Determination of relationships among cancer-related genes using Bayesian networks

Michael Kofi Ahenkan*

Department of Mathematics,
Kwame Nkrumah University of Science and Technology,
P.O. Box BP 56, Kumasi, Ghana
Email: michael.ahenkan@yahoo.com
*Corresponding author

Emmanuel S. Adabor

School of Technology,
Ghana Institute of Management and Public Administration,
P.O. Box AH 50, Achimota, Accra, Ghana
Email: emmanuelsadabor@gimpa.edu.gh

Kwaku F. Darkwah

Department of Mathematics,
Kwame Nkrumah University of Science and Technology,
P.O. Box BP 56, Kumasi, Ghana
Email: fordarkk@gmail.com

Abstract: A network of relationships among cancer-related genes can be reconstructed from high-throughput datasets obtained by deoxyribonucleic acid (DNA) micro-array technologies. However, modelling such biological networks is challenged by the nature of data and the complexities of relationships among biological variables such as genes. In this paper, Bayesian networks are applied to predict novel regulatory relationships among genes in cancer from genomic datasets. The performances of the methods were assessed by standard metrics such as sensitivities and specificities. Furthermore, in order to validate and verify the reliability of the new predicted relationships among the genes, some of the results were examined with experimentally confirmed relationships found by previous research. Interestingly, some predicted regulatory relationships were also found in the literature. This enhances confidence in the newly predicted network of regulatory relationships, which could become hypotheses for further research.

Keywords: gene relationships; regulatory networks; Bayesian network; search techniques; cancer; Bayes theorem; gene network inference; causal networks; genomic data analysis; DNA micro-array technologies.

Reference to this paper should be made as follows: Ahenkan, M.K., Adabor, E.S. and Darkwah, K.F. (2022) 'Determination of relationships among cancer-related genes using Bayesian networks', *Int. J. Computational Biology and Drug Design*, Vol. 15, No. 2, pp.139–154.

Biographical notes: Michael Kofi Ahenkan is an educationist and a researcher with interest in disease modelling and machine learning. He holds a Master of Philosophy in Applied-Mathematics from the Kwame Nkrumah University of Science and Technology.

Emmanuel S. Adabor is currently a Senior Lecturer at the School of Technology, Ghana Institute of Management and Public Administration. His area of research in Applied Mathematics includes the use of computing and domain knowledge to elucidate biological and non-biological systems.

Kwaku F. Darkwah is currently a Professor of mathematics at Kwame Nkrumah University of Science and Technology. His research interest is in Operation research, Mathematical Physics, Numerical Methods and Artificial Intelligence.

1 Introduction

Cancer comprises of several different diseases mainly characterised by uncontrolled or abnormal cell growth. The mechanism accompanying this disease is complicated. However, our ability to understand its dynamics and find the cure for its victims requires research, analysis, knowledge and high quality accurate data. There are over hundreds of different types of cancer seen under the microscopic eye. Knowledge in molecular and genetic basis of cancer clearly shows the existence of many changes in each cancer, thus the molecular defects in an individual's tumour is bound to change over time, a biggest challenge in personalised medicine (Rashbas, 2016). Hence, building large collections of data via cancer patients are of much importance since there is limited number of people with similar cancers.

Although improvements in technology have resulted in generation of large volumes of genomic data, this can be subjected to analysis to enhance understanding. It, however, requires analytical experts to completely extract relevant biological insights from such datasets. For this reason, several computational methods have been proposed that could be applied in reverse-engineering biological networks conveying regulatory relationships among genes in cancer.

Bayesian Network (BN) is becoming a force to reckon with in the scientific community for the task of predicting cellular networks and genetic data analysis. Its probabilistic nature deals with the uncertainties entailed in biological systems and measurements (Adabor et al., 2015). Information from BN models (conveying linear and higher order regulatory relationships among genes accompanying cancer) are easy to interpret due to its graphical nature, thus providing valuable insights into the properties of the data being analysed and giving rise to new models to be produced. Direct causal relationships is being modelled by BN via direct acyclic graph (DAG) with the Causal Markov Assumption been imposed on it (Needham et al., 2007; Friedman et al., 2000).

Clustering algorithms may be used in the analysis and visualisation of genomic datasets. However, the algorithm fails to establish a causal assumption in its inference even though the existence of a relation is clearly established (Bansal et al., 2007). Ordinary differential equation (ODE)-based techniques may succeed in establishing causal interaction in its inference. However, this requires time series data and low

dimensional dataset (Gardner et al., 2003). BN entails solutions to the flaws of both clustering and ODE methods since it is capable of dealing with thousands or even millions of parameters and also establish causal relationships in its predictions (Gregoretti et al., 2010; Greenfield et al., 2010; Gardner et al., 2003). BN is characterised by its applicability to problems. For example, inferred models as a result of perturbations to a DAG may be used to discover cancer genes directly affected by the perturbation (Di Bernado et al., 2005).

Although methods entailed in BN are discussed into details in this work, the primary objective of this paper is to deduce new gene relationships for gaining insights into cancer using Bayesian Network methods. The process involves combinations of BN techniques to unearth new regulatory relationships among genes from breast cancer dataset. New gene relationships found by the methods will provide useful insights for drug discovery processes. Overall, the work provides insights into cancer that will eventually advance efforts to finding new targeted therapies.

2 Method

2.1 Bayesian network

A probabilistic graphical model that consists of random variables together with relationships (edges) among them is a Bayesian Network. It is a DAG. Its edges imply causal relationships when the Markov Assumption is satisfied. The Markov Assumption entails that all d-separations are conditional in-dependencies and every conditional in dependencies entailed by the Markov assumption is identified by d-separation (Neapolitan, 2004).

