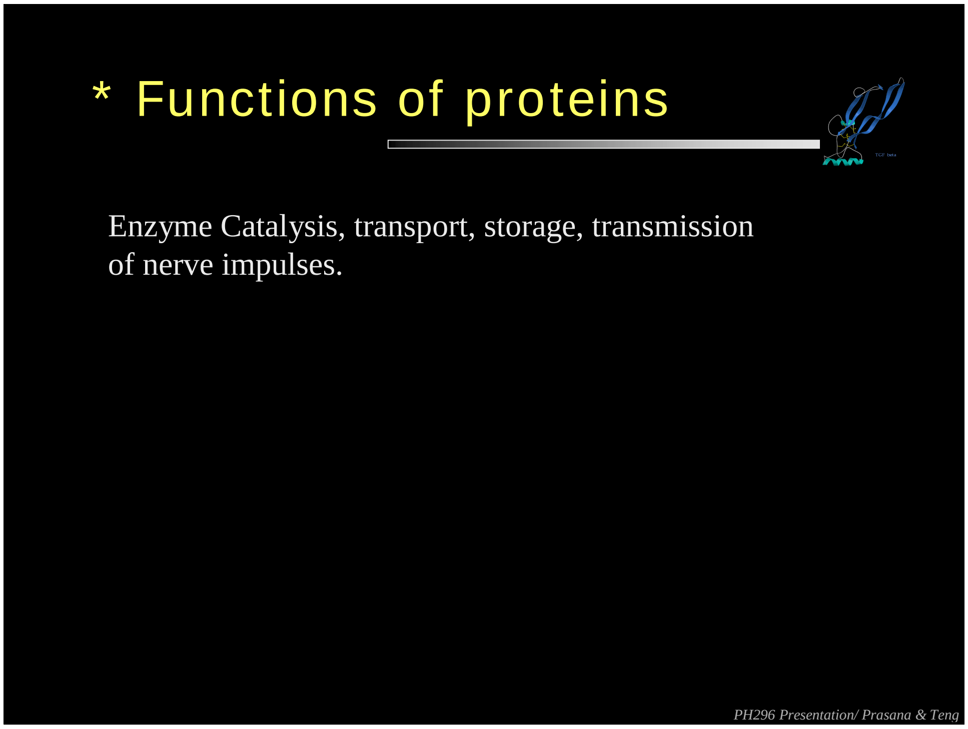
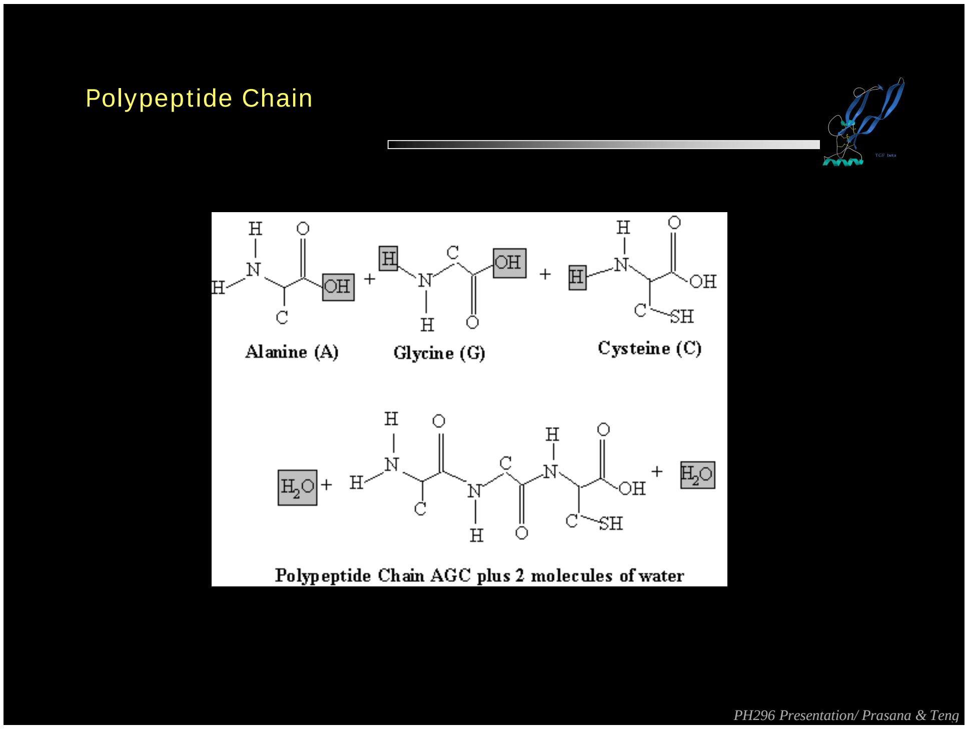


## * Functions of proteins

Enzyme Catalysis, transport, storage, transmission of nerve impulses.

