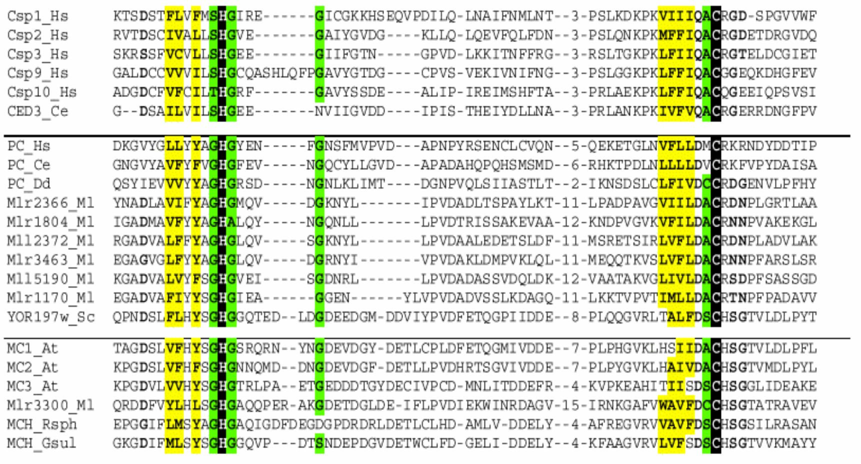
Introduction to Bioinformatics
BINF 630

Dr. D. Andrew Carr

Multiple Sequence Alignments

Multiple Sequence Alignment



```
Csp1_Hs      KTSDSTFLVFMSHGIRE------GICGKKHSEQVPDILQ-LNAIFNMLNT--3-PSLKDKPKVIIIQACRGD-SPGVVWF
Csp2_Hs      RVTDSCIVALLSHGVE-------GAIYGVDG----KLLQ-LQEVFQLFDN--3-PSLQNKPKMFFIQACRGDETDRGVDQ
Csp3_Hs      SKRSSFVCVLLSHGEE-------GIIFGTN-----GPVD-LKKITNFFRG--3-RSLTGKPKLFIIQACRGTELDCGIET
Csp9_Hs      GALDCCVVVILSHGCQASHLQFPGAVYGTDG----CPVS-VEKIVNIFNG--3-PSLGGKPKLFFIQACGGEQKDHGFEV
Csp10_Hs     ADGDCFVFCILTHGRF-------GAVYSSDE----ALIP-IREIMSHFTA--3-PRLAEKPKLFFIQACQGEEIQPSVSI
CED3_Ce      G--DSAILVILSHGEE-------NVIIGVDD----IPIS-THEIYDLLNA--3-PRLANKPKIVFVQACRGERRDNGFPV

PC_Hs        DKGVYGLLYYAGHGYEN-----FGNSFMVPVD---APNPYRSENCLCVQN--5-QEKETGLNVFLLDMCRKRNDYDDTIP
PC_Ce        GNGVYAVFYFVGHGFEV-----NGQCYLLGVD---APADAHQPQHSMSMD--6-RHKTPDLNLLLLDVCRKFVPYDAISA
PC_Dd        QSYIEVVVYYAGHGRSD-----NGNLKLIMT----DGNPVQLSIIASTLT--2-IKNSDSLCLFIVDCCRDGENVLPFHY
Mlr2366_Ml   YNADLAVIFYAGHGMQV-----DGKNYL-------IPVDADLTSPAYLKT-11-LPADPAVGVIILDACRDNPLGRTLAA
Mlr1804_Ml   IGADMAVFYYAGHALQY-----NGQNLL-------LPVDTRISSAKEVAA-12-KNDPVGVKVFILDACRNNPVAKEKGL
Mll2372_Ml   RGADVALFFYAGHGLQV-----SGKNYL-------LPVDAALEDETSLDF-11-MSRETSIRLVFLDACRDNPLADVLAK
Mlr3463_Ml   EGAGVGLFYYAGHGLQV-----DGRNYI-------VPVDAKLDMPVKLQL-11-MEQQTKVSLVFLDACRNNPFARSLSR
Mll5190_Ml   KGADVALVYFSGHGVEI-----SGDNRL-------LPVDADASSVDQLDK-12-VAATAKVGLIVLDACRSDPFSASSGD
Mlr1170_Ml   EGADVAFIYYSGHGIEA------GGEN------YLVPVDADVSSLKDAGQ-11-LKKTVPVTIMLLDACRTNPFPADAVV
YOR197w_Sc   QPNDSLFLHYSGHGGQTED--LDGDEEDGM-DDVIYPVDFETQGPIIDDE--8-PLQQGVRLTALFDSCHSGTVLDLPYT

MC1_At       TAGDSLVFHYSGHGSRQRN--YNGDEVDGY-DETLCPLDFETQGMIVDDE--7-PLPHGVKLHSIIDACHSGTVLDLPFL
MC2_At       KPGDSLVFHFSGHGNNQMD--DNGDEVDGF-DETLLPVDHRTSGVIVDDE--7-PLPYGVKLHAIVDACHSGTVMDLPYL
MC3_At       KPGDVLVVHYSGHGTRLPA--ETGEDDDTGYDECIVPCD-MNLITDDEFR--4-KVPKEAHITIISDSCHSGGLIDEAKE
Mlr3300_Ml   QRDDFVYLHLSGHGAQQPER-AKGDETDGLDE-IFLPVDIEKWINRDAGV-15-IRNKGAFVWAVFDCCHSGTATRAVEV
MCH_Rsph     EPGGIFLMSYAGHGAQIGDFDEGDGPDRDRLDETLCLHD-AMLV-DDELY--4-AFREGVRVVAVFDSCHSGSILRASAN
MCH_Gsul     GKGDIFMLSYSGHGGQVP---DTSNDEPDGVDETWCLFD-GELI-DDELY--4-KFAAGVRVLVFSDSCHSGTVVKMAYY
```

Figure: Conserved catalytic motifs in the caspase-like superfamily of proteases.

Knowledge gain

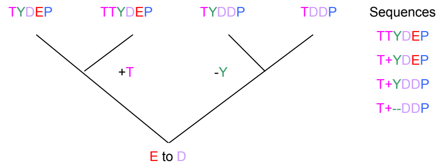
- *In science, “knowledge” is in the patterns/similarities. The interesting questions, however, are in the differences.*

Multiple Sequence Alignment

- It is believed, based on finding similar protein sequences within highly divergent species, that the over time the functional components embedded within the sequences are conserved in order to retain the function.
  - One of the most important elements of sequences is the phylogenetic information that similarities represent.
  - The sequence similarities gives insight into the evolution of families of protein or DNA sequences.
    - Knowledge to be gained:
      - Estimation of evolutionary distance
      - Mutation / Speciation
      - *Note: Multiple Sequence Alignment methods are also employed in assembly of DNA sequences from cloning vectors.*

Multiple Sequence Alignment

- Evolutionary distance:



Sequences

TTYDEP

T+YDEP

T+YDDP

T+--DDP

- Phylogenetic distance is a measure of divergence between two similar sequences.
  - It can be thought of as the number of changes/ substitutions that have occurred, or the number of differences.
    - The simplest estimation of distance is to count the number of base mismatches m between the two sequences when aligned, then present this value as a proportion, or percentage, of the total alignment length n.

