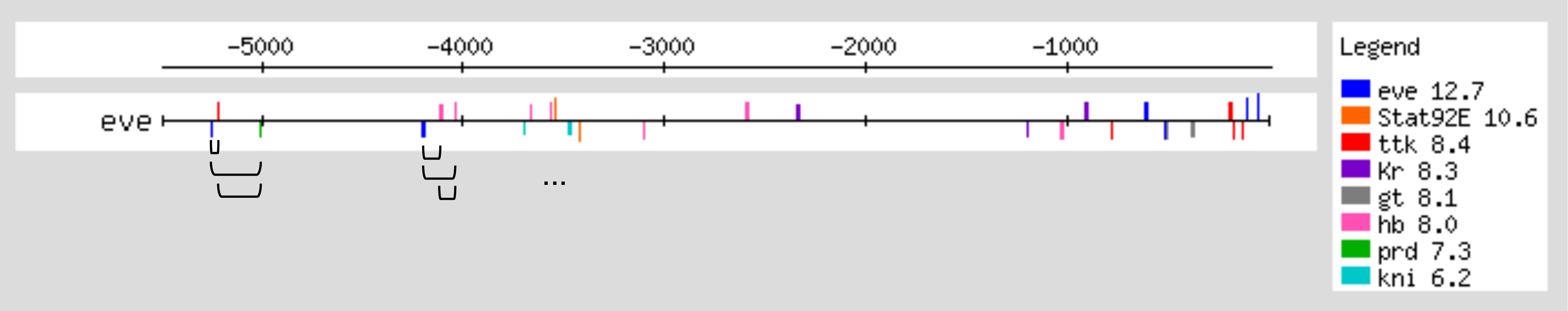
Regulatory Sequence Analysis

Detecting Cis-Regulatory Modules (CRMs)

Jacques.van.Helden@ulb.ac.be
Laboratoire de Bioinformatique des Génomes et des Réseaux (BiGRe)
Université Libre de Bruxelles, Belgique

Principle of CRM prediction

- Detection of regions containing a higher density of predicted TFBS than expected by chance.
- Various programs have been implemented to predict CRMs.
- The RSAT program *matrix-scan* supports CRM prediction, by detecting regions enriched in cis-regulatory elements.
- Main features
  - Detection of homotypic (single motif) or heterotypic (distinct motifs) models.
  - All possible windows are tested within a user-specified width range
    (e.g. From 30 to 300).
  - The user has to specify a threshold on the P-value used for individual site predictions.
  - The program predicts all sites passing this threshold for each the input matrices, and detects regions (windows) having significantly more hits than expected by chance.
  - The enrichment is estimated by using the binomial statistics.

