

# 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

---

---

---

---

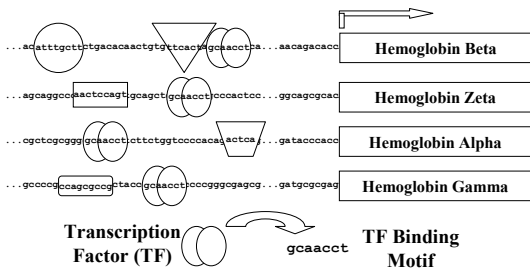
---

---

---

---

## 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

---

---

---

---

---

---

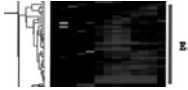
---

---

## Gene Expression Profile Clusters



Gene Expression



Profile Cluster



Upstream Motif Finding

Upstream Regions	Co-expressed Genes
GATGGCTGCACCACGTGTATGC...ACG	Pho 5
CACATCGCATCACGTGACCCAGT...GAC	Pho 8
GCCTCGCACGTGGGTACAGT...AAC	Pho 81
TCTCGTTAGGACCATCACGTGA...ACA	Pho 84
CGCTAGCCACGTGGATCTTGT...AGA	Pho ...

Transcription Start Site  
(TSS)

Copyright of  
X. Shirley Liu  
8/11/2003

---

---

---

---

---

---

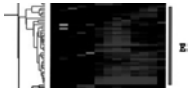
---

---

## Gene Expression Profile Clusters



Gene Expression



Profile Cluster



Upstream Motif Finding

Upstream Regions	Co-expressed Genes
GATGGCTGCACCACGTGTATGC...ACG	Pho 5
CACATCGCATCACGTGACCCAGT...GAC	Pho 8
GCCTCGCACGTGGGTACAGT...AAC	Pho 81
TCTCGTTAGGACCATCACGTGA...ACA	Pho 84
CGCTAGCCACGTGGATCTTGT...AGA	Pho ...

Pho4 binding

Copyright of  
X. Shirley Liu  
8/11/2003

---

---

---

---

---

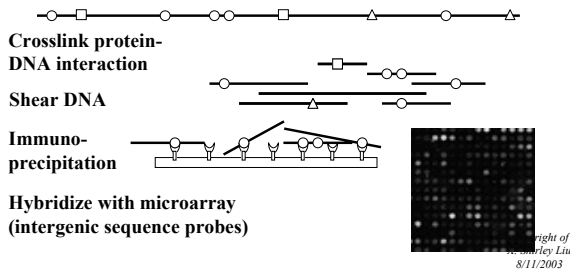
---

---

---

## ChIP-chip Experiments

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




---

---

---

---

---

---

---

---

## 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 CRC AAAW
- Position weight matrix (PWM):  $\begin{matrix} \boxed{A/G} & \boxed{A/T} \end{matrix}$

Pos		Motif Matrix				Con	Segment ATGCAGCT score =
		A	C	G	T		
Pos 12345678	ATGCAGCTG	0.9	0	0	0.1	A	$p(\text{generate ATGCAGCT from motif matrix})$
	AGGGTGGG	0	0.1	0.2	0.7	T	
	ATGCAGCTG	0	0.1	0.2	0.2	G	$p_A^A \times p_A^T \times p_A^G \times p_A^C \times p_A^A \times p_A^G \times p_A^C \times p_A^T$
	TTGGCACGG	0.1	0.4	0.8	0	G	
	ATGGATTTT	0	0.7	0.1	0.2	C	
	ATGCAGCTG	0.8	0	0.2	0	A	
	ATGCAGCTT	0	0.3	0	0.7	T	
	ACTGGATG	0	0	0.8	0.2	G	