      D = m/n (1)

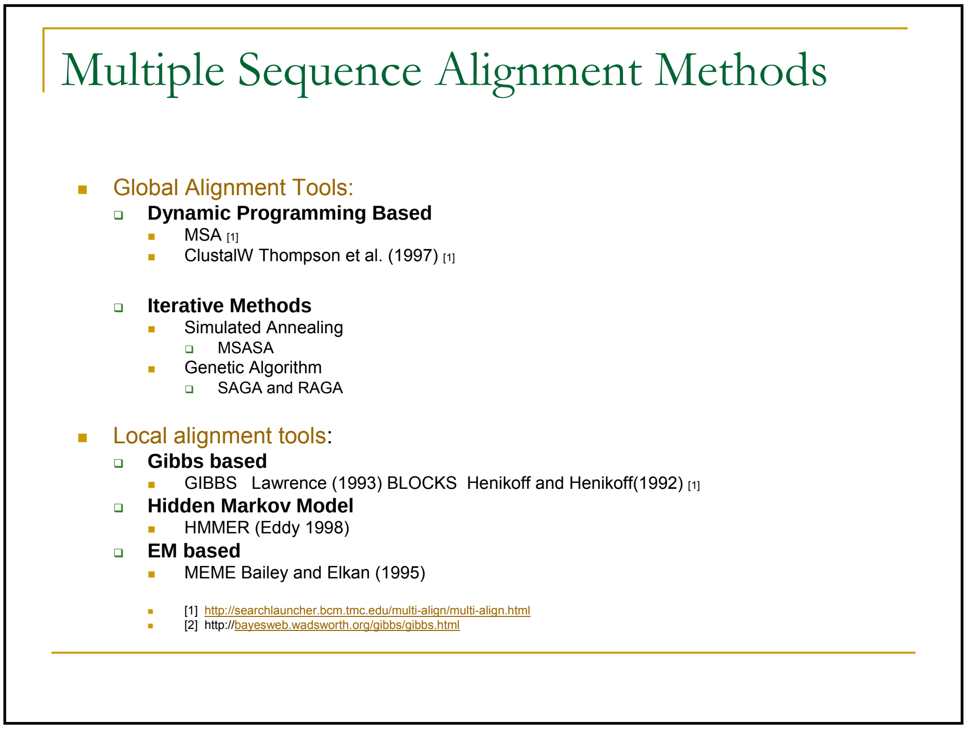
    - Including gaps and indels...

      D = m/([n - g] + [g * penalty]) (2)

      Where g is the total number of gaps within the alignment's consensus sequence.

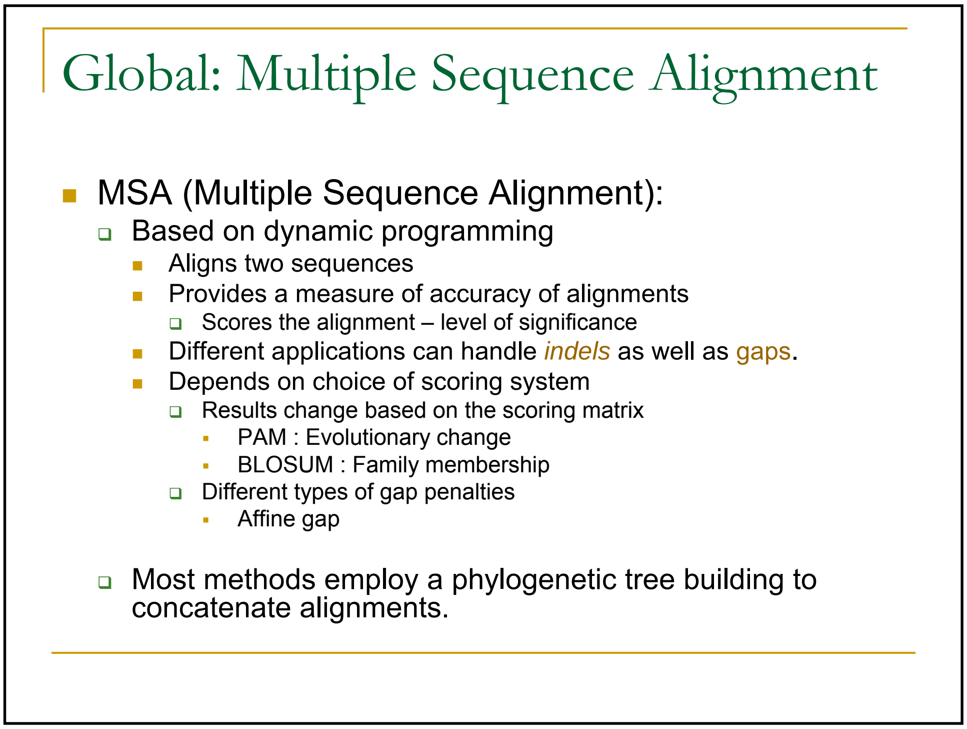
Multiple Sequence Alignment Methods

- Global alignment:
  - Homologous proteins
    - Structural similarity → Functional similarity
    - Common evolutionary origin
- Local alignment:
  - Conserved regions
    - Structural motif → Functional domains
      - Phylogenetic or ancestral similarity



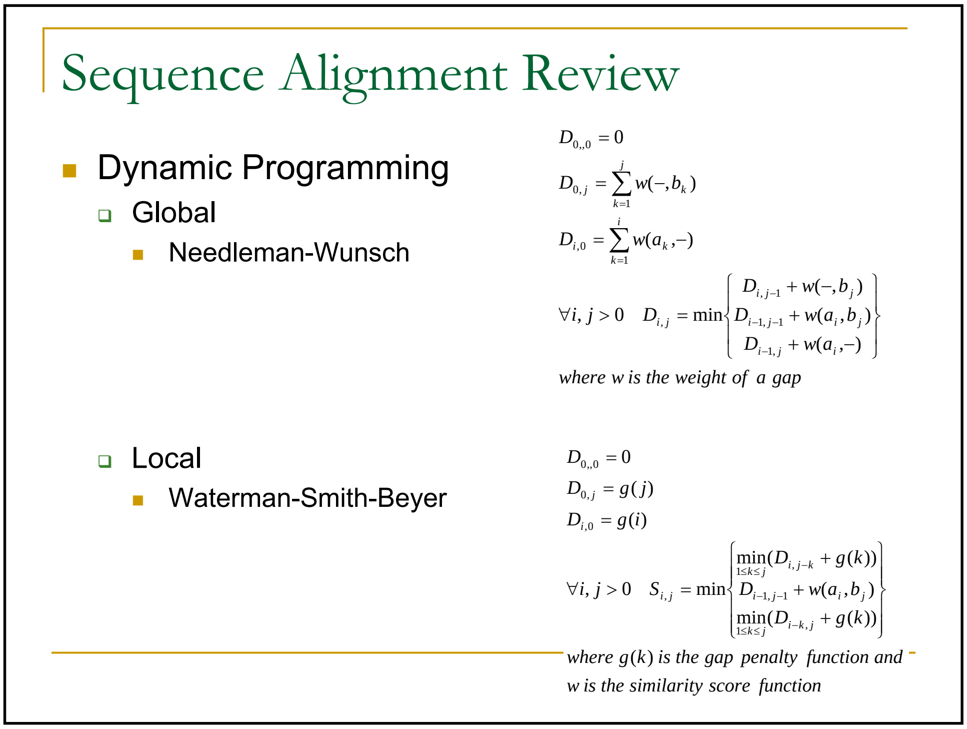
Multiple Sequence Alignment Methods

- Global Alignment Tools:
  - **Dynamic Programming Based**
    - MSA [1]
    - ClustalW Thompson et al. (1997) [1]
  - **Iterative Methods**
    - Simulated Annealing
      - MSASA
    - Genetic Algorithm
      - SAGA and RAGA
- Local alignment tools:
  - **Gibbs based**
    - GIBBS Lawrence (1993) BLOCKS Henikoff and Henikoff(1992) [1]
  - **Hidden Markov Model**
    - HMMER (Eddy 1998)
  - **EM based**
    - MEME Bailey and Elkan (1995)
    - [1] http://searchlauncher.bcm.tmc.edu/multi-align/multi-align.html
    - [2] http://bayesweb.wadsworth.org/gibbs/gibbs.html



