BMC Bioinformatics

Research article

Improving missing value imputation of microarray data by using spot quality weights

Peter Johansson* and Jari Häkkinen

Address: Computational Biology, Department of Theoretical Physics, Lund University, SE-223 62 Lund, Sweden

Email: Peter Johansson* - peter@thep.lu.se; Jari Häkkinen - jari@thep.lu.se

* Corresponding author

Published: 16 June 2006

BMC Bioinformatics 2006, **7**:306 doi:10.1186/1471-2105-7-306

Received: 14 March 2006

Accepted: 16 June 2006

This article is available from: http://www.biomedcentral.com/1471-2105/7/306

Abstract

Background: Microarray technology has become popular for gene expression profiling, and many analysis tools have been developed for data interpretation. Most of these tools require complete data, but measurement values are often missing A way to overcome the problem of incomplete data is to impute the missing data before analysis. Many imputation methods have been suggested, some naïve and other more sophisticated taking into account correlation in data. However, these methods are binary in the sense that each spot is considered either missing or present. Hence, they are depending on a cutoff separating poor spots from good spots. We suggest a different approach in which a continuous spot quality weight is built into the imputation methods, allowing for smooth imputations of all spots to larger or lesser degree.

Results: We assessed several imputation methods on three data sets containing replicate measurements, and found that weighted methods performed better than non-weighted methods. Of the compared methods, best performance and robustness were achieved with the weighted nearest neighbours method (WeNNI), in which both spot quality and correlations between genes were included in the imputation.

Conclusion: Including a measure of spot quality improves the accuracy of the missing value imputation. WeNNI, the proposed method is more accurate and less sensitive to parameters than the widely used kNNimpute and LSimpute algorithms.

Background

During the last decade microarray technology has become an increasingly popular tool for gene expression profiling. Microarrays have been used in numerous biological contexts from studies of differentially expressed genes in tumours [1-4] to identification of cell cycle regulated genes in yeast [5]. A theme in microarray investigations is that they generate large amounts of data, and computer-based visualization and analysis tools must be used in experiment analysis. Tools such as hierarchical clustering [6], multidimensional scaling [7], and principal component analysis [8] are frequently used to visualize data. Machine learning methods like support vector machines [9] and artificial neural networks [10] have been used successfully to classify tumor samples. Common for these methods is that they in their standard versions assume complete data sets.

However, data is usually not complete. Data values may be missing due to poor printing of the arrays and consequently marked as missing during image analysis, but more common is that values are marked to be missing in

