Bioinformatics Topics

■ Today
  ■ Basic biology
  ■ Why text about biology is special
  ■ Text mining case studies
    ■ Microarray analysis
    ■ Abbreviation finding
    ■ Text-enhanced homology search

■ Next week
  ■ Text mining in biological databases
  ■ KDD cup: Information extraction for bio-journals
  ■ Combining text mining and data mining
Basic Biology
Just Enough Molecular Biology

- **Entropy** (the tendency to disorder) always increases (cf. thermodynamics)
- Living organisms have low entropy compared with things like soil.
- They are relatively orderly...
- The most critical task is to maintain the distinction between inside and outside.
In order to maintain low entropy, living organisms must expend energy to keep things orderly.
The functions of life, therefore, are meant to facilitate the acquisition and orderly expenditure of energy.
Just enough.

- The compartments with low entropy are separated from “the world.” Cells are the smallest unit of such compartments.
- **Bacteria are single-cell organisms.**
- **Humans are multi-cell organisms.**
- Low entropy compartments were difficult to get started *de novo*, and so have found ways to pass on the apparatus necessary to perpetuate themselves.
“Entropy-Fighting Apparatus:”

Tasks

- Gather energy from environment
- Use energy to maintain inside/outside distinction
- Use extra energy to reproduce
- Develop strategies for being successful/efficient at the above tasks
  - develop ways to move around
  - develop signal transduction capabilities (e.g. vision)
  - develop methods for efficient energy capture (e.g. digestion)
  - develop ways to reproduce effectively
In order to accomplish these tasks, living compartments on earth have developed three basic technologies:

0. Ability to separate inside from outside (lipids)

1. Ability to build three-dimensional molecules that assist in the critical functions of life (proteins).

2. Ability to compress the information about how (and when) to build these molecules in a linear code (DNA).
Broad Generalization

1. **Lipid** molecules: create compartments that separate inside/outside.
2. **Protein** molecules: do the work, and their 3D structure is critical.
3. **DNA** molecules: store the information
Bioinformatics Schematic of a Cell

- DNA
- Proteins
- Lipid membrane
Lipids

- **Hydrophilic** (water loving) molecular fragment connected to **hydrophobic** fragment.
- Spontaneously form sheets (lipid bilayers, membranes) with hydrophilic ends on the outside, and hydrophobic ends on the inside.
- Create a very stable separation, not easy to pass through except for water and a few other small atoms/molecules.
Lipid bilayer (hydrophobic in, hydrophilic out)
Basics of Lipid structure

Main goal:
separate aqueous compartments effectively.

From
http://cellbio.utmb.edu/
cellbio/
membrane_intro.htm
Protein molecules begin as a sequence of linked subunits

- These subunits are amino acids (also called residues).
- There are 20 different amino acids with different physical and chemical properties.
- The interaction of these properties allows a chain of the amino acids (up to 1000’s long) to fold into a unique, reproducible 3D shape.
20 Amino Acids

- Common, repeating backbone (blue)
- Unique sidechains (yellow)
Shorthand for Protein Sequence

- Specify the sequence of amino acids:
  - Alanine-Tyrosine-Valine
  - ALA-TYR-VAL
  - A-Y-V
Bioinformatics Schematic of a Cell

Lipid membrane

DNA

Proteins
Human DNA
DNA packs in the nucleus to form chromosome.
DNA uses an alphabet of 4 letters (ATCG), more commonly called bases.

Although the 4 letters have interesting chemical structure, for our purposes they are just information carriers.

Long sequences of these 4 letters are linked together to create **GENES** and **CONTROL INFORMATION**.
DNA is a sequence too

- It also has a common backbone, and then specialized sidechains. But there are only 4 specialized sidechains: Adenine, Cytosine, Guanine and Thymidine = A, C, G, and T.

