A controversial topic...

- Mainly Biology
- Speaker from Computer Science
- Simplification
- Computer Science intuition for concepts
Outline

- Regulation: from eukaryotes to prokaryotes
- RNA regulation
- microRNAs
- Finding human microRNAs
RNA Regulation

• Challenging the dogma: the hidden layer of non-protein-coding RNAs in complex organisms.

• RNA regulation: a new genetics.
The central dogma

DNA  →  RNA

transcription

RNA  →  Protein

translation
The central dogma

- DNA → transcription → RNA → translation → Protein

- Essentially correct for prokaryotes
- Few exceptions: infrastructural RNA (rRNAs, tRNAs)
- Control of gene expression by cis-regulatory elements in the 5’ and 3’ ends
- Gene synonymous with proteins
The central dogma

- Essentially correct for prokaryotes
- Few exceptions: infrastructural RNA (rRNAs, tRNAs)
- Control of gene expression by cis-regulatory elements in the 5’ and 3’ ends
- Gene synonymous with proteins
The central dogma

- Essentially correct for prokaryotes
- Few exceptions: infrastructural RNA (rRNAs, tRNAs)
- Control of gene expression by cis-regulatory elements in the 5’ and 3’ ends
- Gene synonymous with proteins
Previous Assumptions

- Also true in multicellular organisms
  - Complexity explained by combinatorics of regulatory elements
- Proportion of protein-coding sequences declines (1.5% in hum.)
  - either cis-regulatory and structural elements or evolutionary debris
Proportion of non-coding DNA

Proportion of non-coding DNA

Prokaryotes

Proportion of non-coding DNA

Simple eukaryotes

Proportion of non-coding DNA

Combinatorics

DNA 5' 3'
DNA

Transcription factor binding sites

protein coding region
DNA

5’ 3’

Transcription factors

A B C

Transcription factor binding sites

protein coding region
From a CS perspective: rules of the form 
if(A and not (B or C)) then gene transcribed
Programming complex organisms

- Two kinds of programming:
  - Specifying the structural and functional components
  - Specifying higher levels of organization (cells and organs)
  - \( \sim 10^{12} \) positionally different cell types in human
Combinatorics and Complexity

• Can combinatorics explain this complexity?
• Can combinatorics explain this complexity?
  • Yes, generating complexity is easy!
  • Most possibilities are meaningless
  • Evolution needs to find the right ones
  • The question is how to control complexity to specify ordered trajectories that lead to organized complex organisms
How does regulation scale?

- Amount of regulation is a non-linear function of the number of genes.
How does regulation scale?

- Amount of regulation is a non-linear function of the number of genes.
How does regulation scale?

- Amount of regulation is a non-linear function of the number of genes.
How does regulation scale?

- Amount of regulation is a non-linear function of the number of genes.
- New genes need to be regulated => linear increase
How does regulation scale?

• Amount of regulation is a non-linear function of the number of genes.

• New genes need to be regulated => linear increase

Diagram:
- $G_1$ → $G_2$
- $G_2$ → $G_3$
- $G_3$ → $G_6$
- $G_1$ → $G_4$
- $G_4$ → $G_5$
- $G_5$ → $G_6$
How does regulation scale?

- Amount of regulation is a non-linear function of the number of genes.
  - New genes need to be regulated => linear increase
  - Might involve new regulators => need also be regulated
How does regulation scale?

- Amount of regulation is a non-linear function of the number of genes.

- New genes need to be regulated => linear increase

- Might involve new regulators => need also be regulated
How does regulation scale?

- Amount of regulation is a non-linear function of the number of genes.

- New genes need to be regulated => linear increase

- Might involve new regulators => need also be regulated
How does regulation scale?

- Amount of regulation is a non-linear function of the number of genes.
- New genes need to be regulated $\Rightarrow$ linear increase
- Might involve new regulators $\Rightarrow$ need also be regulated
- Has impact on other genes
How does regulation scale?

- Amount of regulation is a non-linear function of the number of genes.
- New genes need to be regulated $\Rightarrow$ linear increase
- Might involve new regulators $\Rightarrow$ need also be regulated
- Has impact on other genes
How does regulation scale?

