33rd Annual International Conference of the IEEE EMBS
Boston, Massachusetts USA, August 30 - September 3, 2011

# Discrete Interference Modeling via Boolean Algebra

Gerhard Beckhoff, Department of Computer Science
University of Western Ontario, London, Canada

***Abstract*—Two types of boolean functions are considered, the locus function of $n$ variables, and the interval function of $\nu = n - 1$ variables. A 1-1 mapping is given that takes elements (cells) of the interval function to antidual pairs of elements in the locus function, and vice versa. A set of $\nu$ binary codewords representing the intervals are defined and used to generate the codewords of all genomic regions. Next a diallelic three-point system is reviewed in the light of boolean functions, which leads to redefining complete interference by a logic function. Together with the upper bound of noninterference already defined by a boolean function, it confines the region of interference. Extensions of these two functions to any finite number of $\nu$ are straightforward, but have been also made in terms of variables taken from the inclusion-exclusion principle (expressing "at least" and "exactly equal to" a decimal integer). Two coefficients of coincidence for systems with more than three loci are defined and discussed, one using the average of several individual coefficients and the other taking as coefficient a real number between zero and one. Finally, by way of a malfunction of the mod-2 addition, it is shown that a four-point system may produce two different functions, one of which exhibiting loss of a class of odd recombinants.**

## I. Introduction

The well-known differential equation method introduced by Haldane and improved on by Kosambi, Carter-Falconer, Felsenstein and others, for modeling interference, suffers from the stringent limitation that all calculations rest on the three loci consideration, and that the resulting map function in general cannot be extended to more loci [Karlin and Liberman, (1994)]. An attempt to extend Haldane's approach to to more than three loci will be discussed.

A different approach exposed in a recent textbook [Griffiths et al, (2005)] defines the coefficient of coincidence as the ratio of observed and expected number (or frequency) of the single class of double crossovers of a three-point system. The same can be done for a four-point system, which has three classes of double crossovers. The resulting three coefficients of coincidence are then averaged.

Christiansen [Christiansen, (2000)] in his book on population genetics consistently uses set theory to analyze diallelic systems. This is the closest to employing boolean algebra, since the power set of any finite set forms a boolean algebra. [Mano, (1984)] is an easy readable reference to boolean functions and their map representation.

## II. Boolean Functions

Informally, a *boolean function* of $n$ variables is a mapping $f : B^n \to B$, where $n$ is a positive integer, $B = \{0, 1\}$, and the domain of f, $B^n$, is a set of $2^n$ $n$-bit vectors, called *minterm functions* or just *minterms*, and are denoted $m_i$, where the subscript $i \in \{0, 1, \dots, 2^n - 1\}$ is the decimal equivalent of $m_i$ interpreted as a binary number. The subset of $n$ minterms, having just one zero in the n-bit vector, is the set of *atoms* and can generate the entire set $B^n$, using the operation $\vee$, called *disjunction* or *sum*. Its dual relation $\wedge$ is called *conjunction* or *product*. Its symbol is usually omitted if no confusion can arise. A third, unary operation $^-$ (bar) called *complementation* or *negation* changes a bit to its other value.

There are two notations used to describe boolean functions, the bit notation (a bit is either a 1 or a 0), or the literal notation (a literal is either a variable $a_i$ or $\bar{a}_i$). The arithmetic transform of a boolean function uses the following identities: $\bar{a} = 1 - a$, $a \wedge a = a$, $a \vee b = a + b - ab$.

The *weight* of $m_i$, denoted $|m_i|$, is the number of 1s in the bit sequence, and the *distance* of two minterms, $m_i$ and $m_j$, is the number of positions in which the two minterms differ: $d(m_i, m_j) = |(i)_2 \oplus (j)_2|$, where $(i)_2$ is the binary equivalent of the decimal integer $i$. An odd [even] weighted minterm is called odd [even] minterm. Two graphical representations exist: the $p$-cube *lattice*, which represents minterms by points and connects any two points which are unit distance apart by a line, and the $p$-cube *map*, which represents the minterms by cells. where neighboring cells are unit distance apart. The latter one, also called *Karnaugh map*, is the workhorse in digital design.