2.2 Bayes theorem in connection with relationships among genes

Let X and W be random variables of an event. Then the Bayes rule is defined by equation (1):

$$P(X|W) = \frac{P(W|X)P(X)}{P(W)} \tag{1}$$

The prior likelihood of a particular gene being expressed might be vague (Bolouri, 2011). For example, suppose a gene expression data is discretised into say on, off, low and high. Expression level of a particular gene may be evaluated as say 22, 25, 26 and 27% which implies the absolute probabilities is less informative since the gene is equally roughly expressed with respect to all the levels.

On the other hand, the query of a gene expression level given the activity level of its potential regulators may be very informative. Suppose gene X has a 95% likelihood of being highly expressed, that is P (X = high|Y = high, Z = high), then a hypothesis can be suggested as Y and Z activates the expression X. Also P (X, Y, Z) ≤ P (X|Y, Z) since more information is given in the conditional case. BNs present relationships among variables in sequential connection, common cause and effect relationships. All of these relationships are encoded in a joint probability distribution over the set of variables.

2.3 Joint probability distribution over variables

The number of model parameters needed to define the joint probability distribution (JPD) of variables of a BN structure increases rapidly with the number of variables. Models may be represented in a compact manner by exploiting conditional independence between variables in the BN structure with fewer parameters. The JPD over variables of a BN can be expressed as a product of continuous probability distributions where each node is described in terms ofits causal node (parent).

Let $\boldsymbol{X} = [\xi_1, \ldots, \xi_n]$ where $\xi_1, \ldots, \xi_n$ represents nodes in a DAG.

Then,

$$P(\xi_n|\xi_{n-1}) = \frac{P(\xi_n \Lambda \xi_n - 1)}{P(\xi_n - 1)} \tag{2}$$

Re-arranging equation (2) results in the following:

$$P(\xi_n \Lambda \xi_{n-1}) = P(\xi_n|\xi_{n-1})\, P(\xi_{n-1}) \tag{3}$$

which implies $P(\xi_n \Lambda \xi_{n-1}) = P(\xi_n)\, P(\xi_{n-1})$ given existence of statistical independence. Hence, for a structure X with n variables (nodes) and a set of parameters, θ, then equation (4) can be used to compute the JPD of $\boldsymbol{X}$.

$$P(\boldsymbol{X}|\theta) = \prod_{i=1}^{n} P(\xi_i|\mathbf{pa}(\xi_i), \theta) \tag{4}$$

where pa refers to the parent of any node (s) in $\boldsymbol{X}$.

2.4 Learning BN structure

A BN structure-learning problem involves finding the best network structure, G, given the dataset D. Scoring metric and search technique are the techniques used for learning BN structures. Monte-Carlo techniques may be used to locate network structures in a large search space while scoring metrics evaluate the structures being identified by the searchers.

2.4.1 Scoring metric

The Bayesian scoring metric responsible for evaluating the score of a DAG is derived by taking the log of equation (1). That is, given that D is an assumed multinomial sample, then

$$\text{Score}\ (G : D) = \log P\ (G|D) P\ (G) - \log P\ (D) \tag{5}$$

and $P\ (D|G) = \int P(D \mid G, Q) P(Q \mid G) \mathrm{d}Q$ is the sum or integral of the query of data (D) given its structure (G).

The Score $(G : D) \leq 0$ is a Bayesian scoring function which discriminates between simple and complex structure Darwiche and Provan (1996). Thus, it avoids over fitting which occurs when parameters of a model are too many than necessary. The first term in equation (5) is the log-likelihood function which maximises the probability of seeing the

data while the second term (log$P$ (D)) is a penalty function to select simpler models over complex competing models. By assuming that data does not discriminate between equivalent structures, Heckerman and coworkers developed the Bayesian Dirichlet equivalence (BDe) metric for scoring networks (Heckerman, 2008).

2.4.2 Bayesian Dirichlet equivalence (BDe)

The BDe assumes further that, the metric has the property of score equivalence thus the parameters follows a Dirichlet distribution. Its orders are determined by the effective sample size and assumed local joint probability. This metric seeks to avoid overfitting the model by penalising G for every additional edge used in the computation (i.e., penalises for complexity). The BDe has been used in the past for such studies (Cooper and Herskovits, 1992; Heckerman, 1995; Adabor and Acquaah-Mensah, 2019). This formular is given as,

$$P(G,D) = P(G)\prod_{i=1}^{n}\prod_{j=1}^{q_i}\frac{\Gamma(N_{ij}^{'})}{\Gamma(N_{ij}^{'}+N_{ij})}\prod_{k=1}^{\Pi}\frac{\Gamma\left(N_{ijk}^{'}\right)+N_{ijk}}{\Gamma(N_{ijk}^{'})} \tag{6}$$

where N', assumed as a local joint probability is given as,

$$N_{ijk}^{'} = N'P(x_i = k, \Pi_i = j \mid B_{sc}^{e}, \xi)\ ,\ N'_{ij} = \sum_{k=1}^{r_i} N_{ijk}^{'}\ ,\ N_{ij} = \sum_{k=1}^{r_i} N_{ijk}$$

- Π_i refers to the parents of x_i for each node i in the DAG
- B_{sc}^{e} is the event of complete DAG
- ξ refers to prior knowledge
- G refers to the network structure
- D is the data-set
- r_i refers to the number of states of variable x_i and q_i refers to the number of states of Π_i
- n, N' is the number of variables and equivalent sample size respectively
- N_{ijk} is the number of cases in a database where x_i = k and $\Pi_i = j$
- Γ is the Gamma function which is the generalisation of the factorial function such that Γ (x) = (x − 1)!
- $P(G) = c\prod_{i=1}^{n} k^{\delta_i}$, is the function which incorporates the penalty factor given as $0 < k \leq 1$, c a normalisation constant and δ_i is the number of nodes in the symmetric difference between the parents of x_i in the prior network.

