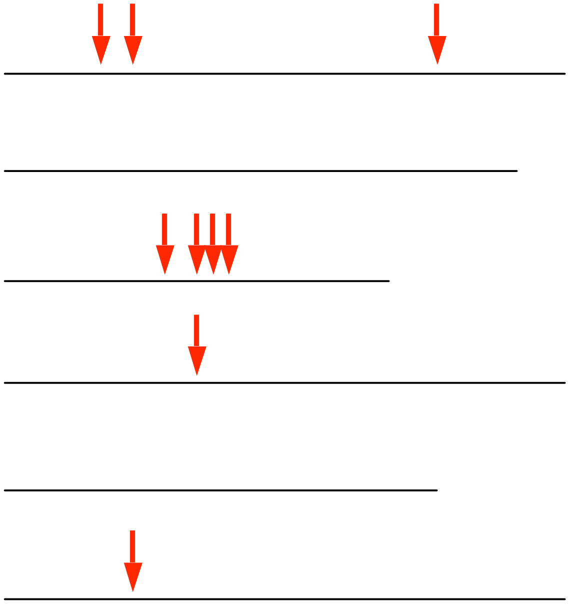
Regulatory Sequence Analysis

String-based pattern matching

Jacques van Helden
Jacques.van.Helden@ulb.ac.be

Word counting - Occurrences or matching sequences

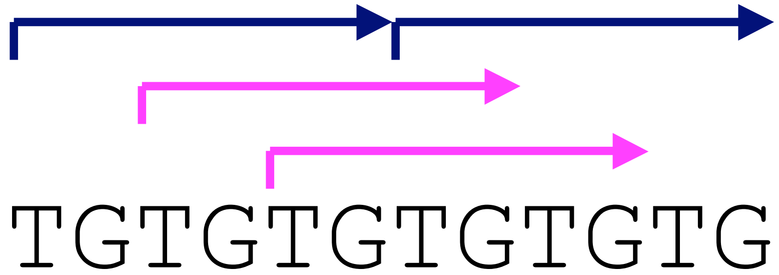
- If a sequence contains multiple occurrences of a given pattern, one can score either
  - ***all of them***, or
  - only count the ***first occurrence per sequence***. In this case, each sequence is scored as "matching" the pattern or not.



	All occurrences	First occurrence
Seq 1	3	true
Seq 2	0	false
Seq 3	4	true
Seq 4	1	true
Seq 5	0	false
Seq 6	1	true
Total	**9 occ**	**4 mseq**

Treatment of self-overlap

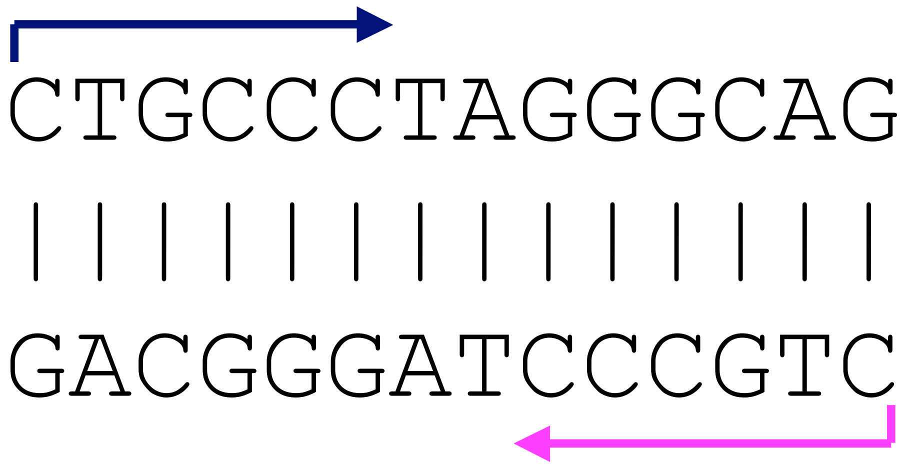
Mutually overlapping occurrences of the same word.



