

String-based pattern matching

Jacques van Helden

<https://orcid.org/0000-0002-8799-8584>

Aix-Marseille Université, France



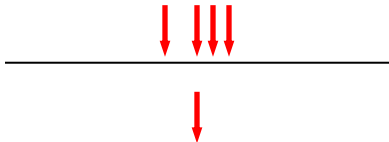



Theory and Approaches of Genome Complexity (TAGC)

Institut Français de Bioinformatique (IFB)

<http://www.france-bioinformatique.fr>

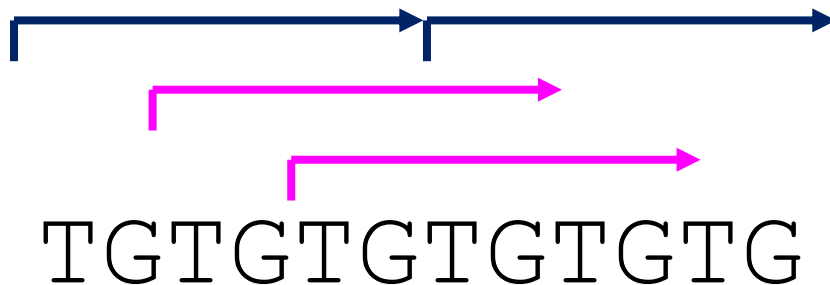
Word counting - Occurrences or matching sequences

- 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

- Some words are self-overlapping.
- For such words, one can count
 - either the **renewing occurrences** only (2 occurrences in the example below)
 - or all occurrences (2 **renewing** and 2 **overlapping** in the example below).
- The choice of the counting mode strongly affects the subsequent statistics (dependency/independency).



2 or 4 occurrences of TGTGTG ?

Single or double strand count

- A particularity of DNA sequence is their double-strand structure.
- Words can be counted either on a single strand, or on both, depending on the nature of the expected biological signal.
 - In RNA sequences, single-strand counts are generally suited.
 - In DNA sequences, both-stands counts can be relevant for cis-regulatory signals, because many transcription factor act in an orientation-independent way.



1 or 2 occurrences of CTGCCC ?

Symmetries in DNA sequences

This English sentence is a palindrome: it is symmetrical relative to the central letter (R). Palindromic sentences have the same succession of letters when read in either direction.

WASITARATISAW

The following sequence contains a textual palindrome

5'-ATGGGC CGGGTA-3'

However, there is no symmetry in the corresponding DNA molecule.

5'-ATGGGC CGGGTA-3'
||||||| |||||
,5'-TACCCG GCCCAT-3'

The following sequence contains no textual palindrome

5'-ATGGGC GCCCAT-3'

However, there is a “reverse complementary palindromic” symmetry in the corresponding DNA molecule: the molecule has the same succession of nucleotides when “read” on either strands (always from 5' to 3' end).