This metric seeks to avoid over-fitting the model by penalising G for every additional edge used in the computation (i.e., penalises for complexity). The BDe is used in this work since it prefers simple over complex models (DAG's), thus applying Occam's

theory which states that, one should opt for simpler models over more complex models other things being equal.

2.5 Search methods

The search space of a network structures with many variables is almost close to infinite. Thus, for a given DAG with n nodes the number of possible structures can be evaluated by the formula $f(n) = \sum_{i=1}^{n} (-1)^{i+1} \binom{n}{i} 2^{(n-1)} f(n-i)$, for $n > 2, f(0) = 1$ and $f(1) = 1$ (Robinson, 1997). Computing f (2), f (3), f (5) and f (10) gives 3, 25, 29000 and 4.2×10^{18} respectively.

However, our main aim is to find the best network structure that maximises the likelihood of seeing the data. This can be done by exhaustively computing the full posterior distribution for only simple models. In BN structure learning, new structures are obtained from perturbations or changes to an existing structure. Local perturbations are done at each iteration, that is, an edge might be removed, reversed, or new edges introduced between existing variables while cyclic networks are avoided simultaneously. In this work, two main search methods were considered in learning the BN of breast cancer data. These two methods namely Simulated annealing and Greedy Hill Climbing are considered as a result of their superior performance (Chickering, 1996; Yu et al., 2004).

2.5.1 Greedy Hill climbing search

The algorithm begins with a prior (initial) network in the search space, considers its neighbours and perform local perturbations to check for a high scoring network among its neighbours. A high scoring structure found from the local search becomes the new reference point for another search (Klebaner, 2005). This process is repeated until there are no neighbouring structures with a higher score than the current network. In order to avoid suboptimal results, this method is improved by techniques such as repeated random restarts. This involves repeating the search algorithm several times each starting with a different randomly generated initial network in the search space. A threshold is set with respect to the number of restarts or time for the search to terminate.

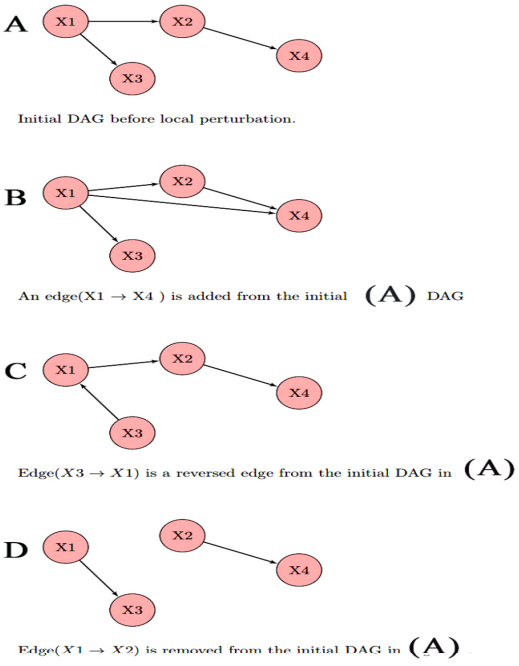
2.5.2 Simulated annealing search

This algorithm contrasts the earlier method allowing the possibility selecting networks that may not necessarily have the best scores in the search process. Particularly, it starts with an initial network and assigns a temperature variable, T, to simulate the heating process. The initial temperature value is set high enough and a lengthier time is allowed for the temperature to reduce to its minimum as the algorithm runs. Worse network scores than our current score are accepted at the initial stages while the temperature is still high. However, there is a downward chance of accepting less fit networks as the temperature decreases which allows the algorithm to slowly focus on optimal networks in the search space.

2.6 Evaluating the performance of methods

The results obtained from the search methods constitute two major parts namely Greedy and Simulated Annealing search techniques. However, each method consists of two local perturbations: Random-Local-Moves and All-Local-Moves. All-Local-Moves has its flaws, it is computationally expensive and consumes large memory compared to Random-Local-Moves. All-Local-Moves involves composing a list of all available local moves by adding, reversing and deleting an edge as pictorially described in Figure 1. However, Random-Local-Moves select one of these (possible local moves) at random while prioritising an acyclic structure during these local perturbations. We use these techniques due to their broad usage in BNs as they have proven to lead to expected results by researchers over the years (Adabor and Acquaah-Mensah, 2019). The methods along a quantile discretisation are implemented in Bayesian network inference with java objects, Banjo (Sladeczek et al., 2008). Experiments were run for 5 h. Predicted results by the methods were visualised using Cytoscape (Shannon et al., 2003).

Figure 1 Possible perturbations to a DAG structure (see online version for colours)



In the context of BN's, moving from a point to its neighbouring point in a search space is done by performing these local perturbations or searchers to the DAG by adding, removing and deleting an edge as expressed in Figure 1.

