Conservative Extensions of Linkage Disequilibrium Measures from Pairwise to Multi-Loci and Algorithms for Optimal Tagging SNP Selection*

Fumei Lam† and Ryan Tarpine‡ and Sorin Istrail§
Center for Computational Molecular Biology
Department of Computer Science
Brown University, Providence, RI 02912

February 7, 2011

We present results on two classes of problems. First, concerns the long standing open problem of finding unifying principles for the Linkage Disequilibrium (LD) measures in common use in population genetics (Lewontin 1964 [**?**], Hedrick 1987 [**?**], Delvin and Rich 1995 [**?**]). Two desirable properties were proposed in the extensive literature on the topic whose mutual consistency remained an open problem and has been at the heart of statistical and algorithmic difficulties with haplotype and GWAS analysis. The first axiom is: *(1.) The extendability of LD measures to multi-loci as a unique conservative extension of the pairwise values.* All the LD measures in common use are pairwise measures. One can talk about a "curse of the pair-wise" as despite significant attempts, it is not clear how to naturally extend them to multiple loci. Usually extensions from pair-wise to multi-loci need an ad hoc aggregation function of the set of pairs, which in turn introduces substantial bias. The second axiom is: *(2.) The Interpretability of Intermediate Values.* This property, satisfied by the LD measure r^2, is substantiated by the Pritchard-Przeworski r^2 sampling theorem [**?**]. In this paper we resolve this mutual consistency problem; we introduce a new LD measure, *Directed Informativeness* $\overrightarrow{\mathcal{I}}$ (the directed graph theoretic counterpart of the *Informativeness* measure introduced by Halldorson et al [**?**]) and show that it satisfies both of the axioms. We show also the "minimum informative subset of tagging SNPs" based on $\overrightarrow{\mathcal{I}}$ can be computed exactly in polynomial time for realistic genome-wide data. In addition, we show that many LD measures can be expressed in terms of Directed Informativeness: Lewontin's D', r^2, the Pearson's χ^2, Yule's Q, haplotype entropy ΔS, and Informativeness I. In particular r^2, the LD measure of choice for quantifying and comparing LD in the context of mapping is equal to the Directed Informativeness "squared" ($r^2 = \overrightarrow{\mathcal{I}}\overleftarrow{\mathcal{I}}$).

Second, we present polynomial time algorithms for optimal genome-wise tagging SNPs selection for a number of commonly used LD measures, under the bounded neighborhood assumption for linked pairs of SNPs. An open problem in the area is the search for a quality measure for tagging SNPs selection that unifies the LD-based methods such as LD-select (the "maximally informative" set of SNPs) Calrson et al 2004 [**?**] (implemented in Tagger, de Bakker et al 2005 [**?**]) and the information-theoretic ones such as Informativeness, directly related to the retrieval of haplotype diversity. We show that the explicit objective function of the optimization as defined by the LD-select algorithm is the *Minimal Dominating Set (MDS)* on r^2-SNP graphs and show that we can compute MDS in polynomial time

*Supported by National Science Foundation

†Present address: University of California Davis, Department of Computer Science, Davis, fumeilam@gmail.com

‡ryan@cs.brown.edu

§Corresponding author: Sorin Istrail sorin@cs.brown.edu

for this class of graphs. Although in the LD-select the "maximally informative" is obtained through a greedy algorithm, and therefore better referred to as "locally maximally informative," in fact, we show that Tagger (LD-select) performs very close to the "global maximally informative" optimum.

1 The Directed Informativeness Equations

The relationships of various LD measures to Directed Informativeness are presented in the following equations. Detailed definitions and proofs are presented in the paper. In the equations, N is the effective population size, s, t are two SNPs, and T^s is a contingency table.

$$r^2(s,t) = \overrightarrow{\mathcal{I}}(s,t)\overrightarrow{\mathcal{I}}(t,s).$$

$$\chi^2(s,t) = \frac{\overrightarrow{\mathcal{I}}(s,t)\overrightarrow{\mathcal{I}}(t,s)}{N} \qquad \Delta S = \frac{\overrightarrow{\mathcal{I}}(s,t)\overrightarrow{\mathcal{I}}(t,s)}{2}.$$

$$D'(s,t) = \begin{cases} \overrightarrow{\mathcal{I}}(s,t)\frac{|E(G_t)|}{D_{max}} & \text{if } D \geq 0 \\ \overrightarrow{\mathcal{I}}(s,t)\frac{|E(G_t)|}{D_{min}} & \text{if } D < 0, \end{cases}$$

$$Q(s,t) = \frac{\sum_{1\leq i<j\leq n}\overrightarrow{\delta_{ij}}(s,t)}{\sum_{1\leq i<j\leq n}|\overrightarrow{\delta_{ij}}(s,t)|} = \overrightarrow{\mathcal{I}}(s,t)\left(\frac{|E(G_t)|}{|E(G_s)\cap E(G_t)|}\right)$$

$$I(s,d) = \overrightarrow{\mathcal{I}}(s,d)\frac{permanent(T^s)}{determinant(T^s)}$$

2 Desiderata for Linkage Disequilibrium Measures

Linkage Disequilibrium (LD) is of fundamental importance for population genetics and evolutionary studies. In the past 10 years we have been witnessing a renaissance in interest in LD study, with a large literature focused on the study of empirical patterns of LD in the human genome; one of the major driving hopes is that these LD patterns and measurements hold the key to developing powerful computational tools for mapping of complex disease loci through genome-wide association studies. Surveys papers on LD include significant warnings about the LD measures extensively used in studies. Such warnings read like: *Gametic Disequilibrium Measures: Proceed With Caution ... it is not obvious whether different measures gave similar information, or whether different measures may give complementary information.*, Hedrick [**?**]; *... the choice of linkage disequilibrium measure can have a substantial impact on the accuracy and interpretability of the ... mapping method."* Delvin and Risch [**?**]; *"Although the measure D captures the intuitive concept of LD, its numerical value is of little use for measuring the strength of and comparing levels of LD. ...Comparing different reports on the extent of LD is complicated by the fact that several measures are in common use, and although all are based on Lewontin's D, they have very different properties and measure different things.* Ardlie, Kruglyak and Seielstad [**?**].