Global: Multiple Sequence Alignment

- MSA (Multiple Sequence Alignment):
  - Based on dynamic programming
    - Aligns two sequences
    - Provides a measure of accuracy of alignments
      - Scores the alignment – level of significance
    - Different applications can handle *indels* as well as gaps.
    - Depends on choice of scoring system
      - Results change based on the scoring matrix
        - PAM : Evolutionary change
        - BLOSUM : Family membership
      - Different types of gap penalties
        - Affine gap
  - Most methods employ a phylogenetic tree building to concatenate alignments.



Sequence Alignment Review

- Dynamic Programming
  - Global
    - Needleman-Wunsch

$$D_{0,,0} = 0$$

$$D_{0,j} = \sum_{k=1}^{j} w(-,b_k)$$

$$D_{i,0} = \sum_{k=1}^{i} w(a_k,-)$$

$$\forall i,j > 0 \quad D_{i,j} = \min\left\{\begin{matrix} D_{i,j-1} + w(-,b_j) \\ D_{i-1,j-1} + w(a_i,b_j) \\ D_{i-1,j} + w(a_i,-) \end{matrix}\right\}$$

where w is the weight of a gap

  - Local
    - Waterman-Smith-Beyer

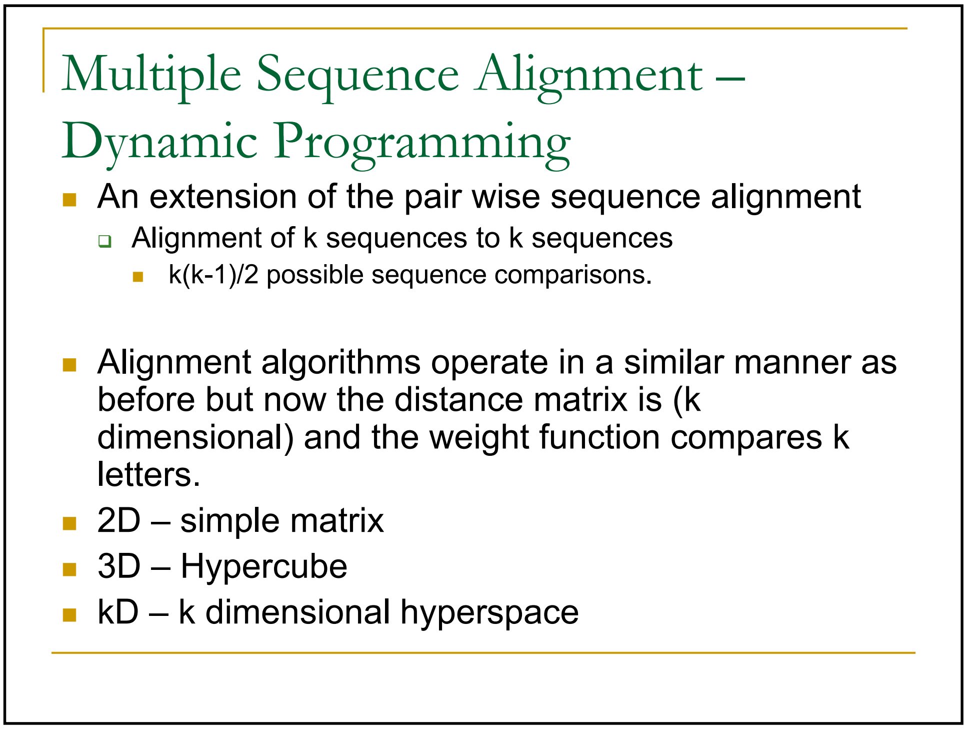
$$D_{0,,0} = 0$$

$$D_{0,j} = g(j)$$

$$D_{i,0} = g(i)$$

$$\forall i,j > 0 \quad S_{i,j} = \min\left\{\begin{matrix} \min\limits_{1\le k\le j}(D_{i,j-k} + g(k)) \\ D_{i-1,j-1} + w(a_i,b_j) \\ \min\limits_{1\le k\le j}(D_{i-k,j} + g(k)) \end{matrix}\right\}$$

where $g(k)$ is the gap penalty function and
w is the similarity score function



Multiple Sequence Alignment – Dynamic Programming

- An extension of the pair wise sequence alignment
  - Alignment of k sequences to k sequences
    - k(k-1)/2 possible sequence comparisons.

- Alignment algorithms operate in a similar manner as before but now the distance matrix is (k dimensional) and the weight function compares k letters.
- 2D – simple matrix
- 3D – Hypercube
- kD – k dimensional hyperspace

MSA: Dynamic Programming

- Assume that we are trying to align three sequence a,b, and c.
- Also assume that we have a cost function w(x,y,z) that computes the cost of comparing x, y, and z in sequences a, b, and c respectively.

$$w(x,y,z)=\begin{cases}0 & x=y=z\\ 1 & 2\,of\,3\,symbols\,are\,the\,same\\ 2 & x\neq y\neq z\end{cases}$$

MSA: Dynamic Programming

Then our distance matrix D can be described by:

$$\forall i,j,k>0\quad D_{i,j,k}=\min\begin{Bmatrix}D_{i-1,j-1,k-1}+w(a_i,b_j,c_k)\\ D_{i,j-1,k-1}+w(-,b_j,c_k)\\ D_{i-1,j,k-1}+w(a_i,-,c_k)\\ D_{i-1,j-1,k}+w(a_i,b_j,-)\\ D_{i-1,j,k}+w(a_i,-,-)\\ D_{i,j-1,k}+w(-,b_j,-)\\ D_{i,j,k-1}+w(-,-,c_k)\end{Bmatrix}$$

MSA: Dynamic Programming

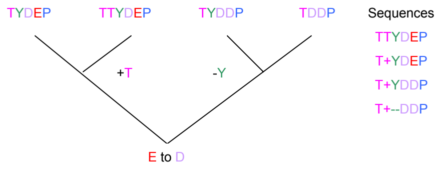
- Computationally expensive
  - Long sequences
  - Large number of sequences
- Computational cost
  - For two sequences of lengths n and m
    - Needleman-Wunsch
      - O(nm) = O(n^2) for n >= m
    - Waterman-Smith-Beyer
      - O(nm(n+m) = O(n^3) for n >= m
    - Gotoh's algorithm
      - O(nm) = O(n^2) for n >= m
  - For three sequences of lengths n, m and p
    - Most algorithms
      - O(nmp) = O(n^3) for n >= m >= p

MSA: Dynamic Programming

- For k sequences we get O(n^k) where n is the longest sequence
  - If we compare 10 sequences of length <= 300
    - 300^{10} comparisons….
      - 5,904,900,000,000,000,000,000,000
    - np-complete task

