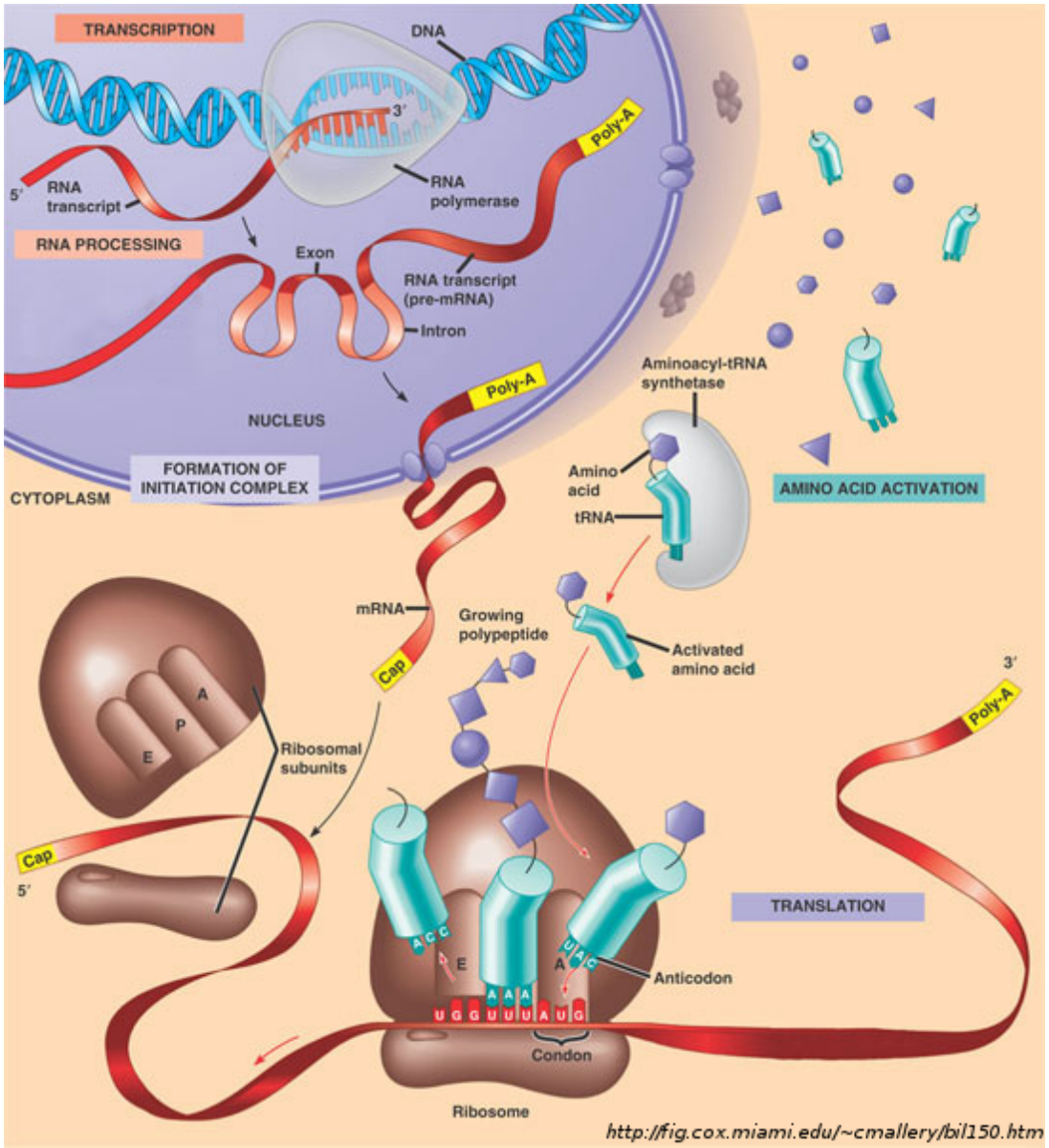
Discovery of regulatory motifs in *Saccharomyces cerevisiae* as a first step of understanding the gene regulation of baker's yeast

