CSEP 590A

Summer 2006

Lecture 4

MLE, EM, RE, Expression

FYI, re HW #2: Hemoglobin History

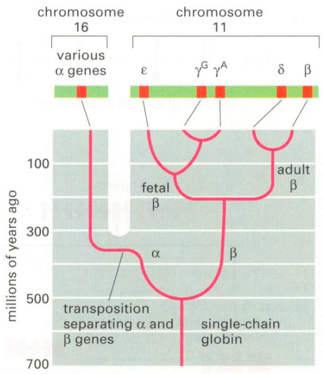


Figure 8–76 An evolutionary scheme for the globin chains that carry oxygen in the blood of animals. The scheme emphasizes the β-like globin gene family. A relatively recent gene duplication of the γ-chain gene produced γ^G and γ^A, which are fetal β-like chains of identical function. The location of the globin genes in the human genome is shown at the top of the figure. Alberts et al., 3rd ed.,pg389

Tonight

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

- Next week: Motif description & discovery

MLE

Maximum Likelihood Estimators

Probability Basics, I

	Ex.	Ex.
Sample Space	$\{1, 2, \dots, 6\}$	$\mathbb{R}$
Distribution	$p_1, \dots, p_6 \ge 0; \sum_{1 \le i \le 6} p_i = 1$	$f(x) >= 0; \int_{\mathbb{R}} f(x)dx = 1$
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 $x_1, x_2, \dots, x_n$ is from a parametric distribution f(x|θ), estimate θ.
- E.g.:

$$\begin{aligned} f(x) &= \frac{1}{\sqrt{2\pi\sigma^2}} e^{-(x-\mu)^2/(2\sigma^2)} \\ \theta &= (\mu, \sigma^2) \end{aligned}$$

Maximum Likelihood Parameter Estimation

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

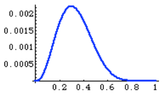
$$L(x_1, x_2, \dots, x_n) = \prod_{i=1}^{n} f(x_i \mid \theta)$$

- 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



$$\begin{aligned}
L(x_1, x_2, \ldots, x_n \mid \theta) &= (1-\theta)^{n_0}\theta^{n_1} \\
\log L(x_1, x_2, \ldots, x_n \mid \theta) &= n_0 \log(1-\theta) + n_1 \log\theta \\
\tfrac{\partial}{\partial\theta} \log L(x_1, x_2, \ldots, x_n \mid \theta) &= \tfrac{-n_0}{1-\theta} + \tfrac{n_1}{\theta}
\end{aligned}$$

Setting to zero and solving:

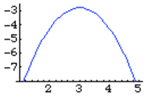
$$\boxed{\theta = \tfrac{n_1}{n}}$$

(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

$$\begin{aligned}
L(x_1, x_2, \ldots, x_n|\theta) &= \prod_{1\le i\le n} \frac{1}{\sqrt{2\pi}} e^{-(x_i-\theta)^2/2} \\
\ln L(x_1, x_2, \ldots, x_n|\theta) &= \sum_{1\le i\le n} -\frac{1}{2}\ln 2\pi - \frac{(x_i-\theta)^2}{2} \\
\tfrac{d}{d\theta} \ln L(x_1, x_2, \ldots, x_n|\theta) &= \sum_{1\le i\le n} (x_i - \theta) \\
&= \left(\textstyle\sum_{1\le i\le n} x_i\right) - n\theta = 0
\end{aligned}$$

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

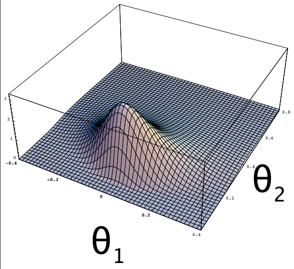


$$\boxed{\hat\theta = \left(\textstyle\sum_{1\le i\le n} x_i\right)/n = \bar{x}}$$

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

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

$$\boxed{\hat\theta_1 = \left(\textstyle\sum_{1\le i\le n} x_i\right)/n = \bar{x}}$$



Ex. 3, (cont.)

