Impacts of the Different Spline Orders on the B-spline Association Estimator

Zeyneb Kurt, Nizamettin Aydin, and Gökmen Altay

***Abstract*—Gene Network Inference (GNI) algorithms enable searching the interactions among the several cell molecules. Many application fields such as computational biology and pharmacology utilize the GNI algorithms to illustrate the interaction networks of the cell molecules. Association score estimation is the most crucial step of the GNI applications. B-spline is a popular approach, which efficiently estimates the interaction scores between the variable (gene) pairs. In this study inference performance of the B-spline estimator according to the selected spline order is examined. In addition to evaluating B-spline performance according to the spline order, influences of using a frequently used pre-processing operation Copula Transform on the performance of B-spline is also examined. Conservative Causal Core network (C3NET) GNI algorithm is used in the experiments. At the overall analysis, B-spline estimator with the spline order 2 gave the best inference performance among the selected spline orders from 1 to 10.**

I. Introduction

Gene Network Inference (GNI) algorithms can visualize and illustrate the interactions between the gene pairs or gene product pairs. They are frequently utilized in the genomics applications, which mostly require investigation of gene networks [1-8]. Estimation accuracy of the association score estimators substantially affects the accuracy of the resulting network in the GNI algorithms. Association estimators had been used not only in GNI applications, but also in several application fields such as economy, signal processing, statistics, etc. In this study, B-spline association estimator, which is proposed by Daub et al. [3], is examined rather than GNI algorithms.

B-spline is used in several GNI applications [3, 9-13]. Daub et al., [3] claimed that the performance of the B-spline approach changes by the spline order. However they did not present a comprehensive study regarding performance variation according to the spline order. They only used the spline order from 1 to 5. In this study impact of the changing spline order is examined in a more comprehensive manner by changing the spline order from 1 up to 10. Moreover in this study impacts of a commonly used pre-processing operation on the inference performance of B-spline are investigated. Some of the GNI applications may perform several pre-processing operations before the association estimation step. Copula Transform (CT) is a frequently used pre-processing operation, which utilize the ranking values of the data samples rather than original values. In this study effect of the CT on the performance of B-spline is also searched as mentioned previously. To our knowledge there is no study in the literature, which examines the influences of the spline order and CT together on the estimation performance of the B-spline. We aim to make researchers aware how spline order and CT affects the performance of the B-spline estimator. In this sense, we utilize two different artificial datasets and Conservative Causal Core network (C3NET) [6] GNI algorithm in the experiments. B-spline with using CT outperforms the case without using CT for almost all of the spline orders. Furthermore, the best performed spline order is observed as 2 for both of the cases for almost all of the scenarios.

Organization of the study is as follows, B-spline estimator is explained in Section 2, C3NET GNI algorithm is mentioned in Section 3, experimental results are given in Section 4; finally conclusion is given in Section 5.

II. B-spline Association Estimator

B-spline is a Mutual Information (MI)-based association estimator. MI-based methods can measure any type of the relationships between the variables. Hence they can efficiently measure not only the linear relationships, but also the nonlinear ones. Before the explanation of the B-spline estimator, the terms entropy and MI are mentioned below.

Entropy is an ambiguity measure regarding a random variable, which requires binning (discretization) operation of that random variable. After the discretization, entropy is obtained by using the empirical probability distribution of each bin as given in (1):

$$H(X) = -\sum_{x_i \in X} P(X = x_i)\log\left(P(X = x_i)\right) \tag{1}$$

where H denotes entropy, $P(X = x_i)$ corresponds the probability density of the cell *i*.

MI score between two random variables, *X* and *Y*, is obtained by their individual and joint entropies as given in (2):

$$MI(X,Y) = H(X) + H(Y) - H(X,Y)$$
$$= -\sum_{x_i \in X}\sum_{y_j \in Y} P(X = x_i, Y = y_j)\log\frac{P(X = x_i, Y = y_j)}{P(X = x_i)P(Y = y_j)} \tag{2}$$

Manuscript received June 20, 2013.

Z. Kurt and N. Aydin are with The Department of Computer Engineering at Yildiz Technical University, Istanbul, Turkey, (e-mails: zeyneb@ce.yildiz.edu.tr and naydin@yildiz.edu.tr).