- A sequence of these subunits is also specified as a string:

  e.g., ACTTAGGACATTTTTTAG

- This is a shorthand for the chemical structure, which is not important right now.
DNA encodes Protein (and RNA)

- Each of the twenty protein amino acids can be specified by 3 consecutive DNA bases.
- The Ribosome “reads” a sequence of DNA bases (three at a time) and creates the corresponding protein chain—which folds itself based on the amino acid properties.
- See:  http://ntri.tamuk.edu/cell/ribosomes.html
- The 64 mappings of 3 bases to 1 amino acid is called the GENETIC CODE and is universal (on earth...).
Genetic Code (T=U here)
(e.g. Tyrosine = UAU or UAC)

<table>
<thead>
<tr>
<th>One-letter code</th>
<th>Amino acid</th>
<th>Three-letter code</th>
<th>Genetic code</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>Alanine</td>
<td>Ala</td>
<td>GC*</td>
</tr>
<tr>
<td>C</td>
<td>Cysteine</td>
<td>Cys</td>
<td>UGU, UGC</td>
</tr>
<tr>
<td>D</td>
<td>Aspartic Acid</td>
<td>Asp</td>
<td>GAA, GAC</td>
</tr>
<tr>
<td>E</td>
<td>Glutamic Acid</td>
<td>Glu</td>
<td>GAA, GAG</td>
</tr>
<tr>
<td>F</td>
<td>Phenylalanine</td>
<td>Phe</td>
<td>UUU, UUC</td>
</tr>
<tr>
<td>G</td>
<td>Glycine</td>
<td>Gly</td>
<td>GG*</td>
</tr>
<tr>
<td>H</td>
<td>Histidine</td>
<td>His</td>
<td>CAU, CAC</td>
</tr>
<tr>
<td>I</td>
<td>Isoleucine</td>
<td>Ile</td>
<td>AUU, AUC, AUA</td>
</tr>
<tr>
<td>K</td>
<td>Lysine</td>
<td>Lys</td>
<td>AAA, AAG</td>
</tr>
<tr>
<td>L</td>
<td>Leucine</td>
<td>Leu</td>
<td>UUA, UUG, CU*</td>
</tr>
<tr>
<td>M</td>
<td>Methionine</td>
<td>Met</td>
<td>AUG</td>
</tr>
<tr>
<td>N</td>
<td>Asparagine</td>
<td>Asn</td>
<td>AAU, AAC</td>
</tr>
<tr>
<td>P</td>
<td>Proline</td>
<td>Pro</td>
<td>CC*</td>
</tr>
<tr>
<td>Q</td>
<td>Glutamine</td>
<td>Gln</td>
<td>CAA, CAG</td>
</tr>
<tr>
<td>R</td>
<td>Arginine</td>
<td>Arg</td>
<td>CG*, AGA, AGG</td>
</tr>
<tr>
<td>S</td>
<td>Serine</td>
<td>Ser</td>
<td>UC*, AGU, AGC</td>
</tr>
<tr>
<td>T</td>
<td>Threonine</td>
<td>Thr</td>
<td>AC*</td>
</tr>
<tr>
<td>V</td>
<td>Valine</td>
<td>Val</td>
<td>GU*</td>
</tr>
<tr>
<td>W</td>
<td>Tryptophan</td>
<td>Trp</td>
<td>UGG</td>
</tr>
<tr>
<td>Y</td>
<td>Tyrosine</td>
<td>Tyr</td>
<td>UAU, UAC</td>
</tr>
</tbody>
</table>
Myoglobin: Gene and Protein

cctgcatatat gtaactaagag gagaacaaca acaatggttct tgtctgaagg
tgaatggcag cctggtttctgc atgttttgggc taaagttgaa gctgacgctcg
ctgggtcatgg tcaggacatcg ttgattcgac tgttcaaaatc tcatccggaa
actctggaaa aattcgatcg tttcaacatc ctgaaaaactg aagctgaatat
gaagacctct ctaagatctga aaaaaacatgg tggttaccggtg ttaactgcccc
taggtgctat ctttaagaaa aaggggcacag tgaagctgta gctcaaaaaccg
cctgcatatat cgtgctgctac taacataaaag atcccgatca aatacctgga
attcatctct gcagctgctac tccatgtttct gcattctaga catccaggta
acgccggtgc tgtaccgttc aaaaagctct ctaagctgta gctcaaaaaccg
cgtaaagata tggctgctaa ctgggttacc aaggttaatg aggtacc

