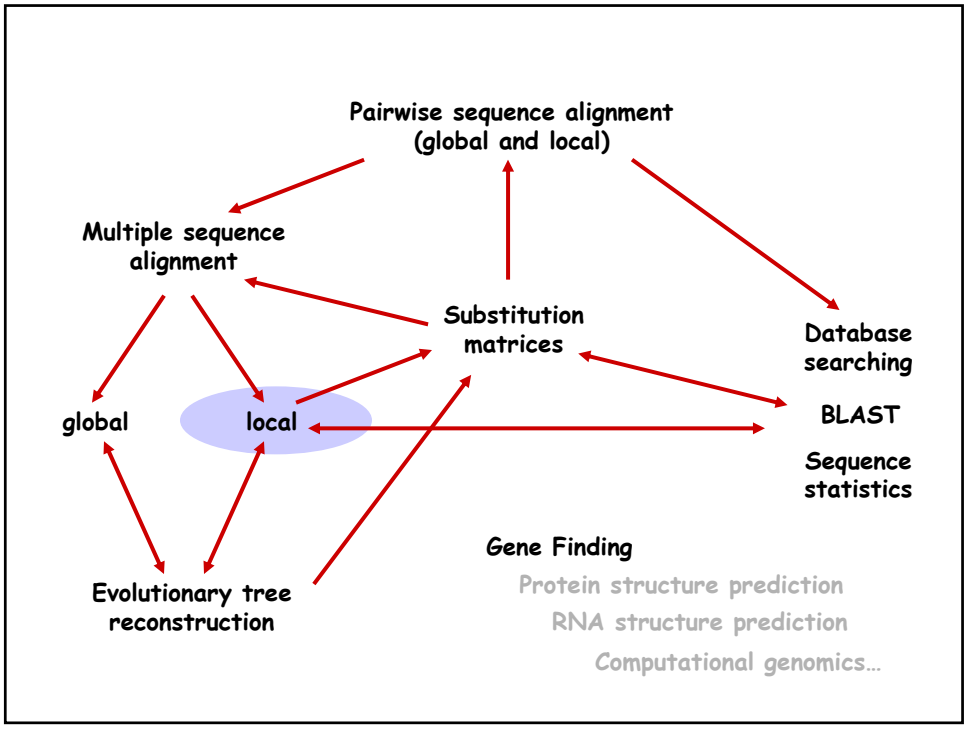
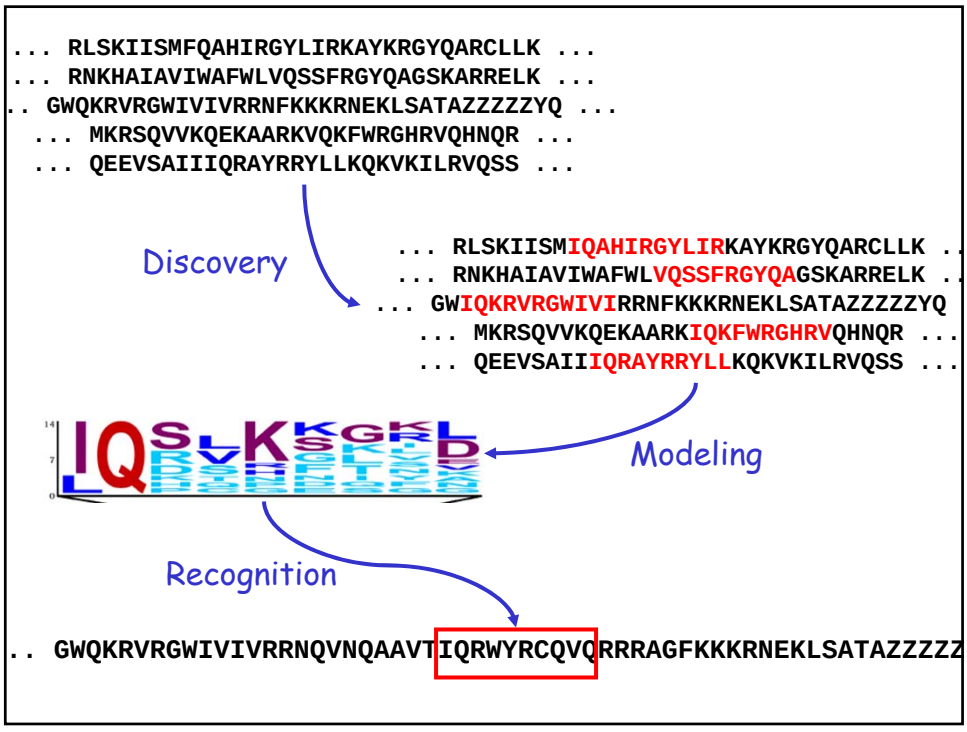
