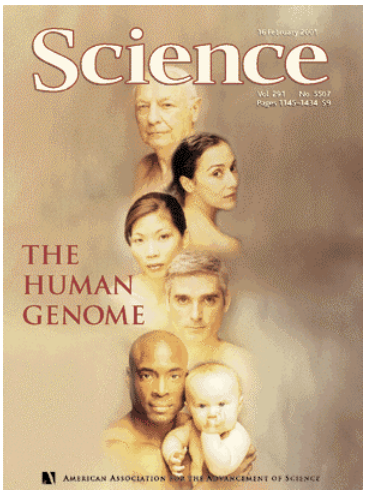
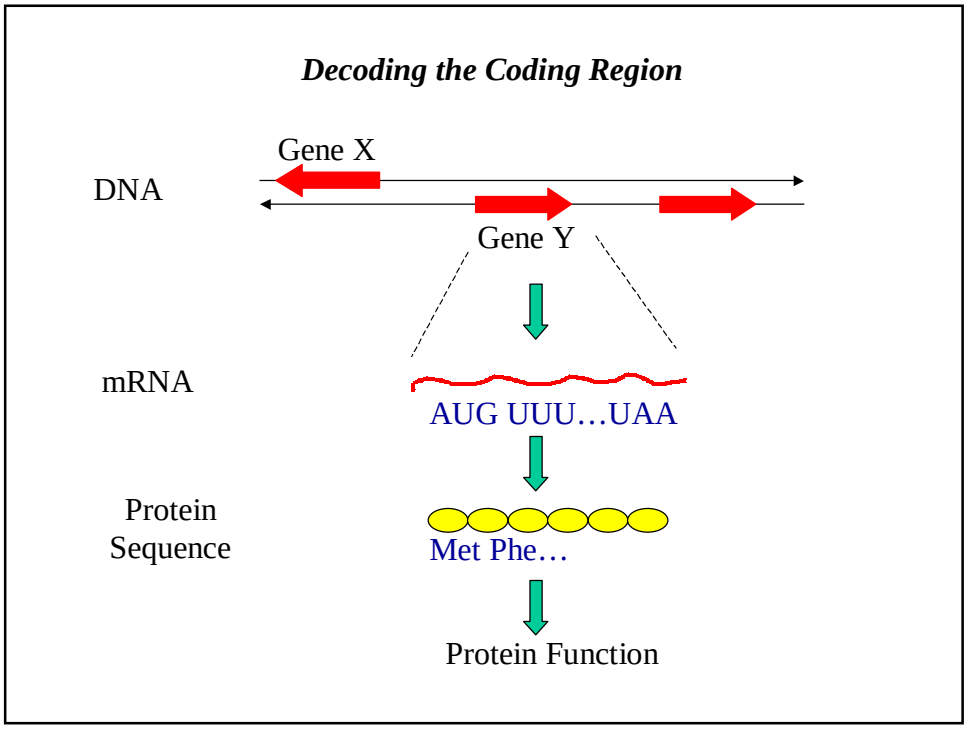
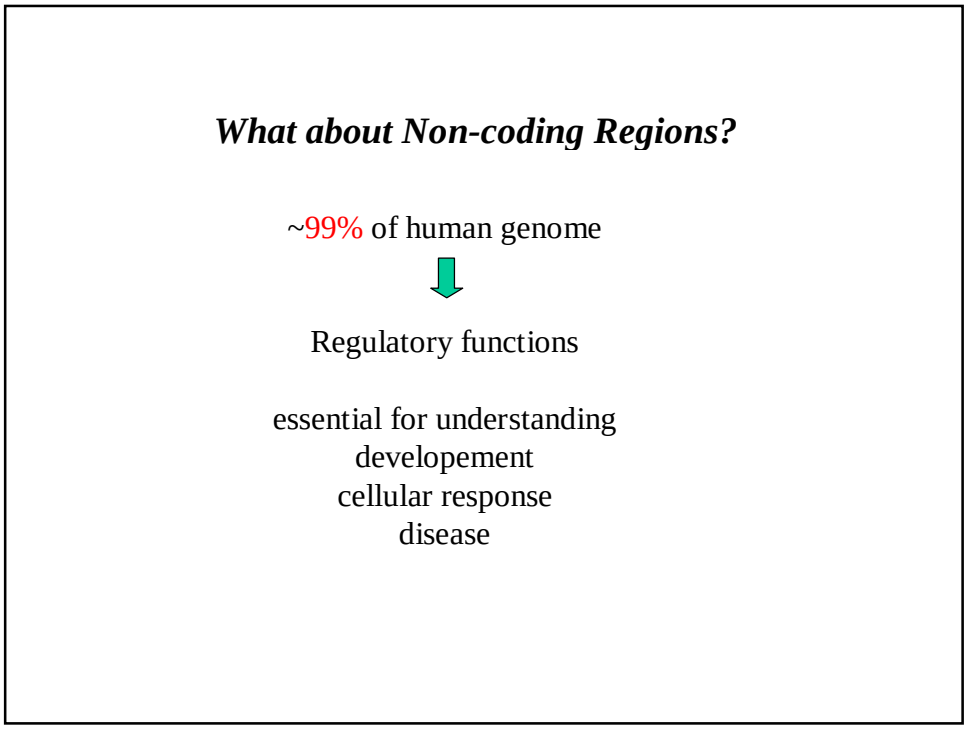
Building a Dictionary for DNA:

