CAP 5510: Introduction to Bioinformatics

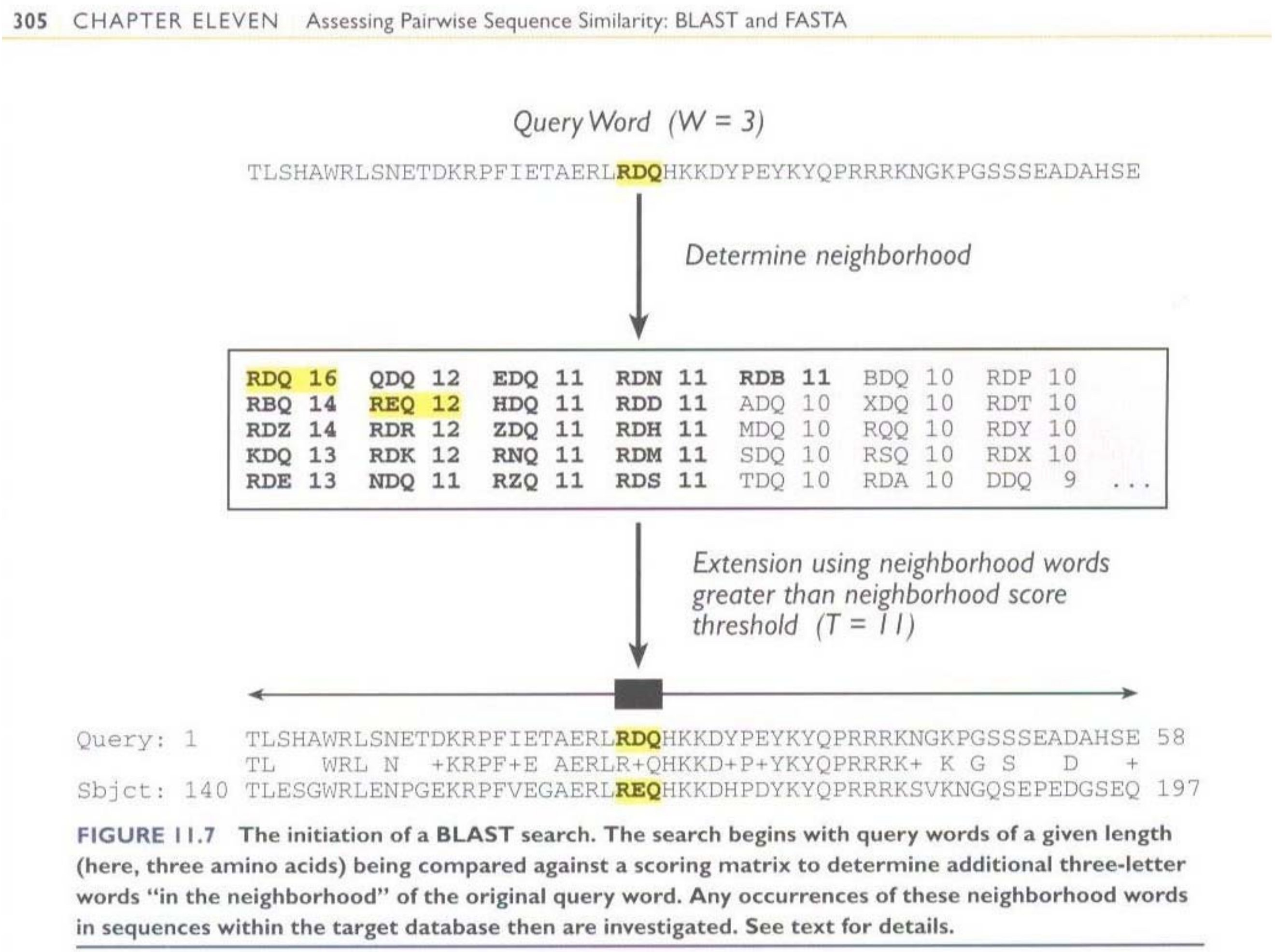
Giri Narasimhan

ECS 254; Phone: x3748

giri@cis.fiu.edu

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

BLAST



305 CHAPTER ELEVEN Assessing Pairwise Sequence Similarity: BLAST and FASTA

Query Word (W = 3)
TLSHAWRLSNETDKRPFIETAERLRDQHKKDYPEYKYQPRRRKNGKPGSSSEADAHSE
Determine neighborhood

