Supplement to "Subnetwork State Functions Define Dysregulated Subnetworks in Cancer" *

Salim A. Chowdhury[1], Rod K. Nibbe[2,4] Mark R. Chance[3,4], Mehmet Koyutürk[1,4]

(1) Dept. of Electrical Engineering & Computer Science, (2) Dept. of Pharmacology

(3) Dept. of Physiology & Biophysics, (4) Center for Proteomics & Bioinformatics

Case Western Reserve University, Cleveland, OH, USA.

In this supplementary document, we prove two theorems that support the algorithms described in [1] and we present experimental results from systematic study of the performance of CRANE when the value of the tuneable parameters are varied. First, in Theorem 1, we provide bound on the score of a state function that can be obtained by extending that state function, in terms of simple statistics of the smaller state function. We use this result to develop efficient search algorithms that grow subnetworks in a bottom-up fashion, effectively pruning the search space using the bound provided by this theorem. Then we provide two lemmas that are used in the proof of this theorem. Subsequently, in Theorem 2, we provide a tight bound on the information that any subnetwork state function can provide on the phenotype ($J(f_{\mathcal{S}}, C)$), for given number of phenotype and control samples. This result is useful for obtaining a uniform criterion to score subnetwork state functions. Specifically, by normalizing $J(.)$ by the bound provided by this theorem, we obtain a scoring criterion that ranges from 0 to 1 regardless of the number of phenotype and control samples in the dataset. Finally, we provide the pseudo-code for the subnetwork search algorithm implemented by CRANE.

In this document, we use the following conventions for notational convenience:

- q denotes $p(1) = P(C = 1)$, that is the probability that the sample is associated with the phenotype of interest.
- z denotes $p(f_{\mathcal{S}}) = P(F_{\mathcal{S}} = f_{\mathcal{S}})$, that is the probability that subnetwork $\mathcal{S}$ is in state $f_{\mathcal{S}}$ in a given sample.
- s denotes $p(1|f_{\mathcal{S}}) = P(C = 1|F_{\mathcal{S}} = f_{\mathcal{S}})$, that is the probability that a sample with state $f_{\mathcal{S}}$ for the genes in $\mathcal{S}$ is associated with the phenotype of interest.
- r denotes $p(1|f_{\mathcal{R}}) = P(C = 1|F_{\mathcal{S}} = f_{\mathcal{R}})$, that is the probability that a sample with state $f_{\mathcal{R}}$ for the genes in $\mathcal{R}$ is associated with the phenotype of interest.
- $\mathcal{T}$ denotes $\mathcal{R} \setminus \mathcal{S}$ and $f_{\mathcal{T}}$ denotes the state of $\mathcal{T}$ that is consistent with $f_{\mathcal{R}}$.
- γ denotes $P(F_{\mathcal{T}} = f_{\mathcal{T}}|F_{\mathcal{S}} = f_{\mathcal{S}})$, that is the probability of observing state $f_{\mathcal{T}}$ for subnetwork $\mathcal{T}$, given that subnetwork $\mathcal{S}$ is in state $f_{\mathcal{S}}$.
- θ denotes $P(C = 1|F_{\mathcal{S}} = f_{\mathcal{S}}, F_{\mathcal{T}} \neq f_{\mathcal{T}})$, that is the probability that a sample is associated with the phenotype of interest, given that subnetwork $\mathcal{S}$ is in state $f_{\mathcal{S}}$, but subnetwork $\mathcal{T}$ is not in state $f_{\mathcal{T}}$ in that sample.

For other notation and definitions, please refer to the main article [1].

*The main article is to be published in the Proceedings of *14th Int'l Conf. on Research in Computational Molecular Biology (RECOMB'2010).*

THEOREM **1** *Consider a subnetwork $\mathcal{S} \subseteq \mathcal{V}$ and associated state function $f_{\mathcal{S}}$. For any $\mathcal{R} \supset \mathcal{S}$, the following bound holds:*

$$J(f_{\mathcal{R}}; C) \leq p(f_{\mathcal{S}}) \max_{c \in \{0,1\}} \left\{ p(c|f_{\mathcal{S}}) \log \frac{1}{p(c)} \right\} = J_{bound}(f_{\mathcal{S}}, C). \tag{1}$$

PROOF. This theorem and its proof are based on a more general result by Smyth & Goodman [2] in the context of association rule mining.

