CAP 5510: Introduction to Bioinformatics

Giri Narasimhan

ECS 254; Phone: x3748

giri@cis.fiu.edu

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

BLAST & FASTA

- FASTA

  [Lipman, Pearson '85, '88]

- Basic Local Alignment Search Tool

  [Altschul, Gish, Miller, Myers, Lipman '90]

Rules of Thumb

- Results of searches using different scoring systems may be compared directly using normalized scores.
- If S is the (raw) score for a local alignment, the **normalized** score S' (in bits) is given by

$$S' = \frac{\lambda - \ln(\mathrm{K})}{\ln(2)}$$

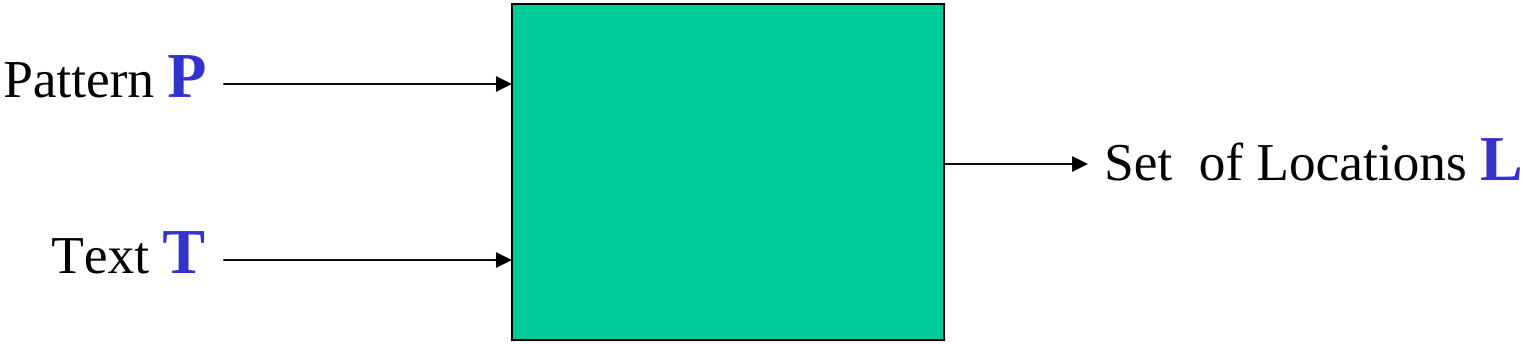
  The parameters depend on the scoring system.

- **Statistically significant normalized score**,

$$S' > \log\left(\frac{N}{E}\right)$$

  where E-value = E, and N = size of search space.

String Matching Problem



Pattern **P**

Text **T**

Set of Locations **L**

(Approximate) String Matching

Input: Text **T** , Pattern **P**

Question(s):

Does **P** occur in **T?**

Find one occurrence of **P** in **T.**

Find all occurrences of **P** in **T.**

Count # of occurrences of **P** in **T.**

Find longest substring of **P** in **T.**

Find closest substring of **P** in **T.**

Locate direct repeats of **P** in **T.**

Many More variants

Applications:

Is **P** already in the database **T**?

Locate **P** in **T.**

Can **P** be used as a primer for **T**?

Is **P** homologous to anything in **T**?

Has **P** been contaminated by **T**?

Is *prefix*(**P**) = *suffix*(**T**)?

Locate tandem repeats of **P** in **T.**

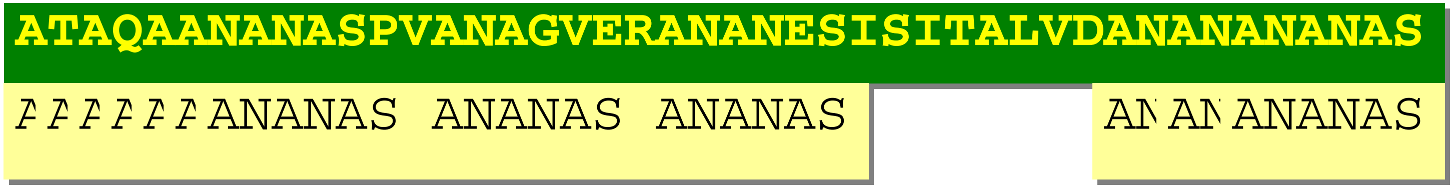
Input: Text **T**; Pattern **P**

Output: All occurrences of **P** in **T**.