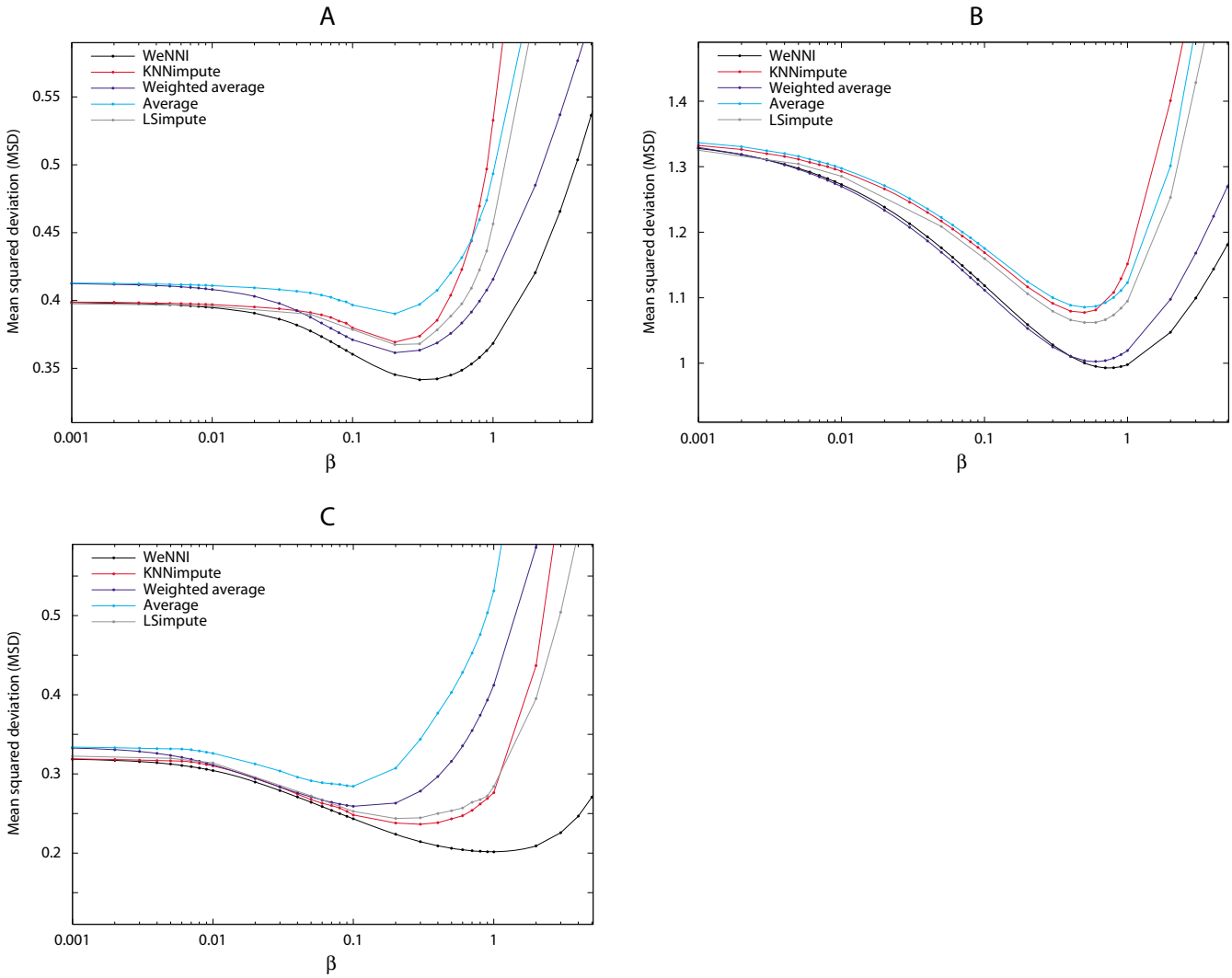

Figure 1

WeNNI is the most accurate imputation method. Performance of the five imputation methods with varying β. As explained in the Methods section, larger β changes weights to smaller values. In non-weighted methods β is the SNR cutoff. The increase in MSD for large β is an effect from too many missing values, which implies imputation breaks down. The standard error of means are within the line thicknesses. (A) *Breast cancer data*. WeNNI (black line) has the lowest MSD and the weighted methods perform better than the non-weighted methods. All methods have a minimum MSD around $\beta = 0.2$. (B) *Melanoma data*. WeNNI (black line) has the lowest MSD and the weighted methods perform better than the non-weighted methods. All methods have a minimum MSD around $\beta = 0.6$. (C) *Mycorrhiza data*. WeNNI (black line) retains the lowest MSD, whereas KNNimpute (red line) performs better that the weighted reporter average method. This may be explained as an effect of a different experimental design as discussed in the text. The minimum MSD is found in a β range 0.3–1 for the different methods.

a quality filtering pre-processing step. Common filter criteria are to mark spots with small area, spots with noisy background, spots with low intensity, or combinations of these [11]. One strategy to keep data complete is to remove reporters having missing values, but this may lead to an unnecessarily large loss of data. In particular when working with large data sets, reporters rarely have a complete set of values over all experiments. Another strategy is to keep reporters with not too many missing values and modify the subsequent analysis to handle incomplete data. However, it may not be feasible to modify the analysis tool, and therefore a popular approach is to impute the missing data in an intermediate step before analysis.

A common method to impute missing values is to replace missing values with the reporter average, *i.e.*, the average