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

$$\boxed{\hat\theta_2 = \left(\textstyle\sum_{1\le i\le n} (x_i - \hat\theta_1)^2\right)/n = \bar{s}^2}$$

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

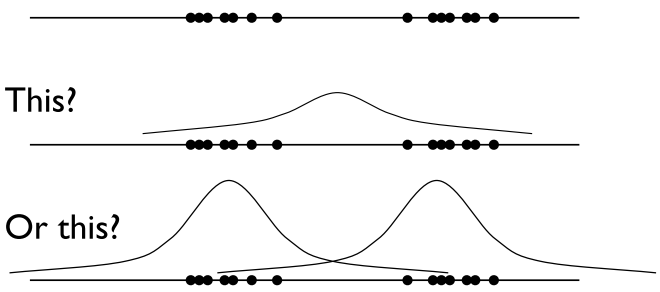
$$\boxed{\hat\theta_2 = \textstyle\sum_{1\le i\le n} \frac{(x_i-\hat\theta_1)^2}{n-1}}$$

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

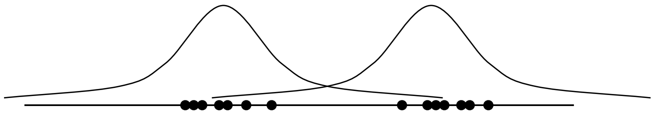
EM

The Expectation-Maximization Algorithm

More Complex Example



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.	$f(x\|\mu_1, \sigma_1^2)$	$f(x\|\mu_2, \sigma_2^2)$