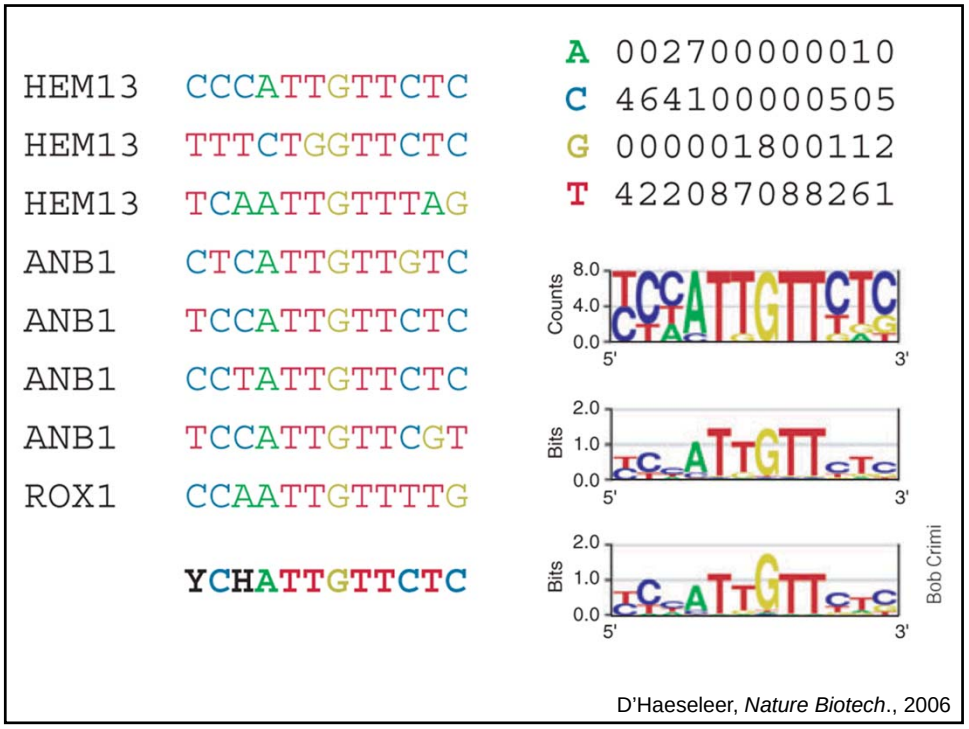
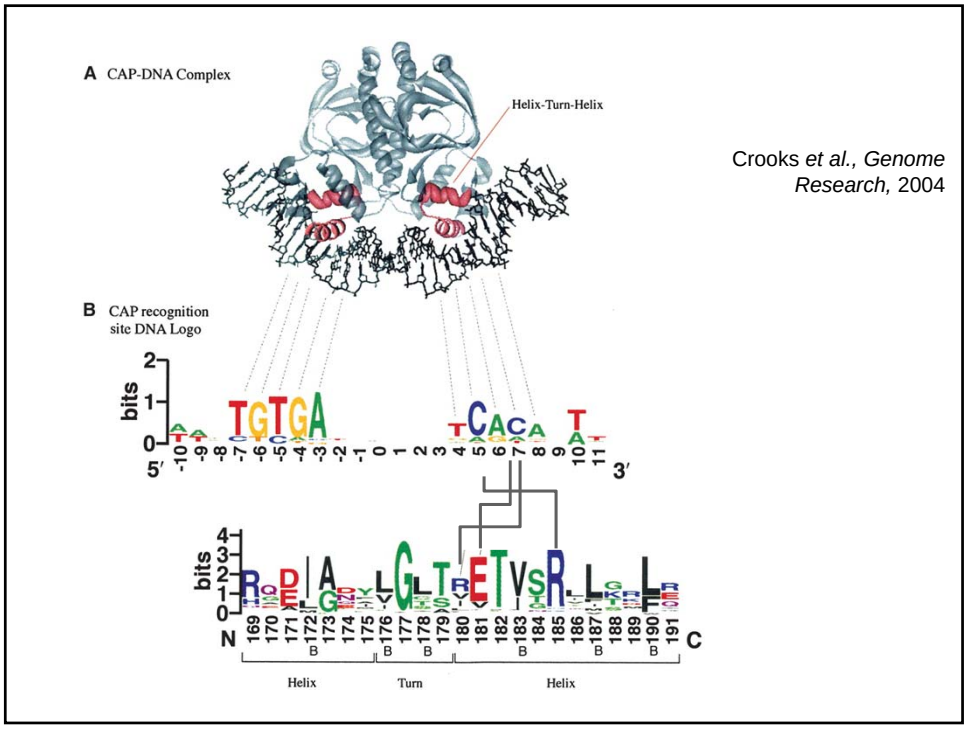

