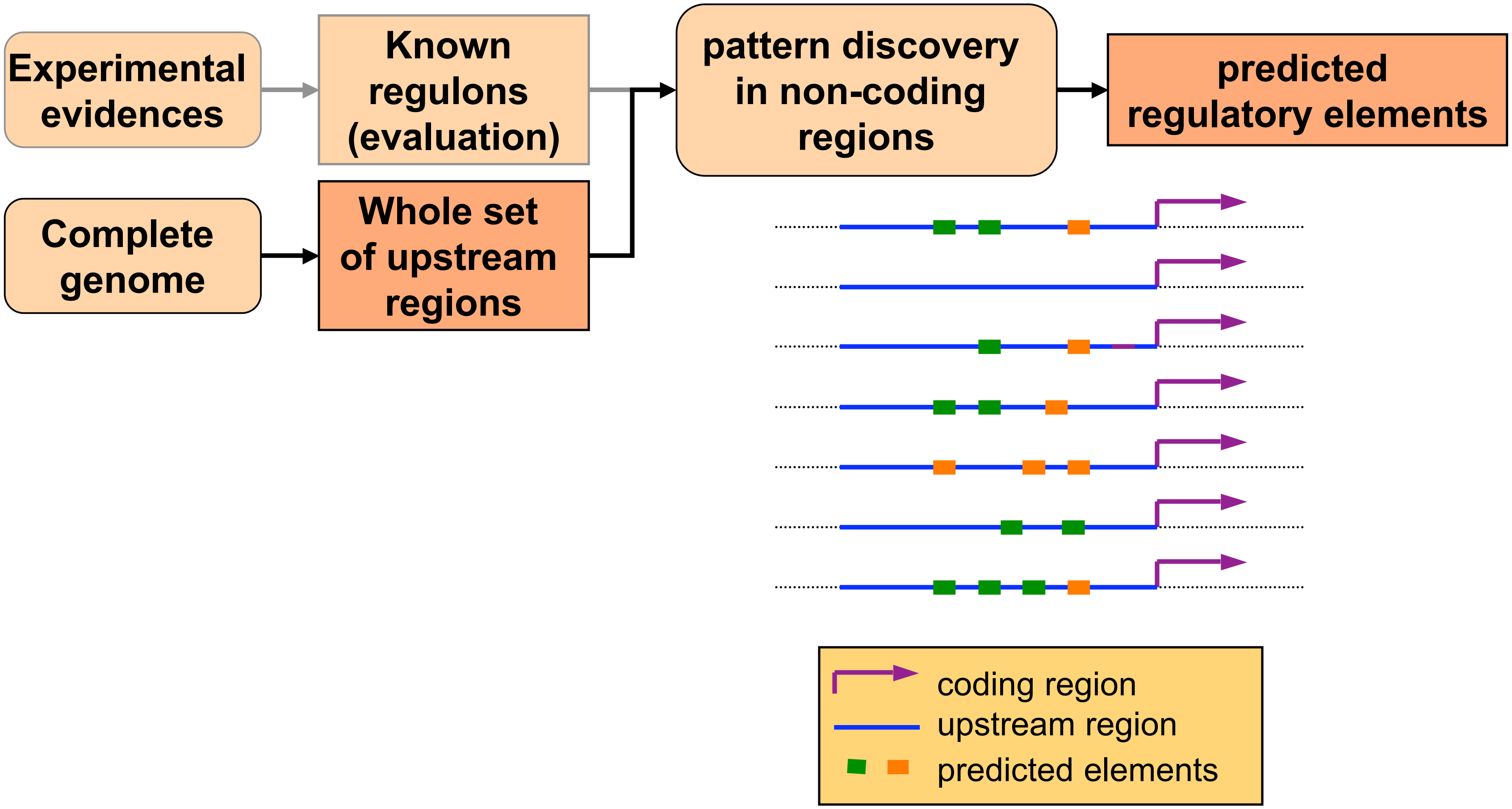
*Regulatory sequence analysis*

# *Genome-scale pattern discovery*

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

# *Genome-scale pattern discovery*



Experimental evidences

Known regulons (evaluation)

