automated identification of certain kinds of facts from text. Typically, populating a relational table from the identified facts.

Example:

### Phosphorylation

<table>
<thead>
<tr>
<th>Kinase</th>
<th>Protein</th>
</tr>
</thead>
<tbody>
<tr>
<td><em>ABC-kinase</em></td>
<td><em>X</em></td>
</tr>
</tbody>
</table>

“... *BES1* is phosphorylated and appears to be destabilized by the glycogen synthase kinase-3 (GSK-3) *BIN2*...”

[Cowie&Lehnert96, Cardie97]

---

© Hagit Shatkay, 2002/3, All Rights Reserved
Information Extraction (cont.)

What it takes:

- Identify the relevant sentences
- Parse to extract the relationships
- Obtain domain-specific information
- Assume “well-behaved” fact sentences

However:

- Missing an ill-formed fact is acceptable if the database is big and redundant.
- Using co-occurrence relationships alone does not require parsing or good fact-structure.
Information Retrieval

Setting:
- A lot of documents
- Specific information need about a subject (posed as a query)

Goal:
Retrieve (automatically) exactly those documents satisfying the information need.

Means:
- Boolean query, Index based (e.g. “Gene and CDC”)
  - Polysemy (Not interested in “Center for Disease Control”)
  - Synonymy (PR55, won’t be retrieved)
- Similarity query, Vector based.
- Categorization.
Information Retrieval (cont.)

The Vector Model

- **DB**: Database of documents.
- **Vocabulary**: \(\{t_1, \ldots, t_M\}\) \(\{\text{Terms in DB}\}\)
- **Document** \(d \in DB\): Vector, \(<w_1^d, \ldots, w_M^d>\), of weights.

Some Weighting Schemes:

\[
W_i^d = \begin{cases} 
1 & \text{if } t_i \in d \\
0 & \text{otherwise} 
\end{cases}
\]

Binary

\[
W_i^d = f_i^d = \text{# of times } t_i \text{ occurs in } d
\]

TF

\[
W_i^d = \frac{f_i^d}{f_i} \quad (\text{where } f_i = \text{# of docs containing } t_i)
\]

TF X IDF

(one version...)

**Example:**

Medium body beans have made Colombia famous for its flavorful coffee with a slightly dry acidity.

\(<..., \text{bean, beer, cat, coffee, colombia, ...} >\)

\(<..., 1, 0, 0, 1, 1, ..., >\)
Information Retrieval (cont.)

Vector-Based similarity

Document \( d = \langle w_1^d, \ldots, w_M^d \rangle \in DB \)

Query \( q = \langle w_1^q, \ldots, w_M^q \rangle \) (\( q \) could itself be a document in DB...)

\[
\text{Sim}(q,d) = \text{cosine}(q,d) = \frac{q \cdot d}{|q||d|}
\]
Information Retrieval (cont.)

Vector-Based similarity

Document $d = < w_1^d, ..., w_M^d > \in DB$

Query $q = < w_1^q, ..., w_M^q >$ (q could itself be a document in DB...)

$$\text{Sim}(q, d) = \cosine(q, d) = \frac{q \cdot d}{|q||d|}$$
Information Retrieval (cont.)

Vector-Based similarity

Document \( d = <w_1^d, \ldots, w_M^d> \in DB \)
Query \( q = <w_1^q, \ldots, w_M^q> \) (q could itself be a document in DB...)

\[
\text{Sim}(q, d) = \cosine(q, d) = \frac{q \cdot d}{||q|| ||d||}
\]
Information Retrieval (cont.)

Vector-Based similarity

**Document** $d = \langle w_1^d, \ldots, w_M^d \rangle \in DB$

**Query** $q = \langle w_1^q, \ldots, w_M^q \rangle$ (q could itself be a document in DB...)

$$\text{Sim}(q, d) = \cosine(q, d) = \frac{q \cdot d}{|q||d|}$$

Probabilistic similarity measure:

$d = \langle w_1^d, \ldots, w_M^d \rangle$ is viewed as a probability distribution over terms, (a *language model*).