G. Altay is with The Department of Biomedical Engineering at Bahcesehir University, Istanbul, Turkey, (corresponding author phone: +90 212 381 5706, e-mail: gokmen.altay@bahcesehir.edu.tr)

978-1-4799-3163-7/13/$31.00 ©2013 IEEE

where $H(X, Y)$ denotes joint entropy. The joint entropy is:

$$H(X,Y) = -\sum_{x_i \in X} P(X = x_i, Y = y_j)\log\big(P(X = x_i, Y = y_j)\big) \tag{3}$$

Joint entropy is obtained from the joint probability distributions of the cells.

B-spline performs soft-binning rather than hard-binning unlike the other MI-based methods, which requires binning operation. In classical binning or discretization operation, each sample can belong to only one sample and this is also called as hard-binning. However in soft-binning, as B-spline does, each sample can belong to several bins at the same time. The spline order denotes the number of bins each sample belongs in. A detailed explanation of B-spline is given below.

B-spline (BS) is a different form of spline functions. It is a kind of polynomial function, which is split into several pieces.

To shape and define the general form of the polynomial function, several *control points* are used in BS functions. To decide how and where the control points are changing, some *knot points* should be used. Knots are alternatively called as connection points of the piece-wised splines. Knot points should be given in a vector, whose elements should not be in decreasing order. Number of the knot points in the knot vector is obtained by summing the spline order and the number of control points.

BS functions are utilized in the several applications [9-13] of the bioinformatics field. Daub *et al*, proposed using BS in the association estimation step of their GNI application [3]. They claimed that a more commonly used estimator, Kernel Density Estimator (KDE) outperforms BS when the spline order is 1. However, when the spline order is changed from 1 to 2, significant performance increment is achieved by BS. And also BS outperforms the KDE with the spline order greater than 1. Moreover it outperformed another commonly used estimator named Best Upper Bound (BUB) even if the spline order is 1. Because BS estimator fulfils soft-binning and provides each sample can belong to several bins concurrently, it can estimate MI with a higher accuracy than the methods that uses hard-binning. In classical binning manner, due to several noises, the data samples that are close to the border of bins may be confused and assigned to neighbor bins accidentally. BS avoids users from this case by using soft-binning. Indicator function of (classical) hard-binning is given in (4). BS function counterpart of indicator function is given in (6).

$$\theta_i(x_u) = \begin{cases} 1, & if\ x_u \in a_i \\ 0, & otherwise \end{cases} \tag{4}$$

The hard-binning indicator function is used as given in (5) to obtain the empirical probability distribution of the bins:

$$\hat{p}(a_i) = \frac{1}{N}\sum_{u=1}^{N}\theta_i(x_u)\,. \tag{5}$$

In the BS association estimator, each data point has a membership score for different bins. Each sample can be belong to more than one bin concurrently by using this membership scores or weights, $\widetilde{B}_{i,k}$. BS functions aim to obtain the weights, $\widetilde{B}_{i,k}$, for each bin i, for a given spline order k. The probability distribution of the i-th bin, $p(a_i)$, becomes as:

$$\hat{p}(a_i) = \frac{1}{N}\sum_{u=1}^{N}\widetilde{B}_{i,k}(x_u)\,. \tag{6}$$

By using (6), individual and joint entropies are calculated to obtain MI between the gene pairs from (2). In this study, effects of the spline order, k, according to the CT pre-processing operation is investigated by changing k from 1 to 10.

III. C3NET GNI Algorithm

In this study BS association estimator is evaluated with varying spline orders, according to the network inference performance of C3NET GNI algorithm. It is a recently proposed popular algorithm with a complete R package available in CRAN [14]. Moreover, it is denoted that it outperforms the other popular GNI algorithms (such as ARACNE [4] and CLR [15]) over various topologies [6, 14].

C3NET involves two main steps after the association estimation phase. In the first step, the non-significant associations, whose score is below a particular threshold value, are deleted from the network. In the second step, the maximum valued interaction of each node (gene) is protected; the rest is eliminated from the gene network.

IV. Experimental Results

B-spline (BS) estimator is evaluated and compared according to different spline orders and with respect to Copula Transform (CT) by using two different synthetic datasets, which are generated by SynTReN[16]. The true networks of these datasets are taken from *E.Coli* sub-networks that are real biological networks. The 1st dataset involves 100 genes and 100 samples; the 2nd one includes 100 genes and 1000 samples. Hence the effect of the number of utilized samples on the inference performance is also examined in this study.

