Modeling the evolution of protein sequences

Most previous work: DNA level

long time ago .....ATTCGTATGCCTATCAC.....
now .....TCTCGTGTATTTCTGAC.....

Advantages:
- Lots of data
- Only 4 x 3 = 12 possible substitutions
- Don't need to consider complicated selective pressure

Disadvantages:
- Fast substitutions: saturation
- Tells little about the encoded proteins

Looking at the protein level

Advantages:

Can tell you about the protein

Selective pressure: can look deeper in evolutionary time

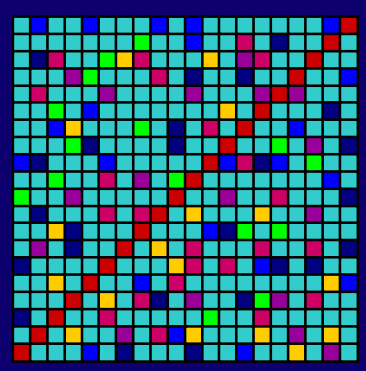
Disadvantages:

Need more complicated models

Interpreting evolutionary changes require a model

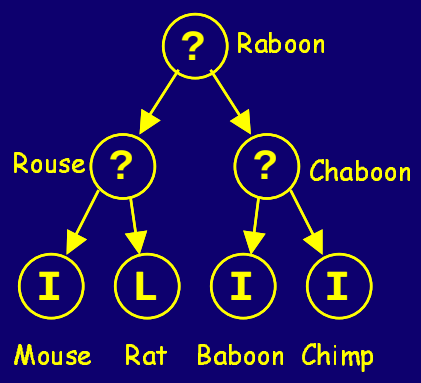
In evolution:
what is the rate at which
Ts become Rs?

e.g. 0.00005 / my



Substitution Matrix

Using current sequences to develop the *evolutionary model*

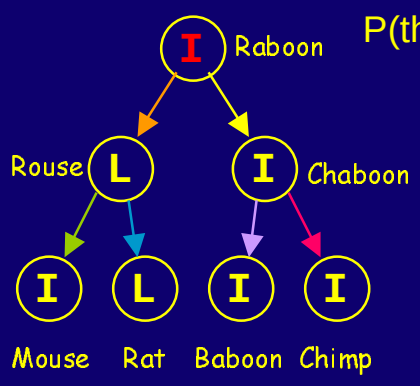



```
Mouse:   ...TLSPGLKIVSNPL...
Rat:     ...TLTPGLKLVSDTL...
Baboon:  ...TVSPGLRIVSDGV...
Chimp:   ...TISPGLVIVSENL...
```

Each location

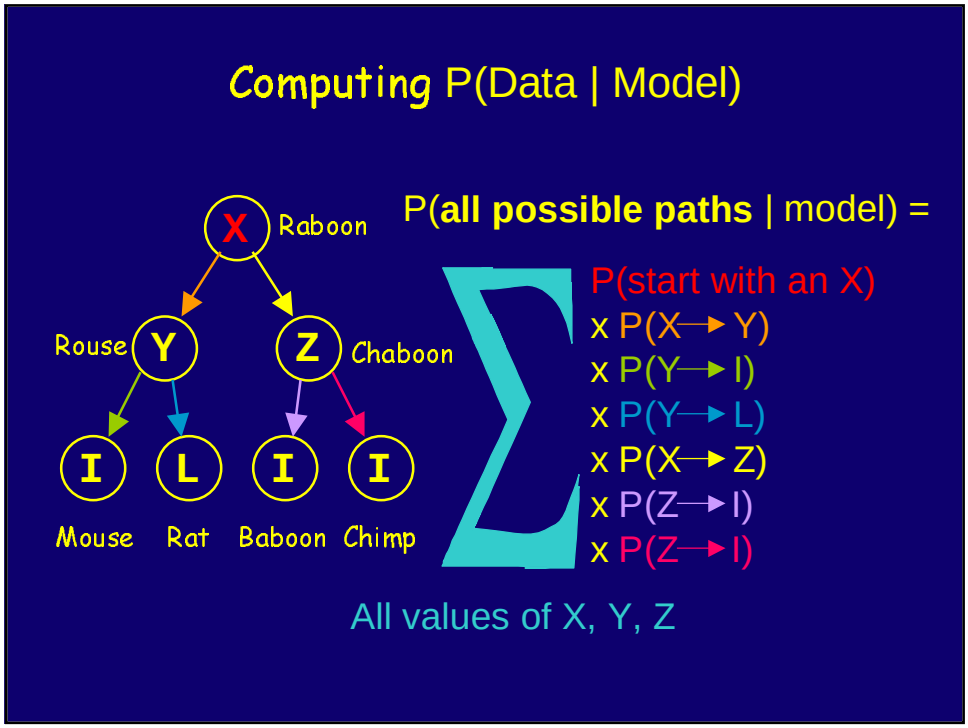
We need to find the best model for the data

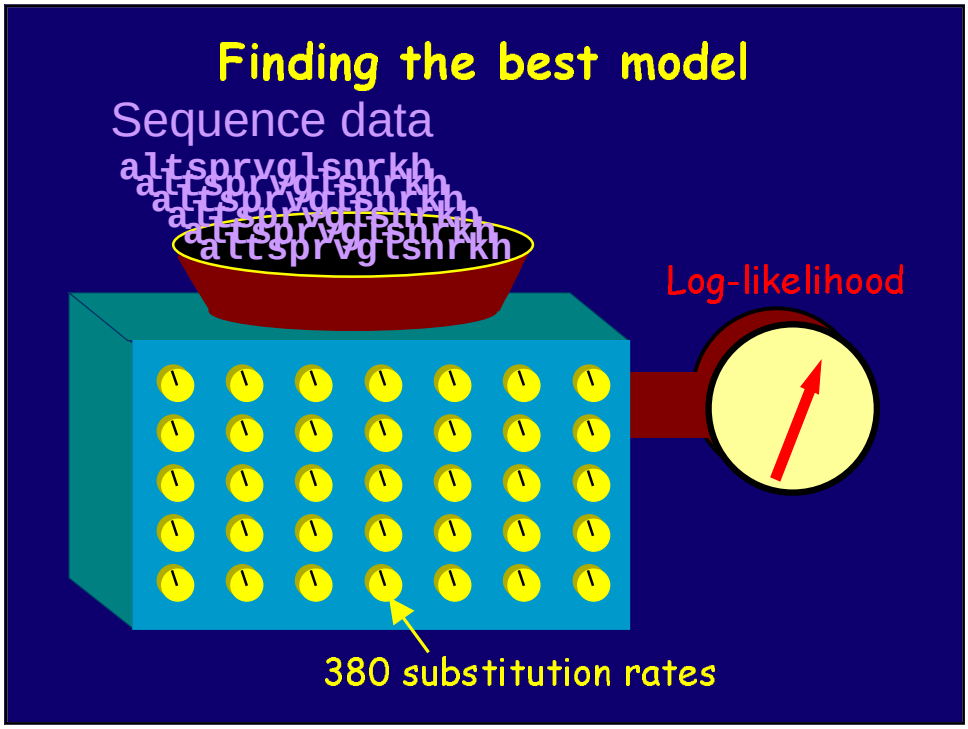
Computing P(Data | Model)