RDQ 16	QDQ 12	EDQ 11	RDN 11	RDB 11	BDQ 10	RDP 10
RBQ 14	REQ 12	HDQ 11	RDD 11	ADQ 10	XDQ 10	RDT 10
RDZ 14	RDR 12	ZDQ 11	RDH 11	MDQ 10	RQQ 10	RDY 10
KDQ 13	RDK 12	RNQ 11	RDM 11	SDQ 10	RSQ 10	RDX 10
RDE 13	NDQ 11	RZQ 11	RDS 11	TDQ 10	RDA 10	DDQ 9 . . .

Extension using neighborhood words greater than neighborhood score threshold (T = 11)

```
Query: 1   TLSHAWRLSNETDKRPFIETAERLRDQHKKDYPEYKYQPRRRKNGKPGSSSEADAHSE 58
           TL   WRL N  +KRPF+E AERLR+QHKKD+P+YKYQPRRRK+ K G S   D   +
Sbjct: 140 TLESGWRLENPGEKRPFVEGAERLREQHKKDHPDYKYQPRRRKSVKNGQSEPEDGSEQ 197
```

FIGURE 11.7 The initiation of a BLAST search. The search begins with query words of a given length (here, three amino acids) being compared against a scoring matrix to determine additional three-letter words "in the neighborhood" of the original query word. Any occurrences of these neighborhood words in sequences within the target database then are investigated. See text for details.

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

  $S' = (\lambda S - \ln K)/\ln 2$

  The parameter λ scales for the scoring system, while K scales for the search space size.
- **Statistically significant normalized score**,

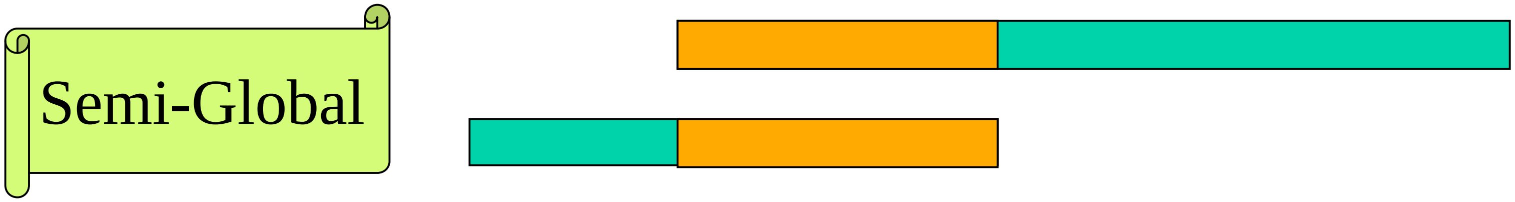
  $$S' > \log\left(\frac{N}{E}\right)$$

  where E-value = E, and N = size of search space.
- Read **http://www.ncbi.nlm.nih.gov/Education/BLASTinfo/glossary2.html** for information about the various terms being used here.