BASE COUNT
155 a 108 c 115 g 129 t

MVLSEGEWQQLLVHWAKVEADVAGHGQDILIRLFKSHPETLEKFDFRFKHLKTEAEMKASED
LKKHGVVLTLGAILKKKGHEAELKPLAQSHATKHKIPKYLEFISEAIIHVLHSRHPG
NFGADAQGAMKALELFKDIAAKYKELGYQG
Why We Care: Diseases

**CFTR.** The gene encoding a chloride ion channel is defective in patients with cystic fibrosis

*IMAGE CREDIT: Q. Alaw姜i, Columbia University, NY, USA. Adapted by K. Sulliff, SCIENCE.*

**OBES.** The obese (Ob) mutation in the mouse provides a useful model system for studying human obesity

*IMAGE CREDIT: Jeff Friedman, Rockefeller University, New York. NY, USA. Reprinted from SCIENCE.*
Genes: Statistics

- The set of all genes required for an organism is the organism’s **GENOME**.
- The human genome has 3,000,000,000 bases divided into 23 linear segments (chromosomes).
- A gene has on average 1340 DNA bases, thus specifying a protein of about 447 amino acids.
- Humans have about 35,000 genes = 40,000,000 DNA bases = 3% of total DNA in genome.
- Humans have another 2,960,000,000 bases for control information. (**e.g. when, where, how long, etc...**)
Main focus used to be

- Sequence analysis (human genome project)
- Structure analysis (what is 3d structure of proteins?)

Increasingly, the focus is:

- Function analysis

This is where text mining can help.
Biological Structure and Function

- Sequence & Structure
  - Precise representation as 1D and 3D objects.
- Function: somewhat fuzzy
  - Often represented as text
What are Functions of Genes?

- **Signal transduction**: sensing a physical signal and turning into a chemical signal.
- **Structural support**: creating the shape and pliability of a cell or set of cells.
- **Enzymatic catalysis**: accelerating chemical transformations otherwise too slow.
- **Transport**: getting things into and out of separated compartments.
What are the Functions of Genes?

- **Movement**: contracting in order to pull things together or push things apart.
- **Transcription control**: deciding when other genes should be turned ON/OFF.
- ** Trafficking**: affecting where different elements end up inside the cell.
Why So Few Human Genes?

- Complexity is not a function of the number of genes.
  - Control information critical.
- Complexity is a function of the number of genes, and mustard weed is more complex than we are.
- Number of genes is not estimated correctly.
How Many Genes Do You Have?

- [link](http://www.ensembl.org/Genesweep/)
- Bet how many human genes there are
- Winner to be decided May 2003?

**Results**

- Bets: 165
- Mean: 61,710
- Lowest: 27,462
- Highest: 153,478
Basic Biology: Summary

- Three “technologies”: lipids, proteins, DNA
- Biology needs text mining / NLP
- Biology is an information-intensive science.
  - A lot of the information is in text.
  - Biology is a natural application area for text mining/processing.
- Function is key for understanding biology.
  - There are formal and precise representations for sequence and structure.
  - Text is still the main representation for function.
Microarray Analysis
Microarrays

- Measure the **expression** of genes
- 2-color arrays compare 2 conditions, **control** and **experimental**
- Upregulated = red, downregulated = green
- **Example Application:** clinical diagnosis
A cDNA Microarray
(Source: C. Benning)
Common Analysis Procedure

- Quality control (did the experiment work?)
- Cropping (select affected genes)
- Clustering (group genes)
- Manual exploration of data
- Sense making
Clustering: Example (Eisen et al.)
Each biologist only know a few genes well.

Wading through search results is tedious and time consuming.

Relating measurements with existing knowledge is a key part of microarray analysis.
Two Approaches

- Cluster on numeric data, then interpret textually
- Cluster on textual data, then interpret numerically
MedMiner:
First Numbers, then Text

- Identify group of genes based on experimental data
- MedMiner
  - Identifies significant keywords
  - Creates a list of relevant contexts
Key words

MedMine
r (Tanabe et al.)
MedMiner (Tanabe et al.), cont.

