CSE 527

Autumn 2006

Lectures 6-7
MLE, EM, Expression

Outline

- MLE: Maximum Likelihood Estimators
- EM: the Expectation Maximization Algorithm
- Bio: Gene expression and regulation

- Next: Motif description & discovery

MLE

Maximum Likelihood Estimators

Probability Basics, I

	Ex.	Ex.
Sample Space		
	$\{1, 2, \ldots, 6\}$	$\mathbb{R}$
Distribution		
	$p_1, \ldots, p_6 \geq 0; \sum_{1 \leq i \leq 6} p_i = 1$	$f(x) >= 0; \int_{\mathbb{R}} f(x)dx = 1$
e.g.		
	$p_1 = \cdots = p_6 = 1/6$	$f(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-(x-\mu)^2/(2\sigma^2)}$

pdf, not probability

Probability Basics, II

	Ex.	Ex.
Expectation	$E(g) = \sum_{1 \le i \le 6} g(i)p_i$	$E(g) = \int_{\mathbb{R}} g(x)f(x)dx$
Population		
mean	$\mu = \sum_{1 \le i \le 6} ip_i$	$\mu = \int_{\mathbb{R}} xf(x)dx$
variance	$\sigma^2 = \sum_{1 \le i \le 6} (i-\mu)^2 p_i$	$\sigma^2 = \int_{\mathbb{R}} (x-\mu)^2 f(x)dx$
Sample		
mean	$\bar{x} = \sum_{1 \le i \le n} x_i/n$	
variance	$\bar{s}^2 = \sum_{1 \le i \le n} (x_i - \bar{x})^2/n$	

Parameter Estimation

- Assuming sample $x_1, x_2, ..., x_n$ is from a parametric distribution f(x|θ), estimate θ.
- E.g.:

$$f(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-(x-\mu)^2/(2\sigma^2)}$$

$$\theta = (\mu, \sigma^2)$$

Maximum Likelihood Parameter Estimation

- One (of many) approaches to param. est.
- *Likelihood* of (indp) observations $x_1, x_2, ..., x_n$

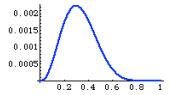
$$L(x_1, x_2, \ldots, x_n \mid \theta) = \prod_{i=1}^{n} f(x_i \mid \theta)$$

- As a function of θ, what θ maximizes the likelihood of the data actually observed
- Typical approach: $\frac{\partial}{\partial\theta} L(\vec{x} \mid \theta) = 0$ or $\frac{\partial}{\partial\theta} \log L(\vec{x} \mid \theta) = 0$

Example 1

n coin flips, $x_1, x_2, ..., x_n$; n_0 tails, n_1 heads, $n_0 + n_1 = n$;

θ = probability of heads



$$L(x_1, x_2, \ldots, x_n \mid \theta) = (1-\theta)^{n_0}\theta^{n_1}$$

$$\log L(x_1, x_2, \ldots, x_n \mid \theta) = n_0 \log(1-\theta) + n_1 \log\theta$$

$$\frac{\partial}{\partial\theta} \log L(x_1, x_2, \ldots, x_n \mid \theta) = \frac{-n_0}{1-\theta} + \frac{n_1}{\theta}$$

Setting to zero and solving:

$$\boxed{\theta = \frac{n_1}{n}}$$

(Also verify it's max, not min, & not better on boundary)

Ex. 2: $x_i \sim N(\mu, \sigma^2),\ \sigma^2 = 1,\ \mu$ unknown

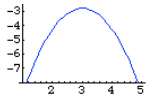
$$L(x_1, x_2, \ldots, x_n|\theta) = \prod_{1\le i\le n} \frac{1}{\sqrt{2\pi}} e^{-(x_i-\theta)^2/2}$$

$$\ln L(x_1, x_2, \ldots, x_n|\theta) = \sum_{1\le i\le n} -\frac{1}{2}\ln 2\pi - \frac{(x_i-\theta)^2}{2}$$

$$\frac{d}{d\theta}\ln L(x_1, x_2, \ldots, x_n|\theta) = \sum_{1\le i\le n} (x_i - \theta)$$

