Tumor-specific Causal Inference model

We designed the TCI algorithm to infer the functional-impact relationships between SGAs and DEGs for a given tumor using a bipartite causal Bayesian network (CBN) in which causal edges are only allowed to point from SGAs to DEGs. We assumed that each DEG is likely regulated by one aberrant pathway in an individual tumor and that such a pathway is likely perturbed by a single SGA due to the well-known mutual exclusivity among SGAs perturbing a common pathway (32-34).

Let **T** = {*T1*, *T2*, …, *Tt*, …, *TN*} denote the tumor set that contains *N* tumor samples, where *t* indexes over the tumors included in the tumor set. Let *SGA\_SETt* = {*A1*, *A2*, …, *Ah*, …, *Am*} denote a subset of *m* genes with genome alterations in tumor *t* (i.e., the SGAs), where *h* indexes over the variables in *SGA\_SETt*. Let *DEG\_SETt* = {*E1*, *E2*, …, *Ei*, …, *En*} denote *n* genes that are differentially expressed in tumor *t*, where *i* indexes over the variables in *DEG\_SETt* (i.e., the DEGs). We further included a variable *A0* to collectively represent factors other than SGAs (e.g., tumor microenvironment or hypoxia) that may cause differential gene expression.

For a given tumor *t*, a model *M* is a bipartite CBN in which, for each variable *Ei* in *DEG\_SETt*, there is exactly one arc into *Ei* from a variable in *SGA\_SETt*. In model *M*, a given *Ah* can have zero, one, or more arcs emanating from it to the variables in *DEG\_SETt*. Within the tumor *t*, TCI searches for a model *M* that best explains the data by assigning the *Ah* in *SGA\_SETt* that is the most probable cause for each *Ei* in *DEG\_SETt* . Pseudocode for the TCI algorithm is given below. Also shown is pseudocode for the algorithm FindGlobalDrivers, which produces results used by TCI, and thus must be called prior to calling TCI.

**FindGlobalDrivers**(*D*: dataset): set;

let *SGA\_SET* be the set of all SEGs across all tumors

let *DEG\_SET* be the set of all DEGs across all tumors

let *G* be an array of global drivers, one for each DEG variable

//a global driver of a DEG is the single best SGA that predicts that DEG over all tumors

for *Ei* ∊ *DEG\_SET* do

*bestScore* := 0;

for *Ah* ∊ *SGA\_SET* do

*score* := *eglobal*(*h*, *i*); //see Equation 5 below

if *score* > *bestScore* then

*bestSGA* := *Ah*;

*bestScore* := *score*;

end for;

*G*(*Ei*).*bestGlobalSGA* := *bestSGA*;

*G*(*Ei*).*bestGlobalScore* := *bestScore*;

end for;

return *G*;

**TCI**(*t*: tumor; *D*: dataset; *G*: array of global drivers): CBN;

let *M* be a CBN;

initialize *M* to contain no arcs;

*SGA\_SETt* := SGAs in tumor *t*;

*DEG\_SETt* := DEGs in tumor *t*;

for *Ei* ∊ *DEG\_SETt* do

*bestScore* := 0;

for *Ah* ∊ *SGA\_SETt* do

*score* :=*elocal*(*h*, *i*); //see Equation 8 below

if *score* > *bestScore* then

*bestSGA* := *Ah*;

*bestScore* := *score*;

end for;

place the arc *bestSGA* 🡪 *Ei* into *M*;

end for;

return *M*;

TCI employs a Bayesian framework to infer causal inference. In a tumor *t*, TCI scores an arc *Ah* 🡪 *Ei* between SGA *Ah* and DEG *Ei* based on the posterior probability of the arc, which can be derived as follows using Bayes rule:

, (1)

where is the data about variables *Ah* and *Ei*; is the marginal likelihood of the data given that *Ah* is modeled as the cause of *Ei*; and is the prior probability that *Ah* is the cause of *Ei*. Since the posterior probability shown on the left side of Equation 1 is proportional to the numerator on the right side, TCI simply uses the numerator as the arc score. In the text below, we discuss how to derive the numerator in both a global and tumor-specific manner.

The *tumor-specific* nature of TCI is reflected in the following points: 1) Each tumor has a unique *SGA\_SETt* and the *DEG\_SETt*; thus, the TCI-inferred CBN structure *M* is tumor-specific. 2) The prior probability of a model [which is based on the component priors ] is tumor-specific. 3) Calculation of the marginal likelihood of an arc consists of two components: one component is computed using the data from the tumors in which *Ah* =1, (i.e., “tumors like me”), and the other component is derived using a global model. As such, the posterior probability of the edge is specific to a given tumor, and the same edge may have different posterior probabilities in different tumors depending on the composition of *SGA\_SETt.*

***Tumor-specific model priors.*** While it would be ideal to define the prior probability for each edge using prior knowledge, we usually have very limited information about it. Despite of the lack of specific prior knowledge, we use the prior probability to incorporate the following general knowledge (on which frequency-oriented approaches are based) regarding the probability that a gene is a driver: the more frequently a gene is perturbed in a cancer cohort, the more likely it is a driver in an individual tumor. Furthermore, we also consider the tumor-specific context of the SGAs in each tumor.

