Redefinition of Mutual Information in the Fuzzy Sets Framework for Computational Genomics

Silvana Badaloni, Marco Falda, Paolo Massignan, and Francesco Sambo

Dept. of Information Engineering, University of Padova
Via Gradenigo 6/A - 35131 Padova, Italy
{name.surname}@unipd.it, massignan@tele2.it

Abstract. Mutual Information is a measure of correlation between two discrete random variables: the aim of this work is to provide a new definition of Mutual Information using concepts from Fuzzy Sets theory, to extend it to continuous variables. With this approach, we extended the model on which the well known REVEAL algorithm for Reverse Engineering of gene regulatory networks is based and we designed a new flexible version of it, called FuzzyReveal, able to avoid the loss of information caused by the binarization of the continuous biological variables.The predictive power of our new version of the algorithm is promising, being both significantly higher than the one of REVEAL and comparable with a state-of-the-art algorithm on a set of simulated problems.

1 Introduction

One of the main goals of studies on Genomics is to understand the mechanism of genetic regulation, which can be modelled as a gene regulatory network, a graph in which nodes represent genes or proteins and two or more nodes are connected if a regulatory relation exists between them. A widely used approach for inferring regulatory relations is based on the analysis of the Shannon Entropy and on the Mutual information of gene expression signals. This mechanism constitutes the basis of REVEAL [1], a well-known Reverse Engineering algorithm. This approach exploits a boolean model to represent gene regulatory networks in which each gene is modelled with a boolean variable True/False; its main aim is to gather boolean relations between time series of quantized gene expression values. However, the Boolean model on which the classical REVEAL algorithm is based is limited: a large amount of information is lost, when a real signal is approximated with just the two symbols 0 and 1.

In order to represent a real signal in a symbolic qualitative way, fuzzy methodologies can provide the basis for a more flexible model. In the present paper we will provide a new definition of Mutual Information in the fuzzy framework that will be used to extend in a fuzzy direction the REVEAL algorithm. In [2] the relationship between the notions of probability and fuzziness is deeply studied: in particular, an interpretation of fuzzy set theory in terms of conditional events and coherent conditional probabilities is proposed. We will apply this theory to re-define Mutual Information, which will be used in the core of the REVEAL algorithm: the modified algorithm will be called FuzzyReveal.

K. Madani et al. (Eds.): Computational Intelligence, SCI 343, pp. 73–83.
springerlink.com

The paper is organized as follows: in Section 2 the concept of classical Mutual Information is recalled and the REVEAL algorithm described, in Section 3 first Conditional Probability is defined in terms of membership functions, then Mutual Information is rewritten in the new setting and the classical REVEAL algorithm updated accordingly. Section 4 reports an example of application.

2 Mutual Information and the REVEAL Algorithm

Given a discrete random variable x, taking values in the set X, its Shannon Entropy [3] is defined as

$$H(x) = -\sum_{\bar{x}\in X} p(\bar{x}) \log p(\bar{x}),$$

where $p(\bar{x})$ is the probability mass function $p(\bar{x}) = Pr(x = \bar{x})$, $\bar{x} \in X$. The joint entropy of a pair of variables x, y, taking values in the sets X, Y respectively, is

$$H(x, y) = -\sum_{\bar{x}\in X, \bar{y}\in Y} p(\bar{x}, \bar{y}) \log p(\bar{x}, \bar{y})$$

while the conditional entropy of x given y is defined as

$$H(x|y) = H(x, y) - H(x).$$

The Mutual Information of x, y is defined as $MI(x, y) = H(x) - H(x|y)$ and can be explicitly expressed as

$$MI(x, y) = \sum_{\bar{x}\in X, \bar{y}\in Y} p(\bar{x}, \bar{y}) \log \frac{p(\bar{x}, \bar{y})}{p(\bar{x})p(\bar{y})} \geq 0 \tag{1}$$

When the two variables are independent, the joint probability distribution factorizes and the MI vanishes:

$$p(\bar{x}, \bar{y}) = p(\bar{x})p(\bar{y}) \;\Rightarrow\; MI(x, y) = 0.$$

Mutual Information is therefore a measure of dependence between two discrete random variables and is used by the REVEAL algorithm [1] to infer causal relations between genes: for each gene in the genome, a time series of its rate of expression (called *gene profile*) is gathered from multiple DNA-microarray experiments; an example is depicted in Figure 1, with time samples on the x-axis and intensity of gene expression on the y-axis.