To analyse relationship among genes in cancer, we obtained published datasets of gene expression values along with the relevant genes (GSE40066 and GSE25011). Dataset with record number GSE40066 from Gene Expression Omnibus consists of 43 samples from mammary gland tissues and 8 samples from cardiac blood tissues. GSE25011 was the results of analysis breast tumour samples preserved via RNA stabilisation methods. These comprised of 623 genes and 86 arrays. GSE40066 and

GSE25011 contains 32 and 100 regulatory relationships respectively (reference networks). The methods were evaluated by comparing the recovered network relationships from the data with a reference networks obtained from the literature. Quality solutions are guaranteed if both networks are similar (Huynh-Thu et al., 2010; Husmeir, 2003). Reference and recovered networks are identical if they have the same links or relationships. Furthermore, the following definitions, which explain the variables of sensitivity and specificity are:

- Let TP = Given that a link is found in both the reference and recovered networks, then it is considered as TP.
- Let FN = If a link exists in the reference DAG but not found in the recovered DAG, then it is considered as FN.
- Let TN = If an edge absent in both networks, it is considered as TN.
- Let FP = If a link exists in a recovered network but non-existent in the reference network, then it is considered as FP.

The sensitivity is defined as:

$$\text{Sensitivity} = \frac{TP}{TP + FN}, \tag{7}$$

and the specificity is expressed as

$$\text{Specificity} = \frac{TN}{TN + FP}. \tag{8}$$

By these definitions, a sensitivity of 100% implies that all links present in the reference network are found in the predicted network by the methods. High sensitivity is desirable. Also, a specificity of 100% indicates that no spurious link is found in the predicted network. Nevertheless, if both obtain a perfect score simultaneously, then it implies the discovered and reference BNs are identical

3 Results for GSE40066 and GSE25011

This section presents a summary result of all BN methods used for dataset GSE40066 and GSE25011. These are simulated annealing with all-local-moves (SAA), simulated annealing with random-local-moves (SAR), Greedy Hill Climbing method with all-local-moves (GA) and Greedy Hill Climbing method with random-local-moves (GR).

3.1 Summary results of BN methods for GSE40066

The searchers initiated with SAA technique, this approach predicted 56 gene relationships (see Figure 2) after having achieved the second best network score among the various searchers (see Table 1). SAA predicted 9 gene relationships from the reference network, notable among these predicted gene relationships are gene ETS1

regulates IKZF1 and BCL11B regulates PRKD1, which were also inferred by all the search methods. SAR on the other hand achieved the highest network score among all the BN search methods (see Table 1). The optimal structure predicted 55 edges (see Figure 3) including 2 gene relationships from the reference network. That is gene ETS1 regulates genes 1KZF1 and GATA3.Greedy Hill Climbing method with All-Local-Moves (GA)method predicted 34 gene relationships in its optimal network structure. Although it recorded the least network score and number of predictions, the search method recovered all gene relationships from the reference network along with two additional edges (see Figure 4 and Table 1). GR technique predicted 48 gene relationships in its optimal structure after the search commenced (see Figure 5). With a next to least network score among the search methods, it predicted 4 edges from the initial network including gene ETS1 regulates IKZF1 (see Table 1).

Table 1 Summary statistics of Bayesian network methods for dataset GSE40066

Search techniques	*SAA*	*SAR*	*GA*	*GR*
Number of nodes	53	53	53	53
Number of edges	56	55	34	48
Additional edges	24	23	2	16
Scores	–2712.4840	–2665.9390	–3153.5007	–2724.3634
Recovered edges	3	4	0	2

Summary statistics of bayesian network methods for dataset GSE25011

Search techniques	*SAA*	*SAR*	*GA*	*GR*
Number of nodes	623	623	623	623
Number of edges	253	301	44	55
Additional edges	153	201	–56	–45
Scores	–74218.4766	–72664.2617	–76471.9513	–75642.3006
Recovered edges	5	7	0	0

SAA is simulated annealing with all-local-moves, SAR is simulated annealing with random-local-moves, GA is Greedy Hill Climbing method with all-local-moves, GR is Greedy Hill Climbing method with random-local-moves.

3.2 Summary results of BN methods for GSE25011

SAR, SAA, GR and GA predicted 301, 253, 55 and 44 gene relationships respectively in their final DAG structures. The highest performing technique (SAR) with a network score of –72664.2617 predicted 201 additional edges. It also recorded the highest number of recovered edges followed by SAA which recorded 5 recovered relationships. SAA which the second best performing network also predicted 153 additional edges. However GA and GR recorded 0 recovered edges with a network score of –76471.9513 and –75642.3006 respectively (see Table 1).

Figure 2 Predicted network for simulated annealing with all-local-moves for GSE40066 (see online version for colours)

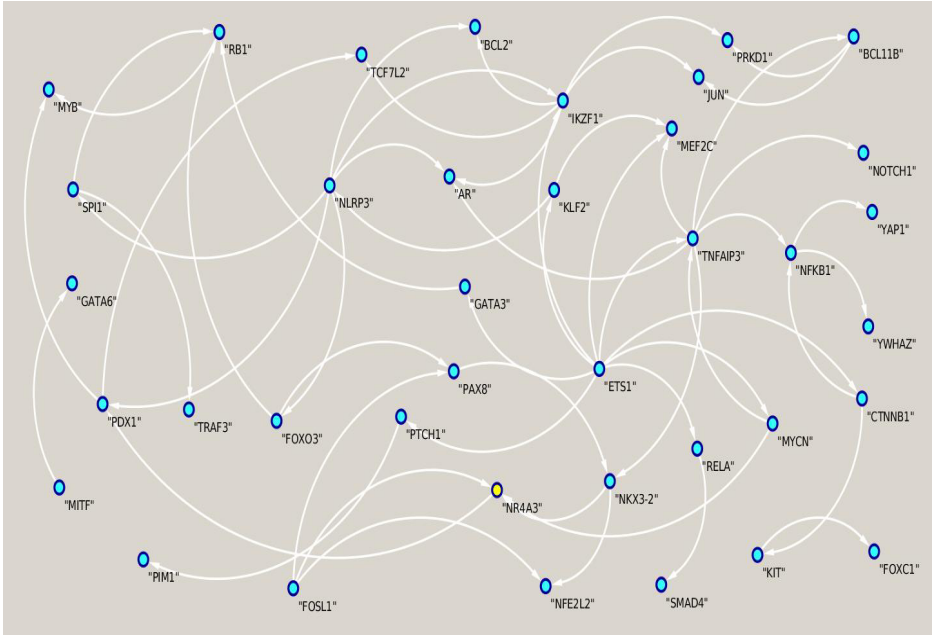


Figure 3 Predicted network for simulated annealing with random-local-moves for GSE40066 (see online version for colours)

