Motif discovery

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

The rationale beyond motif discovery

- We have a set of sequences supposed to share some regulatory signal
  - Examples:
    - upstream sequences from a set of co-expressed genes
    - upstream sequences from a set of genes involved in the same metabolic pathway
    - sequences around the termination sites of all the genes of a given organism
    - bacterial genomes containing restriction sites for a given restriction enzyme
- We don't know a priori de motif
- General approach: detect so-called “exceptional” motifs i.e. motifs whose instances discard from the expectation of the background model
- Criteria for exceptionality
  - Over-representation
  - Under-representation (avoided signals)
  - Positional bias
- Motif descriptions
  - String-based (consensus, strict consensus, regular expression)
  - Position-specific scoring matrices (motif profiles)
  - Hidden Markov Models (HMM, not seen in this course)

Motif discovery : typical cases

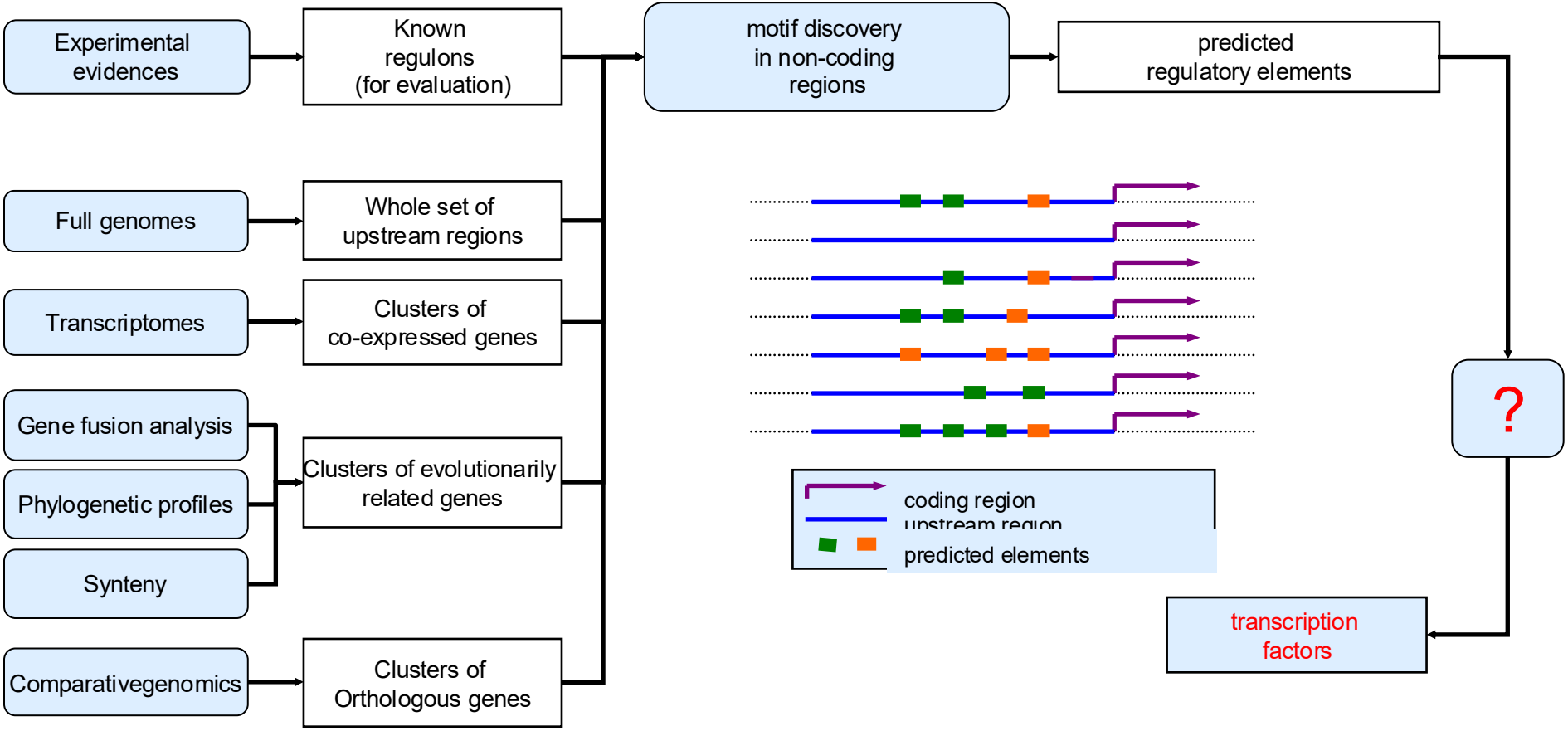
- Small sequence set
  - e.g. family of 20 co-regulated genes, obtained from DNA chip experiment
    → identify putative regulatory sites
- Sorted sequence lists
  - e.g. intergenic fragment sorted by affinity for a given transcription factor, based-on ChIP-seq results.
- Genome-scale motif discovery
  - In full genomes
    - Identify over-represented motifs in full genomes
    - Identify under-represented motifs in full genomes (e.g. organism-specific restriction sites in bacterial genomes)
  - In all upstream sequences
    - identify transcription initiation signals
    - identify binding sites for general transcription factors
  - In all downstream sequences
    - identify 3' maturation signals