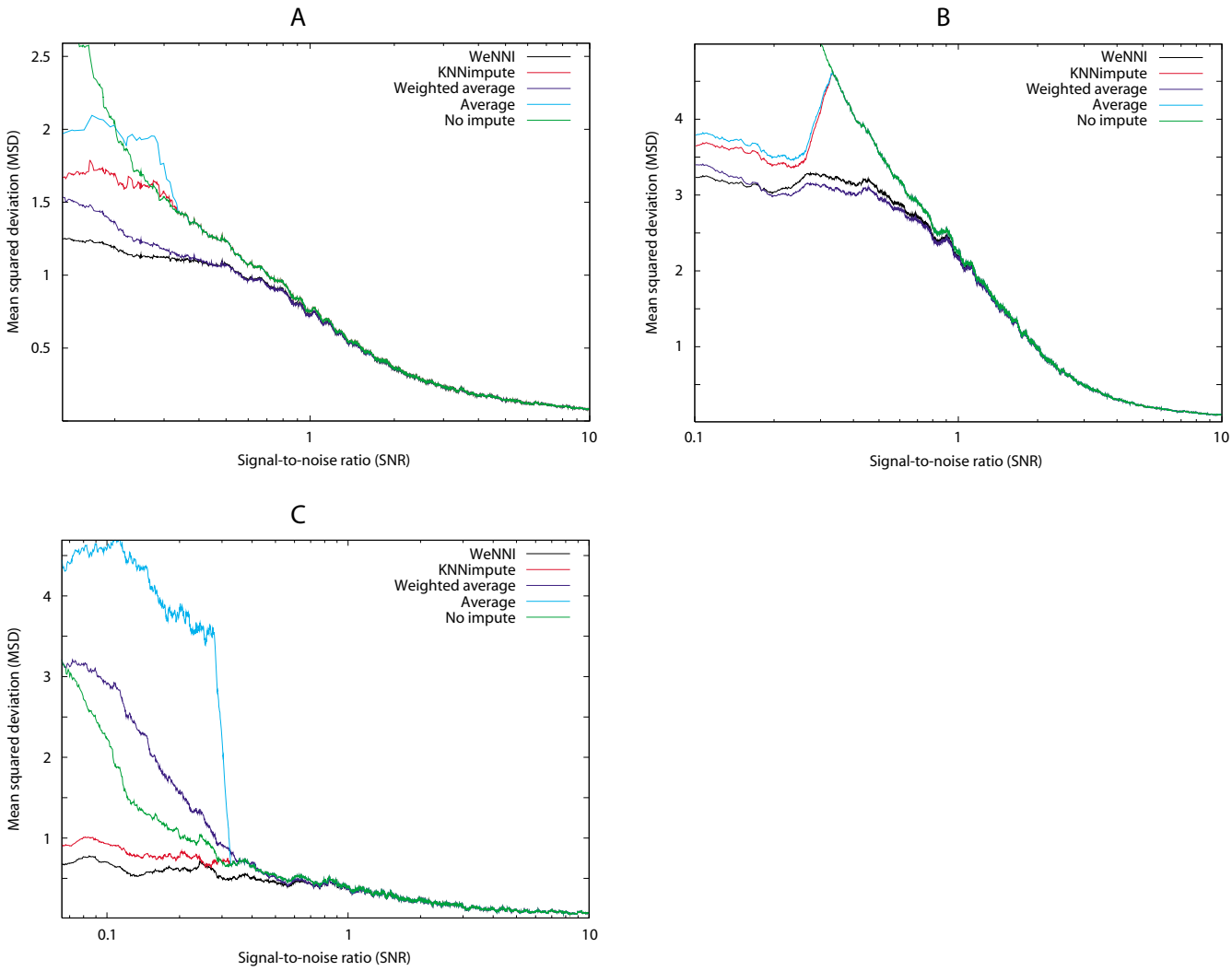

Figure 2

WeNNI is most accurate over all ranges of spot quality. The contribution to MSD for specific SNR for the different imputation methods applied to the three different data sets using $\beta = 0.3$. These plots were created using a sliding window containing 1% of all spots. Spots with small SNR (low quality) have the largest impact on MSD. (A) In the *breast cancer data* a weighted scheme is clearly essential and WeNNI is most accurate over all ranges of SNR (B) In the *melanoma data* a weighted scheme is clearly essential and the weighted reporter average show best performance for an SNR range 0.2–1. (C) In the *mycorrhiza data* the breakdown of the average reporter methods is very prominent. For the SNR range 0.07–0.4 it is even better to use no impute (green line) than the average methods. The breakdown of the reporter average methods are discussed in the text.

for the particular reporter over all experiments. Troyanskaya *et al*. showed that this method is not sufficient as it neglects correlations in data [12]. They also suggested a method KNNimpute, that was shown to reconstruct missing values well. In KNNimpute, for each reporter the most similar reporters are found and the weighted average of these reporters is used as the imputation value. Other imputation methods have been suggested [13-18] using the same basic idea that the imputation value is taken as an average over the neighbouring reporters.

As far as we know, all suggested imputation methods are binary in the sense that each spot is considered either missing or present. Hence they depend on a cutoff, *e.g.*, in intensity, separating poor spots from good spots. Tuning this cutoff is a balance act – a too liberal cutoff means

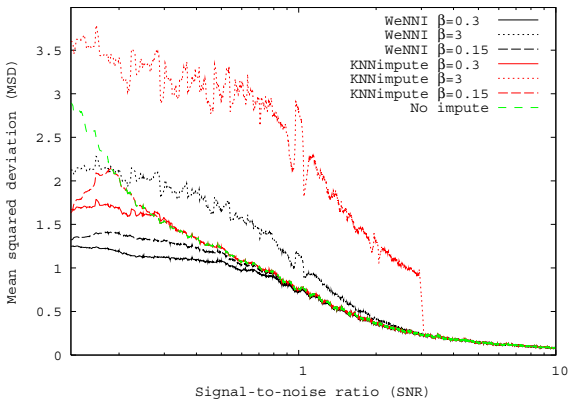

Figure 3
Comparison of WeNNI and KNNimpute. MSD contributions from specific SNR and different β for the *breast cancer data* set. This plot was created using a sliding window containing 1% of all spots.

noisy spots are kept in data, which may complicate subsequent analysis. On the other hand being too strict means spots containing information are marked as missing values and information is thrown away.

We suggest a more balanced approach, in which a spot quality weight is built into the imputation methods: good quality spots have more impact on the imputation of other spots, and are themselves subject to less imputation than spots with poorer quality. To examine the effects of this approach we extended two widely used methods, average imputation and KNNimpute [12], to handle continuous weights. We applied the two resulting methods to three data sets containing replicate measurements and found that weighted methods perform better that non-weighted.

Results and discussion

As outlined in the Methods section, we devised two imputation methods using spot quality weights. These methods are generalizations of two non-weight based methods and we evaluated the methods with replicate data sets. We used the mean squared deviation (MSD) to compare the performance of the two suggested methods, their non-weighted counterparts, and LSimpute [17]. We did the comparisons varying the spot quality threshold for missing values for the non-weight based methods. Correspondingly, for the weighted methods we varied the weight tuning parameter β in the calculations of the weights (see Methods).

In Figure 1 we present how the performance varied with a changing β in the three data sets. The plots show that WeNNI has the lowest MSD for all three data sets, the weighted methods outperform their non-weighted counterparts, and the minimum MSD is within the β range 0.1–1 for all methods.

The overall MSD is larger for the melanoma data set compared to the two other data sets, which may be due to that the melanoma data was generated a few years earlier than the other data.

An interesting finding was that weighted reporter average outperformed KNNimpute and LSimpute in the breast cancer and melanoma data sets. This result was unexpected since the weighted reporter average method neglects correlations between reporters. Moreover, the assumption for using reporter average is in general problematic, since the expression of a reporter in one experimental condition does not always reflect the expression of the reporter in another condition. For the mycorrhiza data used here the situation is even worse, since the cyclic experimental design [20] makes the expression value in one experiment auti-correlated to the reporter's average over the other experiments. For the nearest neighbours imputation methods however this problem does not arise because imputations are calculated as an average over the same experiment. These results imply one should consider the experimental design and choose imputation method carefully.

For small β all methods showed approximately equal performance. This result was expected, because for small β most weights are close to unity. In consequence, only a small fraction of the spots are imputed and make a minor contribution to the MSD. Moreover, the weights are effectively binary for small β, and the weighted methods become identical to their non-weighted counterparts.

To examine the difference between the weighted methods and their non-weighted counterparts, we plotted MSD as a function of SNR (Figure 2). As expected, spots with small SNR contributed most to MSD. The discrepancy between mycorrhiza data and the other two data sets also showed up here – the breakdown of the reporter average methods in the mycorrhiza data is very prominent (Figure 2C). The melanoma and breast cancer data showed very similar patterns for the different methods and the weighted methods performed better than their counterparts for all SNR. In some ranges of SNR, weighted reporter average even surpassed WeNNI, but overall WeNNI imputed the values most accurately.

