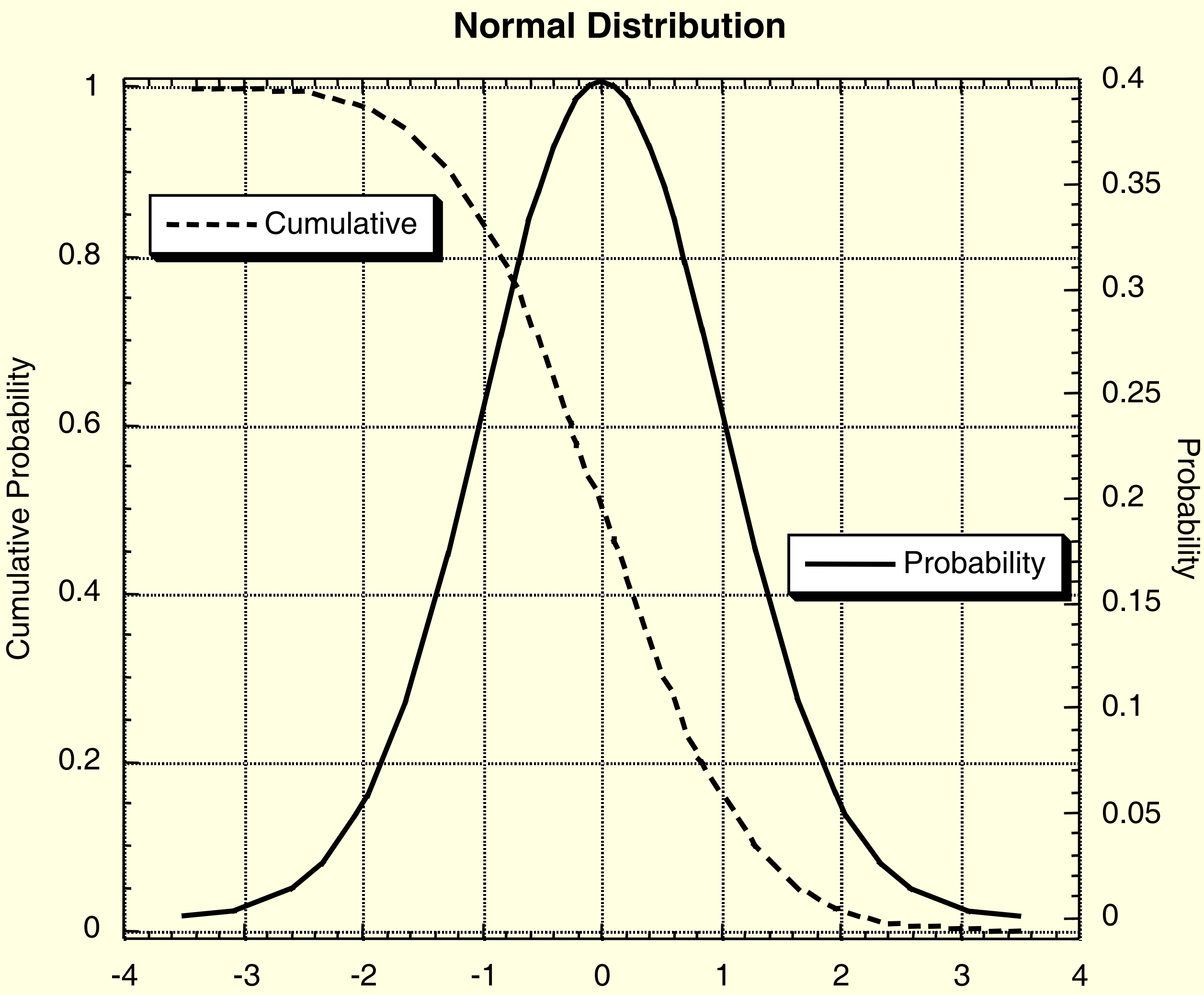
Preliminary Syllabus

Sep 30	Introduction & Genome Assembly
Oct 2	Sequence Comparison
Oct 7	Gene Modeling
Oct 9	Gene Function Identification – Read intro to HMM on blackboard
Oct 14	OCTOBER BREAK
Oct 16	Comparative Genomics
Oct 21	Protein-Protein Interactions
Oct 25	Pathway Resources and Analysis
Oct 28	Structural Genomics / Protein Structure Prediction
Nov 4	Protein Modeling
Nov 8	EXAM

