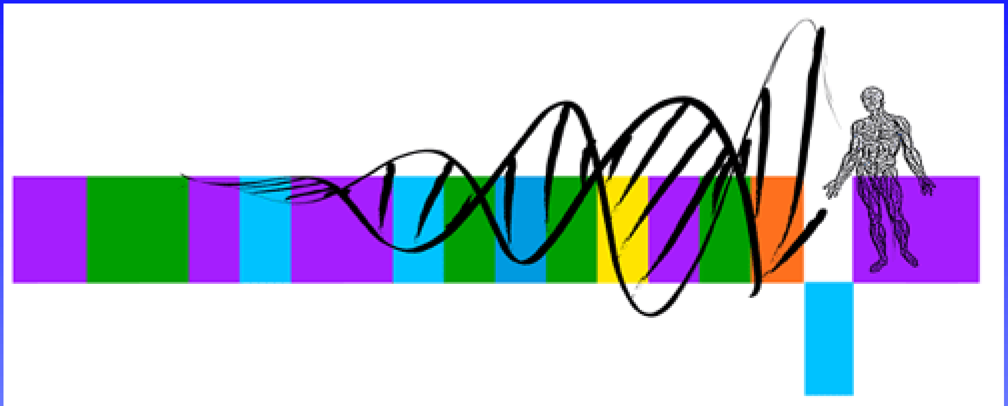


Sequence Analysis

II:

Sequence Patterns and Matrices

George Bell, Ph.D.
WIBR Bioinformatics and Research Computing

Sequence Patterns and Matrices

- Multiple sequence alignments
- Sequence patterns
- Sequence matrices
- Identifying regulatory sites
- Finding overrepresented patterns and profiles
- Gene finding

Why use DNA patterns and matrices?

- To help search the genome for …
  - Transcription start sites
  - Splice junctions (exon-intron boundaries)
  - Transcription factor binding sites
  - microRNA targets
  - Active sites for chromatin regulators
  - Gene regions encoding protein motifs
  - RNA folding patterns (hairpins, etc.)

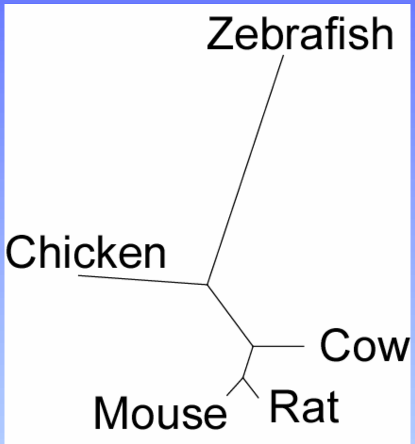
Multiple sequence alignments (MSAs)

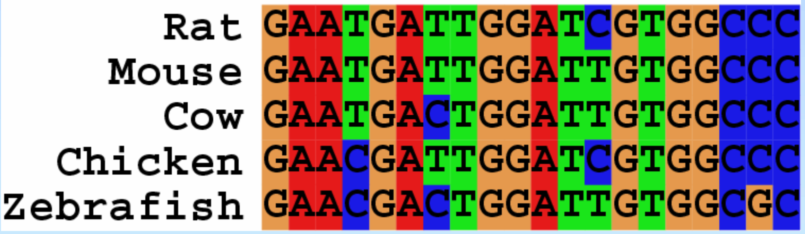
- Global MSA is computationally difficult
- As a result, MSA algorithms use approximate methods
- Independent of the chosen algorithm, choice of scoring matrix is important
- Aligning contigs vs. genes
- Aligning similar vs. divergent sequences

Global progressive MSA

- An MSA method that uses phylogenetic information to determine alignment order

1. Perform all pairwise alignments
2. Use alignment scores to create a tree
3. Align most similar pair of sequences and create consensus.
4. Align next most similar pair of sequences and create consensus … repeat until done





```
      Rat GAATGATTGGATCGTGGCCC
    Mouse GAATGATTGGATTGTGGCCC
      Cow GAATGACTGGATTGTGGCCC
  Chicken GAACGATTGGATCGTGGCCC
Zebrafish GAACGACTGGATTGTGGCGC
```