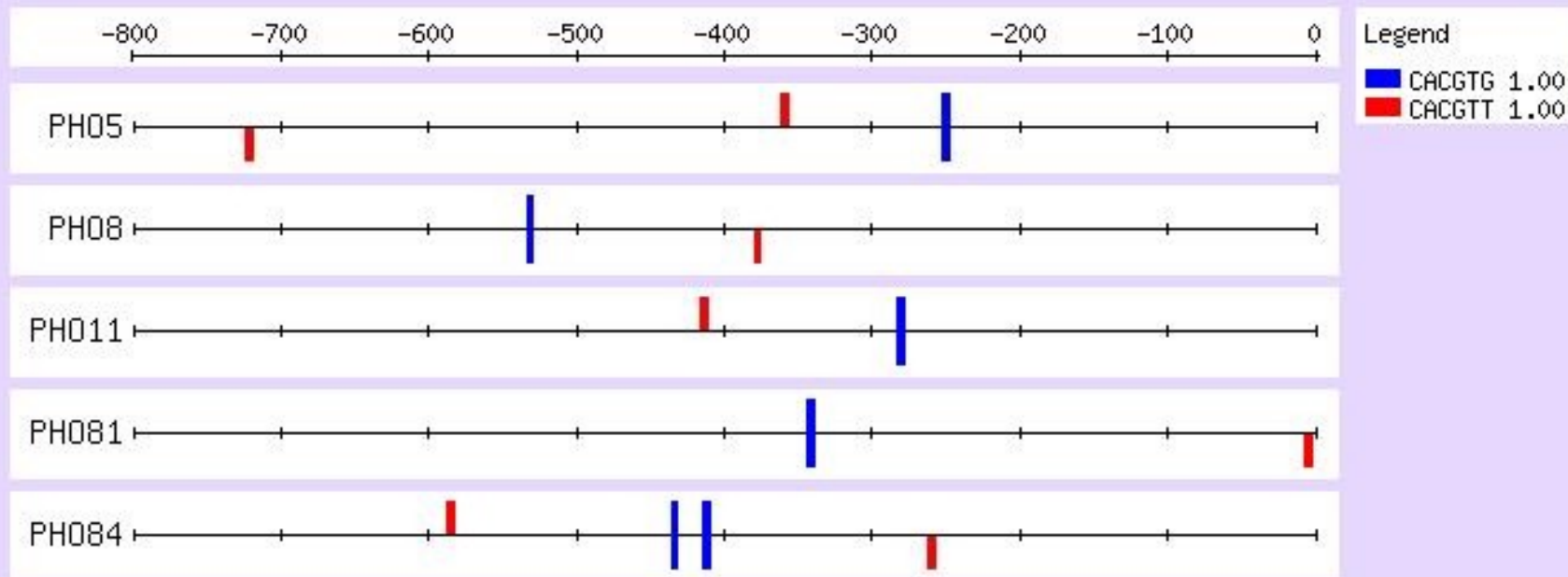
5'-ATGGGC GCCCAT-3'
||||||| . |||||
,5'-ATGGGC GCCCAT-3'