Rules of Thumb

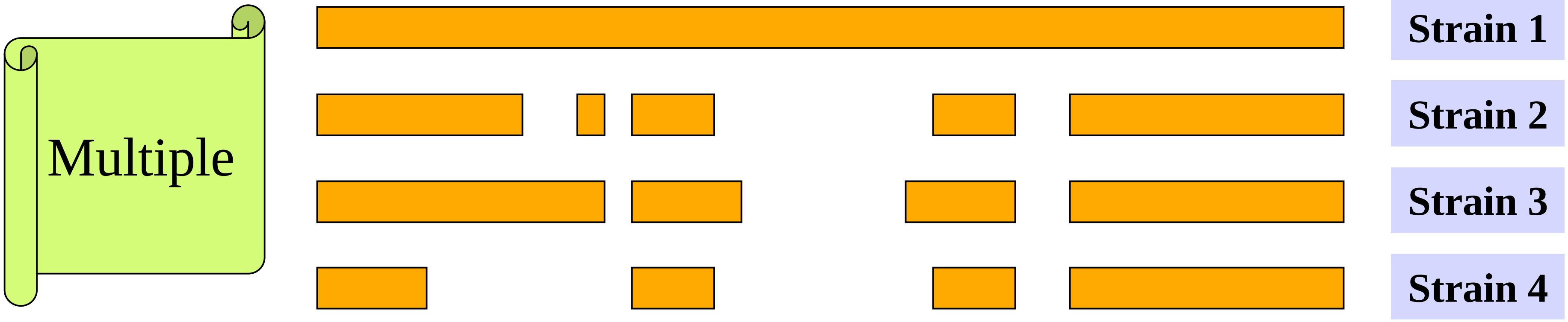
- Most sequences with significant similarity over their entire lengths are homologous.
- Matches that are > 50% identical in a 20-40 aa region occur frequently by chance.
- Distantly related homologs may lack significant similarity. Homologous sequences may have few absolutely conserved residues.
- “A homologous to B” & “B homologous to C” ⇒ “A homologous to C”.
- Low complexity regions, transmembrane regions and coiled-coil regions frequently display significant similarity without homology.
- Greater evolutionary distance implies that length of a local alignment required to achieve a statistically significant score also increases.

