Evaluation of predicted regulatory elements

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

Regulatory Sequence Analysis

The impossible choice of the "right" testing set

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

Typical evaluation sets

- Different sets of sequences can be used to assess the accuracy of predictions
- Positive control: quantify the capability of the program to detect known regulatory elements
  - Annotated sites (e.g. sites from TRANSFAC) in their original context (the promoter sequences).
  - Annotated sites implanted in other context
    - Biological sequences (random selection).
    - Artificial sequences.
  - Artificial sites implanted in artificial sequences.
- Negative control: quantify the capability of the program to return a negative answer when there are no regulatory elements.
  - Artificial sequences
    (generated according to a Bernoulli or a Markov model)
  - Biological sequences without common regulation
    (random selection of genes)

Evaluation sets

Usage	Context	Sites	Pros	Cons
Positive control: evaluation of sensitivity, specificity, accuracy.	Artificial sequences (e.g. generated with a Markov model)	Artificial sites (e.g. generated from a PSSM)	Control on all the parameters (number of sites, motif variations, sequence composition, sequence length). Useful to check theoretical models.	Performances might differ between artificial sets and real conditions
	Artificial sequences	Implanted biological sites	All the "positive" sites (implanted) are known.	Performances mainly reflect the fit between random model of the predictor and of the sequence generator.
	Biological sequences	Biological sites in their context	All the true sites are available for the predictor, even if they are not annotated yet.	Answer can be obtained from databases. Programs can be over-fitted because parameters were estimated with the same DB. Some real sites can be absent from the annotation -> FFP.
	Biological sequences	Implanted biological sites	All the "positive" sites (implanted) are supposedly known.	The number of implanted sites might differ from natural conditions. Annotation-based: under-estimation (many sites are not annotated).
Negative control: estimation of the rates of false positives.	Artificial sequences	None	Control on the sequence composition (background model).	Performances mainly reflect the fit between random models of predictor and of sequence generator, resp..
	Random selection of biological sequences	None	Indicates the rate of false positive in real conditions.	

This table is far from complete, you can add pros and cons as an exercise.

We could say that the best set depends on the question to be addressed.

Example: biological sites implanted in foreign biological sequences

- Down et al. (2005). Nucleic Acids Res. 33(5):1445-1453. NestedMICA: sensitive inference of over-represented motifs in nucleic acid sequences.
- Motifs
  - Jaspar annotations for 4 human transcription factors (HLF,c-Fos, CREB, HFH-1)
- Sequences
  - Random selections of genes, 100 promoters per set.
  - For each factor, different sequences sizes are tested.
- Implanted sites
  - Zero or one occurrence per sequence (zoops).
  - One implant in 50% of the sequences.
- Pattern discovery software
  - NestedMICA (the new program presented in the article)
  - MEME (used with default parameters)

Table 1. Discovery of the HLF motif from sets of 100 synthetic sequences of various lengths

Length	100	150	200	300	400	500	600	700
MEME	y	y	n	n	n	n	n	n
N'MICA	y	y	y	y	y	y	y	n

'y' indicates that the correct motif was found, and 'n' indicates failure.

Table 2. Discovery of the c-FOS motif from sets of 100 synthetic sequences of various lengths

Length	200	300	400	500	600
MEME	y	y	n	n	n
N'MICA	y	y	y	y	n

'y' indicates that the correct motif was found, and 'n' indicates failure.

Table 3. Discovery of the HFH-1 motif from sets of 100 synthetic sequences of various lengths

Length	800	1000	1200	1400	1600
MEME	y	y	y	n	n
N'MICA	y	y	y	n	n

'y' indicates that the correct motif was found, and 'n' indicates failure.

Example: biological sites implanted in foreign biological sequences

- Down et al. (2005). Nucleic Acids Res. 33(5):1445-1453. NestedMICA: sensitive inference of over-represented motifs in nucleic acid sequences.

- Question: criterion to say "yes" or "no" ?
  - Visual inspection ?
  - Quantitative criterion ?

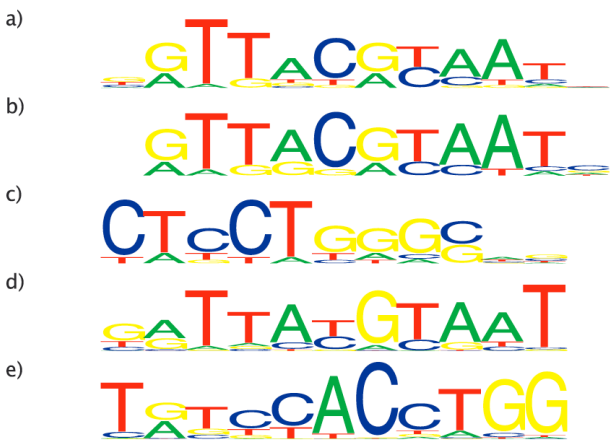


Figure 4. (**a**) The original HLF motif from JASPAR. (**b**) Results for searching for HLF in a set of 150 base sequences using MEME. (**c**) MEME with 200 base sequences. (**d**) NestedMICA with 600 base sequences. (**e**) NestedMICA with 700 base sequences.

Table 1. Discovery of the HLF motif from sets of 100 synthetic sequences of various lengths

Length	100	150	200	300	400	500	600	700
MEME	y	y	n	n	n	n	n	n
N'MICA	y	y	y	y	y	y	y	n

'y' indicates that the correct motif was found, and 'n' indicates failure.

Table 2. Discovery of the c-FOS motif from sets of 100 synthetic sequences of various lengths

Length	200	300	400	500	600
MEME	y	y	n	n	n
N'MICA	y	y	y	y	n

'y' indicates that the correct motif was found, and 'n' indicates failure.

Table 3. Discovery of the HFH-1 motif from sets of 100 synthetic sequences of various lengths

Length	800	1000	1200	1400	1600
MEME	y	y	y	n	n
N'MICA	y	y	y	n	n

'y' indicates that the correct motif was found, and 'n' indicates failure.

Evaluation of pattern matching results

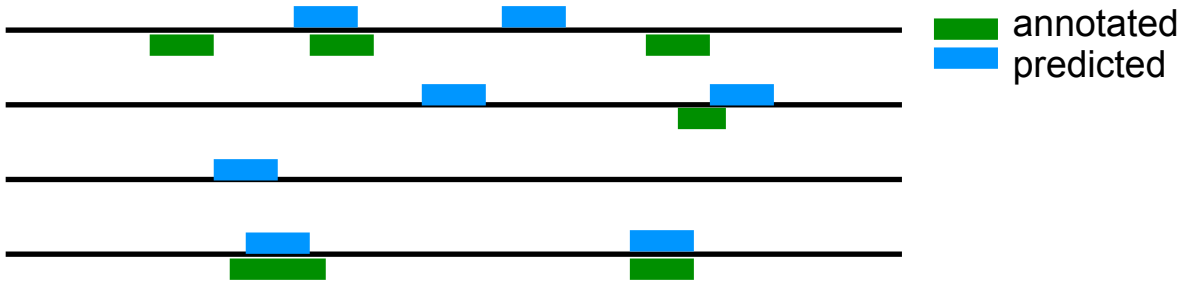
Jacques.van.Helden@ulb.ac.be
Université Libre de Bruxelles, Belgique
Laboratoire de Bioinformatique des Génomes et des Réseaux (BiGRe)
http://www.bigre.ulb.ac.be/

Annotated sites

- The evaluation of pattern matching relies on a collection of annotated sites (locations on the sequences) considered as the ***true*** answer.
  - Each site is defined by its starting and ending position (the strand is not considered here).
- The rest of the sequence is considered as a ***false*** answer.
  - In typical conditions, a large fraction of the positions are annotated as false.

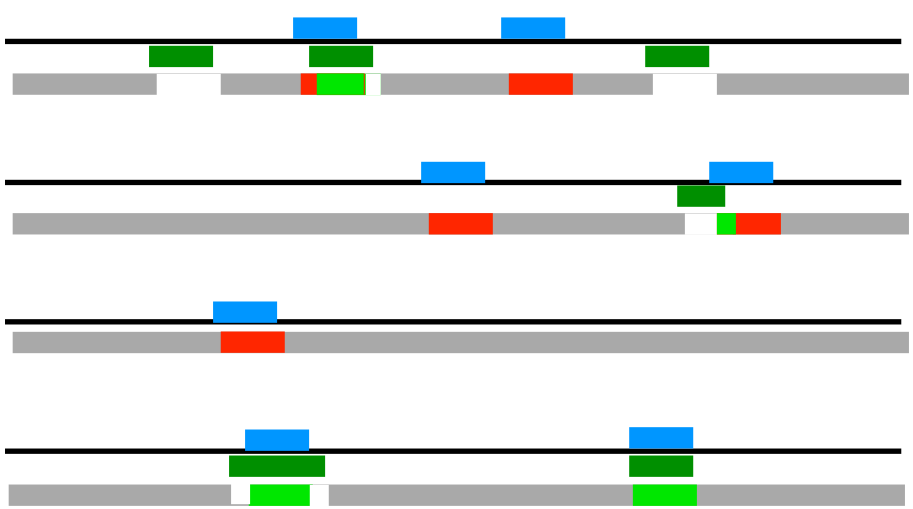
Comparison between annotated and predicted sites

- The annotated and predicted sites are compared.



Comparison at the nucleotide level

- Annotated and predicted sites can be compared **at the nucleotide level**.
- Each predicted nucleotide is considered as a match if it falls within an annotated site.

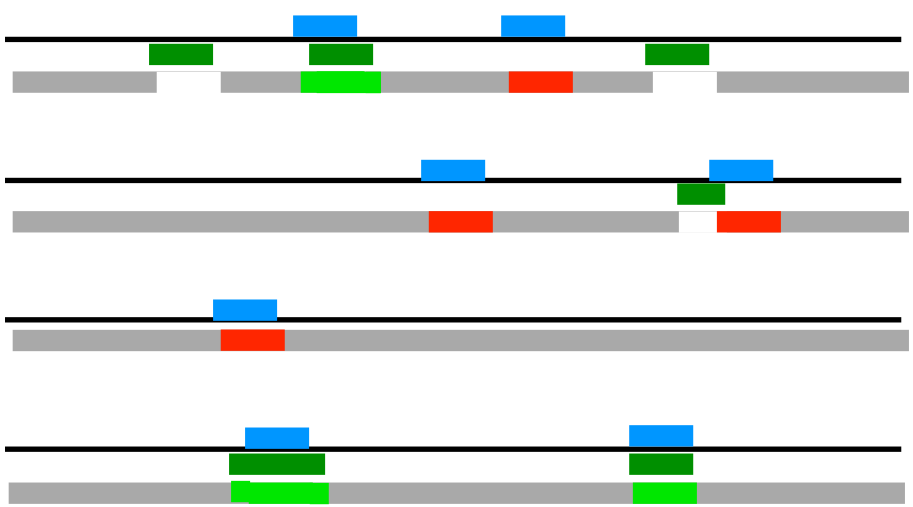


	Predicted: True	Predicted: False
Annotated: False	TP	FN
Annotated: True	FP	TN

TP	**True Positive**	**Annotated and predicted.**
FP	**False Positive**	**Predicted but not annotated**
TN	True Negative	Neither annotated nor predicted
FN	**False Negative**	**Annotated but not predicted**

Comparison at the site level

- Annotated and predicted sites can be compared **at the site level**.
- Each predicted site is considered as a match (as a whole) if it overlaps with an annotated site.
- A threshold can be imposed on the minimal number of overlapping nucleotides in order to consider that a predicted site does or not match an annotated site.



Annotated \ Predicted	True	False
False	TP	FN
True	FP	TN

TP	True Positive	Annotated and predicted.
FP	False Positive	Predicted but not annotated
TN	True Negative	Neither annotated nor predicted
FN	False Negative	Annotated but not predicted

Comparing feature sets

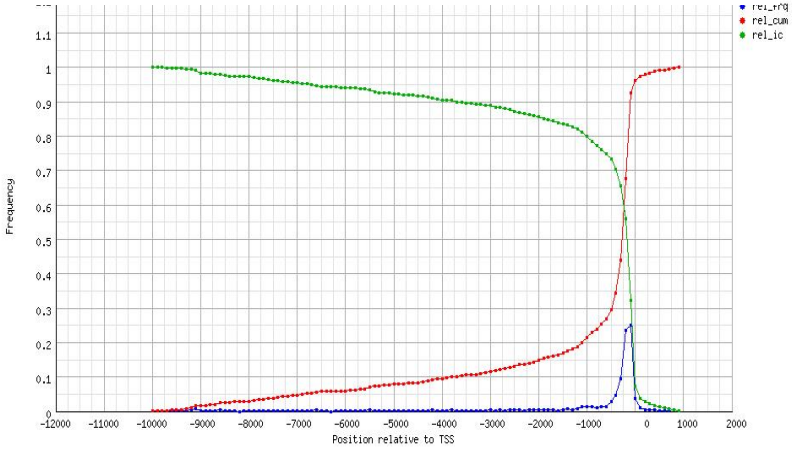
- The RSAT program ***compare-features**** allows to compare two or more sets of features.
  - Computes the intersection, the union and the differences between feature sets.
  - Returns comparison statistics.
- It can be used to compare annotated and predicted transcription factor binding sites.

* Program developed by Jean Valéry Turatsinze & Jacques van Helden

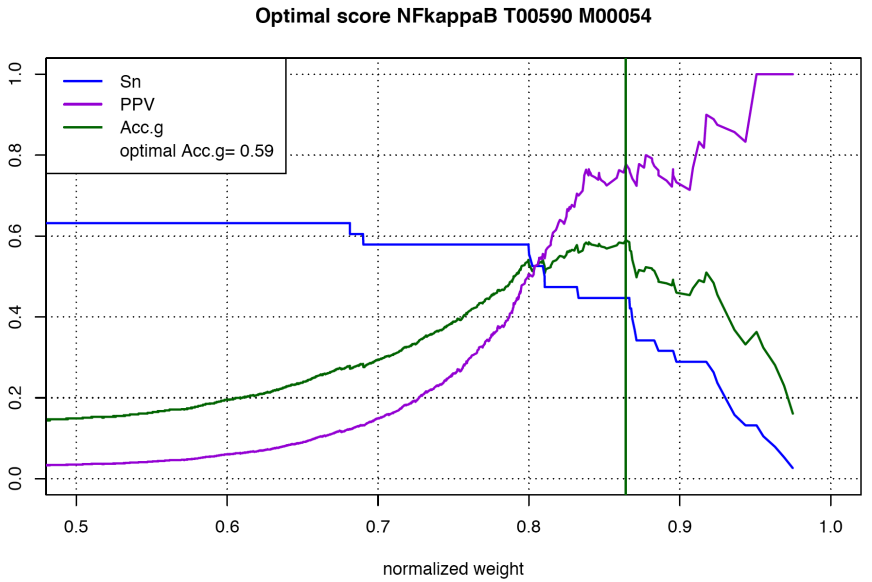
Evaluation of pattern matching results

- Evaluation, at the site level, for pattern matching results in human promoters with an NFkB matrix.
- Statistics
  - Sensitivity
    - *Sn=TP/(TP+FN)=(true predictions)/(annotated sites)*
  - Positive Predictive Value
    - *PPV=TP/(TP+FP)=(true predictions)/(total predictions)*
  - Accuracy
    - *Acc.a = (Sn + PPV)/2*
    - *Acc.g = sqrt(Sn * PPV)*
- Notes
  - The predictions were restricted to 500bp, because this is the best annotated interval in the reference database (TRANSPRO).
  - This is an illustration only, for one of the best examples.
  - NFkB is one of the best annotated factors in TRANSPRO.
  - It is not representative of overall performances.
  - Predictions give better results for NFkB than for other fact (*in preparation*).

Positions of human sites annotated in TRANSPRO (relative to TSS).

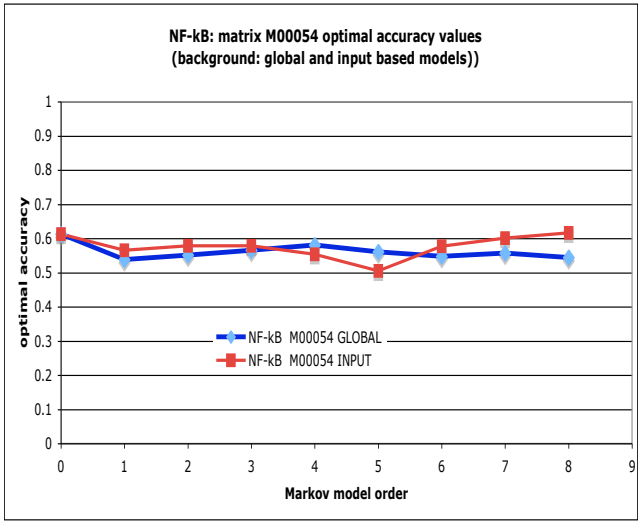


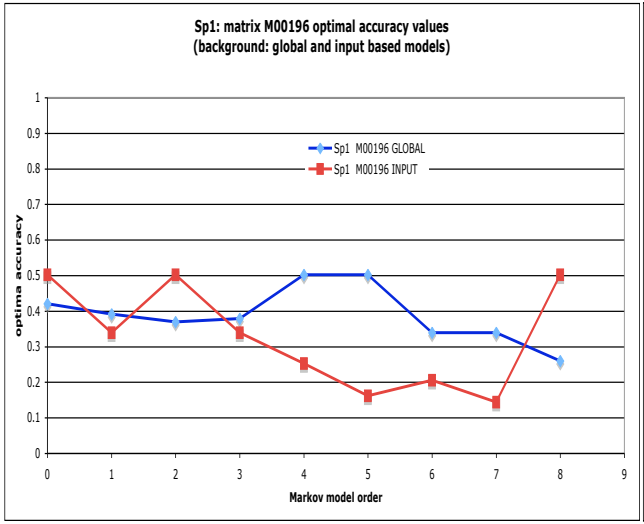
Predictions: NFkB, 500bp upstream of TSS, mkv=0

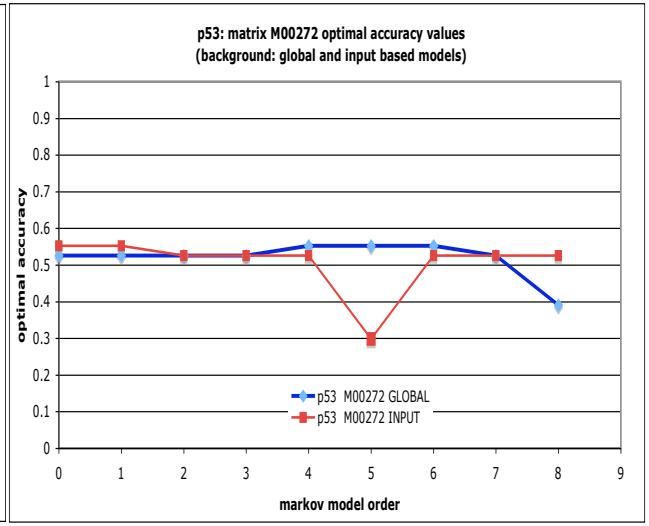


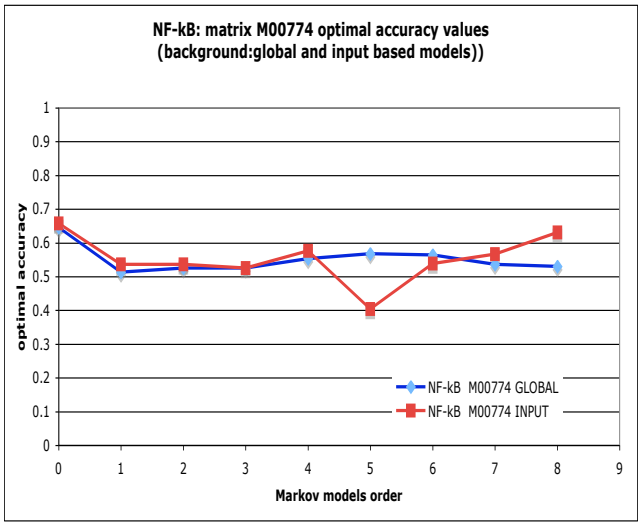
Jean Valéry Turatsinze (2005). In prep.

