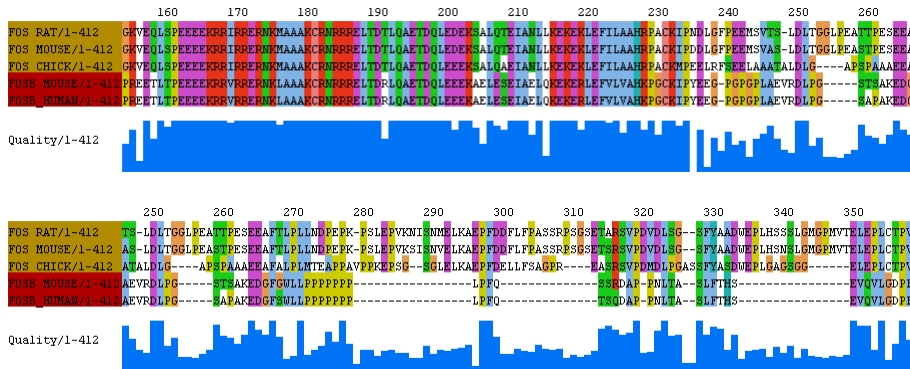


# CMSC 838T – Lecture 4

- ◆ **Multiple sequence alignment (MSA)**
  - Alignment containing multiple DNA / protein sequences
  - Look for conserved regions → similar function



CMSC 838T – Lecture 4

## Multiple Sequence Alignments - Motivation

- ◆ **Identify highly conserved residues**
  - Likely to be essential sites for structure / function
  - More precision from multiple sequences
  - Better structure / function prediction, pairwise alignments
- ◆ **Building gene / protein families**
  - Use conserved regions to guide search
- ◆ **Basis for phylogenetic analysis**
  - Infer evolutionary relationships between genes
- ◆ **Develop primers & probes**
  - Use conserved region to develop
    - Primers for PCR
    - Probes for DNA microarrays

CMSC 838T – Lecture 4

# Multiple Sequence Alignment (MSA)

## ◆ Outline

- Basic concepts & terms
- Global alignment
  - Optimal – dynamic programming (MSA)
  - Progressive – pairwise (PILEUP, CLUSTALW)
  - Iterative progressive (MULTALIN)
  - Block-based (DIALIGN)
- Local alignment (**motif finding**)
  - Patterns (MOTIF, PROTOMAT)
  - Statistical profiles (HMMER2, PSI-BLAST)
- Viewing & editing multiple sequence alignments



CMSC 838T – Lecture 4

# Terminology (for Proteins)

## ◆ Family

- Group of proteins of similar biochemical function with (roughly) > 50% sequence identity when aligned
- Family is transitive, even if sequence identity < 50%
  - $A \rightarrow B$  and  $B \rightarrow C$  implies  $A \rightarrow C$
- 1940 protein families in Protein DataBank (v1.61, Nov 2002)

## ◆ Superfamily

- Group of protein families related by distant yet detectable sequence similarity
- 1100 protein superfamilies in Protein DataBank (v1.61)

CMSC 838T – Lecture 4

## Terminology (for Protein Sequence)

- ◆ **Block**
  - Ungapped conserved sequence pattern (in protein family)
- ◆ **Motif**
  - Conserved sequence pattern found in multiple proteins with similar biochemical activity, usually near active site
- ◆ **Module**
  - Conserved sequence (contiguous) of one or more motifs, considered fundamental unit of structure or function
- ◆ **(Homologous) Domain**
  - Extended sequence pattern suggesting common evolutionary origin (contains one or more motifs, may contain gaps)
- ◆ **Multi-domain (chimeric) protein**
  - Encoded by (artificial) gene containing multiple domains