Using the notation introduced at the beginning of this document, we can write $J(f_{\mathcal{R}}; C)$ and $J_{\text{bound}}(f_{\mathcal{S}}, C)$ as follows:

$$J(f_{\mathcal{R}}; C) = z\gamma(r \log(r/q) + (1-r) \log((1-r)/(1-q))). \tag{2}$$

$$J_{\text{bound}}(f_{\mathcal{S}}; C) = z \max\{s \log(1/q), (1-s) \log(1/(1-q))\}. \tag{3}$$

We will show that, for fixed $\mathcal{S}$ and $f_{\mathcal{S}}$, the maximum value that $J(f_{\mathcal{R}}; C)$ attains cannot exceed $J_{\text{bound}}(f_{\mathcal{S}}; C)$ (for any choice of $\mathcal{T}$ and $f_{\mathcal{T}}$). First, by definition of conditional probability, we note the following equality:

$$s = \gamma r + (1-\gamma)\theta. \tag{4}$$

Since s is fixed, this equation represents a constraint that must be satisfied by r, γ, and θ. Thus, we will bound $J(f_{\mathcal{R}}; C)$ subject to this constraint. Note also that we can write this constraint as

$$\gamma = \frac{s-\theta}{r-\theta} = \frac{\theta - s}{\theta - r}. \tag{5}$$

Without loss of generality, we assume $s > q$, that is the observation of state function $f_{\mathcal{S}}$ increases the probability of a sample being associated with the phenotype ($f_{\mathcal{S}}$ "indicates" phenotype). Since we consider only two classes for the samples (phenotype or control), if the assumption does not hold (i.e., if $f_{\mathcal{S}}$ "indicates" control), then the following arguments still hold if we simply interchange the labels of sample classes.

Given that $s > q$, five different cases are possible: (i) $q < s < r$, (ii) $q < s = r$, (iii) $q < r < s$, (iv) $q = r < s$, and (v) $r < q < s$. We consider each case separately.

Case (i): $q < s < r$. In this case, the probability of phenotype given the state of the larger subnetwork is larger than the probability of phenotype given the state of the smaller subnetwork (and thus the additional part of the larger subnetwork provides additional evidence indicating that the sample might be associated with phenotype).

Since $s < r$, we have $r > \gamma r + (1-\gamma)\theta$ from (4) and thus $r > \theta$. Therefore, since $0 \leq \theta < s < r \leq 1$, we can write by Lemma 1 (see below) that $\gamma \leq s/r$, without putting any additional constraint on r. Consequently, from (2), we obtain

$$J(f_{\mathcal{R}}; C) \leq z\frac{s}{r}(r \log(r/q) + (1-r) \log((1-r)/(1-q))) \tag{6}$$

and thus

$$J(f_{\mathcal{R}}; C) \leq zs(\log(r/q) + (1/r - 1) \log((1-r)/(1-q))). \tag{7}$$

Since $q < r \leq 1$, the second term in parenthesis is negative. Consequently, noting $r \leq 1$, we obtain

$$J(f_{\mathcal{R}}; C) \leq zs \log(1/q) \leq J_{\text{bound}}(f_{\mathcal{S}}, C). \tag{8}$$

This proves the theorem for case (i).

Case (ii): $q < s = r$. In this case, the probability of phenotype given the state of the larger subnetwork is equal to the probability of phenotype given the state of the smaller subnetwork (and thus the additional part of the larger subnetwork does not provide additional information).

Noting $\lambda \leq 1$ and replacing r with s, we can write

$$J(f_{\mathcal{R}}; C) \leq z(s \log(s/q) + (1-s) \log((1-s)/(1-q))). \tag{9}$$

Since $1 - s < 1 - q$, the second term in parentheses is negative, so we have

$$J(f_{\mathcal{R}}; C) \leq z(s \log(s/q)) \leq J_{\text{bound}}(f_{\mathcal{S}}, C). \tag{10}$$

This proves the theorem for case (ii).

