A genetic algorithm-based gene set selection using subpathway activations for patient stratification in terms of subtype and survival outcome

Bonil Koo[1], Dohoon Lee[2], Sangseon Lee[3], Inyoung Sung[1] and Sun Kim[1,4,5]

[1]Interdisciplinary Program in Bioinformatics, Seoul National University, [2]Bioinformatics Institute, Seoul National University, [3]BK21 FOUR Intelligence Computing, Seoul University, [4]Department of Computer Science and Engineering, Seoul National University, [5]Institute of Engineering Research, Seoul National University

Introduction

Advances in high throughput technologies such as the next generation sequencing technology allow researchers can measure high throughput transcriptomic molecular profile. Utilizing this valuable transcriptomic information is now at the core of research in biology and medicine. A routine practice is to compute and use differentially expressed genes (DEGs) for molecular biology research. This is also true even for clinical practice and it is important to select a small number of gene set with clinical relevance. For example, commonly used subtypes of breast cancer, PAM50 subtypes, are defined in terms of expression quantities of 50 genes. However, the use of the PAM50 subtype is not satisfactory enough to accurately predict the prognosis in each individual patient. Another example is to measure metastatic potentials of ER positive and node-negative breast cancers by combining expression quantities of 21 genes, OncoType DX, which is now a common practice for determining the need for chemotherapy around the world as a commercial product. However, there are still a lot of room for improvements for determining clinically relevant gene sets. For instance, some patients with LumA breast cancer subtype have higher potentials of metastasis while some patients with aggressive Basal subtype have lower metastasis potentials.

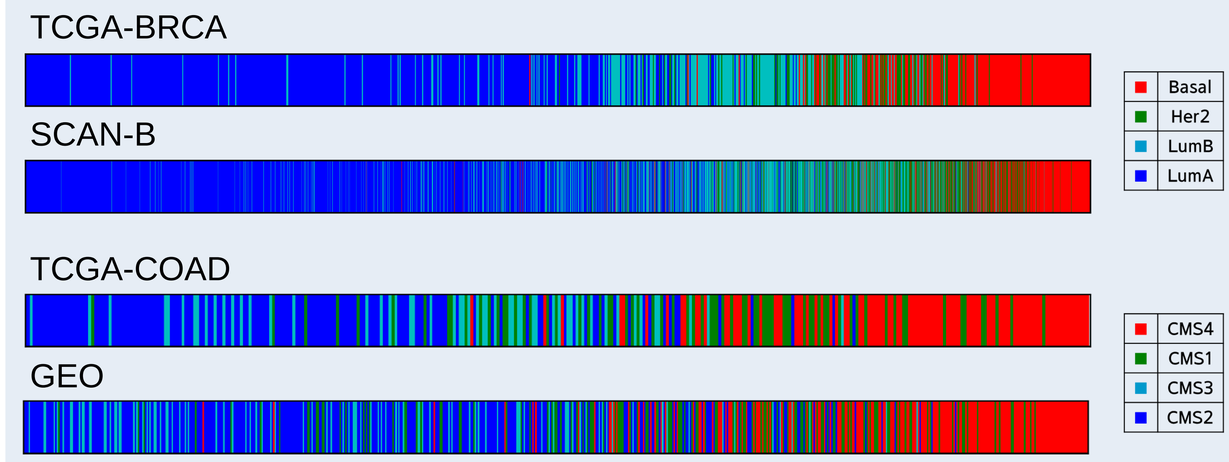
Data sets

Cancer type	Data set	Subtype				Total
		LumA	LumB	Her2	Basal	
Breast cancer	TCGA-BRCA	562 (53.9%)	207 (19.8%)	82 (7.9%)	192 (18.4%)	1,043
	SCAN-B	1,709 (53.7%)	767 (24.1%)	348 (10.9%)	360 (11.3%)	3,184

Cancer type	Data set	Subtype				Total
		CMS1	CMS2	CMS3	CMS4	
Colon cancer	TCGA-COAD	72 (19.7%)	142 (38.9%)	53 (14.5%)	98 (26.8%)	365
	GSE39582	89 (17.2%)	232 (45.0%)	69 (13.4%)	126 (24.4%)	516
	GSE17536	31 (19.9%)	65 (41.7%)	20 (12.8%)	40 (25.6%)	156
	GSE17537	9 (17.6%)	15 (29.4%)	11 (21.6%)	16 (31.4%)	51