Hedi Peterson

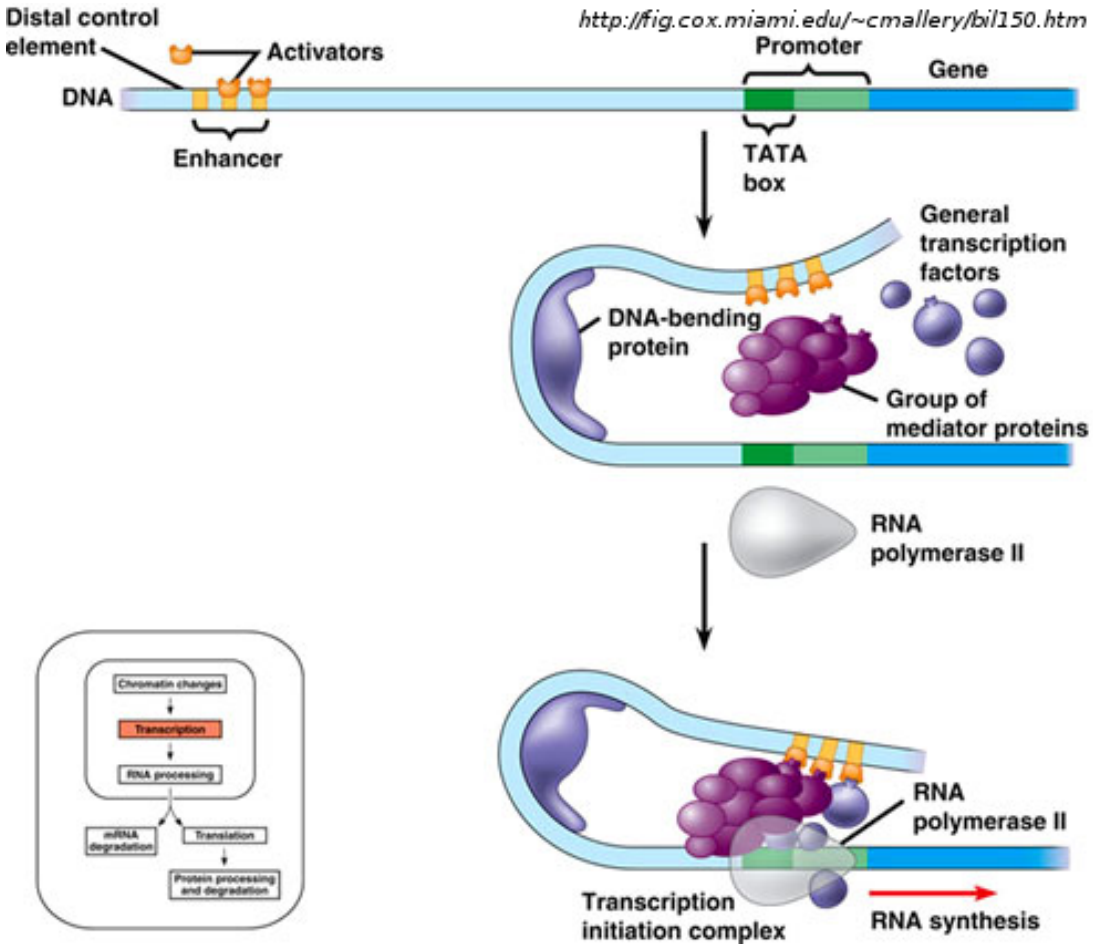
Institute of Molecular and Cell Biology
University of Tartu

29.09.2006
Voore

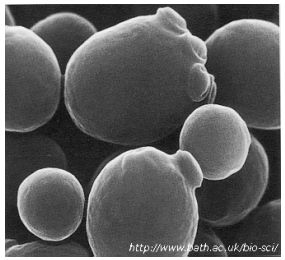
Biology: DNA <-> RNA -> Protein



Transcription (DNA -> RNA)



Background



Saccharomyces cerevisiae

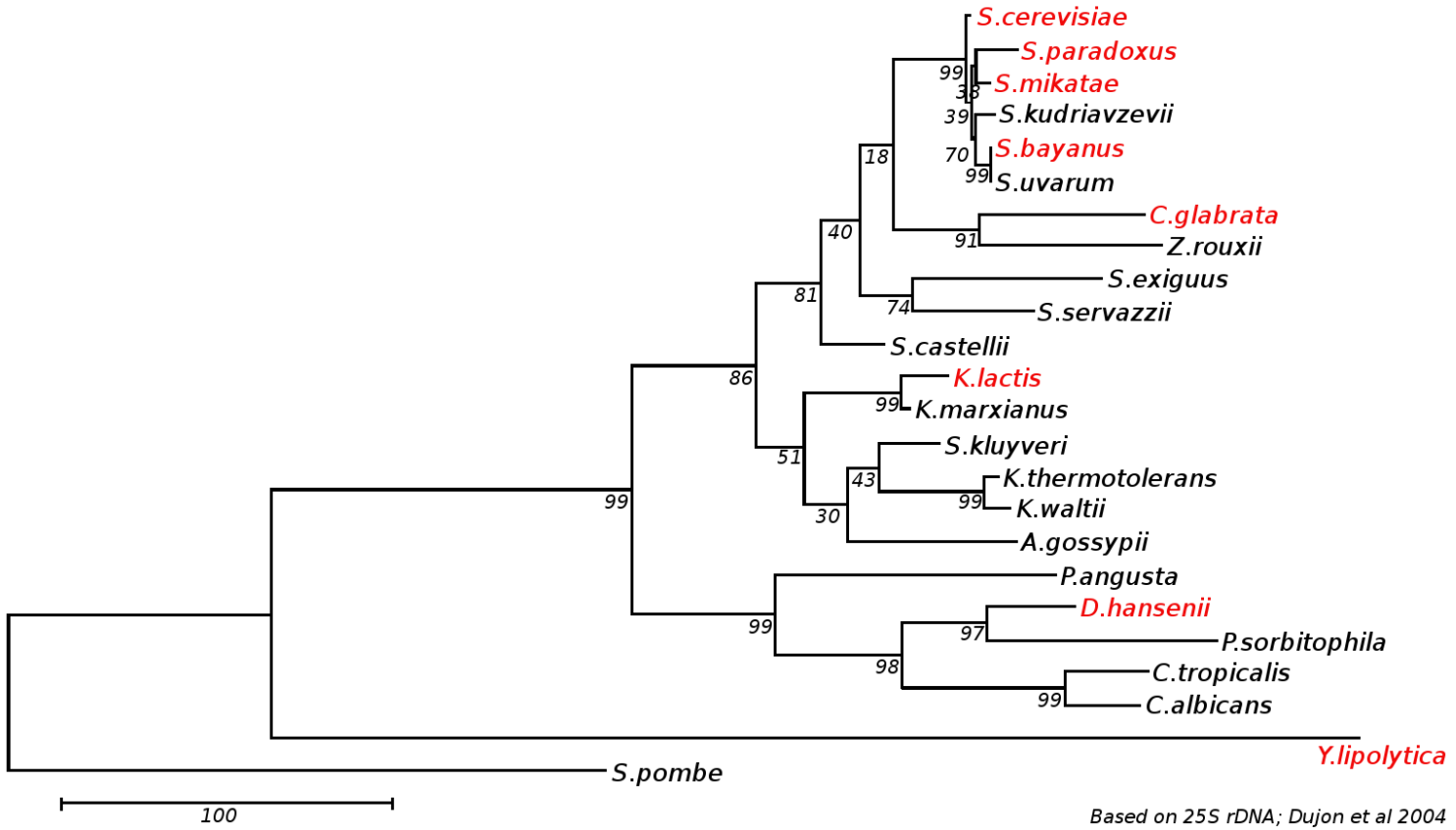
- 6604 predicted open reading frames
- 4403 described, 1377 undescribed and 824 dubious (SGD)
- Approximately one third of genes have no funcionality or no regulator attached