Case (iii): $q < r < s$. In this case, the observation of the state of the larger subnetwork increases the probability of phenotype compared to background, but not to the extent that the smaller subnetwork does.

The proof here is very similar to that in case (ii). Let $y(x) = x\log(x/q) + (1-x)\log((1-x)/(1-q))$. Then we have $y'(x) = \log(x/q) - log((1-x)/(1-q))$. Therefore, for $x > q$, since $x/q > 1$ and $(1-x)/(1-q)$, $y'(x)$ is always positive and y is an increasing function of x. Consequently, for $q < r < s$, we have:

$$r\log(r/q) + (1-r)\log((1-r)/(1-q)) \leq s\log(s/q) + (1-s)\log((1-s)/(1-q)). \tag{11}$$

Once this inequality is etablished, the rest of the proof for case (iii) follows the proof for case (ii).

Case (iv): $q = r < s$. In this case, the probability of phenotype given the state of the larger subnetwork is equal to background, thus the additional part of the larger subnetwork takes away all the evidence provided by the smaller subnetwork in favor of phenotype.

By definition of $J(.)$, $J(f_{\mathcal{R}}; C) = 0$ (both r/q and $(1-r)/(1-q)$ are equal to 1 in (2)). Thus, $J(f_{\mathcal{R}}; C)$ trivially satisfies the bound, proving the theorem for this case.

Case (v): $r < q < s$. In this case, the additional part of the larger subnetwork reverses the direction of evidence provided by the smaller subnetwork, that is the state function of the larger subnetwork increases the probability of the sample being associated with control.

The proof in this case is very similar to that for case (i). Since $r < s$, using Equation 4 we have $r < \gamma r + (1-\gamma)\theta$ and thus $r < \theta$. Therefore, since $0 \leq r < s < \theta \leq 1$, we can write by Lemma 2 (see below) that $\gamma \leq (1-s)/(1-r)$, without putting any additional constraint on r. Consequently, from (2), we obtain

$$J(f_{\mathcal{R}}; C) \leq z\frac{1-s}{1-r}(r\log(r/q) + (1-r)\log((1-r)/(1-q))) \tag{12}$$

and thus

$$J(f_{\mathcal{R}}; C) \leq z(1-s)(\log((1-r)/(1-q)) + (r/(1-r))\log(r/q)). \tag{13}$$

Since $r < q$, the second term in parenthesis is negative and also $1 - r \leq 1$; therefore

$$J(f_{\mathcal{R}}; C) \leq z(1-s)\log(1/(1-q)) \leq J_{\text{bound}}(f_{\mathcal{S}}, C). \tag{14}$$

This proves the theorem for case (v). ∎

LEMMA **1** *For* $0 \leq x < a < b \leq 1$, $\dfrac{(a-x)}{(b-x)} \leq \dfrac{a}{b}$.

PROOF. Let $x_1 < x_2$. Since $b - a > 0$, we have $x_1(b-a) < x_2(b-a)$. Adding $x_1x_2 + ab$ to both sides of the inequality, we obtain $(a-x_2)(b-x_1) < (a-x_1)(b-x_2)$. Consequently, $x_1 < x_2$ implies

$$\frac{a-x_2}{b-x_2} < \frac{a-x_1}{b-x_1}, \tag{15}$$

and therefore the maximum of $\dfrac{(a-x)}{(b-x)}$ for the interval $0 \leq x < a$ occurs at $x = 0$, which is equal to $\dfrac{a}{b}$. ∎

LEMMA **2** *For* $0 \leq b < a < x \leq 1$, $\dfrac{(x-a)}{(x-b)} \leq \dfrac{1-a}{1-b}$.

