Pharm Bioinformatics

Lawrence Hunter, Ph.D., Director
Center for Computational Pharmacology
UCHSC School of Medicine

http://compbio.uchsc.edu
Larry.Hunter@uchsc.edu
Use of computational tools to
- Generate and/or test biological hypotheses
- Store, retrieve and/or integrate biological data
- Communicate (or find) published scientific results

Today’s lecture:
- Discrimination learning: a key bioinformatics technique
- Applications throughout the drug discovery pipeline
- A quick survey of publicly available data sources relevant to pharmacology
Modeling

- Using computation to “interpolate” from existing data to useful conclusions about unobserveds

- Different kinds:
  - Physicochemical: based in causal knowledge, e.g. quantum dynamics calculations
  - Statistical: based on empirical correlations, but without known causation (e.g. cigarettes and lung cancer)
    - Parameter estimation: model structure is given/assumed
    - Model search: find best model as well as fit parameters
Discrimination Learning

- Induction over a set of labelled examples to produce a method to distinguish one class from another
  - e.g. active vs. inactive compounds
  - A.K.A. “supervised learning”

Need:
- Training data (must be representative!)
- Representation (what aspects of examples are relevant?)
- Kind of rule to be induced (e.g. linear discriminant)
- Algorithm for induction...

Goal: Generalization (predictions for unseen data)
Some examples...

Given a set of gene expression values for responders and non-responders to a drug, induce a discriminator for who will respond based on an assay of gene expression.

Given binding affinities to a target for a large set of potential leads, induce a discriminator that predicts the affinity of an untested compound.

Given a set of carcinogenic compounds, induce a discriminator that predicts carcinogenicity.
Most important aspect of inductive learning

Most common: *fixed length feature vector.*
- Feature: observable value related to task
- Fixed length vector: a particular list of features which has the same meaning from example to example

Some feature sets (like sequences) have variable length or positions can have variable meaning

How to map to vectors?
- Descriptor sets for chemistry, e.g. ClogP, path length
Linear discrimination picture
Imagine we set a pseudovariable $Y$ to be +1 when an example is positive, and -1 otherwise.

Define a vector $X$ to represent each example, where each $X_i$ is the value of the $i$th feature for that example, $i \in F$.

Do multiple regression over the examples to calculate the least squares fit of coefficients $a$ and $c_i$ to

$$\hat{y} = a + \sum_{i \in F} c_i X_i$$

For new examples to be tested, if $\hat{y} > 0$, then call it positive.
Linear isn’t good enough

Many simple concepts (e.g. XOR, equality) can’t be linearly discriminated.
Other statistical discriminators

- Can specify other functions to draw discrimination line
  - Quadratic discriminants
  - Log/Exp, other polynomial functions
- Need to have
  - Parameterized discrimination function
  - Method to find best parameters for discrimination based on training data
    - Usually least squares
    - What to minimize?
Neural Networks

- Method for inducing arbitrary non-linear functions as discrimination lines
- Metaphor to real neurons. Made of nodes and arcs
How do NN's work?

Each node has a value, and each arc a weight.
Values of input nodes are set by each example.
Non-input node are the weighted sum of their inputs put through a "squashing" function.

\[ n_4 = f(n_1w_1 + n_2w_2 + n_3w_3) \]
NN discrimination picture

Feature 1

Feature 2
Where do weights come from?

- Have training examples with known outputs.
- Use the "backpropagation" algorithm:
  - Start with small random weights
  - For each example, calculate the predicted outcome.
  - Calculate the error (predicted - actual outcome)
  - Change weights to reduce the error
    - Calculate partial derivative of each weight w.r.t the error.
    - Gradient descent optimization of weights to find minimum error
  - Use “learning rate” to avoid moving too far (can estimate this, too: conjugate gradient descent).
\[ \Delta w = f'(\text{sum}) \times e + \times n_i \times L \]

- **\( \Delta w \)**: Change in weight
- **\( f'(\text{sum}) \)**: Derivative of squashing function
- **\( e \)**: Error
- **\( n_i \)**: Input
- **\( L \)**: Learning rate
NN has problems, too.