To apply REVEAL algorithm, gene profiles are then quantized in two levels, 0 (underexpressed) and 1 (overexpressed), and Mutual Information is computed between all possible pairs of genes. In the specific case probabilities are computed as the frequencies of the symbols 0 or 1 within a given sequence; since the sum of the probabilities being 0 or 1 must be equal to unity, $p(1) = 1 - p(0)$ and the formula for the entropy becomes:

$$H(x) = -p(0) \cdot \log(p(0)) - (1 - p(0)) \cdot \log(1 - p(0))$$

The joint probability is computed as a the probability of co-occurrence of two symbols.

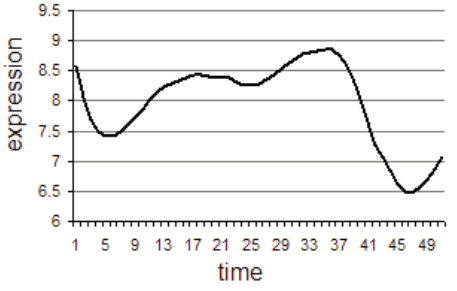


Fig. 1. Example of time series representing the expression of a gene

Example 1. Consider two random variables x and y, representing the quantization of two time series in two levels, 0 and 1; for each variable, consider two sequences of 10 symbols: $x' = \{0, 1, 1, 1, 1, 1, 1, 0, 0, 0\}$ and $y' = \{0, 0, 0, 1, 1, 0, 0, 1, 1, 1\}$. Then for variable x we obtain

$$p(0) = 0.4 \text{ and } p(1) = 0.6 = 1 - p(0)$$

that means 40 % of zeros and 60% of ones respectively. As for joint probabilities, in one case out of 10 $\exists i : x'_i = 0 \wedge y'_i = 0$, therefore

$$p(0, 0) = 0.1$$

The remaining combinations of symbols are $p(0, 1) = 0.3$, $p(1, 0) = 0.4$ and $p(1, 1) = 0.2$.

The algorithm infers causal relations between pairs whose MI is above a given threshold.

3 Fuzzy Extension of the REVEAL Algorithm

The classical REVEAL Algorithm is based on a Boolean model, therefore it has to approximate a real signal with just two symbols 0 and 1; it is clear that in this way much information is lost.

Using the Fuzzy Sets framework it is possible to obtain a more flexible and expressive model.

3.1 Membership Functions and Conditional Probability

In this paper the point of view of Coletti and Scozzafava [4,5] has been adopted.

Let x be a random quantity with range X, the family $\{x = \bar{x}, \bar{x} \in X\}$ is obviously a partition of the sample space Ω [6]; let then φ be any property related to the random quantity x: notice that a property, even if expressed by a proposition, does not single out an event, since the latter needs to be expressed by a non-ambiguous statement that can be either true or false. For this reason the event referred by a property will be indicated with E_φ, meaning "You claim E_φ" (in the sense of De Finetti [6]).

Coletti and Scozzafava state that a membership function can be defined as a Conditional Subjective Probability between two events E_φ and $x = \bar{x}$, meaning that "You believe that E_φ holds given $x = \bar{x}$".

$$\mu_{E_\varphi}(\bar{x}) = P(E_\varphi | x = \bar{x})$$

The membership degree $\mu_{E_\varphi}(\bar{x})$ is just the opinion of a real (or fictitious) person, for instance, a "randomly" chosen one, which is uncertain about it, whereas the truth-value of that event $x = \bar{x}$ is well determined in itself. Notice that conditional probability between events E_φ and $x = \bar{x}$ can be directly introduced rather than being defined as the ratio of the unconditional probabilities $P(E_\varphi \wedge \bar{x})$ and $P(x = \bar{x})$. From the same paper we report also the following example.

