АЛГОРИТМИ ЗА ОЦЕНЯВАНЕ НА ЛИПСВАЩИ СТОЙНОСТИ В МАТРИЦИ ЗА ИЗРАЗЯВАНЕТО НА ГЕНИ

Елена Костадинова

НАУЧЕН СЕМИНАР

на катедра КСТ и групата по Биоинформатика

23 октомври 2008 г.

Missing Values in Microarray Experiments

Reasons

Insufficient resolution

Image corruption

Dust or scratches on the slides

Experimental errors

Solutions

Repeat the experiment

Replace the missing value by zero

Replace the missing value by an average expression over the row

Methods considering the correlation structure of the genes in the data

Some Existing Imputation Methods

Singular **V**alue **D**ecomposition (SVD)

K-**N**earest **N**eighbour Method (KNN)

Sequential **KNN** Imputation (SKNN)

Bayesian **P**rincipal **C**omponent **A**nalysis (BPCA)

Local **L**east **S**quare Techniques (LLS)

Projection **O**nto **C**onvex **S**ets (POCS)

***Fixed** number of neighbours*

Different Solutions to Missing Value Problem

Varying number of candidate estimation profiles

Iterated **LLS** Imputation Method (ILLSimpute)

Adaptive **M**ultiple Imputation (AMimpute)

Dynamic **T**ime **W**arping Based Imputation Algorithms (DTWimpute)

Multiple missing value estimations

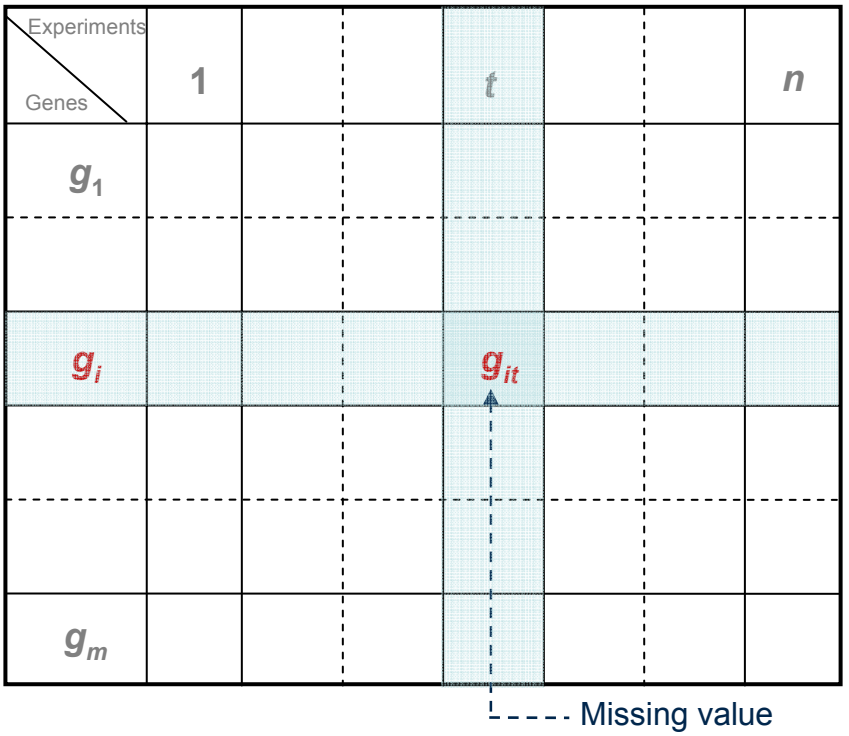
Multiple **L**east **S**quares based Imputation Method

Adaptive **M**ultiple Imputation (AMimpute)

Multiple **KNN** Imputation Algorithm (MI-KNNimpute)

