Genomic Signal Processing: The Sequel
From an EE perspective
Professor Gail L. Rosen

Outline of Class
✓ Paper Discussions
✓ Progress Reports
✓ Final Projects

Paper Review
✓ Everyone must read paper and turn in notes about each paper
✓ Discussion Leader: 30 minute presentation on topic (slides or no slides) – turn in presentation if do one
✓ Reviewer will begin “discussion”
✓ The stage will open up for people to ask questions, chime in with further derivations
✓ All to help understanding of paper

Deadlines
✓ Progress Reports due 5/12
✓ Final Project Presentations 6/11 or 6/12 (Wednesday or Thursday evening)
✓ Everyone should have their date… if they do not, email me.

Syllabus Highlights
✓ Project Topics– 4/14
✓ Start thinking about a topic (please feel free to meet with me).
✓ Schedule meetings with me to check feasibility of topic.
✓ Final Project -- Exploratory research project for YOU to learn about the State-of-the-Art in the field (Due. June 11th).

Classic SP for Biology Applications
✓ Most Popular: Speech Signal Processing
✓ Pattern Recognition / Hidden Markov Models: Aligning sequences, classifying similar genes, gene prediction
✓ Boolean Networks: Modeling Genetic Regulatory Networks
How have we historically looked at Biology?

Historical Understanding of Biology

- Beginnings of Medicine: 2000 B.C. (Asia), 500 B.C. (Hippocrates)
- Discovery of DNA: 1950 (Wilkins and Franklin), 1953 (Watson and Crick)
- Feedback Regulation in Metabolism: 1957 (Umbarger, Brown) (Yates, Pardee)
- 1970's: major breakthroughs

DNA Composition

- A – Adenine
- T – Thymine
- C – Cytosine
- G – Guanine

- 4 Nucleotides (bases)
- 3 Bonds for G-C
- 2 Bonds for A-T
- Helical twist

DNA Structure

- In mRNA
- T → U
- Base pairs (bp)
- 1953
- Phosphate Backbone
- G → C
- A → T

Directional Reading

- Third (3') and Fifth (5') Carbon Atoms in Sugar ring.

We have the bases -- now what?

- What is a gene?
Genetic Code

- Marshall Nirenberg (60's) discovers the genetic code
- 3 nucleotides produce one amino acid

Standard Genetic Code

- 64 Codons map to:
  - 20 amino acids and start/stop codons

Genetic codes can vary among species

Transcription

Translation

DNA/RNA

DNA
- Stable: Double-strand, helix
- Function: Stores genetic information
- Replication catalyst: DNA Polymerase -- conducts Proofreading

RNA
- Stable: Single-stranded, many shapes
- Function: Stores information, catalysts, larger structures (tRNA, Ribosomal)
- Replication catalyst: RNA Polymerase -- no proofreading
Genetic Code

Codon Positions

✓ Positions in Open Reading Frames (ORFs) : Biology
Windows/Frames : Signal Processing

Frame Offset
0
1
2

ATGTACACATTGAAAATGA
ATGTACACATTGAAAATGA
ATGTACACATTGAAAATGA

✓ Ribosome “slippage” in gene coding region could mean that a gene may be:
1) Misinterpreted
2) Not stopped
3) Truncated early

Prokaryotes (without nucleus) vs. Eukaryotes (with nucleus)

✓ Transcription and Translation different
✓ Motifs are different
✓ Eukaryotes have nucleus -- transcription occurs inside and translation outside
✓ Eukaryotes: Introns spliced out of mRNA
✓ Prokaryotes: Exons only

mRNA differences – Prokaryote vs. Eukaryote

Prokaryote
Coding region = gene

Eukaryote

“gene” “gene” “gene”

Operon

Long mRNA that codes for several proteins

mRNA differences – Prokaryote vs. Eukaryote

Eukaryote

“coding” “coding” “coding”

gene
gene
gene

Single mRNA -> Single Protein

Prokaryote

Alternative Splicing (Eukaryotes)

(a) Alternative splicing of exons (e.g., primary transcript)

(b) Alternative splicing of intron/exon junctions (e.g., pre-mRNA transcript)

(c) Exons retained mode (e.g., primary transcript)

(d) Exon skipping mode (e.g., pre-mRNA transcript)
Histogram of Prokaryotic protein-lengths

Fitting Hin Data

Protein length (amino acids)
Protein length (amino acids)

Mja, Methanococcus jannaschii; Hin, Haemophilus influenzae; Soc, Saccharomyces cerevisiae; Cel, Caenorhabditis elegans

Gene Length Distributions (Eukaryotes)

What are the comparative genome sizes of humans and other organisms being studied?