$q$ is viewed as a sample generated from a distribution.

$$\text{Sim}(q, d) = \Pr(q|d).$$
Information Retrieval (cont.)

Vector-Based similarity

Document \( d = \langle w_1^d, \ldots, w_M^d \rangle \in DB \)
Query \( q = \langle w_1^q, \ldots, w_M^q \rangle \) (q could itself be a document in DB...)

\[
\text{Sim}(q, d) = \cosine(q, d) = \frac{q \cdot d}{|q| |d|}
\]

Probabilistic similarity measure:
\( d = \langle w_1^d, \ldots, w_M^d \rangle \) is viewed as a probability distribution over terms,
(a language model).
\( q \) is viewed as a sample generated from a distribution.
\[
\text{Sim}(q, d) = \Pr(q|d).
\]

[Salton89, Witten et al99] Introductory IR.
Information Retrieval (cont.)

Text Categorization

Placing documents in their “right drawer”, making them easy-to-find for the user.

Either manually by indexers, or automatically, through clustering or classification.
Overview

✔ Text, sources and methods

➡ Applications in the Bio-Medical Literature
  • Functional relations among genes through IR
  • Conclusion
Information Extraction

T. Leek, MSc thesis, 1997 [Leek97]

• **Gene localization on chromosomes.**
  
  
  • *HMMs* characterize sentences describing localization relationships.
  
  • Gene names and Chromosomes are identified through heuristics.
  
  • Words denoting location (e.g. *mapped, located, discovered*) and methods (e.g. *southern, blot*) are pre-defined.
  
  • Trained and tested on sentences from OMIM abstracts.
  
  • Scored based on *correct population* of the relation slots.

---

```
start
method1  verb1  Gene  verb2  Chrom  method2
end
```

“Southern analysis shows that HKE4 is located on human chromosome 6.”
Information Extraction (cont.)

M. Craven et al [Craven & Kumlien99, Ray & Craven01]

- **Protein sub-cellular localization** and **gene-disorder associations**.
  
  Examples: Protein: Enzyme UBC6 *localized to* Endoplasmic Reticulum.

  - Gene: PSEN1 *associated with* Alzheimer Disease

- **HMMs** characterize sentences describing sub-cellular localization, and disease association.
  (Other models for sub-cellular localization in [Craven & Kumlien99])

- HMMs’ states represent *structural segments* (e.g. `NP_segment`)

- Training: Sentences whose gene/protein/location/disease words are tagged, based on information from YPD and OMIM.
  (Protein and localization lexicon was provided in [Craven & Kumlien99])

- Scored on correct classification of relevant/irrelevant sentences.
Information Extraction (cont.)

A. Blaschke, A. Valencia et al, 1999 [Blaschke et al 99]

- **Protein-protein interactions.**
  
  *Example:* Protein: Spatzle *Activates* Protein: Toll

- Based on *co-occurrence* of the form “... *p1*...*I1*...*p2*...” within a sentence, where *p1*, *p2* are *proteins* and *I1* is an *interaction term*.

- Protein names and interaction terms (e.g. *activate, bind, inhibit*) are provided as a “dictionary”.

- Does not use formal modeling or machine-learning.

- Applied to two systems in Drosophila:
  
  The Pelle system (6 proteins, 9 interactions) and
  
  The cell-cycle control system (91 proteins), without quantitative analysis of the results.

© Hagit Shatkay, 2002/3, All Rights Reserved
Information Extraction (cont.)

T. Jenssen, E. Hovig et al, 2001 (PubGene) [Jenssen et al 01]

• **Gene-gene relations.**

  *Example:* ![Gene-gene relations](NR4A2 -- NR4A3)

• Based on *co-occurrence* of the form “... $g1...g2...$” within a Pubmed abstract, where $g1$, $g2$ are *gene names*.