In Figure 3, we demonstrate the effect of varying β for WeNNI and KNNimpute using the breast cancer data. In

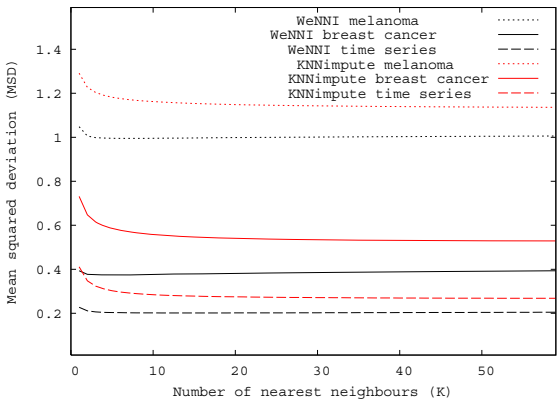

Figure 4
WeNNI and KNNimpute are insensitive to number of neighbours used. Performance of WeNNI and KNNimpute is plotted against the number of nearest neighbours for all three data sets using $\beta = 1$.

KNNimpute, only spots with smaller SNR than the cutoff β are imputed, and consequently the performance for SNR larger than β follows the no impute curve. For KNNimpute a choice of $\beta = 0.3$ was close to optimal. Using a smaller β deteriorated the imputation in two ways. Spots with SNR between the used β and the optimal value 0.3 were not imputed. In the plot we can see that the quality of these spots is so bad that preferably they should be imputed. More importantly, since these spots were not considered missing they were used in the imputation of values with very small SNR, which made the imputation less accurate. Moreover, when we used a too large β, the spots with SNR in the range 0.3–3 were imputed and their deviation from the replicate became larger than if they were not imputed. Also, the imputation of the spots with very small SNR became worse, since less information was used in the imputation. Choosing β corresponds to setting a cutoff in quality control criteria, and Figure 3 illustrates how a suboptimal cutoff level will lead to less reliable data. For WeNNI the cutoff is smoothened by the usage of continuous weights, and consequently WeNNI is more robust with respect to β.

In Figure 4, we illustrate how the performance of WeNNI and KNNimpute depends on the number of neighbours, K, used in the imputation. We notice that both methods are insensitive to changing K. For a small number of neighbours, both methods are insufficient. Troyanskaya *et al*. suggested K to be in the range between 10 and 20 neighbours for KNNimpute [12]. Our results agree with this finding and also show that the imputation of our data sets was accurate for a larger number of nearest neighbours.

When comparing non-weighted imputation methods, it is natural to calculate the comparison measure over imputed values only. Including non-imputed values in the evaluation makes no sense, as these values are not modified and thus independent of the imputation method. This is also the way imputation methods are compared in the literature [12-18]. However, for a weighted method every expression value is modified, and it is sensible to include all values in the calculation of MSD. In Table 1 we compare LSimpute.

KNNimpute, and WeNNI using both MSD and MSD_imputed. MSD_imputed is calculated as MSD but over imputed values (as defined by binary methods) only. We note that MSD_imputed is larger than MSD for all methods and data sets, which is expected because MSD_imputed is calculated over poor spots only and poor spots are expected to deviate more from their duplicates. Moreover, for MSD_imputed the difference between the methods is more apparent, which is a consequence from comparing poor spots only. In MSD all spots are included in the comparison and as good quality spots are modified to lesser degree, the difference between the methods looks smaller. We note that WeNNI is the most accurate method also using MSD_impute, in other words. WeNNI has the best performance even when values not imputed by non-weighted methods are excluded from the comparison.

For the largest data set, *breast cancer data* with approximately 55000 spots, a typical WeNNI run takes approximately 10 CPU minutes on a off-the-shelf computer (AMD Athlon 3700+ processor and 1 GB RAM), whereas kNNI is twice as fast. These two algorithms are implemented in the same C++ code base and differs only in the calculation of the imputation values. A comparison with LSimpute is not fair since LSimpute is an adaptive algorithm and is implemented in Java.

Spot quality weights and expression value imputation

The starting point for imputing expression values in this report is that the weight of a spot should depend on its quality, as best estimated from data. Here, we used a straight forward SNR based weight as it was not our aim to study quality of spots. The SNR based quality weights were introduced in [21], and many different studies of quality measures have been described [11,26-28]. These papers concentrate on studying how the quality of spots should be defined.

Analyses in microarray projects are commonly based on spot intensities, and for that reason we examined if using

Table 1: Comparisons of WeNNI, KNNimpute, and LSimpute adaptive using two different measures. MSD is the mean squared deviation calculated over all spots, whereas MSD_imputed is calculated over spots with SNR smaller than β, *i.e.*, the spots imputed in non-weighted methods. β was chosen to yield the lowest MSD for LSimpute adaptive. WeNNI is more accurate than LSimpute and KNNimpute, even though β was tuned to optimise the performance of LSimpute.

Data set	Measure	β	WeNNI	KNNimpute	LSimpute adaptive
Breast cancer	MSD	0.2	0.345	0.369	0.368
	MSD_imputed		1.59	1.81	1.75
Melanoma	MSD	0.6	0.995	1.08	1.05
	MSD_imputed		3.41	3.77	3.64
Mycorrhiza	MSD	0.2	0.216	0.241	0.244
	MSD_imputed		0.840	0.902	0.954

intensities instead of SNR changes the findings in this paper. We found that using this simpler quality weight (Eq. 4), the performance was almost as good as when using the SNR based weights (data not shown). The fact that the imputed expression value on average gets closer to its pristine replicate value, indicates that the SNR based weight may be a slightly better estimate of the spot quality.