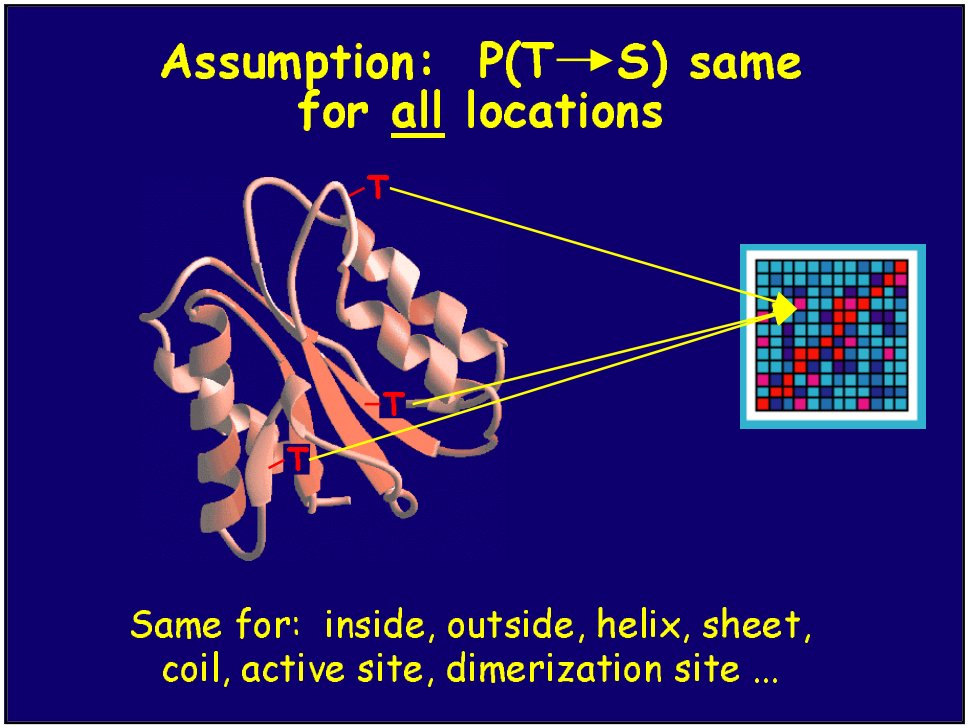


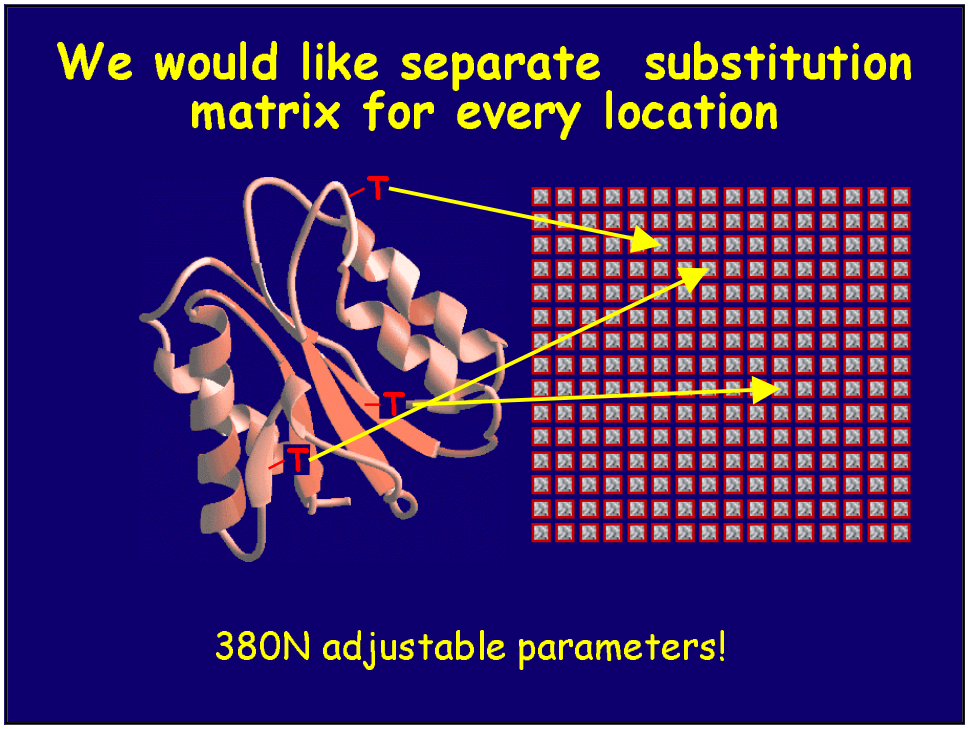
P(this particular path | model) =

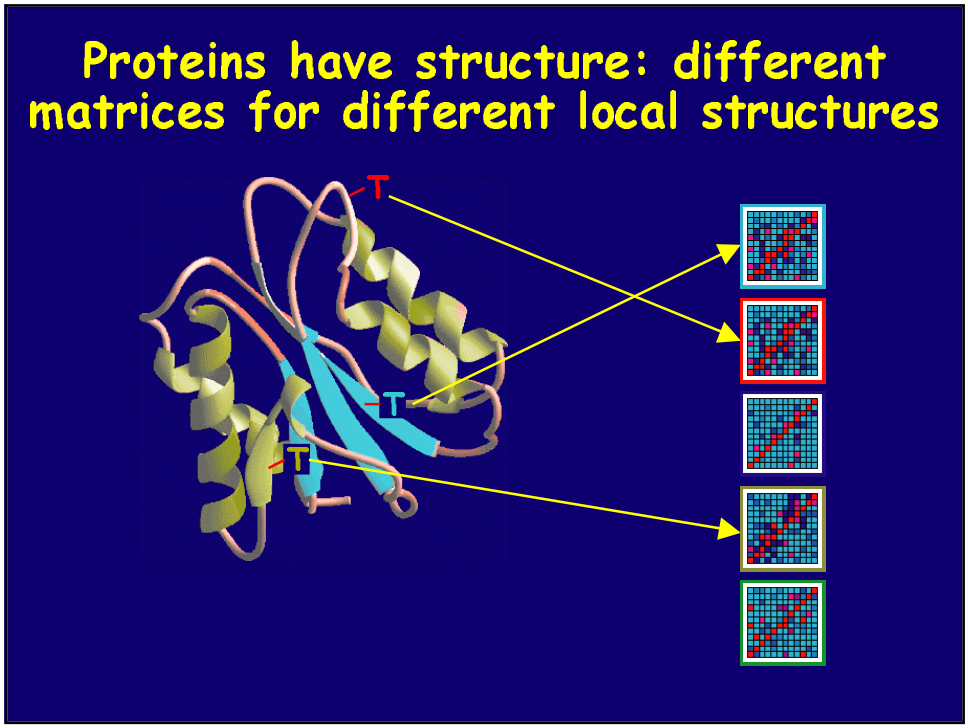
P(start with an I)
x P(I → L)
x P(L → I)
x P(L → L)
x P(I → I)
x P(I → I)
x P(I → I)