• Gene names are provided as a “dictionary”, harvested from *HUGO*, *LocusLink*, and other sources.

• Does not use formal modeling or machine-learning.

• Applied to 13,712 named human genes and millions of PubMed abstracts (Most extensive!)

• No extensive quantitative results analysis.

[Pearson01] Discussion of this approach.
Information Extraction (cont.)

Summary

• Extracts specific kinds of stated entity-entity relations.

• Identifies relations by either:
  - Syntactic/semantic model for “relational” sentence (NLP),
  - Co-occurrence of entity/relation terms.

• Requires a dictionary of entities and relations terms, or well-defined rules for identifying them.

[Fukuda et al 98, Rindflesch et al 00, Friedman et al 01]

PSB proceedings (2000 and on).
The latter is hard to satisfy, as nomenclature for genes/proteins suffers from:

1. Non-uniformity, Incompleteness
2. Synonymy/Aliases (AGP1, aka, Amino Acid Permease1)
3. Polysemy: What might be a gene name in one context is a storage device in another…

Genes A and B may share a similar function but not explicitly related to each other through a publication.
Information Retrieval

Finding functional relations among genes [Shatkay et al 00,02]

Functionally Related Genes

Similar bodies of relevant literature

Per Kernel:
Find related docs. + descriptive terms

Information Retrieval Engine

Bodies of Literature (+ Term Sets.)

Find Similarity among Bodies of Literature

Sets of Inter-related Genes

© Hagit Shatkay, 2002/3, All Rights Reserved
Overview

✓ Text, sources and methods
✓ Applications in the Bio-Medical Literature
→ Functional relations among genes through IR
   – The Information Retrieval Model
   – Themes and How to find them
   – From Documents to Genes and Back
   – Experiments and Results
• Conclusion
The IR Model Used [Shatkay&Wilbur00]

- **Collection** → **Set** of documents from one **broad domain** (e.g. yeast genes, AIDS, food…)
- **Information need** → Documents with a unifying **theme**
- **Query** → Example document
- **Means** → Probabilistic Similarity Search

Documents are generated by a **hidden, stochastic** process
Modeling Themes

**Theme T:**

A collection of documents discussing a common (specific) subject.

Characterized by a family of \( M \) Bernoulli Distributions, \( Pr(t_i \in d | d \in T) \).

**Term Distribution for the theme Coffee**

![Graph showing term distribution for coffee]

© Hagit Shatkay, 2002/3, All Rights Reserved
Modeling Themes (cont.)

Model Components for theme \( T \):

- \( \Pr(\text{doc. } d \text{ is in theme } T) \): \( P_d \overset{\text{def}}{=} \Pr(d \in T) \).
- \( \Pr(\text{Term } t_i \text{ to occur in a theme doc. } d) \): \( P_i \overset{\text{def}}{=} \Pr(t_i \in d \mid d \in T) \).
- \( \Pr(\text{Term } t_i \text{ to occur in an off-theme doc. } d) \): \( Q_i \overset{\text{def}}{=} \Pr(t_i \in d \mid d \not\in T) \).
- \( \Pr(\text{Term } t_i \text{ to occur in any doc. } d) \): \( \text{DB}_i \overset{\text{def}}{=} \Pr(t_i \in d) \).
  \( \text{DB}_i \approx (\text{# of docs containing } t_i) / |DB| \)
- \( \Pr(\text{Term } t_i \text{ to be generated according to } \text{DB}_i) \): \( \lambda_i \)

\[
\text{Model } R = \{ P_d, \{ P_i \}, \{ Q_i \}, \{ \text{DB}_i \}, \{ \lambda_i \} \}
\]
Stochastic Generation of document $d$:

For each term, $t_i$:

- Theme / Off-Theme?
- From DB or P/Q?
- Include $t_i$ in $d$?
GenTheme: Finding Themes

Task: Starting from a single document $d$, find a theme model $R$, that maximizes the likelihood, $Pr(DB|R)$.

