A Study on Allelic Recombination

Carlos Cotta
Dept. Lenguajes y Ciencias de la Computación, Universidad de Málaga,
ETSI Informática, Campus de Teatinos, 29071-Málaga, Spain

ccottap@lcc.uma.es

Abstract- Allelic representations are based on characterizing points of the search space as variable-size feature sets. Recombination processes are studied here from the point of view of this kind of representations. We focus on the structure of the information units manipulated during the process, and in the algorithmic aspects of this manipulation. In this sense, we provide a generic algorithmic template whose sufficiency is established. Moreover, the syntactic properties of the information units manipulated are analyzed and exemplified. This is done within the framework of Forma Analysis.

1 Introduction

Any of the well-known reproductive operators that can be found in an evolutionary algorithm can be generally characterized as a process in which information is extracted from one or more solutions (the *parents*) in order to construct one or more solutions (the *descendants*). The construction of these descendants can be accomplished by exclusively using this information or, on the contrary, some exogenous information can be used. The latter would be the case of, e.g., mutation operators; these are specifically intended to introduce such new information in the evolutionary pool. The scenario is different in the case of recombination operators; while such exogenous information can be used in recombination, it is very common to direct the process towards combining information present in the parents, without introducing the so-called *implicit mutation*. Actually, this methodological principle has been linked to good performance in some contexts, e.g., in the travelling salesman problem [7, 11], or in bayesian network inference [1] among others.

The term *transmitting* [10] has been coined to denote this way of recombining in which using exogenous information is avoided. In some sense, transmission captures the classical role of recombination: exploiting available information, combining valuable portions of solutions that were independently discovered. It is thus worth taking a deeper look at the innards of recombination from a transmitting perspective.

Traditionally, the syntactic-algorithmic properties of a recombination operator (i.e., the shape and structure of the information pieces it manipulates, and the computational pattern used for this) have been studied from a *genetic* perspective. This means that *genes* are taken as the central part of the analysis (e.g., see [4].) A different view is assumed in this work: *alleles* will be the basic constituents around which we articulate the analysis. This alternative view implies several differences in the *modus operandi* of the process, and results in a different –and sometimes simpler– algorithmic template.

2 Preliminaries

Forma Analysis [8] has been used as the formal framework for carrying out the study. For this reason, a brief overview of the main notions required in the subsequent analysis is provided in this section. The definition of these notions is conveniently adapted to the purposes of this work.

First of all, let $\mathcal{S}$ be a discrete search space, and let $\Xi = \{\psi_1, \cdots, \psi_n\}$ be a set of n independent equivalence relations defined over $\mathcal{S}$. Let $\mathcal{C}_\psi$ be the set of equivalence classes induced by ψ, and let Ξ_η be the equivalence relation that induces equivalence class η. Also, let $[x]_\psi$ be the equivalence class to which a certain $x \in \mathcal{S}$ belongs under ψ. Now, let it hold for Ξ that

$$\forall x, y \in \mathcal{S} : \exists \psi \in \Xi : [x]_\psi \neq [y]_\psi. \tag{1}$$

This way, Ξ is said to *cover* $\mathcal{S}$, since any two solutions can be distinguished by at least one equivalence relation in Ξ. This implies in turn that any $x \in \mathcal{S}$ can be univocally referred to (i.e., represented) by enumerating the equivalence classes to which it belongs under any $\psi \in \Xi$, i.e., $x = \{[x]_{\psi_1}, \cdots, [x]_{\psi_n}\}$. Any of these equivalence relations in Ξ is termed a *basic forma* [8]. Basic formae will be usually chosen so as to capture relevant features of solutions with respect to the problem under consideration. Notice thus that equivalence relations model genes, and equivalence classes model alleles for the corresponding gene.

Let us now define the notion of *dynastic potential* as follows:

Definition 1. Let $P \subseteq \mathcal{S}$ be a set of solutions (the parents). Their dynastic potential $\Gamma(P)$ is:

$$\Gamma(P) \triangleq \bigcap_{\psi \in \Xi} \bigcup_{x \in P} [x]_\psi, \tag{2}$$

i.e., the set of all solutions that can be built just using information comprised in any parent $x \in P$. ■

Related to the above concept, the *similarity set* of two solutions is defined as:

Definition 2. Let $P \subseteq \mathcal{S}$ be a set of solutions. Their similarity set $\Sigma(P)$ is:

$$\Sigma(P) \triangleq \bigcap_{\psi \in \Xi, x \in P, P \subseteq [x]_\psi} [x]_\psi, \tag{3}$$

i.e., the intersection of common alleles for all parents. ■

Given two alleles ξ and ξ', they are said to be *compatible* if $\xi \cap \xi' \neq \emptyset$. Let η and ζ be two such compatible alleles. If for any $x \in \eta$, $y \in \zeta$, it holds that $\eta \cap \zeta \cap \Sigma(\{x, y\}) \neq \emptyset$, then Ξ is said to be *separable*. If the intersection of any arbitrary set of alleles $\{\eta_1, \cdots, \eta_n\}$ –where $\eta_i \in \mathcal{C}_{\psi_i}$– is non-empty, Ξ is said to be *orthogonal*. This latter property is stronger than separability [10].

Let the notation $\xi \triangleright \Psi$ denote that given $\Psi = \cap_{j=1}^{s} \theta_j$ –where each θ_j is induced by a different $\psi \in \Xi$–, there exist some θ_j such that $\xi \equiv \theta_j$, $\{\xi, \theta_j\} \subseteq \mathcal{C}_\psi$.