Global Alignment: Dynamic Programming

- Solutions:
  - Sum of Pairs *Carrillo and Lipman* (1988)
    - Look at all pairs
    - Create parsimony tree
    - Build alignment along the tree
    - Take the best SP score
    - Problems:
      - Biased toward similar sets of sequences.
      - Order weighting important



Sequences
TTYDEP
T+YDEP
T+YDDP
T+--DDP

Global Alignment: Dynamic Programming

- Progressive Methods
  - Dynamic Sequence alignment based on hierarchical sequence similarity.
  - Challenge:
    - To choose the correct
      - Sequence weighting
      - Scoring matrix
      - Gap penalties
  - Goal get the correct series of evolutionary changes
  - Programs:
    - CLUSTALW
      - Gaps between conserved regions down weighted
      - Already occurring gaps down weighted
      - New Gaps are costly
    - T-Coffee
      - Employs locally aligned (conserved regions/domains) to help find alignment

Global MSA: Genetic Algorithms

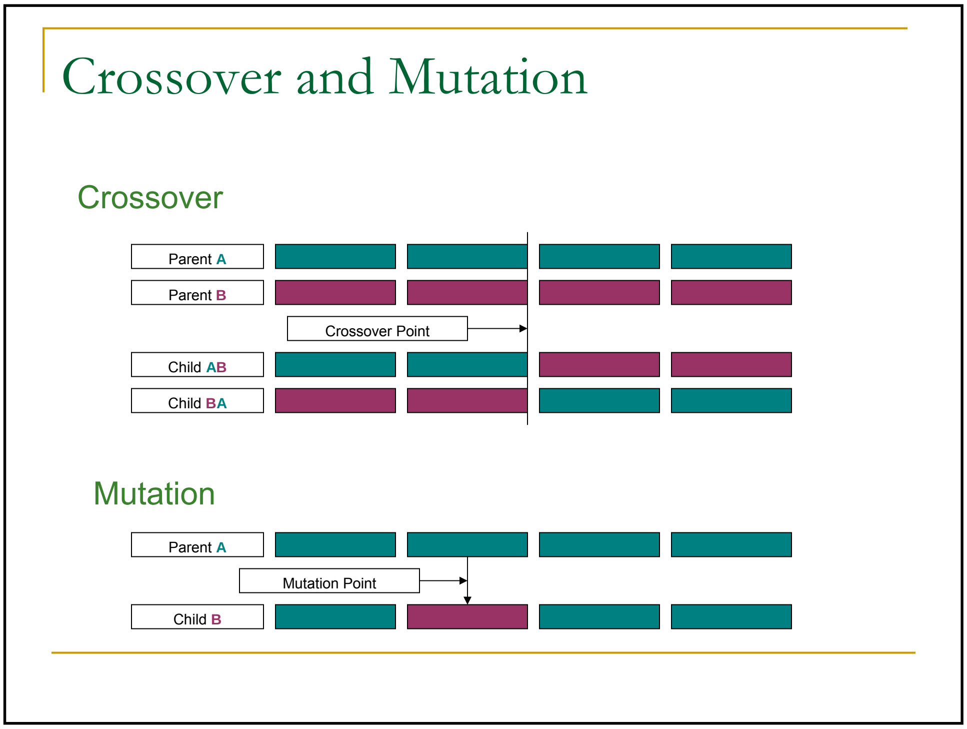
- SAGA Notredame and Higgins (1996)
  - Basic steps
    - 1) Pair wise global alignment
    - 2) Phylogentic tree
    - 3) Weight sequences
    - 4) Iterative refinement of MSA via GA
      - Check score and if not good enough return to 2

Global MSA: Genetic Algorithms

- GA developed by computers scientists as a machine learning application.

- Typical GA flow:
  - Initialize the study population
  - Evaluate the population
    - while (not termination condition)
      - Alter the population at this step
        - Reproduction – combination of members
        - Mutation – random changes in the population
        - Crossover – moving information between members
      - Re-evaluate the population

