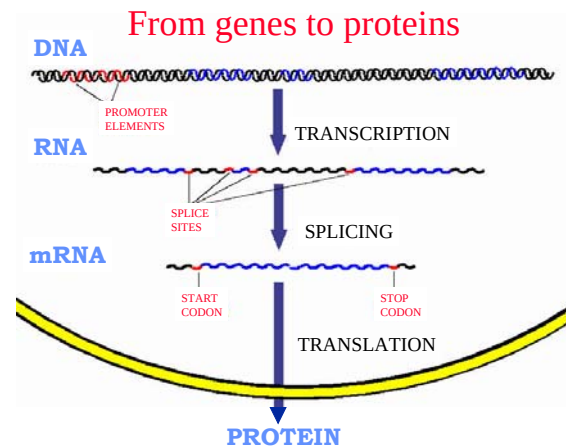


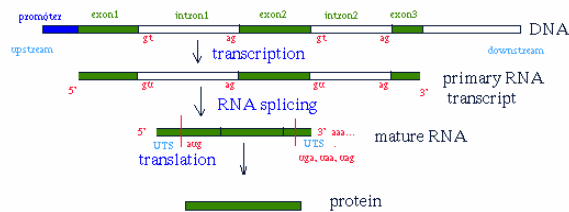
Structural Genomics

Iosif Vaisman

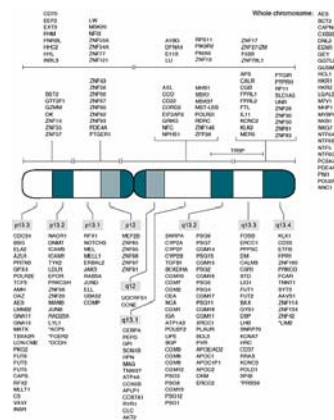
Email: ivaisman@gmu.edu



From genes to proteins



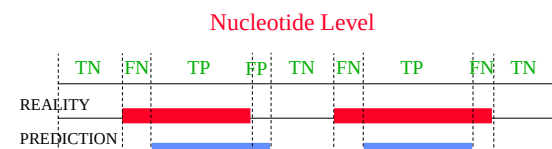
Chromosome 19 gene map



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?

Measures of Prediction Accuracy

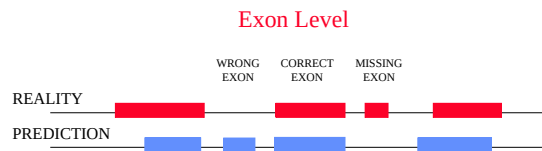


	REALITY	
	c	nc
PREDICTION	c	TP FP
	nc	FN TN

$$\text{Sensitivity } S_n = TP / (TP + FN)$$

$$\text{Specificity } S_p = TP / (TP + FP)$$

Measures of Prediction Accuracy



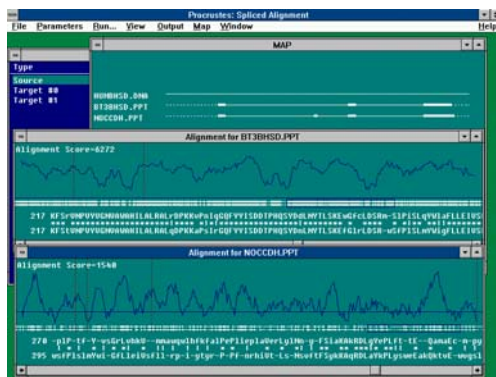
Sensitivity $S_n = \frac{\text{number of correct exons}}{\text{number of actual exons}}$

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

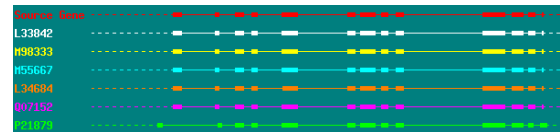
Spliced Alignment (Procrustes)

- 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

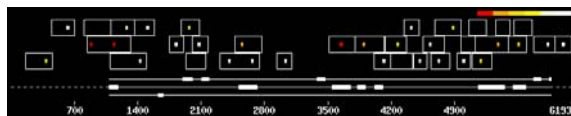
Spliced Alignment (Procrustes)



Predicted Exon Assembly (Procrustes)



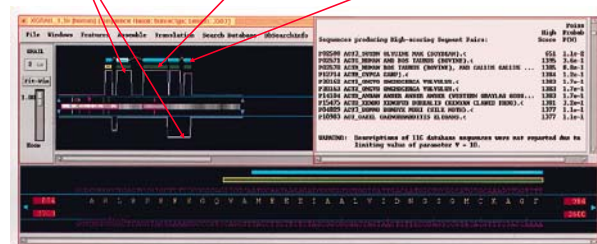
PCR Primers Prediction (GenePrimer)



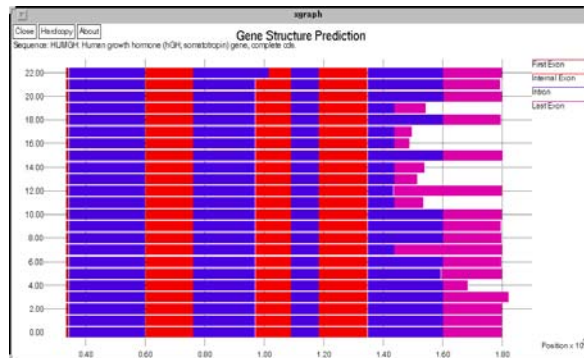
Exon 1085..1182 (98) hit using first 2 primers
 Exon 1628..1676 (49) missed
 Exon 1900..2001 (102) hit using first 8 primers
 Exon 2110..2184 (75) missed
 Exon 2516..2722 (207) hit using first 4 primers
 Exon 3385..3472 (88) missed
 Exon 3546..3746 (201) hit using first primer
 ...