Example 2. Is Mary young? It is natural to think that You have some information about possible values of Mary's age, which allows You to refer to a suitable membership function of the fuzzy subset of "young people" (or, equivalently, of "young ages"). For example, for You the membership function may be put equal to 1 for values of the age less than 25, while it is put equal to 0 for values greater than 40; then it is taken as decreasing from 1 to 0 in the interval from 25 to 40. This choice of the membership function implies that, for You, women whose age is less than 25 are "young", while those with an age greater than 40 are not. The real problem is that You are uncertain on being or not "young" those women having an age between 25 and 40: the interest is in fact directed toward conditional events such as $E_{young}|x = \bar{x}$, with

$$E_{young} = \{\text{You claim that Mary is young}\}$$

$$\{x = \bar{x}\} = \{\text{the age of Mary is } \bar{x}\}$$

where $\bar{x}$ ranges over the interval [25, 40]. It follows that You may assign a subjective probability $P(E_{young}|x = \bar{x})$ equal to 0.2 without any need to assign a degree of belief of 0.8 to the event E_{young} under the assumption $x \neq \bar{x}$ (i.e., the age of Mary is not $\bar{x}$), since an additivity rule with respect to the conditioning events does not hold.

3.2 Fuzzy Mutual Information

The objective is to apply Coletti and Scozzafava theory to temporal gene profiles, therefore we introduce a set of properties Φ which describe qualitative aspects of the profiles, such as their "height" (*high*, *low*) or their "growth" (*increasing*, *decreasing*). Notice that the formula for Fuzzy Mutual Information that will be obtained is independent of the specific set chosen.

Exploiting the disintegration formula, the probability $\widetilde{P}$ of a single event for a property $\varphi \in \Phi$ can be written as

$$\begin{aligned}\widetilde{P}(E_{\varphi\in\Phi}) &= \sum_{\bar{x}\in X} P(E_\varphi | x = \bar{x}) \cdot P(x = \bar{x}) \\ &= \sum_{\bar{x}\in X} \mu_{E_\varphi}(\bar{x}) \cdot P(x = \bar{x})\end{aligned}$$

Since the Mutual Information relates two events (in our case relates two gene profiles) let, without loss of generalization, be $\Phi = \{\pi, \rho\}$. In the following we will write E_φ as $x = \varphi$.

The conjunctive probability for $x = \pi \wedge y = \rho$ is now required. According to [2,5] there is not an unique definition for the conditional probability $P(x = \pi \wedge y = \rho | x = \bar{x} \wedge y = \bar{y})$, called in the following p, for brevity. The probability p can assume any value such that

$$\max\{\mu_{E_\pi}(\bar{x}) + \mu_{E_\rho}(\bar{y}) - 1, 0\} \leq p \leq \min\{\mu_{E_\pi}(\bar{x}), \mu_{E_\rho}(\bar{y})\}$$

since it satisfies the coherence hypotheses [2]. Notice that the bounds for p are indeed T-Norms between the membership functions $\mu_{E_\pi}(\bar{x})$ and $\mu_{E_\rho}(\bar{y})$: p may in fact range between the Lukasievicz T-Norm and the minimum; in this work we show the results for the minimum, but good performance was achieved with many other values, such as product, Lukasievicz or the average between Lukasievicz and minimum. The probability that $x = \pi \wedge y = \rho$ can be defined, again in virtue of the disintegration property, as

$$\begin{aligned}
&\widetilde{P}(x = \pi, y = \rho) = \\
&= \sum_{\bar{x} \in X} \sum_{\bar{y} \in Y} P(x = \pi \wedge y = \rho | x = \bar{x} \wedge y = \bar{y}) \\
&\quad \cdot P(x = \bar{x} \wedge y = \bar{y}) \\
&= \sum_{\bar{x} \in X} \sum_{\bar{y} \in Y} p \cdot P(x = \bar{x}, y = \bar{y})
\end{aligned}$$

The Fuzzy Mutual Information function can now be defined in a similar way w.r.t. the one defined in the Probabilistic setting (Formula 1) by replacing the probability distributions P with distributions $\widetilde{P}$ defined according to Coletti-Scozzafava's theory.

