A Novel DNA Sequence Vector Space over an extended Genetic Code Galois Field

*Robersy Sánchez 1,2, Ricardo Grau 2,3 and Eberto R. Morgado 3


2Faculty of Mathematics Physics and Computation, Central University of Las Villas, Villa Clara, Cuba.

3Center of Studies on Informatics, Central University of Las Villas, Villa Clara, Cuba

(Received December 19, 2005)

Abstract

In the vector space of DNA sequences over the Galois field of the 64 codons \(GF (64)\), recently published, deletions and insertions (indel) could not be analyzed. Now, in order to include these kinds of mutations, we have defined a new Galois field over the set of extended triplets \(X_1X_2X_3 (C_{125})\), where \(X_1 \in \{O, A, C, G, U\}\). Taking the polynomial coefficients \(a_0, a_1, a_2 \in GF (5)\) and the bijective function \(f : GF (5) \rightarrow \{O, A, C, G, U\}\), where \(f (0) = O, f (1) = A, f (2) = C, f (3) = G, f (4) = U\), bijection \(\Psi\) is induced such that \(\Psi(a_0 + a_1x + a_2x^2) = (f(a_1), f(a_2), f(a_0)) = (X_1X_2X_3)\). The field \((C_{125}, +, \cdot)\) allows the definition of a novel \(N\)-dimensional vector space \((S)\) over the field \(GF (5^3)\) on the set of all \(125^N\) sequences of extended triplets in which all possible DNA sequence alignments of length \(N\) are included. Here the “classical gap” produced by alignment algorithm corresponds to the neutral element “O”. In the vector space \(S\), all mutational events that take place in the molecular evolution process can be described by means of endomorphisms, automorphisms and translations. In particular, the homologous (generalized) recombination between two homologous DNA duplexes involving a reciprocal exchange of DNA sequences –e.g. between two chromosomes that carry the same

* Robersy Sánchez: robersy@uclv.edu.cu

Corresponding address: Apartado postal 697. Santa Clara 1. CP 50100. Villa Clara. Cuba
genetic loci—algebraically corresponds to the action of two automorphism pairs over two paired DNA duplexes.

1. Introduction

A new $N$-dimensional vector space of DNA sequences over the Galois field of the 64 codons ($GF(64)$) was recently presented [SAN 05]. This vector space was derived taking into account the order of the bases proposed in the Boolean lattice of the four DNA bases [SAN 04] [SAN 04a]. The isomorphism $\varphi: B(\mathcal{X}) \rightarrow (\mathbb{Z}_2)^3$ between the Boolean lattices of the four DNA bases $B(\mathcal{X})$ and $((\mathbb{Z}_2)^2, \lor, \land)$ ($\mathbb{Z}_2 = \{0,1\}$), and the biological importance of base positions in the codons were used to state a partial order in the codon set. As a result every codon was represented in the field $GF(64)$ as a binary sextuplet.

In this vector space, gene point mutations were considered linear transformations or translations of the wild type gene, however deletions and insertions (indel) could not be considered. Now, in order to include indel mutations, we have defined a new Galois field over the set of elements $X_1X_2X_3 (C_{125})$, where $X_i \in \{O, A, C, G, U\}$. This set can be called the extended triplet set and the elements $X_1X_2X_3$, the extended triplets. At present, the starting base order used here comes from the recently reported $Z_{64}$-algebra of the genetic code [SAN 05a]. In this $Z_{64}$-algebra, the base order $\{A, C, G, U\}$ was obtained by considering the genetic code as a non-dimensional code scale of amino acid interaction energies in proteins.

Like in previous articles, we have kept in mind the biological importance of base position in the codon to state a codon order in the genetic code. The importance of the base position is suggested by the error frequency (accepted mutations) found in codons. Errors on the third base are more frequent than on the first base, and, in turn, these are more frequent than errors on the second base [WOE 85] [FRI 64] [PAR 89]. These positions, however, are very conservative with respect to changes in polarity of coded amino acids [ALF 69].

The principal aim of this work is to show that all mutational events that take place in the molecular evolution process can be described by means of endomorphisms, automorphisms and translations of a novel $N$-dimensional vector space over the Galois field $GF(5^3)$. The new vector space defined over the set of all $125^N$ sequences of extended triplets includes all possible DNA sequence alignments of length $N$. Here the “classical gaps” produced by alignment algorithms correspond to the neutral element “O”.
2. Theoretical model

The concepts of Galois fields and vector space over a finite field are the mathematical basic ideas used in our model. A mathematical background about these structures can be found in the references [RED 67] [WAE 70] [SAN 05].

