CAP 5510: Introduction to Bioinformatics

Giri Narasimhan

ECS 254; Phone: x3748

giri@cis.fiu.edu

www.cis.fiu.edu/~giri/teach/BioinfS07.html

Multiple Alignments

- Global
  - ClustalW, ClustalX
  - MSA
  - T-Coffee
- Local
  - BLOCKS
  - eMOTIF
  - GIBBS
  - HMMER
  - MACAW
  - MEME
- Other
  - Profile Analysis from msa (UCSD)
  - SAM HMM (from msa)

Multiple Alignments: CLUSTALW

* identical

: conserved substitutions

. semi-conserved substitutions

```
gi|2213819      CDN-ELKSEAIIEHLCASEFALR------------MKIKEVKKENGDKK 223
gi|12656123     ----ELKSEAIIEHLCASEFALR------------MKIKEVKKENGD-  31
gi|7512442      CKNKNDDDNDIMETLCKNDFALK------------IKVKEITYINRDTK 211
gi|1344282      QDECKFDYVEVYETSSSGAFSLLGRFCGAEPPPHLVSSHHELAVLFRTDH 400
                    : .   : *  . . *:*               . :*:
```

Red: AVFPMLW (Small & hydrophobic)

Blue: DE (Acidic)

Magenta: RHK (Basic)

Green: STYHCNGQ (Hydroxyl, Amine, Basic)

Gray: Others

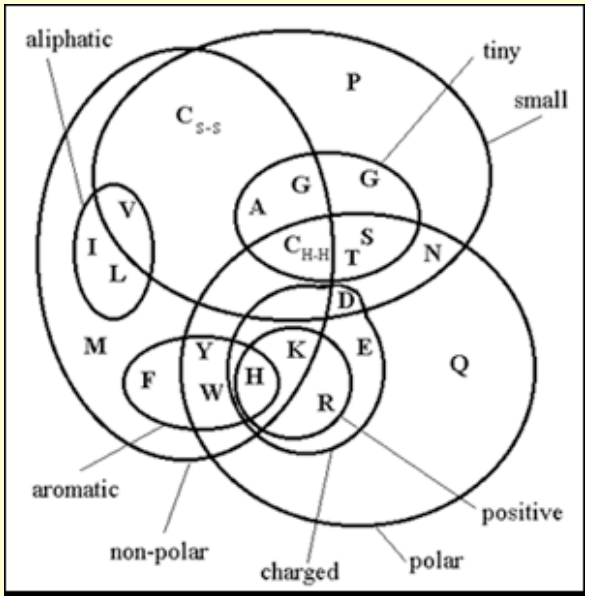


Figure 1. A Venn diagram showing the relationship of the 20 naturally occurring amino acids to a selection of physio-chemical properties thought to be important in the determination of protein structure.

Multiple Alignments

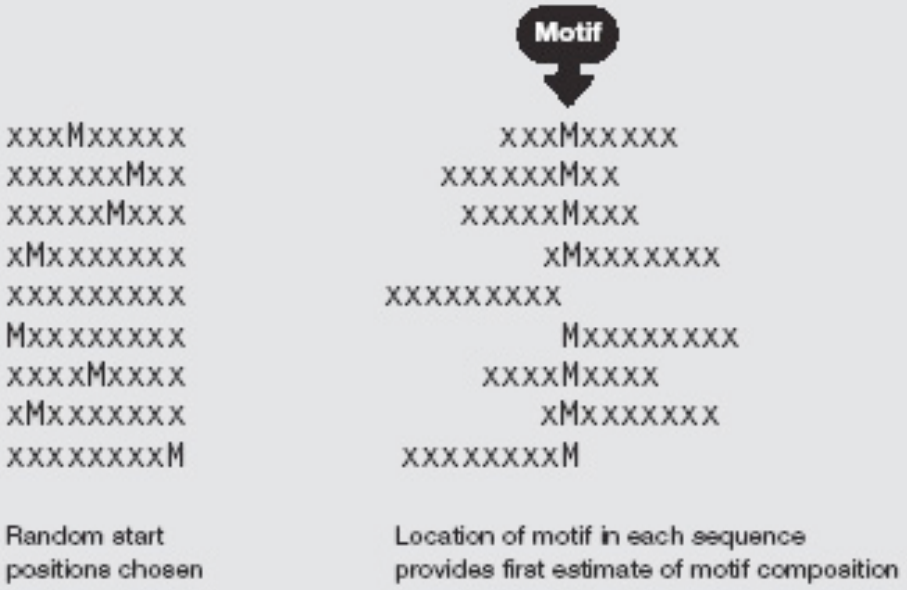
- **Family alignment for the ITAM domain (Immunoreceptor tyrosine-based activation motif)**

```
CD3D_MOUSE/1-2   EQLYQPLRDR EDTQ-YSRLG GN
Q90768/1-21      DQLYQPLGER NDGQ-YSQLA TA
CD3G_SHEEP/1-2   DQLYQPLKER EDDQ-YSHLR KK
P79951/1-21      NDLYQPLGQR SEDT-YSHLN SR
FCEG_CAVPO/1-2   DGIYTGLSTR NQET-YETLK HE
CD3Z_HUMAN/3-0   DGLYQGLSTA TKDT-YDALH MQ
C79A_BOVIN/1-2   ENLYEGLNLD DCSM-YEDIS RG
C79B_MOUSE/1-2   DHTYEGLNID QTAT-YEDIV TL
CD3H_MOUSE/1-2   NQLYNELNLG RREE-YDVLE KK
CD3Z_SHEEP/1-2   NPVYNELNVG RREE-YAVLD RR
CD3E_HUMAN/1-2   NPDYEPIRKG QRDL-YSGLN QR
CD3H_MOUSE/2-0   EGVYNALQKD KMAEAYSEIG TK
Consensus/60%    -.lYpsLspc pcsp.YspLs pp
```

