CLIMP

Cluster-based Imputation of Missing Values in Microarray Data

Nils Gehlenborg · December 2003
gehlenbo@cs.washington.edu

Outline

1. Motivation
2. Algorithm
   - key idea
   - a bit more detail
3. Other approaches
4. Results
5. Discussion
6. Conclusion

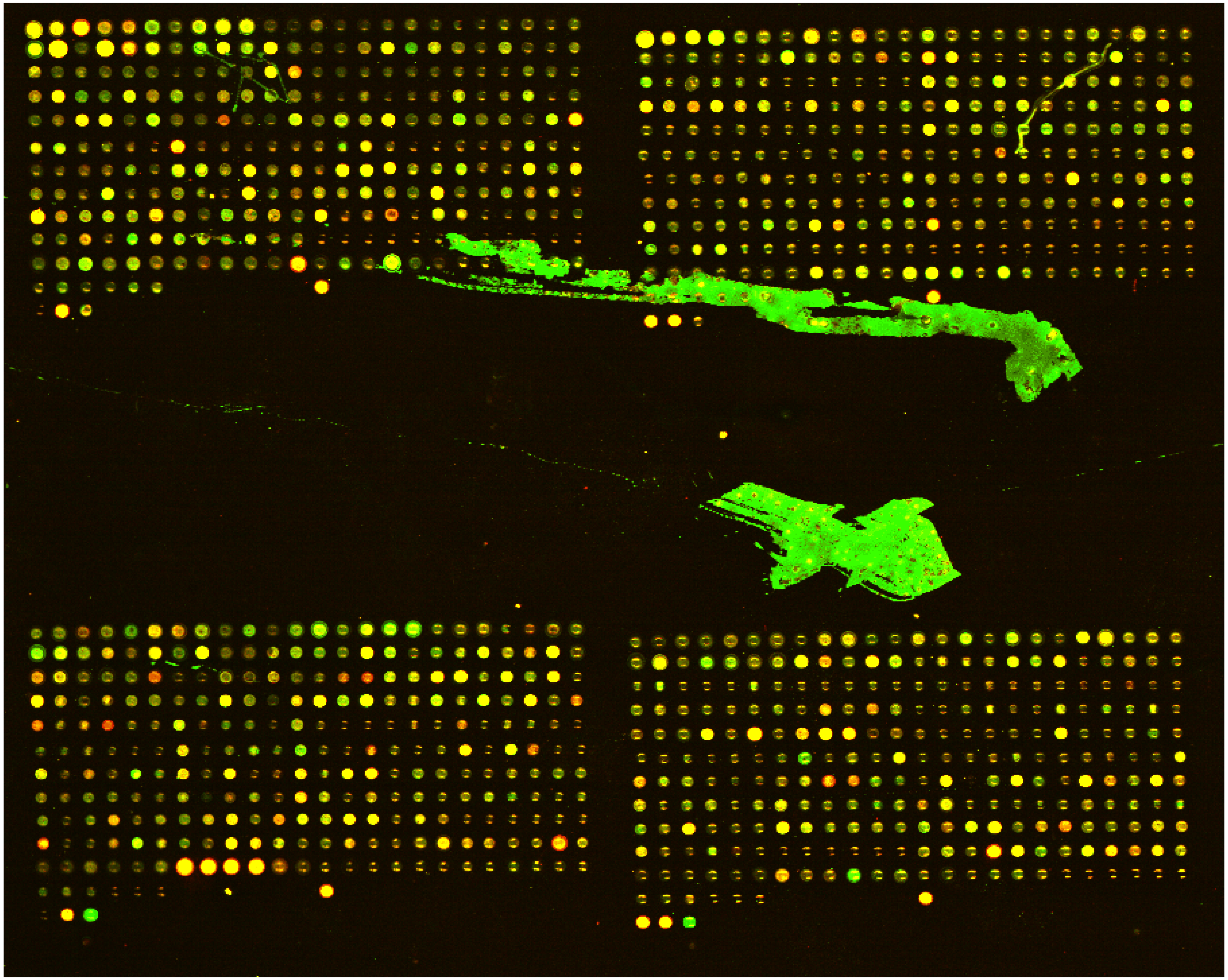
Motivation

- Missing values cause a lot of trouble.
  - similarity/dissimilary measures
  - principal component analysis (PCA)
  - SVMs
  - clustering
- Missing values are inconvenient.
- There is an expensive solution.
  - repeat experiments → more complexity and not perfect
- There are cheap (destructive) solutions.
  - casewise deletion → possibly no valid cases
  - pairwise deletion → genes become more similar

