

# On the monotonicity of HIPP

Lu Gram

July 25, 2007

## 1 Overview

This project investigated three different monotonicity properties of the *HIPP problem* (Haplotype Inference By Pure Parsimony). Given any one of these properties, novel algorithms for solving HIPP could be developed. Unfortunately, counterexamples were found for all three.

## 2 Definitions

We define a *haplotype* to be a string from  $\{0, 1\}^*$  and a *genotype* to be a string from  $\{0, 1, 2\}^*$ . For two haplotypes  $h_1, h_2$  of equal length, we define the *product* of  $h_1$  and  $h_2$  to be the genotype  $g$  where

$$g[i] = \begin{cases} h_1[i] & \text{if } h_1[i] = h_2[i] \\ 2 & \text{o/w} \end{cases}$$

Let  $H$  be a set of haplotypes. The *span* of  $H$ ,  $\langle H \rangle$ , is defined to be the set of all genotypes that are the product of some pair of haplotypes in  $H$ . If  $G$  is a set of genotypes, we say  $H$  *spans* or *explains*  $G$ , if  $G \subseteq \langle H \rangle$ . The *HIPP problem* can be stated as the following:

*Given a set of genotypes  $G$ , find a set of haplotypes of least size  $H$  spanning  $G$*

We now define a *projection*  $P$  to be a mapping of the form  $(s_1, \dots, s_n) \mapsto (s_{i_1}, \dots, s_{i_k})$  where  $(i_1, \dots, i_k)$  is a strictly increasing, non-empty subsequence of  $(1, \dots, n)$ . For a set of sequences of length  $n$ ,  $S$ , we define  $S_P$  to be  $\{P(s) : s \in S\}$ . In this case, we say  $S$  is the *extension* of  $S_P$ . Taking this one step further, if  $U$  is a set of sets of sequences then define  $U_P$  to be  $\{H_P : H \in U\}$ . In the following, we assume that all sequences in a given set have the same length.

## 3 Monotonicity property 1

One reasonably first ask: If  $H$  is a solution to  $G$ , is  $H_P$  necessarily a solution to  $G_P$ ? Phrased otherwise: If  $U$  is the set of all solutions to  $G$  and  $V$  the set of

all solutions to  $G_P$ , is  $U_P \subseteq V$ ? If this were true then the algorithm provided under the section “Monotonicity property 3” would solve HIPP. The following provides a counterexample to this question:

000	is minimal for	000
010		010
101		101
110		220

since 000, 010, 101 are clearly required to produce 000, 010, 101, but 110 is additionally needed to produce 220. But if we remove the last column of our genotypes, we get the genotypes, 00, 01, 10, 11 which can be explained by 00, 01, 10 alone - hence the projection of our solution  $\{00, 01, 10, 11\}$  is not itself a solution.

## 4 Monotonicity property 2

One can then ask: If  $H$  is a solution to  $G_P$  where  $P$  is a projection and  $G$  is a set of genotypes, does there then exist  $H'$  such that  $H = H'_P$  and  $H'$  is a solution to  $G$ ? Phrased otherwise: If  $U$  is the set of all solutions to  $G$  and  $V$  the set of all solutions to  $G_P$ , does  $U_P \supseteq V$ ?

If this were true, we could solve the HIPP problem in the following way: Start with the first column in  $G$ ,  $G_0$  and find the smallest set  $H_0$  that spans  $G_0$ . Then repeatedly extend our current set of genotypes  $G_i$  with an additional column to obtain  $G_{i+1}$  and find the smallest extension of  $H_i$ ,  $H_{i+1}$  that spans  $G_{i+1}$ . At each step, our monotonicity would guarantee that there exists an extension to  $H_i$  that solves  $G_{i+1}$  and since  $H_{i+1}$  has the smallest cardinality among all extensions that span  $G_{i+1}$ ,  $H_{i+1}$  would be a solution. When we arrive at  $G$ , we will have solved HIPP.

