Whole-genome analysis of GCN4 binding in *S.cerevisiae*

Lillian Dai

Alex Mallet

Gcn4/DNA diagram (CREB symmetric site and AP-1 asymmetric site: Song Tan, 1999) removed for copyright reasons.

What is GCN4 ?

- *S.cerevisiae* transcription factor
- Primary regulator of the transcriptional response to amino acid starvation
- Has well-known binding motif: TGAsTCa*
  - Based on examining intergenic sequences
- Is known to bind both in intergenic regions and in ORFs**

* *"Transcriptional Regulatory Code of A Eukaryotic Genome"*, Harbison et al, Nature, Sep 2004

** *"Gcn4 occupancy of open reading frame regions [...]"*, Topalidou et al, EMBO, 4(9) 2003

Our project

- Investigating differences in GCN4 binding in intergenic regions versus ORFs
  - Does it bind to a different motif ?
  - How often do motifs occur in both types of regions ?
  - Are there differences in motif strength between intergenic regions and ORFs ?
  - How strongly does GCN4 bind in the different regions ?
- Similar to other studies eg Lieb *et al** looked at Rap1 binding

* *"Promoter-specific binding of Rap1 revealed by genome-wide maps of protein-DNA association"*, Lieb et al, Nature Genetics **28** (2001)

Dataset

- Whole-genome ChIP-ChIP binding data for GCN4 in *S. cerevisiae* from Gifford/Young lab under amino acid starvation conditions
  - Probes located (approx.) 300 basepairs apart
- “Binding call” data from Gifford lab
- Annotated *S.cerevisiae* genome from the *Saccharomyces* Genome Database (SGD)

MEME Algorithm

- One Occurrence Per Sequence (**OOPS**) $w-mer$

$$\theta = \begin{bmatrix} B_A & P_{A,1} & \cdots & P_{A,w} \\ B_G & P_{G,1} & \cdots & P_{G,w} \\ B_T & P_{T,1} & \cdots & P_{T,w} \\ B_C & P_{C,1} & \cdots & P_{C,w} \end{bmatrix}$$

- Zero or One Occurrence Per Sequence (**ZOOPS**)

  γ prior probability of a sequence containing a motif sequence

$$\max \Pr\left(X \middle| \phi\right)$$

Procedure:

E-step: compute $Z^{(t)} = \underset{(Z|X,\phi^{(t)})}{E}[Z]$

M-step: solve

$$\phi^{(t+1)} = \arg\max_{\phi} \underset{(Z|X,\phi^{(t)})}{E}[\log \Pr(X, Z|\phi)]$$

Converge to local maximum of $\Pr\left(X \middle| \phi\right)$

$X = \{X_1, X_2, ..., X_n\}$ Input sequence

$Z_{ij} = 1$ Motif starts in position j in sequence X_i

$\phi = [\theta\ w\ \gamma]$ Parameters

MEME Inputs

- Probe IP/WCE ratio greater than <3, 4, 6>
- Sequence lengths <400, 600, 800>
- - mod <zoops>
- - nmotifs <1, 10>
- - minsites <20>
- - minw <unspecified, 7, 12>
- - maxw <unspecified, 11, 18>
- - bfile <none, markov order 5>
- - revcomp

- Also, generated our own 5th-order Markov model for coding regions

Intergene Motif, Harbison, et. al. 2004*

meme GCN4_YPD.fsa -dna -minsites 20 -revcomp -mod zoops -bfile yeast.nc.6.freq -minw 7 -maxw 11

N= 59 strands

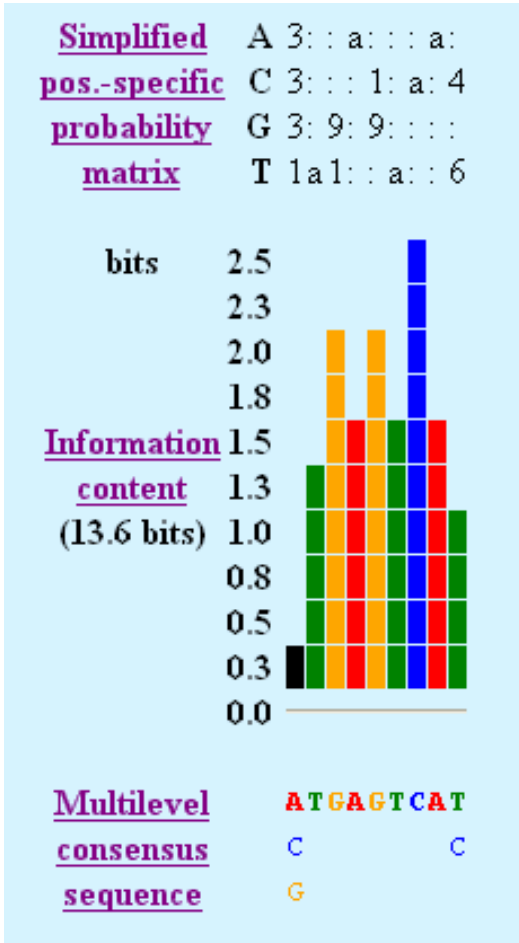
Letter frequencies in dataset:

A 0.322 C 0.178 G 0.178 T 0.322

Background letter frequencies (from yeast.nc.6.freq):

A 0.324 C 0.176 G 0.176 T 0.324

NAME	STRAND	START	P-VALUE	SITES	
iYNL104C	+	97	1.85e-06	TTTGAAAGAA	CTGAGTCAC TTACACGTAA
iYBR113W	-	487	1.85e-06	CCCGGATTGG	CTGAGTCAC CTTCATCGCG
iYHR161C	+	409	1.85e-06	AAAAGCCAGG	CTGAGTCAC GTCAGTTGCT
iYGL126W	-	1	7.11e-06	CCAACTTTTC	CTGAGTCAT
iYOR221C	-	341	7.11e-06	CGCGACTGCA	CTGAGTCAT CAACAACAAG
iYJR016C	+	132	7.11e-06	TACTATATTA	CTGAGTCAT CTGGAGAGGA
iYOR130C	+	229	7.11e-06	CGAGCTCAAG	GTGAGTCAC GATGCAGAAC
iYOL064C	+	150	7.11e-06	TATTGCTCGT	CTGAGTCAT TCGCGCATTT
iYNL005C	+	614	7.11e-06	TCAACGAATG	GTGAGTCAC CATTTAATGC
iYDL171C	+	573	7.11e-06	CTACCAGGGT	CTGAGTCAT CAAAGAAAAA
iYDL198C	-	139	7.11e-06	ACAAAAACTC	GTGAGTCAC TGTGCATTTG
iYCL030C	+	202	7.11e-06	ATAAAAAAAC	GTGAGTCAC TGTGCATGGG
iYER068W	-	375	7.11e-06	TTGATGTAGA	CTGAGTCAT TCGGATAAGA
iYER055C	-	192	7.11e-06	AAGCTTCCAA	GTGAGTCAC CTCTACCGTT
iYOL141W	-	43	7.11e-06	TGTACTTTAA	GTGAGTCAC ATAGCGAGCT



*Transcriptional Regulatory Code of A Eukaryotic Genome, Nature, 2004

Intergene Motif

meme intergenicover4.fsa -minsites 20 -dna -revcomp -mod zoops -
bfile yeast.nc.6.freq -minw 7 -maxw 11

N=71 strands

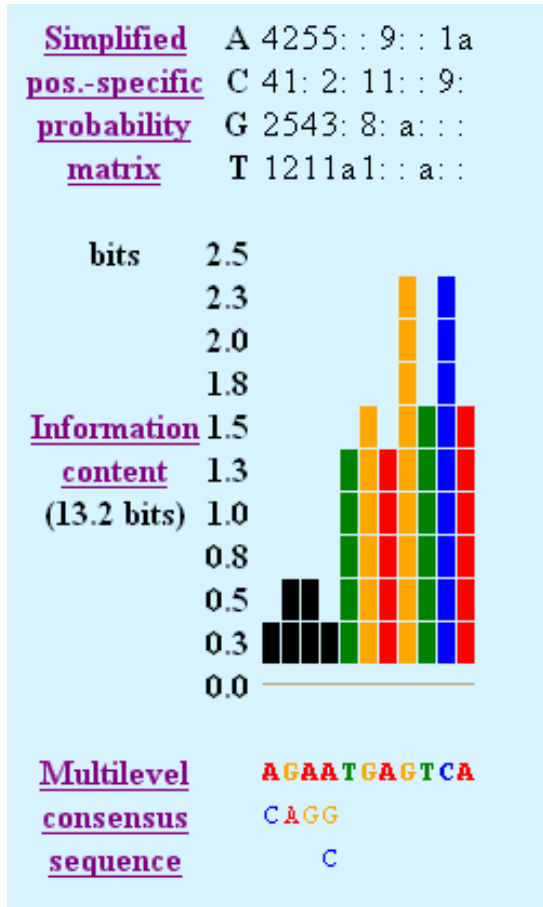
Letter frequencies in dataset:

A 0.296 C 0.204 G 0.204 T 0.296

Background letter frequencies (from yeast.nc.6.freq):

A 0.324 C 0.176 G 0.176 T 0.324

NAME	STRAND	START	P-VALUE	SITES
5.33798	-	418	5.70e-08	GTGTGGTTTC CGGGTGAGTCA TACGGCTTTT
5.33823	-	393	5.70e-08	GTGTGGTTTC CGGGTGAGTCA TACGGCTTTT
12.839902	+	52	1.62e-07	TTACGAAACG CGGATGAGTCA CTGACAGCCA
12.839552	+	402	1.62e-07	TTACGAAACG CGGATGAGTCA CTGACAGCCA
12.839577	+	377	1.62e-07	TTACGAAACG CGGATGAGTCA CTGACAGCCA
12.198181	-	444	3.24e-07	AGAGTCGGAC CGAGTGAGTCA GCGTGATCGG
12.198156	-	469	3.24e-07	AGAGTCGGAC CGAGTGAGTCA GCGTGATCGG
8.422719	+	378	1.09e-06	TACAAAAGCC AGGCTGAGTCA CGTCAGTTGC
4.704348	-	309	1.19e-06	TTTCATGTTC GGGATGAGTCA TATGCATGAC
16.822432	-	358	1.80e-06	TTCAGTTTAC AGAATGAGTCA AATGTTACAT
11.38203	+	465	1.80e-06	GCTATAGATT AGAATGAGTCA ACGAGCCATT
16.822357	-	433	1.80e-06	TTCAGTTTAC AGAATGAGTCA AATGTTACAT
7.156416	+	382	1.80e-06	AAAAAGAGTC AGAATGAGTCA GCCGGATAAC
5.295198	-	388	2.10e-06	TTTTTGATGT AGACTGAGTCA TTCGGATAAG
2.466735	-	192	3.48e-06	TCACCCGGAT TGGCTGAGTCA CCTTCATCGC



In-gene Motif

meme ingenebindingover4.fsa -dna -revcomp -mod zoops -bfile yeast.coding.6.freq -minw 7 -maxw 11

N= 15 strands

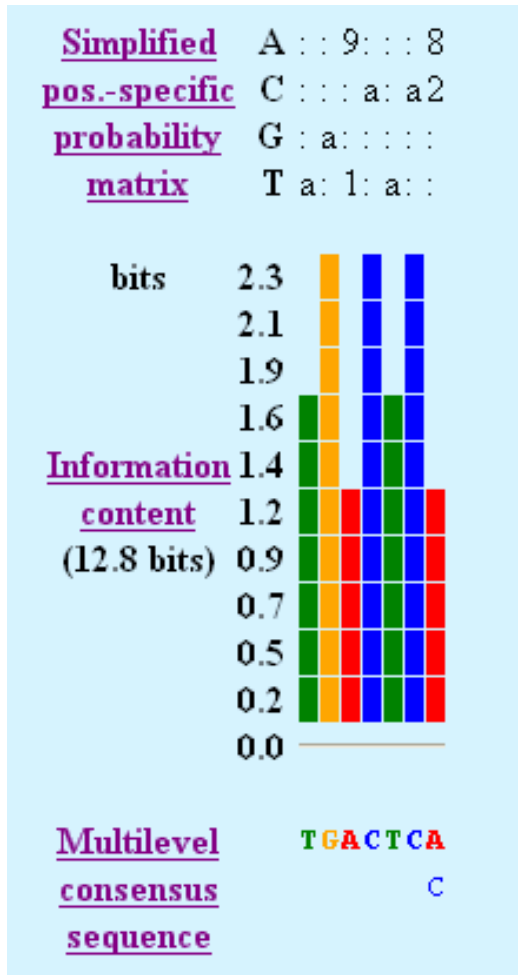
Letter frequencies in dataset:

A 0.291 C 0.209 G 0.209 T 0.291

Background letter frequencies (from yeast.coding.6.freq):

A 0.302 C 0.198 G 0.198 T 0.302

NAME	STRAND	START	P-VALUE	SITES	
13.668041	+	350	6.45e-05	ACTCCTGGTA	TGACTCA TCACTTGAAG
7.288316	-	348	6.45e-05	CAGAAACAAA	TGACTCA TTATATATGA
7.272091	+	453	6.45e-05	ACGTCCAGTA	TGACTCA GGAAAAGTTG
4.1420214	-	417	6.45e-05	AGCAGCATGA	TGACTCA CACATTACCA
14.51924	+	305	6.45e-05	GCAGTAATCA	TGACTCA GTCTCTGGAA
13.396266	-	659	6.45e-05	ACAGGTAGGA	TGACTCA CTGCGTAATT
16.25093	-	553	6.45e-05	TGTTAGCAGA	TGACTCA TCAGACCAAA
4.104410	+	693	6.45e-05	AAATGCACAG	TGACTCA CGAGTTTTTG
2.136145	+	355	6.45e-05	ATCACGAAGC	TGACTCA TCTGTGATTT
15.58527	+	452	6.45e-05	GCTCGCTATG	TGACTCA CTTAAAGTAC
2.136170	+	330	6.45e-05	ATCACGAAGC	TGACTCA TCTGTGATTT
5.342598	+	420	1.07e-04	TTTTTTCACC	TGACTCC GGTCCTGACG
12.404010	-	425	1.07e-04	TAGCGAGGGA	TGACTCC GTAGATACTT
9.254609	-	394	1.71e-04	ATTTCCTGCT	TGTCTCA CTATTAACGC
4.1490164	+	693	2.13e-04	CCCCTTTTCC	TGTCTCC TATTGTGCGA



In-gene Motif

meme ingenebindingover3.fsa -dna -revcomp -mod zoops -bfile yeast.coding.6.freq
-nmotifs 10 -minw 7 -maxw 11

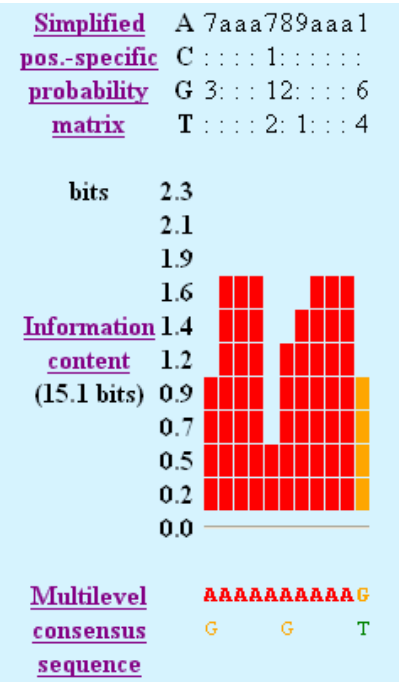
N= 49 strands

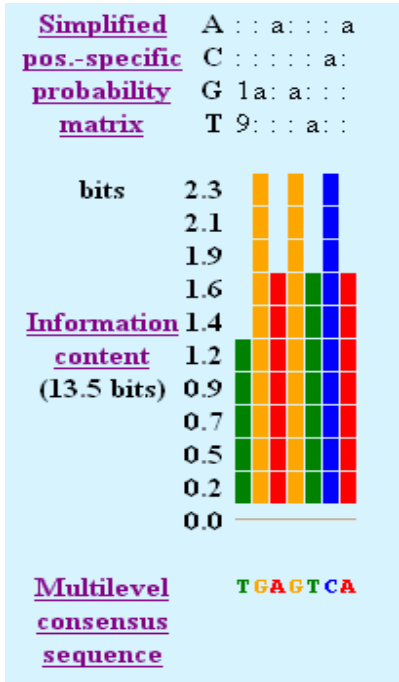
Letter frequencies in dataset:

A 0.292 C 0.208 G 0.208 T 0.292

Background letter frequencies (from yeast.coding.6.freq):

