Geometric Crossover for Biological Sequences

Alberto Moraglio, Riccardo Poli and Rolv Seehuus

No Institute Given

Abstract. This paper extends a geometric framework for interpreting crossover and mutation [4] to the case of sequences. This representation is important because it is the link between artificial evolution and biological evolution. We define and theoretically study geometric crossover for sequences under edit distance and show its intimate connection with the biological notion of sequence homology.

1 Introduction

Evolutionary algorithms (EAs) mimic, in a simplified manner, natural evolution. However, very few theoretical results are available which apply equally well to both forms of evolutionary search.

One important cause of the lack of connection between evolutionary computation theory and evolutionary biology is that they focus on different kinds of genotypes (different solution representations), namely DNA strands (variable-length strings or sequences) and binary strings. Most importantly, even if DNA strands and binary strings appear to be very similar at a first sight, the crossover operator for binary strings is just a caricature of the biological recombination acting on DNA strands. The main difference is that DNA strands align on the basis of their contents (at meiosis) before exchanging genetic material and do not align only positionally as it is the case for binary strings. Such an alignment is flexible in that two DNA strands can stretch and fold to better align with each other. Moreover, DNA strands do not need to be aligned on the extremities. After alignment, the two DNA strands cut in one or more regions in which they match well and exchange DNA segments. This last phase is present in crossovers for EAs, in which, however, typically no alignment process based on content takes place.

Geometric crossover and geometric mutation [4] are representation-independent search operators that generalise by abstraction many pre-existing search operators for the major representations used in EAs, such as binary strings, real vectors, permutations and syntactic trees. They are defined in geometric terms using the notions of line segment and ball. These notions and the corresponding genetic operators are well-defined once a notion of distance in the search space is well-defined. This way of defining search operators as function of the search space is opposite to the standard way [5] in which the search space is seen as a function of the search operators employed. This viewpoint greatly simplifies the relationship between search operators and fitness landscape and allows different search operators to share the same search space thereby clarifying their roles.

Is biological recombination geometric? In this paper we are able to answer this question in the affirmative by extending the geometric framework mentioned above to sequences under edit distance. This has the remarkable consequence that the theory of geometric crossover applies to biological crossover as well, bridging the gap between biological evolution and artificial evolution. Our results reveal a deep connection between crossover for binary strings and biological recombination, showing that standard EA crossover is less of a caricature than it appears at first.

The paper is organised as follows. In section 2, we introduce the geometric framework. In section 3, we show that in the case of sequences endowed with edit distances geometric crossover is a form of homologous crossover which performs the alignment on sequence contents before mixing genetic material. We prove various properties of this crossover and, in section 4, extend it to weighted alignments and alignment with gaps. In section 5 we argue that biological recombination is geometric and discuss the consequences of this.

2 Geometric framework

2.1 Geometric preliminaries

In the following we give necessary preliminary geometric definitions and extend those introduced in [4] and [2]. The following definitions are taken from [6].

The terms *distance* and *metric* denote any real valued function that conforms to the axioms of identity, symmetry and triangular inequality. A simple connected graph is naturally associated to a metric space via its *path metric*: the distance between two nodes in the graph is the length of a shortest path between the nodes. Similarly, an edge-weighted graph with strictly positive weights is naturally associated to a metric space via a *weighted path metric*.

In a metric space (S, d) a *closed ball* is the set of the form $B(x; r) = \{y \in S | d(x, y) \leq r\}$ where $x \in S$ and r is a positive real number called the radius of the ball. A *line segment* (or closed interval) is the set of the form $[x; y] = \{z \in S | d(x, z) + d(z, y) = d(x, y)\}$ where $x, y \in S$ are called extremes of the segment. Metric ball and metric segment generalise the familiar notions of ball and segment in the Euclidean space to any metric space through distance redefinition. These generalised objects look quite different under different metrics. Notice that a metric segment does not coincide to a shortest path connecting its extremes (*geodesic*) as in an Euclidean space. In general, there may be more than one geodesic connecting two extremes; the metric segment is the union of all geodesics.

We assign a structure to the solution set by endowing it with a notion of distance d. $M = (S, d)$ is therefore a solution *space* and $L = (M, g)$ is the corresponding fitness landscape. Notice that d is arbitrary and need not have any particular connection or affinity with the search problem at hand.