Reasons for missing values

- Arbitrarily missing values.
  - no spot intensity measured
  - negative background corrected spot intensity
  - array handling
  - "low quality spot" (cDNA arrays image analysis)
  - ...
- Systematically missing values.
  - array production
  - ...

Example

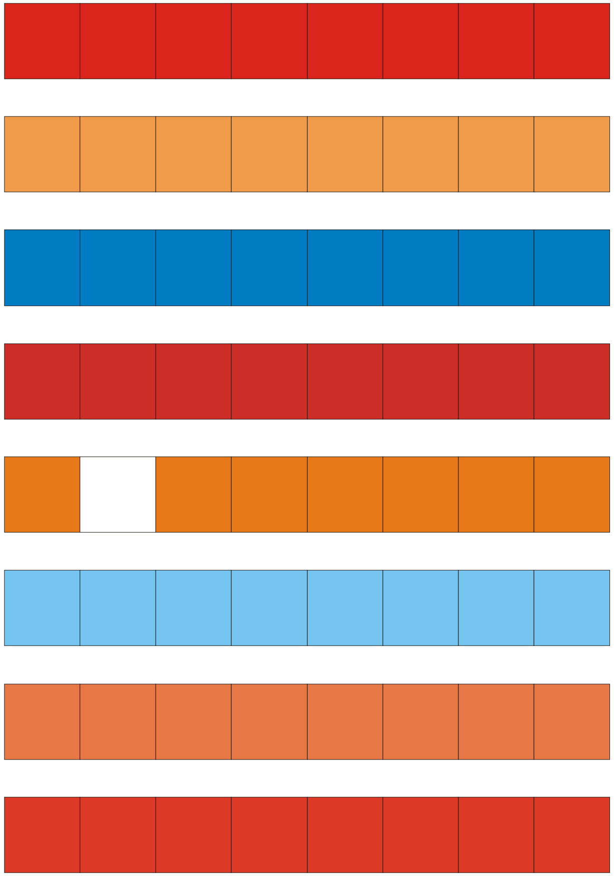


(edited from Stanford Microarray Database)

Starting points

- Image(s) of scanned microarray.
  - find reasons for missing values
  - identification of systematic errors
  - extremely complex to analyze
- Annotated image analysis output.
  - identification of systematic errors
  - different for different types of microarrays
- Expression matrix.
  - least information, but most general
  - probably most wide-spread format

