

Bioinformatics Approaches for Studying Transcription Regulation and Protein-DNA Interactions

Xiaole Shirley Liu

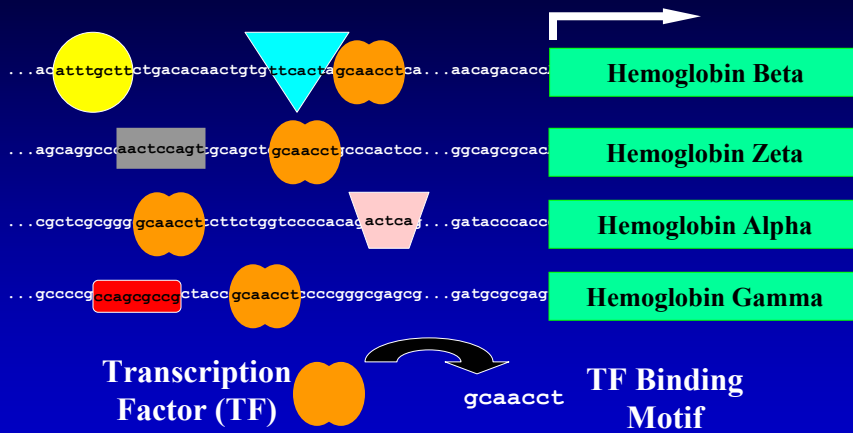
8/11/2003

Outline

- **Biology of transcription regulation**
- **Scan for known TF motif sites**
- **De novo method**
 - Regular expression enumeration
 - Position weight matrix update
 - Using microarray to guide motif search
- **Practical issues in motif finding**
 - Lower organisms
 - Higher eukaryotes

*Copyright of
X. Shirley Liu
8/11/2003*

Biology of Transcription Regulation



Motif can only be computational discovered when there are enough cases for machine learning

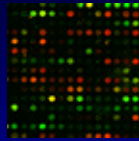
Copyright of
X. Shirley Liu
8/11/2003

Goal of Understanding Regulation

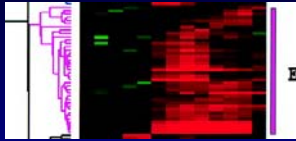
- Which TFs are involved in the regulation?
- What are the binding motifs of these TFs?
- Does a TF enhance / repress gene expression?
- Which genes are regulated by this TF?
- Are there binding partner / competitor for a TF?
- Why there is disease when a TF went wrong?

Copyright of
X. Shirley Liu
8/11/2003

Gene Expression Profile Clusters



Gene Expression



Profile Cluster



Upstream Motif Finding

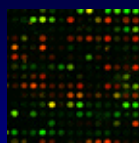
Upstream Regions **Co-expressed Genes**

GATGGCTGCACCACGTGTATGC . . . ACG	Pho 5
CACATCGCATCACGTGACCAGT . . . GAC	Pho 8
GCCTCGCACGTGGTGGTACAGT . . . AAC	Pho 81
TCTCGTTAGGACCATCACGTGA . . . ACA	Pho 84
CGCTAGCCACGTGGATCTTGT . . . AGA	Pho ...

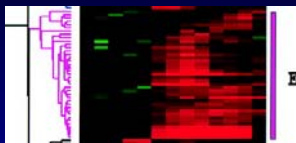
Transcription Start Site (TSS)

Copyright of
X. Shirley Liu
8/11/2003

Gene Expression Profile Clusters



Gene Expression



Profile Cluster



Upstream Motif Finding

Upstream Regions **Co-expressed Genes**

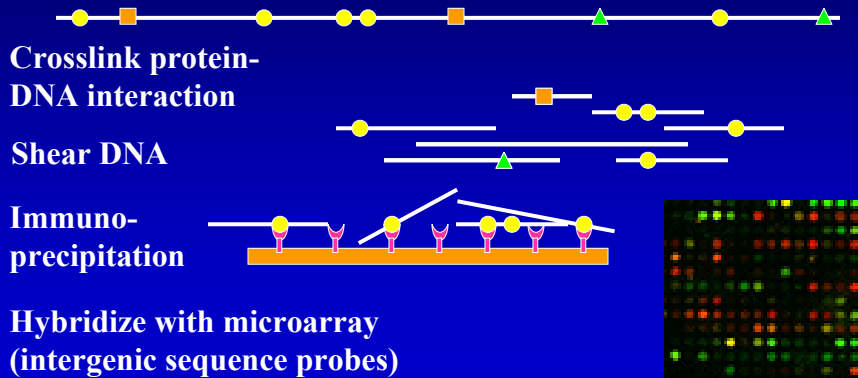
GATGGCTGCAC CACGTG TATGC . . . ACGATG TCTCGC
CACATCGCAT CACGTG ACCAGT . . . GACATGG GACGGC
GCCTCG CACGTG GTGGTACAGT . . . AACATG ACTAA
TCTCGTTAGGACCAT CACGTG A . . . ACAATG GAGGCG
CGCTAGCC CACGTG GATCTTGT . . . AGAATGG CCCTAT