Now, let a recombination operator X be defined as a function $X : \mathcal{S}^k \times \mathcal{S} \to [0, 1]$, where $X(\{x_1, \cdots, x_k\}, z)$ is the probability that X generate z as a descendant when the parents $x_1, \cdots, x_k$ are recombined. The notion of *dynastic span* can then be defined as:

Definition 3. Let $P \subseteq \mathcal{S}$ be a set of solutions. Their dynastic span $\Gamma^1_X(P)$ with respect to a recombination operator X is

$$\Gamma^1_X(P) \triangleq \{z \mid X(P, z) > 0\}, \tag{4}$$

i.e., the set of all potential solutions that could be generated when recombining P using operator X. ■

A recombination operator is said to be transmitting if, and only if, $\Gamma^1_X(P) \subseteq \Gamma(P)$, for any $P \subseteq \mathcal{S}$, that is, it only generates solutions comprising features exhibited by at least one of the parents.

For the sake of simplicity, we will consider here the case of $k = 2$, i.e., two-parent recombination, and will use the notation $X(x, y, z)$ to denote $X(\{x, y\}, z)$. Similarly, we will use $\Gamma(x, y)$ and $\Sigma(x, y)$ to denote $\Gamma(\{x, y\})$ and $\Sigma(\{x, y\})$.

3 Analysis of Allelic Recombination

We will now present the syntactic and algorithmic aspects that have been studied. As a first step, the classical situation of genetic recombination will be considered; subsequently, this situation will be generalized to the allelic context.

3.1 The Basis of Genetic Transmission

As mentioned in Section 1, transmitting recombination can be cast as an iterative process in which information pieces are incrementally extracted from the parents, and used to construct the offspring. Let us firstly consider the shape of these information pieces.

Let the chosen representation be expressed by means of a set of genes (equivalence relations) Ξ. Obviously, this gene set must cover the search space $\mathcal{S}$. Now, let $x, y \in \mathcal{S}$ be the solutions to be recombined. It is clear that, according to Ξ, the minimal amount of information that can be transmitted from any parent to the descendant correspond to the alleles for specific genes. In other words, let $\{[x]_\varphi, [y]_\varphi\}$ be the alleles that are present in the parents for gene φ; Any of these alleles could be transmitted to the descendant z, i.e., it could be chosen –at least in principle– whether $z \in [x]_\varphi$ or $z \in [y]_\varphi$. The use of the conditional tense in the last sentences is important, for these minimal information pieces cannot be freely combined in general. The reason is twofold:

1. Unless the representation were orthogonal, some combinations of alleles will be incompatible. For example, assume two genes φ and φ' for which $\mathcal{C}_\varphi = \{\eta, \eta'\}$, and $\mathcal{C}_{\varphi'} = \{\zeta, \zeta'\}$. Assume also that allele η' is incompatible with allele ζ' (i.e., $\eta' \cap \zeta' = \emptyset$). If it were the case that $x \in \eta \cap \zeta'$ and $y \in \eta' \cap \zeta$, and in a previous step allele η' had been transmitted to the descendant, then the transmission of allele ζ' would be forbidden (or equivalently, the transmission of allele ζ would be enforced). In practice, this means that $\eta' \cap \zeta$ is in this case a non-divisible information piece.

2. Even if a set of alleles induced by different genes is compatible, their intersection may be incompatible with an allele outside this set. In particular, it may be incompatible with any of the alleles that for a certain gene were in the parents. This can be alternatively formulated as the fact that they are compatible in general, but incompatible within the dynastic potential of the solutions being recombined, i.e., $\eta \cap \zeta \neq \emptyset$ but $\Gamma(x, y) \cap \eta \cap \zeta = \emptyset$.

These considerations lead to the notion of *genetic compatibility sets*, the smallest (possibly non-elementary) pieces of information that can be *freely* manipulated without jeopardizing the feasibility of the descendant within the dynastic potential [3]. More formally, they can be defined as follows:

Definition 4. The genetic compatibility set of allele η induced by gene φ is the transitive closure of:

$$\eta \triangleright K(\Psi, \eta, x, y) \tag{5}$$

$$[\Gamma(x, y) \cap \Psi \cap K(\Psi, \eta, x, y) \cap \varpi(\eta', x, y) = \emptyset] \Rightarrow \Rightarrow \eta' \triangleright K(\Psi, \eta, x, y), \tag{6}$$

where $\varpi(\eta', x, y)$ is $[y]_{\Xi_{\eta'}}$ if $x \in \eta'$ or $[x]_{\Xi_{\eta'}}$ otherwise, and Ψ is the partially constructed descendant. ■

Thus, the genetic compatibility set of an allele is the intersection of all alleles in x for which the corresponding alleles in y would make the descendant be infeasible within the dynastic potential of the parents.

The application of this definition to a specific representation will provide the syntactic details of the these freely-manipulable minimal information units. It is then possible to define an algorithmic pattern for genetic transmission based on these units: unspecified genes are successively selected in the descendant, and the genetic compatibility set associated with the allele that for that gene is present in any of the parents is injected in the descendant. More formally:

Definition 5. The *gene transmission* operator GT is a procedure for transmitting genes defined as:

1. $\Psi = \Sigma(x, y)$
2. **while** $|\Psi| > 1$ **do**
   (a) Select an unspecified gene ψ in Ψ, i.e., a gene for which $\Psi \cap [x]_\psi \neq \emptyset$ and $\Psi \cap [y]_\psi \neq \emptyset$.
   (b) Select $\xi \in \{[x]_\psi, [y]_\psi\}$.
   (c) $\Psi = \Psi \cap K(\Psi, \xi, x, y)$.
3. **return** z (the unique solution in Ψ).