$$= \left(\sum_{1\le i\le n} x_i\right) - n\theta = 0$$

And verify it's max, not min & not better on boundary

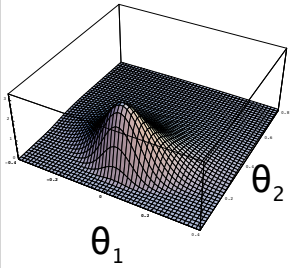


$$\boxed{\hat{\theta} = \left(\sum_{1\le i\le n} x_i\right)/n = \bar{x}}$$

Ex 3: $x_i \sim N(\mu, \sigma^2),\ \mu, \sigma^2$ both unknown

$$\ln L(x_1, x_2, \ldots, x_n|\theta_1, \theta_2) = \sum_{1\le i\le n} -\frac{1}{2}\ln 2\pi\theta_2 - \frac{(x_i-\theta_1)^2}{2\theta_2}$$

$$\frac{\partial}{\partial\theta_1}\ln L(x_1, x_2, \ldots, x_n|\theta_1, \theta_2) = \sum_{1\le i\le n} \frac{(x_i-\theta_1)}{\theta_2} = 0$$



$$\boxed{\hat{\theta}_1 = \left(\sum_{1\le i\le n} x_i\right)/n = \bar{x}}$$

Ex. 3, (cont.)

$$\ln L(x_1, x_2, \ldots, x_n|\theta_1, \theta_2) = \sum_{1\le i\le n} -\frac{1}{2}\ln 2\pi\theta_2 - \frac{(x_i-\theta_1)^2}{2\theta_2}$$

$$\frac{\partial}{\partial\theta_2}\ln L(x_1, x_2, \ldots, x_n|\theta_1, \theta_2) = \sum_{1\le i\le n} -\frac{1}{2}\frac{2\pi}{2\pi\theta_2} + \frac{(x_i-\theta_1)^2}{2\theta_2^2} = 0$$

$$\boxed{\hat{\theta}_2 = \left(\sum_{1\le i\le n}(x_i-\hat{\theta}_1)^2\right)/n = \bar{s}^2}$$

A consistent, but *biased* estimate of population variance. (An example of *overfitting.*) Unbiased estimate is:

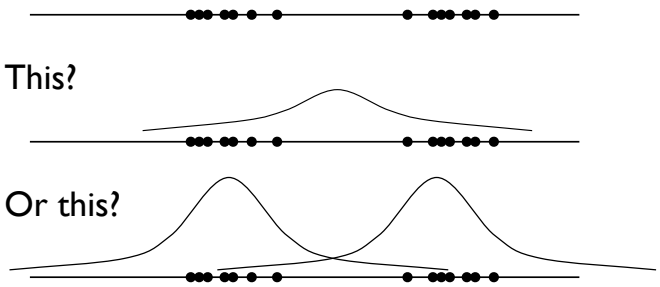
$$\boxed{\hat{\theta}_2 = \sum_{1\le i\le n} \frac{(x_i-\hat{\theta}_1)^2}{n-1}}$$

Moral: MLE is a great idea, but not a magic bullet

EM

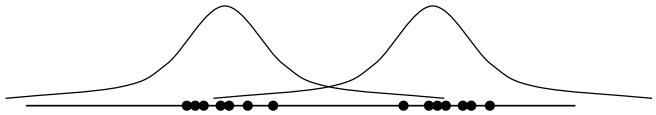
The Expectation-Maximization Algorithm

More Complex Example



(A modeling decision, not a math problem...)

Gaussian Mixture Models / Model-based Clustering



Parameters θ

means	μ_1	μ_2
variances	σ_1^2	σ_2^2
mixing parameters	τ_1	$\tau_2 = 1 - \tau_1$
P.D.F.	$f(x\|\mu_1, \sigma_1^2)$	$f(x\|\mu_2, \sigma_2^2)$