<table>
<thead>
<tr>
<th>correlation</th>
<th>Link to Abstract</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Overexpression of MDM2 (≥10-fold)</strong> was significantly correlated with adriamycin resistance and decreased duration of CR1. (2000)**</td>
<td>PMID 10637478</td>
</tr>
<tr>
<td>The chi2 test was performed to describe the correlation between the Ki-67 index and p53, MDM2, and p21 protein expression. (1999)</td>
<td>PMID 10632343</td>
</tr>
<tr>
<td><strong>A strong correlation</strong> was observed between the Ki-67 index &gt;10% and both MDM2 and p21 proteins. (1999)**</td>
<td></td>
</tr>
<tr>
<td><strong>Accumulation of p53 and MDM2 overexpression correlated with the grade of malignancy.</strong> (1999)</td>
<td></td>
</tr>
<tr>
<td><strong>No correlation</strong> was found between p53 accumulation and the histopathology of gastric cancer. (1999)</td>
<td></td>
</tr>
<tr>
<td><strong>p53 accumulation and MDM2 overexpression did not correlate with tumor size, nodal status, presence of metastases, age or survival.</strong> (1999)</td>
<td></td>
</tr>
</tbody>
</table>
PubGene: First Text, then Numbers

- **Compile** a list of all genes
- Compute **co-occurrence** of genes in medline articles
- Display **network(s)** of selected genes
- Color-code nodes to indicate degree of **up/downregulation**
Text Cluster Analysis (Jenssen et al.)

1H expression levels

8H expression levels

Highly upregulated at 1H
Why Text about Biology is Special
Biological Terminology: A Challenge

- Large number of entities (genes, proteins etc)
- Evolving field, no widely followed standards for terminology -> Rapid Change, Inconsistency
- Ambiguity: Many (short) terms with multiple meanings (eg, CAN)
- Synonymy: ARA70, ELE1alpha, RFG
- High complexity -> Complex phrases
What are the concepts of interest?

- Genes (D4DR)
- Proteins (hexosaminidase)
- Compounds (acetaminophen)
- Function (lipid metabolism)
- Process (apoptosis = cell death)
- Pathway (Urea cycle)
- Disease (Alzheimer’s)
Complex Phrases

- Characterization of the repressor function of the nuclear orphan receptor retinoid receptor-related testis-associated receptor/germ nuclear factor
## Inconsistency

- No consistency across species

<table>
<thead>
<tr>
<th>Species</th>
<th>Protease</th>
<th>Inhibitor</th>
<th>Signal</th>
</tr>
</thead>
<tbody>
<tr>
<td>Fruit fly</td>
<td>Tolloid</td>
<td>Sog</td>
<td>dpp</td>
</tr>
<tr>
<td>Frog</td>
<td>Xolloid</td>
<td>Chordin</td>
<td>BMP2/BMP4</td>
</tr>
<tr>
<td>Zebrafish</td>
<td>Minifin</td>
<td>Chordino</td>
<td>swirl</td>
</tr>
</tbody>
</table>
Rapid Change

Mouse Genome Nomenclature Events 8/25

In 1 week, 166 events involving change of nomenclature

L. Hirschmann
Abbreviation Mining
(Chang, Schütze & Altman)
Abbreviations in Biology

- Two problems
  - “Coreference”/Synonymy
  - What is PCA an abbreviation for?
  - Ambiguity
    - If PCA has >1 expansions, which is right here?
- Only important concepts are abbreviated.
- Effective way of jump starting terminology acquisition.
Ambiguity Example
PCA has >60 expansions

"p-chloroamphetamine" "p-chloroaniline" "p-coumaric acid"
"p.rothrombin c.omplex a.ctivity" "para-chloramphetamin" "parietal
cell antibodies" "parietal cell autoantibodies" "paroxysmal cerebellar
ataxia" "passive cutaneous anaphylactic" "patient care appraisal"
"patient controlled analgesia" "patient controlled anesthesia" "pca"
"pentachloroanisole" "percent cortical area" "perchloracetic acid"
"perchloric acid" "percutaneous coronary angioplasty" "percutaneous
coronary atherectomy" "pericallosal artery" "peritoneal
carcinomatosis" "peritoneal carcinosis" "personal care attendant"
"phenazine-1-carboxylic acid" "phenylciclopentylacetic acid"
"phenylcyclohexylamine" "physical capacity assessment" "pig coronary
artery" "plate count agar" "pneumococcal capsular antigen" "pole
climbing avoidance" "polyclonal activator" "polyclonal antibody"
"polyclonal antisera" "polycyclic aromatic content" "porous coated
anatomic" "porous coated total hip arthroplasty" "porous-coated hip
arthroplasties" "porous-coated patellar component" "porta-caval
anastomosis" "portable clinical analyzer" "portacaval anastomosis"
"portacaval shunt" "post chigger attachment" "postconceptional age"
"posterior cerebral arteries" "posterior communicating artery"
"posterior cortical atrophy" "posterior crico-arytenoid" "potassium
channels activators" "presence of parietal cell" "primary cardiac
arrest" "primary congenital aphakia" "principal component analyses"
"procoagulant" "procoagulant activities" "procoagulant cellular
activity" "procoagulatory activity" "prostatic carcinoma" "protein c
activator" "prothrombin complex activity" "protocatechuic acid"
"protocatechuic acid" "protocaval anastomosis" "pulmonary corpora
amylacea" "pyroglutamic acid" "pyrroloidine carboxylic acid"
"pyrroloidine-2-carboxylic acid" "pyrroloidine-5-carboxylic acid"
"pyrrolidine-5-carboxylate"
Problem 1: Ambiguity

