Biol 478/595 Intro to Bioinformatics

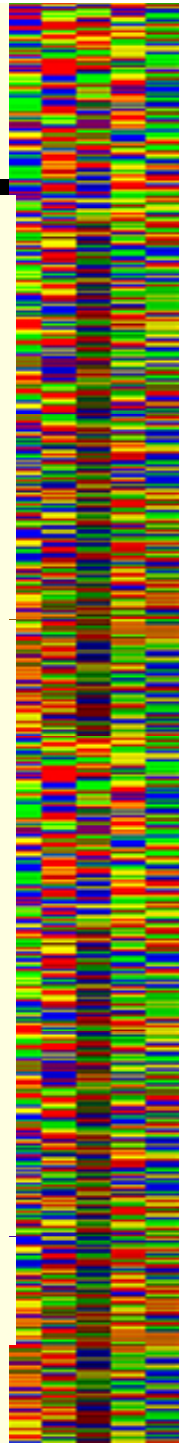
September					
	M 1		Labor Day		
4	W 3	MG	Database Searching		Ch. 6
5	F 5	MG	Database Searching	Hw1	
6	M 8	MG	Scoring Matrices		Ch 3 and Ch 4
7	W 10	MG	Pairwise Alignment		
8	F 12	MG	Pairwise Alignment	Hw2	
9	M 15	MG	Pairwise Alignment		Ch 9
10	W 17	MG	Genome Sequencing		
11	F 19	MG	Gene Finding/Annotation	Hw3	
12	M 22	MG	Gene Finding/Annotation		Ch. 11
13	W 24	MG	Quiz – Gene Finding/Annotation		Ch. 7
14	F 26	MG	Sequence Motifs	No Hw	
X	M 29	Both	Exam		

- ***Homework 4 is not due until 10/3***
- ***Midterm next Monday***
  - ***Monday, September 29, 2008 8:00 PM to 10:00 PM LILY, G432***

Genomics - Gene Modeling

Search by Content

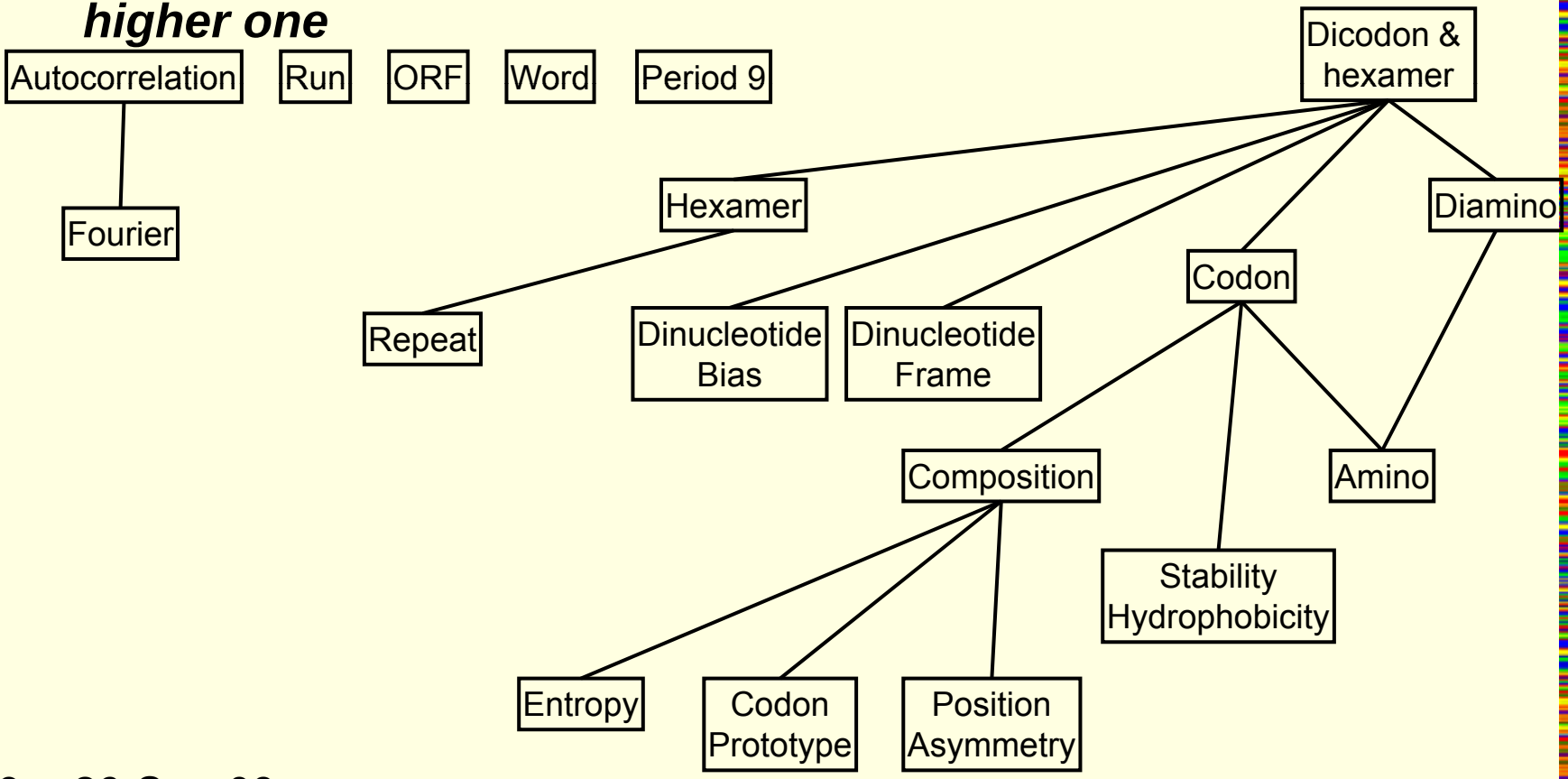
- ***Other methods (see Fickett and Tung, "Assessment of protein coding measures", Nucleic Acids Res. 20, 6441-6450, 1992)***
  - Runs (R&Y stronger in noncoding, W&S in coding)
  - N-word counts, most commonly hexamer
  - Stability (tendency to mutate to same amino acid residue)
  - Other base asymmetry measures
  - Periodicity, such as period 9
  - Global patterns (GC content, CpG islands)
  - Open reading frame (genes have longer ORF)
  - Exon length
  - Intron length



Genomics - Gene Modeling

Search by Content

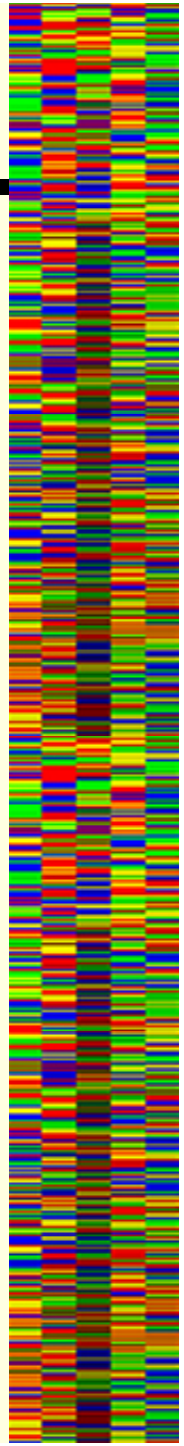
- ***Relationship between coding measures (from Fickett and Tung), lower function connected by lines is a function of the higher one***