Input: Kernel $d$ (PubMed ID), and document collection $DB$

Output:
- Top (10) documents, with highest theme probability, $Pr(d \in T| d, DB, R)$
- Top (10) key terms, with highest ratio $Pr(t_i \in d|d \in T)/Pr(t_i \in d|d \not\in T)$.

Method: Expectation Maximization (EM)
GenTheme in context: (From Genes to Documents)

Kernels → GenTheme → Documents discussing g → Terms summarizing g’s function

Mapping → PubMed queries

G

DB
Broad Domain

For each g ∈ G
Finding Related Genes
(From Documents back to Genes)

**Similar** document sets represent **related** genes.

Thus, Find **similar** sets of documents
Similarity between sets of documents

• Space \( S \) of relevant documents:
  \[
  S \approx \bigcup_i S_i = \{ID_1, \ldots, ID_{M_r}\} \quad |S| = M_r
  \]

• Represent each set \( S_i \), (kernel \( k_i \)) as \( M_r \)-dimensional vector:
  \[
  \langle v_i^1, \ldots, v_i^{M_r}\rangle
  \]
  \[
  v_i^j = \begin{cases} 
  1 & \text{if } ID_j \in S_i \\
  0 & \text{otherwise}
  \end{cases}
  \]

• Similarity between two vectors: The cosine of the angle between them.

Set, \( S_i \rightarrow \) similar sets \( \{S_i^1, \ldots, S_i^p\} \)
Kernel, \( k_i \rightarrow \) similar kernels \( \{k_i^1, \ldots, k_i^p\} \)
Gene, \( g_i \rightarrow \) related genes \( \{g_i^1, \ldots, g_i^p\} \)
From Genes to Documents and Back

Summary

**genes**  **kernels**
Experiments and Results

**Domain:** Yeast genes

- Accessible information sources (*SGD, YPD*)
- Grouping and functional analysis presented by *Spellman et al [1998]*
- Hence, the quality of our results is easily verifiable
Experimental Setting

- **Kernel documents**: 344 PubMed abstracts. (Curated references from SGD as representatives for 408 cell-cycle regulated genes).
- **Database**: 33,700 Pubmed abstracts, generated through iteratively neighboring the 344 abstracts.
- **GenTheme algorithm**: Produce a theme for each kernel
- **Cosine-based method**: Find groups of related genes based on related kernels
- Compare the results with the functionality of genes according to Spellman et al. [Spellman et al 98].

<table>
<thead>
<tr>
<th>Biological Function</th>
<th>G1</th>
<th>S</th>
<th>G2</th>
<th>M</th>
<th>M/G1</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Replication Initiation</strong></td>
<td>CDC45</td>
<td>ORC1</td>
<td></td>
<td>CDC47 CDC54 <strong>MCM2</strong> MCM6</td>
<td>CDC6 CDC46 MCM3</td>
</tr>
<tr>
<td><strong>Fatty Acids/Lipids/Sterols/Membranes</strong></td>
<td>EPT1 LPP1 PSD1 SUR1 SUR2 SUR4</td>
<td>AUR1 ERG3 LCB3</td>
<td><strong>ERG2 ERG5 PMA1 PMA2 PMP1</strong></td>
<td>ElO1 FAA1 FAA3 FAA4 FAS1</td>
<td></td>
</tr>
<tr>
<td><strong>Nutrition</strong></td>
<td>BAT2 PHO2</td>
<td><strong>AGP1 BAT1 GAP1</strong></td>
<td>DIP5 FET3 FTR1 MEP3 PFK1 PHO3 PHO5 PHO11 PHO12 PHO84 RGT2 SUC2 SUT1 VAO1 VCX1 ZRT1</td>
<td>AUA1 GLK1 HXT1 HXT2 HXT4 HXT7</td>
<td></td>
</tr>
</tbody>
</table>
# 1. Qualitative Evaluation