Pho4 binding

Copyright of
X. Shirley Liu
8/11/2003

ChIP-chip Experiments

- Chromatin immunoprecipitation + microarray (Chromatin IP, ChIP-array, ChIP-chip)
- Detects *in vivo* protein-DNA interaction



right of
X. Shirley Liu
8/11/2003

Computational Motif Finding

- **Input data:**
 - Upstream sequences of gene expression profile cluster or ChIP-chip selected sequences
 - 20-800 sequences, each 300-5000 bps long
- **Output: enriched sequence patterns (motifs)**
- **Challenges:**
 - False positive sequences, variable sites / sequence
 - Motif sites have substitution from consensus
 - Many non-functional repeats
 - Many motifs may be involved

Copyright of
X. Shirley Liu
8/11/2003

Scan for Known TF Motif Sites

- TRANSFAC database: experimental TF sites
- Motif representation:

– Regular expression: Consensus CACAAAA
 Degenerate CRCAAAW
 IUPAC A/G A/T

Copyright of
 X. Shirley Liu
 8/11/2003

Scan for Known TF Motif Sites

- TRANSFAC database: experimental TF sites
- Motif representation:

– Regular expression: Consensus CACAAAA
 Degenerate CRCAAAW
 – Position weight matrix (PWM): A/G A/T

Motif Matrix

Pos	A	C	G	T	Con
1	0.9	0	0	0.1	A
2	0	0.1	0.2	0.7	T
3	0	0.1	0.7	0.2	G
4	0.1	0.1	0.8	0	G
5	0	0.7	0.1	0.2	C
6	0.8	0	0.2	0	A
7	0	0.3	0	0.7	T
8	0	0	0.8	0.2	G

Segment ATGCAGCT score =

$$\frac{p(\text{generate ATGCAGCT from motif matrix})}{p(\text{generate ATGCAGCT from background})}$$

$$p_0^A \times p_0^T \times p_0^G \times p_0^C \times p_0^A \times p_0^G \times p_0^C \times p_0^T$$

Copyright of
 X. Shirley Liu
 8/11/2003

De novo Sequence Motif Finding

- **Goal: look for common sequence patterns enriched in the input data (compared to the genome background)**
- **Regular expression enumeration**
 - Pattern driven approach
 - Enumerate patterns, check significance in dataset
- **Position weight matrix update**
 - Data driven approach
 - Initialize random matrices, use dataset to refine
- **Using microarray measures to guide motif search**
 - Motif occurrence best correlated with expression

*Copyright of
X. Shirley Liu
8/11/2003*

Regular Expression Enumeration

- **For every oligonucleotide w check for over-representation:**
 - Expected w occurrence in data
 - Consider genome sequence + current data size
 - Observed w occurrence in data
 - Over-represented w is potential TF binding motif
- **Exhaustive, guaranteed to find global optimum, and can find multiple motifs**
- **Not as flexible with base substitutions, long list of similar good motifs, and limited with motif width**

*Copyright of
X. Shirley Liu
8/11/2003*

RE Enumeration Derivatives

- oligo-analysis, spaced dyads $w_1 \cdot n_s \cdot w_2$ (J van Helden)
- IUPAC alphabet, Markov background (M Tompa)
- Suffix trie (A Brazma) to represent all intergenic sequences, prune nodes with two few sites
- 2-bit encoding (P Baldi), fast index access
- WINNOWER (P Pevzner), find cliques in a graph
- MobyDick (H Bussemaker), build long motifs from shorter ones

*Copyright of
X. Shirley Liu
8/11/2003*

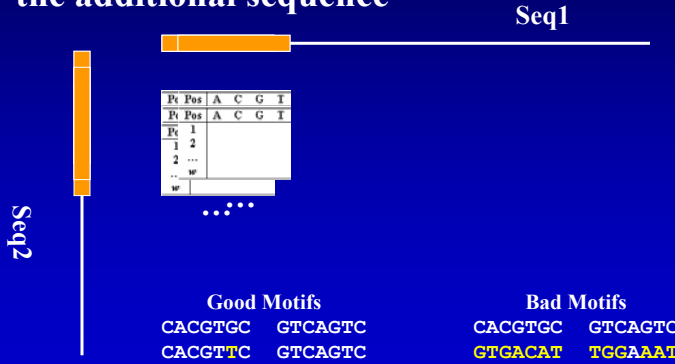
De novo Sequence Motif Finding

- **Goal:** look for common sequence patterns enriched in the input data (compared to the genome background)
- **Regular expression enumeration**
 - Pattern driven approach
 - Enumerate patterns, check significance in dataset
- **Position weight matrix update**
 - Data driven approach
 - Initialize random matrices, use dataset to refine
- **Using microarray measures to guide motif search**
 - Motif occurrence best correlated with expression

*Copyright of
X. Shirley Liu
8/11/2003*

Consensus

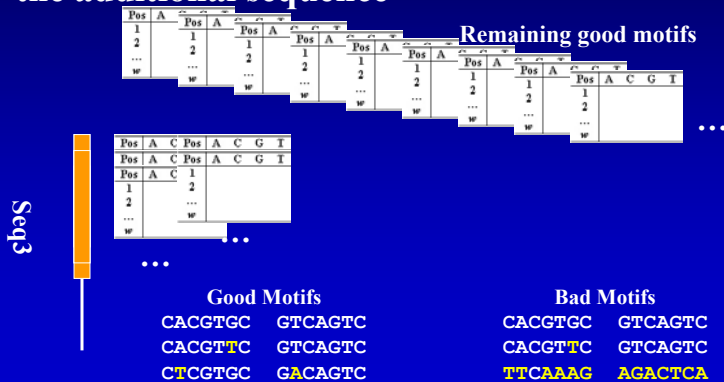
- Starting from the 1st sequence, add one sequence at a time to look for the best motifs obtained with the additional sequence