Aim

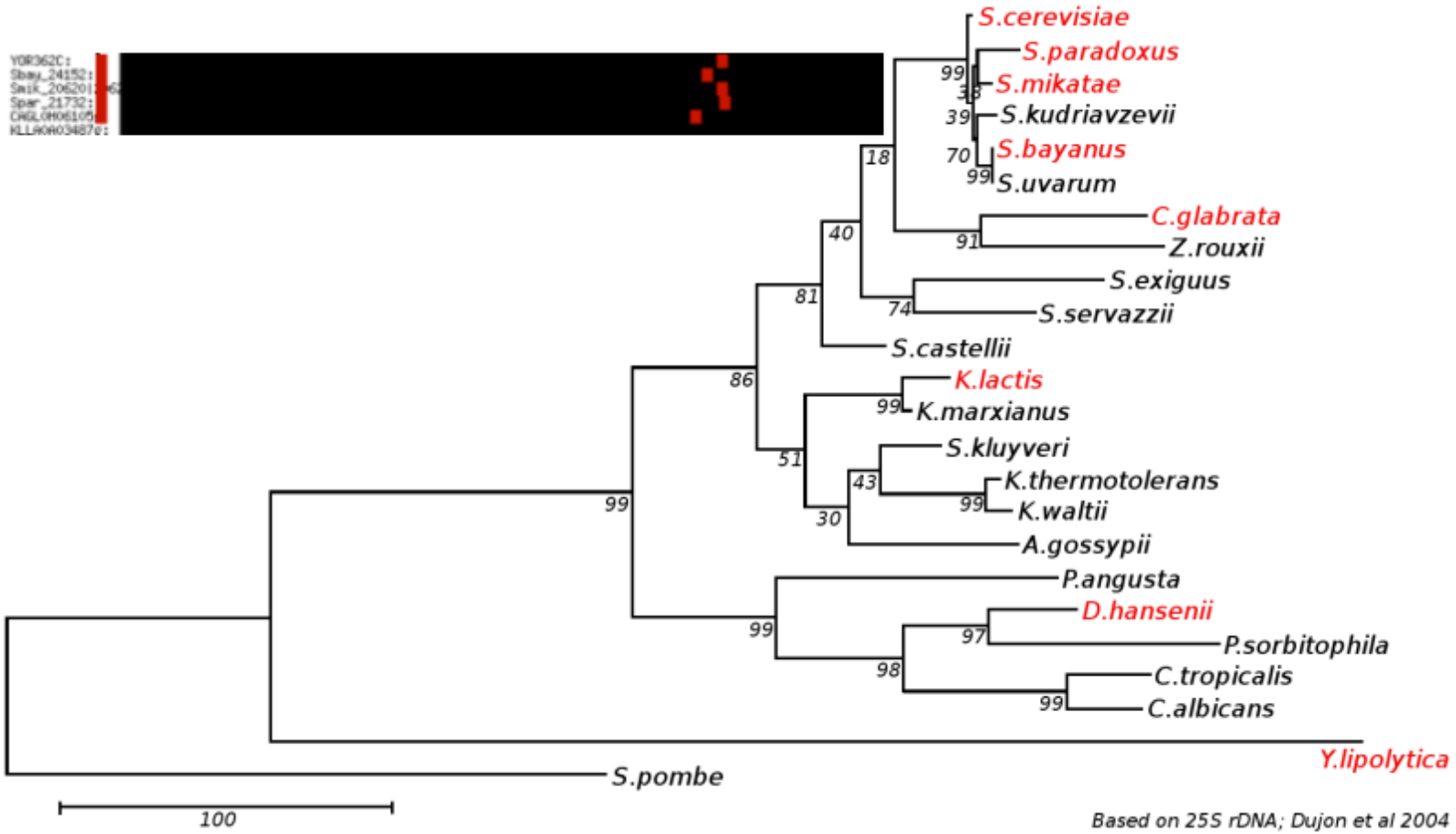
- Predict regulatory motifs
- Using different gene groups and regulatory motifs, find similar genes