GRAIL gene identification program

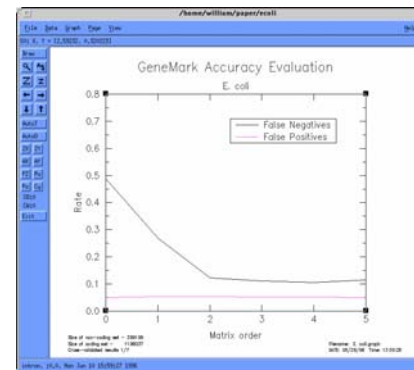
POSSIBLE EXONS REFINED EXON POSITIONS FINAL EXON CANDIDATES



Suboptimal Solutions for the Human Growth Hormone Gene (GeneParser)



GeneMark Accuracy Evaluation



Sequence-structure correlations

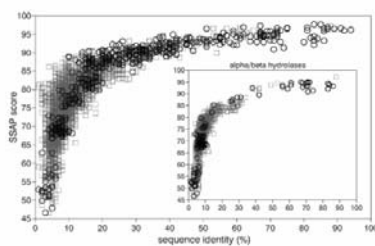
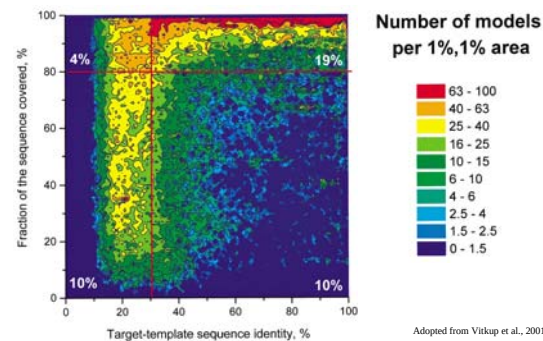


Fig. 1. Correlation between structure similarity (measured by the SSAP structure comparison algorithm, 0-100) and sequence similarity (measured by sequence identity) for all pairs of homologous domain structures in the CATH domain database.

Redfern and Orengo, 2005

Model structure coverage in sequence space



Adopted from Viskup 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.

History of Structural Genomics

1995	SG project proposed in Japan	2000 Sep. NIGMS Protein Structure Initiative starts in US with 7 Centers
1997 Apr.	SG pilot project starts at RIKEN Inst.	2000 Nov. International Conference on SG (ICSG 2000) , Yokohama, Japan / International SG Task Force Meeting / OECD/GSF Meeting
1997	SG studies initiated through DOE, NIGMS in US	2001 Jan. OECD/CSTP/GSF, Paris, France – Further Study on SG
1998/99	Initial SG projects start in Canada, Germany, US	2001 Apr. 2nd International SG Meeting , Airlie House, US – Start of ISGO
1999 June	Call for SG pilot projects issued by NIGMS/NIH	2001 Sep. NIGMS Protein Structure Initiative adds 2 new centers
2000 Jan.	OECD Committee for Scientific and Technological Policy (CSTP) proposes to initiate SG studies	2002 Mar. European Commission announces funding of Structural Proteomics in Europe (SPINE)
2000 Apr.	1st International SG Meeting , Hinxton, UK	2002 Apr. National project on Protein Structural and Functional Analyses starts in Japan
2000 June	OECD/Global Science Forum (GSF) and SG Workshop , Florence, Italy	2002 Oct. ISGO International Conference on SG (ICSG 2002) , Berlin, Germany
2000 Sep.	SG: From Gene to Structure to Function , Cambridge, UK	

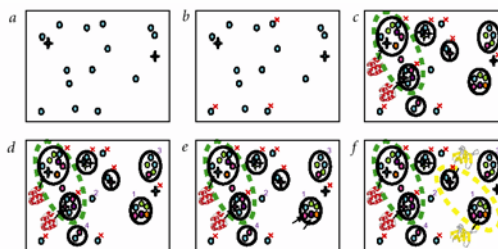
Heinemann, 2002

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

Phil Bourne, 2005

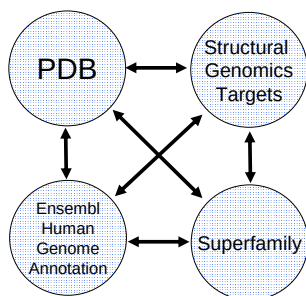
Target selection



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

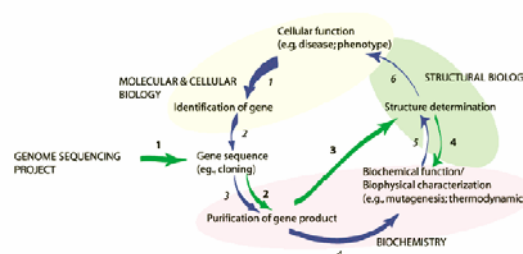
S.Brenner, 2000

Coverage of the Human Genome By Structure



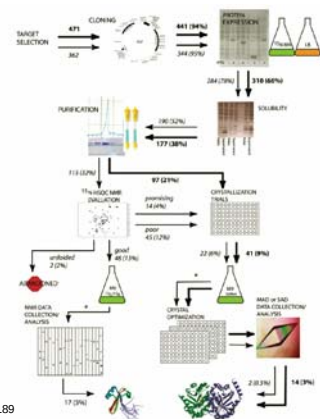
Xie and Bourne, 2005

Structural genomics shortcuts



Yee et al., Acc. Chem. Res. 2003, 36, 183-189

M. thermoautotrophicum structural genomics project



Yee et al., Acc. Chem. Res. 2003, 36, 183-189

Structural genomics target database

