Iterative method for identification of essential proteins based subcellular localization, orthology and PPI networks

A brief introduction of SON

SON 是一种基于融合亚细胞定位信息、蛋白质同源信息及 PPI 拓扑特征的关键蛋白质识别算法，在关键蛋白质的识别上具有十分良好的性能，也是目前的 state-of-the-art method，其构成如下：

$$pr(\mathrm{i}) = (1-\alpha) * \mathrm{NOS}(\mathrm{i}) + \alpha * ((1-\beta) * \mathrm{NSL}(\mathrm{i}) + \beta * \mathrm{NNC}(\mathrm{i}))$$

其中 NOS 为蛋白质同源信息测度、NSL 为亚细胞定位信息测度、NNC 为 PPI 拓扑特征测度。

NNC

NNC 为 NC 的归一化形式：

$$\mathrm{NNC} = \frac{NC}{Max\_NC}$$

NOS

NOS 为蛋白质同源属性，

$$\mathrm{NOS(i)} = \frac{OS(i)}{Max\_OS}$$

其中 OS(i)为蛋白质 i 的同源物种的数目（同源属性是指不同物种的蛋白质具有相似的序列或结构， 例如人的蛋白质与酵母菌的蛋白质有一部分属于同源蛋白质，具有相似的功能，而该部分蛋白质则可能具有关键性，参与同源的物种越多，越倾向于为关键蛋白质）。

NSL

关键蛋白质倾向于出现在某些亚细胞位置中，所以出现在这些位置的蛋白质可能为关键蛋白质，SON 在计算时对每个亚细胞位置进行评分，由于一个蛋白质可能出现在多个亚细胞位置中（如转录蛋白等），因此蛋白质的亚细胞位置得分为该蛋白的所有位置得分的总和，

$$SL(\mathrm{i}) = \sum_{j \in N(i)} LCC(\mathrm{j})$$

其中N(i)为蛋白质的位置集合，LCC(j)为亚细胞位置得分。

而 LCC 的计算中采取的方法则是按照计算该位置出现的关键蛋白质的个数来评价该位置得分，该位置出现的关键蛋白质越多，得分越高，反之越低。

SON 的计算方法为实现通过 NNC 对所有蛋白质节点进行排序，然后取排名在前 5%的蛋白质为关键蛋白质，倒数 5%的作为非关键蛋白质，并统计这些蛋白质中的各个位置出现的频数fl与hl，

$$\mathrm{LCC(l)} = \begin{cases} 1 - \dfrac{\mathrm{fl}}{hl}, & \mathrm{fl} < \mathrm{hl} \\ \dfrac{\mathrm{fl}}{hl} - 1, & \mathrm{fl} \geq \mathrm{hl} \end{cases}$$

Our method

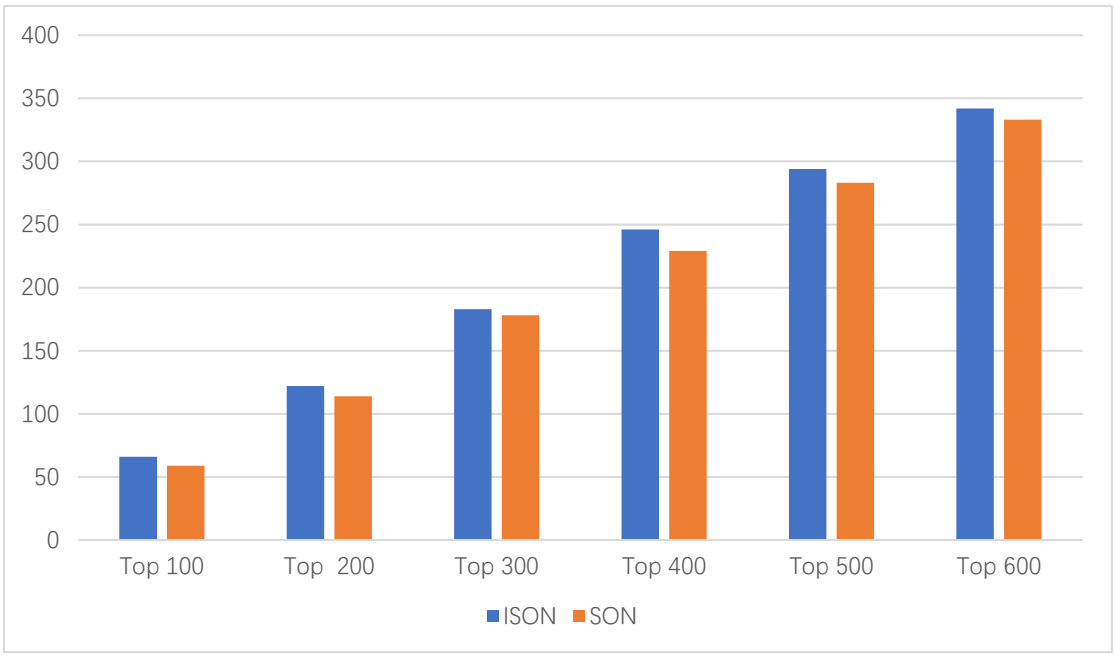
SON 的问题在于在计算 NSL 时只用了 NC 即 PPI 拓扑特征对蛋白质的关键性进行判别而 PPI 网络存在高假阳性，通过 PPI 网络得到的结果并不理想，而运用此方法对计算亚细胞位置得分也不准确，因此我们运用如下迭代方法计算：

```
Compute NNC, NOS first

for i = 1:itr

    if i == 1
        use NNC to compute NSL, compute pri = (1- α)* NOS (i)+  α *((1-
β )* NSL (i)+  β* NNC (i))
    else
        compute pri
        [~,I] = sort(pri,'descend');
        use I to compute NSL_i
    end

    use NSL_itr to compute pri

end
```

由于我们并未收集到 Orthologous proteins dataset，因此我们只比较 SON 算法的前半部分

$$pr(\mathrm{i}) = (1 - \beta) * \mathrm{NSL\,(i)} + \beta * \mathrm{NNC\,(i)}$$

只迭代运行一次，用来计算 NSL 的可靠排序结果来自另一篇报告 SGC，$\beta = 0.3$为 SON 设定参数，为方便表示我们将本文算法命名为 ISON（Iterative SON）.


400
350
300
250
200
150
100
50
0
Top 100
Top 200
Top 300
Top 400
Top 500
Top 600
ISON
SON