Bioinformatics
History and Introduction

Luce Skrabanek
24 January 2013
http://chagall.med.cornell.edu/BioinfoCourse/
What is bioinformatics?

• The application of statistics and computer science to the field of molecular biology

• Uses computers to store, manage, manipulate, analyze biological data

• Applications:
  – Unravel biological meaning and develop biological and evolutionary insights
  – Drug discovery and development; personalized medicine
Required skills

• Biology
• Computer science
• Mathematics/Statistics
• Information technology

• All needed to some degree but some needed more for some applications than others
• Often, not all required skills are present in one person. Need teams with programmer, biologist, IT person, statistician, even graphic designer
Bioinformatics is predictive

• Runs the gamut of predictive capability

• Prediction is based on what we already know (i.e., biological knowledge)
  – What is the function of a protein?
  – How does a protein function?
  – What factors does a protein need to function?
  – How do multiple proteins interact with one another and affect each other’s function?

• Bioinformatics does not replace experiments
Applications of bioinformatics

• Most basic: use web-based tools
  – Primarily need biology
• Use Unix-based tools
  – Above, plus need ability to use Unix, write wrappers in Perl/Python, write shell scripts
• Use Unix tools for high-throughput data
  – Above, plus an understanding of data storage and scalability
• Algorithm development
  – Computer-science focus, will usually partner with a biologist
• Making data/methods public
  – Creating databases/web pages (IT)
Interplay of disciplines

• Advances in **experimental biology** give us new and different types of data to analyze, which may require completely new ways of dealing with that data (e.g., data from high-throughput experiments)

• Advances in **biological understanding** allow computational **predictive models** to incorporate more information and become more complex

• Advances in **computer technology** allow us to deal with greater amounts of data in reasonable time frames
The scales of knowledge

- Historically, dealing with one protein at a time
- Nowadays, we are dealing with large amounts of noisy data – requires a new kind of scientist: “data scientist”
  - Translates data into predictive insights
  - Feels comfortable dealing with noisy incomplete data
  - Explores data to generate hypotheses, exploration drives direction of research
  - Can handle the challenge of dealing with petabytes of data

Luce Skrabanek
Scalability

• Huge volumes of data available to us
  – Complete genomes, NGS

• Necessary computational resources now available to deal with these amounts of data
  – 8 GB (~human genome) can be stored on an iPod
  – Tree of life can be stored in 1TB
  – Raw data from 1 NGS experiment = 1TB

• Tools and techniques have to be efficient and scalable
Biology of Systems

• Huge amounts of ‘parts’ data
  – Sequence (both nucleotide and protein)
  – Structure
  – Function
  – Biochemical information
  – Protein-protein interactions, complexes
  – Protein-DNA complexes
  – Kinetics of reactions

→ Integrated together into “Systems Biology”
  • The study of the interactions between the components of a biological system
  • How do those interactions give rise to the functions and behaviors that we see?
Mathematical modeling

• Biological systems can be represented by ODEs
  – compartments
  – stochastic methods for low concentration components

• Systems modeling can:
  – effectively integrate “parts” information
  – help reveal non-intuitive “emergent” properties
  – teach us how cells store information and ‘compute’

• Quantitative models of pathways and networks
  – predict cellular responses to external stimuli
  – model effects of perturbations on the system
  – predict how to ‘correct’ disease states
    • identify control points in the system
Where do we get our data from?

- Experimental data from collaborators
- Databases

http://www.oxfordjournals.org/nar/database/c/
Some important databases

- Archives of data stored in a uniform, consistent, efficient, searchable way (database and query tools)
- Primary databases vs. secondary databases
  - Sequence and annotation vs. curated results of analyses
- PubMed
- GenBank/EBI/DDBJ vs. RefSeq
- Short Reads Archive
- UniProtKB
- PDB
- HPRD
Caution!

