# Estimating Mutual Information in Genetics

Christoph Schmitz
UMIC Research Centre
RWTH Aachen University
52056 Aachen, Germany
schmitz@umic.rwth-aachen.de

Anke Schmeink
UMIC Research Centre
RWTH Aachen University
52056 Aachen, Germany
schmeink@umic.rwth-aachen.de

Rudolf Mathar
Chair of Theoretical Information Technology
RWTH Aachen University
52056 Aachen, Germany
mathar@ti.rwth-aachen.de

***Abstract*—In gene mapping, the science of finding connections between the genotype and the phenotype, the revealing of complex, nondeterministic connections is an important problem. Tools from information theory, especially the mutual information, have proven to be valuable. This arises the need to estimate the mutual information from a set of samples, and to know the distribution of the estimator. In this work, the established maximum likelihood estimator for the mutual information is examined using simulated data, and it is compared to an approximation. Additionally, another estimator based on preprocessing the data using B-splines is considered, and compared to the conventional estimator and to the well-established chi-square test for independence.**

## I. Background

Since the discovery of the deoxyribonucleic acid (DNA) as carrier of the genetic information, and the decipherment of the coding of the proteins, there has been great interest in connections between the genome of an individual and its outward appearance, the phenotype. This is especially true for all kinds of diseases. Since then, several results have been achieved in gene mapping, the science of finding such dependencies. The dependencies that were the easiest to find are the so-called Mendelian traits, where a genetic marker and a phenotype trait are connected in a deterministic way.

However, there are other more complex dependencies which are harder to discover. Methods from information theory have proven to be valuable tools for this purpose in recent years, see for example [1] and [2]. This work presents some methods, which are based on the concept of mutual information, and deals with the problems that arise in the practical application.

### A. The Genetic Code

As explained in [2] and [3], the deoxyribonucleic acid (DNA) as carrier of genetic information consists of two complementary strands of nucleotides, a large number of bases attached to a backbone. Four different values are possible at each locus, coded by the four bases Adenine, Guanine, Cytosine and Thymine. Figure 1 shows a sketch of a short section of the DNA. Each individual possesses two sets of DNA, one inherited from the mother and one from the father.

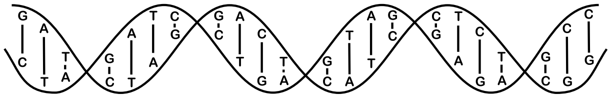

Fig. 1. The Deoxyribonucleic Acid (DNA)

The human DNA consists of about 3 billion base pairs and is largely identical for all individuals. There are about ten million positions, called single nucleotide polymorphisms (SNP), where different alleles appear with a significant probability. These positions are responsible for the differences between human individuals that are caused genetically. Usually, two different values do appear on such a locus, for example C and T. The allele with the higher probability (e.g., C) is called the major allele, the other one is the minor allele. Considering the value of the SNP on both sets of DNA in this case yields the four different genotypes CC, CT, TC and TT, in which it is often not possible to distinguish between CT and TC. If the minor allele T causes some disease or phenotype trait in general, there are two main ways in which this can occur. If only those individuals carrying this allele on both sets of DNA (genotype TT) are affected, the allele is called recessive. If the genotypes CT and TC are also affected, the allele is called dominant.

### B. Mutual Information

Let $X$ and $Y$ be two discrete random variables with supports $\mathcal{X} = \{x_1, \ldots, x_m\}$ and $\mathcal{Y} = \{y_1, \ldots, y_{m'}\}$, respectively, and a distribution given by

$$\begin{aligned} P(X = x_i) &= p_{i\bullet}, \\ P(Y = y_j) &= p_{\bullet j}, \\ P(X = x_i, Y = y_j) &= p_{ij}. \end{aligned}$$

Mutual information as a measure of dependency between $X$ and $Y$ is defined, e.g., in [4] as

$$I(X;Y) = \sum_{i=1}^{m} \sum_{j=1}^{m'} p_{ij} \log_2 \frac{p_{ij}}{p_{i\bullet} p_{\bullet j}}.$$

The choice of the base of the logarithm alters the result just by a scaling factor. Throughout this work the logarithm to the basis two will always be used. The mutual information $I(X;Y)$ is zero if and only if $X$ and $Y$ are stochastically independent.