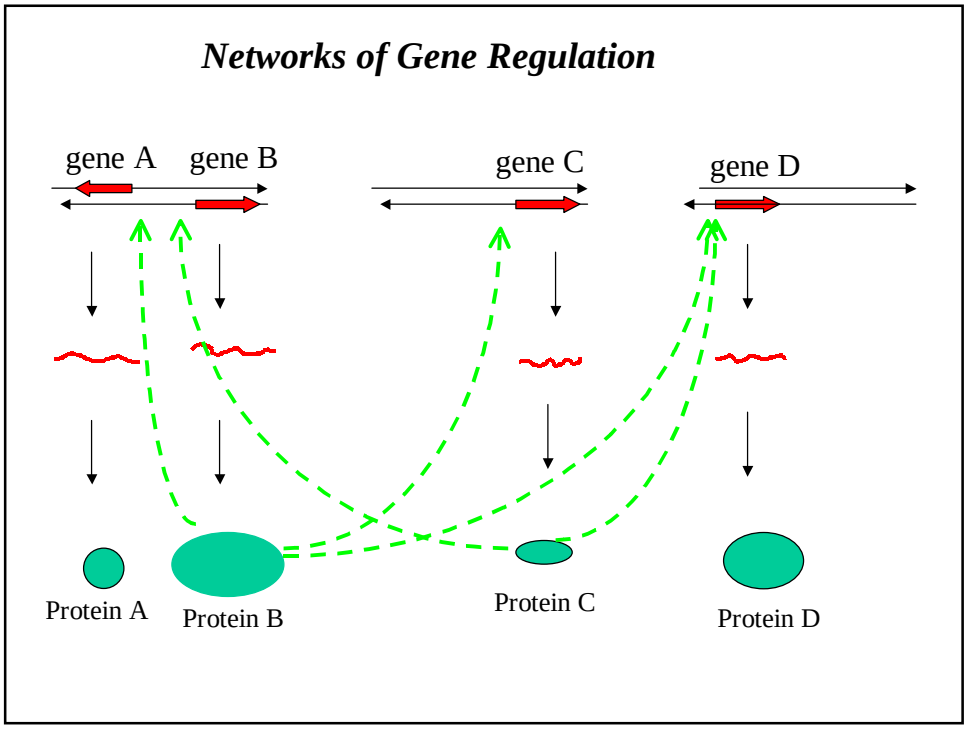
Decoding the Regulatory Regions of a Genome