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

---

---

---

---

---

---

---

---

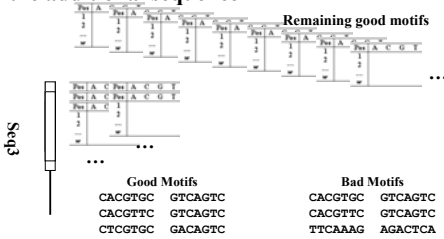
---

---



## Consensus

- Starting from the 1<sup>st</sup> 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 1<sup>st</sup> 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
  - $\theta_0$ : non-motif (genome background) probability
  - $\theta$ : motif probability matrix parameter
  - $\pi$ : unknown variable, site locations
- Problem:  $P(\theta, \pi | \text{seq}, \theta_0)$
- Approach: alternately estimate
  - $\pi$  by  $P(\pi | \theta, \text{seq}, \theta_0)$
  - $\theta$  by  $P(\theta | \pi, \text{seq}, \theta_0)$
  - EM and Gibbs differ in the estimation methods

Copyright of  
X. Shirley Liu  
8/11/2003

---

---

---

---

---

---

---

---

---

---

---

---



# Gibbs Sampling

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

Copyright of  
X. Shirley Liu  
8/11/2003

---

---

---

---

---

---

---

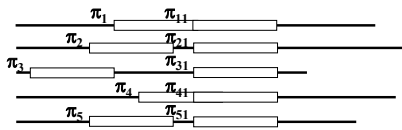
---

# Gibbs Sampler

- Randomly initialize a probability matrix

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

$\theta$  estimated with counts



Initial  $\theta_1$

Pos	A	C	G	T
1				
2				
...				
w				

Copyright of  
X. Shirley Liu  
8/11/2003

---

---

---

---

---

---

---

---

# Gibbs Sampler

- Take out one sequence with its sites from current motif



$\theta_1$  Without  
 $\pi_{11}$  Segment

Pos	A	C	G	T
1				
2				
...				
w				

Copyright of  
X. Shirley Liu  
8/11/2003

---

---

---

---

---

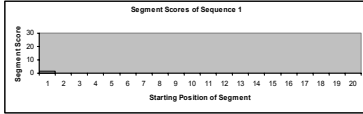
---

---

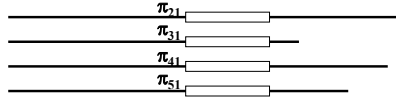
---

# Gibbs Sampler

- Score each possible segment of this sequence



Segment (1-6): 1.5      Sequence 1



$\theta_1$  Without  
 $\pi_{11}$  Segment

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

Copyright of  
X. Shirley Liu  
8/11/2003

---

---

---

---

---

---

---

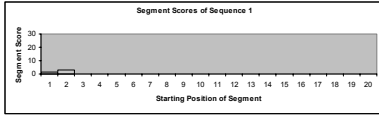
---

---

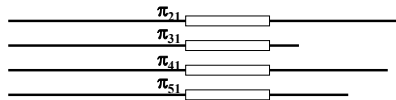
---

# Gibbs Sampler

- Score each possible segment of this sequence



Segment (2-7): 3      Sequence 1



$\theta_1$  Without  
 $\pi_{11}$  Segment

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

Copyright of  
X. Shirley Liu  
8/11/2003

---

---

---

---

---

---

---

---

---

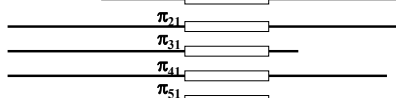
---

# Gibbs Sampler

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



$\pi_{12}$



Modified  $\theta_1$

$\theta$  estimated with counts

Copyright of  
X. Shirley Liu  
8/11/2003

---

---

---

---

---

---

---

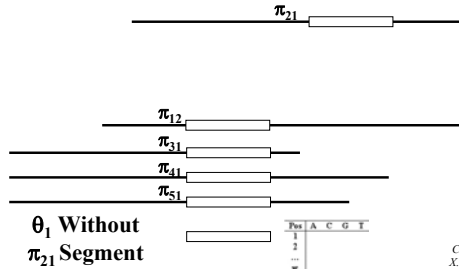
---

---

---

## Gibbs Sampler

- Repeat the process until motif converges



Copyright of  
X. Shirley Liu  
8/11/2003

---

---

---

---

---

---