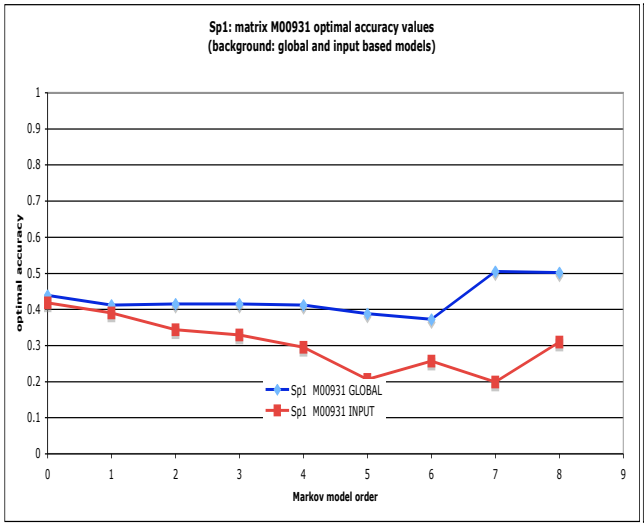
Effect of the matrix and of the background model

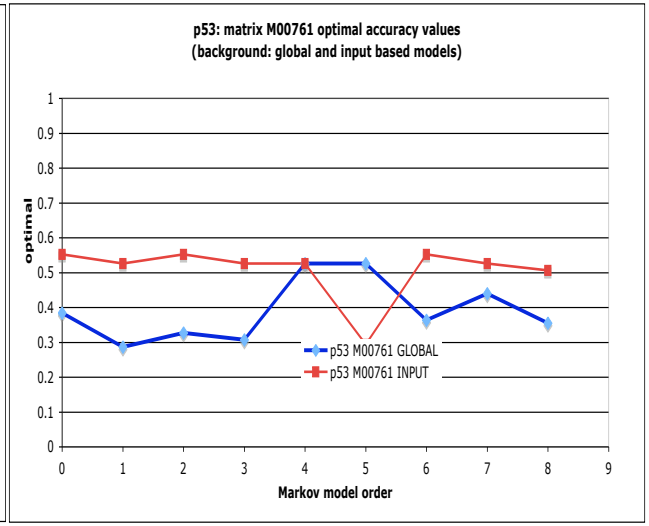








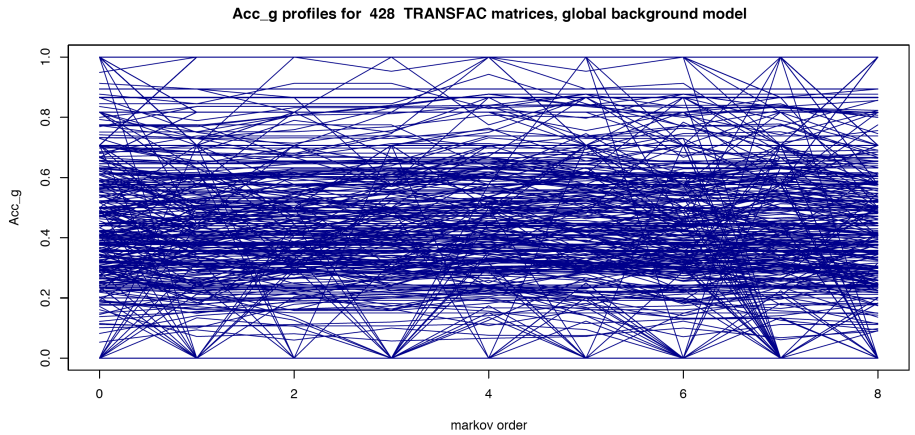




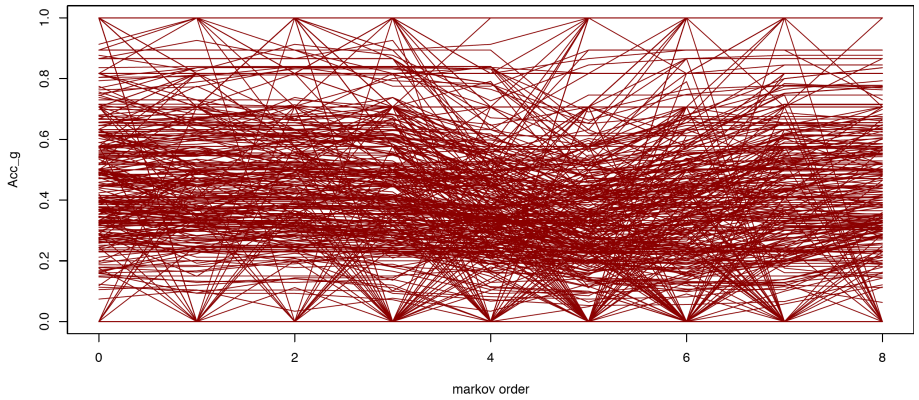
Jean Valéry Turatsinze (2005). In prep.

Accuracy profiles for 428 TRANSFAC matrices

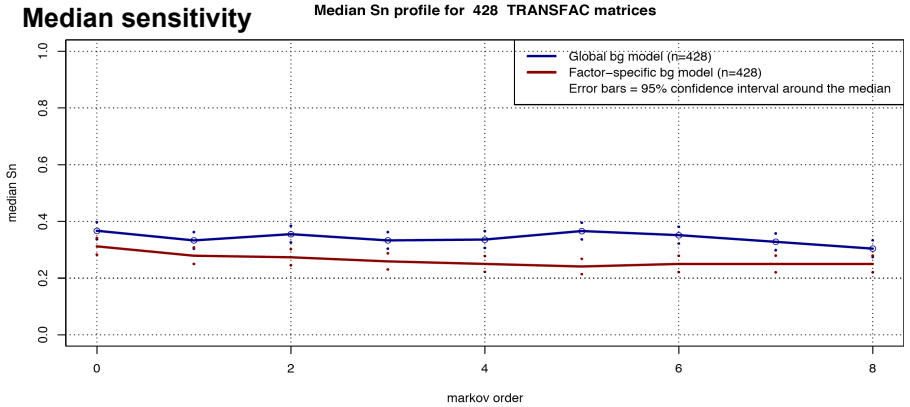
All accuracy profiles, global background



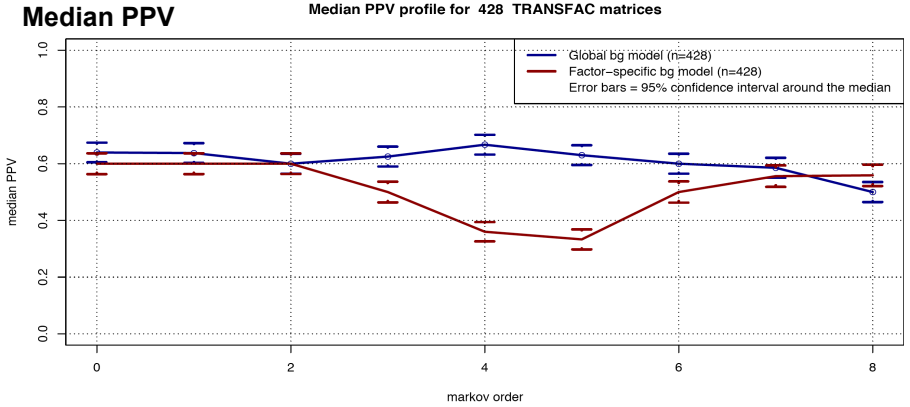
All accuracy profiles, factor-specific background



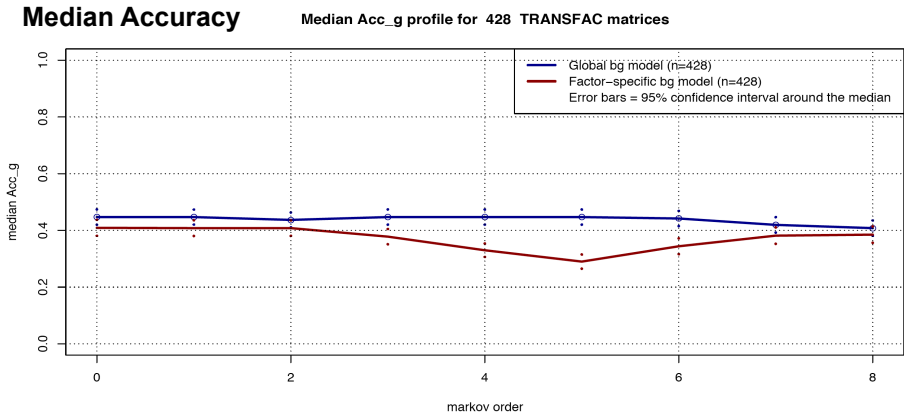
Median sensitivity



Median PPV



Median Accuracy



Evaluation of pattern discovery results

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

Formats of pattern discovery results (string-based approaches)

Collection of words (can be IUPAC) with a score assigned to each

Word pair	*F(W)*	*Match. Seq.*	*occ*	*E(W)*	*P-value*	*E-value*	*sig*	*Overlaps (discarded)*	*Rank*
CACGTG|CACGTG	0.000164	9	13	1.42	4e-09	8.4e-06	5.08	0	1
CCACAG|CTGTGG	0.000265	8	11	2.30	3e-05	6.2e-02	1.21	0	2
ACGTGA|TCACGT	0.000368	9	13	3.19	3e-05	6.3e-02	1.20	6	3
AACTGT|ACAGTT	0.000610	10	17	5.28	3.8e-05	8.0e-02	1.10	0	4
ACTGTG|CACAGT	0.000374	9	12	3.24	0.00015	3.0e-01	0.52	0	5
GCTTCC|GGAAGC	0.000421	7	12	3.65	0.00042	8.6e-01	0.06	0	6
GCCACA|TGTGGC	0.000307	7	10	2.66	0.00045	9.4e-01	0.03	0	7
AGTCAT|ATGACT	0.000489	8	13	4.24	0.00046	9.6e-01	0.02	0	8

Collection of words, assembled in several motifs

```
;cluster # 1       seed: CACGTG       3 words      length
TCACGT..     ..ACGTGA     1.20
.CACGTG.     .CACGTG.     5.08
..ACGTGA     TCACGT..     1.20
TCACGTGA     TCACGTGA     5.08   best consensus

;cluster # 2       seed: CCACAG       4 words      length 8
GCCACA...    ...TGTGGC    0.03
.CCACAG..    ..CTGTGG.    1.21
..CACAGT.    .ACTGTG..    0.52
...ACAGTT    AACTGT...    1.10
GCCACAGTT    AACTGTGGC    1.21   best consensus

; Isolated patterns: 2
GCTTCC       GGAAGC       0.06
AGTCAT       ATGACT       0.02
```

Formats of pattern discovery results (matrix-based approaches)

Position-specific scoring matrix with the sites used in the alignment

```
MATRIX 1
number of sequences = 5
unadjusted information = 12.264
sample size adjusted information = 28.1942
ln(p-value) = -40.0503   p-value = 4.03996E-18
ln(expected frequency) = -3.91122   expected frequency = 0.0200161
A |   1   2   0   5   0   0   0   0   0   0
C |   3   0   5   0   5   0   0   0   1   2
G |   0   3   0   0   0   5   0   5   4   3
T |   1   0   0   0   0   0   5   0   0   0
  1|1   :   1/546   CACACGTGGG
  2|2   :   2/516   CACACGTGGG
  3|5   :  -3/265   TGCACGTGGC
  4|3   :   4/385   AGCACGTGGG
  5|4   :  -5/455   CGCACGTGCC
```

Motif comparisons

- How can we compare the results of different pattern discovery programs **at the motif level** ?
- Comparison between annotated binding sites and discovered motifs
  - String against string
  - Matrix against string
- Comparison between annotated PSSM and discovered motifs
  - String against matrix
  - Matrix against matrix

String(s) against string(s)

- The matching score should take into account the following information:
  - number of matching/non-matching positions
  - significance of matches between IUPAC codes
    - C against C (perfect match) a good score
    - C against S should have a lower score
    - C against N should have a null score
  - partial overlaps
  - prior residue frequencies
    - in yeast promoters for example, A against A is more probable than C against C.

Matrix against matrix

- Circumvent the problem
  - String)to-string comparison with
    - Collection of sites used to build the annotated matrix
    - Collection of sites used to build the predicted matrix

Matrix against string(s)

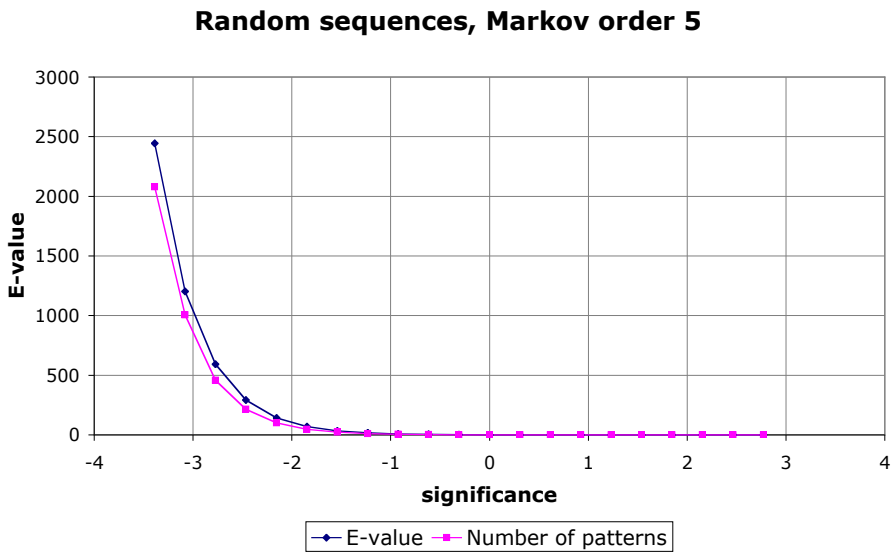
- Collection of sites used to build the matrix against collection of sites
- Consensus from the collection of strings against matrix consensus (1 string to 1 string)
- Build a matrix from the string-based pattern and matrix-to-matrix comparison

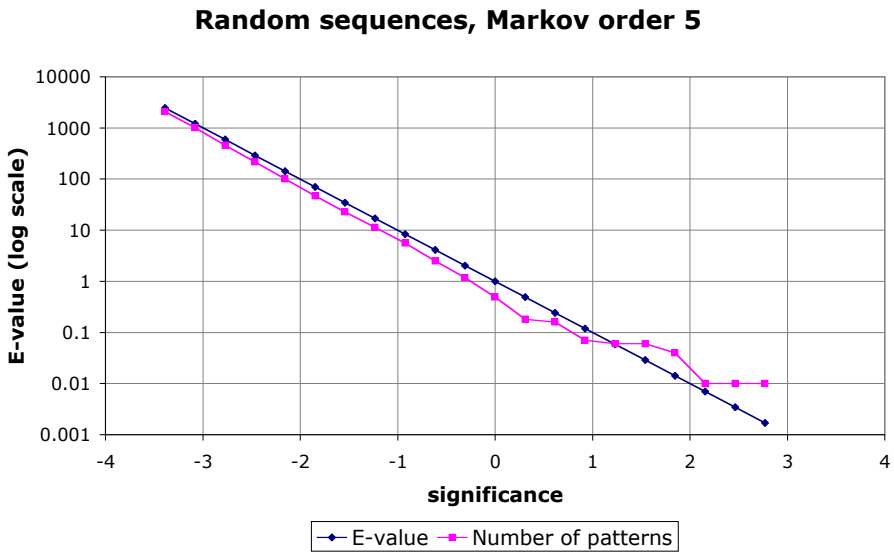
Rates of false positives in random sequences

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

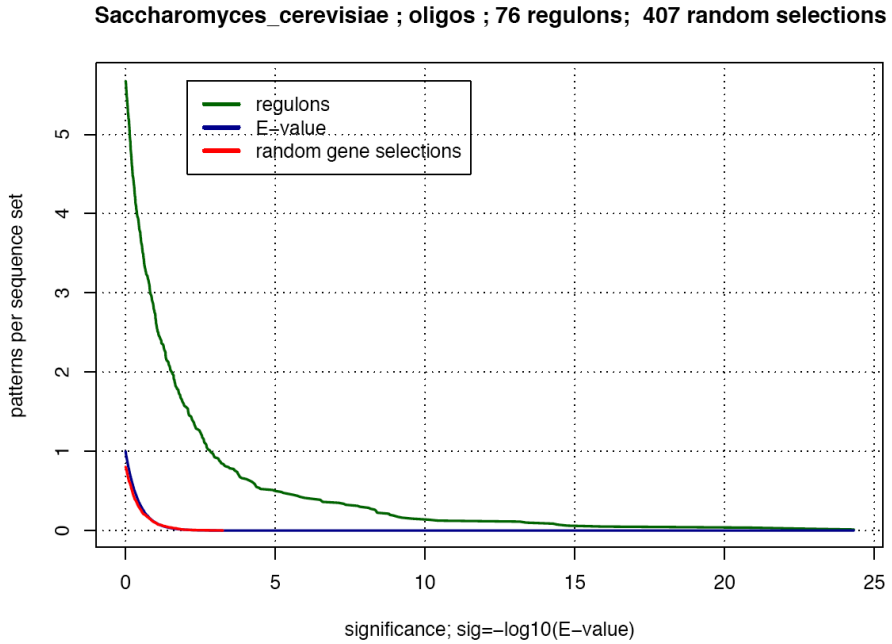
A trivial negative control: rate of false positives in random sequences

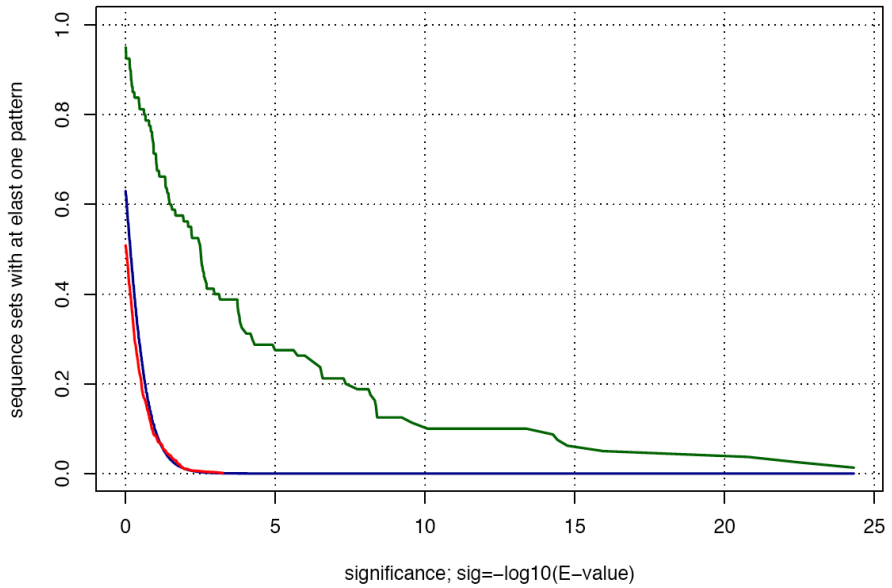
- A first control is to measure the number of significant patterns in a set of random sequences.
  - Random sequences can be generated according to different background models (Bernoulli or Markov chains of various orders).
- Test
  - Measure the number of patterns above each score value (empirical E-value).
  - Compare it with the theoretical E-value (calculated with the binomial)
- Weakness
  - This test mainly checks the appropriateness of the scoring statistics for the chosen background model.
  - It does not tell much about the behaviour of the program with real biological sequences.