Number of genes that encode regulatory proteins vs. Total number of genes

LogR vs. LogG graph


RNA regulation: a new genetics.
How does regulation scale?

Double-logarithmic plot, slope = 1.96 => quadratic increase

Number of genes that encode regulatory proteins

Total number of genes


Marc A. Schaub
A limit to combinatorics?

• As the system becomes more complex, an increasing proportion is devoted to regulation.

• At some point, the fractional cost of additional regulation will exceed the benefits of new functions.

• Close to the upper limit of bacterial genome sizes.
A limit to combinatorics?

Evolution of subcellular structures:
- Unicellular world
- Single-cell eukaryotes (protista)
- Eubacteria
- Archaea

Time (mya):
-4,000 → -3,000 → -2,000 → -1,000 → Present
A limit to combinatorics?

Diagram showing the evolution of subcellular structures over time (in mya) with a focus on unicellular world, single-cell eukaryotes (protista), Eubacteria, and Archaea. The graph suggests a question mark for a ceiling.
A limit to combinatorics?

Evolution of RNA-regulatory networks

Entry and expansion of introns, evolution of the splicesome

Unicellular world

Evolution of subcellular structures

Multicellular world

Animals

Plants

Fungi

Multicellular eukaryotes

Sponges

Ceiling?

Complexity

Time (mya)

-4,000 -3,000 -2,000 -1,000 Present

Single-cell eukaryotes (protista)

Eubacteria

Archaea
A limit to combinatorics?

Evolution of RNA-regulatory networks

Entry and expansion of introns, evolution of the splicesome

Eukaryotes found a solution

Ceiling?

Multicellular world

Unicellular world

Evolution of subcellular structures

Animals
Plants
Fungi
Multicellular eukaryotes
Sponges

Single-cell eukaryotes (protista)
Eubacteria
Archaea

Time (mya)

-4,000
-3,000
-2,000
-1,000
Present
Outline

• Regulation: from eukaryotes to prokaryotes
• RNA regulation
• microRNAs
• Finding human microRNAs
Background: Splicing

DNA

transcription

RNA

translation

Protein
Background: Splicing

**DNA**

**transcription**

**RNA**

**Protein**
Background: Splicing

- DNA
- RNA

transcription → Exons

Protein
Background: Splicing

DNA  -------->  RNA  -------->  Protein

transcription

Exons  <---  Introns

Exons  ---  Exons  ---  Exons
Background: Splicing

- DNA
- RNA
- Protein

*transcription*

Exons

Introns
Background: Splicing

DNA → transcription → RNA → translation → Protein

- DNA
- RNA
- Exons
- Introns

transcription

translation
Alternate Splicing

DNA → transcription → RNA → translation → Protein

Exons

Introns

transcription

translation
Alternate Splicing

DNA → RNA → Protein

transcription

Exons

Introns

translation
Alternate Splicing

DNA → Exons → Introns → Protein

transcription

RNA → Exons

translation
Alternate Splicing

- DNA
- RNA
- Protein

Transcription

Translation

- Exons
- Introns

- One gene != one protein
Hypothesis

- 98% of the transcribed sequence are introns
- ~ half of the human genome is transcribed
- Complex eukaryotes have more extensive introns
- Introns assumed to be degraded after splicing
Hypothesis

• 98% of the transcribed sequence are introns
• ~ half of the human genome is transcribed
• Complex eukaryotes have more extensive introns
• Introns assumed to be degraded after splicing
• Could be genetically active
• Would be surprising if evolution had not tried this
Introns and combinatorics

- DNA
- RNA
- Exons
- Introns
- Protein

Transcription

Translation
Introns and combinatorics

DNA → transcription → RNA → translation → Protein

if(A and not (B or C))

Exons

Introns
Introns and combinatorics

DNA → transcription → RNA → translation → Protein

if(A and not (B or C))
Introns and combinatorics

DNA → transcription → RNA → translation → Protein

if(A and not (B or C))

Exons

Introns
Introns and combinatorics

DNA

RNA

Protein

Transcription

Translation

if(A and not (B or C))

Exons

Introns

Condition of another gene:
if(((A and not (B or C)) and D)
Introns and combinatorics