Definition 1. *Given two events x and y and a set of symbols Φ their Fuzzy Mutual Information is defined as*

$$\begin{aligned}
&\widetilde{MI}(x, y) = \\
&\sum_{\varphi \in \Phi} \sum_{\varphi' \in \Phi} \widetilde{P}(x = \varphi, y = \varphi') \cdot \log \frac{\widetilde{P}(x = \varphi, y = \varphi')}{\widetilde{P}(x = \varphi) \cdot \widetilde{P}(y = \varphi')}
\end{aligned}$$

This definition will be used to extend the classical REVEAL algorithm.

3.3 The Algorithm

The structure of the FuzzyReveal algorithm is similar to the classic REVEAL, but the extended definition of Fuzzy Mutual Information is used; its pseudo-code is reported in listing Algorithm 1.

Algorithm 1. Fuzzy Reveal

Input: $\mathcal{G} = \{x_1, \ldots, x_G\}$ a set of profile sequences, Φ the set of symbols, N the number of pairs to return
Output: the first N top-rated pairs
begin
  foreach $\bar{g}$ ***in*** $\mathcal{G}$ **do**
    foreach φ ***in*** Φ **do**
      $\triangleright$ compute the membership function of the profile $x = \bar{g}$ w.r.t. the property φ
    end
  end
  $Rank \leftarrow \emptyset$
  foreach x, y ***in*** $\mathcal{G} : x \neq y$ **do**
    foreach π, ρ ***in*** Φ **do**
      $\triangleright$ compute $\widetilde{P}(x = \pi, y = \rho)$
      $\triangleright$ compute $\widetilde{MI}(x, x)$ and $\widetilde{MI}(x, y)$
      $\triangleright$ compute $r_{xy} = \widetilde{MI}(x, y)/\widetilde{MI}(y, x)$
    end
    $Rank \leftarrow Rank \cup r_{ij}$
  end
  $\triangleright$ sort the pairs $\langle x, y \rangle$ according to $Rank$
  return *the first* N *pairs*
end

The parameter N can be set hypothesizing a scale-free topology for the underlying network: scale-free networks are sparse, with a number of edges that usually lies between V and $2V$, where V is the number of nodes [7].

4 Example of Application

The properties that have been considered to evaluate the proposed Fuzzy Mutual Information are:

- the value of the profile x at a given point $\bar{x}$ (high or low);
- the growth behavior of the profile x (increasing or decreasing).

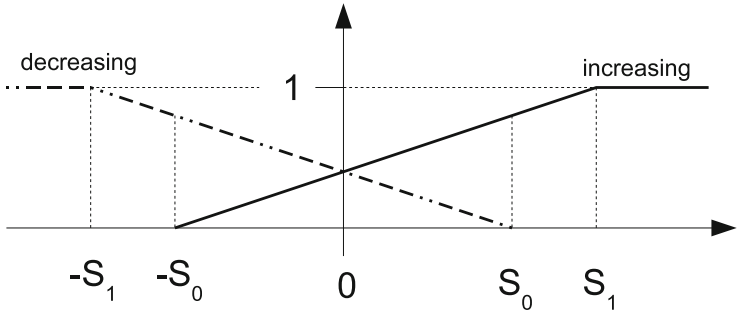
For each of these four events a membership function has been provided.

Definition 2. *the set* Φ' *is the set of qualitative aspects* $\{"high", "low", "increasing", "decreasing"\}$.