*1) Application in Gene Mapping:* Gene mapping means finding connections between the genotype of an individual and the outward appearance, the phenotype. While this is easy for the so-called Mendelian traits, where a phenotype trait depends

978-1-61284-402-2/11/$26.00 ©2011 IEEE

on a single genetic marker in a deterministic manner, more complex connections are harder to find, especially if they are not deterministic. The information-theoretic approach given in [1] interprets the genotype as a random variable $X$ and the phenotype as a random variable $Y$. While $X$ will often be ternary (as there are three distinguishable genotypes), $Y$ might be binary or have more outcomes. It might even be continuous if the regarded trait is a continuous measure. Among other methods, the mutual information $I(X;Y)$ is a measure of dependency that reveals whether $X$ and $Y$ are independent or not.

*2) The Estimation Problem:* Applying the method mentioned above usually requires to calculate the mutual information from some samples, without knowledge about the actual distribution. A suitable estimator for this purpose is the maximum likelihood estimator that is presented in [5] and [6]. Let $N$ be the number of samples, $(\tilde{x}_1, \tilde{y}_1), \ldots, (\tilde{x}_N, \tilde{y}_N)$ the observed outcomes, and

$$\begin{aligned} n_{i+} &= \left|\{k \in \{1, \ldots, N\} \mid \tilde{x}_k = x_i\}\right|, \\ n_{+j} &= \left|\{k \in \{1, \ldots, N\} \mid \tilde{y}_k = y_j\}\right|, \\ n_{ij} &= \left|\{k \in \{1, \ldots, N\} \mid \tilde{x}_k = x_i, \tilde{y}_k = y_j\}\right| \end{aligned}$$

the frequencies of occurrence. Using the relative frequencies as an estimation for the probabilities yields the estimator

$$\hat{I}_N(X;Y) = \sum_{i=1}^{m} \sum_{j=1}^{m'} \frac{n_{ij}}{N} \log_2 \frac{n_{ij} N}{n_{i+} n_{+j}}.$$

This estimator is a random variable that depends on the drawn samples. It is not difficult to prove that it converges almost surely to the mutual information $I(X;Y)$ when the number of samples rises to infinity, this is for example shown in [6].

The distribution of the estimator is not known, but according to [1] and [5] there exists a second-order approximation

$$\hat{I}_N(X;Y) \approx \frac{1}{2 \ln 2} \sum_{i=1}^{m} \sum_{j=1}^{m'} \frac{(n_{ij} - n_{i+} n_{+j})^2}{n_{i+} n_{+j} N^2}$$

based on the Taylor series around $\hat{I}_N(X;Y) = 0$, which corresponds to the case where $X$ and $Y$ are stochastically independent. This yields an approximative gamma distribution

$$\tilde{I}_N \sim \Gamma\left(\frac{1}{2}(m-1)(m'-1), \frac{1}{N \ln 2}\right)$$

for the independent case, which for example is useful to compute confidence intervals for hypothesis testing. The properties of the gamma distribution imply that the cumulative distribution function $F_{\tilde{I}_N}(z)$ of $\tilde{I}_N$ is reciprocally scaled with $N$ on the $z$ axis, so

$$F_{\tilde{I}_N}(z) = F_{\tilde{I}_1}(zN)$$

holds, and doubling the number of samples bisects all quantiles. An extension of the approximation to weakly dependent random variables is given in [5].