- Departmental expert
  - Dr. John Grefenstette



Crossover and Mutation

Crossover

Parent A

Parent B

Crossover Point

Child AB

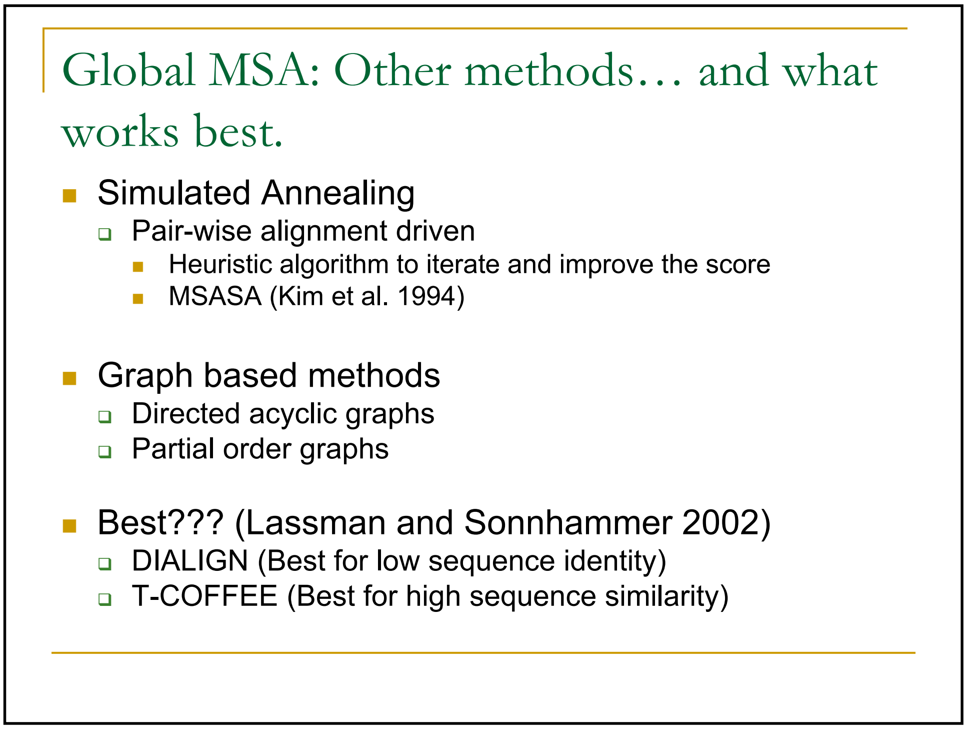
Child BA

Mutation

Parent A

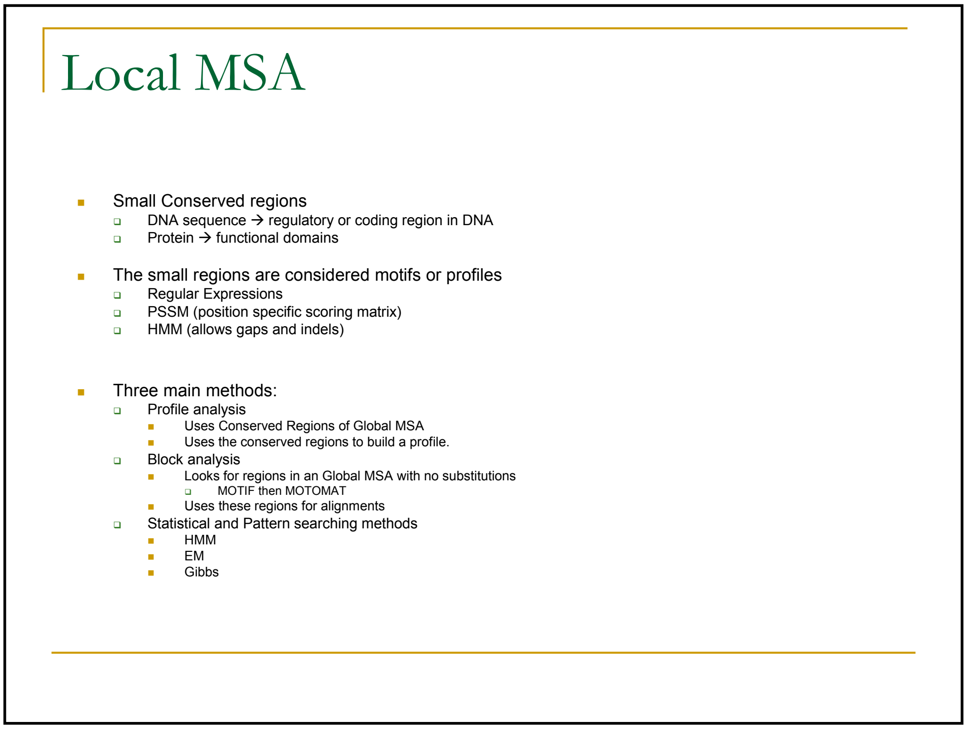
Mutation Point

Child B



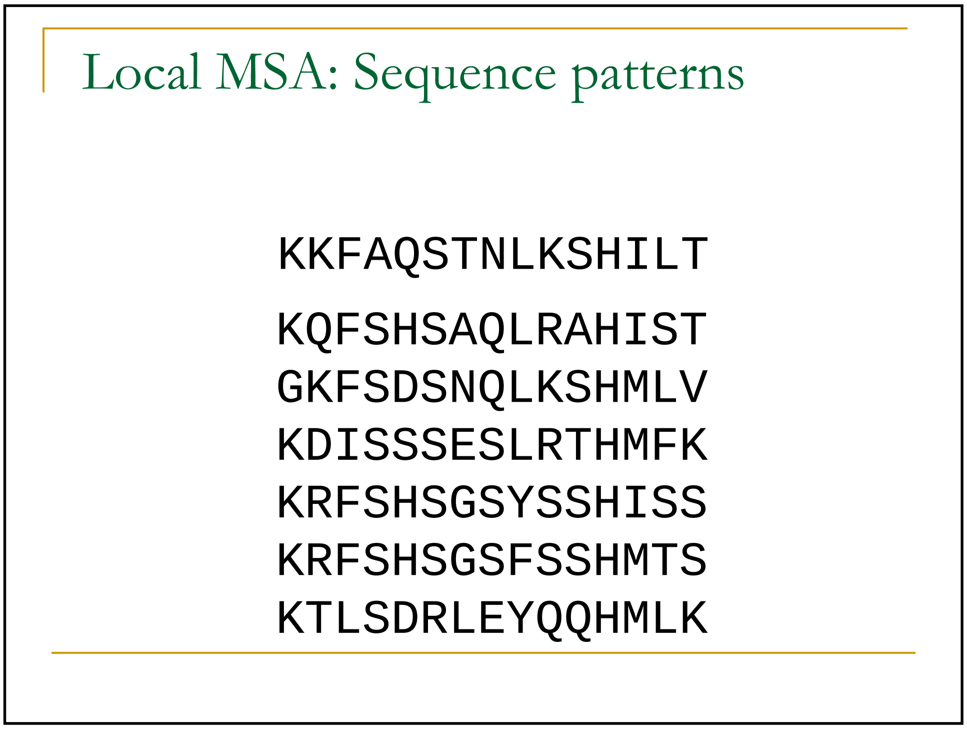
Global MSA: Other methods… and what works best.

- Simulated Annealing
  - Pair-wise alignment driven
    - Heuristic algorithm to iterate and improve the score
    - MSASA (Kim et al. 1994)
- Graph based methods
  - Directed acyclic graphs
  - Partial order graphs
- Best??? (Lassman and Sonnhammer 2002)
  - DIALIGN (Best for low sequence identity)
  - T-COFFEE (Best for high sequence similarity)



Local MSA

- Small Conserved regions
  - DNA sequence → regulatory or coding region in DNA
  - Protein → functional domains
- The small regions are considered motifs or profiles
  - Regular Expressions
  - PSSM (position specific scoring matrix)
  - HMM (allows gaps and indels)
- Three main methods:
  - Profile analysis
    - Uses Conserved Regions of Global MSA
    - Uses the conserved regions to build a profile.
  - Block analysis
    - Looks for regions in an Global MSA with no substitutions
      - MOTIF then MOTOMAT
    - Uses these regions for alignments
  - Statistical and Pattern searching methods
    - HMM
    - EM
    - Gibbs



Local MSA: Sequence patterns

```
KKFAQSTNLKSHILT

KQFSHSAQLRAHIST
GKFSDSNQLKSHMLV
KDISSSESLRTHMFK
KRFSHSGSYSSHISS
KRFSHSGSFSSHMTS
KTLSDRLEYQQHMLK
```

Local MSA: Protein Motif Databases

- PROSITE http://www.expasy.org/prosite/
  - Method:
    - CLUSTAL MSA based
- BLOCKS http://blocks.fhcrc.org/
  - Method:
    - MSA based
- PFAM http://www.sanger.ac.uk/Software/Pfam/
  - Method:
    - MSA and HMM models of protein motifs

Local MSA: PROSITE

- PROSITE
  - Current version contains
    - 1446 documentation entries that describe
      - 1331 patterns
      - 4 rules
      - 650 profiles/matrices
- Cytochrome P450 cysteine heme-iron ligand signature
  - [FW] - [SGNH] - x - [GD] - {F} - [RKHPT] - {P} - C - [LIVMFAP] - [GAD]
    - ***C is the heme iron ligand***

Regular Expressions

Patterns described in a standard way are known as *regular expressions*

x	ANY		
[ ]	OR	[ILV]	I or L or V
{ }	NOT	{DE}	not D or E
( )	repetitions	x(2,3)	x-x or x-x-x
-	separator		
<	N-terminal		
>	C-terminal		
.	END		

Regular Expressions

[AC]-x-V-x(4)-{ED}

[Ala or Cys]-any-Val-any-any-any-any-{any but Glu or Asp}

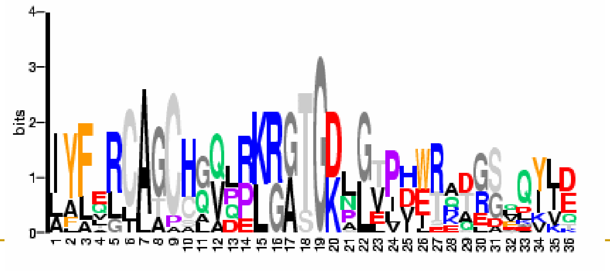
```
...LKHVAYVFQALIYWIK...
...AVEMAGVKYLQVQHGS...
...LYTGAIVTNNDGPYMA...
...KEYKCKVEKELTDICN...
```

BLOCKS

- Blocks are multiply aligned un-gapped segments corresponding to the most highly conserved regions of proteins

- **Block IPB003143A**
- ID D1_heme; BLOCK AC IPB003143A; distance from previous block=(7,88)

