

Doing the Time Warp

Sequence Comparison
and its Applications

Transform Brainstorm

TREE

FREE

FRET

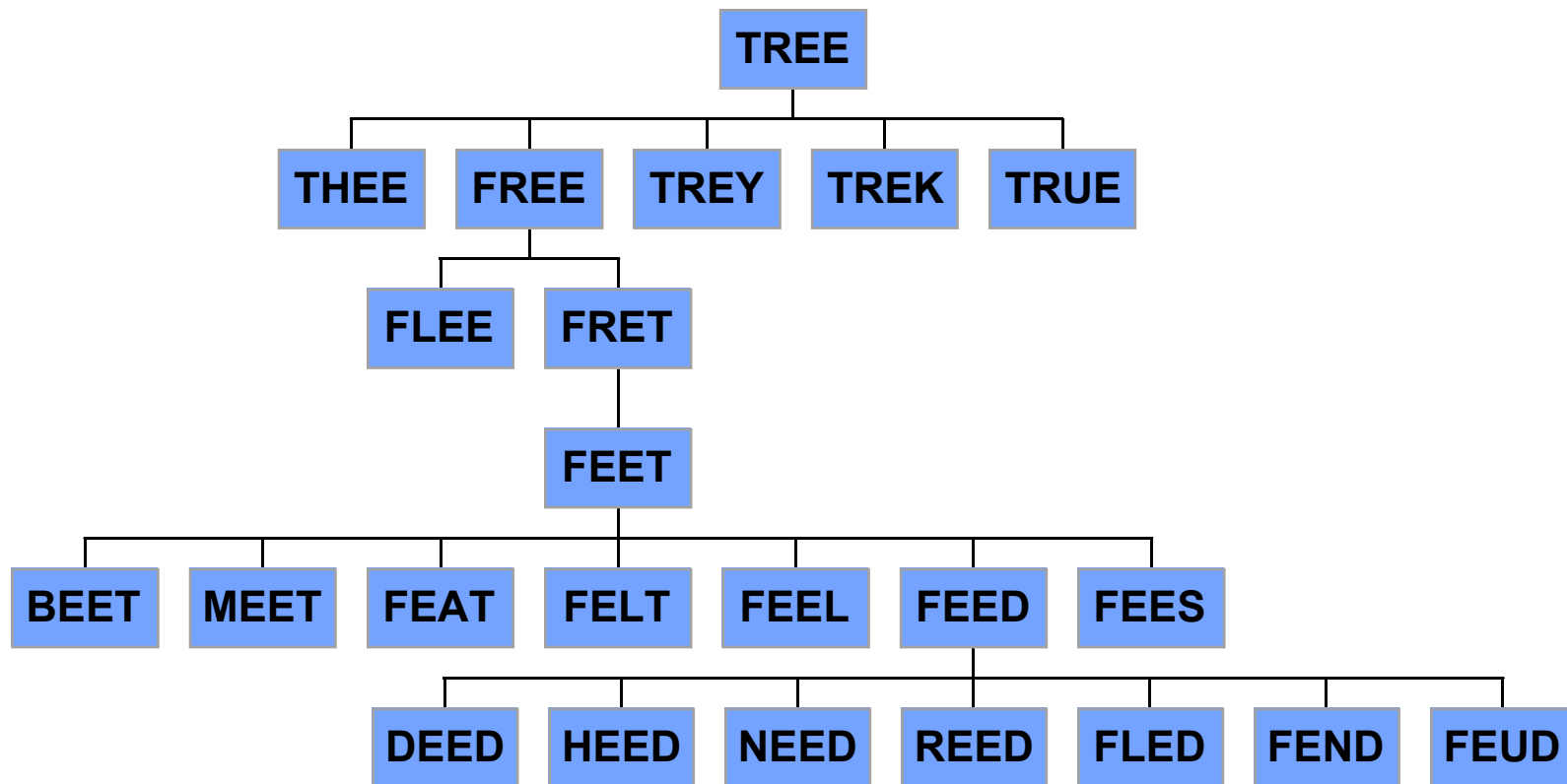
FEET

FEED

REED

“Linguistically constrained” edit distance ≤ 5

A Linguistic Tree



Sequences, Strings and Such

- Word: An element of a language
 - TREE
- String: A contiguous ordered list of characters from an alphabet
 - ACBADDCABD
- Substring: A contiguous part of a string
 - DCA
- Sequence: Same as string
- Subsequence: Ordered part of a sequence, not necessarily contiguous
 - BDA