In imputation of expression values, as in any transformation of data (*e.g.*, LOWESS normalisation or centralisation), one must be careful to not destroy the biological signal in the data. In our three data sets, we noticed that when WeNNI is used, the deviation from the pristine replicate is on average smaller than when not doing the transformation, in other world, on average an expression value is closer to its replicate after the transformation. The effect is measurable even for the naïve weight used here.

The goal of a weight is to catch the "true" quality of the spot, and as such it is important to define spot quality weight calculation to suit the data at hand, prior knowledge, and expertise. One important aspect of applying prior knowledge into weight calculation is that initial prescreening of array data should still be done before imputation, or any subsequent analysis. In this screening step bad spots are removed, and known malfunction in data (arrays) should be communicated with zero weights.

Conclusion

Virtually every analysis of microarray data is preceded by a filtering step, in which each spot is required to fulfil certain quality control criteria. If the spot fails to meet the quality requirements it is marked as a missing value. This is equivalent to accompanying each expression value with a binary weight, and enforces an abrupt cutoff in quality control criteria. We have generalised two widely used imputation methods to use continuous weights. Our finding that the weighted imputation methods outperformed their non-weighted counterparts, suggests that using continuous weights is superior to using binary weights. Our suggested improvement – to use continuous weights – is generic in the sense that most imputation methods can be generalised to use continuous weights.

The weighted nearest neighbours imputation method presented in this paper. WeNNI, outperformed all other tested methods for the three different data sets used in this study. WeNNI performs accurate imputation of expression values and is insensitive to the parameter values used, *i.e.*, the number of nearest neighbours and β. An increasing β corresponds to having a more strict spot quality control criteria. For a non-weighted method it means that more values are considered missing and consequently imputed. Our results suggest that the usage of a continuous weight makes the imputation less sensitive to the choice of β. The findings in this manuscript are based on comparisons of replicate data, however replicate data may not be available in every experimental setting and the scientific investigator cannot evaluate the impact of different parameter values. The results in this study show that the choice of parameters is not crucial, and suggest a value around 10 for nearest, neighbours and a β in the range 0.1–1.

Methods

Data sets and pre-processing

To evaluate the imputation methods, we used three data sets, i) *Melanoma data*. The melanoma data set was obtained from a panel of 61 human cell lines [19]. For each experiment, 19,200 reporters were printed in duplicates. Identification of individual spots on scanned arrays was done with ImaGene 4.0 (BioDiscovery, E1 Segundo, CA, USA). ii) *Breast cancer data*. The breast cancer data set is a subset of a larger ongoing study. We selected the 55 experiments that had been hybridised at the Swegene DNA Microarray Resource Centre in Lund, Sweden, and were from tumours mutated either in *BRCA1* or in *BRCA2*. Each array contained 55,488 spots and except a small number of control spots each reporter was printed in

duplicate. Identification of individual spots on scanned arrays was done with GenePix Pro 4.0 (Axon Instruments, Union City, CA, USA). iii) *Mycorrhiza data*. The mycorrhiza data set was generated to study ectomycorrhizal root tissue [20]. In order to avoid any bias from using dye swap replicates, we used half of the arrays from the study. We used the 10 arrays denoted R3 between ECM's at different time points, and R1 between ECM and REF (Figure 2 in [20]). Each array contained 10,368 spots and except a small number of control spots each reporter was printed four times. Identification of individual spots on scanned arrays was done with GenePix Pro 3.0.6.89) (Axon Instruments, Union City, CA, USA).

For each spot, we used the mean spot intensity, I_{fg}, the mean background intensity, I_{bg}, and the standard deviation of the background intensity, σ_{bg}. For each spot we calculated the signal-to-noise ratio (SNR) [11] as

$$\frac{1}{\mathrm{SNR}^2} = \frac{1}{\mathrm{SNR}_t^2} + \frac{1}{\mathrm{SNR}_c^2} = \frac{\sigma_{bg,t}^2}{(I_{fg,t} - I_{bg,t})^2} + \frac{\sigma_{bg,c}^2}{(I_{fg,c} - I_{bg,c})^2}. \quad (1)$$

Subscripts t and c denotes treatment and control, respectively. As expression value, x, we used the logarithm to base 2 of the ratio of the signal in the treatment sample and the signal in the control sample

$$x = \log_2\left(\frac{I_{fg,t} - I_{bg,c}}{I_{fg,c} - I_{bg,c}}\right), \quad (2)$$

where spots with non-positive signal in either treatment or control were marked as invalid.

We applied a liberal filter to the data. In the melanoma data set we kept reporters having less than 50% invalid values in both duplicates. The remaining data was split into two replicate data sets. This was also done for the two other data sets, with the exception that the mycorrhiza data was split into four replicate data sets. Each data set was then centralised experiment by experiment such that the average expression value for an experiment was zero.

After filtering, the melanoma data consisted of two replicate data sets each having 61 experiments and 17, 549 reporters, the breast cancer data consisted of two explicate data sets each having 55 experiments and 23,764 reporters, and the mycorrhiza data consisted of four replicate data sets each having 10 experiments and 2,052 reporters.