The authors of published criticism about measures of LD often accompanied their criticism with commentary on criteria or "axioms" that an ideal such measure should satisfy. Some of these criteria are presented as mathematical properties and some are presented only informally. Given a set of axioms, the question of whether there are LD measures satisfying all of them becomes a mathematical challenge; the solution of the challenge, if obtained, could lead to constructing a new measure, or be

an impossibility proof. Such an axiomatic approach is often employed in mathematics in the search for robust concepts of solution for various problems; e.g., results such as Arrow's impossibility theorem in voting theory [**?**] and impossibility results in clustering [**?**] demonstrate the importance of reasoning about such sets of axioms and understanding trade-offs. The aim of our approach is to develop an axiomatic framework by formalizing two such desiderata about LD measures proposed in the literature, and then to study the problem of finding LD measures mutually satisfying both of them.

An example of a list of desiderata is presented by Hedrick in [**?**]: *... There are a number of criteria that can be used to determine the most appropriate measure of gametic disequilibrium for a given situation. Several possible criteria are that should have (1) a simple biological interpretation, (2) statistical tests available or easily developed, (3) be directly related mathematically to evolutionary factors such as recombination, selection, genetic drift, gene flow, etc. and (4) be standardized to allow comparison across loci or populations. Obviously, all these criteria (and probably more) are important characteristics for a disequilibrium measure. "* Hedrick 1987 [**?**]. Two classic papers containing most comprehensive comparison of LD measures are Hedrick 1987 [**?**], and Delvin and Risch [**?**].

2.1 Axiom 1: The extendability of LD measures to multi-loci as a unique conservative extension of the pairwise values.

All the extensively used LD measures are defined pairwise. When, however, one would want to measure LD for a genomic region with three of more loci, one has only pairwise measures to work with, and then a certain aggregation function is required to aggregate the pairwise values in the region, all or some of them selected somehow. The choice of the aggregation function is necessarily *adhoc* (there are many options on general principles) and the effect of this adhoc choice is hard to evaluate. For example, one could use a weighted sum of all the pairwise values, the sum of pairs, the min-max of all pairwise values, a graph theoretic representation of them and looking at a graph concept e.g., all connected nodes within a threshold [**?**].
Our reformulation of this axiom is the following: an LD measure defined pairwise, that generalizes to many sites, as a conservative unique extension of the pairwise values. To our knowledge, none of the LD measures in use have such an unique conservative extension.

We call this major unresolved difficulty, the *curse of pairwise* (of the continuous mathematics approach). The name is inspired by our discussions with Andy Clark during our collaboration on the Minimum Informative Subset of SNPs Problem, when we solved the problem posed by him [**?**]. The problem was the following. One would want to genotype the minimum set of informative SNPs to be typed for a region, such that the information about all the remaining SNPs could be inferred from the subset of the SNPs typed. Formulating the problem with pairwise LD measures, involved considering arbitrary subsets of SNPs and expending the analytical expression even with all the three terms is already hopelessly complex, let alone the arbitrary subset sizes. So the problem was outside the reach of extending analytically/continuous mathematics terms to cover all the subsets. It turns out that a new information theoretic measure was needed, called *Informativeness*, that quantifies how much information about a target SNP t exists in a set of SNPs S. The optimization problem asked for finding the minimum subset of SNPs $S' \subset S$ that contains the same amount of information as in the entire set of SNPs S. It turns out that this problem reduced to a well-studied problem in computer science, although notorious for its computational complexity as well, the Set Cover Problem. Based on that insight, exact algorithms were obtained in [**?**] that work in polynomial time for genome-wide data, and compute the globally minimum informative subset of SNPs. Discrete mathematics/combinatorics/computer science came to the rescue for the curse of the pairwise of the analytical/continuous mathematics approach. We present in this paper a contribution of the same type.

This paper introduces a sister measure to Informativeness, called *Directed Informativeness*. The approach is based on graph theory and algorithmic results. It turns out that finding the Minimum Informative Subset under Directed Informativeness is NP-complete as well, but again we have a polynomial time and practical algorithm (a generalization of the one designed for Informativeness) that works for genome-wide data. However, the new measure has unexpected properties unifying a number of aspects of the extensively used LD measures, and as a consequence, it is consistent with the two axioms we propose. In particular, like Informativeness, Directed Informativeness can be uniquely extended to multi-loci as a conservative-to-all-pairwise-values extension.

2.2 Axiom 2: The Interpretability of Intermediary Values

- **Criticism of D'**:
  "For values of $D' < 1$ the relative magnitude of values of D' has no clear interpretation. Moreover, estimates of D' are inflated in small samples, even for SNPs with common alleles, but especially for SNPs wirh rare alleles. So, high values can be obtained even when markers are in fact in linkage equilibrium. Because the magnitude of D' depends strongly on sample size, samples are difficult to compare. ... intermediate values of D' should not be used for comarison of the strength of LD between studies, or to measure the extent of LD." Ardlie, Kruglyak, Seielstad[**?**]

- **Intermediate values of r^2 are easily interpretable**.
  "Consider two loci: one locus functionally associated with with disease and the other is a nearby marker in LD with the susceptibility locus. To have the same power to detect the association between the disease and the marker locus, the sample size must be increased by roughly $\frac{1}{r^2}$ when compared with the sample size for detecting association with the susceptibility locus itself." [**?**].

  Pritchard and Przeworski show the relationship between r^2, Pearson correlation coefficient χ^2 and effective population size N [**?**]. As a corollary, the directed informativeness between two SNPs can also be related to the χ^2 Pearson correlation coefficient and effective population size.