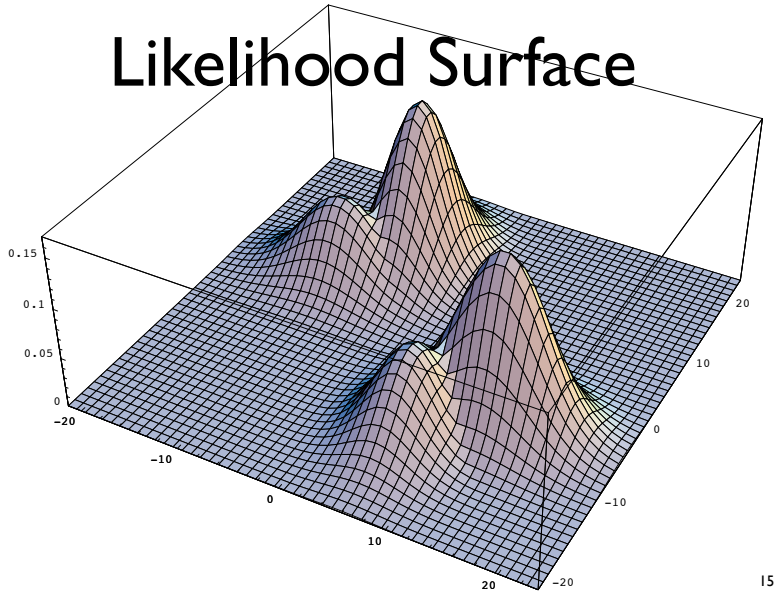
Likelihood

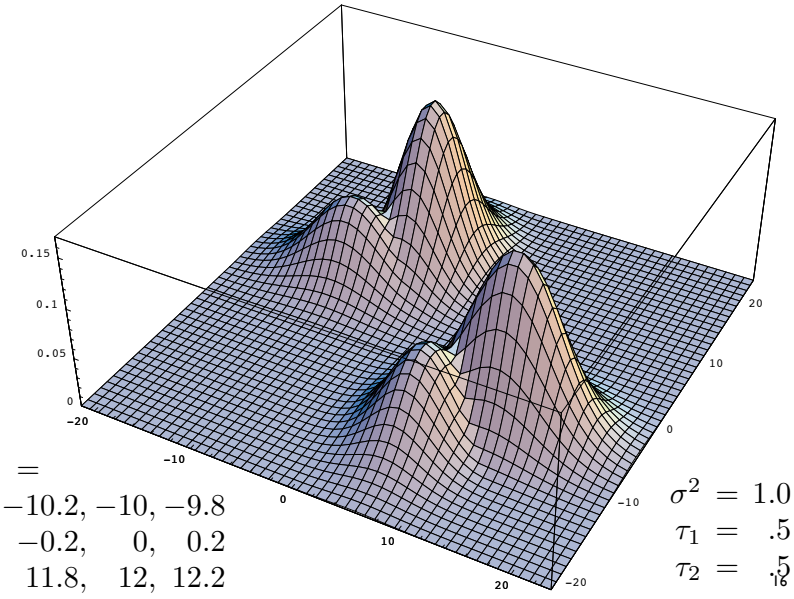
$$L(x_1, x_2, \ldots, x_n|\mu_1, \mu_2, \sigma_1^2, \sigma_2^2, \tau_1, \tau_2)$$

$$= \prod_{i=1}^{n} \sum_{j=1}^{2} \tau_j f(x_i|\mu_j, \sigma_j^2)$$

No closed-form max

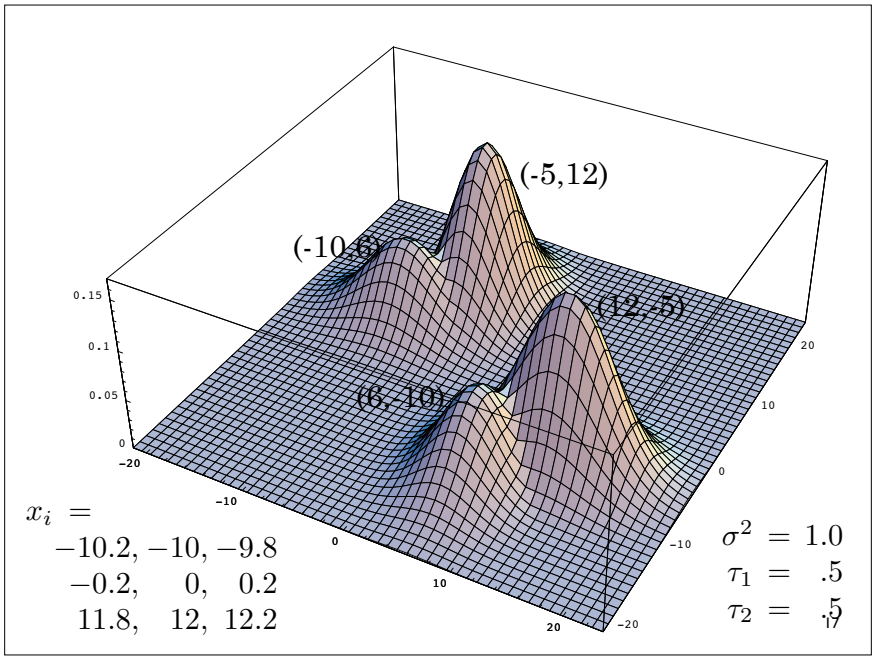
Likelihood Surface





$x_i =$
$-10.2, -10, -9.8$
$-0.2, \quad 0, \quad 0.2$
$11.8, \quad 12, \quad 12.2$