Genomics - Gene Modeling

Extrinsic methods (search by signal)

- ***Try to identify sequence signals relevant to the presence, absence, frame, and content of genes***
- ***Signals***
  - promoters
  - terminators
  - polyA sites
  - Cap signals
  - splice junctions
- ***Sequence matches***
  - expressed genes (ESTs)
  - protein databases
  - closely related genomes (translated DNA vs translated DNA)



Genomics - Gene Modeling

Search by site

- *Eukaryotic transcription initiation site*

	-3	-2	-1	0	1	2	3	4	5	6	7	8	9	10	11
A	16	4	90	1	91	69	92	57	40	14	21	21	21	17	20
C	37	12	0	2	0	0	1	1	11	35	38	33	30	28	26
G	39	5	1	1	1	0	5	11	40	39	33	33	33	36	36
T	8	79	9	96	8	31	2	31	9	12	8	13	16	19	18

	G	T	A	T	A	A	A	A	G	G	C	G	G	G	G
	S	T	A	T	A	W	A	W	R	S	S	N	N	S	S

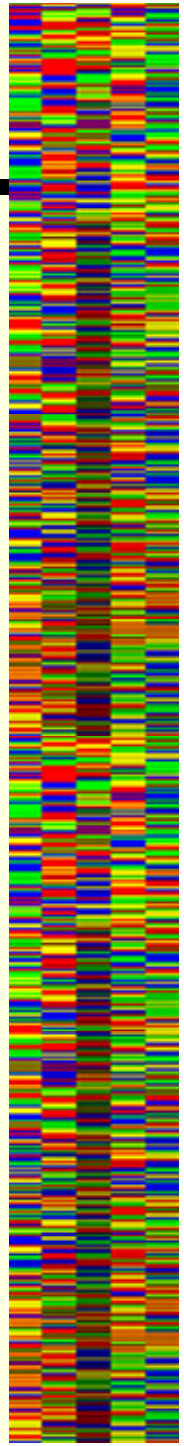
%frequency per position

Y = pyrimidine = C or T
R = purine = A or G
S = strong = G or C
W = weak = A or T

Genomics - Gene Modeling

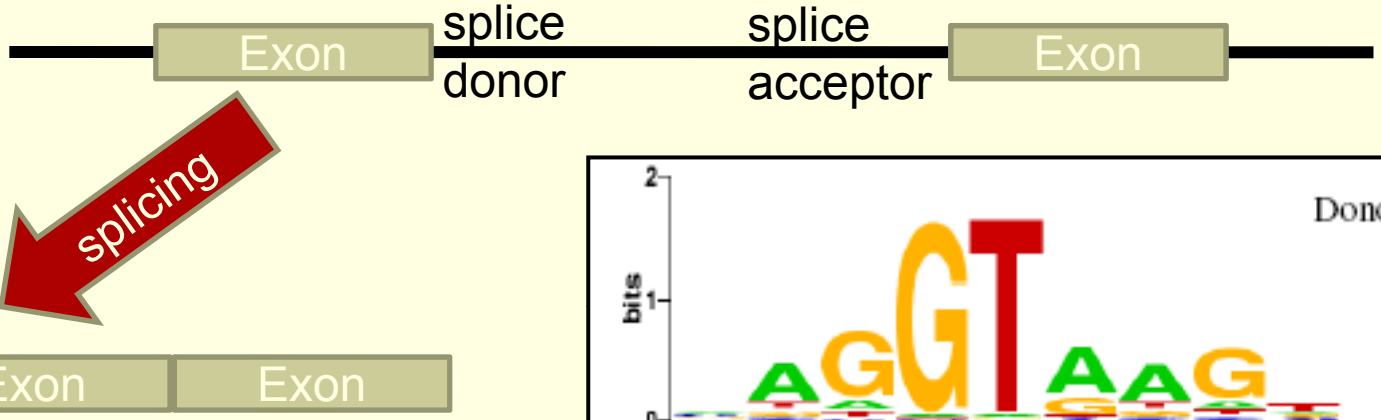
Search by Site - Splice sites

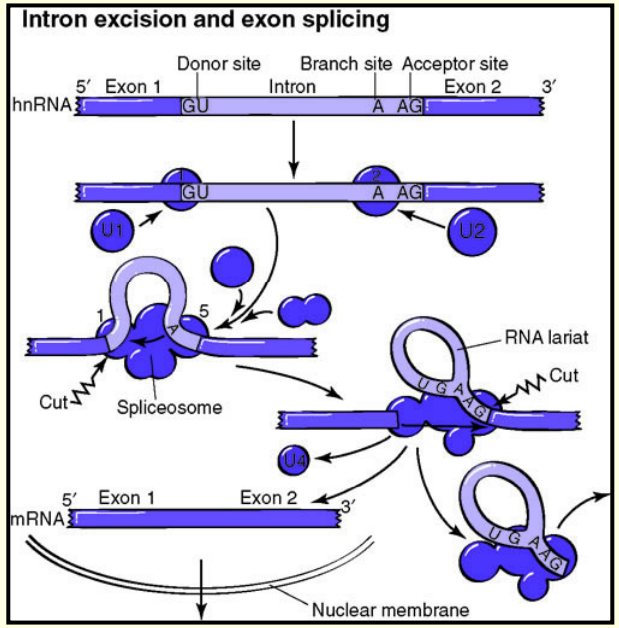
- ***The splicing of introns is a multi step process of RNA maturation which takes place in the nucleus***
  - generate mature mRNA molecules for transport to the cytoplasm.
  - Involves a complex of several factors such as snRNP (small nuclear ribonucleoprotein particles) and hnRNPs (heterogeneous nuclear ribonucleoprotein particles). This complex assembly is called the spliceosome.
- ***Introns usually begin with GU (donor splice site) and end with AG dinucleotides (acceptor splice site).***
- ***The branch point signal typically is located 10-50 bases upstream from the acceptor splice site (the lariat region).***