Pattern significance in regulons - Saccharomyces cerevisiae





- As a control, we compare the significance of patterns discovered in
  - regulons (*positive control*)
  - random gene selections (*negative control*)
- In the yeast *Saccharomyces cerevisiae*
  - The rate of false positive corresponds remarkably well with the theoretical expectation.
  - When the score increases,
    - the sensitivity (patterns discovered in regulons) decreases,
    - the specificity increases (less patterns in random selections)

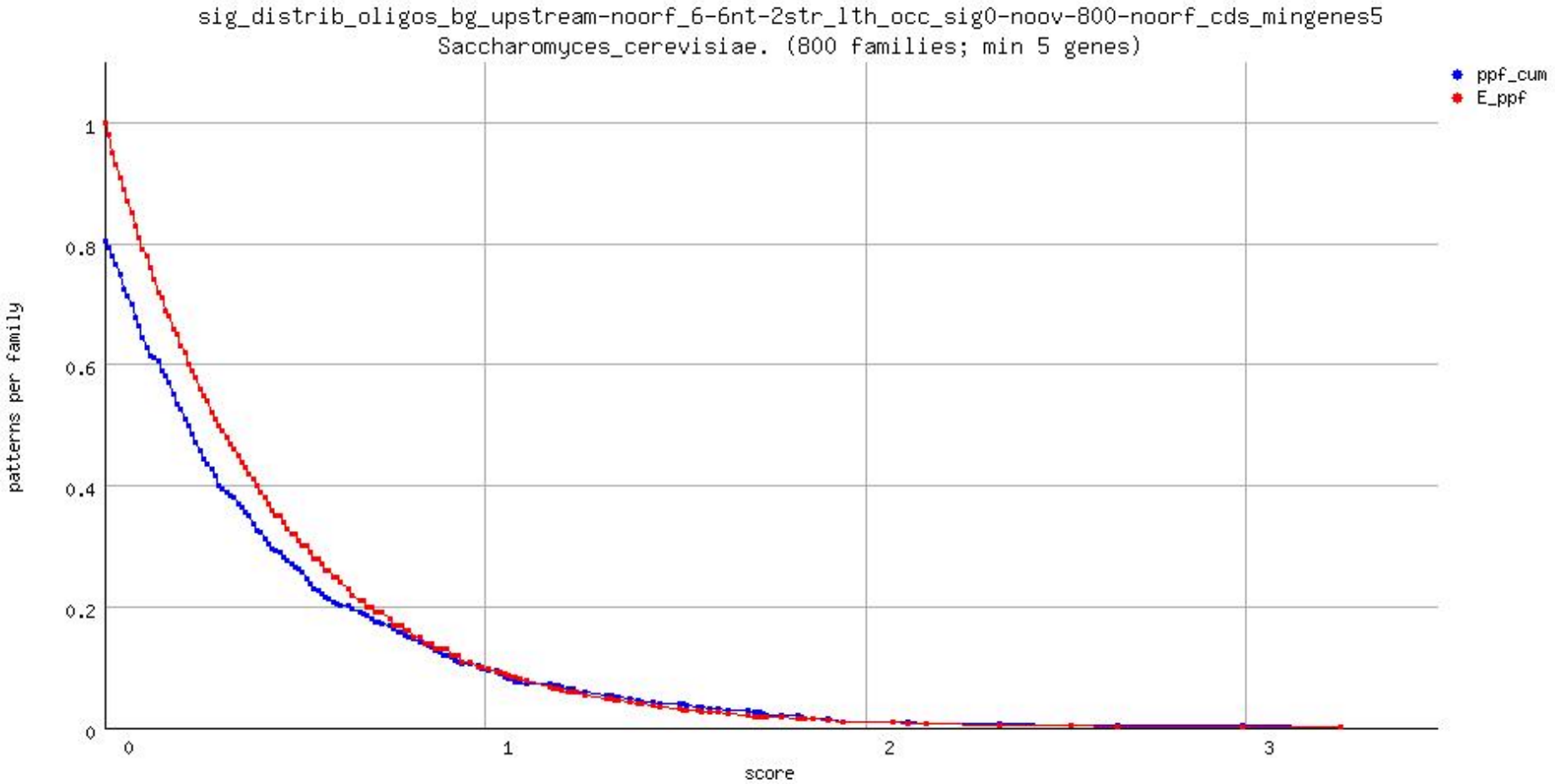
O.Sand & J.van Helden, in prep.

Rates of false positives estimated in different organisms with random gene selections

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

Patterns per sequence set - Saccharomyces cerevisiae

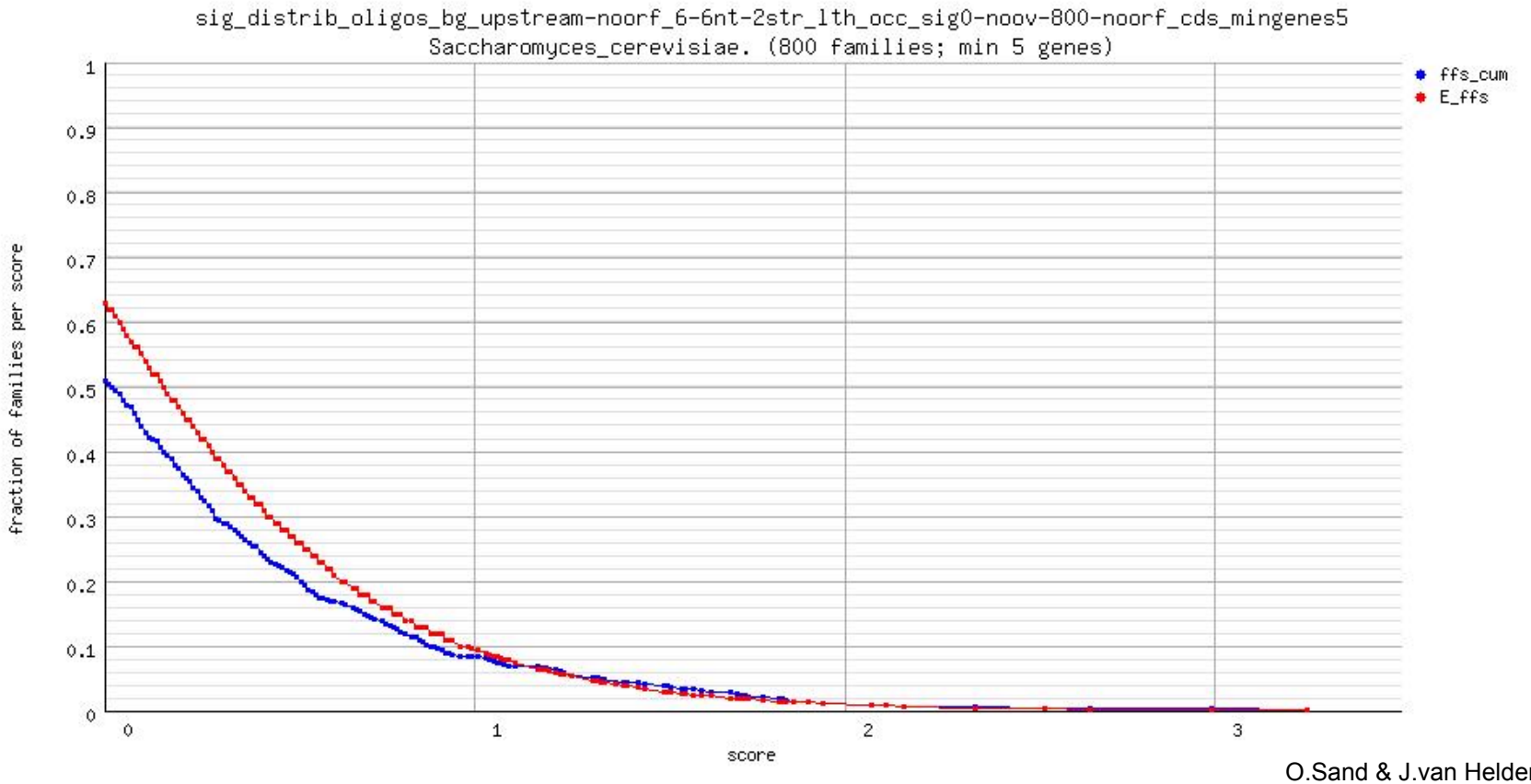
- Number of patterns selected per sequence set, as a function of the score (sig)
  - Red: expected number of false positives (**E-value**)
  - Blue : observed number of false positive
    (patterns discovered in random gene selections)



O.Sand & J.van Helden, in prep.

Sequence sets per score - Saccharomyces cerevisiae

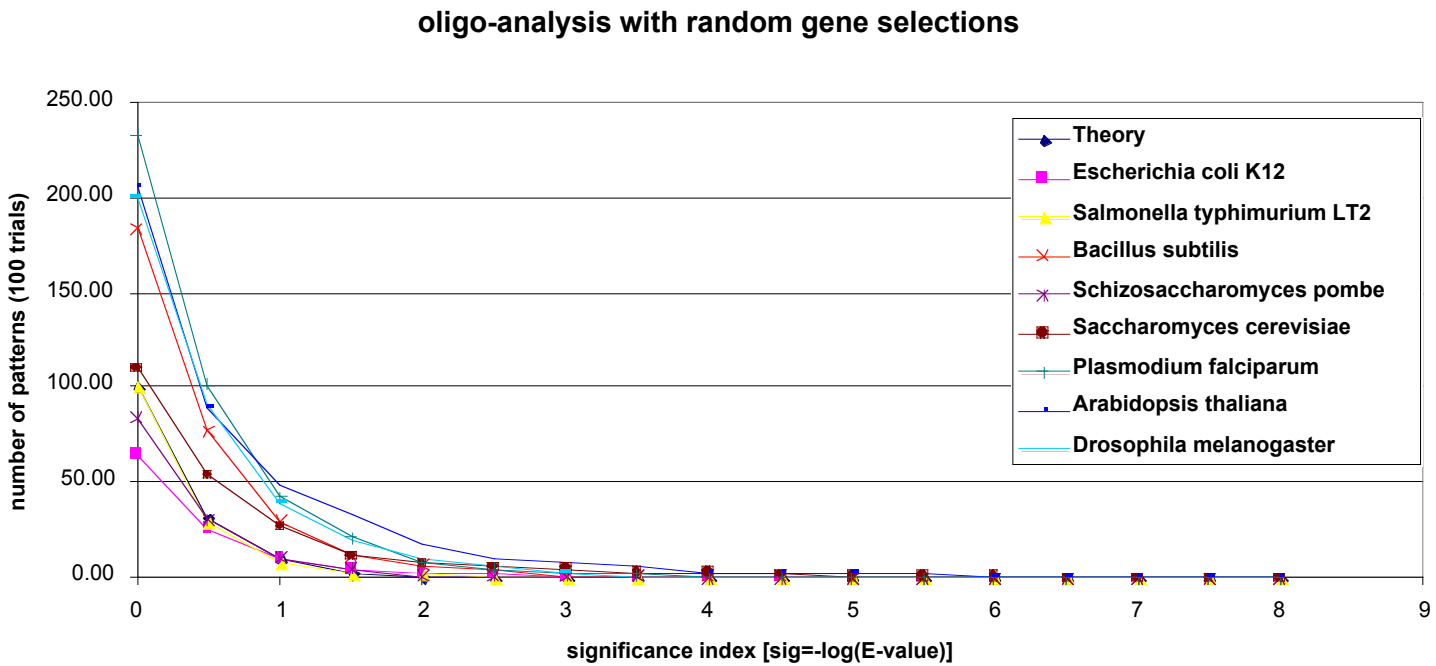
- Number of sequence sets with at least one discovered pattern
  - Red: expected number of false positives:
    - Family-wise error rate (**FWER**) FWER = 1 - (1-P-value)^N
    - Probability to have at least one false positive in the sequence set.
  - Blue : observed number of false positive
    (random gene selections with at least one discovered pattern)



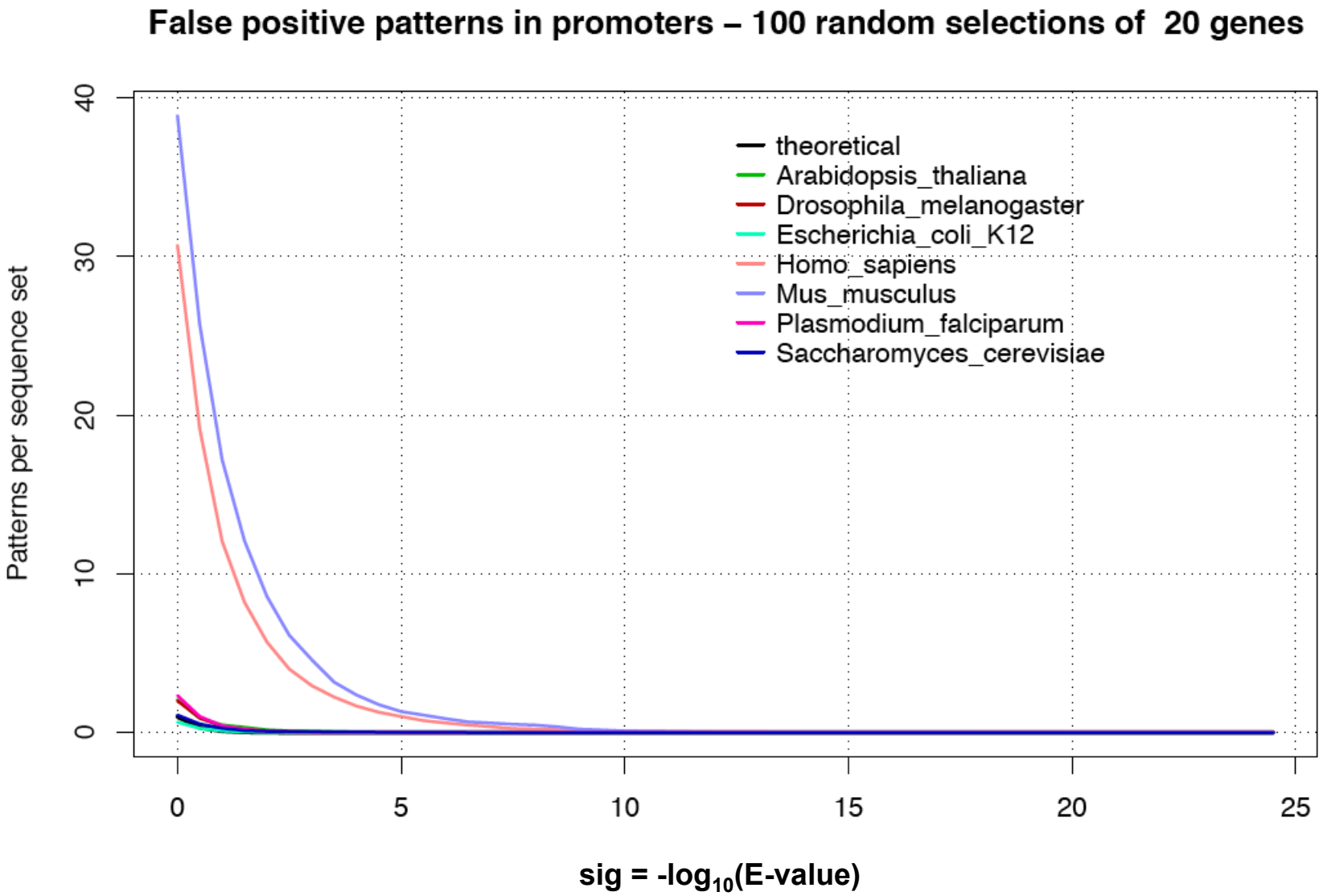
O.Sand & J.van Helden, in prep.

Rate of false positive in different organisms

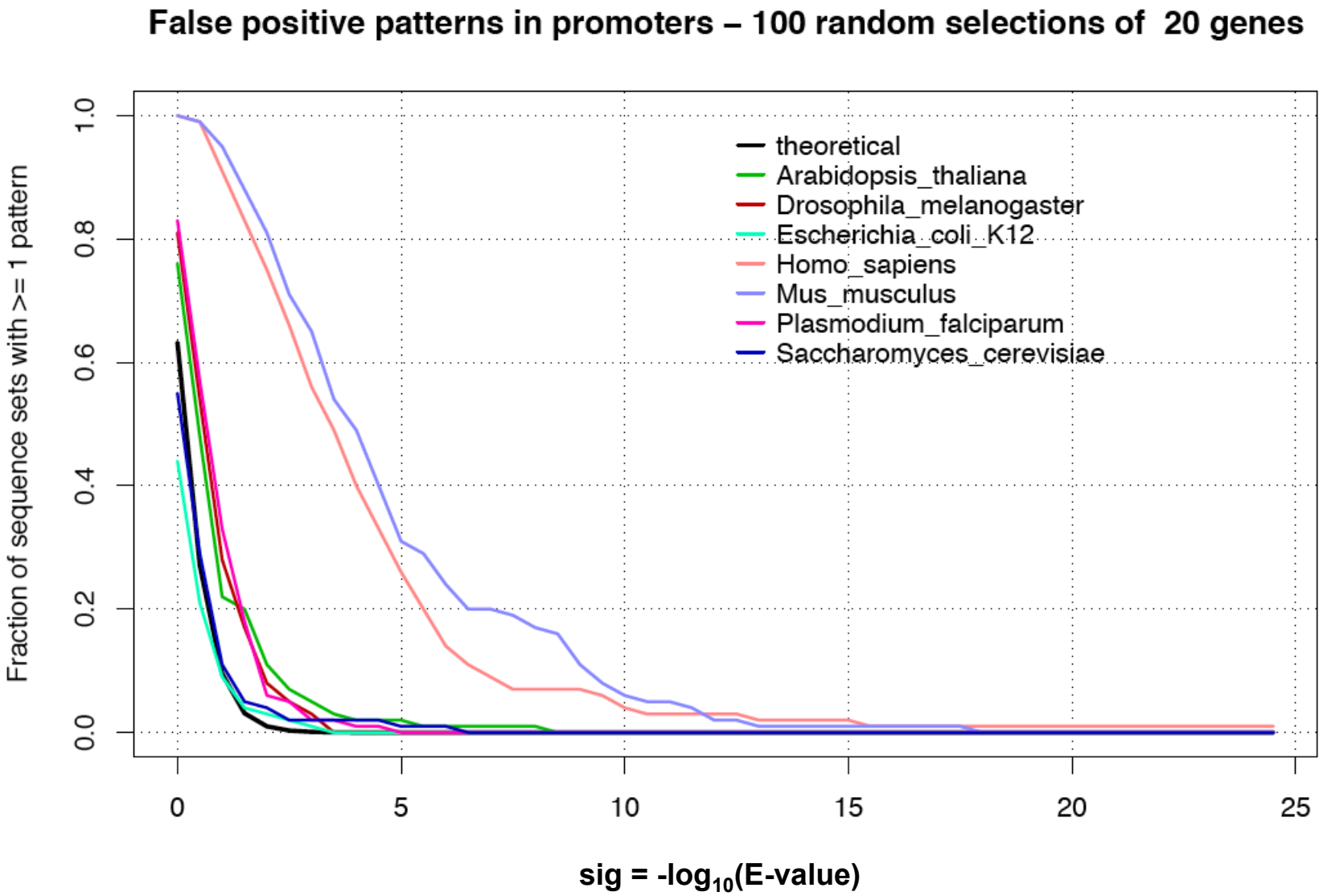
- The analysis of random gene selections allows to evaluate the rate of false positive returned by a pattern discovery program.
- The rate of false positive is good for microbes (bacteria and yeasts), but increases for higher organisms.
- This is likely to result from the higher heterogeneity of genomic sequences in these organisms. We are currently developing more elaborate background models to treat this problem.



