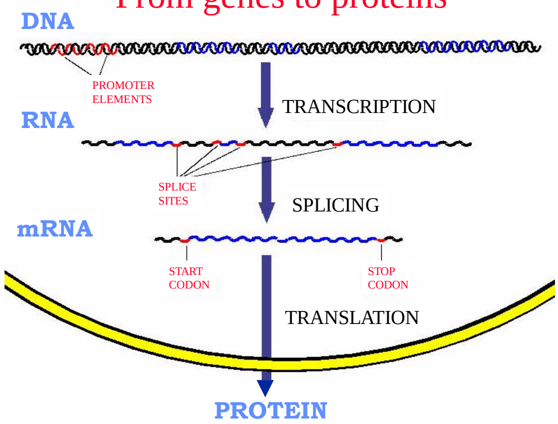
Bioinformatics Methods

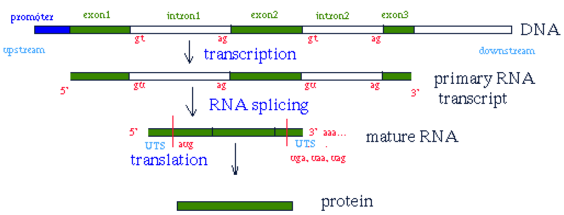
Iosif Vaisman

Email: ivaisman@gmu.edu

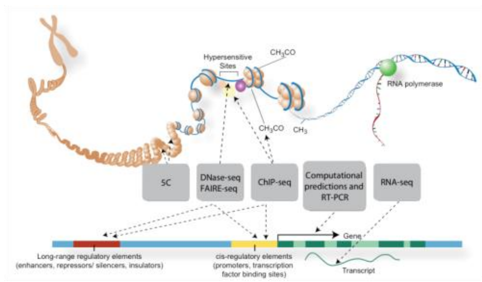
From genes to proteins



From genes to proteins

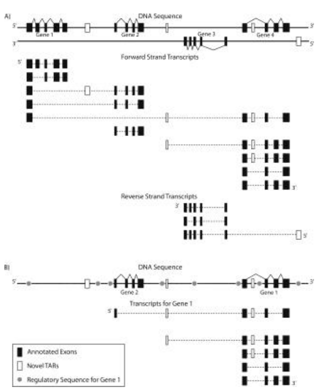


Encyclopedia of DNA Elements (ENCODE)



Darryl Leja (NHGRI), Ian Dunham (EBI), http://genome.ucsc.edu/ENCODE/

Genomic region (ENCODE)



Gerstein M B et al. Genome Res. 2007

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

Gerstein M B et al. Genome Res. 2007

Gene concept problems

Table 1. Phenomena complicating the concept of the gene

Phenomenon	Description	Issue
Gene location and structure		
Intronic genes	A gene exists within an intron of another (Henikoff et al. 1986)	Two genes in the same locus
Genes with overlapping reading frames	A DNA region may code for two different protein products in different reading frames (Contreras et al. 1977)	No one-to-one correspondence between DNA and protein sequence
Enhancers, silencers	Distant regulatory elements (Spilianakis et al. 2005)	DNA sequences determining expression can be widely separated from one another in genome. Many-to-many relationship between genes and their enhancers.
Structural variation		
Mobile elements	Genetic element appears in new locations over generations (McClintock 1948)	A genetic element may be not constant in its location
Gene rearrangements/structural variants	DNA rearrangement or splicing in somatic cells results in many alternative gene products (Early et al. 1980)	Gene structure is not hereditary, or structure may differ across individuals or cells/tissues
Copy-number variants	Copy number of genes/regulatory elements may differ between individuals (Iafrate et al. 2004; Sebat et al. 2004; Tuzun et al. 2005)	Genetic elements may differ in their number
Epigenetics and chromosome structure		
Epigenetic modifications, imprinting	Inherited information may not be DNA-sequence based (e.g., Dobrovic et al. 1988); a gene's expression depends on whether it is of paternal or maternal origin (Sager and Kitchin 1975)	Phenotype is not determined strictly by genotype
Effect of chromatin structure	Chromatin structure, which does influence gene expression, only loosely associated with particular DNA sequences (Paul 1972)	Gene expression depends on packing of DNA. DNA sequence is not enough to predict gene product.