CMSC 838T – Lecture 4

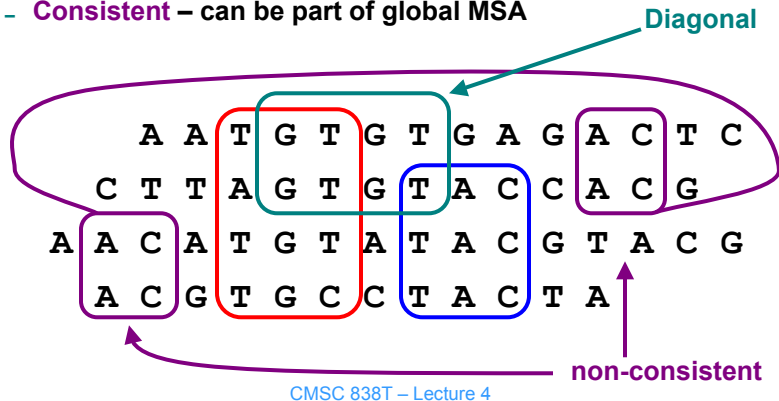
## Terminology (for Protein Structure)

- ◆ **Motif / super-secondary structure**
  - Combination of several secondary structural elements, folding adjacent polypeptide chains into specific 3D configurations
- ◆ **Fold**
  - Similar to motif, but usually larger combination of secondary structural units
  - 701 folds in Protein DataBank (v1.61, Nov 2002)
- ◆ **Domain**
  - Segment of polypeptide chain that can fold into 3D structure irrespective of other segments (multiple domains in protein)
- ◆ **Class**
  - Classify domains according to secondary structure
  - Examples: mainly- $\alpha$ , mainly- $\beta$ ,  $\alpha / \beta$ ,  $\alpha + \beta$ , membrane

CMSC 838T – Lecture 4

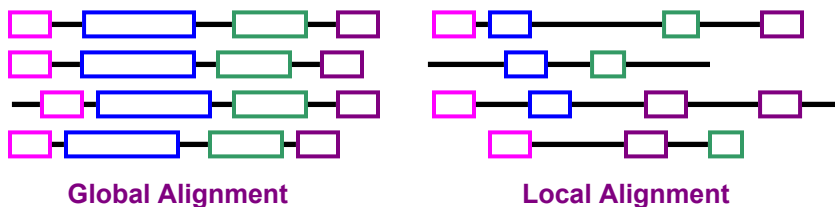
## Multiple Sequence Alignment - Block

- ◆ Ungapped conserved sequence pattern
- ◆ Types of blocks
  - **Exact** – composed of identical segments
  - **Uniform** – found in every sequence
  - **Consistent** – can be part of global MSA



## MSA - Global vs. Local Alignment

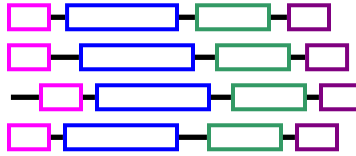
- ◆ **Protein structure**
  - Derived from a limited number of building blocks (domains) that have been mixed and shuffled through evolution
  - Proteins can thus share a **global** or **local** relationship
- ◆ **Global sequence alignment**
  - Alignment over entire sequence (near same length)
- ◆ **Local sequence alignment**
  - Alignment over parts(s) of sequence



## Multiple Sequence Alignment - Approaches

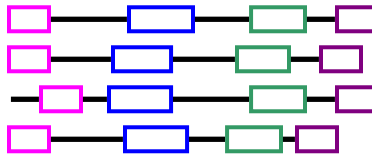
### ◆ Progressive global alignment

- If sequences related over entire length



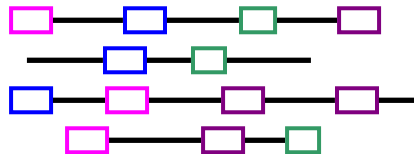
### ◆ Block-based global alignment

- If related by large consistent blocks



### ◆ Local alignment

- If related by small non-consistent blocks



CMSC 838T – Lecture 4

## Multiple Sequence Alignments - Issues

### ◆ No single “correct” answer

- Ideally, find single evolutionary correct alignment
- In practice, evolutionary history must be inferred
- Try find sequence alignment → good structure alignment