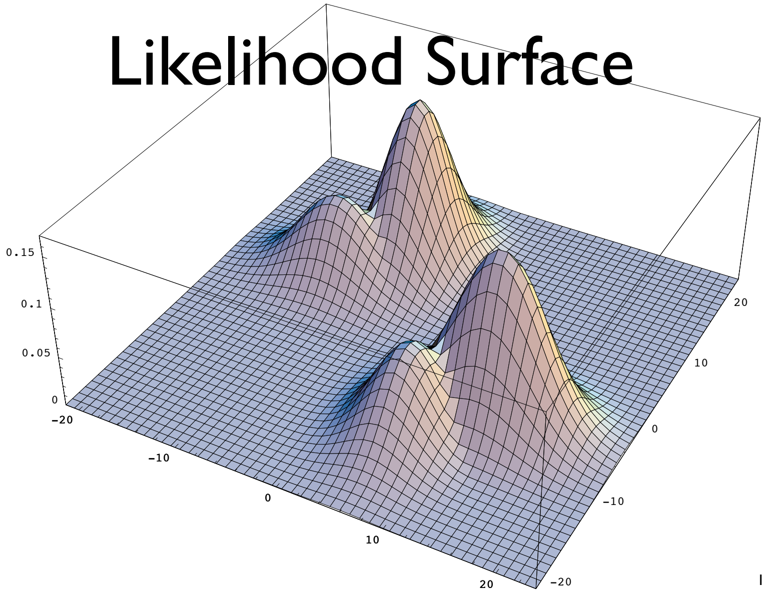
Likelihood

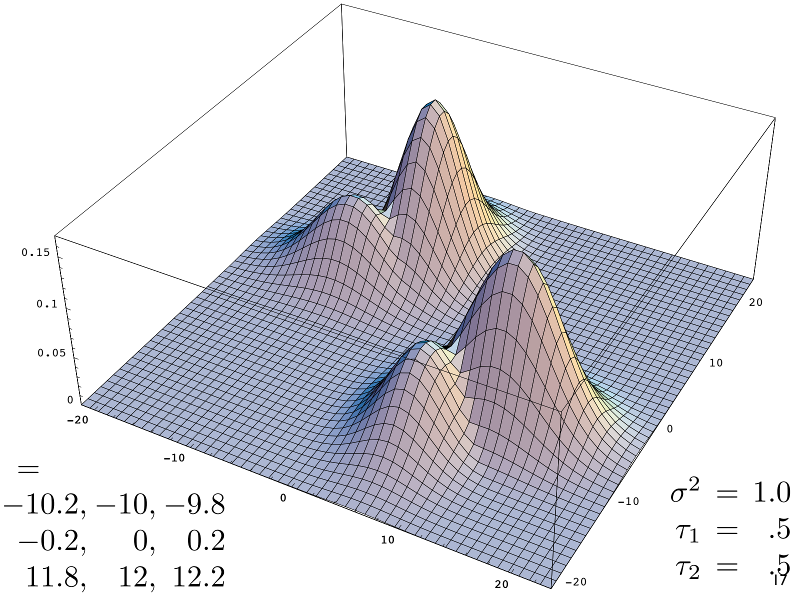
$$L(x_1, x_2, \ldots, x_n | \mu_1, \mu_2, \sigma_1^2, \sigma_2^2, \tau_1, \tau_2)$$

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

No closed-form max

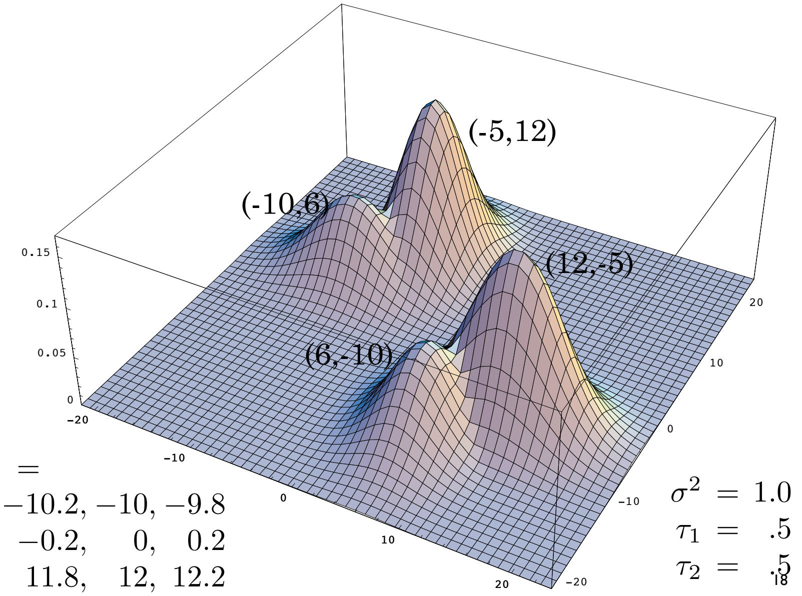
Likelihood Surface





$$x_i = \begin{array}{ccc} -10.2, & -10, & -9.8 \\ -0.2, & 0, & 0.2 \\ 11.8, & 12, & 12.2 \end{array}$$

$$\begin{aligned} \sigma^2 &= 1.0 \\ \tau_1 &= .5 \\ \tau_2 &= .5 \end{aligned}$$



$$x_i = \begin{array}{ccc} -10.2, & -10, & -9.8 \\ -0.2, & 0, & 0.2 \\ 11.8, & 12, & 12.2 \end{array}$$

$$\begin{aligned} \sigma^2 &= 1.0 \\ \tau_1 &= .5 \\ \tau_2 &= .5 \end{aligned}$$

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})

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) — $E = 0 \cdot P(0) + 1 \cdot P(1)$

$$\left.\begin{aligned} \boxed{P(A|D) = \frac{P(D|A)P(A)}{P(D)}} & \\ P(D) &= P(D|A)P(A) + P(D|B)P(B) \\ &= f_1(x_i|\theta_1)\,\tau_1 + f_2(x_i|\theta_2)\,\tau_2 \end{aligned}\right\} \text{Repeat for each } x_i$$

The M-Step

Goal is to find MLE θ of:

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

x_i's are known;

Would be easy *if* z_{ij}'s also known, but they aren't.