Overfitting or poor generalization

- Discrimination line follows training examples too closely. “Memorizing the training set”
- Not enough data to fit all the parameters (each weight is a parameter)
- Stop training before error minimum to get better generalization?
Overfitting picture
Support Vector Machines (SVMs)

- Recent (Vapnik, 1992) discrimination method with good formal properties
  - Fast to train (polynomial time)
  - Closed form error bounds
  - Avoids the "curse of dimensionality"

- Works by using a kernel (similarity) function
- Uses similar examples (neighbors) to determine predictions.
How do SVMs work?

- Use kernel function to find *maximum margin hyperplane* dividing classes.
  - Finding the hyperplane is a quadratic optimization problem
- Maximum margin has good generalization properties
- Examples far from the decision boundary have no effect.
  - Examples near the boundary are "support vectors"
- The class of possible shapes of the dividing line in the data space (and therefore the accuracy of the discrimination and potential overfitting) depends on the kernel function.
SVM picture

Feature 1

Feature 2
SVM problems

- **Interpretability**
  - SVMs produce no explicit model or generalization

- **Dependence on kernel function**
  - Similarity is complex, and may be hard to define well
  - Use of generic kernels (e.g. Euclidean distance) often doesn’t perform as well as other methods (e.g. NNs)
General challenges

⚠️ The curse of dimensionality
  - Many possibly interacting factors. Number of interactions goes up with $F!$
  - Kernel methods escape the curse by requiring a similarity function (which must take interactions into account)

🐝 Too many parameters (models) and too little data.
  - General rule: prefer 10x data to parameters, need at least 1x (practical minimum is 3x).
  - Model search requires even more data.
Statistical evidence of interactions

Gene-gene relationships in expression data
- Correlations among expression levels
- Gene-gene interaction terms which predict phenotype

Look for pairs which together predict phenotype
- Lots of pairs to test
- Use multiple testing corrections
High order interactions

Not just pairs...

Many possible high order interactions!

Discretize and association mine

Risks in discretizing and finding high order associations in too little data, but unique predictive models and “biological reality”
Induction and drug discovery

Why build predictive models from data? *Exploit the results of high throughput instrumentation*

Target identification from expression arrays:
- Expression levels of which genes are correlated with disease / therapeutic response / progression / etc.?

Optimization from combichem data:
- Which aspects of small molecule structure are correlated with binding affinity to target?

ADME/Tox predictions
HT instruments and bioinformatics are so intertwined, I invented a term for them together

From the Greek \( \text{βίος} \) (life) and \( \text{γνωσίς} \) (knowing)

Two kinds of biagnostic machines:

- *Instruments* that produce data about a living things in molecular detail and with genomic breadth
- *Bioinformatics systems* that bring to bear existing knowledge in the computational analysis of data
Biognostic Instruments

- 15k gene expression arrays for $500 / array
- SNP genotyping 10k markers for $1000
- Combichem synthesis and affinity assay of 50k compounds / day
So Much Wonderful Data...

- More than 12,000,000 biomedical journal articles in Medline
- 600,000 new articles per year, accelerating at 10% per year
- ~800 Genomes complete or in progress (17 Eukaryotes done)
- Proteomics, Structural genomics, HapMap projects underway
Data about Databases

*Nucleic Acids Research* publishes an annual database issue. 2003 issue lists more than 300 high quality databases.

- [http://nar.oupjournals.org/content/vol31/issue1/](http://nar.oupjournals.org/content/vol31/issue1/)

Small excerpt from the A's:

- AARSDB: Aminoacyl-tRNA synthetase sequences
- ABCdb: ABC transporters
- AceDB: C. elegans, S. pombe, and human sequences and genomic information
- ACTIVITY: Functional DNA/RNA site activity
- ALFRED: Allele frequencies and DNA polymorphisms
What can be discovered about a gene by a database search?