K-Nearest Neighbours (KNN) Imputation Algorithm

Weighted K-Nearest Neighbours



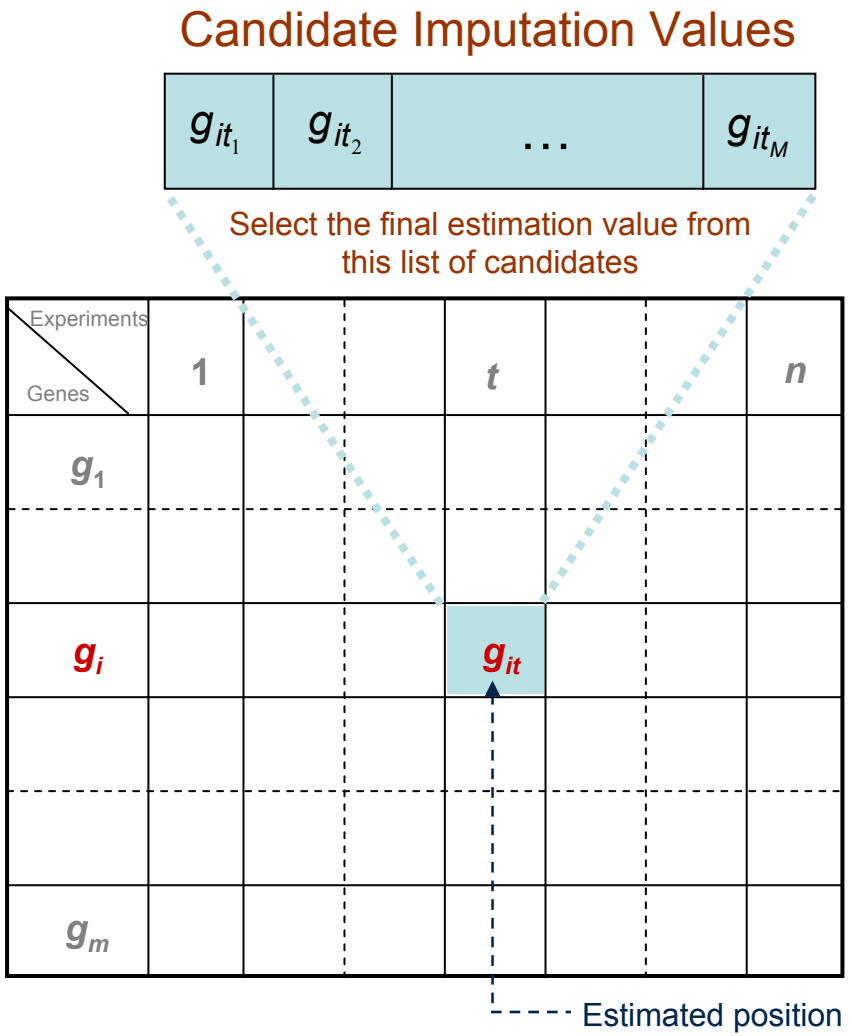
- Consider gene g_i with a missing value at experiment t;
- Calculate Euclidean distance to each gene g_j, which has a value at position t;
- Rank the genes in increasing order of their Euclidean distances;
- Select the first K genes:
- **Estimate**: $$g_{it} = \sum_{j=j_1}^{j_K} w_j g_{jt}$$

g_{jt}: value at position t from gene g_j

w_j : weight of gene g_j based on its Euclidean distance.

Troyanskaya et al. (2001). Missing value estimation methods for DNA microarrays. *Bioinformatics*, 17, 520-525.

Multiple KNN Imputation Algorithm (MI-KNNimpute)



- Consider gene g_i with a missing value at experiment t and a set of M positive integers
$$\{K_p \mid p=1,\ldots, M\};$$
- Construct a set of M gene estimation lists, one for each integer K_p;
- Generate a list of M candidate imputation values $\{g_{it_p} \mid p=1,\ldots, M\}$, one for each integer K_p, by calculating
$$g_{it_p} = \sum_{j=1}^{k_p} w_{k_j} g_{k_j t};$$
- Select a value from $\{g_{it_p} \mid p=1,\ldots, M\}$ to fill in the missing value.