- “Senses” of an abbreviation are usually not related.
- Long form often occurs at least once in a document.
- Disambiguating abbreviations is easy.
Problem 2: “Coreference”

- **Goal:** Establish that abbreviation and long form are coreferring.
- **Strategy:**
  - Treat each pattern $w^*(c^*)$ as a hypothesis.
  - Reject hypothesis if well-formedness conditions are not met.
  - Accept otherwise.
Dynamic Programming

- Align the abbreviation with the preceding text using dynamic programming.
- Associate costs with each alignment that reflect well-formedness of the abbreviation.
Medline excerpt: According to a system proposed by the European group for the immunological classification of leukemia (EGIL) ....

Align: “EGIL” with preceding text

European group for the immunological classification of leukemia
### Dynamic Programming Alignment costs

<table>
<thead>
<tr>
<th>long form</th>
<th>abbreviation</th>
<th>cost</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\varepsilon$</td>
<td>character c</td>
<td>100.0</td>
</tr>
<tr>
<td>$c_1$</td>
<td>$c_2 \ (c_1! = c_2)$</td>
<td>100.0</td>
</tr>
<tr>
<td>non-initial c</td>
<td>first c</td>
<td>100.0</td>
</tr>
<tr>
<td>initial c</td>
<td>$\varepsilon$</td>
<td>5.0</td>
</tr>
<tr>
<td>initial c</td>
<td>c</td>
<td>0.0</td>
</tr>
<tr>
<td>non-initial c</td>
<td>$\varepsilon$</td>
<td>0.1</td>
</tr>
<tr>
<td>non-initial c</td>
<td>c</td>
<td>1.0</td>
</tr>
</tbody>
</table>
Evaluation: Precision

- Algorithm tested on a dictionary of abbreviations available from the China Medical Tribute (452)
- 406 (90%) correct
- Error analysis:
  - Syllable boundaries
  - “Morphology”
  - Semantics
  - Suboptimal length/wellformedness trade-off
Errors: Syllable Boundaries

phosphatidylinositol manno-oligosaccharides

amplitude-integrated electroencephalography
pr---------o--M------M------P--------s--
precursors of matrix metalloproteinase

N-------A-----P-------R------T-a-s-e----
nicotinate phosphoribosyltransferase

C--------I-----------------N -------I--
cervical intraepithelial n-eoplasia
Errors: Semantics

antiphospholipid anticardiolipin antibodies

glucose-6-phosphate dehydrogenase-deficient
Errors: Incorrect Tradeoff
Length vs. Well-Formedness

P___O________P__C_______
pulmonary complications
P___O________P_________C____________
Postoperative pulmonary complications

P_________P__R__O_______M__________
premature rupture of the membranes
P________P_________R__O_______M__________
Preterm premature rupture of the membranes
Recall

- Analyze all of Medline (37 gigabytes)
- Identify all possible candidates
- 375 correctly identified out of 452 (83%)
- Errors:
  - Precision errors
  - Abbreviation not in Medline
  - Narrow scope of defining context
Errors: Abbreviations not in Medline

- VATS: video assisted thorascoscopy (vs. video assisted thorascopysurgery)
- VVR: ventricular volume reduction
Errors: Narrow Scope of Defining Context

- We only mine text segments for abbreviations that match regular expression.
- This regular expression was too narrowly defined.