Instead, maximize *expected* likelihood of visible data

$$E(L(x_1, \ldots, x_n \mid \theta)),$$

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

M-step Details

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

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

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

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

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)

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

Relative entropy

Relative Entropy

- AKA Kullback-Liebler Distance/Divergence, AKA Information Content
- Given distributions P, Q

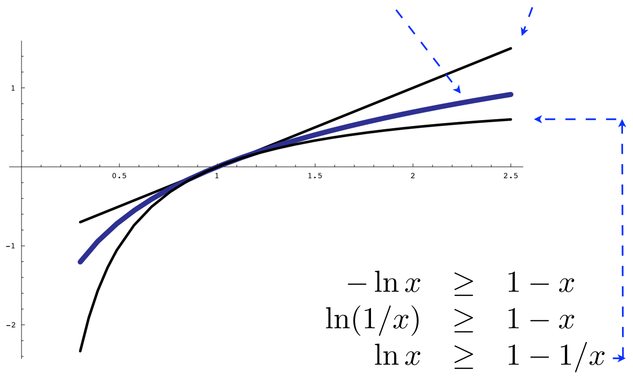
$$H(P||Q) = \sum_{x\in\Omega} P(x) \log \frac{P(x)}{Q(x)}$$

Notes:

Let $P(x)\log\frac{P(x)}{Q(x)} = 0$ if $P(x) = 0$ [since $\lim_{y\to 0} y \log y = 0$]

Undefined if $0 = Q(x) < P(x)$

$\ln x \quad \leq \quad x - 1$



$$\begin{aligned} -\ln x &\geq 1 - x \\ \ln(1/x) &\geq 1 - x \\ \ln x &\geq 1 - 1/x \end{aligned}$$

Theorem: $H(P||Q) \geq 0$

$$\begin{aligned} H(P||Q) &= \textstyle\sum_x P(x) \log \frac{P(x)}{Q(x)} \\ &\geq \textstyle\sum_x P(x) \left(1 - \frac{Q(x)}{P(x)}\right) \\ &= \textstyle\sum_x (P(x) - Q(x)) \\ &= \textstyle\sum_x P(x) - \sum_x Q(x) \\ &= 1 - 1 \\ &= 0 \end{aligned}$$

Furthermore: H(P||Q) = 0 if and only if P = Q

Bottom line: “bigger” means “more different”