Copyright of
X. Shirley Liu
8/11/2003

Consensus

- Starting from the 1st sequence, add one sequence at a time to look for the best motifs obtained with the additional sequence



Copyright of
X. Shirley Liu
8/11/2003

Consensus

- Starting from the 1st sequence, add one sequence at a time to look for the best motifs obtained with the additional sequence
- G Stormo, algorithm runs very fast
- Sequence order plays a big role in performance
 - First two sequences better contain the motif
 - Sites stop accumulating at the first bad sequence
 - Newer version allowing [0-n] is much slower

*Copyright of
X. Shirley Liu
8/11/2003*

Expectation Maximization and Gibbs Sampling Model

- **Objects:**
 - Seq: sequence data to search for motif
 - θ_0 : non-motif (genome background) probability
 - θ : motif probability matrix parameter
 - π : unknown variable, site locations
- **Problem:** $P(\theta, \pi \mid \text{seq}, \theta_0)$
- **Approach:** alternately estimate
 - π by $P(\pi \mid \theta, \text{seq}, \theta_0)$
 - θ by $P(\theta \mid \pi, \text{seq}, \theta_0)$
 - EM and Gibbs differ in the estimation methods

*Copyright of
X. Shirley Liu
8/11/2003*

Expectation Maximization

- E step: $\pi \mid \theta, \text{seq}, \theta_0$

TTGACGACTGCACGT

TTGAC P_1

TGACG P_2

GACGA P_3

ACGAC P_4

CGACT P_5

GACTG P_6

ACTGC P_7

CTGCA P_8

...

$P_1 = \text{likelihood ratio} =$

$$\frac{P(\text{TTGAC} \mid \theta)}{P(\text{TTGAC} \mid \theta_0)}$$

Pos	A	C	G	T
1	0.7	0.1	0.01	0.2
2	0.01	0.01	0.8	0.1
3	0.32	0.02	0.3	0.18
4	0.03	0.42	0.1	0.47
5	0.2	0.5	0.1	0.2

$$p_0T \times p_0T \times p_0G \times p_0A \times p_0C$$

$$= 0.3 \times 0.3 \times 0.2 \times 0.3 \times 0.2$$

Copyright of
X. Shirley Liu
8/11/2003

Expectation Maximization

- E step: $\pi \mid \theta, \text{seq}, \theta_0$

TTGACGACTGCACGT

TTGAC P_1

TGACG P_2

GACGA P_3

ACGAC P_4

CGACT P_5

GACTG P_6

ACTGC P_7

CTGCA P_8

...

- M step: $\theta \mid \pi, \text{seq}, \theta_0$

- $\text{argmax}_\theta p(\pi \mid \theta, \text{seq}, \theta_0)$

$P_1 \times \text{TTGAC}$

$P_2 \times \text{TGACG}$

$P_3 \times \text{GACGA}$

$P_4 \times \text{ACGAC}$

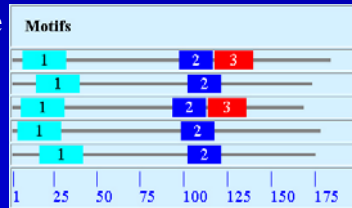
...

- θ reflects weighted average of π

Copyright of
X. Shirley Liu
8/11/2003

EM Derivatives

- **First EM motif finder (C Lawrence)**
 - Deterministic algorithm, guarantee local optimum
- **Random projection (J. Buhler)**
 - Sample h / w positions, hash words agreeing at h into bucket
 - Run EM on buckets with enough size
- **MEME (TL Bailey)**
 - Prior probability allows 0-n site / sequence
 - Parallel running multiple EM with different seed
 - User friendly results



Copyright of
X. Shirley Liu
8/11/2003

Gibbs Sampling

- **Stochastic process, although still may need multiple initializations**
 - Sample π from $P(\pi | \theta, \text{seq}, \theta_0)$
 - Sample θ from $P(\theta | \pi, \text{seq}, \theta_0)$
- **Collapsed form:**
 - θ estimated with counts, not sampling from Dirichlet
 - Sample site from one seq based on sites from other seqs
- **Converged motif matrix θ and converged motif sites π represent stationary distribution of a Markov Chain**

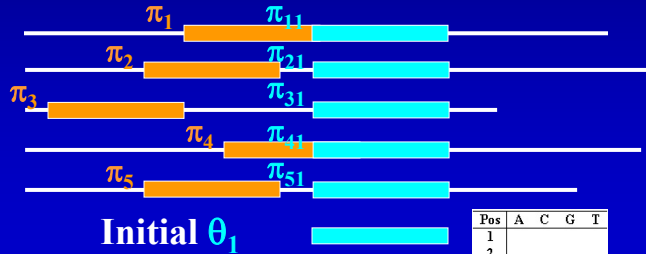
Copyright of
X. Shirley Liu
8/11/2003

Gibbs Sampler

- Randomly initialize a probability matrix

$$p_{A1} = \frac{n_{A1} + s_A}{n_{A1} + s_A + n_{C1} + s_C + n_{G1} + s_G + n_{T1} + s_T}$$