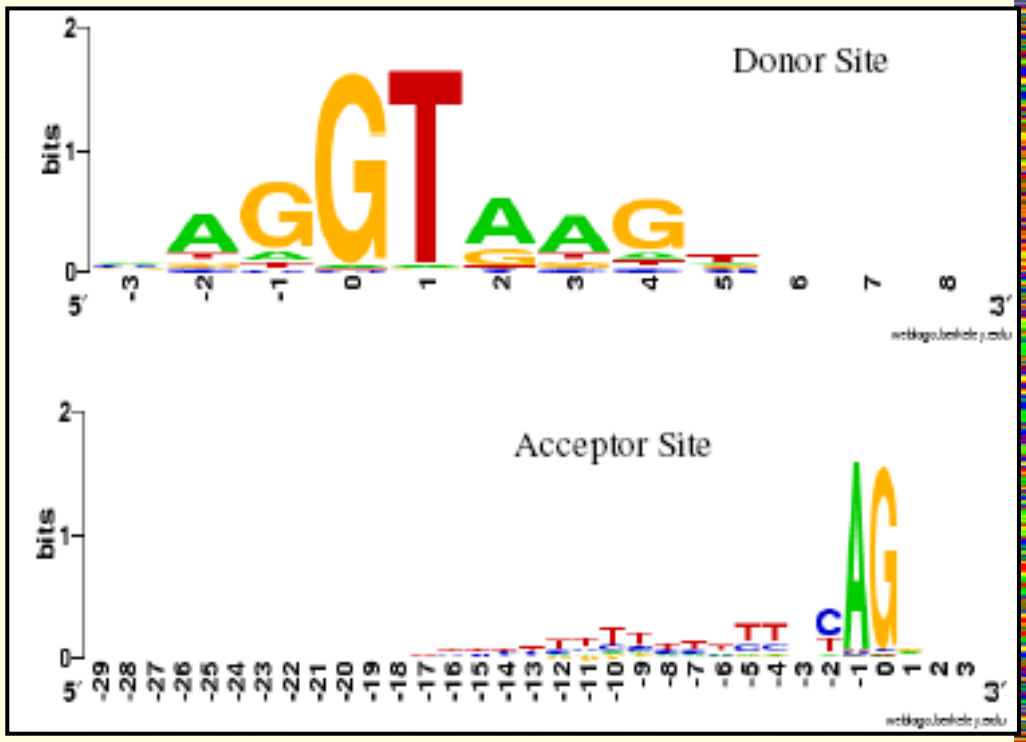


Genomics - Gene Modeling

Splice signals







mouse splice junction

Genomics - Gene Modeling

Search by Site – Splice junction

- *Donor site*

A	28	59	8	0	0	54	74	5	16
C	40	14	5	0	0	2	8	6	18
G	17	13	81	100	0	42	11	85	21
T	14	14	6	0	100	2	8	4	45
	C	A	G	G	T	A	A	G	T

or Weight Matrix

- *Acceptor site*

A	10	8	6	6	9	9	8	9	6	6	23	2	100	0	28
C	31	36	34	34	37	38	44	41	44	40	28	79	0	0	14
G	14	14	12	8	9	10	9	8	6	6	26	1	0	100	47
T	44	43	48	52	45	44	40	41	45	48	23	18	0	0	11
	T	T	T	T	T	T	T	T	T	T	N	C	A	G	G

Genomics - Gene Modeling

Search by Site – splice signals

- ***Branch point signal***

```
A  1  0 39 99 11
C 76  8 15  1 45
G  2  0 42  0  6
T 21 91  4  0 38
  C  T  G  A  C
```

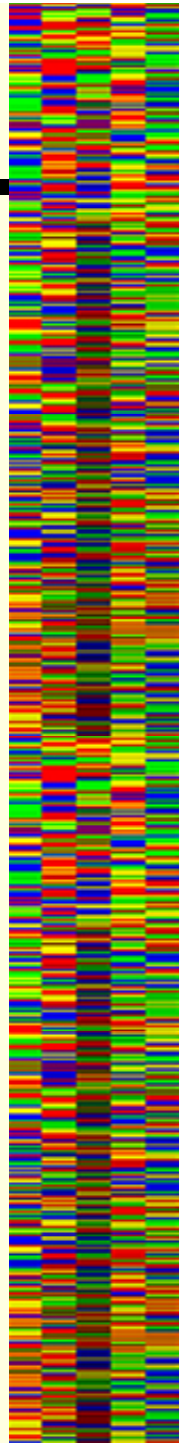
```
A -5.8 -6.8 -0.5  0.8 -2.3
C  0.8 -2.5 -1.3 -5.5    0
G -4.5 -6.5 -0.1 -6.5 -2.9
T -1.4  0.7 -3.8 -6.8 -0.5
    C    T    G    A    C
```

Consensus: CTGAC

Regular Expression: [CT]T[AG]A[CT]
YTRAY

Y = pyrimidine = C or T
R = purine = A or G
S = strong = G or C
W = weak = A or T

Log-odds assuming 45% AT, 55% GC

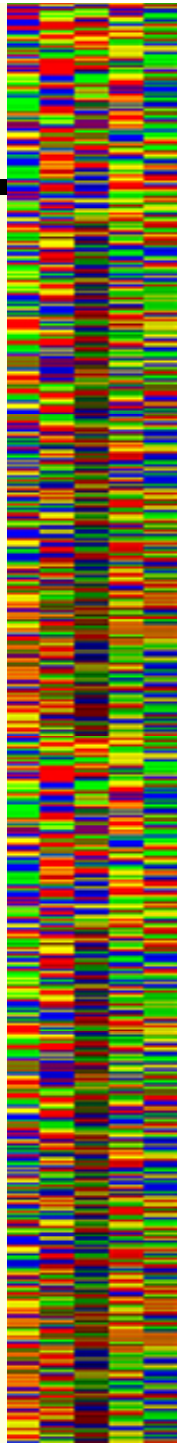


Genomics - Gene Modeling

Search by Site

- *Eukaryotic translation initiation site*

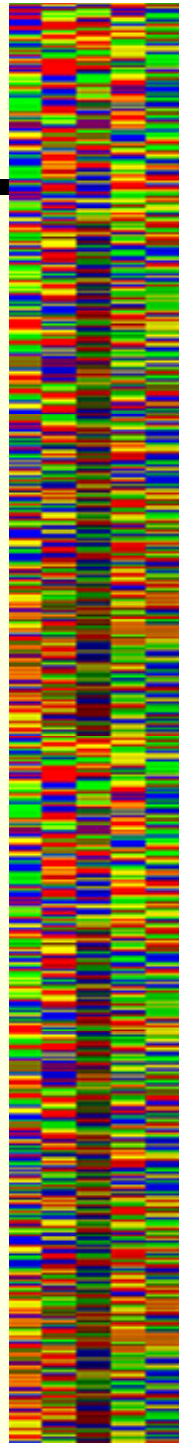
	-6	-5	-4	-3	-2	-1	+1	+2	+3
A	18	19	24	68	23	15	100	0	0
C	21	40	58	2	55	53	0	0	0
G	47	23	12	30	16	23	0	0	100
T	13	18	6	0	7	9	0	100	0
	G	C	C	A	C	C	A	T	G



