String Matching

String matching: definition of the problem (text,pattern)

- Exact matching: depends on what we have: text or patterns
  - The patterns ---> Data structures for the patterns
    - 1 pattern ---> The algorithm depends on |p| and |Σ|
    - k patterns ---> The algorithm depends on k, |p| and |Σ|
  - Extensions
  - Regular Expressions
  - The text ----> Data structure for the text (suffix tree, ...)
- Approximate matching:
  - Dynamic programming
  - Sequence alignment (pairwise and multiple)
  - Sequence assembly: hash algorithm
- Probabilistic search: Hidden Markov Models

Pairwise and multiple alignment

Pairwise alignment

Edit distance:

match=0 mismatch=1 indel=1

$$d(AC,CTACT)=\text{minimum}\begin{cases} d(A,CTAC)+1 \\ d(A,CTA)\ldots.+1 \\ d(AC,CTA)+1 \end{cases}$$

Similarity:

match=1 mismatch=-1 indel=-2

$$s(AC,CTACT)=\text{maximum}\begin{cases} s(A,CTAC)-2 \\ s(A,CTA) \pm 1 \\ s(AC,CTA)-2 \end{cases}$$

Pairwise alignment

Connect to

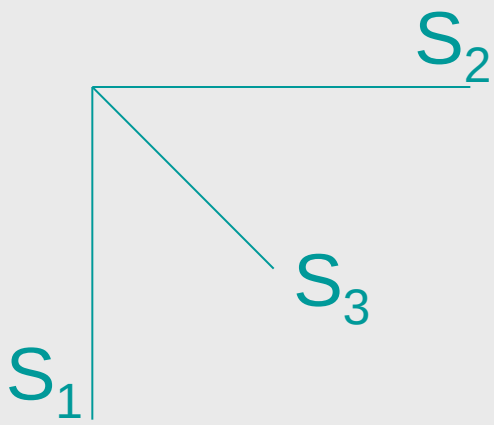
http://alggen.lsi.upc.es

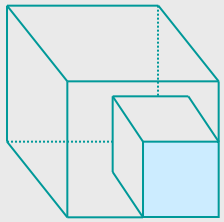
Links to TEACHING EMBER LePA

Pairwise to multiple alignment

What happens with three strings?

Let n be their lenght, then the cost becomes





A
C
A

-1

$O(n^3)$ “$O(2^3)$” “$O(3^2)$”

And with k strings? $O(n^k\, 2^k\, k^2)$

Multiple alignment

algorithmics genetics group

Programs of multialignment use different heuristics:

- Clustal (Progressive alignment)
  http://www.ebi.ac.uk/clustalw
- TCoffee (Progressive alignment + data bases)
  http://igs-server.cnrs-mrs.fr/Tcoffee_cgi/index.cgi
- HMM (Hidden Markov Models)

Multiple alignment

Connect to

http://alggen.lsi.upc.es/

and follow the links TEACHING EMBER.