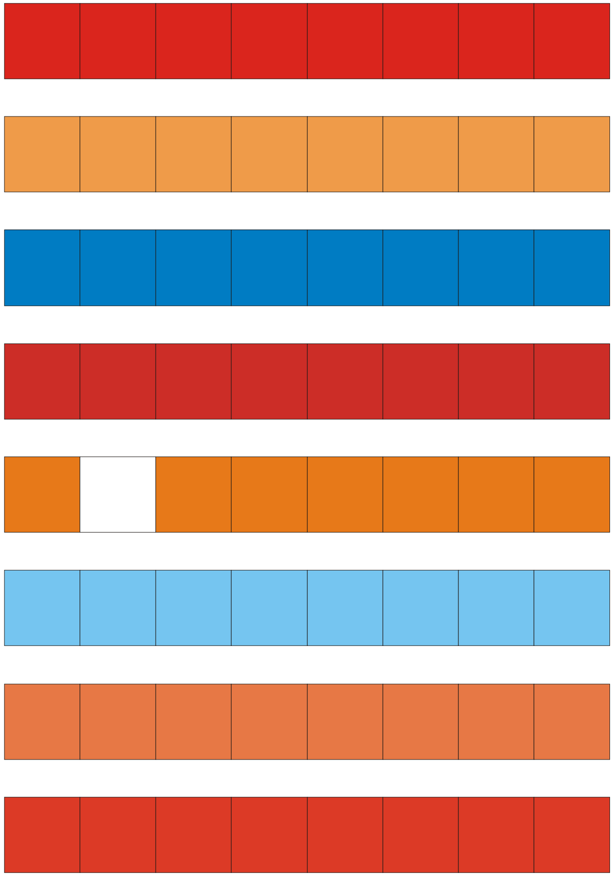
Problem



columns = conditions
rows = genes
color = expression profile

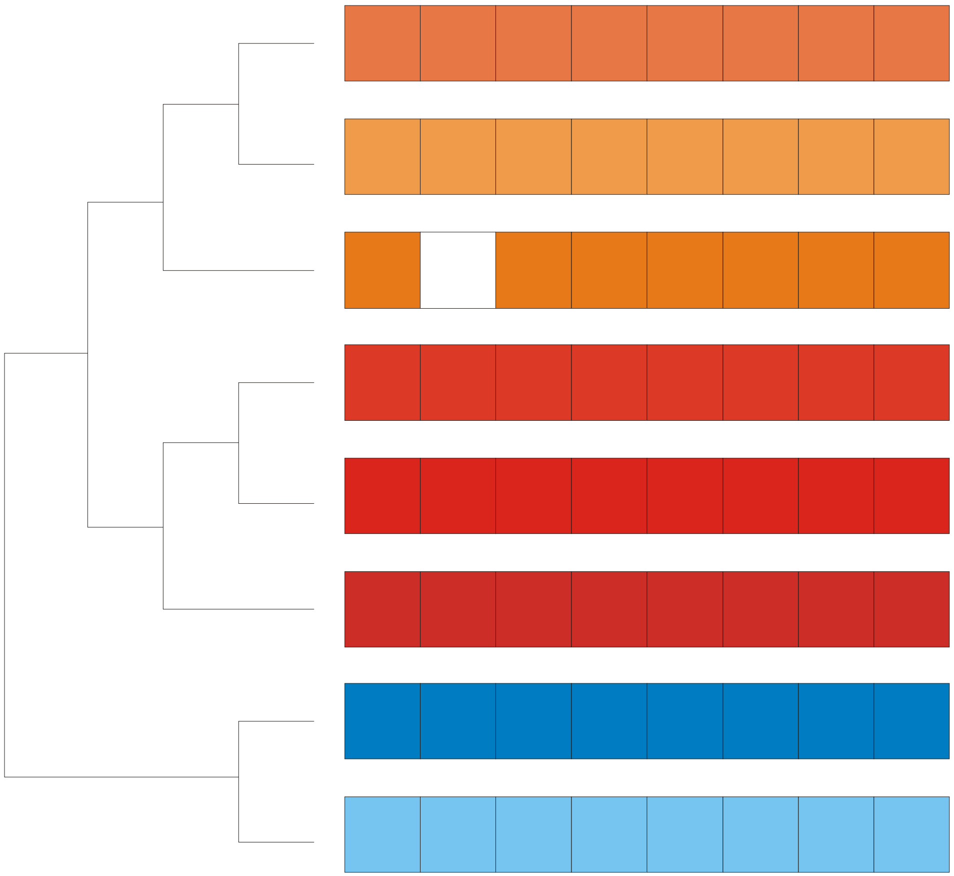
- Given an expression matrix with missing values, how do we estimate (impute) the missing values?

