BMC Bioinformatics

BioMed Central

Oral presentation

Open Access

Functionally informative tag SNP selection using a pareto-optimal approach: playing the game of life

Phil Hyoun Lee, Jae-Yoon Jung and Hagit Shatkay

Address: School of Computing, Queen's University, Kingston, ON, K7L 3N6, Canada

from Fifth International Society for Computational Biology (ISCB) Student Council Symposium
Stockholm, Sweden 27 June 2009

Published: 19 October 2009

BMC Bioinformatics 2009, **10**(Suppl 13):O5 doi: 10.1186/1471-2105-10-S13-O5

This article is available from: http://www.biomedcentral.com/1471-2105/10/S13/O5

Introduction

Major interest in current epidemiology, medicine, and pharmarco-genomics is focused on identifying single nucleotide polymorphisms (SNPs) that underlie the etiology of common and complex diseases. However, due to the tremendous number of SNPs on the human genome, there is a clear need to prioritize SNPs to expedite genotyping and analysis overhead associated with disease-gene studies. Tag SNP selection and Functional SNP selection are the two main approaches for addressing the SNP selection problem. However, little was done so far to effectively combine these distinct and possibly competing approaches. Here we present a new multi-objective optimization framework for identifying SNPs that are both informative tagging and have functional significance.

Methods

Our SNP selection algorithm is based on the notion of Pareto optimality [1], which has been extensively used for addressing multi-objective optimization problems in game theory, economics and engineering. We describe the details of its three main steps as follows.

STEP 1. Computing Linkage Disequilibrium of SNPs

To efficiently compute the score of tagging informativeness, we calculate the pair-wise LD between all pairs of candidate SNPs in advance. As a measure of pair-wise LD, following Carlson et al. [2], we currently use the coefficient of determination, r^2.

Input: a set of SNPs, $V = \{ X_1, ..., X_p \}$;
a set of functional significance scores, $E = \{ e_1, ..., e_p \}$;
a haplotype dataset D;
the maximum number of SNPs to select, k.

Output: sets of Pareto optimal solutions, $\mathcal{PO} = \{ S_1, ... \}$.

Algorithm:

Compute $LD = \{ ld_{11}, ..., ld_{pp} \}$;

$\mathcal{PO} \leftarrow \emptyset$;
$m \leftarrow 0$;
While $(m < M)$
  $T \leftarrow T_0$;
  $S_c \leftarrow S_p$; a randomly selected k SNPs from V.
  Compute $f(S_c|D, E) = (f_1(S_c|D),\ f_2(S_c|E))$;
  While $(T > T_{min})$
    $S_n =$ neighbor (S_c).
    Compute $f(S_n|D, E) = (f_1(S_n|D),\ f_2(S_n|E))$.
    If $(\ \exists S_i \in \mathcal{PO},\ S_n \succ S_i\ )$
      remove $\forall S_i \in \mathcal{PO}$ s.t. $S_n \succ S_i$;
      $\mathcal{PO} \leftarrow \mathcal{PO} \cup \{S_n\}$.
    Else if $\forall S_i \in \mathcal{PO},\ S_i \not\succ S_n$
      $\mathcal{PO} \leftarrow \mathcal{PO} \cup \{S_n\}$.
    EndIf
    $P_{accept}(\ S_c,\ S_n,\ T\ )$
    $\leftarrow \min\left\{1, exp\left(\max_{j \in \{1,2\}} \frac{(f_j(S_n) - f_j(S_c))}{T}\right)\right\}$.
    If $(S_n \succ S_c$ **or** $P_{accept} > random\ )$
      $S_c \leftarrow S_n$.
    EndIf
    $T \leftarrow r_c \cdot T$.
  EndWhile
  $m \leftarrow m + 1$.
EndWhile

Figure 1
The Multi-Objective SA Algorithm.

