Sequence Alignment

iCB2 – Introduction to Computational Biology and Bioinformatics

November 13, 2015

Emidio Capriotti

http://biofold.org/

Biomolecules Folding and Disease

Institute for Mathematical Modeling
of Biological Systems
Department of Biology

HEINRICH HEINE
UNIVERSITÄT DÜSSELDORF

Sequence Analysis

Sequence Analysis is one the first research field in Bioinformatics and Computational biology which consists in processing DNA, RNA or protein sequence through with wide range of analytical methods to understand characterize the function, structure, or evolution.

The main methodology for sequence analysis consist in the comparison of sequences performed using

1. Pattern matching
2. Dot Plot
3. Sequence alignment

from wikipedia

Sequence Alignment

The sequence alignment is a method of sequence comparison based detection of similarity between 2 or multiple biomolecules.

When sequences are small and similar enough the alignment can be executed manually, in more complex cases an optimization procedure is needed to find the best alignment.

Alignment can be either Pairwise or Multiple and Global or Local

The main issues

How to find the optimal alignment that maximizes the matching between sequences?

How the score is defined. Are gaps allowed?

What is the best algorithm?

How the result is evaluated?

A simple score

A basic alignment score consists in assigning 1 to matching residues or nucleotides otherwise -1

```
AGATCAGAAATG
--AT-AG-AAT-
--:-:-:-:-:-
```

```
AGATCAGAAATG
--AT-AGAA-T-
--:-:-:::-:-
```

The optimization

Given two sequences of length *m* and *n* a brute force algorithm that calculates all the possible pairwise alignment will have the exponential complexity

$$\binom{n+m}{m} = \frac{(m+n)!}{n!m!}$$

"*Divide and conquer*" approach allow to reduce the computational complexity of the problem.

The basic assumption: the optimal alignment between two strings should include the optimal alignment of the smaller prefixes.

Divide and Conquer

Given two sequence *P* and *Q*, the knowing the optimal alignment between the prefix $P_{1,i-1}$ and $Q_{1,j-1}$ the best alignment between the positions *i* in *P* and *j* in *Q* is the optimal alignment between positions prefixes *i-1,j-1* or *i-1,j* or *i,j-1*.

```
          i
P:  AGATCAGAAATG
Q:  ATAGAAT
          j
```

```
          i
AGATCAG
--AT-AG
      j
```

```
          i
AGATCAG-
--AT-A-G
      j
```

```
          i
AGATCA-G
--AT-AG-
      j
```

Examples from CAP5937: Introduction to Bioinformatics: http://people.cs.missouri.edu/~chengji/

The scoring matrix

More sophisticated methods to score the mis-matches in protein sequence alignment based on the observation of mutation rates in curated alignments.

Most famous are the PAM (Point accept mutations) and BLOSUM

There are numbers associated to each matrix.
PAM higher the number high the divergence between the sequence included for the calculation of the mutation rates. (PAM250 most divergent).

For BLOSUM is the opposite higher the number higher the similarity

$$sim(i,j) = \log \frac{P(i,j)}{P(i)P(j)}$$

BLAST Algorithm

The first revolution in the bioinformatics was the development of the Basic Local Alignment Search Tool (BLAST).

The most recent paper *Altschul SF et al.* (1997) NAR PMID: 9254694 more than 41,000 citations in SCOPUS.

BLAST allows to compare large dataset of sequences in short amount time with respect to standard alignment programs.

BLAST implements a heuristic method, to finds similar sequences by locating short matches between the two sequences.

This process of finding initial words is called seeding. The alignment of words by default of 3 letters is used to calculate a local alignment

How to run BLAST

First step consist in creating a database in appropriate format using
formatdb -i database

One example

blastpgp -i fastafile -d database -o output.txt

```
BLASTP 2.2.18 [Mar-02-2008]


Query= sp|P04637|P53_HUMAN Cellular tumor antigen p53 OS=Homo sapiens
GN=TP53 PE=1 SV=4
         (393 letters)

Database: uniprot_sprot.fasta
           547,599 sequences; 195,014,757 total letters

Searching..................................................done
                                                           Score    E
Sequences producing significant alignments:                 (bits) Value

sp|P04637|P53_HUMAN Cellular tumor antigen p53 OS=Homo sapiens G...   813   0.0
sp|P13481|P53_CHLAE Cellular tumor antigen p53 OS=Chlorocebus ae...   746   0.0
```