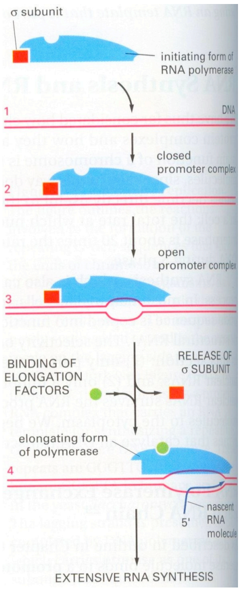
Gene Expression & Regulation

Gene Expression

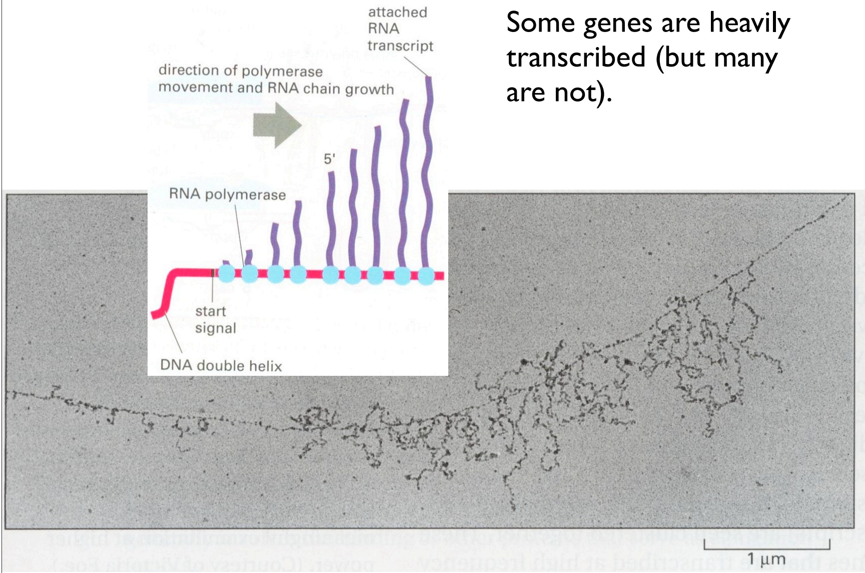
- Recall a *gene* is a DNA sequence
- To say a gene is *expressed* means that it
  1. is *transcribed* from DNA to RNA
  2. the mRNA is *processed* in various ways
  3. is *exported* from the nucleus (eukaryotes)
  4. is *translated* into protein
- A key point: not all genes are expressed all the time,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

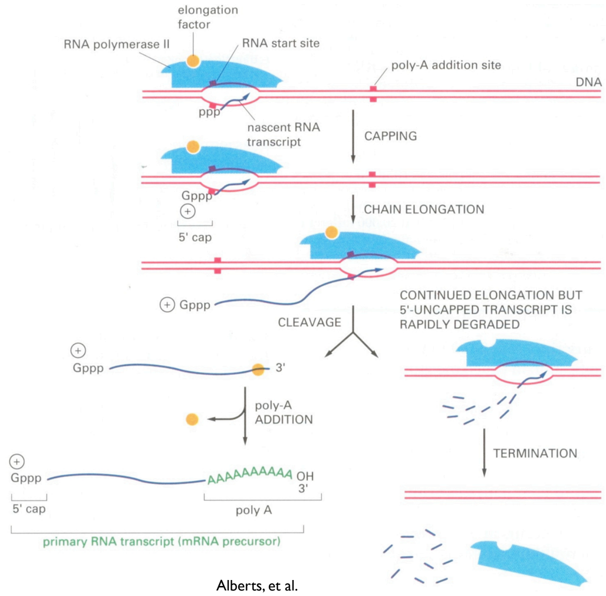
3' Processing: Poly A

(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)

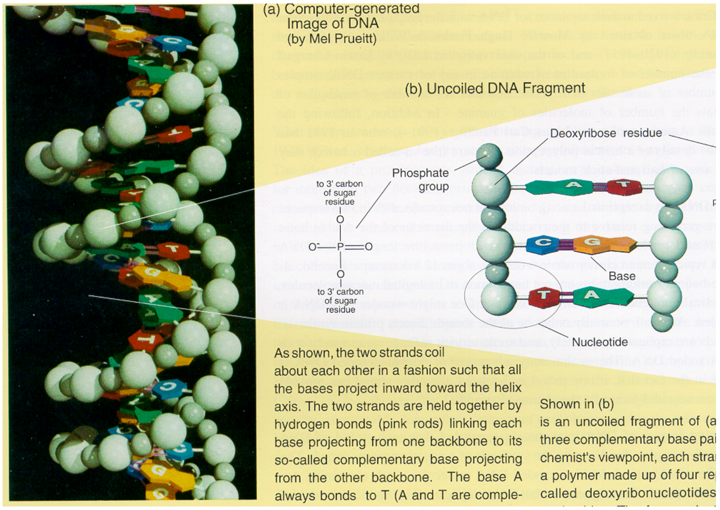
Regulation

- 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 10%?, of all human proteins) modulate transcription of protein coding genes

