# Pattern Recognition and Gene Analysis

Manu Madhavan

Lecture 13

# Recap

- Gene Expression
- Microarray technology

# Outline

- Pattern Recognition
- Patterns in gene/protein sequences
- Representation: Regular Expression
- ML, ANN, HMM methods

# Pattern Recognition in Bioinformatics

- Only about 5% of the genome contains useful patterns of nucleotides, or genes, that code for proteins.
- The initiation of translation or transcription process is determined by the presence of specific patterns of DNA or RNA, or motifs.
- Research on detecting specific patterns of DNA sequences such as genes, protein coding regions, promoters, etc., leads to uncover functional aspects of cells.
- **Comparative genomics** focus on comparisons across the genomes to **find conserved patterns over the evolution, which possess some functional significance**.

# Pattern Representation: RE

- Literal match: `re.find('GAATT')`
- Character set: 'CC[GA][TC]GG'

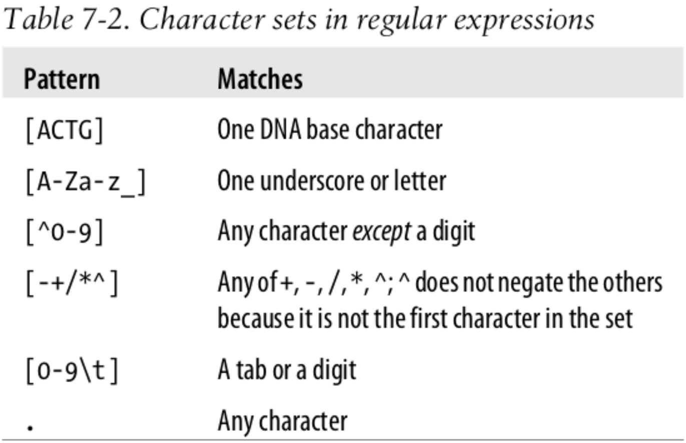
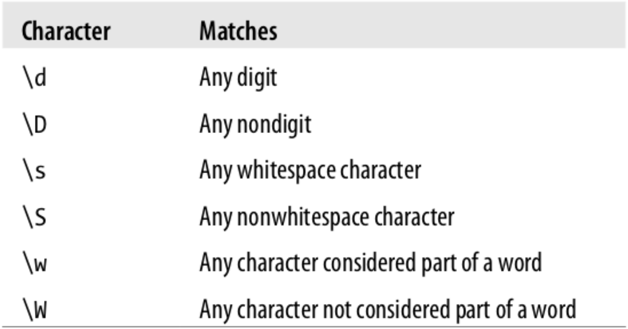


*Table 7-2. Character sets in regular expressions*

| Pattern | Matches |
|---|---|
| [ACTG] | One DNA base character |
| [A-Za-z_] | One underscore or letter |
| [^0-9] | Any character *except* a digit |
| [-+/*^] | Any of +, -, /, *, ^; ^ does not negate the others because it is not the first character in the set |
| [0-9\t] | A tab or a digit |
| . | Any character |

# Pattern Representation: RE

*Table 7-3. Character classes in regular expressions*



| Character | Matches |
|---|---|
| \d | Any digit |
| \D | Any nondigit |
| \s | Any whitespace character |
| \S | Any nonwhitespace character |
| \w | Any character considered part of a word |
| \W | Any character not considered part of a word |

# Pattern Representation: RE

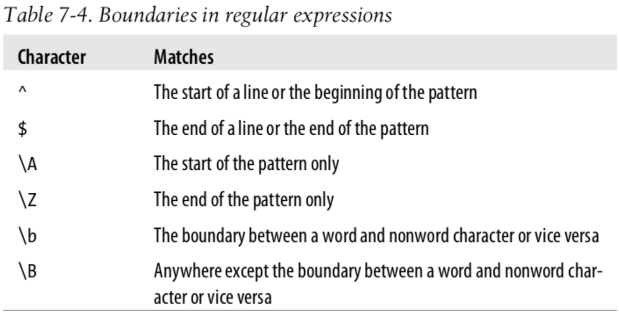


*Table 7-4. Boundaries in regular expressions*

| Character | Matches |
|---|---|
| ^ | The start of a line or the beginning of the pattern |
| $ | The end of a line or the end of the pattern |
| \A | The start of the pattern only |
| \Z | The end of the pattern only |
| \b | The boundary between a word and nonword character or vice versa |
| \B | Anywhere except the boundary between a word and nonword character or vice versa |

