Regulatory sequence analysis

Pattern discovery

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

Pattern discovery : goal

- We have a set of sequences
- We suspect that they share some functional signal
- We don't know the pattern of this signal
- General approach: detect unexpected patterns
  - Over-representation
  - Under-representation (avoided signals)
  - Positional bias
- Pattern descriptions
  - String-based descriptions
  - Position-specific scoring matrices (motif profiles)

Pattern discovery : typical cases

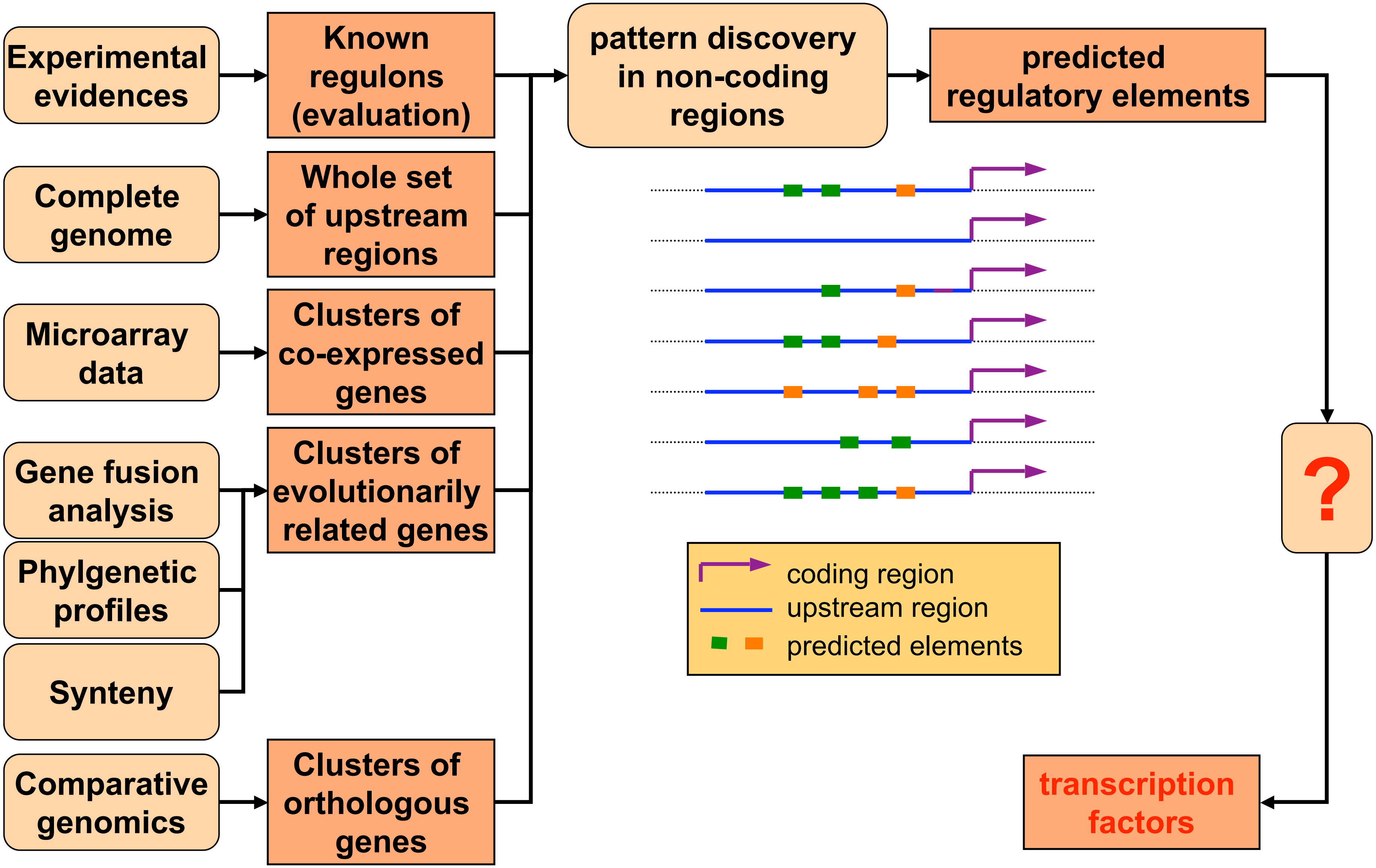
- Small sequence set
  - e.g. family of 20 co-regulated genes, obtained from DNA chip experiment
    → identify putative regulatory sites
- Sorted sequence lists
  - e.g. intergenic fragment sorted by affinity for a given transcriptin factor, on the basis of a ChIP-chip experiment.
- Genome-scale pattern 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

Pattern 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
  - Let us assume we receive a set of sequences supposed to be 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) on the basis of these sequences ?
    - This is a problem of ***pattern discovery***
("ab initio" motif detection)
  - Can we afterwards report the instances of these discovered motifs in the input sequences ?
    - This is a problem of ***pattern matching***.
  - Can we predict the transcription factor that would bind the discovered motifs ?
    - By comparison with a library of known factors
      - ***Pattern comparison***
    - From the genome only
      - This is a difficult problem.