## III. Interval Functions

Biology is not an exact science. It is based on probability theory and uses statistical methods. Roughly speaking, minterms are now considered cells indexed in bit notation but containing probabilistic values (frequencies) of acquired (observed) data. The variables in the literal notation are random variables and and their cell denotation may be interpreted as expected values of the cell contents.

The boolean variables $a_i$ of an $n$-variable boolean function can be identified as genes of a diallelic system having $n$ loci, and therefore the function will be called *locus function*. When considering probabilities, these are random variables $Pr(a_i) = R(a_i)$ with the properties (1) $\prod_{i=0}^{n-1} R(a_i) = R(\prod_{i=0}^{n-1} a_i) = R(m_k)$ for some $k$ determined by the literals in the product, and (2) $R(m_i) = R(m_j)$ for $i + j = 2^n - 1$; i.e., the pair $(m_i, m_j)$ is antidual (one component is the bit-by-bit complement of the other).

Recombination, however, occurs between loci and can be detected only at the two loci marking the region. Therefore, interval functions will be introduced having as atoms intervals (an interval is the smallest region marked by neighboring loci, $A_i = [i, i - 1], i = 1, 2, \dots, \nu = n - 1$, and its probability is $Pr(A_i) = r_i$.

978-1-4244-4122-8/11/$26.00 ©2011 IEEE

The following identity is used to transform (compress) the locus function of dimension $n$ into the dimension of the interval function

$$\begin{aligned} A_i &= a_i \oplus a_{i-1} \\ &= a_i \bar{a}_{i-1} \vee \bar{a}_i a_{i-1} \quad i = 1, 2, \ldots, \nu \end{aligned} \tag{1}$$

By mod2-addition ($\oplus$) of the basic codewords representing intervals it is possible to find the codewords of the entire genomic region, whether as a single segment (one or several connected intervals), or as a multiple segment region.

Let $M_k$ be an interval minterm. To find the corresponding locus minterms $m$, replace each $A_i$ by expression (Eq.1) and multiply out. For the case $\nu = 2$, let $k = 2$. Then

$$\begin{aligned} M_2 &= A_2 \bar{A}_1 = (a_2 \oplus a_1)(\overline{a_1 \oplus a_0}) \\ &= (a_2 \bar{a}_1 \vee \bar{a}_2 a_1)(a_1 a_0 \vee \bar{a}_1 \bar{a}_0) \\ &= a_2 \bar{a}_1 \bar{a}_0 \vee \bar{a}_2 a_1 a_0 = (m_4, m_3) \end{aligned}$$

Note that one interval minterm always maps into two allele minterms.

Conversely, two allele minterms always map into one interval minterm. To find the corresponding interval minterm of $m_k$, add (mod-2) the bit-notation (binary number) of $m_k$ to its by one bit shifted version (it does not matter in which direction the number is shifted). The right-most and left-most bits, which do not have an addend, will be discarded. The result is a $\nu$-bit interval minterm, since each $A_i$ satisfies Eq.1. For the last example, $m_4$, and $m_3$,

$$\begin{array}{cccc||cccc||cccc} a_2 & a_1 & a_0 & & 1 & 0 & 0 & & 0 & 1 & 1 & \\ & a_2 & a_1 & a_0 & & 1 & 0 & 0 & & 0 & 1 & 1 \\ \cline{2-3} \cline{6-7} \cline{10-11} & A_2 & A_1 & & & 1 & 0 & & & 1 & 0 & \end{array}$$

An equivalent code in bit notation is the set of all even weighted n-bit vectors, where the ones mark the end points of the corresponding region. The interval codewords can generate the entire code of $2^\nu - 1$ codewords (not including the empty region), by using the $\oplus$ operation. For n=3, these codewords are $A_3 = (1100)$, $A_2 = (0110)$, $A_1 = (0011)$, and the composite regions are $A_{32} = A_3 \oplus A_2 = (1010)$, $A_{31} = A_3 \oplus A_1 = (1111)$, $A_{21} = A_2 \oplus A_1 = (0101)$, $A_{321} = A_3 \oplus A_2 \oplus A_1 = (1001)$.