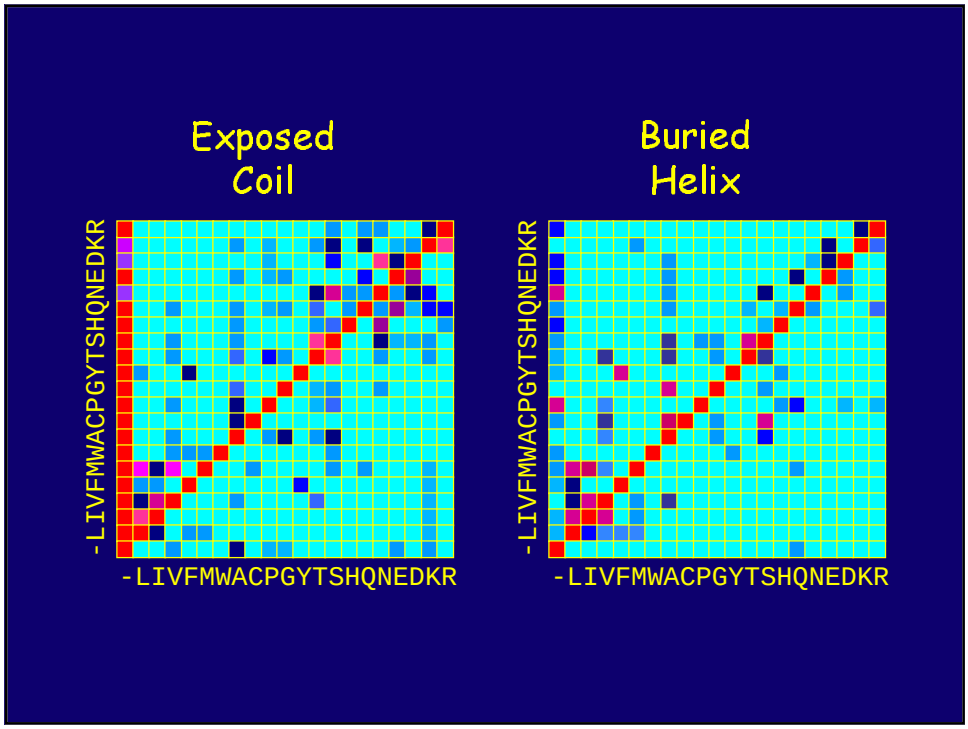

Computing P(Data | Model)
X
Raboon
Rouse
Y
Z
Chaboon
I
L
I
I
Mouse
Rat
Baboon
Chimp
P(all possible paths | model) =
P(start with an X)
x P(X→Y)
x P(Y→I)
x P(Y→L)
x P(X→Z)
x P(Z→I)
x P(Z→I)
All values of X, Y, Z

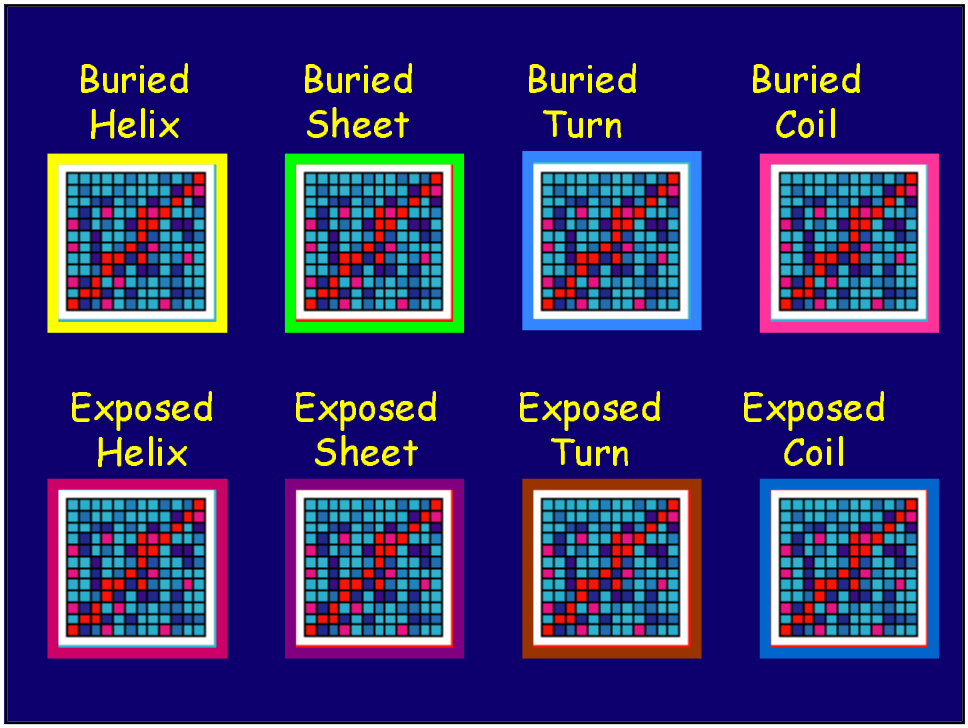

Finding the best model
Sequence data
altsprvglsnrkh
altsprvgtsnrkh
altsprvgtsnrkh
altsprvgtsnrkh
altsprvgtsnrkh
Log-likelihood
380 substitution rates