RSAT tool: 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

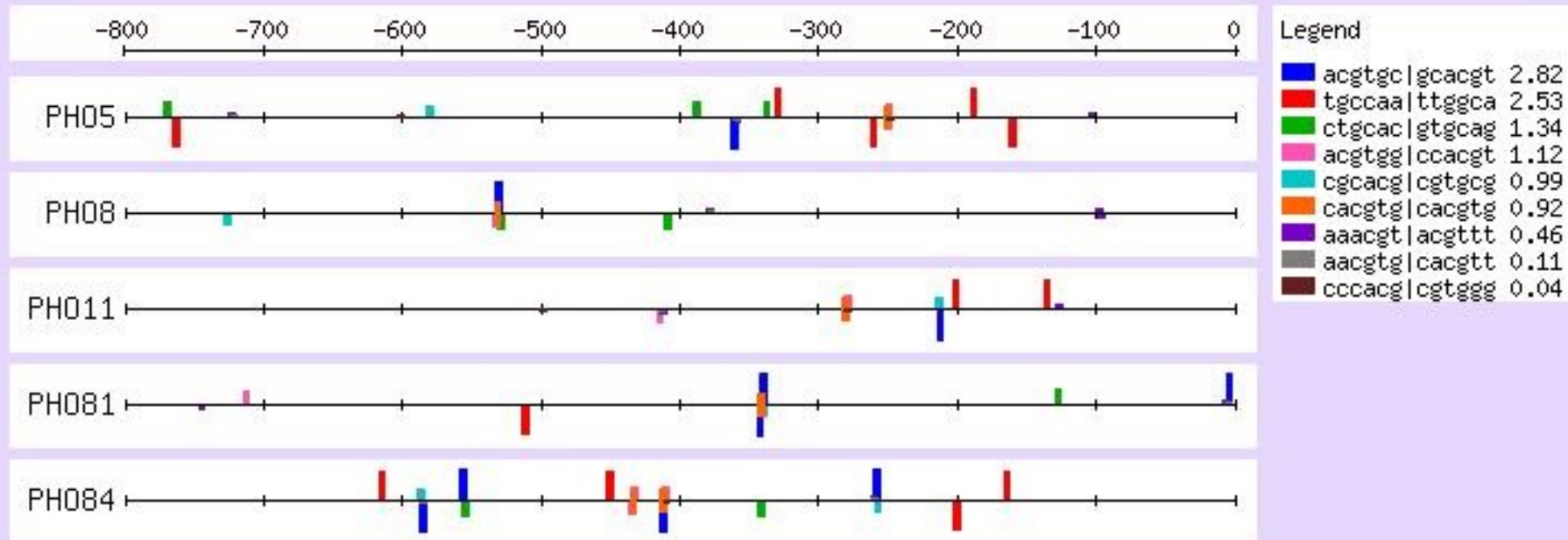
Matching simple patterns

- 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).



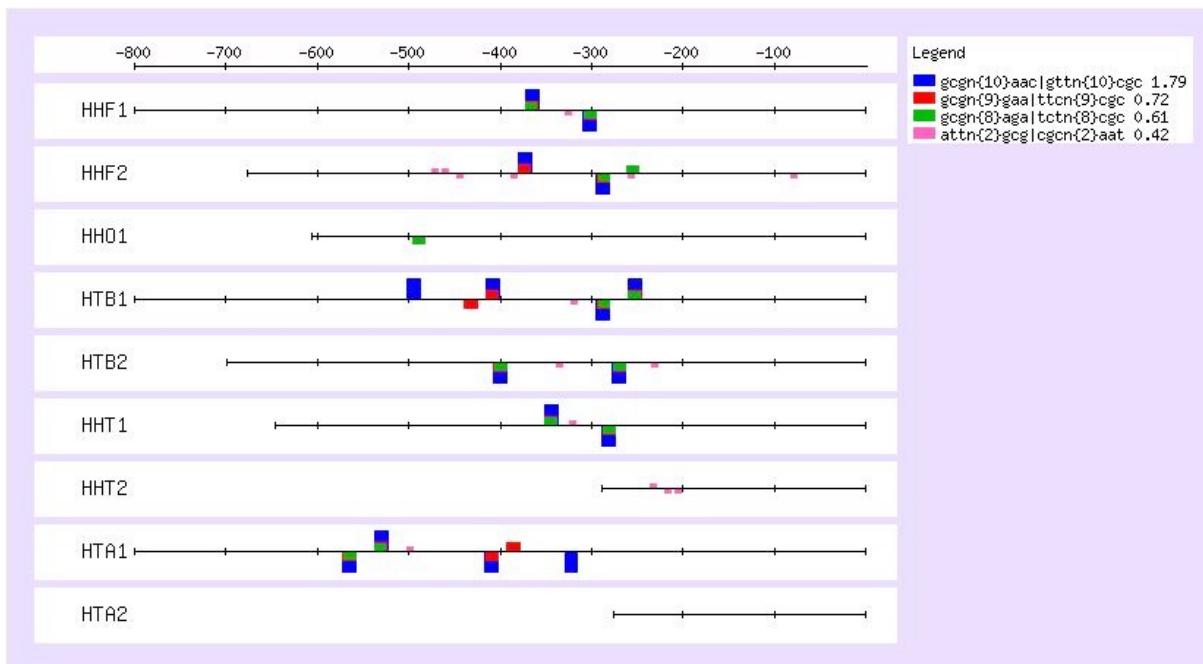
Assigning scores to patterns

- 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.



Matching a collection of overlapping patterns

- Pattern matching results can be improved by matching a collection of mutually overlapping patterns (oligonucleotides or spaced dyads)
 - Multiple patterns can be used to represent fragments of a larger binding site, or the variants arising from the degeneracy of the consensus.
 - Specific weights can be assigned to the elements of the collection, to represent their relative importance for the binding.