θ estimated with counts

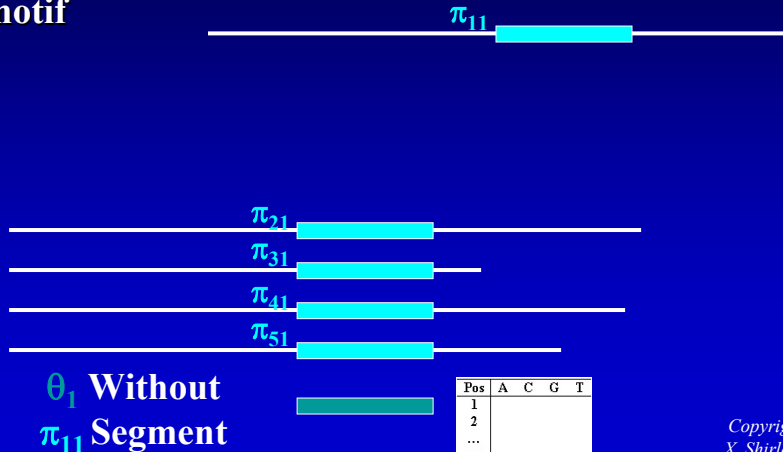


Pos	A	C	G	T
1				
2				
...				
1P				

Copyright of
X. Shirley Liu
8/11/2003

Gibbs Sampler

- Take out one sequence with its sites from current motif

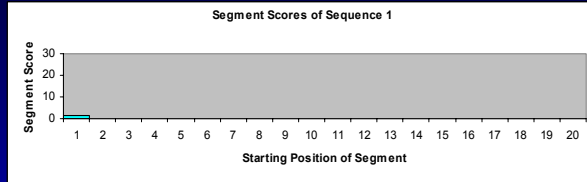


Pos	A	C	G	T
1				
2				
...				
1P				

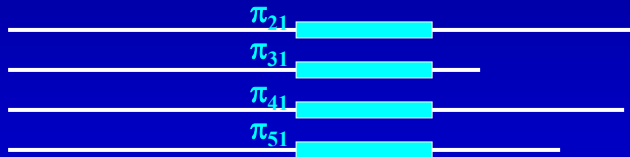
Copyright of
X. Shirley Liu
8/11/2003

Gibbs Sampler

- Score each possible segment of this sequence



Segment (1-6): 1.5 Sequence 1



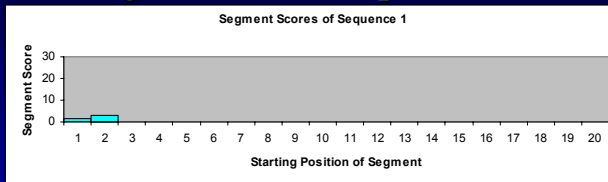
θ_1 Without
 π_{11} Segment

Pos	A	C	G	T
1				
2				
...				
w				

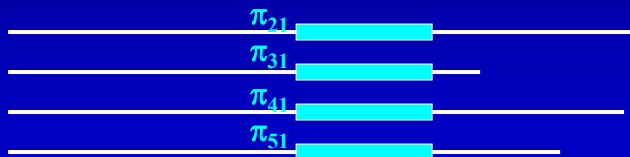
Copyright of
X. Shirley Liu
8/11/2003

Gibbs Sampler

- Score each possible segment of this sequence



Segment (2-7): 3 Sequence 1



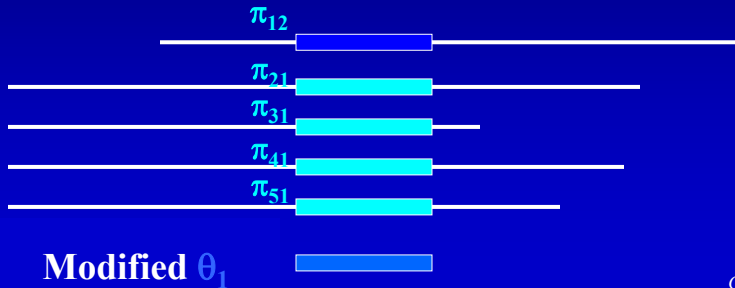
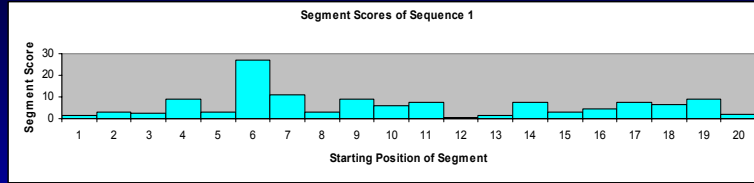
θ_1 Without
 π_{11} Segment