Types of Sequence Alignments - 2



Semi-Global

- ❑ Semi-global Alignment: end segments may not be similar



Multiple

Strain 1

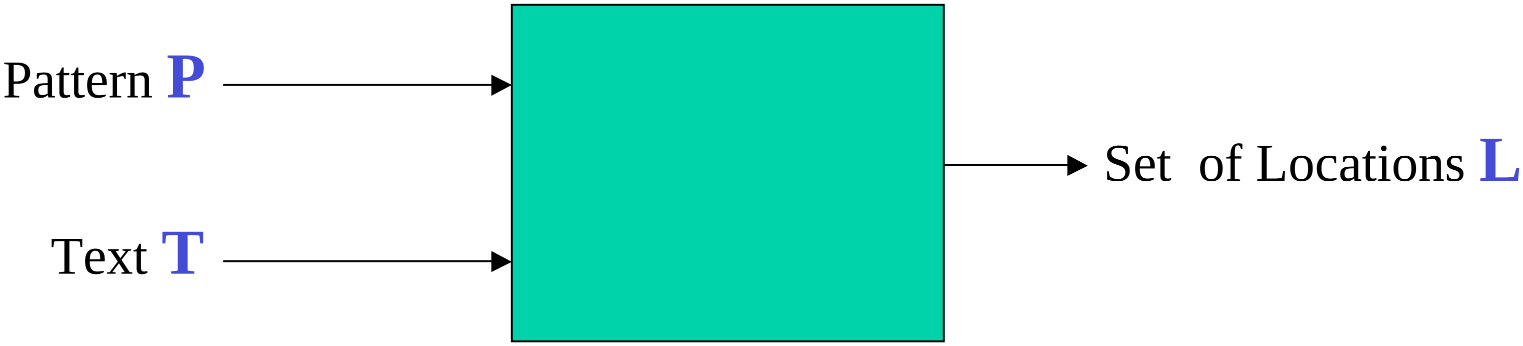
Strain 2

Strain 3

Strain 4

- ❑ Multiple Alignment: similarity between sets of sequences

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.**

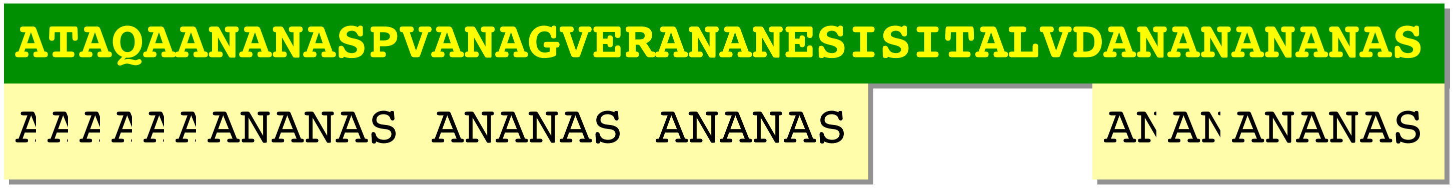
Input: Text **T**; Pattern **P**

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



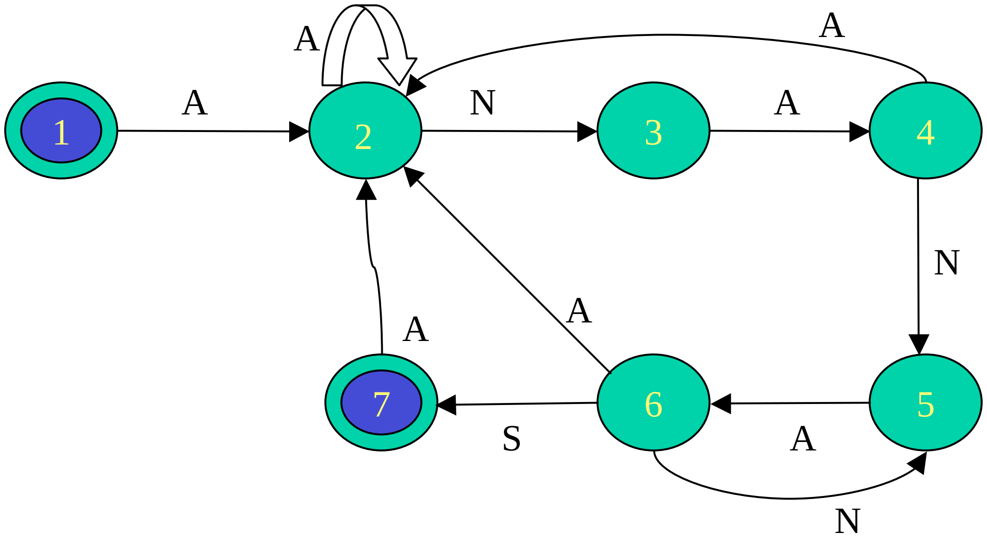
ATAQAANANASPVANAGVERANANESISITALVDANANANANAS

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
1	1	2	0	0
2	3	0	0	0
3	1	4	0	0
4	5	0	0	0
5	1	4	6	0
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:

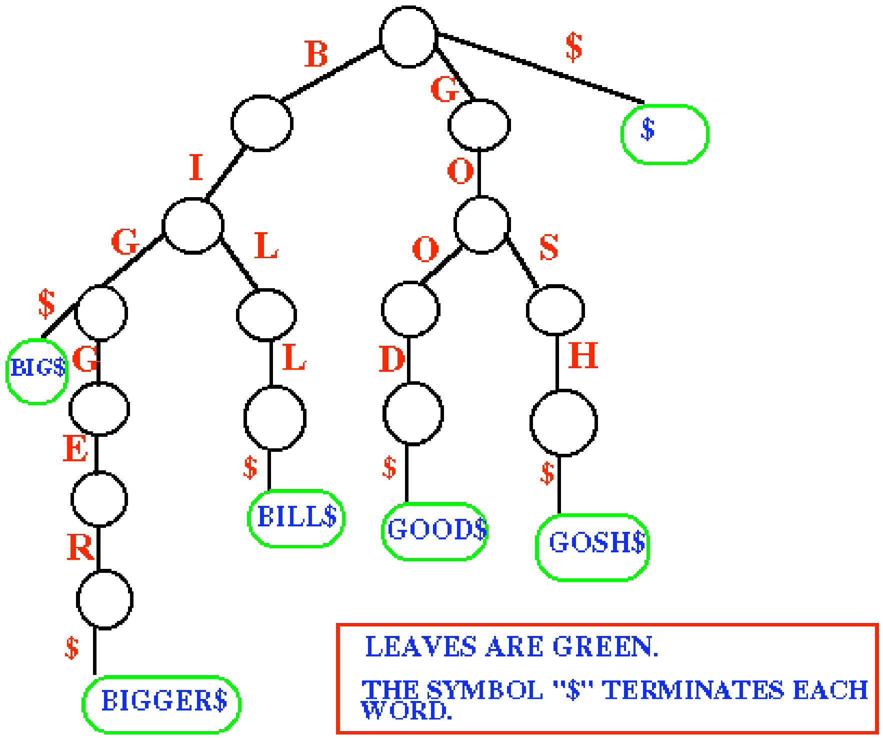
BIG

BIGGER

BILL

GOOD

GOSH



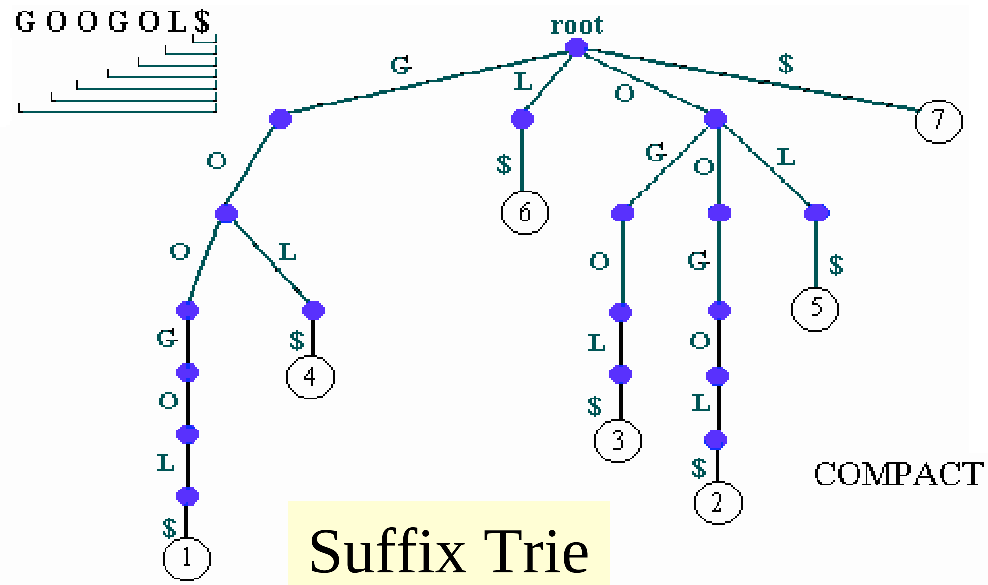
LEAVES ARE GREEN.

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 $.