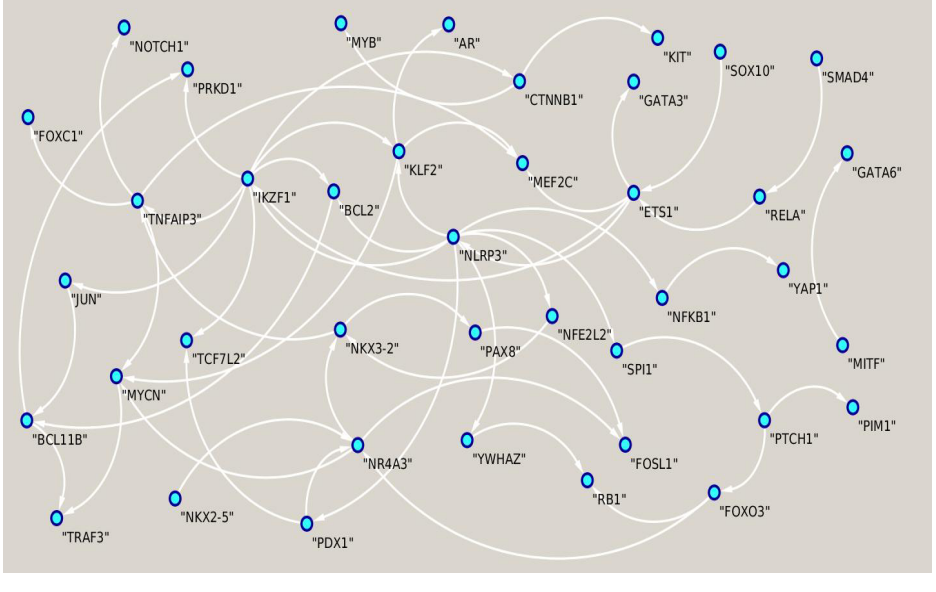


Figure 4 Predicted network for Greedy Hill Climbing method with all-local-moves for GSE40066 (see online version for colours)

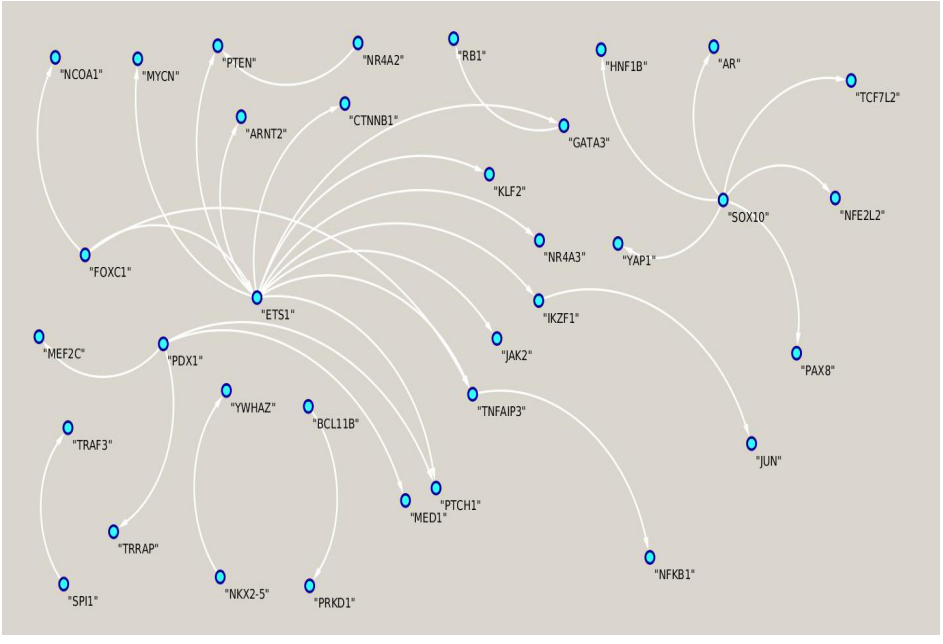
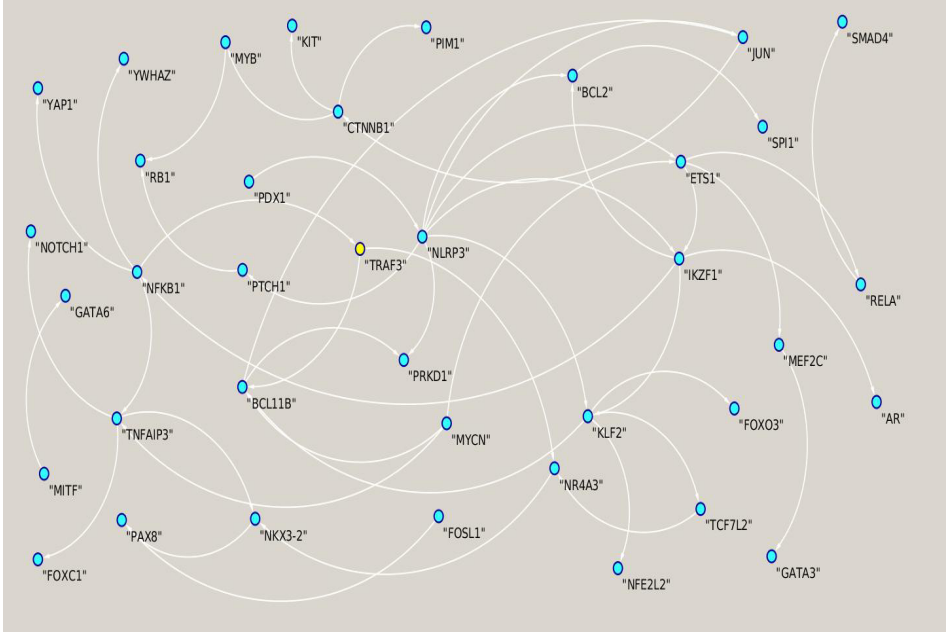