• Errors in the databases
  – wrong positions of genes
  – exon-intron boundary errors
  – contaminating sequences
  – sequence discrepancies/variations
  – frameshift errors
  – annotation errors
  – spelling mistakes
  – incorrectly joined contigs
Mouse build 30

Mouse BLAT Results

BLAT Search Results

<table>
<thead>
<tr>
<th>ACTIONS</th>
<th>QUERY</th>
<th>SCORE</th>
<th>START</th>
<th>END</th>
<th>QSIZE</th>
<th>IDENTITY</th>
<th>CHRO</th>
<th>STRAND</th>
<th>START</th>
<th>END</th>
<th>SPAN</th>
</tr>
</thead>
<tbody>
<tr>
<td>browser details NM_009375.1</td>
<td>4659</td>
<td>1</td>
<td>4696</td>
<td>8301</td>
<td>99.9%</td>
<td>15</td>
<td>-</td>
<td>68019482</td>
<td>68056718</td>
<td>37237</td>
<td></td>
</tr>
<tr>
<td>browser details NM_009375.1</td>
<td>2583</td>
<td>5668</td>
<td>8301</td>
<td>8301</td>
<td>99.5%</td>
<td>15</td>
<td>+</td>
<td>68256929</td>
<td>68370132</td>
<td>113204</td>
<td></td>
</tr>
<tr>
<td>browser details NM_009375.1</td>
<td>700</td>
<td>4697</td>
<td>5400</td>
<td>8301</td>
<td>100.0%</td>
<td>15_random</td>
<td>-</td>
<td>18845817</td>
<td>18860431</td>
<td>14615</td>
<td></td>
</tr>
</tbody>
</table>

Chr 15

Chr 15_random
My protein doesn’t begin with an M?

• Nuclear-encoded proteins always begin with a methionine

• Some entries in protein databases don’t!
  – Some part of the 5’ UTR was misinterpreted as coding sequence
  – mRNA was not fully sequenced at the 5’ end

• In both cases, we can use comparative alignment information, either from closely related paralogs or other ESTs/mRNA sequences to try to identify the correct gene structure
Consequences of errors in trusted data sources

• Automated annotators (e.g., function, genome) rely on curated information sources to make predictions

• False research directions, false conclusions
  – Willersley 2002 – contaminant misidentified as lateral transfer from prokaryotes into humans; Sekyere 2003 – non-existent human melanotransferrin gene
  – deCODE overestimated genetic diversity in Icelanders because of errors in mtDNA sequencing studies

• Impediment to improving current prediction models
  – Harder to improve if we don’t know where the errors in current knowledge lie
Some sample error rates

- **GenBank**: 0.29% - 3.1% sequencing errors
  - Wesche et al, DNA Seq 2004

- **Enzyme annotations**: ~0% (SwissProt) – 5-63% (NR/TrEMBL/KEGG)

- **GO database** (standard database used for automated functional annotation prediction): 28-30%
  - Jones et al, BMC Bioinfo 2007
Methods of error introduction

• Entry errors in primary databases
  – Direct deposit into database by experimentalists
  – Incorrect transfer (from literature, other databases) by curators

• Analysis errors in secondary databases
  – Incorrectly derived data

• Inherited errors in secondary databases
  – Sequence similarity based annotation strategies may introduce new annotation errors, by incorrectly identifying similar sequences
Propagation of errors

- The quality of existing sequence annotations impacts on the quality of future sequence annotations through the commonly used practice of basing sequence annotations on sequence similarity.

- Corrections to such errors rarely occur, and when they do, the correction usually does not get propagated.
  - Some exceptions: SwissProt/UniProtKB, a curated database, does get regularly updated and corrected

**MORAL:** as a submitter, quality check when entering data; as a user, keep possibility of errors in mind
Examining bioinformatics approaches to a biological problem

miRNA target prediction
miRNAs: an introduction

• Small RNA molecules, 19-24 bases in length, that are involved in regulation of numerous biological processes
  – Target mRNA for degradation
  – Repress translation

• Derived from longer primary miRNA transcripts
Introduction contd.