# Pattern Representation: RE

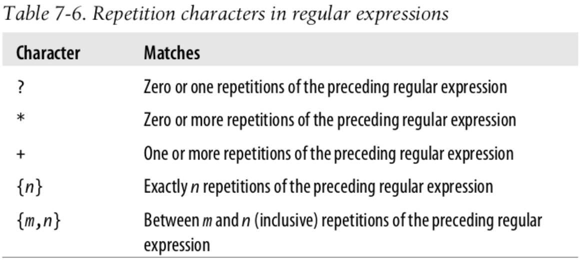


*Table 7-6. Repetition characters in regular expressions*

| Character | Matches |
|---|---|
| ? | Zero or one repetitions of the preceding regular expression |
| * | Zero or more repetitions of the preceding regular expression |
| + | One or more repetitions of the preceding regular expression |
| {*n*} | Exactly *n* repetitions of the preceding regular expression |
| {*m*,*n*} | Between *m* and *n* (inclusive) repetitions of the preceding regular expression |

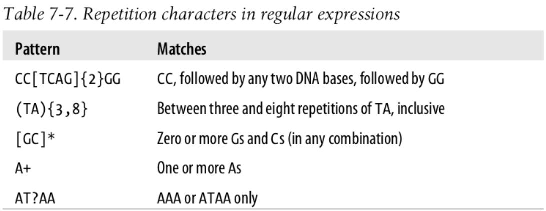


*Table 7-7. Repetition characters in regular expressions*

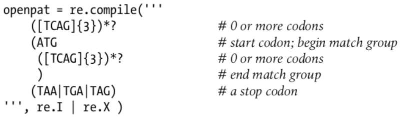
| Pattern | Matches |
|---|---|
| CC[TCAG]{2}GG | CC, followed by any two DNA bases, followed by GG |
| (TA){3,8} | Between three and eight repetitions of TA, inclusive |
| [GC]* | Zero or more Gs and Cs (in any combination) |
| A+ | One or more As |
| AT?AA | AAA or ATAA only |

# Pattern Representation: RE

- Write a regular expression for ORF pattern

# Pattern Representation: RE

- Write a regular expression for ORF pattern



```
openpat = re.compile('''
    ([TCAG]{3})*?            # 0 or more codons
    (ATG                     # start codon; begin match group
     ([TCAG]{3})*?           # 0 or more codons
     )                       # end match group
    (TAA|TGA|TAG)            # a stop codon
''', re.I | re.X )
```

# Pattern Representation: RE

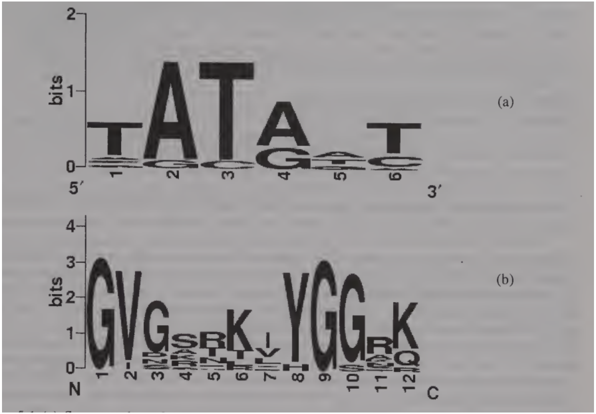
Example: PROSITE

- **RE**-Regular Expression
- The standard IUPAC one-letter code is used for amino acids
- Each element is separated by '-'
- Symbol 'x' is sued for a position where any amino acid is accepted
- More than one accepted amino acid: listed between '

  '
- To specify not acceptable: use { and }

# Probabilisitc Patterns

- Identifying most prominent consensus sequence (by identifying the patterns at each position)



# Pattern characterisation and classification

- The first is to use the sequence pattern to identify structural, and consequently functional features that are common to a set of proteins
- variations are possible
- Especially for new patterns of unknown structure and function, that the conservation is a result of chance and has no biological.significance
- p-score statistics
- how well a particular pattern is diagnostic of membership in a specific sequence family
  - Specificity: $\frac{TN}{TN+FP}$
  - Sensitivity: $\frac{TP}{TP+FN}$
  - Positive Predict Value (PPV): $\frac{TP}{TP+FP}$

# Pattern Discovery

- The first task is to understand and decide on the type of patterns that the process will result in.
- For example, we may be interested, say, in only repeating patterns, in which identical, or similar residues repeat at regular fixed intervals along the sequence.
- Measure the fitness of the pattern
- Methods: Classification and clustering

# ANN Based pattern Discovery

- Use an ensemble of neural networks to identify the different patterns

# HMM Based pattern Discovery

- For pattern identification
- Profile-HMM