PROOF. Let $x_1 > x_2$. Since $a - b > 0$, we have $x_1(a-b) > x_2(a-b)$. Adding $x_1x_2 + ab$ to both sides of the inequality, we obtain $(x_1-a)(x_2-b) > (x_1-b)(x_2-a)$. Consequently, $x_1 > x_2$ implies

$$\frac{x_1-a}{x_2-a} > \frac{x_1-b}{x_2-b}, \tag{16}$$

and therefore the maximum of $\dfrac{(x-a)}{(x-b)}$ for the interval $a < x \leq 1$ occurs at $x = 1$, which is equal to $\dfrac{1-a}{1-b}$. ∎

THEOREM **2** *For a given gene expression dataset, let the fraction of phenotype samples be q. Then, for any subnetwork $\mathcal{S} \subseteq \mathcal{V}$,*

$$0 \leq J(f_{\mathcal{S}}, C) \leq \max\{-q \log q, -(1-q)\log(1-q)\}. \tag{17}$$

PROOF. Assume that q and z are fixed. Then we can write $J(f_{\mathcal{S}}; C)$ as a function of s:

$$J(s) = z(s \log(s/q) + (1-s)\log((1-s)/(1-q))). \tag{18}$$

Taking the derivative of this function with respect to s, we obtain

$$J'(s) = z \log \frac{s(1-q)}{q(1-s)}. \tag{19}$$

Observe that $J'(s)$ assumes its zero at $s = q$. Furthermore, for $s > q$, since $s/q > 0$ and $(1-s)/(1-q) < 0$, $J'(s)$ is always positive and J is an increasing function of s. Similarly, for $s < q$, $J'(s)$ is always negative and J is a decreasing function of s. Consequently, $J(s)$ is always non-negative and it assumes its maximum at one of the boundaries of the range of values that s can take. Therefore, for fixed q, if we bound $J(s)$ at the boundaries that are enforced by z, we can write the bound on J as a function of z. The maxima of this function over all values of z will provide a bound on J over all possible values of z and s for fixed q. We analyze the cases $z \geq q$ and $z \leq q$ separately.

Case A: $z \geq q$, that is the state function is observed at least as commonly as the phenotype of interest. In this case, since the number of phenotype samples in which the state function is observed can be at most equal to the number of all phenotype samples, we have $s \leq q/z$. On the other hand, if $z \leq 1 - q$, then it is possible that none of the samples that exhibit the state function are associated with the sample, and therefore $s \geq 0$. Finally, when $z \leq 1 - q$ (which is only possible if $z \geq 1/2$, the s will be minimized if all samples that are not associated with the phenotype exhibit the state function, and therefore we have $s \geq 1 - (1-q)/z$. Consequently, we have three boundary cases for s:

1. $s = q/z$, subject to $q \leq z \leq 1$.
2. $s = 0$, subject to $q \leq z \leq 1 - q$.
3. $s = 1 - (1-q)/z$, subject to $\max\{q, 1-q\} \leq z \leq 1$.

We consider each of these boundary cases separately.

Case A1: Letting $s = q/z$ in (18), we obtain

$$J_{A1}(z) = z\left(\frac{q}{z}\log\frac{1}{z} + \frac{z-q}{z}\log\frac{z-q}{z(1-q)}\right), \tag{20}$$

and therefore $J_{A1}(z) = (z-q)\log((z-q)/(1-q)) - z\log z$. Consequently, $J'_{A1}(z) = \log((z-q)/(z(1-q))) \leq 0$ for $q \leq z \leq 1$ and therefore $J_{A1}(z) \leq J(q) = -q \log q$, proving the bound for this case.

Case A2: Letting $s = 0$ in (18), we obtain $J_{A2}(z) = -z\log(1-q)$ and therefore $J_{A2}(z) \leq J(1-q) = (1-q)\log(1-q)$ for $q \leq z \leq 1-q$, proving the bound for this case.