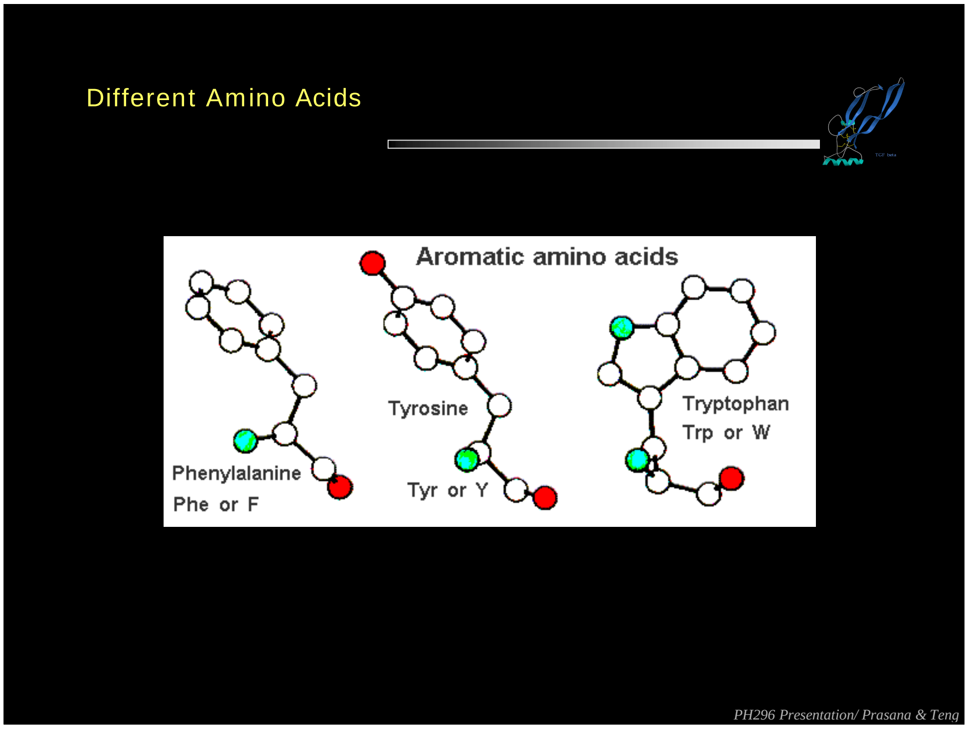
*PH296 Presentation/ Prasana & Teng*



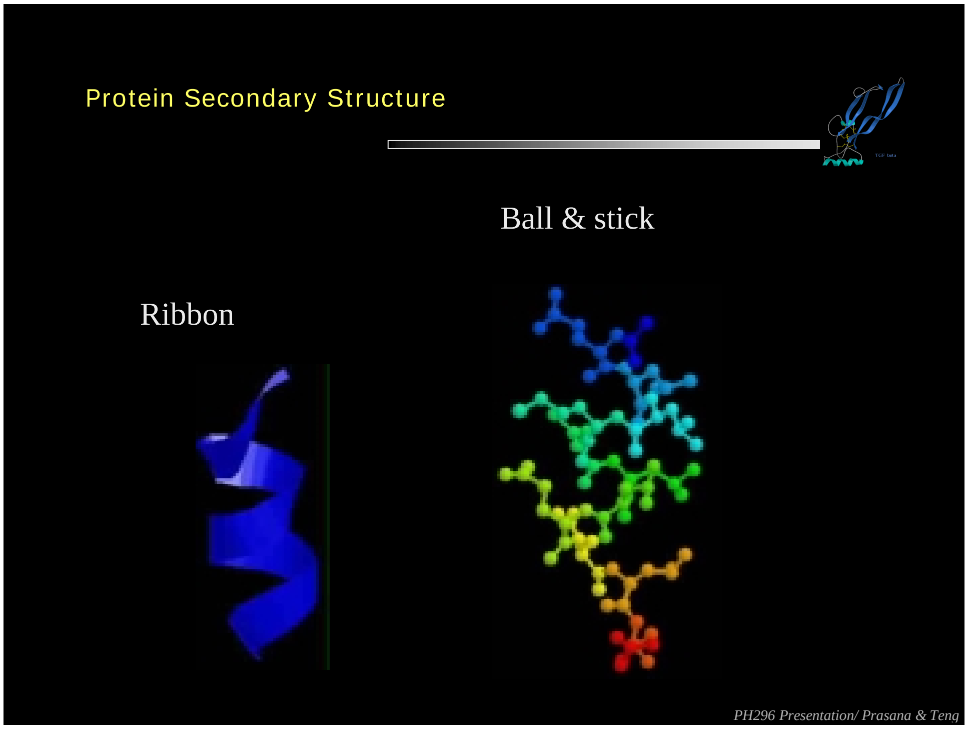
## Polypeptide Chain

Polypeptide Chain AGC plus 2 molecules of water

*PH296 Presentation/ Prasana & Teng*



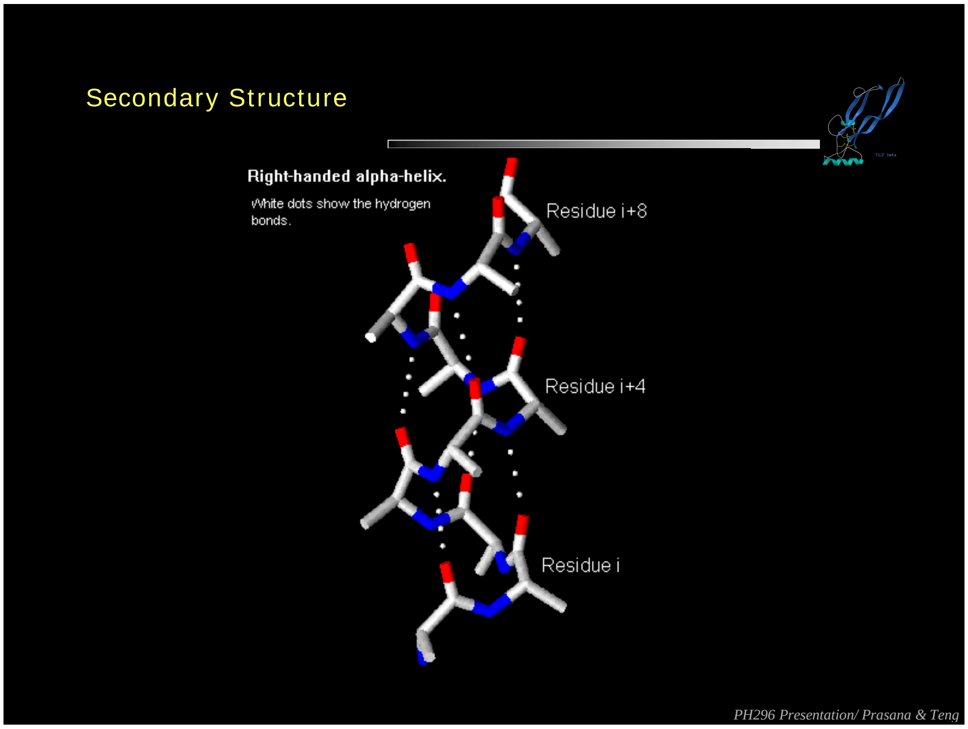
## Different Amino Acids



## Protein Secondary Structure

Ball & stick

Ribbon

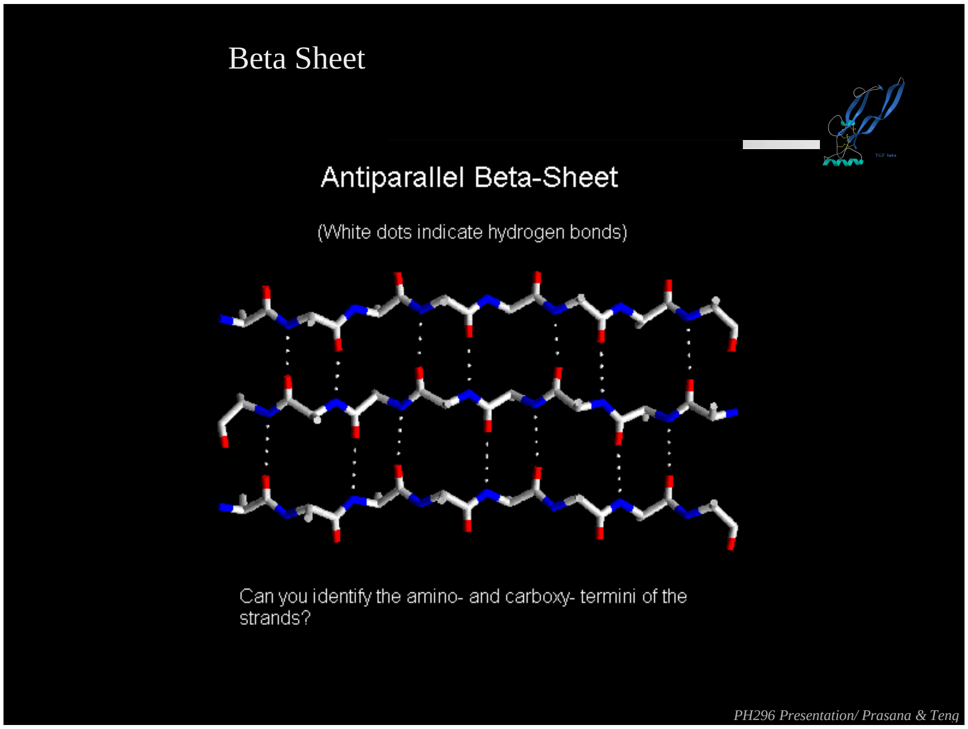


## Secondary Structure

**Right-handed alpha-helix.**

White dots show the hydrogen bonds.

*PH296 Presentation/ Prasana & Teng*



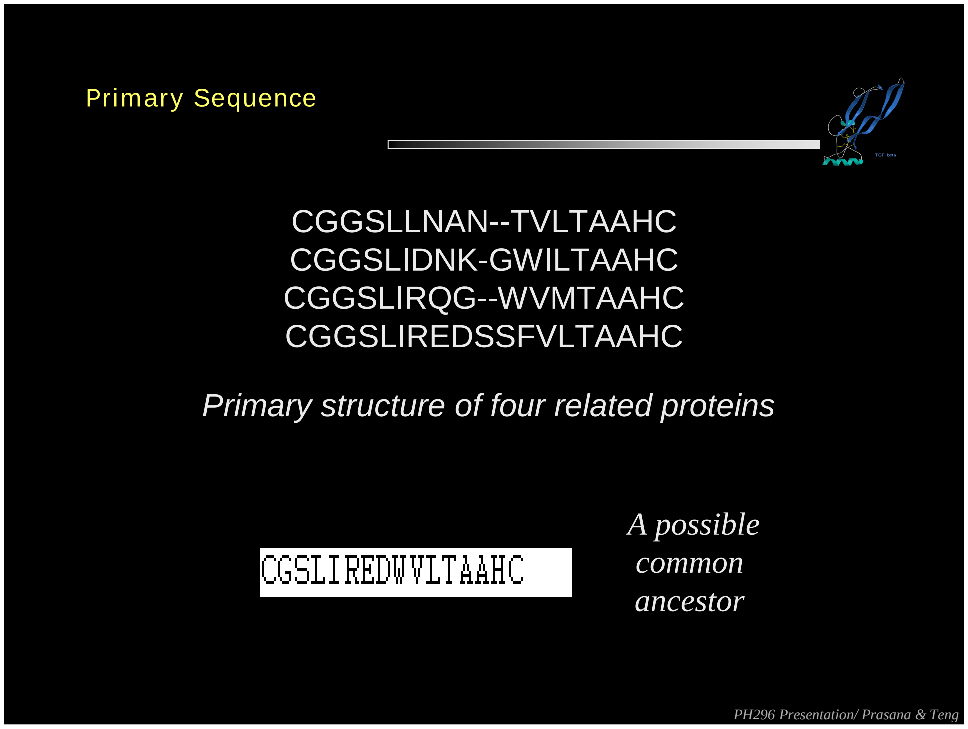
## Beta Sheet

### Antiparallel Beta-Sheet

(White dots indicate hydrogen bonds)

Can you identify the amino- and carboxy- termini of the strands?