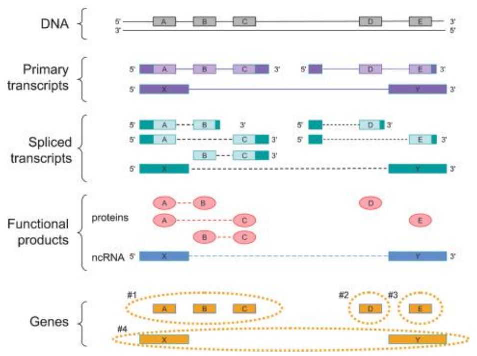
Gerstein M B et al. Genome Res. 2007

Gene concept problems

Post-transcriptional events		
Alternative splicing of RNA	One transcript can generate multiple mRNAs, resulting in different protein products (Berget et al. 1977; Gelinas and Roberts 1977)	Multiple products from one genetic locus; information in DNA not linearly related to that on protein
Alternatively spliced products with alternate reading frames	Alternative reading frames of the INK4a tumor suppressor gene encodes two unrelated proteins (Quelle et al. 1995)	Two alternative splicing products of a pre-mRNA produce protein products with no sequence in common
RNA *trans*-splicing, homotypic *trans*-splicing	Distant DNA sequences can code for transcripts ligated in various combinations (Borst 1986). Two identical transcripts of a gene can *trans*-splice to generate an mRNA where the same exon sequence is repeated (Takahara et al. 2000).	A protein can result from the combined information encoded in multiple transcripts
RNA editing	RNA is enzymatically modified (Eisen 1988)	The information on the DNA is not encoded directly into RNA sequence
Post-translational events		
Protein splicing, viral polyproteins	Protein product self-cleaves and can generate multiple functional products (Villa-Komaroff et al. 1975)	Start and end sites of protein not determined by genetic code
Protein *trans*-splicing	Distinct proteins can be spliced together in the absence of a *trans*-spliced transcript (Handa et al. 1996)	Start and end sites of protein not determined by genetic code
Protein modification	Protein is modified to alter structure and function of the final product (Wold 1981)	The information on the DNA is not encoded directly into protein sequence
Pseudogenes and retrogenes		
Retrogenes	A retrogene is formed from reverse transcription of its parent gene's mRNA (Vanin et al. 1980) and by insertion of the DNA product into a genome	RNA-to-DNA flow of information
Transcribed pseudogenes	A pseudogene is transcribed (Zheng et al. 2005, 2007)	Biochemical activity of supposedly dead elements

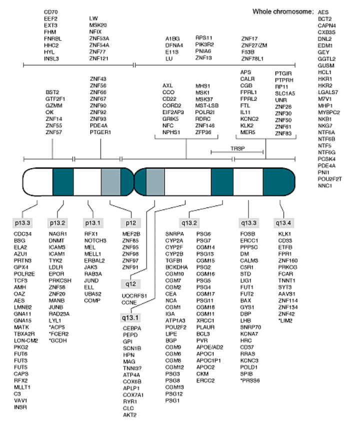
Gerstein M B et al. Genome Res. 2007

ENCODE definition of gene



Gerstein M B et al. Genome Res. 2007

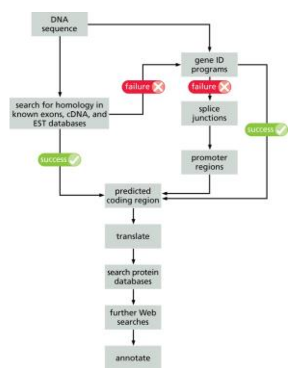
Chromosome 19 gene map



Computational Gene Prediction

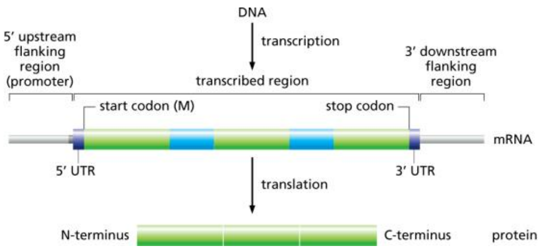
- Where the genes are unlikely to be located?
- How do transcription factors know where to bind a region of DNA?
- Where are the transcription, splicing, and translation start and stop signals?
- What does coding region do (and non-coding regions do not) ?
- Can we learn from examples?
- Does this sequence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