To define an informative prior that can represent the biological foundations of different genome alterations in the tumor cells can help us effectively correct model bias and thus make accurate predictions (29,35). Therefore, we need to specify the model prior for each SGA *Ah* in each tumor *t* by comparing its alteration frequency in the tumor cohort against normal cells. In our paper, we used additional genomic information for both SM and SCNA to derive the prior probability of each edge using existing prior knowledge. We calculated and collected the following SGA information for each gene *h*: (1) the MutSigCV *p* value for *h* among the tumors in *D* from TCGA, and (2) the copy number amplification and deletion of *h* in a normal population without cancer from 1000 genome project (http://www.internationalgenome.org/) (64,65). Such information can be applied to help account for mutation and copy number alterations that are due to differences in gene lengths and chromosome locations which doesn’t depend on SGA frequency.

For a tumor *t* and an arbitrary DEG *Ei*, we defined the prior probability of *Ah* being a parent of *Ei* using a multinomial distribution with a parameter vector , where . Here, *θ*0 is a user-defined parameter representing the prior belief that the non-SGA factor *A0* being the cause of *Ei*, and represents the prior probability of *Ah* being the cause of *Ei*. In this study, we set = 0.1. We assumed that where is a tumor-specific Dirichlet parameter vector governing the distribution of . For a tumor *t*, we calculated the prior probability as follows:

(2)

where *h’* indexes over the *m* variables in *SGA\_SETt*; is a Dirichlet parameter and is MutSigCV *p* value for *Ah*.

***Marginal likelihood function P(D|M)***. In TCI, the overall marginal likelihood of the data *D* given model *M* is a product of the marginal likelihood of each arc` in *M*. The term is the marginal likelihood function of *M*, which can be derived by marginalizing out model parameters as follows:

where represents the parameters (probabilities) associated with causal Bayesian network structure *M*. Since our data (*A* and *E*) are discrete variables, the parameters and their prior distributions are multinomial and Dirichlet respectively. In addition, TCI uses the Bayesian Dirichlet equivalent uniform (BDeu)(66) scoring measure to derive the marginal likelihood for each arc in *M* as a function that is expressed in terms of the products and ratios of gamma functions. Taken together, we have the following:

, (4)

where *j* indexes over the states of the cause of *Ei* in *M* (i.e., some variable *Ah*); *qi* is the number of possible values of *Ah* (in our case, it is 2, because *Ah* is modeled as a binary variable); *k* is variable which indexes over the states of *Ei*; *ri* denotes the total possible states of *Ei* (in our case, it is set to 2); *Nijk* is the number of tumors in dataset *D* in which node *Ei* has value *k* and its cause *Ah* in *M* has the value denoted by *j*; *αijk* is a parameter in a Dirichlet distribution that represents prior belief about *P*(*Ei* | *cause*(*Ei*)); Γ is the gamma function; ; and .

Let represent the function that calculates the Bayesian score of the edge in tumor *t* (i.e., the numerator of the Eq (1)) over all the data. Then, can be defined as follows:

. (5)

where can be calculated using Equation (2).

The tumor-specific calculation of the marginal likelihood uses the data that are most relevant to the given tumor (aka “tumors like me”) to infer if the hypothesis of a candidate causal edge is supported among these tumors. To do so, we modified the Bayesian scoring function shown in Equation 5 by dividing the training data *D* into a tumor-specific subset (the subset of tumors with *Ah* = 1 for the current tumor *t*) and the remaining data (the subset of tumors with *Ah* = 0). Let denote the subset of tumors in which *Ah*=1 and denote the subset of tumors in which *Ah*=0, such that . Let *G*(*i*) represent the SGA with maximal Bayesian score for *Ei* derived from Equation 5 at the whole cohort level (referred to as “global driver”) and let denote the prior probability that *G*(*i*) is the cause of *Ei* in tumor *t*.

We can calculate the and for tumors with *Ah*=1 and those with *Ah*=0, respectively, as follows:

(6)

(7)

(8)

where is the number of tumors in in which *Ei* has value *k* and *Ah* = 1; is the number of tumors in that *Ei* has value *k*; and *AG(i)* has the value indexed by *j*. Finally, the posterior probability of a causal edge can be calculated with the following equation:

(9)