Pos	A	C	G	T
1				
2				
...				
w				

Copyright of
X. Shirley Liu
8/11/2003

Gibbs Sampler

- Sample site from one seq based on sites from other seqs

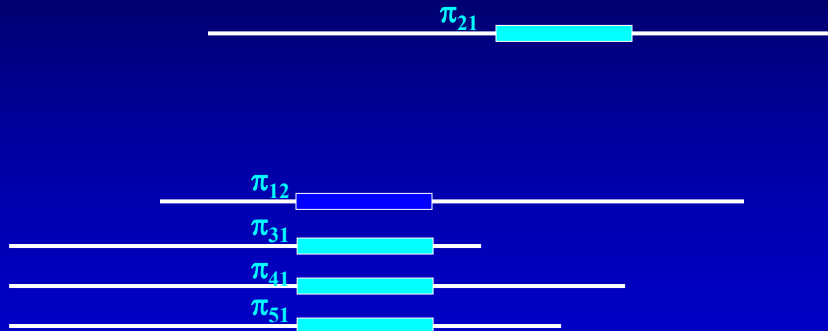


θ estimated with counts

Copyright of
X. Shirley Liu
8/11/2003

Gibbs Sampler

- Repeat the process until motif converges



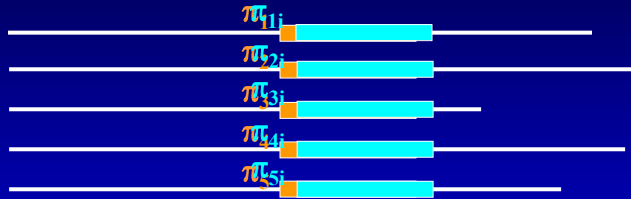
θ_1 Without
 π_{21} Segment

Pos	A	C	G	T
1				
2				
...				
π				

Copyright of
X. Shirley Liu
8/11/2003

Gibbs Sampler

- Column shift



- Metropolis algorithm:

- Propose π^* as π shifted 1 column to left or right
- Calculate motif score $u(\pi)$ and $u(\pi^*)$
- Accept π^* with prob = $\min(1, u(\pi^*) / u(\pi))$

Copyright of
X. Shirley Liu
8/11/2003

Gibbs Sampling Derivatives

- Gibbs Motif Sampler (JS Liu)

- Add prior probability to allow 0-n site / seq
- Sample motif positions to consider

- AlignACE (GM Church)

- Mask out one motif to find more different motifs

- BioProspector (XS Liu)

- Use background model with Markov dependencies
- Sampling with threshold (0-n sites / seq), new scoring function
- Can find two-block motifs with variable gap

Copyright of
X. Shirley Liu
8/11/2003

Position Weight Matrix Update

- **Advantage**
 - Can look for motifs of any widths
 - Flexible with base substitutions
- **Disadvantage:**
 - No guaranteed global optimum

Break!!

*Copyright of
X. Shirley Liu
8/11/2003*

Using microarray measures to guide motif search

- Motif occurrence best correlated with variations in gene expression
- **REDUCE**
- **GMEP**
- **MDscan**
- **Motif Regressor**

*Copyright of
X. Shirley Liu
8/11/2003*

REDUCER

- Incorporating TFBS copy number with microarray values (H Bussemaker)
 - Single microarray experiment (no clustering)
 - Enumerate all possible w -mers
 - Check w -mer copy number in each upstream seq
 - Check downstream expression of every gene
 - See whether there is any correlation

*Copyright of
X. Shirley Liu
8/11/2003*

Example

- Good motif

	Copy #	Expr
Seq1	5	5.31
Seq2	0	0.75
Seq3	4	3.86
Seq4	2	2.47
Seq5	0	0.42
Seq6	1	1.83
...		

- Bad motif

	Copy #	Expr
Seq1	5	1.31
Seq2	0	2.75
Seq3	4	2.86
Seq4	2	1.47
Seq5	0	0.42
Seq6	1	0.83
...		

*Copyright of
X. Shirley Liu
8/11/2003*

REDUCER

- Incorporating TFBS copy number with microarray values (H Bussemaker)
 - Enumerate all possible w -mers
 - Check w -mer copy number in each upstream seq
 - Check downstream expression of every gene
 - See whether there is any correlation
 - More upstream sites, more expression \rightarrow inducer
 - More upstream sites, less expression \rightarrow repressor
- Exhaustive, multiple motifs, global optimum
- Many similar motifs, limited in motif width

Copyright of
X. Shirley Liu
8/11/2003

Genome Mean Expression Profiles

- M Eisen, single microarray experiment
- For each w -mer, find the G genes whose upstream contain the w -mer and calculate GMPEP

$$GMPEP(m) = \frac{\sum_{g \in G} N_{mg} \cdot E_g}{\sum_{g \in G} N_{mg}}$$

- Randomly pick G genes, calculate their expression distribution $N(\mu, \sigma)$
- Report w -mer as potential motif if GMPEP is extremely distributed on $N(\mu, \sigma)$

Copyright of
X. Shirley Liu
8/11/2003

MDscan

- ChIP-chip results insights:
 - High ChIP ranking => true targets
 - Highest ChIP ranking => contain more sites
- Basic strategy:
 - Search TF motif from highest ranking targets first (high signal / background ratio)
 - Refine candidate motifs with all targets
- Deterministic algorithm, very fast (XS Liu)