Methods:

- Naïve Method
- Rabin-Karp Method
- FSA-based method
- Knuth-Morris-Pratt algorithm
- Boyer-Moore
- Suffix Tree method
- Shift-And method

Naive Strategy



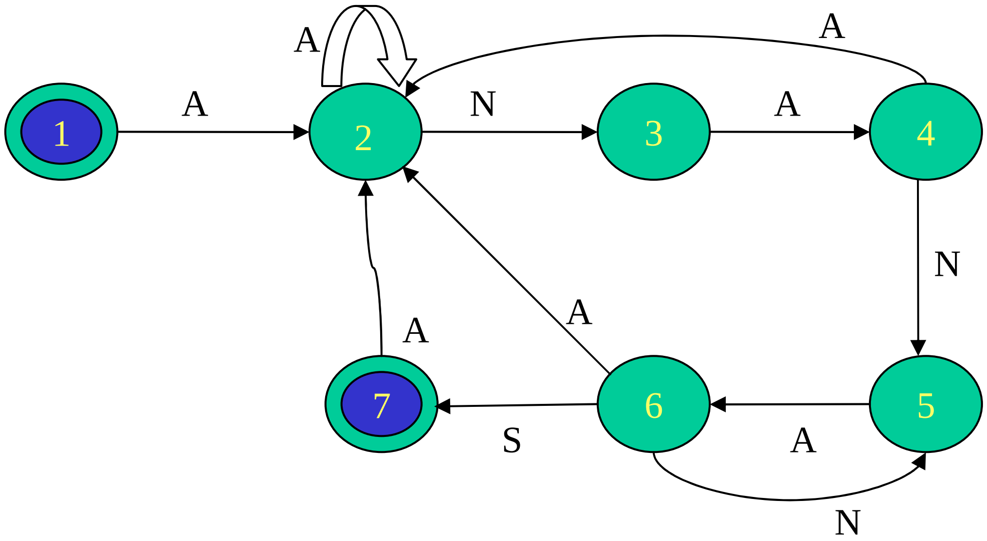
ATAQAANANASPVANAGVERANANESISITALVDANANANANAS

A A A A A A ANANAS ANANAS ANANAS

AN AN ANANAS

Finite State Automaton

ANANAS



Finite
State
Automaton

ATAQAANANASPVANAGVERANANESISITALVDANANANANAS

State Transition Diagram

		A	N	S	*
-	0	1	0	0	0
A	1	1	2	0	0
AN	2	3	0	0	0
ANA	3	1	4	0	0
ANAN	4	5	0	0	0
ANANA	5	1	4	6	0
ANANAS	6	1	0	0	0

Input: Text **T**; Pattern **P**

Output: All occurrences of **P** in **T**.

Sliding Window Strategy:

```
Initialize window on T;
While (window within T) do
    Scan: if (window = P) then report it;
    Shift: shift window to right    (by ?? positions)
endwhile;
```

Tries

Storing:

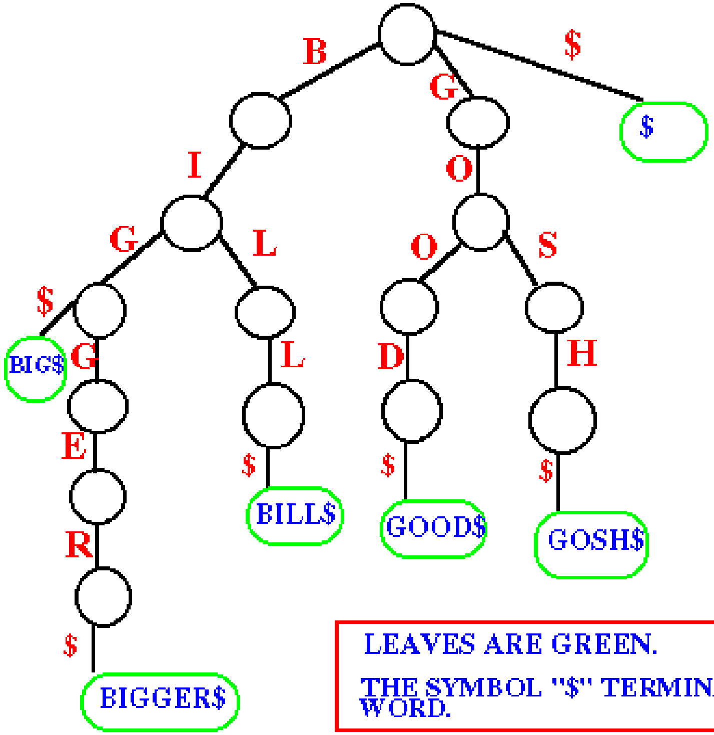
BIG

BIGGER

BILL

GOOD

GOSH



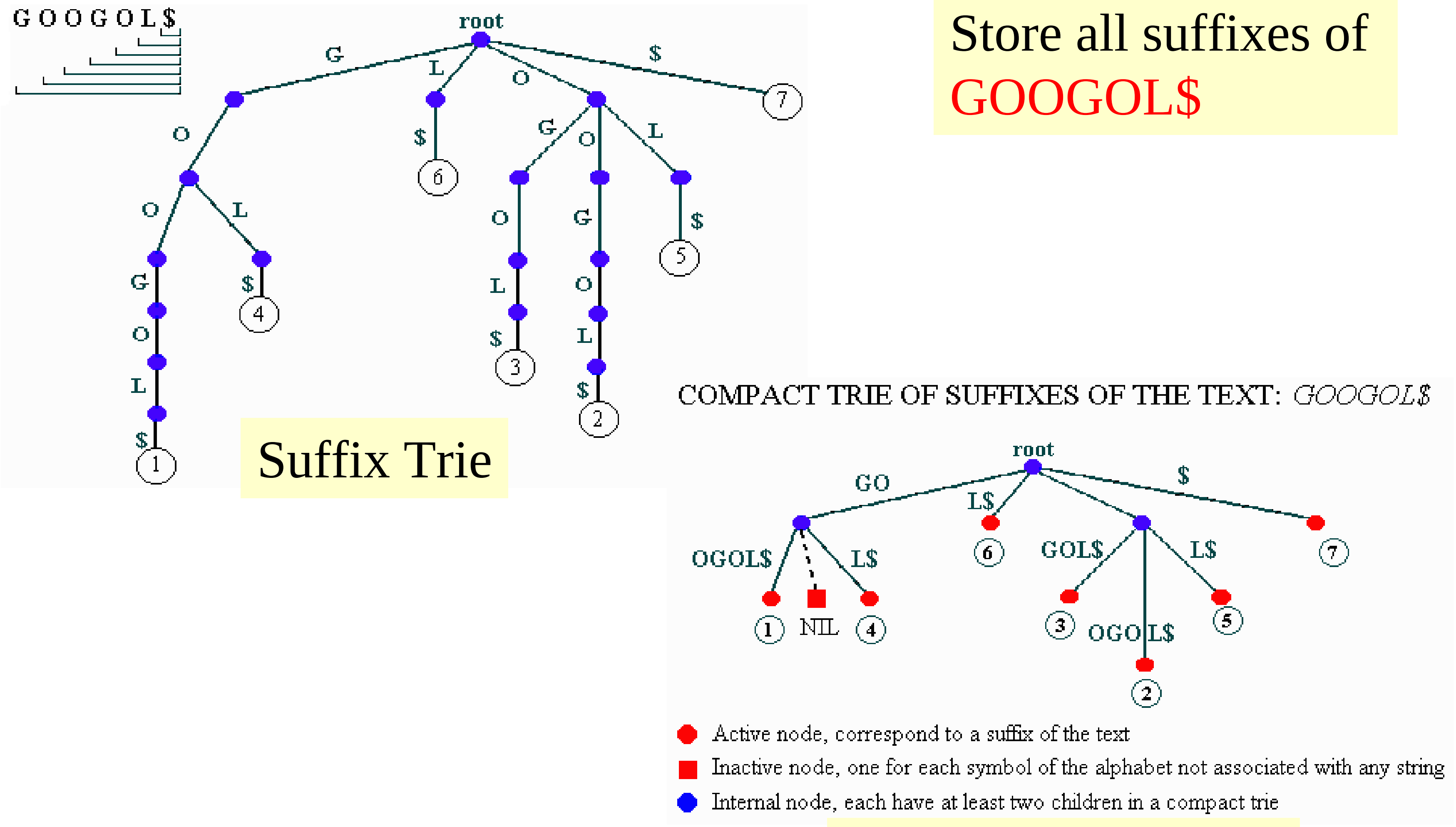
LEAVES ARE GREEN.