The Double Helix



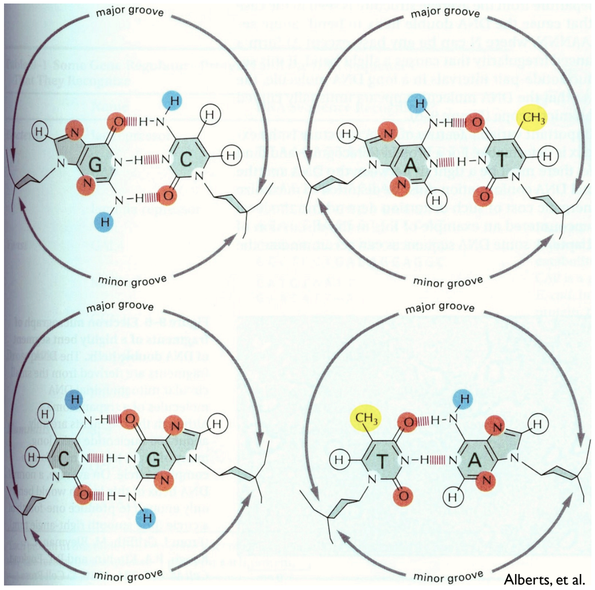
As shown, the two strands coil about each other in a fashion such that all the bases project inward toward the helix axis. The two strands are held together by hydrogen bonds (pink rods) linking each base projecting from one backbone to its so-called complementary base projecting from the other backbone. The base A always bonds to T (A and T are comple-

Shown in (b) is an uncoiled fragment of (a three complementary base pai chemist's viewpoint, each stra a polymer made up of four re called deoxyribonucleotides

Los Alamos Science

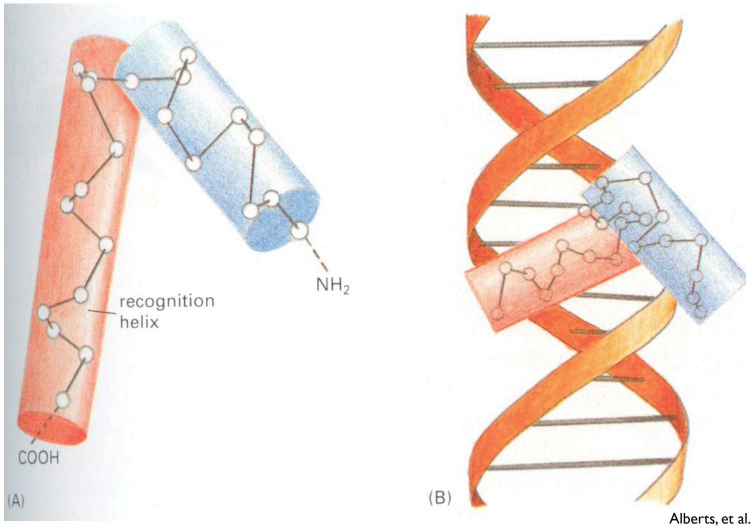
In the groove

Different patterns of potential H bonds at edges of different base pairs, accessible esp. in major groove



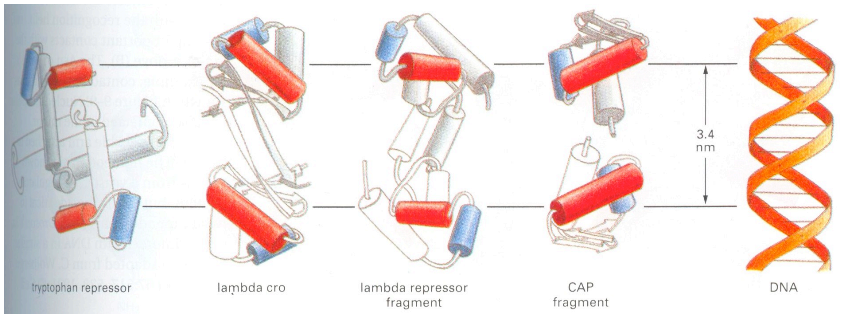
Alberts, et al.