## IV. Three-Point System

The extreme values of interference for a three-point system ($\nu = 2$) are commonly stated as $r^o = r_1 + r_2 - 2r_1 r_2$ for no interference NI, and $r^c = r_1 + r_2$ for complete interference CI (the superscripts o and c stand for odd and complete, respectively). While the first value can be rewritten as a Boolean function, $r^o = r_1 \oplus r_2$, the second one can not. In the 2-cube map, $r^o$ occupies the two cells $1_d = 01$ and $2_d = 10$ (d stands for decimal). Including the only free cell $3_d = 11$ (the zero cell $0_d = 00$ is always excluded, since they contain nonrecombinants), yields another logic function $r^c = r_2 \vee r_1$ and on the logic level this function together with the function $r^o$ define the boundaries of intermediate interference functions.

As an example, taken from [Wu, Ma, and Casella (2007)], consider a backcross $a_2 a_1 a_0 / \bar{a}_2 \bar{a}_1 \bar{a}_0 \times \bar{a}_2 \bar{a}_1 \bar{a}_0 \times \bar{a}_2 \bar{a}_1 \bar{a}_0$. The total of eight groups of genotypes produced are shown in the locus 3-cube map and in compressed form in the interval 2-cube map, using the technique outlined earlier. The gene order is assumed to be 2, 1, 0 and the two intervals $A_2 = [2, 1]$, $A_1 = [1, 0]$ and the composite of the two intervals, $A_{21} = [2, 0]$ have respective recombination frequencies $r_2$, $r_1$, and $r_{21}$.

$$\begin{array}{c|c|c|c|c|} & \multicolumn{4}{c}{a_1} \\ \cline{2-5} & .31 & .10 & .11 & .01 \\ \cline{2-5} a_2 & .05 & .02 & .38 & .02 \\ \cline{2-5} & \multicolumn{4}{c}{a_0} \end{array} \Rightarrow \begin{array}{c|c|c|} & \multicolumn{2}{c}{r_1} \\ \cline{2-3} & .69 & .12 \\ \cline{2-3} r_2 & .69 & .03 \\ \cline{2-3} \end{array} \tag{2}$$

The entries of the allele 3-cube map are the ratios $n_{ij}/N$, where $n_{ij}$ is the number of genotypes containing $i$ recombinants in $A_2$ and $j$ recombinants in $A_1$, and $N$ is the sample size, here $N = 100$, after deleting the observations missing in the two regions. The entries of the interval matrix are the antidual pairs.

Note that the class (0,7) of nonrecombinants maps into interval cell 0 and the recombinant classes into 01, 10, and 11; this is the lower region of the three decimal numbers defining the cells of the locus function. Another choice would be would be to map class (0,7) into 7 and the recombinant classes into decimal 6, 5, 4. Erasing the leftmost bit of the corresponding binary numbers will put the cells in reverse order. Although inconsequential, this mapping would be closer to Mather's formula, which states that recombination fractions $r$ lie in the range $0 \leq r \leq 0.5$. As a consequence of dealing with antidual classes, the functions and polynomials related to the interval map are all symmetric.