Complete genome

Whole set of upstream regions

pattern discovery in non-coding regions

predicted regulatory elements

coding region

upstream region

predicted elements

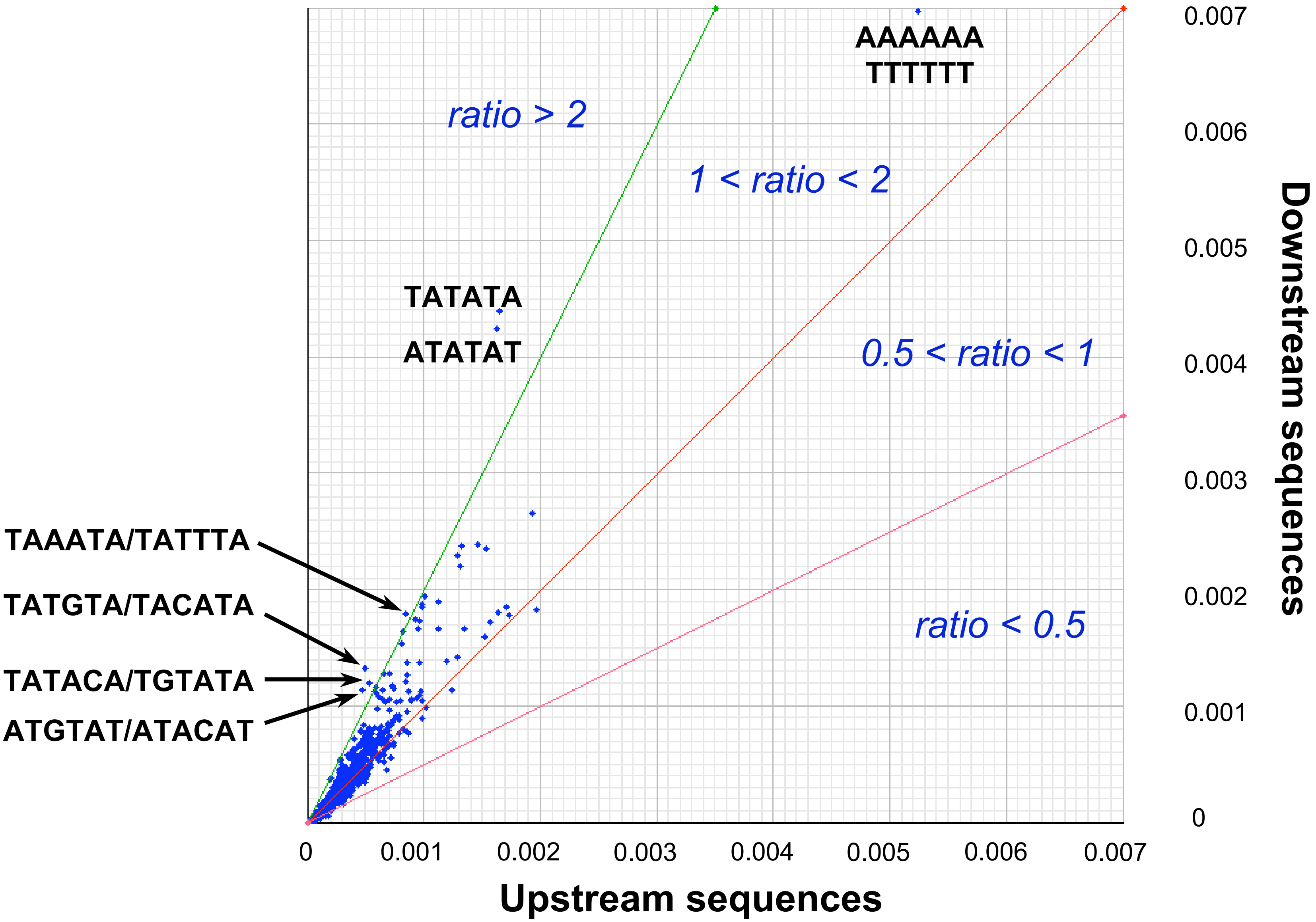
# *Genome-scale pattern discovery*

- Goal : extraction of functional signals involved in general mechanisms :
  - 5'-end signals (initiation of transcription)
  - 3'-end signals (termination of transcription, RNA cleavage and maturation)
- 3' end signal analysis
  - 6217 downstream sequences
  - 200 bp from the stop codon
- Problem: how to estimate expected word frequencies ?
  - The family now includes all yeast genes

# *Expected frequencies: external reference*

- Downstream sequences vs whole genome frequencies
  - problem of interpretation
    - may reflect merely differences between non-coding and coding sequences, which represent 73% of the genome
- Downstream versus upstream sequences
  - problem of interpretation:
    a word may be significant because
    - over-represented in downstream sequences
    - under-represented in upstream sequences

# *Downstream vs upstream sequences*



# *Expected frequencies: internal reference*

- Estimation of expected word frequencies
  - on basis of the downstream sequences themselves
- Markov chain models
  - The expected frequency of each k-letter word is estimated on basis of sub-word frequencies

e.g.: $$\text{exp}\{\text{GATAAG}\} = \frac{\text{obs}\{\text{GATAA}\} \text{ x } \text{obs}\{\text{ATAAG}\}}{\text{obs}\{\text{ATAA}\}}$$