2.2 Geometric crossover definition

The following definitions are *representation-independent* therefore crossover is well-defined for any representation. It is only *function of the metric d* associated with the search space being based on the notion of metric segment.

Definition 1. *(Image set) The* image set $Im[OP]$ *of a genetic operator* OP *is the set of all possible offspring produced by* OP *with non-zero probability.*

Definition 2. *(Geometric crossover) A binary operator is a geometric crossover under the metric* d *if all offspring are in the segment between its parents.*

Definition 3. *(Uniform geometric crossover) Uniform geometric crossover* UX *is a geometric crossover where all* z *laying between parents* x *and* y *have the same probability of being the offspring:*

$$f_{UX}(z|x,y) = \frac{\delta(z \in [x;y])}{|[x;y]|}$$

$$Im[UX(x,y)] = \{z \in S | f_{UX}(z|x,y) > 0\} = [x;y].$$

A number of general properties for geometric crossover and mutation have been derived in [4].

2.3 Geometric crossover landscape

Geometric operators are defined as functions of the distance associated to the search space. However, the search space does not come with the problem itself. The problem consists only of a fitness function to optimise, that defines what a solution is and how to evaluate it, but it does not give any structure on the solution set. The act of putting a structure over the solution set is part of the search algorithm design and it is a designer's choice. A fitness landscape is the fitness function plus a structure over the solution space. So, for each problem, there is one fitness function but as many fitness landscapes as the number of possible different structures over the solution set. In principle, the designer could choose the structure to assign to the solution set completely independently from the problem at hand. However, because the search operators are defined over such a structure, doing so would make them decoupled from the problem at hand, hence turning the search into something very close to random search.

In order to avoid this one can exploit problem knowledge in the search. This can be achieved by carefully designing the connectivity structure of the fitness landscape. For example, one can study the objective function of the problem and select a neighbourhood structure that couples then distance between solutions and their fitness values. Once this is done problem knowledge can be exploited by search operators to perform better than random search, even if the search operators are problem-independent (as in the case of geometric crossover and mutation).

Under which conditions is a landscape well-searchable by geometric operators? As a rule of thumb, geometric mutation and geometric crossover work well on landscapes where the closer pairs of solutions, the more correlated their fitness values. Of course this is no surprise: the importance of landscape smoothness has been advocated in many different context and has been confirmed in uncountable empirical studies with many neighbourhood search meta-heuristics [7].

3 Geometric crossover for sequences

In this section, we extend the geometric framework to the case of sequences. In particular we will focus on edit distances that associate with sequence homology.

3.1 Preliminaries: sequences, edit distance and alignments

A sequence is a variable length string of characters. In particular, DNA strands are sequences of characters from the alphabet $\Sigma_{dna} = \{a, c, t, g\}$. The *edit distance* between two sequences is defined as the minimum number of edit operations – insertions, deletions, and substitutions – needed to transform the first string into the second. The edit distance is a metric in that it respects all the metric axioms. Hence, the space of sequences endowed with edit distance is a metric space. There are a number of extensions to the simple edit distance such as weighted edit distance, block-edit distance, reversals and transpositions distances (see Sections 4 and 5 for a discussion on their use). The edit distance between two sequences is a measure of their syntactic dissimilarity. This syntactic dissimilarity is intimately connected with the notion of sequence alignment.

An *alignment* of two sequences is obtained by first appropriately inserting spaces (which we represent with dashes), either into or at the ends of the two sequences, and then placing the two resulting sequences one above the other so that every character or space in one sequence is aligned with a character or space in the other sequence. The *score of an alignment* is the number of aligned characters that are different in the two sequences. There may be more that one optimum alignment between two sequences. The score of an optimum alignment of two sequences equals their edit distance. Changing the scoring system, one can obtain optimal alignments associated to weighted edit distances and block-edit distances. Edit distances and optimal alignments can be computed efficiently using dynamic programming.

The (edit) *transcript* T associated to an alignment q is a vector that specifies what edit move to apply to parent 1 to reach parent 2 for each position. For each alignment q there is only one transcript T and vice versa. For example, $T = (\texttt{RIMDMDMMI})$ and $q = \begin{pmatrix} \texttt{v-intner-} \\ \texttt{wri-t-ers} \end{pmatrix}$, where R, I and D stand for replace, insert, delete and match, respectively, while M is a just place holder.