■

The sufficiency of this algorithmic pattern for spanning the dynastic potential of the parents has been already established in [2]. Next subsection will provide analogous notions and algorithmic templates in the context of allelic representations.

3.2 The Allelic Viewpoint on Transmission

The allelic perspective provide an alternative framework within which recombination can be carried out. Before defining what allelic recombination means, it is necessary to revisit the concept of allelic representation. This topic will be dealt firstly.

As before, let $\Xi = \{\psi_1, \cdots, \psi_n\}$ be the set of genes considered. For each gene $\psi_i \in \Xi$, an allele $\psi_i^0 \in \mathcal{C}_{\phi_i}$ is identified. We will refer to this allele as the *negative* allele for gene ψ_i. Intuitively, this notion is somewhat analogous to a default value for the corresponding gene. Subsequently, Let us define the global allele set $\mathcal{A}_\Xi$ as

$$\mathcal{A}_\Xi \triangleq \bigcup_{i=1}^{n} \left[\mathcal{C}_{\psi_i} \setminus \{\psi_i^0\}\right] \ . \tag{7}$$

We can now describe a solution $x \in \mathcal{S}$ by enumerating the subset of alleles in $\mathcal{A}_\Xi$ it exhibits:

$$\mathcal{A}_\Xi(x) \triangleq \{\eta \in \mathcal{A}_\Xi \mid x \in \eta\} \ . \tag{8}$$

Of course, this subset can be expressed by means of its incidence vector on $\mathcal{A}_\Xi$, that is, a binary n-dimensional vector $\overrightarrow{v} = \langle v_1, \cdots, v_n \rangle$, $v_i = 1$ if, and only if, the i^{th} allele in $\mathcal{A}_\Xi$ (under an arbitrary enumeration) is exhibited by x. Such a vector induces a genetic representation, although it is not equivalent to the original representation from which alleles are extracted (as a matter of fact, it would be a non-separable representation, even if the original representation was orthogonal.)

In a similar way a solution is represented by a set of alleles, a set of alleles $\Omega \subset \mathcal{A}_\Xi$ defines a solution $\mathcal{S}(\Omega) = \{x\}$ where

$$\mathcal{S}(\Omega) \triangleq \mathcal{S}_1(\Omega) \cap \mathcal{S}_2(\Omega) \tag{9}$$

$$\mathcal{S}_1(\Omega) \triangleq \bigcap_{\eta \in \Omega} \eta \tag{10}$$

$$\mathcal{S}_2(\Omega) \triangleq \bigcap_{\mathcal{C}_{\psi_i} \cap \Omega = \emptyset} \psi_i^0 \tag{11}$$

i.e., the intersection of the specified alleles with the negative alleles for unspecified genes. Obviously, it is possible that $\mathcal{S}(\Omega) = \emptyset$ (in other words, it represents no solution.) This can be due to two reasons:

1. $\mathcal{S}_1(\Omega) = \emptyset$. The solution tried to incorporate incompatible alleles, and hence Ω is said to be *incompatible*.
2. $\mathcal{S}_2(\Omega) = \emptyset$, or $\mathcal{S}_1(\Omega) \cap \mathcal{S}_2(\Omega) = \emptyset$. Ω is in this case *underspecified*, i.e., more non-negative alleles are required to represent a single solution.

A clear difference between the genetic and the allelic context can be observed at this point. In the former, any allele set Ω for which $|\mathcal{S}_1(\Omega)| > 1$ would be considered as a partially specified solution. On the contrary, such a set would be a well-defined solution in the allelic context as long as $\mathcal{S}(\Omega) = \{x\}$.

Having conveniently defined the allelic representation, it is now possible to consider a recombination operator manipulating this information analogously to GT. To do so, it is first necessary to define the notion of *allelic compatibility set*:

Definition 6. The allelic compatibility set of allele η is

$$K^{\mathcal{A}}(\Omega, \eta, x, y) \triangleq \{\zeta |\ \zeta \triangleright K(\mathcal{S}_1(\Omega), \eta, x, y)\} \ , \tag{12}$$

where Ω is the current allele set. ■

Now, a generic template for allelic recombination can be defined as follows:

Definition 7. The *allele transmission* operator AT is a procedure for transmitting alleles defined as:

1. $\Omega = \emptyset$.
2. $\mathcal{B} = \mathcal{A}_\Xi(x) \cup \mathcal{A}_\Xi(y)$.
3. **while** $\mathcal{B} \neq \emptyset$ **do**

(a) **if** $\mathcal{S}(\Omega) = \{z\}$
- **go to** 4 with probability p.

(b) Select $\eta \in \mathcal{B}$.

(c) Compute $K^{\mathcal{A}}(\Omega, \eta, x, y)$.

(d) **for each** $\eta' \triangleright K^{\mathcal{A}}(\Omega, \eta, x, y)$ **do**:

i. $\mathcal{B} = \mathcal{B} \setminus \{\eta', \varpi(\eta', x, y)\}$

ii. **if** $\eta' \in \mathcal{A}_\Xi$ (i.e., η' is non-negative)
- $\Omega = \Omega \cup \{\eta'\}$.

4. **return** z (the unique solution in $\mathcal{S}(\Omega)$).

■

This algorithmic template suffices to generate any solution in the dynastic potential of x and y as shown below.