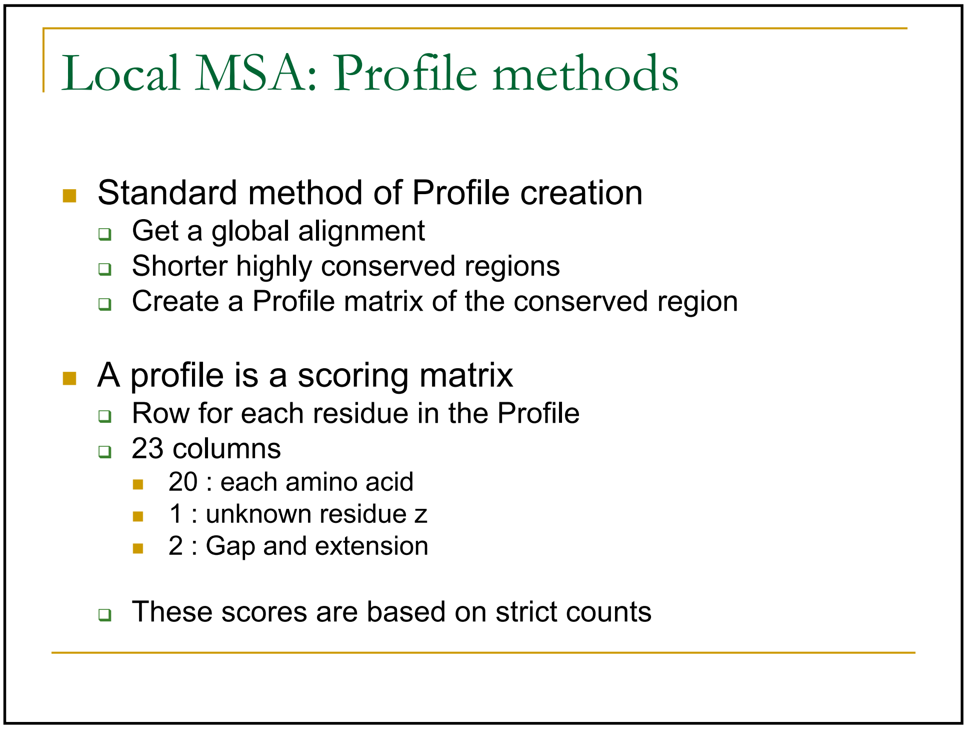
```
DE Cytochrome d1, heme region BL ACG; width=36; seqs=12; 99.5%=1578; strength=1225
NIRF_PSEAE|Q51480  (  9) LLLTLLAGCSQQPPLRGSGDLGVLIERADGSVQILD 59
NIRS_PARDE|Q51700 (  89) IYFERCAGCHGVLRKGATGKALTPDLTRDLGFDYLQ 32
NIRS_PSEAE|P24474 (  67) IYFQRCAGCHGVLRKGATGKPLTPDITQQRGQQYLE 29
NIRF_PSEST|Q52521  (  8) LAAVGLLTACAQQPLRGTGDLGVVVERATGSLQIIE 88
NIRS_PARPN|P72181 (  89) IYFERCAGCHGVLRKGATGKALTPDLTRDLGFDYLQ 32
NIRS_PSEST|P24040 (  42) IYFERCAGCHGVLRKGATGKNLEPHWSKTEADGKKT 51
Q9R9J9             (  8) AALLLIAAPALADELRGTGDLGLIVEREAGSLLVVD 100
Q44012            (  53) IYFERCAGCHGVLRKGATGKSLTPDITRARGTEYLK 36
Q9F0W9            (  58) IYFQRCAGCHGVLRKGATGKPLTPDITQSRGQAYLE 32
```



PSSM of IPB003143A (D1_heme;) 12 sequences.

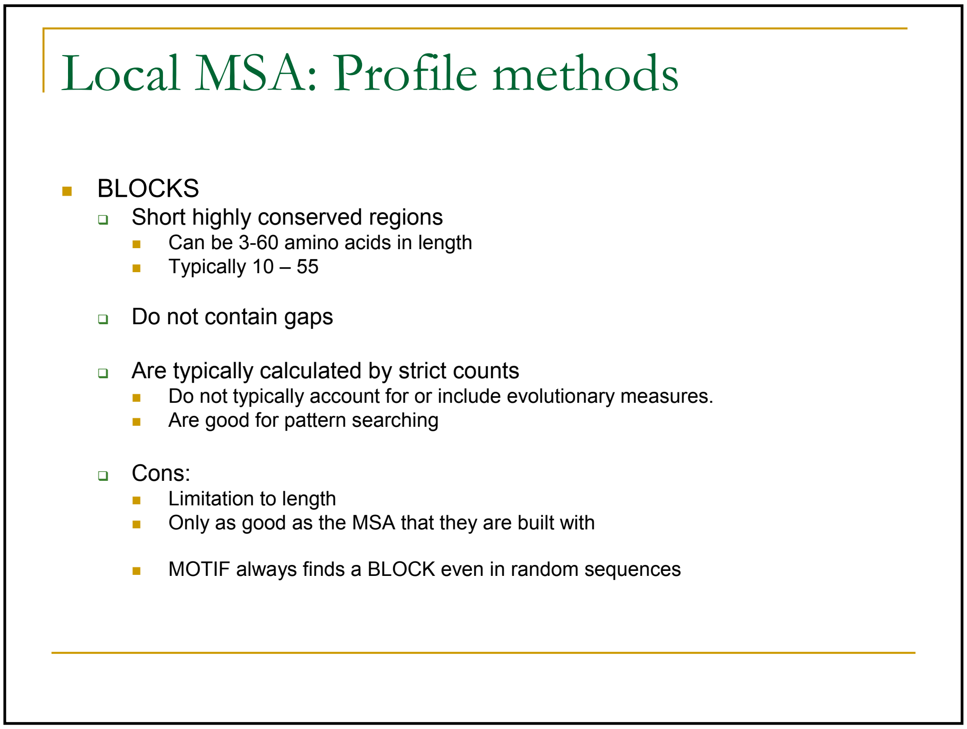
Profile Method of Local MSA

- Two types of profile method
  - Weighted
    - Strictly drawn from Dayhoff PAM matrices based on amino acid frequency
  - Evolutionary
    - Rate of evolution is taken into account
    - Probability of the change between residues
      - Calculated as a log odds score of the Dayhoff method
  - The method for computation of the