TGTGTGTGTGTG

2 or 4 occurrences of TGTGTG ?

Single or double strand count



```
CTGCCCTAGGGCAG
||||||||||||||
GACGGGATCCCGTC
```

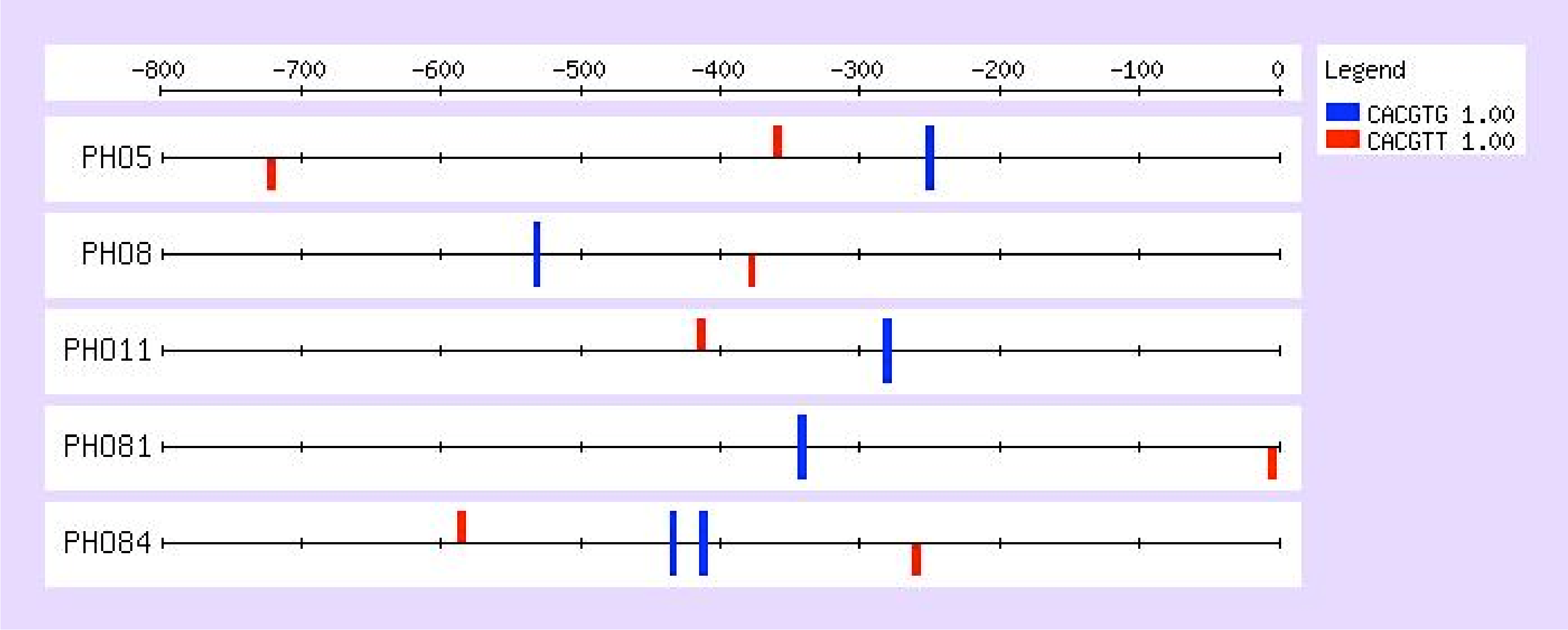
1 or 2 occurrences of CTGCCC ?