Introns

DNA

RNA

Protein

transcription

translation

if(A and not (B or C))

Exons

Introns

Condition of another gene:
if(((A and not (B or C)) and D) equivalent to
if(intron2 and D)
Regulation revisited

DNA

transcription

RNA

translation

Protein
Regulation revisited

DNA

RNA

Protein

- Alternate splicing
Regulation revisited

- Alternate splicing

DNA → transcription → RNA → splicing → translation → Protein
Regulation revisited

- Alternate splicing
- Introns could have a role in regulation
Regulation revisited

- Alternate splicing
- Introns could have a role in regulation
- RNA regulation not limited to introns
- Alternate splicing
- Introns could have a role in regulation
- RNA regulation not limited to introns
- Not only transcription regulation, also RNA degradation
Regulation revisited

- Alternate splicing
- Introns could have a role in regulation
- RNA regulation not limited to introns
- Not only transcription regulation, also RNA degradation
- Alternate splicing needs to be controlled too
Implications

- Genetic operating system radically different in eukaryotes
- Quantum shift in regulatory sophistication:
  - RNA signal ~ 22 nt
  - 2 orders of magnitude smaller than a protein
Implications

• New definition of gene

• “Transcription unit” (Okazaki et al.)

• “A complete chromosomal segment responsible for making a functional product” (Snyder et al.)

• Alternate splicing: could produce different proteins
Evidence

- Comparative genomics:
  - Significant conservation in non-protein coding sequences
  - Not consistent with neutral drift
  - In other cases, the level of sequence divergence is higher than expected
- Positive selection

I should have gotten coffee before this lecture...
Evidence

- Non-coding transcripts:
  - Several identified in human
  - Linked to diseases
  - Estimate: 7% of all transcripts
  - Tip of the iceberg
  - Later: how to search for microRNA
Evidence

- RNA interference (RNAi):
  - Fragments of double-stranded RNA
  - Post transcriptional silencing
  - Can be used for gene knockdown
  - Nobel Price
Bioinformatic analysis

• Sequence interactions: RNA-RNA, RNA-DNA

• Possibility to identify the elements and their targets from the sequence

• But:
  • not necessarily canonical base-pairing
  • BLAST performs poorly on short sequences
Outline

• Regulation: from eukaryotes to prokaryotes
• RNA regulation
• microRNAs
• Finding human microRNAs
microRNAs

- Noncoding RNAs
- ~22 nucleotides long
- Bind to the 3’ end of an mRNA transcript
- Lead to its degradation
microRNAs

- Transcribed as pri-microRNAs sequences
- Processed in the nucleus into ~70 nt long stem-loop structures
microRNAs

- Processed to mature microRNAs in the cytoplasm
- Interaction with Dicer
- Creates two short RNA molecules
microRNAs

- Processed to mature microRNAs in the cytoplasm
- Interaction with Dicer
- Creates two short RNA molecules
microRNAs

- Processed to mature microRNAs in the cytoplasm
- Interaction with Dicer
- Creates two short RNA molecules
microRNAs

- One molecule integrated into the RNA-induced silencing complex
- Binds to complementary mRNA sequences
- Initiates their degradation
microRNAs

- One molecule integrated into the RNA-induced silencing complex
- Binds to complementary mRNA sequences
- Initiates their degradation
microRNAs

- One molecule integrated into the RNA-induced silencing complex
- Binds to complementary mRNA sequences
- Initiates their degradation
microRNAs

- One molecule integrated into the RNA-induced silencing complex
- Binds to complementary mRNA sequences
- Initiates their degradation
microRNAs

- One molecule integrated into the RNA-induced silencing complex
- Binds to complementary mRNA sequences
- Initiates their degradation
microRNAs and Cancer

- Lu et al. Nature. 2005:
  - Analysis of the expression of 217 microRNAs in 334 samples, including multiple tumors
  - microRNAs surprisingly informative
  - Successful classification of poorly differentiated cancers
  - General downregulation of microRNAs in cancer
Outline

• Regulation: from eukaryotes to prokaryotes
• RNA regulation
• microRNAs
• Finding human microRNAs
Identifying microRNAs

- Identification of hundreds of conserved and nonconserved human microRNAs
  Bentwich et al., Nature Genetics 2005
Previously identified microRNAs

- 222 (as of June 2005):
  - 86 by random cloning and sequencing
  - 43 by computational approaches
  - 93 homologous to microRNAs in other species
microRNAs identified by Bentwich et al.