CT is a frequently used pre-processing operation in GNI applications, as mentioned previously. It utilizes the ranking values of the samples of each variable (gene), instead of their original values. The ranking values of the samples become between the interval (0,1] by dividing those ranking values with the maximum ranking value (i.e. number of samples). Then 0.5 is subtracted from the normalized ranking values. Hence copula transformed values of the samples are provided being between the range (-0.5, 0.5].

In this study, evaluation of the inference performance is achieved by using F-score metric, which is obtained by (7):

$$F = \frac{2pr}{p + r} \tag{7}$$

where p denotes precision and r denotes recall. Precision and recall metrics are given in the (8):

$$p = \frac{TP}{TP + FP} \text{ and } r = \frac{TP}{TP + FN} \quad (8)$$

where TP is True Positive, FP is False Positive, and FN is False Negative. TPs are the number of the edges that actually exist in the true network and can be inferred by the GNI algorithm. FPs are the edges found by the GNI algorithm, however they do not actually exist in the true net. FNs actually exist in the true net, but they could not be inferred by the GNI algorithm.

For the 1st dataset without using CT, numbers of TPs, FPs, and FNs; *p*, *r*, and F-score metrics of the BS estimator according to the varying spline order from 1 to 10 are given in Table 1. The barplot of the F-score values according to spline orders is also given in Fig.1.

TABLE I
EVALUATION OF B-SPLINE ACCORDING TO SPLINE ORDER FOR THE 1ST DATASET WITHOUT USING CT

Spline order	**TP**	**FP**	**FN**	***p***	***r***	**F-score**
1	33	50	96	0.3976	0.2558	0.3113
2	43	44	86	0.4943	0.3333	0.3982
3	32	55	97	0.3678	0.2481	0.2963
4	16	76	113	0.1739	0.1240	0.1448
5	9	89	120	0.0918	0.0698	0.0793
6	1	98	128	0.0101	0.0078	0.0088
7	1	98	128	0.0101	0.0078	0.0088
8	1	98	128	0.0101	0.0078	0.0088
9	1	98	128	0.0101	0.0078	0.0088
10	2	97	127	0.0202	0.0156	0.0175

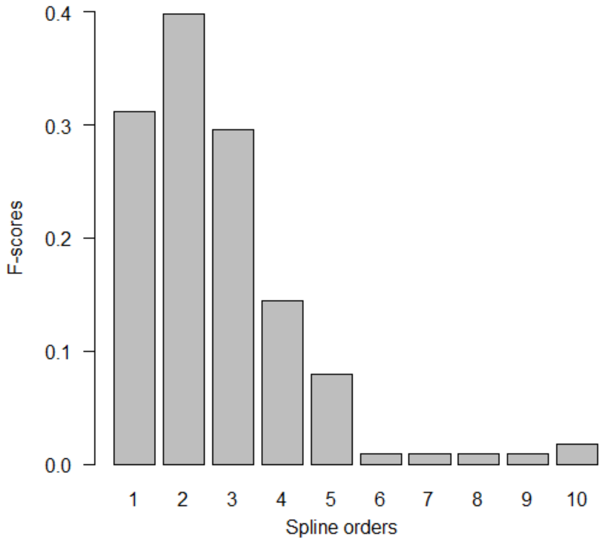

Fig. 1. Evaluation of B-spline according to spline order for the 1st dataset WITHOUT using CT.

For the 1st dataset without using CT, BS estimator with the spline order 2 (BS2) gives the best F-score among all of the spline orders. BS estimator dramatically gives worse results when the spline order is greater than 4. When the spline order is changed from 1 to 2, the performance is increased significantly.

For the 1st dataset performance evaluation of the BS according to spline order with using CT is given in Table 2 and Fig.2. In this case, the best result is taken when the spline order is 3 (BS3). BS with spline order 2 (BS2) gives the second best result. And also result of the BS2 is similar to the result of BS3. With CT, when the spline order is greater than 4, we do not get dramatically worst results, unlikely in the case of without using CT.