THE SYMBOL "$" TERMINATES EACH WORD.

In this figure, the strings either start with B or G. Therefore, the root of the trie is connected to 3 edges called B, G and $.

Suffix Tries & Compact Suffix Tries

Store all suffixes of GOOGOL$



COMPACT TRIE OF SUFFIXES OF THE TEXT: *GOOGOL$*

● Active node, correspond to a suffix of the text

■ Inactive node, one for each symbol of the alphabet not associated with any string

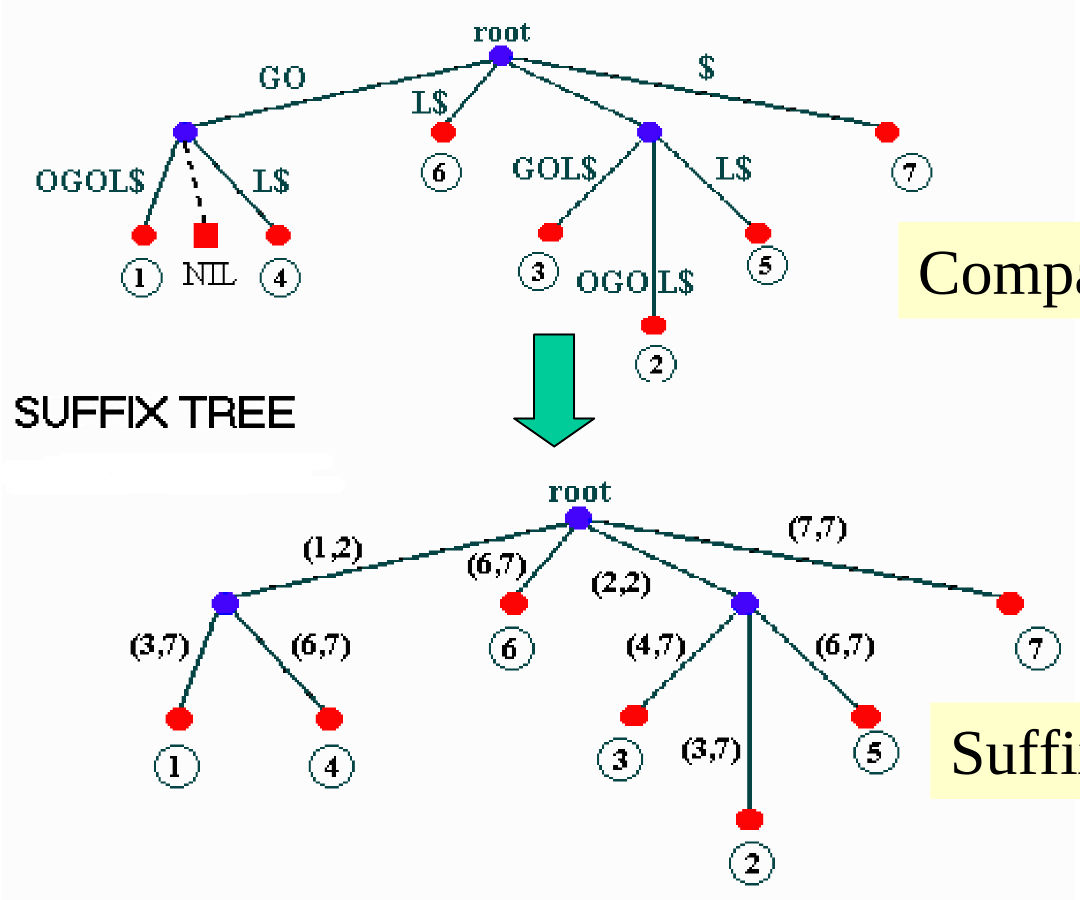
● Internal node, each have at least two children in a compact trie