Our starting point is the base order \{A, C, G, U\} derived from the \(Z_{64}\)-algebra of the genetic code [SAN 05a]. In order to analyze indel mutations, this alphabet may be extended including the new symbol “O” to denote base omissions (gaps) in DNA sequence alignments. As a result, a new triplet set can be built with elements \(X_1X_2X_3\) where \(X_i \in \{O, A, C, G, U\}\). This set can be called the extended triplet set \(C_{125}\) and the elements \(X_1X_2X_3\), the extended triplets.

Now, considering the order in the set \{O, A, C, G, U\} and the biological importance of the base position in the codon, it is possible to establish an order in the extended triplet set, i.e. from triplet OOO to UUU. First, in the triplets \(X_1X_2X_3\) keeping invariables bases \(X_1\) and \(X_2\), the third base \(X_3\) is consecutively changed until all possibilities are exhausted. Next, a similar variation is applied to the first base and finally to the second one, i.e. the variations are introduced from the less biologically relevant base to the most relevant base in the codon. Then, the ordered triplet set shown in the Table 1 was obtained.

2.1. Nexus between the Galois Field Elements and the Set of Codons

As one can see in Table 1, a bijection is suggested between the orders in the extended triplet set and the \(GF(5^3)\) elements. In particular, there is a bijective function \(f: GF(5) \rightarrow \{O, A, C, G, U\}\), between the elements of \(GF(5)\) and the letters \(X_k \in \{O, A, C, G, U\}\). This function is explicitly given by the equalities:

\[
f(0) = O, f(1) = A, f(2) = C, f(3) = G, f(4) = U
\]

Next, taking into account the biological importance of base positions in codons, we can state the bijective function \(\Psi : GF(5^3) \rightarrow C_{125}\) between the extended triplet set and the polynomial representation of \(GF(5^3)\) elements:

\[
\Psi(a_0 + a_1x + a_3x^2) = (f(a_1), f(a_2), f(a_0)) = X_1X_2X_3
\]
are also the coefficients of the polynomials in the biological meaning of the corresponding base. Notice that coefficients of the terms with a maximal degree correspond to the base in the first codon position, and finally, the coefficient of the term of degree 0 is assigned to the base in the second codon position. The coefficient of the term with degree 1 corresponds to the integer digit sequence corresponds to the integer representation in base 5 of the triplet index number (see Table 1). So, as an example, the following bijections are given:

The polynomial coefficient \( a_2 \) of the terms with a maximal degree \( a_2x^2 \) corresponds to the base in the second codon position. The coefficient of the term with degree 1 corresponds to the first codon position, and finally, the coefficient of the term of degree 0 is assigned to the third codon position. That is, the degree of the polynomial terms decreases according to the biological meaning of the corresponding base. Notice that coefficients \( a_i \) correspond –for every triplet– to the integer digits of its 3-tuple vector representation in \( GF(5^3) \). The reverse of this integer digit sequence corresponds to the integer representation in base 5 of the triplet index number (see Table 1). So, as an example, the following bijections are given:

<table>
<thead>
<tr>
<th>a</th>
<th>O</th>
<th>A</th>
<th>C</th>
<th>G</th>
<th>U</th>
</tr>
</thead>
<tbody>
<tr>
<td>I</td>
<td>II</td>
<td>III</td>
<td>I</td>
<td>II</td>
<td>III</td>
</tr>
<tr>
<td>0</td>
<td>000</td>
<td>OOO</td>
<td>25</td>
<td>001</td>
<td>OAO</td>
</tr>
<tr>
<td>2</td>
<td>200</td>
<td>OOC</td>
<td>27</td>
<td>201</td>
<td>OAC</td>
</tr>
<tr>
<td>3</td>
<td>300</td>
<td>OOG</td>
<td>28</td>
<td>301</td>
<td>OAG</td>
</tr>
<tr>
<td>4</td>
<td>400</td>
<td>OOU</td>
<td>29</td>
<td>401</td>
<td>OAU</td>
</tr>
<tr>
<td>5</td>
<td>010</td>
<td>AO0</td>
<td>30</td>
<td>011</td>
<td>AAO</td>
</tr>
<tr>
<td>6</td>
<td>110</td>
<td>AOA</td>
<td>31</td>
<td>111</td>
<td>AAA</td>
</tr>
<tr>
<td>7</td>
<td>210</td>
<td>AOC</td>
<td>32</td>
<td>211</td>
<td>AAC</td>
</tr>
<tr>
<td>8</td>
<td>310</td>
<td>AOG</td>
<td>33</td>
<td>311</td>
<td>AAG</td>
</tr>
<tr>
<td>9</td>
<td>410</td>
<td>A00</td>
<td>34</td>
<td>411</td>
<td>AAU</td>
</tr>
<tr>
<td>10</td>
<td>020</td>
<td>CO0</td>
<td>35</td>
<td>021</td>
<td>CAO</td>
</tr>
<tr>
<td>11</td>
<td>120</td>
<td>C0A</td>
<td>36</td>
<td>121</td>
<td>CAA</td>
</tr>
<tr>
<td>12</td>
<td>220</td>
<td>C0C</td>
<td>37</td>
<td>221</td>
<td>CAC</td>
</tr>
<tr>
<td>13</td>
<td>320</td>
<td>C0G</td>
<td>38</td>
<td>321</td>
<td>CAG</td>
</tr>
<tr>
<td>14</td>
<td>420</td>
<td>C0U</td>
<td>39</td>
<td>421</td>
<td>CAU</td>
</tr>
<tr>
<td>15</td>
<td>030</td>
<td>G00</td>
<td>40</td>
<td>031</td>
<td>GA0</td>
</tr>
<tr>
<td>16</td>
<td>130</td>
<td>G0A</td>
<td>41</td>
<td>131</td>
<td>GAA</td>
</tr>
<tr>
<td>17</td>
<td>230</td>
<td>G0C</td>
<td>42</td>
<td>231</td>
<td>GAC</td>
</tr>
<tr>
<td>18</td>
<td>330</td>
<td>G0G</td>
<td>43</td>
<td>331</td>
<td>GAG</td>
</tr>
<tr>
<td>19</td>
<td>430</td>
<td>G0U</td>
<td>44</td>
<td>431</td>
<td>GAU</td>
</tr>
<tr>
<td>20</td>
<td>040</td>
<td>U00</td>
<td>45</td>
<td>041</td>
<td>UAO</td>
</tr>
<tr>
<td>21</td>
<td>140</td>
<td>U0A</td>
<td>46</td>
<td>141</td>
<td>UAA</td>
</tr>
<tr>
<td>22</td>
<td>240</td>
<td>U0C</td>
<td>47</td>
<td>241</td>
<td>UAC</td>
</tr>
<tr>
<td>23</td>
<td>340</td>
<td>U0G</td>
<td>48</td>
<td>341</td>
<td>UAG</td>
</tr>
<tr>
<td>24</td>
<td>440</td>
<td>U0U</td>
<td>49</td>
<td>441</td>
<td>UAU</td>
</tr>
</tbody>
</table>

* In this table it is possible to see the bijection between the triplet set and the set of 3-tuples in \((Z_5)^3\), which are also the coefficients of the polynomials in the \( GF(5^3) \). The corresponding integer number of every 3-tuples is also shown. I. Triplet index number. II. Polynomial coefficients. III. Extended triplets.
In particular, we will use the bijective function \( f[\cdot] \) such that \( f: s \rightarrow GF(5^3) \), between the subset of the integer number \( s = \{0, 1 \ldots 124\} \) and the elements of \( GF(5^3) \). According to the above example \( f[7] = 2 + x \), \( f[44] = 4 + 3x + x^2 \) and \( f[117] = 2 + 3x + 4x^2 \).

### 2.2. Vector Spaces over the Genetic Code Galois Field

Now, by means of the function \( \Psi \), a product operation can be define in set \( C_{125} \). Let \( \Psi^{-1} \) be the inverse function of \( \Psi \) then, for all pair of codons \( X_1Y_1Z_1 \in C_{125} \) and \( X_2Y_2Z_2 \in C_{125} \), their product “•” will be:

\[
X_1Y_1Z_1 \cdot X_2Y_2Z_2 = \Psi[\Psi^{-1}(X_1Y_1Z_1) \Psi^{-1}(X_2Y_2Z_2) \mod g(x)]
\]

That is to say, the product between two triplets is obtained from the product of their corresponding polynomial module \( g(x) \), where \( g(x) \) is an irreducible polynomial of second degree over \( GF(5) \). Since there are 40 irreducible polynomials of second degree, 40 possible variants are available to choose the product between two extended triplets. It is not difficult to prove that the set of codons \( C_{125} \setminus \{OOO\} = C_{125}^* \) with the operation product “•” is an Abelian group \( (C_{125}^*, \cdot) \). Likewise, a sum operation is defined by using the sum operation in \( GF(5^3) \).

In this field, the sum is carried out by means of the polynomial sum in the usual fashion with polynomial coefficients reduced by module 5.