“Post”-definition

ACA2p (Arabidopsis Ca2+-ATPase, isoform 2 protein)

Non-standard term

benzodiazepine receptor (peripheral) (BZRP)
Evaluation: recall/precision
No syllable boundaries
w/ syllable boundaries corrected
Welcome to our Biomedical Abbreviation Server!

We have scanned 11,447,996 PubMed citations for abbreviations and put them in a database. The database currently has 2,074,367 abbreviations.

Search for an abbreviation in the database.
Abbreviation: [ ] SEARCH

- Example: CDK

Search for an abbreviation with a keyword.
Keyword: [ ] SEARCH

- Example: oncogene

Show the abbreviations in a PubMed citation.
PubMed ID: [ ] SEARCH

- Example: 10226534

Search for abbreviations in some text.
We observed an increase in mitogen-activated protein kinase (MAPK) activity.
I found 462 hits for pca.

[Next 10] Start at #

<table>
<thead>
<tr>
<th>Abbreviation</th>
<th>Long Form</th>
<th>Score</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>PCA</td>
<td>0.91</td>
</tr>
<tr>
<td>2</td>
<td>PCA</td>
<td>0.91</td>
</tr>
<tr>
<td>3</td>
<td>PCA</td>
<td>0.91</td>
</tr>
<tr>
<td>4</td>
<td>PCA</td>
<td>0.91</td>
</tr>
<tr>
<td>5</td>
<td>PCA</td>
<td>0.91</td>
</tr>
<tr>
<td>6</td>
<td>PCA</td>
<td>0.91</td>
</tr>
<tr>
<td>7</td>
<td>PCA</td>
<td>0.91</td>
</tr>
<tr>
<td>8</td>
<td>PCA</td>
<td>0.91</td>
</tr>
<tr>
<td>9</td>
<td>PCA</td>
<td>0.91</td>
</tr>
<tr>
<td>10</td>
<td>PCA</td>
<td>0.91</td>
</tr>
</tbody>
</table>

Home
Approach 2

- The algorithm shown only considers the best alignment. If (best score > \( \theta \)) accept else reject.

- Alternative
  - Generate a set of good alignments
  - Build feature representation
  - Classify feature representation
Features for Classifier

- Describes the abbreviation.
  - Lower Abbrev

- Describes the alignment.
  - Aligned
  - Unused Words
  - AlignsPerWord

- Describes the characters aligned.
  - WordBegin
  - WordEnd
  - SyllableBoundary
  - HasNeighbor
# Weights of Abbreviation Features

<table>
<thead>
<tr>
<th>Feature</th>
<th>Weight</th>
</tr>
</thead>
<tbody>
<tr>
<td>CONSTANT</td>
<td>-9.70</td>
</tr>
<tr>
<td>LowerAbbrev</td>
<td>-1.21</td>
</tr>
<tr>
<td>Aligned</td>
<td>3.67</td>
</tr>
<tr>
<td>UnusedWords</td>
<td>-5.82</td>
</tr>
<tr>
<td>AlignsPerWord</td>
<td>0.70</td>
</tr>
<tr>
<td>WordBegin</td>
<td>5.54</td>
</tr>
<tr>
<td>WordEnd</td>
<td>-1.40</td>
</tr>
<tr>
<td>SyllableBoundary</td>
<td>2.08</td>
</tr>
<tr>
<td>HasNeighbor</td>
<td>1.50</td>
</tr>
</tbody>
</table>
Discussion

- Overall an easy problem
- Could learn the parameters of dynamic programming from training set.
  - Minimize cost: $\alpha$ align-cost + $(1-\alpha)$ recognition-cost
- Related work: see resources
Text-Enhanced Homology Search
(Chang, Raychaudhuri, Altman)
Sequence Homology Detection

- Obtaining sequence information is easy; characterizing sequences is hard.
- Organisms share a common basis of genes and pathways.
- Information can be predicted for a novel sequence based on sequence similarity:
  - Function
  - Cellular role
  - Structure
Evaluation: China Medical Tribune

- List of 452 biomedical abbreviations with expansions
- One model randomly picked from converged subset.
- Evaluation of precision: Test algorithm on set of 452
- Evaluation of recall: Run algorithm on medline
PSI-BLAST

- Used to detect protein sequence homology. (Iterated version of universally used BLAST program.)
- Searches a database for sequences with high sequence similarity to a query sequence.
- Creates a profile from similar sequences and iterates the search to improve sensitivity.
At each iteration, could find non-homologous (false positive) proteins.