TABLE II
EVALUATION OF B-SPLINE ACCORDING TO SPLINE ORDER FOR THE 1ST DATASET WITH USING CT

Spline order	**TP**	**FP**	**FN**	***p***	***r***	**F-score**
1	34	50	95	0.4048	0.2636	0.3193
2	40	43	89	0.4819	0.3101	0.3774
3	42	42	87	0.5000	0.3256	0.3944
4	28	55	101	0.3374	0.2171	0.2642
5	36	45	93	0.4444	0.2791	0.3429
6	26	56	103	0.3171	0.2016	0.2465
7	24	62	105	0.2791	0.1861	0.2233
8	27	64	102	0.2967	0.2093	0.2455
9	24	62	105	0.2791	0.1861	0.2233
10	29	61	100	0.3222	0.2248	0.2648

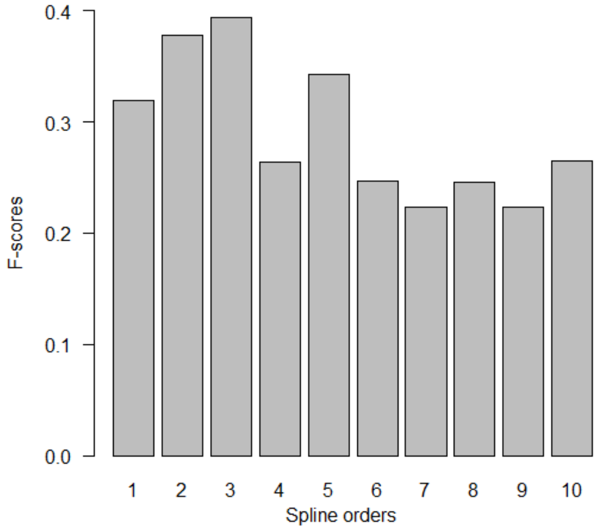

Fig. 2. Evaluation of B-spline according to spline order for the 1st dataset WITH using CT.

Finally, for the 1st dataset using CT increases the performance in terms of F-score for all of the spline orders, except 2, from 1 to 10. F-score values of BS2 with using CT and without using CT is very close to each other. Moreover, in both cases, (with or without using CT), BS2 gives better performance than the most of the rests. Hence BS2 seems as a proper choice for the 1st dataset, which has fewer samples than the 2nd one. Overall comparison of the BS according to spline order with using CT is illustrated in Fig.3. "COP" means the method is used with CT.

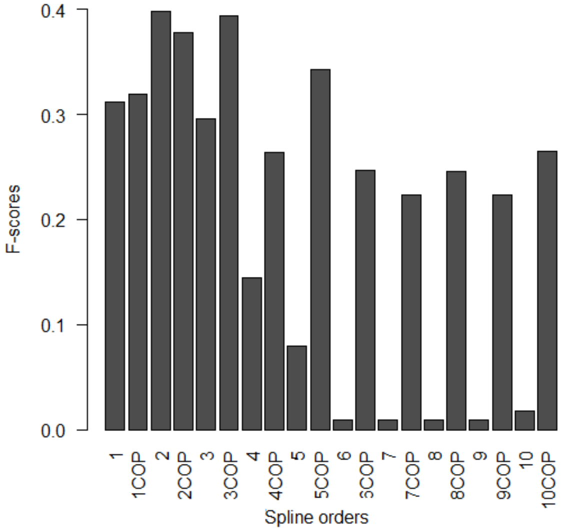

Fig. 3. Overall evaluation of the BS estimator for the 1st dataset

For the 2nd dataset numbers of TPs, FPs, and FNs; *p*, *r*, and F-score metrics of the BS estimator according to the varying spline order from 1 to 10, without and with using CT operation are given in Table 3 and Table 4, respectively. The corresponding barplots of the F-score values are also given in Fig.4 and Fig 5, respectively. In the 2nd dataset number of samples is 1000. For this dataset without using CT, BS2 gives the best result. After that, BS1 (BS with spline order 1) gives the second best result. Those two spline orders were also the best ones for the 1st dataset. BS3 without using CT did not perform badly for the 1st dataset; its F-score value was close to the F-score value of BS1. However for the 2nd dataset F-score value of BS2 is not close to the F-score values of BS1 and BS3. Without using CT, the most promising spline order seemed as 2, because its F-score value is the greatest one among the rests.