Copyright of
X. Shirley Liu
8/11/2003

Similarity defined by m -match

For a given w -mer and any other random w -mer

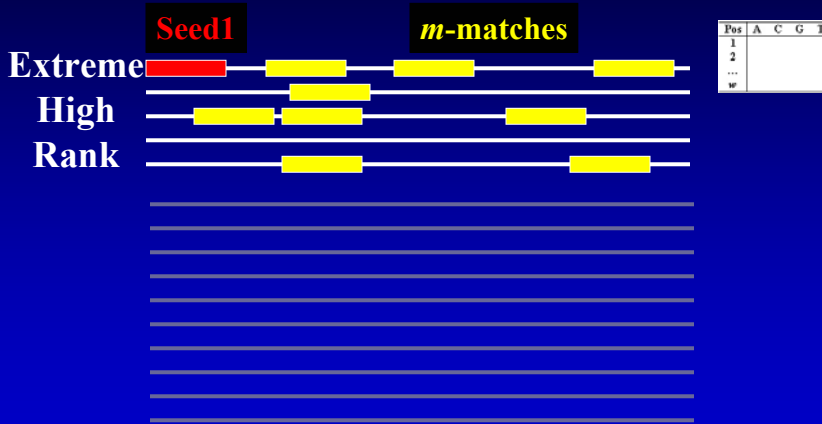
TGTAACGT	8-mer	
TGTAACGT	matched 8	} m -matches for TGTAACGT
AGTAACGT	matched 7	
TGCAACAT	matched 6	
TGACACGG	matched 5	
AATAACAG	matched 4	

Pick a reasonable m to call two w -mers similar

w	5	6	7	8	9	10	11	12	13	14	15
m	5	5	6	6	7	7	8	8	9	9	10

Copyright of
X. Shirley Liu
8/11/2003

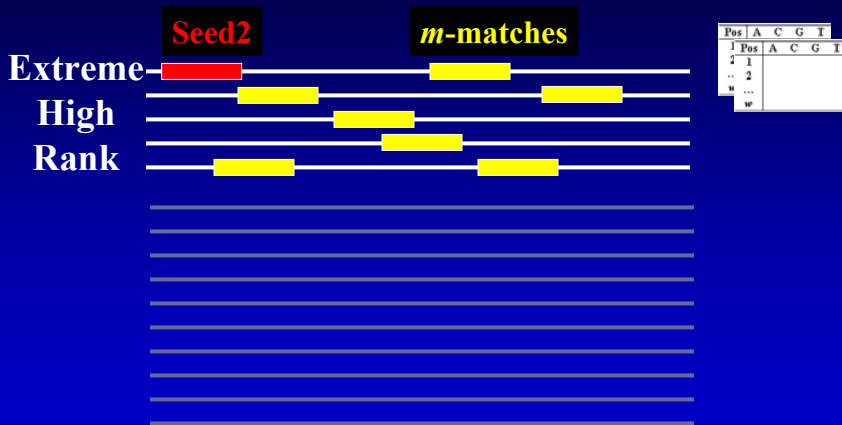
MDscan Algorithm: Finding candidate motifs



All CHIP-selected targets

Copyright of
X. Shirley Liu
8/11/2003

MDscan Algorithm: Finding candidate motifs



All CHIP-selected targets

Copyright of
X. Shirley Liu
8/11/2003

MDscan Algorithm: Scoring candidate motifs

- Motif scoring function:

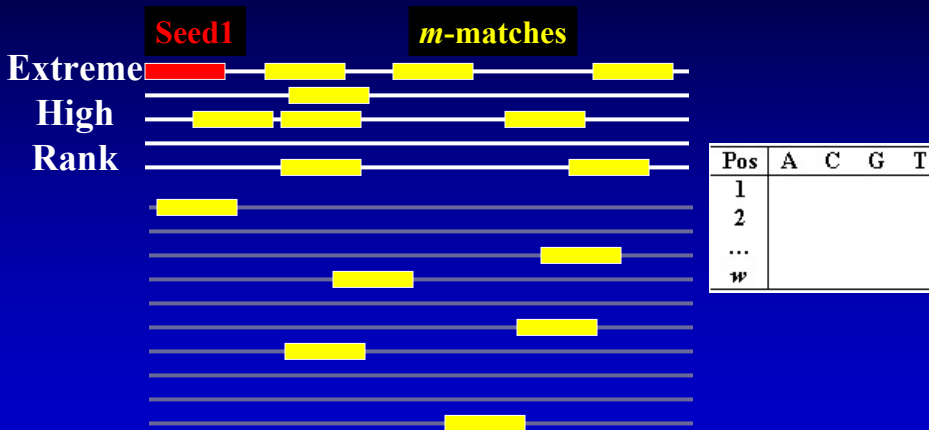
$$\frac{\log(x_m)}{w} \times \left[\sum_{i=1}^w \sum_{j=A}^T p_{ij} \log p_{ij} - \frac{1}{x_m} \sum_{s=1}^{x_m} \log(p_0(s)) \right]$$

Motif Signal
Positions
Specific (unlikely in
Abundant
Conserved
genome background)

- Prefer: conserved motifs with many sites, but are not often seen in the genome background
- Keep best 30-50 candidate motifs