3.2 Homologous Crossover and Geometric Crossover

Homologous crossover for sequences has been introduced by [8] in the context of linear GP. We formalise and generalise it, we prove that it is geometric crossover and then list some of its properties.

Definition 4. *(Alignment-based homologous crossover operators)*

1. *Let Q be the set of all optimal alignments of two sequences S_1 and S_2 under simple edit distance. Homologous crossover picks a random optimal alignment $q \in Q$ with a given probability distribution over Q. Let $\overline{S}_1$ and $\overline{S}_2$ be the two sequences aligned with gaps according to q.*
2. *Let l be the length of q and m be a mask drawn from $\{0,1\}$ with a given probability distribution. m specifies for each position of q from which parent to copy the corresponding character to produce an aligned offspring $\overline{S}_3$*
3. *The actual offspring S_3 is obtained by remove the dashes from $\overline{S}_3$.*

Example 1. If $S_1 = \texttt{agcacaca}$ and $S_2 = \texttt{acacacta}$ and the chosen optimal alignment is $q = \begin{pmatrix} \texttt{agcacac-a} \\ \texttt{a-cacacta} \end{pmatrix}$ then $l = 9$, $\overline{S}_1 = \texttt{agcacac-a}$ and $\overline{S}_2 = \texttt{a-cacacta}$. If $m = \texttt{111100000}$ we obtain the offspring $\overline{S}_3 = \texttt{a-cacac-a}$. After gap removal we obtain $S_3 = \texttt{acacaca}$.

Theorem 1. *All alignment-based homologous crossover operators are geometric crossovers under edit distance.*

Proof. An optimal edit transcript T contains a smallest set E of edit moves to transform u in v. $|E| = d(u,v)$. The edit moves in E are independent because they can be applied in any order and transform u into v. Any intermediate sequence z obtained by applying a subset $E' \subseteq E$ of edit moves to u is on a shortest path between u and v because z is $d(u,z) = |E'|$ moves away from u and $d(z,v) = |E \setminus E'|$ moves to v hence $d(u,z) + d(z,v) = d(u,v)$. A mask m selects a subset of edit moves $E_m \subseteq E$ from the transcript T to apply to u and produce the offspring z. Hence z is on the shortest path.

Theorem 2. *Every sequence O in the segment between two sequences $P1$ and $P2$ under edit distance is reachable by homologous alignment-based crossover applied to the parent sequences $P1$ and $P2$.*

Proof. We need to prove that for each $O \in [P_1, P_2]_e d$ there exists an optimal alignment q of $P1$ and $P2$ and a mask m that applied to a gives O. We prove it by constructing q and m given any O.

If $O \in [P_1, P_2]_e d$ then there exists a shortest path sp between $P1$ and $P2$ in the search space of sequences endowed with the edit distance such that $O \in sp$. Then there exists a transcript T such as all the edit moves in T are the same of the set of edit moves that generate sp. The transcript T may comprise also one or more M characters that do not correspond to any edit move. The transcript T is optimal by construction because the number of edit moves in T (non-M characters) is exactly $ed(P1, P2)$.

Given T, $P1$ and $P2$, it is possible to build the unique alignment q of $P1$ and $P2$ associated with T. The alignment q is optimal because T is optimal. Consider now the crossover mask m of the same length of the transcript T obtained by setting at 1 the loci corresponding to those edit moves in the transcript T that in the path sp transform $P1$ into O. The crossover mask m applied to the optimal alignment q produces O.

Theorems 1 and 2 establish that a crossover is an alignment-based homologous crossover if and only if it is a geometric crossover under simple edit distance.

3.3 Optimal alignments and segment subsets

The family of crossovers introduced in the previous section can be seen as an extension to sequences of the family of alignment-based crossovers for fixed-length binary strings. [4] proved that for binary strings, uniform crossover, where crossover masks are obtained by flipping n times a unbiased coin, picks offspring with uniform probability distribution on the line segment between parents under Hamming distance. In this section we introduce a generalisation of uniform crossover based on masks for sequences and show that, unlike the binary string case, this crossover, surprisingly, does not pick offspring uniformly in the segment between parents under edit distances.