Genomics - Gene Modeling

Search by Site

- ***Consensus sequences***
  - CCAAT-box
    - Y Y Y R R C C A W W S R -212 .. -57
  - GC-box
    - W R K R G G Y R K R K Y Y K -164 .. +1
  - cap-site
    - K C W K Y Y Y Y +1 .. +5
  - Information about composite regulatory elements, transcription factors and eukaryotic promoters are collected in the following databases:
    - TRANSFAC, http://www.gene-regulation.com/pub/databases.html (Wingender et al., 1996).
    - TFD, http://www.ifti.org/ootfd/ (Ghosh, 1993)
    - EPD, epd promoter, (Bucher, 1988)

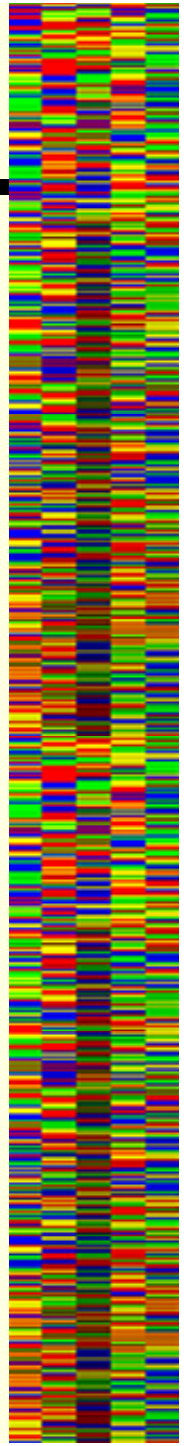


Genomics - Gene Modeling

Search by Site

- ***Polyadenylation site***

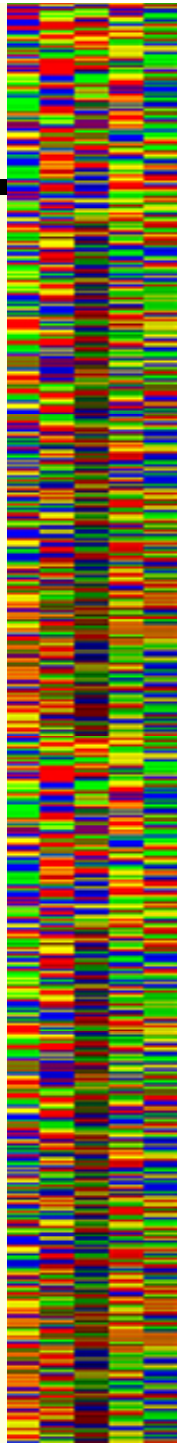
- ***Polyadenylation (cleavage of pre-mRNA 3' end and synthesis of poly-(A) tract) is a very important early step of pre-mRNA processing.***
- ***Sites***
  - AATAAA, located 15-20 nucleotides upstream from the poly-(A)
  - ATTAAA, is nearly as active as the canonical sequence.
  - An additional signal with consensus YGTGTTYY (diffusive GT-rich sequence) was revealed in region from 20 to 30 nucleotides downstream of poly-(A) site (site of cleavage) (McLauchlan et al., 1985).



Genomics - Gene Modeling

Search by Sites

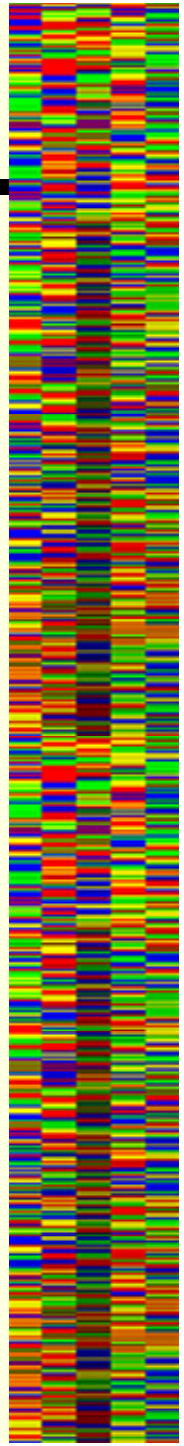
- ***Methods for identifying sites (weakest to strongest)***
  - Consensus sequence
  - Regular expression
  - Log-odds matrix / window analysis (PSSM)
  - Neural network or Hidden Markov model



Genomics - Gene Modeling

Known genes and proteins

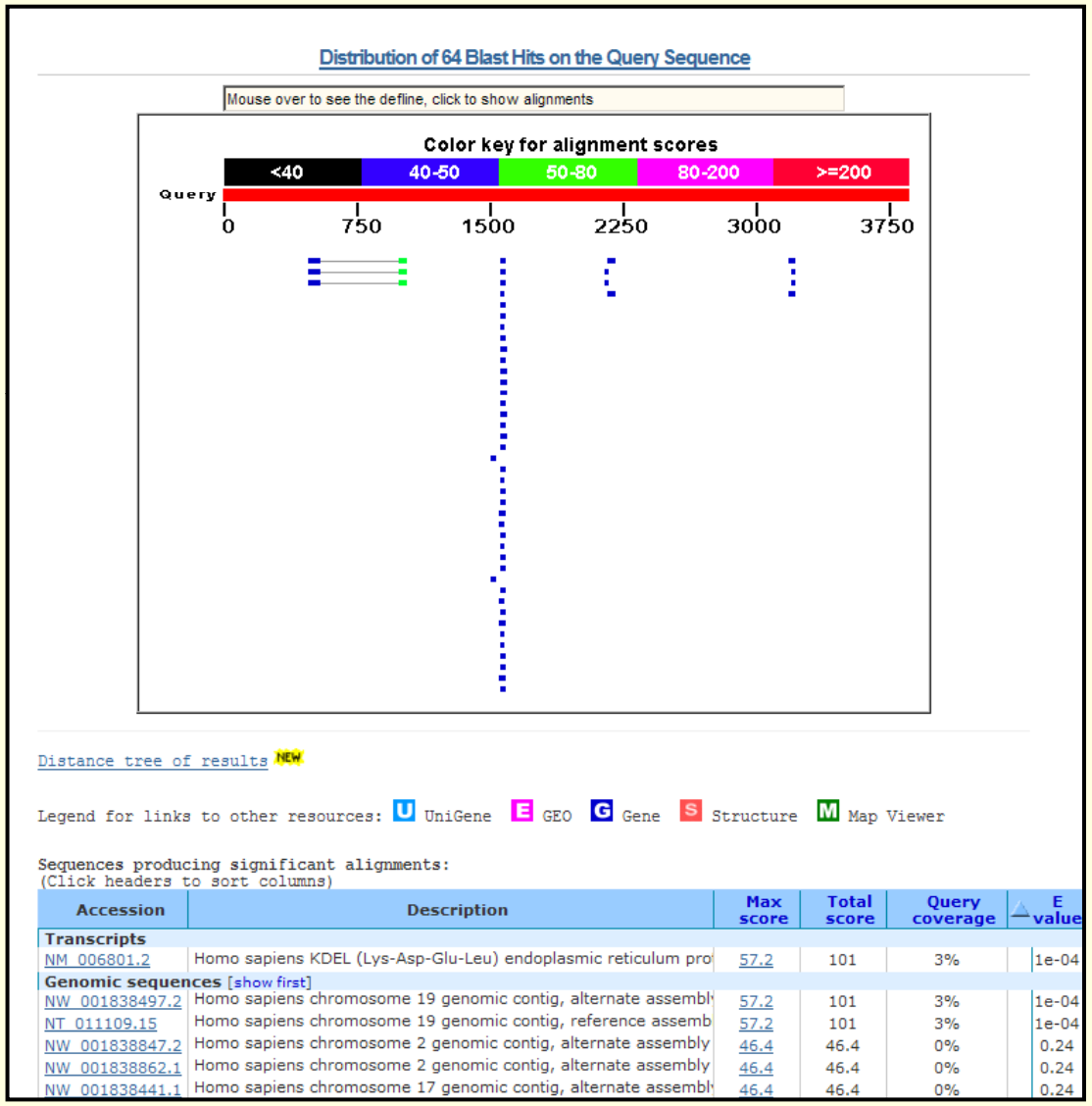
- ***Because we now know many genomes and many proteins, use these as queries to identify genes in new sequences***
- ***cDNA (EST) sequences can be sued in the same way***
- ***Problems***
  - Organism may have novel genes
  - Evolutionary divergence may make it difficult to find homologs
  - Gene databases may contain bad gene models or other "wrong" sequences
  - Database annotation may be incorrect (database pollution), or disagree between hits
  - EST collection may be incomplete
  - Sequencing and assembly errors make matches hard to detect
  - Genome may contain pseudogenes