The recombination frequencies are related to the 'minterms' $g_{ij}$ via the matrix equation $\mathbf{r} = M\mathbf{g}$ and $\mathbf{g} = M^{-1}\mathbf{r}$, where

$$\begin{aligned} \mathbf{r}^t &= \left|\, 1 \quad r_1 \quad r_2 \quad r_{21}) \,\right| \\ \mathbf{g}^t &= \left|\, g_{00} \quad g_{01} \quad g_{10} \quad g_{11} \,\right| \\ M &= \begin{vmatrix} 1 & 1 & 1 & 1 \\ & 1 & 1 & \\ & & 1 & 1 \\ & 1 & 1 & \end{vmatrix} \quad M^{-1} = \begin{vmatrix} 2 & -1 & -1 & -1 \\ & 1 & -1 & 1 \\ & -1 & 1 & 1 \\ & 1 & 1 & -1 \end{vmatrix} \end{aligned}$$

The square matrix $M$ is not orthogonal, but can be made so by replacing the recombination frequencies $r_k$ by $\lambda_k = 1 + 2r_k$, where $k = 2, 1, 21$, and then can be written as a Kronecker product of a $2 \times 2$ matrix (see [Schnell,(1961)]).

## V. Extensions

The extension of the two logic function and their complements to $\nu$ loci yields:

$$r^o_\nu = r_1 \oplus r_2 \oplus \cdots \oplus r_\nu = \bigoplus_{i=1}^{\nu} r_i \tag{3}$$

$$\bar{r}^o_\nu = 1 \oplus r_1 \oplus \cdots \oplus r_\nu = \bigoplus_{i=0}^{\nu} r_i, \quad r_0 = 1 \tag{4}$$

$$r^c_\nu = r_1 \vee r_2 \vee \cdots \vee r_\nu = \bigvee_{i=1}^{\nu} r_i \tag{5}$$

$$\bar{r}^c_\nu \;=\; \overline{\bigvee_{i=0}^{\nu} r_i} = \bigwedge_{i=1}^{\nu} \bar{r}_i \qquad (6)$$

The corresponding arithmetic polynomials can be written as

$$\begin{aligned}
r^o_\nu &= S_{\nu,1} - 2S_{\nu,2} + 4S_{\nu,3} - + \cdots + (-2)^{\nu-1} S_{\nu,\nu} \\
&= \sum_{i=1}^{\nu} (-2)^{k-1} S_{\nu,k} && (7) \\
\bar{r}^o_\nu &= \prod_{i=1}^{\nu} (1 - 2r_i) && (8) \\
r^c_\nu &= S_{\nu,1} - S_{\nu,2} + S_{\nu,3} - + \cdots + (-1)^{\nu-1} S_{\nu,\nu} \\
&= \sum_{k=1}^{\nu} (-2)^{k-1} S_{\nu,k} && (9) \\
\bar{r}^c &= \prod_{i=1}^{\nu} (1 - r_i) && (10)
\end{aligned}$$

where

$$S_{\nu,k} \;=\; \sum_{1 \le i_1 < i_2 < \cdots < i_k \le \nu} r_{i_1} r_{i_2} \cdots r_{i_k} \qquad (11)$$

that is, the summation in Eq.11 is extended over all $k$ combinations $\{i_1, i_2, \ldots, i_k\}$ of the $\nu$ indices.

In genetic parlance the general inclusion-exclusion principle [Charalambides, (2005)] states: the probabilities $q_{\nu,k}$ and $p_{\nu,k}$ that crossover occurs in at least $k$ and in exactly $k$ intervals, respectively, are given by

$$\begin{aligned}
q_{\nu,k} &= \sum_{i=k}^{\nu} (-1)^{i-k} \binom{i-1}{k-1} S\nu, i \quad 1 \le k \le \nu && (12) \\
p_{\nu,k} &= \sum_{i=k}^{\nu} (-1)^{i-k} \binom{i}{k} S_{\nu,i} \quad 0 \le k \le \nu && (13)
\end{aligned}$$

where $S_{\nu.0} = 1$ and $S\nu, i$ is given by Eq.11.