Case A3: Letting $s = 1 - (1-q)/z$ in (18), we obtain

$$J_{A3}(z) = z\left(\frac{q+z-1}{z}\log\frac{q+z-1}{qz} + \frac{1-q}{z}\log\frac{1}{z}\right) \tag{21}$$

and therefore $J_{A3}(z) = (q+z-1)\log((q+z-1)/q) - z\log z$. Consequently, $J'_{A3}(z) = \log((q+z-1)/qz)$. $J'_{A3}(z)$ assumes its zero at $z = 1$, corresponding to a minimum at $J(1) = 0$. Therefore, if $q \leq 1-q$, then $J_{A3}(z)$ attains its maximum at $z = 1-q$, which gives $J_{A3}(z) \leq J_{A3}(q) = -(1-q)\log(1-q)$. Otherwise ($q > 1-q$ and hence $q > 1/2$), $J_{A3}(z)$ attains its maximum at $z = q$, which gives $J_{A3}(z) \leq J_{A3}(1-q) = (2q-1)\log((2q-1)/q) - (1-q)\log(1-q) \leq -(1-q)\log(1-q)$ since $(2q-1)/q \leq 1$ for $1/2 \leq q \leq 1$. This proves the bound for this case.

Case B: $z \leq q$, that is the state function is observed at most as commonly as the phenotype of interest. In this case, s can attain the value 1 if all samples that exhibit the state function are associated with the phenotype of interest, thus $s \leq 1$. On the other hand, for $z \leq 1-q$, s can be as low as 0 if all samples that exhibit the state function are samples that are not associated with the phenotype. Finally, if $z \geq 1-q$, then s has to be at least $1-(1-q)/z$ since at most this fraction of samples that exhibit the state function can be samples that are not associated with the phenotype. Consequently, we have three boundary cases for s:

1. $s = 1$, subject to $0 \leq z \leq q$.
2. $s = 0$, subject to $0 \leq z \leq \min\{1-q, q\}$.
3. $s = 1-(1-q)/z$, subject to $1-q \leq z \leq q$.

We consider each of these boundary cases separately.

Case B1: Letting $s = 1$ in (18), we obtain $J_{B1}(z) = -z \log q$ and therefore $J_{B1}(z) \leq -q \log q$ for $0 \leq z \leq q$, proving the bound for this case.

Case B2: Letting $s = 0$ in (18), we obtain $J_{B2}(z) = -z \log(1-q)$ and therefore $J_{B2}(z) \leq -(1-q)\log(1-q)$ for $q \leq z \leq 1-q$, proving the bound for this case.

Case B3: Observe that $J_{B3}(z) = J_{A3}(z)$. As we know from case A3, $J_{B3}(z)$ is a decreasing function of z and $J_{A3}(1-q) \leq -(1-q)\log(1-q)$, so $J_{B3}(1-q) \leq -(1-q)\log(1-q)$, proving the bound for this case. ∎

Algorithm 1 CRANE-EXTENDSTATEFUNCTION $((\mathcal{S},f_{\mathcal{S}}),\mathcal{T}, j^*, b, d)$: Extends a subnetwork and associated state function. Invoked for each $g_i \in \mathcal{V}$ and $\hat{E}_i \in \{0,1\}$ as CRANE-EXTENDSTATEFUNCTION$((\{g_i\},\{\hat{E}_i\}), \emptyset, j^*, b, d)$, where j^*, b, and d are user-defined.

Global: ▷ $\mathcal{V}$: Set of genes, C: Phenotype vector
▷ E: Gene expression associated with $\mathcal{V}$ and C, $\mathcal{E}$: PPI dataset associated with $\mathcal{V}$
Inputs: ▷ $(\mathcal{S},f_{\mathcal{S}})$: Subnetwork/state-function pair to be extended
▷ j^*: Threshold on information provided by a state function on phenotype
▷ d: Maximum number of genes in a subnetwork
▷ b: Maximum number of immediate extensions of a subnetwork/state-function pair
Input/Output: ▷ $\mathcal{T}$: Set of state functions informative of phenotype.