- A little or a lot, depending on the gene
  - *Evolutionary information*: homologous genes, taxonomic distributions, allele frequencies, synteny, etc.
  - *Genomic information*: chromosomal location, introns, UTRs, regulatory regions, shared domains, etc.
  - *Structural information*: associated protein structures, fold types, structural domains
  - *Expression information*: expression specific to particular tissues, developmental stages, phenotypes, diseases, etc.
  - *Functional information*: enzymatic/molecular function, pathway/cellular role, localization, role in diseases
Using a database

How to get information out of a database:
- Summaries: how many entries, average or extreme values
- Browsing: no targeted information to retrieve
- Search: looking for particular information

Searching a database:
- Must have a key that identifies the element(s) of the database that are of interest.
  - Name of gene
  - Sequence of gene
  - Other information
- Helps to have particular *informational goals*
Searching sequence databases

- Start from sequence, find information about it
- Many kinds of input sequences
  - Could be amino acid or nucleotide sequence
  - Genomic or mRNA/cDNA or protein sequence
  - Complete or fragmentary sequences
- Exact matches are rare (even uninteresting in many cases), so often goal is to retrieve a set of similar sequences
- BLAST is fast approximate sequence matcher
NCBI and Entrez

One of the most useful and comprehensive sources of databases is the NCBI, part of the National Library of Medicine.

NCBI provides interesting summaries, browsers for genome data, and search tools.

Entrez is their database search interface


Can search on gene names, sequences, chromosomal location, diseases, keywords...
Entrez is a retrieval system for searching several linked databases. It provides access to:

- **PubMed**: The biomedical literature (PubMed)
- **Nucleotide**: sequence database (GenBank)
- **Protein**: sequence database
- **Structure**: three-dimensional macromolecular structures
- **Genome**: complete genome assemblies
- **PopSet**: population study data sets
- **OMIM**: Online Mendelian Inheritance in Man
- **Taxonomy**: organisms in GenBank
- **Books**: BookShelf online books
- **3D Domains**: domains from Entrez Structure
- **UniSTS**: markers and mapping data
- **SNP**: single nucleotide polymorphisms
- **CDD**: conserved domains
- **Journals**: journals in Entrez
- **UniGene**: gene-oriented clusters of transcript sequences
- **PMC**: full-text digital archive of life sciences journal literature

**NEW:** [NCBI Web Site](http://www.ncbi.nlm.nih.gov): NCBI Web site search
LocusLink

NCBI’s *curated* source of information about genes (in humans and model organisms)

Best place to start looking for information about a gene of interest

Assembles pointers to all kinds of information in the NCBI databases:

- sequence, structure, polymorphisms, synteny, phenotypes, literature
Homo sapiens Official Gene Symbol and Name (HGNC)

MAPK1: mitogen-activated protein kinase 1

LocusID: 5594

Overview

**RefSeq Summary:** The protein encoded by this gene is a member of the MAP kinase family. MAP kinases, also known as extracellular signal-regulated kinases (ERKs), act as an integration point for multiple biochemical signals, and are involved in a wide variety of cellular processes such as proliferation, differentiation, transcription regulation and development. The activation of this kinase requires its phosphorylation by upstream kinases. Upon activation, this kinase translocates to the nucleus of the stimulated cells, where it phosphorylates nuclear targets. Two alternatively spliced transcript variants encoding the same protein, but differing in the UTRs, have been reported for this gene.

**Locus Type:** gene with protein product, function known or inferred

**Product:** mitogen-activated protein kinase 1

**Alternate Symbols:** ERK, p38, p40, p41, ERK2, ERT1, MAPK2, PRKM1, PRKM2, P42MAPK, p41mapk

**Alias:** protein tyrosine kinase ERK2

mitogen-activated protein kinase 2 extracellular signal-regulated kinase 2
Function: GeneRIF

Submit GeneRIF (All Pubs)

EC Number: 2.7.1.-

GeneRIF: Gene References into Function:

11840343  • mediates regulation of p73 by c-Abl

12082091  • Role of ERK2 in BRCA1-induced apoptosis

12058028  • role in stabilizing p21(Cip1) by phosphorylation

12200131  • ITGB1 activated by ERK1/2, p38 MAPK after hypoxia

12460991  • activation by Pro33 polymorphism of integrin beta3

11943771  • mediates activation of neutrophils by lipopolysaccharide

12076252  • p38MAPK is activated by phosphorylated ATF6 and induces HSPA5 binding