  So "interpretability" of intermediate values is defined as analytical guidance regarding the sample size via the Pearson correlation coefficient χ^2 and effective population size N. r^2 is the flagship measure in this respect.

3 Directed Informativeness and the Minimum Informative Subset Tagging SNPs Problem

A *Single Nucleotide Polymorphism* (SNP) is a position in the genome at which two or more different alleles occur in the population, each with frequency above a certain threshold.

The goal of association studies is to correlate genetic variation with the occurence of disease.The difficulty arises in the large number of candidate sites and the combinatorial explosion in the number of subsets of SNPs when multiple sites are considered. In chromosome-wide studies, whole genome scans are performed and it is desirable for cost efficiency reasons to select only a subset of SNPs which accurately represent the genetic variation of the entire population [**?**, **?**, **?**].

We introduce a measure for association which extends the *minimum informative subset*, a concept from data compression that has been studied in tagging and in the analysis of haplotype block robustness [**?**, **?**, **?**].

3.1 Informativeness

We first introduce the graph theoretic measure of informativeness introduced in [**?**], which aims to capture how well a set of SNPs can predict a target SNP (our notation largely follows that of [**?**]). The central idea is to define a measure of informativeness to quantify the accuracy of predicting an untyped SNP by a set of proximal SNPs. The input is an $n \times m$ matrix M representing n binary (0/1) haplotypes of length m, together with an $n \times 1$ disease vector t. The vector t takes values 0 and 1 to distinguish between 'case' and 'control' individuals. Consider the $n \times (m+1)$ matrix M', obtained from the matrix M by appending column t. The value in the ith row and jth column of M' is denoted by $M'_{i,j}$. For a column s of M' and haplotypes i, j, let $D^s_{i,j}$ denote the event that SNP s differs in positions i and j ($M_{i,s} \neq M_{j,s}$).

Definition 1 *The* informativeness *of SNP s with respect to the disease vector t is*

$$I(s,t) = Prob_{i \neq j}(D^s_{i,j} \mid D^t_{i,j}).$$

We can also interpret the informativeness of a SNP with respect to the disease vector in terms of graph theory. Associate with each column s of M' a complete bipartite graph G_s on n vertices, with bipartition defined by the alleles of s. Each vertex in G_s corresponds to a row of M', and there are edges between any two rows which differ at column s, i.e.,

$$E(G_s) = \{\{i,j\} \mid M_{i,s} \neq M_{j,s}\}.$$

Let $V(G_{s,0})$ denote the vertices in graph G_s corresponding to allele 0 and $V(G_{s,1})$ denote the vertices in G_s corresponding to allele 1. For $1 \leq i < j \leq n$, let

$$\delta_{ij}(s,t) = \begin{cases} 1 & \text{if } M_{i,s} \neq M_{j,s} \text{ and } M_{i,t} \neq M_{j,t} \\ 0 & \text{otherwise.} \end{cases}$$

Then the informativeness of s with respect to t can be expressed in terms of the bipartite graphs G_s and G_t as

$$I(s,t) = \frac{\sum_{1 \leq i < j \leq n} \delta_{ij}(s,t)}{|E(G_t)|}.$$

In [**?**], this measure was used to detect how well a target SNP can be predicted from a set of tagging SNPs and to solve the *k most informative SNPs problem* by observing that the minimum informative subset problem is equivalent to the minimum set cover problem. The advantage of this measure is that it can easily be generalized to define informativeness for subsets of SNPs (multiple sites) [**?**], therefore satisfying Axiom (1) of the desired properties for a linkage disequilibrium measure.

For $S' \subseteq S$, let

$$I(S',t) = Prob_{i \neq j}\left(D^{S'}_{i,j} | D^t_{i,j}\right) = \frac{|E(S') \cap E(t)|}{|E(t)|}$$

3.2 Directed Informativeness

Using the graph-theoretic interpretation of informativeness as a starting point, we modify the graph under consideration to define a measure we call *directed informativeness* and relate this measure to existing measures of linkage disequilibrium. For site s, denote the major allele by 0 and the minor allele

by 1. We create a *directed* bipartite graph $\overrightarrow{G_s}$ for site s, with vertex set $\{1, 2, \ldots n\}$ and directed edge set

$$E(\overrightarrow{G_s}) = \{(i, j) \mid M_{i,s} = 0 \text{ and } M_{j,s} = 1\}.$$

In this directed bipartite graph, all the edges are between different alleles and are oriented from the major allele to the minor allele. This addition of edge orientations in the graph will allow us to make a connection between the graph theoretic interpretation of informativeness with existing linkage disequilibrium measures. Note that by considering the underlying undirected graph of $\overrightarrow{G_s}$ for each site s, we obtain the undirected graph G_s defined in [**?**]. However, the pattern of intersection between the directed edges will play an important role in our extended definition of directed informativeness.

For $1 \leq i < j \leq n$, let

$$\overrightarrow{\delta_{ij}}(s,t) = \begin{cases} 1 & \text{if } M_i^s = M_i^t = 0 \text{ and } M_j^s = M_j^t = 1 \\ -1 & \text{if } M_i^s = M_j^t = 0 \text{ and } M_j^s = M_i^t = 1 \\ 0 & \text{otherwise.} \end{cases}$$

Definition 2 *The* directed informativeness *of SNP s with respect to SNP t is defined as*

$$\overrightarrow{\mathcal{I}}(s,t) = \frac{\sum_{1 \leq i < j \leq n} \overrightarrow{\delta_{ij}}(s,t)}{|E(G_t)|}$$

4 Directed Informativeness and the Conservative Extension to Multi-Loci Axiom

One problem with the LD measures r^2 and D' is that they are not adequate for SNP subset selection/tagging SNPs and do not extend to multiple SNPs in a canonical way. In contrast, we demonstrate how the directed informativeness measure can be extended uniquely to multiple sites while satisfying the desired properties.