if $|\mathcal{S}| == d$ **then**
  if $J(f_{\mathcal{S}};C) \geq j^*$ **then**
    $\mathcal{T} \leftarrow \mathcal{T} \cup \{(\mathcal{S}, f_{\mathcal{S}})\}$
  end if
  return
end if
$\mathcal{Q} \leftarrow \emptyset$; $g_i \leftarrow$ most recently added gene to $\mathcal{S}$
for each $g_k : (g_i, g_k) \in \mathcal{E}$ **and** $g_k \notin \mathcal{S}$ **do**
  $\mathcal{S}' \leftarrow \mathcal{S} \cup \{g_k\}$
  for each $\hat{E}_k \in \{0,1\}$ **do**
    $f_{\mathcal{S}'} \leftarrow f_{\mathcal{S}'} \cup \hat{E}_k$; *redundant* $\leftarrow$ **false**
    for $g_j \in \mathcal{S}'$ **do**
      if $(J(f_{\mathcal{S}'\setminus\{g_j\}};C) \geq J(f_{\mathcal{S}'};C)$ **then**
        redundant $\leftarrow$ **true**; **break**
      end if
    end for
    if (**not** *redundant*) **and** $((J_{\max}(f_{\mathcal{S}'};C) > j^*))$ **then**
      $\mathcal{Q} \leftarrow \mathcal{Q} \cup \{(\mathcal{S}', f_{\mathcal{S}'})\}$
    end if
  end for
end for
if $\mathcal{Q} = \emptyset$ **then**
  if $J(f_{\mathcal{S}};C) \geq j^*$ **then**
    $\mathcal{T} \leftarrow \mathcal{T} \cup \{(\mathcal{S}, f_{\mathcal{S}})\}$
  end if
  return
end if
$\mathcal{Q}_b \leftarrow$ set of top b subnetwork/state-function pairs in Q with respect to $J(f_{\mathcal{S}'};C)$
for each $(\mathcal{S}', f_{\mathcal{S}'}) \in \mathcal{Q}_b$ **do**
  CRANE-EXTENDSTATEFUNCTION$((\mathcal{S}', f_{\mathcal{S}'}), \mathcal{T}, j^*, b, d)$
end for

Effect Of Parameters on Performance of CRANE

Here, we investigate the systematic effect of parameters used to configure CRANE on classification performance of identified subnetworks. We vary the parameter under test and fix the other parameters. The classification performance is reported using the average 'F Measure' for top 10 subnetworks. The tuneable parameters of CRANE are:

- d: d is the maximum number of genes belonging to a subnetwork. CRANE stops extending a subnetwork when the number of genes in the subnetwork rearches d. In other words, d determines the depth of the search.
- b: b is the number of state functions selected by CRANE at each iteration with maximum $J(.)$ value. So b determines the breadth of the search.
- j^{**}: j^{**} is the minimum J value each subnetwork state function must satisfy in order to be considered extensible. Thus this parameter allows pruning of larger state functions using statistics of smaller state functions.
- α: α is the fraction of the entries in the normalized gene expressin matrix that is set to H (high expression). The rest of the $(1-\alpha)$ entries of the gene expression matrix is set to L (low expression).