Lemma 1. $z \in \Gamma(x, y) \Rightarrow \mathcal{A}_\Xi(z) \subseteq \mathcal{A}_\Xi(x) \cup \mathcal{A}_\Xi(y)$.
Proof: The proof is simple. Since z belongs to $\Gamma(x, y)$, for any allele $\xi \ni z$ it must hold that $x \in \xi$ or $y \in \xi$. In particular, this will be true for alleles $\xi \in \mathcal{A}_\Xi(z)$, and thus $\xi \in \mathcal{A}_\Xi(x)$ or $\xi \in \mathcal{A}_\Xi(y)$. □

Proposition 1. $z \in \Gamma(x, y) \Rightarrow AT(x, y, z) > 0$.
Proof: According to Lemma 1, every allele exhibited by z is initially in $\mathcal{B}$, since z belongs to $\Gamma(x, y)$. It thus suffices to have $0 < p < 1$ in step 3a, and have for the allele selection in step 3b that the probability of selecting any allele $\eta \in \mathcal{B}$ be $q(\mathcal{B}, \eta) > 0$ [1]. AT would generate z with probability bounded by

$$p(1-p)^{|\mathcal{A}_\Xi(z)|} Q(\mathcal{B}, z) \leqslant AT(x, y, z) \leqslant Q(\mathcal{B}, z) \quad (13)$$

where

$$Q(\mathcal{B}, z) = \prod_{i=1}^{|\mathcal{A}_\Xi(z)|} \sum_{\xi \in \mathcal{R}_i} q(\mathcal{B}_i, \xi) \quad (14)$$

and

$$\mathcal{R}_1 = \mathcal{A}_\Xi(z),\ \mathcal{B}_1 = \mathcal{A}_\Xi(x) \cup \mathcal{A}_\Xi(y) \quad (15)$$

$$\mathcal{R}_{i+1} = \mathcal{R}_i \setminus \{\xi\},\ \mathcal{B}_{i+1} = \mathcal{B}_i \setminus \{\xi\} \quad (16)$$

where ξ is the allele selected in each step. The lower limit in Equation (13) models the situation in which $|\mathcal{S}(\Omega)| = 1$ for every $\Omega \subseteq \mathcal{R}_1$, and $\mathcal{B}_{|\mathcal{A}_\Xi(z)|+1} \neq \emptyset$. As to the upper limit, it refers to the situation in which $\mathcal{S}(\Omega) = \emptyset$ for every $\Omega \subset \mathcal{R}_1$, and $\mathcal{B}_{|\mathcal{A}_\Xi(z)|+1} = \emptyset$. Both limits are greater than zero, so the proof is completed. □

Proposition 1 shows that any solution of the dynastic potential can be generated by operator AT. The converse is not true though: AT can generate solutions not belonging to the dynastic potential of the parents, as shown in the following proposition.

[1] E.g., having $q(\mathcal{B}, \eta) = |\mathcal{B}|^{-1}$ would be enough.

Proposition 2. $AT(x, y, z) > 0 \not\Rightarrow z \in \Gamma(x, y)$.
Proof: The proof is done by providing a counterexample to the statement $AT(x, y, z) > 0 \Rightarrow z \in \Gamma(x, y)$. Let Ξ be an orthogonal representation. This means that for any set of alleles $\Omega \subset \mathcal{A}_\Xi$, each $\xi \in \Omega$ induced by a different gene, it holds that $|\mathcal{S}(\Omega)| = 1$ (it cannot be $|\mathcal{S}(\Omega)| = 0$ since in that case there would be an incompatible set of alleles, and hence the representation would not be orthogonal.)

Let x and y be such that $\{x, y\} \in \xi$, with $\xi \in \mathcal{A}_\Xi$, i.e., x and y exhibit a common non-negative allele. The condition in step 3a of Definition 7 is always true; hence, the operator can exit without having transmitted ξ. Since in that case $z \notin \xi$, and since it can be easily seen that $\Gamma(x, y) \subseteq \xi$, it follows that $z \notin \Gamma(x, y)$. Thus, the fact that $AT(x, y, z) > 0$ does not imply that $z \in \Gamma(x, y)$. □

Notice that the result of proposition 2 is not negative; on the contrary, it points out the higher flexibility of AT. In fact, it is possible to enforce strict membership to the dynastic potential of the parents by simply changing the exit condition in step 3a ensuring that

$$\forall \eta \triangleright S_2(\Omega)\ :\ \{x, y\} \cap \eta \neq \emptyset\ . \quad (17)$$

Going back to the pseudocode in Definition 7, the most appreciable difference between AT and GT is the fact that the latter never transmits negative alleles, at least explicitly. On the contrary, such alleles are implicitly assumed upon completion of the operator execution. The relevance of this fact will be clear when some examples are tackled in Section 4. Previously, it is important to establish two results:

Lemma 2. Orthogonality of non-negative alleles implies that $K^{\mathcal{A}}(\Omega, \eta, x, y) \subseteq \mathcal{A}_\Xi(x) \cup \mathcal{A}_\Xi(y)$ for any η and Ω.
Proof: Essentially, this lemma states that the orthogonality of non-negative alleles is a sufficient condition for ensuring the absence of negative alleles in $K^{\mathcal{A}}(\Omega, \eta, x, y)$. In this situation, any allele set $\Upsilon \subset \mathcal{A}_\Xi(x) \cup \mathcal{A}_\Xi(y)$ is compatible (notice that all alleles in Υ are non-negative,) as long as at most one allele per gene is included in this set.