TABLE I. EVALUATION OF B-SPLINE ACCORDING TO SPLINE ORDER FOR THE 2ND DATASET WITHOUT USING CT

Spline order	**TP**	**FP**	**FN**	***p***	***r***	**F-score**
1	55	28	49	0.6627	0.5289	0.5882
2	69	14	35	0.8313	0.6635	0.7380
3	26	64	78	0.2889	0.2500	0.2680
4	14	80	90	0.1489	0.1346	0.1414
5	5	91	99	0.0521	0.0481	0.0500
6	8	91	96	0.0808	0.0769	0.0788
7	1	98	103	0.0101	0.0096	0.0099
8	2	97	102	0.0202	0.0192	0.0197
9	1	98	103	0.0101	0.0096	0.0099
10	1	98	103	0.0101	0.0096	0.0099

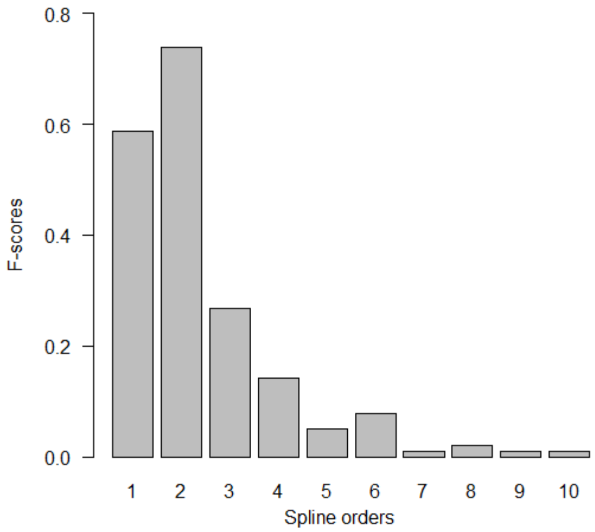

Fig. 4. Evaluation of B-spline according to spline order for the 2nd dataset WITHOUT using CT

TABLE II. EVALUATION OF B-SPLINE ACCORDING TO SPLINE ORDER FOR THE 2ND DATASET WITH USING CT

Spline order	**TP**	**FP**	**FN**	***p***	***r***	**F-score**
1	58	13	46	0.8169	0.5577	0.6629
2	67	5	37	0.9306	0.6442	0.7614
3	65	7	39	0.9028	0.6250	0.7386
4	61	11	43	0.8472	0.5865	0.6932
5	62	10	42	0.8611	0.5962	0.7046
6	44	29	60	0.6028	0.4231	0.4972
7	38	40	66	0.4872	0.3654	0.4176
8	28	50	76	0.3590	0.2692	0.3077
9	22	59	82	0.2716	0.2115	0.2378
10	24	58	80	0.2927	0.2581	0.2581

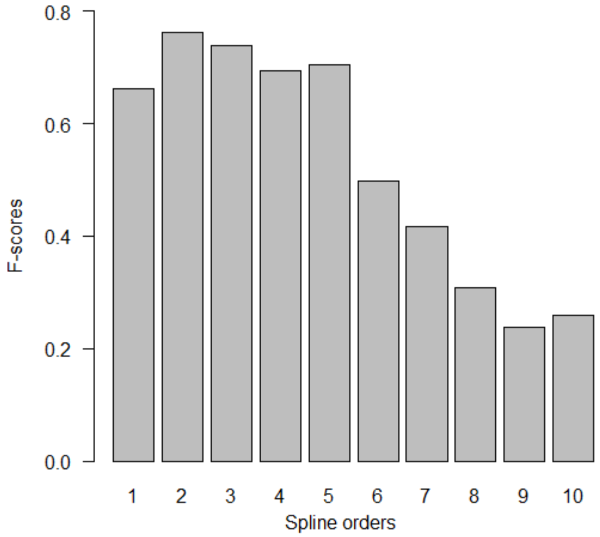

Fig. 5. Evaluation of B-spline according to spline order for the 2nd dataset WITH using CT

For the 2nd dataset by using CT all of the spline orders exhibit significant performance increment, as similar to experiments of the 1st dataset. BS2 has the best F-score value for both cases, i.e. with or without using CT. With CT the second best spline order is 3. Results of BS2 and BS3 are close to each other with using CT for both of the datasets. For most of the spline orders F-score values are increased significantly or become similar when the sample size is increased even without using CT.