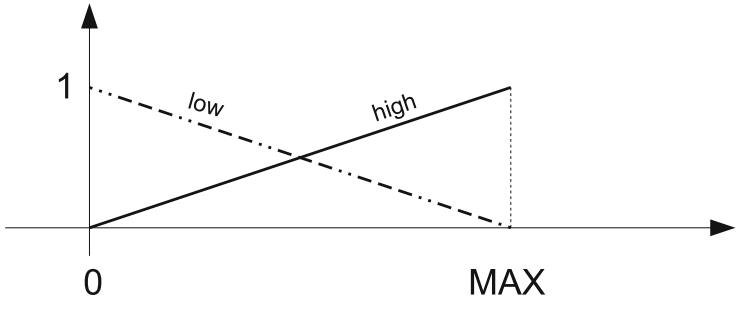
The membership functions for these qualitative aspects have been defined as

$$\mu_{high}(x) = \frac{x}{MAX}$$

$$\mu_{low}(x) = 1 - \mu_{high}(x)$$



(a) definitions for "increasing" and "decreasing".



(b) definitions for "high" and "low".

Fig. 2. Membership functions describing the growth behavior of a profile and its expression

$$\mu_{increasing}(x) = \begin{cases} 1 & \textit{if } x > S_1 \\ \frac{x+S_0}{S_1+S_0} & \textit{if } -S_0 \leq x \leq S_1 \\ 0 & \textit{otherwise} \end{cases}$$

$$\mu_{decreasing}(x) = \begin{cases} 1 & \textit{if } x < -S_1 \\ \frac{x-S_0}{S_1+S_0} & \textit{if } -S_1 \leq x \leq S_0 \\ 0 & \textit{otherwise} \end{cases}$$

where MAX is the maximum among all samples; S_1, S_2 are thresholds that shape the trapezoids (Figure 2), and they are applied to the angular coefficients of the series.

With this approach, a set of numerical values that represent the time series can be quantified using fuzzy levels, as in Figure 3.

Algorithm 1 has been evaluated using the Precision and Recall measures, defined as

$$Precision = \frac{TP}{TP + FP}$$

$$Recall = \frac{TP}{TP + FN}$$

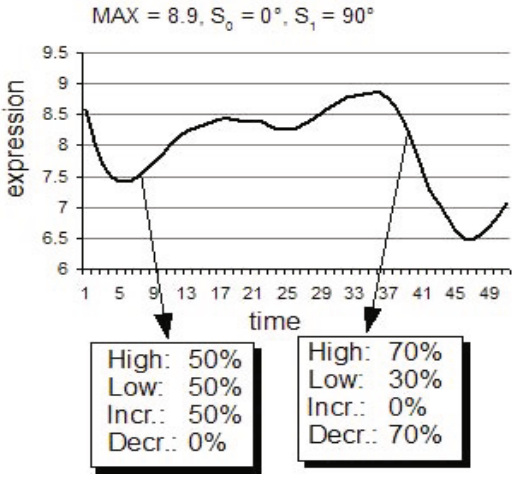


Fig. 3. Fuzzy values for two time series points

where TP represents the number of relations among genes that have been correctly identified by the algorithm (t*rue positives*), FP are the relations found by the algorithm but not representing real relations among genes (*false positives*), and finally FN are the real relations that the algorithm has not been able to find (*false negatives*).

To evaluate and compare the performance of different reverse-engineering approaches, transcriptional networks whose interactions are perfectly known should be used; since at present no biological network is known with sufficient precision to serve as a standard, quantitative assessment of reverse engineering algorithms can be accomplished using synthetic networks [8] or simulation studies [9,10].

To show the application of the new definition of Mutual Information we have generated two datasets for 12 genes and 50 time-points using an *ad hoc* simulator [11]. In Figure 4 the results are shown, together with the performances of a "state-of-the-art" algorithm (Aracne, [12]) and the classical version of REVEAL. It is possible to notice that there is a statistically significant improvement w.r.t. the classical algorithm[1]; this is mainly due to the fact that the classical REVEAL is based on Boolean networks, and so it uses just two values to represent gene expressions, while our approach allows describing time series intensity in a much better way and computing better similarity measures, since a whole range of values from 0 to 1 is used.

