Optimal haplotype block free selection of tagging SNPs for genome-wide association studies

Supplemental material 1: Algorithm Complexity

We will now present notation to show how the k-MIS can be solved efficiently, when the size of each neighborhood is bounded by a constant w. Enumerate the n SNPs from 1 to n. For ease of exposition, suppose that all SNPs within distance $\lfloor \frac{w}{2} \rfloor$ of s are used to predict s. Define the corresponding *assignment* A_s as follows

$$A_s[i] = \begin{cases} 1 & \text{if SNP } s - \lfloor \frac{w}{2} \rfloor + i \in S' \\ 0 & \text{otherwise} \end{cases} \quad (1)$$

Correspondingly, define the subset of SNPs $S(A_s)$ to contain all SNPs s' such that $A_s[s' + \lfloor \frac{w}{2} \rfloor - s] = 1$.

Theorem 1 *The k-MIS problem can be solved $O(nk2^w)$ time, and $O(k2^w)$ space, if the size of all neighborhoods is bounded by a constant w.*

Proof: A solution to the k-MIS problem can be described by an $O(n)$ size bit-vector, such that $B[i] = 1$ if SNP i is selected, and 0 otherwise. At most k entries are 1 in any solution. The

solution also implies an assignment A_s for each SNP s as $A_s[i] = B[s - \lfloor \frac{w}{2} \rfloor + i]$.

Let A_s^0 and A_s^1 be the vectors obtained by removing the rightmost element of A_s and moving the other elements by one to the right and adding a 0 or 1 as the leftmost element. Note that in any solution $A_s[i] = B[s - \lfloor \frac{w}{2} \rfloor + i] = B[(s-1) - \lfloor \frac{w}{2} \rfloor + (i+1)] = A_{s-1}[i+1]$. Therefore, depending on whether the $A_{s-1}[0]$ is 0 or 1, $A_{s-1} = A_s^0$, or $A_{s-1} = A_s^1$.

Let $I_w(s, l, A_s)$ be the score of most informative subset of l SNPs chosen from SNPs 1 through s, such that A_s described the assignment for SNP s. The score obtained for informing SNP s is exactly $I(S(A_s), s)$, and $I_w(s, l, A_s)$ is given by $I(S(A_s), s)$ plus score of the best assignment for SNPs 1 through $s-1$ that is consistent with A_s.

By the argument above, there are only two possibilities for the assignment to SNP $s-1$, described by A_s^0, or A_{s-1}^1. Finally, the assignment to SNPs 1 through $s-1$ cannot use SNP $s + \lfloor \frac{w}{2} \rfloor$. Therefore, the number of SNPs available to SNPs 1 through $s-1$ are $l-1$ if $A_s[w] = 1$, and l otherwise. Thus

$$I_w(s, l, A_s) = I(S(A_s), s) +$$

$$\max(I_w(s-1, l - A_s[w], A_s^0), I_w(s-1, l - A_s[w], A_s^1)$$

Figure 2 describes the algorithm for computing this recurrence using dynamic programming.

The score of the optimal assignment for choosing k SNPs from the whole set can be retrieved as $\max_{A_n} I_w(s, k, A_n)$ and the optimal assignment can be retrieved via a backward traversal of the dynamic program. The space saving trick of Hirschberg (Hirschberg 1975) can be used to reduce the space requirements to $O(k2^w)$. □

We note that as most neighborhoods will be smaller than the maximum size w efficiency gains can be made in implementation.

References

Hirschberg, D. S. 1975. A linear space algorithm for computing maximal common subsequence.

Communications of the ACM **18**: 341-343.