Sequence patterns

Pattern: an expression describing all possible combinations of bases in a sequence

- Generally derived from a MSA
- Ex1. EcoRI enzyme site: GAATTC
- Ex2. Codons for proline: CC[ACTG]; CCN
- Ex3. TATA box: TATA[AT][AGT][GA]
- Ex4. TFBS for GATA4: AGATA[AGT][AC]AGGGA
- Ex 5. Gene region encoding your favorite protein motif => better to use protein pattern!

More complex patterns

- May want to consider:
  - Mismatches
  - Insertions
  - Deletions
  - Alphabet reflecting ambiguity
- Ex: Patscan (Argonne National Laboratory) syntax
  - Pattern[Mismatches,Deletions,Insertions]
  - Ex: RRRRRYYYYY[3,2,1]

    (R = purine; Y = pyrimidine)

Pattern considerations

- Is there reliable data behind it?
- Is it specific and sensitive?
- How many matches would you expect by chance?
- Patterns don’t represent the different probabilities of each combination of bases, just whether they can occur or not.
- DNA or protein?

Pattern searching programs

- Check examples or help for syntax
- EMBOSS:
  - fuzznuc: nucleic acid pattern search
  - fuzzpro: protein pattern search
  - dreg: regular expression search of a nucleotide sequence
- PatScan
- Perl (programming language) regular expressions

Sensitivity and Specificity

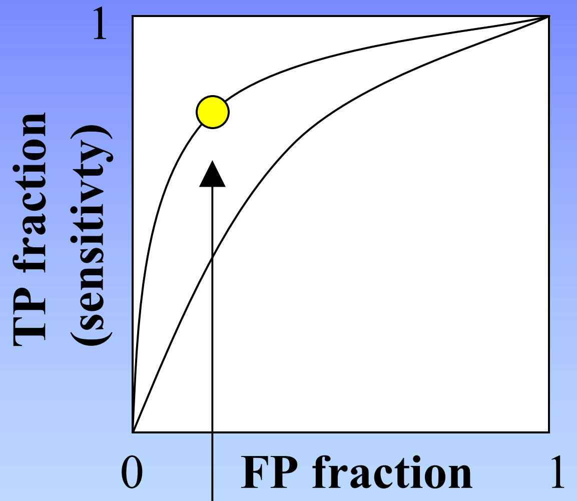
Proportion of true sites correctly identified:

$$\text{sensitivity} = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FN}}$$

Proportion of false sites correctly identified:

$$\text{specificity} = \frac{\mathrm{TN}}{\mathrm{TN} + \mathrm{FP}}$$

ROC plot:



Aim for “optimal” sensitivity

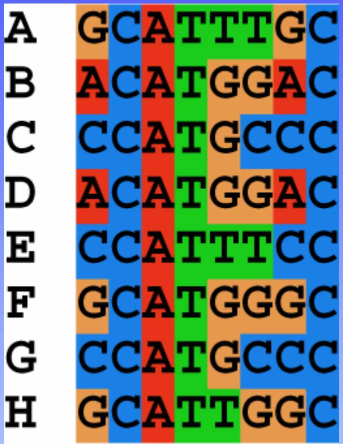
Matrix Representations

Matrix: a probabilistic representation of bases in a sequence

- Generally derived from a MSA
- Related to concept of "profile" (but no gaps allowed in MSA)
- Maintains meaning when transposed
- Position-specific scoring matrix (PSSM) assumes each position is independent
- Handling "zero" probabilities with pseudocounts

Creating a matrix (PSSM)

1. Create alignment



```
A  GCATTTGC
B  ACATGGAC
C  CCATGCCC
D  ACATGGAC
E  CCATTTCC
F  GCATGGGC
G  CCATGCCC
H  GCATTGGC
```