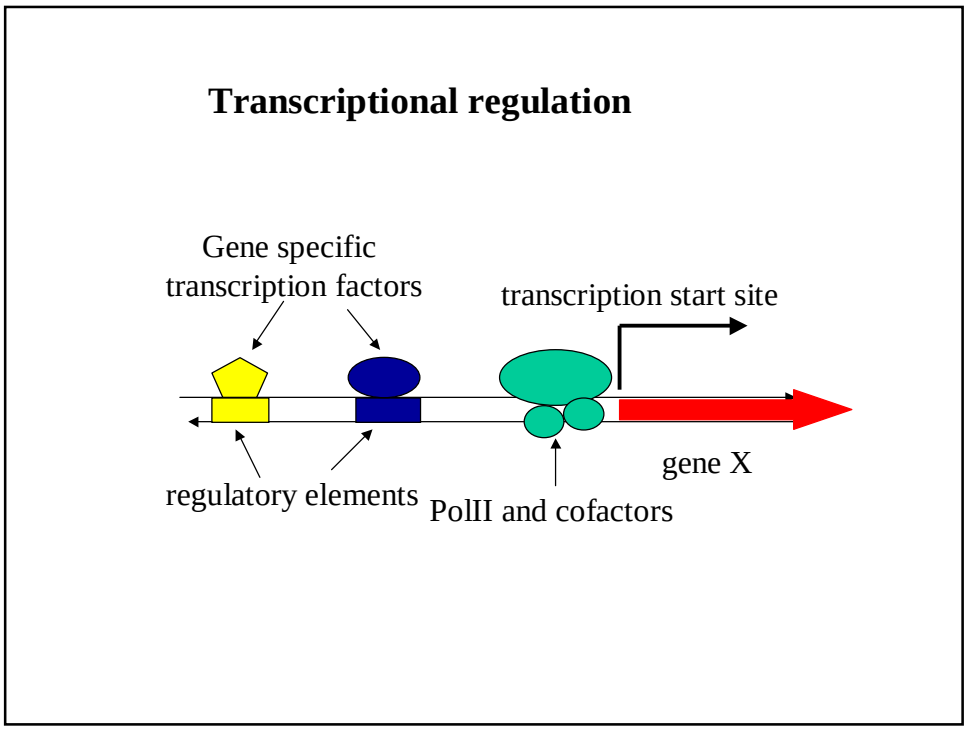
The Book of Man

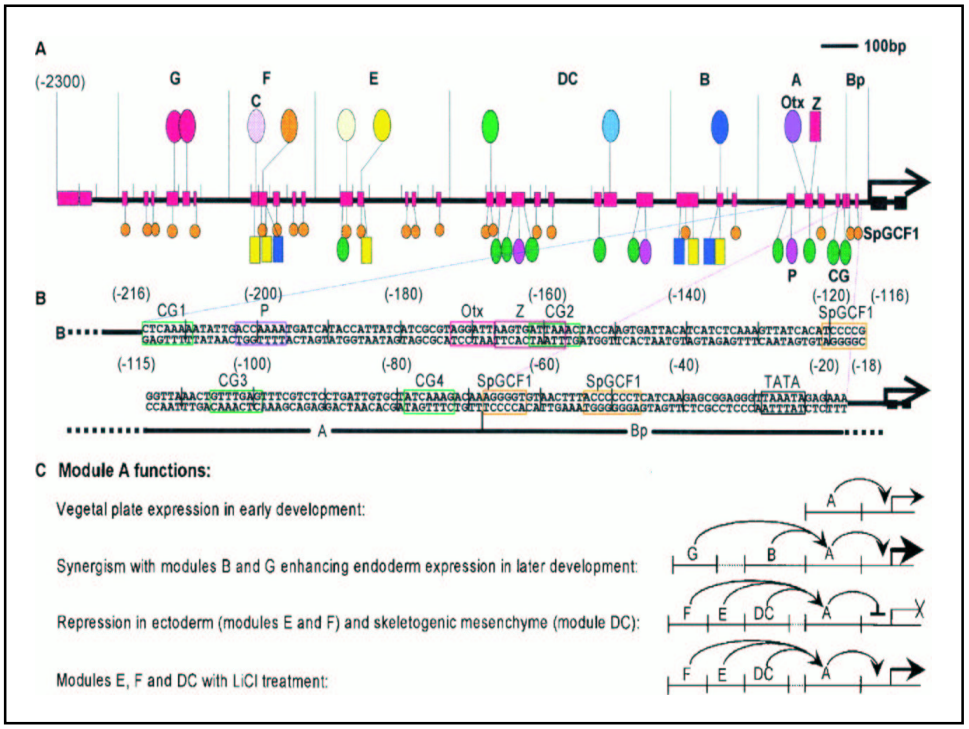



Decoding the Coding Region
Gene X
DNA
Gene Y
mRNA
AUG UUU…UAA
Protein
Sequence
Met Phe…
Protein Function


What about Non-coding Regions?
~99% of human genome
Regulatory functions
essential for understanding
developement
cellular response
disease


Networks of Gene Regulation
gene A
gene B
gene C
gene D
Protein A
Protein B
Protein C
Protein D


Transcriptional regulation
Gene specific
transcription factors
transcription start site
gene X
regulatory elements
PolII and cofactors



Measure genome-wide gene expression

mRNA from cell extract

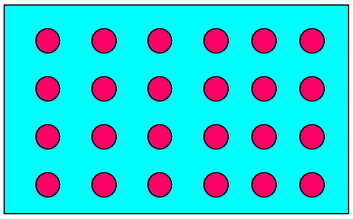
↓ reverse transcribe

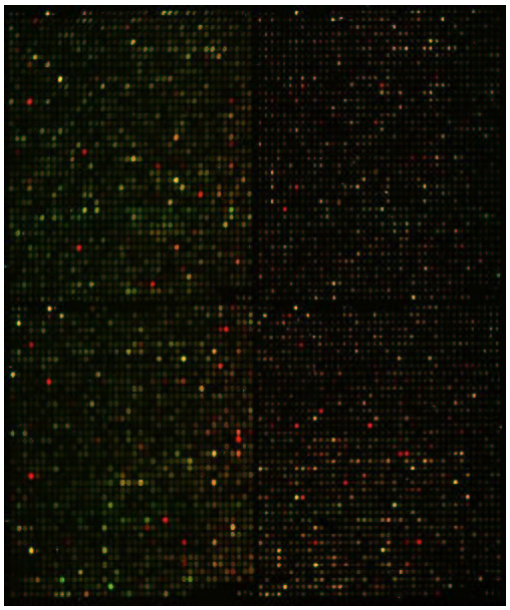
cDNA with label (e.g., fluorescence)

↓ hybridize with the array

DNA microarray

DNA probes for all the genes





complete genome sequences

+

genome-wide expression data
(e.g., from DNA microarray)

↓

opportunity to decipher a cell's
Transcriptional program

Computational approaches to decoding the regulatgory regions

- Statistical analysis and pattern discovery
  deciphering the book of life
- Quantitatively model gene expression data
  matching sequence features to gene expression
- Comparing regulatory regions across species

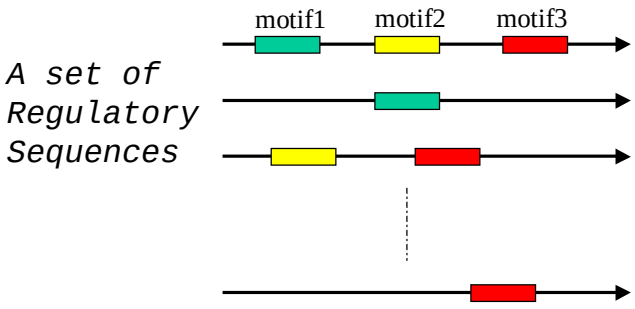
Statistical analysis

building a dictionary for the genome – delineating biologically meaningful "word" from the sequence