Helix-Turn-Helix DNA Binding Motif



Alberts, et al.

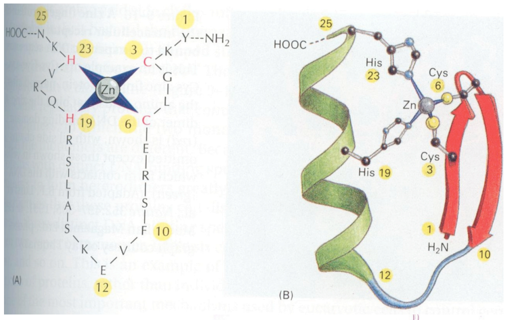
H-T-H Dimers

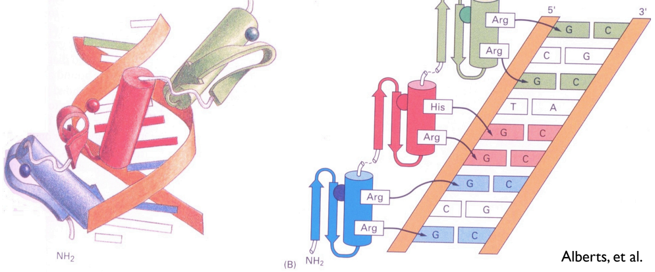


Alberts, et al.

Bind 2 DNA patches, ~ 1 turn apart
Increases both specificity and affinity

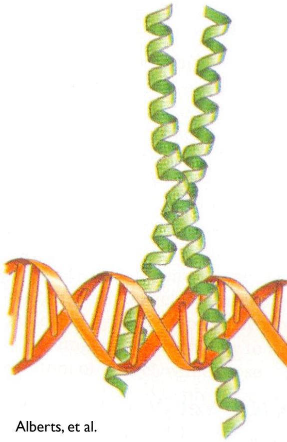
Zinc Finger Motif

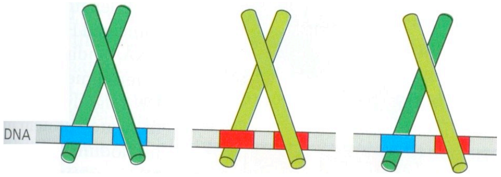




Alberts, et al.

Leucine Zipper Motif





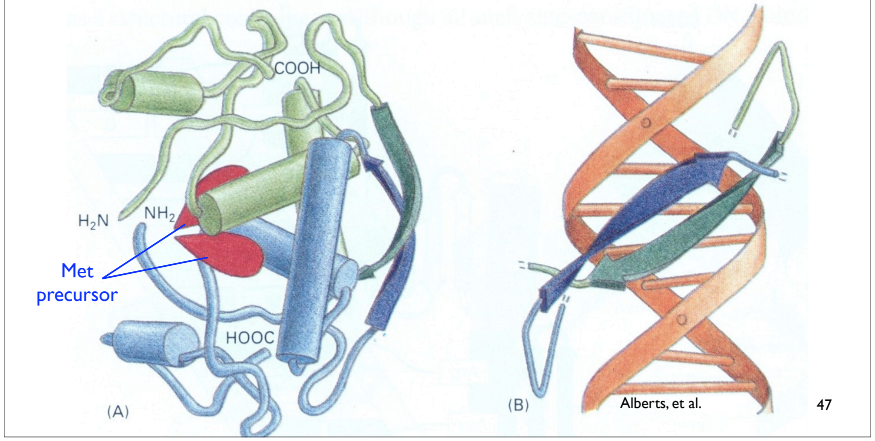
Homo-/hetero-dimers
and combinatorial
control

Alberts, et al.

Bacterial Met Repressor

a beta-sheet DNA binding domain

Negative feedback loop:
high Met level ⇒ repress Met synthesis genes



Alberts, et al.

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 week: using MLE/EM to find regulatory motifs in biological sequence data