TABLE I
DISTRIBUTION TABLE FOR THE INDEPENDENT CASE

| | 0 | 1 | $\sum$ |
|---|---|---|---|
| CC | 0.55296 | 0.08704 | 0.64 |
| CT/TC | 0.27648 | 0.04352 | 0.32 |
| TT | 0.03456 | 0.00544 | 0.04 |
| $\sum$ | 0.864 | 0.136 | 1 |

TABLE II
DISTRIBUTION TABLE FOR THE DEPENDENT CASE

| | 0 | 1 | $\sum$ |
|---|---|---|---|
| CC | 0.576 | 0.064 | 0.64 |
| CT/TC | 0.256 | 0.064 | 0.32 |
| TT | 0.032 | 0.008 | 0.04 |
| $\sum$ | 0.864 | 0.136 | 1 |

## II. RESULTS AND DISCUSSION

### A. Maximum Likelihood Estimator

In the example constructed for the simulation, the ternary random variable $X$ represents the genotype. The minor allele T is assumed to appear with a probability of 0.2 on each DNA set. So the genotypes CC, CT/TC and TT have the probabilities 0.64, 0.32 and 0.04, respectively. The phenotype trait $Y$ is binary, thus, there are six bins into which the observations can fall. The value $Y = 1$ is assumed to appear with probability 0.136 under all genotypes in the independent case. Table I shows the entire distribution table for $X$ and $Y$. In the dependent case, the allele T is assumed to be dominant, and the probability for $Y = 1$ is assumed to be 0.1 for the genotype CC, and 0.2 for the genotypes CT/TC and TT. The distribution table for the dependent case is shown in Table II. The probablities were chosen in such a way that the marginal distributions are identical in the independent and the dependent case. The true value of the mutual information in the dependent case is 0.0136.

For both cases, the distribution of the estimator $\hat{I}_N(X;Y)$ is computed by simulation for a different number of samples $N$. The cumulative distribution functions for 100 and for 1000 samples are shown in Figure 2. The marked thresholds are chosen such that the type I error and the type II error are equal. The achievable error level for 100 samples is 0.3566 $(\pm 0.0005)$, for 1000 samples it is 0.0339 $(\pm 0.0002)$ (all numerical values are given rounded up to the fourth decimal place in this work, and the standard deviation is given rounded up to one significant decimal place). As expected, the selectivity is higher for a larger number of samples, the decision between the independent and the dependent case can be made with a lower error probability.

### B. Comparison with the Approximation

For the independent case, Figure 3 shows a comparison of the simulated distribution of the estimator and the approximative distribution given above. While there can be seen quite a deviation between the two distributions for ten samples (which indeed is a really small number of samples for six bins), they are already almost identical for 100 samples.