The comparison with Aracne is acceptable, since no statistically significant difference is observed in Precision and Recall.

5 Related Works

Soft computing tools, such as fuzzy sets, evolutionary strategies and neurocomputing, have been found to be helpful in providing low cost, acceptable solutions in the presence of various types of uncertainties when analysing gene regulatory networks data [13].

[1] Exact Wilcoxon two-sample tests, p-value < 0.05.

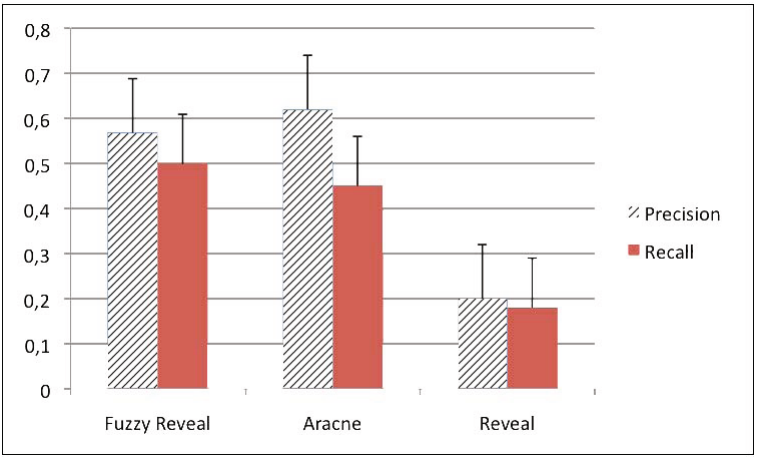


Fig. 4. Precision and recall measures of FuzzyReveal, Aracne and Classical Reveal algorithms

In [14], Mutual Information is computed between quantized profiles of gene expression, which are assigned to a fuzzy set of clusters: each profile can belong to different clusters, with different degrees of preference. More recently, Mutual Information extended using Fuzzy Sets and Rough Sets has been exploited to develop a new concept of equivalence partition matrix that allows for efficiently approximate the true marginal and joint distributions of continuous features from high dimensional microarray gene expression data [15].

Fuzzy rough sets and fuzzy Mutual Information have also been used in [16] for feature reduction, when selecting potential cancer genes from DNA-microarray experiments: given a subset of features, new features are added to the subset if their addition significantly increases the Mutual Information.

A Fuzzy Mutual Information measure is proposed in [17], where the authors follow the approach pioneered by De Luca and Termini; according to De Luca and Termini [18] an entropy function must satisfy four characteristic axioms, and in this way a Mutual Information function can be built. This is similar to what is done by Shannon using Probability theory instead of Fuzzy sets theory.

Other approaches to extend the REVEAL algorithm are present in the literature: [19] avoids the computation of Mutual Information, shifting the paradigm to the domain of *consistent pairs*, *i.e.* pairs $\langle regulators, regulated\ gene \rangle$ for which the same discrete value appears in the regulated gene every time the same combination of values appears in the set of regulators. In [20] this approach is further brought, defining a fitness function for putative causal relations, which allows the algorithm to rank pairs $\langle regulators, regulated\ gene \rangle$ in terms of distance from a consistent pair; the algorithm chooses, for each gene, the pair $\langle regulators, regulated\ gene \rangle$ closer to a consistent pair. The fitness function is specifically designed to tolerate quantization noise and variable regulatory delays.

6 Conclusions

In this paper we have considered the problem of Reverse Engineering and we have applied Coletti and Scozzafava results in order to replace expression profiles with qualitative descriptions. These descriptions are defined on a set of qualitative properties, and can assume different membership degrees w.r.t. a given property. Since the qualitative description comes from a random variable whose domain is finite, all classical results of Information Theory can be applied. We have extended the classical Mutual Information in a fuzzy direction and we have included it in the REVEAL algorithm thus obtaining the FuzzyReveal algorithm.