$\sigma^2 = 1.0$
$\tau_1 = .5$
$\tau_2 = .5$



$x_i =$
$-10.2, -10, -9.8$
$-0.2, 0, 0.2$
$11.8, 12, 12.2$

$\sigma^2 = 1.0$
$\tau_1 = .5$
$\tau_2 = .5$

A What-If Puzzle

Likelihood

$$L(x_1, x_2, \ldots, x_n | \overbrace{\mu_1, \mu_2, \sigma_1^2, \sigma_2^2, \tau_1, \tau_2}^{\theta}) = \prod_{i=1}^{n} \sum_{j=1}^{2} \tau_j f(x_i | \mu_j, \sigma_j^2)$$

- Messy: no closed form solution known for finding θ maximizing L
- But *what if* we knew the *hidden data*?

$$z_{ij} = \begin{cases} 1 & \text{if } x_i \text{ drawn from } f_j \\ 0 & \text{otherwise} \end{cases}$$

EM as Egg vs Chicken

- *IF* z_{ij} known, could estimate parameters θ
- *IF* parameters θ known, could estimate z_{ij}
- But we know neither; (optimistically) iterate:
  - E: calculate *expected* z_{ij}, given parameters
  - M: calc "MLE" of parameters, given $E(z_{ij})$

Simple Idea: "Classification EM"

- If $z_{ij} < .5$, pretend it's 0; $z_{ij} > .5$, pretend it's 1
  i.e., *classify* points as component 0 or 1
- Now recalc θ, assuming that partition
- then recalc z_{ij}, assuming that θ
- then re-recalc θ, assuming new z_{ij}
- etc., etc.

Full EM

x_i's are known; θ unknown. Goal is to find MLE θ of:

$$L(x_1, \ldots, x_n \mid \theta)$$ (hidden data likelihood)

Would be easy *if* z_{ij}'s were known, i.e., consider:

$$L(x_1, \ldots, x_n, z_{11}, z_{12}, \ldots, z_{n2} \mid \theta)$$ (complete data likelihood)

But z_{ij}'s aren't known.

Instead, maximize *expected* likelihood of visible data

$$E(L(x_1, \ldots, x_n, z_{11}, z_{12}, \ldots, z_{n2} \mid \theta)),$$

where expectation is over distribution of hidden data (z_{ij}'s)

The E-step

- Assume θ known & fixed
- A (B): the event that x_i was drawn from f_1 (f_2)
- D: the observed datum x_i
- Expected value of z_{i1} is P(A|D) — $E = 0 \cdot P(0) + 1 \cdot P(1)$

$$\boxed{P(A|D) = \frac{P(D|A)P(A)}{P(D)}}$$

$$P(D) = P(D|A)P(A) + P(D|B)P(B)$$
$$= f_1(x_i|\theta_1)\,\tau_1 + f_2(x_i|\theta_2)\,\tau_2$$

Repeat for each x_i

Complete Data Likelihood

Recall:

$$z_{1j} = \begin{cases} 1 & \text{if } x_1 \text{ drawn from } f_j \\ 0 & \text{otherwise} \end{cases}$$

so, correspondingly,

$$L(x_1, z_{1j} \mid \theta) = \begin{cases} \tau_1 f_1(x_1 \mid \theta) & \text{if } z_{11} = 1 \\ \tau_2 f_2(x_1 \mid \theta) & \text{otherwise} \end{cases}$$

Formulas with "if's" are messy; can we blend more smoothly?
Yes, many possibilities. Idea 1:

$$L(x_1, z_{1j} \mid \theta) = z_{11} \cdot \tau_1 f_1(x_1 \mid \theta) + z_{12} \cdot \tau_2 f_2(x_1 \mid \theta)$$

Idea 2:

$$L(x_1, z_{1j} \mid \theta) = (\tau_1 f_1(x_1 \mid \theta))^{z_{11}} \cdot (\tau_2 f_2(x_1 \mid \theta))^{z_{12}}$$