2. Count frequencies; add pseudocounts

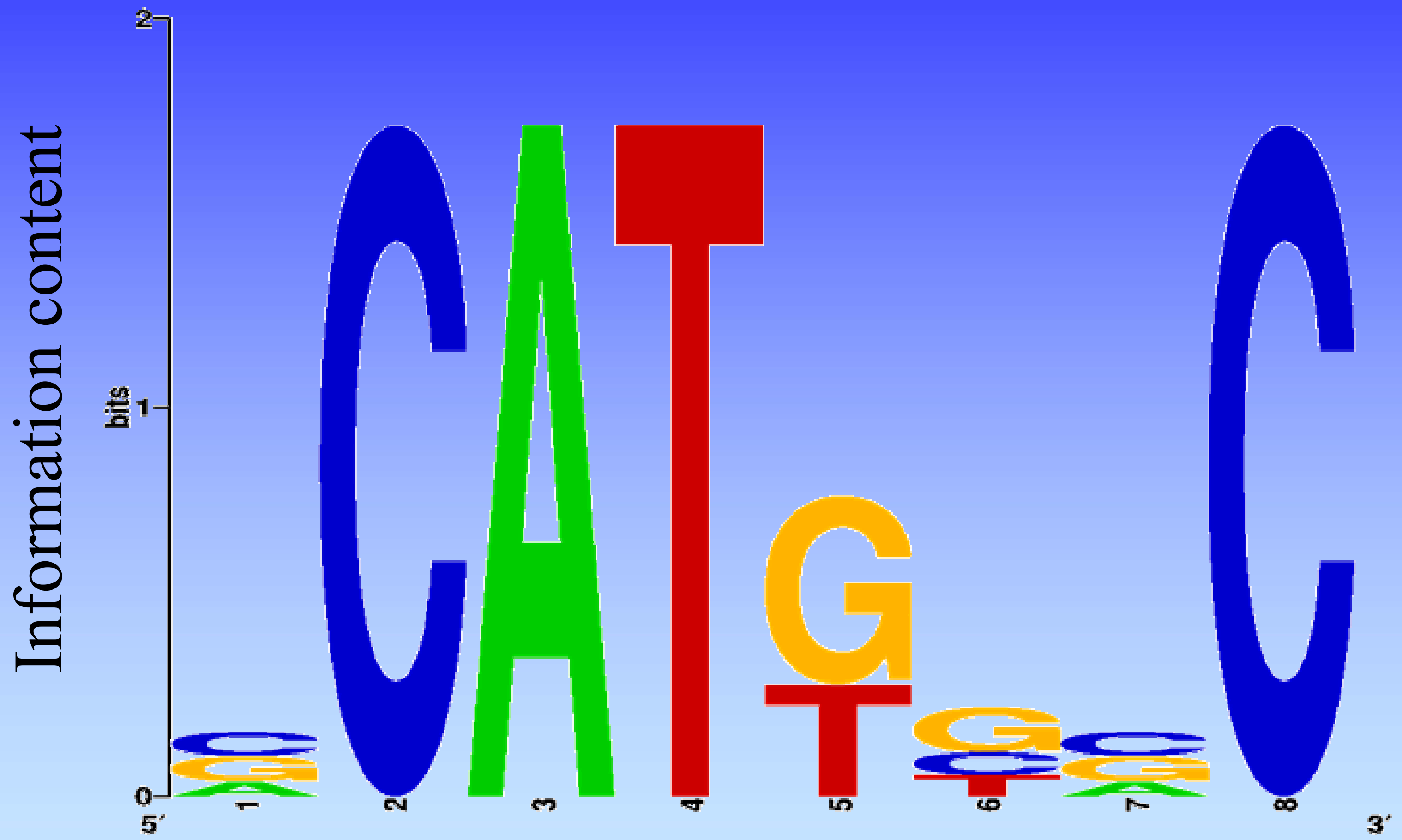
	1	2	3	4	5	6	7	8
A	2	ψ	8	ψ	ψ	ψ	2	ψ
C	3	8	ψ	ψ	ψ	2	3	8
G	5	ψ	ψ	ψ	5	4	3	ψ
T	ψ	ψ	ψ	8	3	2	ψ	ψ