Then, for all pairs of codons \( X_1Y_1Z_1 \in C_{125} \) and \( X_2Y_2Z_2 \in C_{125} \), their sum “+” will be:

\[
X_1Y_1Z_1 + X_2Y_2Z_2 = \Psi[\Psi^{-1}(X_1Y_1Z_1) + \Psi^{-1}(X_2Y_2Z_2) \mod 5]
\]

As a result, the set of codons \( (C_{125}, +) \) with operation “+” is an Abelian group and the set \( (C_{125}, +, \cdot) \) is a field isomorphic to \( GF(5^3) \). After that, the product of a codon \( XYZ \in C_{125} \) can be defined by the element \( \alpha_i \in GF(5^3) \). For all \( \alpha_i \in GF(5^3) \) and for all \( XYZ \in C_{125} \), this operation will be defined as:
\[ \alpha_i (XYZ) = \Psi^i \Psi^{-i} (XYZ) \mod 5 \]

This operation is analogous to the multiplication rule of a vector by a scalar. So, \((C_{125}, +)\) can be considered a one-dimensional vector space over \(GF (5^3)\). The canonical base of this space is the triplet OOA. This structure can be called the vector space of extended triplets over \(GF (5^3)\). Such structure can be extended to the \(N\)-dimensional sequence space \((S)\) consisting of the set of all \(125^N\) DNA alignment sequences with \(N\) extended triplets. Obviously, this set is isomorphic to the set of all \(N\)-tuples \((x_1, \ldots, x_N)\) where \(x_i \in C_{125}\). Then, set \(S\) can be represented by all \(N\)-tuples \((x_1, \ldots, x_N) \in (C_{125})^N\). As a result, the \(N\)-dimensional vector space of \(S\) over \(GF (5^3)\) will be the direct sum

\[ S = (C_{125})^N = C_{125} \oplus C_{125} \oplus \ldots \oplus C_{125} \text{ (}N\text{ times)} \]

The sum and product in \(S\) are carried out by components (Redéi, 1967). That is, for all \(\alpha \in GF (5^3)\) and for all \(s, s' \in S\) we have:

\[ s + s' = (s_1, s_2, \ldots, s_N) + (s_1', s_2', \ldots, s_N') = (s_1 + s_1' + s_2 + s_2', \ldots, s_N + s_N') \]

\[ \alpha s = \alpha (s_1, s_2 \ldots s_N) = (\alpha s_1, \alpha s_2 \ldots \alpha s_N) \]

Next, it can be proved that \((S, +)\) is an Abelian group with the \(N\)-tuple \(s_e = (OOO, OOO \ldots OOO)\) as its neutral element. The canonical base of this space is the set of vectors:

\[ e_1 = (OOA, OOO, \ldots, OOO), \quad e_2 = (OOO, OOA, \ldots, OOO), \ldots, \quad e_N = (OOO, OOO, \ldots, OOA) \]

As a result, every sequence \(s \in S\) has a unique representation:

\[ s = \alpha_1 e_1 + \alpha_2 e_2 + \ldots + \alpha_N e_N \quad (\alpha_i \in GF (5^3)) \]

It is usually said that the \(N\)-tuple \((\alpha_1, \alpha_2, \ldots, \alpha_N)\) is the coordinate representation of \(s\) in the canonical bases \(\{e_i \in C_{125}, i = 1, 2, \ldots, N\}\) of \(S\).
3. Results and Discussion

Evidently, the Galois field of codons is not unique. Actually, we have obtained forty isomorphic Galois fields, each one with the product operation defined from one of the forty irreducible polynomials. It is convenient, however, to choose a more biologically significant Galois field.

The most attractive irreducible polynomials are the primitive polynomials. If $\alpha$ is a root of a primitive polynomial then their powers $\alpha^n$ ($n = 1, \ldots, 124$) are the elements of the multiplicative group of $GF(125)$, i.e. $\alpha$ is a group generator. As it was shown in [SAN 05], a product operation in a Galois field generated by a primitive polynomial is carried out in a very simple way (see Table 2). Just twenty of the forty irreducible polynomials are primitives.

In [SAN 05a] the sum operation is a manner to consecutively obtain all codons from the codon AAC in such a way that the genetic code will represent a non-dimensional code scale of amino acid interaction energy in proteins. Here, in order to consecutively obtain all codons from the codon AAC, primitive polynomials are chosen with root $\alpha = 2 + x + x^2$—corresponding to codon AAC. Only primitive polynomial $g(x) = 2 + 3 x^2 + x^3$ has the root $\alpha$, in this way the product operation is unique.

Notice that, in the vector space $S$, all $125^N$ possible DNA alignment sequences of length $N$ are represented. Here the “classical gap”, produced by alignment algorithms corresponds to the neutral element “O”. The neutral element appears from algebraic operations with codons. For instance, in the additive group $(C_{125}, +)$, the inverse of codons $X_1AX_3$ coding to hydrophilic amino acids are the codons $(-X_1)U(-X_3)$ that in turn code to hydrophobic amino acids. The sum of a $X_1AX_3$ codon to a $X_1UX_3$ codon produces an extended triplet $X_1OX_3$. Then, this sum introduces, at least, one base deletion in the obtained extended triplet. In general, indel mutations found in the molecular evolution process can be described by means of algebraic operations in $(C_{125}, +, \cdot)$, i.e. any deletion or insertion presented in any mutant DNA sequence is described by means of algebraic transformations of the corresponding wild type gene.
Table 2. Logarithm table of the elements of the $GF(5^3)$ generated by the primitive polynomial $g(x) = 2 + 3x^2 + x^3$.