### ◆ Typical alignment target

- Protein family with ~30% identity

### ◆ Computationally expensive

- “Optimal” solution exponential in number of sequences
- Greater reliance on (greedy) heuristics

### ◆ Benefits from user interaction

- Select which sequences to include in alignment
- Select which regions to align
- Edit resulting alignments

CMSC 838T – Lecture 4

## Scoring Multiple Sequence Alignments

### ◆ Sum of pairs (SP)

- Align sequences in columns
- Score all possible pairwise combinations in column
- Requires  $\frac{N \times (N-1)}{2}$  comparisons for N sequences
- Total score =  $\sum$  score for each column



### ◆ Issues

- Assumes each sequence descended from N-1 sequences
- No probabilistic justification for SP scoring
- Relative penalty for mismatch goes **down** for more matches, should go up
- Example: mismatched **T**

$A \times A = 4$	$A \ A$	$A \ A$	$A \ A$	$A \ A$	$A \ A$	$A \ A$	$A \ A$
$A \times T = 0$	$A \ T$	$A \ A$	$A \ T$	$A \ A$	$A \ T$	$A \ A$	$A \ T$
	12 4	24 12	40 24				
	$\Delta=67\%$	$\Delta=50\%$	$\Delta=40\%$				

CMSC 838T – Lecture 4

## Multiple Sequence Alignment (MSA)

### ◆ Outline

- Basic concepts & terms
- Global alignment
  - Optimal – dynamic programming (MSA)
  - Progressive – pairwise (PILEUP, CLUSTALW)
  - Iterative progressive (MULTALIN)
  - Block-based (DIALIGN)
- Local alignment (**motif finding**)
  - Patterns (MOTIF, PROTOMAT)
  - Statistical profiles (HMMER2, PSI-BLAST)
- Viewing & editing multiple sequence alignments



CMSC 838T – Lecture 4

## Global MSA - Approach

- ◆ **General approach to global MSA**
  1. Find sequences to align (e.g., result of pairwise search)
  2. Locate region(s) of similar length to include in alignment
  3. Apply global alignment algorithm
  4. Refine alignment (repeat as needed)
    1. Inspect resulting alignment
      - Identify conserved physical / chemical properties
    2. Remove seriously misaligned sequences
    3. Reapply algorithm
    4. Add back remaining sequences
      - While preserving key features of alignment
  5. If looking for local alignment, reduce sequence length to highly conserved regions, align to conserved region

CMSC 838T – Lecture 4

## Global MSA - Dynamic Programming (DP)

- ◆ **“Optimal” dynamic programming** [Sankoff+ 1983]
  - Assume  $k$  sequences of length  $n$
  - Attempt to maximize sum-of-pairs (SP) score
  - Build  $F$ , a  $k$ -dimensional table of length  $n+1$  ( $n^k$  elements)
  - Recursive formula  $\rightarrow F(i) = \max(F(i-1) + SP(\text{column}_i))$
- ◆ **Complexity**
  - $O(n^k)$  entries to fill
  - Each entry combines  $O(2^k)$  other entries
  - Total cost =  $O(2^k n^k)$
- ◆ **Bounded search (MSA)** [Carillo & Lipman 1988]
  - Apply heuristic alignment, use resulting SP to bound search
  - Significant speed improvement, still limited to small values of  $k$



CMSC 838T – Lecture 4

# Global MSA – Progressive Global Alignment

## ◆ Motivation

- Reduce cost by building global alignment incrementally

## ◆ Approach

1. Compute **distance** between all pairs of sequences
2. Build simple **guide tree** reflecting distance between sequences
  - Use **UPGMA** (PILEUP) OR **neighbor-joining** (CLUSTALW)
3. Align sequences following guide tree, starting at leaves
  - Align **consensus** sequences OR **profiles**
  - Use optimal or heuristic pairwise algorithms
  - Attempt to place gaps between conserved regions



## ◆ Problems

- Greedy approach dependent on initial pairwise alignments
- Cannot fix early mistakes (gaps cannot be removed)

CMSC 838T – Lecture 4

# Global MSA – CLUSTALW

## ◆ Algorithm

[Thompson+ 1994]