*PH296 Presentation/ Prasana & Teng*



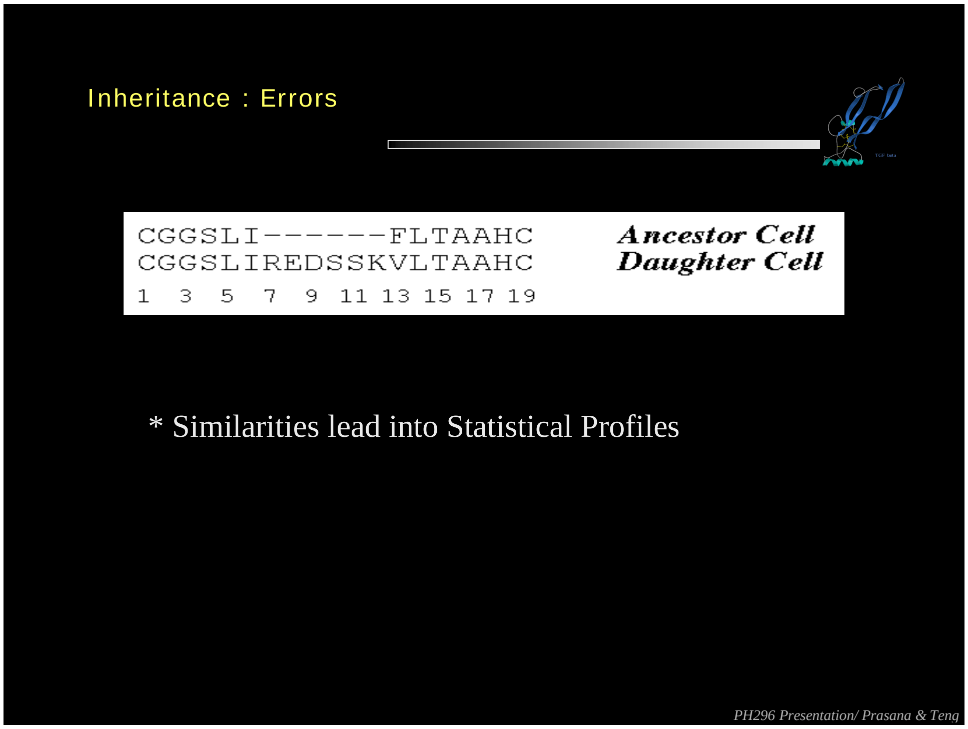
## Primary Sequence

CGGSLLNAN--TVLTAAHC
CGGSLIDNK-GWILTAAHC
CGGSLIRQG--WVMTAAHC
CGGSLIREDSSFVLTAAHC

*Primary structure of four related proteins*

CGSLIREDWVLTAAHC

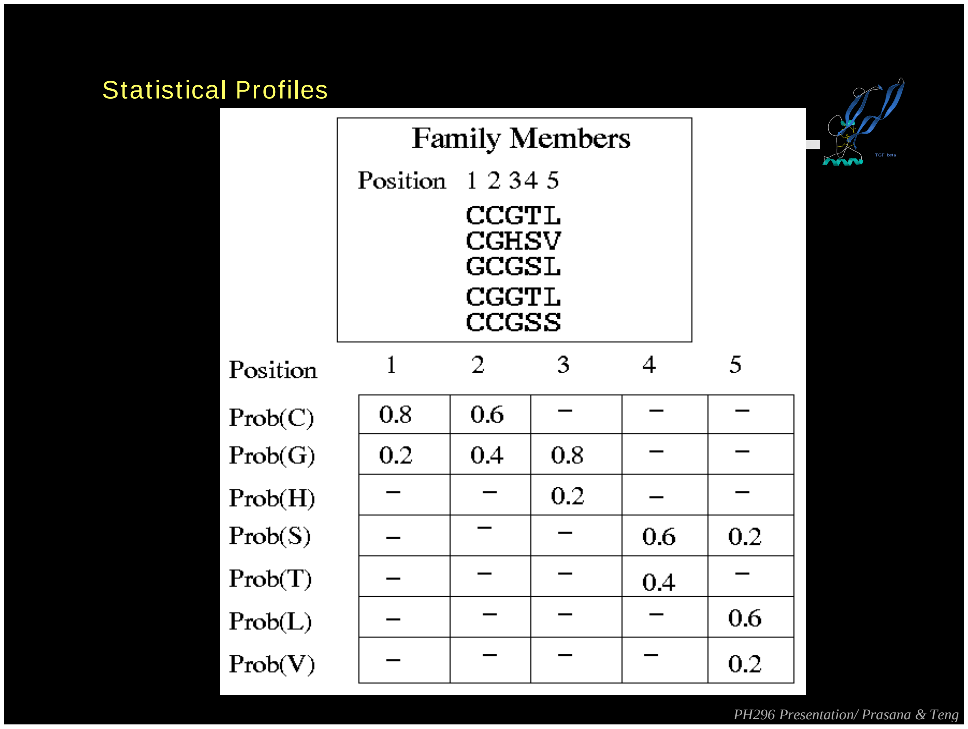
*A possible common ancestor*



## Inheritance : Errors

```
CGGSLI------FLTAAHC      Ancestor Cell
CGGSLIREDSSKVLTAAHC      Daughter Cell
1  3  5  7  9 11 13 15 17 19
```

* Similarities lead into Statistical Profiles



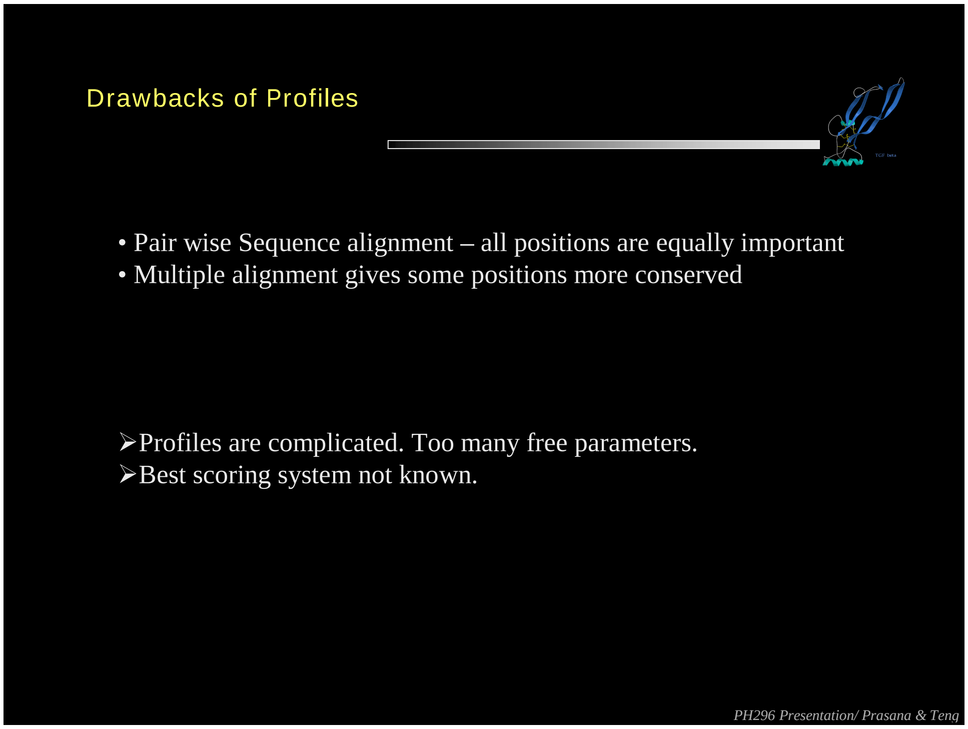
## Statistical Profiles

Family Members

Position 1 2 34 5

CCGTL<br>
CGHSV<br>
GCGSL<br>
CGGTL<br>
CCGSS

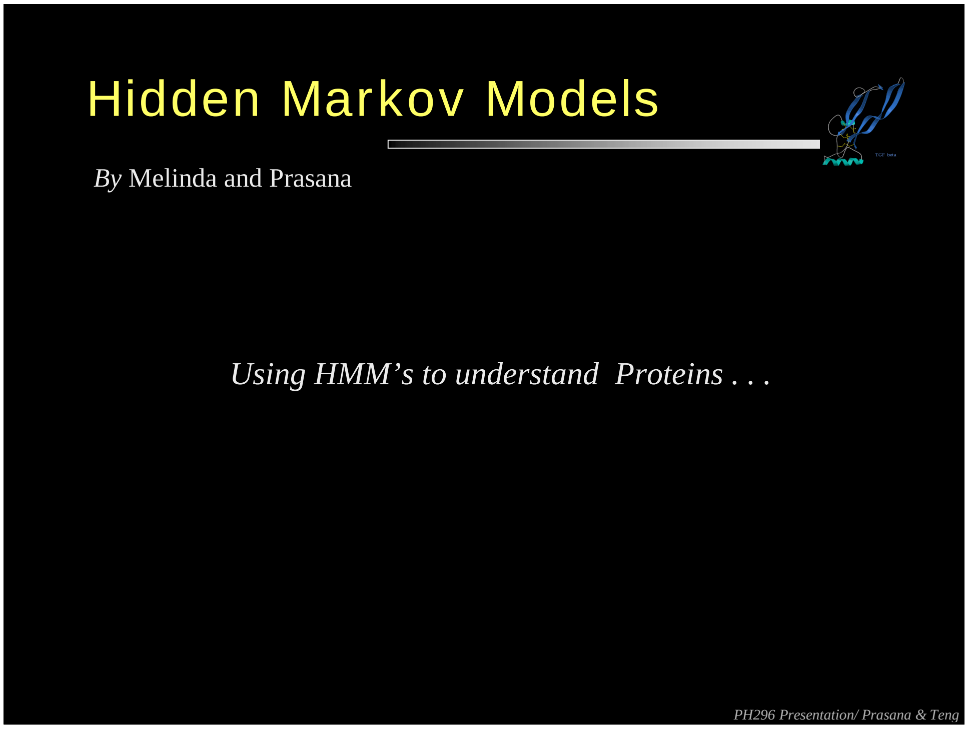
| Position | 1 | 2 | 3 | 4 | 5 |
|---|---|---|---|---|---|
| Prob(C) | 0.8 | 0.6 | – | – | – |
| Prob(G) | 0.2 | 0.4 | 0.8 | – | – |
| Prob(H) | – | – | 0.2 | – | – |
| Prob(S) | – | – | – | 0.6 | 0.2 |
| Prob(T) | – | – | – | 0.4 | – |
| Prob(L) | – | – | – | – | 0.6 |
| Prob(V) | – | – | – | – | 0.2 |



## Drawbacks of Profiles

- Pair wise Sequence alignment – all positions are equally important
- Multiple alignment gives some positions more conserved

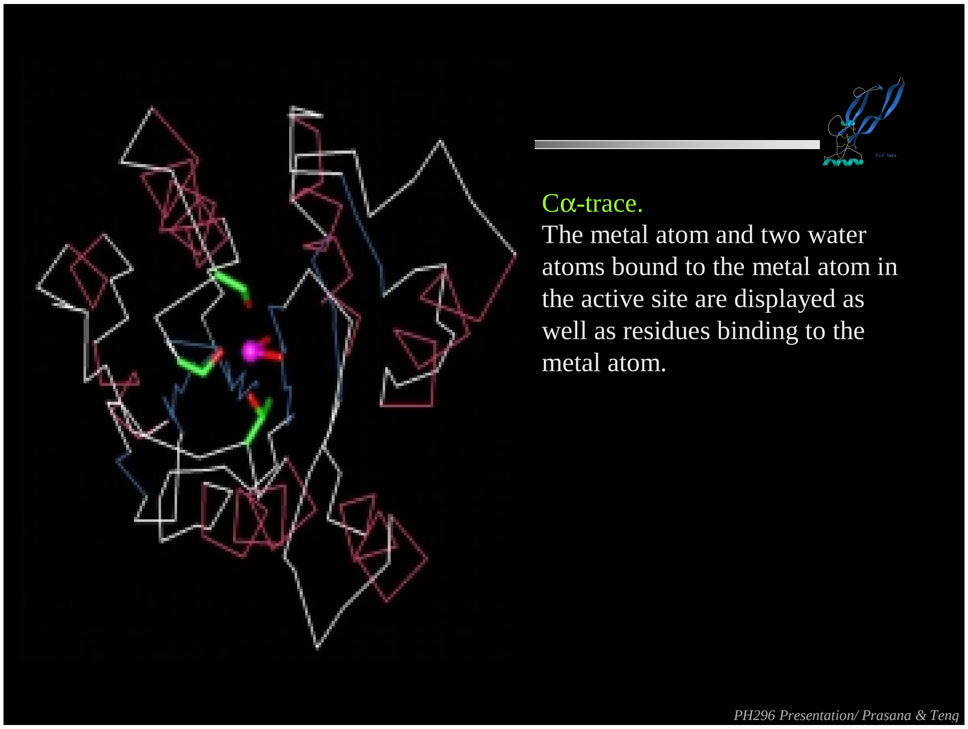
- ➢Profiles are complicated. Too many free parameters.
- ➢Best scoring system not known.