Genomics - Gene Modeling

Known genes and proteins

- *BLAST (nucleotide) may not show much*

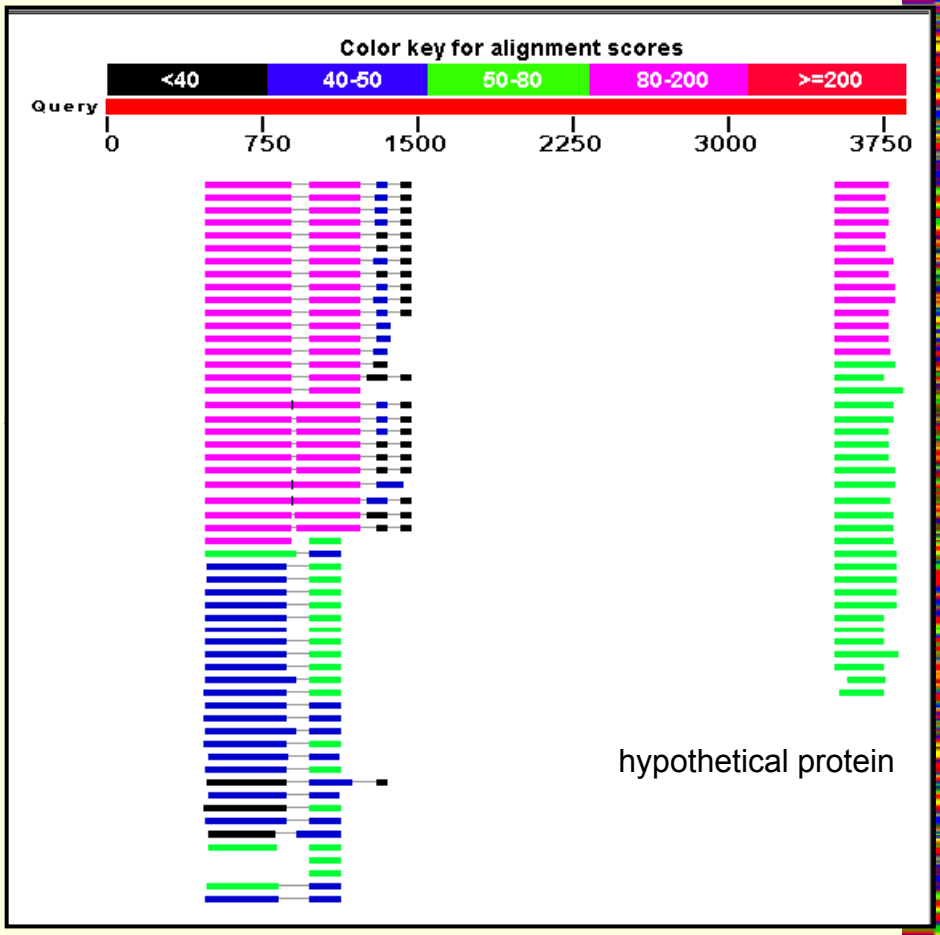


Accession	Description	Max score	Total score	Query coverage	E value
Transcripts					
NM_006801.2	Homo sapiens KDEL (Lys-Asp-Glu-Leu) endoplasmic reticulum pro	57.2	101	3%	1e-04
Genomic sequences [show first]					
NW_001838497.2	Homo sapiens chromosome 19 genomic contig, alternate assembl	57.2	101	3%	1e-04
NT_011109.15	Homo sapiens chromosome 19 genomic contig, reference assemb	57.2	101	3%	1e-04
NW_001838847.2	Homo sapiens chromosome 2 genomic contig, alternate assembly	46.4	46.4	0%	0.24
NW_001838862.1	Homo sapiens chromosome 2 genomic contig, alternate assembly	46.4	46.4	0%	0.24
NW_001838441.1	Homo sapiens chromosome 17 genomic contig, alternate assembl	46.4	46.4	0%	0.24

Genomics - Gene Modeling

Known Genes/Proteins

- ***BLASTX***
  - DNA query translated
  - protein database
- ***Finds matches to known proteins and gene models***
- ***May miss alternative exons***