False discovery rate - patterns per sequence set



Sequence sets with >= 1 selected patterns



False positive rate - summary

- In lower organisms (bacteria, yeast), the rate of false positive observed in random selections of promoters follows pretty well the theoretical expectation, as calculated with the binomial distribution.
- This rate of false positive increase spectacularly with promoters of higher organisms (human, mouse). This suggest that words distributions do not follow the binomial distribution.

Regulatory Sequence Analysis

Score distributions in positive and negative control sets

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

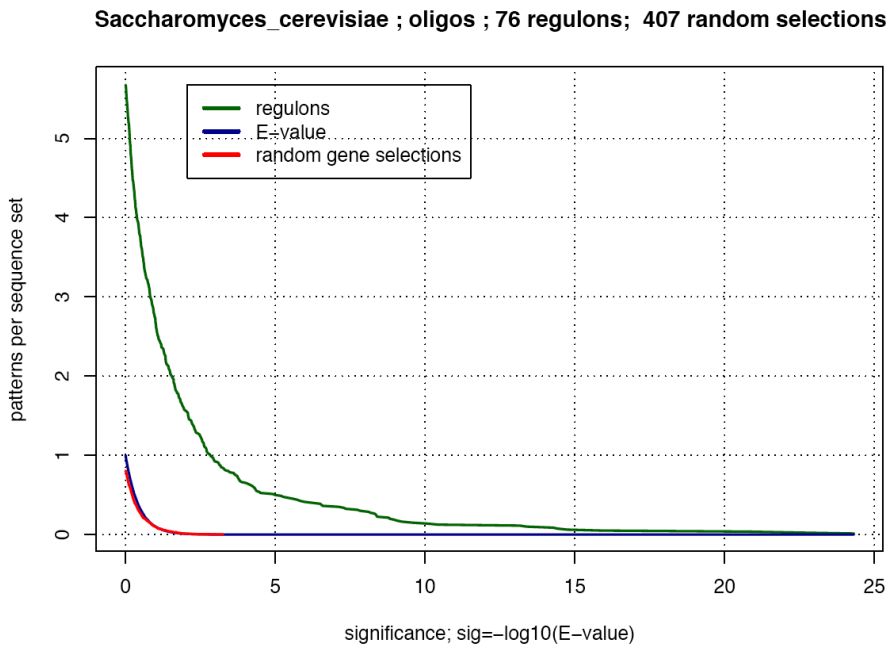
Negative control

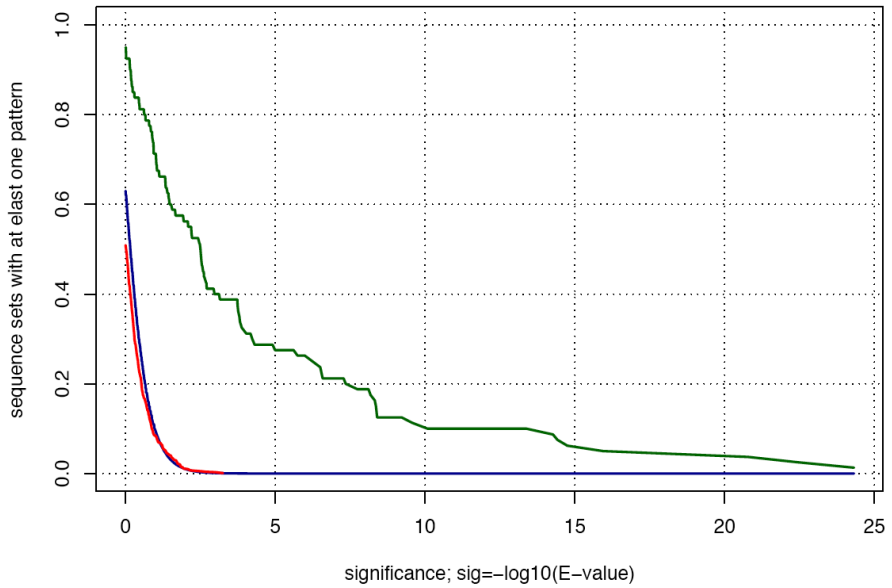
- Randomly generated sequences.
  - Which generating model ?
    - Bernoulli (independent succession of nucleotides)
    - Markov chain of order *k* (the probability of each nucleotide depends on the *k* preceding ones)
  - Problem
    - This control evaluates the fitting of our program with the chosen random mode
    - This is not always indicative of its behaviour on real biological sequences.
- Random gene selection
  - Select a random set of genes.
  - Retrieve their promoters.
  - Apply exactly the same procedure to these promoters as you did for regulons.

Positive control

- Measure the significance of motifs discovered in
  - Annotated regulons (TRANSFAC, RegulonDB).
    - Strength: reliable information
    - Weakness: annotation represents a fraction of publications, which represent a fraction of the real target genes
  - High throughput regulons (Lee et al)
    - Strength: supposed to be exhaustive and homogeneous
    - Weakness: noise

Pattern significance in regulons - Saccharomyces cerevisiae

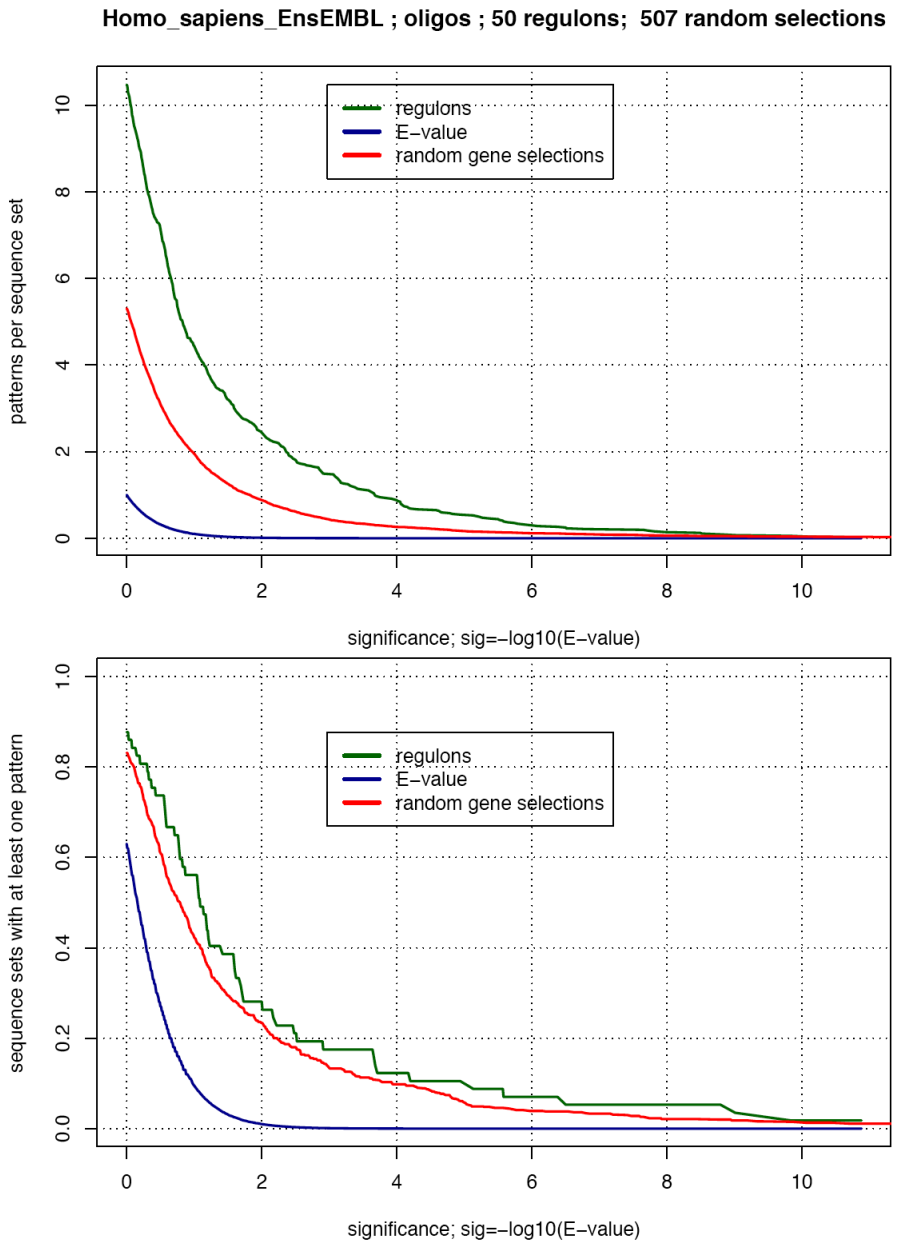




- As a positive control, we compare the significance of patterns discovered in regulons (*positive control*) and random gene selections (*negative control*)
- In the yeast *Saccharomyces cerevisiae*
  - The rate of false positive corresponds remarkably well with the theoretical expectation.
  - When the score increases,
    - the sensitivity (patterns discovered in regulons) decreases,
    - the specificity increases (less patterns in random selections)

O.Sand & J.van Helden, in prep.

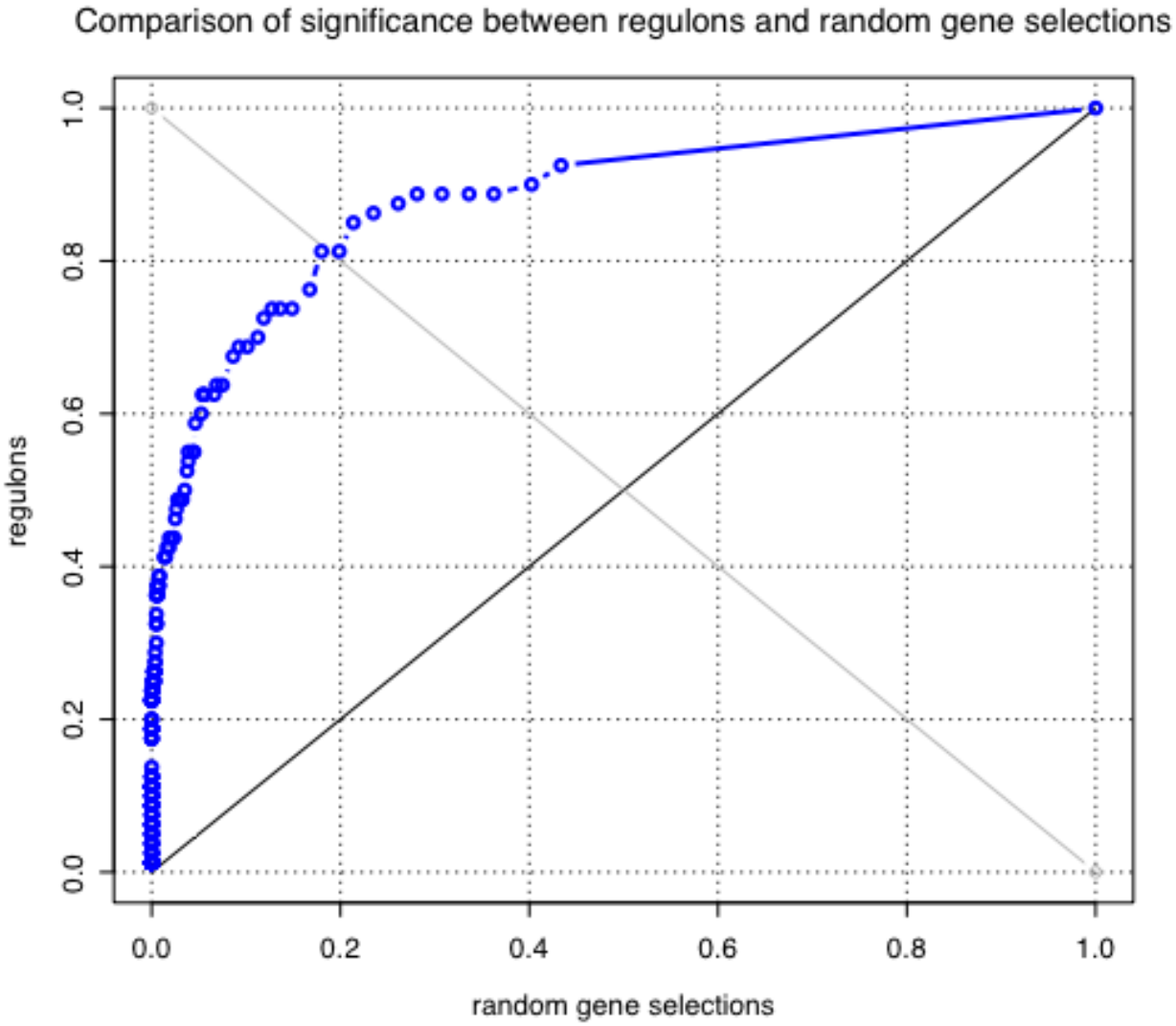
Pattern significance in regulons - Homo sapiens



- In Homo sapiens
  - The rate of false positive is much higher than the theoretical expectation
  - The number of patterns detected in regulons is still higher, but the significance score is quite inefficient to distinguish between reliable motifs and false positives.
  - This indicates that the background model is unadequate to treat the complexity of human promoters.

O.Sand & J.van Helden, in prep.

Receiver Operating Characteristic (ROC) curve



- X axis: 1 - specificity
  - In our case, we plot the significance in random gene selections
- Y axis: sensitivity
  - In our case, we plot the significance in regulons
- The area under the ROC curve (AUC) indicates the accuracy of the predictions.

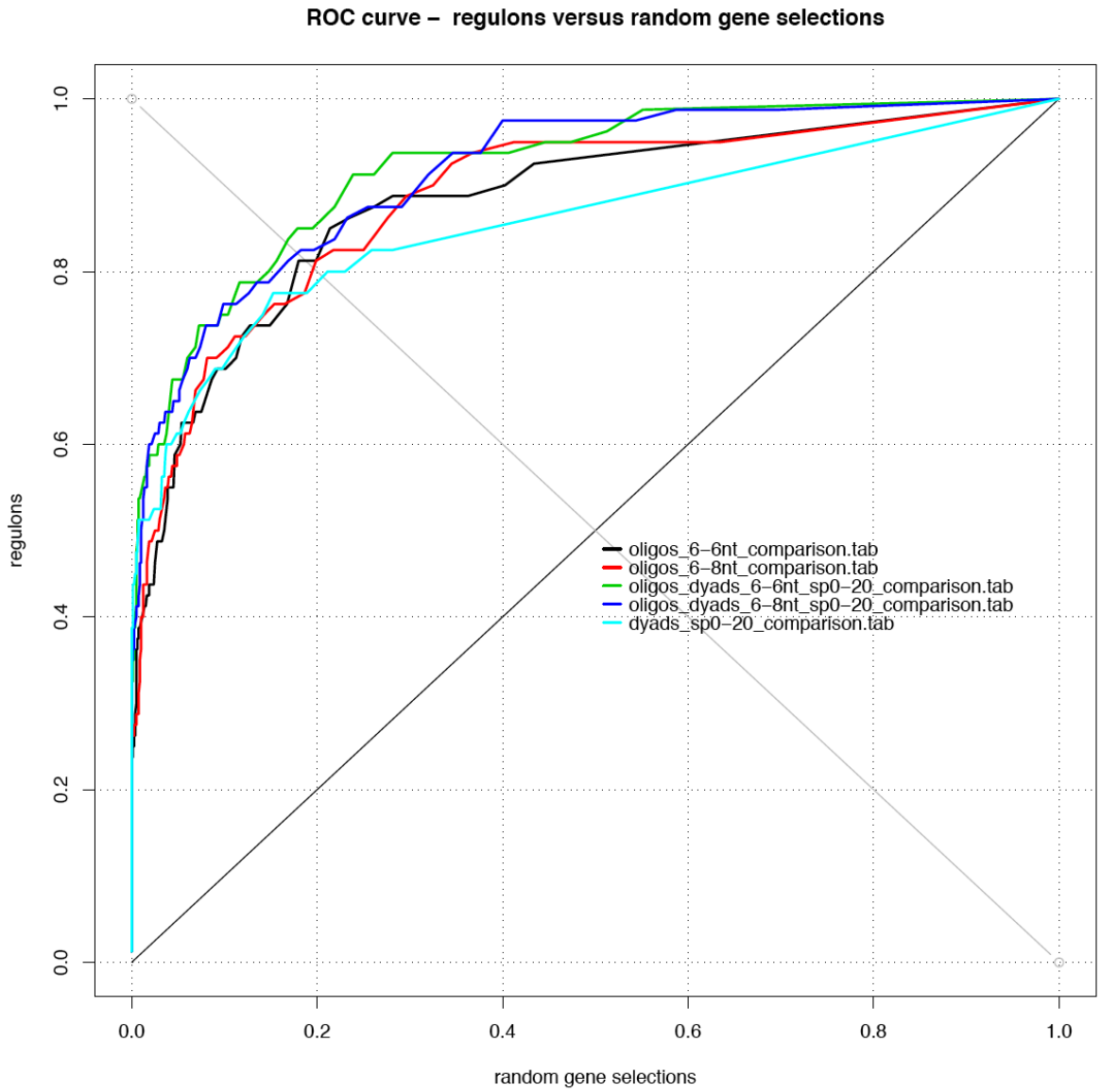
More info on ROC curves:
http://en.wikipedia.org/wiki/ROC_Curve
http://www.anaesthetist.com/mnm/stats/roc/

Regulatory Sequence Analysis

Using ROC curves to select optimal parameters

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

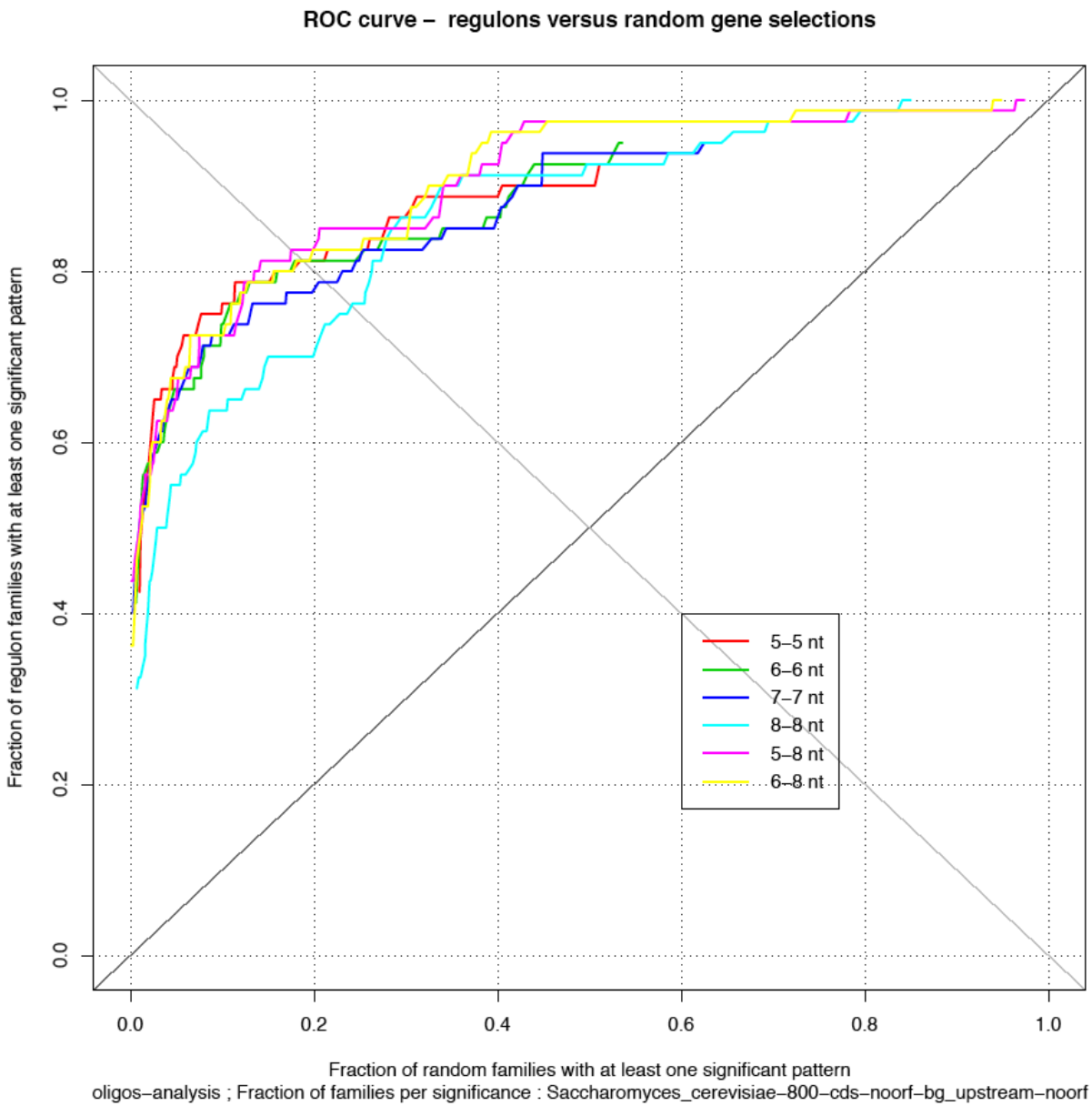
Oligo-analysis and dad-analysis : selection of optimal parameters



- Analysis of 79 yeast regulons (Y axis), versus 79 random gene selections (X axis).
  - Oligo-analysis and dyad-analysis give better results together than any of them alone.
  - Including 7nt and 8nt slightly reduces the accuracy.