A counterexample to the proposed property is given by  $G = \{102201, 122221, 222002, 221000\}$  and  $P$  being the projection that removes the last character. By simply considering all possible ways of spanning  $G_P$ , one can show that there are only two possible spanning sets of minimum size (5),  $H_1 = \{10000, 10110, 01100, 11001, 10100\}$  and  $H_2 = \{11011, 01100, 10100, 10010, 01000\}$ . The following two diagrams show how they explain  $G_P$  (vertices are haplotypes, two vertices are linked iff their product lies in  $G_P$  and the edges are labeled with their resp. products):

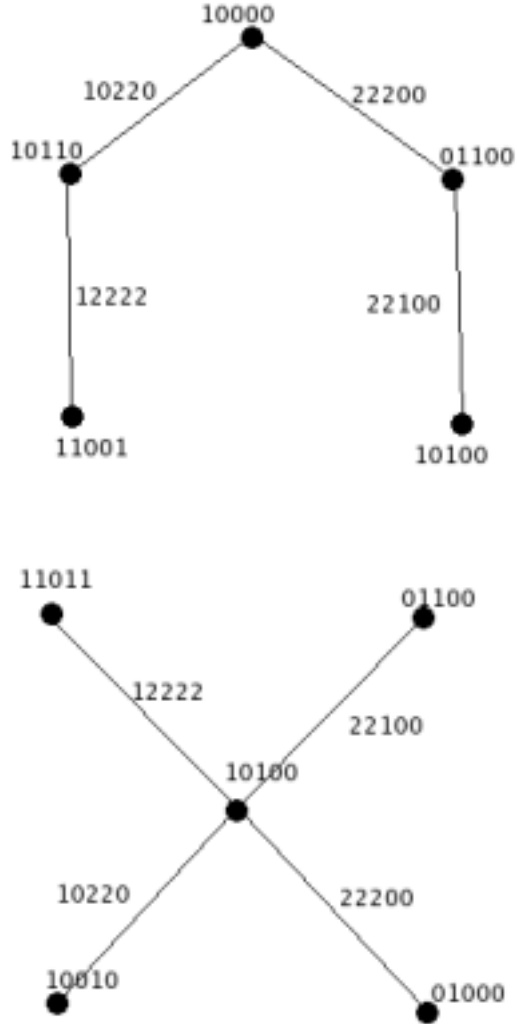


Fig. 1. Two minimal explanations for  $G_P$ ,  $H_1$  and  $H_2$ .

Now, we can span  $G$  by extending every haplotype in  $H_1$  by just one extra character, so  $G$  can be spanned by a set of size 5 (namely  $\{100001, 101101, 110011, 011000, 101000\}$ ). Since we know the minimal spanning set for  $G_P$  has size 5, we conclude that 5 is the size of the smallest explanation for  $G$  as well.

$G$  is obtained from  $G_P$  by extending every sequence in  $G_P$  with one character. Hence, we can explain  $G$  using an extension to any solution to  $G_P$ . In our case, we extend 12222 with a 1 and 22100 with a 0. These two extensions can be handled using  $H_1$  without introducing additional vertices because the explanations for 12222 and 22100 involve four distinct haplotypes. However, in

$H_2$  10100 is used to explain both 12222 and 22100 and so any smallest extension to  $H_2$  explaining  $G$  must extend 10100 with both 0 and 1. It is easily seen that every other haplotype must be extended as well, so any extension to  $H_2$  has at least 6 haplotypes, more than the global minimum, 5. But  $H_2$  solves  $G_P$  and no extension to  $H_2$  solves  $G$ . So the second monotonicity property does not hold either.

## 5 Monotonicity property 3

We now know that if  $U$  contains all the solutions to  $G$  and  $V$  all solutions to  $G_P$  then  $U_P$  is not necessarily either subset or superset of  $V$ . The remaining question to ask is: Is  $U_P \cap V$  necessarily non-empty? In other words, does there always exist *some* solution to  $G_P$  that has an extension solving  $G$ ?

If this was the case, then we could modify our proposed algorithm above to provide a correct solution thus: Pick the first column of  $G$ ,  $G_0$ . Instead of finding just one solution, find all solutions to  $G_0$ ,  $S_0$  say. Now extend  $G_0$  with one column to obtain  $G_1$ . Now find all smallest extensions of solutions in  $S_0$ ,  $S'_0$  say. By the above property, some solution in  $S_0$  can be extended to a solution to  $G_1$ , so let  $S_1$  be the non-empty subset of  $S'_0$  containing solutions to  $G_1$ . Now we can extend  $G_1$  as well and repeat our procedure until we reach  $G$  in which case we have found a set of solutions to  $G$ .

