Statistical Modelling for Data from Experiments with Short Hairpin RNAs

Frank Klawonn[1,2], Torsten Wüstefeld[3,4], and Lars Zender[3,4]

[1] Department of Computer Science
Ostfalia University of Applied Sciences
Salzdahlumer Str. 46/48, D-38302 Wolfenbuettel, Germany
[2] Bioinformatics and Statistics
Helmholtz Centre for Infection Research
Inhoffenstr. 7, D-38124 Braunschweig, Germany
[3] Chronic Infection and Cancer
Helmholtz Centre for Infection Research
Inhoffenstr. 7, D-38124 Braunschweig, Germany
[4] Gastroenterology, Hepatology and Endocrinology
"Rebirth" Cluster of Excellence
Hannover Medical School
Carl-Neuberg-Str. 1, 30625 Hannover, Germany

Abstract. This paper delivers an example of applying intelligent data analysis to biological data where the success of the project was only possible due to joint efforts of the experts from biology, medicine and data analysis. The initial and seemingly obvious approach for the analysis of the data yielded results that did not look plausible to the biologists and medical doctors. Only a better understanding of the experimental setting and the data generating process enabled us to develop a more suitable model for the underlying experiments and to provide results that are coherent with what could be expected from our knowledge and experience.

The data analysis problem we discuss here is the identification of significant changes in experiments with short hairpin RNA. A simple Monte Carlo test yielded incoherent results and it turned out that the assumptions on the underlying experiments were not justified. With a Bayesian approach incorporating necessary prior knowledge from the biologists, we could finally solve the problem.

1 Introduction

A fundamental part of intelligent data analysis [1] is the combination of expertise in data analysis and in the domain from which the data originate. Both partners, the data analysis expert and the data expert must cooperate and develop a basic understanding of the other's scientific field. This is usually a learning process that takes time and can lead to failures in the initial phase that are seldom reported.

This paper describes an application where biological data from the so called third generation microRNA based shRNA (shRNAmir) technology were analyzed. The data come from mouse experiments and a goal of this project is the characterization of new

P.R. Cohen, N.M. Adams, and M.R. Berthold (Eds.): IDA 2010, LNCS 6065, pp. 79–90, 2010.
© Springer-Verlag Berlin Heidelberg 2010

cellular signalling networks that are essential for regenerative processes of the liver. The results can lead to new pharmacological strategies for the treatment of patients with chronic liver damages.

The data are similar, but not identical to standard microarray experiments. In microarray experiments, measurements for two or more conditions are taken for the expression of genes and one is interested in identifying those genes with a significant change of expression. In contrast to standard microarray experiments our measurements are based on simple counting procedures requiring a statistical evaluation which is not as obvious as it seems at first sight.

This paper describes the whole process of modelling the problem with failures and success as a case study in intelligent data analysis. Section 2 describes the biological background and Section 3 provides a more abstract and formal definition of the problem. Sections 4 and 5 discuss two approaches that failed to explain the observed data in the end, leading to the finally successful model derived in Section 6. The final conclusions address open problems and future work that will be based on the ideas described in this paper.

2 Biological Background

Our research group is taking advantage of genetic approaches to study the regulation of liver regeneration. The liver has a tremendous potential to regenerate upon tissue damage by toxins or infection. It is unique that, in contrast to many other epithelial organs, differentiated hepatocytes, which normally reside in the G0 phase of the cell cycle, can, upon liver damage, re-enter the cell cycle and give rise to new hepatocytes. However, when chronic liver damage occurs (e.g. chronic viral hepatitis), there is eventually an exhaustion of the regenerative capacity of hepatocytes and only partial compensation by a stem cell compartment (bipotential liver progenitor cells). The consequence is chronic liver failure, which represents a major health problem worldwide. A unique system for conducting multiplex in vivo RNA interference (RNAi) screens for new positive and negative regulators of liver regeneration was developed. Combining a well characterized mouse model of liver repopulation with third generation microRNA based short hairpin RNA (shRNAmir) technology, we show that mouse livers can be stably repopulated with complex shRNAmir libraries [2,3]. RNA interference is a naturally occurring process, where the presence of double stranded RNA leads to a targeted degradation of a cellular messenger RNA which is sequence complementary to one of the two RNA strands. Since its discovery RNA interference is being used routinely to knock down any gene of choice in vitro as well as in vivo. The RNAi pathway can be harnessed in experimental systems by introducing shRNAs into a cell, which after processing by the internal enzymatic machinery releases a double stranded RNA such as an siRNA, which finally releases one strand. This strand can find a sequence complementary messenger RNA and triggers the degradation of the respective messenger RNA, thus reducing or abolishing the amount of corresponding protein.