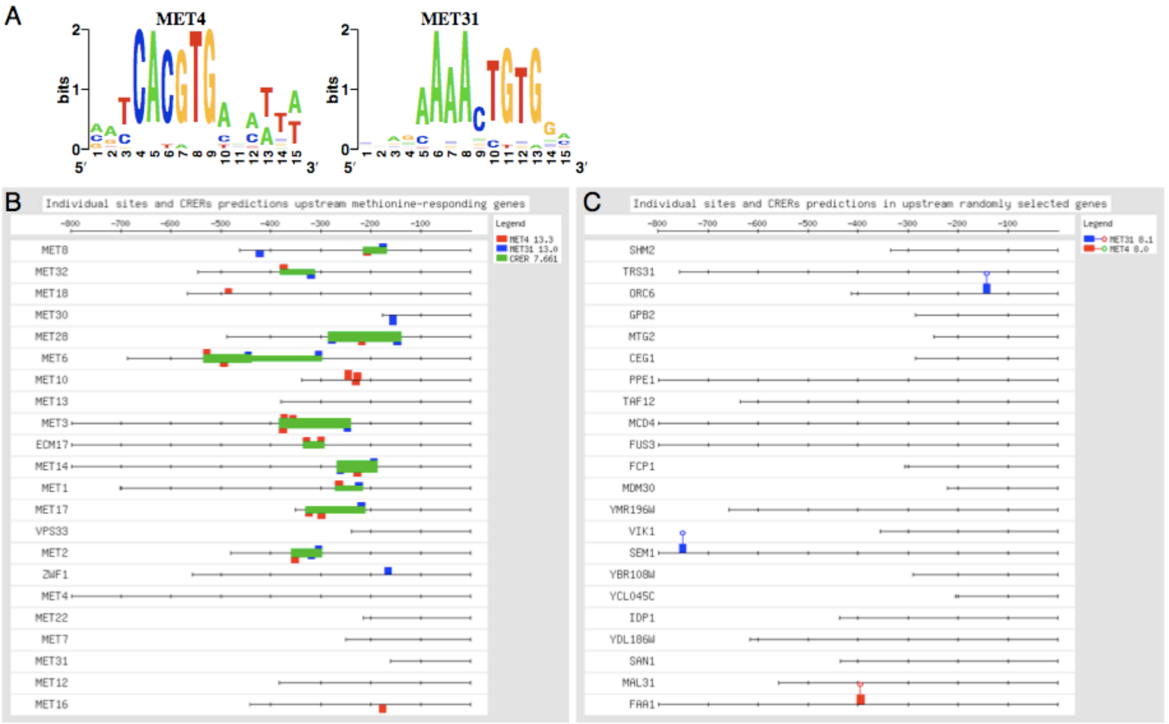


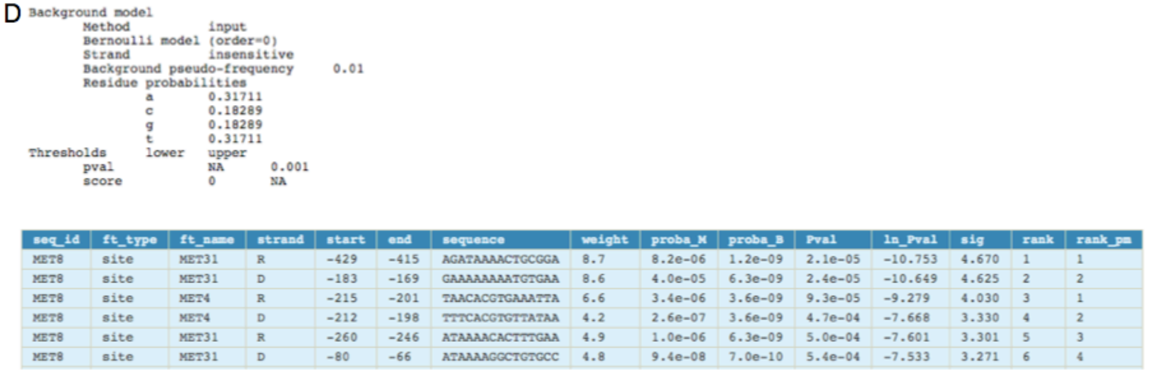
- Turatsinze, J. V., Thomas-Chollier, M., Defrance, M. & van Helden, J. (2008). Using RSAT to scan genome sequences for transcription factor binding sites and cis-regulatory modules. Nat Protoc 3, 1578-88.

Cis-regulatory element enriched regions (CRERs)

as putative cis-regulatory modules (CRMs)

- Example of CRER detection.
- Detection of methionine-responding genes in the yeast *Saccharomyces cerevisiae*.
- **A:** matrices
    - MET4: motif bound by the complex Met4p/Cbf1P/Met28p.
    - MET31: motif bound by either Met31p or Met32p (two homologous transcription factors).
- **B:** predicted sites and CRERs in upstream non-coding sequences of MET genes.
- **C:** predicted sites and CRERs in random selections of yeast genes.
- **D:** examples of sites reported by matrix-scan.



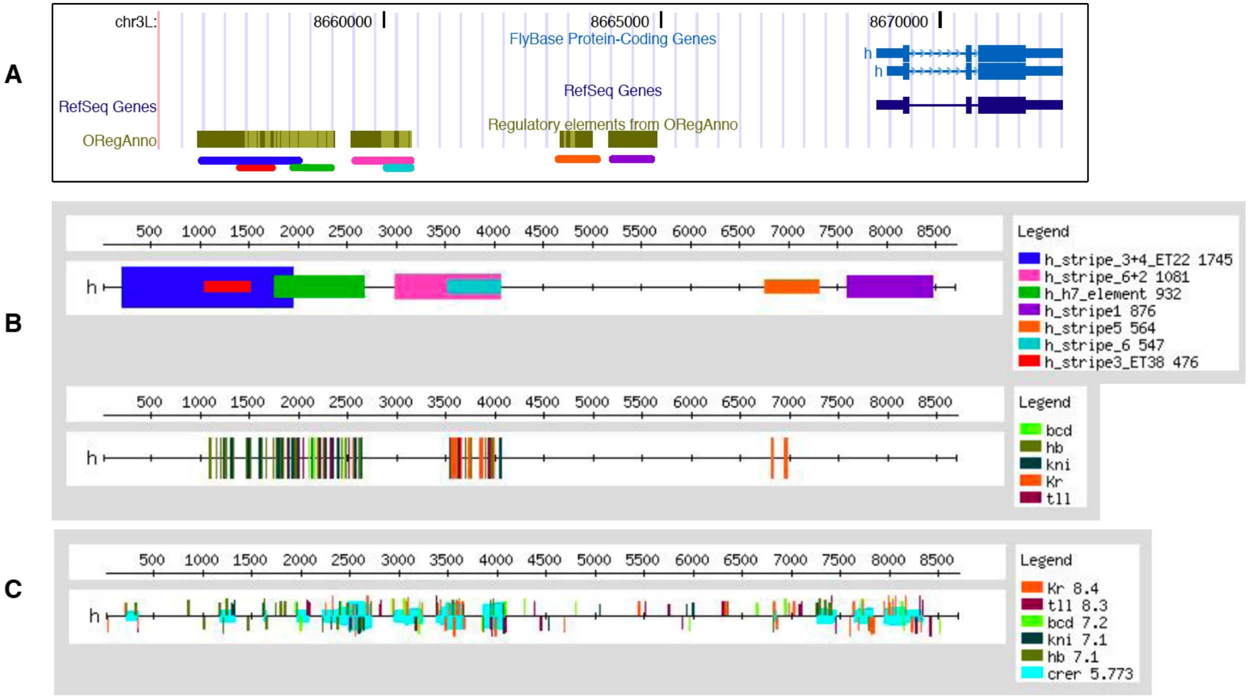


```
Background model
        Method                input
        Bernoulli model (order=0)
        Strand                insensitive
        Background pseudo-frequency     0.01
        Residue probabilities
                a             0.31711
                c             0.18289
                g             0.18289
                t             0.31711
Thresholds      lower         upper
        pval                  NA        0.001
        score                 0         NA
```