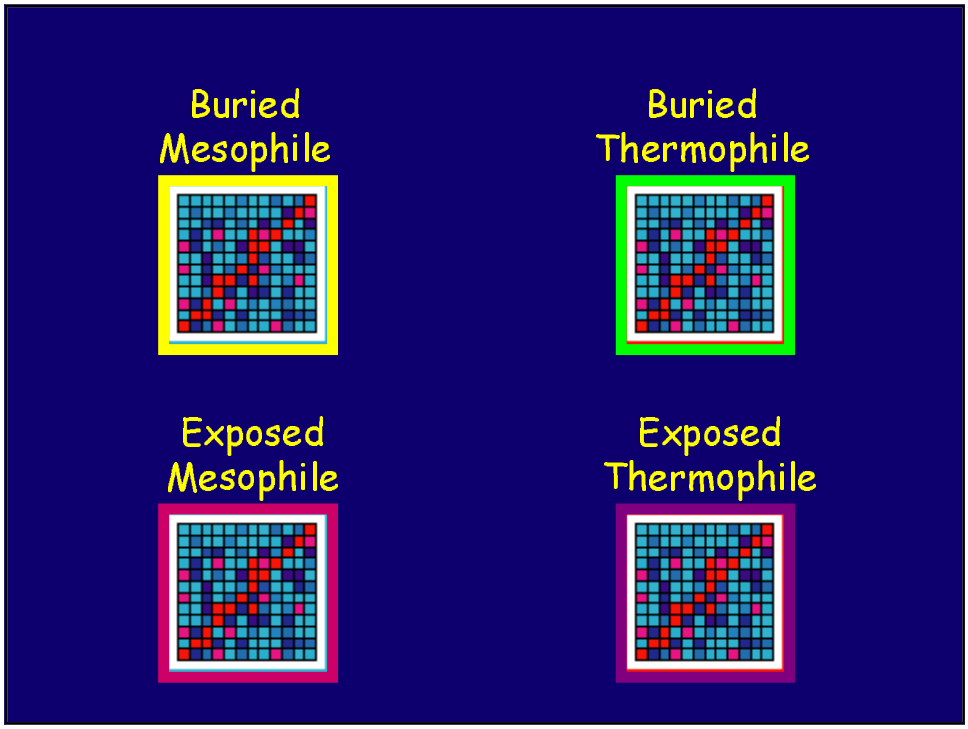

Assumption: P(T→S) same for all locations
T
T
T
Same for: inside, outside, helix, sheet, coil, active site, dimerization site ...


We would like separate substitution matrix for every location
T
T
T
380N adjustable parameters!


Proteins have structure: different matrices for different local structures
T
T
T


Exposed Coil
Buried Helix
-LIVFMWACPGYTSHQNEDKR
-LIVFMWACPGYTSHQNEDKR
-LIVFMWACPGYTSHQNEDKR
-LIVFMWACPGYTSHQNEDKR


Buried Helix
Buried Sheet
Buried Turn
Buried Coil
Exposed Helix
Exposed Sheet
Exposed Turn
Exposed Coil

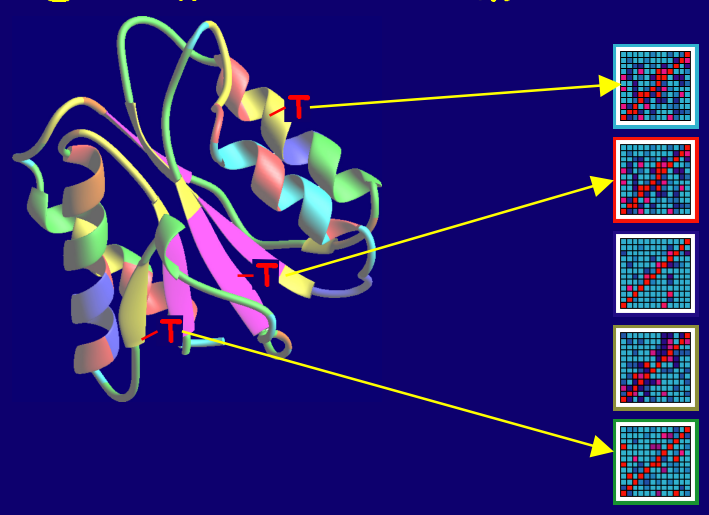

Buried Mesophile
Buried Thermophile
Exposed Mesophile
Exposed Thermophile

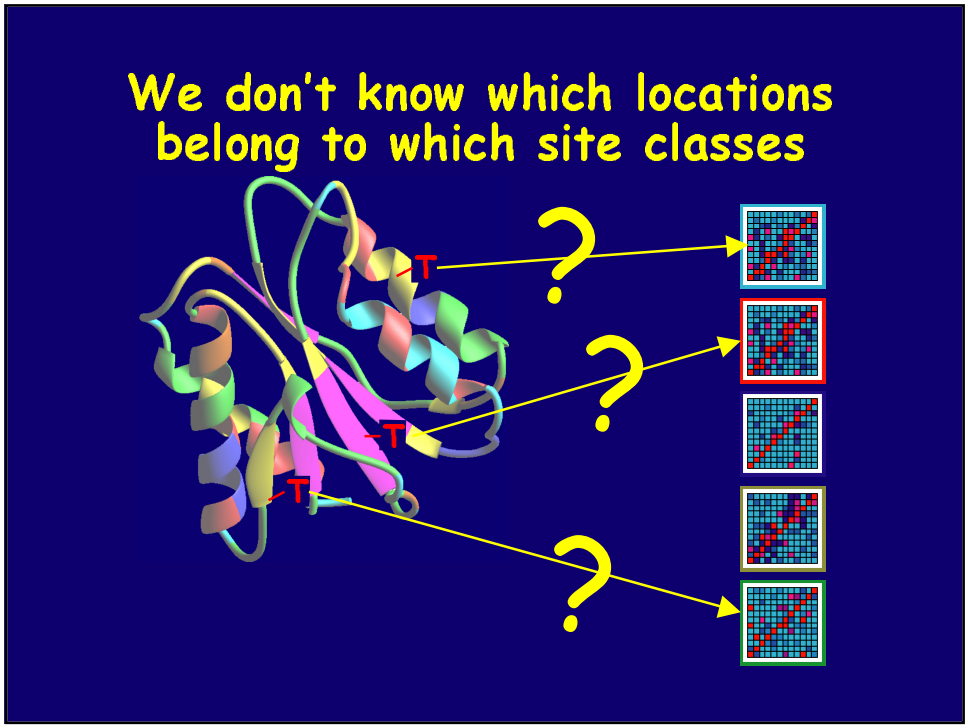
Is this enough?