Kernel PMID: 8702485, Gene: ELO1, Function: Fatty Acids/Lipids/Sterols/Membranes

<table>
<thead>
<tr>
<th>Keywords</th>
<th>Genes</th>
<th>Function</th>
</tr>
</thead>
<tbody>
<tr>
<td>Fatty acid</td>
<td>OLE1</td>
<td><em>Fatty Acid, Sterol Metabolism</em></td>
</tr>
<tr>
<td>Fatty</td>
<td>FAA4</td>
<td>FA/Lipids/Sterols/Membranes</td>
</tr>
<tr>
<td>Lipids</td>
<td>FAA3</td>
<td>FA/Lipids/Sterols/Membranes</td>
</tr>
<tr>
<td>Acid</td>
<td>SUR2</td>
<td>FA/Lipids/Sterols/Membranes</td>
</tr>
<tr>
<td>Carbon, Grown,</td>
<td>ERG2</td>
<td>FA/Lipids/Sterols/Membranes</td>
</tr>
<tr>
<td>Medium, Synthase,</td>
<td>FAA1</td>
<td>FA/Lipids/Sterols/Membranes</td>
</tr>
<tr>
<td>Strains, Deficient</td>
<td>PSD1</td>
<td>FA/Lipids/Sterols/Membranes</td>
</tr>
<tr>
<td></td>
<td>CYB5</td>
<td><em>Fatty Acid, Sterol Metabolism</em></td>
</tr>
<tr>
<td></td>
<td>PGM1</td>
<td><em>Carbohydrate Metabolism</em></td>
</tr>
</tbody>
</table>
Another Example:

Starting point: **2233722** (Curated in SGD for HXT2):

Top ranking docs: **10336421, 7862149, 8594329, 2660462**, …

Top ranking terms: *glucose transport, glucose, transporter*, etc.

Related Genes:

![Gene network diagram]

- Spellman Verified
- YPD Verified
- Spurious Relation
K. PMID: 6323245, Gene MCM2, Function: Replication Initiation

<table>
<thead>
<tr>
<th>Keywords</th>
<th>Genes</th>
<th>Function</th>
</tr>
</thead>
<tbody>
<tr>
<td>ARS, Autonom. replicating,</td>
<td>CDC10</td>
<td>Site selection, Morphogenesis</td>
</tr>
<tr>
<td>Replicating sequence,</td>
<td>PHO3</td>
<td>Nutrition</td>
</tr>
<tr>
<td>Autonomously,</td>
<td>EST1</td>
<td>DNA Synthesis</td>
</tr>
<tr>
<td>Minichromosomes,</td>
<td>MIF2</td>
<td>Chromatin</td>
</tr>
<tr>
<td>Replicating,</td>
<td>PHO12</td>
<td>Nutrition</td>
</tr>
<tr>
<td>Centrometric, leu2, Plasmids,</td>
<td>POL3</td>
<td>DNA Synthesis</td>
</tr>
<tr>
<td>ura3</td>
<td>DHS1</td>
<td>DNA Repair</td>
</tr>
<tr>
<td></td>
<td>SNQ2</td>
<td>DNA Repair</td>
</tr>
<tr>
<td></td>
<td>SMC3</td>
<td>Chromatin</td>
</tr>
<tr>
<td></td>
<td>EXG2</td>
<td>Cell Wall Synthesis</td>
</tr>
</tbody>
</table>

© Hagit Shatkay, 2002/3, All Rights Reserved
2. Quantitative Evaluation of Summaries

- **Kernels:** 105 abstracts discussing the biological function of distinct yeast genes.

- **Expert Thesaurus Construction:**
  - List 5 top ranking terms for each of the 105 kernels. 330 Terms, alphabetically sorted.
  - 4 Independent judges assign terms to one of 22 functional categories (+ “uninformative”).