- Calculate evolutionary distances from alignment scores
- Performs pairwise alignment of **profiles** (probabilities of residues at each position) using dynamic programming
- Later calculates consensus sequence from profile

## ◆ Heuristics for improving multiple alignments

- Weight sequences to compensate for biased representation
- Scoring matrix chosen based on expected similarity from tree
  - E.g., nearby → BLOSUM 80, distant → BLOSUM 50
- Gap penalty modified by residue (function) at position
  - E.g., Higher gap penalty for hydrophobic residues
- Gap penalty higher if first gap in column & nearby gaps
- Dynamically adjust guide tree to defer poor alignments

CMSC 838T – Lecture 4

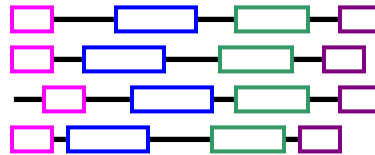
## Global MSA – Iterative Progressive Alignment

- ◆ **Motivation**
  - Avoid problems with initial pairwise alignments by iterating
- ◆ **Star approach** [Barton & Sternberg, 1987]
  1. Align the two sequences w/ highest alignment score
  2. Align sequence w/ highest alignment score to result, repeat
  3. Remove sequence X, align X to profile of remaining sequences
  4. Repeat step 3 for each sequence until scores converge
- ◆ **Tree-based approach (MULTALIN)** [Corpet 1988]
  - Recalculate pairwise scores during alignment
  - Recalculate tree from new pairwise scores, repeat
- ◆ **Problem**
  - Computation intensive

CMSC 838T – Lecture 4

## Global MSA – Block-based Global Alignment


- ◆ **Motivation**
  - Align based on conserved regions despite large gaps
- ◆ **Approach**
  - Search for ungapped conserved regions (**blocks**)
    - Pairwise alignment → weighted diagonals (DIALIGN)
    - Suffix trees → common subsequences
    - Dot matrix plots → diagonals
  - Use blocks as anchors to align segments
    - Find consistent set of uniform / near-uniform blocks
    - Align sequences to produce maximum SP weight
- ◆ **Problem**
  - Finding optimal consistent non-uniform blocks is NP-hard



CMSC 838T – Lecture 4

# Multiple Sequence Alignment (MSA)

## ◆ Outline

- Basic concepts & terms
- Global alignment
  - Optimal – dynamic programming (MSA)
  - Progressive – pairwise (PILEUP, CLUSTALW)
  - Iterative progressive (MULTALIN)
  - Block-based (DIALIGN)
- Local alignment (**motif finding**) 
  - Patterns (MOTIF, PROTOMAT)
  - Statistical profiles (HMMER2, PSI-BLAST)
- Viewing & editing multiple sequence alignments

CMSC 838T – Lecture 4

## Local MSA (Motif-finding) - Approach

### ◆ Motivation

- Find local regions of high similarity (motifs)
- Align based on motifs

### ◆ Approach

- Find motifs
  - Patterns
  - Blocks
  - Statistical profiles
    - ◆ Position-specific scoring matrix (PSSM)
    - ◆ Hidden Markov model (HMM)
- Align sequences
  - Preserve motifs as much as possible

CMSC 838T – Lecture 4

## Terminology (for Protein Sequences)

- ◆ **Pattern**
  - **Deterministic** syntax describing well-conserved region
- ◆ **Profile**
  - **Probabilistic** syntax describing well-conserved region
  - Score-based representations
    - Position-specific scoring matrix (PSSM)
    - Hidden Markov model (HMM)
- ◆ **Pattern & profile**
  - Can be used to search for motifs / domains of biological significance that characterize protein family

CMSC 838T – Lecture 4

## Significance of Patterns / Motifs

- ◆ **DNA**
  - Recognition sites of restriction endonucleases
  - Codons specifying the amino acid sequence of a protein
  - Intron splice sites
  - Promoter
  - Binding sites for regulatory proteins which activate or repress transcription
- ◆ **Proteins**
  - Presence of active sites
  - Prediction of protein secondary structure
  - Presence of signals used to localize the protein in the cell