Overall comparison of the spline orders according to using CT is illustrated in Fig.6. "COP" means the method is used with CT. Performances of the all of the spline orders increased by using CT without any exception for the 2nd dataset.

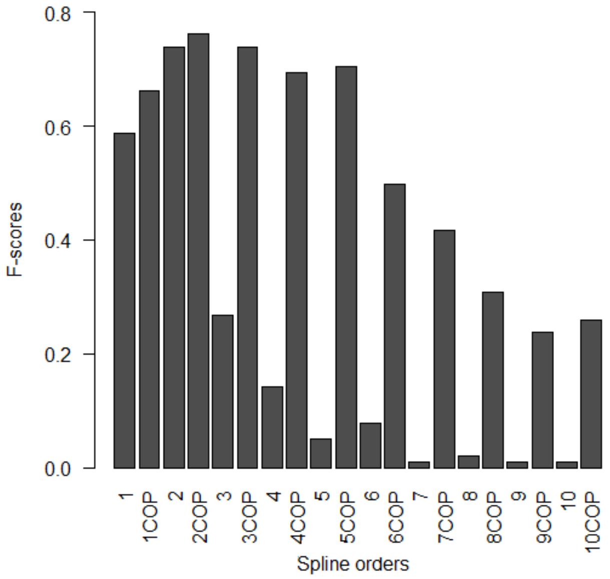

Fig. 6. Overall evaluation of the BS estimator for the 2nd dataset

F-score changing graphic for both of the cases and for both of the datasets is given in Fig.7.

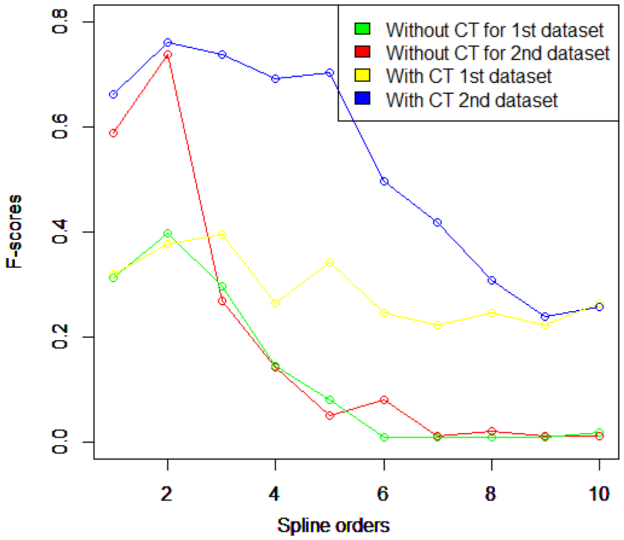

Fig. 7. Overall evaluation of the BS estimator for both cases, for both datasets

At the end of the experiments from Fig. 7, it is observed that for most of the spline orders of the BS estimator, *using CT* and *increasing number of the samples* improve the inference performance. For the 2nd dataset, performance becomes significantly larger with using CT and more samples, especially for the larger spline orders. At overall analysis, spline order 2 seems as the best choice for the both cases (with or without CT) and for the both datasets. For the best performed spline order, i.e. 2, performance changing caused by the CT is the least one among the others, especially for the 2nd dataset.

V. Conclusion

In the proposed study, impacts of the spline order and influences of the sample numbers and CT on the inference performance of BS estimator are examined. Spline order is changed from 1 to 10 during the experiments. C3NET GNI algorithm is used to obtain the gene networks. To our knowledge, in the literature this is the only study, which comprehensively evaluates the impacts of the varying spline orders on the BS estimator according to the using CT and varying sample numbers.

Two different artificial datasets are used in the experiments, whose true nets belong to subnets of E.Coli bacteria. In both of the datasets there are 100 genes. The first dataset includes 100 samples, while the second one involves 1000 samples. At the end of the experiments, it is noted that, CT improves the inference performance of the BS for almost all of the spline orders in terms of F-score. Different spline orders perform nearly similar with using CT for both of the datasets. Moreover, increasing the number of the samples also improves the performances of the BS for all of the spline orders. At the overall analysis, the best performed spline order is found as 2, with or without using CT and with large or small number of samples.