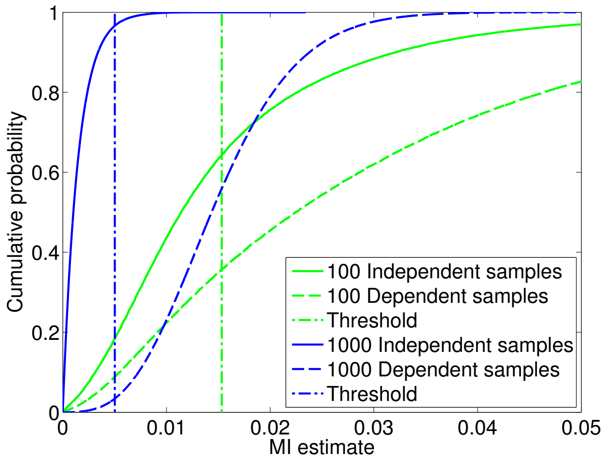

Fig. 2. Independent Case vs. Dependent Case

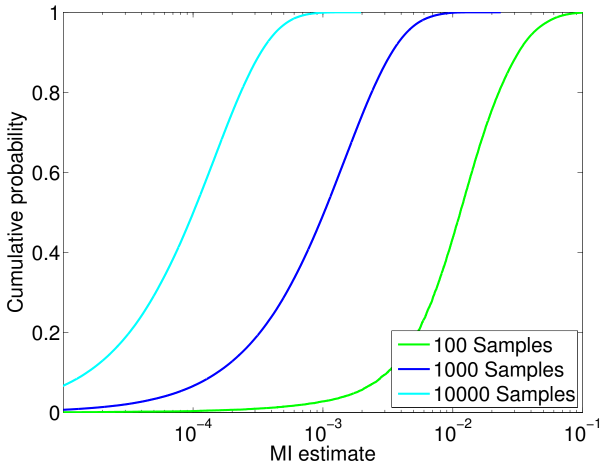

Fig. 4. Distribution of the MI Estimate on a Logarithmic Scale

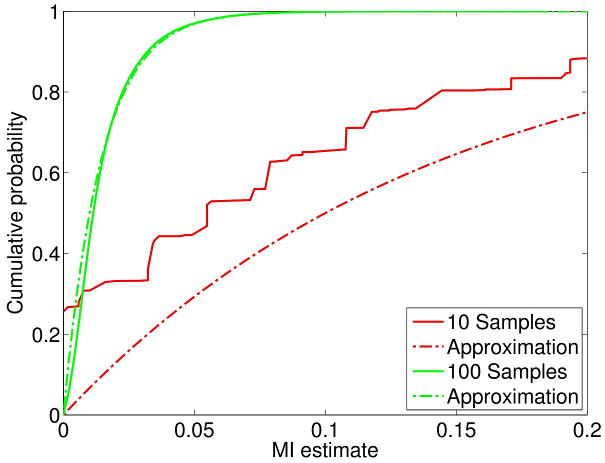

Fig. 3. Comparison of Results and Approximation

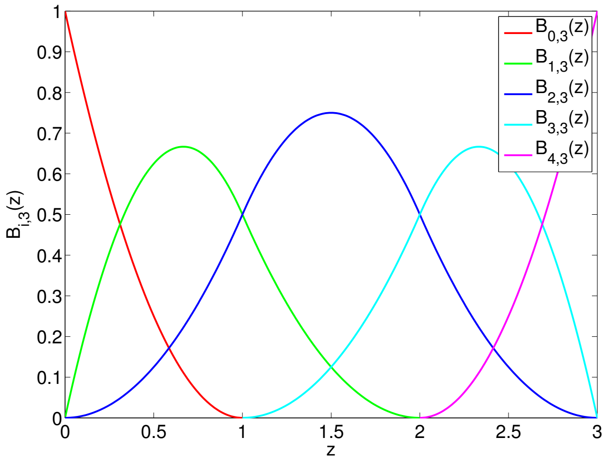

Fig. 5. B-Spline Functions

The scaling property of the approximation, which is already mentioned in [5], is reflected in Figure 4, where the cumulative distribution function of the simulated estimator is plotted on a logarithmic scale for different numbers of samples. Each time the number of samples is multiplied by ten, the curve is shifted to the left by one order of magnitude, apart from that its shape does not change substantially. This matches the convergence property that was shown earlier, as well as the bias property of the estimator that is mentioned in [6]. Some more comparisons between the simulation and the approximation will be presented later, in the course of the comparison with the chi-square test.

### C. Usage of B-Spline Functions

A new approach for the estimation of mutual information, especially for continuous data, is given in [7]. Each observation is not counted in just one bin as before, but in several bins using B-spline functions as weighting functions. These counts, which are not restricted to integers any more, are used to calculate the estimator as before. For two-dimensional data, the weighting functions for both dimensions are multiplied to obtain the weighting for each observation. The recursive definition of the functions is presented in [7]. For $M = 5$ bins and spline order $k = 3$, these functions $B_{i,k}(z)$ are shown in Figure 5.

To compare this approach to the conventional estimator $\hat{I}_N(X;Y)$, a continuous example is constructed. In the independent case, $(X,Y)$ is assumed to be uniformly distributed on $[0,1]^2$, while in the dependent case the joint probability density function is

$$f_{XY}(x,y) = (1 + (2x-1)(2y-1))\, \mathrm{I}_{[0,1]^2}(x,y)$$

on the same support. In both cases the marginal distributions of $X$ and $Y$ are uniform distributions on $[0,1]$. Figure 6 shows the results for 100 samples and five bins for $X$ and for $Y$. The conventional estimator (the same as before, of course applied to the observations after discretization by binning) is denoted by $k = 1$, while $k = 3$ marks the spline estimator. The value $\beta = 0$ designates the independent case, and $\beta = 1$ the dependent case. While the error level for the conventional estimator is $0.2305$ $(\pm 0.0004)$, the spline approach achieves