- 89 new human microRNAs
- 53 that are not conserved beyond primates
- Suggested total number: at least 800
General Approach

• Combine
  • bioinformatic predictions
  • microarray analysis
  • sequence-directed cloning

No. They didn’t like my loop. Maybe I should mutate.
Methods

1. Scan the entire genome for hairpin structures
2. Select hairpins
   1. Stability
   2. Structural Features
3. Microarray sampling
4. Confirmation by sequencing

That's funny. There is a page about me on Wikipedia. But none about Marc!
Methods

1. Scan
   - Scan the entire genome

2. Select
   - 1000 nucleotide windows
   - 150 nucleotide overlap

3. Microarray

4. Sequencing
1. Scan
   - Scan the entire genome
   - Fold each window

2. Select
   RNA structure prediction:
   Vienna RNA package
   Method: free energy minimization
   Now outperformed by ContraFold

3. Microarray

4. Sequencing

Minimum free energy fold
Methods

1. Scan
   - Scan the entire genome
   - Fold each window

2. Select
   RNA structure prediction:
   Vienna RNA package
   Method: free energy minimization
   Now outperformed by ContraFold

3. Microarray

4. Sequencing
Unlike the other programs in our comparison, CONTRAfold's use of the maximum expected accuracy algorithm for parsing... lower (/C0) sensitivity/specificity than CONTRAfold. p-values were calculated using the sign test.

C.B.Do et al.

Unlike the other programs in our comparison, CONTRAfold's use of the maximum expected accuracy algorithm for parsing predictions between the two methods. Sensitivity/ Specificity values were calculated using the sign test.

Methods

1. Scan
   - Scan the entire genome
   - Fold each window

2. Select
   RNA structure prediction: Vienna RNA package
   Method: free energy minimization
   Now outperformed by ContraFold

3. Microarray

4. Sequencing
Methods

1. Scan
   - Scan the entire genome
   - Fold each window

2. Select
   - **Extract hairpins**
     - at least 6 bp
     - at least 55 nt
     - loop < 20 nt

3. Microarray

4. Sequencing
Methods

1. Scan
   - Remove if
   - repetitive elements

2. Select
   - in protein coding regions

3. Microarray

4. Sequencing

I read on Facebook that you’re in a relationship with p53!

It’s already over...
Hairpin conservation:

- measure of evolutionary conservation for each nucleotided using seven species (from UCSC phastCons data)
- Tag hairpin as conserved if average conservation score over any 15-nt sequence in the hairpin stem above a certain threshold
Methods

1. Scan
   - Stability score
   - energetically stable
     \( \iff \)
     appears in many folding configurations

2. Select
   - Use information from the Vienna package

3. Microarray

4. Sequencing
Methods

1. Scan
   - Scoring using structural features:
     - stability score
     - hairpin length
     - loop length
     - free energy per nucleotide
     - number of matching bp
     - bulge size
     - sequence features ...

2. Select

3. Microarray

4. Sequencing
Methods

1. Scan
   • “Optimal predictor”

2. Select
   • Known human microRNA
   • 10,000 hairpins, randomly selected in non-protein-coding regions
   • Find the combination of features that best separates both sets

3. Microarray

4. Sequencing
   • Issues?
Methods

1. Scan

2. Select

3. Microarray

4. Sequencing

- Hairpin scoring algorithm performance

Identification of hundreds of conserved and nonconserved human microRNAs
Bentwich et al., Nature Genetics 2005
Methods

1. Scan

2. Select

3. Microarray

4. Sequencing

- Hairpin scoring algorithm performance

Identification of hundreds of conserved and nonconserved human microRNAs
Bentwich et al., Nature Genetics 2005
Methods

1. Scan

2. Select

• Hairpin scoring algorithm performance

3. Microarray

4. Sequencing

Identification of hundreds of conserved and nonconserved human microRNAs
Bentwich et al., Nature Genetics 2005
Detect expression of microRNA