Using our in vivo RNAi screening platform, we are characterizing new cellular signalling networks which regulate the proliferation of hepatocytes during chronic liver damage. It is the ultimate goal of our work to translate the obtained genetic information

into new pharmacological strategies which can increase the liver's regenerative potential during chronic liver damage. Such therapies are holding the great promise to prolong patients' survival until they are eligible for definite treatment by liver- or hepatocyte transplantation.

For our experiments we used the mouse as a model organism. The mouse genome consists of approximately 30000 genes. For this study we used a focused shRNA library with 631 shRNAs targeting 301 genes. Therefore we have in average a coverage of 2 shRNAs per gene. The 301 genes were chosen based on frequent deletions in human HCCs (hepatocellular cancers).

631 shRNAs were introduced into mouse livers. The first half of those livers (n=6) were harvested directly after intrahepatic shRNA delivery. The second half of the population (n=6) underwent a protocol for chronic liver damage (intraperitonal CCl4 treatment) after shRNA delivery into the livers was accomplished. CCl4 induces cell death with subsequent compensatory proliferation of surviving hepatocytes. As mentioned above, in this setting hepatocytes containing an shRNA which confers a proliferative advantage will expand, whereas hepatocyes containing an shRNA whose gene knockdown confers a disadvantage under the conditions of chronic liver damage will be reduced in number over time. To quantify the representation of each shRNA in the whole population, we are using a PCR amplification protocol of all shRNAs in the population. PCR products containing the individual shRNA sequences are then subjected to deep sequencing. In average we are applying 8 - 12 million sequence reads per biological sample. Deep sequencing analysis yields the total number of sequence reads for each hairpin, which together with the total number of applied reads can be used to calculate the percent of representation for each shRNA in the population. If this procedure is done for the starting population (livers directly after shRNA delivery) and for the population after manipulation, both populations can be compared to find out whether a certain shRNA is enriched, stays unchanged or is depleted in the system. However, a straight forward analysis of shifts in shRNA representation is hampered by the fact, that strong changes of single hairpins mask smaller changes or suggest changes in unchanged hairpins. Therefore we needed to establish a specific analysis method for this approach to take the experimental setting into account.

The newly used statistical approach helped us to define bona fide candidates. Already preliminary experiments verified, that one highly enriched hairpin influences the hepatocyte proliferation under chronic liver damage in a positive way, recognized by several biological parameters, like survival.

3 Problem Formalization

In our experiments, short hairpin RNA (shRNA) [4] is attached to genes. Most of the genes will be marked by one specific hairpin, but some of the genes can also be marked by more than one hairpin. This is not just redundancy, but also related to different functions of the gene. We use a few hundred different types of hairpins in our experiments. The number of different types of hairpins will be denoted by h. We deal with a pool of more than 10^{12} genes. Some hairpins can be easier adapted to the corresponding genes, for others it is more difficult. Therefore, when marking the genes with the hairpins, we

cannot say in advance, how successful the process is for the different types of genes. Therefore, we draw a sample – the sample size is usually a few million – from the pool of 10^{12} genes and count, how often we find each of the hairpins. Let m_i be the counts for hairpin i ($i \in \{1, \ldots, h\}$). The sample size is therefore

$$m = \sum_{i=1}^{h} m_i.$$

From the theoretical point of view, we draw m balls (genes) from an urn with more than 10^{12} balls of h different colours (types of hairpins) without replacement. Due to the large number of genes in the initial pool compared to the sample we draw, we can neglect the fact that we draw the sample without replacement and consider it as an experiment with replacement. In this way, we can assume that our sample originates from a multinomial distribution with h possible outcomes. We do not know the probabilities for the outcomes, but we draw a sample of size m. Of course, we could estimate these probabilities by $\hat{p}_i = \frac{m_i}{m}$.

After some time, the distribution of the hairpins might have changed and we repeat the experiment again. We do not necessarily draw a sample of exactly the same size. We draw now a sample of size n instead of m from the possibly changed multinomial distribution. We could estimate the probabilities for this multinomial distribution in the same way as before as $\hat{q}_i = \frac{n_i}{n}$ where n_i is now the count for hairpin i for the second sample. This implies $n = \sum_{i=1}^{h} n_i$.