<table>
<thead>
<tr>
<th>Organism</th>
<th>Estimated size</th>
<th>Estimated gene number</th>
<th>Average gene density</th>
<th>Chromosomes</th>
</tr>
</thead>
<tbody>
<tr>
<td>Homo sapiens (human)</td>
<td>2900 million bases</td>
<td>30,000 genes</td>
<td>1 gene per 100,000 bases</td>
<td>46</td>
</tr>
<tr>
<td>Rattus norvegicus (rat)</td>
<td>1,775 million bases</td>
<td>30,000 genes</td>
<td>1 gene per 100,000 bases</td>
<td>42</td>
</tr>
<tr>
<td>Mus musculus (mouse)</td>
<td>2500 million bases</td>
<td>30,000 genes</td>
<td>1 gene per 100,000 bases</td>
<td>40</td>
</tr>
<tr>
<td>Drosophila melanogaster (fruit fly)</td>
<td>180 million bases</td>
<td>12,500 genes</td>
<td>1 gene per 9,000 bases</td>
<td>8</td>
</tr>
<tr>
<td>Caenorhabditis elegans (Caenorhabditis)</td>
<td>125 million bases</td>
<td>25,500 genes</td>
<td>1 gene per 4,000 bases</td>
<td>10</td>
</tr>
<tr>
<td>Saccharomyces cerevisiae (yeast)</td>
<td>97 million bases</td>
<td>19,500 genes</td>
<td>1 gene per 6,000 bases</td>
<td>12</td>
</tr>
<tr>
<td>Escherichia coli (bacterium)</td>
<td>4.7 million bases</td>
<td>3200 genes</td>
<td>1 gene per 1,000 bases</td>
<td>1</td>
</tr>
<tr>
<td>H. influenzae (bacterium)</td>
<td>1.8 million bases</td>
<td>1700 genes</td>
<td>1 gene per 1,000 bases</td>
<td>1</td>
</tr>
</tbody>
</table>

Genome size does not correlate with evolutionary status, nor is the number of genes proportionate with genome size. Largest Genome size - Proteus vulgaris, Marbled lungfish @ 130 Gbp

Genome size increases exponentially, but not the number of genes

<table>
<thead>
<tr>
<th>Organism</th>
<th>Year</th>
<th>Millions bases sequenced</th>
<th>Predicted genes</th>
<th>Genes per million bases</th>
</tr>
</thead>
<tbody>
<tr>
<td>H. influenzae</td>
<td>1955</td>
<td>1.8</td>
<td>1,800</td>
<td>1030</td>
</tr>
<tr>
<td>E. coli</td>
<td>1997</td>
<td>4.6</td>
<td>4,300</td>
<td>900</td>
</tr>
<tr>
<td>M. genitalium</td>
<td>1995</td>
<td>0.58</td>
<td>468</td>
<td>806</td>
</tr>
<tr>
<td>S. cerevisiae</td>
<td>1996</td>
<td>12</td>
<td>5,800</td>
<td>483</td>
</tr>
<tr>
<td>C. elegans</td>
<td>1998</td>
<td>97</td>
<td>19,000</td>
<td>195</td>
</tr>
<tr>
<td>D. melanogaster</td>
<td>2005</td>
<td>116</td>
<td>14,100</td>
<td>122</td>
</tr>
<tr>
<td>A. thaliana</td>
<td>2000</td>
<td>115</td>
<td>25,500</td>
<td>220</td>
</tr>
<tr>
<td>Human Chr. 21</td>
<td>2000</td>
<td>34</td>
<td>226</td>
<td>7</td>
</tr>
<tr>
<td>Human Chr. 22</td>
<td>1999</td>
<td>34</td>
<td>545</td>
<td>16</td>
</tr>
<tr>
<td>Human (draft)</td>
<td>2001</td>
<td>2,693</td>
<td>31,780</td>
<td>12</td>
</tr>
<tr>
<td>Human (finished)</td>
<td>2004</td>
<td>2,850</td>
<td>25,000</td>
<td>9</td>
</tr>
<tr>
<td>Mouse (draft)</td>
<td>2002</td>
<td>2,372</td>
<td>30,000</td>
<td>13</td>
</tr>
</tbody>
</table>
C+G Content Distribution in Humans

Why are eukaryotic genomes larger?