Patterns with ambiguous IUPAC code

FadR binding motif descriptions from RegulonDB

- Source

- <http://regulondb.ccg.unam.mx/regulon?term=ECK120011824&format=jsp&organism=ECK12&type=regulon>

Genome-scale pattern matching

Genome-scale pattern matching

- Knowing the consensus binding site for a given transcription factor, one would be tempted to use this information for predicting its target genes in the whole genome.
- This approach is however very inaccurate, because
 - The consensus poorly reflects the binding specificity
 - Binding is not synonymous of regulation
- As an experiment, we counted the number of occurrences for the consensus of various yeast transcription factors (source: TRANSFAC + our annotations). For each one of the
 - 800bp upstream sequences, clipped to prevent upstream ORFs.
 - For each gene, we counted the number of occurrences of each consensus.

Occurrences of consensus per promoter

- Saccharomyces cerevisiae, 5864 genes, 800bp (clipped if upstream ORF)

| factor | sequence | mean | >= 1 | >=2 | >=3 | >=4 | >=5 | >=6 | 7 |
|-------------|---------------------------|------|------|------|-----|-----|-----|-----|----|
| ABF1 | TCNNNNNNNACG | 0.62 | 788 | 196 | 43 | 13 | 5 | 0 | 0 |
| ABF1.1 | RTCRYNNNNNACG | 0.15 | 61 | 5 | 2 | 0 | 0 | 0 | 0 |
| ADR1 | GGRGK | 1.52 | 2187 | 1283 | 725 | 414 | 230 | 113 | 72 |
| CBF1 | RTCACRTG | 0.07 | 70 | 10 | 6 | 0 | 0 | 0 | 0 |
| CCBF | RNNYCACGAAAA | 0.01 | 1 | 1 | 1 | 1 | 1 | 0 | 0 |
| GAL4 | CGGNNNACWNTCSTCCGARS | 0.00 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| GATA_L5 | GATAA | 1.63 | 2588 | 1322 | 685 | 322 | 168 | 89 | 40 |
| GATA_L6 | GATAAG | 0.28 | 266 | 46 | 14 | 7 | 2 | 1 | 0 |
| GCN4 | SKRTGASTCAYMS | 0.00 | 2 | 0 | 0 | 0 | 0 | 0 | 0 |
| GCN4.1 | SNSNNNNNRTGACTCATNS | 0.00 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| GCR1 | RGCTTCCWC | 0.03 | 14 | 1 | 0 | 0 | 0 | 0 | 0 |
| GFII | RTCACRTG | 0.07 | 70 | 10 | 6 | 0 | 0 | 0 | 0 |
| HAP1 | aaCttCCGWTAWCtCCNtNCNNNNT | 0.00 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| HAP2 | YCINNCCAATNANM | 0.01 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| HAP3 | YCINNCCAATNANM | 0.01 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| HAP4 | YCINNCCAATNANM | 0.01 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| MATa1 | TGATGTANNT | 0.04 | 4 | 0 | 0 | 0 | 0 | 0 | 0 |
| MATalpha2 | CRTGTNNW | 0.74 | 1088 | 345 | 87 | 15 | 6 | 2 | 1 |
| MCM1 | WTWCCYAAWNNGGTAA | 0.00 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| MET4_core | CACGTG | 0.16 | 401 | 49 | 49 | 8 | 8 | 2 | 2 |
| MET4_L8 | TCACGTGA | 0.04 | 101 | 2 | 2 | 0 | 0 | 0 | 0 |
| MIG1 | KANWWWWATSYGGGGW | 0.00 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| PHO4 | CACGTKBG | 0.04 | 17 | 2 | 0 | 0 | 0 | 0 | 0 |
| PHO4_both | CACGTK | 0.32 | 460 | 118 | 61 | 28 | 10 | 3 | 3 |
| PHO4_high | CACGTG | 0.16 | 401 | 49 | 49 | 8 | 8 | 2 | 2 |
| PHO4_medium | CACGTT | 0.16 | 74 | 5 | 0 | 0 | 0 | 0 | 0 |