11861509  • ERK1/2 activation is a regulator of progesterone synthesis in hGL cells

12536241  • IL-2 decreased the expression of ERK2, whereas polysaccharide K did not.
Gene Ontology™:

- MAP kinase activity
- ATP binding activity
- Protein serine/threonine kinase activity
- Induction of apoptosis
- Cell cycle
- Chemotaxis
- Response to stress
- Protein amino acid phosphorylation
- Synaptic transmission
- Signal transduction
- Transferase activity

Evidence Source Pub

<table>
<thead>
<tr>
<th>Term</th>
<th>Source</th>
<th>Evidence</th>
<th>Pub</th>
</tr>
</thead>
<tbody>
<tr>
<td>MAP kinase activity</td>
<td>Tas</td>
<td>GOA</td>
<td>pm</td>
</tr>
<tr>
<td>ATP binding activity</td>
<td>Iea</td>
<td>GOA</td>
<td></td>
</tr>
<tr>
<td>Protein serine/threonine</td>
<td>N</td>
<td>GOA</td>
<td></td>
</tr>
<tr>
<td>kinase activity</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Induction of apoptosis</td>
<td>Tas</td>
<td>GOA</td>
<td>pm</td>
</tr>
<tr>
<td>Cell cycle</td>
<td>Iea</td>
<td>GOA</td>
<td></td>
</tr>
<tr>
<td>Chemotaxis</td>
<td>Tas</td>
<td>GOA</td>
<td>pm</td>
</tr>
<tr>
<td>Response to stress</td>
<td>Tas</td>
<td>GOA</td>
<td>pm</td>
</tr>
<tr>
<td>Protein amino acid</td>
<td>Iea</td>
<td>GOA</td>
<td></td>
</tr>
<tr>
<td>Phosphorylation</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Synaptic transmission</td>
<td>Tas</td>
<td>GOA</td>
<td>pm</td>
</tr>
<tr>
<td>Signal transduction</td>
<td>Tas</td>
<td>GOA</td>
<td>pm</td>
</tr>
<tr>
<td>Transferase activity</td>
<td>Iea</td>
<td>GOA</td>
<td></td>
</tr>
</tbody>
</table>

Relationships

Mouse Homology Maps:

NCBI vs. MGD  16 9.82 cM Mapk1 Hs Mm

Map Information

Chromosome: 22
Cytogenetic: 22q11.21 22q11.2
Markers: Chr. 22 D22S1248 D22S1248

Links

LocusLink: Home
MAPK1 Index
Overview
Function
Relationships
Map
RefSeq
Related Seqs
Links

LocusLink:
Collaborators
Download
FAQ
Help
Statistics

RefSeq:
About
Download
FAQ
NCBI Reference Sequences (RefSeq)

Category: Reviewed

1. mRNA: NM_002745
   Protein: NP_002736  mitogen-activated protein kinase 1
   Domains: Protein kinase domain, score: 534
            Tyrosine kinase, catalytic domain, score: 319
            Serine/Threonine protein kinases, catalytic domain, score: 616

   Transcript Variant: This variant (1) contains a different 3' UTR region, however, encodes the same protein as compared to variant 2.

   GenBank: BC017832, M84489, Z11695
   Source:

2. mRNA: NM_138957
   Protein: NP_620407  mitogen-activated protein kinase 1
   Domains: Protein kinase domain, score: 535
            Tyrosine kinase, catalytic domain, score: 319
            Serine/Threonine protein kinases, catalytic domain, score: 615

   Transcript Variant: This variant (2) contains a different 3' UTR region, however, encodes the same protein as compared to variant 1.