Idea



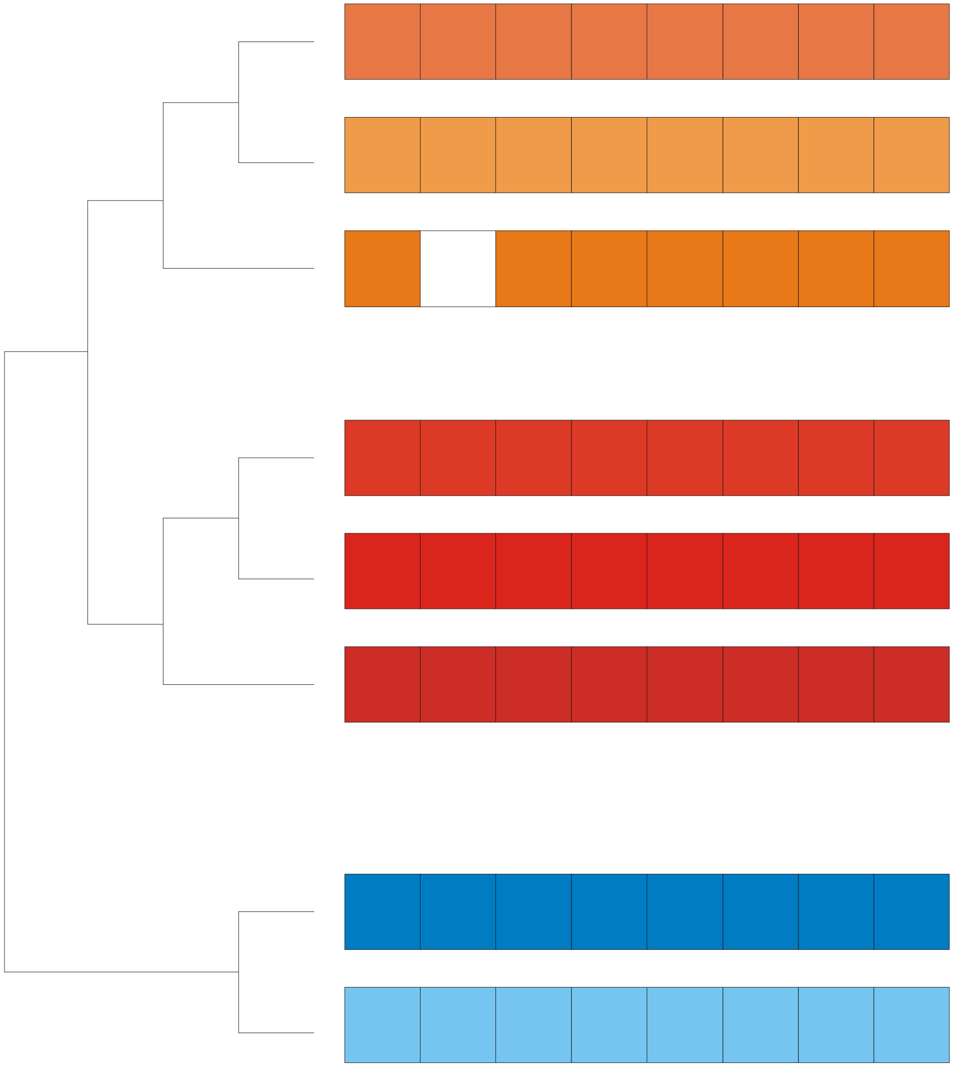
- Estimate missing values from similar genes, taking into account the correlation structure.
- How do we find similar genes?

Clustering



- How many clusters are there?
- Define an upper bound for cluster size!