Simple Modular Architecture Research Tool

Multiple Alignment

A. Estimate the amino acid frequencies in the motif columns of all but one sequence. Also obtain background.



How to Score Multiple Alignments?

- Sum of Pairs Score (SP)
  - Optimal alignment: $O(d^N)$ [Dynamic Prog]
  - Approximate Algorithm: Approx Ratio 2
    - Locate Center: $O(d^2N^2)$
    - Locate Consensus: $O(d^2N^2)$

Consensus char: char with min distance sum

Consensus string: string of consensus char

Center: input string with min distance sum

Multiple Alignment Methods

- Phylogenetic Tree Alignment (NP-Complete)
  - Given tree, task is to label leaves with strings
- Iterative Method(s)
  - Build a MST using the distance function
- Clustering Methods
  - Hierarchical Clustering
  - K-Means Clustering

Multiple Alignment Methods (Cont'd)

- Gibbs Sampling Method
  - Lawrence, Altschul, Boguski, Liu, Neuwald, Winton, *Science*, 1993
- Hidden Markov Model
  - Krogh, Brown, Mian, Sjolander, Haussler, *JMB*, 1994

Multiple Sequence Alignments (MSA)

- Choice of Scoring Function
  - Global vs local
  - Gap penalties
  - Substitution matrices
  - Incorporating other information
  - Statistical Significance
- Computational Issues
  - Exact/heuristic/approximate algorithms for optimal MSA
  - Progressive/Iterative/DP
  - Iterative: Stochastic/Non-stochastic/Consistency-based
- Evaluating MSAs
  - Choice of good test sets or benchmarks (BAliBASE)
  - How to decide thresholds for good/bad alignments