Motif discovery: from sequences to motifs

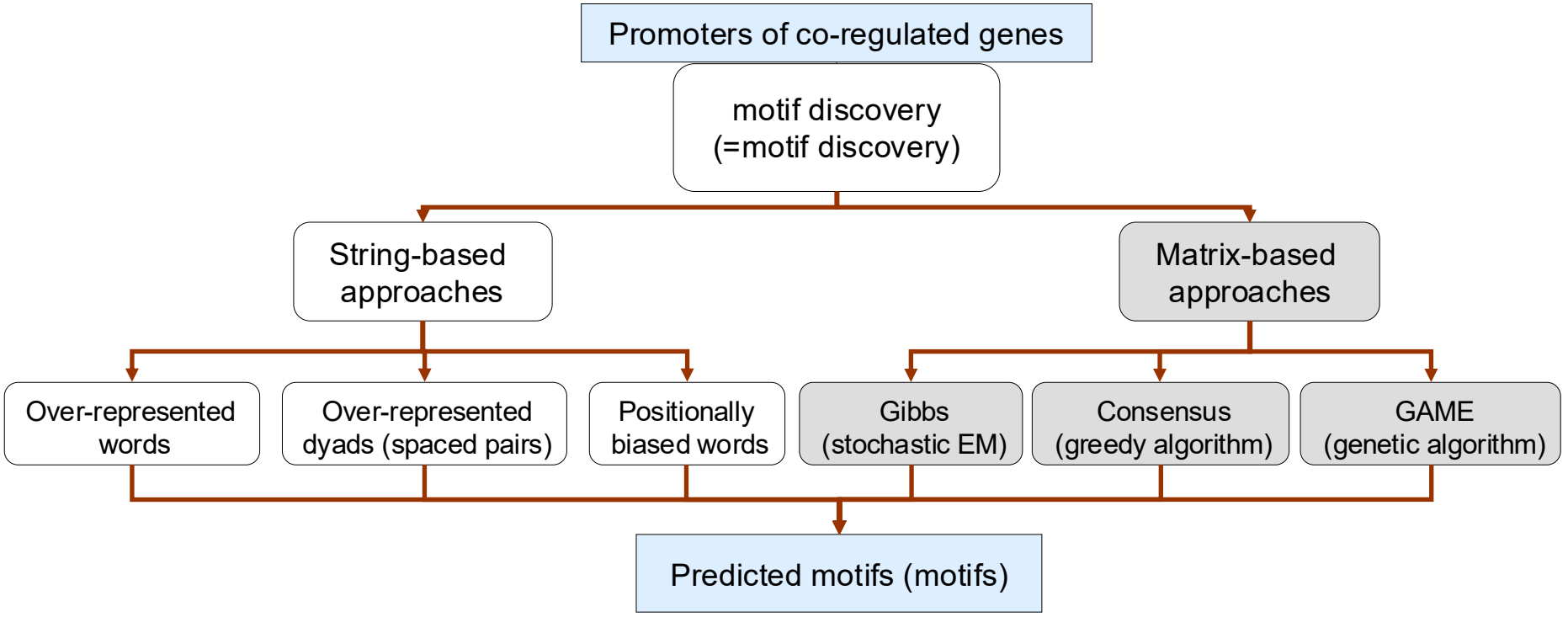
```
>YAR071W; upstream from -800 to -1; size: 800
GCAGCCTCTACCATGTTGCAAGTGCGAACCATACTGTGGCCACATAGATTACAAAAAAAG
TCCAGGATATCTTGCAAACCTAGCTTGTTTTGTAAACGACATTGAAAAAAGCGTATTAAG
GTGAAACAATCAAGATTATCTATGCCGATGAAAAATGAAAGGTATGATTTCTGCCACAAA
TATATAGTAGTTATTTTATACATCAAGATGAGAAAATAAAGGGATTTTTTCGTTCTTTTA
TCATTTTCTCTTTCTCACTTCCGACTACTTCTTATATCTACTTTCATCGTTTCATTCATC
GTGGGTGTCTAATAAAGTTTTAATGACAGAGATAACCTTGATAAGCTTTTTCTTATACGC
TGTGTCACGTATTTATTAAATTACCACGTTTTCGCATAACATTCTGTAGTTCATGTGTAC
TAAAAAAAAAAAAAAAAAAGAAATAGGAAGGAAAGAGTAAAAAGTTAATAGAAAACAGAA
CACATCCCTAAACGAAGCCGCACAATCTTGGCGTTCACACGTGGGTTTAAAAAGGCAAAT
TACACAGAATTTCAGACCCTGTTTACCGGAGAGATTCCATATTCCGCACGTCACATTGCC
AAATTGGTCATCTCACCAGATATGTTATACCCGTTTTGGAATGAGCATAAACAGCGTCGA
ATTGCCAAGTAAAACGTATATAAGCTCTTACATTTCGATAGATTCAAGCTCAGTTTCGCC
TTGGTTGTAAAGTAGGAAGAAGAAGAAGAAGAAGAGGAACAACAACAGCAAAGAGAGCAA
GAACATCATCAGAAATACCA
>YBR092C; upstream from -446 to -1; size: 446
TTTGTATAACTAAATAATATTGGAAACTAAATACGAATACCCAAATTTTTTATCTAAAT
TTTGCCGAAAGATTAAAATCTGCAGAGATATCCGAAACAGGTAAATGGATGTTTCAATCC
CTGTAGTCAGTCAGGAACCCATATTATATTACAGTATTAGTCGCCGCTTAGGCACGCCTT
TAATTAGCAAAATCAAACCTTAAGTGCATATGCCGTATAAGGGAAACTCAAAGAACTGGC
ATCGCAAAAATGAAAAAAAGGAAGAGTGAAAAAAAAAAAATTCAAAAGAAATTTACTAAA
TAATACCAGTTTGGGAAATAGTAAACAGCTTTGAGTAGTCCTATGCAACATATATAAGTG
CTTAAATTTGCTGGATGGAAGTCAATTATGCCTTGATTATCATAAAAAAAATACTACAGT
AAAGAAAGGGCCATTCCAAATTACCT
>YBR093C; upstream from -800 to -1; size: 800
TTTTACACATCGGACTGATAAGTTACTACTGCACATTGGCATTAGCTAGGAGGGCATCCA
AGTAATAATTGCGAGAAACGTGACCCAACTTTGTTGTAGGTCCGCTCCTTCTAATAATCG
CTTGTATCTCTACATATGTTCTATTTACTGACCGAAAGTAGCTCGCTACAATAATAATGT