We are now interested in those hairpins i for which the numbers have changed significantly, corresponding to up- or down-regulated genes.

4 The Seemingly Obvious Statistical Model and a Monte Carlo Test

In order to identify those hairpins for which the number has changed significantly from the initial to the final sample, we could apply a statistical test with the null hypothesis that the initial and the final sample originate from multinomial distributions with the same underlying probabilities, i.e. the null hypothesis would be $p_i = q_i$ for all $i \in \{1, \ldots, h\}$.

This test can be easily implemented as a Monte Carlo test [5]. We choose the combination of probabilities for the multinomial distribution that would generate the two samples with highest probability. The maximum likelihood estimator for this problem is obtained by joining the two samples and estimate the probabilities as $\hat{r}_i = \frac{m_i + n_i}{m+n}$. Then we draw two samples from a (pseudo-)random number generator for this multinomial distribution of size m and n. We now obtain simulated estimations $\hat{p}_i^{\text{sim}}$ and $\hat{q}_i^{\text{sim}}$ and can compare these with the estimates $\hat{p}_i$ and $\hat{q}_i$ from the original sample. If

$$\hat{p}_i^{\text{sim}}, \hat{q}_i^{\text{sim}} \in [\hat{p}_i, \hat{q}_i] \quad \text{or} \quad \hat{p}_i^{\text{sim}}, \hat{q}_i^{\text{sim}} \in [\hat{q}_i, \hat{p}_i] \tag{1}$$

holds, then $\hat{p}_i^{\text{sim}}$ and $\hat{q}_i^{\text{sim}}$ variate less than $\hat{p}_i$ and $\hat{q}_i$. In other words, if this is not the case, the difference between $\hat{p}_i$ and $\hat{q}_i$ can be explained by simple random variations in the two samples from the multinomial distribution with the same probabilities.

Of course, we have to repeat this test a large number of times, say 100,000 times. We can then check, how often in these 100,000 simulations condition (1) is satisfied for each hairpin. The proportion of the simulations where this condition is satisfied can be viewed as a (simulated) p-value. We carry out multiple testing here, since we run the test for all h hairpins in parallel. Therefore, a correction for multiple testing must be incorporated into the p-values. We use the simple Bonferroni correction [6] where we have to multiply the obtained p-values with the number of tests we have carried out, i.e. with the number of hairpins h.

Even after Bonferroni correction, more than 90% of the hairpins have a p-value smaller than 0.001. That would mean that more than 90% of the hairpins (or genes) have changed significantly from the initial sample to the final sample. This is in contradiction to all experiences biologists have and does not seem plausible. But what could cause this effect?

To explain this effect, we have to go back to our initial considerations that we do actually draw our samples from very large ($> 10^{12}$), but finite hairpin pools (or poulations). We have made the implicit assumption that the overall size of the pool remains stable which is an incorrect assumption. In order to illustrate the effect of a changing pool size, let us consider a simplified example with much smaller samples and hairpin pools. Assume, our original hairpin pool contains only three different types of hairpins, 1000 of each. So we have diminished the pool size to 3000 instead of the original more than $> 10^{12}$ hairpins in the pool. We draw a sample of size 30 from this pool. In the ideal case, we would obtain 10 representatives from each type of hairpin. Now assume that before we draw the final sample, the first and the second type of hairpin have not changed their quantity and remain at the level of 1000. But the third type of hairpin has increased from 1000 to 4000. So the final sample will be drawn from a pool of hairpins with 1000, 1000 and 4000 replicates from each type. If the final sample has the same size as the initial sample, in our example 30, we would expect in the ideal case to draw 5 hairpins of the first, 5 of the second and 20 of the third type of hairpin. So the counts for the initial sample were (10,10,10) and for the final sample (5,5,20) giving the impression that the quantities of all hairpins have changed (under the wrong assumption that the size of the pool has not changed).

For our real-world data this would mean that if a single hairpin with a high number in the initial sample would change significantly in quantity, the proportions of all other hairpins will be affected, even though they might not have changed in quantity. Therefore, we must take a possible change of the hairpin pool size in our model into account.