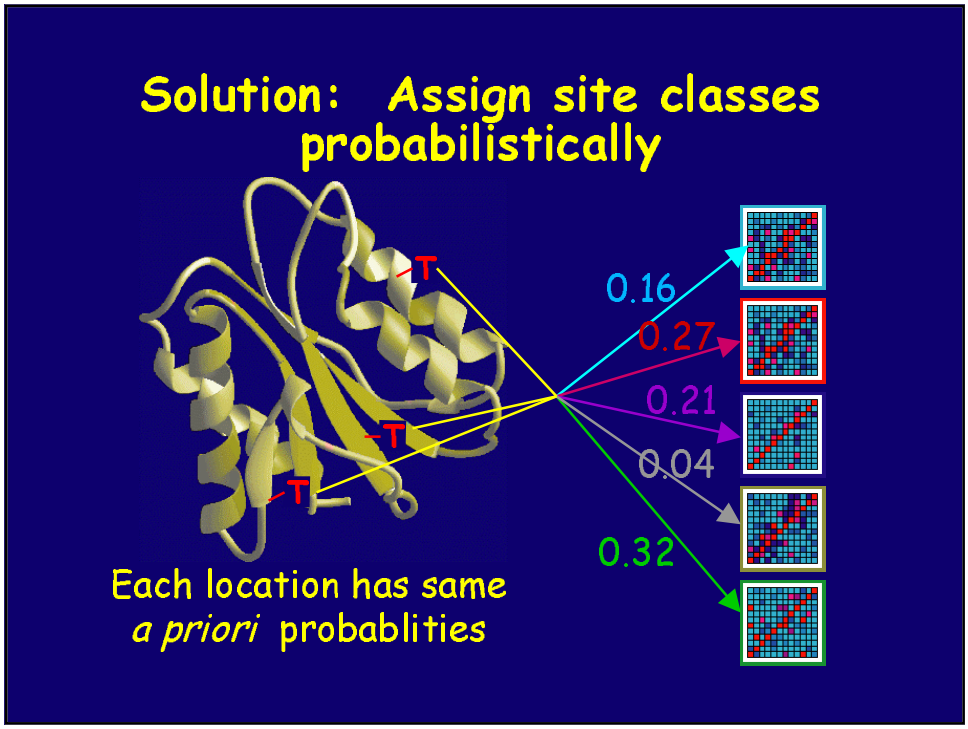
Assumes all locations in a given local structure evolve identically

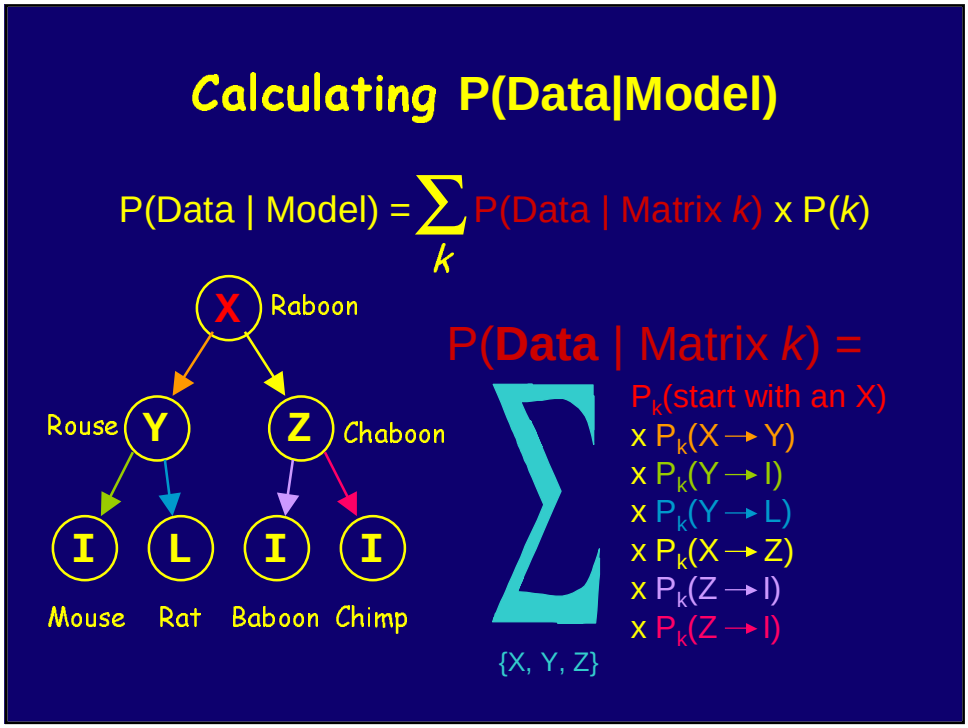
- Ignores complex nature of structural constraints
- Ignores functional constraints
  - active sites
  - dimerization sites
- Ignores any other type of selective pressure
- Designation between local structure categories somewhat arbitrary
- What about proteins of unknown structure?

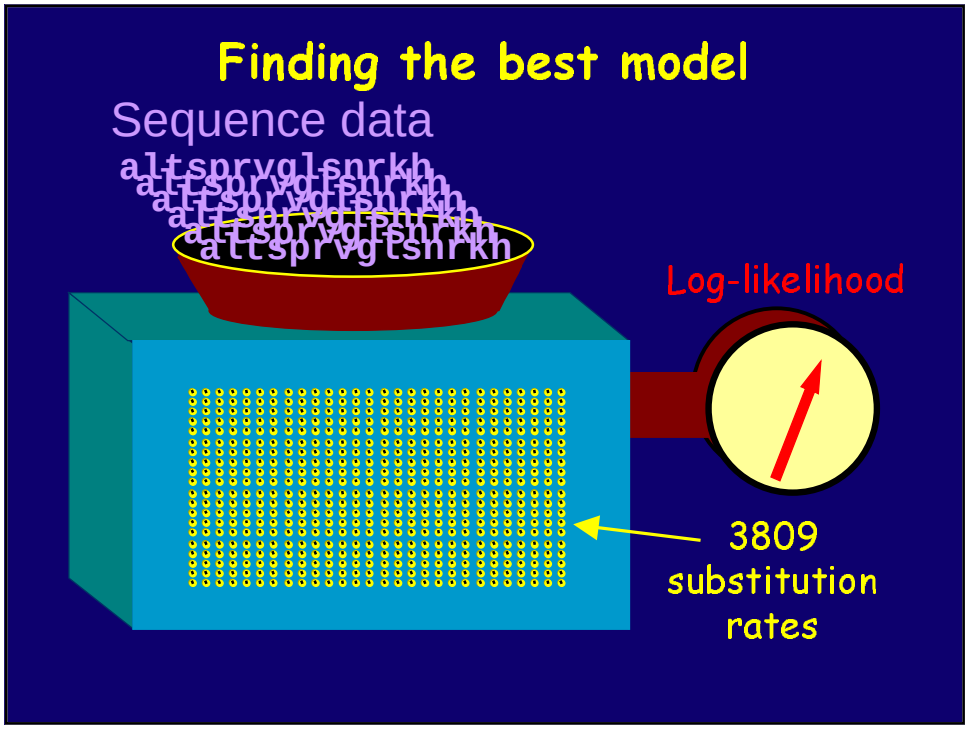
Different site classes in the protein, Each with its own matrices




We don't know which locations belong to which site classes
T
T
T
?
?
?