seq_id	ft_type	ft_name	strand	start	end	sequence	weight	proba_M	proba_B	Pval	ln_Pval	sig	rank	rank_pm
MET8	site	MET31	R	-429	-415	AGATAAAACTGCGGA	8.7	8.2e-06	1.2e-09	2.1e-05	-10.753	4.670	1	1
MET8	site	MET31	D	-183	-169	GAAAAAAAATGTGAA	8.6	4.0e-05	6.3e-09	2.4e-05	-10.649	4.625	2	2
MET8	site	MET4	R	-215	-201	TAACACGTGAAATTA	6.6	3.4e-06	3.6e-09	9.3e-05	-9.279	4.030	3	1
MET8	site	MET4	D	-212	-198	TTTCACGTGTTATAA	4.2	2.6e-07	3.6e-09	4.7e-04	-7.668	3.330	4	2
MET8	site	MET31	R	-260	-246	ATAAAACACTTTGAA	4.9	1.0e-06	6.3e-09	5.0e-04	-7.601	3.301	5	3
MET8	site	MET31	D	-80	-66	ATAAAAGGCTGTGCC	4.8	9.4e-08	7.0e-10	5.4e-04	-7.533	3.271	6	4

Thomas-Chollier, M., Sand, O., Turatsinze, J. V., Janky, R., Defrance, M., Vervisch, E., Brohee, S. & van Helden, J. (2008). RSAT: regulatory sequence analysis tools. Nucleic Acids Res 36, W119-27.

Using matrix-scan to detect Transcription Factor Binding Sites (TFBS) and Cis-regulatory Element Enriched Regions (CRERs)



- Example of CRER detection: upstream region of the Drosophila gene *hairy*

A

- Annotations in the ORegAnno database (displayed with the UCSC genome browser)

B

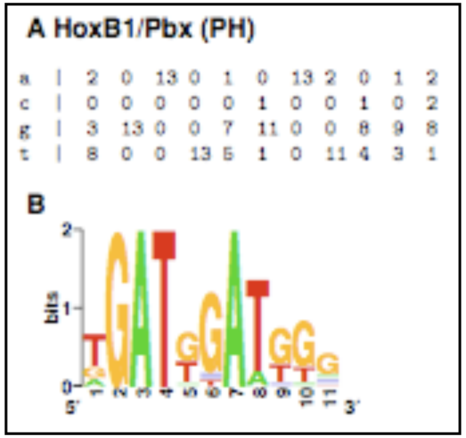
- Cis-regulatory moduels annotated in FlyReg (top)
- individual sites annotated in ORegAnno (bottom).

C

- Scanning with 5 PSSM to detect individual binding sites + CRERs

Multi-genome CRER detection

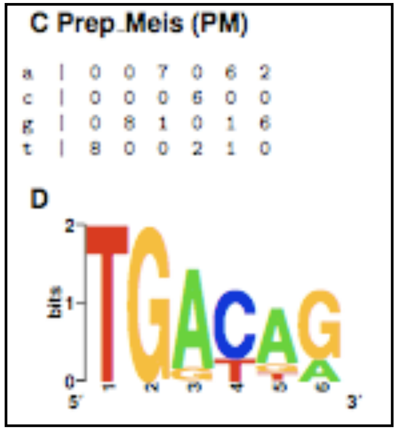
- Morgane Thomas-Chollier (PhD thesis VUB/ULB, May 2008)
- Detection of CRERs with two matrices (HoxB1/Pbx and Prep/Meis) in the introns of the gene HoxA2.