CMSC 838T – Lecture 4

## Local MSA – Patterns

### ◆ Motivation

- Align sequences based on regular patterns

### ◆ Pattern syntax (PROSITE)

- Single residue **A** - Wildcard **x**
- Set of residues **[ACD]** - Wildcard length **x(3)**
- Excluded residues **{FHW}** - Varying lengths **x(3,6)**

### ◆ Example

- Docking of a kinase to a receptor

x(3)-[DE]-[AVLI]-x(4)-[RKH]-[VFHW]-x(3)

X	X	X	D	A	X	X	X	X	R	Y	X	X	X
			E	V					K	F			
				L					H	W			
				I					H				

CMSC 838T – Lecture 4

## Local MSA – Pattern Discovery

### ◆ Pattern-driven approach (MOTIF)

1. Fetch patterns from PROSITE database (expert curated)
2. Find all possible triplets in patterns ( $aa_1 d_1 aa_2 d_2 aa_3$ )
3. Look for (triplet) matches in sequence (usually  $d \leq 20$ )

### ◆ Sequence-driven approach (ASSET, BLOCKMAKER)

1. Search for patterns

- Perform pairwise alignment
- Store common positions (**diagonals**)
- Remove different positions
- Repeat for all sequences

A	C	T	G	A
A	T	T	G	C
A	-	T	G	-

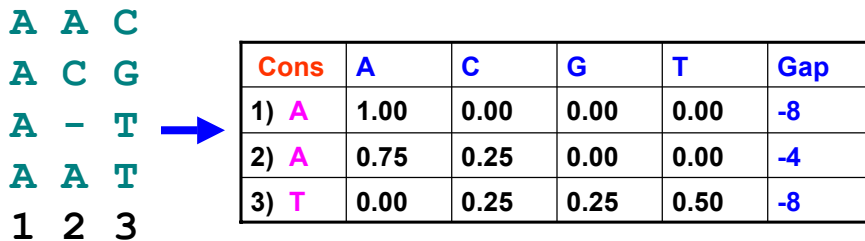
2. Merge overlapping patterns
3. Verify statistical significance with Gibbs sampling
4. Select best patterns

CMSC 838T – Lecture 4

## Local MSA - Statistical Profile

### ◆ Position-specific scoring matrix (PSSM)

- Summary representation for (aligned) conserved region
- Stores probability of element at each position in sequence
- Entries usually stored in log-odds form
- Weight entries by 1) average proportion, 2) evolutionary dist.
- **Consensus** → most likely base / residue at each position

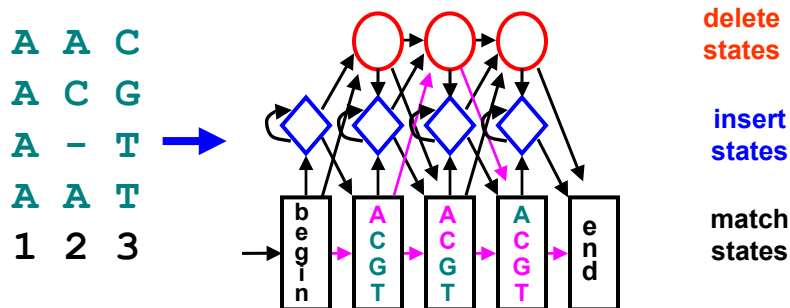


CMSC 838T – Lecture 4

## Local MSA - Statistical Profile

### ◆ Hidden Markov model (HMM)

- Statistical summary representation for conserved region
- Model stores probability of match, mismatch, insertions, and deletions at each position in sequence
- Alignment of conserved region not necessary, but helpful