*Copyright of
X. Shirley Liu
8/11/2003*

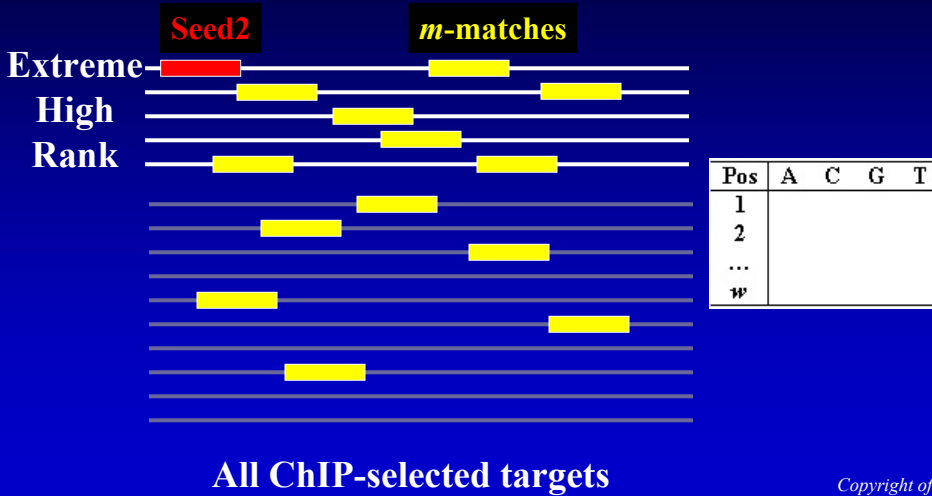
MDscan Algorithm: Update motifs with remaining seqs



All ChIP-selected targets

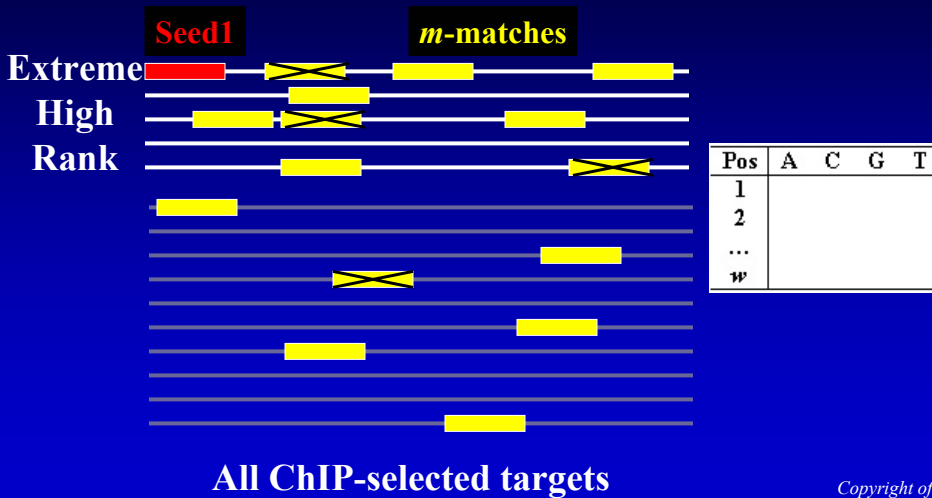
*Copyright of
X. Shirley Liu
8/11/2003*

MDscan Algorithm: Update motifs with remaining seqs



Copyright of
X. Shirley Liu
8/11/2003

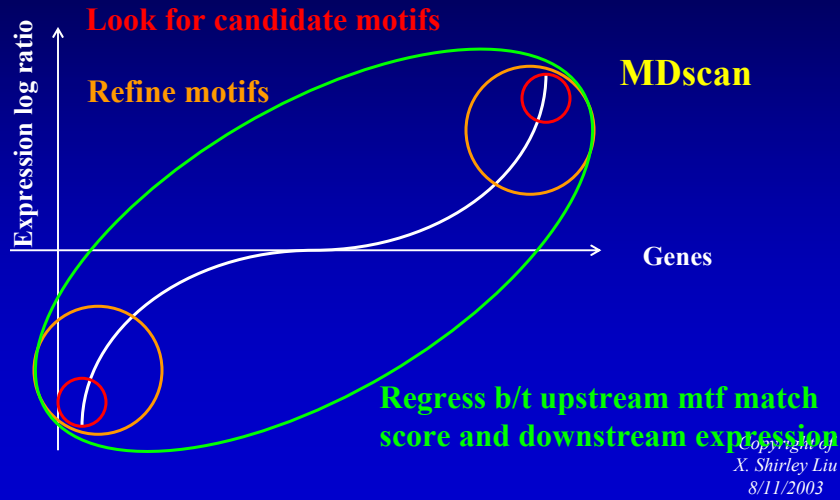
MDscan Algorithm: Refine the motifs



Copyright of
X. Shirley Liu
8/11/2003

Motif Regressor

- EM Conlon

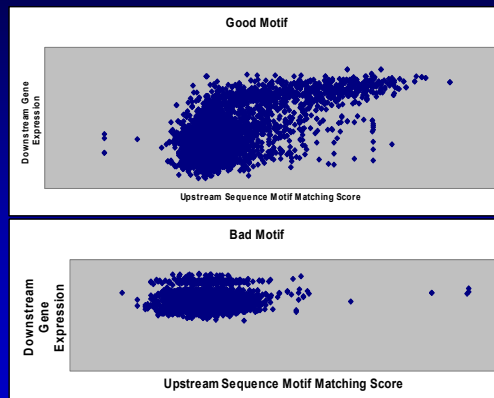


Motif Regressor Rational

- For each TF:

	Upstream Seq Mtf Match	Downstream Gene Exp
Gene1	3.2	1.8
Gene2	2.8	0.3
Gene3...		