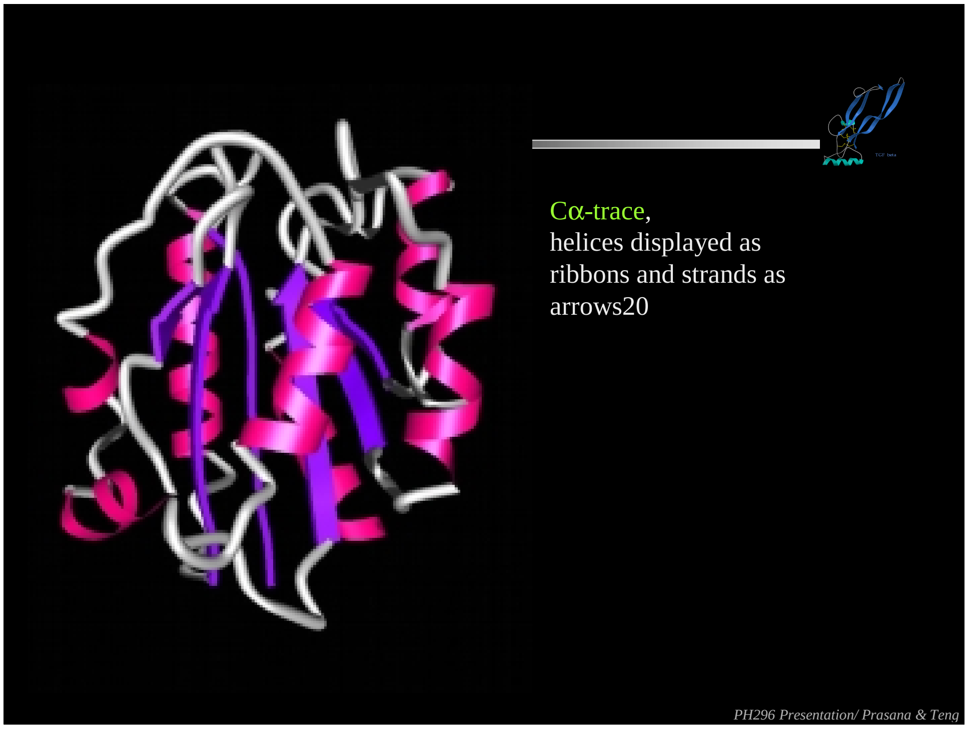
# Hidden Markov Models

*By* Melinda and Prasana

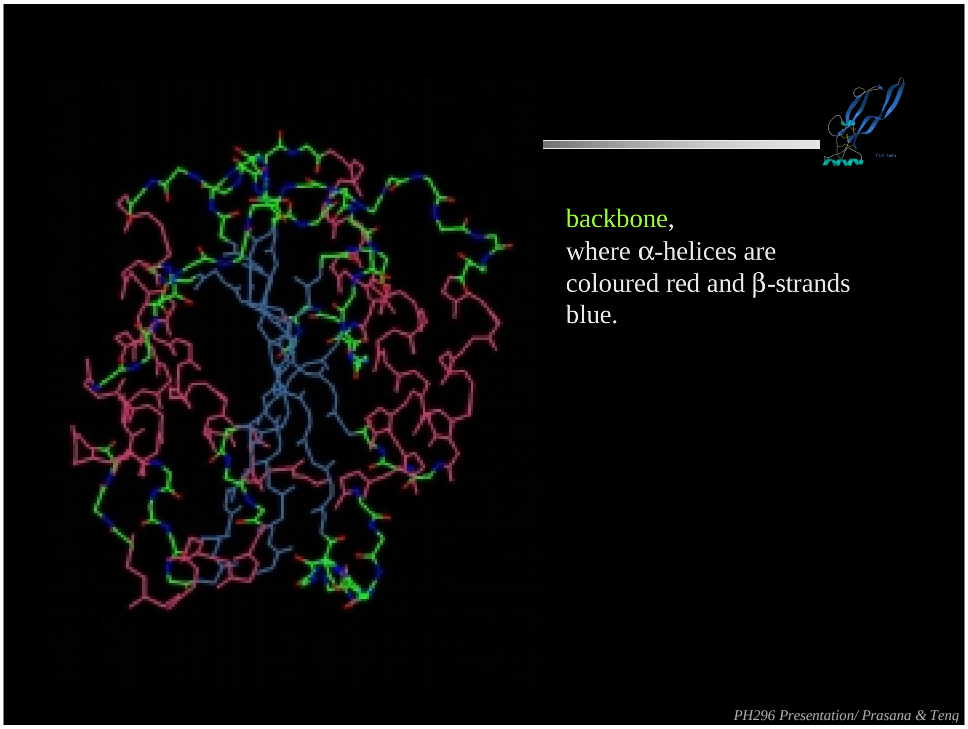
*Using HMM’s to understand  Proteins . . .*



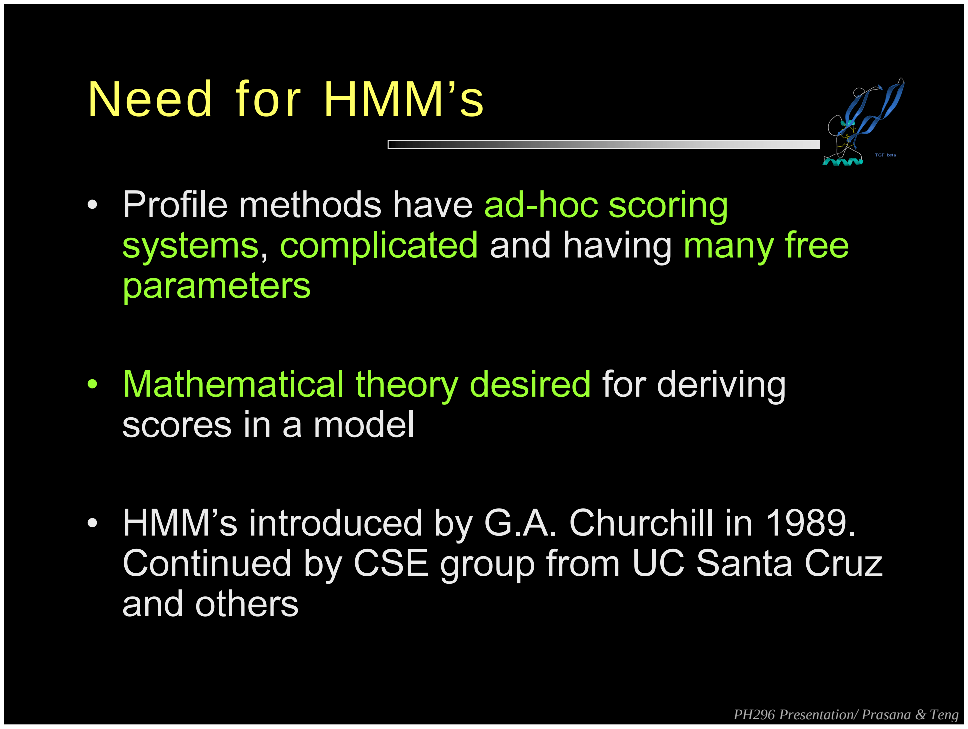
Cα-trace.
The metal atom and two water atoms bound to the metal atom in the active site are displayed as well as residues binding to the metal atom.



Cα-trace,
helices displayed as
ribbons and strands as
arrows20

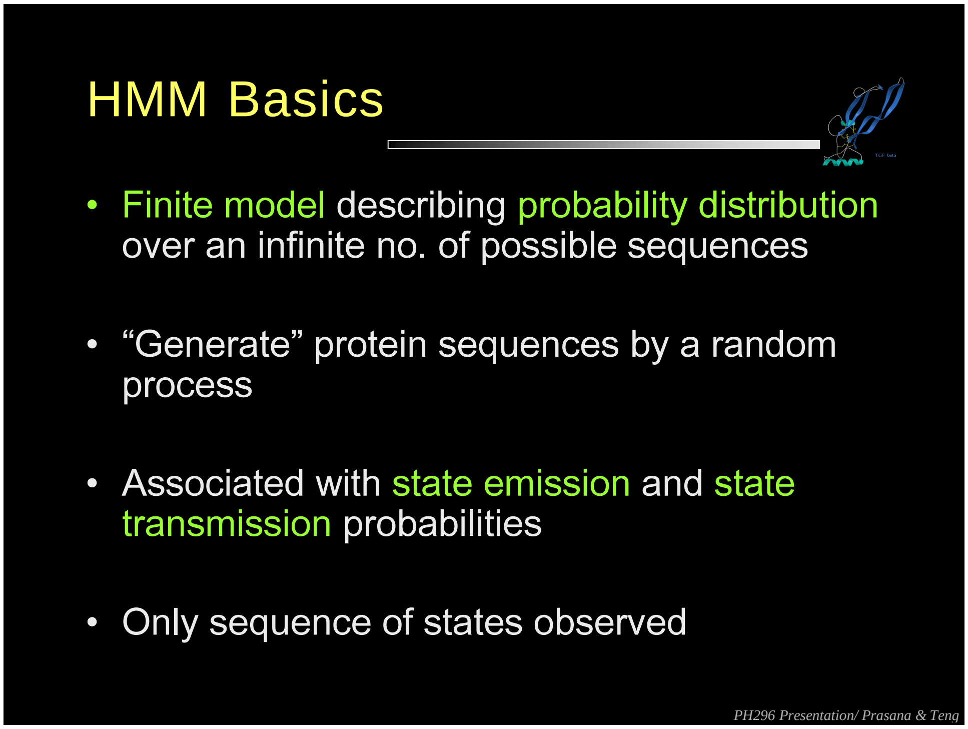


backbone,
where α-helices are
coloured red and β-strands
blue.