M-step Details

(For simplicity, assume $\sigma_1 = \sigma_2 = \sigma; \tau_1 = \tau_2 = .5 = \tau$)

$$L(\vec{x}, \vec{z} \mid \theta) = \prod_{1 \le i \le n} \frac{\tau}{\sqrt{2\pi\sigma^2}} \exp\left(- \sum_{1 \le j \le 2} z_{ij} \frac{(x_i - \mu_j)^2}{2\sigma^2} \right)$$

$$E[\log L(\vec{x}, \vec{z} \mid \theta)] = E\left[\sum_{1 \le i \le n} \left(\log \tau - \frac{1}{2} \log 2\pi\sigma^2 - \sum_{1 \le j \le 2} z_{ij} \frac{(x_i - \mu_j)^2}{2\sigma^2} \right) \right]$$

$$= \sum_{1 \le i \le n} \left(\log \tau - \frac{1}{2} \log 2\pi\sigma^2 - \sum_{1 \le j \le 2} E[z_{ij}] \frac{(x_i - \mu_j)^2}{2\sigma^2} \right)$$

Find θ maximizing this as before, using $E[z_{ij}]$ found in E-step. Result:

$\boxed{\mu_j = \sum_{i=1}^n E[z_{ij}] x_i / \sum_{i=1}^n E[z_{ij}]}$ (intuit: avg, weighted by subpop prob)

2 Component Mixture

$\sigma_1 = \sigma_2 = 1;\ \tau = 0.5$

		mu1	-20.00		-6.00		-5.00		-4.99
		mu2	6.00		0.00		3.75		3.75
x1	-6	z11		5.11E-12		1.00E+00		1.00E+00	
x2	-5	z21		2.61E-23		1.00E+00		1.00E+00	
x3	-4	x31		1.33E-34		9.98E-01		1.00E+00	
x4	0	z41		9.09E-80		1.52E-08		4.11E-03	
x5	4	z51		6.19E-125		5.75E-19		2.64E-18	
x6	5	z61		3.16E-136		1.43E-21		4.20E-22	
x7	6	z71		1.62E-147		3.53E-24		6.69E-26	

EM Summary

- Fundamentally a max likelihood parameter estimation problem
- Useful if analysis is more tractable when 0/1 hidden data z known
- Iterate:
  - E-step: estimate E(z) given θ
  - M-step: estimate θ maximizing E(likelihood) given E(z)

EM Issues

- Under mild assumptions (sect 11.6), EM is guaranteed to increase likelihood with every E-M iteration, hence will converge.
- *But* may converge to *local*, not global, max. (Recall the 4-bump surface...)
- Issue is probably intrinsic, since EM is often applied to NP-hard problems (including clustering, above, and motif-discovery, soon)
- Nevertheless, widely used, often effective

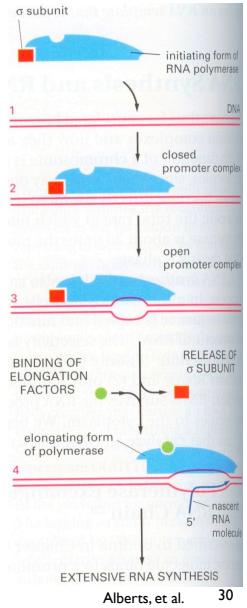
Gene Expression & Regulation

Gene Expression

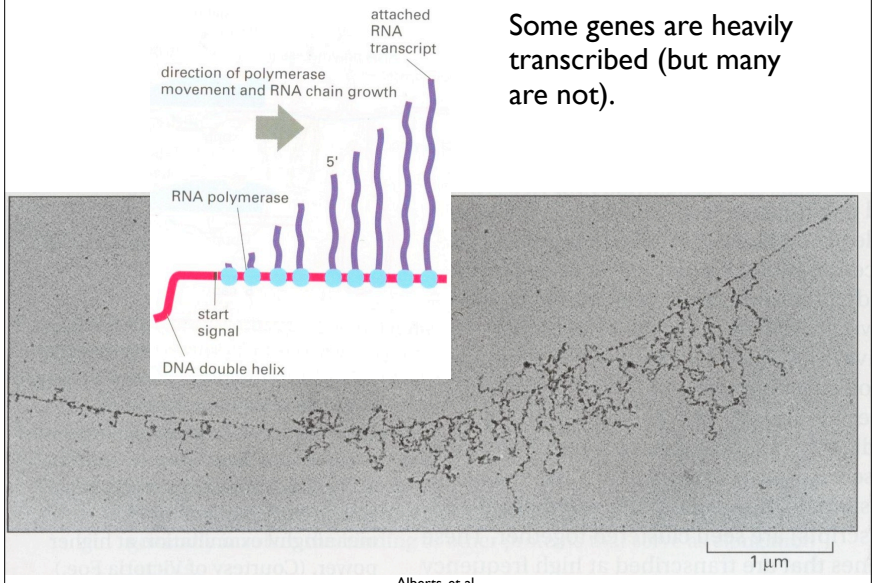
- Recall a *gene* is a DNA sequence for a prot
- To say a gene is *expressed* means that it
  1. is *transcribed* from DNA to RNA
  2. the mRNA is *processed* in various ways
  3. is *exported* from the nucleus (eukaryotes)
  4. is *translated* into protein
- A key point: not all genes are expressed all the time, in all cells, or at equal levels