# Oligo-analysis with Markov chain models

- Analysis of a set of 6217 downstream sequences, 200bp each
- Detection of over-represented words, and grouping by sequence similarity

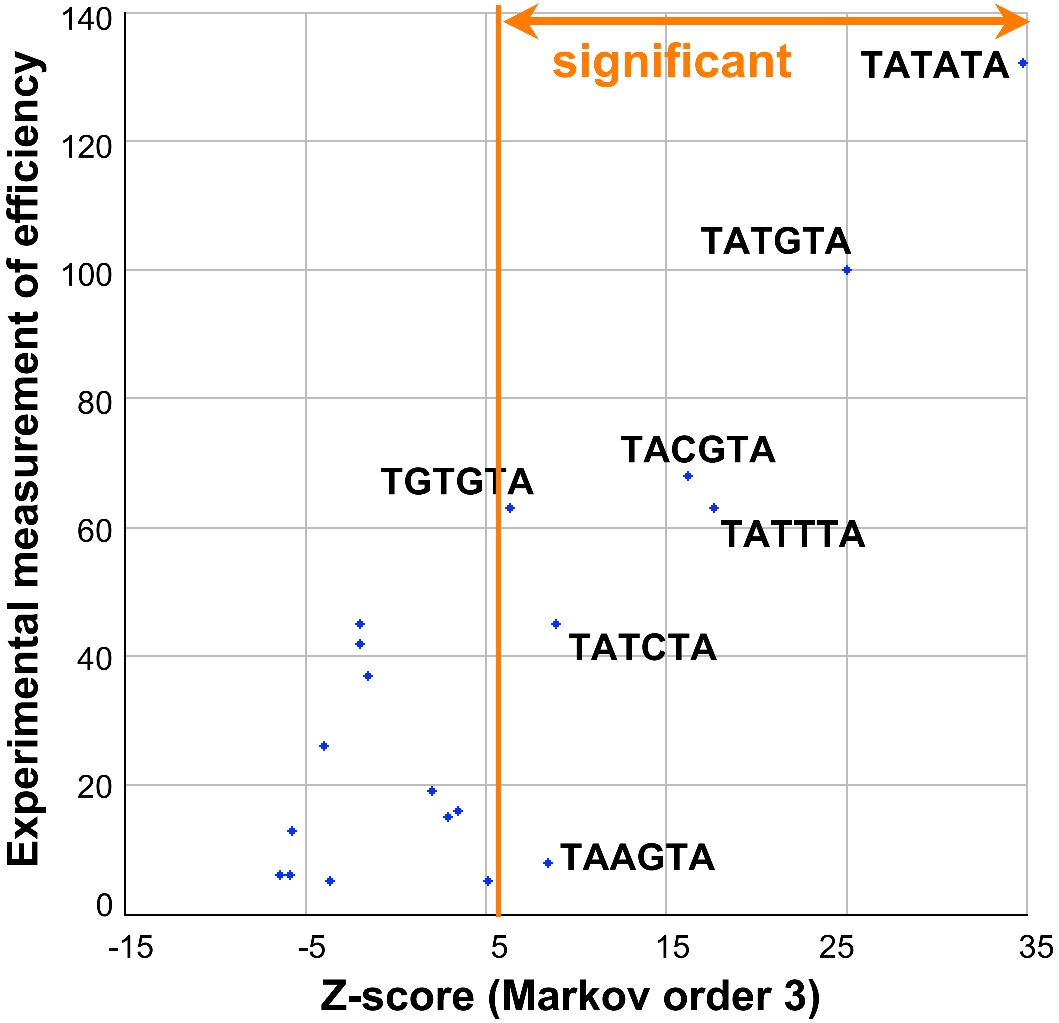
| | |
|---|---|
| ATATAT. | 27.0 |
| ATACAT. | 15.5 |
| ATGTAT. | 11.9 |
| ATAAAT. | 9.9 |
| ATAGAT. | 9.9 |
| ATTTAT. | 9.8 |
| GTATAT. | 8.2 |
| ATATGT. | 7.8 |
| ACATAT. | 7.7 |
| ATATAC. | 7.4 |
| .TATATA | 34.9 |
| .TACATA | 27.7 |
| .TATGTA | 25.0 |
| .TAAATA | 22.0 |
| .TATTTA | 17.7 |
| .TAGATA | 11.9 |
| .TGTATA | 8.6 |
| .TATACA | 7.3 |
| .CATATA | 3.5 |