Figure 5 Predicted network for Greedy Hill Climbing method with random-local-moves for GSE40066 (see online version for colours)



3.3 Sensitivity and specificity of methods

Regarding the sensitivities and specificities of the methods, Greedy Hill Climbing method with All-Local-Moves achieved the highest specificity score and a perfect sensitivity score for GSE40066 (see Table 2). This is as a result of predicting all the reference edges and only two False Positive edges Husmeier (2003). This high performance support the suitability of the use of Bayesian Network to be applied to the task of finding relationships among genes in cancer as is the objective of the study.

Table 2 Sensitivity and specificity for dataset GSE40066

Search techniques	*GA*	*SAA*	*GR*	*SAR*
Total no. Of predicted links	34	56	48	55
Total number of true positive	32	9	4	2
Total number of false negative	0	23	28	30
Total number of true negative	33	33	33	33
Total number of false positive	2	24	16	23
Specificity score	0.943	0.556	0.541	0.589
Sensitivity score	1	0.281	0.125	0.063

Sensitivity and specificity for dataset GSE25011

Search techniques	*GA*	*SAA*	*GR*	*SAR*
Total no. of predicted links	44	253	55	301
total number of true positive	23	56	31	81
total number of false negative	77	44	69	19
total number of true negative	32	32	32	32
total number of false positive	21	153	24	201
Specificity score	0.604	0.421	0.571	0.137
Sensitivity score	0.23	0.56	0.31	0.81

GA is Greedy Hill Climbing method with all-local-moves, SAA is simulated annealing with all-local-moves, GR is Greedy Hill Climbing method with random-local-moves and SAR is simulated annealing with random-local-moves.

Furthermore, we discuss sensitivity and specificity scores of all the various search techniques. The main objective of this work is to use BNs to unearth non-existent regulatory gene relationships using reliable prior structures as our basis to initiate our search. For datasetGSE40066, GA, SAA, GR and SAR inferred 34,56,48 and 55 general relationships respectively in their final predicted networks. GA recorded high sensitivity and specificity scores (1 and 0.943 respectively) because it predicted only two gene regulatory relationships which were false positive in its final structure. Other techniques based on SAA, GR and SAR predicted high number of new regulatory relationships (24,16 and 23 respectively) although they had relatively lower sensitivities recorded (see Table 2). Nevertheless, algorithms with low sensitivities (such as SAA, GR and SAR)

have high network scores (see Table 1). Table 2 also indicates a large range of variation of sensitivity/specificity scores for GSE40066, thus the average specificity and sensitivities are fairly spread roughly at 0.655 and 0.368 respectively.

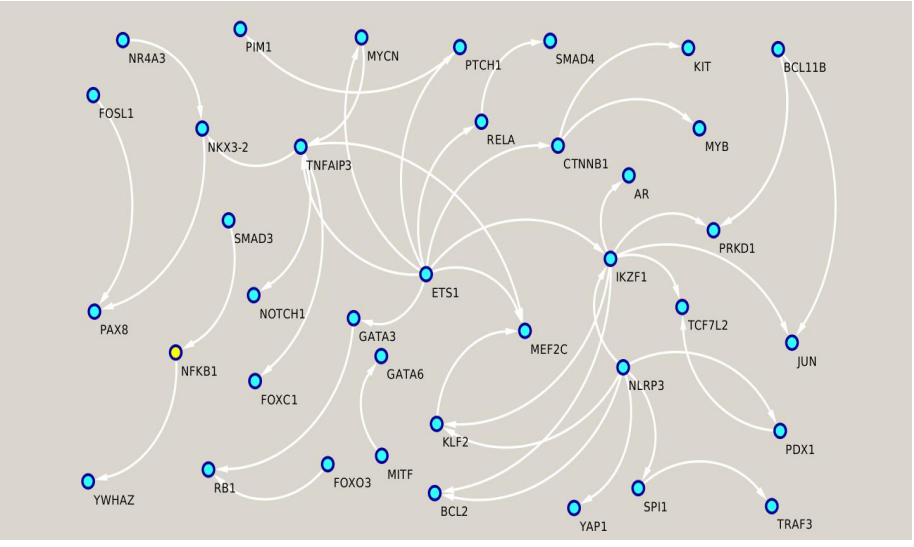
Dataset GSE25011 predicted 44,253,55 and 301 edges for GA, SAA, GR and SAR techniques respectfully. Scores obtained by the various searchers were relatively consistent to dataset GSE40066 considering their respective ranks of performances in both datasets (see Table 1). With the exception of SAR, all the searchers recorded a fair rate of actual negative (specificities), reason for such low outcome is attributed to its prediction of high false positives in its final structure. However, SAR also retrieved a high number of gene relationships in its inferred network, thus, resulted in a high sensitivity score (0.81) (see Table 2). SAA recorded a fair rate of sensitivity (0.56) whereas GA and GR recorded sensitivities of 0.23 and 0.31 respectfully.