Multiple tissues (placenta, testis, thymus, brain, prostate)

Two probes per candidate microRNA

Signal ranging from 400 - 65,000

Signal > 2’500 => positive, but not necessarily reliable
• 3000 randomly chosen 35-mers to determine reliability of the signal

• High correlation between the maximal signal and the probe’s cytosine content

• Filter: score > 2500, Cytosine content < 35%

• 6% of the background probes pass => P-value = 0.06
Methods

1. Scan
   - Validate the sequence of the predicted microRNAs using a novel high-throughput sequence-directed cloning method

2. Select
   - Biotinylated capture oligonucleotide designed for the microRNA to be cloned
   - Fish out the complementary sequences

3. Microarray

4. Sequencing
   - Amplify, clone, sequence
Methods

1. Scan
2. Select
3. Microarray
4. Sequencing

Identification of hundreds of conserved and nonconserved human microRNAs
Bentwich et al., Nature Genetics 2005
Methods

1. Scan

2. Select

3. Microarray

4. Sequencing

Identification of hundreds of conserved and nonconserved human microRNAs
Bentwich et al., Nature Genetics 2005
1. Scan

2. Select

3. Microarray

4. Sequencing

Biotinylated capture oligonucleotide

Identification of hundreds of conserved and nonconserved human microRNAs
Bentwich et al., Nature Genetics 2005
Methods

1. Scan

2. Select

3. Microarray

4. Sequencing

Identification of hundreds of conserved and nonconserved human microRNAs
Bentwich et al., Nature Genetics 2005
Identification of hundreds of conserved and nonconserved human microRNAs
Bentwich et al., Nature Genetics 2005
Results

1. Scan ~ 11 million hairpins
2. Select 434,239
   Choose ~5,300
3. Microarray 323 candidates with significant signal
   Choose 169
4. Sequencing 88 validated microRNAs
## Specificity of Selection

1. **Scan**  
   - ~11 million hairpins

2. **Select**  
   - 434,239
   - Choose ~5,300

3. **Microarray**  
   - 323
   - Choose 169

4. **Sequencing**  
   - 88
   - 1 validated miRNA

### Control Group

- Not selected hairpins
  - ~7,500
- Choose 190
- 563
- Choose 190
Sensitivity

1. Scan

2. Select

86% of known microRNA precursors

14% do not meet the criteria, or

have folds that overlap windows

3. Microarray

4. Sequencing
Sensitivity

1. Scan
   86% of known microRNA precursors
   \textit{14% do not meet the criteria, or have folds that overlap windows}

2. Select
   Still 86%
   \textit{What is the issue here?}

3. Microarray

4. Sequencing
Sensitivity

1. Scan
   - 86% of known microRNA precursors
     - 14% do not meet the criteria, or have folds that overlap windows

2. Select
   - Still 86%
     - What is the issue here?

3. Microarray
   - What is the training set?
   - What is the test set?

4. Sequencing
   => we cannot say much about specificity
Results

• Additional sequence validation of 69 adjacent microRNAs
  • adjacent to successfully sequenced microRNAs
  • bioinformatically predicted
  • not present on the microRNA
Results

• Two clusters:
  • 54 new microRNAs on chromosome 19, expressed in placenta
  • 10 microRNAs on chromosome X close near gene FMR1, expressed in testis
• Low conservation score
Results

- 36 conserved microRNAs
- 32 found by later, independent, bioinformatic studies
- 8 validated in these studies

So, what’s the most likely way for Stanford not too loose the Big Game?
How many are there?

- 89 new microRNAs => 311 known in human
- Estimations, based on the method:
  - 442 or 460 conserved miRNAs
  - > 159 nonconserved miRNAs

Hmm... Canceled because of heavy snow? An earthquake?
Conclusion

- Different regulation mechanisms in Eukaryotes than in Prokaryotes
- RNA plays a key role in regulation
- Tip of the iceberg
  => many discoveries still need to be made
  => will require experimental and computational efforts

- Combination of computational and experimental method to find microRNAs
Thank you!

That was funny. They still know so little about us...

At least they know that they don't know.

Yes. Let's worry about Saturday now...