Quality weight

The basis for weight calculations are two weight formulae inspired by previous work [21-24]. We used an SNR based weight defined as

$$w = \frac{1}{1 + \frac{\beta^2}{\mathrm{SNR}_t^2} + \frac{\beta^2}{\mathrm{SNR}_c^2}}. \quad (3)$$

This weight is defined to be bound within zero and unity. The free parameter β is used to tune the distribution of weights. For a small β all weights are close to unity, except when zero or negative intensities have been measured which implies a zero weight. For a large β all weights are close to zero. In non-weighted (binary) methods we marked expression values to be missing when the corresponding continuous weight was less than 0.5. In this way β defined a cutoff for when a value is considered to be missing.

To cross check that the findings in this paper do not depend on SNR, we also used a simple weight based on intensity only:

$$w = \frac{1}{1 + \frac{\beta^2}{(I_{fg,t} - I_{bg,t})^2} + \frac{\beta^2}{(I_{fg,c} - I_{bg,c})^2}}. \quad (4)$$

This weight is also bound to be within zero and unity, and β has the same function here as for the SNR based weight above.

Imputation methods

We compared five imputation methods; three non-weight based methods, reporter average, KNNimpute, and LSimpute_adaptive; and two weight based, weighted reporter average and weighted nearest neighbours imputation (WeNNI).

Reporter average methods

The widely used reporter average imputation method is intuitive and easy to implement. Assuming the expression level of a reporter in one experiment to be similar to the expression level in other experiments the expression value is imputed as the average of the reporter's expression value over all experiments. Similarly to Andersson *et al*. [21], we extended the reporter average by using continuous spot quality weights between zero and unity. A spot with a weight equal to unity is not imputed, whereas for a spot with weight equal to zero the expression value is imputed to be the weighted reporter average. A spot having an intermediate weight is imputed as a linear combination of the extreme cases above. These three cases are covered in the imputation equation

$$x'_{re} = w_{re}x_{re} + (1 - w_{re})\,\hat{x}_{re}\,, \quad (5)$$

in which x_{re} is the expression value in reporter r and experiment e, w_{re} is the quality weight, and $\hat{x}_{re}$ is the weighted reporter average

$$\hat{x}_{re} = \frac{\sum_{i=1}^{M} w_{ri}x_{ri}}{\sum_{i=1}^{M} w_{ri}}, \quad (6)$$

where M is the number of experiments.

The use of the spot quality weight is twofold. First, the weight is used in the calculation of the reporter average. Second, the weight is used in the calculation of the imputed expression value – poor quality spots are changed more than good quality spots.

KNNimpute

The currently most popular imputation method that goes beyond reporter averaging. KNNimpute, has been shown to be a very good method for imputation of missing values [12]. The main idea of KNNimpute is to look for the K most similar reporters when a value is missing for a reporter. Two reporters n and m are considered to be similar when the Euclidean distance.

$$d_{nm}^2 = \frac{1}{M}\sum_{i=1}^{M}(x_{ni} - x_{mi})^2, \quad (7)$$

between their expression patterns is small. These K reporters are used to calculate a weighted average of the values in the experiment of interest. The weighted average is calculated as

$$\hat{x}_{re} = \frac{\sum_{i=1}^{K} \frac{x_{ie}}{d_{ri}}}{\sum_{i=1}^{K} \frac{1}{d_{ri}}}, \quad (8)$$

where x_{ie} is the value of the ith nearest reporter, d_{ri} is the distance between reporter r and reporter i and K is the number of neighbours to use in the calculation. This weighted average is used as imputation value of missing values.

Weighted Nearest Neighbours Imputation [WeNNI]

KNNimpute is binary in the sense that each value is regarded as either missing or present. In WeNNI, we smooth out this sharp border between missing and present values by assigning a continuous quality weight to each value, where a zero weight means the value is completely missing and a larger weight means the value is more reliable. In the special case when all weights are either 0 or 1. WeNNI is equivalent to KNNimpute.

The WeNNI method consists of two steps. First, we calculate distances between the reporters taking the weights into account. Second, we calculate a weighted average of the values of the nearest neighbours. We expanded the Euclidean distance used in KNNimpute to include quality weights. The weights were included in such a way that spots with large weights are more important for the distance measure than spots with low weights. We calculated the distance d_{nm} between reporter n and reporter m as

$$d_{nm}^2 = \frac{\sum_{i=1}^{M} w_{ni}w_{mi}(x_{ni} - x_{mi})^2}{\sum_{i=1}^{M} w_{ni}w_{mi}}, \quad (9)$$

where M is the number of experiments. A weighted average of the nearest neighbours is calculated as

$$\hat{x}_{re} = \frac{\sum_{i=1}^{L} \frac{w_{ie}x_{ie}}{d_{ri}}}{\sum_{i=1}^{L} \frac{w_{ie}}{d_{ri}}}, \quad (10)$$

where L is defined by

$$\sum_{i=1}^{L} w_{ie} \le K < \sum_{i=1}^{L+1} w_{ie}. \quad (11)$$

In the second step, we take the imputed value as a linear combination of the original value and the value suggested by the neighbours

$$x'_{re} = w_{re}x_{re} + (1 - w_{re})\,\hat{x}_{re}\,. \quad (12)$$

As for weighted reporter average above, when the quality weight is zero, we ignore the original value. When the weight is unity, we trust the original value and ignore the value suggested by the neighbours.

LSimpute

Bø *et al*. showed that LSimpute_adaptive is a very good method for imputation of missing values [17]. The method is based on the least squares principle, which means the sum of squared errors of a regression model is minimised and the regression model is used to impute missing values. The method utilises correlations both between reporters and experiments.

In the comparisons made in this report, we used the LSimpute_adaptive algorithm implemented in the publicly available LSimpute program [17].