According to the pseudocode shown in Definition 7, every time an allele η is included in Ω, both η and $\varpi(\eta, x, y)$ are extracted from $\mathcal{B}$, the set of alleles available for recombination (see step 3(d)i in Definition 7.) This way, any allele $\zeta \in \mathcal{B}$ extracted in subsequent steps will be compatible with Ω, since it will correspond to a different gene. Notice finally that the fact that a negative allele ψ_i^0 were in the allelic compatibility set of η would mean two things:

$$\Omega \cap \mathcal{C}_{\psi_i} = \emptyset \quad (18)$$

$$\forall \xi \in [\mathcal{A}_\Xi(x) \cup \mathcal{A}_\Xi(y)] \cap \mathcal{C}_{\psi_i} : \xi \cap \bigcap_{\zeta \in \Omega} \zeta = \emptyset \quad (19)$$

Equation (18) indicates that ψ_i is an unspecified gene in Ω, and Equation (19) indicates that all alleles available for this gene are incompatible with Ω (if any of them were not, there would not be the need for including the negative allele for that gene.) However, this incompatibility is impossible, since non-negative alleles are orthogonal according to the statement of the lemma. Hence, it is impossible to have $\psi_i^0 \in K^{\mathcal{A}}(\Omega, \eta, x, y)$. □

Theorem 1. The orthogonality of non-negative alleles implies that the procedure described in Definition 7 can be cast as follows:

1. $\Omega = \emptyset$; $\mathcal{B} = \mathcal{A}_\Xi(x) \cup \mathcal{A}_\Xi(y)$.
2. **while** $\mathcal{B} \neq \emptyset$ **do**
   (a) **if** $\mathcal{S}(\Omega) = \{z\}$
      - **go to** 3 with probability p.

   (b) Select $\eta \in \mathcal{B}$.
   (c) $\Omega = \Omega \cup \{\eta\}$.
3. **return** z (the unique solution in $\mathcal{S}(\Omega)$).

Proof: As shown in Lemma 2, $K^{\mathcal{A}}(\Omega, \eta, x, y) \subseteq \mathcal{A}_\Xi(x) \cup \mathcal{A}_\Xi(y)$ as long as non-negative alleles are orthogonal. Given $\zeta \in K^{\mathcal{A}}(\Omega, \eta, x, y)$, $(\zeta \neq \eta)$ it is clear that $\varpi(\zeta, x, y)$ is a negative allele, since in case it were not, and provided that

$$S_1(\Omega \cup \left[K^{\mathcal{A}}(\Omega, \eta, x, y) \setminus \{\zeta\}\right]) \cap \varpi(\zeta, x, y) = \emptyset \ , \quad (20)$$

it would follow that non-negative alleles would not be orthogonal, since only non-negative alleles are involved in this expression, thus contradicting the proposition premise. Furthermore, if these non-negative alleles are not included in Ω in that precise step, it will be still possible to include them later. Moreover, $S(\Omega)$ will not be reduced to a singleton until these are included, fact that will necessarily happen *a fortiori*. □

This is an important result for it shows that it is not necessary computing compatibility sets when non-negative alleles are orthogonal: any allele that belonged to any such compatibility set would be finally included at any rate. This does simplify the mechanics of recombination. It must be also taken into account that mere separability of non-negative alleles does not suffice to guarantee the result of Lemma 2: plain orthogonality is required.

4 Examples

The notions presented Section 3 will be exemplified here. We consider three different problems, all of them based on search spaces comprised in the power set of some element set. In such search spaces the difference between the allelic and the genetic representation becomes more evident.

4.1 Allelic and Genetic Representation of Sets

We will firstly provide some general ideas on representing sets [9]. Let $\mathcal{E} = \{\epsilon_1, \cdots, \epsilon_n\}$ be a set of n elements, and let $2^{\mathcal{E}}$ be the power set of $\mathcal{E}$. Now, let $f : \mathcal{E} \times 2^{\mathcal{E}} \to \{0, 1\}$ be a function that computes the incidence of an element $\epsilon \in \mathcal{E}$ in a set $s \subseteq \mathcal{E}$, i.e., $f(\epsilon, s) = 1 \Leftrightarrow \epsilon \in s$. An equivalence relation ψ_ϵ can then be defined for each $\epsilon \in \mathcal{E}$, such that

$$\psi_\epsilon(s_1, s_2) \triangleq [f(\epsilon, s_1) = f(\epsilon, s_2)] \ . \quad (21)$$

This way, each ψ_ϵ divides the search space in two classes ψ_ϵ^0 and ψ_ϵ^1, respectively comprising the subsets of $\mathcal{E}$ including or excluding element ϵ. It is then clear that the set of genes $\Xi = \{\psi_\epsilon \ | \epsilon \in \mathcal{E}\}$ covers the search space, and can be used for representing solutions. Thus, any $s \in 2^{\mathcal{E}}$ can be described as $\{\psi_1^{i_1}, \psi_2^{i_2}, \cdots, \psi_n^{i_n}\}$, where the superscript vector $\langle i_1, i_2, \cdots, i_n \rangle$ is precisely the incidence vector of s on $\mathcal{E}$.

As in [9], let each allele η be associated to two sets $\xi^+(\eta)$ and $\xi^-(\eta)$ defined as

$$\xi^+(\eta) \triangleq \{v \mid \psi_v^1 \triangleright \eta\}, \quad \xi^-(\eta) \triangleq \{v \mid \psi_v^0 \triangleright \eta\} \ . \quad (22)$$

According to this definition, $\xi^+(\eta)$ is the set of all elements whose membership to any set $s \in \eta$ is enforced, and $\xi^-(\eta)$ comprise those elements whose membership to any $s \in \eta$ is forbidden.

The fact that each ψ_ϵ induces exactly two alleles ψ_ϵ^0 and ψ_ϵ^1 provides a natural way of defining the allelic representation: alleles ψ_ϵ^0 could be considered negative. It then follows that for any compatible non-negative allele set Ω,

$$S(\Omega) = \{z\} \Rightarrow z = \xi^+(S_1(\Omega)) \quad (23)$$

The converse would not hold in general, since z could be an infeasible set.