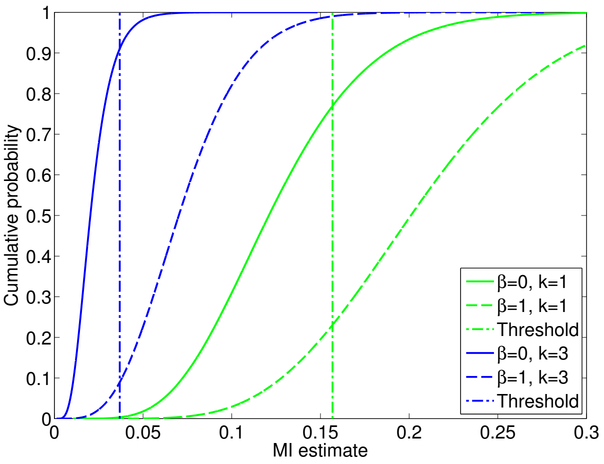

Fig. 6. Comparison for Continuous Data

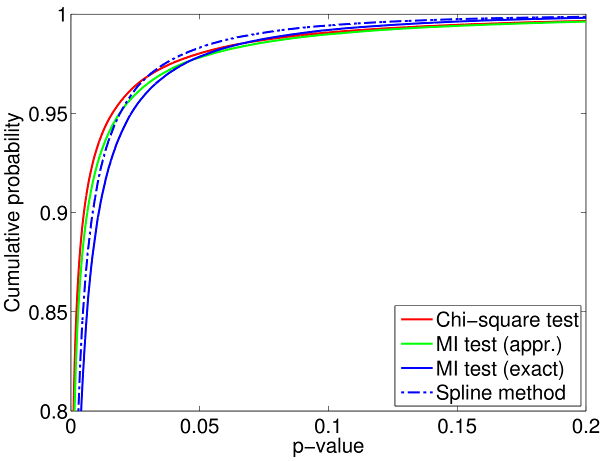

Fig. 7. Comparison for Discrete Data

an error level of $0.0896$ ($\pm 0.0003$), which is clearly better.

*1) Application to Discrete Data:* To apply this new method to discrete data, the observations have to be interpreted as continuous. An observation falling into a certain bin has to be assigned some value, rationally a value that is inside this bin. Here, two ideas are examined. The first one is to assign the central point of the bin deterministically to each observation, the second one is to assign a value somewhere in the bin, drawn according to a rectangular distribution. The second idea proved clearly worse in all examined examples, thus it was not pursued any further.

For the same example that was already used before (ternary genotype $X$, binary phenotype $Y$, 1000 samples), Figure 7 compares four different methods. As their test-statistics can not be compared directly, the distribution of the p-value for the dependent case is used instead. This value reflects the probability that an observation at least as unbalanced as the actual sample appears if the null hypothesis (independence) is valid. While the p-value would be uniformly distributed on $[0, 1]$ for independent samples, smaller p-values appear with higher probabilities for dependent samples, as the null hypothesis seems less likely. The higher the probabilities for small p-values, the better the test works for deciding whether the samples are independent or not.

The first method is the well-established chi-square test for independence, which was already extensively used in genetic studies, e.g. in [8]. The second and the third are both applications of the mutual information approach with splines, once the distribution of the estimator for the independent case was approximated, and once it was simulated, too. Using the approximation, this approach performs slightly worse than the chi-square test, otherwise it performs better, at least for p-values of about 0.07 and higher. For the fourth method, the spline method, the distribution of the estimator in the independent case had of course to be simulated, as there is no approximation known. This approach is clearly the best one in a wide range of p-values, including the value 0.05 which is often relevant.