- Gribskov@purdue.edu – Lilly G-233

Sequence Comparison

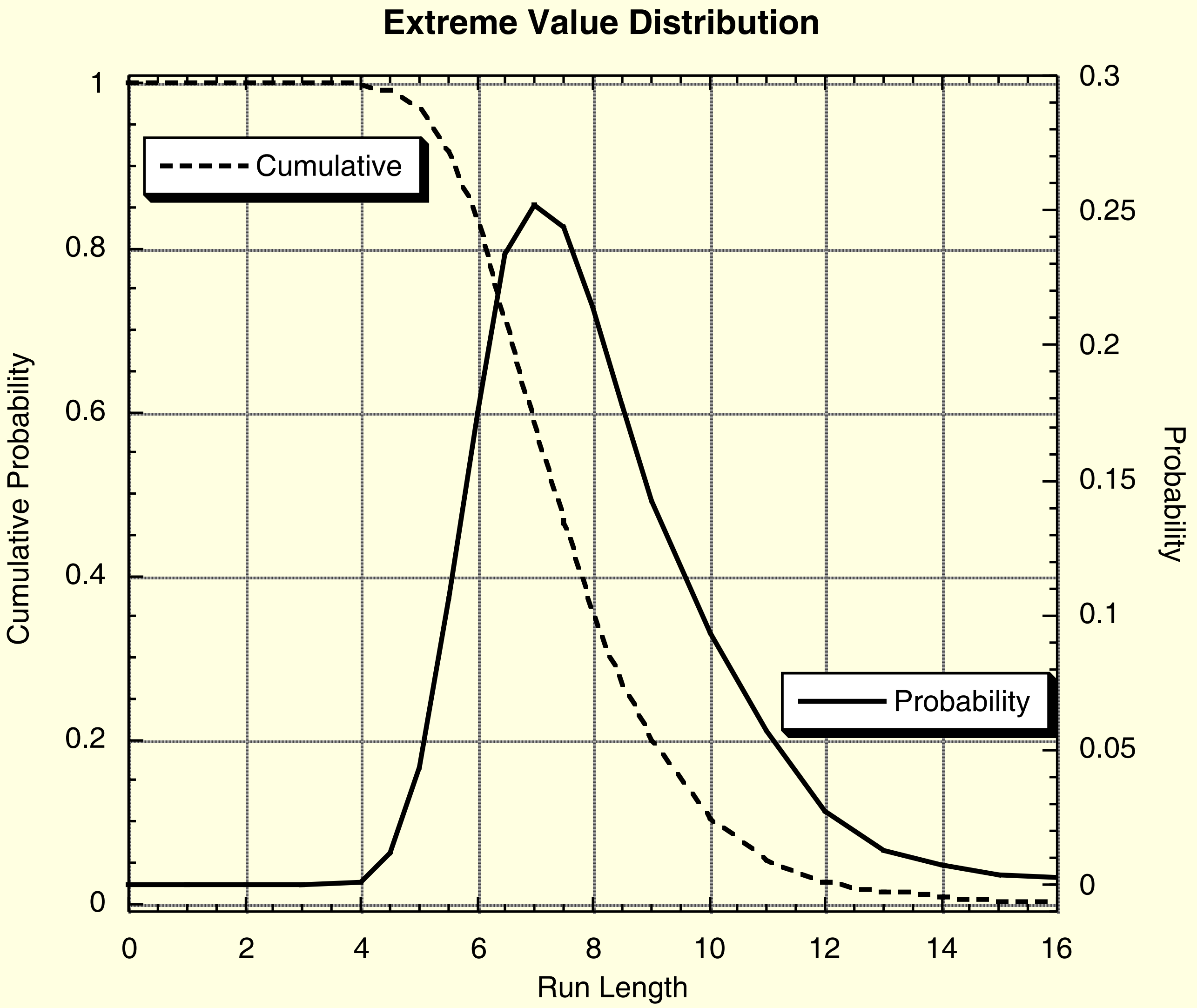
Genomics



Gribskov 3.2

Sequence Comparison

Genomics



Gribskov 3.3

Gene Modeling

Gene Modeling

Goals

- Gene modeling begins with an uncharacterized genomic sequence and predicts the transcriptional and translational products of each gene, including
  - **Gene location, direction, and/or frame**
  - **5’ and 3’ untranslated regions**
  - **Introns and exons**
  - **Possibly includes regulatory elements**

- Gene modeling is notoriously difficult, especially in eukaryotes, but it is widely felt that current methods produce largely correct models, i.e. have errors in only 30% or so of eukaryotic genes and 10% of prokaryotic genes.
  - **Most common errors are in 5’ end of gene and small exons**
  - **Difficult to distinguish errors from true genetic variation**
    - splice variants
    - pseudogenes

Gene Modeling

Basic Approaches

- Prokaryotic genes are obviously easier
  - **No introns**
  - **Simpler signals**
  - **Often better DNA sequence**
- Eukaryotic genes are very challenging
  - **Exons/introns may be very small (less than 10 bases)**
  - **Introns may be very large (greater than 1 Mbase)**
  - **Signals are poorly known and more complex**
  - **DNA sequence may be more poorly assembled**

Gene Modeling

Basic Approaches

- extrinsic – comparison to other known genes
  - **sequence comparisons to known proteins, cDNAs**
  - **genome comparison**
- intrinsic – properties of the sequence caused by the fact that it codes a protein
  - **ORF length**
  - **GC content**
  - **word frequencies**
- hybrid

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)**

Gene Modeling

Genomics

Sequence Motifs - Consensus Sequence

- Feature is represented as the majority or plurality character at each position

```
GCGGTGATAATGGTTGCATG
TTGGGTATATTTGACTATGG
ATGCATACACTATAGGTGTG
TGCAGTAAGATACAAATGGC
ATGGTTATAGTATGCCCATG
     TATAAT    GCGTG
```