4.2 Vertex Cover

As a first example, let us consider a well-known combinatorial problem, Vertex Cover. This problem is defined on the basis of an undirected graph $G(V, E)$. The search space comprises in this case all subsets of vertices $V' \subseteq V$ for which it holds that

$$(v, w) \in E \Rightarrow \{v, w\} \cap V' \neq \emptyset \ , \quad (24)$$

i.e., every edge in E has at least one of its endpoints in V'. Determining whether there exist a vertex cover of an arbitrary size for an arbitrary graph G can be shown to be NP-complete, and finding the smallest vertex cover is NP-hard [6].

To compute the genetic compatibility sets for this problem one has to consider Equation (5) in the first place:

$$\psi_w^k \triangleright K(\Psi, \psi_w^k, s_1, s_2) \quad k \in \{0, 1\} \quad (25)$$

Subsequently, let us consider how Equation (6) applies to this problem. A partially specified solution Ψ is valid (i.e., can be extended to be a feasible solution) as long as no $v, w \in \xi^-(\Psi)$ exist such that $(v, w) \in E$. We also know that for any unspecified gene ψ_z, $\Psi \cap \psi_z^k \neq \emptyset$ for $k \in \{0, 1\}$ (otherwise, either ψ_z^0 or ψ_z^1 would belong to the genetic compatibility set of other gene specified in Ψ, so ψ_z would not be actually unspecified.) It thus follows that for all $w \in \xi^-(\Psi)$ and $(v, w) \in E$, $\psi_v^1 \triangleright \Psi$ (i.e., $v \in \xi^+(\Psi)$).

After having stated this, assume that a ψ_w^1 allele is being transmitted. Including vertex w in the final solution cannot make Ψ be invalid (actually it makes all edges $(v, w) \in E$ be covered,) so no further gene value need be transmitted as a part of the genetic compatibility set, i.e., Equation (25) suffices for this case. We now consider the situation in which a ψ_w^0 allele is transmitted. In this case, it is known that for all $(v, w) \in E$, either $\psi_v^1 \triangleright \Psi$ or ψ_v is unspecified. In the latter case, transmitting also ψ_v^0 would make Ψ be invalid, so ψ_v^1 must be transmitted. Thus,

$$K(\Psi, \psi_w^0, s_1, s_2) = \psi_w^0 \cap \bigcap_{(v,w)\in E,\ v\notin\xi^+(\Psi)} \psi_v^1 \qquad (26)$$

In the context of the allelic representation the situation is different: one never has to decide whether transmit ψ_w^0 or ψ_w^1, but which of the ψ_v^1 alleles in the set of available alleles $\mathcal{B}$ has to be transmitted. Since non-negative alleles are orthogonal in this case (for any subset of non-negative alleles Ω, $\xi^+(S_1(\Omega))$ is a vertex cover or can be extended with more vertices to be a vertex cover,) AT simply has to transmit one-allele-at-a-time (the allele selected from $\mathcal{B}$ in step 2b) according to Theorem 1.

4.3 Clique

The Clique problem is also defined on the basis of an undirected graph $G(V, E)$. The search space is in this case composed of all subsets of vertices $V' \subseteq V$ for which it holds that

$$v, w \in V' \Rightarrow (v, w) \in E \ , \qquad (27)$$

i.e., V' is a fully-connected subset. As it was the case for Vertex Cover, determining whether a clique of arbitrary size exists for G is NP-complete, and finding the largest clique in a graph is NP-hard [6].

We consider the same representation presented for Vertex Cover. Again, we have in the first place that $\psi_w^k \triangleright K(\Psi, \psi_w^k, s_1, s_2)$, for $k \in \{0, 1\}$. Now, the requirement for validity of Ψ is having $\xi^+(\Psi)$ being fully connected. Since this set is not modified when an allele ψ_w^0 is transmitted, the previous equation suffices for this case. The situation is different when an allele ψ_w^1 is transmitted. In this situation, it is clear that

$$(v, w) \notin E \Rightarrow \left(\Psi \cap \psi_w^1 \cap \psi_v^1 = \emptyset\right) \ . \qquad (28)$$

Thus,

$$K(\Psi, \psi_w^1, s_1, s_2) = \psi_w^1 \cap \bigcap_{(v,w)\notin E,\ v\notin\xi^-(\Psi)} \psi_v^0 \qquad (29)$$

Moving to the allelic context, notice that ψ_v^1 alleles are not orthogonal (if $(v, w) \notin E$, then $\psi_w^1 \cap \psi_v^1 = \emptyset$), so Theorem 1 cannot be applied. However, ψ_v^0 alleles are indeed orthogonal (for any subset of ψ_v^0-like alleles Ω, $V \setminus \xi^-(S_1(\Omega))$ is a clique, or can be converted in a clique by extracting more vertices.) Hence, considering ψ_v^1 alleles as negative alleles, each solution s is represented as a set of ψ_v^0 alleles, with $v \in V \setminus s$. By transmitting these alleles one-at-a-time, it is ensured that any solution within the dynastic potential can be generated.

4.4 Dag structure learning

A directed acyclic graph (DAG) is a digraph $DAG(V, E)$, where E verifies that no $v, v' \in V$ exist such that there are directed paths from v to v' and vice versa. DAGs are important data structures since they can be used to represent many relevant entities such as *field programmable gate arrays* (FPGAs), or Bayesian networks among others.