Pattern discovery: groups of functionally related genes



Experimental evidences

Known regulons (evaluation)

Complete genome

Whole set of upstream regions

Microarray data

Clusters of co-expressed genes

Gene fusion analysis

Phylgenetic profiles

Synteny

Clusters of evolutionarily related genes

Comparative genomics

Clusters of orthologous genes

pattern discovery in non-coding regions

predicted regulatory elements

?

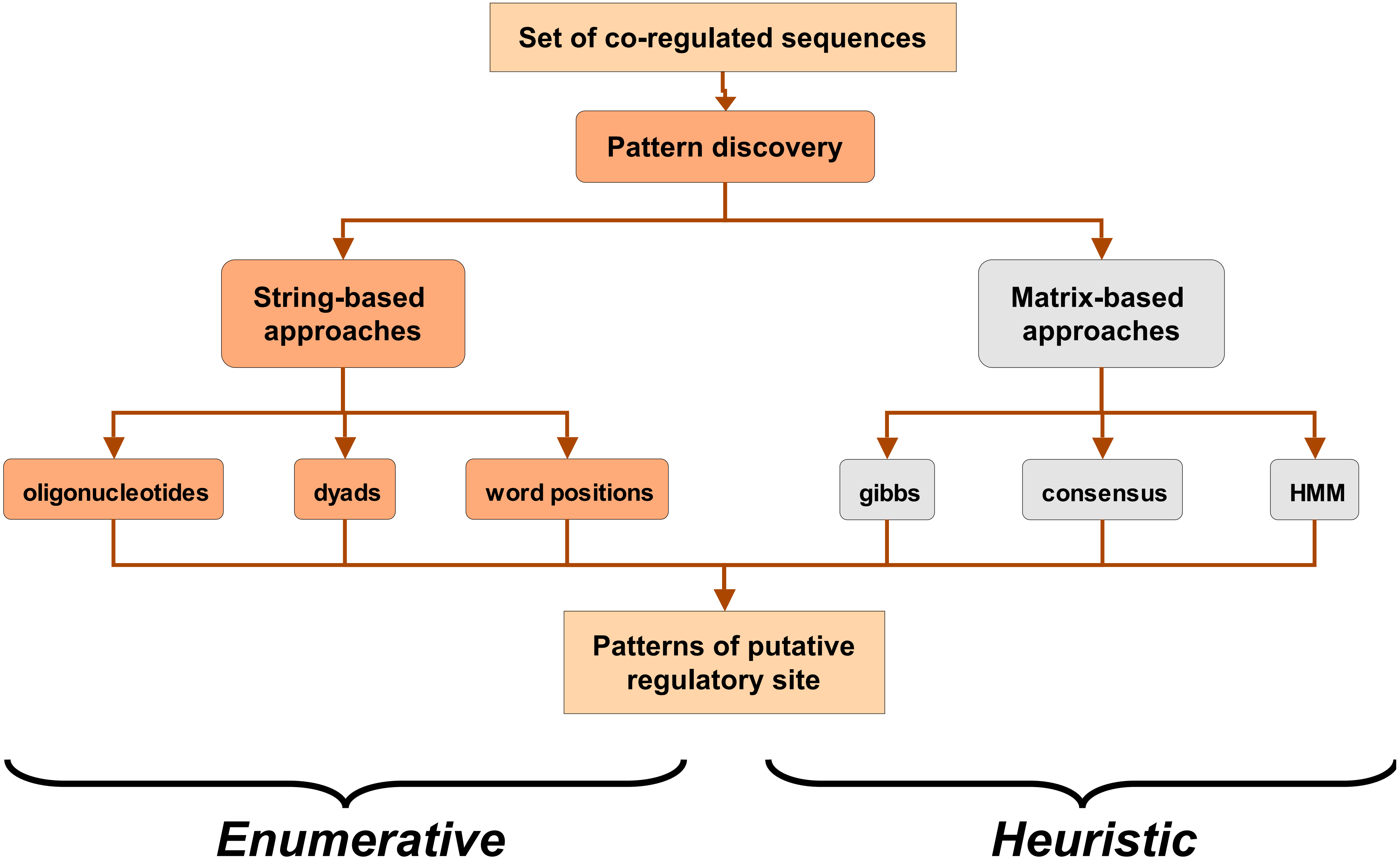
transcription factors

coding region

upstream region

🟩 🟧 predicted elements

Pattern discovery: approaches



Set of co-regulated sequences

Pattern discovery

String-based approaches

Matrix-based approaches

oligonucleotides

dyads

word positions

gibbs

consensus

HMM

Patterns of putative regulatory site

Enumerative

Heuristic

Pattern 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 pattern
  - Heuristic algorithms
    - Iterative optimization of the pattern
    - Gibbs sampler (gibbs, alignACE)
- Hidden Markov models (YEBIS, MEME)