CMSC 838T – Lecture 4

## Local MSA - HMMs

- ◆ **HMM construction** (HMMER2, PSI-BLAST)
  1. Initialize model with estimated amino acid transition probabilities, using PAM / BLOSUM / **Dirichlet** mixtures
  2. For each “training” sequence containing conserved region
    - Find all possible paths for sequence through model
    - **Forward-backward** algorithm =  $O(\text{size} \times \text{sequences})$
    - Increase weight (probability) of each path taken
- ◆ **HMM properties**
  - Ideally use 20-100 training sequences to build model
    - With better initialization, smaller “**training set**” sufficient
  - Well grounded in probability theory (statistical significance)
  - Explicit gap penalties not needed (automatically trained)
  - Can extract consensus sequence (w/ dynamic programming)

CMSC 838T – Lecture 4

## Calibrating Profiles for PSSM & HMM

- ◆ **Profiling methods (PSSM, HMM)**
  - Training set used to build profile may be biased / skewed
    - Over-represented sequences (common motifs)
    - Under-represented sequences (rare residues)
  - Resulting profile matches training set, not desired motif
- ◆ **Weighting / calibration**
  - Differentially weight sequences to compensate for non-representative sampling in training set
  - Similar sequences → lower weights
  - Rare sequences → higher weights
  - Maximum discrimination → set of weights that best differentiate between real matches and background noise
- ◆ **Simulated annealing to avoid local maxima**

CMSC 838T – Lecture 4

## Local MSA - Patterns vs. Profiles

### ◆ Patterns

- Easy to understand
- Human readable
- Can account for long, variable-length gaps

### ◆ Profiles

- More sensitive
- Can be automatically constructed
- Requires sufficient “training” sequences (minimum 20)
- Can estimate statistical significance


CMSC 838T – Lecture 4

## Using Multiple Sequence Alignments

### ◆ Can search based on MSA pattern / profile

- Pattern
- Position specific scoring matrix (PSSM)
- Hidden Markov model (HMM)

### ◆ PSI-BLAST (Position Specific Iterative BLAST)

1. Perform pairwise search, returns multiple sequences
  2. Perform multiple sequence alignment, report E-values
  3. User selects / deletes sequences
  4. Build HMM from sequences
  5. Search database using HMM, returns multiple sequences
  6. Repeat until process stabilizes
- 

CMSC 838T – Lecture 4

## Searching Based on MSA Profile

- ◆ **Advantages**
  - Searches based on domain, not sequences
  - Greatly improved sensitivity in practice
- ◆ **Dependent on user selection / deletion of sequences**
  - Once included in profile, sequence will score well
  - Including false positives (mismatches) reduces accuracy
    - Can use pairwise alignment to compare sequences
    - Demonstrate sequence can mutate into other sequences

CMSC 838T – Lecture 4

## Multiple Sequence Alignment (MSA)

- ◆ **Outline**
  - Basic concepts & terms
  - Global alignment
    - Optimal – dynamic programming (MSA)
    - Progressive – pairwise (PILEUP, CLUSTALW)
    - Iterative progressive (MULTALIN)
    - Block-based (DIALIGN)
  - Local alignment (**motif finding**)
    - Patterns (MOTIF, PROTOMAT)
    - Statistical profiles (HMMER2, PSI-BLAST)
  - Viewing & editing multiple sequence alignments ←

CMSC 838T – Lecture 4

# Viewing & Editing Multiple Alignments

## ◆ Motivation

- Use multiple sequence alignment as starting point
- Improve usability / readability
  - Format alignments
  - Add annotations
- Improve alignments manually with expert knowledge
  - Find biologically significant regions

## ◆ Multiple sequence alignment tools

- Viewers
  - ClustalX, Jalview, Cinema, Sequence logos
- Editors / annotation
  - SeqVu, MACAW

CMSC 838T – Lecture 4

# Viewing Multiple Sequence Alignments

## ◆ Coloring scheme

- Helps better visualize conserved regions

## ◆ Example color code

- AVFPMILW: **RED**, Small
  - (small + hydrophobic (including aromatic -Y))
- DE: **BLUE**, Acidic
- RHK: **MAGENTA**, Basic
- STYHCNGQ: **GREEN**, Hydroxyl + Amine + Basic – Q
- Others: Grey