| | |
|---|---|
| AAAAAA | 18.28 |
| AAATAA | 16.65 |
| AATAAA | 14.09 |
| AAGAAA | 9.27 |
| AACAAA | 9.02 |
| AAAGAA | 8.17 |
| AAACAA | 7.69 |

| | |
|---|---|
| TTTTTT | 16.87 |
| TTATTT | 16.74 |
| TTTATT | 13.25 |
| TTTCTT | 9.42 |
| TTTGTT | 8.72 |
| TTCTTT | 8.46 |

| | |
|---|---|
| ACATAC. | 12.21 |
| ACACAC. | 11.15 |
| .CACACA | 13.00 |
| .CATACA | 8.81 |

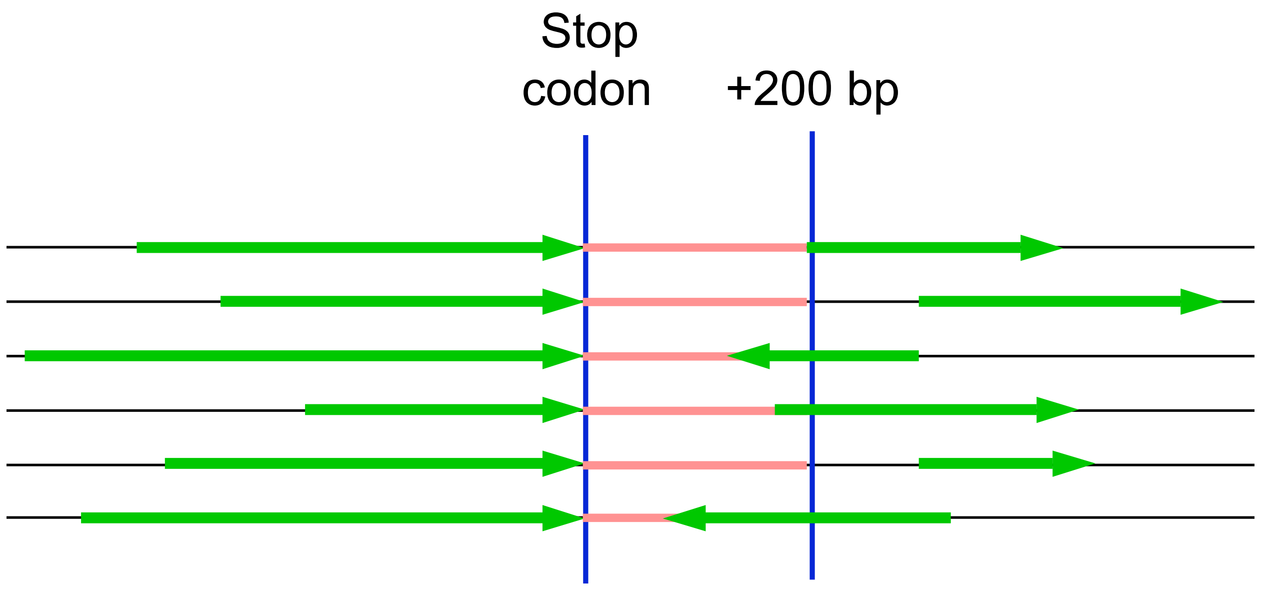
# *Comparison with experimental values*



- Irniger and Braus (1994) performed a saturation mutagenesis and measured the the efficiency of all single-base mutants of TATGTA.
- High Z-score values from Markov 4 model correlate pretty well with experimental efficiency