Gene Modeling

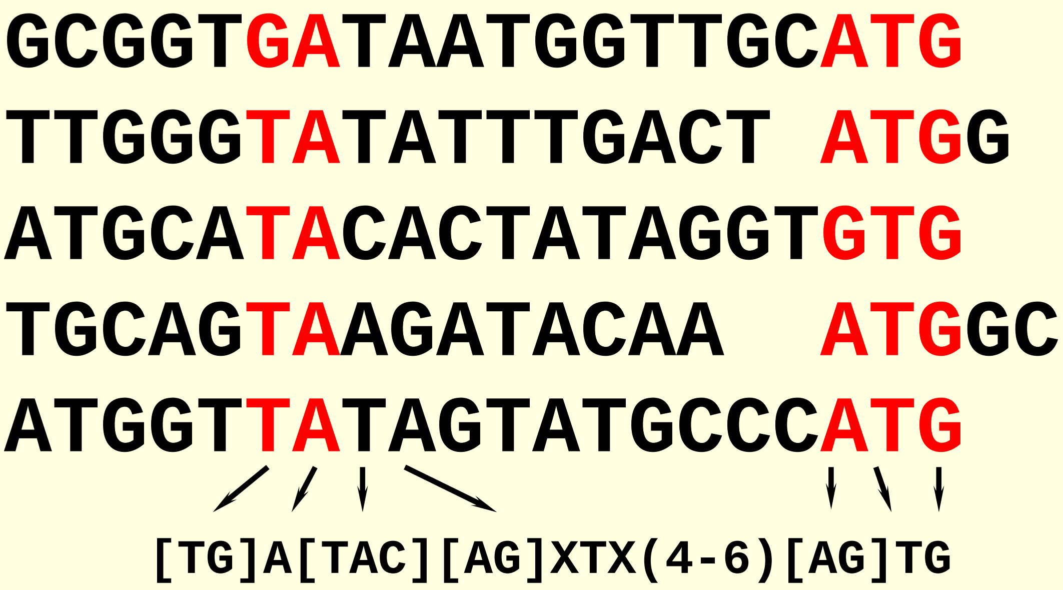
Sequence Motifs - Consensus Sequence

- Advantages
  - **Concise**
  - **Simple to detect**
  - **Easily remembered and displayed**
- Disadvantages
  - **Most information is lost – poor ability to find signals**
  - **Difficult to evaluate partial match**
  - **Very sensitive to alignment**

Gene Modeling

Sequence Motifs - Regular Expression

- Feature represented by logical combination of characters



```
GCGGTGATAATGGTTGCATG
TTGGGTATATTTGACT ATGG
ATGCATACACTATAGGTGTG
TGCAGTAAGATACAA  ATGGC
ATGGTTATAGTATGCCCATG

[TG]A[TAC][AG]XTX(4-6)[AG]TG
```

Gene Modeling

Sequence Motifs - Regular Expression

- Advantages
  - **Fairly concise and easy to understand**
  - **Well known algorithms for matching, *O(n log n)***
  - **Fairly easy to display**
  - **Can accept gaps**
- Disadvantages
  - **Still loses information, better than consensus**
  - **Rigid**
  - **Difficult to evaluate partial matches**

Gene Modeling

Sequence Motifs - Regular Expression Methods

- PROSITE Release 19.35, of 19-Sep-2006
  - **Constant updates**
  - **1331 different patterns, 4 rules and 650 profiles/matrices).**
  - **1446 documentation entries**
- Signatures derived by hand
- Relatively "fragile"
- Hulo N., Bairoch A., Bulliard V., Cerutti L., De Castro E.,Langendijk-Genevaux P.S., Pagni M., Sigrist C.J.A.
  The PROSITE database.
  Nucleic Acids Res. 34:D227-D230 (2006)

Gene Modeling

Sequence Motifs - Regular Expression Methods

- PROSITE “language”
- Each position is separated from the next by a hyphen “-”
- X means any residue
- [ ] surround ambiguities, e.g. [ALT] means ala, leu or thr
- { } surround forbidden residues, {AM} means neither ala nor met
- ( ) surround repeat counts
  - **(3) means exactly three repeats**
  - **(2-4) means 2 to 4 repeats**
- < and > indicate the beginning or end of the sequence, respectively
- . ends the pattern

Sequence Motifs - Regular Expression Methods

- PROSITE – tabulated results are useful for training new methods
  - **True positives (T) - Sequences that have the feature and match the signature**
  - **False positives (F) - Sequences that do not have the feature but match the signature**
  - **False Negatives (N) - Sequences that have the feature but do not match the signature**
  - **True negatives - Sequences that have the feature but do not match the signature**
  - **Potential (P) - likely to be a true positive**
  - **Maybe (?) - might have the feature, but unclear**