• First discovered in *C. elegans* in 1993
  – lin4, let7
• Subsequently found to be widespread
• About 1600+ human miRNAs known
  – Many have been associated with specific diseases
    • Identification of specific miRNA species may be a possible tool for cancer diagnosis
  – TarBase – database of experimentally verified miRNA targets
  – miRBase – database of published miRNA sequences and targets
miRNA

3' UTR

adapted from Li et al, Mamm Genome 2009
Direct experimental identification of target genes

- Reporter constructs (luciferase, GTP) with the 3’ UTR of the predicted target gene: measure change in intensities after introduction of miRNA
- RT-PCR: track changes in mRNA levels after introduction of miRNA
- Does not directly identify MREs (miRNA recognition elements)
  - Site-directed mutagenesis studies
  - Restoration of complementarity by mutating miRNA sequence
Indirect experimental identification of target genes (target “enrichment”)

• For miRNAs involved in RNA degradation
  – Microarrays: differential gene expression with/without miRNA [secondary and non-specific effects]
  – RNA-Seq: next gen sequencing method of measuring differential expression [secondary and non-specific effects]
  – HITS-CLIP (high-throughput sequencing of RNAs isolated by crosslinking immunoprecipitation): can identify specific binding sites for targeted miRNAs [non-functional sites] [Chi et al Nature 2009]

• For miRNAs involved in translation repression
Computational prediction of miRNA targets

Main types of target site duplex

Mazière and Enright, Drug Disc Today, 2007
Computational prediction of miRNA targets

• Problems:
  – Short length: normal (Karlin-Altschul) alignment statistics no longer hold; even over the short alignment length, can have bulges and loops
  – Not all 3’ UTRs characterized: critically important to have the whole 3’ UTR
    • 30% of human genes lack definitive 3’ UTR boundaries
  – Unavailability of appropriate species for conservation analysis
Computational prediction of miRNA targets

• Different trade-off between sensitivity and specificity depending on requirements
  – Large-scale genome studies often sacrifice sensitivity for specificity – they want fewer false positives, at the cost of more false negatives – to ensure (fewer) predictions of better quality
  – A researcher interested in a single gene or pathway who wants to know how miRNAs impact their systems may want to have a very high sensitivity (few false negatives) because they will want to identify (and test) as many real regulation points as possible
Pipeline to identify miRNA targets in *Drosophila* - miRanda

Find complementary sequence matches in 3’ UTRs
(Modified Smith-Waterman algorithm)

Calculate free energy (stability) of miRNA/UTR binding
($\Delta G$ Kcal / mol)

Estimate evolutionary conservation
(Sequence conservation; relative positioning within the 3’ UTR)

73 known *Drosophila* miRNAs

Enright et al, Genome Biology, 2003
miRNAs are very small (21-22nt)
  - Enormous number of potential targets with complementary sequence
  - BLAST does not scale.

Low-complexity sequences
  - Signal to noise problem

Standard sequence analysis packages generally not applicable
  - Looking for complementarity, not similarity
    - i.e. A:U G:C not A:A G:G etc.
  - Wobble pairing permitted
    - G:U and U:G base pairs

Small number of known cases to work with
Sequence matching algorithm

• Modified Smith-Waterman algorithm
  – Instead of looking for matching nucleotides, finds complementary nucleotides
  – Allows GU ‘wobble’ pairs (but downweights them)
  – Scoring system weighted so that complementarity to the first 11 bases of the miRNA is more greatly rewarded
  – Non-complementarity also more heavily penalized in that region
  – Known miRNAs bind 3’ UTRs at multiple sites
    • Additive scoring system for all target sites predicted in a UTR
• Calculate free energy of binding (Vienna RNA package)
Evolutionary conservation

- Used conservation as a way of keeping only the most likely miRNA target candidates
- Used *Drosophila pseudooscura* and *Anopheles gambiae* as closely related species:
  - Required $\geq 80\%$ sequence similarity of target site with *D. pseudooscura*
  - Required $\geq 60\%$ sequence identity with *A. gambiae*
- Also, required that the location of the target site in the 3’ UTR be equivalent between species
Control sequences