$$\psi \approx \frac{\sqrt{n_{seq}}}{n_{\Psi}} = \frac{\sqrt{8}}{17} = 0.167$$

3. Calculate log-odds scores: $\log_2 (freq_{obs} / freq_{exp})$

	1	2	3	4	5	6	7	8
A	-2.6	-2.6	2	-2.6	-2.6	-2.6	-2.6	-2.6
C	0.6	2	-2.6	-2.6	-2.6	-2.6	0.6	2
G	1.3	-2.6	-2.6	-2.6	1.3	1	0.6	-2.6
T	-2.6	-2.6	-2.6	2	0.6	-2.6	-2.6	-2.6

Sequence logos

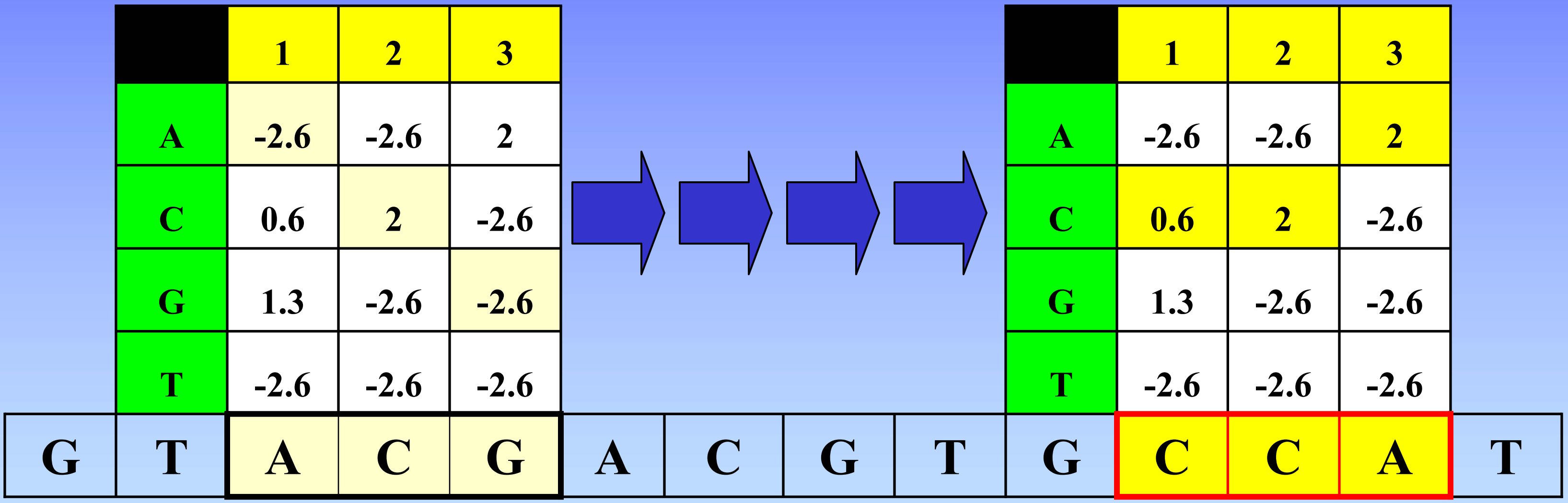


For DNA, maximum bits = 2 (for perfect consensus)

Searching with matrices

- Slide matrix along sequence(s), take sum of log-odds scores for base at each sequence position:

Example with sample matrix of length 3, showing scores at two positions



	1	2	3
A	-2.6	-2.6	2
C	0.6	2	-2.6
G	1.3	-2.6	-2.6
T	-2.6	-2.6	-2.6

	1	2	3
A	-2.6	-2.6	2
C	0.6	2	-2.6
G	1.3	-2.6	-2.6
T	-2.6	-2.6	-2.6

G	T	A	C	G	A	C	G	T	G	C	C	A	T

$$\Sigma = (-2.6) + 2 + (-2.6)$$
$$= -3.2$$

$$\Sigma = 0.6 + 2 + 2$$
$$= 4.6$$

Highest score wins

Hints for identifying regulatory sites

- Mask repetitive sequence first (RepeatMasker) to remove “non-functional noise”
- What specific area(s) of the sequence or genome can you search (instead of all of it)?
- Look at conservation: functional regulatory sites tend to be conserved
- ENCODE project (1% of human genome)

Identifying over-represented patterns

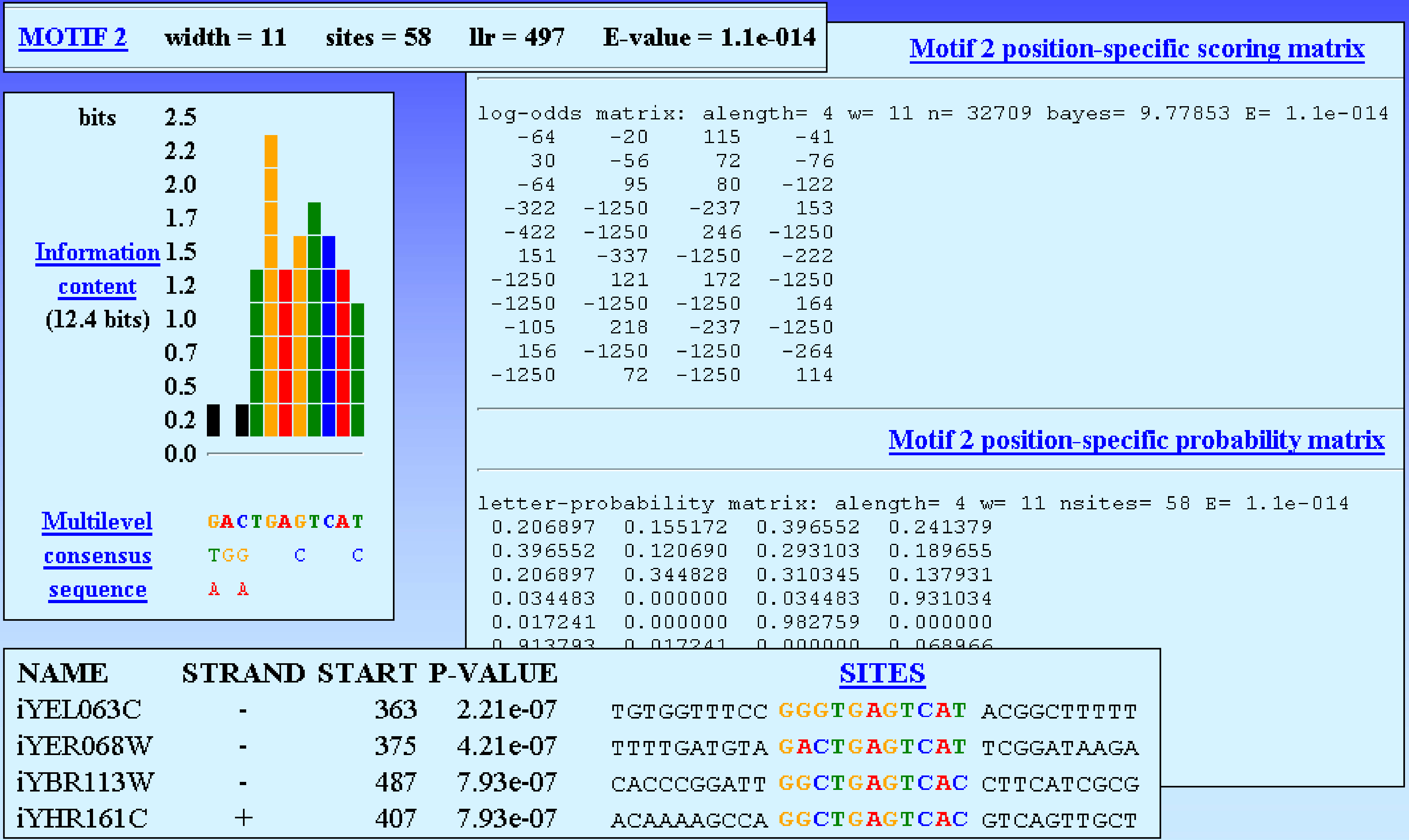
1. Count oligos of each sequence of expected length.
2. Calculate expected frequencies.
3. Rank observed/expected values.
4. Repeat for oligos of another length.