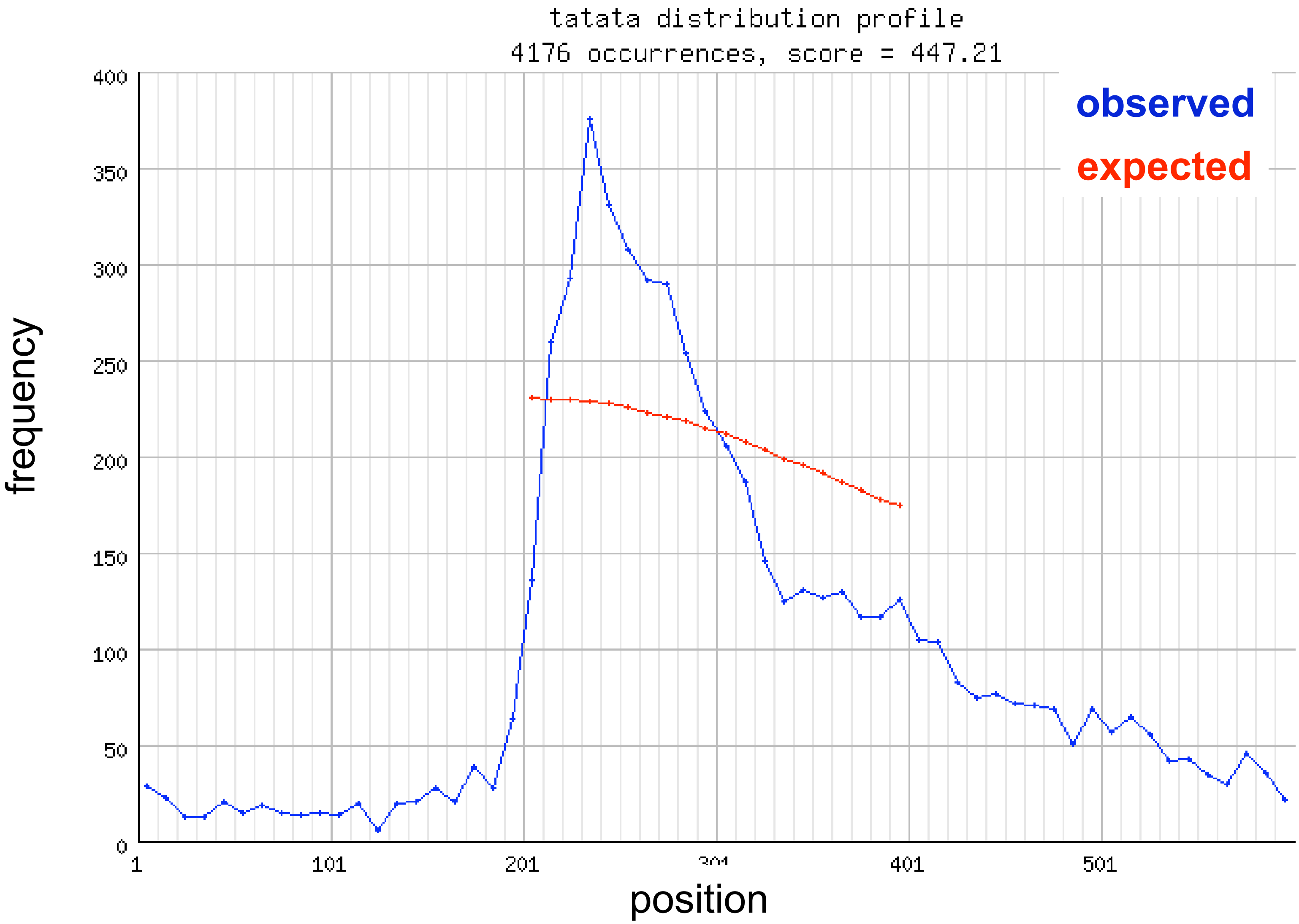
# *Position analysis*

- Measure the positional distribution of each word
- Perform a test of homogeneity and select all words with a significant bias
- Significance of the non-homogeneity is estimated with a $\chi^2$ test
- Note : in our case, homogeneous is not flat, because sequences are clipped when there is a downstream ORF closer than 200 bp