![Diagram showing terms and judges]

Terms:
- chromatin
- fatty
- telomere
- acid

Judges:
- 1
- 2
- 22
- U

© Hagit Shatkay, 2002/3, All Rights Reserved
2. Quantitative Evaluation (cont.)

Expert Thesaurus - Example:

<table>
<thead>
<tr>
<th>Function</th>
<th>Associated Terms</th>
</tr>
</thead>
<tbody>
<tr>
<td>Chromatin</td>
<td>Chromatids, Chromatin, Chromosome, sister Chromatids, telomere, telomeric</td>
</tr>
<tr>
<td>Secretion</td>
<td>Acid phosphatase, coatomer, endoplasmic, endoplasmic reticulum, er, golgi apparatus, golgi complex, golgi transport, golgi</td>
</tr>
</tbody>
</table>

Count how many of the top 5 summary terms are assigned to a thesaurus entry matching the function discussed in the query document; *average over 105 queries*.

<table>
<thead>
<tr>
<th>Correct</th>
<th>Wrong</th>
<th>Don’t Care</th>
</tr>
</thead>
<tbody>
<tr>
<td>3.27 terms</td>
<td>1.12 terms</td>
<td>0.61 terms</td>
</tr>
</tbody>
</table>
General

- Similarity Search + EM: Independent of explicit query terms (synonymy, polysemy).
- Provides a key-term list, justifying the relevance of retrieved documents.

Application Specific

- Independent of explicit gene/protein names, makes no assumption about standard nomenclature.
- Independent of sentence structures (e.g. “A interacts with B”).
- Can foreshadow putative, undiscovered relationships.
Results Summary

Starting from \textit{informative kernel documents} our method:

- ☺ Provides an efficient way for establishing \textit{putative functional relationships among genes}
- ☺ Provides references to the relevant literature
- ☺ Generates a \textit{summary} explaining the discovered relationships (complements direct analysis methods)

- 😞 Performance depends on \textit{informative kernels}.
  \textbf{Challenge:} Automate the kernel-picking process.
  (At times: compose your own kernel and use it…)

© Hagit Shatkay, 2002/3, All Rights Reserved
Acknowledgements

**Collaborators**
Stephen Edwards (Rosetta)
Mark Boguski (Fred Hutchinson Cancer Research Center)
John Wilbur (NCBI)

**Yeast Experts (NCBI):**
Jan Fassler
Ken Katz
Steven Sullivan
Tyra Wolfsberg

**EM Discussions:**
Luis Ortiz (Brown, UPenn)
Conclusion

Information Extraction:
• Extracts well-defined facts from the literature.
• Requires domain vocabulary or rules to identify these facts.
• Finds explicitly stated facts.
• Looks for facts stated within a sentence, a paragraph or a single document (Fine granularity)

Information Retrieval:
• Looks for relevant documents.
• Does not give a tidy fact statement (Coarse granularity)
• Can find relations among documents or document collections.
• Can create a coherent context for performing Information Extraction.
• Can foreshadow putative, yet-undiscovered relationships.
• Less sensitive to vocabulary and terminology.
Conclusion (cont.)

Challenges:

- Reduce dependency on vocabularies and nomenclature.
- Automate fact-finding about gene-disease interaction.
- Reconstruct metabolic, signaling or regulatory pathways.
- Augment analysis of large-scale experiments with data from the literature. (e.g. [Chang et al 01]).
- Establish evaluation standards for evaluating the utility of literature mining tools.

No single method can address all the needs.

A combined approach is likely to get us closer to our goal.
Bibliography


**Bibliography** (cont.)


**Bibliography (cont.)**


The HXT2 gene of Saccharomyces cerevisiae is required for high-affinity glucose transport.

Kruckeberg AL, Bisson LF.