- probabilistic segmentation model -- Mobydick
- analysis of the yeast genome

  codon structure
  regulatory elements
  co-regulated genes

a model for regulatory sequences



Example: ATTGTATATAAAGTGTACGCGTGACC

chapterptgpbqdrftezptqtasctmvivwpecjsnisrmbtqlmlfvetl
loomingsfkicallxjgkmekysjerishmaeljplfsomeylqyearstvh
njbagoaxhjtjcokhvneverpmqpmindhowzrbdlzjllonggbhqi
preciselysunpvskepfdjktcgarwtnxybgcvdjfbnohavinglittl
ezorunozsoyapmoneyyvugsgtsqintmyteixpurseiwfmjwgj
nyyveqxwftlamnbxkrsbkyandrnothingcgparticularwtzao
qsjtnmtoqsnwvxfiupinterestztimebymonlnshoreggditho
ughtyxfxmhqixceojjzdhwouldsailpcaboutudxsbsnewtpg
gvjaasxmsvlittleplvcydaowgwlbzizjlnzyxandzolwcudthjd
osbopxkkfdosxardgcseebbthefzrsskdhmawateryjikzicim
ypartmofprtheluworldvtoamfutitazpisagwewayrqbkiosh
avebojwphiixofprmalungipjdrivingpkuyoikrwxoffodhicb
nimtheixyucpdzacemspleenqbpcrmhwvddyaiwnandada
bkpgzmptoregulatingeetheslcirculationvsuctzwvfyxstuzr
dfwvgygzoejdfmbqescwheneverpitfindfmyselfcgrowingne
ostumrydrrthmjsmgrimcczhjmgbkwczoaboutjbwanbwzq
thehrjvdrcjjgmouthuutwheneveritddfouishlawwphxnae

Moby Dick: CHAPTER 1
Loomings

Call me Ishmael. Some years ago- never mind how long precisely-having little or no money in my purse, and nothing particular to interest me on shore, I thought I would sail about a little and see the watery part of the world. It is a way I have of driving off the spleen and regulating the circulation. Whenever I find myself growing grim about the mouth; whenever it is a damp, drizzly November in my soul; whenever I find myself involuntarily pausing before coffin warehouses, and bringing up the rear of every funeral I meet; and especially whenever my hypos get such an upper hand of me, that it requires a strong moral principle to prevent me from deliberately stepping into the street, and methodically knocking people's hats off- then, I account it high time to get to sea as soon as I can.

Our Model:

Probabilistic Segmentation/Maximum likelihood

A probabilistic dictionary
Words
probabilities

$$\left\{\begin{array}{ll} A & \rightarrow P_A \\ C & \rightarrow P_C \\ G & \rightarrow P_G \\ T & \rightarrow P_T \\ GC & \rightarrow P_{GC} \\ TATAA & \rightarrow P_{TATAA} \end{array}\right\}$$

A | G | T | A | T | A | A | G | C
A | G | T A T A A | G C
A | G | T A T A A | G | C

maximizing the likelihood function

$$Z = \sum_{Seg} P_{w_1} P_{w_2} P_{w_3} \ldots P_{w_n}$$

Dictionary Construction

Fitting step: given the entries in the dictionary, find P_W by maximizing the likelihood function. Starting with a simple dictionary with all possible words

Adding new words: do statistical test on longer words based on the current dictionary, add the ones that are over-represented re-assign P_W by maximizing the likelihood function

Iterate the above

ξ_w Number of occurrences of word w anywhere in the sequence

$N_w(Seg)$ Number of counts of word w in a given segmentation

$\xi_{the} = 2$

$N_{the} = 1$

$$Z = \sum_{Seg} \prod (p_w)^{N_w(Seg)}$$

"free energy" $f = \frac{-\log(Z)}{L}$

"energy" $p_w = e^{-E_w}$

$$< N_w >= L \frac{\partial f}{\partial E_w}$$

self consistent eqs. $p_w = \frac{\langle N_w \rangle}{\sum_{w'} \langle N_{w'} \rangle}$

find interaction parameters which minimize the free energy

recursion relations for the likelihood function

sequence ⟶ ACGGTAAC

position ⟶ 1 2 L-1 L

$$Z(L) = Z(L-1)P_c + Z(L-2)P_{AC} + Z(L-3)P_{AAC} + \ldots$$

derivatives of Z can be calculated
effectively by defining greens functions

ATTTTACGCGTTTGT

$$\partial Z / \partial p_{ACGCGT} =$$
$$Z(ATTTT) * Z(TTGT) \equiv G(X = 6, l = 6)$$

Solution of the segmentation model

The likelihood function and its derivatives can be computed in a time ~ $O(L_g D_m)$ using dynamical programming

Solving the fixed point equations by Newton's

Assign a probability P_W to each word $Q_w = \frac{<N_w>}{\xi_w}$

Assign a quality factor to each word

$$\xi_{the} = 2$$
$$N_{the} = 1$$

Ditionary1		Ditionary2		Dictionary3	
e	0.065239	e	0.048730	e	0.042774
t	0.055658	s	0.042589	s	0.040843
a	0.052555	a	0.040539	a	0.038595
o	0.050341	t	0.040442	i	0.036897
n	0.049266	i	0.038550	t	0.036871
i	0.048101	d	0.038547	d	0.036323
s	0.047616	o	0.036486	l	0.035336
h	0.047166	l	0.036300	c	0.034818
r	0.043287	g	0.034509	m	0.034650
l	0.041274	r	0.034496	y	0.034482
d	0.039461	c	0.033916	b	0.034396
u	0.034742	m	0.033724	r	0.034105
m	0.034349	n	0.033321	p	0.034044
g	0.034001	y	0.033227	w	0.033819
w	0.033967	p	0.033156	n	0.033817
c	0.032934	f	0.032863	g	0.033676
f	0.032597	b	0.032780	f	0.033534
y	0.031776	w	0.032009	o	0.033206
p	0.031711	h	0.031494	h	0.033200
b	0.031409	v	0.030727	k	0.032103
v	0.028268	k	0.030445	v	0.031498
k	0.028113	u	0.030379	j	0.031209
j	0.026712	j	0.029268	u	0.031186
q	0.026561	z	0.028905	z	0.031003
z	0.026542	x	0.028404	x	0.030544
x	0.026357	q	0.028123	q	0.030244
		th	0.009954	the	0.005715
		in	0.006408	ing	0.003237
		er	0.004755	and	0.003128
		an	0.004352	in	0.002968
		ou	0.003225	ed	0.002547
		on	0.003180	to	0.002496
		he	0.003108	of	0.002486
		at	0.002851	en	0.001331
		ed	0.002804	an	0.001313
		or	0.002786	th	0.001270
		en	0.002538	er	0.001250
		to	0.002511	es	0.001209
		of	0.002475	at	0.001181
		st	0.002415	it	0.001171
		nd	0.002297	that	0.001165