Species used

- *S.cerevisiae*, *S.bayanus*, *S.paradoxus*, *S.mikatae*, *C.glabrata*, *K.lactis*, *D.hansenii*, *Y.lipolytica*



Species used



Regulatory motifs

- Transcription factor binding sites
  - discrete motifs
  - alphabet {A,C,G,T}
  - length: 6-20 bp
  - location: 50-500 bp from ATG
- 600bp as promoter length

Used datasets

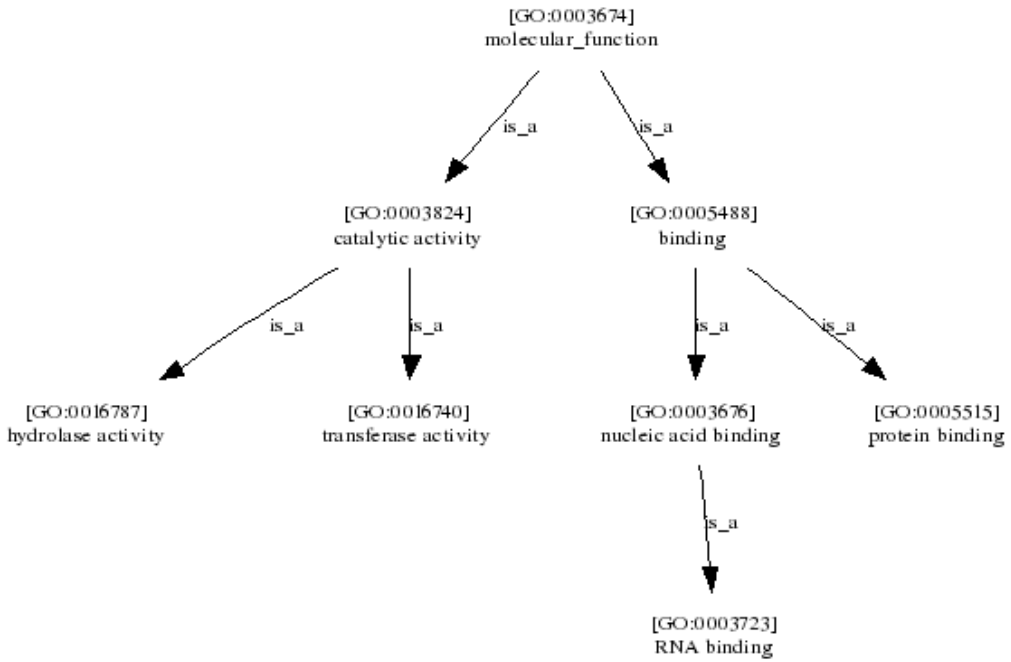
- Experimentally validated transcription factor binding sites
- Gene Ontology (GO)
- Protein-protein interactions (PPI)

TF target genes

- Experimantely validated
- 143 different trancription factors
- 4248 targets
- On average 84.8 targets per TF
- On average 2.9 TF per gene

Gene Ontology

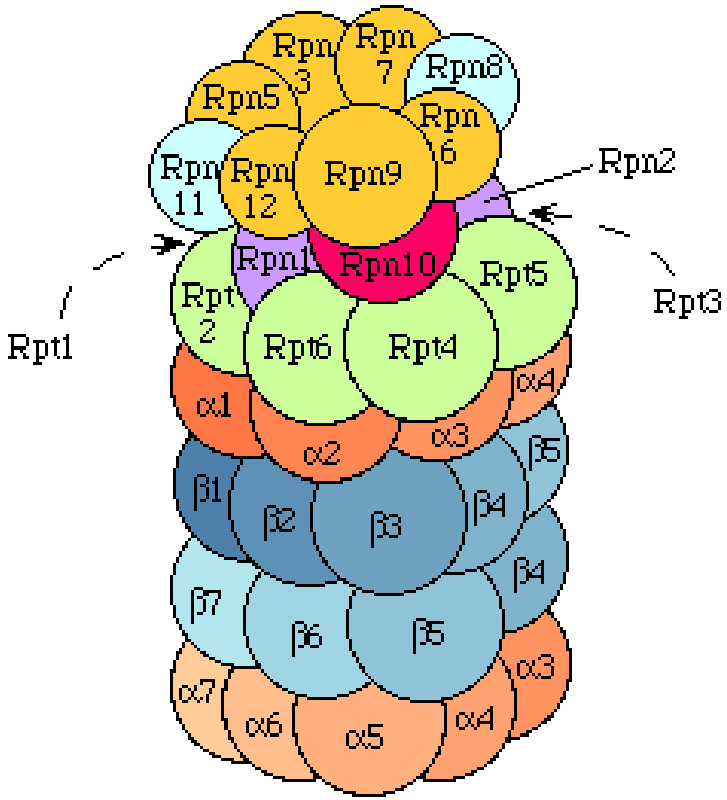
- All annotations are equal
- 3977 different GO groups
- size 1 to 658 genes in a GO group