Multiple **KNN** Imputation Algorithm (MI-KNNimpute)

***Three* estimates to impute the final missing value:**

1) ***the average of all candidate values***
2) ***the closest candidate value to the gene expression average***
3) ***the closest candidate value to the mean of the two non-missing neighbours of the missing value in the gene***

KNNimpute versus MI-KNNimpute

KNNimpute Algorithm

Preliminary fixed number of candidate genes for estimation

⇩

Only **one** candidate estimation value for each missing entry

⇩

Danger of using in the estimation **profiles rather distant** from the target gene

MI-KNNimpute Algorithm

Varying number of candidate genes for estimation

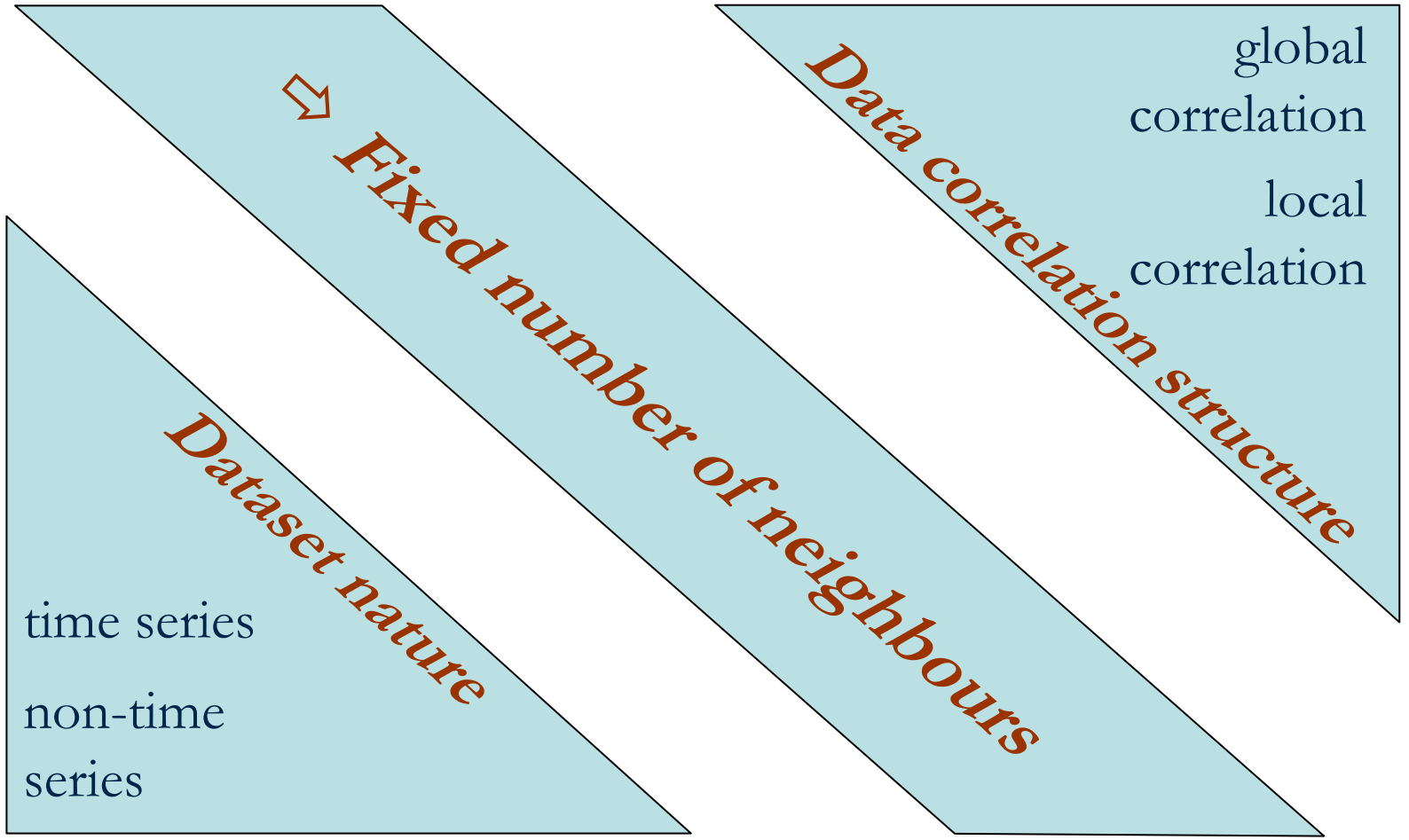
⇩

A **list** of candidate estimation value for each missing entry

⇩

Select the **most suitable** value according to some criteria

Existing Imputation Algorithms' Limitations



Hybrid **I**mputation **A**lgorithm (HYimpute)

Use a **set** of **different** imputation methods

Different aspects of the estimated data are taken into account during the imputation process

