From genes to proteins

**Encyclopedia of DNA Elements (ENCODE)**

Gene definitions

- Definition 1910s: Gene as a distinct locus
- Definition 1940s: Gene as a blueprint for a protein
- Definition 1950s: Gene as a physical molecule
- Definition 1960s: Gene as transcribed code
- Definition 1970s–1980s: Gene as open reading frame (ORF) sequence pattern
- Definition 1990s–2000s: Annotated genomic entity, enumerated in the databanks (current view, pre-ENCODE)
- A current computational metaphor: Genes as “subroutines” in the genomic operating system
Gene concept problems

ENCODE definition of gene

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?
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

Suboptimal Solutions for the Human Growth Hormone Gene (GeneParser)

Measures of Prediction Accuracy

Nucleotide Level

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

Sensitivity $S_n = TP / (TP + FN)$
Specificity $S_p = TP / (TP + FP)$

Exon Level

<table>
<thead>
<tr>
<th>REALITY</th>
<th>PREDICTION</th>
</tr>
</thead>
<tbody>
<tr>
<td>REALITY</td>
<td>PREDICTION</td>
</tr>
</tbody>
</table>

Sensitivity $S_n = \frac{\text{number of correct exons}}{\text{number of actual exons}}$
Specificity $S_p = \frac{\text{number of correct exons}}{\text{number of predicted exons}}$

Transposons

mRNA
normal splicing
3' pre-mRNA
alternative splicing

Zvelebil & Baum, 2007
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
Model structure coverage in sequence space

![Image](image.png)

Adopted from Volkov et al., 2001

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.

Target selection

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

Coverage of the Human Genome By Structure

![Diagram](diagram.png)

Xie and Bourne, 2005

PDB

Structural Genomics Targets

Ensembl Human Genome Annotation

Superfamily