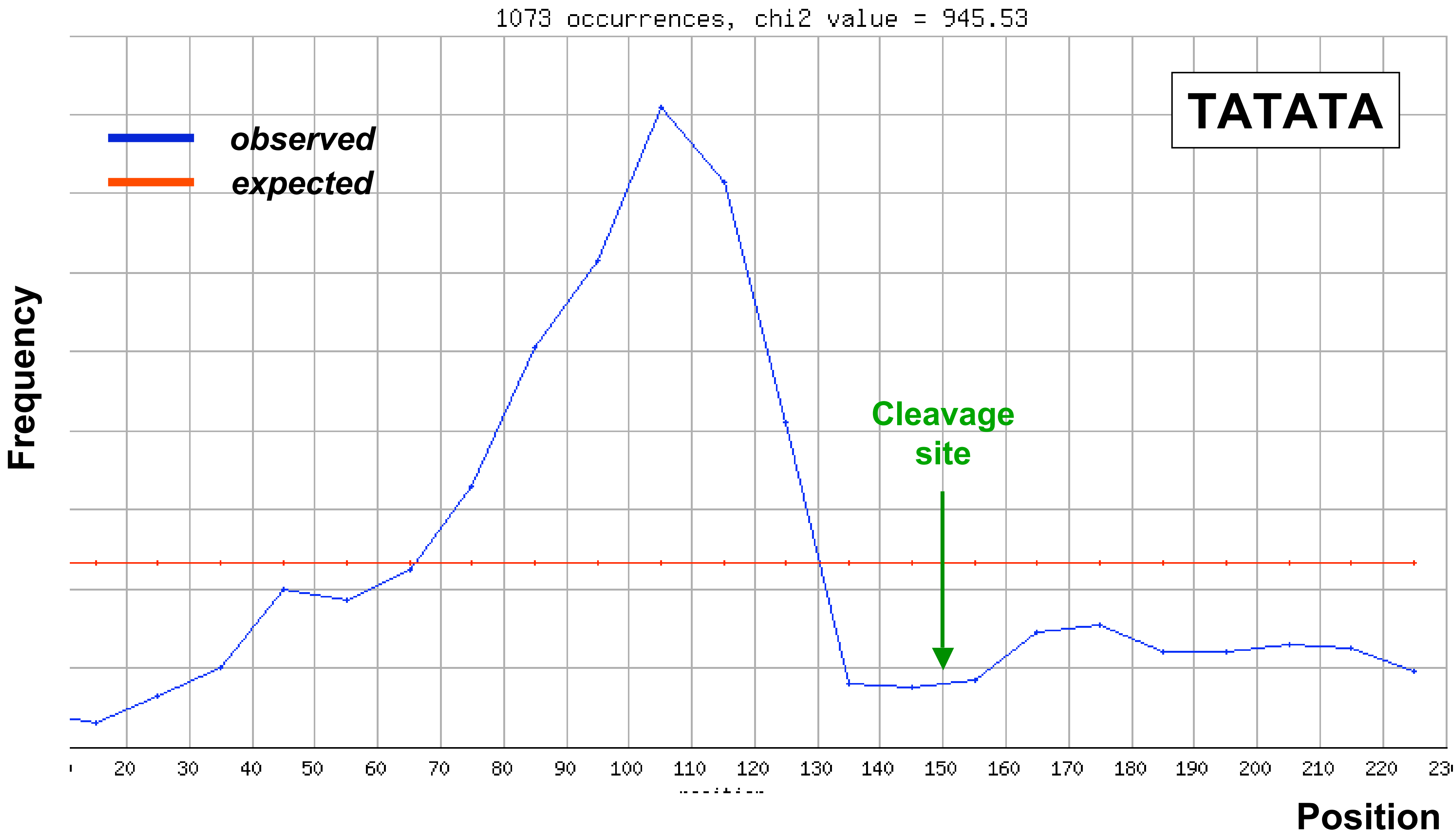
# *Word position distribution*

- Positions relative to the stop codon

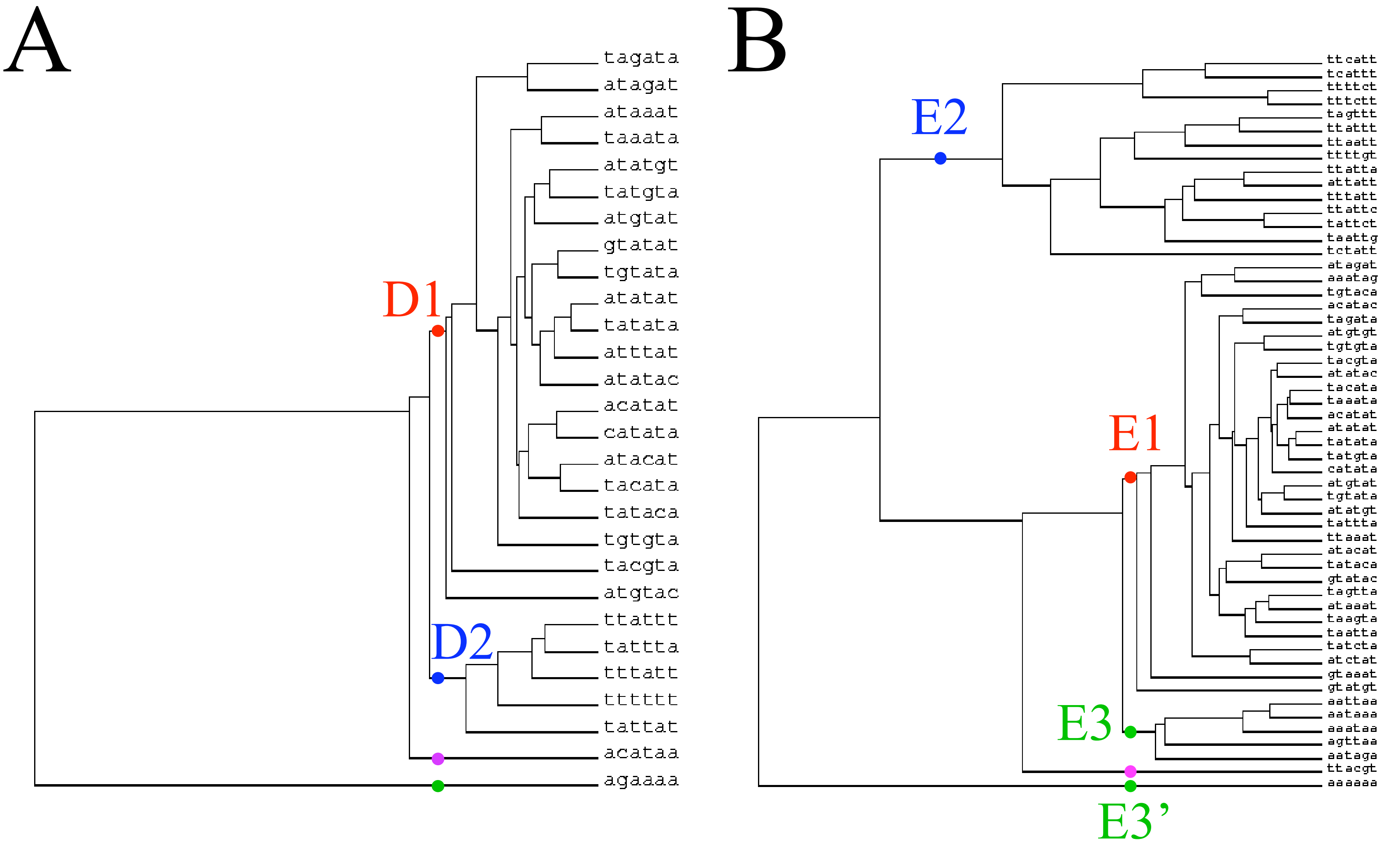


# *Position analysis : profiles of word distribution*