Transcription

- RNA *polymerase* complex
  - E. coli: 5 proteins (2α, β, β', σ)
    σ is *initiation factor*; finds promoter, then released/replaced by *elongation factors*
  - Eukaryotes: 3 pols, each >10 subunits
- attaches to DNA, melts helix, makes RNA copy (5' → 3') of template (3' → 5') at ~30nt/sec



Alberts, et al.



Some genes are heavily transcribed (but many are not).

Alberts, et al.

5' Processing: Capping

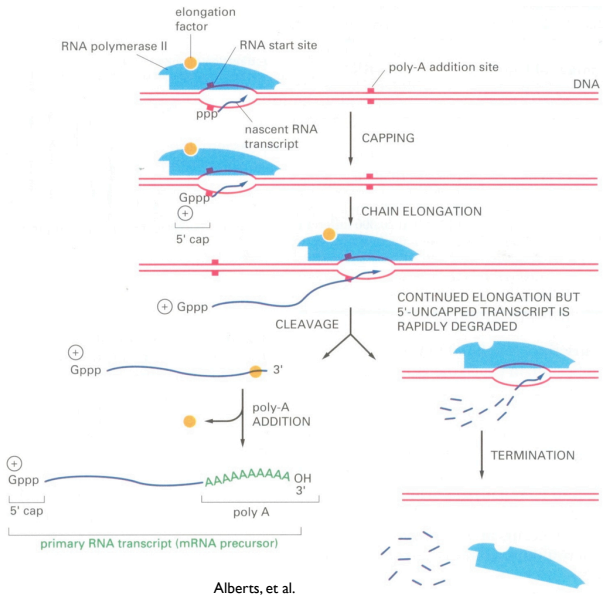
- methylated G added to 5' end, and methyl added to ribose of 1st nucleotide of transcript
- probably helps distinguish protein-coding mRNAs from other RNA junk
  - prevents degradation
  - facilitates start of translation

3' Processing: Poly A
(Eukaryotes)

- Transcript cleaved after AAUAAA (roughly)
- pol keeps running (until it falls off) but no 5' cap added to strand downstream of poly A site, so it's rapidly degraded
- 10s - 100s of A's added to 3' end of transcript - its "poly A tail"

More processing: Splicing

- Also in eukaryotes, most genes are spliced: protein coding exons are interrupted by non-coding introns, which are cut out & degraded, exons spliced together
- More details about this when we get to gene finding



Alberts, et al.

Nuclear Export

- In eukaryotes, mature mRNAs are actively transported out of the nucleus & ferried to specific destinations (e.g., mitochondria, ribosomes)

Regulation

- In most cells, pro- or eukaryote, easily a 10,000-fold difference between least- and most-highly expressed genes
- Regulation happens at all steps. E.g., some transcripts can be sequestered then released, or rapidly degraded, some are weakly translated, some are very actively translated, some are highly transcribed, some are not transcribed at all
- Below, focus on 1st step only: transcriptional regulation

DNA Binding Proteins

A variety of DNA binding proteins ("transcription factors"; a significant fraction, perhaps 5-10%, of all human proteins) modulate transcription of protein coding genes

Summary

- Learning from data:
  - MLE: Max Likelihood Estimators
  - EM: Expectation Maximization (MLE w/hidden data)
- Expression & regulation
  - Expression: creation of gene products
  - Regulation: when/where/how much of each gene product; complex and critical
- Next: using MLE/EM to find regulatory motifs in biological sequence data