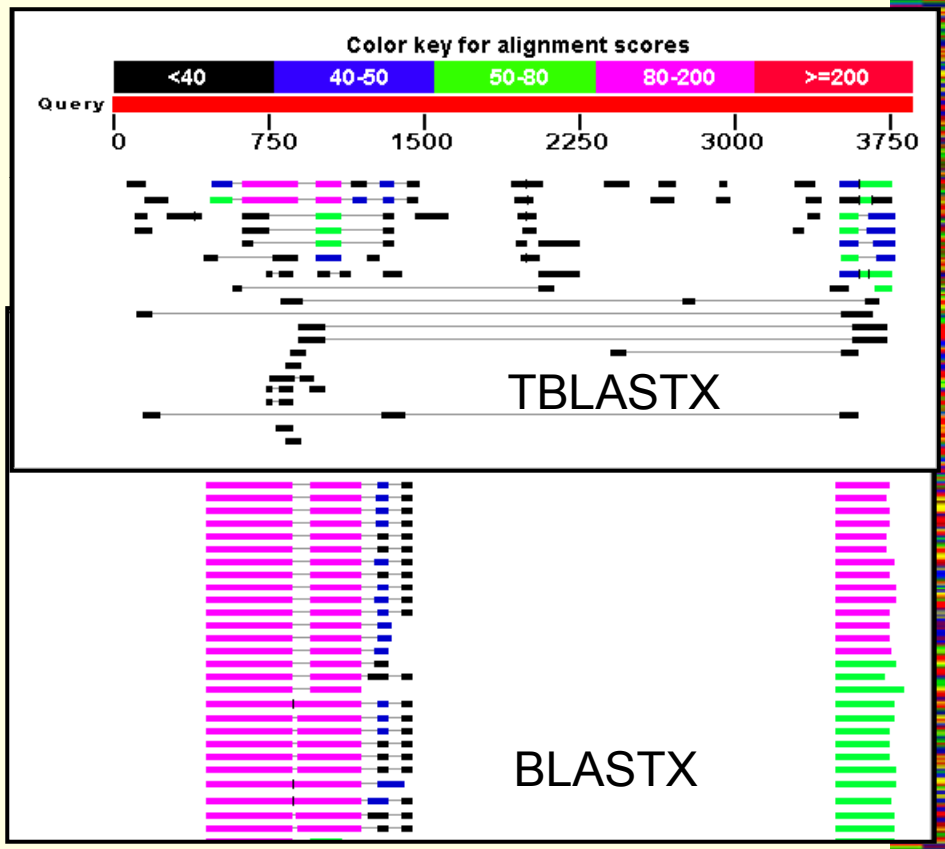


KDEL receptor A
ERD2 (ER lumen protein retaining receptor 2)

Genomics - Gene Modeling

Known Genes/Proteins

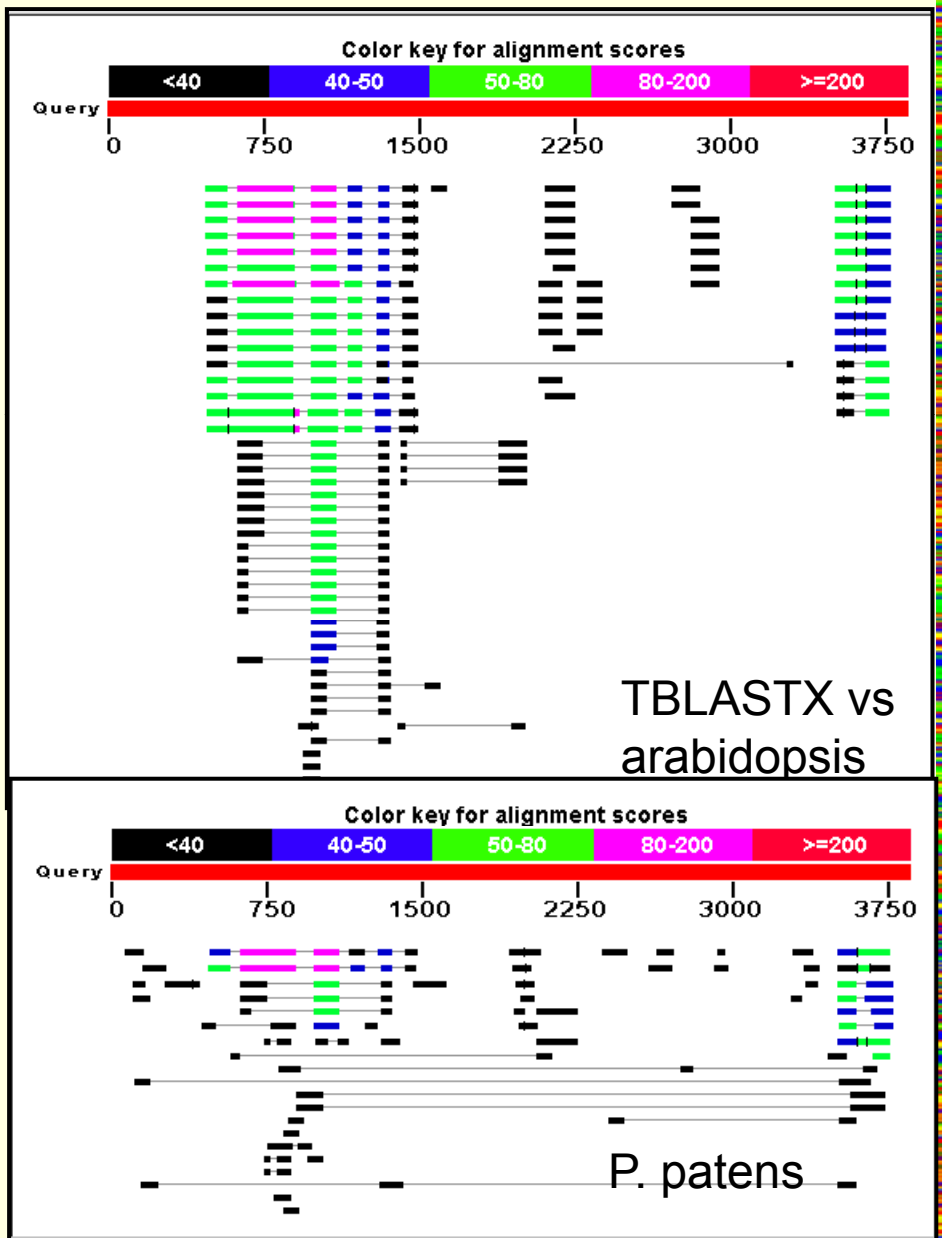
- ***TBLASTX***
  - DNA query translated
  - DNA database translated
- ***Finds matches to genes that may have been missed in annotation of database genomes***
- ***May find alternative exons***



Genomics - Gene Modeling

Known Genes/Proteins

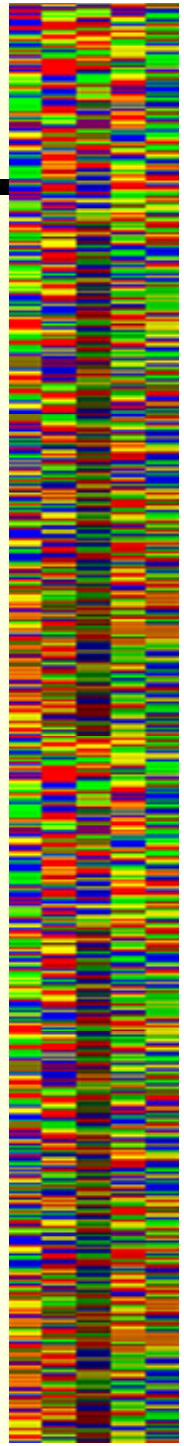
- *TBLASTX generally works best at an intermediate distance*
- *Too close – entire sequence matches*
- *Too far – nothing matches*
- *Just right – exons only match*