Let $S = \{s_1, s_2, \ldots s_k\}$ and $T = \{t_1, t_2, \ldots t_l\}$ be disjoint subsets of loci. For each site $s_i \in S$ and $t_j \in T$, consider the associated directed graphs G_{s_i} and G_{t_j}. Each of the graphs G_{s_i} and G_{t_j} are defined on vertex set $\{1, 2, \ldots n\}$. Let G_T be the graph with vertex set $\{1, 2, \ldots n\}$ and with edge set $E(G_{t_1}) \cup E(G_{t_2}) \cup \ldots \cup E(G_{t_l})$. Note that G_T is a graph (not a multigraph), and any edge (i, j) appearing in two or more graphs $G_{t_j} (1 < i < l)$ appears only once in G_T. Let $E(G_T)$ denote the edge set of graph G_T. Then the directed informativeness of SNP subset S with respect to SNP t is

$$\overrightarrow{\mathcal{I}}(S,t) = \frac{\sum_{e \in E(G_s)} \overrightarrow{\delta}_e(S,t)}{|E(G_t)|}$$

5 Directed Informativeness and the Interpretability of Intermediary Values Axiom

We now relate the directed informativeness measure $\mathcal{I}$ with the widely used linkage disequilibrium measure r^2. This direct relationship provides insight into the observation in [**?**] that "optimizing for either [informativeness or r^2] optimizes very well for the other. These results thus suggest that informativeness and haplotype r^2, although distinct measures, are closely related in practice."

Note that while r^2 is a symmetric measure with respect to the pair of sites considered, the directed informativeness measure is not (i.e., $\overrightarrow{\mathcal{I}}(s,t) \neq \overrightarrow{\mathcal{I}}(t,s)$). The following theorem shows that the two measures are related by a natural product symmetrization.

Theorem 5.1 *For any two SNPs s and t, linkage disequilibrium measure r^2 between s and t is equal to*

$$r^2(s,t) = \overrightarrow{\mathcal{I}}(s,t)\overrightarrow{\mathcal{I}}(t,s).$$

Proof. Recall that $r^2(s,t) = \frac{(p_{00}p_{11}-p_{01}p_{10})^2}{p_{0+}p_{+0}p_{1+}p_{+1}}$. For each $i,j \in \{0,1\}$, let $C_{ij} = p_{ij}n$, where n is the size of the sample. We will show that

(1) $C_{00}C_{11} - C_{01}C_{10} = \sum_{1\leq i<j\leq n} \overrightarrow{\delta_{ij}}(s,t)$

(2) $C_{0+}C_{1+}C_{+0}C_{+1} = |E(G_s)||E(G_t)|$

To prove (1), observe that each haplotype occurence of 00 and 11 in columns i and j contribute $+1$ to both the left and right hand sides of the equation, while each haplotype occurence of 01 and 10 contribute -1 to both the left and right hand sides of the equation. Furthermore, all remaining haplotype pairs contribute 0 to both equations.

To prove (2), note that

$$C_{0+}C_{1+} = |E(G_s)|$$

and

$$C_{+0}C_{+1} = |E(G_t)|.$$

This proves the theorem. □

Pritchard and Przeworski show the relationship between LD measure r^2, Pearson correlation coefficient χ^2 and effective population size N [**?**]. As a corollary, the directed informativeness between two SNPs can also be related to the χ^2 Pearson correlation coefficient and effective population size.

Corollary 5.2 *For any two SNPs s and t,*

$$\chi^2(s,t) = \frac{\overrightarrow{\mathcal{I}}(s,t)\overrightarrow{\mathcal{I}}(t,s)}{N}.$$

6 Directed Informativeness, Informativeness and Contingency Tables

Tests for association are often based on differences in allele frequencies. One approach is based on contingency tables of allele frequencies compared with disease phenotypes. The contingency table T^s corresponding to SNP s is an $r \times 2$ matrix defined by

$$T^s_{ij} = |\{k \,|\, x(k) = i, y(k) = j\}|.$$

The following lemma describes the relationship between the (undirected) informativeness measure and the values in the contingency table.

Lemma 6.1 *let T^s be the contingency table corresponding to SNP s as defined above. The informativeness of SNP s with relation to disease vector d is equal to*

$$I(s,d) = \frac{T^s_{00}T^s_{11} + T^s_{10}T^s_{01}}{(T^s_{00} + T^s_{10})(T^s_{01} + T^s_{11})} = \frac{permanent(T^s)}{(T^s_{00} + T^s_{10})(T^s_{01} + T^s_{11})}.$$

In contrast, the directed informativeness measure can be expressed in terms of the contingency table using the *determinant* in place of the permanent.

Lemma 6.2 *The directed informativeness of SNP s with relation to disease vector d is equal to*

$$\overrightarrow{\mathcal{I}}(s,d) = \frac{T^s_{00}T^s_{11} - T^s_{10}T^s_{01}}{(T^s_{00} + T^s_{10})(T^s_{01} + T^s_{11})} = \frac{determinant(T^s)}{(T^s_{00} + T^s_{10})(T^s_{01} + T^s_{11})}.$$

7 Formalizing LD-Select/Tagger

A SNP Threshold Graph (STG) is constructed as follows: Given a set S of n SNPs ($|S| = n$), construct a set V of vertices where each vertex corresponds to a single SNP in S ($|V| = n$). There exists an edge between two vertices u and v if and only if there is linkage disequilibrium (LD) above a certain threshold τ between the SNPs represented by u and v.

LD-Select[**?**] and Tagger (without multimarker tests)[**?**] are essentially greedy heuristics which attempt to find the minimum dominating set for this graph. A dominating set for a graph $G = (V, E)$ is a subset $V' \subseteq V$ such that every vertex not in V' is connected to at least one member of V' by an edge. The minimum dominating set problem is to find the smallest such set. In the context of SNP selection, finding the smallest possible set V' means finding the smallest set of tagging SNPs such that every SNP is either in the set or is in LD above τ with at least one SNP in the set.