Computational Gene Prediction



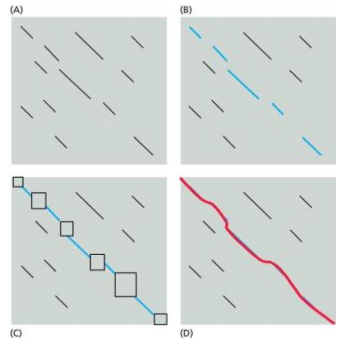
Zvelebil & Baum, 2007

Computational Gene Prediction



Zvelebil & Baum, 2007

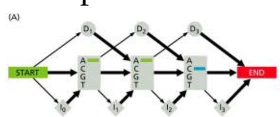
Computational Gene Prediction



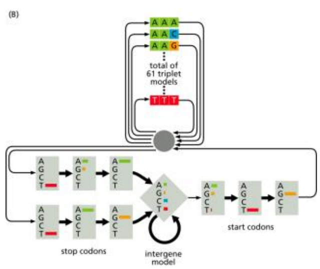
LAGAN algorithm

Zvelebil & Baum, 2007

Computational Gene Prediction



EcoParse Hidden Markov Model for individual codon



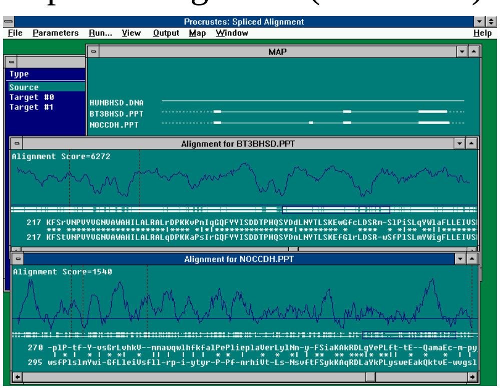
Complete HMM

Zvelebil & Baum, 2007

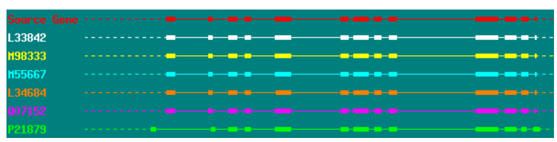
Spliced Alignment (Procrustes)

- 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 usge statistics)
- Construction of all possible chains of candidate exons
- Finding a chain with the maximum global similarity to the target protein

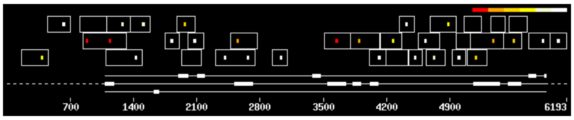
Spliced Alignment (Procrustes)



Predicted Exon Assembly (Procrustes)

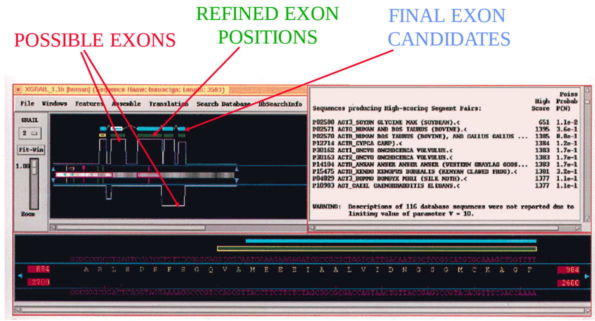


PCR Primers Prediction (GenePrimer)

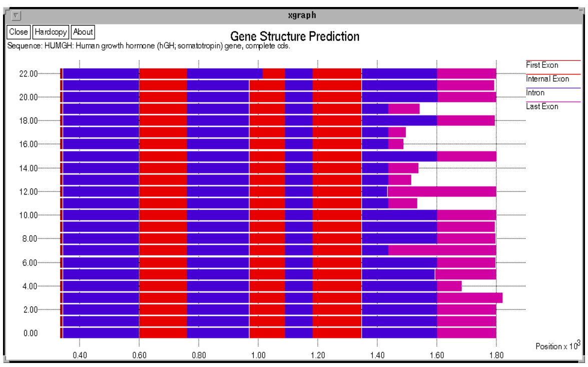


Exon 1085..1182 (98) hit using first 2 primers
Exon 1628..1676 (49) missed
Exon 1900..2001 (102) hit using first 8 primers
Exon 2110..2184 (75) missed
Exon 2516..2722 (207) hit using first 4 primers
Exon 3385..3472 (88) missed
Exon 3546..3746 (201) hit using first primer
...

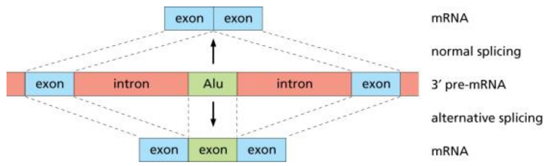
GRAIL gene identification program



Suboptimal Solutions for the Human Growth Hormone Gene (GeneParser)



