**Structural Genomics**

Iosif Vaisman

Email: ivaisman@gmu.edu

---

**From genes to proteins**

From genes to proteins

Computational Gene Prediction

- Where the genes are unlikely to be located?
- How do transcription factors know where to bind a region of DNA?
- Where are the transcription, splicing, and translation start and stop signals?
- What does coding region do (and non-coding regions do not)?
- Can we learn from examples?
- Does this sequence look familiar?

---

Measures of Prediction Accuracy

<table>
<thead>
<tr>
<th>REALITY</th>
<th>PREDICTION</th>
</tr>
</thead>
<tbody>
<tr>
<td>TN</td>
<td>TP</td>
</tr>
<tr>
<td>FP</td>
<td>FN</td>
</tr>
</tbody>
</table>

Sensitivity: $S_n = \frac{TP}{TP + FN}$

Specificity: $S_p = \frac{TN}{TN + FP}$

---

<table>
<thead>
<tr>
<th>REALITY</th>
<th>PREDICTION</th>
</tr>
</thead>
<tbody>
<tr>
<td>TP</td>
<td>FP</td>
</tr>
<tr>
<td>FN</td>
<td>TN</td>
</tr>
</tbody>
</table>

---

**Chromosome 19 gene map**

**DNA**

**RNA**

**mRNA**

**PROTEIN**

**TRANSCRIPTION**

**SPlicing**

**TRANSLATION**

**PROMOTER ELEMENTS**

**SPLICE SITES**

**START CODON**

**STOP CODON**

**SPLICE SITES**

**PROTEIN**

**From genes to proteins**

**Chromosome 19 gene map**

**DNA**

**RNA**

**mRNA**

**PROTEIN**

**TRANSCRIPTION**

**SPlicing**

**TRANSLATION**

**PROMOTER ELEMENTS**

**SPLICE SITES**

**START CODON**

**STOP CODON**

**SPLICE SITES**

**PROTEIN**
Measures of Prediction Accuracy

**Exon Level**

<table>
<thead>
<tr>
<th>REALITY</th>
<th>PREDICTION</th>
</tr>
</thead>
<tbody>
<tr>
<td>WRONG EXON</td>
<td>CORRECT EXON</td>
</tr>
</tbody>
</table>

Sensitivity

\[
S_n = \frac{\text{number of correct exons}}{\text{number of actual exons}}
\]

Specificity

\[
S_p = \frac{\text{number of correct exons}}{\text{number of predicted exons}}
\]

Spliced Alignment (Procrustes)

- New genomic sequence
- Selection of candidate exons
  - AUG → GU initial exons
  - AG → GU internal exons
  - AG → UAA or UAG or UGA terminal exons
- Filtration (based on the codon usage statistics)
- Construction of all possible chains of candidate exons
- Finding a chain with the maximum global similarity to the target protein

Predicted Exon Assembly (Procrustes)

PCR Primers Prediction (GenePrimer)

Exon 1085..1182 (98) hit using first 2 primers
Exon 1628..1676 (49) missed
Exon 1900..2001 (102) hit using first 8 primers
Exon 2110..2184 (75) missed
Exon 2516..2722 (207) hit using first 4 primers
Exon 3385..3472 (88) missed
Exon 3546..3746 (201) hit using first primer
...

GRAIL gene identification program
**Suboptimal Solutions for the Human Growth Hormone Gene (GeneParser)**

![Graph](image)

**GeneMark Accuracy Evaluation**

![Graph](image)

**Sequence-structure correlations**

![Graph](image)

**Model structure coverage in sequence space**

![Graph](image)

**Structural Genomics Project**

- Organize known protein sequences into families.
- Select family representatives as targets.
- Solve the 3D structure of targets by X-ray crystallography or NMR spectroscopy.
- Build models for other proteins by homology to solved 3D structures.

**History of Structural Genomics**

<table>
<thead>
<tr>
<th>Year</th>
<th>Event</th>
</tr>
</thead>
<tbody>
<tr>
<td>1995</td>
<td>SG project proposed in Japan</td>
</tr>
<tr>
<td>1997</td>
<td>SG pilot project starts at RIKEN, Japan</td>
</tr>
<tr>
<td>1998/99</td>
<td>Initial SG projects start in Canada, Germany, US</td>
</tr>
<tr>
<td>2000</td>
<td>OECD/ISTP/PSF, Paris, France - Further Study on SG</td>
</tr>
<tr>
<td>2002</td>
<td>National project on Protein Structural and Functional Analyses starts in Japan</td>
</tr>
</tbody>
</table>

*Heinemann, 2002*
Goals of structural genomics

- Provision of enough structural templates to facilitate homology modeling of most proteins
- Structures of all proteins in a complete proteome
- Structural elucidation of a complete biological pathway
- Structural elucidation of a complete disease

Target selection

- realm of interest
- family exclusion - impossible
- family exclusion - known
- prioritization
- selection
- analysis and interpretation

Coverage of the Human Genome By Structure

Structural genomics shortcuts

M. thermoautotrophicum structural genomics project

Structural genomics target database