Pairwise sequence alignment
(global and local)
Multiple sequence
alignment
Substitution
matrices
Database
searching
global
local
BLAST
Sequence
statistics
Gene Finding
Protein structure prediction
RNA structure prediction
Computational genomics…
Evolutionary tree
reconstruction


... RLSKIISMFQAHIRGYLIRKAYKRGYQARCLLK ...
... RNKHAIAVIWAFWLVQSSFRGYQAGSKARRELK ...
.. GWQKRVRGWIVIVRRNFKKKRNEKLSATAZZZZZYQ ...
... MKRSQVVKQEKAARKVQKFWRGHRVQHNQR ...
... QEEVSAIIIQRAYRRYLLKQKVKILRVQSS ...
Discovery
... RLSKIISMIQAHIRGYLIRKAYKRGYQARCLLK ..
... RNKHAIAVIWAFWLVQSSFRGYQAGSKARRELK ..
... GWIQKRVRGWIVIRRNFKKKRNEKLSATAZZZZZYQ
... MKRSQVVKQEKAARKIQKFWRGHRVQHNQR ...
... QEEVSAIIIQRAYRRYLLKQKVKILRVQSS ...
Modeling
Recognition
.. GWQKRVRGWIVIVRRNQVNQAAVTIQRWYRCQVQRRRAGFKKKRNEKLSATAZZZZZ



D'Haeseleer, *Nature Biotech*., 2006



Crooks *et al., Genome Research,* 2004

Local Multiple Sequence Alignment
Probabilistic Framework

- Discovery
  - Given multiple sequences, often unaligned, find a conserved pattern (e.g., the Pax domain)
- Representation
  - Given a local MSA for the Pax domain, construct probabilistic model
- Recognition (using model)
  - Given a new sequence, does it contain the Pax domain?
  - Find all sequences with Pax domains in the data base.

Local MSA Methods

- Discovery:
  - Hidden Markov Models (HMMs)
  - Gibb's sampler
  - PSI BLAST
- Modeling:
  - ➢ Position Specific Scoring Matrices (PSSMs)
  - HMMs
- Recognition:
  - Depends on model

Position Specific Scoring Matrices

PSSM's, profiles, weight matrices, templates…

Assume pattern has already been discovered.

Input: local MSA, *k x n* matrix

*A[i,j]: j*th symbol in *i*th sequence

Output: Scoring matrix, |Σ| x n

S[i,j]: score of symbol *i* at position *j*

Position Specific Scoring Matrices

PSSM's, profiles, weight matrices, templates…

Example:

WEIRD
WEIRD
WEIRE
WEIQH

k sequences

w

See spreadsheets…

Position Specific Scoring Matrices

PSSM's, profiles, weight matrices, templates…

Given an ungapped local multiple alignment of *w* residues in *k* sequences, the frequency of amino acid *i* at position *j* is

$$q[i, j] = \frac{c[i, j]}{k}$$

where *c[i,j]* is the number of instances of aa *i* at site *j.*

The propensity of amino acid *i* at position *j is*

$$P[i, j] = \frac{q[i, j]}{p(i)}$$

where *p(i)* is the background frequency of amino acid *i.*

From this, we obtain, a position specific scoring matrix

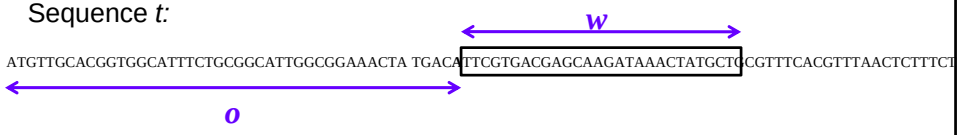
$$S[i, j] = \log_2 P[i, j]$$

Scoring a potential new instance of the pattern:

Given a sequence *t,* a window of length *w* starting at offset position *o* is scored as follows*:*

$$Score[t, o] = \sum_{j=0}^{w-o} S[t[j+o], j]$$

Sequence *t:*



This score can be interpreted as a log likelihood ratio…

Bayes Rule

likelihood

posterior probability

prior probability

$$P(Ha \mid D) = \frac{P(D \mid Ha)P(Ha)}{P(D)}$$

probability of data independent of hypothesis

Bayes Rule

$$P(Ha \mid D) = \frac{P(D \mid Ha)P(Ha)}{P(D)}$$

$$\frac{P(Ha \mid D)}{P(H_0 \mid D)} = \frac{P(D \mid Ha)P(Ha)}{P(D \mid H_0)P(H_0)}$$

prior probability
is hard to estimate

Bayes Rule

$$P(Ha \mid D) = \frac{P(D \mid Ha)P(Ha)}{P(D)}$$

$$\frac{P(Ha \mid D)}{P(H_0 \mid D)} = \frac{P(D \mid Ha)P(Ha)}{P(D \mid H_0)P(H_0)}$$

$$\frac{P(Ha \mid D)}{P(H_0 \mid D)} = \frac{P(D_1 \mid Ha)P(D2 \mid Ha)P(D_3 \mid Ha)....P(Ha)}{P(D_1 \mid H_0)P(D_2 \mid H_0)P(D_3 \mid H_0).....P(H_0)}$$

Bayes Rule

$$P(Ha \mid D) = \frac{P(D \mid Ha)P(Ha)}{P(D)}$$

$$\frac{P(Ha \mid D)}{P(H_0 \mid D)} = \frac{P(D \mid Ha)P(Ha)}{P(D \mid H_0)P(H_0)}$$

$$\frac{P(Ha \mid D)}{P(H_0 \mid D)} = \frac{P(D_1 \mid Ha)P(D2 \mid Ha)P(D_3 \mid Ha)....P(Ha)}{P(D_1 \mid H_0)P(D_2 \mid H_0)P(D_3 \mid H_0).....P(H_0)}$$

$$\frac{P(Ha \mid D)}{P(H_0 \mid D)} \approx \frac{\prod_i P(D_i \mid Ha)}{\prod_i P(D_i \mid H_0)}$$

Hypothesis testing using a likelihood ratio

How likely is the data under the alternate hypothesis compared with the likelihood under the null hypothesis?

An example: Suppose you observe 6 heads in 8 coin tosses. Is the coin biased? Let H_a be the probability that the coin is biased and let H_0 be the hypothesis that the coin is fair.

P (toss yields heads): H_a: q=0.75, H_o: 0.5

Likelihood ratio: $\frac{P(D \mid H_a)}{P(D \mid H_0)} = \frac{P(6 \text{ heads in 8 tosses} \mid q)}{P(6 \text{ heads in 8 tosses} \mid 0.5)}$

Note: There are other ways to test a hypothesis; e.g., a *p-value.*

Hypothesis testing using a (log) odds ratio

How likely is the data under the alternate hypothesis compared with the likelihood under the null hypothesis?

Likelihood ratio: $\frac{P(D \mid H_a)}{P(D \mid H_0)} = \frac{P(6 \text{ heads in 8 tosses} \mid 0.75)}{P(6 \text{ heads in 8 tosses} \mid 0.5)}$

$$\frac{(0.75)^6(0.25)^2}{(0.5)^6(0.5)^2} = 2.85$$

Observing 6 heads in 8 coin tosses is 2.85 times as likely if $q = 0.75$ than if the coin is fair.

The log odds ratio is $\log_2(2.85)=1.51$.

Note: the sample size is very small!

A PSSM is a log odds scoring matrix

Note that the score of a window of length *w* at position *o* in *t*, is a log likelihood ratio of the form

$$S[t, L] = \log_2 \frac{P[data \mid H_a]}{P[data \mid H_0]}$$

where the *data* is the subsequence at *o*, H_a is the alternate hypothesis that *t* contains the pattern and H_0 is the null hypothesis (no pattern, background frequencies)

$$\begin{aligned} S[t,o] &= \sum_{j=0}^{w-o} S[t[j+o], j] \\ &= \sum_{j=0}^{w-o} \log_2 P[t[j+o], j] \\ &= \sum_{j=0}^{w-o} \log_2 \frac{q[t[j+o], j]}{p(t[j+o])} \\ &= \log_2 \frac{\prod_{j=0}^{w-o} q[t[j+o], j]}{\prod_{j=0}^{w-o} p(t[j+o])} \\ &= \log_2 \frac{P[data \mid H_a]}{P[data \mid H_0]} \end{aligned}$$

Pseudocounts

Example:

AAAAA
CCCCC
DDDDD
. . .
YYYYY
WEIRD
WEIRD
WEIRE
WEIQH

$$q[i, j] = \frac{c[i, j] + b}{k + |\Sigma| b}$$

The pseudocount, ***b,*** avoids the problem of zero entries in the frequency matrix (and negative infinity in the log odds scoring matrix.)

Frequently, ***b = 1****,* is chosen.

Also, see Durbin, 5.6