Evaluation method

In order to validate the imputation methods we did as follows for each of the three data sets. We split the data into replicate data sets: two sets for the melanoma and breast cancer data, and four sets for the mycorrhiza data. We imputed the data in one of the replicate data sets and compared the imputed data, x', to the other pristine replicate data, y. For the mycorrhiza data, we compared the imputed data to the (non-weighted) average of the three pristine replicate data sets. We measured the quality of the method using the mean squared deviation

$$\mathrm{MSD} = \frac{1}{N}\sum_{i=1}^{N}(x_i' - y_i)^2. \qquad (13)$$

where the sum runs over all expression values in all replicate data sets, except spots in the pristine data set that were marked as invalid in the data pre-processing step described above. The fraction of spots not used in the summation were: 6% for the melanoma data, 7% for the breast cancer data, and 8% for the mycorrhiza data.

The motivation for this choice of MSD as evaluation metric is threefold. First, in the weighted methods the imputed value is a linear combination of the value suggested by the neighbours and the original value. Hence, comparing with the original value would introduce an information leak, making the evaluation unfair. Second, introducing artificial missing values randomly may not be optimal [15,25], since it assumes missing values to occur uncorrelated. By using replicates we could avoid this problem and mark spots as missing values depending on their quality. Third, we avoided any bias that could be introduced by imputing both replicates and comparing the imputed values. By considering the zero impute method (missing values are set to zero), it is easy to understand that a bias could be introduced. If both replicate spots are imputed, *i.e.*, both set to zero, they would have no deviation and the evaluation would obviously be flattering.

Availability and requirements

Project name: WeNNI

Project home page: http://lev.thep.lu.se/trac/baseplugins/wiki/WeNNI

Operating system: All that supports GCC 4.0 or later (tested on GNU/Linux and MacOS X 10.4).

Other requirements: Gnu Scientific Library, GSL. http://www.gnu.org/software/gsl/

License: GNU General Public License

Authors' contributions

Both authors developed weighted methods, designed and performed comparisons of methods, and wrote the manuscript.

Acknowledgements

We thank Patrik Edén for valuable discussions. The mycorrhiza data set was kindly provided by Tomas Johansson at the Department of Ecology, Lund University, Sweden. The melanoma data set was kindly provided by Sandra Pavey and Nicholas Hayward at the Queensland Institute of Medical Research, Australia. The breast cancer data set was kindly provided by Johan Vallon-Christersson at the Swegene DNA Microarray Resource Center at the BioMedical Center in Lund, Sweden, supported by the Knut and Alice Wallenberg Foundation through the Swegene consortium. J.H. was in part supported by the Knut and Alice Wallenberg Foundation through the Swegene consortium.

References

1. DeRisi J, Penland L, Brown PO, Bittner ML, Meltzer PS, Ray M, Chen Y, Su YA, Trent JM: **Use of a cDNA microarray to analyse gene expression patterns in human cancer.** *Nat Genet* 1996, **14(4):**457-460.
2. Golub TR, Slonim DK, Tamayo P, Huard C, Gaasenbeek M, Mesirov JP, Coller H, Loh ML, Downing JR, Caligiuri MA, Bloomfield CD, Lander ES: **Molecular classification of cancer: class discovery and class prediction by gene expression monitoring.** *Science* 1999, **286(5439):**531-537.
3. Sørlie T, Perou CM, Tibshirani R, Aas T, Geisler S, Johnsen H, Hastie T, Eisen MB, van de Rijn M, Jeffrey SS, Thorsen T, Quist H, Matese JC, Brown PO, Botstein D, Eystein Lonning P, Borresen-Dale AL: **Gene expression patterns of breast carcinomas distinguish tumor subclasses with clinical implications.** *Proc Natl Acad Sci USA* 2001, **98(19):**10869-10874.
4. Hedenfalk I, Duggan D, Chen Y, Radmacher M, Bittner M, Simon R, Meltzer P, Gusterson B, Esteller M, Kallioniemi OP, Wilfond B, Borg A, Trent J: **Gene-expression profiles in hereditary breast cancer.** *N Engl J Med* 2001, **344(8):**539-548.
5. Spellman PT, Sherlock G, Zhang MQ, Iyer VR, Anders K, Eisen MB, Brown PO, Botstein D, Futcher B: **Comprehensive identification of cell cycle-regulated genes of the yeast Saccharomyces cerevisiae by microarray hybridization.** *Mol Biol Cell* 1998, **9(12):**3273-3297.
6. Eisen MB, Spellman PT, Brown PO, Botstein D: **Cluster analysis and display of genome-wide expression patterns.** *Proc Natl Acad Sci USA* 1998, **95(25):**14863-14868.
7. Khan J, Simon R, Bittner M, Chen Y, Leighton SB, Pohida T, Smith PD, Jiang Y, Gooden GC, Trent JM, Meltzer PS: **Gene expression profiling of alveolar rhabdomyosarcoma with cDNA microarrays.** *Cancer Res* 1998, **58(22):**5009-5013.
8. Pomeroy SL, Tamayo P, Gaasenbeek M, Sturla LM, Angelo M, McLaughlin ME, Kim JYH, Goumnerova LC, Black PM, Lau C, Allen JC, Zagzag D, Olson JM, Curran T, Wetmore C, Biegel JA, Poggio T, Mukherjee S, Rifkin R, Califano A, Stolovitzky G, Louis DN, Mesirov JP, Lander ES, Golub TR: **Prediction of central nervous system embryonal tumour outcome based on gene expression.** *Nature* 2002, **415(6870):**436-442.
9. Furey TS, Cristianini N, Duffy N, Bednarski DW, Schummer M, Haussler D: **Support vector machine classification and validation of cancer tissue samples using microarray expression data.** *Bioinformatics* 2000, **16(10):**906-914.
10. Khan J, Wei JS, Ringnér M, Saal LH, Ladanyi M, Westermann F, Berthold F, Schwab M, Antonescu CR, Peterson C, Meltzer PS: **Classification and diagnostic prediction of cancers using gene expression profiling and artificial neural networks.** *Nat Med* 2001, **7(6):**673-679.
11. Chen Y, Kamat V, Dougherty ER, Bittner ML, Meltzer PS, Trent JM: **Ratio statistics of gene expression levels and applications to microarray data analysis.** *Bioinformatics* 2002, **18(9):**1207-1215.
12. Troyanskaya O, Cantor M, Sherlock G, Brown P, Hastie T, Tibshirani R, Botstein D, Altman RB: **Missing value estimation methods for DNA microarrays.** *Bioinformatics* 2001, **17(6):**520-525.