Find a trade-off between imputation methods' conflicting performances

HYimpute: the Best of Three Different Imputation Techniques

STEP I - *Construct a list of imputation values generated by three different imputation methods:*

1) *Row average*
2) *KNNimpute*
3) *The Average of the two non-missing neighbours of the missing value*

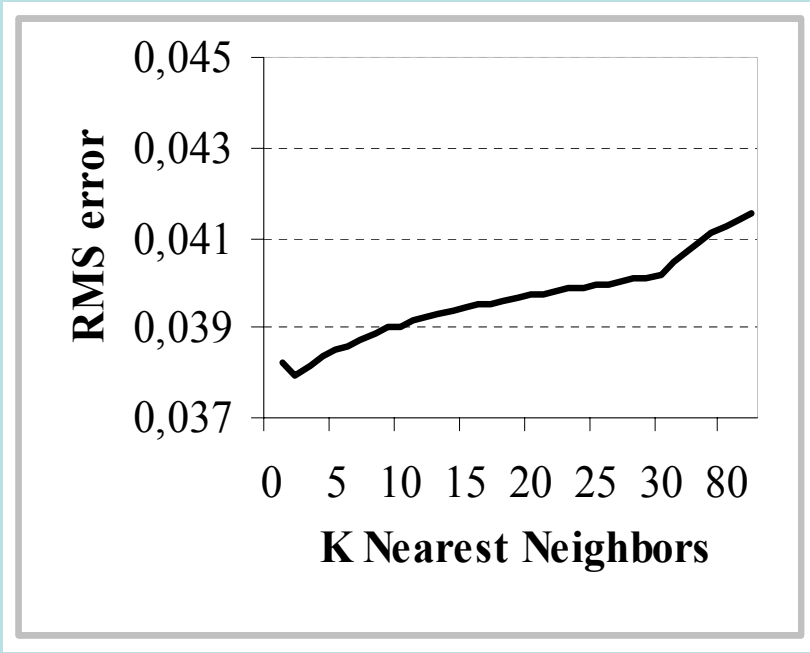
STEP II - *Calculate the final estimate:*

1) *Weighted average of all candidate imputation values*
2) *Selecting the most suitable value according to some well-defined criteria*

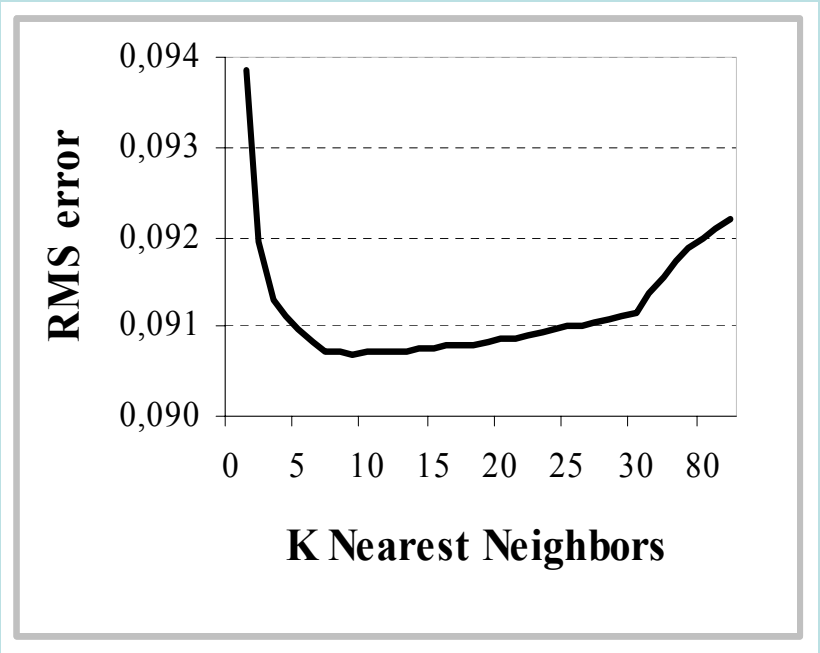
Experimental Setup

- **Test data corpus:** three time series datasets, a study of the global cell-cycle control in fission yeast *S. pombe* (Rustici *et al.*, 2004);
- **Pre-process each data set by removing rows containing missing values.**
- **Create test data sets by randomly deleting 1%, 5%, 10%, 15% and 20%.**
- Evaluate the accuracy of the estimation as the **R**oot **M**ean **S**quared (RMS) difference between the imputed matrix and the complete matrix, divided by the average value of the complete matrix.