Protein-protein interactions

Three experimental datasets

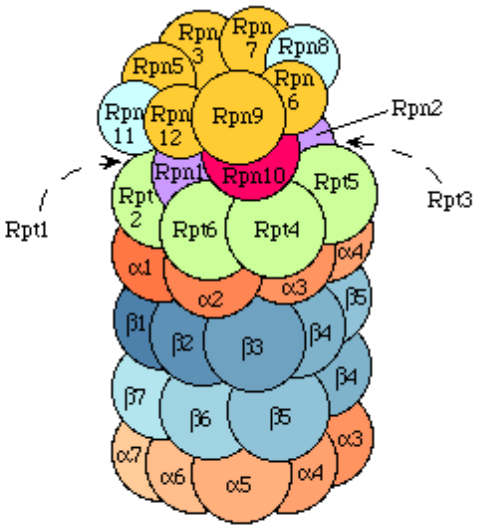
- Kemmeren *et al* 1309
- Krogan *et al* 2186
- Gavin *et al* 1709
- In total 3334 different genes



26S Proteasome (Saccharomyces cerevisiae)

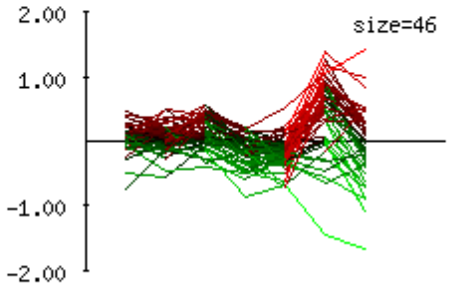
KEGG

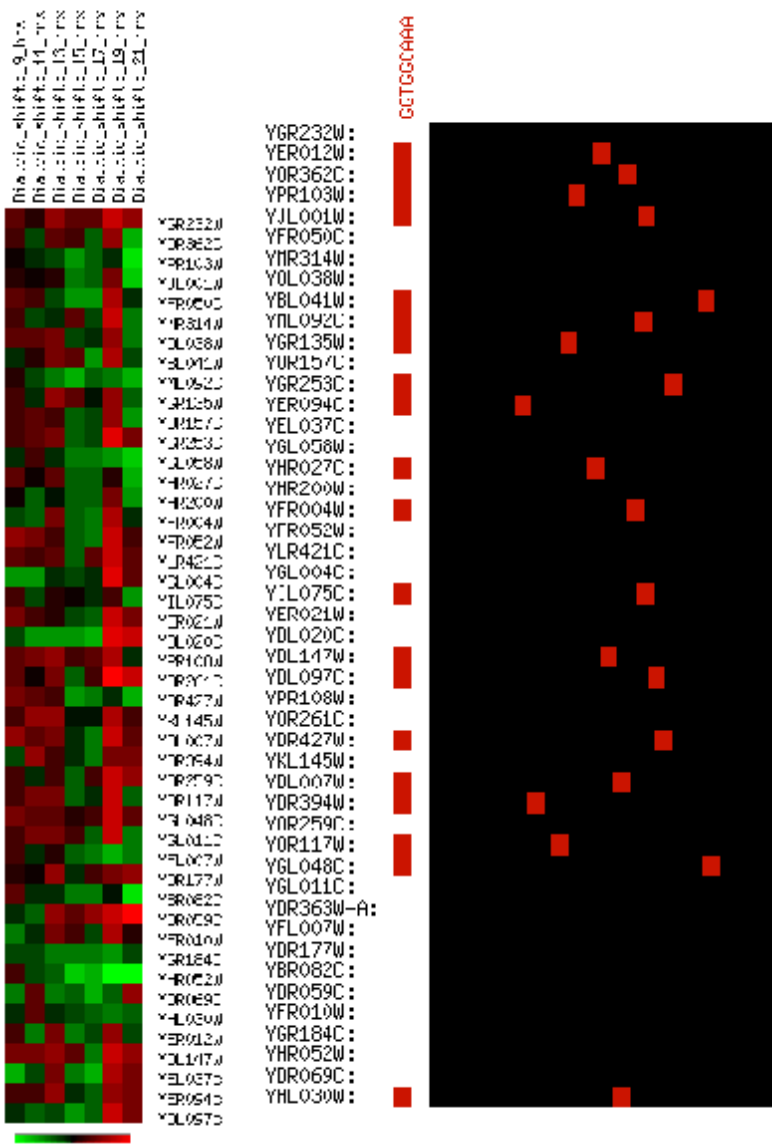
Protein-protein interactions



26S Proteasome (Saccharomyces cerevisiae)

KEGG





Steps to do

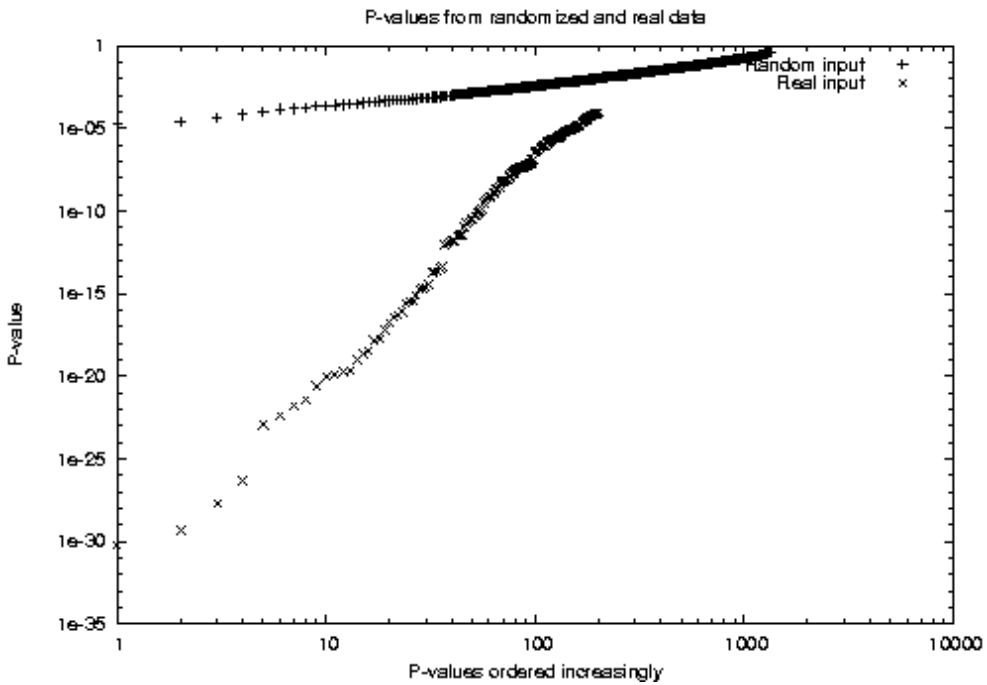
- Pattern discovery using different gene groups
- Expansion of gene groups based on PPI and motifs