• 100 sets of random 73 miRNAs generated
  – Conserved *D. melanogaster* miRNA nucleotide frequencies
• Analysis run independently for each set
• Results and counts averaged over all 100 sets
• Overall estimated false positive (FP) rate: 35%
  – Number of random hits / number of “real” hits
• If only count targets that have \( \geq 2 \) conserved sites in a UTR, the FP rate drops to 9%
Validation

• Initial validation: application to experimentally verified targets
  – 9/10 known target genes for three miRNAs correctly identified
  – BUT biased in favor of this since the method (i.e., parameter ranges chosen) is based on the background knowledge derived from these known target sites

• For 73 *Drosophila* miRNAs, 535 predicted target genes (out of ~9,805 3’ UTRs (from ~13,500 genes in the genome))
  – 231 have ≥ 1 predicted target interaction site
  – Target genes include many transcription factors and other genes involved in development

→ See both one-to-many and many-to-one relationships
  → One miRNA binds to many targets
  → One target is bound by many miRNAs
Pipeline to identify mammalian targets: TargetScan

79 conserved mammalian miRNAs

Find “seed matches” in the 3’ UTR of human/mouse/rat genes (match bases 2-8 of the miRNA exactly)

Extend the seed matches; optimize base pairing of remainder (RNAfold)

Evaluate the folding free energy (RNAeval)

Assigns scores to each UTR based on number and free energy of matches
Controls

• Shuffled sequences - have fewer matches than the real miRNA
• Selected so as to preserve all relevant compositional features
  – Expected frequency of seed matches to the UTR dataset
  – Expected frequency of matching to the 3’ end of the miRNA
  – Observed count of seed matches in the UTR dataset
  – Predicted free energy of the RNA duplex
• Each shuffled control sequence also has the same length and base composition as the parent
• Signal:noise ratio = 3.2:1
  – 5.7 targets with real miRNAs versus 1.8 targets found with control sequences = ~3.9 “real” targets per miRNA
  – Approximately a 31% FP rate
Validation

• Luciferase reporter assays used to test 15 (out of >400) predicted targets
  – Experimental support for 11/15

• Mammalian miRNA targets have diverse functions (unlike plants, where miRNAs are almost exclusively involved in developmental processes)
  – Enriched in developmental function, transcription
  – Also in nucleic acid binding and transcriptional regulator activity
Examine results

- Added in dog and chicken conservation
- Looked at flanking sequence of control and real matches in the UTRs

Lewis et al, Cell, 2005
Modify model - TargetScanS

- Targets identified by **conserved** complementarity to nucleotides 2-7 of the miRNA
- Require either a conserved adenosine at nucleotide 1 or a match at m8
  - t1A anchor may be important for binding by the RISC complex
- Often, a conserved t9A anchor at nucleotide 9, when there is an m8 match
- Don’t look past nucleotide 9 anymore
- Don’t calculate free energy anymore
- Don’t up-weigh multiple hits on the same UTR anymore
- Potentially, thousands of mammalian targets
Many other programs...

- Many programs are claimed to be able to discover miRNA targets in mammals
  - miRanda  Enright et al, SKI  Whole sequence complementarity searching
  - TargetScan Lewis et al, MIT  Seed complementarity
  - TargetScanS Lewis et al, MIT  Seed complementarity
  - PicTar Rajewsky et al, NYU  Thermodynamics
  - PITA  Kertesz et al, Weizmann/RU  Thermodynamics
  - RNA22  Rigoutsos et al, IBM  Thermodynamics and multiple pattern occurrence
  - EIMMo  Gaidatzis et al, UBasel  Evolutionary conservation
  - DIANA-MicroT  Hatzigeorgiou et al, Upenn  Thermodynamics and evolutionary conservation

- Different algorithms / models give different results
- Different miRNA-mRNA target site duplexes may be better predicted by one algorithm than another
User frustration