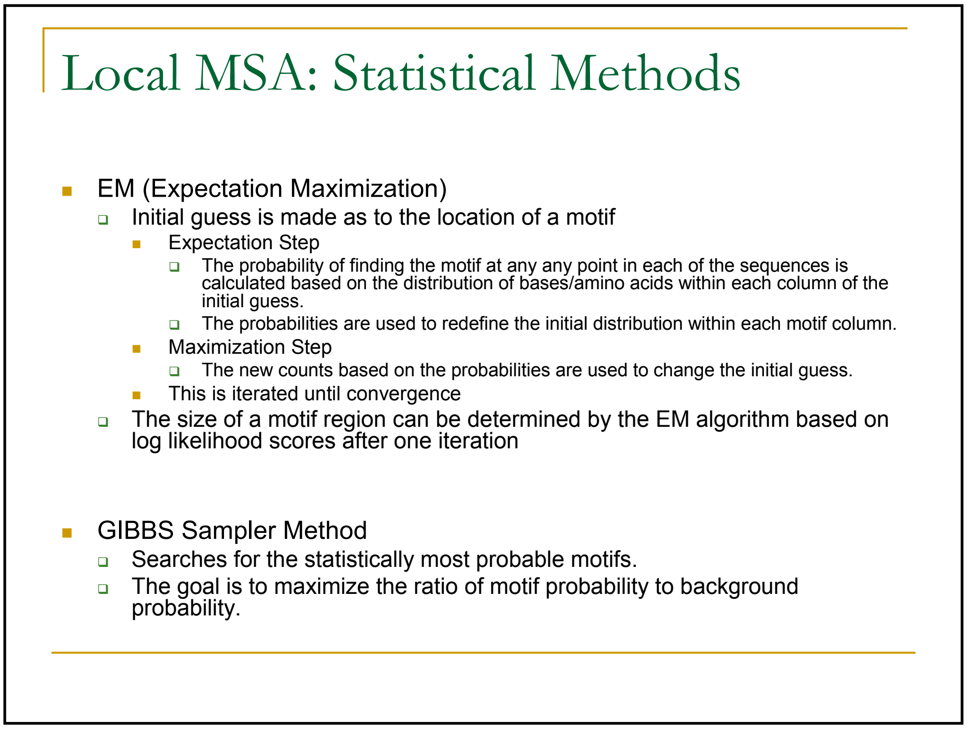
Local MSA: Profile methods

- Standard method of Profile creation
  - Get a global alignment
  - Shorter highly conserved regions
  - Create a Profile matrix of the conserved region
- A profile is a scoring matrix
  - Row for each residue in the Profile
  - 23 columns
    - 20 : each amino acid
    - 1 : unknown residue z
    - 2 : Gap and extension
  - These scores are based on strict counts



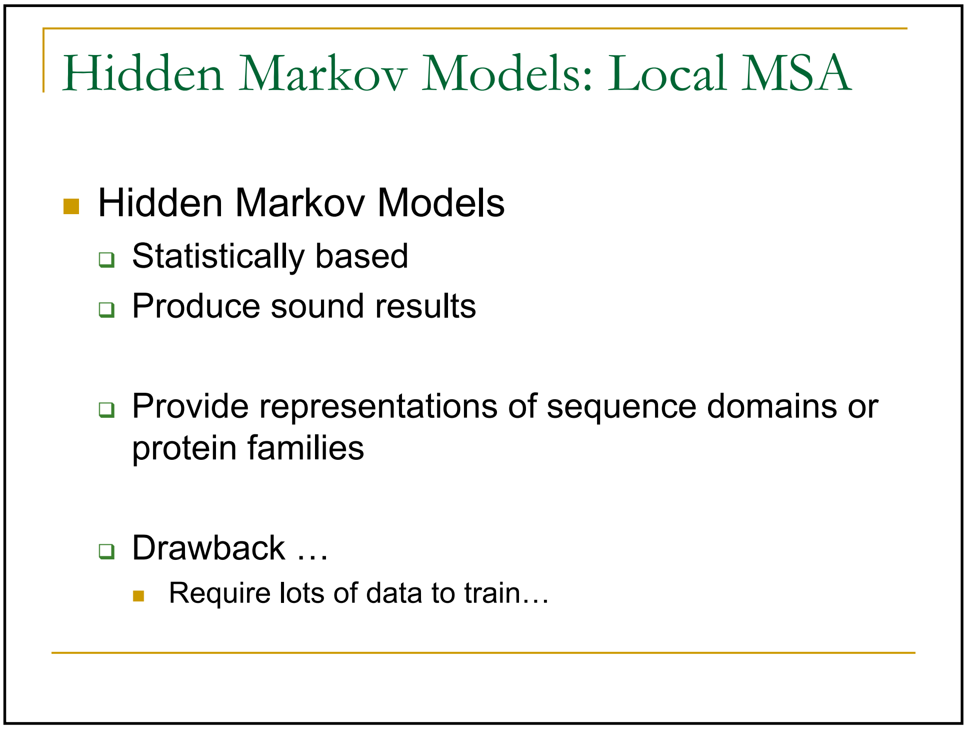
Local MSA: Profile methods

- BLOCKS
  - Short highly conserved regions
    - Can be 3-60 amino acids in length
    - Typically 10 – 55
  - Do not contain gaps
  - Are typically calculated by strict counts
    - Do not typically account for or include evolutionary measures.
    - Are good for pattern searching
  - Cons:
    - Limitation to length
    - Only as good as the MSA that they are built with
    - MOTIF always finds a BLOCK even in random sequences



Local MSA: Statistical Methods

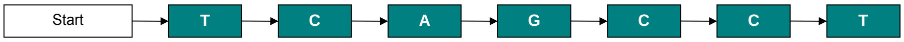
- EM (Expectation Maximization)
  - Initial guess is made as to the location of a motif
    - Expectation Step
      - The probability of finding the motif at any any point in each of the sequences is calculated based on the distribution of bases/amino acids within each column of the initial guess.
      - The probabilities are used to redefine the initial distribution within each motif column.
    - Maximization Step
      - The new counts based on the probabilities are used to change the initial guess.
    - This is iterated until convergence
  - The size of a motif region can be determined by the EM algorithm based on log likelihood scores after one iteration
- GIBBS Sampler Method
  - Searches for the statistically most probable motifs.
  - The goal is to maximize the ratio of motif probability to background probability.



Hidden Markov Models: Local MSA

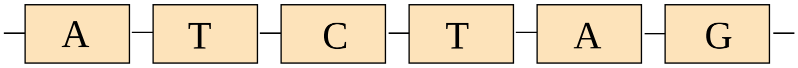
- Hidden Markov Models
  - Statistically based
  - Produce sound results
  - Provide representations of sequence domains or protein families
  - Drawback …
    - Require lots of data to train…

Markov Chains



- Probability for each state
  - is based only on a set number of preceding characters
- # of preceding characters = ***order*** of the Markov Model
- Probability of a sequence (order = 2):
  - P(s) = P{T} P{T,C} P{T,C,A} P{C,A,G} P{A,G,C} P{G,C,C} P{C,C,T}

Hidden Markov Models

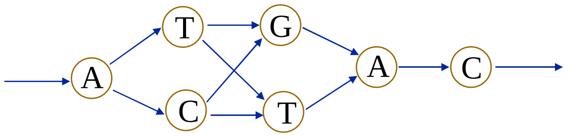


Observed frequencies	A 0.7 T 0.3	A 0.1 T 0.9	C 0.8 G 0.2	A 0.4 T 0.6	A 0.8 T 0.2	C 0.3 G 0.7

Probabilistic model - true state is unknown

Hidden Markov Models

States -- well defined conditions
Edges -- transitions between the states

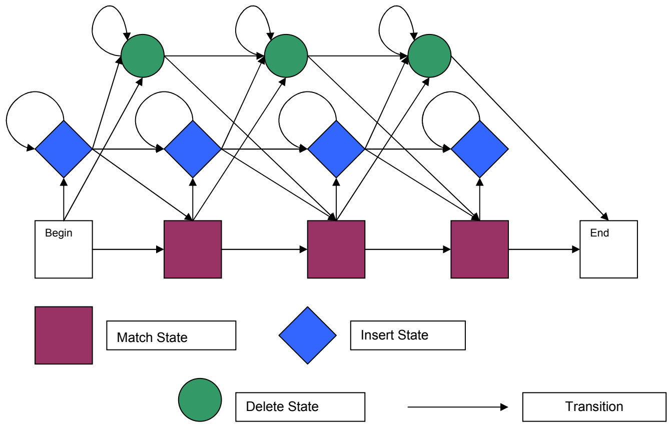


ATGAC
ATTAC
ACGAC
ACTAC

Each transition assigned a probability.