Suffix Trie

Compact Suffix Trie

Suffix Tries to Suffix Trees

COMPACT TRIE OF SUFFIXES OF THE TEXT: *GOOGOL$*



Compact Suffix Trie

SUFFIX TREE

Suffix Tree

Key:	G	O	O	G	O	L	$
	1	2	3	4	5	6	7

Suffix Trees

- Linear-time construction!
- String Matching, Substring matching, substring common to k of n strings
- All-pairs prefix-suffix problem
- Repeats & Tandem repeats
- Approximate string matching

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

Multiple Alignments

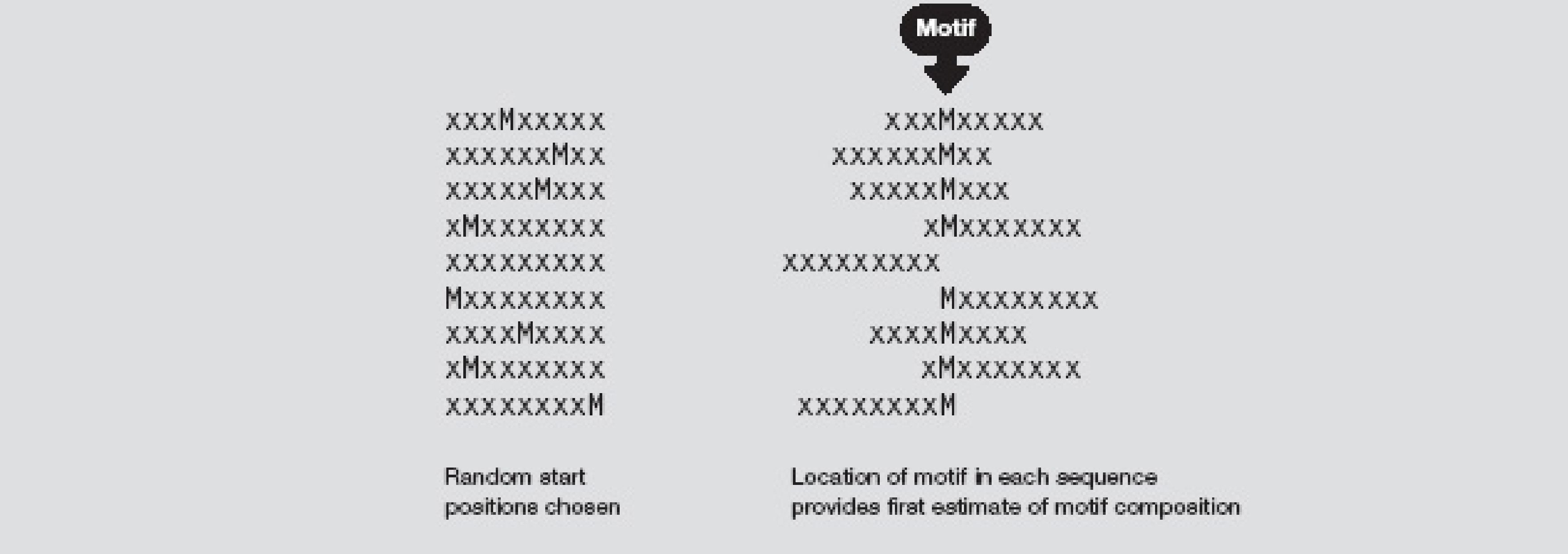
- ❑ **Family alignment for the ITAM domain (Immunoreceptor tyrosine-based activation motif)**