Software : dna-pattern

- Specialized program for pattern matching in DNA sequences
  - Supports IUPAC code for partly specified nucleotides (e.g. TSWNATTK)
  - Supports spaces of fixed or variable length within the patterns (e.g. $GGGWn_{\{0,30\}}WCCC$)
  - Single or both strands
  - Allow substitutions but no insertion or deletion
- Extract neighbourhood of the match (flanking bases)
- Return
  - matching positions
  - match count per sequence
- Sliding window
  - Detection of regions containing combinations of multiple patterns
  - A specific weight can be associated to each pattern

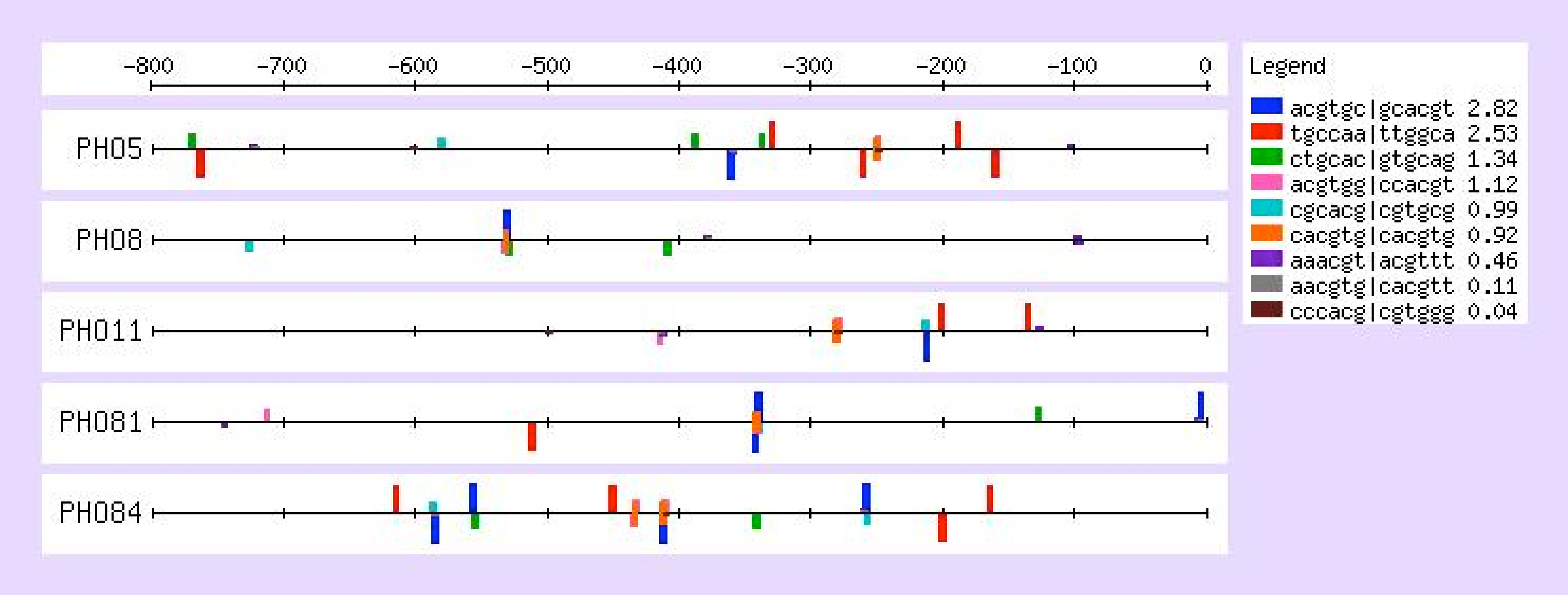
Matching simple patterns

- A simple string-based pattern matching is usually poorly informative.
  - spurious matches are expected to be found anywhere
  - the presence of the consensus does not necessarily mean that the factor binds
  - some patterns have a higher significance than other ones (e.g. the core of the consensus).



Assigning scores to patterns

- Pattern-specific scores can improve the interpretation by highlighting the most significant patterns.
- Scores can be assigned arbitrarily (e.g. on the basis of prior biological knowledge) or reflect the significance calculated by pattern discovery programs.



Sliding windows - scoring mutually overlapping matches

Sliding windows - scoring successions of matches