False positives create a poor profile, leading to more false positives.
Addressing Profile Drift

- **PROBLEM**: Sequence similarity is only one indicator of homology.
  - More clues, e.g. protein functional role, exists in the literature.
- **SOLUTION**: we incorporate MEDLINE text into PSI-BLAST.
1. Search Database
2. Construct Profile
3. Examine Literature
Modification to PSI-BLAST

- Before including a sequence, measure similarity of literature. Throw away sequences with least similar literatures to avoid drift.

- Literature obtained from SWISS-PROT gene annotations to MEDLINE (text, keywords).

- Define domain-specific “stop” words (< 3 sequences or > 85,000 sequences) = 80,479 out of 147,639.

- Use similarity metric between literatures (for genes) based on word vector cosine.
Evaluation

- Created families of homologous proteins based on SCOP (gold standard site for homologous proteins--http://scop.berkeley.edu/)
- Select one sequence per protein family:
  - Families must have $\geq$ five members
  - Associated with at least four references
  - Select sequence with worst performance on a non-iterated BLAST search
Evaluation

- Compared homology search results from original and our modified PSI-BLAST.
- Dropped lowest 5%, 10% and 20% of literature-similar genes during PSI-BLAST iterations
Results

- 46/54 families had identical performance.
- 2 families suffered from PSI-BLAST drift, avoided with text-PSI-BLAST.
- 3 families did not converge for PSI-BLAST, but converged well with text-PSI-BLAST.
- 2 families converged for both, with slightly better performance by regular PSI-BLAST.
<table>
<thead>
<tr>
<th>Superfamily</th>
<th>Query Sequence</th>
<th>Words</th>
<th># Seqs</th>
<th>Convergence</th>
<th>Precision</th>
<th>Recall</th>
</tr>
</thead>
<tbody>
<tr>
<td>EGF/Laminin</td>
<td>C1R_HUMAN</td>
<td>1661</td>
<td>5</td>
<td>yes</td>
<td>no</td>
<td>0.11</td>
</tr>
<tr>
<td>Acid proteases</td>
<td>POL_HV2RO</td>
<td>1271</td>
<td>22</td>
<td>yes</td>
<td>no</td>
<td>0.6</td>
</tr>
<tr>
<td>PLP-dependent transferases</td>
<td>GLYC_RABIT</td>
<td>1052</td>
<td>21</td>
<td>no</td>
<td>yes</td>
<td>N/A</td>
</tr>
<tr>
<td>Thioredoxin-like</td>
<td>CAQS_RABIT</td>
<td>1516</td>
<td>13</td>
<td>no</td>
<td>yes</td>
<td>N/A</td>
</tr>
<tr>
<td>Glucocorticoid receptor-like</td>
<td>CYSR_CHICK</td>
<td>1738</td>
<td>10</td>
<td>no</td>
<td>yes</td>
<td>N/A</td>
</tr>
<tr>
<td>(DNA-binding domain)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>EF-hand</td>
<td>SCP_NERDI</td>
<td>963</td>
<td>31</td>
<td>yes</td>
<td>yes</td>
<td>0.92</td>
</tr>
<tr>
<td>Glycosyltransferases</td>
<td>CHLY_HEVBR</td>
<td>1007</td>
<td>20</td>
<td>yes</td>
<td>yes</td>
<td>1.0</td>
</tr>
<tr>
<td>Snake toxin-like</td>
<td>CD59_HUMAN</td>
<td>2435</td>
<td>23</td>
<td>yes</td>
<td>yes</td>
<td>1.0</td>
</tr>
</tbody>
</table>
Discussion

- Profile drift is rare in this test set and can sometimes be alleviated when it occurs.
- Overall PSI-BLAST precision can be increased using text information.
Resources

- Pac Symp Biocomput. 2001;:374-83. PMID: 11262956
- http://abbreviation.stanford.edu