We can represent DAGs as a set of edges, i.e., $\mathcal{E} = E$ (rather than $\mathcal{E} = V$ as in the previous examples). Then, $\Xi = \{\psi_{ij} \mid i, j \in V\}$, each ψ_{ij} dividing the space of DAGs in two classes: those DAGs including edge (i, j), and those without it. As it was the case for Vertex Cover and Clique, it is easy to see that this representation is not orthogonal (e.g., $\psi_{ij} \cap \psi_{ji} = \emptyset$). It is then necessary to consider the shape of compatibility sets.

Again, $K(\Psi, \psi_{ij}^0, d_1, d_2) = \psi_{ij}^0$, since excluding an edge from a DAG cannot make it infeasible, i.e., it cannot introduce a cycle. This risk of introducing cycles only exist when an edge (i, j) is transmitted to the DAG. In this case, cycle avoidance implies that edges from j –and from any k for which a directed path from j to k exist– to i must be forbidden in the descendant. Let C_D be the adjacency matrix of a digraph D, and let C_D^∞ be the transitive closure of C_D, then

$$K(\Psi, \psi_{ij}^1, d_1, d_2) = \psi_{ij}^1 \cap \bigcap_{C^\oplus(r,s)=1} \psi_{sr}^0 \qquad (30)$$

where $C^\oplus = C^\infty_{\xi^+(\Psi)}$ XOR $C^\infty_{\xi^+(\Psi\cap\psi_{ij}^1)}$.

We will use this DAG structure learning problem to exemplify an interesting difference in the dynamics of GT and AT. Recall firstly that GT would work in this case by selecting an unspecified gene, and determining which of the available alleles for that gene will be transmitted to the descendant. Since any time a ψ_{ij}^1 allele is transmitted a number of other ψ_{sr}^0 alleles are enforced (i.e., several edges are forbidden,) it is clear that GT has a bias to produce offspring with a lower number of

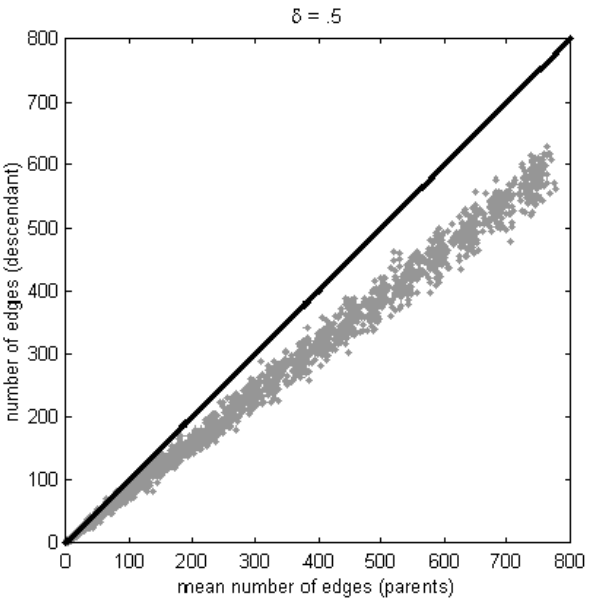

Figure 1: Descendant number of edges for GT as a function of the parents' mean number of edges. Parent density fixed to $\delta = .5$

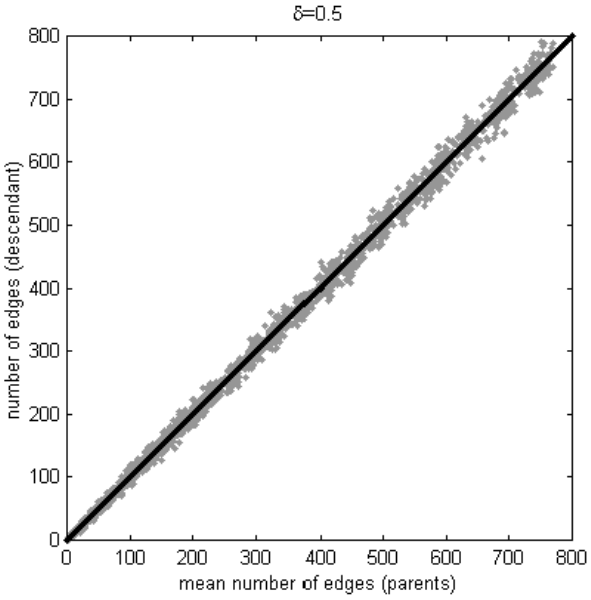

Figure 2: Descendant number of edges for AT as a function of the parents' mean number of edges. Parent density fixed to $\delta = .5$

edges than the parents' average. This can be seen in Figure 1.

As the number of vertices grows, the deviation from the parents' average also grows (empirical modelling [5] suggests that the descendant's mean number of edges is proportional to $\mu_e^{0.96}$, where μ_e is the parents' mean number of edges.)

In the allelic context, AT would work by selecting any of the available ψ_{ij}^1 alleles until no such allele remains, or an early exit is decided. It is not difficult to adjust this exit criterion so as to have a desired number of edges in the descendant. This is shown in Figure 2; to be precise, the exit criterion has been here selected as reaching a number of edges following a binomial distribution $(\nu,\ \phi)$ where $\phi = 1/2$ and $\nu/2$ approximates the parents' mean number of edges. This results in a recombination operator without density bias.

5 Conclusions

This work has presented an analysis of recombination from the allelic point of view. It has been shown that the syntax of the the information pieces manipulated by an allelic operator can be simpler than those of its genetic counterpart. In turn, such a simplification results in the possibility of using a simpler algorithmic template, whose functioning is also more flexible than that of the genetic operator.