However, the last monotonicity property does not hold either. Consider  $G_P = \{12222, 22100, 10220, 22200, 21022, 22220\}$ ,  $G = \{122221, 221000, 102200, 222001, 210220, 222201\}$  and  $P$  removes the last character as usual.  $H = \{11011, 10100, 01100, 10010, 01000\}$  solves  $G_P$ , but any smallest extension of  $H$  spanning  $G$  must extend *all* haplotypes in  $H$  with both 0 and 1, so any extension of  $H$  contains at least 10 haplotypes.

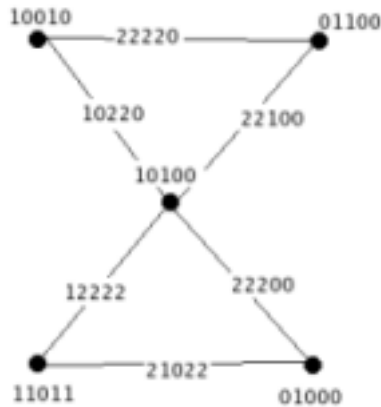
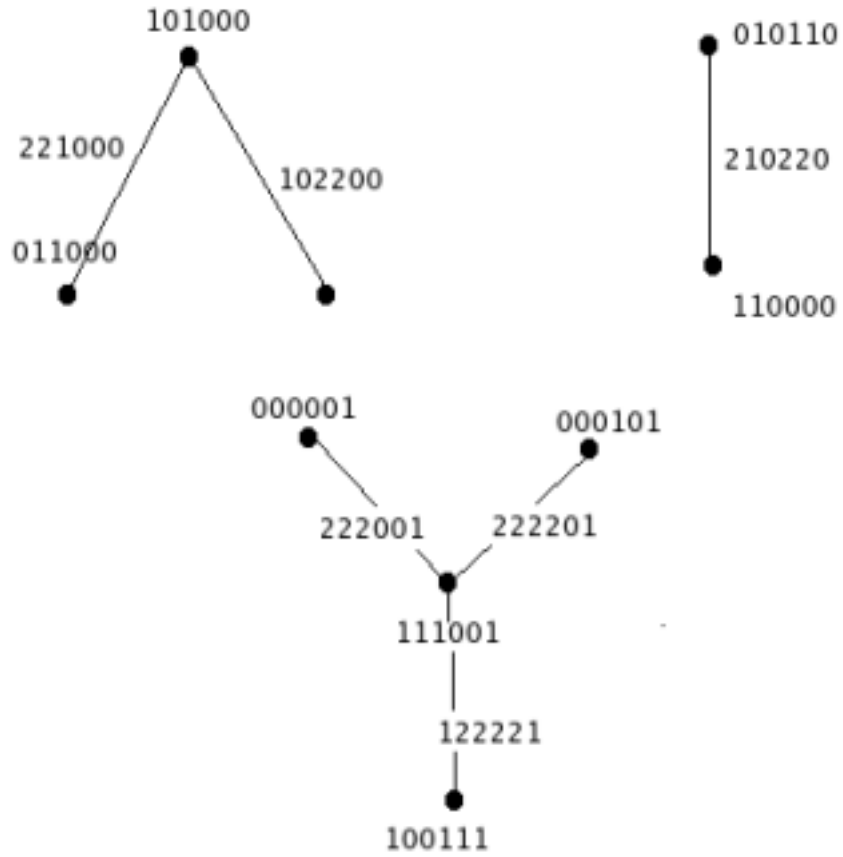


Fig.2. Only solution to  $G_P$ .

However,  $G$  can be explained using just nine haplotypes thus:



*Fig.3. Explanation for  $G$  using just 9 haplotypes.*

So no extensions of the only minimal spanning set for  $G_P$  are minimal. Hence, the HIPP problem is not monotone in any of the senses defined above.