Sequence Motifs - Regular Expression Methods

- A PROSITE entry

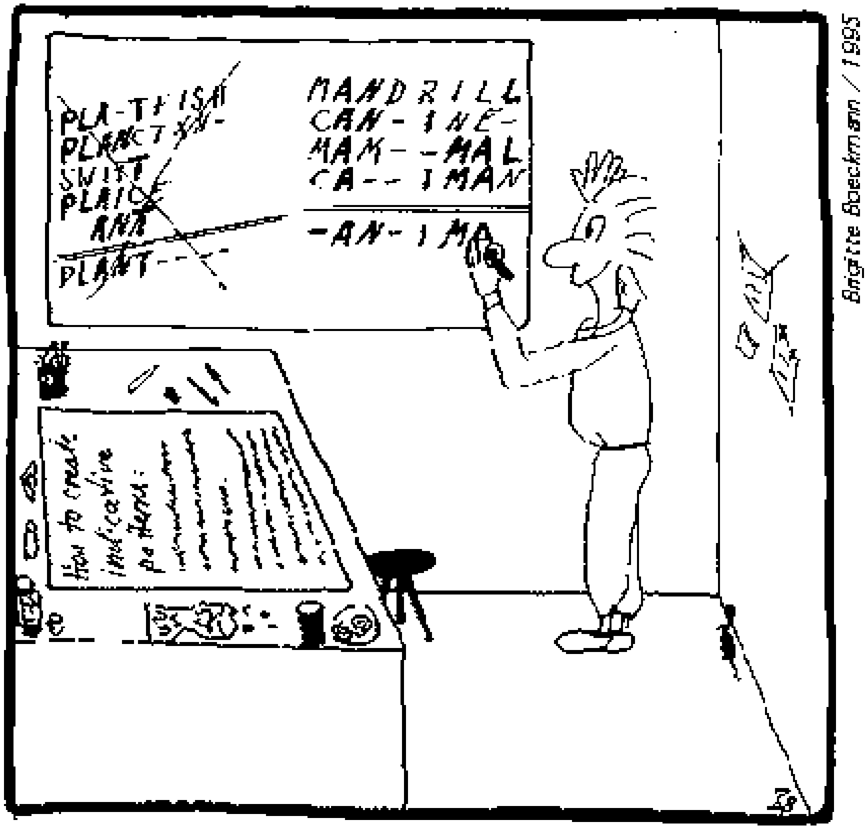
```
ID   CNMP_BINDING_2; PATTERN.
AC   PS00889;
DT   OCT-1993 (CREATED); OCT-1993 (DATA UPDATE); OCT-1993 (INFO UPDATE).
DE   Cyclic nucleotide-binding domain signature 2.
PA   [LIVMF]-G-E-x-[GAS]-[LIVM]-x(5,11)-R-[STAQ]-A-x-[LIVMA]-x-[STACV].
NR   /RELEASE=26,33329;
NR   /TOTAL=56(34); /POSITIVE=55(33); /UNKNOWN=0(0); /FALSE_POS=1(1);
NR   /FALSE_NEG=1(1);
CC   /TAXO-RANGE=??EP?; /MAX-REPEAT=2;
DR   P03020, CRP_ECOLI , T; P29281, CRP_HAEIN , T; P06170, CRP_SALTY , T;
DR   Q00194, CGCC_BOVIN, T; P29973, CGCC_HUMAN, T; P29974, CGCC_MOUSE, T;
DR   P05207, KAP2_PIG  , N;
DR   P31324, KAP3_MOUSE, P;
DR   P29956, XANB_XANCP, F;
3D   2GAP; 3GAP; 1CGP;
DO   PDOC00691;
```

Gene Modeling

Genomics

Sequence Motifs - Regular Expression Methods

How we develop Prosite patterns!



Gribskov 3.17

Gene Modeling

Sequence Motifs - Regular Expression Methods

- PROSITE

- Steps to defining a signature (manual)
  - **1. Align sequences**
  - **2. Find a four or five residue sequence that is part of a known important region (core pattern)**

    Active site, substrate binding, prosthetic group, etc.
  - **3. Scan SWISS-PROT and see what matches**
  - **4. If only true positives are found, stop. Otherwise, add to the signature and return to step 3.**

Gene Modeling

Sequence Motifs - Regular Expression Methods

- PROSITE

- Generation of signature - "Walker type" ATP binding sites

```
malk        SGCGKS.TLL
hisp        SGSGKS.TFL
oppd        SGSGKSQSRL
ecatpa      AGVGKT.VNM
bovatpb     AGVGKT.VFI
```

[SA]-G-[CSV]-G-K-[ST]-X(0,1)-[TSV]-[LMI]

- Simplest method - combine observed residues at each position

Gene Modeling

Genomics

Sequence Motifs - PSSM

- Position Specific Scoring Matrix, or weight matrix, is calculated based on observed frequencies in a column

```
GCGGTGATAATGGTTGCATG
TTGGGTATATTTGACTATGG
ATGCATACACTATAGGTGTG
TGCAGTAAGATACAAATGGC
ATGGTTATAGTATGCCCATG
```