- ❑

```
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



Random start positions chosen

Location of motif in each sequence provides first estimate of motif composition

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)

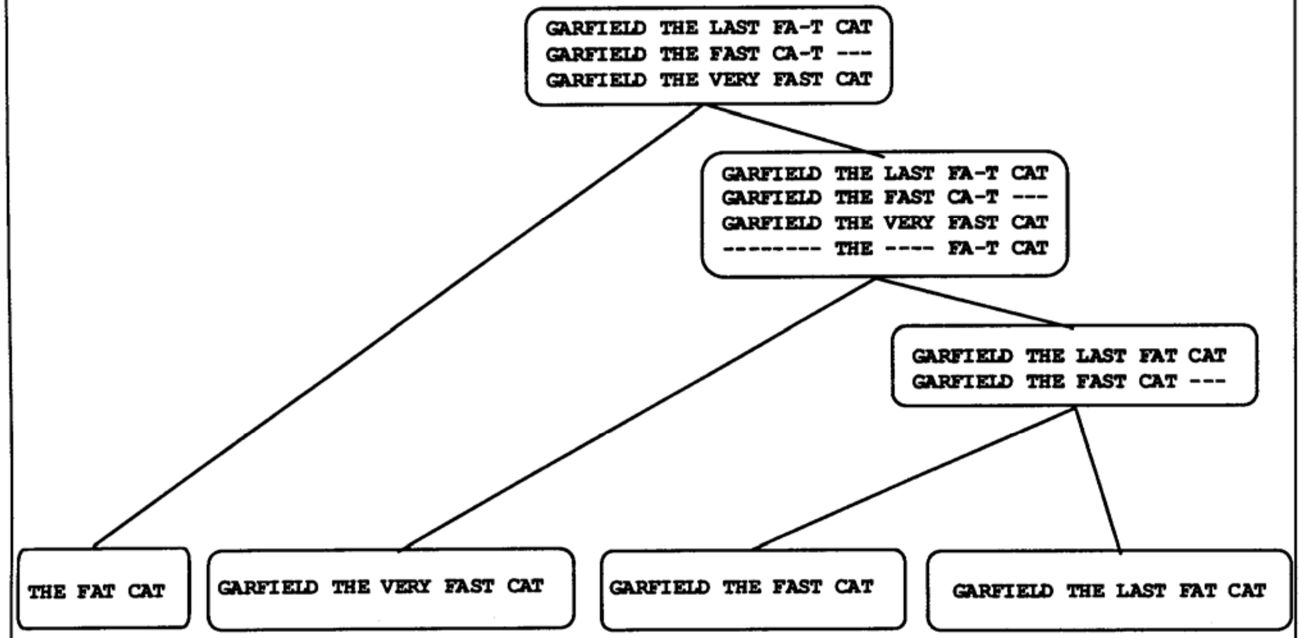
- Gibbs Sampling Method
  - Lawrence, Altschul, Boguski, Liu, Neuwald, Winton, *Science*, 1993
- Hidden Markov Model
  - Krogh, Brown, Mian, Sjolander, Haussler, *JMB*, 1994

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.