Results

Our method is a strategy of selecting clinically consistent gene sets as ordering patients. We tested our methods extensively for breast cancer and colon cancer with well-defined subtypes using TCGA data. Then, identified gene sets were applied to additional data sets (SCAN-B and GEO data sets) to stratify patients in the order of clinical relevance in each of subtypes. Therefore, gene sets selected in our framework will be useful for molecular subtyping and prognosis prediction.



Workflow

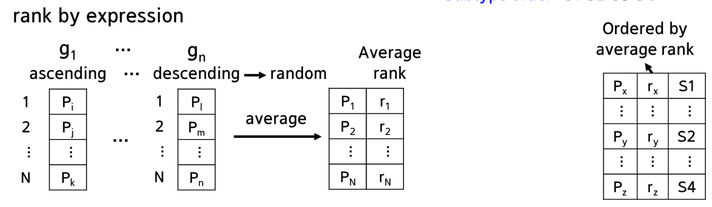
In this study, a computational framework using genetic algorithm with a novel fitness function and differentially activated subpathways is proposed for determining clinically significant gene sets for cancer subtypes. Specifically, differentially activated subpathways are computed among multiple subtypes by using MIDAS [1]. Then, in genetic algorithm, our method is performed by selecting genes in KEGG pathways and fitness function is composed of two terms which are calculated using order of samples according to ranking of gene expression value. One is kendall tau-b correlation coefficient for patient stratification in terms of pre-defined order of cancer subtypes, and the other is clinical related term which orders patients in terms of clinical outcome such as overall survival. Differentially activated subpathways were used in crossover and mutation steps. In crossover step, nodes in differentially activated subpathways are combined together in units for subpathways and the others cross uniformly. In addition, different mutation probabilities are set for nodes that are in differentially activated subpathways and nodes that are not in mutation step.

Calculate the fitness value

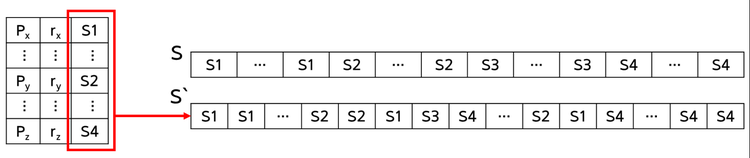
Fitness value
(stratification score) + λ x (survival score)

1. Sample (patient) ordering
P: patient, g: selected genes, N: # samples, S: subtype
r_k: average rank of P_k
rank by expression