Words	<Nw>	quality factor
abominate	2.0000	1.0000
achieved	2.0000	1.0000
aemploy	2.0000	1.0000
affrighted	2.0000	1.0000
afternoon	2.0000	1.0000
afterwards	5.0000	1.0000
ahollow	2.0000	1.0000
american	3.0000	1.0000
anxious	2.0000	1.0000
apartment	2.0000	1.0000
appeared	4.0000	1.0000
astonishment	4.0000	1.0000
attention	2.0000	1.0000
avenues	2.0000	1.0000
bashful	2.0000	1.0000
battery	2.0000	1.0000
beefsteaks	2.0000	1.0000
believe	2.0000	1.0000
beloved	2.0000	1.0000
beneath	6.0000	1.0000
between	12.0000	1.0000
boisterous	3.0000	1.0000
botherwise	2.0000	1.0000
bountiful	2.0000	1.0000
bowsprit	2.0000	1.0000
breakfast	5.0000	1.0000
breeding	2.0000	1.0000
bulkington	3.0000	1.0000
bulwarksb	2.0000	1.0000
bumpkin	2.0000	1.0000
business	6.0000	1.0000
carpenters	2.0000	1.0000

Building a dictionary for chrIV

"discovered" genetic code

Words	<Nw>	quality factor
gaa	17519.0000	0.5232
aaa	16298.1000	0.3407
gat	14533.6000	0.6406
aat	14095.1000	0.4361
att	11351.0000	0.3601
aag	11285.1000	0.3608
tta	10186.9000	0.4469
ttt	9865.0000	0.3080
ttg	9800.2000	0.4083
caa	9748.4000	0.3199
aac	8939.7000	0.4049
tct	8626.5000	0.5041
ggt	8232.7000	0.5778
gtt	8012.3000	0.4316
aga	7982.4000	0.2690
atg	7615.5000	0.3212
gct	7409.3000	0.5653
gag	7381.9000	0.4979
tca	7352.7000	0.3367
gac	7313.2000	0.5756
act	7298.2000	0.4173
tat	7169.2000	0.3043
ata	7006.6000	0.2996
aca	6671.1000	0.3040
ttc	6584.9000	0.3108
.	.	.
.	.	.
taa	422.4000	0.0191
tag	206.5000	0.0167
ga	146.3000	0.0017
ttga	93.9000	0.0110
atga	91.1000	0.0100
aa	56.8000	0.0004

build a dictionary for all the regulatory regions in the yeast genome
~6000 promoters

a dictionary of ~1200 words
many known regulatory elements

Words	<Nw>	quality factor
acgcgtcgcgt	5.8	0.9587
tcggcggcta	16.1	0.8927
tagccgccga	18.4	0.8384
cgcgacgcgt	7.3	0.8130
tattacccg	54.4	0.8120
tgggcggcta	12.9	0.8043
tagccgccca	14.7	0.7731
tttgccaccg	11.2	0.7477
gttacccg	71.4	0.7439
gcgatgagct	19.9	0.7372
tcacgtgat	26.9	0.7092
cggtggcaaa	12.7	0.7080
gcgatggg	16.6	0.6900
accccgcg	10.3	0.6873
cattacccg	24.0	0.6870
aggacgcc	14.4	0.6864
ccgggtga	21.8	0.6817
cggcgcgg	11.4	0.6706
gtcacccgg	7.1	0.6480
cgggtaata	40.1	0.6471
gcccgggc	5.8	0.6448
gtcacgtg	36.9	0.6358
gacgggat	13.2	0.6271
gggggggag	11.9	0.6246
cgagccgg	8.1	0.6239
gtgacccgc	4.9	0.6175
cccaccccc	4.9	0.6156
tcccgtcc	10.4	0.6102
cgcgccgg	6.1	0.6099
tctcggtc	12.5	0.5971
ccccccggt	5.9	0.5864
ggggaagg	22.1	0.5819
ccggagtg	9.3	0.5807
tgtgcgtg	18.2	0.5692
gagcccag	11.3	0.5666

Table 1. Known cell cycle sites and some metabolic sites that match words from our genomewide dictionary

MCB	ACGCGT	AAACGCGT ACGCGTCGCGT CGCGACGCGT TGACGCGT
SCB	CRCGAAA	ACGCGAAA
SCB'	ACRMSAAA	ACGCGAAA ACGCCAAA AACGCCAA
Swi5	RRCCAGCR	GCCAGCG GCAGCCAG
SIC1	GCSCRGC	GCCCAGCC CCGCGCGG
MCM1	TTWCCYAAWNNGGWAA	TTTCCNNNNNNGGAAA
NIT	GATAAT	TGATAATG
MET	TCACGTG	RTCACGTG TCACGTGM CACGTGAC CACGTGCT
PDR	TCCGCGGA	TCCGCGG
HAP	CCAAY	AACCCAAC
MIG1	KANWWWWATSYGGGGW	TATATGTG CATATATG GTGGGGAG
GAL4	$CGGN_{11}CCG$	$CGGN_{11}CCG$

our dictionary vs. known TF binding sites

Yeast promoter database 443 non-redundant sites
(Zhu and Zhang, cold spring harbor)