Definition 5. *(Uniform alignment-based homologous crossover) Uniform homologous crossover is an alignment-based crossover operator that chooses optimal alignments and crossover masks with uniform probability.*

In Table 1, we enumerate all possible offspring under homologous crossover of the sequences “vint” and “writ”. For these sequences there are three possible optimal alignments. The edit distance between the sequences is 3. This can be seen also from the edit transcript associated to each optimal alignment in which there are 3 non-M characters. These characters describe the edit operations and the location of their application on the alignment to transform the first sequence into the second one. In the first column, all the possible crossover masks are shown. For space limitations we report only the bits corresponding to the three non-`M` symbols, thereby obtaining 8 effective crossover masks. The entry at the intersection of a row (effective crossover mask) and a column (optimal alignment) contains the offspring obtained by the application of the mask on the alignment. Alignment-based uniform crossover returns any of the offspring in the table at random with uniform probability ($\frac{1}{24}$). However, some offspring can be generated by more than one alignment, and so they have higher chances to be picked. “vint” and “writ”, for example, are produced with a probability $\frac{3}{24}$, while “vit”, “wrint”, “vrit” and “wint” are returned with probability $\frac{2}{24}$.

The image set of an optimal alignment q is the set of offspring that can be generated by homologous crossover using any mask m over q.

Theorem 3. *Consider the image sets $Im(q_1) \dots Im(q_n)$ of homologous crossover applied to all optimal alignment $q_1 \dots q_n$ of the sequences $P1$ and $P2$. The union of $Im(q_1) \dots Im(q_n)$ is $[P1, P2]$ but they do not form a partition of $[P1, P2]$.*

Proof. For theorem 1, the image set of any optimal alignment is subset of the segment. For theorem 2, any sequence z in the segment $[P1, P2]$ can be generated by homologous crossover. Hence, there must exist at least an alignment such as its image set includes z. This means that every point in the segment is at least in $Im(q_i)$, hence the union of all $Im(q_i)$ is the segment $[P1, P2]$. Proof by counterexample: example 2 shows that all $Im(q_i)$ do not form a partition of the segment $[P1, P2]$ because their intersections are non-empty.

Table 1. Possible offspring under uniform alignment-based homologous crossover.

	Alignment 1	Alignment 2	Alignment 3
mask	`mm*m*`	`mm*m*`	`mmm*`
transcript	`IRMDM`	`RIMDM`	`RRRM`
parent 1	`-vint`	`v-int`	`vint`
parent 2	`wri-t`	`wri-t`	`writ`
000	`-vint`	`v-int`	`vint`
001	`-vi-t`	`v-i-t`	`viit`
010	`-rint`	`vrint`	`vrnt`
011	`-ri-t`	`vri-t`	`vrit`
100	`wvint`	`w-int`	`wint`
101	`wvi-t`	`w-i-t`	`wiit`
110	`wrint`	`wrint`	`wrnt`
111	`wri-t`	`wri-t`	`writ`

Theorem 4. *Uniform alignment-based homologous crossover is not the uniform geometric crossover under edit distance.*

Proof. Proof by counterexample: example 2 shows that the frequency of some offspring sequences under uniform homologous crossover is higher than others. So the probability is not uniformly distributed over the segment.

The non-uniformity of this crossover is the result of the same offspring sequence being generated by multiple different optimal alignments. Parent sequences, for example, are in this category because they can be generated by all optimal alignments using masks 0...0 and 1...1. Other offspring sequences can be generated more than once when two optimal transcripts share non-`M` characters at the same positions. For example, if two transcripts have a `D` at position 1, then the mask 0X...X where X...X is either 0...0 or 1...1 will produce the same offspring with both alignments. The mask 1X...X will have the same effect.

3.4 Bounds on offspring size

In this section we explore how offspring and parent sizes are related in homologous crossover.

Theorem 5. *Given two parent sequences $P1$ and $P2$ of length l_1 and l_2 with $l_1 \leq l_2$ and edit distance ed, the length l_3 of any offspring sequence O obtained by homologous recombination is bounded as follows:*

1. *Edit distance ed known: $(l_1 + l_2 - ed)/2 \leq l_3 \leq (l_1 + l_2 + ed)/2$*
2. *Edit distance ed not known: $l_1/2 \leq l_3 \leq l_1/2 + l_2$*
3. *Parents of same length $l_1 = l_2 = l$: $l/2 \leq l_3 \leq 3l/2$*
4. *Non-empty parents imply non-empty offspring*