In general graphs, the dominating set problem is NP-complete. However, SNP Threshold Graphs have certain properties which allow them to be solved optimally in polynomial time.

7.1 LD Measure assumptions

In order to construct the graph, we make the following assumptions about the LD measure under consideration:

1. The measure is commutative: $\forall s.\forall t.\, LD(s,\, t) = LD(t,\, s)$. Pairwise measures usually fulfill this criterion; e.g., r^2.
2. The LD between a SNP and itself is always above the threshold τ (i.e., $\forall s.\, LD(s,\, s) \geq \tau$)
3. Long-range LD (LD between SNPs hundreds of kilobases apart) is not meaningful and can be ignored. This induces a "neighborhood" around each SNP, judging SNPs to beyond the edges of the neighborhood to be ignored because they are too far away. To simplify things, a fixed neighborhood size w is often chosen, such as 13 or 21[**?**].

7.2 SNP Selection Graph Properties

These assumptions yield a graph with the following properties:

1. The graph is undirected (since the LD measure is commutative).
2. There is a linear ordering on the vertices. Each vertex uniquely represents one SNP, so the ordering on the SNPs gives an ordering on the vertices.

3. Because we are not interested in long-range LD, there are restrictions on which vertices may be connected by edges. This is more meaningful than just a limit to the degree of nodes; based on the vertex ordering, we know that if the neighborhood is centered on SNP i and w is odd, then vertex i can only have edges to vertices $i - (w-1)/2$ through $i + (w-1)/2$.

This final property is key in finding the minimum dominating set efficiently. LD-Select and Tagger perform well in practice because the greedy approximation algorithm they use is known find an $O(\log d)$-approximation for graphs of maximum degree d, and the degree is bounded by both the window size and the actual extent of LD between SNPs.

The final property allows us to apply the dynamic programming (DP) algorithm given by Halldorsson et al. [**?**] with a novel measure of informativeness to compute the minimum dominating set. We call our method MIS-DS because it utilizes the minimum informative subset (MIS) algorithm to solve the dominating set problem.

7.3 Minimum Dominating Set Algorithm

The DP algorithm of Halldorsson et al. requires an upper bound k on the number of tagging SNPs in order to run (since it computes a dynamic programming matrix, it must know the dimensions of the matrix in advance). An obvious upper bound would be n, the total number of SNPs, but this is unnecessarily wasteful. Instead, we first run a greedy heuristic such as LD-Select [**?**] to establish a tighter upper bound. We then use the size of the resulting set as k. We run the DP algorithm with the following informativeness measure I:

$$I_\tau(S', t) = \begin{cases} 1 & \exists s \in S'.\, LD(s, t) \geq \tau \\ 0 & \text{otherwise} \end{cases}$$

where $LD(s, t)$ computes r^2 between SNPs s and t.

I_τ is a binary measure of whether SNP t is in LD above τ with at least one of the SNPs in the set S'. The DP algorithm finds in time $O(nk2^w)$ the set S' of size k which maximizes $\iota_k = \sum_t I_\tau(S', t)$, which in this case counts the number of SNPs which are either in the set S' or have LD above τ with a SNP in S'. As long as k is equal to or greater than the size of the minimum dominating set, this value will be n, the total number of SNPs. This allows us to find the size of the minimum dominating set regardless of how well the greedy heuristic performed. The DP matrix contains implicitly the maximum informativess ι' possible for all upper bounds $0 \leq k' \leq k$. Therefore, we can check whether $k-1$ would have gotten the same result ι, and if so, whether $k-2$ would have done the same, etc. Finding the maximum informativeness takes time $O(2^w)$ for each k' (because we must examine the DP matrix for each assignment A_s, of which there are 2^w), so in total time $O(k2^w)$ we can find the smallest $k^\star$ such that $\iota_{k^\star} = n$. Then backtracking in the DP matrix will yield the set $S^\star$ such that $|S^\star| = k^\star$ and $S^\star$ is a dominating set, i.e., the smallest possible set of tagging SNPs.

The MIS-DS algorithm is a globally optimal solver for the dominating set problem that works on arbitrarily large genome-wide data and runs in polynomial time. It is not restricted to r^2; it can be used with any pairwise measure of LD that fulfils the criteria above, such as D'[**?**], Q[**?**], and so on.

7.4 Comparison of LD-Select/Tagger vs Optimal

First 100 SNPs with MAF ≥ 0.05 of chromosome 22 from HapMap3 release 2 phasing data. The number of tagging SNPs required to capture all of the SNPs within the r^2 threshold:

r^2 threshold	0.9	0.8	0.7	0.6	0.5	0.4	0.3	0.2	0.1
Tagger pairwise	64	55	49	45	39	34	27	18	13
MIS-DS $w = 17$	67	61	54	50	45	39	32	24	18

First 50 SNPs:

r^2 threshold	0.9	0.8	0.7	0.6	0.5	0.4	0.3	0.2	0.1
Tagger pairwise	40	34	31	29	26	22	17	13	9
MIS-DS $w = 17$	40	35	31	30	27	24	19	14	10

First 40 SNPs:

r^2 threshold	0.9	0.8	0.7	0.6	0.5	0.4	0.3	0.2	0.1
Tagger pairwise	33	28	26	25	23	20	16	14	9
MIS-DS $w = 17$	35	33	29	28	26	24	20	15	12

Considering that the MIS-DS algorithm is optimal within the window size, we immediately assumed that Tagger's smaller result set was due to a lack of this constraint. Examining some of Tagger's results confirmed this to be true: in more than one case, a tagging SNP was capturing another SNP with 11 SNPs in between. This would require a centered window size of 25 (12 to the left, 12 to the right, and the center position), which is not feasible with our current implementation. In terms of physical distance, the SNPs are about 20 kb apart, so the LD is certainly significant. Future research will involve expanding the algorithm to handle larger window sizes.