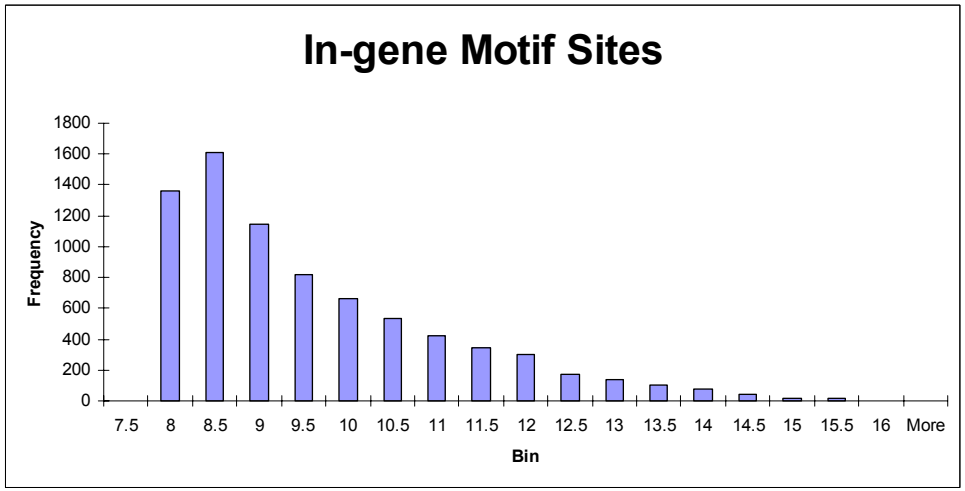
A 0.302 C 0.198 G 0.198 T 0.302



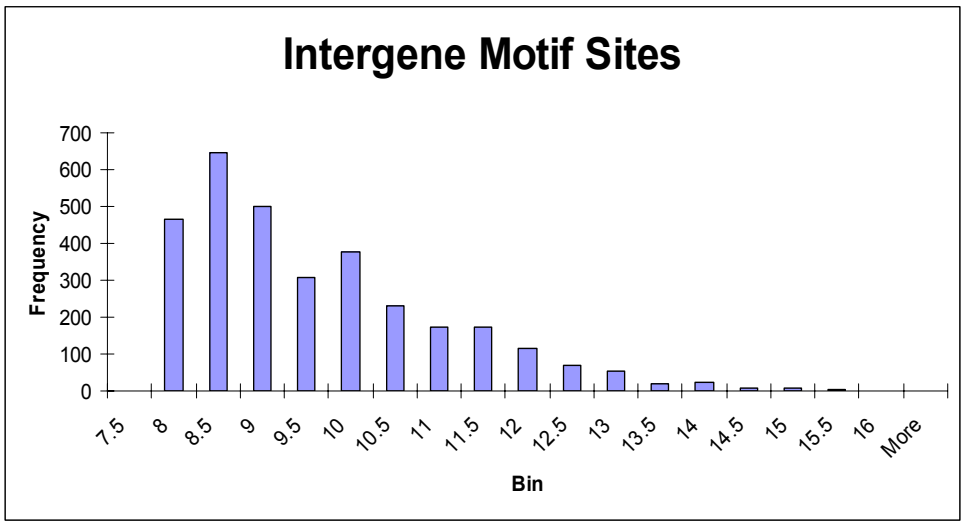


NAME	STRAND	START	P-VALUE	SITES
12.838427	+	682	6.45e-05	TGGTTCAACG TGAGTCA AGTCCTTGAA
7.272066	-	478	6.45e-05	CAACTTTTCC TGAGTCA TACTGGACGT
15.150658	+	368	6.45e-05	GGTACAAAGA TGAGTCA TCAGATATTC
16.74463	-	463	6.45e-05	CGATTGTCGA TGAGTCA GAATGCACGC
6.204361	-	552	6.45e-05	AAGAATTTCA TGAGTCA TGCGCAAGAA
8.141907	+	61	6.45e-05	ATCCAAAAAG TGAGTCA TTCATCTACT
15.758447	-	125	6.45e-05	GCGACTGCAC TGAGTCA TCAACAACAA
4.935102	+	371	6.45e-05	CAGAATTCTG TGAGTCA TACTTCCCAG
3.68174	+	753	6.45e-05	TAAAAAAACG TGAGTCA CTGTGCATGG
5.295348	-	238	6.45e-05	TGATGTAGAC TGAGTCA TTCGGATAAG
4.835164	+	483	6.45e-05	AAGCTTGTTA TGAGTCA TCTCATCGTT
10.633618	+	414	6.45e-05	ATTTGATCTT TGAGTCA CCACAATTGT
8.118307	+	489	6.45e-05	GAATTGCTGA TGAGTCA TCCGAAACTT
10.268276	-	526	6.45e-05	AGAAAGCATA TGAGTCA CGACGTGACC
12.233460	+	308	6.45e-05	AGCTGGGAAT TGAGTCA AGTCAATCAA
13.668066	-	325	6.45e-05	CTTCAAGTGA TGAGTCA TACCAGGAGT
10.633643	+	389	6.45e-05	ATTTGATCTT TGAGTCA CCACAATTGT
12.838952	+	157	6.45e-05	TGGTTCAACG TGAGTCA AGTCCTTGAA
2.324360	-	342	6.45e-05	ATAGTCAGAA TGAGTCA TTGTAAATAG
13.668041	-	350	6.45e-05	CTTCAAGTGA TGAGTCA TACCAGGAGT