Sequence Alignment

- Use **gaps** to help align **matching** characters
- Compute similarity between words A and B

$$1, \quad a = b$$

$$d(a, b) = \mu, \quad a \neq b$$

$$\delta, \quad a = "--" \text{ or } b = "--"$$

$$D(A, B) = \max_{A^*, B^*} \sum_{i=1}^n d(A_i^*, B_i^*)$$

Example

TREE
REED

$$1 + 3\mu$$

TREE -
- REED

$$3 + 2\delta$$

T - REE
REE - D

$$3\mu + 2\delta$$

TRE - E
- REED

$$2 + \mu + 2\delta$$

Applications

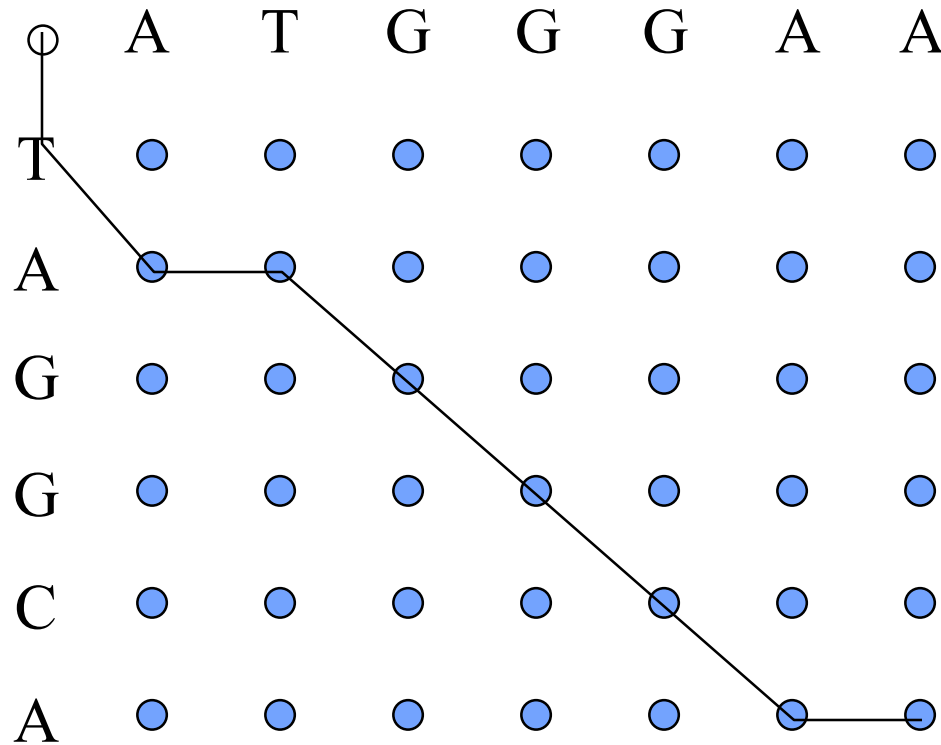
- Molecular biology
 - protein structure
 - gene function
 - drug development