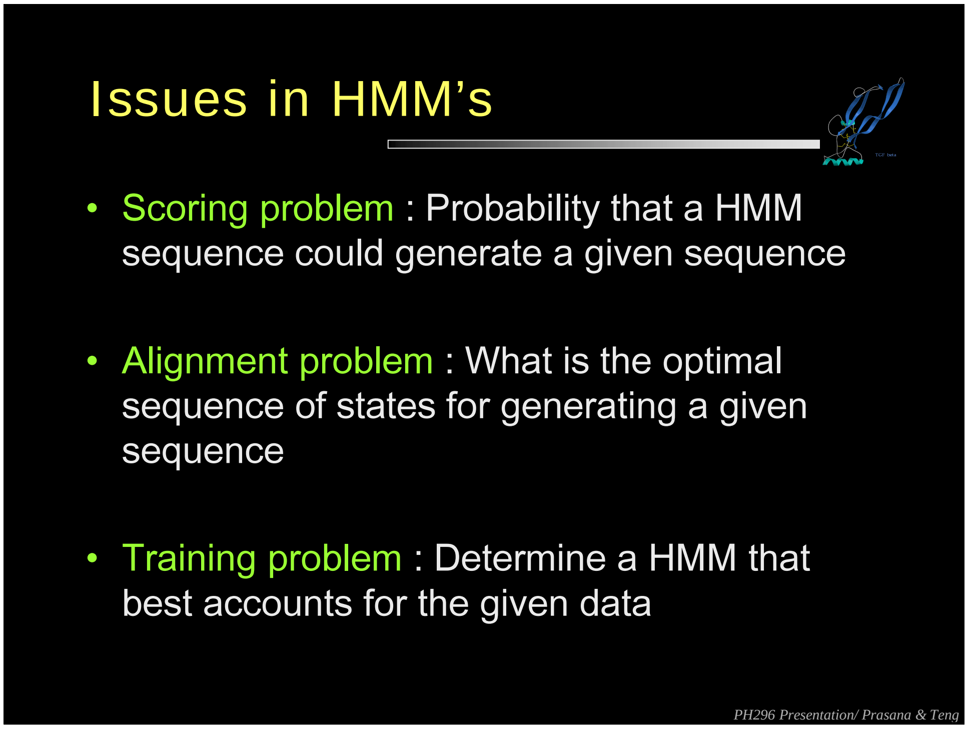
## Need for HMM’s

- Profile methods have ad-hoc scoring systems, complicated and having many free parameters
- Mathematical theory desired for deriving scores in a model
- HMM’s introduced by G.A. Churchill in 1989. Continued by CSE group from UC Santa Cruz and others



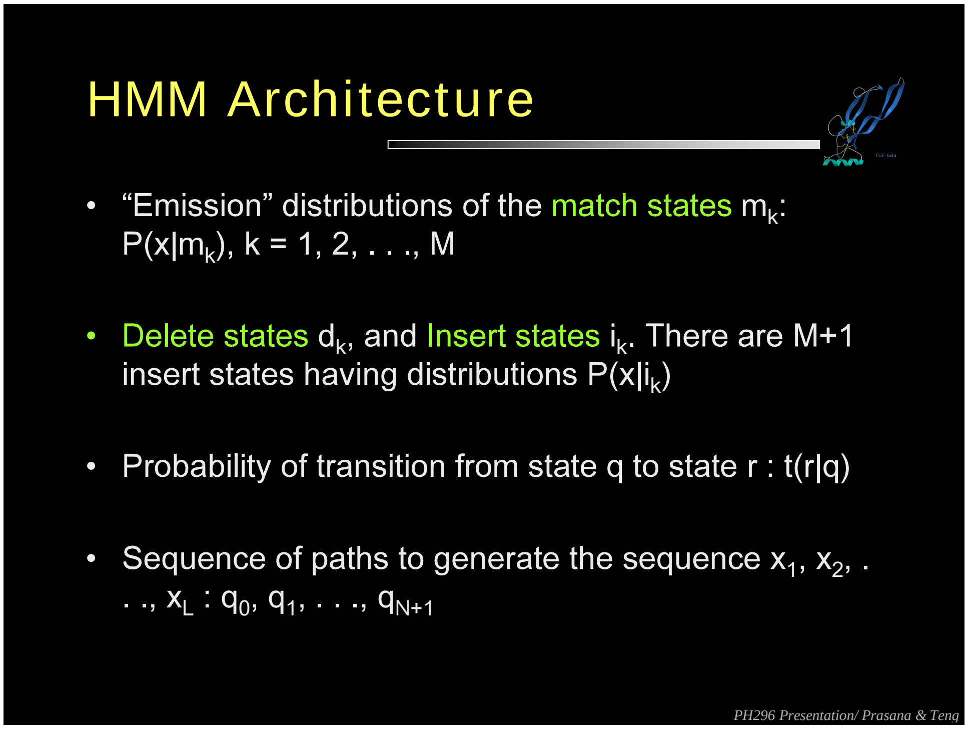
## HMM Basics

- Finite model describing probability distribution over an infinite no. of possible sequences
- “Generate” protein sequences by a random process
- Associated with state emission and state transmission probabilities
- Only sequence of states observed



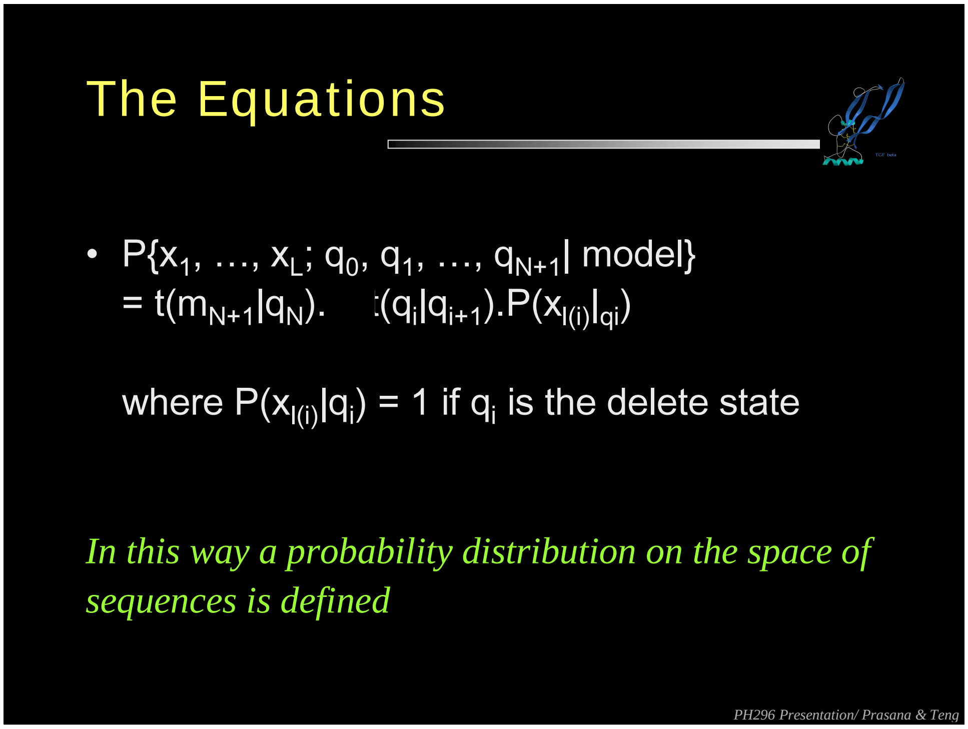
## Issues in HMM’s

- Scoring problem : Probability that a HMM sequence could generate a given sequence
- Alignment problem : What is the optimal sequence of states for generating a given sequence
- Training problem : Determine a HMM that best accounts for the given data



## HMM Architecture

- “Emission” distributions of the match states $m_k$: $P(x|m_k)$, k = 1, 2, . . ., M
- Delete states $d_k$, and Insert states $i_k$. There are M+1 insert states having distributions $P(x|i_k)$
- Probability of transition from state q to state r : $t(r|q)$
- Sequence of paths to generate the sequence $x_1, x_2, . . ., x_L : q_0, q_1, . . ., q_{N+1}$



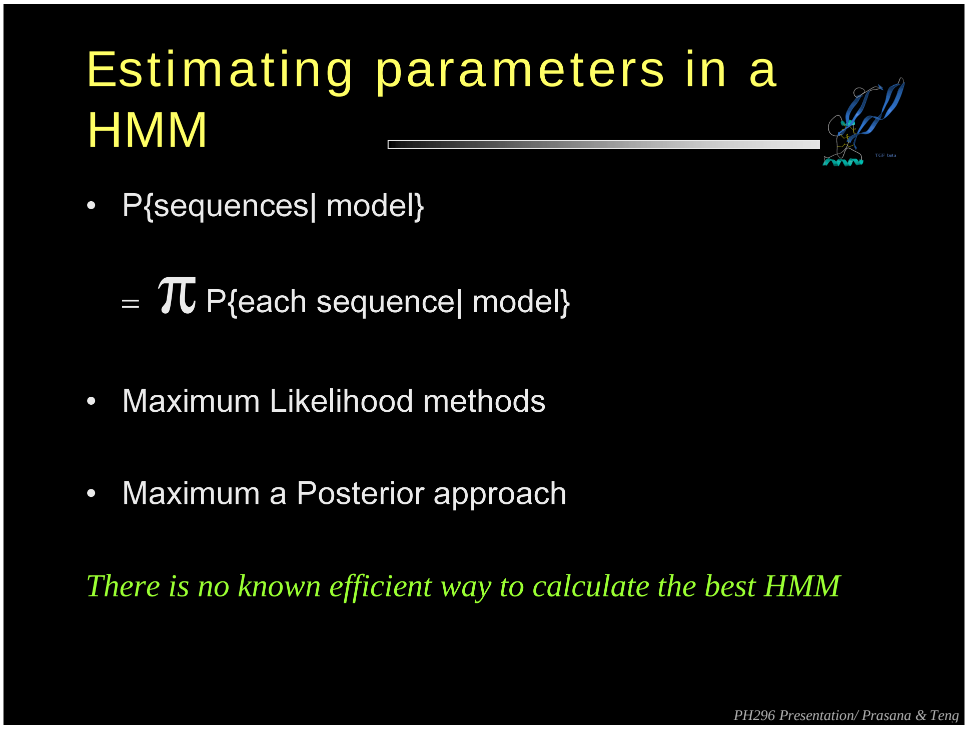
## The Equations

- $P\{x_1, \ldots, x_L; q_0, q_1, \ldots, q_{N+1}| \text{ model}\}$
  $= t(m_{N+1}|q_N). \quad t(q_i|q_{i+1}).P(x_{l(i)}|_{qi})$