- Gene structure
- No operons (genetic elements to produce mRNA)
- Introns
- Extensive 5' and 3' regulatory regions
- Spacer regions
- Repetitive DNA
- Non-coding functional DNA

Genes - only ~3% of Human Genome

- What is in between? Signal sites for ribosomes but currently speculated as “junk” (insertions from viruses and leftover elements that got weeded out due to mutation.
- Where are the Genes?
  - Start/Stop codons are a clue (not always)
  - Transition probabilities (HMM’s successful)
  - Use genes from other species to find genes in a new species?

Haploid vs. Diploid genomes

- E. Coli: Only one copy of each gene (Haploid)
- Humans: Two copies of each gene (Diploid) … Pairs of Chromosomes

E. Coli shows new phenotypes sooner (Adapts faster to its environment)

Bacterial Mating

- Plasmids

Insertion of DNA by Virii
Transposons

- Move from place to place within DNA
- "Transposable Elements"
- "Jumping Genes" or Mobile Elements

Nucleotides to Proteins

- Genetic Code: Nucleotides (Gene) -> Amino Acids
- Amino Acids -> Proteins (many-to-one mapping)

Four Structures towards Proteins

- Primary Structure: Polypeptides
- Secondary Structure: Alpha helices and Beta sheets
- Tertiary Structure: 3D
- Quaternary: Multiple polypeptides

Challenges in Protein Structure Prediction

- The number of possible structures is extremely large
- The physical basis of protein structural stability is not fully understood
- The amino acid sequence itself may not fully specify the tertiary structure. (Environment)
- Simulating protein folding is computationally intensive (thus the Folding@Home project)

OK - we have a protein. Now what?

- Protein-Protein Interactions (Gene Regulatory Networks)
  - Enzymes, proteins, etc. trigger the production of proteins
  - How do they know when?!
- Protein-Protein Interactions (Metabolic Pathways)
  - Energy Production/ Glycolysis
  - Pathways building more complex functions
- Gene Expression Analysis (how much protein generated from a gene)

The Genetic Machinery (Central Dogma of Genomics)

- DNA -> Genetic Code -> Protein
The Context for Gene Regulation
(Nucleus from Skin Cell creates tadpole)

Unfertilized Egg → UV Light to Destroy Nucleus → Insert Nucleus from Frog Skin Cells → Fertilize → Tadpole

Two types of gene classes

Genes Common to all Cells
- Houskeeping Genes
  - DNA Polymerase
  - RNA Polymerase
  - Ribosomes

Genes characteristic to cell
- E.g. Red Blood Cell Hemoglobin
- E.g. Insulin Pancreas

Cell Control of Protein Generation
- Controlling when and how often a given gene is transcribed
- Controlling how the primary transcript is spliced or otherwise processed
- Selecting which mRNAs are activated by the proteins
- Activating or Inactivating Proteins

What we know about the regulation

- Regulatory Region Promoter ~50bp+Initiation Site
- 10 bp in Prokaryotes
- 10000 bp in Eukaryotes

These sites take several "input" gene regulatory proteins
- Repressors turn off transcription
- Activators turn on transcription

 Constitutive Gene Expression (Tryptophan example)

Promoter → Tryptophan Operon

Will bind and block RNA Polymerase
If enough Tryptophan Binds

Tryptophan is always present in low-levels (continual expression) – Constitutive
Tryptophan is an essential amino acid for the human diet

Induced/Facultative Gene Expression

- Induced Gene Expression
  - Responding to Stimulus Examples:
    - SIEGE: Smoking Induced Epithelial Gene Expression Database (on airways)
    - Interferon-induced gene expression (Interferon released only upon "intruder" to vertebrate’s immune system)

Tryptophan is an essential amino acid for the human diet
Prokaryotic vs. Eukaryotic Transcription