7.288316	+	348	6.45e-05	TCATATATAA TGAGTCA TTTGTTTCTG
7.272091	-	453	6.45e-05	CAACTTTTCC TGAGTCA TACTGGACGT
4.1420214	+	417	6.45e-05	TGGTAATGTG TGAGTCA TCATGCTGCT
14.51924	-	305	6.45e-05	TTCCAGAGAC TGAGTCA TGATTACTGC
13.396266	+	659	6.45e-05	AATTACGCAG TGAGTCA TCCTACCTGT
16.25093	+	553	6.45e-05	TTTGGTCTGA TGAGTCA TCTGCTAACA
4.104410	-	693	6.45e-05	CAAAAACTCG TGAGTCA CTGTGCATTT
2.136145	-	355	6.45e-05	AAATCACAGA TGAGTCA GCTTCGTGAT
15.58527	-	452	6.45e-05	GTACTTTAAG TGAGTCA CATAGCGAGC
2.136170	-	330	6.45e-05	AAATCACAGA TGAGTCA GCTTCGTGAT
7.883568	+	361	1.07e-04	AAACACCAGT GGAGTCA ATGGCGATGT
7.625334	+	615	1.07e-04	TTCAACGCCT GGAGTCA GACCCTGCGC
5.342598	-	420	1.07e-04	CGTCAGGACC GGAGTCA GGTGAAAAAA
12.404010	+	425	1.07e-04	AAGTATCTAC GGAGTCA TCCCTCGCTA

GCN4 Motif Strength and Density



Average score	9.379596
Max score	15.241
Min score	7.623
Median score	8.901
Mode	7.916
Number of occurrences	7768
Number of coding bases	8659538
Average distance between motifs	1114.771



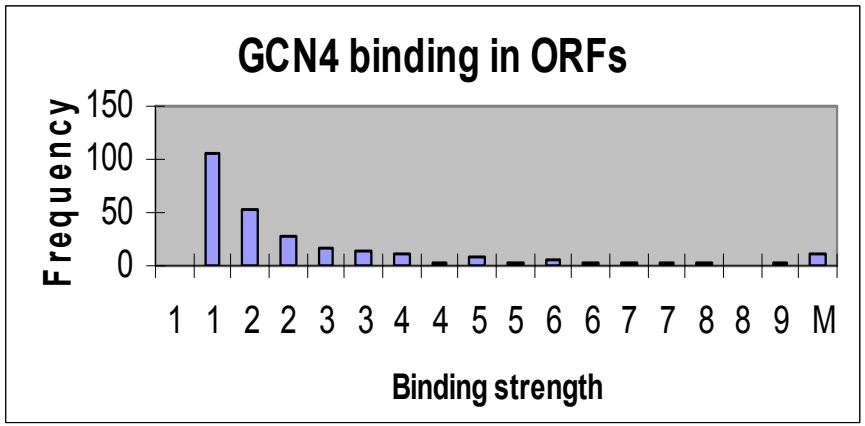
Average score	9.378928
Max score	15.241
Min score	7.623
Median score	8.9825
Mode	7.916
Number of occurrences	3168
Number of intergenic bases	3497052
Average distance between motifs	1103.867