Gene Modeling

Sequence Motifs - PSSM

- Position specific scoring matrix (PSSM)
- Feature is represented as a matrix with a score for every possible character
- A simple weight matrix for the bacterial promoter -10 region, values here are simply % frequencies

A	2	95	26	59	51	1
C	9	2	14	13	20	3
G	10	1	16	15	13	0
T	79	3	44	13	17	96
	T	A	T	A	A	T

Gene Modeling

Sequence Motifs - PSSM

- Advantages
  - **Preserves first order information, i.e. assumes that positions are independent**
  - **Flexible, can model all regular expression type signatures**
  - **Accommodates partial matches, with known method for evaluating significance of matches**
- Disadvantages
  - **Difficult to display, impossible to remember**

Gene Modeling

Sequence Motifs - PSSM

- Log-odds matrix - as we have already learned, a log-odds statistic is one of the most powerful discriminators. Weight matrices are often in log-odds form.
  $w_{ij} = ln \, (f_{obs} / f_{exp})$ $score = \Sigma w$ over width of pattern

- What should one use for the background model, f_{exp} ?
  - **Database composition**
  - **Global composition of query sequence**
  - **Local composition of query sequence**
  - **Combination of query and database sequences**

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

Gribskov 3.24

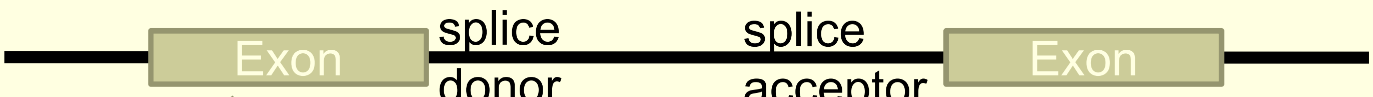
Genomics - Gene Modeling

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

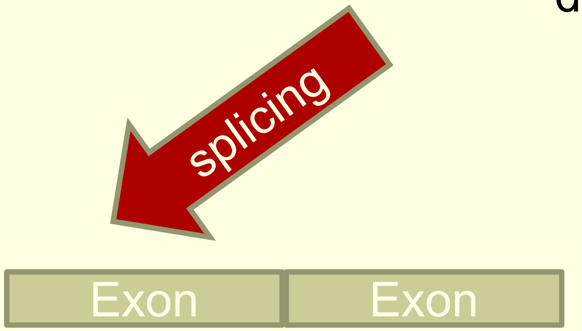


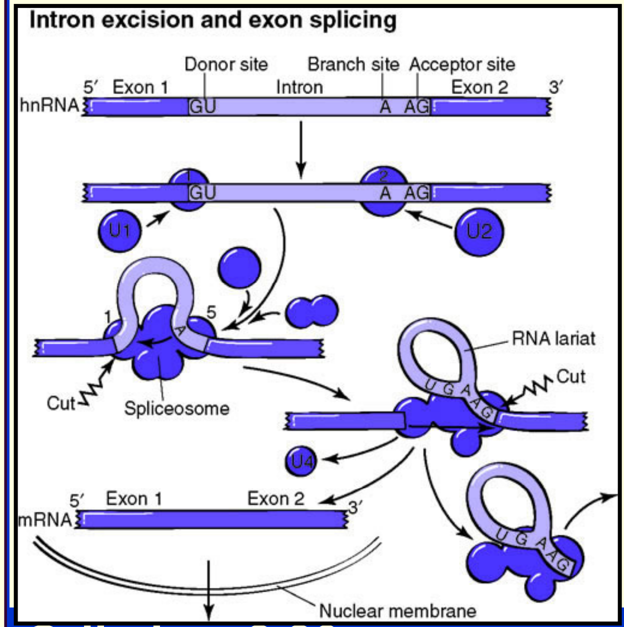
Exon

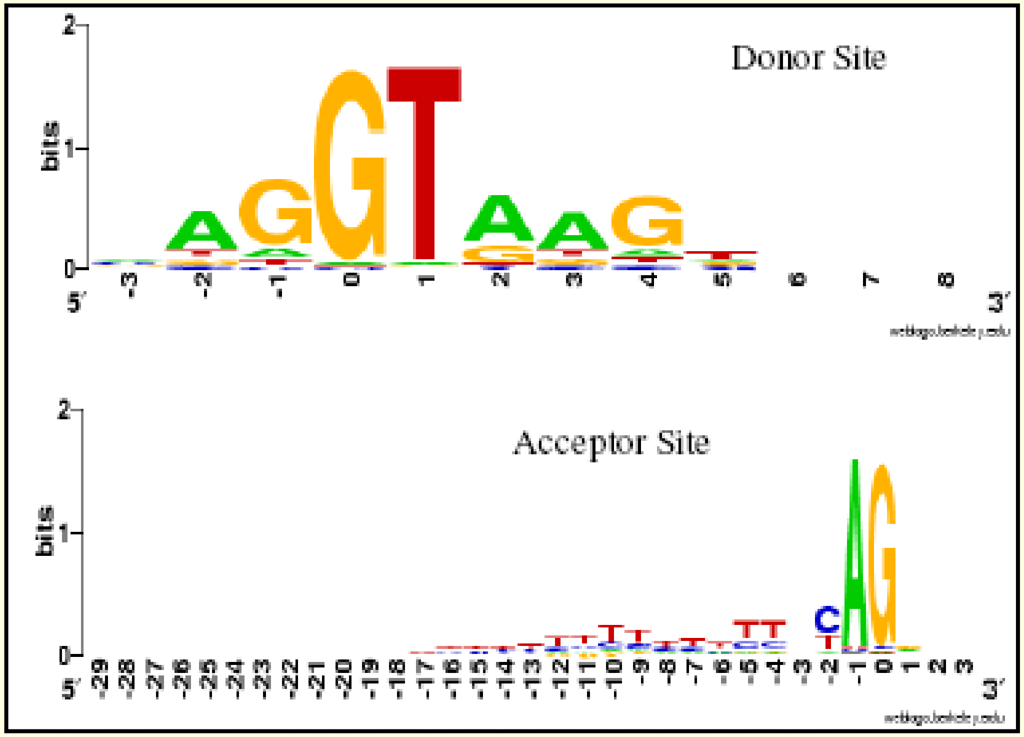
splice donor

splice acceptor

Exon







mouse splice junction

Gribskov 3.26

Genomics - Gene Modeling

Genomics

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
```

- 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
```

Position Specific Scoring Matrix (PSSM)

