Bioinformatics

- **Comparative analysis of seven multiple protein sequence alignment servers:**

  clues to enhance reliabillity of predictions

Motivation

- Evaluate the reliability of seven multiple alignment servers currently available on the Internet in terms of
    - power(sensitivity)
    - confidence(selectivity).

Why multiple alignment

- the detection of common patterns in protein families
- suggesting primers for polymerase chain (PCR) of fragments of homologous genes
- understanding molecular evolution
- predicting secondary and tertiary structures

System and methods

- **Online server**: ClustalW, MAP,
  PIMA, Block Maker,
  MSA
- **Email server**: MEME, Match-Box

System and methods-source

- common core of each test family is defined as a set of SCRs, initiated by superimposing the backbone of the major elements of secondary structure of the less similar pair of sequence
- each SCR was extended as far as the root mean square(RMS) computed between
- α-carbons on the whole SCR remains<1.8A
- the other proteins of the family are progressively aligned and the common SCR limited to the set for which all the pairwise comparisons produce an RMS of < 1.8A.

System and methods–11 output

- PIMA:
  - PIMA_ML(maximum linkage)
  - PIMA-SB(sequential branching)
- Block Maker
  - Gibbs method
  - Gibbs method

System and methods-11 output

- MatchBox
  - Reliability <=4, MB1
  - Reliability <=5, MB2
  - Reliability <=9, MB3
- MEME
- ClustalW
- MAP
- MSA

Definition and formula

- **pSCRs**: predicted structurally conserved regions are defined as the segments aligned in all the sequences and not disrupted by gaps.
- **SCR**: structurally conserved regions.
- **S**: cumulated length of the SCRs.
- **s**: the cumulated length of the pSCRs.
- **I**: the cumulated length correctly predicted.