A HoxB1/Pbx (PH)

```
a   |   2   0   13  0   1   0   13  2   0   1   2
c   |   0   0   0   0   0   1   0   0   1   0   2
g   |   3   13  0   0   7   11  0   0   8   9   8
t   |   8   0   0   13  5   1   0   11  4   3   1
```

B



C Prep_Meis (PM)

```
a   |   0   0   7   0   6   2
c   |   0   0   0   5   0   0
g   |   0   8   1   0   1   6
t   |   8   0   0   2   1   0
```

D

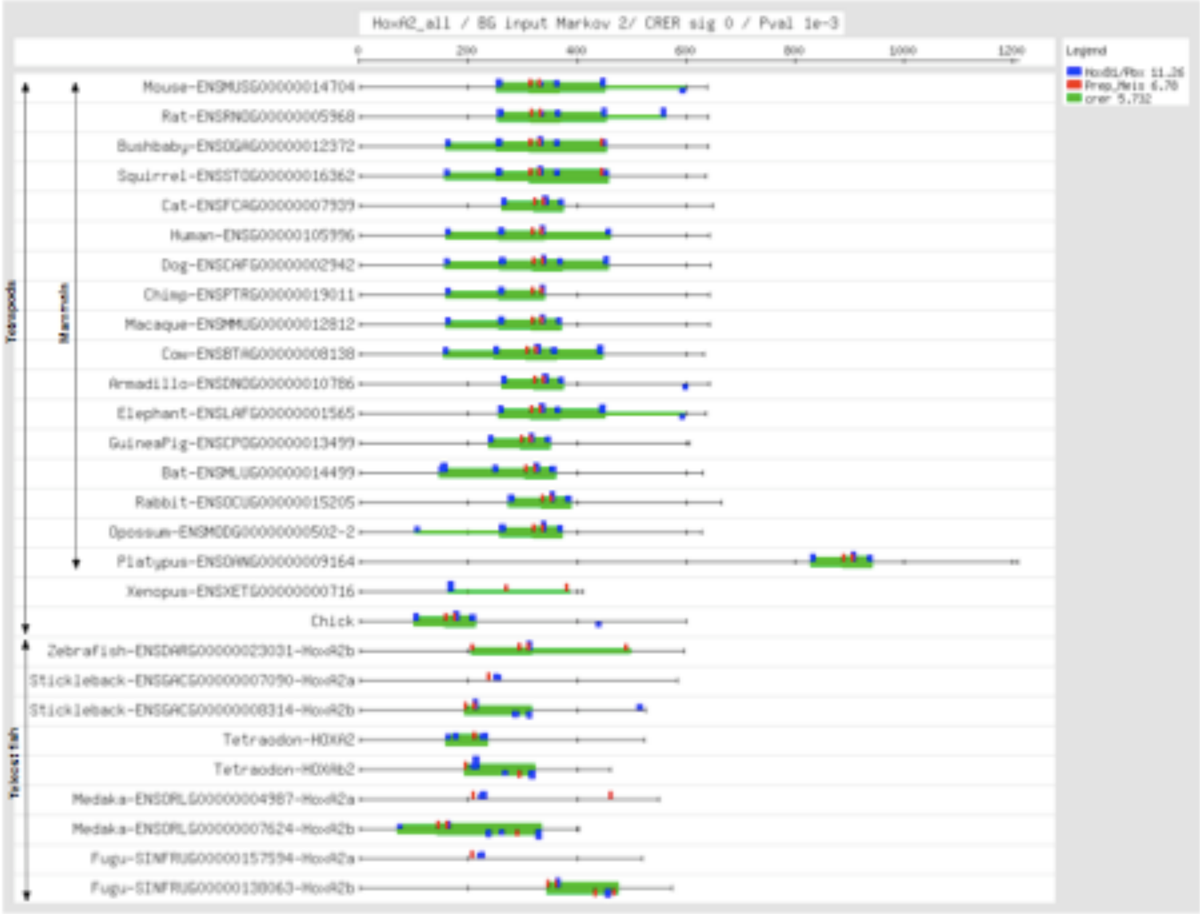


Figure 6.3: Cross-species predictions with matrix-scan in the HoxA2 intron. Predictions of TFBSs and CRERs in the HoxA2 intron of various vertebrate species. The height of each site is proportional to its weight score. CRER heights are proportional to their significance score. The numbers in the legend correspond to the highest weights for PH and PM matrices, and to the highest significance for the CRERs.