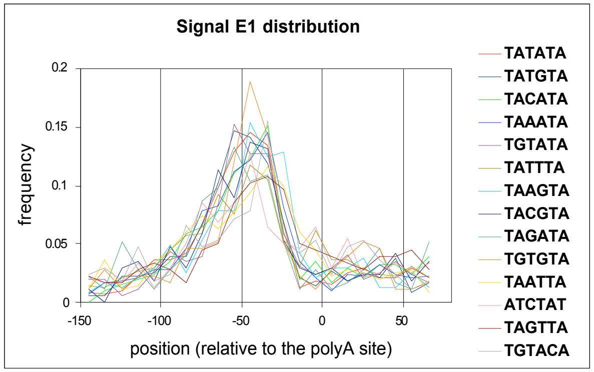
- Positions relative to the cleavage site

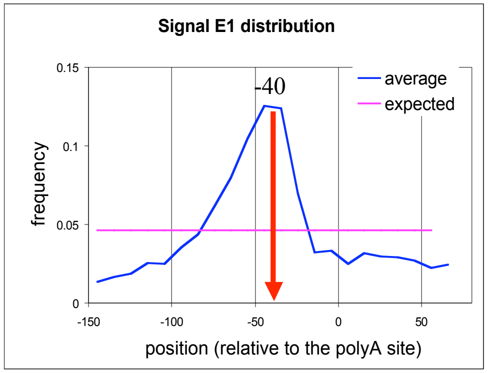


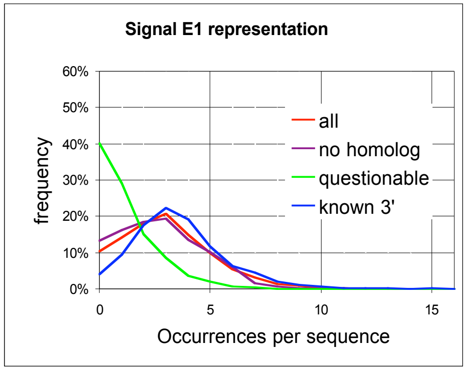
# *Word clustering according to position profiles*