Comparing the polynomial $r^c_\nu$ (Eq.9) with $q_{\nu,1}$ (Eq.13 for $k = 1$) shows that the two polynomials are identical: $r^c_\nu = q_{\nu,1}$. Hence crossover must occur in at least one interval when there is complete interference. Similarly, the complement of the noninterference function $r^o_\nu$ can be defined as

$$\begin{aligned}
\bar{r}^o_\nu &= 1 \oplus r_1 \oplus \cdots \oplus r_\nu = \bigoplus_{i=0}^{\nu} r_i, \quad r_0 = 1 && (14) \\
\bar{r}^o_\nu &= \prod_{i=1}^{\nu} (1 - 2r_i) && (15)
\end{aligned}$$

and the functions of complete interference as

$$\begin{aligned}
\bar{r}^c_\nu &= \bar{r}_1 \wedge \bar{r}_2 \wedge \ldots \wedge \bar{r}_\nu = \bigwedge_{i=1}^{\nu} \bar{r}_i && (16) \\
\bar{r}^c_\nu &= \prod_{i=1}^{c} (1 - r_i) = \sum_{i=1}^{\nu} (-1)^i S_{\nu,i} \quad S_{\nu,0} = 1 && (17)
\end{aligned}$$

The relation between the probabilities $q_{\nu,k}$ and $p_{\nu,k}$ is given by

$$\begin{aligned}
q_{\nu,k} &= \sum_{i=k}^{\nu} p_{\nu,i} && (18) \\
1 - q_{\nu,k} &= \sum_{i=0}^{k-1} p_{\nu,i}, \quad 1 \le k \le \nu && (19)
\end{aligned}$$

and the relation between Eq.18 and Eq.19 can be shown via the identity $\sum_{i=0}^{\nu} p_i = 1$.

The functions NI and CI can now be rewritten as

$$\begin{aligned}
r^o_\nu &= \sum_{i=1}^{\lfloor \nu/2 \rfloor} p_{\nu,2i-1} && (20) \\
r^c_\nu &= q_{\nu,1} && (21)
\end{aligned}$$

These connections to the area of the Inclusion Exclusion Principle, may be useful in narrowing the interference bounds (e.g., by Bonferroni type inequalities).

### *A. Interference Measure*

This subject will be general and short because of the author's limited knowledge of the intricacies of data collection and classification. No special examples will be provided.

Two interference measures will be considered, both under the assumption that the number of odd recombinant classes will not change under interference or will change only incrementally. Then only even recombinant classes need to be considered. The first interference measure considered is based on the calculation of the coefficient of coincidence $C$ by [Griffiths et al, (2005)] for the single class of double recombinants in a 3-point system. The four-point system has three classes of double recombinants, located in cells 3, 5, 6, or in literal notation, respectively, in $\bar{r}_3 r_2 r_1$, $r_3 \bar{r}_2 r_1$ and $r_2 r_1 \bar{r}_1$. The idea is to individually find $C$ for each even recombinant class and then to take their average. The recombination fractions $r_i$, and the observed frequency of double recombination classes can be calculated from the data and the expected frequency of each class is the literal notation, just stated. There are two dissatisfactory points to this definition, the use of averaging and the exponential explosion of even recombinant classes.

Another definition of interference index is an extension of Haldane's formula. The function Eq.8 contains only even minterms and the function Eq.10 is the literal notation of the zero cell. Hence the difference of the two functions is a function of all non-zero even cells, that is, the difference is a CI function and the NI function would be zero:

$$I^c_\nu = \prod_{i=1}^{\nu} (1 - 2r_i) - \prod_{i=1}^{\nu} (1 - r_i) \qquad (22)$$

$$I^o_\nu = \prod_{i=1}^{\nu} (1 - r_i) - \prod_{i=1}^{\nu} (1 - r_i) = 0 \qquad (23)$$

Replacing 2 by a real number $\gamma$, where $2 \ge \gamma \ge 1$, would produce an infinity of interference functions. Usually, the range of this variable is between 1 and 0. This range shift can be accomplished by adding 1 to $\gamma$. Thus let $\delta = \gamma + 1$. Then

$$I_\nu = \prod_{i=1}^{\nu}(1-\delta r_i) - \prod_{i=1}^{\nu}(1-r_i) \tag{24}$$

Incidentally, the variable $\delta$ also appears in the Haldane differential equation for a three-point system, when the assumption $r^c = r_1 + r_2$ is replaced by $r^c = r_1 + r_2 - r_1 r_2$. The function $I_\nu$ depends on both the number of intervals $\nu$, and the real number $\delta$, but extracting the latter number from the polynomial will be very difficult if not impossible.