Among the advantages of using an allelic approach we can cite the possibility of exerting a stronger control on the information exchange, as it was exemplified in the domain of DAGs. Furthermore, the allelic operator can be gracefully endowed with heuristic information in this context. This has been shown in [1], where a heuristic measure was successfully used for selecting the next edge to be transmitted (recall step 3b in Definition 7) in a Bayesian-network inference problem. This policy is more flexible than the treatment of individual genes that would be carried out in the genetic context.

Future work will be directed to a quantitative study of AT properties. For example, the closed form of the distribution probability defined over the dynastic potential by some instances of GT is known. Finding analogous expressions for AT must be the next step.

Acknowledgement

The author is partially supported by the *Ministerio de Ciencia y Tecnología* (MCyT) and *Fondo Europeo para el Desarrollo Regional* (FEDER) under grant TIC2002-04498-C05-02.

Bibliography

[1] C. Cotta and J. Muruzábal. Towards a more efficient evolutionary induction of bayesian networks. In J.J. Merelo et al., editors, *Parallel Problem Solving from Nature VII*, volume 2439 of *Lecture Notes in Computer Science*, pages 730–739. Springer-Verlag, Berlin Heidelberg, 2002.

[2] C. Cotta and J.M. Troya. Information processing in transmitting recombination. *Applied Mathematics Letters*. In Press.

[3] C. Cotta and J.M. Troya. On the influence of the representation granularity in heuristic forma recombination. In J. Carroll et al., editors, *ACM*

Symposium on Applied Computing 2000, pages 433–439. ACM Press, 2000.

[4] C. Cotta and J.M. Troya. Using dynastic exploring recombination to promote diversity in genetic search. In M. Schoenauer et al., editors, *Parallel Problem Solving from Nature VI*, volume 1917 of *Lecture Notes in Computer Science*, pages 325–334. Springer-Verlag, Berlin, 2000.

[5] C. Cotta and J.M. Troya. Analyzing directed acyclic graph recombination. In B. Reusch, editor, *Computational Intelligence: Theory and Applications*, volume 2206 of *Lecture Notes in Computer Science*, pages 739–748. Springer-Verlag, Berlin Heidelberg, 2001.

[6] M.R. Garey and D.S Johnson. *Computers and Intractability: A Guide to the Theory of NP-Completeness*. Freeman and Co., San Francisco CA, 1979.

[7] K. Mathias and D. Whitley. Genetic operators, the fitness landscape and the traveling salesman problem. In R. Männer and B. Manderick, editors, *Parallel Problem Solving From Nature II*, pages 259–268. Elsevier, Amsterdam, 1992.

[8] N.J. Radcliffe. Equivalence class analysis of genetic algorithms. *Complex Systems*, 5:183–205, 1991.

[9] N.J. Radcliffe. Genetic set recombination. In D. Whitley, editor, *Foundations of Genetic Algorithms II*, pages 203–219, San Mateo CA, 1992. Morgan Kauffman.

[10] N.J. Radcliffe. The algebra of genetic algorithms. *Annals of Mathematics and Artificial Intelligence*, 10:339–384, 1994.

[11] T. Starkweather, S. McDaniel, K. Mathias, D. Whitley, and C. Whitley. A comparison of genetic sequencing operators. In R.K. Belew and L.B. Booker, editors, *Proceedings of the Fourth International Conference on Genetic Algorithms*, pages 69–76, San Mateo CA, 1991. Morgan Kauffman.

Update – January 2004

There is a caveat in the proof of Lemma 2. To be precise, Equation (19) does not fully characterize the situations in which a negative allele is in $K^{\mathcal{A}}(\Omega, \eta, x, y)$. The correct expression should be:

$$\forall \xi \in [\mathcal{A}_\Xi(x) \cup \mathcal{A}_\Xi(y)] \cap \mathcal{C}_{\psi_i} : \Gamma(x,y) \cap \xi \cap \bigcap_{\zeta \in \Omega} \zeta = \emptyset \tag{31}$$

The additional term $\Gamma(x, y)$ is required to ensure that the compatibility of non-negative alleles holds within the dynastic potential of the solutions being recombined. It is possible that non-negative alleles be orthogonal in general, but not within an arbitrary dynastic potential. Consider the following example:

Example 1. Let $\Xi = \{\psi_1, \cdots, \psi_n\}$, were $n = 2k$ for some $k > 0$ and $\mathcal{C}_{\psi_i} = \{\psi_i^0, \psi_i^1\}$. Now, let the search space $\mathcal{S}$ be such that:

$$\forall x \in \mathcal{S} : \sum_{i=1}^{n} \sigma(x,i) = 2k,\ k \geqslant 0 \tag{32}$$

where $x \in \psi_i^{\sigma(x,i)}$, $1 \leqslant i \leqslant n$. It can be seen that any $S_1\left(\{\psi_{i_1}^1, \cdots, \psi_{i_m}^1\}\right)$ is non-empty (it comprises all solutions belonging to an even number $m' \geqslant m$ of positive alleles, for example $\cap_{i=1}^{n} \psi_i^1$.) Hence, non-negative alleles are orthogonal. If we consider x and y such that $|\mathcal{A}_\Xi(x) \cup \mathcal{A}_\Xi(y)|$ is odd, we can see that $S\left(\mathcal{A}_\Xi(x) \cup \mathcal{A}_\Xi(y)\right) = \emptyset$ since no solution can belong to an odd number of non-negative alleles. Hence, orthogonality does not imply compatibility within an arbitrary dynastic potential.

In the light of Equation (31), the statement of Lemma 2 (and subsequently that of Theorem 1) must be modified so as to include the tag "within an arbitrary dynastic potential" whenever "orthogonality" is referred. The examples shown in Sections 4.2 and 4.3 verify this property of orthogonality within an arbitrary dynastic potential.