Subtype order: S1 S2 S3 S4

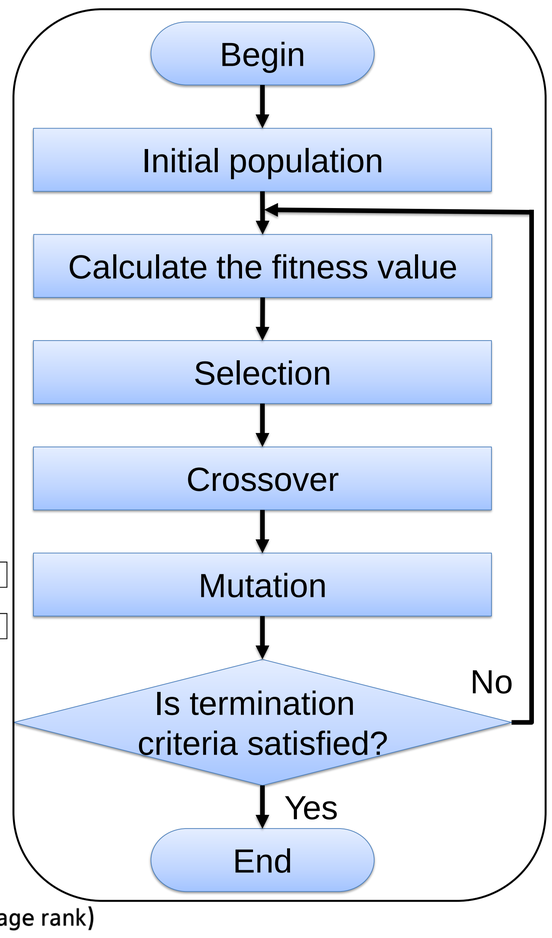


2. Calculate stratification score



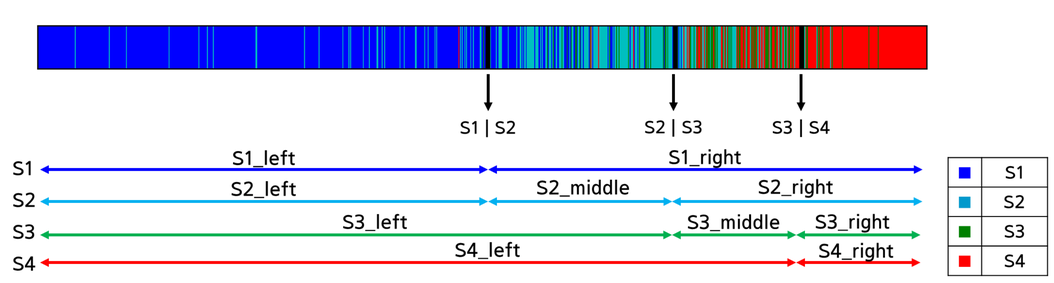
stratification score
= kendall tau correlation between S and S` for stratification

S: subtype vector when completely stratified
S`: subtype vector from ordered samples



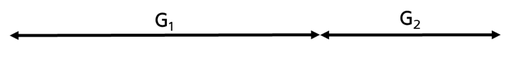
3. Calculate survival score

(1) Determine boundary between subtypes -> SVM (feature: average rank)



(2) Score

$Score(G_1, G_2) = c \times l$



Cox's proportional hazard regression

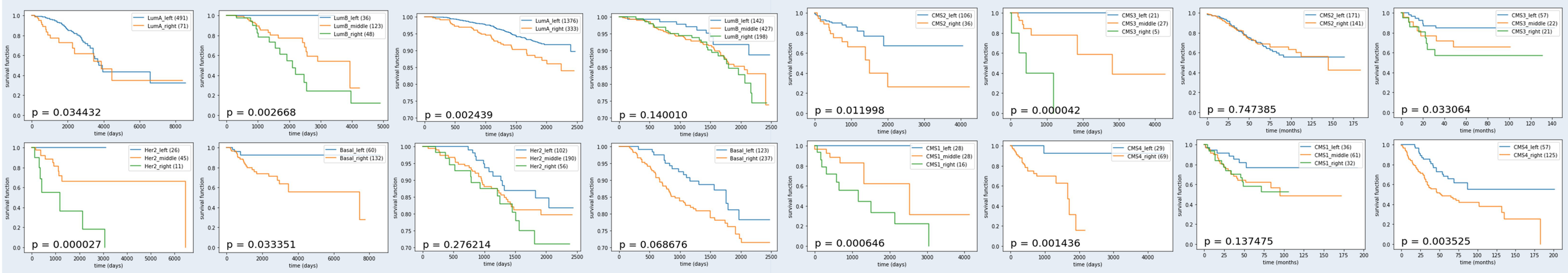
$$c = \begin{cases} 1 & if\ hazard\ is\ higher\ in\ G_2\ than\ G_1 \\ -1 & otherwise \end{cases}$$

Logrank test

u: upper bound

$$l = \begin{cases} -\log_{10}(p\ value) & if\ p\ value > u \\ -\log_{10} u & otherwise \end{cases}$$

survival score = $\text{average}(Score(S1_{left}, S1_{right}), Score(S2_{left}, S2_{middle}), Score(S2_{middle}, S2_{right}), Score(S3_{left}, S3_{middle}), Score(S3_{middle}, S3_{right}), Score(S4_{left}, S4_{right}))$



[1] Lee, Sangseon, Youngjune Park, and Sun Kim. "MIDAS: mining differentially activated subpaths of KEGG pathways from multi-class RNA-seq data." *Methods* 124 (2017): 13-24.