As for future directions, the application of this approach to real Genomics data will be the next step of the research. This will allow to have a better evaluation of performances with respect to both current biomedical experiments and noise typical of these data sets. A further possible improvement of this study could be the integration of a learning module for the automatic definition of the membership functions that describe the properties of the profiles.

References

1. Liang, S., Fuhrman, S., Somogyi, R.: Reveal: a general reverse engineering algorithm for inference of genetic network architectures. In: Pacific Symposium on Biocomputing, pp. 18–29 (1998)
2. Coletti, G., Scozzafava, R.: Conditional probability, fuzzy sets, and possibility: a unifying view. Fuzzy Sets and Systems 144, 227–249 (2004)
3. Shannon, C.E.: A mathematical theory of communication. The Bell Systems Technical Journal 27, 379–423 (1948)
4. Coletti, G., Scozzafava, R.: Probabilistic Logic in a Coherent Setting. Kluwer Academic Publishers, Dordrecht (2002)
5. Coletti, G., Scozzafava, R.: Conditional probability and fuzzy information. Computational Statistics & Data Analysis 51, 115–132 (2006)
6. de Finetti, B.: Teoria della Probabilità, Einaudi, Torino (1970)
7. Reka, A., Barabási, A.L.: Statistical mechanics of complex networks. Reviews of Modern Physics 74, 47–97 (2002)
8. Grilly, C., Stricker, J., Pang, W.L., et al.: A synthetic gene network for tuning protein degradation in saccharomyces cerevisiae. Mol. Syst. Biol. 3, 127 (2007)
9. Smith, V.A., Jarvis, E.D., Hartemink, A.J.: Evaluating functional network inference using simulations of complex biological systems. Bioinformatics 18, 216–224 (2002)
10. Mendes, P., Sha, W., Ye, K.: Artificial gene networks for objective comparison of analysis algorithms. Bioinformatics 19, 122–129 (2003)
11. Di Camillo, B., Toffolo, G., Cobelli, C.: A gene network simulator to assess reverse engineering algorithms. Ann. N Y Acad. Sci. 1158, 125–142 (2009)
12. Margolin, A.A., Nemenman, I., Basso, K., Wiggins, C., Stolovitzky, G., Dalla Favera, R., Califano, A.: Aracne: an algorithm for the reconstruction of gene regulatory networks in a mammalian cellular context. BMC Bioinformatics 7 (2006)
13. Mitra, S., Hayashi, T.: Bioinformatics with soft computing. IEEE Transactions on Systems, Man, and Cybernetics 36, 616–635 (2006)
14. Zhou, X., Wang, X., Dougherty, E.R., Russ, D., de Suh, E.: Gene clustering based on clusterwide mutual information. Journal of Computational Biology 11, 147–161 (2004)

15. Maji, P., Pal, K.S.: Fuzzy-rough sets for information measures and selection of relevant genes from microarray data. IEEE Transactions on Systems, Man, and Cybernetics. Part B, Cybernetics (2010) (retrieved on April 2010) (to appear)
16. Xu, F., Miao, D., Wei, L.: Fuzzy-rough attribute reduction via mutual information with an application to cancer classification. Computers and Mathematics with Applications (2008)
17. Ding, S.F., Xia, S.X., Jin, F.X., Shi, Z.Z.: Novel fuzzy information proximity measures. Journal of Information Science 33, 678–685 (2007)
18. De Luca, A., Termini, S.: A definition of a non-probabilistic entropy in the setting of fuzzy sets theory. Information and Control 20, 301–312 (1972)
19. Nam, D., Seo, S., Kim, S.: An efficient top-down search algorithm for learning boolean networks of gene expression. Machine Learning 65, 229–245 (2006)
20. Sambo, F., Di Camillo, B., Falda, M., Toffolo, G., Badaloni, S.: CNET: an algorithm for the inference of gene regulatory interactions from gene expression time series. In: Proceedings of the 14th Workshop on Intelligent Data Analysis in Medicine and Pharmacology IDAMAP 2009, Verona, Italy, pp. 23–28 (2009)