7.5 Tagger "best N"

It is well understood that by choosing tagging SNPs using an r^2 threshold τ the same power is approximately achieved by increasing the sample size by a factor of $1/\tau$. In this sense, the experimenter understands the tradeoff between using a low threshold versus a high threshold–a low threshold means fewer tagging SNPs to genotype but requires a larger sample size. This is only true when every SNP of interest is in LD above the threshold with a tagging SNP. When using Tagger's "best N" method (the "Max tags" feature of Haploview), the experimenter is told exactly how many SNPs were captured within the threshold, but nothing is known about the SNPs that were not captured. It is possible that some of them are in LD just below the threshold with the chosen tagging SNPs. It is also possible that they are not in significant LD with any of them; moreover, it is possible that if the experimenter had set the threshold slightly lower, then the "best N" tagging SNPs would have been very different. This is because lowering the threshold adds edges to the SNP Threshold Graph, which can considerably increase the degrees of certain nodes, and this changes the minimum dominating set–the set could be much smaller and contain different vertices.

A better algorithm would not require a binary measure of whether LD between two SNPs is above a certain level or not–it would take into account the fractional LD value as is. One possibility is using the MIS DP algorithm with the informativeness measure:

$$I_{\max}(S', t) = \max_{s \in S'} LD(s, t)$$

$I_{\max}$ computes the maximum LD between SNP t and any SNP in set S'. Maximizing $\sum_t I_{\max}(S', t)$ would then attempt to make sure every SNP is in some level LD with tagging SNPs, rather than focusing entirely on LD above an arbitrary threshold.

8 Comparison of Directed Informativeness with (Undirected) Informativeness

8.1 Prediction vs Distinguishability

The measure of informativeness was defined in order to capture distinguishability between haplotypes. Given a sample of haplotypes, find the smallest set of SNPs such that these SNPs are capable of distinguishing between the original haplotypes–if two haplotypes were different when all the known SNPs were genotyped, then these haplotypes will still be different when only considering the tagging SNPs. For a single SNP s, G_s can be used to infer the allele of s in each haplotype because it is a complete bipartite graph. It is clear which haplotypes contain one allele and which contain the other (although one cannot be sure exactly which contain the allele 0 and which contain 1, this much detail is not necessary to compute association with a disease). However, once the measure is extended to compute the informativeness of a set of SNPs S', the union of the edges $E(S') = \bigcup_{s \in S'} E(s)$ is no longer sufficient to infer the alleles of the SNP t which is "predicted" by S', because the graph is no longer bipartite.

Directed informativeness, on the other hand, captures the ability to impute non-genotyped SNPs directly from the tagging SNPs. The directed edges of $\overrightarrow{G}_s$ make clear exactly which haplotypes contain 0 and which contain 1. For a set of SNPs S', the directedness of the edges in $E(S')$ makes it possible to determine for each vertex whether its allele is 0 or 1 simply by comparing its outdegree and indegree: if the outdegree is greater (or equal), then the allele is 0. This is because each edge represents a SNP $s \in S'$ for which the allele was 0 at this vertex and 1 at its endpoint. If the outdegree is greater, then there is more evidence that the allele is 0 than 1. If the indegree and outdegree are equal, then the predicted allele is still 0 since 0 is the major allele, by definition more likely if no other information is known.

8.2 Non-monotonicity vs Monotonicity

Because informativeness is calculated by counting edges in an undirected graph, adding more tagging SNPs can never decrease informativeness. That is, $\forall s.\, I(S',\, t) \leq I(S' \cup \{s\},\, t)$. This is not the case with directed informativeness. If the edges contributed by a tagging SNP have the opposite orientation with respect to edges in a predicted SNP's graph, then those edges will decrease the directed informativeness, not increase it. This means that we cannot assume every tagging SNP in the neighborhood of a non-genotyped SNP t should be used to predict it, because if any of those tagging SNPs have a different allele pattern than t then they could potentially decrease the directed informativeness and hinder imputing the SNP. Otherwise even tagging every SNP would fail to produce total directed informativeness because each SNP would interfere with its neighbors.

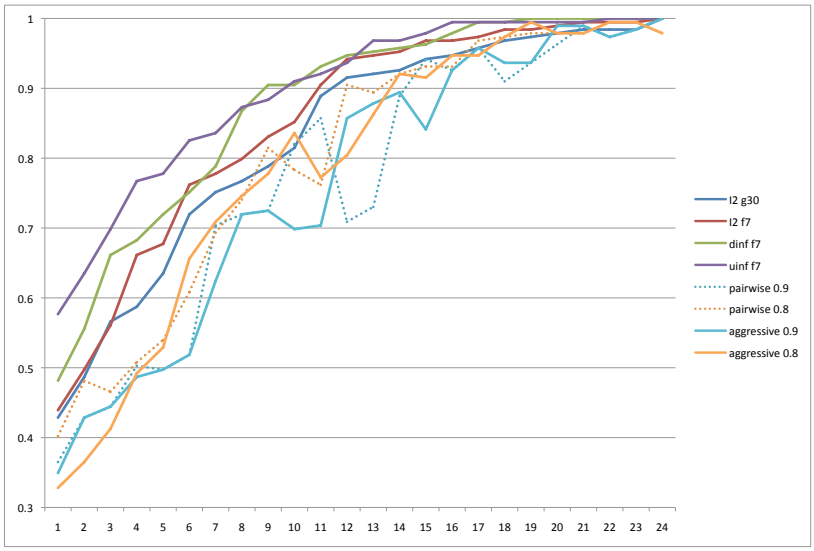
8.2.1 Ramifications to Minimum Informative Subset problem

Non-monotonicity means that we must change the definition of the minimum informative subset problem in order to make it relevant to directed informativeness. We define it as:

$$\overrightarrow{\mathcal{I}}\,(S',\, T) = \sum_{t \in T} \max_{S'_t \subseteq S' \cap N(t)} \overrightarrow{\mathcal{I}}\,(S'_t,\, t)$$

Rather than using all the tagging SNPs in the neighborhood of t to predict t, we use the subset of the tagging SNPs in the neighborhood which maximizes directed informativeness. We modified the algorithm of Halldorsson et al. to use this metric to compute the minimum directed informative subset.