Gene	Genomic	Total	SNPselector				TAMAL			
Symbol	Locus	SNP #	k	SA_{FI}	SA_0	RS	k	SA_{FI}	SA_0	RS
ADRB2	5q31-q32	153	41	100.0	100.0	33.3	17	66.6	100.0	50.0†
APEX1	14q11.2-q12	83	27	100.0	100.0	100.0	19	100.0	100.0	100.0
ATR	3q22-q24	181	36	100.0	100.0	100.0	20	100.0	100.0	50.0
CDKN1A	6p21.2	116	34	100.0	100.0	100.0	20	100.0	100.0	100.0
CYP1A1	15q22-q24	49	34	100.0	100.0	100.0	10	100.0	100.0	75.0
CYP1B1	2p21	172	51	100.0	100.0	100.0	28	100.0	100.0	100.0
NQO1	16q22.1	86	6	100.0	100.0	50.0	8	100.0	100.0	50.0
EPHX1	1q42.1	148	27	80.0	25.0	14.2†	23	25.0	·	·
ERCC2	19q13.3	210	27	100.0	100.0	100.0	30	50.0	50.0	·
ERCC4	16p13.3-p13.11	289	41	100.0	100.0	44.4	49	50.0	50.0	20.0†
ERCC5	13q22,13q33	261	43	88.8	25.0	·	43	11.1	·	·
GSTP1	11q13	70	27	100.0	100.0	100.0	14	100.0	100.0	50.0
LIG4	3q33-q34	107	27	100.0	100.0	100.0	27	20.0	100.0	25.0†
MBD1	18q21	65	24	100.0	100.0	100.0	19	100.0	100.0	50.0
MGMT	10q26	550	36	100.0	100.0	71.4	81	20.0	25.0†	·
MMP9	20q11.2-q13.1	111	33	100.0	100.0	100.0	16	100.0	100.0	33.3
MTHFR	1p36.3	206	42	100.0	100.0	100.0	24	50.0	100.0	50.0
MTR	1q43	372	27	75.0	100.0	100.0	33	14.2	33.3	33.3
MTRR	5p15.3-p15.2	212	31	100.0	100.0	100.0	32	33.3	50.0	75.0†
NBN	8q21	355	21	100.0	100.0	100.0	38	67.0	100.0	50.0†
POLB	8p11.2	143	25	100.0	100.0	100.0	18	100.0	100.0	100.0
RAD23B	9q31.2	197	12	100.0	100.0	100.0	29	20.0	16.6	·
SOD2	6q25.3	188	31	20.0	·	·	27	25.0	25.0†	20.0†
SULT1A1	16p12.1	180	39	100.0	100.0	100.0	6	33.3	100.0†	100.0†
TP53	17p13.1	307	46	50.0	100.0	100.0	11	50.0	33.3	66.6†
XPC	3p25	237	35	100.0	100.0	100.0	29	20.0	·	·
XRCC1	19q13.2	152	46	80.0	33.3	·	46	20.0	·	·
XRCC2	7q36.1	253	13	100.0	100.0	100.0	25	33.3	33.3	50.0†
XRCC3	14q32.3	158	11	100.0	100.0	100.0	37	50.0	100.0	33.3
EXO1	1q42-q43	283	35	33.3	100.0	·	36	20.0	100.0	66.6†
HDAC5	17q21	111	21	100.0	100.0	100.0	13	100.0	100.0	100.0
POLI	18q21.1	239	23	75.0	100.0	100.0	24	25.0	100.0	·
REV1	2q11.1-q11.2	307	53	100.0	100.0	100.0	32	50.0	100.0	50.0

Figure 2
Evaluation results of three Pareto optimal search algorithms against two compared systems.

STEP 2. Retrieving Functional Significance of SNPs

We currently use the FS score of SNPs obtained from F-SNP [3], which assesses the deleterious functional effects of SNPs, using 16 bioinformatics tools, with respect to protein translation, splicing regulation, transcriptional regulation, and post-translational modification.

STEP 3. Selecting Functionally Informative Tag SNPs

Our selection algorithm is based on multi-objective simulated-annealing (SA) search. We also introduce two heuristics for generating a new neighboring solution to guide efficient search while expediting convergence. Figure 1 summarizes the proposed algorithm.

Conclusion

We applied our system to 34 disease-susceptibility genes for lung cancer, which is one of the most extensively-studied cancer types due to its high mortality rate [4]. Our algorithm always finds a collection of Pareto optimal SNP subsets that performs better than the subsets selected by other SNP selection approaches, with respect to both tagging informativeness and

functional significance (shown in Figure 2). Moreover, we clearly show that our system improves upon general-purpose search algorithms for identifying Pareto optimal solutions (p-values are 1.37e-004, 3.11e-015, 2.43e-149 and 3.89e-179).

References

1. Kirman AP: **Pareto as an economist.** *The New Palgrave: A Dictionary of Economics* 1987, **5:**804–808.
2. Carlson CS, Eberle MA, Rieder MJ, Yi Q, Kruglyak L and Nickerson DA: **Selecting a maximally informative set of single-nucleotide polymorphisms for association analyses using linkage disequilibrium.** *Am J Hum Genet* 2004, **74 (1):**106–120.
3. Lee P and Shatkay H: **F-SNP: computationally predicted functional SNPs for disease association studies.** *Nucleic Acids Res* 2008, **(36 Database issue):**D820–D824.
4. Zhu Y, Hoffman A, Wu X, Zhang H, Zhang Y, Leaderer D and Zheng T: **Correlating observed odds ratios from lung cancer case-control studies to SNP functional scores predicted by bioinformatics tools.** *Mutation Research* 2008, **639:**80–88.