

Gene exclusion / inclusion criterion

This concerns the paper *Finding disease specific changes in the co-expression of genes* published in *Bioinformatics 20 (Suppl. 1), 2004: i194-i199*. The exclusion / inclusion criterion given on page i196, i.e. formula (2), is not correct and here we give the correct version.

For either excluding a gene k from I or including a gene k' not yet in I we can decompose the score $S(I)$ in the following manner:

$$S(I) = \frac{A_k^{(1)} + B_k^{(1)}}{A_k^{(2)} + B_k^{(2)}} := \frac{A_1 + B_1}{A_2 + B_2} .$$

Further let $A_1/A_2 =: a$ and $B_1/B_2 =: b$. If we neglect the effects of re-fitting the parameters, we can write the new approximate score \tilde{S} in case of excluding / including a gene as:

$$S(I) = s , \quad \tilde{S}(I \setminus k) = a \quad (\text{exclusion})$$

$$S(I) = a , \quad \tilde{S}(I \cup k') = s \quad (\text{inclusion})$$

Then the criteria for excluding / including a gene take the form:

$$b > s \implies s > a \quad (\text{exclusion})$$

$$b < s \implies a > s \quad (\text{inclusion})$$

and can be rewritten as:

$$C_k = \pm (b - s) > 0 \tag{1}$$

for exclusion and inclusion, respectively. That this is indeed the case can be seen as follows:

$$\left. \begin{aligned} b &= s \pm \delta \text{ with } \delta \in \mathbf{R}_+ \\ s &= \frac{A_1 + B_1}{A_2 + B_2} = \frac{aA_2 + bB_2}{A_2 + B_2} \end{aligned} \right\} \implies$$

$$s = \{aA_2 + (s \pm \delta)B_2\} / \{A_2 + B_2\} \iff$$

$$sA_2 + sB_2 - aA_2 - sB_2 = \pm \delta B_2 \iff$$

$$A_2(s - a) = \pm \delta B_2 \iff$$

$$s \gtrless a ,$$

where we have used that sums of squares and their quotients are non-negative.

Including the tuning parameter in equation (1) finally leads to:

$$C_k(\alpha) = \pm \alpha (b - s) \mp (1 - \alpha) 1/|I| > 0 \tag{2}$$

... which **should** have been formula (2) in the paper. Here one can also clearly see that a penalty is given when a gene is excluded ($+-$ combination of signs) and a bonus when a gene is included ($-+$ combination of signs). The size of the penalty / bonus does depend on how many genes are in the current group via $1/|I|$, i.e. it mainly prevents the groups from getting too small.