Figure 1: Comparison of haplotype discrimination according to the tagging SNPs of different measures



9 Empirical Results

For all of the following tests we used as our dataset the 189 unique haplotypes generated by taking the first 40 SNPs with a minor allele frequency of at least 0.05 of chromosome 22 of the CEU population in HapMap3 release 3.

9.1 Haplotype Discrimination

One measure-agnostic method for testing the level of information captured by a set of tagging SNPs is haplotype discrimination. Given a set of m unique haplotypes, are the tagging SNPs sufficient to distinguish between all m, or do they collapse into a smaller set? The following table shows what fraction of those haplotypes are still unique after viewing only the tagging SNPs chosen by different algorithms:

We chose 33 to be the upper bound for the number of tagging SNPs because this was the number of SNPs selected by Tagger when run without a limit with the threshold $r^2 \geq 0.9$. Lower thresholds and aggressive tagging both require fewer tags.

The measure $\mathcal{I}^2$ is defined by $\mathcal{I}^2(S, t) = \overrightarrow{\mathcal{I}}(S, t)\overrightarrow{\mathcal{I}}(t, S)$. This is the measure created by taking the pairwise identity with r^2 and extending it to multiple loci. Tagging SNPs for $\mathcal{I}^2$ are chosen by using the minimum directed informative subset algorithm with $\mathcal{I}^2$ as the measure. We discuss $\mathcal{I}^2$ below.

Figure 2: A comparison of the ability to impute non-genotyped alleles by looking into a dataset where the alleles are known using the tagging SNPs chosen by different measures. The X axis is the number of tagging SNPs chosen. The Y axis shows the fraction of the alleles of the original dataset corrected imputed.

9.2 Haplotype Imputation

Another measure of how representative tagging SNPs are is whether by looking at only the tagging SNPs can the untyped SNPs be inferred. There are two methods for imputation: looking up the missing SNPs in a reference panel where these SNPs are known, and computing the missing alleles using only the genotyped SNPs themselves. We tested both types of imputation, using the 207 unique haplotypes generated by taking the first 100 SNPs in the method described above. Since the Tagger algorithm involves some degree of randomness, for each test we ran Tagger 10 times and used the results of the run which yielded the highest mean r^2 as reported by Tagger itself.

9.2.1 Using reference panel

For the first type of imputation we tested each measure in a method similar to Halldorsson et al.: an untyped SNP of an individual is inferred by looking at the typed SNPs in its neighborhood. If there are haplotypes in the training set that had the same allele as the individual on all the tagging SNPs, then a majority vote is taken to estimate which allele the individual has. If the votes are split 50/50, or if there are no training haplotypes which share every allele, then votes are taken from the haplotypes which differ by at most one allele. If even these votes are split or if no training haplotypes match, then the process continues by counting votes from haplotypes which differ on up to two alleles (and so on).

Given the tagging SNPs assigned by each measure, we tested the percentage of SNPs correctly imputed by this method. The results can be seen in figure 2.

(Undirected) informativeness is the best measure for imputation using a reference panel because of its ability to distinguish haplotypes. This property is exactly what is needed in order to look up the correct haplotype to observe the missing SNPs. The SNPs which are tagged maximizing informativeness are like a key into the reference panel: they are sufficient to find the exact haplotype, if present, because they distinguish between all of the different known haplotypes.

9.2.2 Tagging SNPs only

To impute non-genotyped SNPs using only the tagging SNPs, for $\mathcal{I}^2$ we used the method detailed above in section 8.1. For tagging SNPs chosen by Tagger, there are two cases. If a SNP s is captured by a single tagging SNP t, then we assume for each individual that his allele for s is the same as the allele genotyped for t. If s is captured by a haplotype, then we compare the individual's haplotype (i.e., the alleles of the test SNPs) to the test value given by Tagger. If they are equal, then the predicted SNP is said to have the allele 1, otherwise 0. The results of this analysis can be seen in figure 3.

It is evident that no measure is best for all numbers of tagging SNPs. $\mathcal{I}^2$ does the best for few tagging SNPs, but aggressive tagging with Tagger with an r^2 threshold of 0.7 results in more accurate imputation for a large range (28-42) of tagging SNPs. However, 39 SNPs is sufficient to capture all of the SNPs according that threshold, so allowing for more SNPs beyond that point does not result in any improvement–any increase or decrease beyond 39 SNPs is entirely due to randomness in the Tagger algorithm. More tagging SNPs allow higher r^2 thresholds to perform better, but each performs optimally only within a certain range. A zoomed-in view of the performance with high numbers of tagging SNPs can be seen in figure 4. $\mathcal{I}^2$ performs consistently, and near-optimally, across all ranges.

Why $\mathcal{I}^2$.

Figure 3: A comparison of the ability to impute non-genotyped alleles using only the tagging SNPs chosen by different measures.