---

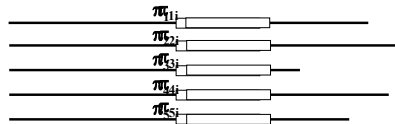
---

---

---

## Gibbs Sampler

- Column shift



- Metropolis algorithm:
  - Propose  $\pi^*$  as  $\pi$  shifted 1 column to left or right
  - Calculate motif score  $u(\pi)$  and  $u(\pi^*)$
  - Accept  $\pi^*$  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			• Bad motif		
	Copy #	Expr		Copy #	Expr
Seq1	5	5.31	Seq1	5	1.31
Seq2	0	0.75	Seq2	0	2.75
Seq3	4	3.86	Seq3	4	2.86
Seq4	2	2.47	Seq4	2	1.47
Seq5	0	0.42	Seq5	0	0.42
Seq6	1	1.83	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 GMEP**

$$GMEP(m) = \frac{\sum_g N_{mg}}{\sum_g N_{mg}} \cdot E_g$$

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

Copyright of  
X. Shirley Liu  
8/11/2003

---

---

---

---

---

---

---

---

# MDscan

- **ChIP dip 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* **nr** and any other random *w* **nr**

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

Pick a reasonable *m* to call two *w*- **nr**s similar

<i>w</i>	5	6	7	8	9	10	11	12	13	14	15
<i>m</i>	5	5	6	6	7	7	8	8	9	9	10

Copyright of  
X. Shirley Liu  
8/11/2003

---

---

---

---

---

---

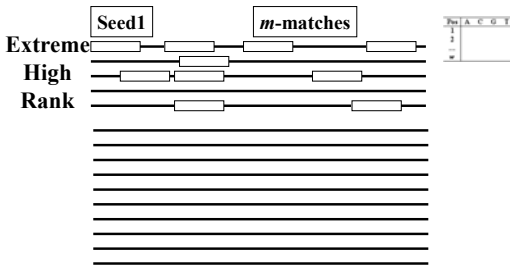
---

---

---

---

# MDscan Algorithm: Finding candidate motifs



All ChIP selected targets

Copyright of  
X. Shirley Liu  
8/11/2003

---

---

---

---

---

---

---

---

---

---

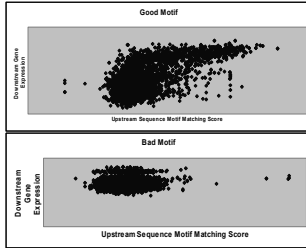




# Motif Regressor Rational

- For each TF:

	Upstream	Downstream
	Seq Mtf Match	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

Copyright of  
X. Shirley Liu  
8/11/2003

---

---

---

---

---

---

---

---

---

---

# Motif Regressor Strategy

- Rank genes by  $\log_2$  (expression fold change)
- Try MDscan (width 5 bp) on induced and repressed genes separately
  - Find 50 candidate motifs from top 100 genes
  - Refine candidate motifs with top 500 genes
  - Report  $\leq 30$  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

---

---

---

---

---

---

---

---

---

---

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

## 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  $g$   $\leftarrow$   $Y_g$   $\leftarrow$   $\alpha$   $\leftarrow$  Baseline expression  
 $\leftarrow$   $\beta_m$   $\leftarrow$  Regression coefficient  
 $\leftarrow$   $S_{mg}$   $\leftarrow$  Upstream motif-match score  
 $\leftarrow$   $\epsilon_g$   $\leftarrow$  Error term

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)
- Motifs are usually very long (10 – 20 bases)
- There are many two-block motifs
  - Sigma factor motifs: two blocks with a variable gap
  - Many HTH proteins bind 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  $\{-80, 200\}$  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

---

---

---

---

---

---

---

---