- Prokaryotes – only need to generate RNA Polymerase to initiate transcription
- Eukaryotes – very complicated
  - Need to generate 3 RNA polymerases AND
  - Transcription Factors
  - Transcription factors help place RNA polymerase at promoter, pull apart DNA strands, and release RNA polymerase when done
  - Because of “bends”, regulatory regions may be 1000 of bases away and control transcription initiation

P vs. E Transcription Initiation

- Prokaryotes:
  - Bacterial genes controlled by a single activator or repressor protein
  - Expression level coordinated by single protein
- Eukaryotes:
  - Committee of regulatory proteins – thought of as combinatorial control or Boolean logic
  - Expression level controlled by single protein (if protein is last element needed in combinatorial control)

Eyeless (or too many), Ey, Gene

- Expression of Ey in Development of wings, Legs, etc.
- A) Scanning electron micrograph of an ectopic eye (arrowhead) in the head region formed by the antennal disc.
- B) Overview of a fly with an ectopic eye under the wing (arrow) and on the antenna (arrowhead).
- C) Higher magnification of (A). The ectopic eye (to the left) contains hexagonal ommatidia and interommatidial bristles. The organization of the facets in the ectopic eye is very similar to the pattern in the normal eye (to the right). Some facets, however, are fused and some irregularities are observed.
- D) Higher magnification of the ectopic eye under the wing shown in (B) (arrow). The ectopic eye protrudes out of the thoracic body wall (ventral pleura). The organization of the facets and interommatidial bristles is similar to that of the ectopic eye shown in (C).

Pathway Diagrams: Qualitative

- Genetic Regulatory Network
- Metabolic Pathways (Glycolysis)

Microarrays: Quantitative attempts for Gene Expression

- E. Imaging

Comparative Genomics

- Phylogeny
Tying it all together

<table>
<thead>
<tr>
<th>Structure analysis</th>
<th>Sequence analysis</th>
<th>Function analysis</th>
</tr>
</thead>
<tbody>
<tr>
<td>nucleic acid structure prediction</td>
<td>genome comparison</td>
<td>metabolic pathway modeling</td>
</tr>
<tr>
<td>protein structure prediction</td>
<td>phylogeny</td>
<td>gene expression profiling</td>
</tr>
<tr>
<td>protein structure classification</td>
<td>gene &amp; promoter prediction</td>
<td>protein interaction prediction</td>
</tr>
<tr>
<td>protein structure comparison</td>
<td>motif discovery</td>
<td>protein subcellular localization prediction</td>
</tr>
<tr>
<td></td>
<td>sequence database searching</td>
<td></td>
</tr>
<tr>
<td></td>
<td>sequence alignment</td>
<td></td>
</tr>
</tbody>
</table>

Application Areas for Projects

Part II: Databases

- Nucleotide Sequences (Genbank, EMBL, etc.)
- Protein sequences (Uniprot)
- Human Sequence patterns (STRBase)
- Macromolecular 3D structure (MMDB)
- Gene Expression data (Omnibus, NCI, Stanford, etc.)
- Metabolic Pathways (KEGG, MetaCyc, etc.)

A brief history of biological databases

1965 M. O. Dayhoff *et al.* publish “Atlas of Protein Sequences and Structures”
1982 EMBL initiates DNA sequence database, followed within a year by GenBank (then at LANL) and in 1984 by DNA Database of Japan
1988 EMBL/GenBank/DDBJ agree on common format for data elements
2002 Gene Expression Omnibus


- Created in 1988 as part of the National Library of Medicine at NIH
- Establish public databases
- Research in computational biology
- Develop software tools for sequence analysis
- Disseminate biomedical information

Growth of GenBank database

1965 M. O. Dayhoff *et al.* publish “Atlas of Protein Sequences and Structures”
1982 EMBL initiates DNA sequence database, followed within a year by GenBank (then at LANL) and in 1984 by DNA Database of Japan
1988 EMBL/GenBank/DDBJ agree on common format for data elements
2002 Gene Expression Omnibus


- Created in 1988 as part of the National Library of Medicine at NIH
- Establish public databases
- Research in computational biology
- Develop software tools for sequence analysis
- Disseminate biomedical information

Growth of GenBank database