Progressive MSA: CLUSTALW

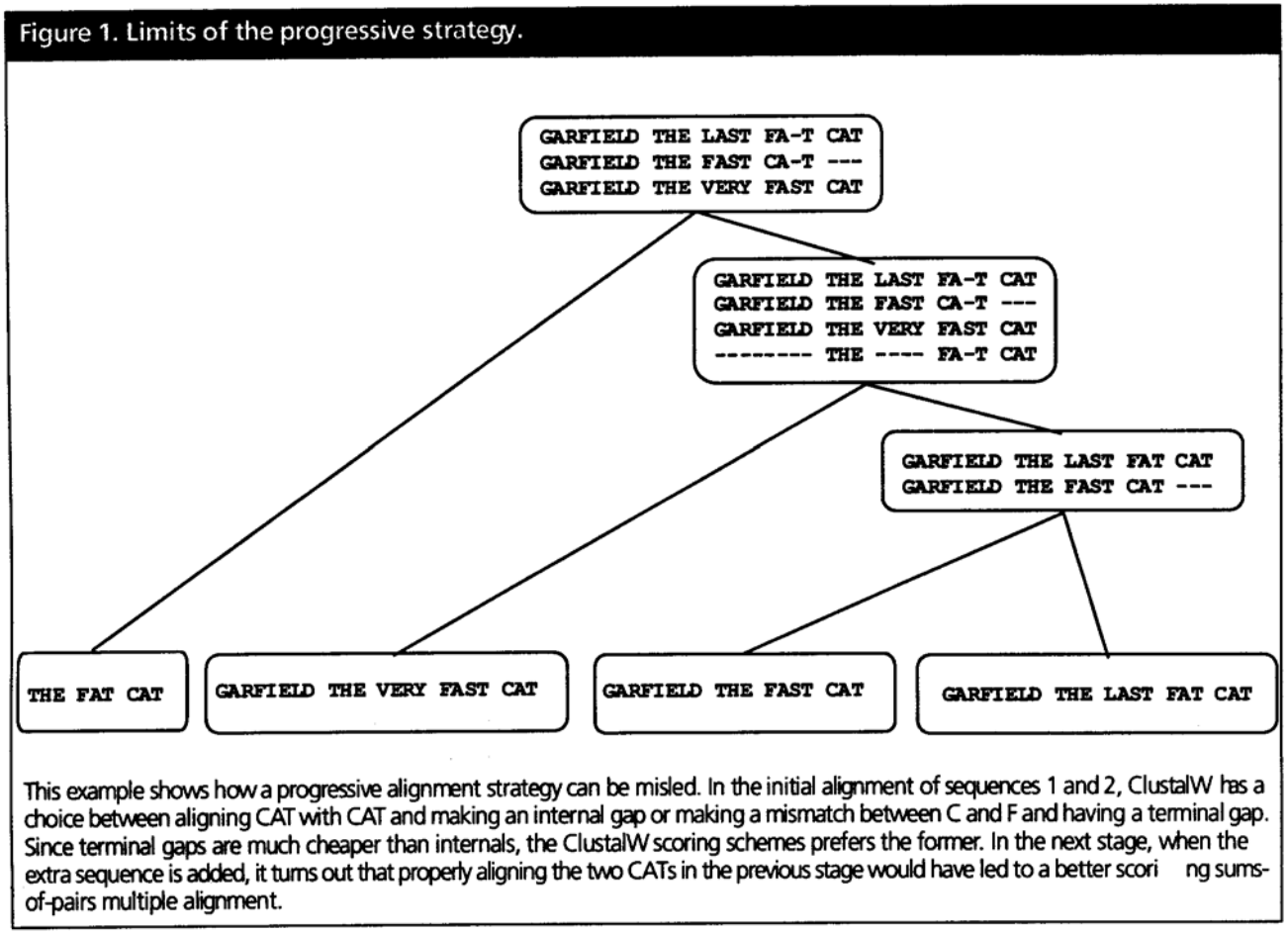


Figure 1. Limits of the progressive strategy.

This example shows how a progressive alignment strategy can be misled. In the initial alignment of sequences 1 and 2, ClustalW has a choice between aligning CAT with CAT and making an internal gap or making a mismatch between C and F and having a terminal gap. Since terminal gaps are much cheaper than internals, the ClustalW scoring schemes prefers the former. In the next stage, when the extra sequence is added, it turns out that properly aligning the two CATs in the previous stage would have led to a better scori ng sums-of-pairs multiple alignment.

C. Notredame, *Pharmacogenomics*, **3**(1), 2002.

Software for MSA

Table 1. Some recent and less recent available methods for MSAs.

MSA	Exact	http://www.ibc.wustl.edu/ibc/msa.html	[28]
OMA	Iterative DCA	http://bibiserv.techfak.uni-biefield.de/oma	[61]
MultAlin	Progressive	http://www.toulouse.inra.fr/multalin.html	[41]
ComAlign	Consistency-based	http://www.daimi.au.df/~ ocaprani	[75]
Praline	Iterative/progressive	jhering@nimr.mrc.ac.uk	[48]
Prrp	Iterative/Stochastic	ftp://ftp.genome.ad.jp/pub/genome/saitama-cc/	[47]
HMMER	Iterative/Stochastic/HMM	http://hmmer.wustl.edu/	[68]
GA	Iterative/Stochastic/GA	czhang@watnow.uwaterloo.ca	[52]

C. Notredame, *Pharmacogenomics*, **3**(1), 2002.

MSA: Conclusions

- Very important
  - Phylogenetic analyses
  - Identify members of a family
  - Protein structure prediction
- No perfect methods
- Popular
  - Progressive methods: CLUSTALW
  - Recent interesting ones: Prrp, SAGA, DiAlign, T-Coffee
- Review of Methods [C. Notredame, *Pharmacogenomics*, **3**(1), 2002]
  - CLUSTALW works reasonably well, in general
  - DiAlign is better for sequences with long insertions & deletions (indels)
  - T-Coffee is best available method