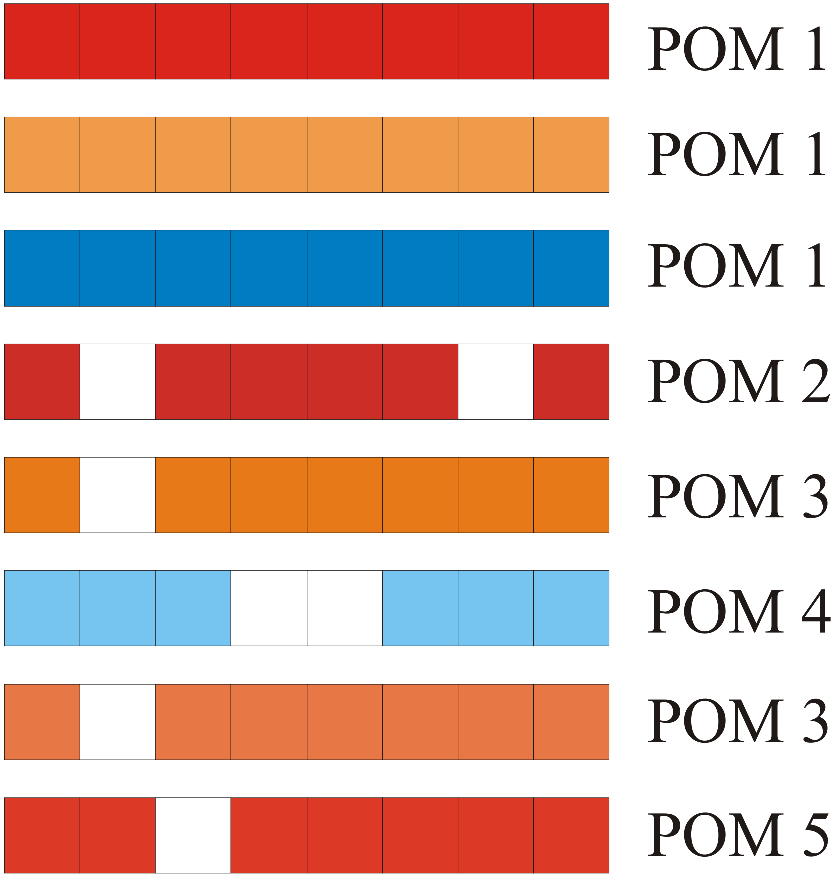
Estimation



maximal cluster size = 5

- Use genes in cluster for estimation.

Details



POM 1

POM 1

POM 1

POM 2

POM 3

POM 4

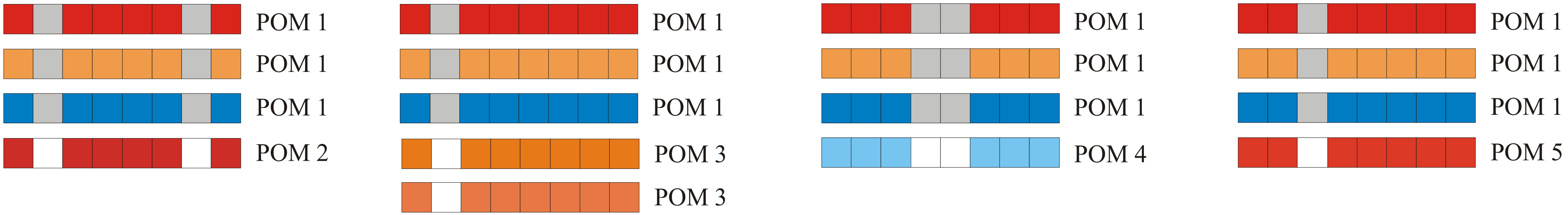
POM 3

POM 5

- Clustering for each *pattern of missingness* (POM).
  - POM = pattern of missing values in a row = a set of columns
  - length of a POM = cardinality of set of columns

Details

- *Base matrix* = all rows with POM of length 0 (here: POM 1).
- Cluster base matrix with all rows have the same POM.
  - leave out missing conditions
  - use hierarchical clustering with complete-linkage for dense clusters



- Compute missing value as rank-weighted average from base matrix genes in corresponding cluster.
- Cluster size below threshold?
  - use *k* nearest neighbors

Other (constructive) methods

- Simple methods
  - fill in zeros
  - fill in column- or row-averages
- Troyanskaya *et al.* 2001
  - *k* nearest neighbors (KNN)
  - singular value decompostion (SVD)
- Oba *et al.* 2003
  - Bayesian Principal Component Analysis (BPCA)
- Zhou *et al.* 2003
  - (non)-linear regression with Bayesian gene selection