5 A Modified Approach

Assume the initial hairpin pool contains k_i^{init} replicates of hairpin i. We do not know these numbers and cannot even estimate them from the sample because we do not know the overall pool size $k^{\text{init}} = \sum_{i=1}^{h} k_i^{\text{init}}$. Since the samples we draw are quite large, we can at least assume that

$$\hat{p}_i \approx \frac{k_i^{\text{init}}}{k^{\text{init}}}$$

holds. The same applies to the final sample that contains the unknown number of k_i^{final} replicates of hairpin i. But we can also assume that

$$\hat{q}_i \;\approx\; \frac{k_i^{\text{final}}}{k^{\text{final}}} \tag{2}$$

holds where $k^{\text{final}} = \sum_{i=1}^{h} k_i^{\text{final}}$

Assume that hairpin i changes from the initial to the final sample by the (unknown) regulation factor c_i, i.e. $k_i^{\text{final}} = c_i k_i^{\text{init}}$. With equation (2), we obtain

$$\hat{q}_i \;\approx\; \frac{c_i k_i^{\text{init}}}{\sum_{j=1}^{h} c_j k_j^{\text{init}}}. \tag{3}$$

When we extend the right hand side of equation (3) by the factor $\frac{\frac{1}{k^{\text{final}}}}{\frac{1}{k^{\text{final}}}}$, we get

$$\hat{q}_i \;=\; \frac{c_i \hat{p}_i}{\sum_{i=1}^{h} c_j \hat{p}_j} \qquad (i \in \{1, \ldots, h\}) \tag{4}$$

where we have replaced approximately in equation (2) by equal. We should choose the regulation factors c_i in such a way that equation (4) is satisfied.

Without any restrictions on the regulation factors c_i, one possible solution would be $c_i = \frac{\hat{q}_i}{\hat{p}_i}$. But this would mean that we explain the changes in the relative frequencies of the hairpins in the two samples by assuming that each hairpin has changed proportionally to the change of the measurements which does not go along with the considerations and the simple example we have provided in the previous section.

From the experience of the biologists we know that most of the regulation factors should be roughly 1. Therefore, we should try to find a solution for the c_i with as little deviations from 1 as possible. This can be formulated as an optimization problem. Minimize the objective function

$$L(c_1, \ldots, c_h) \;=\; \sum_{i=1}^{h} (1 - c_i)^2 \tag{5}$$

under the constraints (4).

To solve this problem, we replace all variable c_i in the objective function (5) by using equation (4) from which we obtain

$$\frac{\hat{q}_i}{\hat{q}_j} \;=\; \frac{c_i \hat{p}_i}{c_j \hat{p}_j}.$$

This implies

$$c_i \;=\; \frac{\hat{p}_j \hat{q}_i}{\hat{p}_i \hat{q}_j} c_j$$

and for $j = 1$, we finally get

$$c_i \;=\; \frac{\hat{p}_1 \hat{q}_i}{\hat{p}_i \hat{q}_1} c_1. \tag{6}$$

This simplifies the objective function (5) to

$$L \;=\; (1-c_1)^2 + \sum_{i=2}^{h}\left(1-\frac{\hat{p}_1\hat{q}_i}{\hat{p}_i\hat{q}_1}c_1\right)^2 \;=\; \sum_{i=1}^{h}\left(1-\frac{\hat{p}_1\hat{q}_i}{\hat{p}_i\hat{q}_1}c_1\right)^2. \tag{7}$$

In order to find the minimum of this quadratic function, we compute the root of the derivative.

$$\frac{dL}{dc_1} \;=\; -2\sum_{i=1}^{h}\left(1-\frac{\hat{p}_1\hat{q}_i}{\hat{p}_i\hat{q}_1}c_1\right)\frac{\hat{p}_1\hat{q}_i}{\hat{p}_i\hat{q}_1} \;=\; -2\sum_{i=1}^{h}\left(\frac{\hat{p}_1\hat{q}_i}{\hat{p}_i\hat{q}_1}-\frac{\hat{p}_1^2\hat{q}_i^2}{\hat{p}_i^2\hat{q}_1^2}c_1\right) \;=\; 0 \tag{8}$$

This leads to

$$c_1 \;=\; \frac{\hat{q}_1}{\hat{p}_1}\cdot\frac{\sum_{i=1}^{h}\frac{\hat{q}_i}{\hat{p}_i}}{\sum_{i=1}^{h}\frac{\hat{q}_i^2}{\hat{p}_i^2}}. \tag{9}$$