Proof. Trivial edit distance bounds: (i) $d(a,b) \geq |l(a) - l(b)|$ and (ii) $d(a,b) \leq max(l(a), l(b))$. From bound (i) applied to $P1$ and $P3$: $d(P1, P3) \geq |l1 - l3|$ that breaks into two cases: (1) $l_1 - l_3 \leq 0 \rightarrow l_1 \leq l_3 \leq d(P1, P3) + l_1$ (worst case upper bound) (2) $l_1 - l_3 \geq 0 \rightarrow l_1 - d(P1, P3) \leq l_3 \leq l_1$ (worst case lower bound). Analogously, applying bound (i) to P2 and P3 we obtain other two alternative cases: (3) $l_2 - l_3 \leq 0 \rightarrow l_2 \leq l_3 \leq d(P2, P3) + l_2$ (worst case upper bound) (4) $l_2 - l_3 \geq 0 \rightarrow l_2 - d(P2, P3) \leq l_3 \leq l_2$ (worst case lower bound).

Let us consider the upper bound for l_3. Both the conditions (1) and (3) must hold true, so $2l_3 \leq d(P1, P3) + d(P2, P3) + l_1 + l_2$. For all $P3$: $d(P1, P3) + d(P2, P3) = d(P1, P2) = ed$. Hence for all $P3$: $l_3 \leq (l_1 + l_2 + ed)/2$. If the distance ed between parents $P1$ and $P2$ is unknown we can use bound (ii) to bound it: $ed \leq max(l_1, l_2) \rightarrow ed \leq l_2$. Hence for all $P3$ in the worst case we have: $l_3 \leq l_1/2 + l_2$. In case $l_1 = l_2 = l$ we have for all $P3$: $l_3 \leq 3l/2$.

Let us consider the lower bound for l_3. Both the conditions (2) and (4) must hold true, so $l_1 + l_2 - (d(P1, P3) + d(P2, P3)) \leq 2l_3$. For all $P3$: $d(P1, P3) + d(P2, P3) = d(P1, P2) = ed$. Hence for all $P3$: $(l_1 + l_2 - ed)/2 \leq l_3$. If the distance ed between parents $P1$ and $P2$ is unknown we can use bound (ii) to bound it: $ed \leq max(l_1, l_2) \rightarrow ed \leq l_2$. Hence for all $P3$ in the worst case we have: $l_1/2 \leq l_3$. In case $l_1 = l_2 = l$ we have for all $P3$: $l/2 \leq l_3$.

Homologous crossover cannot produce empty offspring from non-empty parents. This can be shown by using the second inequality: $l_1/2 \leq l_3 \leq l_1/2 + l_2$. Independently from the distance between parents the minimum lower bound of the length of any offspring is half of the length of the shortest parent. When such parent is not empty ($l_1 \geq 1$) then $l_3 \geq 1/2$. Since the length is an integer we have $l_3 \geq 1$. So even for parents of length 1 the offspring are non-empty.

Under geometric crossover, the more different the parents are, the more "unrelated", or "innovative", the offspring become. From the previous theorem, the size of the offspring is bounded by: $(l_1 + l_2 - ed)/2 \leq l_3 \leq (l_1 + l_2 + ed)/2$. Hence, the bigger the difference between the parents the bigger the range of the size of possible offspring. Note, however, that when using weighted edit distances it is possible to create situations were an empty offspring can be returned.

4 Extensions of Homologous Crossover

4.1 Weighted Edit Distances and Geometric Crossover

Extending homologous crossover to the case of weighted edit distances is crucial to capture more realistic details of real biological sequences. Weighted edit distances allow to specify relative preferences in the alignment before recombination such as character mismatches vs. sequence interruptions (spaces), positional preferences (for example, matches at the extremities vs. matches at the centre of the sequences) or preferences on the mismatching pairs (for example, preferring a mismatch (a, t) to a mismatch (a, c)).

The following theorem is a very general and useful result that connects weighted edit moves for any solution representation and metric spaces.[1]

[1] This is a fairly simple result. However, it appears that this is not been proved in published literature, leading to significant confusion, particularly in the bio-informatics literature, in which edit distances and scoring matrices are extensively used.

Theorem 6. *Any weighted edit distance with strictly positive weights on edit moves is a metric.*

Proof. A space of configurations endowed with an edit distance with strictly positive weights can be represented by a weighted graph in which nodes are syntactic configurations and weighted edges represent (reversible) weighted edit moves transforming one configuration into neighbour configuration. Any graph with strictly positive weights on edges is a metric space [6] hence an edit distance with strictly positive weights on edit moves, that is isomorphic to such a graph, is a metric.