Evaluation

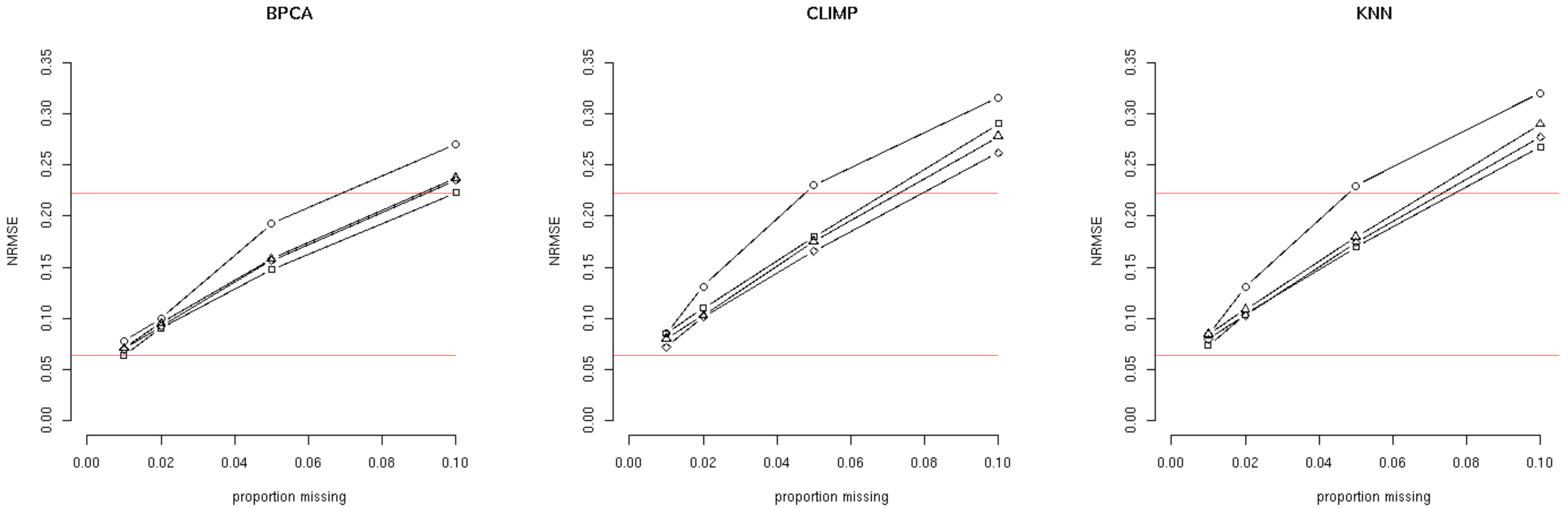
- Comparison of CLIMP, KNN and BPCA.
- Data sets:
  - Spellman *et al.* 1998, yeast cell cycle α-factor- and *cdc15*-based synchronization (18 and 15 conditions)
- Parameters to be chosen:
  - upper and lower bound (here: 35 and 20)
  - *k* (here: 17)
  - clustering algorithm (here: complete-linkage)
  - distance measure (here: Euclidean)
- Amount of missing data:
  - 1%, 2%, 5%, 10%