	# of matches	Expected (standard deviation)
Our dictionary	**114**	25 (4.8)
Scrambled dictionary	**33**	14 (3.3)
Brazma et al.	**30**	9 (2.9)

Compare to other motif-finding algorithms

Wconsensus Gibbs-Sampler MEME

✓ do not need to group genes

✓ handle large data set (~10^7) many motifs (~10^3)

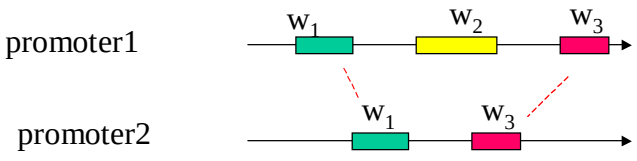
✓ exact solution

✓ need to generalize to handle fuzzy motifs

Measure similarity between regulatory regions

Protein sequence: homologous proteins with similar function detected by sequence alignment

Regulatory regions: genes likely to be regulated similarly generally cannot be detected by alignment



P = prob. by chance the two sequences share a set of motifs

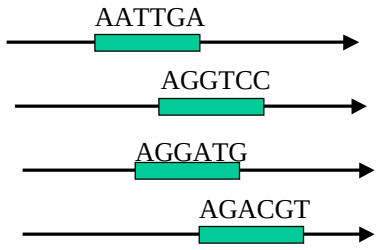
Similarity search using PUP2 (proteasome subunit) as a target

ORF_name	gene	function
YNR032C-A	?	
YDR427W	RPN9	Subunit of the regulatory particle of the proteasome
YHR027C	RPN1	Subunit of 26S Proteasome (PA700 subunit)
YOR260W	gcd1	translation initiation factor eIF2b gamma subunit negative regulator in the general control of amino acid biosynthesis
YGL010W	?	
YDL147W	RPN5	Subunit of the regulatory particle of the proteasome
YDL126C	CDC48	microsomal ATPase
YIL075C	RPN2	RPN2p is a component of the 26S proteosome
YLR095C	?	
YAL054C	acs1	inducible acetyl-coenzyme A synthetase
YDL007W	RPT2	(putative) 26S protease subunit
YER012W	pre1	22.6 kDa proteasome subunit
YER094C	PUP3	20S proteasome subunit beta3_sc
YER177W	bmh1	Homolog of mammalian 14-3-3 proteins
YJL001W	PRE3	Subunit of 20S proteasome
YNL155W	?	
YOR052C	?	
YOR362C	PRE10	proteasome component YC1 (protease yscE subunit 1)
YOR117W	RPT5	26S protease regulatory subunit
YPR109W	?	
YLR341W	?	
YOR073W	?	
YOR078W	?	
YMR153W	NUP53	Component of karyopherin docking complex of the nuclear pore complex

Similarity search using MET32 as target

ORF	gene	function
YDR254W	CHL4	Protein necessary for stability of ARS-CEN plasmids suggested to be required for kinetochore function
YLR092W	SUL2	high affinity sulfate permease
YJL060W	?	
YDR502C	sam2	S-adenosylmethionine synthetase
YIR017C	MET28	Transcriptional activator of sulfur amino metabolism
YIR018W	YAP5	bZIP protein\; transcription factor
YDL059C	RAD59	A mutation in this gene results in RADiati sensitivity and recombination defects
YJL101C	GSH1	gamma-glutamylcysteine synthetase
YAL018C	?	
YJR010W	met3	ATP sulfurylase
YNL037C	IDH1	alpha-4-beta-4 subunit of mitochondrial isocitrate dehydrogenase
YFR030W	met10	subunit of assimilatory sulfite reductase
YDR007W	trp1	n-(5'-phosphoribosyl)-anthranilate isomera
YER042W	?	
YAR064W	?	
YEL043W	?	
YER091C-A	?	
YER092W	?	
YGR155W	CYS4	Cystathionine beta-synthase
YIL046W	MET30	Met30p contains five copies of WD40 motif interacts with and regulates Met4p
YDR090C	?	
YDR438W	?	
YKR069W	met1	siroheme synthase
YPL189W	?	
YIL126W	STH1	helicase related protein, snf2 homolog

Generalizing the segmentation model
To include fuzzy motifs



position frequency matrix

	1	2	3	4	5	6
A	4	1	1	1	0	1
C	0	0	0	1	1	1
G	0	3	2	0	2	1
T	0	0	1	2	1	1

$\Leftarrow f_{i\sigma}$

including fuzzy words in our dictionary

dictionary

$w \rightarrow P_w$
$m \rightarrow P_m$ and $f_{i\sigma}$

likelihood

$$Z = \sum_{Seg} \prod_{w} (p_w)^{N_w(Seg)} \prod_{m} (p_m)^{N_m(Seg)} \prod_{i} (f_{i\sigma}(m))^{n_{i\sigma}(m)}$$

self consistent equations

$$p_w = \frac{\langle N_w \rangle}{\sum_{w'} \langle N_{w'} \rangle}$$

$$f_{i\sigma} = \frac{\langle n_{i\sigma} \rangle}{\sum_{\sigma'} \langle n_{i\sigma'} \rangle}$$

the Pavzner challenge

20 sequences each 600pb long
implant one motif 15 bp, 4 mutations
MEME, Gibbs, Consensus fail, all need to find alignment path

out segmentation algorithm with matrix
easily find length 15 motif with 0.75 polarization
even if the sequence is 2000 bp long

local minimum
funnel shape landscape

Analytical theory??