Pattern discovery

- Using Sequence Pattern Exhaustive Search (SPEXS) algorithm (*Vilo*)
- Input: promoter sequences of somehow related genes (functionally or regulatory related)
- Background: all promoter sequences (*S.cerevisiae* or from all 8 yeasts)
- p-value threshold: 1.0e-04

Updating confidence threshold

- Input: randomly chosen promoter sequences
- Confidence threshold updated so, that expectance of false-positive motifs is smaller than 0.01%.
- New p-value threshold is between 9.1e-05 and 9.3e-06



Results - TF

- Known target gene sets
- For 138 out of 143 TF we found at least one motif
- 7381 distinct motifs

TF	Nr	Known TFBS	S.cer mot.	p-value	Ort. mot.	p-value
ACE2	83	TGCTGGT	TGCTGGC	3.95396e-07	AACCAGCA	3.02497e-13
ADR1	43	GGRGK	TTGGGGTA	1.22318e-07	TTGGGGTA	3.88413e-15
ASH1	5	YTGAT	*NA*	*NA*	GTCTCCCACATCACCA	2.66303e-17
CBF1	56	RTCACRTG	CACGTGA	4.49396e-39	CACGTG	4.44005e-83
CIN5	133	TTACRTAA	TTACATAA	4.20119e-12	TTACATAAT	3.55377e-14
CUP2	2	TCTTTTGCTG	TTCTTTTGCT	2.61691e-10	AATTAGTAAGC	6.91919e-11
GAT3	59	*NA*	TACTTCGAAGC	1.64527e-26	TACTTCGAAGC	2.10251e-31
IFH1	194	*NA*	TCCGTACA	2.41217e-27	TACTAAC	2.35358e-50

Results - GO

- For 2006 GO groups from 3977 we found at least one motif

GO description	p-value	pattern	TF
retrotransposon nucleocapsid	2.57319e-136	TGTTGGAATA	Mot3
nucleolus	3.73244e-53	AAAATTTT	
peptidase activity	1.09457e-34	GCAAGGATTGATAAT	
cytosolic ribosome (sensu Eukaryota)	1.89356e-31	CCGTACA	Rap1
amino acid biosynthesis	5.62027e-31	TGACTCA	Gcn4
proteasome complex (sensu Eukaryota)	5.39292e-28	GGTGGCAAA	Rpn4
hydrolase activity, acting on ester bonds	1.75131e-25	ATAATGTAATA	Hcm1?
transferase activity, transferring phosphorus-containing groups	1.10842e-23	GATTGATAATG	
membrane-enclosed lumen	1.74176e-23	GCGATGAG	Esr1/Mec1?
DNA replication	4.16142e-21	ACGCGT	Mbp1
helicase activity	1.53663e-14	CCTCGACTAA	Xbp1
aspartate family amino acid catabolism	1.91911e-11	AGCACGTGAC	Pho4

Expansion of gene groups

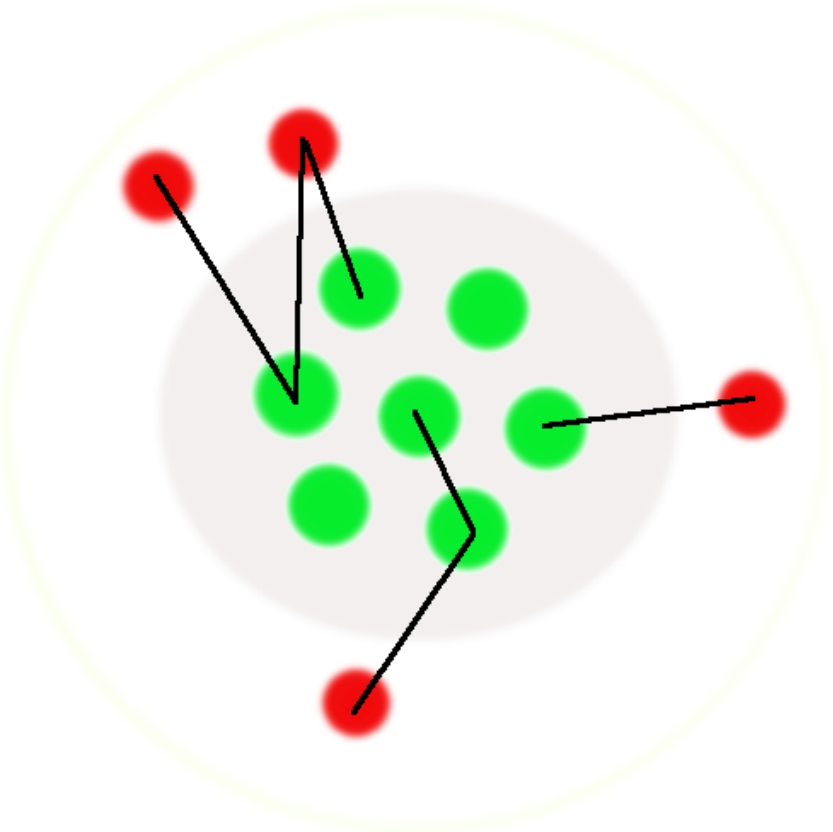
- For a input geneset (e.g GO:0008652) we found all interacting proteins
- We describe genes with the geneset specific motif (TGACTC)