With equation (6) we obtain the solution

$$c_i \;=\; \frac{\hat{q}_i}{\hat{p}_i}\cdot\frac{\sum_{j=1}^{h}\frac{\hat{q}_j}{\hat{p}_j}}{\sum_{j=1}^{h}\frac{\hat{q}_j^2}{\hat{p}_j^2}}. \tag{10}$$

From this equation it is clear that the regulation factors c_i only depend on the ratios of the relative frequencies $\hat{p}_i$ and $\hat{q}_i$, but not on the absolute frequencies. Therefore, a change from an initial count for hairpin i of $m_i = 2$ to a final count of $n_i = 4$ would be treated in the same way as a change from $m_i = 20,000$ to $n_i = 40,000$. But it is obvious that the chance that the change from $m_i = 20,000$ to $n_i = 40,000$ is a pure random effect is much lower than for the change from $m_i = 2$ to $n_i = 4$. Therefore, this simple model is also not suitable for our purposes.

6 A Bayesian Maximum Likelihood Approach

The approach described in the previous section has introduced a penalty for regulation factors deviating from 1, representing the idea that most of the expression values of genes (or hairpins) will not change. This actually represents prior knowledge on the regulation factors. Bayesian methods are designed to take such prior knowledge into account. Therefore we develop a Bayesian approach here with a prior that reflects the knowledge that normally the regulation factors will be close to 1.

We slightly change the notation in order to handle up- and down-regulations in a symmetric way. If we just use a factor directly, then up-regulation corresponds to values from the infinite interval $(1, \infty)$, whereas down regulations lie in the finite interval $[0, 1)$. Therefore, we use the parametrization e^{c_i} for the regulation factors. In this way, up-regulation is equivalent to $c_i \in (0, \infty)$ and down-regulation to $c_i \in (-\infty, 0)$, so that up- and down-regulation are just a matter of the sign.

We want to estimate the values for the regulation factors. As mentioned before, we have prior knowledge about the possible values for the regulation factors. This knowledge will be given by a prior distribution $f_{\text{prior}}(x)$. We assume that the priors are independent and that all hairpins have the same type of prior. How we choose the prior, will be discussed later on.

Let us assume that the estimates $\hat{p}_i = \frac{m_i}{m}$ for the probabilities of the multinomial distribution in the first sample are more or less correct. This means we have $\ell \cdot m_i$ hairpins in our original pool of more than 10^{12} hairpins. The constant ℓ is unknown, but independent of the hairpin i. Given the true, but unknown regulation factors e^{c_i}, we find a proportion of

$$q_i \;=\; \frac{e^{c_i} \cdot \ell \cdot m_i}{\sum_{j=1}^{h} e^{c_j} \cdot \ell \cdot m_j} \;=\; \frac{e^{c_i} \cdot m_i}{\sum_{j=1}^{h} e^{c_j} \cdot m_j}$$

of hairpin i in the final pool. Then the likelihood for drawing n_i replicates of hairpin i ($i \in \{1, \ldots, h\}$) from our final samples is, including the prior,

$$L(c_1, \ldots, c_h) \;=\; \prod_{i=1}^{h} f_{\text{prior}}(c_i) \cdot q_i^{n_i}. \tag{11}$$

Note that we have omitted the constant factor

$$\binom{n}{n_1! \cdot \ldots \cdot n_h!}$$

that is independent of the values c_i.

The log-likelihood is then

$$\begin{aligned}
\ln(L(c_1, \ldots, c_h)) &= \sum_{i=1}^{h} \left(\ln(f_{\text{prior}}(c_i)) + n_i \cdot \ln(p_i)\right) \\
&= \sum_{i=1}^{h} \left(\ln(f_{\text{prior}}(c_i)) + n_i \cdot c_i + n_i \cdot \ln(m_i) \right. \\
&\qquad \left. - n_i \cdot \ln\left(\sum_{j=1}^{h} e^{c_j} \cdot m_j\right)\right) \\
&= -n \cdot \ln\left(\sum_{i=1}^{h} e^{c_i} \cdot m_i\right) + \sum_{i=1}^{h} n_i \cdot c_i + \sum_{i=1}^{h} \ln(f_{\text{prior}}(c_i)) \\
&\quad + \sum_{i=1}^{h} n_i \cdot \ln(m_i).
\end{aligned} \tag{12}$$