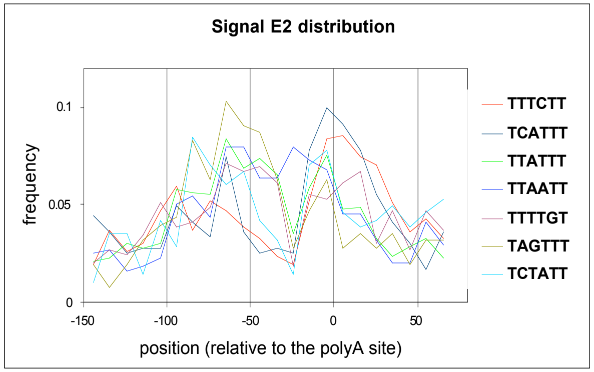


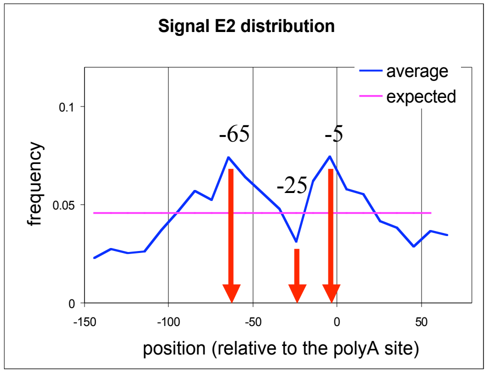
# Signal distribution and representation

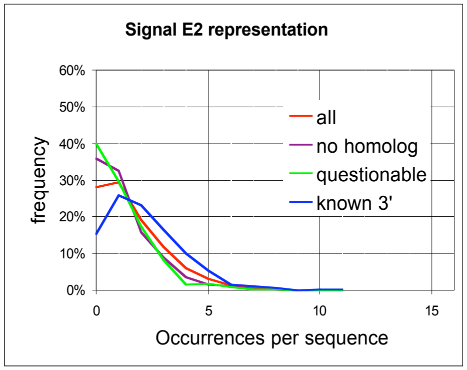


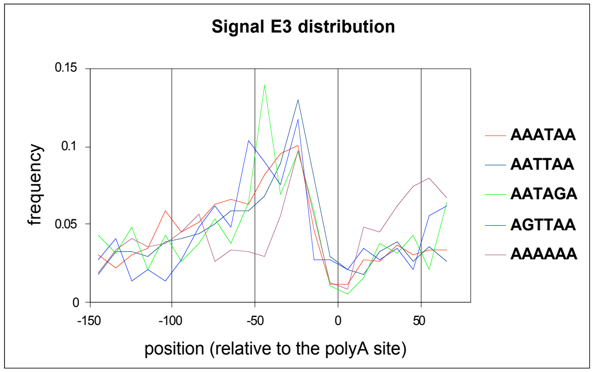


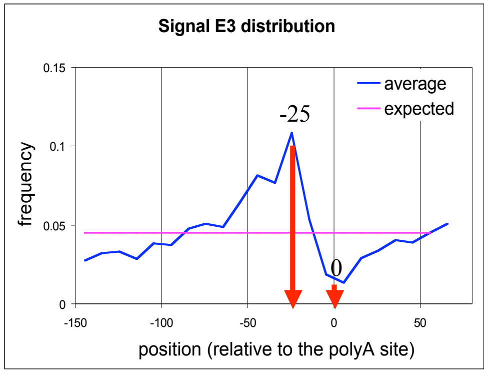


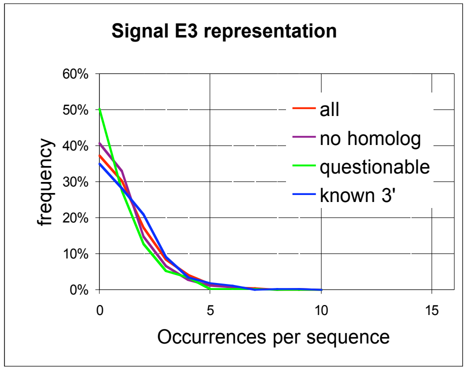












# *Genome-scale pattern discovery - references*

- van Helden, J., del Olmo, M. & Perez-Ortin, J.E. Statistical analysis of yeast genomic downstream sequences reveals putative polyadenylation signals. *Nucleic Acids Res* **28**, 1000-10 (2000).
- Bussemaker, H.J., Li, H. & Siggia, E.D. Regulatory element detection using a probabilistic segmentation model. *Ismb* **8**, 67-74 (2000).
- Bussemaker, H.J., Li, H. & Siggia, E.D. From the cover: building a dictionary for genomes: identification of presumptive regulatory sites by statistical analysis. *Proc Natl Acad Sci U S A* **97**, 10096-100 (2000).
- Brazma, A., Vilo, J., Ukkonen, E. & Valtonen, K. Data mining for regulatory elements in yeast genome. *Ismb* **5**, 65-74 (1997).
- Brazma, A., Jonassen, I., Vilo, J. & Ukkonen, E. Predicting gene regulatory elements in silico on a genomic scale. *Genome Res* **8**, 1202-15 (1998).