or Weight Matrix

Gribskov 3.27

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

Log-odds assuming 45% AT, 55% GC

Gribskov 3.28

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

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**

Methodology

What is Homology?

- Nothing in biology makes sense except in the light of evolution.
  - **Theodosius Dobzhansky (1900-1975)**
  - **…without that light it becomes a pile of sundry facts some of them interesting or curious but making no meaningful picture as a whole.**

- homology - the presence of a similar feature because of descent from a common ancestor

- homoplasy - the presence of a similar feature because of convergence

  - **Homology cannot be observed. We can't actually see the ancestral organisms/molecules and trace descent.**
  - **Homology is an inference, a conclusion we draw based on observed similarity.**
  - **Homology is an all-or-none relationship**

Methodology

Why is homology Important?

- Homology strongly suggests that the molecules have similar structure and function
- There are (very) many ways to fold a polypeptide to place specific chemical groups at specific locations. There is no reason, *a priori*, why proteins with a specific function should have similar 3-D structures.
- Therefore, there is no reason, *a priori*, why unrelated sequences should have any detectable similarity in sequence. Significantly similar molecular sequences are very unlikely to arise by chance - i.e. homoplasy on the molecular level is very unlikely.
- When we see significant similarity, we infer that the sequences/structures are homologous, i.e. at some point in the past they share an identical structure.
- The only thing that keeps the sequences tied to each other is the commonality of structure and function arising from homology.