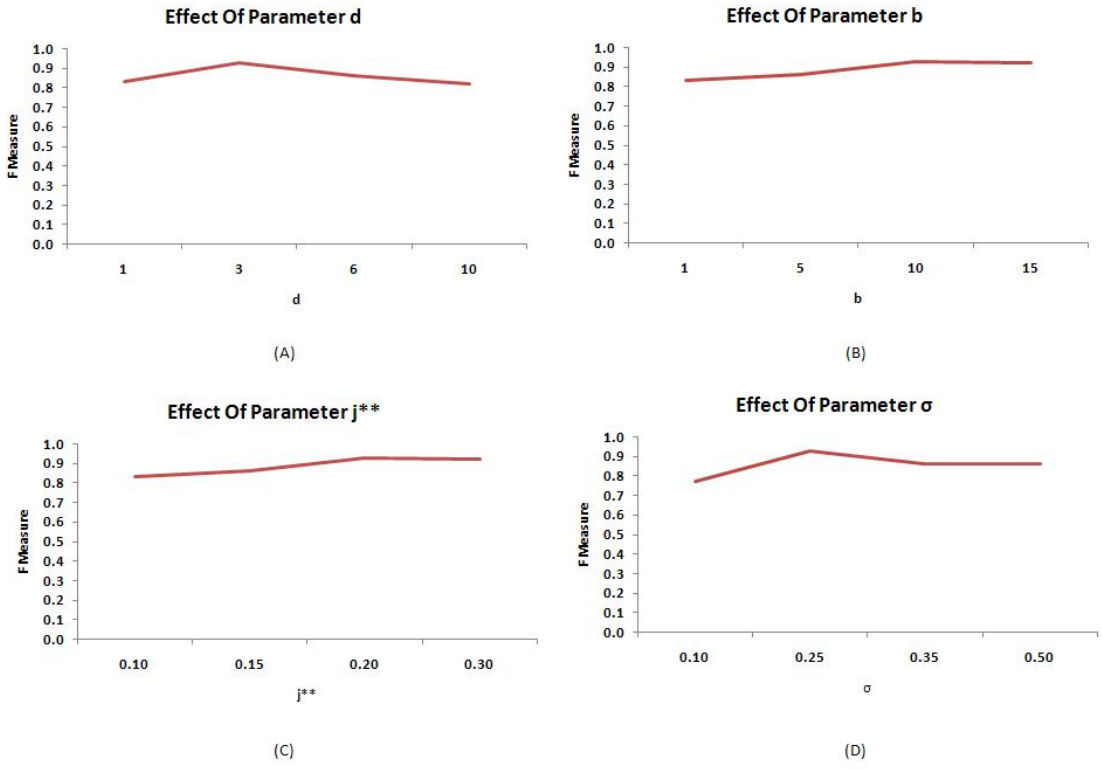

Figure 1: Peformance of CRANE across different value of the tuneable parameters. For all experiments, subnetworks are discovered on $GSE3964$ and tested on samples of $GSE6988$. F Measure of the classifier is calculated by increasing the number of subnetworks gradually from 1 to 10 and average F Measure is reported. (A), (B), (C) and (D) show performance of CRANE for parameters d, b, j^{**} and α respectively. In (A), $b = 10$ and $j^{**} = 0.15$. In (B), $d = 3$ and $j^{**} = 0.15$. In (C), $b = 10$ and $d = 3$. In (D), $b = 10$, $j^{**} = 0.15$ and $d = 3$

.

Figure shows the effect of choosing different value of the above mentioned parameters on the classification performance of CRANE. We discuss the results one at a time for each parameter.

- d: A high value of d indicates larger subnetworks. Larger subnetworks allow us to gain more insights into the mechanistic bases of diseases by incorporating more interactions amongst the genetic products that might be informative. But large subnetworks might also pose a negative impact on the performance of the classifiers by incorporating more dimensionality. As can be seen from the figure, while values of $d > 1$ give impressive classification accuracies of more than 80%, the performance shows slight degradation when $d > 3$. This might be attributed to the curse of the dimensionality associated with the larger subnetworks.

- b: The classifier shows better performance when the breadth of the search, b, is increased. The graph show expected monotonically increasing behavior for all the 4 cases where b takes value ranging within 1 to 15.

- j^{**}: Large value j^{**} ensures that the discovered subnetworks are highly informative of the phenotype and thus, it is quite natural that, subnetworks associated with high j^{**} would show better performance in terms of classifying the samples. The upward trend in the graph of Figure (C) with increased value of j^{**} shows consistency between the experimental results and theoretical aspect.

- α: Although CRANE shows very good performance across different values of α, it's performance reaches the peak when 25% of the genes are expressed across all samples. Thus, it can be hypothesized from this performance trend, that, on average, in a cell, 25% of the genes are expressed.

Thus from the discussion above, it is easy to comprehend that the performance of CRANE across different value of the parameters matches our logical understanding of the algorithm.

References

[1] S. A. Chowdhury, R. K. Nibbe, M. R. Chance, and M. Koyutürk. Subnetwork state functions define dysregulated subnetworks in cancer. In *14th Int'l Conf. on Research in Computational Molecular Biology (RECOMB'2010)*, in press.

[2] P. Smyth and R. M. Goodman. An information theoretic approach to rule induction from databases. *IEEE Trans. on Knowl. and Data Eng.*, 4(4):301–316, 1992.