Expansion of gene groups

- For a input geneset (e.g GO:0008652) we found all interacting proteins
- We describe genes with the geneset specific motif (TGACTC)

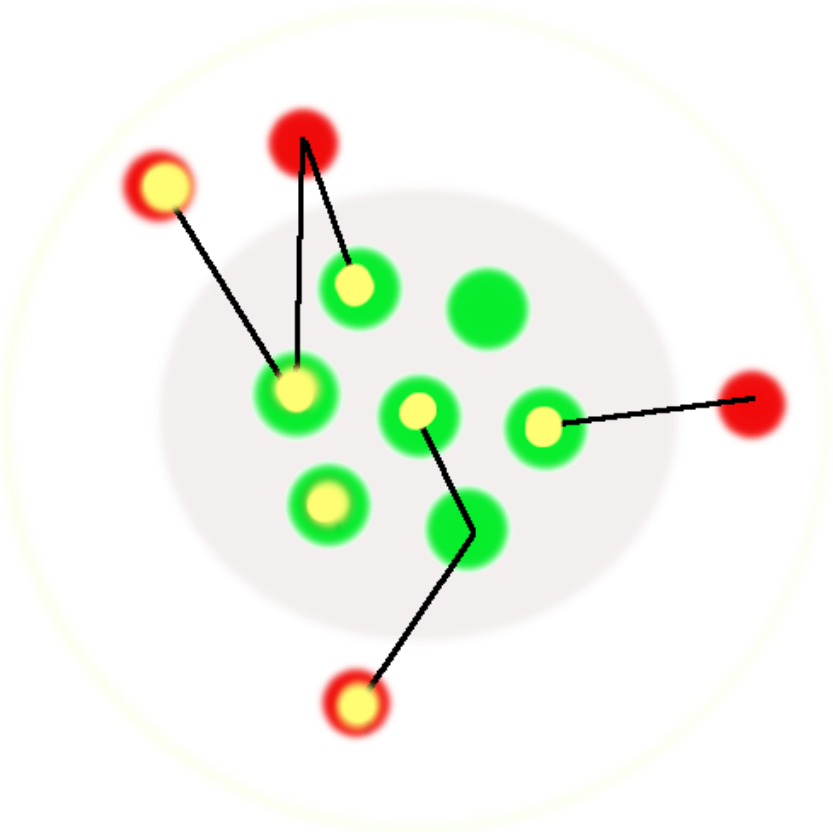
Expansion of gene groups

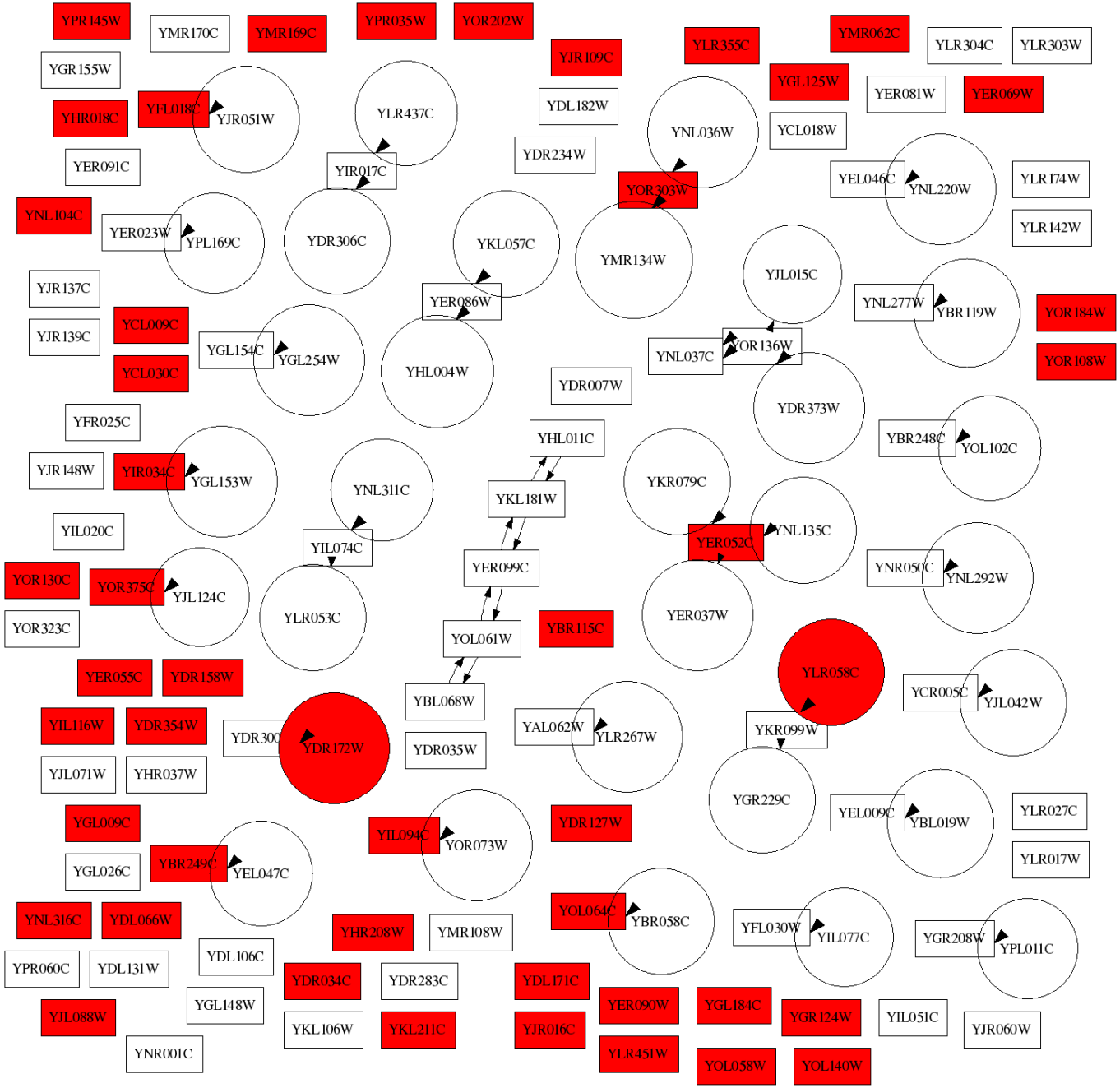
- For a input geneset (e.g GO:0008652) we found all interacting proteins
- We describe genes with the geneset specific motif (TGACTC)



Expansion of gene groups

- For a input geneset (e.g GO:0008652) we found all interacting proteins
- We describe genes with the geneset specific motif (TGACTC)




YPR145W
YMR170C
YMR169C
YPR035W
YOR202W
YJR109C
YLR355C
YMR062C
YLR304C
YLR303W
YGR155W
YFL018C
YJR051W
YLR437C
YDL182W
YGL125W
YER081W
YER069W
YHR018C
YNL036W
YCL018W
YER091C
YIR017C
YDR234W
YEL046C
YNL220W
YLR174W
YOR303W
YNL104C
YER023W
YPL169C
YDR306C
YKL057C
YMR134W
YLR142W
YJL015C
YJR137C
YNL277W
YBR119W
YOR184W
YER086W
YJR139C
YCL009C
YGL154C
YGL254W
YHL004W
YNL037C
YOR136W
YOR108W
YCL030C
YDR007W
YDR373W
YFR025C
YHL011C
YBR248C
YOL102C
YJR148W
YIR034C
YGL153W
YNL311C
YKL181W
YKR079C
YIL020C
YIL074C
YER052C
YNL135C
YOR130C
YOR375C
YJL124C
YER099C
YNR050C
YNL292W
YLR053C
YER037W
YOR323C
YOL061W
YBR115C
YLR058C
YER055C
YDR158W
YBL068W
YCR005C
YJL042W
YIL116W
YDR354W
YDR300
YDR172W
YDR035W
YAL062W
YLR267W
YKR099W
YJL071W
YHR037W
YGR229C
YEL009C
YBL019W
YLR027C
YGL009C
YIL094C
YOR073W
YDR127W
YLR017W
YBR249C
YEL047C