Solution: Assign site classes probabilistically
T
T
T
0.16
0.27
0.21
0.04
0.32
Each location has same
a priori probablities


Calculating P(Data|Model)
P(Data | Model) = $\sum_k$ P(Data | Matrix k) x P(k)
X
Raboon
Rouse
Y
Z
Chaboon
I
L
I
I
Mouse
Rat
Baboon
Chimp
P(Data | Matrix k) =
$\sum_{\{X, Y, Z\}}$
P_k(start with an X)
x P_k(X → Y)
x P_k(Y → I)
x P_k(Y → L)
x P_k(X → Z)
x P_k(Z → I)
x P_k(Z → I)

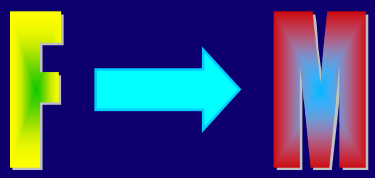

Finding the best model
Sequence data
altsprvglsnrkh
altsprvgtsnrkh
altsprvgtsnrkh
altsprvgtsnrkh
altsprvgtsnrkh
Log-likelihood
3809
substitution
rates

We need:

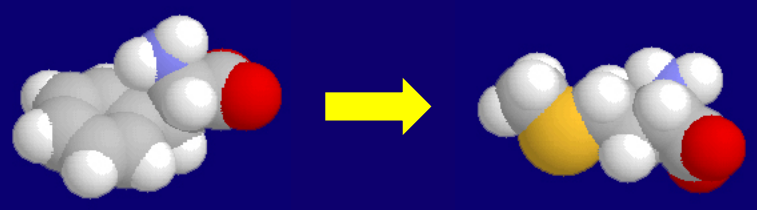
- Simpler substitution model
- Many fewer adjustable parameters
- Biologically-motivated

Bioinformaticists consider alphabets

What is the probability of an F changing to an M?



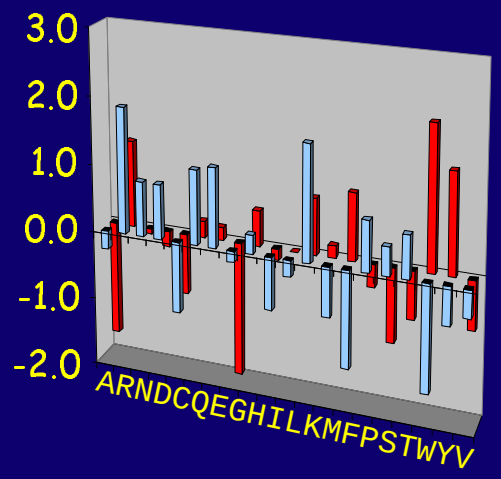
Proteins care about amino-acid properties



What is the rate of substitution amino acids with the properties of Phe by an amino acid with the properties of Met?

Step 1: Represent the amino acids by their physical-chemical properties

	Hydrophobicity	Bulk
Ala	-0.27	-1.67
Arg	1.87	1.27
Asn	0.81	-0.07
Asp	0.81	-0.22
Cys	-1.05	-0.89
Gln	1.10	0.24
Glu	1.17	0.19
Gly	-0.16	-1.96
His	0.28	0.52
Ile	-0.77	-0.16
Leu	-0.24	0.00
Lys	1.70	0.82
Met	-0.73	0.18
Phe	-1.43	0.98
Pro	0.75	-0.33
Ser	0.42	-1.08
Thr	0.63	-0.70
Trp	-1.57	2.10
Tyr	-0.56	1.48
Val	-0.40	-0.71



Step 2: Represent the fitness F_k of any amino acid by a simple (unknown) function of its properties

Linear:

$$F_k(A_i) = \alpha_1 \times \text{hydro}(A_i) + \alpha_2 \times \text{bulk}(A_i)$$

Quadratic:

$$F_k(A_i) = \alpha_1 \times (\text{hydro}(A_i) - \text{hydro}_0)^2 + \alpha_2 \times (\text{bulk}(A_i) - \text{bulk}_0)^2$$

	Hydrophobicity	Bulk
Ala	-0.27	-1.67
Arg	1.87	1.27
Asn	0.81	-0.07
⋮	⋮	⋮

20 fitnesses with 2 to 4 adjustable parameters

Example:

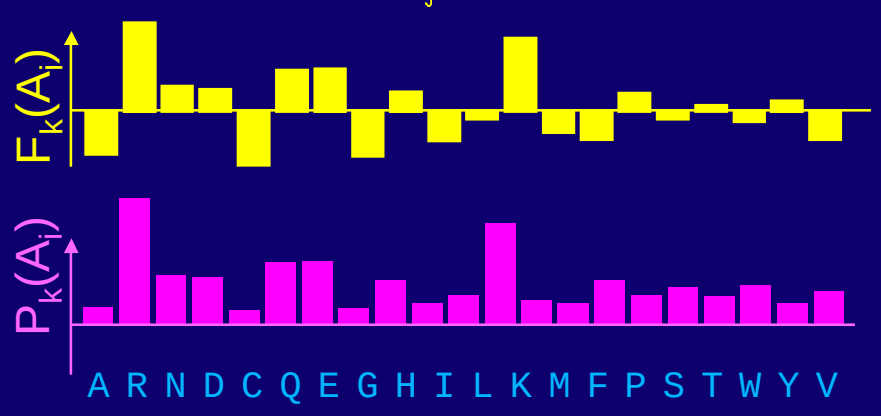
$$F_k(A_i) = 0.5 \times \text{hydro}(A_i) + 0.3 \times \text{bulk}(A_i)$$