The prior should definitely be a symmetric distribution with mean zero, preferring no regulation at all and treating up- and down-regulations in the same way. We should choose an uninformative prior. There are various concepts of uninformative priors. Based on the principle of maximum entropy [7], we would have to choose a Gaussian prior for which we still have to fix the variance σ^2. There are, of course, other ways to define uninformative priors that are based on maximizing the entropy or the KullbackLeibler divergence of the posterior distribution or on the Fisher information

(Jeffrey's prior). For an overview on Bayesian inference and priors, we refer to [8]. To keep things simple, we stick to a Gaussian prior with mean $\mu = 0$.

$$f_{\text{prior}}(x) \;=\; \frac{1}{\sigma\sqrt{2\pi}} e^{\frac{-x^2}{2\sigma^2}}$$

Inserting this prior into the log-likelihood (12), we obtain

$$\begin{aligned}
\ln(L(c_1,\ldots,c_h)) &= -n \cdot \ln\left(\sum_{i=1}^{h} e^{c_i} \cdot m_i\right) + \sum_{i=1}^{h} n_i \cdot c_i \\
&\quad - \sum_{i=1}^{h} \left(\ln(\sigma) + \frac{1}{2}\ln(2\pi) + \frac{c_i^2}{2\sigma^2}\right) \\
&\quad + \sum_{i=1}^{h} n_i \cdot \ln(m_i) \\
&= -n \cdot \ln\left(\sum_{i=1}^{h} e^{c_i} \cdot m_i\right) + \sum_{i=1}^{h} n_i \cdot c_i \\
&\quad - \frac{1}{2\sigma^2}\sum_{i=1}^{h} c_i^2 \\
&\quad + \left(\sum_{i=1}^{h} n_i \cdot \ln(m_i)\right) - h \cdot \ln(\sigma) - \frac{h}{2} \cdot \ln(2\pi).
\end{aligned} \tag{13}$$

The last line of equation (13) does not depend on the unknown parameters c_i, so that it can be neglected for the maximization of the log-likehood. The log-likelihood (and also the likelihood) is maximized when the function

$$T(c_1,\ldots,c_h) \;=\; -n \cdot \ln\left(\sum_{i=1}^{h} e^{c_i} \cdot m_i\right) + \sum_{i=1}^{h} n_i \cdot c_i - \frac{1}{2\sigma^2}\sum_{i=1}^{h} c_i^2 \tag{14}$$

is maximized.

Determining the maximum of this objective function and in this way obtaining the maximum likelihood estimates for the regulation factors c_i is not an easy task. A closed form solution cannot be provided. We apply a gradient method here. We carry out the gradient method twice, using the two obvious and most extreme initializations. The likelihood function consists of two main parts. The priors that are maximized for $c_i = 0$, which is our first initialization, and the factors $q_i^{n_i}$ which are maximized when we choose the raw regulation factors, i.e.

$$e^{c_i} \;=\; \frac{n_i/n}{m_i/m} \;=\; \frac{\hat{q}_i}{\hat{q}_i} \tag{15}$$

which give our second initialization.

In this way, we obtain two (local) maxima of the likelihood function from the two initializations and in the best case these two local maxima should be more or less identical. This provides also a hint, how much we can trust our result.

We still have to specify the value for σ in our Gaussian prior. We estimate σ based on the raw regulation factors. We compute the values c_i based on equation (15) and estimate the standard deviation from these c_i values. Since we have some very extreme raw regulation factors due to very small counts for some hairpins, we do not estimate the standard deviation by the sample standard deviation, but based on the more robust interquartile range IQR of the values c_i, i.e. $\hat{\sigma} = 1.349 \cdot \text{IQR}$.

7 Results

First of all, it should be mentioned that we carry out Laplace correction [9,10] for the counting. This means that we replace the values m_i and n_i by $(m_i + 1)$ and $(n_i + 1)$, respectively. Of course, this changes the sums m and n to $(m + h)$ and $(n + h)$, respectively. Laplace correction is required, because there are experiments where the initial or the final count for some hairpins is zero. This would require $c_i = \pm\infty$ for our second initialization for the gradient method and would also cause problems in the likelihood function (11) when one of the initial counts m_i is zero. Then the likelihood function would become zero automatically when the corresponding final count n_i is nonzero.

To illustrate how our approach helps to obtain a more realistic picture about the regulation factors, we take a look at results from one of our experiments with $h = 400$ hairpins, an initial sample size of $m = 6,682,558$ and a final sample size of $n = 15,105,284$. Table 1 shows the results for some selected hairpins. The second column shows the initial count of the corresponding hairpin, the third column the final count. The fourth column contains the raw factor according to equation (15). Our maximum likelihood estimates based on the two above mentioned initializations for the gradient method can be seen in the last two columns.