Genomics - Gene Modeling

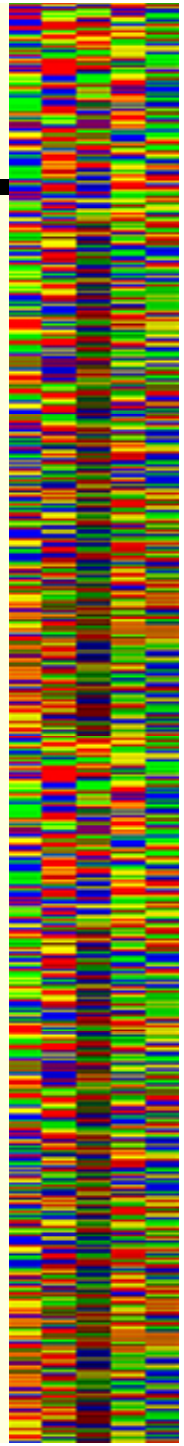
Motif libraries

- *Libraries of known protein motifs are available. Translated sequences that match to these libraries imply the presence of a gene (or pseudogene)*
- *Interpro*
- *Pfam*
- *PROSITE*
- *prints, blocks many others*



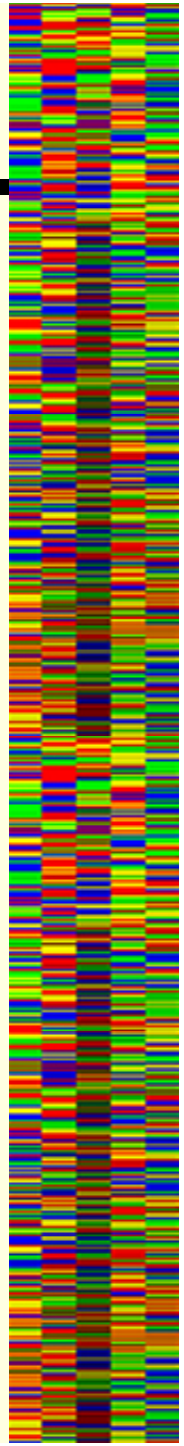
Genomics - Gene Modeling

- ***GRAIL uses a combination of search by content and search by signal approaches to produce a complete gene model based on genomic DNA sequence***
- ***GRAIL uses a neural net approach to combine information from a variety of "sensors" or indicators. Because of the neural net training, it doesn't matter two much if the sensors are highly correlated, they just get lower weights in the final prediction.***
- ***System is trained on real genes from a specific organism***



Genomics - Gene Modeling

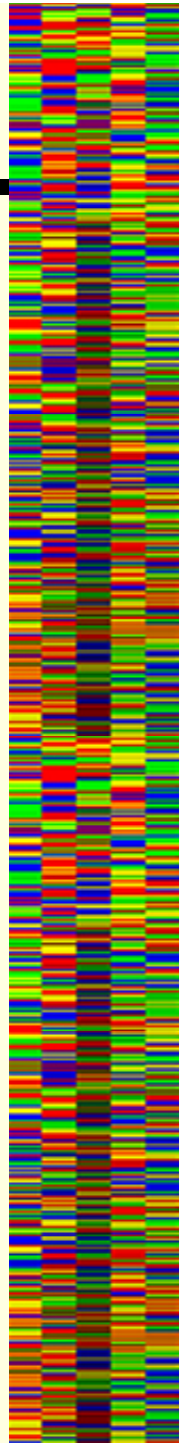
- ***Hexamer in-frame, candidate region and 60 bases left and right***
  - Isochore, candidate region
  - Left and right regions are presumably introns, if they have high coding scores, it may indicate that the candidate region is too small.
- ***Markov chain model (high AT and high GC models)***
- ***Isochore GC content***
- ***Exon GC content***
- ***Coding region length profile***
  - Candidates with lengths corresponding to common lengths get higher scores
- ***Candidate region length***



Genomics - Gene Modeling

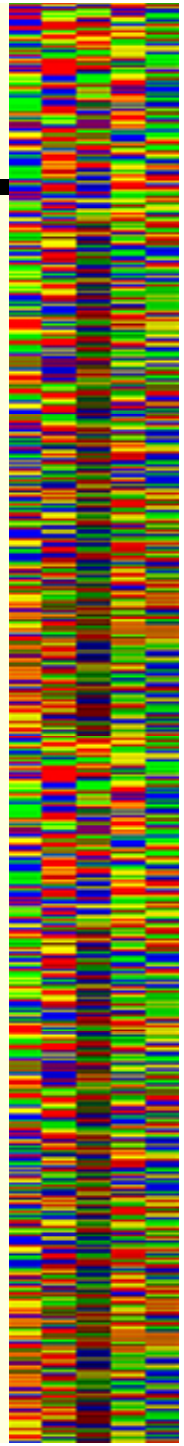
Splice donor site

- ***Splice acceptor site***
- ***Intron vocabulary (2 methods)***
  - Isochore, candidate region
  - Search for "words" that are common in introns but not exons



Genomics - Gene Modeling

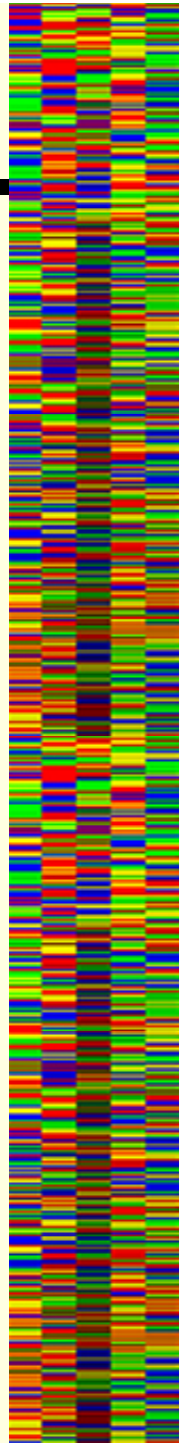
- ***Find candidate regions with specific edge signals, i.e. splice junctions***
- ***Evaluate coding potential for all sensors***
- ***Predict coding region using neural net***
- ***Assemble gene model***
  - 1st coding region starts with ATG
  - last coding region ends with inframe stop codon
  - adjacent coding regions maintain translation frame
  - distance must be at least minimum intron size
  - Uses dynamic programming to optimize combination of coding regions