*2) Application to Mixed Data:* In the mixed model, $X$ represents a ternary genotype as before, with the same distribution as in the example above. The random variable $Y$ now represents a continuous phenotype trait. In the independent case, $Y$ is uniformly distributed on $[0, 1]$ for all genotypes. In the dependent case, the density of $Y$ is

$$f_Y(y) = (1.18 - 0.36y)\mathrm{I}_{[0,1]}(y)$$

for the genotype CC and

$$f_Y(y) = (0.68 + 0.64y)\mathrm{I}_{[0,1]}(y)$$

for the genotypes CT/TC and TT, with $\mathrm{I}_{[0,1]}(y)$ denoting the indicator function. Again, the marginal distributions in the independent and the dependent case are identical. The number of bins for $Y$ is ten, the number of samples is 1000.

The results for this example are shown in Figure 8. Here, the mutual information approach with approximation performs clearly better than the chi-square test, and also outperforms the same approach without approximation in the important range below 0.25. Our new spline approach performs much better than the former methods. Applying the spline method just to the continuous phenotype trait $Y$ already delivers an impressive enhancement, but some more is possible when the spline method is applied to the discrete genotype $X$, too.

## III. Methods

### A. Simulation

To obtain the distribution of the MI estimator, a simple simulation was used. Given the distribution of $X$ and $Y$, the number of samples $N$ and other details about the procedure (e.g. if the conventional or the spline estimator is used), the samples are drawn according to the given distribution, and the estimator is computed. This is repeated several times to gain certain knowledge about the distribution of the estimator. For the simulations in this work, one million cycles of drawing samples and computing the estimator were used. The observed values were stored in a binary tree to allow a sorted output.

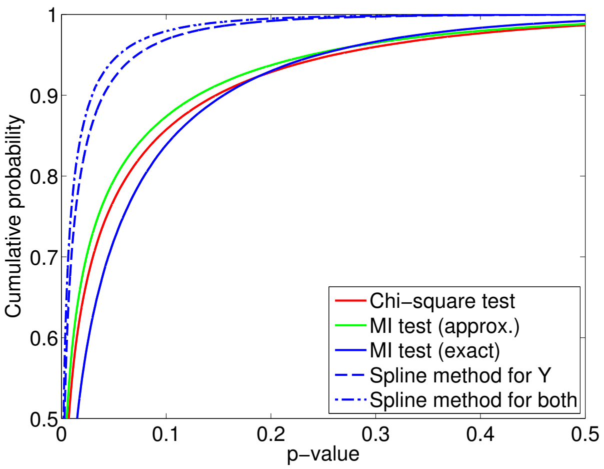

Fig. 8. Comparison for Mixed Data

*1) Error Analysis:* Whenever data is obtained by simulation instead of exact calculation, the accuracy of the simulation is a point that has to be considered. Most of the curves represented here show the cumulative distribution function of the estimator, this means at each point the value of the curve represents the $p$-value, the probability that the MI estimator does not exceed this point. The estimator for this $p$-value is the proportion of the simulation runs in which the MI estimator did not exceed the point.

When $p$ is the true $p$-value at some point and $N'$ is the number of simulation runs, the number of runs $n$ in which the value is not exceeded is $n \sim \mathrm{B}(N', p)$ distributed with expectation $N' \cdot p$ and variance $N' \cdot p \cdot (1-p)$. Thus the estimator $\tilde{p} = n/N'$ has the expectation $\mathrm{E}(\tilde{p}) = p$, the variance

$$\mathrm{Var}(\tilde{p}) = \frac{\mathrm{Var}(n)}{N'^2} = \frac{p \cdot (1-p)}{N'}$$

and hence the standard deviation

$$\sigma_{\tilde{p}} = \sqrt{\mathrm{Var}(\tilde{p})} = \sqrt{\frac{p \cdot (1-p)}{N'}}.$$

As the simulations were done with one million runs each, the standard deviation reaches a maximum of 0.0005 for $p = 0.5$ (and tends to zero when $p$ tends to zero or one). This is far below the accuracy of the plotted curves, thus error bars would be simply invisible if they had been provided.