  where $P(x_{l(i)}|q_i) = 1$ if $q_i$ is the delete state

*In this way a probability distribution on the space of sequences is defined*

PH296 Presentation/ Prasana & Teng



## Estimating parameters in a HMM

- P{sequences| model}

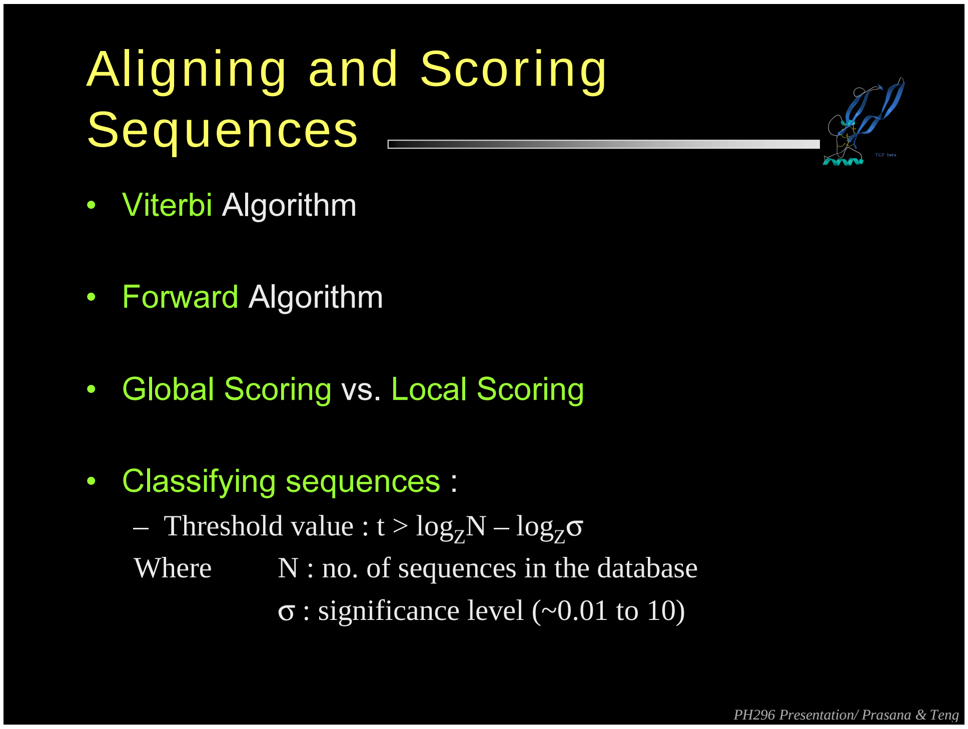
  $= \pi$ P{each sequence| model}

- Maximum Likelihood methods

- Maximum a Posterior approach

*There is no known efficient way to calculate the best HMM*

PH296 Presentation/ Prasana & Teng

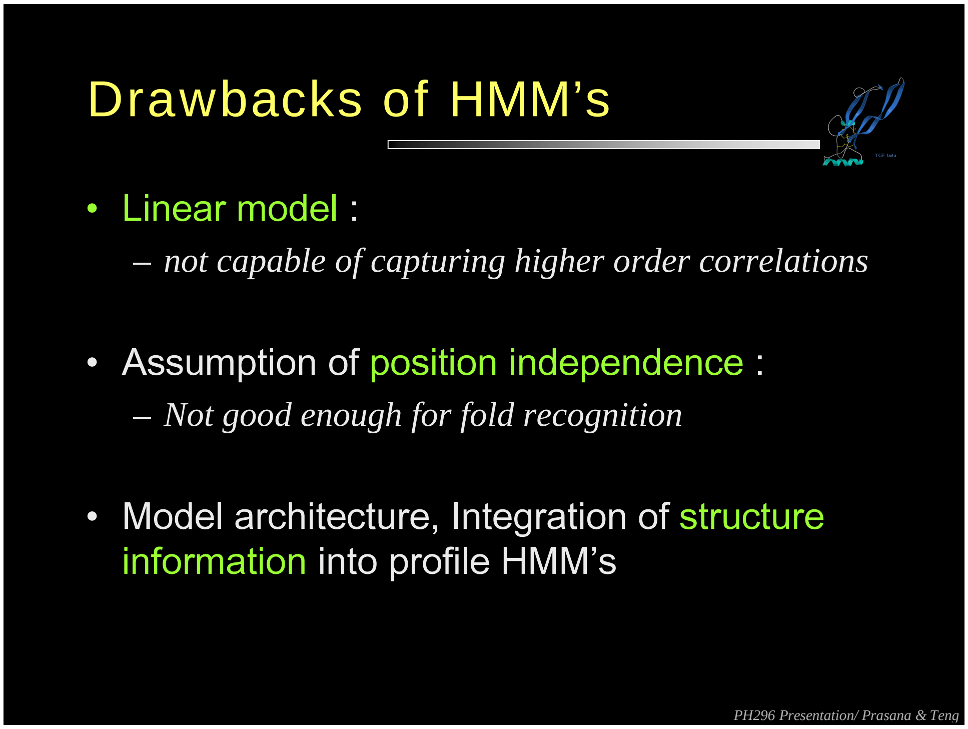


## Aligning and Scoring Sequences

- Viterbi Algorithm
- Forward Algorithm
- Global Scoring vs. Local Scoring
- Classifying sequences :
  - Threshold value : $t > \log_Z N - \log_Z \sigma$

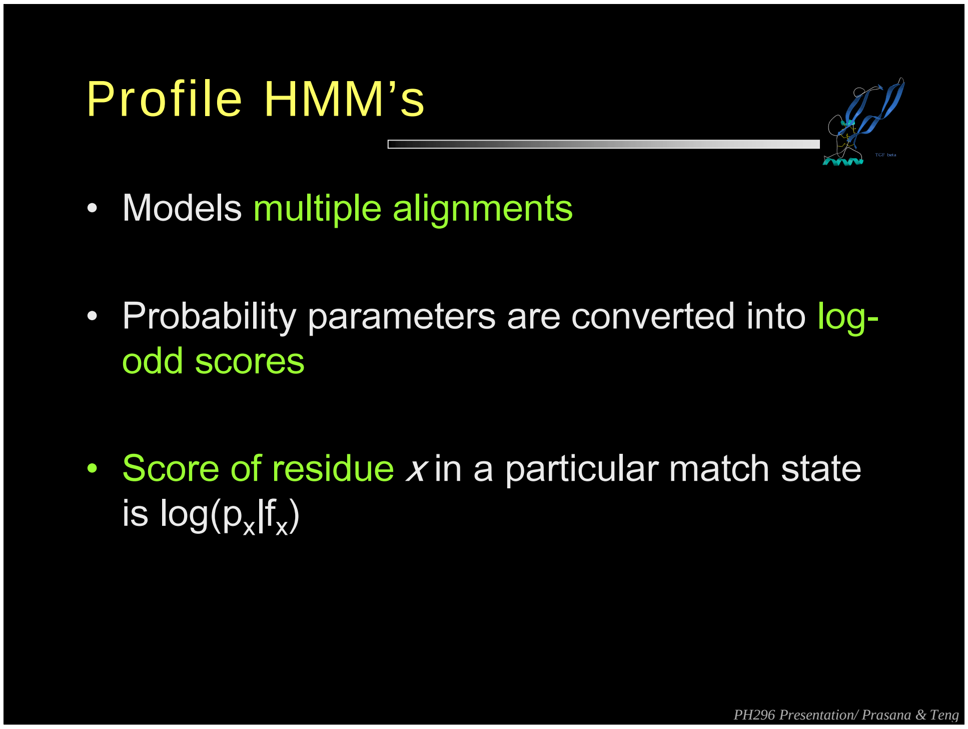
  Where N : no. of sequences in the database

  $\sigma$ : significance level (~0.01 to 10)