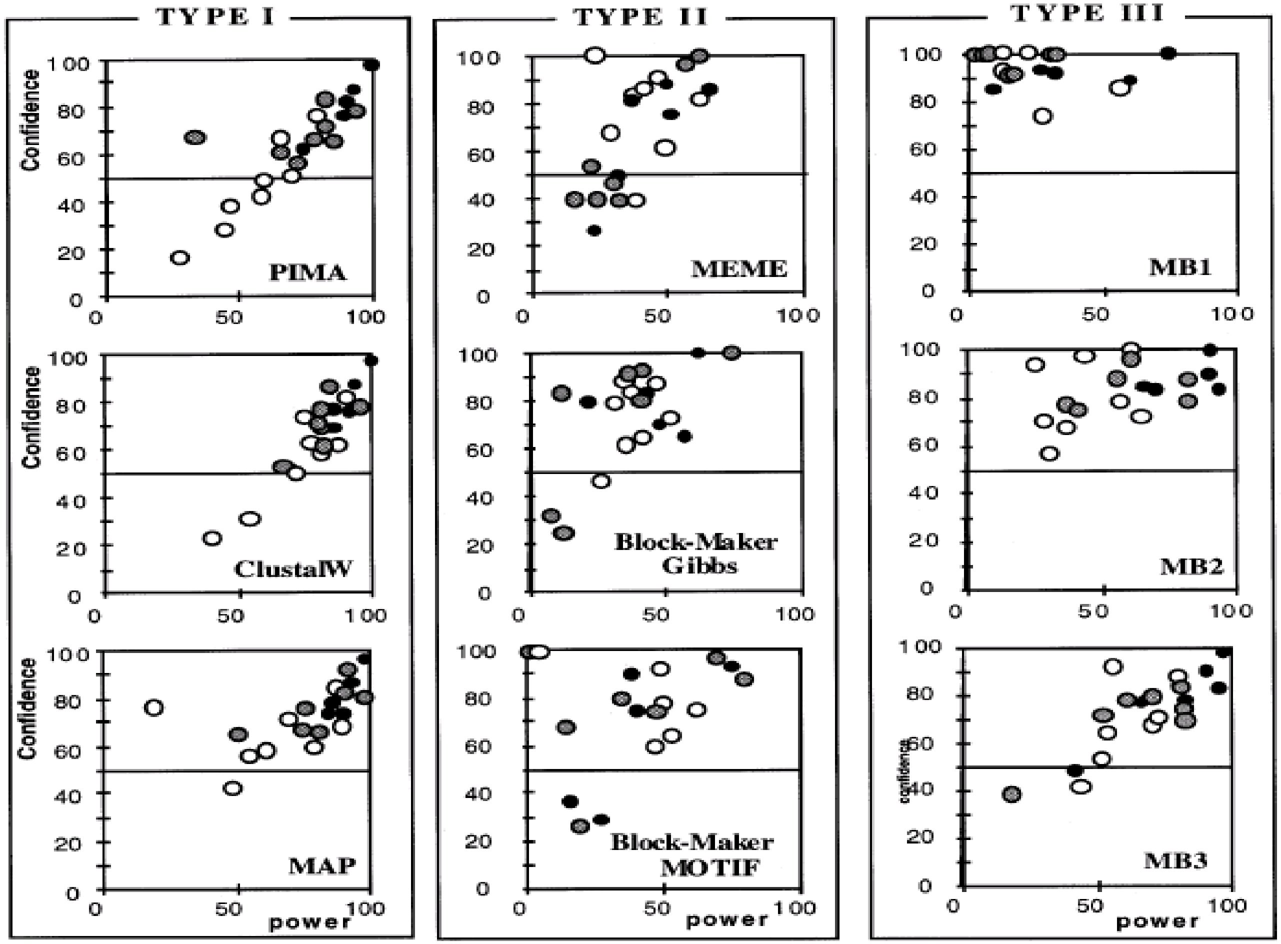
Definition and formula

- The performances of a given method applied on a given family are evaluated by the following relationshiops:

$$\text{Power} = I/S * 100$$

$$\text{Confidence}=I/s * 100$$

Result


TYPE I
TYPE II
TYPE III
Confidence
power
0
20
40
60
80
100
50
PIMA
MEME
MB1
ClustalW
Block-Maker
Gibbs
MB2
MAP
Block-Maker
MOTIF
MB3
confidence

Result

- Identity:
  - black >20%
  - grey 10-20%
  - white <10%

Result

- Type I methods:
  - For ClustalW, MAP, PIMA
  - Power and confidence are linear relationship
  - Best overall rate on Power

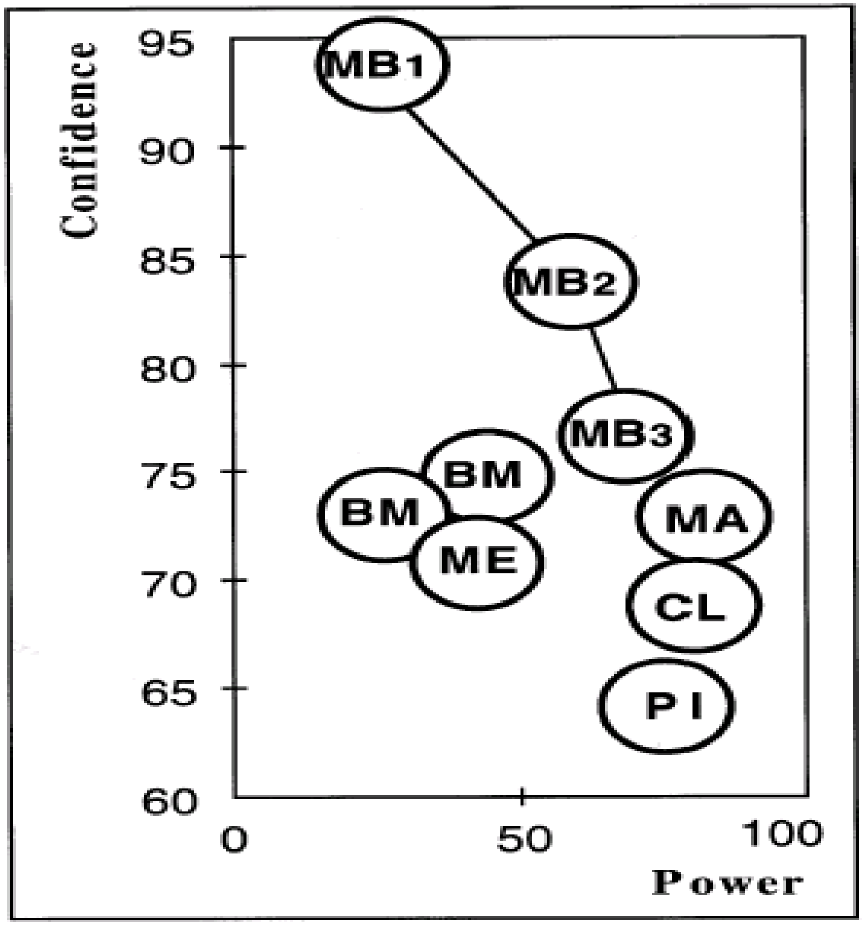
Result

- Type II methods
  - For Block Maker and MEME
  - Power is low
  - Not clearly related to the rate of identity
  - Priori unpredictable

Result

- Type III methods:
  - For Match-Box 1,2,3
  - Confidence is high
  - Low performance in low reliability score

Result


Confidence
95
90
85
80
75
70
65
60
0
50
100
Power
MB1
MB2
MB3
BM
BM
ME
MA
CL
P I

Result

- Type I : large power with low confidence
- Type II : hybrid situation between I, II
- Type III :large confidence variable power

Question?