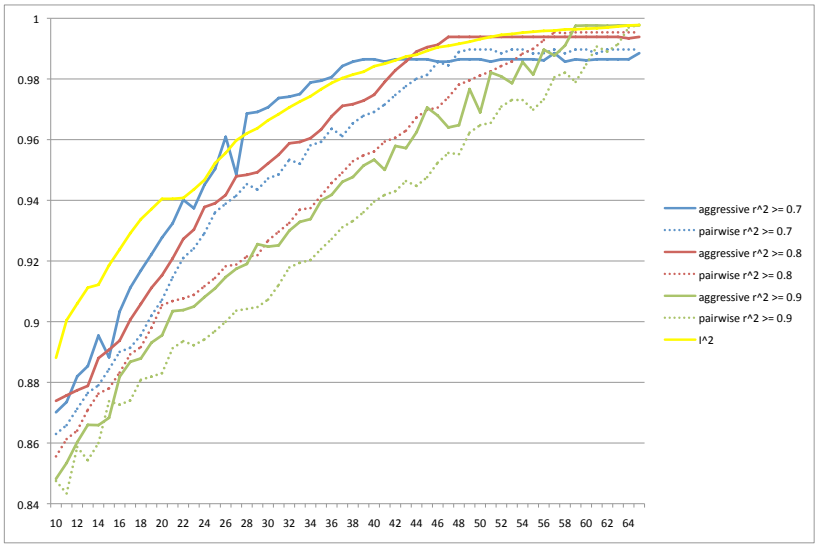
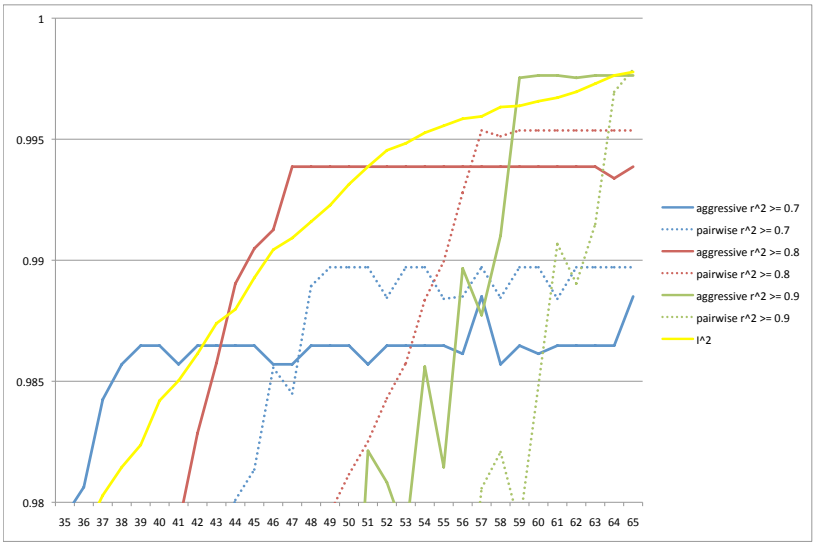


Figure 4: SNP imputation with large numbers of tagging SNPs.



10 Directed Informativeness and Other LD Measures

Not all LD measures can be expressed in terms of Directed Informativeness; e.g., the ideal measure for simple disequilibrium mapping δ, the difference in proportions d, [**?**] are not subject to such a decomposition.

Two other measures of LD can be expressed as well in terms of Directed Informativeness.

10.1 Yule's Q

Similarly, we have the following theorem for linkage disequilibrium measure $Q = \frac{p_{00}p_{11}-p_{01}p_{10}}{p_{00}p_{11}+p_{01}p_{10}}$.

Theorem 10.1 *For any two SNPs s and t, linkage disequilibrium measure Q between s and t is equal to*

$$Q(s,t) = \frac{\sum_{1\le i<j\le n} \overrightarrow{\delta_{ij}}(s,t)}{\sum_{1\le i<j\le n} |\overrightarrow{\delta_{ij}}(s,t)|} = \overrightarrow{\mathcal{I}}(s,t)\left(\frac{|E(G_t)|}{|E(G_s)\cap E(G_t)|}\right)$$

10.2 Haplotype Entropy

Nothnagel, Furst and Rohde define a measure ΔS based on the information theoretic concept of *entropy* [**?**]. At each locus, a sequence can be one of 2^m possible haplotypes and the frequency of haplotype i is given by p_i. If n haplotypes are present, then the entropy is given by

$$S_n = -\sum_{i=1}^{n} p_i \log p_i.$$

If the sites are in linkage disequilibrium, the probabilities p_i can be expressed in terms of marginal allele frequencies at the loci. For $q_i = q_{a_1^i, a_2^i, \ldots a_m^i}$, we have

$$q_{a_1^i, a_2^i, \ldots a_m^i} = \prod_{k=1}^{m} p_k^{1_{(a_k^i=0)}} (1-p_k)^{1_{(a_k^i=1)}}$$

and the corresponding entropy measure is

$$S_E = -\sum_{i=1}^{2^m} q_i \log q_i.$$

Then the Entropy Difference is given by $\Delta S = S_E - S_B$.

This measure can also be related to LD measure r^2 [**?**]. This implies the following relationship between directed informativeness and haplotype entropy.

Corollary 10.2 *For any two SNPs s and t,*

$$\Delta S = \frac{\overrightarrow{\mathcal{I}}(s,t)\overrightarrow{\mathcal{I}}(t,s)}{2}.$$

11 Computational Complexity of the Minimum Directed Informative Set of SNPs/Tagging SNPs Problem

In this section, we establish the complexity of the Minimum Directed Informative SNPs problem.

Lemma 11.1 *The Minimum Directed Informative SNPs problem is NP-complete.*

Proof: The proof follows the proof for the complexity of the Minimum (undirected) Informative SNPs problem, with a reduction from the set cover problem. Given a collection C of subsets of a finite set X, and positive integer $k \leq |C|$, the set cover problem asks if there exist $C' \subseteq C$ with $|C'| \leq k$ such that every element of X belongs to at least one member of C'. We construct a SNP matrix M with $|X|+1$ haplotypes and $|C|+1$ SNPs. For each subset $C_j \in C$, define a SNP $M[*, j]$ such that

$$M[i,j] = \begin{cases} 0 \text{ if } i \leq |X| \text{ and } X_i \in C_j \\ 1 \text{ otherwise} \end{cases}$$

The SNP $t = M[*, |C|+1]$ is dened by the vector $[0, 0, \ldots, 0, 1]$ with exactly $|X|$ zeros and a single one. Then $C' \subseteq C$ covers X if and only if the corresponding subset of SNPs S' are informative with respect to t.

□

A polynomial time algorithm – that is practical for genome-wide data sets – for Directed Informativeness is obtained by generalizing the algorithm used for the Informativeness measure (and available from the authors) [**?**].