Evaluation

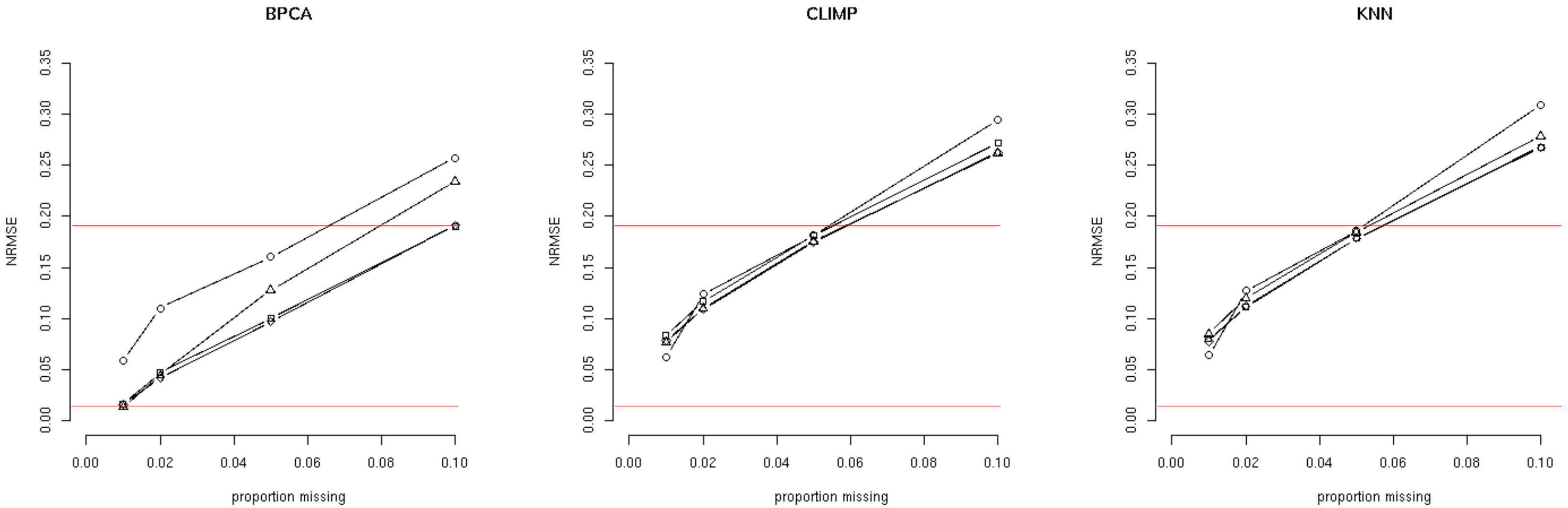
- Different number of genes from each test set: 100, 500, 1000 and 2000 out of ~ 6100.
- Performance evaluated by the normalized root mean squared error (NRMSE) of the estimated matrix (*E*) vs. the original matrix (*O*).
  - $NRMSE = \sqrt{\frac{mean(O-E)^2}{variance(O)}}$
  - if *NRMSE* close to 0, then *E* more accurate (*NRMSE* = 0 → *E* = *O)*
  - if *NRMSE* close to 1, then *E* less accurate

NRMSE on test data

α-factor

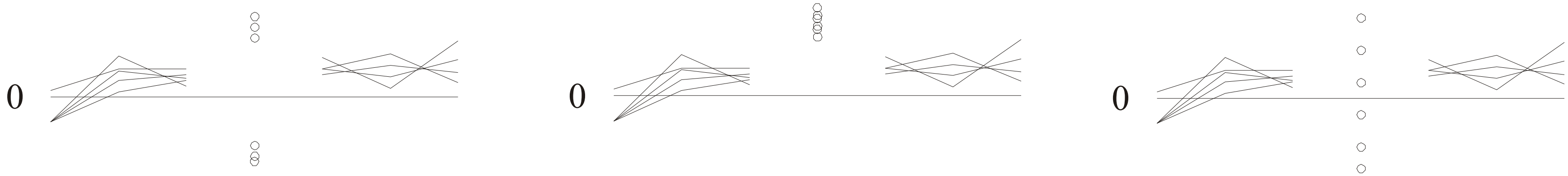


cdc15



Discussion

- CLIMP has some weak spots.
  - base matrix
  - how to find good values for parameters (→ usage of KNN)
  - runtime
- Performance might be increased in several ways.
  - genes with estimated missing values might be added to base matrix
  - analysis of values used for estimation



  - base weighted average on distance not on ranked distance
  - selection of parameters appropriate for given expression matrix

Conclusion

- The bigger the base matrix, the more information, the better the results.
- CLIMP is slower than KNN and BPCA, but time is not an important criterion in missing value estimation.
- Performance of CLIMP is at least equal to that of KNN and might be improved.
- Bayesian methods are likely to remain significantly better.

Handle estimated values with care,

they still might be completely wrong!