13. Ouyang M, Welsh WJ, Georgopoulos P: **Gaussian mixture clustering and imputation of microarray data.** *Bioinformatics* 2004, **20(6):**917-923.
14. Kim KY, Kim BJ, Yi GS: **Reuse of imputed data in microarray analysis increases imputation efficiency.** *BMC Bioinformatics* 2004, **5:**160.
15. Seligal MSB, Gondal I, Dooley LS: **Collateral missing value imputation: a new robust missing value estimation algorithm for microarray data.** *Bioinformatics* 2005, **21(10):**2417-2423.
16. Kim H, Golub GH, Park H: **Missing value estimation for DNA microarray gene expression data: local least squares imputation.** *Bioinformatics* 2005, **21(2):**187-198.
17. Bø TH, Dysvik B, Jonassen I: **LSimpute: accurate estimation of missing values in microarray data with least squares methods.** *Nucleic Acids Res* 2004, **32(3):**e34.
18. Scheel I, Aldrin M, Glad IK, Sorum R, Lyng H, Frigessi A: **The influence of missing value imputation on detection of differentially expressed genes from microarray data.** *Bioinformatics* 2005, **21(23):**4272-4279.
19. Pavey S, Johansson P, Packer L, Taylor J, Stark M, Pollock PM, Walker GJ, Boyle GM, Harper U, Cozzi SJ, Hansen K, Yudt L, Schmidt C, Hersey P, Ellem KAO, O'Rourke MGE, Parsons PG, Meltzer P, Ringnér M, Hayward NK: **Microarray expression profiling in melanoma reveals a BRAF mutation signature.** *Oncogene* 2004, **23(23):**4060-4067.
20. Le Quere A, Wright DP, Soderstrom B, Tunlid A, Johansson T: **Global patterns of gene regulation associated with the development of ectomycorrhiza between birch (Betula pendula Roth.) and Paxillus involutus (Batsch) Fr.** *Mol Plant Microbe Interact* 2005, **18(7):**659-673.
21. Andersson A, Edén P, Lindgren D, Nilsson J, Lassen C, Heldrup J, Fontes M, Borg A, Mitelman F, Johansson B, Hoglund M, Fioretos T: **Gene expression profiling of leukemic cell lines reveals conserved molecular signatures among subtypes with specific genetic aberrations.** *Leukemia* 2005, **19(6):**1042-1050.
22. Fernebro J, Francis P, Edén P, Borg A, Panagopoulos I, Mertens F, Vallon-Christersson J, Akerman M, Rydholm A, Bauer HC, Mandahl N, Nilbert M: **Gene expression profiles relate to SS18/SSX fusion type in synovial sarcoma.** *Int J Cancer* 2006, **118(5):**1165-1172.
23. Andersson A, Olofsson T, Lindgren D, Nilsson B, Ritz C, Edén P, Lassen C, Rade J, Fontes M, Morse H, Heldrup J, Behrendtz M, Mitelman F, Hoglund M, Johansson B, Fioretos T: **Molecular signatures in childhood acute leukemia and their correlations to expression patterns in normal hematopoietic subpopulations.** *Proc Natl Acad Sci USA* 2005, **102(52):**19069-19074.
24. Francis P, Fernebro J, Edén P, Laurell A, Rydholm A, Domanski HA, Breslin T, Hegardt C, Borg A, Nilbert M: **Intratumor versus intertumor heterogeneity in gene expression profiles of soft-tissue sarcomas.** *Genes Chromosomes Cancer* 2005, **43(3):**302-308.
25. Oba S, Sato Ma, Takemasa I, Monden M, Matsubara KI, Ismi S: **A Bayesian missing value estimation method for gene expression profile data.** *Bioinformatics* 2003, **19(16):**2088-2096.
26. Bylesjö M, Eriksson D, Sjödin A, Sjöström M, Jansson S, Antti H, Trygg J: **MASQOT: a method for cDNA microarray spot quality control.** *BMC Bioinformatics* 2005, **6:**250.
27. Tran PH, Peiffer DA, Shin Y, Meek LM, Brody JP, Cho KWY: **Microarray optimizations: increasing spot accuracy and automated identification of true microarray signals.** *Nucleic Acids Res* 2002, **30(12):**e54.
28. Wang X, Hessner MJ, Wu Y, Pati N, Ghosh S: **Quantitative quality control in microarray experiments and the application in data filtering, normalization and false positive rate prediction.** *Bioinformatics* 2003, **19(11):**1341-1347. [Evaluation Studies].
29. Saal LH, Troein C, Vallon-Christersson J, Gruvberger S, Borg A, Peterson C: **BioArray Software Environment (BASE): a platform for comprehensive management and analysis of microarray data.** *Genome Biol* 2002, **3(8):**SOFTWARE0003.