Simonis, unpublished results

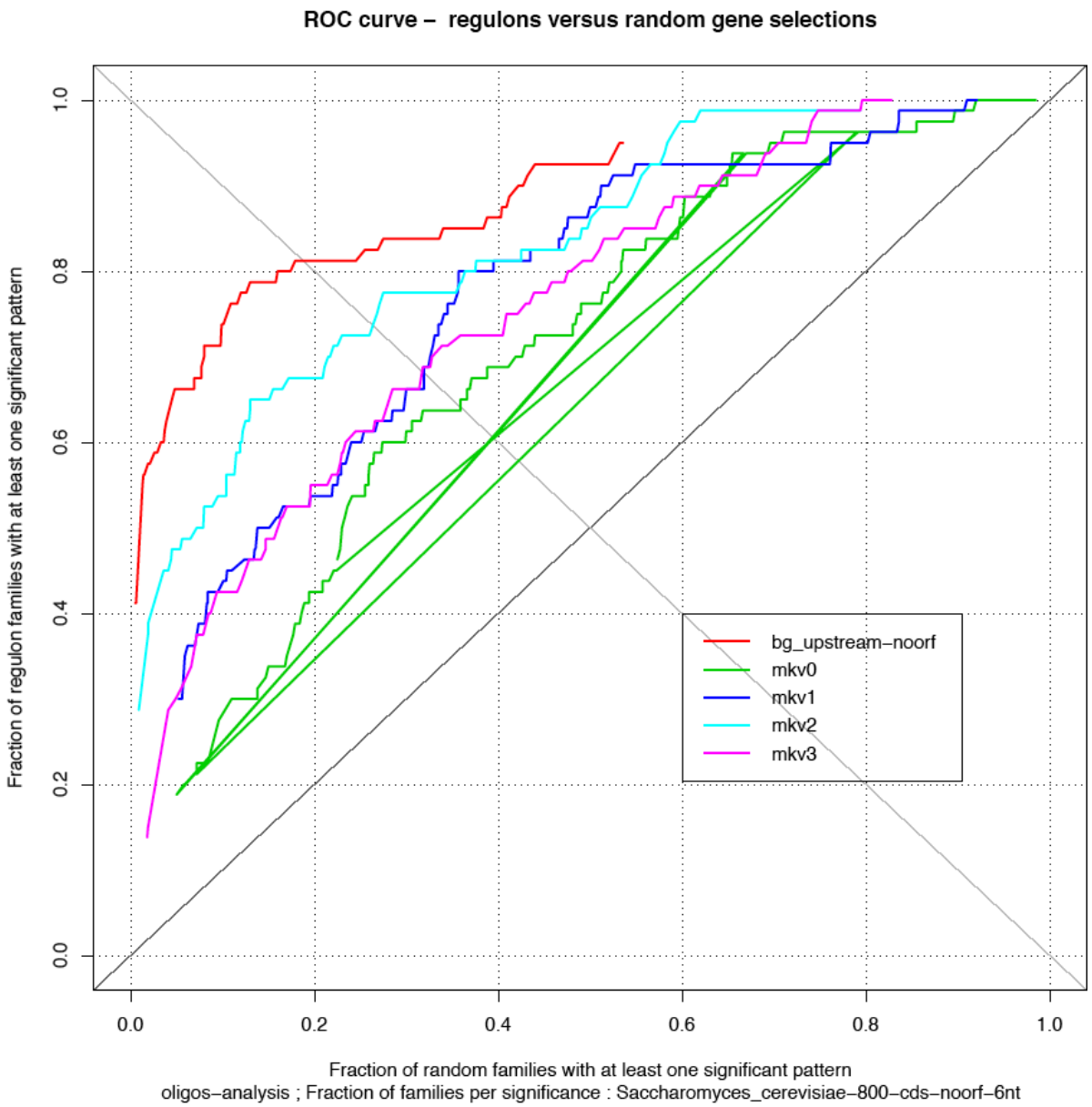
Oligo-analysis: effect of oligonucleotide length



- In the yeast Saccharomyces cerevisiae, the significance score discriminates well between regulons and random gene selections.
- With human promoters, the results are almost identical to those expected from random predictions !

Sand & van Helden, in prep.

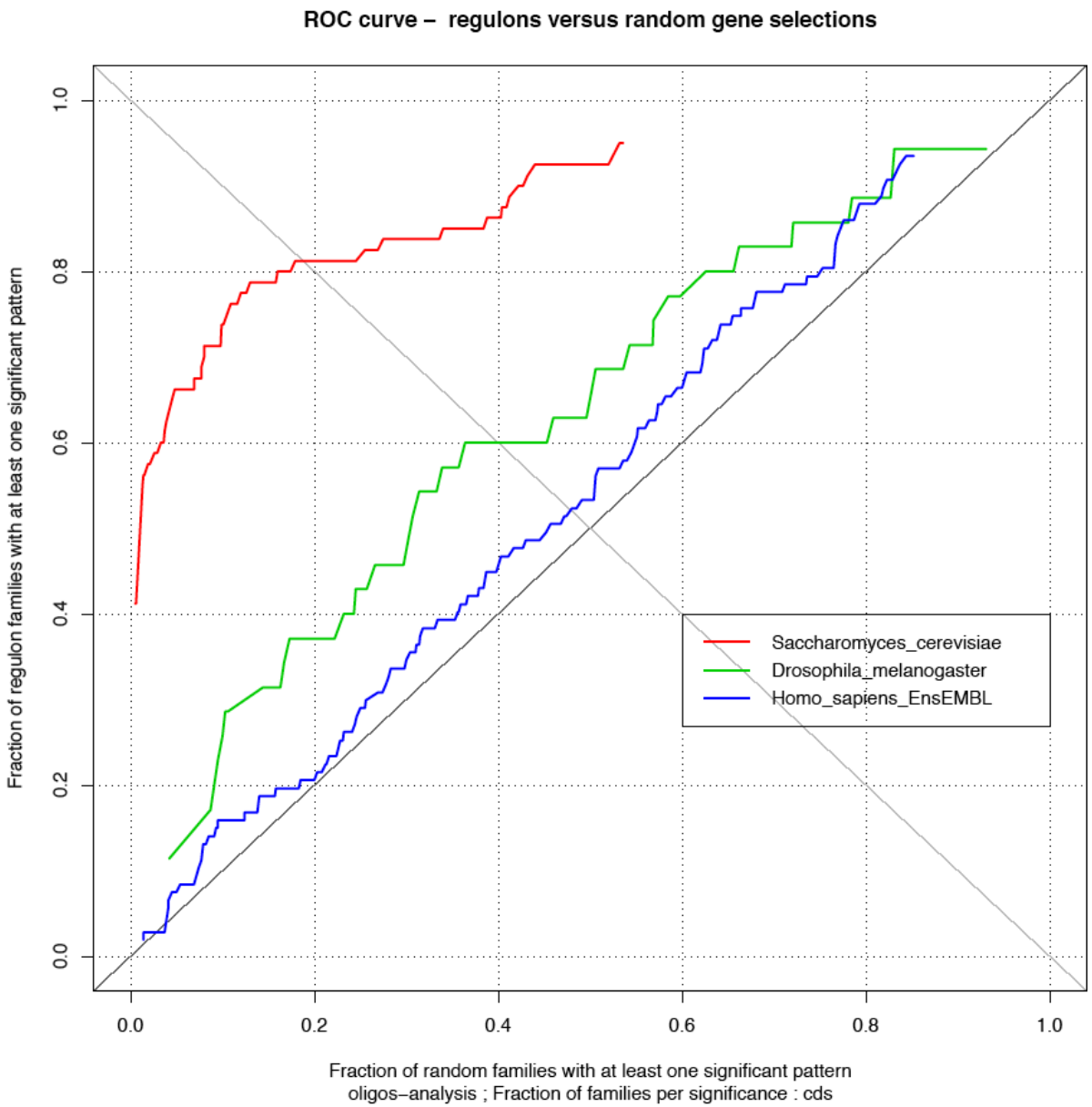
Oligo-analysis: effect of background model



- Prior word frequencies can be estimated with different background models.
  - Upstream-noorf: collection of all upstream sequences for the considered organism (yeast in this case)
  - Markov models trained on the input sequence itself.
  - Note: mkv0 corresponds to a Bernoulli model (independent succession of residues)

Sand & van Helden, in prep.

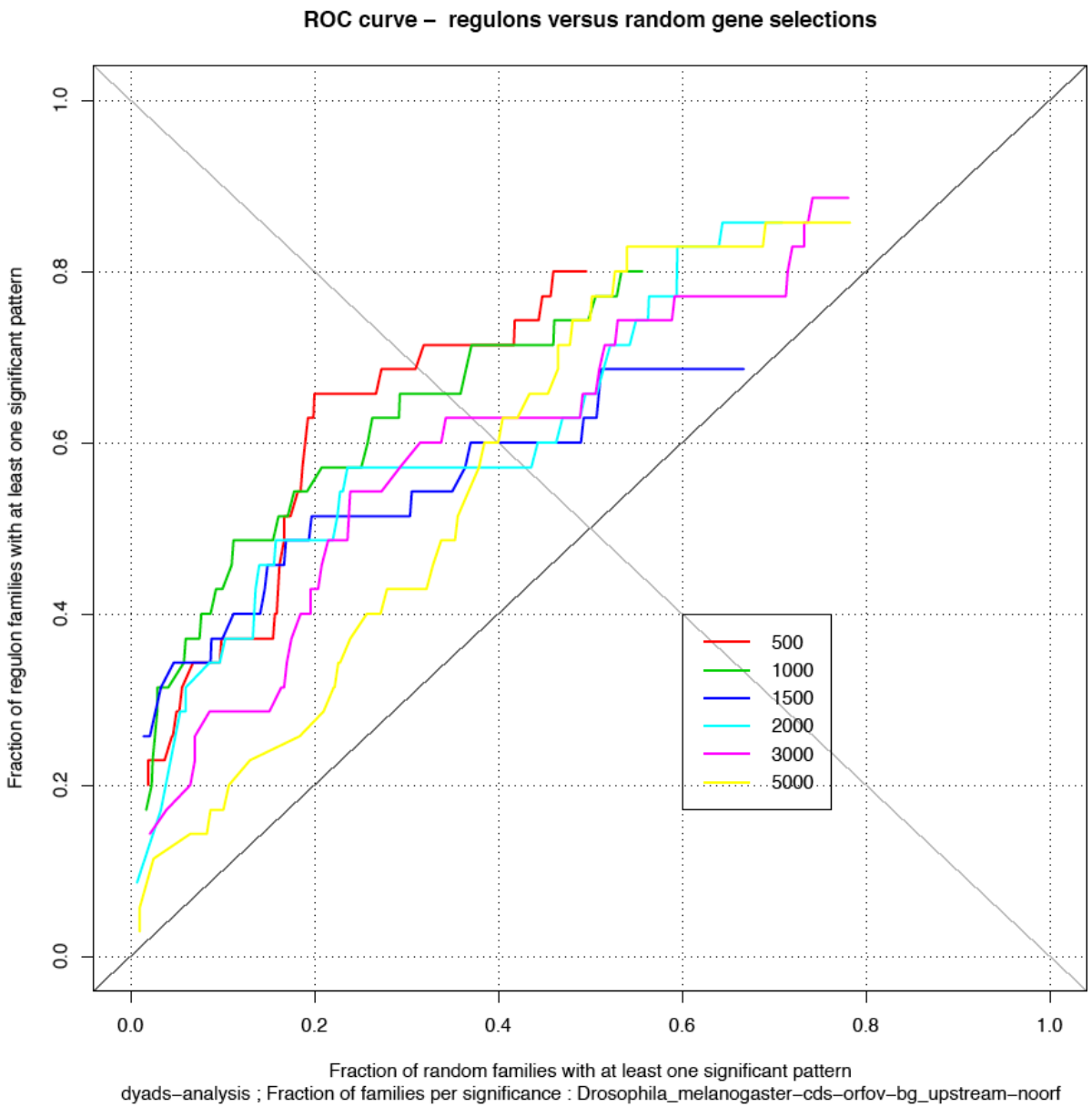
Oligo-analysis: effect of the organism



- In the yeast Saccharomyces cerevisiae, the significance score discriminates well between regulons and random gene selections.
- With human promoters, the results are almost identical to those expected from random predictions !

Sand & van Helden, in prep.

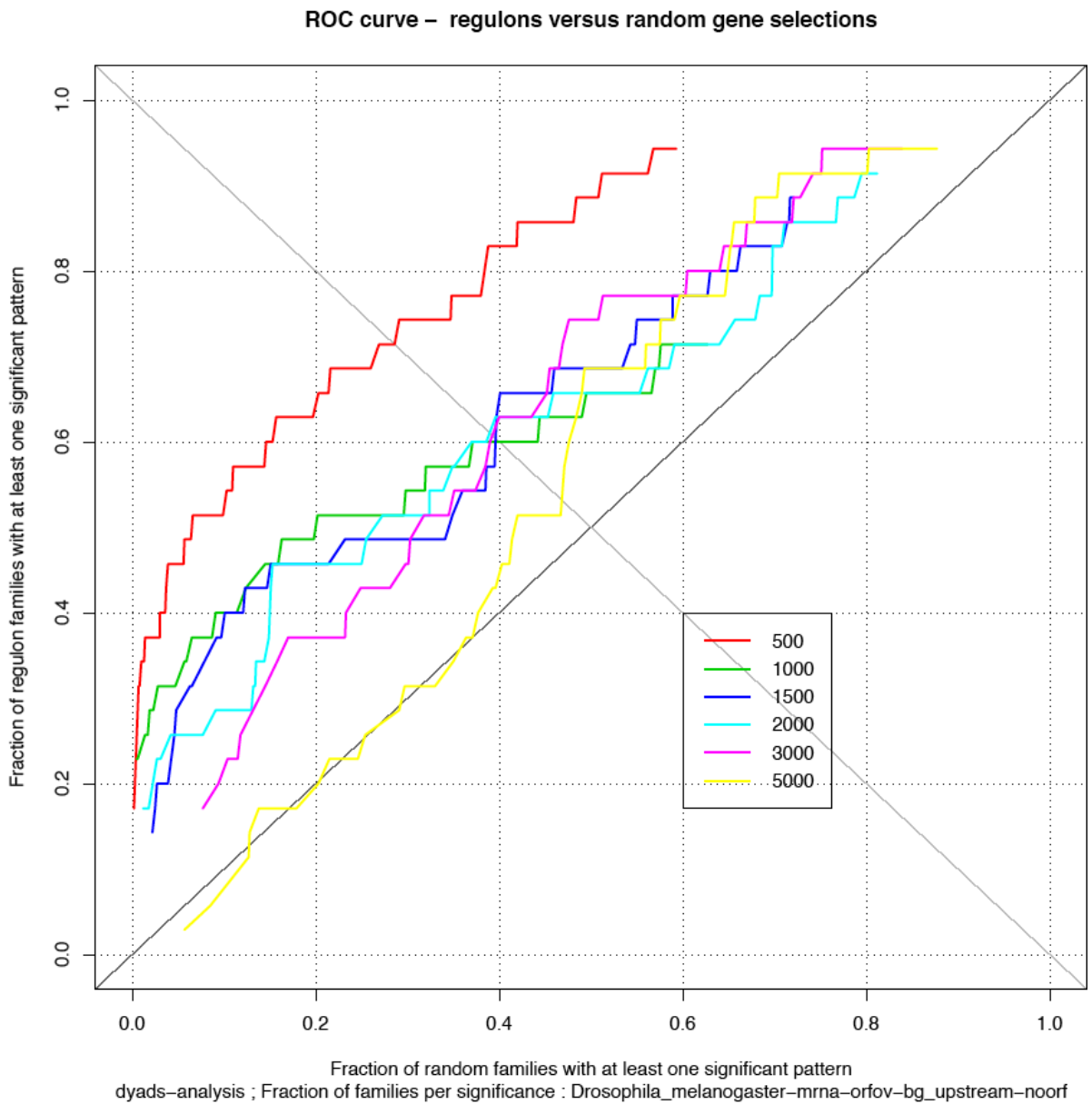
Oligo-analysis: effect of sequence length (from start codon)



- For Drosophila, we compared the results obtained with different sequence lengths (upstream from the start codon).
- Shorter lengths give better results with the testing set (footprints database)

Sand & van Helden, in prep.