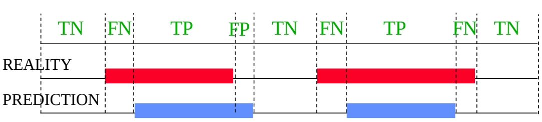
Transposons

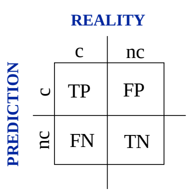


Zvelebil & Baum, 2007

Measures of Prediction Accuracy

Nucleotide Level





Sensitivity

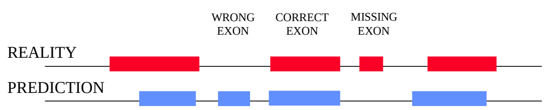
$S_n = TP / (TP + FN)$

Specificity

$S_p = TP / (TP + FP)$

Measures of Prediction Accuracy

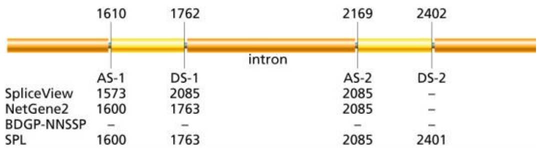
Exon Level



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

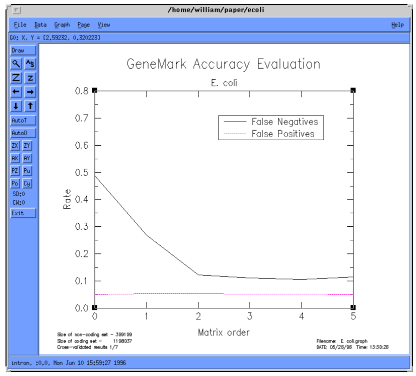
Specificity $S_p = \frac{\text{number of correct exons}}{\text{number of predicted exons}}$

Computational Gene Prediction

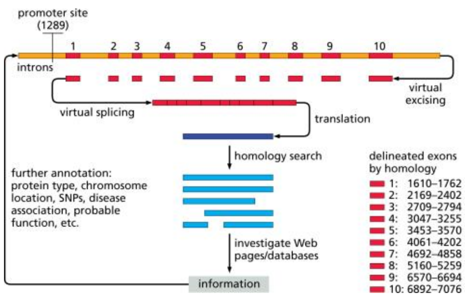


Zvelebil & Baum, 2007

GeneMark Accuracy Evaluation

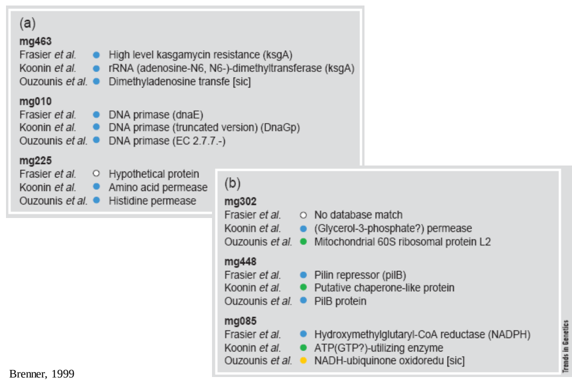


Gene Annotation



Zvelebil & Baum, 2007

Errors in genome annotation



Brenner, 1999

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

Phil Bourne, 2005

Sequence-structure correlations

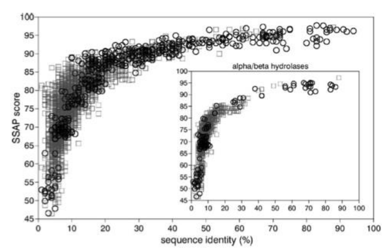
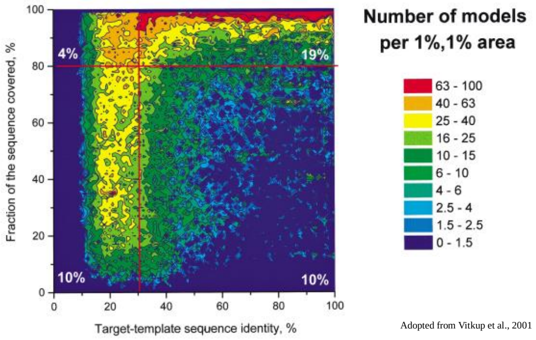


Fig. 1. Correlation between structure similarity (measured by the SSAP structure comparison algorithm, 0–100) and sequence similarity (measured by sequence identity) for all pairs of homologous domain structures in the CATH domain database.

Redfern and Orengo, 2005

Model structure coverage in sequence space

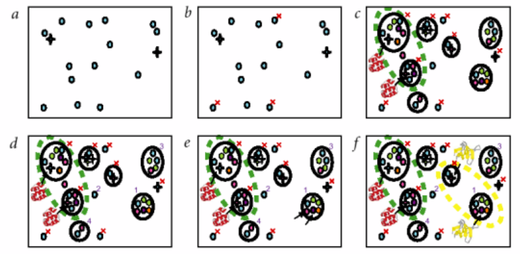


Adopted from Vitkup et al., 2001

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.

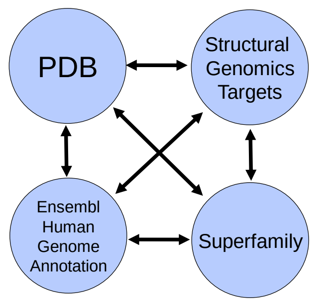
Target selection



a) realm of interest
b) family exclusion - impossible
c) family exclusion - known
d) prioritization
e) selection
f) analysis and interpretation

S.Brenner, 2000

Coverage of the Human Genome By Structure



Xie and Bourne, 2005