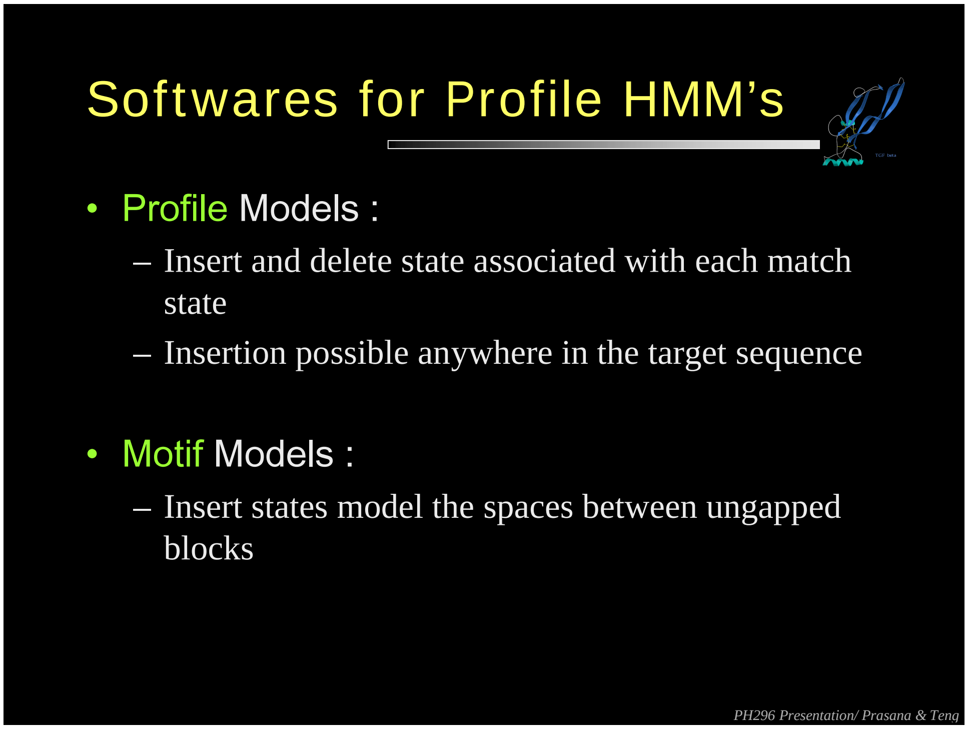
## Drawbacks of HMM’s

- Linear model :
  - *not capable of capturing higher order correlations*
- Assumption of position independence :
  - *Not good enough for fold recognition*
- Model architecture, Integration of structure information into profile HMM’s



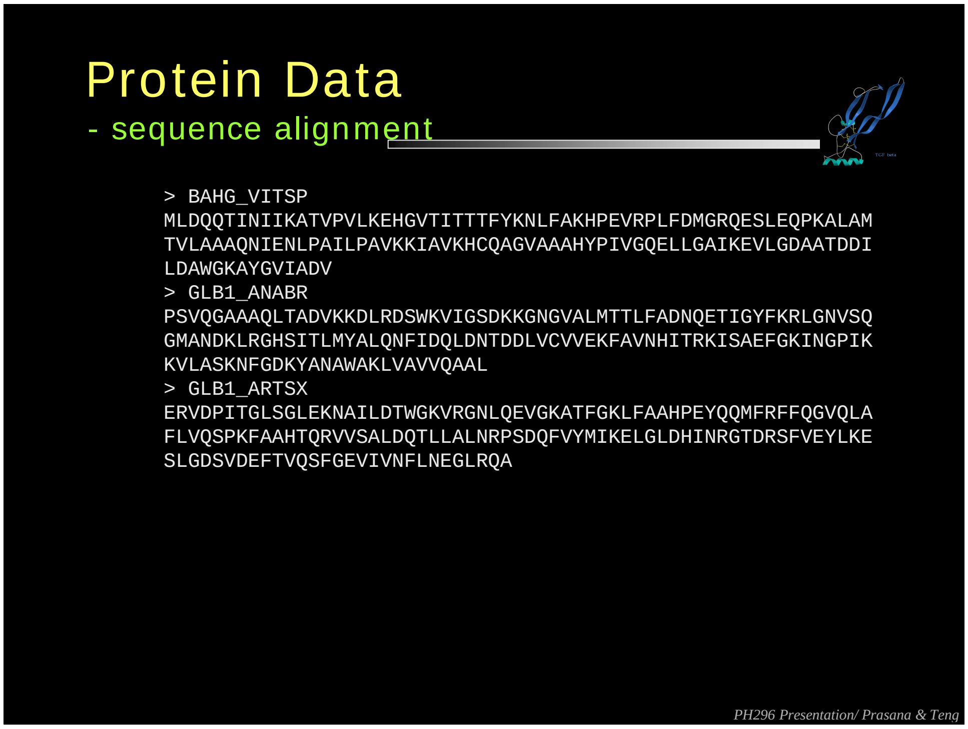
## Profile HMM's

- Models multiple alignments
- Probability parameters are converted into log-odd scores
- Score of residue *x* in a particular match state is log($p_x$|$f_x$)



## Softwares for Profile HMM's

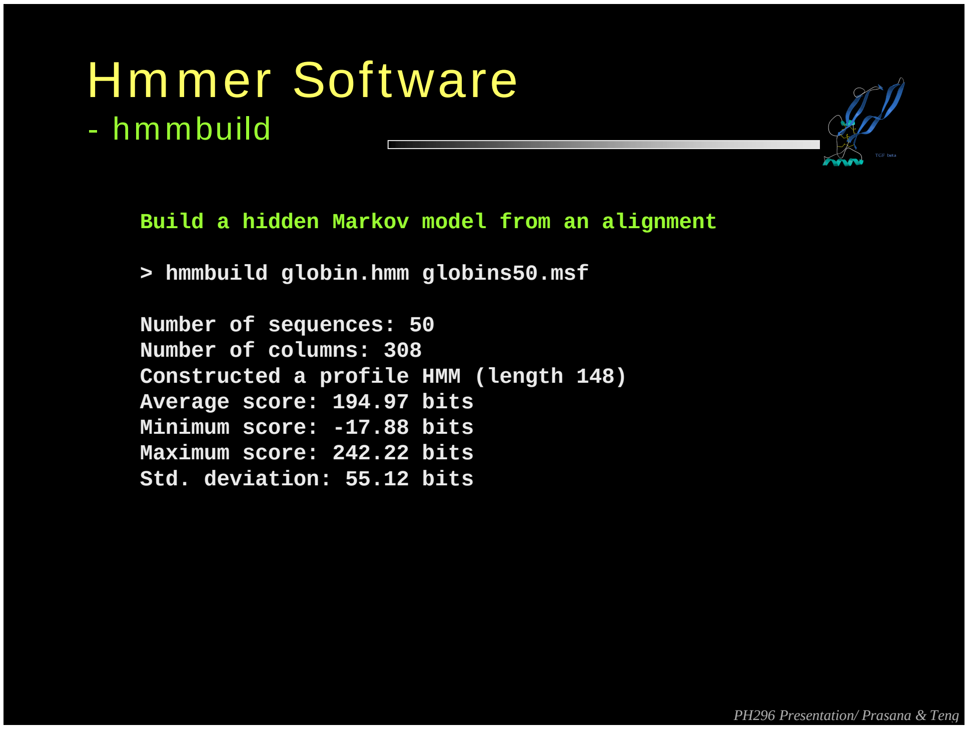
- Profile Models :
  - Insert and delete state associated with each match state
  - Insertion possible anywhere in the target sequence
- Motif Models :
  - Insert states model the spaces between ungapped blocks



# Protein Data

## - sequence alignment

```
> BAHG_VITSP
MLDQQTINIIKATVPVLKEHGVTITTTFYKNLFAKHPEVRPLFDMGRQESLEQPKALAM
TVLAAAQNIENLPAILPAVKKIAVKHCQAGVAAAHYPIVGQELLGAIKEVLGDAATDDI
LDAWGKAYGVIADV
> GLB1_ANABR
PSVQGAAAQLTADVKKDLRDSWKVIGSDKKGNGVALMTTLFADNQETIGYFKRLGNVSQ
GMANDKLRGHSITLMYALQNFIDQLDNTDDLVCVVEKFAVNHITRKISAEFGKINGPIK
KVLASKNFGDKYANAWAKLVAVVQAAL
> GLB1_ARTSX
ERVDPITGLSGLEKNAILDTWGKVRGNLQEVGKATFGKLFAAHPEYQQMFRFFQGVQLA
FLVQSPKFAAHTQRVVSALDQTLLALNRPSDQFVYMIKELGLDHINRGTDRSFVEYLKE
SLGDSVDEFTVQSFGEVIVNFLNEGLRQA
```



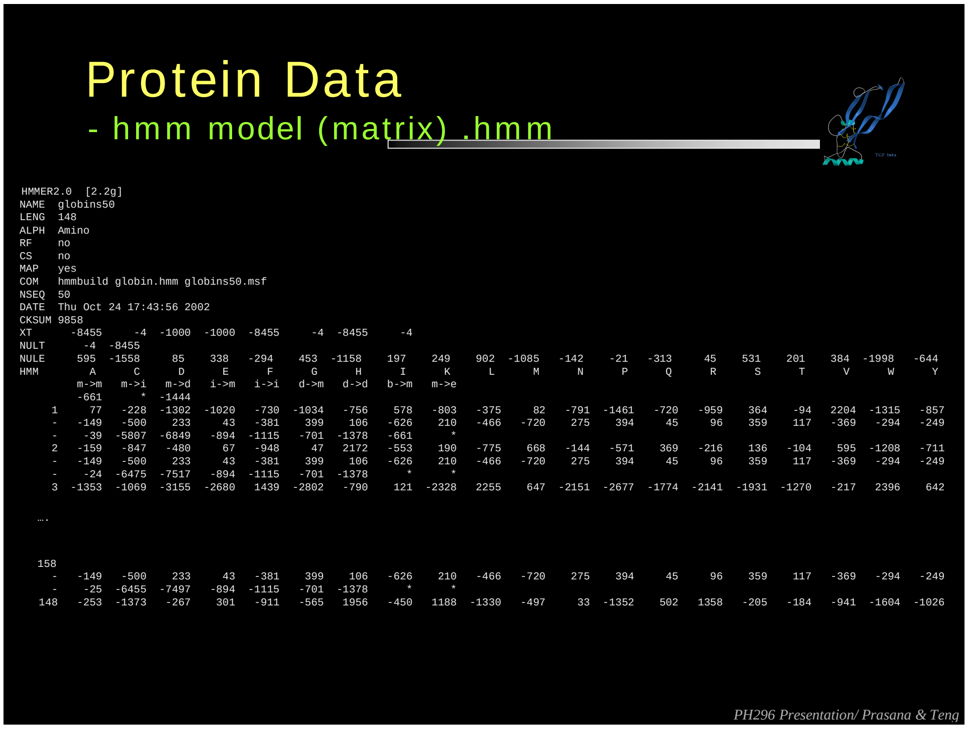
# Hmmer Software

## - hmmbuild

**Build a hidden Markov model from an alignment**

```
> hmmbuild globin.hmm globins50.msf

Number of sequences: 50
Number of columns: 308
Constructed a profile HMM (length 148)
Average score: 194.97 bits
Minimum score: -17.88 bits
Maximum score: 242.22 bits
Std. deviation: 55.12 bits
```