To verify the error analysis, one of the simulations for Figure 2 (1000 dependent samples) was run a second time, the difference of the values from these independent runs is shown in Figure 9. For two independent estimates $\tilde{p}$ and $\tilde{p}'$, each computed by one million runs, the standard deviation of their difference is approximately

$$\sigma_{\tilde{p}'-\tilde{p}} = \sqrt{\frac{\tilde{p} \cdot (1-\tilde{p}) + \tilde{p}' \cdot (1-\tilde{p}')}{10^6}}.$$

The boundaries given by the onefold standard deviation are also plotted in the figure, the outlier percentage seems reasonable.

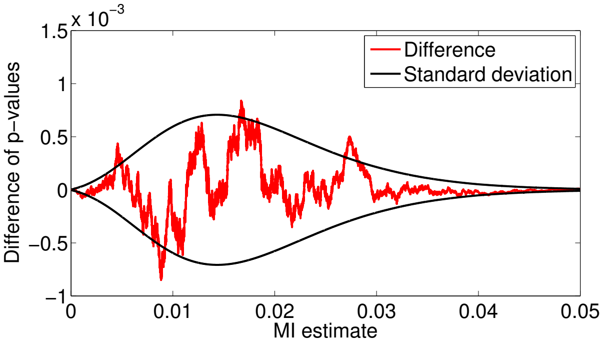

Fig. 9. Error Analysis

## IV. Conclusion

We have seen that there exists an estimator for the mutual information which converges towards the actual value. The distribution of the estimator is not yet known, but an appropriate approximation is available. Thus, the application of the mutual information as a measure of dependency between some drawn samples is feasible, since confidence intervals for hypothesis testing can be calculated.

A new approach using B-spline functions was presented, and it performed better than the conventional approach in the examples presented here, even for discrete data where no additional information is available. Developing some more theoretical background (for example an approximative distribution of the estimator) for the B-spline method leaves room for future work.

## Acknowledgment

This work was supported by the UMIC Research Centre at RWTH Aachen University and by DFG grant SCHM 2643/4-1.

## References

[1] Dawy Z, Goebel B, Hagenauer J, Andreoli C, Meitinger T, Mueller J: **Gene mapping and marker clustering using Shannon's mutual information**. *Computational Biology and Bioinformatics, IEEE/ACM Transactions on* 2006, **3**:47 –56.

[2] Sarkis M, Goebel B, Dawy Z, Hagenauer J, Hanus P, Mueller J: **Gene mapping of complex diseases - A comparison of methods from statistics, information theory, and signal processing**. *Signal Processing Magazine, IEEE* 2007, **24**:83 –90.

[3] Hanus P, Goebel B, Dingel J, Weindl J, Zech J, Dawy Z, Hagenauer J, Mueller JC: **Information and communication theory in molecular biology** 2007, [[http://dx.doi.org/10.1007/s00202-007-0062-6]].

[4] Cover TM, Thomas JA: *Elements of information theory*. John Wiley and Sons, Inc. 1991.

[5] Goebel B, Dawy Z, Hagenauer J, Mueller J: **An approximation to the distribution of finite sample size mutual information estimates**. In *IEEE International Conference on Communications (ICC 2005), Volume 2* 2005:1102 – 1106 Vol. 2.

[6] Paninski L: **Estimation of Entropy and Mutual Information**. *Neural Computation* 2003, **15**(6):1191–1253, [[http://neco.mitpress.org/cgi/content/abstract/15/6/1191]].

[7] Daub CO, Steuer R, Selbig J, Kloska S: **Estimating mutual information using B-spline functions - an improved similarity measure for analysing gene expression data**. *BMC Bioinformatics* 2004, **5**:118, [[http://dx.doi.org/10.1186/1471-2105-5-118]].

[8] de Bakker PIW, Yelensky R, Pe'er I, Gabriel SB, Daly MJ, Altshuler D: **Efficiency and power in genetic association studies**. *Nature Genetics* 2005, **37**(11):1217 – 1223, [[http://dx.doi.org/10.1038/ng1669]].