Methodology

Homology

- Sequences alignments and database searches let us
  - **Find homologous sequences (genes/proteins)**
  - **Map information from known systems to new ones**
    - Gene identification
    - Gene function
    - Metabolic and regulatory systems
- Two common classes of homologs
  - **Orthologs – genes separated by a speciation event, i.e. the same gene in two species**
  - **Paralogs – genes separated by a duplication events, originally the same but now diverged with possibly different functions**

Sequence Comparison

BLAST Basic Idea

- Determine in advance the MSP score you need to be significant, *S*
  - **for example, choose *S* so that you will see fewer than 10 unrelated sequences in the database that score as high**
- Look for matching words of length w that score above a threshold, *T*, such that MSPs of score *S* are unlikely to be missed. These are High-scoring Segment Pairs (HSPs)

Sequence Comparison

Genomics

BLAST procedure

- Step 1: Compile list of high scoring words from query sequence
- Step 2: Scan database for "hits"
- Step 3: Extend regions with 2 hits into MSPs
- Step 4: Dynamic programming alignment around MSPs

Sequence Comparison

BLAST Step 1 - List of High Scoring Words

- Choose a significance level *S*
- Choose a word size, *w*, and cutoff , *T*, so that you are unlikely to miss MSPs with score *S*
- Make a table of all words in the "neighborhood" of the query (DNA sequences use all words)
- Typically 50 words for each residue

Sequence Comparison

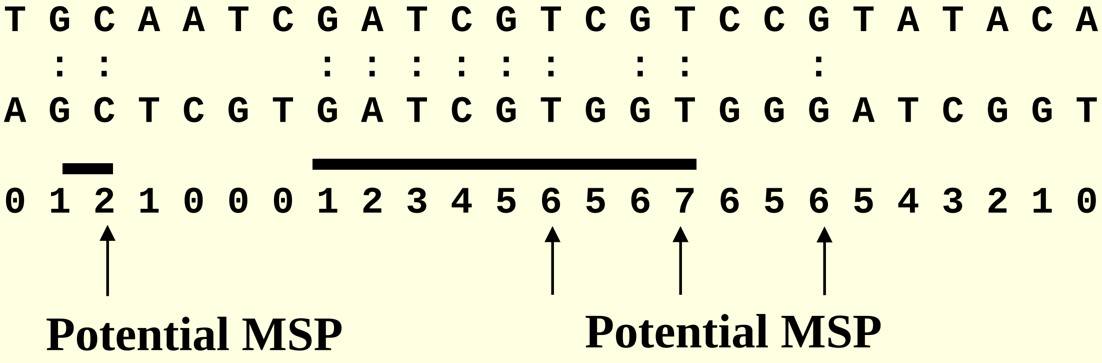
BLAST Step 2 - Scan Database

- Scan only for words in neighborhood
- Use lookup tables (like FASTA) or finite automaton
- Keep data in memory to make it faster

Sequence Comparison

BLAST Step 3 - Extend Words to MSPs

- In BLAST2, a “diagonal” must have two word hits before extension to MSP is attempted.
- In principal, must examine diagonal until score drops to zero
- Shortcut, only check until score drops by X



```
T G C A A T C G A T C G T C G T C C G T A T A C A
  : :         : : : : : :   : :     :
A G C T C G T G A T C G T G G T G G G A T C G G T

0 1 2 1 0 0 0 1 2 3 4 5 6 5 6 7 6 5 6 5 4 3 2 1 0
```

Potential MSP

Potential MSP

Sequence Comparison

Filtering

- Some sequences give spurious matches because of their unusual properties. Such sequences are automatically filtered by BLAST
- Filters remove “low entropy” sequences. These are repetitive sequences that often give anomalous matches in a database search.
  - **Degenerate sequences – e.g., poly A runs**
  - **Dinucleotide, trinucleotide (or longer) repeats**
  - **Transmembrane regions and signal peptides in proteins**