Table 1. Some results from one of our experiments

			estimated factor		
hairpin no.	initial count	final count	raw factor	init. $c_i = 0$	init. $c_i =$ raw factor
1	47	2	-53.12	-18.05	-18.45
2	448	3	-337.55	-108.94	-111.56
3	3940	1534	-5.81	-5.64	-5.79
4	5178	25517	2.18	2.24	2.18
5	18980	43938	1.02	1.05	1.02
6	18385	44546	1.07	1.10	1.07

Negative signs of regulation factors indicate down-regulations. For instance, if equation (15) yields values like 0.5 or 0.25, we would not enter these values in the table, but the values -2.00 and -4.00 instead, respectively.

Most of our data look like the ones in the last two rows where we have more or less no regulation. A certain fraction of the hairpins shows a moderate regulation as for

hairpin 3 and 4 in the table. The unregulated and the moderate (raw) regulation factors are confirmed by our approach.

The first two entries are more extreme concerning the regulation factors. Such extreme regulation factors can only occur when at least one of the two counts for the hairpin is comparatively small. Some of these extreme cases are very interesting from the biological and medical point of view. Although our approach still yields very large regulation factors, they are downsized to roughly one third compared to the raw factors.

Comparing the last two columns, the gradient method seems to yield quite similar results for the two extremely different initializations.

As mentioned already in Section 2, the statistical evaluation helped us to define bona fide candidate hairpins that influence the hepatocyte proliferation under chronic liver damage in a positive way.

8 Conclusions

We have a presented a typical experience in intelligent data analysis. In the beginning, the way how to analyze the data seems to be obvious. But it turns out that the initial simplified understanding of the question to be solved by data analysis and the modelling of the process that generates the data were not sufficient to provide suitable answers. Only with the joint expertise, in our case from biology, medicine and data analysis, a solution can be found in the end.

Our project is still in an initial phase. We are now in the process of analyzing data from repeated experiments and need to find out what causes sometimes extreme variations between experiments.

Apart from the estimation of the regulation factors that we have presented in this paper, we are now developing methods to compute confidence intervals for them.

We are also interested in using other priors. But a sensitivity analysis of our Gaussian prior with the respect to the parameter (standard deviation) σ has shown that the results do not change significantly when we vary σ in a reasonable range. Therefore, we would not expect significant changes when we use other priors.

References

1. Hand, D.J., Berthold, M. (eds.): Intelligent Data Analysis: An Introduction, 2nd edn. Springer, Berlin (2009)
2. Dickins, R.A., Hemann, M.T., Zilfou, J.T., Simpson, D.R., Ibarra, I., Hannon, G.J., Lowe, S.W.: Probing tumor phenotypes using stable and regulated synthetic microRNA precursors. Nat. Genet. 37, 1289–1295 (2005)
3. Silva, J.M., Li, M.Z., Chang, K., Ge, W., Golding, M.C., Rickles, R.J., Siolas, D., Hu, G., Paddison, P.J., Schlabach, M.R.: Second-generation shRNA libraries covering the mouse and human genomes. Nat. Genet. 33, 1281–1288 (2005)
4. Paddison, P., Caudy, A., Bernstein, E., Hannon, G., Conklin, D.: Short hairpin rnas (shrnas) induce sequence-specific silencing in mammalian cells. Genes Dev. 16, 948–958 (2002)
5. Zhu, L.: Nonparametric Monte Carlo Tests and Their Applications. Springer, New York (2005)
6. Shaffer, J.P.: Multiple hypothesis testing. Ann. Rev. Psych. 46, 561–584 (1995)

7. Jaynes, E.T.: Probability Theory: The Logic of Science. Cambridge University Press, Cambridge (2003)
8. O'Hagan, A., Forster, J.: Bayesian Inference, 2nd edn. Oxford University Press, Oxford (2003)
9. Cestnik, B.: Estimating probabilities: A crucial task in machine learning. In: Aiello, L.C. (ed.) Proceedings of the ninth European Conference on Artificial Intelligence, pp. 147–149 (1990)
10. Good, I.J.: The Estimation of Probabilities: An Essay on Modern Bayesian Methods. MIT Press, Cambridge (1965)