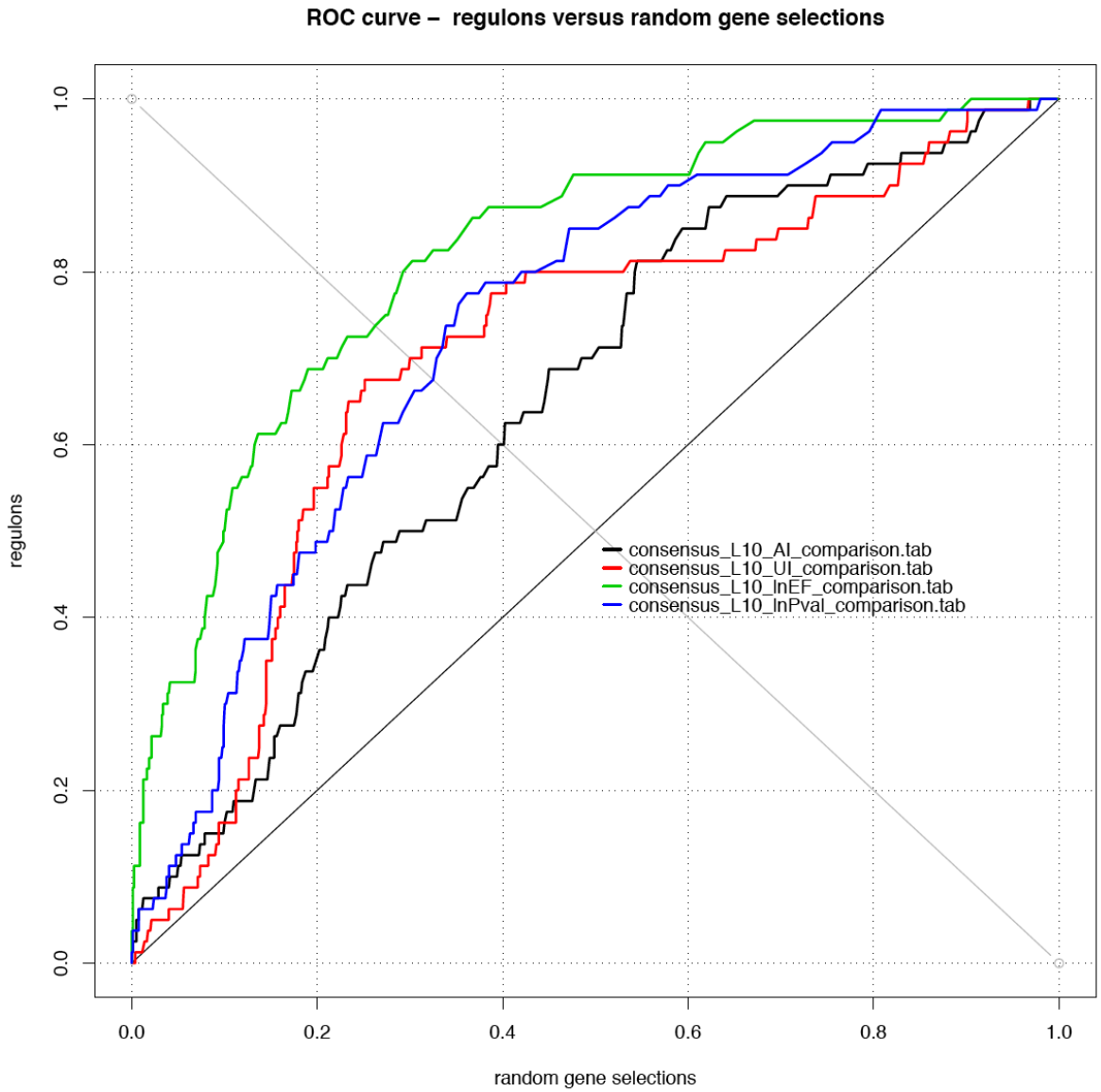
Oligo-analysis: effect of the sequence length (from TSS)



- For Drosophila, we compared the results obtained with different sequence lengths (upstream from the transcription start site).
- Shorter lengths give better results with the testing set (footprints database)
- The fact to use as reference the TSS instead of the start codon increases the performances.

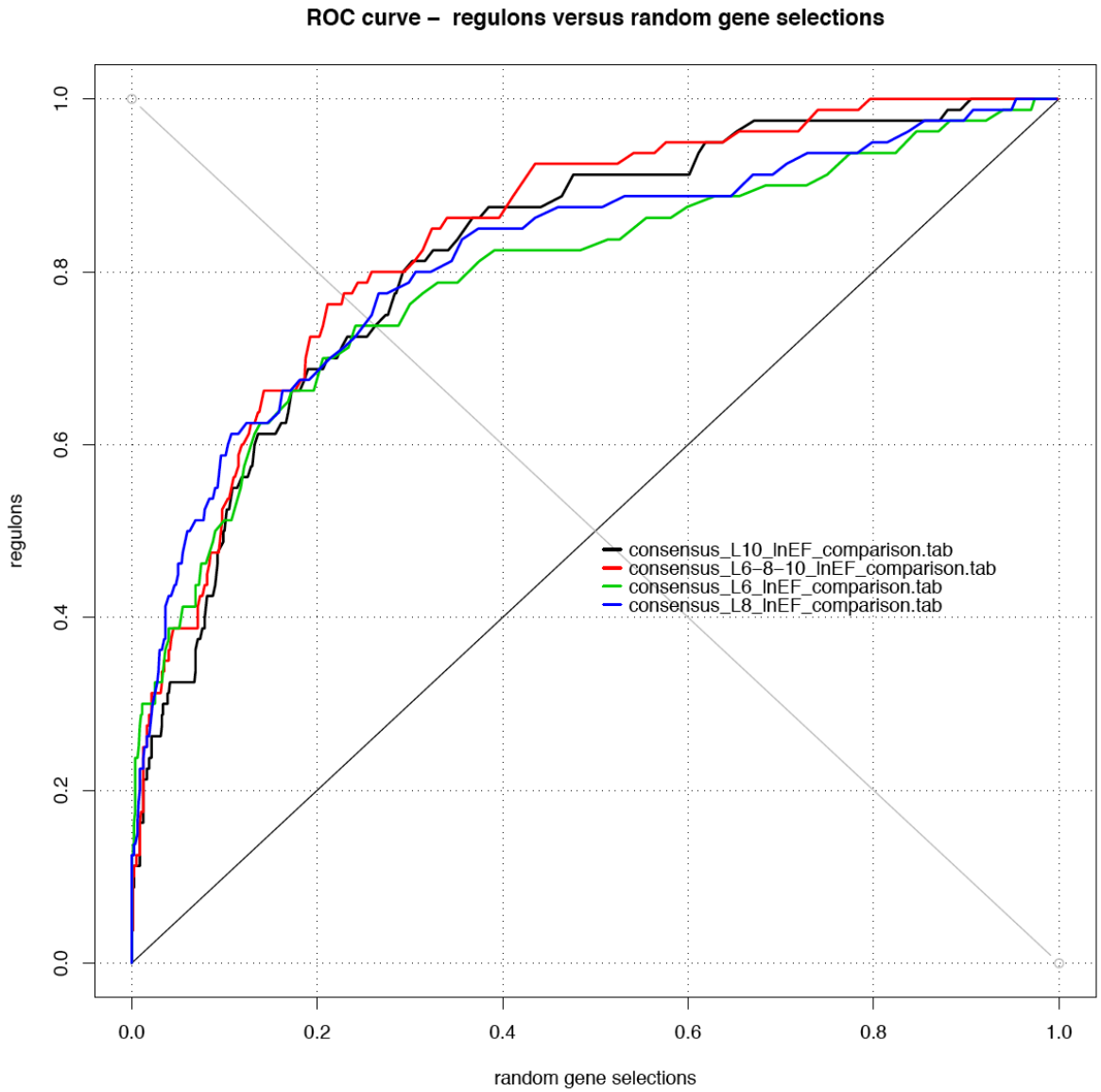
Sand & van Helden, in prep.

Consensus : Selection of optimal scoring function



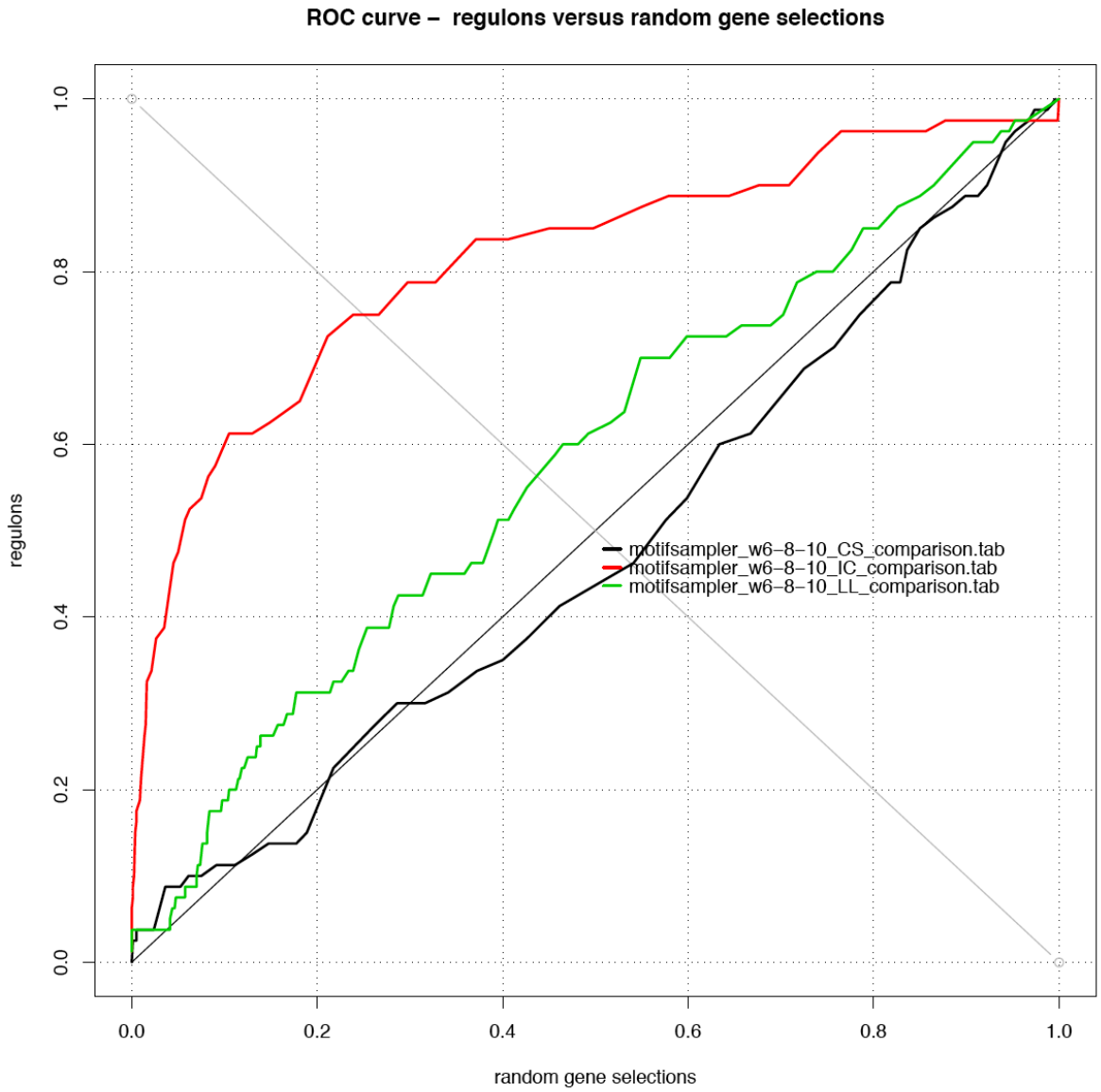
- ROC curves can be used to compare different scoring functions for a given program.
- For ***consensus*** (Hertz, 1990, 1999), E-value is the best scoring function.

Consensus: selection of optimal parameters



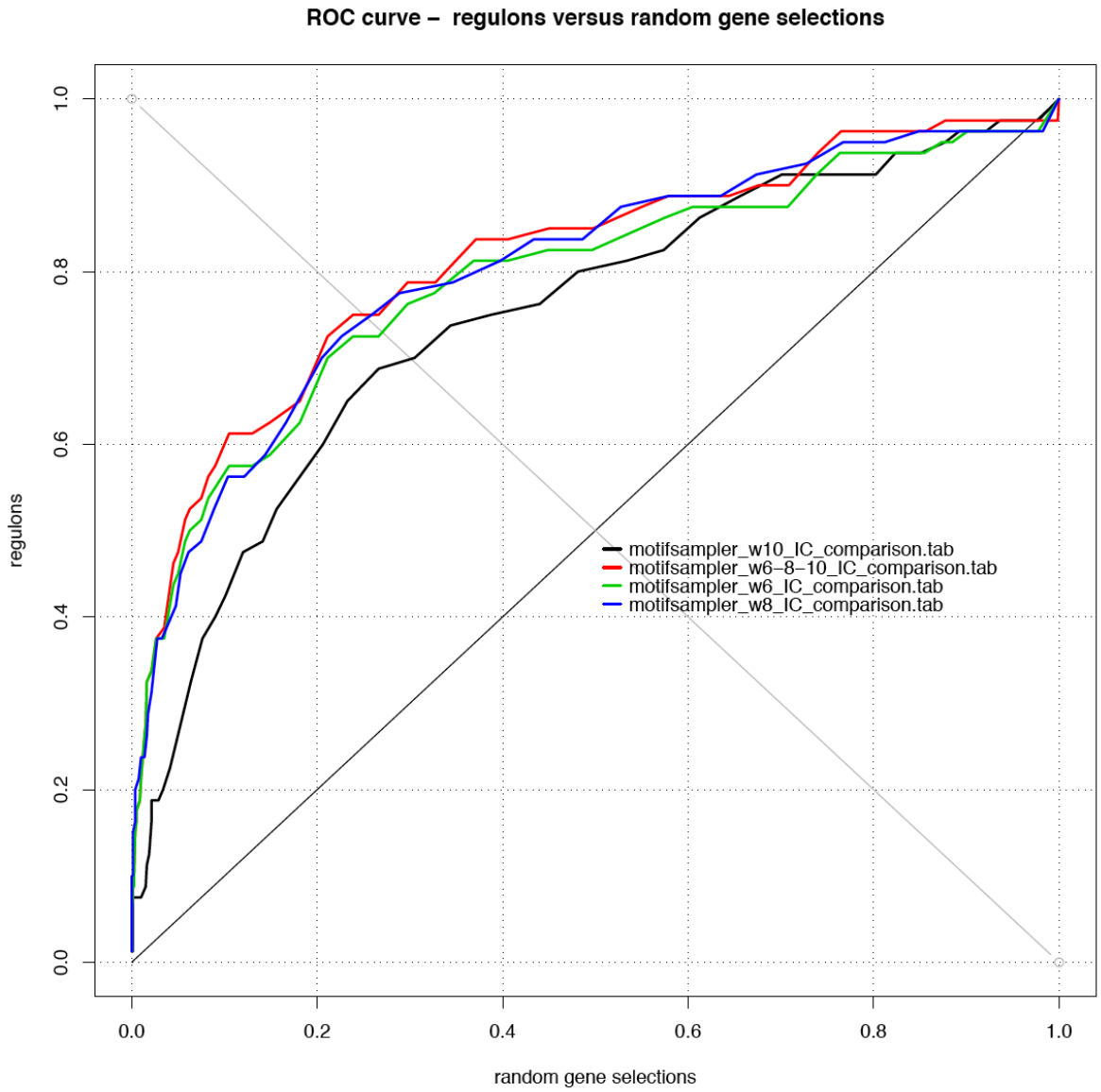
- Having selected the optimal scoring function, the ROC curves can be used to compare other parameter values.
- For ***consensus***, the combination of several matrix widths (6, 8 and 10) gives better results than each width separately.

MotifSampler: selection of optimal scoring function



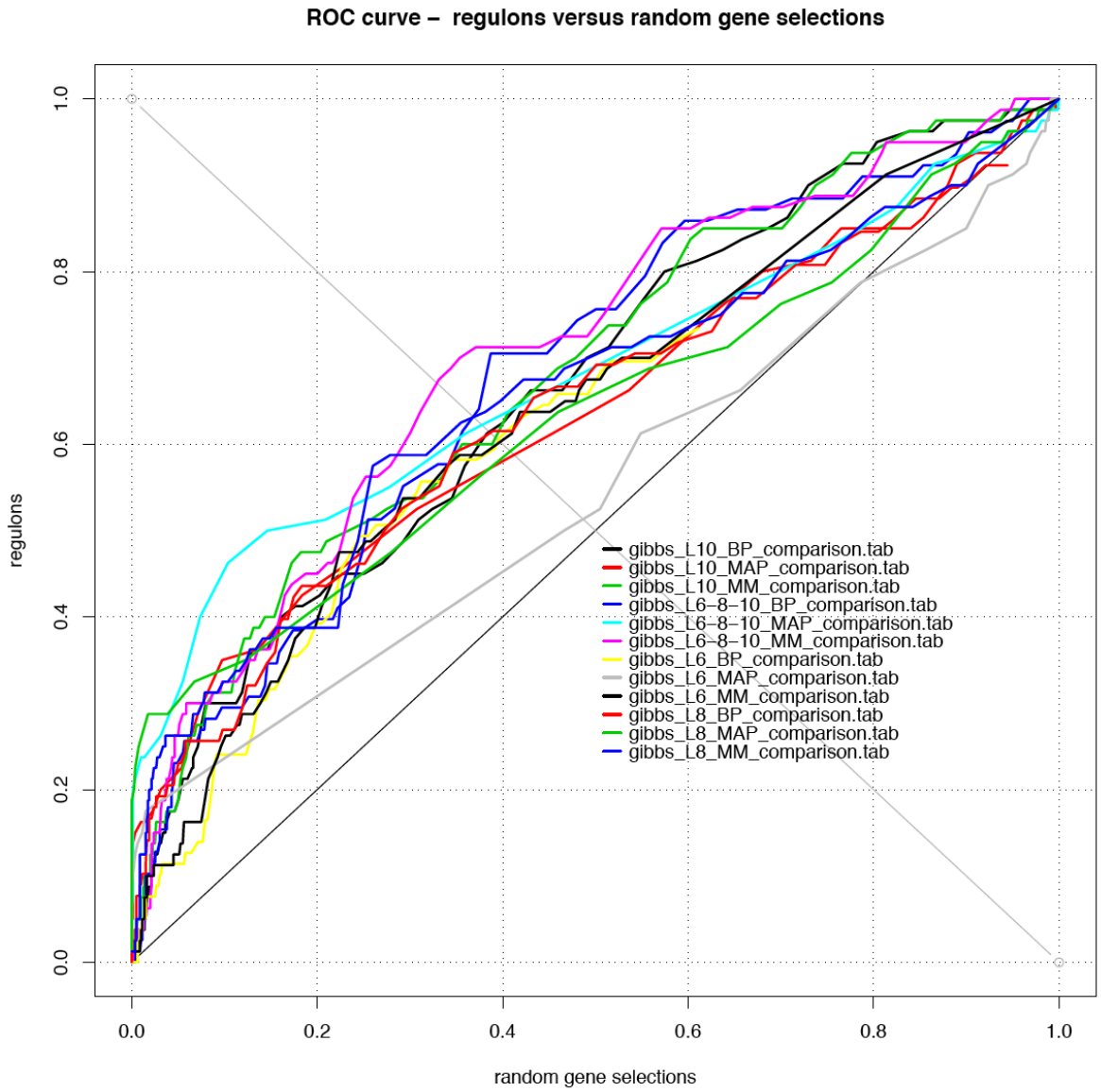
- ROC curves can be used to compare different scoring functions for a given program.
- For the ***MotifSampler*** (Thijs, 2000),
  - the Information Content (IC) is drastically better than
  - the Consensus Score (CS) or
  - Log Likelihood (LL).

MotifSampler: selection of optimal parameters



- Having selected the optimal scoring function, the ROC curves can be used to compare other parameter values.
- For the ***MotifSampler***, a matrix width of 10 gives weaker results than 6, 8, or the combination of 6,8 and 10.

Gibbs: selection of optimal scoring function



- For ***gibbs*** (Neuwald, 1995) returns quite poor results with yeast regulons.
- Note that this program was developed for the detection of protein motifs.
  - It is thus not optimal for DNA motifs.
- Subsequent versions of the gibbs sampler give better results.

Using ROC curves to compare methods

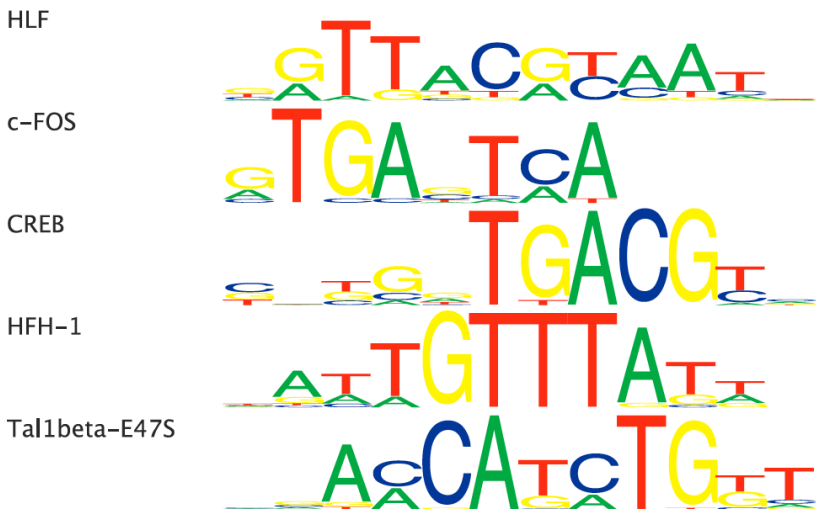
- The preceding slides should in no case be used to compare the different methods.
- Indeed, since I developed one of these methods, the comparison is unfair:
  - I may be (consciously or unconsciously) biased by my own interest.
  - Even assuming honesty, I am not so familiar with the parameters of the programs developed by other people than with my own programs.