Step 3: Probability P_k of finding a given amino acid at a site class k represented by Boltzmann function

$$P_k(A_i) = \frac{\exp(F(A_i))}{\sum_j \exp(F(A_j))}$$

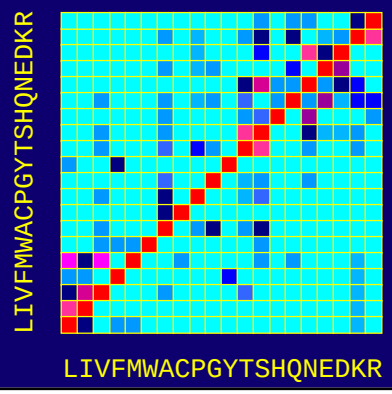
$$F_k(A_i) = 0.5 \times \text{hydro}(A_i) + 0.3 \times \text{bulk}(A_i)$$

$$P_k(A_i) = \frac{\exp(F(A_i))}{\sum_j \exp(F(A_j))}$$



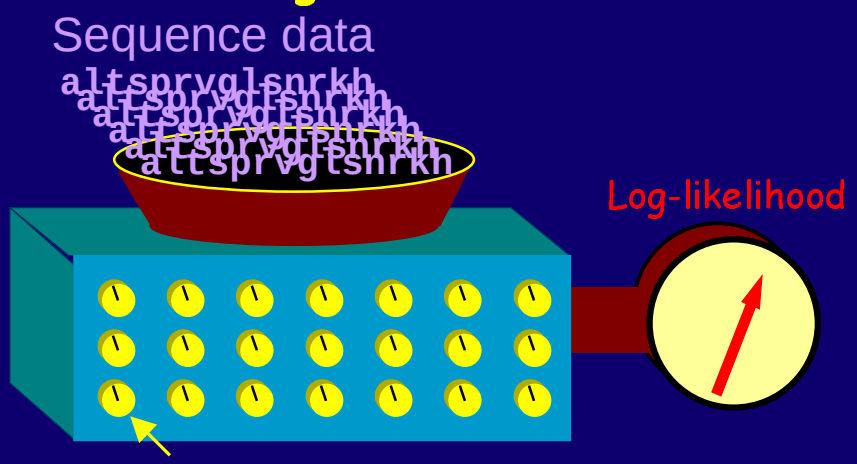
Step 4: Model substitution rates with Metropolis kinetics

$$P(A_i \rightarrow A_j) = \begin{cases} \nu & \text{increase in fitness} \\ \nu\, e^{\Delta F} & \text{decrease in fitness} \end{cases}$$



As a function of 3 to 5 adjustable parameters!

Finding the best model



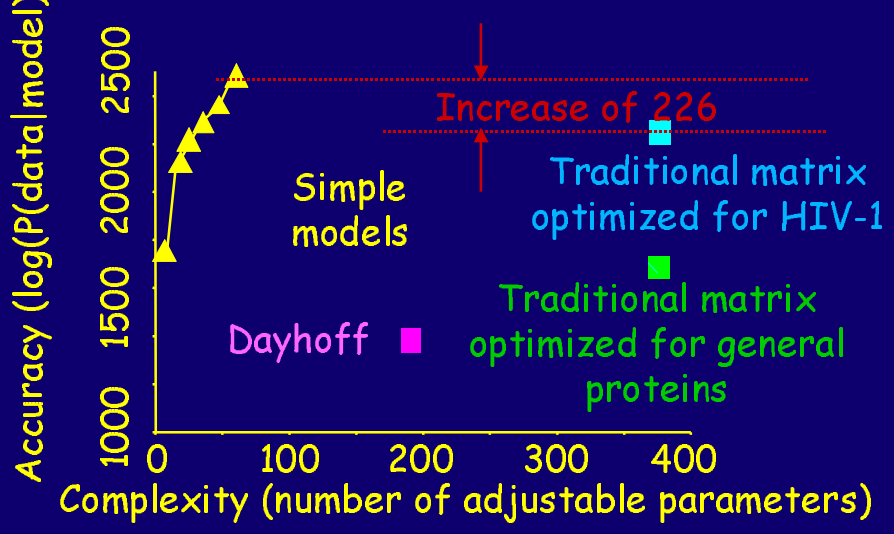
9 site classes: 6 quadratic, 3 linear
(6 x 4) + (3 x 2) + 9 + (9 - 1)
47 adjustable parameters