# Protein Data

## - hmm model (matrix) .hmm

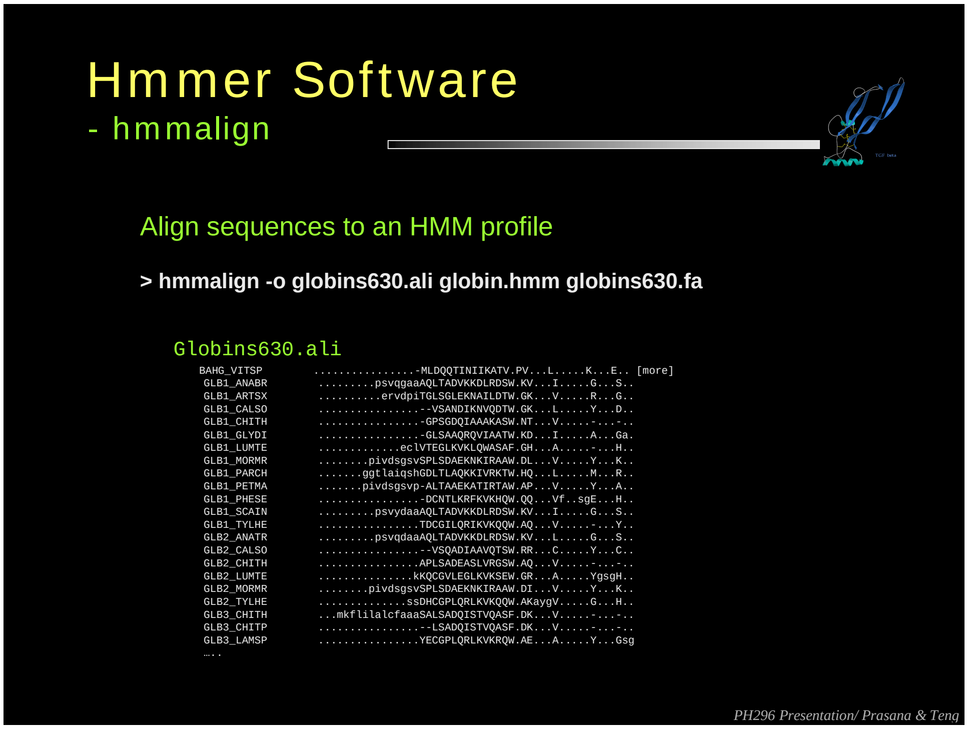
```
HMMER2.0  [2.2g]
NAME  globins50
LENG  148
ALPH  Amino
RF    no
CS    no
MAP   yes
COM   hmmbuild globin.hmm globins50.msf
NSEQ  50
DATE  Thu Oct 24 17:43:56 2002
CKSUM 9858
XT      -8455     -4  -1000  -1000  -8455     -4  -8455     -4
NULT       -4  -8455
NULE      595  -1558     85    338   -294    453  -1158    197    249    902  -1085   -142    -21   -313     45    531    201    384  -1998   -644
HMM         A      C      D      E      F      G      H      I      K      L      M      N      P      Q      R      S      T      V      W      Y
         m->m   m->i   m->d   i->m   i->i   d->m   d->d   b->m   m->e
         -661      *  -1444
     1     77   -228  -1302  -1020   -730  -1034   -756    578   -803   -375     82   -791  -1461   -720   -959    364    -94   2204  -1315   -857
     -   -149   -500    233     43   -381    399    106   -626    210   -466   -720    275    394     45     96    359    117   -369   -294   -249
     -    -39  -5807  -6849   -894  -1115   -701  -1378   -661      *
     2   -159   -847   -480     67   -948     47   2172   -553    190   -775    668   -144   -571    369   -216    136   -104    595  -1208   -711
     -   -149   -500    233     43   -381    399    106   -626    210   -466   -720    275    394     45     96    359    117   -369   -294   -249
     -    -24  -6475  -7517   -894  -1115   -701  -1378      *      *
     3  -1353  -1069  -3155  -2680   1439  -2802   -790    121  -2328   2255    647  -2151  -2677  -1774  -2141  -1931  -1270   -217   2396    642

   ….



   158
     -   -149   -500    233     43   -381    399    106   -626    210   -466   -720    275    394     45     96    359    117   -369   -294   -249
     -    -25  -6455  -7497   -894  -1115   -701  -1378      *      *
   148   -253  -1373   -267    301   -911   -565   1956   -450   1188  -1330   -497     33  -1352    502   1358   -205   -184   -941  -1604  -1026
```

*PH296 Presentation/ Prasana & Teng*



# Hmmer Software

## - hmmalign

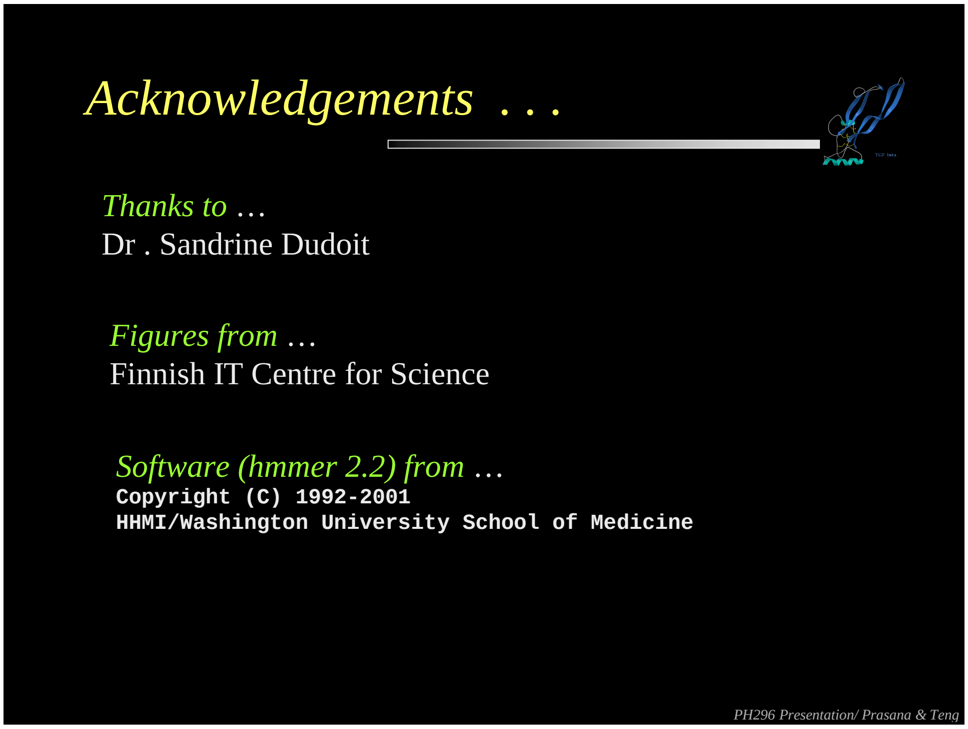
Align sequences to an HMM profile

**> hmmalign -o globins630.ali globin.hmm globins630.fa**

Globins630.ali

```
BAHG_VITSP        ................-MLDQQTINIIKATV.PV...L.....K...E.. [more]
 GLB1_ANABR        .........psvqgaaAQLTADVKKDLRDSW.KV...I.....G...S..
 GLB1_ARTSX        ..........ervdpiTGLSGLEKNAILDTW.GK...V.....R...G..
 GLB1_CALSO        ................--VSANDIKNVQDTW.GK...L.....Y...D..
 GLB1_CHITH        ................-GPSGDQIAAAKASW.NT...V.....-...-..
 GLB1_GLYDI        ................-GLSAAQRQVIAATW.KD...I.....A...Ga.
 GLB1_LUMTE        ............eclVTEGLKVKLQWASAF.GH...A.....-...H..
 GLB1_MORMR        ........pivdsgsvSPLSDAEKNKIRAAW.DL...V.....Y...K..
 GLB1_PARCH        .......ggtlaiqshGDLTLAQKKIVRKTW.HQ...L.....M...R..
 GLB1_PETMA        .......pivdsgsvp-ALTAAEKATIRTAW.AP...V.....Y...A..
 GLB1_PHESE        ................-DCNTLKRFKVKHQW.QQ...Vf..sgE...H..
 GLB1_SCAIN        .........psvydaaAQLTADVKKDLRDSW.KV...I.....G...S..
 GLB1_TYLHE        ................TDCGILQRIKVKQQW.AQ...V.....-...Y..
 GLB2_ANATR        .........psvqdaaAQLTADVKKDLRDSW.KV...L.....G...S..
 GLB2_CALSO        ................--VSQADIAAVQTSW.RR...C.....Y...C..
 GLB2_CHITH        ................APLSADEASLVRGSW.AQ...V.....-...-..
 GLB2_LUMTE        ..............kKQCGVLEGLKVKSEW.GR...A.....YgsgH..
 GLB2_MORMR        ........pivdsgsvSPLSDAEKNKIRAAW.DI...V.....Y...K..
 GLB2_TYLHE        .............ssDHCGPLQRLKVKQQW.AKaygV.....G...H..
 GLB3_CHITH        ...mkflilalcfaaaSALSADQISTVQASF.DK...V.....-...-..
 GLB3_CHITP        ................--LSADQISTVQASF.DK...V.....-...-..
 GLB3_LAMSP        ................YECGPLQRLKVKRQW.AE...A.....Y...Gsg
 …..
```

*PH296 Presentation/ Prasana & Teng*



# *Acknowledgements . . .*

*Thanks to* …
Dr . Sandrine Dudoit

*Figures from* …
Finnish IT Centre for Science

*Software (hmmer 2.2) from* …

```
Copyright (C) 1992-2001
HHMI/Washington University School of Medicine
```