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>$n$</td>
<td>0</td>
<td>93</td>
<td>31</td>
<td>62</td>
<td>25</td>
<td>4</td>
<td>42</td>
<td>105</td>
<td>102</td>
<td>118</td>
<td>74</td>
<td>97</td>
</tr>
<tr>
<td>$n$</td>
<td>71</td>
<td>11</td>
<td>56</td>
<td>73</td>
<td>9</td>
<td>35</td>
<td>12</td>
<td>87</td>
<td>40</td>
<td>43</td>
<td>104</td>
<td>66</td>
</tr>
<tr>
<td>$n$</td>
<td>50</td>
<td>23</td>
<td>30</td>
<td>28</td>
<td>106</td>
<td>29</td>
<td>14</td>
<td>1</td>
<td>20</td>
<td>86</td>
<td>67</td>
<td>8</td>
</tr>
<tr>
<td>$n$</td>
<td>83</td>
<td>120</td>
<td>38</td>
<td>6</td>
<td>80</td>
<td>46</td>
<td>10</td>
<td>79</td>
<td>3</td>
<td>119</td>
<td>95</td>
<td>109</td>
</tr>
<tr>
<td>$n$</td>
<td>84</td>
<td>19</td>
<td>121</td>
<td>116</td>
<td>75</td>
<td>123</td>
<td>99</td>
<td>103</td>
<td>49</td>
<td>48</td>
<td>15</td>
<td>122</td>
</tr>
<tr>
<td>$n$</td>
<td>113</td>
<td>107</td>
<td>55</td>
<td>94</td>
<td>96</td>
<td>78</td>
<td>88</td>
<td>53</td>
<td>64</td>
<td>36</td>
<td>89</td>
<td>101</td>
</tr>
<tr>
<td>$n$</td>
<td>7</td>
<td>52</td>
<td>81</td>
<td>61</td>
<td>13</td>
<td>54</td>
<td>59</td>
<td>98</td>
<td>114</td>
<td>69</td>
<td>39</td>
<td>27</td>
</tr>
<tr>
<td>$n$</td>
<td>34</td>
<td>2</td>
<td>115</td>
<td>26</td>
<td>16</td>
<td>60</td>
<td>32</td>
<td>117</td>
<td>45</td>
<td>51</td>
<td>37</td>
<td>77</td>
</tr>
<tr>
<td>$n$</td>
<td>110</td>
<td>111</td>
<td>41</td>
<td>112</td>
<td>44</td>
<td>90</td>
<td>92</td>
<td>85</td>
<td>65</td>
<td>22</td>
<td>47</td>
<td>33</td>
</tr>
<tr>
<td>$n$</td>
<td>57</td>
<td>68</td>
<td>17</td>
<td>72</td>
<td>108</td>
<td>18</td>
<td>5</td>
<td>100</td>
<td>58</td>
<td>21</td>
<td>70</td>
<td>91</td>
</tr>
<tr>
<td>$n$</td>
<td>24</td>
<td>82</td>
<td>63</td>
<td>76</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

*Here, codon AAC corresponds to the primitive root $\alpha = 2 + x + x^2$, i.e. $f[s] = (2 + x + x^2)^s \mod g(x)$ and $n = \log_{\alpha} f[s]$. The properties of this logarithm function are alike to the classical definition in arithmetic:

1. $\log_{\alpha}(f[x]^{*}f[y]) = (\log_{\alpha}f[x] + \log_{\alpha}f[y]) \mod 124 = (n_x + n_y) \mod 124$
2. $\log_{\alpha}(f[x]/f[y]) = (\log_{\alpha}f[x] - \log_{\alpha}f[y]) \mod 124 = (n_x - n_y) \mod 124$
3. $\log_{\alpha}f[x]^m = m \log_{\alpha}f[x] \mod 124$

3.1. Transformations of the DNA Extended Sequences

Gene mutations can be considered as linear transformations of the wild type gene in the $N$-dimensional vector space of DNA sequences. These linear transformations are endomorphisms and automorphisms. In particular, there are some remarkable automorphisms. Automorphisms are one-one transformations on the group $(C_{125})^N$, such that, for all extended DNA sequences $\alpha$ and $\beta$ in $(C_{125})^N$ and $\alpha \in GF(125)$, we have:

$$f(\alpha \cdot (\alpha + \beta)) = af(\alpha) + af(\beta)$$
That is, automorphisms forecast mutation reversions, and if the molecular evolution process went by through automorphisms then, the observed current DNA sequences would not depend on the mutational pathway followed by the ancestral DNA sequences. In addition, the set of all automorphisms is a group.