regulatory elements detection using quantitative gene expression data

Simple assumption: if word W is a responsive element, a gene's expression level should correlate with the occurrence of W

$$\chi = \left[\log(E_g) - \sum_w F_w N_g(w) \right]^2$$

E_g Expression level for gene g

$N_g(w)$ Number of occurences of W in g's promoter

F_w Fitting parameter for word W

Mechanistic interpretation

✓ RNA concentration is determined by trans. Rate

✓Rate limiting step is the assembly of RNA polymerase

✓Trans. Factors contribute additively to the energetics

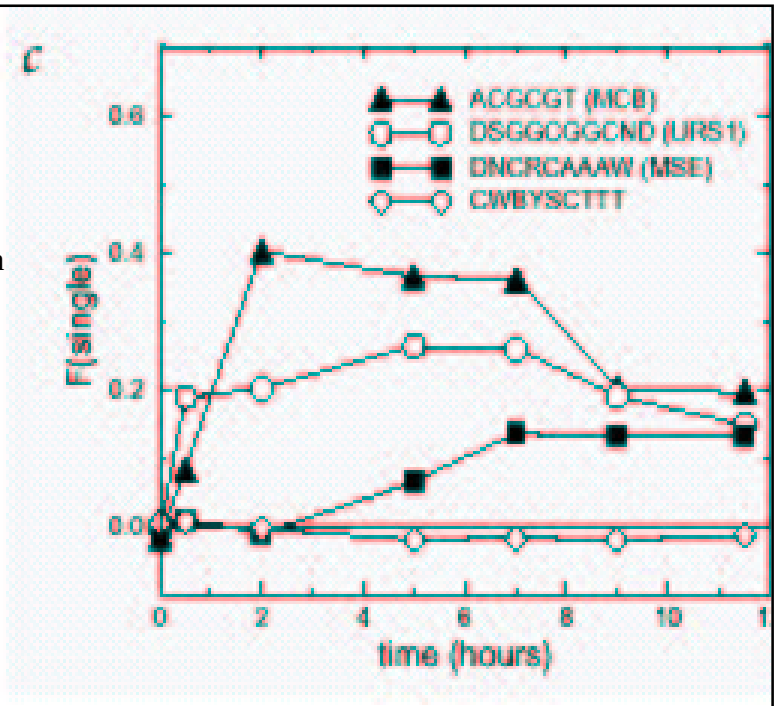
Fitting cell cycle data

Table 2 • Final result of the iterative motif finding procedure

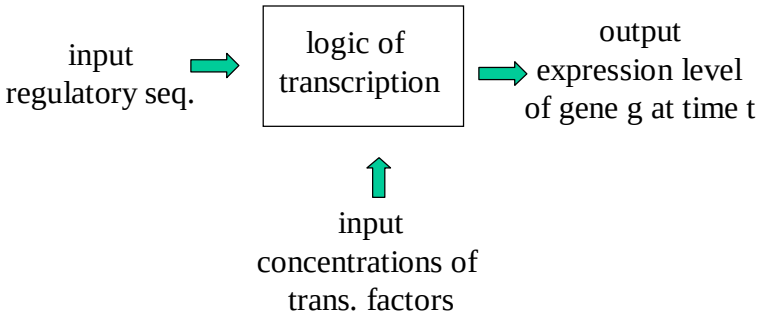
Motif	$\Delta\chi^2$	*F*(single)	*F*(multi)	Matches	ORFs
AAAATTT	0.033534	−0.119555	−0.080316	1564	1331
ACGCGT	0.024535	0.209973	0.211215	327	289
AGGGG	0.019754	0.105028	0.101450	1065	907
CGATGAG	0.022773	−0.249775	−0.200283	251	243
CTCATCG	0.014836	−0.206987	−0.179062	241	235
CCTCGAC	0.008866	0.350493	0.323390	49	48
CCCCT	0.007516	0.061382	0.060757	1146	954
TAAACAA	0.003290	−0.060218	−0.069649	610	565
ATTTTT	0.009661	−0.032125	−0.021880	5260	3167
TGACG	0.008097	0.068384	0.053070	1145	1012
TGAAAA	0.008472	−0.041577	−0.030628	3139	2325

Using a *P* value cutoff of 0.01, a model containing 11 motifs is constructed, using the same expression data as in Table 1.