- Upstream sequence X motif matching score measures:
 - Number of sites
 - Strength of matching



Motif Regressor Strategy

- Rank genes by \log_2 (expression fold change)
- Try MDscan (width 5-17) on induced and repressed genes separately
 - Find 50 candidate motifs from top 100 genes
 - Refine candidate motifs with top 500 genes
 - Report ≤ 30 distinct motifs
- Score each upstream sequence with each motif
- Linear regression to eliminate insignificant motifs

*Copyright of
X. Shirley Liu
8/11/2003*

Linear Regression Example

Person	IQ	Age	Education	Height	Eye color	Spend/week	# of CD
A	120	30	High	171	blue	\$4000	30
B	250	41	PhD	155	brown	\$1500	18
C	150	8	Grade10	115	black	\$100	90
D	180	16	Grade12	140	gray	\$200	15
E	90	4	Preschool	88	green	\$500	26
F	130	17	High	178	black	\$80	500
G	110	21	College	182	blue	\$800	220
...							
Gene	Express	Mtf1	Mtf2	Mtf3	Mtf4	Mtf5	Mtf6
Single Regression		X	X	X	--	--	--

*Copyright of
X. Shirley Liu
8/11/2003*

Motif Regressor Strategy

- Stepwise regression to find multiple motifs that work together
 - Each step find the motif that explains the remaining expression the best
 - Remove its effects from expression
- Multiple regression model: expression explained as the sum of motifs' effects

$$Y_g = \alpha + \sum_{m=1}^M \beta_m S_{mg} + \epsilon_g$$

Expression of gene g → Y_g ← Error term
 Baseline expression → α ← Regression coefficient → β_m ← Upstream motif- match score → S_{mg}

Copyright of
X. Shirley Liu
8/11/2003

Stepwise Regression Example

Person	IQ	Age	Education	Height	Eye color	Spend/week	# of CD
A	120	30	High	171	blue	\$4000	30
B	250	41	PhD	155	brown	\$1500	18
C	150	8	Grade10	115	black	\$100	90
D	180	16	Grade12	140	gray	\$200	15
E	90	4	Preschool	88	green	\$500	26
F	130	17	High	178	black	\$80	500
G	110	21	College	182	blue	\$800	220

Gene	Express	Mtf1	Mtf2	Mtf3	Mtf4	Mtf5	Mtf6
Single Regression		X	X	X	--	--	--
Stepwise Regression		2	1	--			

Copyright of
X. Shirley Liu
8/11/2003

Outline

- **Biology of transcription regulation**
- **Scan for known TF motif sites**
- **De novo method**
 - Regular expression enumeration
 - Position weight matrix update
 - Using microarray to guide motif search
- **Practical issues in motif finding**
 - Lower organisms
 - Higher eukaryotes

*Copyright of
X. Shirley Liu
8/11/2003*

Motif Finding in Bacteria

- **Promoter sequences are short (200-300 bp)**
- **Motif are usually very long (10-20 bases)**
- **There are many two-block motifs**
 - Sigma factors motifs: two blocks with a variable gap
 - Many HTH proteins binds to palindrome motifs
- **Long motifs are usually very degenerate**
 - Often requires each sequence to contain \geq one site
 - Adding orthologous sequences from other species can aid discovery of weak motifs

*Copyright of
X. Shirley Liu
8/11/2003*

Motif Finding in Lower Eukaryotes

- Upstream sequences longer (800 bp), with some simple repeats
- Motif width varies (5 – 17 bases)
- Expression clusters provide decent input sequences quality for TF motif finding
- Motif combination appears, although single motifs are usually significant enough for identification

*Copyright of
X. Shirley Liu
8/11/2003*

Motif Finding in Higher Eukaryotes

- Upstream sequences very long (3KB-20KB), TF motif could appear downstream
- Usually [-800, 200] TSS has the highest TF density, finding TSS is critical (RefSeq)
- Motifs are usually short (6-12 bases), and work in combination, so individual motif may not be significant
- Needs to run RepeatMasker to remove simple repeats
- Gene expression cluster not good enough input

*Copyright of
X. Shirley Liu
8/11/2003*

Comparative Genomics

- **TF sites across species are more conserved than random background due to functional constraints, comparative genomics can narrow down the search space**
- **Many genes across 2 species (WW Wasserman)**
 - Align orthologous sequences (Vista, LAGAN, Pipmaker, Bayes block aligner)
 - get rid of sequence too mutated between species before running motif finding algorithms
- **One gene across multiple species, phylogenetic foot printing**
 - Zoo project (E Rubin)
 - Phylogenetic foot printer (M Tompa)

*Copyright of
X. Shirley Liu
8/11/2003*

Motif Site Clusters

- **Higher eukaryote TF motif sites appear in clusters**
- **Use scanning or de novo methods to find individual TF motifs**
 - Check whether there are site combinations always occurring in close proximity (M Levine)
 - If known one motif, check sites appearing near the known motif sites, e.g. E2F motif
 - Consider comparative genomics as well

*Copyright of
X. Shirley Liu
8/11/2003*

Summary

- **Understanding transcription regulation is important**
- **Scan for known TF motif sites, TRANSFAC**
- **De novo method**
 - Regular expression enumeration
 - Position weight matrix update (consensus, EM, Gibbs)
 - Using microarray to guide motif search (REDUCER, GMEP, MDscan, Motif Regressor)
- **Practical issues in motif finding**
 - Lower organisms
 - Higher eukaryotes (Comparative genomics, motif site clusters)
- **Despite wide computational studies on transcription regulation, we are far from reaching the goal**
- **Questions: xslu@jimmy.harvard.edu**

*Copyright of
X. Shirley Liu
8/11/2003*