For every endomorphism (or automorphism) \( f: (C_{125})^N \rightarrow (C_{125})^N \), there is a \( N \times N \) matrix:

\[
A = \begin{pmatrix}
a_{i1} & \ldots & a_{iN} \\
\vdots & \ddots & \vdots \\
a_{N1} & \ldots & a_{NN}
\end{pmatrix}
\]

where the rows are the image vectors \( f(e_i) \) and \( i = 1, 2, \ldots N \). This matrix will be called the representing matrix of the endomorphism \( f \) with respect to the canonical base \( e_i \) \( \{i = 1, 2, \ldots, N\} \).

As in [SAN 05], single point mutations can be considered local endomorphisms. An endomorphism \( f: S \rightarrow S \) will be called local endomorphism if there exists \( k \in \{1, 2, \ldots, N\} \) and \( a_{ik} \in GF(125) \) \( (i = 1, 2, \ldots, N) \) such that:

\[
f(e_i) = a_{ik}e_k + e_i, \text{ for } i \neq k,
\]

and

\[
f(e_k) = a_{kk}e_k
\]

This means that:

\[
f(x_1, x_2, \ldots, x_n) = (x_1, x_2, \ldots, \sum_{i=1}^{n} x_i a_{ik}e_k, \ldots, x_n)
\]

It is evident that a local endomorphism will be a local automorphism if, and only if, the element \( a_{kk} \) is different from zero. The local endomorphism \( f \) will be considered diagonal if \( f(e_k) = (0, \ldots, a_{kk}, \ldots, 0) = a_{kk}e_k \) and \( f(e_i) = e_i \), for \( i \neq k \). This means that:

\[
f(x_1, x_2, \ldots, x_N) = (x_1, x_2, \ldots, a_{kk}x_k, \ldots, x_N)
\]

The previous concepts allow us to present the following theorem:
Theorem 1. For every single point mutation changing the codon \( \alpha_i \) of the wild type gene \( \alpha = (\alpha_1, \alpha_2, \ldots, \alpha_i, \ldots, \alpha_N) \) (\( \alpha \) different from the null vector) by the codon \( \beta_i \) of the mutant gene \( \beta = (\alpha_1, \alpha_2, \ldots, \beta_i, \ldots, \alpha_N) \), there is:

i. At least a local endomorphism \( f \) such that \( f(\alpha) = \beta \).

ii. At least a local automorphism \( f \) such that \( f(\alpha) = \beta \).

iii. A unique diagonal automorphism \( f \) such that \( f(\alpha) = \beta \) if, and only if, the codons \( \alpha_i \) and \( \beta_i \) of the wild type and mutant genes, respectively, are different of OOO.

Proof: Since genes are included in the vector space over a Galois Field, this proves is similar to those reported in [SAN 05]. □

According to the last theorem, any point mutation presented in the vector space \( (C_{125})^N \) of all DNA alignment sequences of length \( N \) sequences are described by means of automorphisms of the corresponding wild type gene. Specifically, the most frequent mutations can be described by means of diagonal automorphisms [see SAN 05]. For example, the sequence \( \alpha=UAUAUGAGUGAC \) can be considered. Let us suppose that, with successive mutations, this sequence becomes the sequence \( \beta = UGUAUAAGUOAG \). According to Table 1 these sequences correspond in the vector space \( (C_{125})^4 \) to vectors \( \alpha = (f[24], f[108], f[84], f[42]) \) and \( \beta = (f[99], f[106], f[84], f[28]) \). Hence, according to the Theorem, there exists a diagonal endomorphism \( f \), so that \( \beta = f(\alpha) \). Our Galois field is generated by the primitive polynomial \( g(x) = 2 + 3 x^2 + x^3 \). In this field, the root \( \alpha = 2 + x + x^2 \) –corresponding to codon AAC– is a generator of the multiplicative group. Next, by means of Table 2 we can compute:

\[
(f[99], f[106], f[84], f[28]) = (f[24], f[108], f[84], f[42]) \begin{pmatrix}
1 & 0 & 0 & 0 \\
0 & f[61] & 0 & 0 \\
0 & 0 & f[1] & 0 \\
0 & 0 & 0 & f[29]
\end{pmatrix}
\]

On the other hand, mutations can be considered translations of the wild type gene in the \( N \)-dimensional vector space of the DNA extended sequences. In the Abelian group \( (C_{125}, +) \), for two extended triplets \( a, b \in (C_{125}, +) \), equation \( a + x = b \) always has a solution. Then, for all
pair of aligned sequences $\alpha, \beta \in (C_{125}, +)^N$ there is always a sequence $\kappa \in (C_{125}, +)^N$ so that $\alpha + \kappa = \beta$. That is, there exists translation $T: \alpha \to \beta$. Translation $T_k$ with constant $k$ acting on triplet $x$ will be represented as:

$$T_k(x) = x + k$$

Next, given applications: $W \xrightarrow{f} X \xrightarrow{g} Y$, the composition $g \circ f: W \to Y$ of translations $g$ and $f$ is defined by $(g \circ f)(x) = g(f(x))$. It is well known that the set of all translations with composition operations is a group $G$.