Binding call data

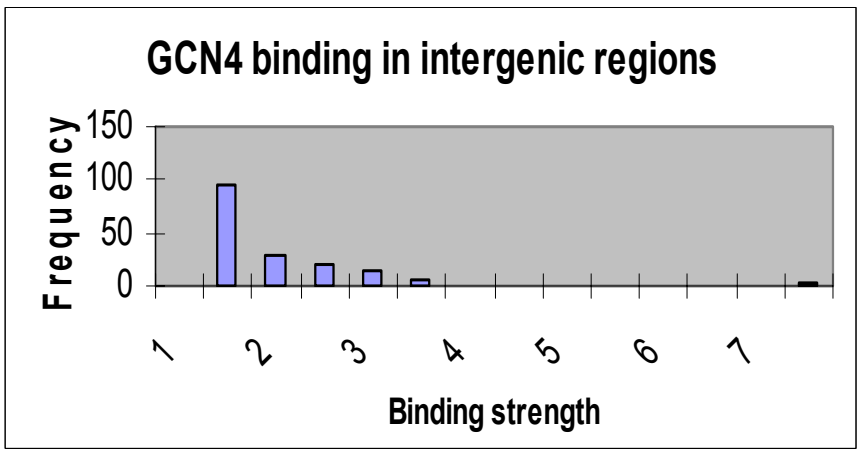
- Maximum likelihood model of actual location where GCN4 bound*
- Tries to find binding location that will best reproduce raw ChIP-ChIP data based on probabilistic model of
  - DNA fragment size distribution
  - Expected signal at a probe, given a GCN4 binding event at another location
- Quite accurate:
  - Majority of occurrences of GCN4 motif are within 100 bases of predicted binding sites with intensity > 1.75 (compared to mean distance of 500 bases for entire genome)

* *"Computationally Increasing Microarray Resolution"*, Rolfe et al, unpublished.

Analyzing binding strength*



Binding calls > 2.0	175
Average binding strength	1.852
Max binding strength	11.017
Min binding strength	1.002
Median binding strength	1.411
Avg distance between binding calls	~50000



Binding calls > 2.0	275
Average binding strength	2.838
Max binding strength	22.710
Min binding strength	1.006
Median binding strength	1.730
Avg distance between binding calls	~13000

*Assumes probe spacing is the same in intergenic regions and ORFs

Conclusions

- GCN4 appears to bind to the same motif in ORFs and intergenic regions
- Distribution and strength of GCN4 motif is same in ORFs as in intergenic regions
- Despite this, GCN4 binds more strongly, and more often, in intergenic regions than in ORFs
- Possible future directions:
  - Further characterize ORF binding locations e.g. by distance from intergenic region
  - Try to correlate binding strength with eg chromatin structure, mediator protein binding sites, chromosome remodeling complex binding sites etc

Thanks to

- Tim Danford
- Alex Rolfe
- Kenzie MacIsaac
- Robin Dowell
- Prof. Gifford

Questions ?

extra

- **Gcn4**, a basic leucine zipper protein, is the primary regulator of the transcriptional response to amino acid starvation (Hinnebusch and Fink, 1983). It is regulated at multiple levels, all of which alter the amount of **Gcn4** present within the cell. **Gcn4** is regulated at the level of protein stability, with its half-life ranging from approximately 2 minutes under growth in rich medium to 10 minutes under amino acid starvation conditions (Shemer et al., 2002). This degradation is mediated by the ubiquitin-conjugating enzymes Rad6 and Cdc34 (Kornitzer et al., 1994), and requires phosphorylation by the nuclear cyclin-dependent kinases Pho85 (Meimoun et al., 2000) or Srb10 (Chi et al., 2001).
- **Gcn4** is also regulated at the level of translation. Modulation of the activity of translation initiation machinery leads to an increase in the synthesis of **Gcn4** protein under conditions of amino acid starvation. Under non-starvation conditions, ribosomal complexes are diverted to ORFs upstream of the **Gcn4** coding region (Hinnebusch, 1984; Hinnebusch, 1997; Hinnebusch et al., 1988; Mueller and Hinnebusch, 1986; Thireos et al., 1984).
- Finally, there is evidence that **Gcn4** is also regulated transcriptionally. In strains carrying mutations that abolish translational regulation, there is nonetheless an increase in the levels both of **Gcn4** protein and mRNA levels following induction of the amino acid starvation response (Albrecht et al., 1998).
- Interestingly, the binding site for **Gcn4**, TGASTCA, is also recognized by the unrelated transcriptional regulator Bas1 (Springer et al., 1996). It is thought that this overlapping specificity serves as a mechanism for cross-regulation of adenine biosynthesis by both regulators.