Correctness of the discovered motifs

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

Example: biological sites implanted in foreign biological sequences

- Down et al. (2005). Nucleic Acids Res. 33(5):1445-1453. NestedMICA: sensitive inference of over-represented motifs in nucleic acid sequences.
- Motifs
  - Jaspar annotations for 4 human transcription factors (HLF,c-Fos, CREB, HFH-1)
- Sequences
  - Random selections of genes, 100 promoters per set.
  - For each factor, different sequences sizes are tested.
- Implanted sites
  - Zero or one occurrence per sequence (zoops).
  - One implant in 50% of the sequences.
- Pattern discovery software
  - NestedMICA (the new program presented in the article)
  - MEME (used with default parameters)



Example: biological sites implanted in foreign biological sequences

- Down et al. (2005). Nucleic Acids Res. 33(5):1445-1453.
- Questions
  - Which criterion was used here to say “yes” or “no” with matrices ?
    - Visual inspection ?
    - Quantitative criterion ?
  - How can we extend this to string-based pattern discovery ?

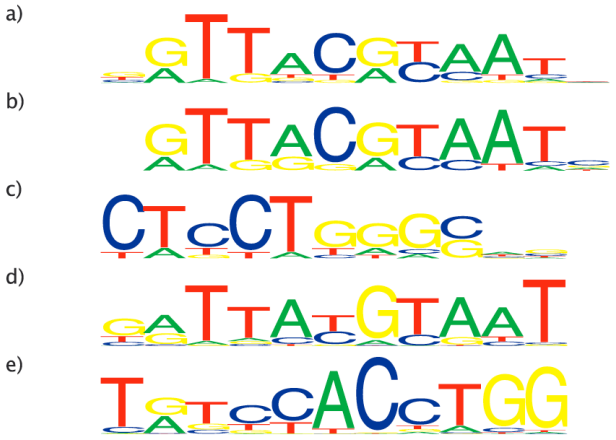


Figure 4. (**a**) The original HLF motif from JASPAR. (**b**) Results for searching for HLF in a set of 150 base sequences using MEME. (**c**) MEME with 200 base sequences. (**d**) NestedMICA with 600 base sequences. (**e**) NestedMICA with 700 base sequences.

Table 1. Discovery of the HLF motif from sets of 100 synthetic sequences of various lengths

Length	100	150	200	300	400	500	600	700
MEME	y	y	n	n	n	n	n	n
N’MICA	y	y	y	y	y	y	y	n

‘y’ indicates that the correct motif was found, and ‘n’ indicates failure.

Table 2. Discovery of the c-FOS motif from sets of 100 synthetic sequences of various lengths

Length	200	300	400	500	600
MEME	y	y	n	n	n
N’MICA	y	y	y	y	n

‘y’ indicates that the correct motif was found, and ‘n’ indicates failure.

Table 3. Discovery of the HFH-1 motif from sets of 100 synthetic sequences of various lengths

Length	800	1000	1200	1400	1600
MEME	y	y	y	n	n
N’MICA	y	y	y	n	n

‘y’ indicates that the correct motif was found, and ‘n’ indicates failure.

Oligo-analysis “yes” and “no” values

oligo-analysis results
Background model: Markov order of ordre 3, trained on the input sequence

HLF

Length	100	150	200	300	400	500	600	800	900	1000
Match with most signif pattern	y	y	y	y	y	n	y	y	y	y
Rank of first matching pattern	1	1	1	1	1	2	1	1	1	1

c-FOS **no annotated sites in Jaspar (only a matrix)**

HFH-1

Length	800	1000	1200	1400	1600	2000
Match with most signif pattern	y	y	y	n	y	n
Rank of first matching pattern	1	1	1	5	1	3

CREB

Length	100	200	300	400	500	600	800
Match with most signif pattern	y	y	y	y	n	n	n
Rank of first matching pattern	1	1	1	1	2	2	2

Table 1. Discovery of the HLF motif from sets of 100 synthetic sequences of various lengths

Length	100	150	200	300	400	500	600	700
MEME	y	y	n	n	n	n	n	n
N’MICA	y	y	y	y	y	y	y	n

‘y’ indicates that the correct motif was found, and ‘n’ indicates failure.

Table 2. Discovery of the c-FOS motif from sets of 100 synthetic sequences of various lengths

Length	200	300	400	500	600
MEME	y	y	n	n	n
N’MICA	y	y	y	y	n

‘y’ indicates that the correct motif was found, and ‘n’ indicates failure.

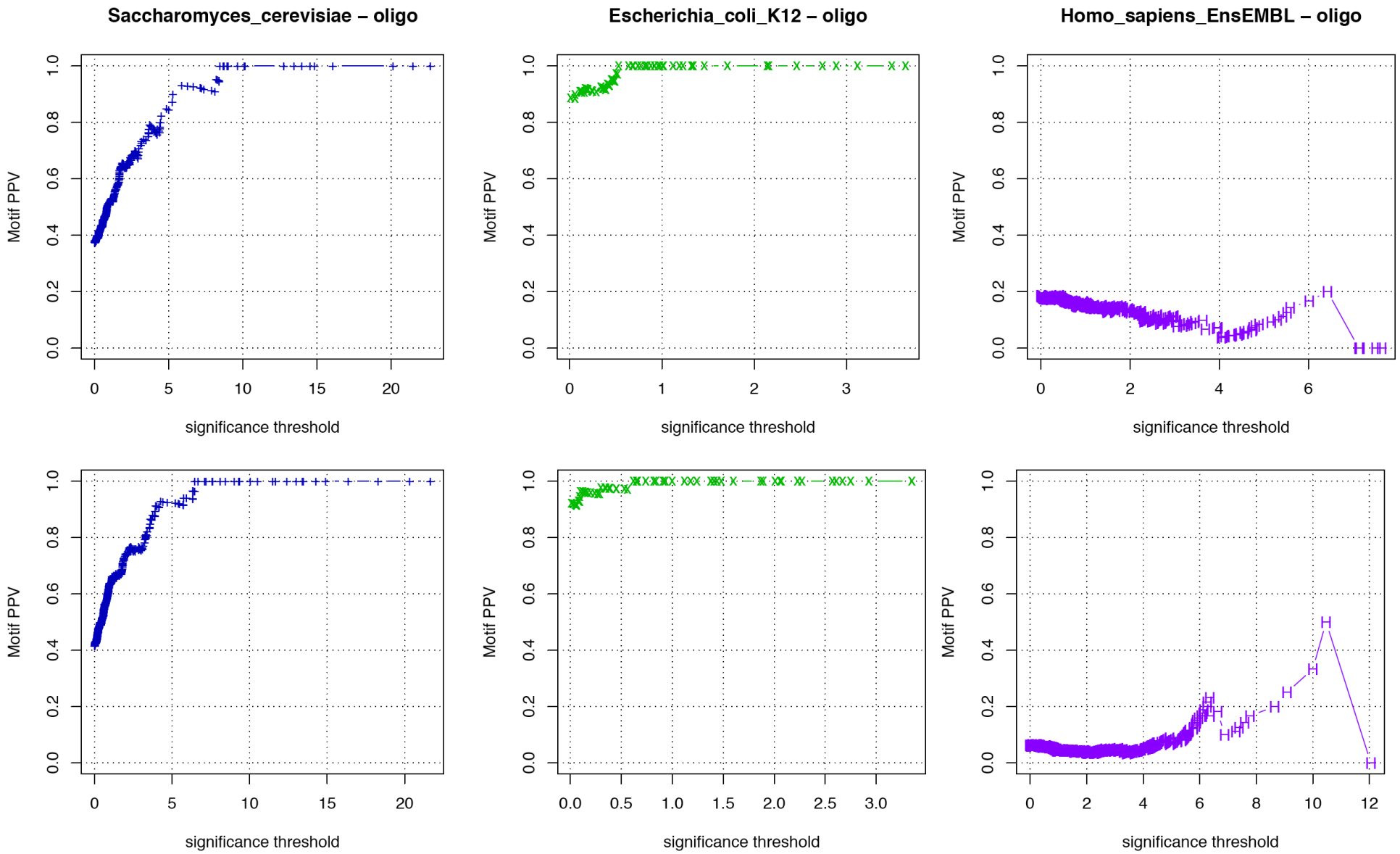
Table 3. Discovery of the HFH-1 motif from sets of 100 synthetic sequences of various lengths

Length	800	1000	1200	1400	1600
MEME	y	y	y	n	n
N’MICA	y	y	y	n	n

‘y’ indicates that the correct motif was found, and ‘n’ indicates failure.

- Analysis of Down’s data set with oligo-analysis
  - For all the sequence lengths, the motif is selected as significant
  - In most conditions, the most significant pattern matches some annotated site(s)
  - When this is not the case, the right motif comes very close to the first rank.
- Some comments on these results
  - The values “yes” and “no” are a bit rough, we could refine them.
  - The fact that the motifs corresponding to implanted sites are found so easily suggests that the testing set might be too “simple”, by comparison with real cases.

PPV versus significance (pooled motifs)



Figures from Olivier Sand, BioSapiens project

Accuracy

- There is always a trade between sensitivity PPV.
  - Stringent threshold on significance are expected to increase PPV at the cost of sensitivity.
  - Relaxing the threshold increases sensitivity at the cost of PPV.
- We thus need a statistics which captures both sensitivity and PPV: the accuracy.
- The literature contains different definitions of accuracy.

Arithmetic mean between Sensitivity and PPV

$$Acc_a = \left(Sn + PPV\right)/2$$

- The arithmetic mean can be misleading for trivial cases
  - Program predicting all possible motifs
    - $Sn \sim 1 \; ; PPV \sim 0 \Rightarrow Acc_a \sim 0.5$
  - Program predicting a single correct motif but missing all the other ones
    - $Sn \sim 0 \; ; PPV \sim 1 \Rightarrow Acc_a \sim 0.5$

Geometric mean between Sensitivity and PPV

$$Acc_g = \sqrt{Sn \cdot PPV}$$

- A more reliable statistics is Acc_g, the *geometric mean* between Sn and PPV
  - Program predicting all the motifs
    - $Sn \sim 1 \; ; PPV \sim 0 \Rightarrow Acc_a \sim 0.5$
  - Program predicting a single correct motif but missing all the other ones
    - $Sn \sim 0 \; ; PPV \sim 1 \Rightarrow Acc_a \sim 0.5$
  - In the cases where $Sn \sim PPV, Acc_a \sim Acc_g$

Sensitivity and PPV

- Arithmetic mean $Acc_a = (PPV + Sn)/2$
  - Perfect matches $Acc_a = (0.24 + 0.56)/2 = 0.40$
  - At most 1 subst $Acc_a = (0.81 + 1.00)/2 = 0.905$
- Geometric mean $Acc_g = sqrt(PPV*Sn)$
  - Perfect matches $Acc_g = sqrt(0.24 * 0.56) = 0.367$
  - At most 1 subst $Acc_g = sqrt(081*1.00) = 0.90$

	rank	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	Perfect matches	At most one substi
	sig	2.92	1.78	1.30	1.27	1.16	1.15	0.81	0.77	0.73	0.67	0.63	0.61	0.48	0.45	0.36	0.28	0.21	0.14	0.10	0.07	0.05		
	Annotated \ predicted	cgtaac	acgcaa	ccccag	caaagc	cggggc	aaggaa	ggtgac	aaacaa	ccccgc	cttggc	ctgtcc	tacgta	atgtaa	tacgca	cgttac	cttagc	ccaagg	aagaaa	cccagg	acttcag	cccgcc		
1	ctTGTTACGCAATCaagggc	6	6	4	4	4	4	4	5	3	4	3	5	4	6	5	4	5	4	4	4	3	3	7
2	caaaGATTACGTAATCgtgc	5	5	3	5	3	4	4	4	3	4	3	6	5	5	4	4	4	4	3	3	3	1	6
3	ccagGGTTACGCAACTcggc	6	6	4	4	4	4	5	4	4	5	3	5	4	6	5	4	5	4	5	5	4	3	10
4	cTATTACGCAATTctaaccgc	5	6	3	4	3	4	4	4	4	3	4	5	4	6	4	4	3	4	3	4	4	2	4
5	cttAGTTACGCAATAattgt	6	6	3	4	3	4	4	4	3	4	3	5	4	6	5	5	4	4	3	5	3	3	7
6	gcGATTGCGCAATAgatcgc	4	5	3	3	3	3	3	3	3	3	3	4	4	5	3	3	3	4	3	3	3	0	2
7	cGGTTGCATAATCaggcgcg	4	5	4	3	4	4	4	4	4	3	3	4	5	5	4	3	3	4	4	5	5	0	5
8	cacaaGATTACATCATAtac	4	4	4	3	2	4	4	4	2	4	3	4	6	3	4	3	4	4	4	3	2	1	1
9	cggcacTGTTACACAACCgt	5	5	3	4	4	3	4	5	3	3	4	4	5	5	5	3	3	3	3	4	3	0	6
10	ggagtaGGTTACATAAGTcg	5	4	3	3	3	4	5	4	3	3	3	5	6	4	5	4	3	4	3	4	3	1	5
11	tctagTGTTACGTGATGttg	6	5	4	3	3	4	4	4	3	3	3	5	5	4	5	4	3	4	3	3	3	1	5
12	gaggGGTTACATCAACcatt	5	4	4	3	4	4	5	4	5	3	3	4	6	3	5	3	3	4	3	3	4	1	5
13	ccgaattCGTTATGTAATGc	5	4	3	4	3	5	3	4	3	3	3	5	6	4	5	4	3	4	3	4	3	1	5
14	tgcaaTGTTCGGTAATAcgc	4	4	3	4	3	4	4	4	3	4	4	4	4	5	4	4	4	3	3	3	3	0	1
15	cGGAGACACCATTatgcaat	4	5	4	3	4	4	5	4	4	3	4	4	5	5	4	3	3	4	3	4	3	0	4
16	aacCATTACGTCTATtatgc	5	5	3	3	2	4	3	4	2	3	3	5	5	4	5	3	3	4	3	3	3	0	5
17	agtaataaCATGTCGCAGTG	4	4	4	4	3	4	4	4	4	4	4	3	5	4	4	3	3	4	4	4	3	0	1
18	acaaaggCGTGTTGCATCAc	4	5	3	5	3	4	4	4	4	4	3	3	5	4	4	4	5	4	4	4	5	0	5
	Perfect matches	4	4	0	0	0	0	0	0	0	0	0	1	4	4	0	0	0	0	0	0	0	17	
	At most one substitution	11	12	0	2	0	1	4	2	1	1	0	9	12	10	9	1	3	0	1	3	2		84

Motif accuracy

- Motifs discovered in promoters of regulons are compared to those annotated in reference database.
- Positive predictive value (discovered patterns)
  - $PPV = TPm/(TPm+FPm)$
  - Fraction of discovered patterns matching at least one annotated sites.
- Sensitivity (sites)
  - $Sn = TPs/(TPs+FNs)$
  - Fraction of annotated sites matched by at least one predicted motif.
- Motif accuracy (geometric)
  - $Acc.g = sqrt(Sn*PPV)$
  - It is important to use the geometric rather than arithmetic accuracy to avoid trivial behaviour
    - e.g. selecting all possible motifs
    - => Sn=1: PPV ~0 ;
      Acc.a ~ 0.5, whereas Acc.g ~ 0

	rank	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	Perfect matches	At most one substi
	sig	2.92	1.78	1.30	1.27	1.16	1.15	0.81	0.77	0.73	0.67	0.63	0.61	0.48	0.45	0.36	0.28	0.21	0.14	0.10	0.07	0.05		
	Annotated \ predicted	cgtaac	acgcaa	ccccag	caaagc	cggggc	aaggaa	ggtgac	aaacaa	ccccgc	cttggc	ctgtcc	tacgta	atgtaa	tacgca	cgttac	cttagc	ccaagg	aagaaa	cccagg	actcag	cccgcc		
1	ctTGTTACGCAATCaagggc	6	6	4	4	4	4	4	5	3	4	3	5	4	6	5	4	5	4	4	4	3	3	7
2	caaaGATTACGTAATCgtgc	5	5	3	5	3	4	4	4	3	4	3	6	5	5	4	4	4	4	3	3	3	1	6
3	ccagGGTTACGCAACTcggc	6	6	4	4	4	4	5	4	4	5	3	5	4	6	5	4	5	4	5	5	4	3	10
4	cTATTACGCAATTctaaccgc	5	6	3	4	3	4	4	4	4	3	4	5	4	6	4	4	3	4	3	4	4	2	4
5	cttAGTTACGCAATAattgt	6	6	3	4	3	4	4	4	3	4	3	5	4	6	5	5	4	4	3	5	3	3	7
6	gcGATTGCGCAATAgatcgc	4	5	3	3	3	3	3	3	3	3	3	4	4	5	3	3	3	4	3	3	3	0	2
7	cGGTTGCATAATCaggcgcg	4	5	4	3	4	4	4	4	4	3	3	4	5	5	4	3	3	4	4	5	5	0	5
8	cacaaGATTACATCATAtac	4	4	4	3	2	4	4	4	2	4	3	4	6	3	4	3	4	4	4	3	2	1	1
9	cggcacTGTTACACAACCgt	5	5	3	4	4	3	4	5	3	3	4	4	5	5	5	3	3	3	3	4	3	0	6
10	ggagtaGGTTACATAAGTcg	5	4	3	3	3	4	5	4	3	3	3	5	6	4	5	4	3	4	3	4	3	1	5
11	tctagTGTTACGTGATGttg	6	5	4	3	3	4	4	4	3	3	3	5	5	4	5	4	3	4	3	3	3	1	5
12	gaggGGTTACATCAACcatt	5	4	4	3	4	4	5	4	5	3	3	4	6	3	5	3	3	4	3	3	4	1	5
13	ccgaattCGTTATGTAATGc	5	4	3	4	3	5	3	4	3	3	3	5	6	4	5	4	3	4	3	4	3	1	5
14	tgcaaTGTTCGGTAATAcgc	4	4	3	4	3	4	4	4	3	4	4	4	4	5	4	4	4	3	3	3	3	0	1
15	cGGAGACACCATTatgcaat	4	5	4	3	4	4	5	4	4	3	4	4	5	5	4	3	3	4	3	4	3	0	4
16	aacCATTACGTCTATtatgc	5	5	3	3	2	4	3	4	2	3	3	5	5	4	5	3	3	4	3	3	3	0	5
17	agtaataaCATGTCGCAGTG	4	4	4	4	3	4	4	4	4	4	4	3	5	4	4	3	3	4	4	4	3	0	1
18	acaaaggCGTGTTGCATCAc	4	5	3	5	3	4	4	4	4	4	3	3	5	4	4	4	5	4	4	4	5	0	5
	Perfect matches	4	4	0	0	0	0	0	0	0	0	0	1	4	4	0	0	0	0	0	0	0	17	
	At most one substitution	11	12	0	2	0	1	4	2	1	1	0	9	12	10	9	1	3	0	1	3	2		84

In the example above

Perfect matches (black)

Sn = 10/18 = 56%

PPV = 5/21 = 24%

Acc_a *= (0.24 + 0.56)/2 = 40%*

Acc_g *= sqrt(0.24 * 0.56)= 37%*

At most 1 substitution (gray)

Sn = 18/18 = 100%

PPV = 17/21= 81%

Acc_a *= (0.81 + 1.00)/2 = 91%*

Acc_g *= sqrt(081*1.00) = 90%*

O.Sand & J.van Helden, in prep.