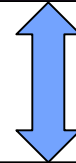
Applications

- Signal processing
 - speech recognition
 - ornithology
 - robot vision

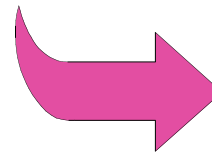
Alignment Paths



Alignment of
two sequences



Path through
matrix



--ATGGGAA
TA--GGCA--

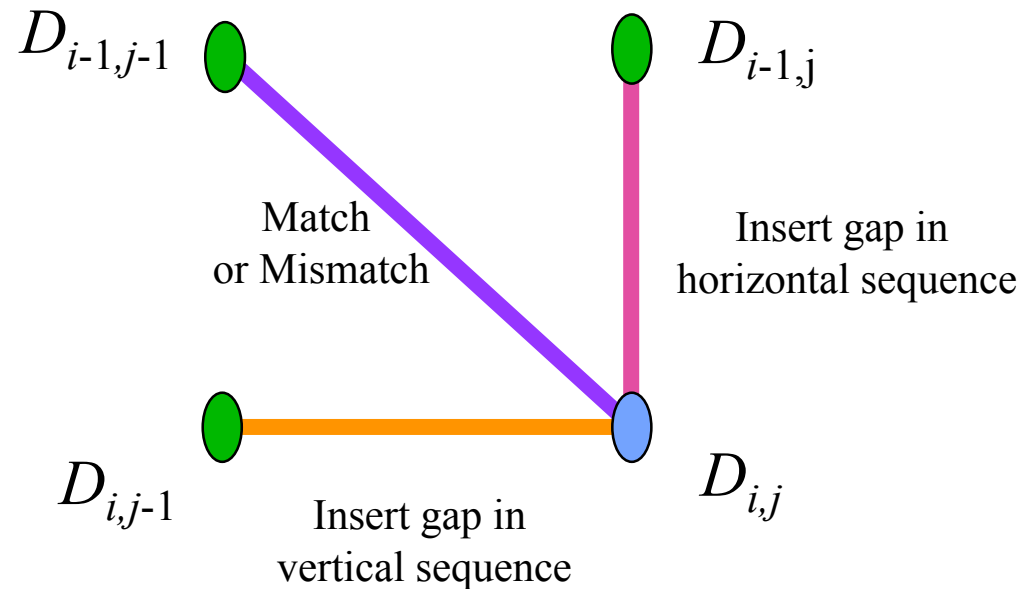
Dynamic Programming

- Bellman's principle: the optimal path from $(1,1)$ to (m,n) *through* the point (i,j) is the concatenation of the optimal path from $(1,1)$ *to* (i,j) with the optimal path *from* (i,j) to (m,n)

$$(i_0, j_0) \xrightarrow{(i,j)} (i_m, j_n) = (i_0, j_0) \longrightarrow (i, j) \oplus (i, j) \longrightarrow (i_m, j_n)$$

Computing the Cost Function

$$D_{i,j} = \max \{ D_{i-1,j} + d(A_i, -), D_{i,j-1} + d(-, B_j), D_{i-1,j-1} + d(A_i, B_j) \}$$



Example

Initialization:

$$D_{1,j} = \delta * j$$

$$D_{i,1} = \delta * i$$

$\mu = -3$ $\delta = -4$

	--	A	C	G	G	C	U	C
--	0	-4	-8	-12	-16	-20	-24	-28
A	-4	1	-3	-7	-11	-15	-19	-23
U	-8	-3	-2	-6	-10	-14	-14	-18
G	-12	-7	-6	-1	-5	-9	-13	-17
G	-16	-11	-10	-5	0	-4	-8	-12
C	-20	-15	-10	-9	-4	1	-3	-7
C	-24	-19	-14	-13	-8	-3	-2	-2
U	-28	-23	-18	-17	-12	-7	-2	-5
C	-32	-27	-22	-21	-16	-11	-6	-1

AUGGCCUC
ACGGC--UC

Related Topics

- Multiple alignment
- Other types of comparisons
 - local
 - structural
 - Multidimensional
- Approximate alignment (BLAST)
- Distribution of alignment scores
 - What is distribution of scores when random sequences are compared?