Department of Viticulture and Enology, University of California, Davis 95616. The HXT2 gene of the yeast Saccharomyces cerevisiae was identified on the basis of its ability to complement the defect in glucose transport of a snf3 mutant when present on the multicopy plasmid pSC2. Analysis of the DNA sequence of HXT2 revealed an open reading frame of 541 codons, capable of encoding a protein of Mr 59,840. The predicted protein displayed high sequence and structural homology to a large family of procaryotic and eucaryotic sugar transporters. These proteins have 12 highly hydrophobic regions that could form transmembrane domains; the spacing of these putative transmembrane domains is also highly conserved. Several amino acid motifs characteristic of this sugar transporter family are also present in the HXT2 protein. An hxt2 null mutant strain lacked a significant component of high-affinity glucose transport when under derepressing (low-glucose) conditions. However, the hxt2 null mutation did not incur a major growth defect on glucose-containing media. Genetic and biochemical analyses suggest that wild-type levels of high-affinity glucose transport require the products of both the HXT2 and SNF3 genes; these genes are not linked. Low-stringency Southern blot analysis revealed a number of other sequences that cross-hybridize with HXT2, suggesting that S. cerevisiae possesses a large family of sugar transporter genes.

PMID: 2233722 [PubMed - indexed for MEDLINE]
Glucose uptake kinetics and transcription of HXT genes in chemostat cultures of Saccharomyces cerevisiae.


Slater Institute, University of Amsterdam, Plantage Muidergracht 12, 1018 TV Amsterdam, The Netherlands.

The kinetics of glucose transport and the transcription of all 20 members of the HXT hexose transporter gene family were studied in relation to the steady state in situ carbon metabolism of Saccharomyces cerevisiae CEN.PK113-7D grown in chemostat cultures. Cells were cultivated at a dilution rate of 0.10 h-1 under various nutrient-limited conditions (anaerobically glucose- or nitrogen-limited or aerobically glucose-, galactose-, fructose-, ethanol-, or nitrogen-limited), or at dilution rates ranging between 0.05 and 0.38 h-1 in aerobic glucose-limited cultures. Transcription of HXT1-HXT7 was correlated with the extracellular glucose concentration in the cultures. Transcription of GAL2, encoding the galactose transporter, was only detected in galactose-limited cultures. SNF3 and RGT2, two members of the HXT family that encode glucose sensors, were transcribed at low levels. HXT8-HXT17 transcripts were detected at very low levels. A consistent relationship was observed between the expression of individual HXT genes and the glucose transport kinetics determined from zero-trans influx of 14C-glucose during 5 s. This relationship was in broad agreement with the transport kinetics of Hxt1-Hxt7 and Gal2 deduced in previous studies on single-HXT strains. At lower dilution rates the glucose transport capacity estimated from zero-trans influx experiments and the residual glucose concentration exceeded the measured in situ glucose consumption rate. At high dilution rates, however, the estimated glucose transport capacity was too low to account for the in situ glucose consumption rate.

PMID: 10336421 [PubMed - indexed for MEDLINE]
Three different regulatory mechanisms enable yeast hexose transporter (HXT) genes to be induced by different levels of glucose.

Ozcan S, Johnston M.

Department of Genetics, Washington University School of Medicine, St. Louis, Missouri 63110.