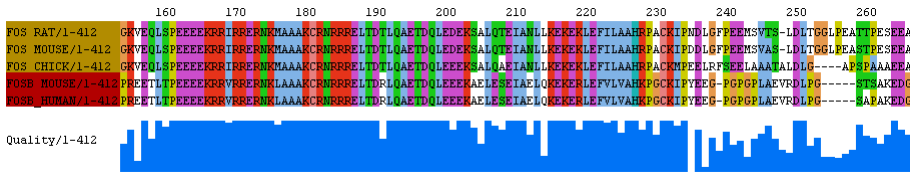
```
FOS_RAT      MMFSGFNADYEASSSRCSASPAGDSLSYHSPADSFSSMGSPVNTQDFCADLSVSSANF 60
FOS_MOUSE   MMFSGFNADYEASSSRCSASPAGDSLSYHSPADSFSSMGSPVNTQDFCADLSVSSANF 60
FOS_CHICK   MMYQGFAGEYEAPSSRCSSASPAGDSLTIYYFPADSFSSMGSPVNSQDFCTDLAVSSANF 60
FOSB_MOUSE  -MFQAFPGDYDS-GSRCSS-SPSAESQ--YLSSVDFGSPPTAAASQE-CAGLGEMPGSF 54
FOSB_HUMAN  -MFQAFPGDYDS-GSRCSS-SPSAESQ--YLSSVDFGSPPTAAASQE-CAGLGEMPGSF 54
*:.:* .:.*: .***** *:.*: * *.*.*.* .:. :*: *:.* .***
```

CMSC 838T – Lecture 4

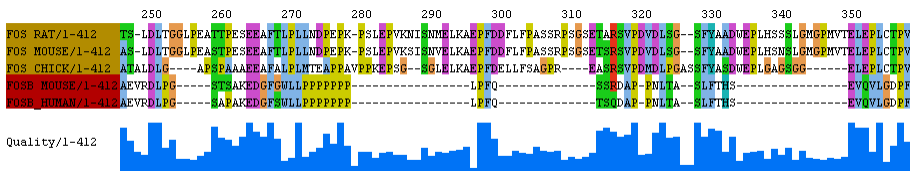
# Viewing MSA - Jalview

## ◆ Examples

### - Conserved region



### - Dissimilar region



CMSC 838T – Lecture 4

# Viewing MSA – Sequence Logos

## ◆ Graphical representation of conserved region

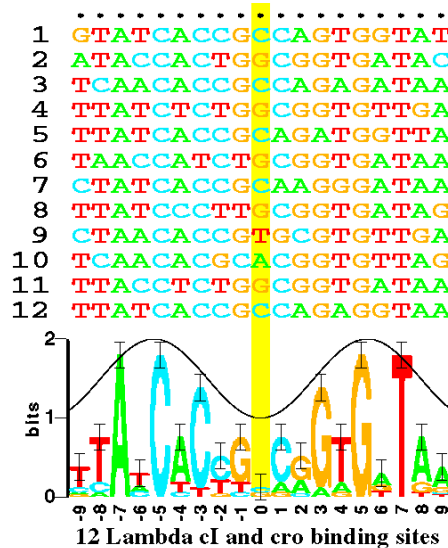
- Good for short sequences
- More information than just consensus sequence

## ◆ Total height of a stack of letters (measured in bits)

- Degree of sequence conservation

## ◆ Relative letter heights

- Frequencies of bases or residues at each position



CMSC 838T – Lecture 4

# Multiple Sequence Alignment (MSA)

## ◆ Summary

- Many multiple sequence alignment algorithms
- Most global alignment algorithms too expensive
  - Exception - progressive pairwise alignment (heuristic)
- Local alignment algs. try to find essential conserved regions
  - Can be very simple (matching motifs)
  - Or use heavy-duty statistical analysis models
- Searches using MSA more sensitive than pairwise alignments
- When using MSA to search / edit motifs
  - Knowledge of biochemistry provides major advantage