The cost of a weighted alignment is the sum of the weights associated to each character alignment. The weight of each couple of characters is symmetric and matching characters have weight 0. An optimal alignment is an alignment with minimal cost. The cost of the optimal weighted alignment between two sequences equals their weighted edit distance where the edit moves allowed correspond to the set of couple of characters corresponding with their alignment weights.

The following theorem extend the geometricity result of homologous crossover to weighted edit distances and weighted alignments.

Theorem 7. *Alignment-based homologous crossover on the optimal alignments under weighted edit distance d_w is geometric crossover under d_w.*

Proof. An optimal edit transcript T contains a set E of edit moves to transform u in v whose cost $w(E) = \sum_{e \in E} w_e$ is minimal. The weighted edit distance is $d_w(u,v) = w(E)$. The edit moves in E are independent because they can be applied in any order and transform u into v. Any intermediate sequence z obtained by applying a subset $E' \subseteq E$ of edit moves to u is on a shortest weighted path between u and v because $d_w(u,z) = w(E')$ and $d(z,v) = w(E \setminus E')$=w(E)-w(E') hence $d(u,z) + d(z,v) = d(u,v)$. A mask m selects a subset of edit moves $E_m \subseteq E$ from the transcript T to apply to u and produce the offspring z. Hence z is on the shortest path.

4.2 Gaps and Geometric Crossover

In this section we extend homologous crossover to the case of edit distances based on replacement move and a block ins/del move. This edit distance allows to specify preference to few big gaps against many small gaps in the alignment before recombination and allows to model loops in the alignments.

Theorem 8. *Alignment-based homologous crossover with one locus for each entire gap on the optimal alignments under weighted edit distance with block moves d_{bw} is geometric crossover under d_{bw} with convex weight gap model.*

Proof. Let us consider a weighted block ins/del edit move such as its weights depends only on the length of the block in a way that shorter blocks have smaller cost per length unit: $l_1 < l_2 \rightarrow w(l_1)/l_1 > w(l_2)/l_2$. An optimal edit transcript must necessarily comprise the largest block ins/del edit move. The crossover mask has to treat each edit move as a unity: for block edit moves there must be only one locus in the crossover mask. The rest follows from theorem 7.

5 Bridging natural and artificial evolution

In this section we discuss the feasibility of homologous crossover as a model of biological recombination and its implications.

Is biological recombination geometric? Most of pre-existing recombination operators for the most-used representations are geometric. So this geometric property unifies by abstraction across representations the notion of "crossoverness" emerged experimentally over the year. The importance of being geometric for a crossover relies in the fact that all geometric operators do the same type of search (convex search). Plus, the connection between geometric crossover and fitness landscape is very intuitive. This question if answered affirmatively would show a deep unity in the way EAs and biological evolution do the search and would allow to apply the geometric framework to study both natural and artificial evolution jointly casting a computational and geometric perspective on natural evolution.

All the details of real biological recombination are unknown and it is focus of active research to elicit them. There are various models for studying different aspects of biological evolution at different levels of granularity.

At genetic level, the model of homologous recombination based on fixed-size strings used in population genetics, is a simple extension of the traditional crossover for binary strings to the multi-valued case and it is geometric under Hamming distance. Unequal crossover at a genetic level happens when the homologous alignment of the strands is not perfect. This can be due to an error in the alignment due to environmental noise (this can be considered as a mutation) or being one of the possible best inexact alignments under edit distance at level of genes. In this case unequal crossover would be geometric.

The reason why strands tend to align according to the edit distance can be understood at a molecular level. Our working hypothesis is that an edit distance, weighted and based on edit moves such as insertion/deletion (to model frame-shift), replacement (to model base mismatch), block-insertion/deletion (to model folds/loops), block-reversal (to model subsequence inversion) and block-transposition (to model subsequence transposition), is expressive enough to model the resulting configuration obtained at the equilibrium of all the forces that lead to the inexact homologous alignment of two chromosomes at a molecular level (before crossing over). The notion of minimum distance connects naturally with the notion of optimal alignment (best trade-off among all forces involved, or chemical equilibrium) of two macromolecules (chromosomes) that as any other chemical reaction tends to evolve toward the state of "minimum free energy". In summary:

1. the geometric crossovers associated with edit distances naturally capture the notion of homology, or inexact alignment based on the sequences contents
2. there is a natural parallel between weighted edit distances and DNA pairing up at a molecular level because the weighs on edit moves can be interpreted in chemical terms as attraction and repulsion forces

3. there are a variety of edit distances that allow to show that pre-existing model of biological crossovers and many variants are still geometric. This shows that assuming that biological recombination is geometric is a realistic assumption even in the lack of full-knowledge about all its details

Is the natural landscape smooth? The natural adaptive landscape of a population is not static but changes over time in response to environmental changes and in response to the change in the population composition adapting to the new environment due to the evolutionary forces. Evolution (adaptation) happens when the adaptive landscape becomes non-flat due to a fitness change in response to a change in the environment.

Despite the inherent fluidity of the natural adaptive landscape, it has a smooth trend: most of the mutations are neutral (Kimura), do not affect the phenotype or quasi-neutral in that affect the phenotype marginally and so its fitness. Very rarely a single mutation is lethal, creating "cracks" in the landscape. The landscape may be rugged and may present various neutral paths but the overall trend is smooth and when evolution (adaptation) in progress, non-flat. Hence, we can safely state that closer genotypes under edit distance (mutation) have more correlated fitness values. Indeed, this is the same principle on which bio-informatics is firmly based upon: similarity of genotypes allows to infer similarity in the phenotype (hence, in fitness) without doing any experimental work just by searching databases of known genotypes by homology [1].

Geometric biological operator + smooth natural landscape = quick adaptation What is the fitness function to optimise in case of natural evolution? Natural evolution, seen as a search algorithm, is trying to optimise the fitness function that is obtained from the adaptive landscape by removing the space structure (see section 2). While doing this optimization, the fitness function is constantly changing, because the adaptive landscape is constantly changing under the effect of population change due to evolution (optimization) itself. The evolution (optimization) ends[2] when the fitness landscape reaches a flat-shape and the fitness function becomes constant. This means that the population is completely adapted to the environment. Hence, the performance of biological evolution, seen as a search algorithm, is in terms of speed of adaptation.

Since we have seen that biological recombination is geometric under edit distance and that smoothness of the landscape is the condition we need to enforce to the landscape to be well-searched by geometric crossover and geometric mutation, we conclude that the biological recombination and mutation are well-matched with the natural fitness landscape. So their performance in terms of adaptation is expected to be much better than pure random search. This is to say that biological evolution is very efficient at doing adaptation.[3]

[2] evolution may also never end because of red-queen dynamics

[3] This offers an answer to the anti-evolutionist William Dembski that has used the no free lunch theorems to criticise the theory of evolution, stating that the No Free Lunch theorems demonstrate that evolution is no better than random chance at selecting optimal outcomes.

Why are biological recombination and natural landscape well-matched? Indeed, it could have been the case that fitness landscape and search operators were unmatched making adaptation non-efficient. But, so is not. How happened? We leave this as an open question and will investigate it in future work.

6 Conclusions

In this paper we have extended the geometric framework to the important case of sequences. We have given a number of theoretical results and started investigating the hypothesis that biological recombination is geometric and discussed its consequences.

References

1. Dan Gusfield. *Algorithms on Strings, Trees and Sequences.* Cambridge University Press, 1997.
2. Alberto Moraglio and Richardo Poli. Geometric landscape of homologous crossover for syntactic trees. In *Proceedings of CEC 2005.*
3. Alberto Moraglio and Richardo Poli. Topological crossover for the permutation representation. In *Proceedings of GECCO 2005.*
4. Alberto Moraglio and Richardo Poli. Topological interpretation of crossover. In *Proceedings of GECCO 2004.*
5. Terry Jones. Evolutionary Algorithms, Fitness Landscapes and Search. In *PhD dissertation, University of New Mexico, 1995..*
6. Deza and Laurent. Geometry of cuts and metrics. In *Springer, 1991..*
7. P. M. Pardalos, M. G. C. Resende. Handbook of Applied Optimization. In *Oxford University Press, 2002..*
8. Michael Defoin Platel, Manuel Clergue, and Philippe Collard. Maximum homologous crossover for linear genetic programming. In *Genetic Programming: 6th European Conference*, Lecture Notes in Computer Science, pages 194–203. Springer-Verlag GmbH, 2003.