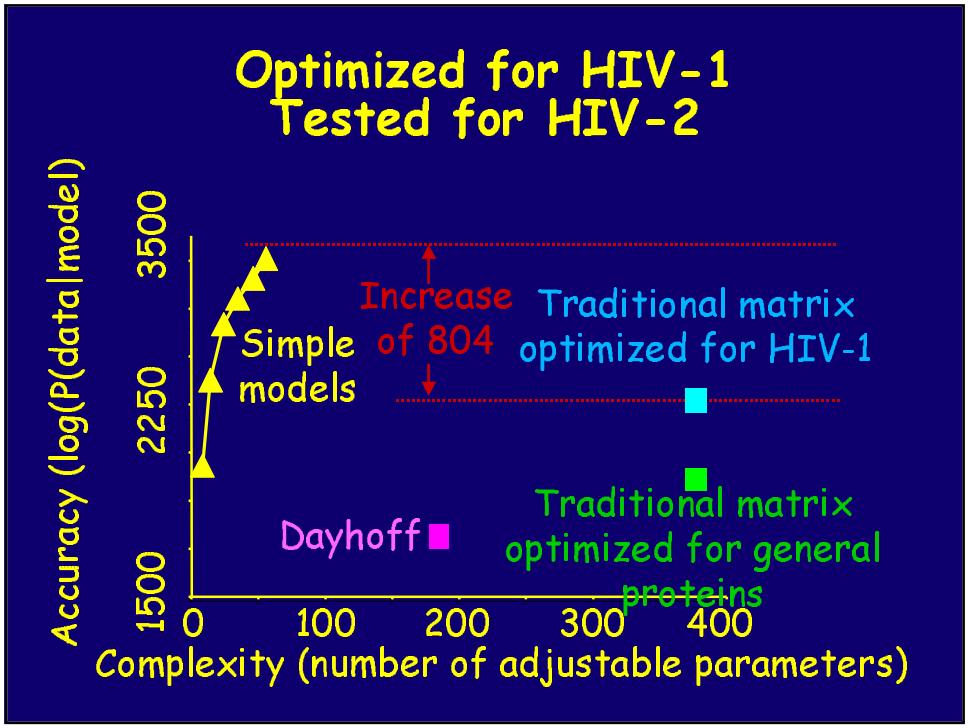
What works better

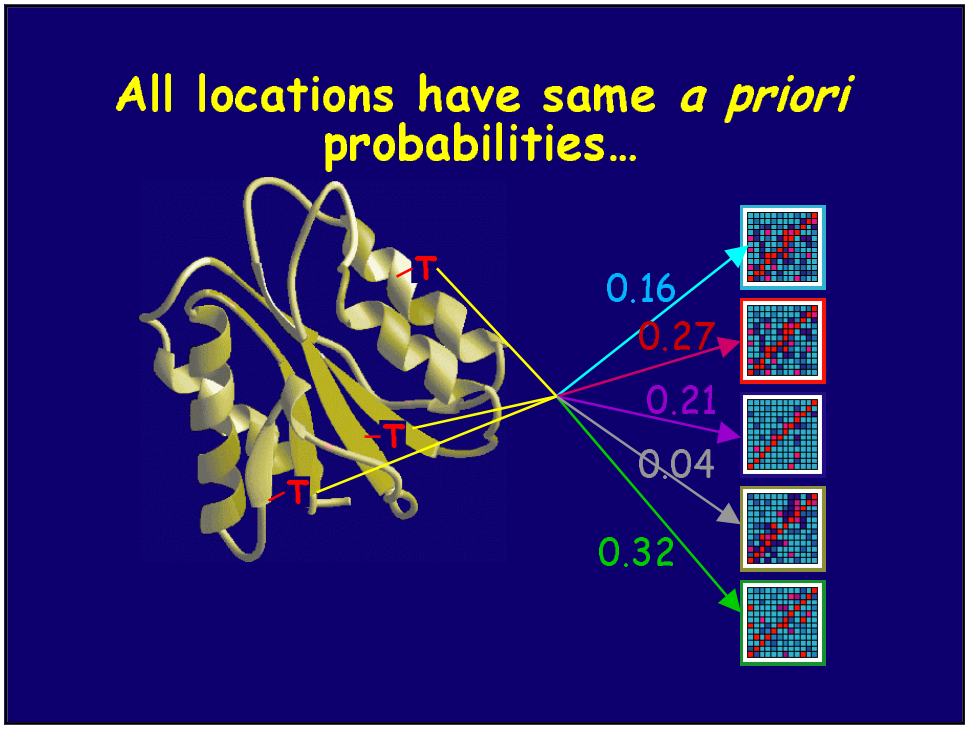
Traditional substitution matrix
- all rates can vary independently
- 380 adjustable parameters
- ignore site heterogeneity

Simple physical-based models
- depends on hydrophobicity, bulk
- Fitness simple functions of these parameters
- Boltzmann distributions
- Metropolis Kinetics

Optimized for HIV-1
Tested for HIV-1




Optimized for HIV-1
Tested for HIV-2
Accuracy (log(P(data|model))
3500
2250
1500
Simple models
Increase of 804
Traditional matrix optimized for HIV-1
Dayhoff
Traditional matrix optimized for general proteins
0
100
200
300
400
Complexity (number of adjustable parameters)

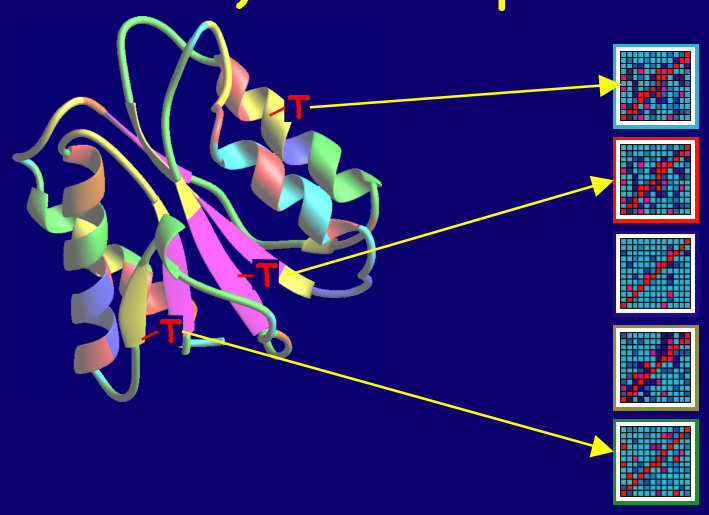

All locations have same *a priori* probabilities…
T
T
T
0.16
0.27
0.21
0.04
0.32

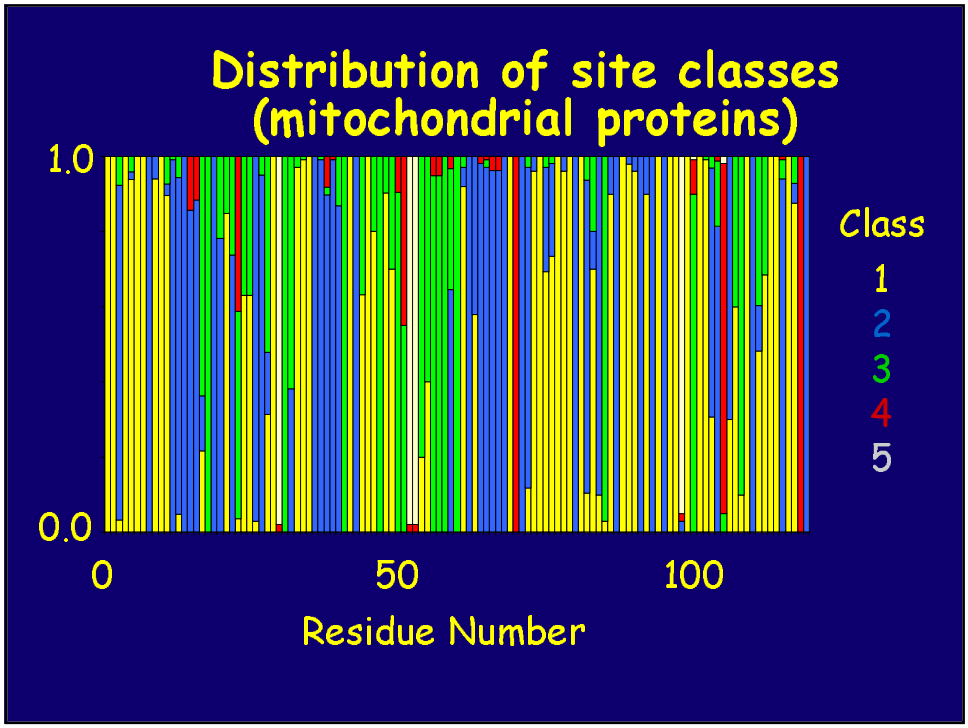
We can assign site classes *a posteriori*

P(Site class k|Data at location l)

$$= \frac{P(\text{Data}|\text{Model for site class } k)\, P(k)}{\sum_{k'} P(\text{Data}|\text{Model for site class } k')\, P(k')}$$

Identify locations under similar (or distinctive) selective pressures





These models are much, much better

Current and New Applications

- Phylogenetic reconstruction
- Reconstructing ancient proteins
- Understanding the effect of structure, function and environment on the selective pressure
- Developing new methods for identifying homologies

Acknowledgments
Matt Dimmic, Michael Thompson, Sara Ingalls, Bin Qian
Darin Taverna, Maricel Kann, Richard Goldstein, Paul Williams,
Nicholas Buchler, Jeff Koshi, Ting-Lan Chiu, Sridhar Govindarajan