HYimpute versus KNNimpute performance for a different number of candidate estimation genes

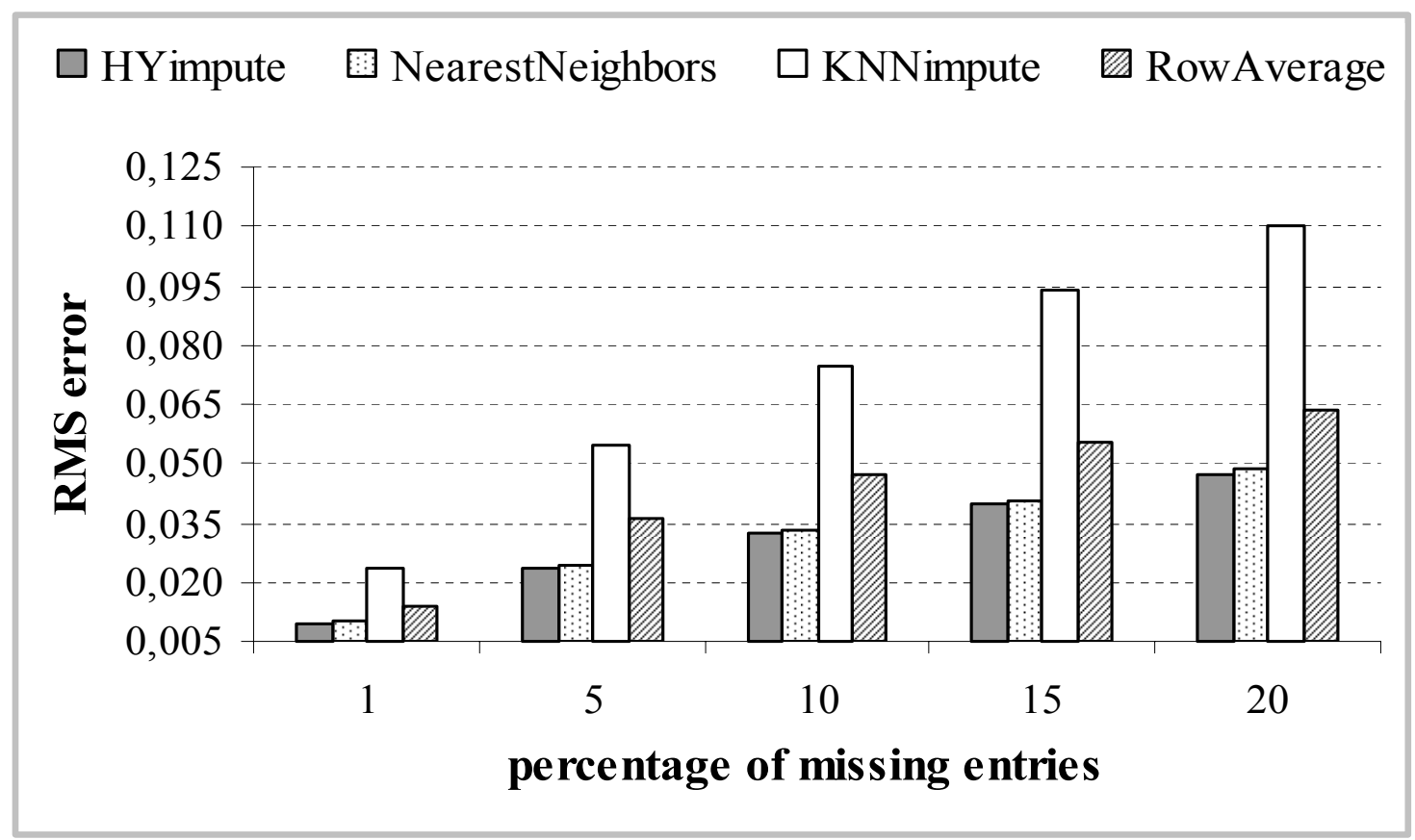


HYimpute

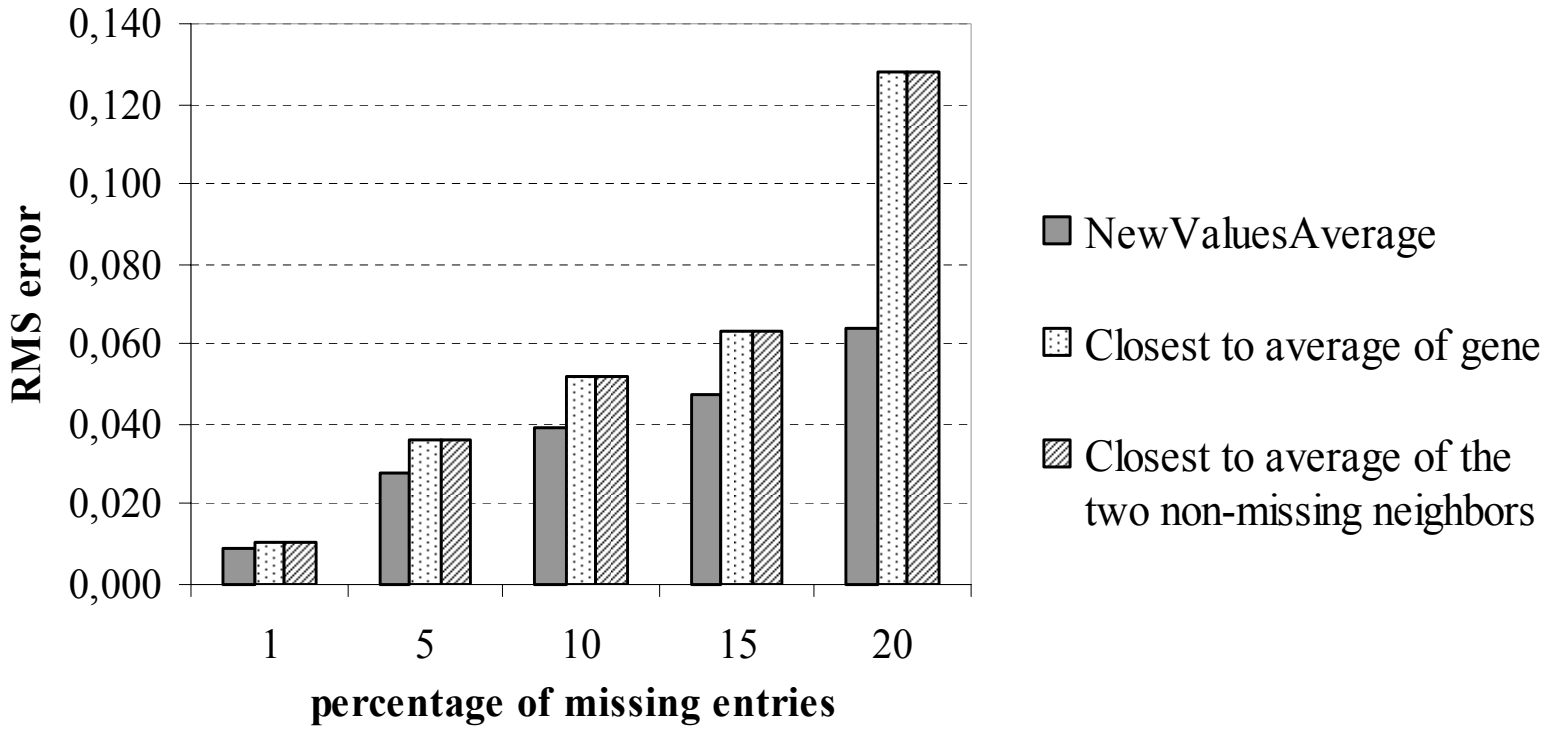


KNNimpute

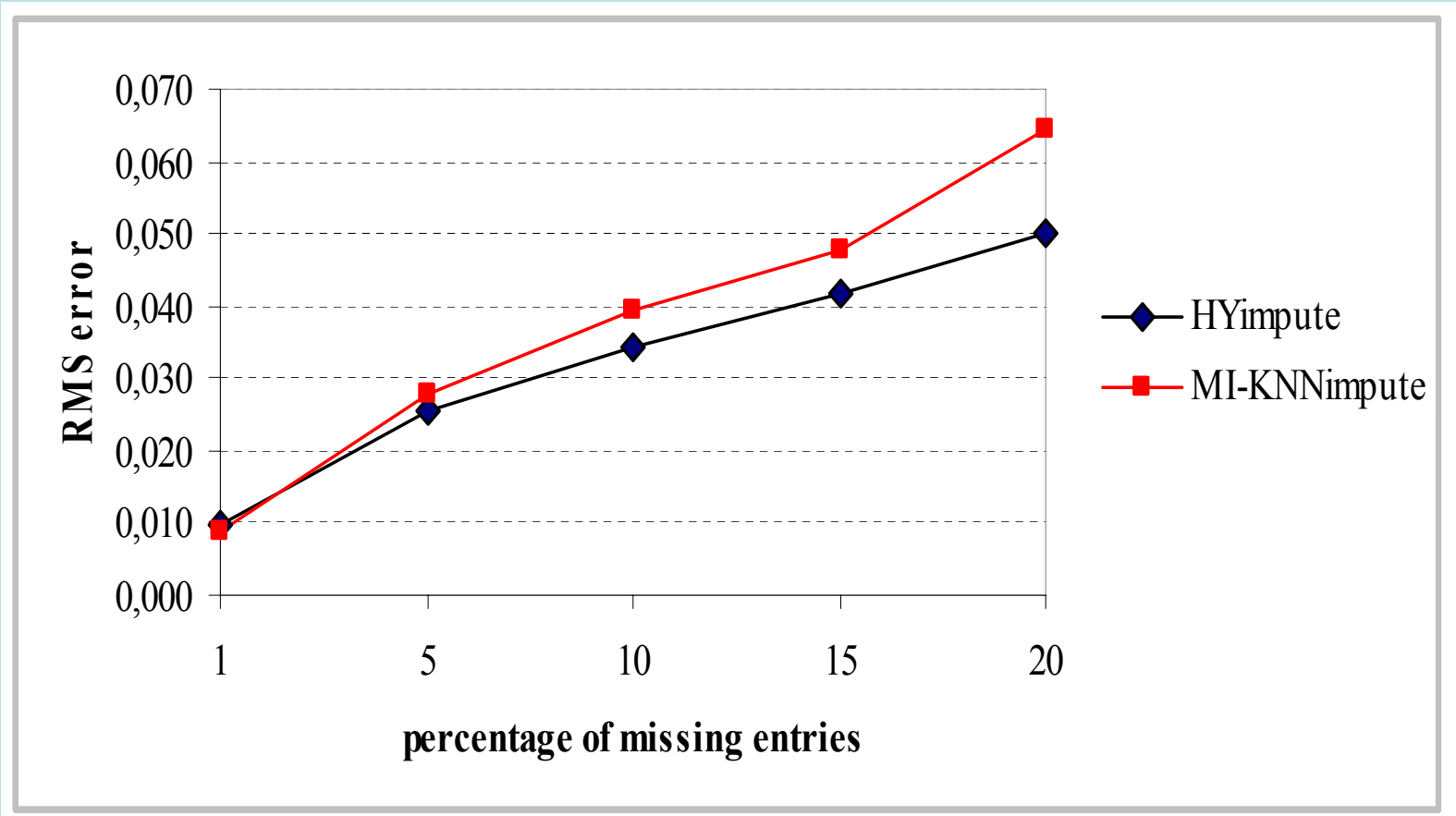
HYimpute versus KNNimpute, Row average and The two non-missing neighbours average



MI-KNNimpute RMS figures for three different final imputation values



HYimpute versus MI-KNNimpute performance



HYimpute Advantages

More robust and accurate missing value estimation by **combining** the **estimates** generated by a **set of different imputation algorithms**

Better performance than **two** other imputation methods: **KNNimpute** and **Multiple KNNimpute**

HYimpute can easily be extended and implemented by selecting **other more elaborated** missing value estimation algorithms.

Future Research

Development and implementation of **novel techniques for integrating** data coming from **multiple** microarray experiments.

Construction and evaluation of machine learning methods and techniques **for integration of heterogeneous biological data.**

ВЪПРОСИ . . .