TGACCTGATGTCAGTCCCCACGCTAATAGCGGCGTGTCGCACGCTCTCTTTACAGGACGC
CGGAGACCGGCATTACAAGGATCCGAAAGTTGTATTCAACAAGAATGCGCAAATATGTCA
ACGTATTTGGAAGTCATCTTATGTGCGCTGCTTTAATGTTTTCTCATGTAAGCGGACGTC
GTCTATAAACTTCAAACGAAGGTAAAAGGTTCATAGCGCTTTTTCTTTGTCTGCACAAAG
AAATATATATTAAATTAGCACGTTTTCGCATAGAACGCAACTGCACAATGCCAAAAAAAG
TAAAAGTGATTAAAAGAGTTAATTGAATAGGCAATCTCTAAATGAATCGATACAACCTTG
GCACTCACACGTGGGACTAGCACAGACTAAATTTATGATTCTGGTCCCTGTTTTCGAAGA
...
```

- Situation
  - We received a set of sequences supposedly co-regulated.
  - We ignore the transcription factors involved in this regulation.
  - We ignore the cis-acting elements (motifs and binding sites).
- Questions
  - Could we discover some signals (motifs) from these sequences only ?
    - This is a problem of ***motif discovery*** (“*ab initio*” motif detection)
  - Can we afterwards report the instances of these discovered motifs in the input sequences ?
    - This is a problem of ***motif matching***.
  - Can we predict the transcription factor that would bind the discovered motifs ?
    - By comparison with a library of known factors
      - ***motif comparison***
    - From the genome only
      - This is a difficult problem.

Motif discovery: groups of functionally related genes



Motif discovery: approaches



Motif discovery approaches

- String-based approaches
  - detection of over-represented words
  - oligo-analysis (single words)
  - dyad-detector (pairs of words separated by a spacer)
- Matrix-based approaches
  - Greedy algorithms (consensus)
    - progressive incorporation of more sequences into the motif
  - Heuristic algorithms
    - Iterative optimization of the motif
    - Gibbs sampler (gibbs, alignACE)
- Hidden Markov models (YEBIS, MEME)