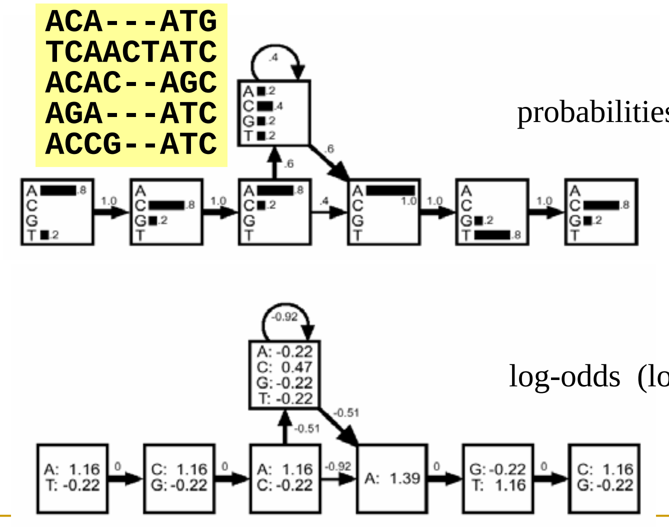
Probability of the sequence:
single path with the highest probability --- *Viterbi* path
sum of the probabilities over all paths -- *Baum-Welch* method

Common Protein Sequence HMM



Adapted representation of the model by of Krogh et al. 1994

Hidden Markov Models



probabilities

log-odds $(\log \frac{P(S)}{0.25^L})$

Adopted from Anders Krogh, 1998

How HMMs are built/ trained

- Start with a basic model structure and a given distribution of probabilities
- Algorithms for training markov models
  - Forward and Backward Algorithms
- Predictive Measures
  - Viterbi
  - Baum-Welch
- Best practices to improve HMM performance
  - Alter prior distribution of amino acids
    - Knowledge added start point can emphasize convergence
  - Alter architecture based on familial knowledge

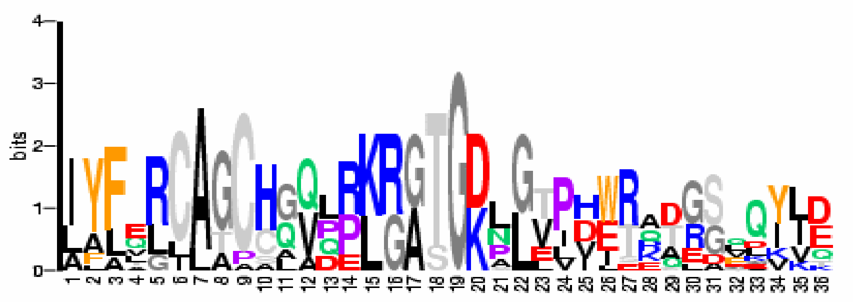
PSSM Calculation…

- The goal of local MSA is to find small regions that are highly conserved.
  - Once these regions are found it is important to consider the underlying distribution by which the motif was created.
  - What manner is the sequence to be compared
    - [FW] - [SGNH] - x - [GD] - {F} - [RKHPT] - {P} - C - [LIVMFAP] - [GAD]
    - Does this contain sufficient variation?
    - Most likely the search will be done with dynamic programming
      - Thus the scoring matrix is important and must contain sufficient variation, but at the same time limit the number of false predictions.
      - To do this we measure the scoring matrix how?
      - Shannon Entropy measure:

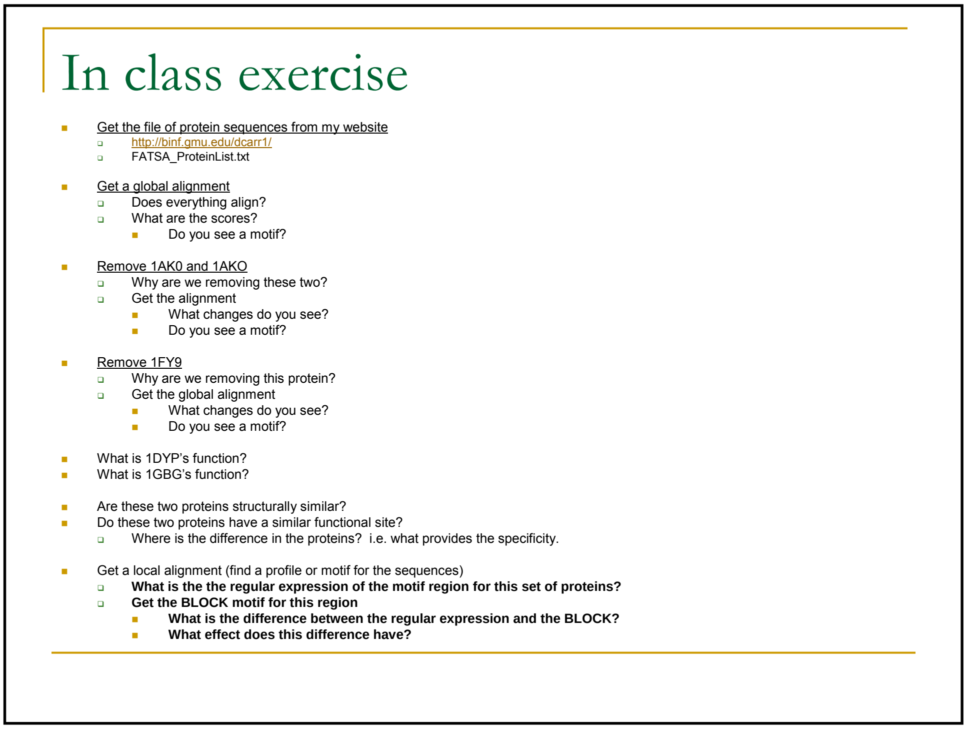
$$H(x) = - \sum_{i=1}^{M} P_i \log_2 P_i$$

- In this case the x is the column in the motif and i is the amino acid/base. P_i is the frequency of the amino acid/base.
- Total score for a scoring matrix can is the sum of all the column scores: $H_{total} = \sum H(x)$

The Logo Display of Shanon Entropy Scores for heme BLOCK.



PSSM of IPB003143A (D1_heme;) 12 sequences.



In class exercise

- Get the file of protein sequences from my website
  - http://binf.gmu.edu/dcarr1/
  - FATSA_ProteinList.txt
- Get a global alignment
  - Does everything align?
  - What are the scores?
    - Do you see a motif?
- Remove 1AK0 and 1AKO
  - Why are we removing these two?
  - Get the alignment
    - What changes do you see?
    - Do you see a motif?
- Remove 1FY9
  - Why are we removing this protein?
  - Get the global alignment
    - What changes do you see?
    - Do you see a motif?
- What is 1DYP's function?
- What is 1GBG's function?
- Are these two proteins structurally similar?
- Do these two proteins have a similar functional site?
  - Where is the difference in the proteins? i.e. what provides the specificity.
- Get a local alignment (find a profile or motif for the sequences)
  - **What is the the regular expression of the motif region for this set of proteins?**
  - **Get the BLOCK motif for this region**
    - **What is the difference between the regular expression and the BLOCK?**
    - **What effect does this difference have?**