This method assumes the pattern is very specific

Identifying over-represented matrices

- Inputs
  - a set of sequences assumed to contain a matrix
  - range of presumed profile width?
  - ≥0 (“zoops”) or ≥1 (“oops”) occurrence per sequence?
- Programs
  - Meme: based on the expectation maximum (EM) algorithm; meme.sdsc.edu
  - AlignACE: based on the Gibbs sampling algorithm; atlas.med.harvard.edu

Meme sample output



MOTIF 2 width = 11 sites = 58 llr = 497 E-value = 1.1e-014

Information content (12.4 bits)

Multilevel consensus sequence

```
GACTGAGTCAT
TGG   C   C
A A
```

Motif 2 position-specific scoring matrix

```
log-odds matrix: alength= 4 w= 11 n= 32709 bayes= 9.77853 E= 1.1e-014
   -64     -20     115     -41
    30     -56      72     -76
   -64      95      80    -122
  -322   -1250    -237     153
  -422   -1250     246   -1250
   151    -337   -1250    -222
 -1250     121     172   -1250
 -1250   -1250   -1250     164
  -105     218    -237   -1250
   156   -1250   -1250    -264
 -1250      72   -1250     114
```

Motif 2 position-specific probability matrix

```
letter-probability matrix: alength= 4 w= 11 nsites= 58 E= 1.1e-014
 0.206897  0.155172  0.396552  0.241379
 0.396552  0.120690  0.293103  0.189655
 0.206897  0.344828  0.310345  0.137931
 0.034483  0.000000  0.034483  0.931034
 0.017241  0.000000  0.982759  0.000000
 0.913793  0.017241  0.000000  0.068966
```

NAME	STRAND	START	P-VALUE	SITES		
iYEL063C	-	363	2.21e-07	TGTGGTTTCC	GGGTGAGTCAT	ACGGCTTTTT
iYER068W	-	375	4.21e-07	TTTTGATGTA	GACTGAGTCAT	TCGGATAAGA
iYBR113W	-	487	7.93e-07	CACCCGGATT	GGCTGAGTCAC	CTTCATCGCG
iYHR161C	+	407	7.93e-07	ACAAAAGCCA	GGCTGAGTCAC	GTCAGTTGCT

Identifying features of genes in genomic DNA

- Splice sites
- Open reading frames
- Promoters
- Codon bias
- Expression information (ESTs, mRNA)
- Protein similarity to known genes
- Conservation across species

Gene finding programs (sample)

- GeneWise (Birney and Durbin, 2000)
- Genscan (Burge and Karlin, 1997)
- Acembly (Thierry-Mieg et al.)
- Twinscan (Korf et al., 2001)
- SGP (Parra et al., 2003)
- GeneID (Parra et al., 2000)

Use all available data and predictions when possible

Summary

- Multiple sequence alignments
- Sequence patterns
- Sequence matrices
- Identifying regulatory sites
- Finding over-represented patterns and matrices
- Gene finding

References

- Bioinformatics: Sequence and Genome Analysis, 2nd ed. David Mount. CSHL Press, 2004.
- Publications describing algorithms and software for
  - multiple sequence alignment
  - pattern and matrix analysis and searching
  - gene finding

Exercises

1. Investigating the mechanisms of miRNA activity through pattern searching

2. Studying transcriptional control with DNA matrices

Both involve computational analysis of data from recently published studies