• Anil Jeqqa, posting on the miRNA Nature forums, reports:
  - “I was looking at and comparing the miRNA target gene predictions from five commonly used algorithms, viz., miRanda, targetScanS, PicTar, microT and mirTarget. Surprisingly, there is so little overlap! And I also did a comparison with the entries in TarBase (that houses about 100 experimentally validated miRNA-gene pairs) and surprisingly almost all of the five prediction algorithms perform quite badly.” (from the miRNA forum on the Nature website, 27 August, 2007)
5 miRNAs vs. Selbach

**Evaluation comparison**

Most programs rely heavily on the evolutionary conservation of the seed target region, incorporating some of the following features:

- Detailed phylogenetic models
- HMM that combines scores for multiple target sites on the same UTR
- Accessibility of the binding site region
- Local concentration of miRNA sequences
- Thermodynamic stability

**Precision** = number correctly predicted / total predicted

**Sensitivity** = number correctly predicted / total real
A single accurate algorithm is better than a combination of predictions. Better specificity of a combination is achieved at a higher price in sensitivity.

Evaluating these predictions:

- vs. TarBase: 61 miRNAs, 150 targets
- Effect of combining/intersecting target predictions
SVMs: Brief introduction

• Positive dataset
• Negative dataset
• Set of features that are likely to discriminate between the two

• Divide data into training set and test set
• Project training set into many dimensional space and find the hyperplane that best divides the training set into the correct categories
• Validate model (i.e., choice of plane) on test set

24 January 2013
Luce Skrabanek
SVMs: Negative Datasets

- Machine learning approaches rely heavily on negative (as well as positive) datasets
  - Negative datasets usually created by generating random sequences
    - May inadvertently include real sites
    - May be so different from real sites that the difference between positive and negative sets is artificially great
- Bandyopadhyay and Mitra created a negative dataset from high-throughput data where you would expect many false positives (either secondary or non-specific effects)
  - Series of steps to filter out potential candidate pairs
    - The miRNA and its target could not both be highly expressed in the same tissue. If an miRNA is highly expressed in a tissue, then its target cannot be
    - The interaction energy between miRNA and target cannot be too strong
    - Target seed site cannot be too conserved
  - 289 negative examples, validated by looking at the Selbach pSILAC dataset
TargetMiner

- Used the 289 negative examples and 289 positive examples
- Extracted 90 features, including:
  - Features from seed-matching site
  - Frequency of single nucleotides in seed-matching site
  - Frequency of single nucleotides outside of seed-matching site
  - Frequency of di-nucleotides in seed-matching site
  - Frequency of di-nucleotides outside of seed-matching site
  - miRNA-mRNA base interaction features in seed region
  - Pairs of consecutive miRNA-mRNA base interaction features in seed region
- From these, chose the 30 that were the most discriminatory
Comparison of TargetMiner with other target prediction algorithms

24 January 2013

Luce Skrabanek

Bandyopadhyay and Mitra Bioinformatics 2009
Further evaluation

• Trained other SVMs with their negative dataset – found that they improved the accuracy

• Trained their 90-feature SVM with a negative dataset made in the traditional way and decreased the accuracy
Future directions

• Advent of experimental data gives excellent benchmarking opportunities as well as providing new data to refine hypotheses
  – SILAC: measures the levels of many proteins concurrently
    • Baek et al Nature 2008
    • Selbach et al Nature 2008
  – HITS-CLIP: identification and sequencing of target sites for miRNAs
    • Chi et al Nature 2009
• Look for target sites outside the 3’UTR
• Combinatorial effect of miRNAs
  – Coordinated regulation by multiple miRNAs (which may also be co-transcribed in the same pri-miRNA)
• See review by Bartel (Cell 2009) for a discussion of other challenges
Important points

• This type of analysis follows the same basic procedure as a ‘normal’ wetlab scientific experiment
  – Background information
  – Hypothesis / model
  – Controls
  – Validation
  – Modify model and repeat

• Many of the techniques used here are well-known, some are modified

• Availability of complete genomes, scalable algorithms and computational resources crucial to this type of analysis

Knowledge of the biology informs the bioinformatics