Genomics - Gene Modeling

Neural Net

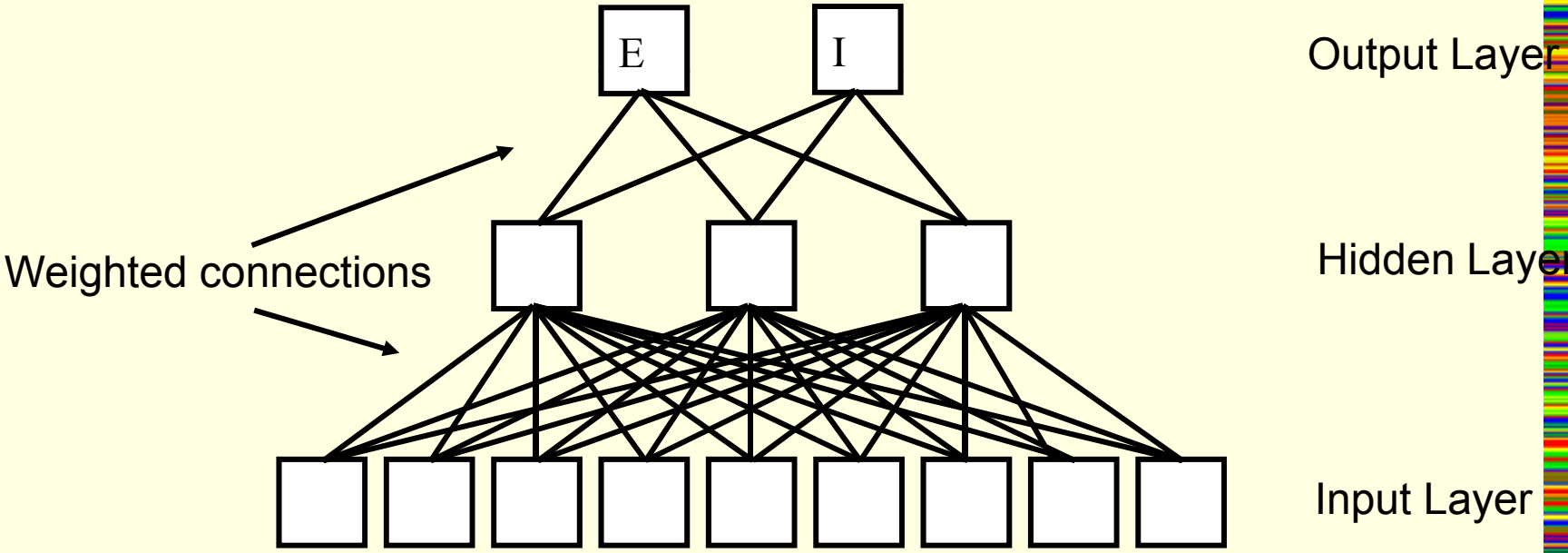
- *Neural net methods provide essentially a black box method for making predictions based on a set of training data. Neural nets explicitly consider interactions between the various inputs - interactions that may be very complex.*
- *Each connection in the net has a weight associated with it.*
- *During training, weights are iteratively adjusted so that the prediction agrees with the training data. This process is typically known as back propagation.*
- *Weight matrices can be looked at as a neural net with no hidden layers (also called a perceptron).*



Genomics - Gene Modeling

Search by Site

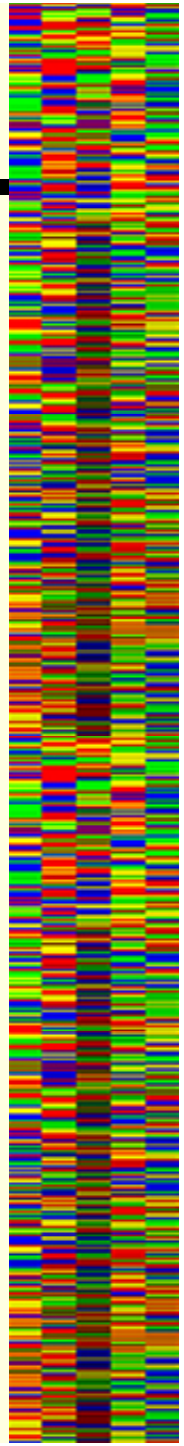
- ***Artificial Neural Network (ANN)***
- ***Must be trained with classified data***



Genomics - Gene Modeling

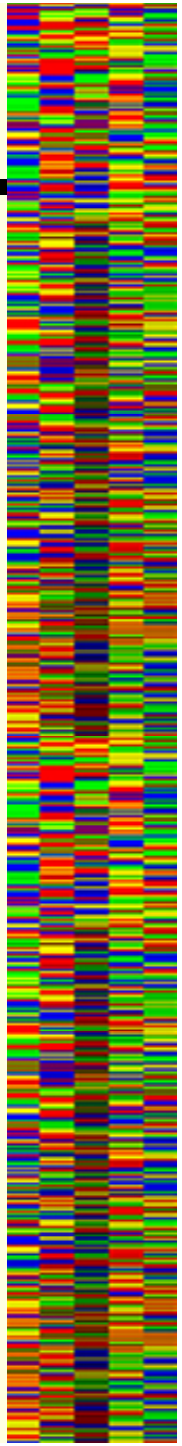
Other things GRAIL does

- *Attempts to correct indels*
- *Detects CpG islands (frequently found at 5' end of genes)*
- *Promoters*
- *Poly-A sites*
- *Repetitive DNA*
- *Protein sequence searching*



Genomics - Gene Modeling

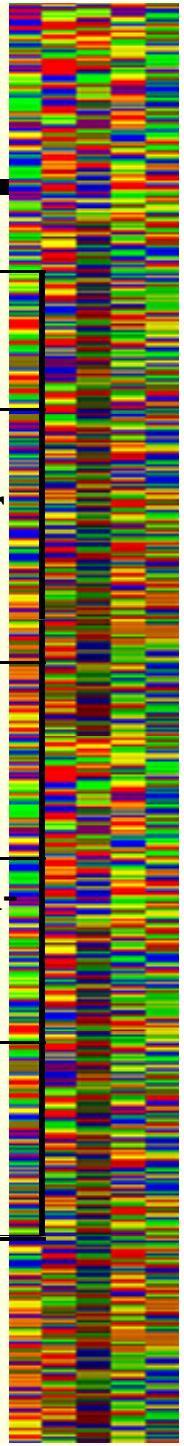
- ***GRAIL II uses candidate region approach described above***
- ***Many other systems***
  - Glimmer
  - Genemark (mostly for bacterial genomes)
  - GeneParser
  - GeneID
  - FGeneH
  - GenLang



Genomics - Gene Modeling

Limitation

Gene prediction method	*Limitation*
Ab initio (Hidden Markov Model (HMM)-based) methods	**Poor sensitivity and specificity, leading to whole genes or exons being missed or wrongly predicted**
Similarity to existing expression sequence tags (ESTs)	**Contaminating ESTs derived from unspliced mRNA, genomic DNA and nongenic transcription**
Similarity to existing gene/proteins	**Unable to distinguish pseudogenes (non-protein coding) and novel genes undetected**
Current approaches result in	**Partial genes, fragmented genes, gene fusions and spurious predictions**



Genomics - Gene Modeling

Things to Remember about gene modeling

- *It is, in general, organism-specific*
- *It works best on genes that are reasonably similar to something seen previously*
- *It finds protein coding regions far better than non-coding regions*
- *In the absence of external (direct) information, alternative forms will not be identified*
- *It is imperfect! (It's biology, after all…)*