### *B. Malfunction as a Model*

Finally, a model based on ideas taken from fault detection of logic circuits is offered to show violation of the assumption that the number of all odd recombinants remain the same. The two operations $\oplus$ and $\vee$ used in the definition of $r^o$ and $r^c$, respectively, differ from each other in just one combination: $1 \oplus 1 = 0$ but $1 \vee 1 = 1$. Assume now that the $\oplus$ operator malfunctions and gets "stuck-at 1"; i.e., the $\oplus$-operation changes to $\vee$-operation. If all $\oplus$-operators malfunction over time in this way, then complete interference occurs.

For $\nu = 3$, the nonzero even recombinants, which are all double recombinants, can occupy three cells of the 3-cube interval map. The NI function is $r_3^o = r_3 \oplus r_2 \oplus r_1$. If the left operator gets stuck-at-1, then the function changes to $r_3 \vee r_2 \oplus r_1$. These two operations are not distributive. If first the $\oplus$ operation is performed, then

$$r_3 \vee (r_2 \oplus r_1) = \vee(1,2,4,7,5,6)$$

where the decimal numbers refer to the cells (minterms) of the new function. the four cells 1,2,4,7, defining the NI function are all present and of the remaining three cells, 5 and 6 are occupied by two double recombinant groups. None of the two functions express complete interference, since both have unoccupied even cells.

The other possibility is that first the $\vee$ operation is performed. Then

$$(r_3 \vee r_2) \oplus r_1 = \vee(1,2,4,6)$$

The three-weighted odd recombinants in cell 7 have shifted to cell 6 and the assumption is violated that no odd class disappears. This may perhaps also a cause for negative interference. But this event is rare, due to the higher priority (strength) of the $\oplus$ operation over the $\vee$ operation.

## VI. Conclusion

A boolean function can be considered a special probability function having the extreme probability values zero and one. By replacing these values by ranges, $0 \leq r \leq 0,5$ and $1 \geq \bar{r} = 1 - r \geq 0.5$, respectively, these logic functions may become quite a useful tool in the analysis of biological phenomena. The aim of this report was to draw attention to this fact by applying ideas from other disciplines.

Another aim of writing this report was to show that complete interference can also be presented by a logic function as noninterference can be. For the three-point system the complete interference formula is not $r^c = r_1 + r_2$, but $r^c = r_1 + r_2 - r_1 r_2$.

Intervals are also used in quantitative trait loci (QTL) mappings and some of the results presented here may find there too useful applications.

## References

[Charalambides, (2005)] Charalambides, C.A. (2005) Combinatorial Methods in Discrete Functions. John Wesley and Sons, New York.

[Christiansen, (2000)] Christiansen F.B. (2000) Population Genetics of Multiloci, John Wiley and Sons, New York.

[Griffiths et al, (2005)] Griffiths A.J.F.,et al. (2005) Introduction to Genetic Analysis. W.H.Freeman and Co., New York

[Karlin and Liberman, (1994)] Karlin and S, U. Liberman (1994) Theoretical Recombination Process Incorporating Interference Effects. Theor. Popul. Biology **46**: 198-231

[Mano, (1984)] Mano, M.(1984) Digital Design. Prentice Hall, Englewood Cliffs,N.J.

[Schnell,(1961)] Schnell, F.W. (1961) General Formulations of Linkage Effects in Inbreeding. Genetics **46**: 947-957

[Wu, Ma, and Casella (2007)] Wu R., C.-X. Ma, G. Casella (2007). Springer, New York.