The differing specificity and sensitivity are attributable to the new regulatory gene relationships obtained by the BNs. Hence, high specificity rate does not necessarily guarantee its corresponding (high) sensitivity rate.

4 Discussion

All the BN methods used in this study predicted new regulatory relationships among genes in cancer. Simulated annealing search techniques predicted the highest number of regulatory relationships among the methods. This was in addition to achieving the high scoring networks in both sets of data (see Table 1). Prominent among the predicted relationships by all the methods are TNFAIP3 regulates NOTCH1, IKZF1 regulates BCL2, KLF2 regulates MEF2C, IKZF1 regulates TCF7L2, IKZF1 regulates KLF2, IKZF1 regulates AR, BCL11B regulates PRKD1, ETS1 regulates TNFAIP3, ETS1 regulates IKZF1, ETS1 regulates GATA3, ETS1 regulates CTNNB1, BCL11B regulates PRKD1 and ETS1 regulates MEF2C. These consensus predictions for GSE40066 by all the methods improves the likelihood of the existence of such relationships accompanying cancer. In other words, the consensus predictions by all methods provide biological insights into cancer (see Figure 6). Interestingly, some of predicted relationships were further confirmed by previous studies. Particularly, ETS1 regulates IKZF1 is supported by prior report in biomedical literature Perotti et al. (2015). A normal cell will divide and stop dividing itself only when it receives a chemical signal interpreted from its nucleus. Cancer cells fails to obey this rule and will proliferate even if no appropriate signals are received. Since the predicted relationships among the variables in cancer, the suggested regulatory relationships are key components in elucidating the aetiology of cancer. From the consensus network by GSE40066 (see Figure 6), ETS1 influence cell specification suggesting an important role in regulating cell motility (Meyer et al., 2007; Remy and Baltzinger, 2000; Tahtakran and Selleck, 2003). As a result of these properties, ETS1 gene is involved in tissue remodelling during metastasis (Brian and Patricia, 2010). Hence the predicted novel indirect regulatory relationships between TNFAIP3, KZF1, GATA3, CTNNB1, MEF2C and ETS1 is particularly interesting and can be used as the basis for further research in tumourigenesis. Nevertheless, since this is a computational study, the consensus predictions by the different methods and the confirmations of some predictions in previous studies give confidence in the new predictions.

Figure 6 Consensus network obtained from various BN methods for GSE40066 (see online version for colours)



Results from the two high-scoring methods, SAA and SAR, predicted 21 similar relationships. Prominent among these predictions are CTNNB1 regulates KIT which is confirmed in a study by Carraro et al. (2016). Examining some of the inferred relations by the BN methods in existing literature convinces us to appreciate the importance of other equally predicted relationships concerning regulation of cancer related genes.

The search space or possible DAG's of a particular BN model with multiple variables is almost close to infinite. Thus, it is very essential for search techniques (Monte-Carlo methods) to be equipped with features enabling them to cover high magnitude of length and possible DAG's within a specified search time. Techniques with these capabilities increases the likelihood of selecting models with relative less error with respect to its global optimal structure within a search space. It can however be observed that certain computational techniques performed better than others judging from their scores (see Table 1). Among the four search techniques (SAR, SAA, GR and GA) GA performed poorly in both dataset results (GSE40066 and GSE25011) judging from its scores and number additional edges. This is largely attributed to its restricted random movement within the search space. GA prioritises high scoring networks and simultaneously performs all the local searchers at each iteration. This exhaustive and tedious process restricts the algorithm to a limited search area within a search time, thus, resulting in low scoring networks. The algorithm (GA) may be more suitable to datasets with less dimensionality.

5 Conclusion

Though mechanisms leading to cancer are complicated, Bayesian Network models helps to elucidate the causality of the disease by finding new regulatory relationships among genes entailed in its optimal predictions. This study found a consensus network of some of these inferred regulatory relationships. In our bid to guarantee the certainty and

accuracy of the results obtained, discovered DAG together with its novel relationships (regulatory genes) were further analysed and examined via a sensitivity and specificity test with respect to a prior DAG from both GSE40066 and GSE25011. Interestingly some predicted gene relationships from GSE40066 were also found in existing literature as being possible cause of cancer considering their role they play thus, increasing further the belief in the newly predicted results by the algorithm. Again, a fraction of new gene relationships was discussed with respect to their role they play leading to the causality of cancer (see Section 4). A very essential feature in drug discovery.

References

Adabor, E.S. and Acquaah-Mensah, G.K. (2019) *Restricted-Derestricted Dynamic Bayesian Network Inference of Transcriptional Regulatory Relationships Among Genes in Cancer*, https://authors.elsevier.com/a/1YdmZ5FQ7aFjle

Adabor, E.S., Acquaah-Mensah, G.K. and Oduro, F.T. (2015) 'SAGA: a hybrid search algorithm for bayesian network structure learning of transcriptional regulatory networks', *J. Biomed. Inform.*, Vol. 53, pp.27–35.

Bansal, M., Belcastro, V., Ambesi-Impiobato, A. and Bernardo, D. (2007) 'How to infer gene networks from expression profiles', *Molecular Systems Biology*, Vol. 3, p.78, http://dx.doi.org/10.1038/msb4100120

Bolouri, H. (2011) *Computational Modeling of Gene Regulatory Networks: A Primer*, Imperial College Press, London.