When the spline order gets larger than 4, complexity of the algorithm is increasing, while the inference performance is decreasing. If researchers want to study with spline orders greater than 4, using CT is proposed especially. However spline order 2 exhibits similar performances with or without using CT, for both small and large datasets. Hence, BS with spline order 2 seemed and suggested as the ideal choice for using in the GNI applications.

References

[1] C. Olsen, P.E. Meyer and G. Bontempi, "On the Impact of Entropy Estimation on Transcriptional Regulatory Network Inference Based on Mutual Information", *EURASIP Journal on Bioinformatics and Systems Biology*, Vol. 2009(308959), 2009.

[2] R.M. Simoes and F. Emmert-Streib, "Influence of Statistical Estimators of Mutual Information and Data Heterogeneity on the Inference of Gene Regulatory Networks", *PLoS ONE*, Vol.6(12), 2011.

[3] C.O. Daub, R. Steuer, J. Selbig and S. Kloska, "Estimating mutual information using B-spline functions - an improved similarity measure for analysing gene expression data", *BMC Bioinformatics*, Vol. 5(118), 2004.

[4] A.A. Margolin, I. Nemenman, K. Basso, C. Wiggins, G. Stolovitzky, R.D. Favera and A. Califano, "ARACNE: An Algorithm for the Reconstruction of Gene Regulatory Networks in a Mammalian Cellular Context", *BMC Bioinformatics*, Vol. 7(Suppl 1):S7, 2006.

[5] A.J. Butte and L.S. Kohane, “Mutual information relevance networks: functional genomic clustering using pairwise entropy measurements”, *Pacific Symposium on Biocomputing*, Vol. 5, pp. 418-429, Jan. 2000.

[6] G. Altay and F. Emmert-Streib, “Inferring the conservative causal core of gene regulatory networks”, *BMC Systems Biology*, vol.4(132), 2010.

[7] G. Altay, M. Asim, F. Markowetz and D.E. Neal, “Differential C3NET reveals disease networks of direct physical interactions”. *BMC Bioinformatics*, Vol. 12(296), 2011.

[8] G. Altay and F. Emmert-Streib, “Revealing differences in gene network inference algorithms on the network-level by ensemble methods”, *Bioinformatics*, Vol. 26(14), pp. 1738-1744, 2010.

[9] T. Çakır, M.M.W.B. Hendriks, J.A. Westerhuis, and A.K. Smilde, “Metabolic network discovery through reverse engineering of metabolome data”, *Metabolomics,* vol. 5, pp. 318–329, DOI 10.1007/s11306-009-0156-4, 2009.

[10] R. Steuer, J. Kurths, C.O. Daub, J. Weise and J. Selbig, “The mutual information: detecting and evaluating dependencies between variables”, *Bioinformatics,* S231-S240, 2002.

[11] Y. Luan and H. Li, “Clustering of time-course gene expression data using a mixed-effects model with B-splines”, *Bioinformatics*, Vol. 19(4), pp. 474-482, 2003.

[12] H. Li, Y. Sun, and M. Zhan, “Analysis of Gene Coexpression by B-Spline Based CoD Estimation”, *EURASIP Journal on Bioinformatics and Systems Biology*, Vol. 2007:49478, 2007.

[13] S. Xu, “*Principles of Statistical Genomics 2013, chapter 23: Time-Course Microarray Data Analysis*”. pp 365-382, Springer New York, ISBN: 978-0-387-70806-5 (Print) 978-0-387-70807-2 (Online).

[14] G. Altay and F. Emmert-Streib, "Structural influence of gene networks on their inference: analysis of C3NET", *Biology Direct*, Vol. 6(31), 2011.

[15] J. J. Faith, B. Hayete, J. T. Thaden, I. Mogno, J. Wierzbowski, G. Cottarel, S. Kasif; J.J. Collins and T.S. Gardner, “Large-scale mapping and validation of Escherichia coli transcriptional regulation from a compendium of expression profiles”, *PLoS Biology*, vol. 5(1), pp. 1-3, 2007.

[16] http://homes.esat.kuleuven.be/~kmarchal/SynTReN/ [last access date: 11.05.2013]