### 3.2. Stabilizer subgroup of the wild type conserved regions

It is well known that in a wild type ORF, normally, not all codon sequences are susceptible to experimental mutations. Usually, conserved variables and hypervariables regions are found in genes. A typical case is the antibody where heavy chain variable domain ($V_H$) and a light chain variable domain ($V_L$) are found. Within $V_L$ and $V_H$ there are "hot spots" of variability. These hot spots of variability were termed hypervariable regions. The hypervariable regions of the heavy and light chains together form the antigen binding site of the immunoglobulin molecule. Next, let $P$ be the subset of mutant DNA sequences conserving the same regions from a wild type DNA coding sequence $\alpha_0 \in (C_{125})^N$. Then, according to the group theory [RED 67], the set $St(\alpha_0)$ of automorphisms $f \in G$ that preserves these regions is a subgroup of $G$, that is:

$$St(\alpha_0) = \{f \in G, \text{ such that: } f(\alpha_0) = \beta \in P\} \subset G$$

This subgroup could be called the stabilizer subgroup in $G$ of the conserved regions of wild type $\alpha_0$. Notice that the stabilizer subgroup $St(\alpha_0)$ is connected with the homologous recombination that involves a reciprocal exchange of DNA sequences –e.g. between two chromosomes that carry the same genetic loci. The homologous recombination algebraically corresponds to the action of two automorphism pairs that could be included in the $St(\alpha_0)$ (see Fig. 1).
Homologous DNA duplexes pair $\alpha$ $\alpha'$ $\beta$ $\beta'$

\[
\begin{array}{c}
\alpha \\
\alpha' \\
\beta \\
\beta'
\end{array}
\quad \begin{array}{c}
f(\alpha) \\
g(\alpha') \\
f^{-1}(\beta) \\
g^{-1}(\beta')
\end{array}
\rightarrow
\begin{array}{c}
Reciprocal recombinants
\end{array}
\]

\[f, g, f^{-1}, g^{-1}\]

**Figure 1.** The homologous (generalized) recombination between two homologous DNA duplexes algebraically corresponds to the action of two automorphism pairs over two paired DNA duplexes. The two automorphism pairs express a reciprocal exchange of DNA sequences and could be included in the subgroup of automorphism $S\alpha_0$.

For instance, the pair $f$ and $f^{-1}$ acts over the homologous strands $\alpha$ and $\beta$ to produce the homologous reciprocal recombinants $f(\alpha)$ and $f^{-1}(\beta)$. Likewise, the pair $g$ and $g^{-1}$ acts over the homologous strands $\alpha'$ and $\beta'$ to produce the homologous reciprocal recombinants $g(\alpha')$ and $g^{-1}(\beta')$. As a result, two reciprocal recombinant DNA sequences are generated. In particular, if homologous recombination results in an exact exchange of genetic information, then the automorphism pairs are diagonal automorphisms. Since evolution could not happen without genetic recombination, this algebraic description is biologically relevant. If it were not possible to exchange material between (homologous) chromosomes, the content of each individual chromosome would be irretrievably fixed in its particular alleles. When mutations occurred, it would not be possible to separate favourable and unfavourable changes [LEW 04]. Hence, the study of the automorphism subgroup involved in this transformation—the homologous recombination—could reveal new rules of molecular evolution process so far unknown.

### 3.3. Finite Abelian group of DNA sequences

The multiple sequence alignment is the cornerstone of Bioinformatic. Now, some subset of DNA alignment sequences with length $N$ will be analysed. By means of multiple sequence alignments, it is possible to find in the DNA genomic sequences small subregions in which there are not introduced gaps.
Figure 2. An example of alignment group: \( S = (Z_5)^{12} \oplus (Z_2)^{15} \oplus (Z_5)^6 \oplus (Z_2)^{12} \oplus (Z_5)^6 \). Building blocks correspond to ungapped sub-sequences (power of \( Z_2 \)).