Brian, L.N. and Patricia, A.L. (2010) *Transcriptional Control of Neural Crest Development*, San Rafael (CA), Morgan & Claypool Life Science, https://www.ncbi.nlm.nih.gov/books/NBK53145/

Carraro, E., Bonetta, S., Bertino, C., Lorenzi, E., Bonetta, S. and Gilli, G. (2016) 'Hospital effluents management: chemical, physical, microbiological risks and legislation in different countries', *Journal of Environmental Management*, Vol. 168, pp.185–199, https://doi.org/10.1016/j.jenvman.2015.11.021

Chickering, D.M. (1996) 'Learning equivalence classes of Bayesian network structures', *Proceedings of the Twelfth Annual Conference on Uncertainty in Artificial Intelligence*, Morgan Kaufmann Publishers Inc., San Francisco, CA, USA, pp.150–157.

Cooper, G.F. and Herskovits, E. (1992) 'A Bayesian method for the induction of probabilistic networks from data', *Mach. Learn*., Vol. 9, pp.309–347.

Darwiche, A. and Provan, G. (1996) 'Query DAGs: A practical paradigm for implementing belief-network inference', *Proceedings of Twelfth Conference on Uncertainty in Artificial Intelligence,* Portland OR, Morgan Kaufmann, San Francisco, pp.203–210.

Di Bernardo, D., Thompson, M. and Gardner, T. (2005) 'Chemogenomic profiling on genome-wide scale using reverse-engineered gene networks', *Nat. Biotechnol.*, Vol. 23, pp.377–383.

Friedman, N., Linial, M., Nachman, I. and Pe'er, D. (2000) 'Using Bayesian networks to analyze expression data', *J. Comput. Biol.*, Vol. 7, pp.601–620.

Gardner, T., Di Bernardo, D., Lorenz, D. and Collis, J. (2003) 'Inferring genetic networks and identifying compound mode of action via expression profiling', *Science*, Vol. 301, pp.102–105.

Greenfield, A., Madar, A., Ostrer, H. and Bonneau, R. (2010) 'DREAM 4: combining genetic and dynamic information to identify biological networks and dynamical models', *PLoS One*, Vol. 5, No. 10, p.e13397, http://dx.doi.org/10.1371/journal.pone.0013397

Gregoretti, F., Belcastro, V. and Di Bernardo, D. (2010) 'A parallel implementation of the network identification by multiple regression (NIR) algorithm to reverse-engineer regulatory gene networks', *PLoS One*, Vol. 5, No. 4, p.e10179.

Heckerman, D. (1995) 'A Bayesian approach to learning causal networks', *Proceedings of Eleventh Conference on Uncertainty in Artificial Intelligence*, Morgan Kaufmann Publishers Inc., San Francisco, CA, USA, pp.285–295.

Heckerman, D. (2008) 'A tutorial on learning with Bayesian networks', in Holmes, D.E. and Jain, L.C. (Eds.): *Innovations in Bayesian Networks. Studies in Computational Intelligence*, Vol. 156, Springer, Berlin, Heidelberg.

Husmeier, D. (2003) 'Sensitivity and specificity of inferring genetic regulatory interactions from microarray experiments with dynamic Bayesian networks', *J. Bioinform*., Vol. 19, No. 17, pp.2271–2282.

Huynh-Thu, V.A., Irrthum, A., Wehenkel, L. and Geurts, P. (2010) 'Inferring regulatory networks from expression data using tree-based methods', *PLoS One*, Vol. 5, No. 9, p.e12776.

Klebaner, F.C. (2005) *Introduction to Stochastic Calculus with Applications*, 2nd ed. Imperial College Press, UK.

Meyer, P.E., Kontos, K., Lafitte, F. and Bontempi, G. (2007) 'Information-theoretic inference of large transcriptional regulatory networks', *EURASIP J. Bioinform Syst Biol*., p.79879, http://dx.doi.org/10.1155/2007/79879

Neapolitan, R.E. (2004) *Learning Bayesian Networks*, KDD '04.

Needham, C.J., Bradford, J.R., Bulpitt, A.J. and Westhead, D.R. (2007) 'A primer on learning in bayesian networks for computational biology', *PLoS Comput Biol*, Vol. 3, No. 8, e129, doi: 10.1371/journal. pcbi.0030129 (2007).

Perotti, E.A., Georgopoulos, K. and Yoshida, T. (2015) *An Ikaros Promoter Element with Dual Epigenetic and Transcriptional Activities*, Retrieved from https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4489883/

Rashbass, J. (2016) *Understanding Cancer-the Importance of Patient Data*, https://publichealthmatters.blog.gov.uk/2016/02/04/understanding-cancer-the-importance-of-patient-data/

Remy, P. and Baltzinger, M. (2000) 'The ETS-transcription factor family in embryonic development: lessons from the amphibian and bird', *Oncogene*, Vol. 19, No. 55, pp.6417–6431, DOI: 10.1038/sj.onc.1204044, PMID: 11175358.

Shannon, P., Markiel, A., Ozier, O., Baliga, N., Wang, J., Ramage, D. and Ideker, T. (2003) 'Cytoscape: A software environment for integrated models of bimolecular interaction networks', *Genome Res*, Vol. 13, pp.2498–2504.

Sladeczek, J., Hartemink, A.J. and Robinson, J. (2008) *Banjo*, Retrieved from http://www.cs.duke.edu/–amink/software/banjo

Tahtakran, S. and Selleck, M.A. (2003) 'Ets-1 expression is associated with cranial neural crest migration and vasculogenesis in the chick embryo', *Gene Expression Patterns: GEP*, Vol. 34, pp.455–458.

Yu, J., Smith, V. and Wang, P. (2004) 'Advances to bayesian network inference for generating causal networks from observational biological data', *Bioinformatics*, Vol. 20, pp.3594–3603.