   GenBank: BC017832
Related Sequences

<table>
<thead>
<tr>
<th>Nucleotide</th>
<th>Type</th>
<th>Protein</th>
</tr>
</thead>
<tbody>
<tr>
<td>AK090869</td>
<td>m</td>
<td>AAH17832</td>
</tr>
<tr>
<td>BC017832</td>
<td>m</td>
<td>AAA58459</td>
</tr>
<tr>
<td>M84489</td>
<td>m</td>
<td>CAA77752</td>
</tr>
<tr>
<td>Z11694</td>
<td>m</td>
<td>CAA77753</td>
</tr>
<tr>
<td>Z11695</td>
<td>m</td>
<td>P28482</td>
</tr>
<tr>
<td>None</td>
<td>p</td>
<td>BL</td>
</tr>
</tbody>
</table>

Additional Links

- **OMIM:** 176948
- **UniGene:** Hs.324473
- **KEGG pathway:** MAPK signaling pathway
- **KEGG pathway:** Sphingoglycolipid metabolism
- **KEGG pathway:** Inositol phosphate metabolism
- **KEGG pathway:** Starch and sucrose metabolism
- **KEGG pathway:** Integrin-mediated cell adhesion
- **KEGG pathway:** Porphyrin and chlorophyll metabolism
- **KEGG pathway:** Benzoate degradation via CoA ligation
- **KEGG pathway:** Nicotinate and nicotinamide metabolism
MAPK signaling pathway - Homo sapiens
Go to: [ Ortholog Table ]
Go to: Homo sapiens

MAPK SIGNALING PATHWAY

04010hsa 12/5/01
### dbSNP Table

<table>
<thead>
<tr>
<th>Config position</th>
<th>dbSNP rs/# cluster id</th>
<th>Heterozygosity</th>
<th>3D</th>
<th>Function</th>
<th>dbSNP allele</th>
<th>Protein residue</th>
<th>Codon position</th>
<th>Amino acid position</th>
<th>Snps In Gene Model Legend</th>
</tr>
</thead>
<tbody>
<tr>
<td>1508071</td>
<td>rs13058</td>
<td>0.078</td>
<td></td>
<td>untranslated region</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1508979</td>
<td>rs1803546</td>
<td>N.D</td>
<td></td>
<td>untranslated region</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1509002</td>
<td>rs1048752</td>
<td>N.D</td>
<td></td>
<td>untranslated region</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1510272</td>
<td>rs743411</td>
<td>0.000</td>
<td></td>
<td>untranslated region</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1511724</td>
<td>rs2266965</td>
<td>0.457</td>
<td></td>
<td>untranslated region</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1511896</td>
<td>rs5755074</td>
<td>N.D</td>
<td></td>
<td>untranslated region</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1512123</td>
<td>rs2266966</td>
<td>0.465</td>
<td></td>
<td>untranslated region</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1513758</td>
<td>rs2298432</td>
<td>0.494</td>
<td></td>
<td>untranslated region</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1513950</td>
<td>rs3188781</td>
<td>N.D</td>
<td></td>
<td>untranslated region</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1515609</td>
<td>rs5755099</td>
<td>N.D</td>
<td></td>
<td>intron</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1517734</td>
<td>rs1803545</td>
<td>0.061</td>
<td>Yes</td>
<td>config reference</td>
<td>C</td>
<td></td>
<td>3</td>
<td>321</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>synonymous change</td>
<td>T</td>
<td></td>
<td>Asp</td>
<td>[D]</td>
<td></td>
</tr>
<tr>
<td>1518059</td>
<td>rs4821263</td>
<td>N.D</td>
<td></td>
<td>intron</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1519247</td>
<td>rs2006893</td>
<td>N.D</td>
<td></td>
<td>intron</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1519298</td>
<td>rs2096687</td>
<td>N.D</td>
<td></td>
<td>intron</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1519696</td>
<td>rs2055108</td>
<td>N.D</td>
<td></td>
<td>intron</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Beyond NCBI

- Lots of good databases available outside of NCBI
- SwissProt / TrEMBL is better annotated [http://us.expasy.org/sprot/](http://us.expasy.org/sprot/)
- UCSC has a nicer genome browser [http://genome.ucsc.edu/](http://genome.ucsc.edu/)
- PDB has better structure services [http://www.rcsb.org/pdb/](http://www.rcsb.org/pdb/)
- And remember the NAR Database issue…
Conclusions

- Bioinformatics is increasingly important throughout the drug discovery pipeline.
- Discrimination techniques are an important tool for predictive modeling.
- There are many useful public repositories of molecular biology data. NCBI is just the start.
- Biol 7711 (fall Q) is open to pharm students with an interest in bioinformatics (need biostats!)