For instance, if multiple sequence alignments of open reading frames (ORF) from gene super families are observed, small blocks of ungaped aligned sequences from different genes can be detected. These small blocks of codon sequences may be called “building blocks” (see Fig 2). Theoretically, a building block will be a set of aligned sequences \( X_1, X_2, \ldots, X_N \) where \( X_i \in \{A, C, G, U\} \) \((i = 1 \ldots N)\) with some evolutionary relationship between them. The building blocks with length \( N \) can be described by vector spaces over \( GF(64) \); in particular, these can be described by the Abelian group \((C_{64})^N, +\) of the \( N \)-dimensional vector space of DNA sequences. Whereas regions with gaps can be described by means of the group \((C_{125})^N, +\).

Notice that groups \((C_{64}, +)\) and \((C_{125}, +)\) are isomorphic to the \( p \)-groups \((p \text{ a prime number})\) \((Z_2)^3\) (a \(2\)-group) and \((Z_5)^3\) (a \(5\)-group) respectively. It is well known that every Abelian group can be written as a direct sum of \( p \)-groups [DUB 63]. Actually, in the set of all alignment sequences with length \( N \), several finite Abelian groups over the subsets of all \( 2^{n_1+n_2+\ldots+n_p} \) possible alignment sequences \((N = n_1+n_2+\ldots+n_p)\) can be defined. An example of these is shown in Fig 1. These groups may be called “alignment groups”.

If a finite group \( G \) is written as a direct sum \( G = G_1 \oplus G_2 \oplus \ldots \oplus G_s \), then endomorphism ring \( \text{End}(G) \) is isomorphic to the ring of matrices \((A_{ij})\), where \( A_{ij} \in \text{Hom}(G_i, G_j) \), with the usual matrix operations. In our case the endomorphism that transform the DNA alignment sequence \( \alpha \) into \( \beta \) \((\alpha, \beta \in G)\) is represented by a matrix with only non-cero elements in the principal diagonal. These diagonal elements are sub-matrices \( A_{ii} \in \text{End}(G_i) \) (0 \( A_{ii} \in \text{Aut}(G_i) \)).
Finally, since the canonical decomposition of an Abelian group $G$ into $p$-groups is unique up to isomorphism [DUB 63], it is possible to characterize alignment groups for a fixed sequence length $N$. That is to say, two alignment groups can have different $p$-group decompositions and simultaneously be isomorphic by holding the same canonical decomposition into $p$-groups. This algebraic description biologically suggests that the same biological architectural principium underlies the alignment groups with the same canonical decomposition into $p$-groups. Here the basic construction materials come from building blocks. It could also correspond to the fact that in the molecular evolution process, the new genetic information frequently comes into being from the rearrangements of existing genetic material in the chromosomes.

4. Conclusions

In this paper, the starting point to analyze deletions and mutations in DNA sequences is the extended triplet set with elements $X_1X_2X_3$, where $X_i \in \{O, A, C, G, U\}$. Taking into account the order in the set $\{O, A, C, G, U\}$ and the biological importance of base positions in the codon, it is possible to establish a bijection between the extended triplet set and the Galois field $GF(5^3)$. This bijection allows us to define the Galois field of the extended triplet set. Over this new field, a new $N$-dimensional vector space is defined in the set of all possible DNA alignment sequences where gene mutations can be considered linear transformations or translations of the wild type gene.

For every single point mutation in the wild type gene, there is at least an automorphism that transforms the wild type in the mutant gene. So, automorphism-group could be a useful tool to study the mutational pathway followed by genes in the $N$-dimensional vector space of all possible DNA alignment sequences.

Besides this, the set $St(\alpha_0)$ of automorphisms that conserve the same regions from a wild type DNA coding sequence $\alpha_0 \in (C_{125})^N$ is a subgroup connected with the homologous recombination that involves a reciprocal exchange of DNA sequences — e.g. between two chromosomes that carry the same genetic loci. The homologous recombination algebraically corresponds to the action of two automorphism pairs that could be included in the $St(\alpha_0)$.
By means of multiple sequence alignments, it is possible to define several finite Abelian groups –alignment groups– over the subsets of all $2^{n_1+m_1+\ldots+n_p+m_p} \cdot \ldots \cdot 2^{n_1+m_1+\ldots+n_n+m_n}$ possible alignment sequences ($N = n_1+\ldots+n_p+m_1+\ldots+m_n$). Two alignment groups can have different $p$-group decompositions and simultaneously be isomorphic holding the same canonical decomposition into $p$-groups. For alignment groups with the same canonical $p$-group, decompositions could underlie the same biological architectural principium.

This algebraic structure of DNA sequences over the extended Galois field of the genetic code is an attempt to formalize the biological interpretation of the multiple sequence alignment results, the corner stone of Bioinformatic. Our first approach leads us to the concept of building blocks in DNA sequences that could be a key to help us understand how DNA genomic sequences are assembled.

**Acknowledgments**

This research was supported within the framework of a VLIR-IUS Collaboration Programme.

**References**