Fitting Sporulation data



Challenge: derive better models to account for quantitative expression data



combine computational approaches with experiments

One example

ER unfolded protein pathway

unfold ← treated by Tm

↓

IRE1 ← dimerize

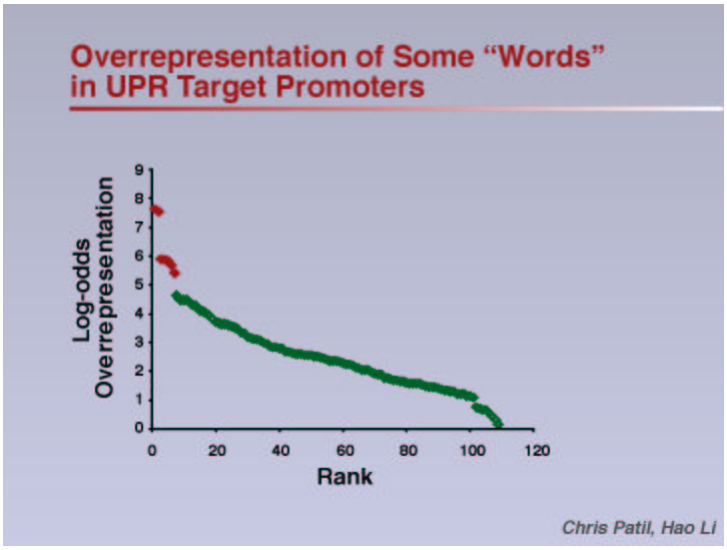
↓

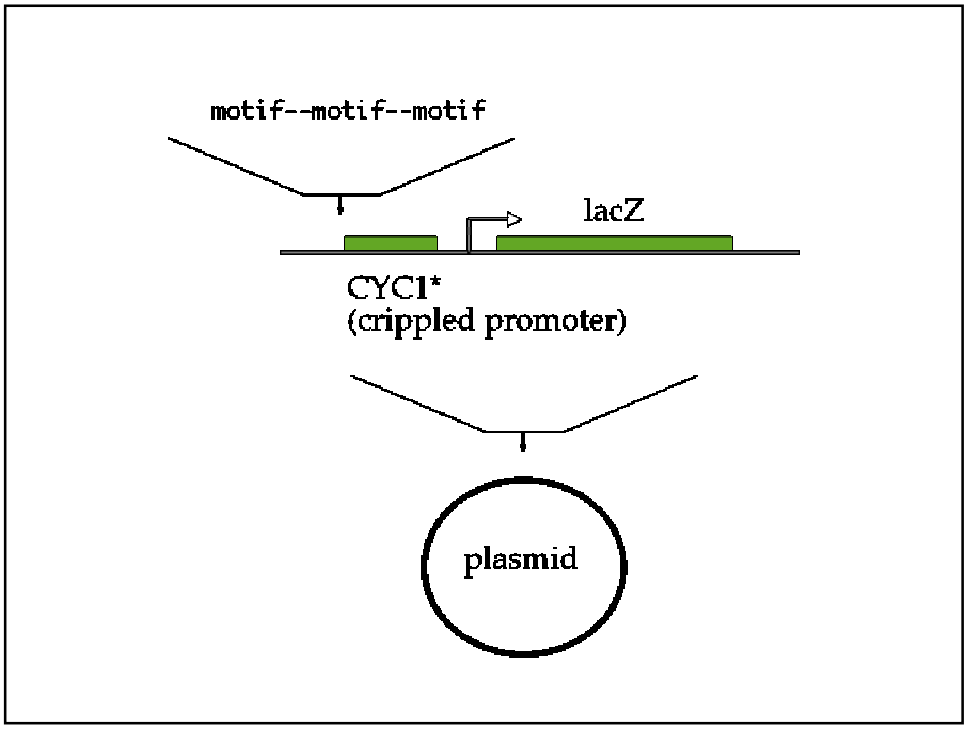
HAC1 ← trans. factor

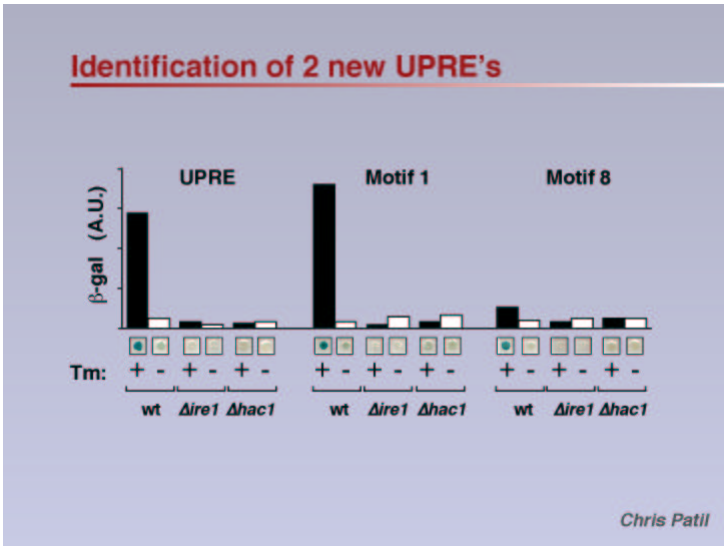
↓

target genes

Problem: only ~20 out of ~300 responding genes have the known UPRE element




motif--motif--motif
lacZ
CYC1*
(crippled promoter)
plasmid


Identification of 2 new UPRE's
UPRE
Motif 1
Motif 8
β-gal (A.U.)
Tm:
+ - + - + -
wt Δire1 Δhac1
Chris Patil

Summary

Complete genome sequences

+

Genomic scale gene expression data

⇩

Opportunities and challenges to decipher
Information in the non-coding regions

Theoretical/computational approaches

- ✓ Statistical analysis pattern discovery– Mobydick algorithm
- ✓ Detecting sequence features using expression data
- ✓ Theoretical predictions can direct new experiments

acknowledgement

Eric Siggia, Rockefeller
Harmen Bussemaker, Univ. Amsterdam
Ming Zhang, Univ. of Chicago
Chris Patil, UCSF
Peter Walter, UCSF
Virgil Rhodius, UCSF
Carol Gross, UCSF

References available from
http://mobydick.ucsf.edu/~haoli