YGL026C
YOL064C
YBR058C
YFL030W
YIL077C
YGR208W
YPL011C
YNL316C
YDL066W
YHR208W
YMR108W
YDL106C
YPR060C
YDL131W
YDR034C
YDR283C
YDL171C
YER090W
YGL184C
YGR124W
YIL051C
YJR060W
YGL148W
YJL088W
YKL106W
YKL211C
YJR016C
YLR451W
YOL058W
YOL140W
YNR001C

Conclusions

- We found regulatory motifs for different input genesets:
  - we improved previously known motifs
  - we found new motifs for transcription factors
  - we found Gene Ontology group specific motifs
- Using protein-protein interactions and regulatory motifs we can expand gene groups with probable candidate genes
  Web-tool PPI-Gviz is publicly available (http://bioinf.ebc.ee/u/peterson/gviz/)
- All discovered motifs are in BiGeR database (http://bioinf.ebc.ee/biger/)

What next

- Motif discovery using non-descrete motifs
- Filtering input genesets
  - GO geneset filtering based on evidence codes
  - Using more protein-protein interaction datasets + filter the interactions based on the gene expression similarity.
  - Use only very similar orthologs from more distant yeasts
- Comparison of newly found motifs with previously described motifs in BiGeR database
- Try to discover specie specific motifs

Acknowledgements

Jaak Vilo, Jüri Reimand, Priit "Lemps" Adler, Jelena Zaitseva, Meelis Kull, Hendrik Nigul, Raivo Kolde, Jaanus Hansen ..

ESF grant 5724, EU STRE ATD, EU NoE ENFIN and Kristjan Jaak scholarship foundation