Validation with 53 yeast regulons - perfect matches

Sand & van Helden, BioSapiens project

row.ID	nb.targets	nb.sites	max.sig.oligo	nb.oligo	sites.matched.oligo	TP.oligo	PPV.oligo	Sn.oligo	Acc.g.oligo	max.sig.dyad	nb.dyad	sites.matched.dyad	TP.dyad	PPV.dyad	Sn.dyad	Ac
MSN4	58	3	**13.45**	26	1	8	0.308	0.333	0.320	**11.66**	61	1	7	0.115	0.333	0.196
MSN2	56	1	**14.82**	30	1	8	0.267	1.000	0.516	**13.01**	58	1	6	0.103	1.000	0.322
ZAP1	52	8	0.57	1	3	1	1.000	0.375	0.612	**13.47**	5	8	5	1.000	1.000	1.000
GCN4	40	18	**22.64**	9	11	6	0.667	0.611	0.638	**21.65**	8	8	6	0.750	0.444	0.577
TEC1	38	0	0.27	1	NA	0	NA	NA	NA	0.16	2	NA	0	NA	NA	NA
RAP1	32	20	**1.67**	2	7	1	0.500	0.350	0.418	0.16	1	0	0	0.000	0.000	0.000
GLN3	31	2	**21.47**	9	1	1	0.111	0.500	0.236	**20.32**	6	1	1	0.167	0.500	0.289
YAP1	31	2	**1.40**	3	0	0	0.000	0.000	0.000	0.78	1	0	0	0.000	0.000	0.000
MIG1	26	15	**6.27**	23	12	7	0.304	0.800	0.493	**4.68**	34	14	12	0.353	0.933	0.574
UME6	26	4	**2.49**	6	2	6	1.000	0.500	0.707	**1.15**	4	2	3	0.750	0.500	0.612
RLM1	25	0	**2.36**	2	NA	0	NA	NA	NA	0.55	1	NA	0	NA	NA	NA
OAF1	24	0	**1.32**	4	NA	0	NA	NA	NA	**4.89**	6	NA	0	NA	NA	NA
HSF1	21	4	**7.12**	5	2	2	0.400	0.500	0.447	**5.70**	7	3	6	0.857	0.750	0.802
PHO2	21	5	**14.53**	6	1	1	0.167	0.200	0.183	**13.40**	4	1	1	0.250	0.200	0.224
DAL80	19	5	**20.13**	13	2	4	0.308	0.400	0.351	**18.29**	10	2	3	0.300	0.400	0.346
INO2	19	3	**8.40**	7	2	3	0.429	0.667	0.535	**6.32**	5	3	5	1.000	1.000	1.000
INO4	19	1	**8.40**	7	0	0	0.000	0.000	0.000	**6.32**	5	1	2	0.400	1.000	0.632
PIP2	19	1	0.17	1	0	0	0.000	0.000	0.000	**5.73**	5	1	3	0.600	1.000	0.775
REB1	19	10	**2.41**	4	7	3	0.750	0.700	0.725	**1.26**	1	5	1	1.000	0.500	0.707
GCR1	18	11	0.77	2	0	0	0.000	0.000	0.000	**1.81**	2	0	0	0.000	0.000	0.000
BAS1	17	2	**16.07**	7	2	3	0.429	1.000	0.655	**14.89**	5	2	2	0.400	1.000	0.632
CBF1	16	1	**9.65**	6	1	1	0.167	1.000	0.408	**8.10**	5	1	1	0.200	1.000	0.447
PDR1	16	9	**13.97**	13	9	8	0.615	1.000	0.784	**16.35**	18	9	18	1.000	1.000	1.000
HAP3	15	2	0.64	3	0	0	0.000	0.000	0.000	NA	0	0	0	0.000	0.000	0.000
MIG2	15	0	**3.74**	12	NA	0	NA	NA	NA	**2.24**	10	NA	0	NA	NA	NA
MOT3	15	2	**1.75**	8	0	0	0.000	0.000	0.000	**3.58**	9	0	0	0.000	0.000	0.000
ROX1	15	25	0.99	2	20	2	1.000	0.800	0.894	0.10	2	0	0	0.000	0.000	0.000
HAP2	14	2	0.42	2	0	0	0.000	0.000	0.000	NA	0	0	0	0.000	0.000	0.000
HAP4	14	1	0.29	3	0	0	0.000	0.000	0.000	NA	0	0	0	0.000	0.000	0.000
MCM1	14	24	NA	0	NA	0	NA	NA	NA	0.67	1	0	0	0.000	0.000	0.000
STE12	13	5	**3.95**	3	5	2	0.667	1.000	0.816	**2.59**	3	5	2	0.667	1.000	0.816
RTG1	12	1	**2.49**	3	NA	0	0.000	NA	NA	**1.56**	6	NA	0	0.000	NA	NA
ADR1	11	4	**3.71**	7	3	2	0.286	0.750	0.463	**1.80**	2	2	2	1.000	0.500	0.707
GCR2	11	0	**1.29**	1	NA	0	NA	NA	NA	0.09	2	NA	0	NA	NA	NA
NDT80	11	1	**2.41**	5	0	0	0.000	0.000	0.000	NA	0	0	0	0.000	0.000	0.000

Validation with 38 regulons from E.coli - perfect matches

Sand & van Helden
BioSapiens project

row.ID	nb.targets	nb.sites	max.sig.oligo	nb.oligo	sites.matched.oligo	TP.oligo	PPV.oligo	Sn.oligo	Acc.g.oligo	max.sig.dyad	nb.dyad	sites.matched.dyad	TP.dyad	PPV.dyad	Sn.dyad	Ac
CRP	104	137	**1.71**	4	67	4	1.000	0.489	0.699	**2.93**	7	93	7	1.000	0.679	0.824
IHF	38	50	0.74	4	16	4	1.000	0.320	0.566	**1.44**	6	16	5	0.833	0.320	0.516
FNR	35	47	0.36	2	19	2	1.000	0.404	0.636	**3.35**	3	23	3	1.000	0.489	0.700
ArcA	25	35	NA	0	0	0	0.000	0.000	0.000	NA	0	0	0	0.000	0.000	0.000
Lrp	22	77	0.49	4	34	4	1.000	0.442	0.664	**2.07**	5	21	5	1.000	0.273	0.522
FIS	17	58	0.52	1	0	0	0.000	0.000	0.000	NA	0	0	0	0.000	0.000	0.000
PurR	17	16	**1.19**	6	8	5	0.833	0.500	0.645	NA	0	0	0	0.000	0.000	0.000
NarL	15	42	**1.46**	5	15	5	1.000	0.357	0.598	0.12	1	6	1	1.000	0.143	0.378
Hns	13	2	0.96	2	1	1	0.500	0.500	0.500	0.09	1	0	0	0.000	0.000	0.000
FruR	12	10	**2.89**	5	10	5	1.000	1.000	1.000	**1.41**	4	8	4	1.000	0.800	0.894
Fur	11	13	0.77	1	6	1	1.000	0.462	0.679	NA	0	0	0	0.000	0.000	0.000
LexA	10	11	**1.12**	3	7	3	1.000	0.636	0.798	**1.38**	2	11	2	1.000	1.000	1.000
SoxS	10	9	0.49	1	0	0	0.000	0.000	0.000	NA	0	0	0	0.000	0.000	0.000
MarA	8	7	0.29	1	0	0	0.000	0.000	0.000	0.45	2	2	1	0.500	0.286	0.378
TyrR	8	19	0.98	2	12	2	1.000	0.632	0.795	**2.23**	3	9	3	1.000	0.474	0.688
ArgR	7	12	NA	0	0	0	0.000	0.000	0.000	0.24	1	3	1	1.000	0.250	0.500
CysB	7	7	NA	0	0	0	0.000	0.000	0.000	NA	0	0	0	0.000	0.000	0.000
CytR	7	12	0.67	4	NA	0	0.000	NA	NA	NA	0	0	0	0.000	0.000	0.000
FlhD	7	4	0.12	2	1	2	1.000	0.250	0.500	0.62	1	1	1	1.000	0.250	0.500
PhoB	7	9	**3.64**	4	5	4	1.000	0.556	0.745	**2.67**	4	6	4	1.000	0.667	0.816
ModE	6	6	0.06	1	0	0	0.000	0.000	0.000	NA	0	0	0	0.000	0.000	0.000
OmpR	6	16	**1.23**	2	4	2	1.000	0.250	0.500	0.45	1	2	1	1.000	0.125	0.354
Rob	6	5	NA	0	0	0	0.000	0.000	0.000	NA	0	0	0	0.000	0.000	0.000
AraC	5	15	**3.12**	2	11	2	1.000	0.733	0.856	**1.89**	1	11	1	1.000	0.733	0.856
FhlA	5	3	**1.98**	3	NA	0	0.000	NA	NA	0.72	2	NA	0	0.000	NA	NA
OxyR	5	5	NA	0	0	0	0.000	0.000	0.000	0.35	2	NA	0	0.000	NA	NA
TrpR	5	5	**3.49**	4	5	4	1.000	1.000	1.000	**2.06**	5	5	5	1.000	1.000	1.000
CpxR	4	4	NA	0	0	0	0.000	0.000	0.000	NA	0	0	0	0.000	0.000	0.000
DnaA	4	11	NA	0	0	0	0.000	0.000	0.000	0.09	1	11	1	1.000	1.000	1.000
FadR	4	6	0.46	1	1	1	1.000	0.167	0.408	NA	0	0	0	0.000	0.000	0.000
GlpR	4	19	**1.32**	1	5	1	1.000	0.263	0.513	0.15	2	7	2	1.000	0.368	0.607
MalT	4	9	0.68	3	9	3	1.000	1.000	1.000	**2.57**	1	5	1	1.000	0.556	0.745
MetJ	4	7	**1.34**	1	6	1	1.000	0.857	0.926	0.10	1	6	1	1.000	0.857	0.926
Mlc	4	6	NA	0	0	0	0.000	0.000	0.000	0.09	1	0	0	0.000	0.000	0.000
NagC	4	8	0.18	1	4	1	1.000	0.500	0.707	NA	0	0	0	0.000	0.000	0.000
NtrC	4	11	**2.14**	2	9	2	1.000	0.818	0.905	**1.12**	4	10	4	1.000	0.909	0.953
DcuR	3	0	**1.53**	2	NA	0	NA	NA	NA	0.12	1	NA	0	NA	NA	NA
DeoR	3	7	0.92	1	1	1	1.000	0.143	0.378	0.14	1	1	1	1.000	0.143	0.378
DsdC	3	0	NA	0	NA	0	NA	NA	NA	NA	0	NA	0	NA	NA	NA
GadW	3	0	**1.13**	5	NA	0	NA	NA	NA	**1.03**	3	NA	0	NA	NA	NA
GntR	3	4	0.88	1	3	1	1.000	0.750	0.866	**2.27**	4	4	4	1.000	1.000	1.000
MetR	3	3	0.43	1	0	0	0.000	0.000	0.000	NA	0	0	0	0.000	0.000	0.000
RcsB	3	3	0.34	1	1	1	1.000	0.333	0.577	0.56	1	0	0	0.000	0.000	0.000

Validation with 93 human regulons - perfect matches

row.ID	nb.targets	nb.sites	max.sig.oligo	nb.oligo	sites.matched.oligo	TP.oligo	PPV.oligo	Sn.oligo	Acc.a.oligo	Acc.g.oligo	max.sig.dyad	nb.dyad	sites.matched.dyad	TP.dyad	PPV.dyad	Sn.dyad	Acc.a.dyad	Ac
T00759_Sp1	76	186	**3.63**	10	121	10	1.000	0.651	0.825	0.807	NA	0	0	0	0.000	0.000	0.000	0.000
T00671_p53	28	36	**2.00**	22	10	15	0.682	0.278	0.480	0.435	NA	0	0	0	0.000	0.000	0.000	0.000
T00133_c-Jun	25	32	0.28	2	0	0	0.000	0.000	0.000	0.000	NA	0	0	0	0.000	0.000	0.000	0.000
T00590_NF-ka	19	25	0.46	5	0	0	0.000	0.000	0.000	0.000	0.17	1	0	0	0.000	0.000	0.000	0.000
T00035_AP-2a	17	29	**2.98**	8	6	4	0.500	0.207	0.353	0.322	0.80	3	6	3	1.000	0.207	0.603	0.455
T00167_ATF-2	17	18	0.71	3	1	1	0.333	0.056	0.194	0.136	NA	0	0	0	0.000	0.000	0.000	0.000
T00163_CREB	16	26	**2.01**	4	1	1	0.250	0.038	0.144	0.098	0.31	1	0	0	0.000	0.000	0.000	0.000
T00581_C.EBF	16	0	**1.37**	2	NA	0	NA	NA	NA	NA	NA	0	NA	0	NA	NA	NA	NA
T00368_HNF-1	15	21	0.67	4	1	1	0.250	0.048	0.149	0.109	0.35	1	2	1	1.000	0.095	0.548	0.309
T00261_ER-al	13	21	**2.29**	2	3	2	1.000	0.143	0.571	0.378	NA	0	0	0	0.000	0.000	0.000	0.000
T00594_RelA	13	16	0.32	2	0	0	0.000	0.000	0.000	0.000	NA	0	0	0	0.000	0.000	0.000	0.000
T00593_NF-ka	12	18	0.29	2	2	2	1.000	0.111	0.556	0.333	NA	0	0	0	0.000	0.000	0.000	0.000
T00874_USF1	12	15	**1.13**	7	1	1	0.143	0.067	0.105	0.098	NA	0	0	0	0.000	0.000	0.000	0.000
T01609_HIF-1	12	18	**1.41**	14	5	4	0.286	0.278	0.282	0.282	NA	0	0	0	0.000	0.000	0.000	0.000
T02758_HNF-4	12	14	0.67	6	6	4	0.667	0.429	0.548	0.535	NA	0	0	0	0.000	0.000	0.000	0.000
T00105_C.EBF	11	0	0.54	4	NA	0	NA	NA	NA	NA	NA	0	NA	0	NA	NA	NA	NA
T00123_c-Fos	10	13	**2.43**	1	0	0	0.000	0.000	0.000	0.000	**1.07**	2	0	0	0.000	0.000	0.000	0.000
T00140_c-Myc	10	17	0.39	2	0	0	0.000	0.000	0.000	0.000	NA	0	0	0	0.000	0.000	0.000	0.000
T00423_IRF-1	10	21	**1.60**	7	0	0	0.000	0.000	0.000	0.000	0.98	2	2	1	0.500	0.095	0.298	0.218
T00112_c-Ets-	9	16	0.52	1	1	1	1.000	0.063	0.531	0.250	NA	0	0	0	0.000	0.000	0.000	0.000
T00641_POU2	9	23	0.20	1	0	0	0.000	0.000	0.000	0.000	**1.31**	2	0	0	0.000	0.000	0.000	0.000
T01345_RXR-a	9	11	**2.55**	11	4	5	0.455	0.364	0.409	0.407	**1.04**	8	4	4	0.500	0.364	0.432	0.426
T00221_E2F	8	17	0.20	1	0	0	0.000	0.000	0.000	0.000	NA	0	0	0	0.000	0.000	0.000	0.000
T00113_c-Ets-	7	10	**2.25**	6	2	2	0.333	0.200	0.267	0.258	0.99	2	0	0	0.000	0.000	0.000	0.000
T00149_COUP	7	12	**1.52**	5	3	2	0.400	0.250	0.325	0.316	0.30	4	0	0	0.000	0.000	0.000	0.000
T00306_GATA	7	50	NA	0	0	0	0.000	0.000	0.000	0.000	NA	0	0	0	0.000	0.000	0.000	0.000
T00764_SRF	7	13	**7.64**	75	2	4	0.053	0.154	0.104	0.091	**12.05**	503	4	22	0.044	0.308	0.176	0.116
T00915_YY1	7	11	0.54	1	0	0	0.000	0.000	0.000	0.000	NA	0	0	0	0.000	0.000	0.000	0.000
T01553_MITF	7	11	0.16	3	0	0	0.000	0.000	0.000	0.000	0.37	2	0	0	0.000	0.000	0.000	0.000
T01950_HNF-1	7	10	**2.26**	7	2	2	0.286	0.200	0.243	0.239	0.83	12	6	3	0.250	0.600	0.425	0.387
T03828_HNF-4	7	11	NA	0	0	0	0.000	0.000	0.000	0.000	NA	0	0	0	0.000	0.000	0.000	0.000
T04759_STAT	7	8	0.86	2	0	0	0.000	0.000	0.000	0.000	0.30	1	0	0	0.000	0.000	0.000	0.000
T00045_COUP	6	8	**1.94**	10	0	0	0.000	0.000	0.000	0.000	0.57	12	0	0	0.000	0.000	0.000	0.000
T00241_Egr-1	6	9	0.54	2	0	0	0.000	0.000	0.000	0.000	0.34	2	1	1	0.500	0.111	0.306	0.236
T00250_Elk-1	6	10	**7.52**	52	3	4	0.077	0.300	0.188	0.152	**10.00**	394	6	17	0.043	0.600	0.322	0.161
T01542_E2F-1	6	12	**3.04**	6	2	5	0.833	0.167	0.500	0.373	0.66	2	0	0	0.000	0.000	0.000	0.000
T01945_NF-AT	6	7	0.25	1	0	0	0.000	0.000	0.000	0.000	NA	0	0	0	0.000	0.000	0.000	0.000
T01951_HNF-1	6	8	0.46	2	1	1	0.500	0.125	0.313	0.250	NA	0	0	0	0.000	0.000	0.000	0.000
T01978_JunD	6	7	0.54	3	1	1	0.333	0.143	0.238	0.218	0.07	1	0	0	0.000	0.000	0.000	0.000
T02338_Sp3	6	13	**1.28**	9	9	5	0.556	0.692	0.624	0.620	NA	0	0	0	0.000	0.000	0.000	0.000
T05324_LXR-a	6	0	**3.29**	5	NA	0	NA	NA	NA	NA	NA	0	NA	0	NA	NA	NA	NA
T00040_AR	5	17	0.09	1	1	1	1.000	0.059	0.529	0.243	NA	0	0	0	0.000	0.000	0.000	0.000
T00168_c-Rel	5	8	NA	0	0	0	0.000	0.000	0.000	0.000	NA	0	0	0	0.000	0.000	0.000	0.000

Synthesis of the (preliminary) results

- Note
  - For human regulons, the analysis was performed with 2kb upstream sequences, and with the default probabilistic model (binomial distribution)
  - We are currently working on alternative models to improve the accuracy of predictions in human.

Organism	Program	PPV	Sn	Acc_a	Acc_g
Escherichia coli K12	*oligo-analysis*	0.804	0.530	0.688	0.656
Saccharomyces cerevisiae	*oligo-analysis*	0.390	0.489	0.454	0.428
Homo sapiens	*oligo-analysis*	0.200	0.092	0.153	0.129

Summary

- Separation of pattern discovery and pattern matching problems.
- Importance of the negative control: random selections of genes.
- Comparison at the level of significance.
  - ROC curves
- Comparison at the level of motif accuracy.
  - Matching of

Perspectives

- Human regulons : need for improvement.
  - Test different upstream lengths.
  - Test different statistical models.
- Comparisons with other programs.
  - Define a fair comparison procedure (CASP-like)
    - Developers should be involved in the assessment.
    - Evaluation should be performed by an external committee.
  - Compare the result of each program with annotated binding sites
  - Compare results returned by the different programs.
  - Test “consensus strategies” : predict with different programs, and extract the most robust predictions.