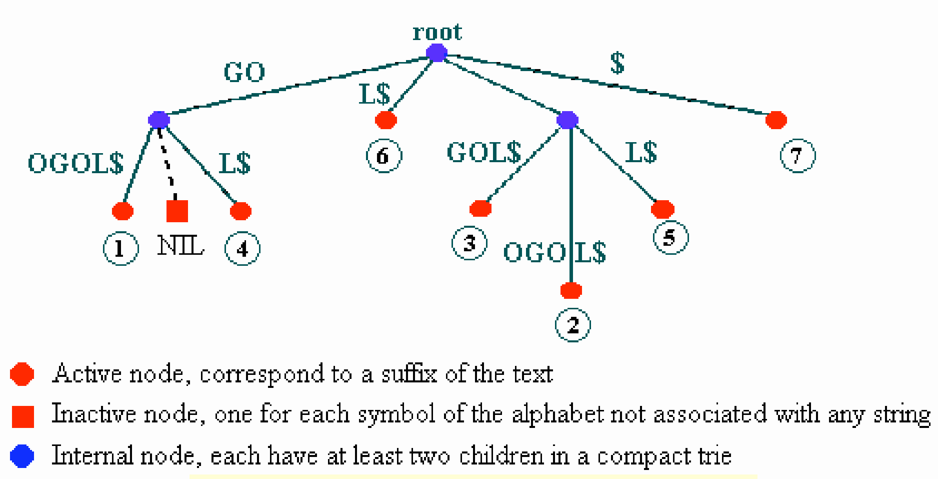
Suffix Tries & Compact Suffix Tries

Store all suffixes of GOOGOL$



Suffix Trie

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



● Active node, correspond to a suffix of the text

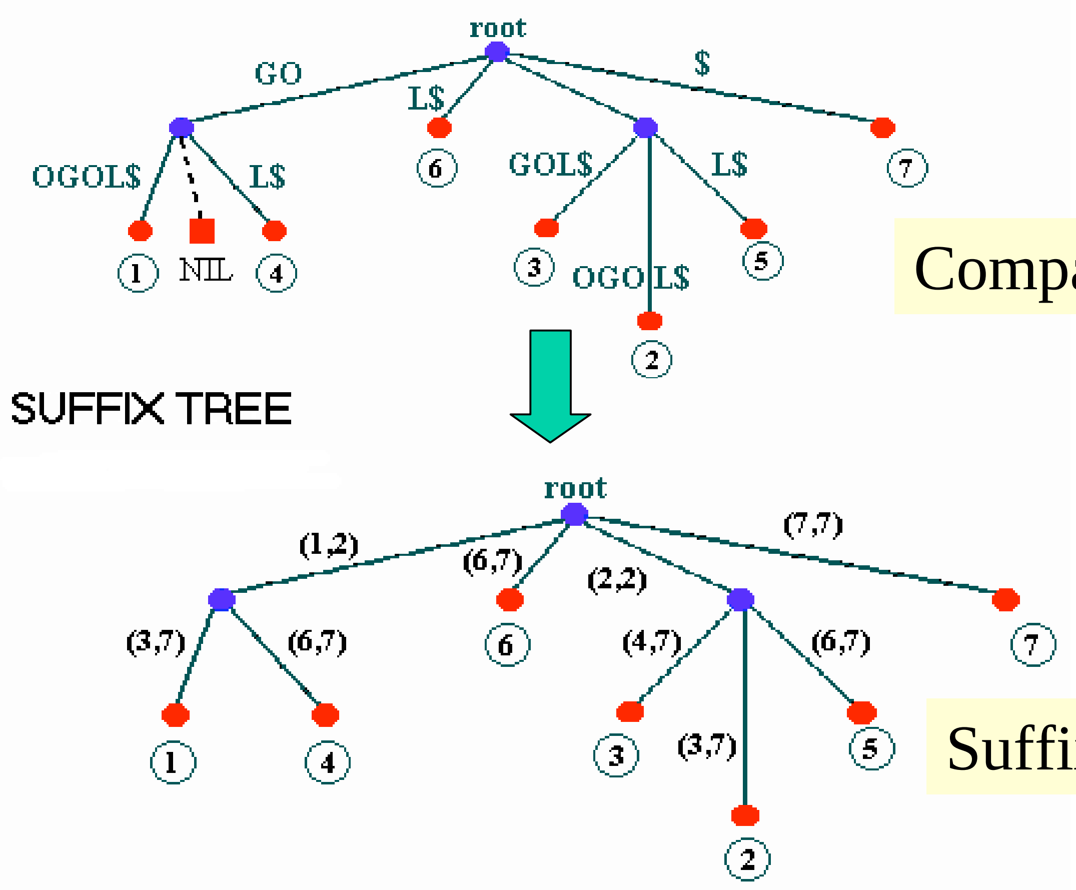
■ Inactive node, one for each symbol of the alphabet not associated with any string

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

Compact Suffix Trie

Suffix Tries to Suffix Trees

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



Compact Suffix Trie

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

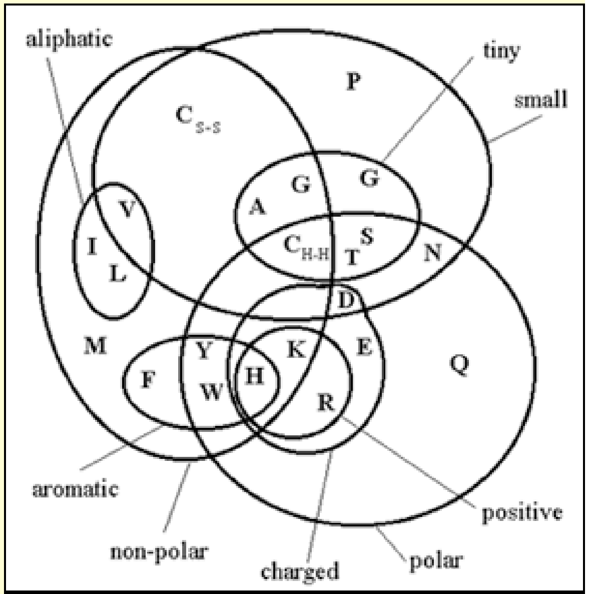


Figure 1. A Venn diagram showing the relationship of the 20 naturally occurring amino acids to a selection of physio-chemical properties thought to be important in the determination of protein structure.

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