```
GARFIELD THE LAST FA-T CAT
GARFIELD THE FAST CA-T ---
GARFIELD THE VERY FAST CAT
```

```
GARFIELD THE LAST FA-T CAT
GARFIELD THE FAST CA-T ---
GARFIELD THE VERY FAST CAT
-------- THE ---- FA-T CAT
```

```
GARFIELD THE LAST FAT CAT
GARFIELD THE FAST CAT ---
```

```
THE FAT CAT
```

```
GARFIELD THE VERY FAST CAT
```

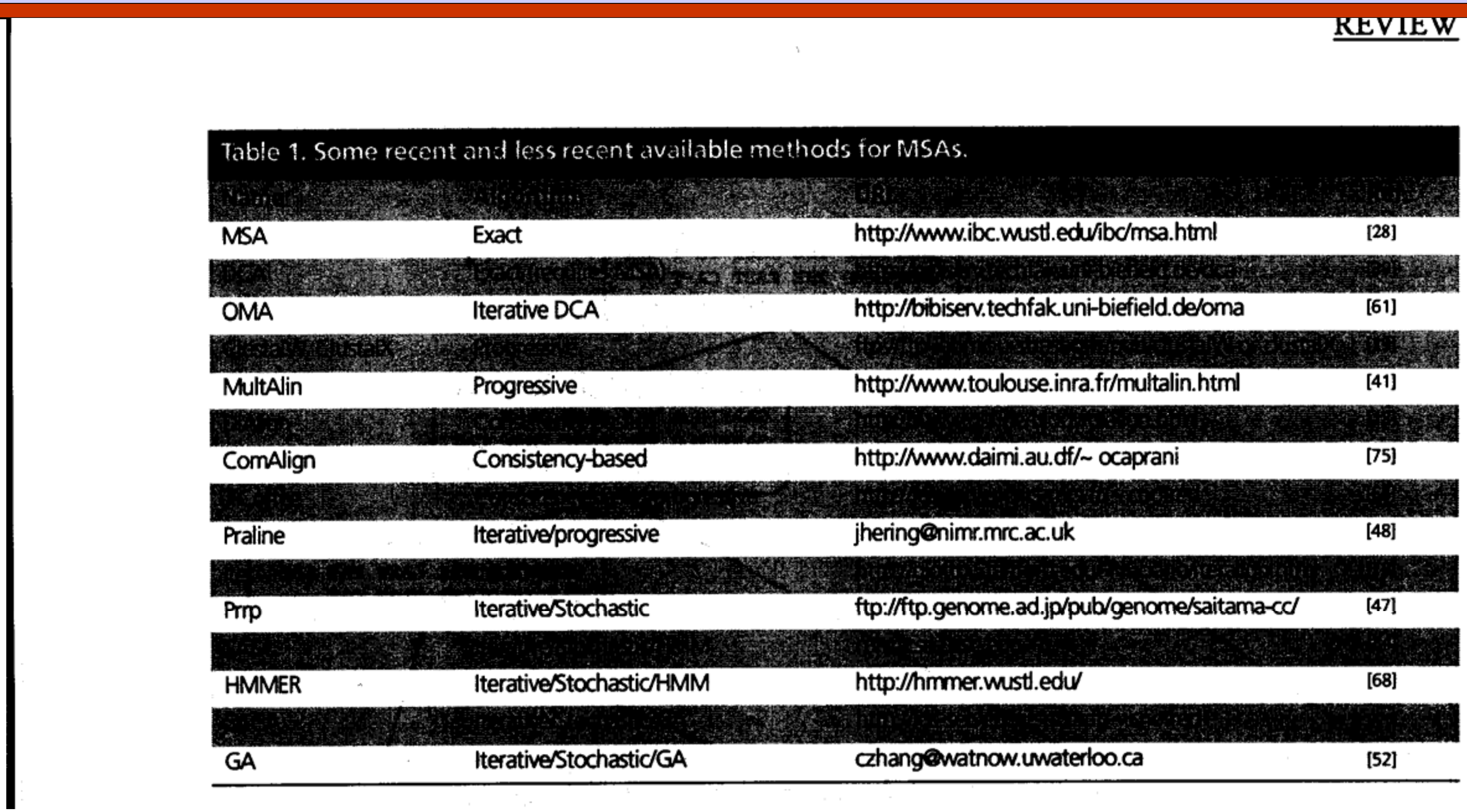
```
GARFIELD THE FAST CAT
```

```
GARFIELD THE LAST FAT CAT
```

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



REVIEW

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

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