The HXT genes (*HXT1* to *HXT4*) of the yeast *Saccharomyces cerevisiae* encode hexose transporters. We found that transcription of these genes is induced 10- to 300-fold by glucose. Analysis of glucose induction of HXT gene expression revealed three types of regulation: (i) induction by glucose independent of sugar concentration (*HXT3*); (ii) induction by low levels of glucose and repression at high glucose concentrations (*HXT2* and *HXT4*); and (iii) induction only at high glucose concentrations (*HXT1*). The lack of expression of all four HXT genes in the absence of glucose is due to a repression mechanism that requires Rgt1p and Ssn6p. *GRR1* seems to encode a positive regulator of HXT expression, since grr1 mutants are defective in glucose induction of all four HXT genes. Mutations in *RGT1* suppress the defect in HXT expression caused by grr1 mutations, leading us to propose that glucose induces HXT expression by activating Grr1p, which inhibits the function of the Rgt1p repressor. HXT1 expression is also induced by high glucose levels through another regulatory mechanism: rgt1 mutants still require high levels of glucose for maximal induction of HXT1 expression. The lack of induction of HXT2 and HXT4 expression on high levels of glucose is due to glucose repression: these genes become induced at high glucose concentrations in glucose repression mutants (*hxk2, reg1, ssn6, tup1, or mig1*). Components of the glucose repression pathway (*Hxk2p* and *Reg1p*) are also required for generation of the high-level glucose induction signal for expression of the HXT1 gene. Thus, the glucose repression and glucose induction mechanisms share some of the same components and may share the same primary signal generated from glucose.

PMID: 7862149 [PubMed - indexed for MEDLINE]
Characterization of AGT1 encoding a general alpha-glucoside transporter from Saccharomyces.

Han EK, Cotty F, Sottas C, Jiang H, Michels CA.

Queens College, Flushing, New York, USA.

Molecular genetic analysis is used to characterize the **AGT1 gene** encoding an alpha-glucoside transporter. **AGT1** is found in many Saccharomyces cerevisiae laboratory strains and maps to a naturally occurring, partially functional allele of the **MAL1** locus. Agt1p is a highly hydrophobic, postulated integral membrane protein. It is 57% identical to **Mal61p**, the maltose permease encoded at **MAL6**, and is also a member of the 12 transmembrane domain superfamily of sugar transporters. Like Mal61p, Agt1p is a high-affinity, maltose/proton symporter, but Mal61p is capable of transporting only maltose and turanose, while Agt1p transports these two alpha-glucosides as well as several others including isomaltose, alpha-methylglucoside, maltotriose, palatinose, trehalose and melezitose. AGT1 expression is maltose inducible and induction is mediated by the Mal-activator. The sequence of the upstream region of AGT1 is identical to that of the maltose-inducible **MAL61** gene over a 469 bp region containing the UASMAL but the 315 bp sequence immediately upstream of AGT1 shows no significant homology to the sequence immediately upstream of **MAL61**. The evolutionary origin of the **MAL1** allele to which **AGT1** maps and the relationship of **AGT1** to other alpha-glucoside fermentation genes is discussed.

PMID: 8594329 [PubMed - indexed for MEDLINE]
Kinetics of growth and glucose transport in glucose-limited chemostat cultures of Saccharomyces cerevisiae CBS 8066.

Postma E, Scheffers WA, van Dijken JP.

Department of Microbiology and Enzymology, Delft University of Technology, the Netherlands.

The glucose transport capacity of Saccharomyces cerevisiae CBS 8066 was studied in aerobic glucose-limited chemostat cultures. Two different transport systems were encountered with affinity constants of 1 and 20 mM, respectively. The capacity of these carriers (Vmax) was dependent on the dilution rate and the residual glucose concentration in the culture. From the residual glucose concentration in the fermenter and the kinetic constants of glucose transport, their in situ contribution to glucose consumption was determined. The sum of these calculated in situ transport rates correlated well with the observed rate of glucose consumption of the culture. The growth kinetics of S. cerevisiae CBS 8066 in glucose-limited cultures were rather peculiar. At low dilution rates, at which glucose was completely respired, the glucose concentration in the fermenter was constant at 110 microM, independent of the glucose concentration in the reservoir. At higher dilution rates, characterized by the occurrence of both respiration and alcoholic fermentation, the residual substrate concentration followed Monod kinetics. In this case, however, the overall affinity constant was dependent on the reservoir glucose concentration.

PMID: 2660462 [PubMed - indexed for MEDLINE]