Beyond Fixed-Resolution Alignment-Free Measures for Mammalian Enhancers Sequence Comparison

Matteo Comin and Davide Verzotto

Abstract—The cell-type diversity is to a large degree driven by transcription regulation, i.e., enhancers. It has been recently shown that in high-level eukaryotes enhancers rarely work alone, instead they collaborate by forming clusters of *cis*-regulatory modules (CRMs). Even if the binding of transcription factors is sequence-specific, the identification of functionally similar enhancers is very difficult. A similarity measure to detect related regulatory sequences is crucial to understand functional correlation between two enhancers. This will allow large-scale analyses, clustering and genome-wide classifications. In this paper we present $Under_2$, a parameter-free alignment-free statistic based on variable-length words. As opposed to traditional alignment-free methods, which are based on fixed-length patterns or, in other words, tied to a fixed resolution, our statistic is built upon variable-length words, and thus multiple resolutions are allowed. This will capture the great variability of lengths of CRMs. We evaluate several alignment-free statistics on simulated data and real ChIP-seq sequences. The new statistic is highly successful in discriminating functionally related enhancers and, in almost all experiments, it outperforms fixed-resolution methods. Finally, experiments on mouse enhancers show that $Under_2$ can separate enhancers active in different tissues. Availability: http://www.dei.unipd.it/~ciompin/main/UnderIICRMS.html

Index Terms—Alignment-free statistics, pattern discovery, regulatory sequences comparison

1 Introduction

One of the fundamental questions in bioinformatics is how to measure the similarity between biological sequences. When dealing with protein sequences or coding genes, this is probably one of the most studied problems, as it relates to the identification of homologous sequences. The use of tools like BLAST [1] to assess the degree of similarity between two sequences is nowadays a standard procedure.

In this paper we focus on the same question, but for regulatory sequences like promoters or enhancers of genes. The similarity between coding sequences has been widely used to estimate functional correlations. For regulatory sequences, it is a general belief that similar binding site contents are expected to drive similar expression patterns.

With the advent of ChIP-seq technologies, large collections of regulatory sequences are now available. One of the most important steps in the analysis of ChIP-seq data is the identification of enhancing sequences that regulate the same cell-type.

There are cases where traditional alignment based methods cannot be applied, for example, when the sequences being compared do not share any statistical significant alignment. This is the case when the sequences come from distant related organisms, or they are functionally related but not orthologous.

Moreover transcription factors binding sites often occur in clusters, also called *cis*-regulatory modules (CRMs). These modules play a key role in the regulation of the transcription process in *human* [28] as well as in *Drosophila* [15]. In addition, the position and orientation of binding sites in CRMs sharing the same cell-specific function may vary, making an alignment of them often impossible.

Lastly, enhancers in high-level eukaryotes rarely work alone; instead, they usually collaborate by forming closely located CRM clusters. It has been recently shown that during embryonic development every gene is regulated on average by three different transcription factors [7]. However, the presence of different enhancer clusters has not been already fully explored by current motif-finding tools. Moreover, the presence of multiple binding sites can make the localization of each enhancer very difficult. For these reasons biologists need first to screen ChIP-seq data sets to select cell-specific regulatory sequences, which are based on "common" contents.

The use of alignment-free methods for comparing sequences has been proved useful for a variety of different tasks. See Vinga and Almeida for a comprehensive review [26]. The idea to describe a sequence by its word content fits very well the model of CRMs, where we assume that a similar function is driven by the binding site content of different enhancers. Several alignment-free methods have been devised for this problem [14], [18], [20].

Almost all alignment-free method are based on statistics of words with a fixed-length k. The problem with this methods is that the performance depends dramatically on the choice of the resolution k [22]. For example in the analysis

- *M. Comin is with the Department of Information Engineering, University of Padova, via Gradenigo 6/A, 35131 Padova, Italy. E-mail: comin@dei.unipd.it.*
- *D. Verzotto is with the Department of Computational and Systems Biology, Genome Institute of Singapore, Singapore 138672. E-mail: verzottod@gis.a-star.edu.sg.*

Manuscript received 28 Dec. 2013; accepted 24 Jan. 2014. Date of publication 18 Feb. 2014; date of current version 4 Aug. 2014.
For information on obtaining reprints of this article, please send e-mail to: reprints@ieee.org, and reference the Digital Object Identifier below.
Digital Object Identifier no. 10.1109/TCBB.2014.2306830

of enhancers using simulated data [14], [20], the best performing k is usually equal to the length of the implanted enhancer. In real cases, where it is not possible to know the enhancer length in advance, the choice of k is critical. Moreover, in the presence of several CRMs, it is simply not feasible to select the k that best fits enhancers of different lengths. For these reasons, in this paper we present a parameter-free alignment-free method, called $Under_2$, based on variable-length words. We will define a similarity measure using variable-length words along with their statistical and syntactical properties, so that "uninformative" words will be discarded.

Another contribution is the definition of a more realistic variant of the patter transfer model, introduced in [21] to construct synthetic regulatory sequences and mimic horizontal gene transfer. We thus define the pattern transfer model revised, in which the exchange of genetic material includes reverse, complement, and reverse-complement patterns, as well as variable-length regions.

The paper is organized as follows. In Section 2 we review alignment-free methods and their applications. In Section 3 we present our contributions, the $Under_2$ statistic and the pattern transfer model revised. We test the performance of several alignment-free measures in both synthetic and real regulatory sequences in Section 4. Conclusions and future work are discussed in Section 5.

2 Previous Work

Historically, one of the first papers that introduces an alignment-free method is due to Blaisdell in 1986 [5]. He propose a statistic called D_2, to study the correlation between two sequences. The initial purpose was to speed up database searches, where alignment-based methods were too slow. The D_2 similarity is the correlation between the number of occurrences of all k-mers appearing in two sequences. Let A and B be two sequences from an alphabet Σ. The value A_w is the number of times w appears in A, with possible overlaps. Then the D_2 statistic is

$$D_2 = \sum_{w \in \Sigma^k} A_w B_w.$$

This is the inner product of the word vectors A_w and B_w, each one representing the number of occurrences of words of length k, i.e., k-mers, in the two sequences. However, it was shown by Lippert et al. [19] that the D_2 statistic can be biased by the stochastic noise in each sequence. To address this issue another popular statistic, called D_2^z, was introduced in [18]. This measure was proposed to standardize the D_2 in the following manner:

$$D_2^z = \frac{D_2 - \mathbb{E}(D_2)}{\mathbb{V}(D_2)},$$

where $\mathbb{E}(D_2)$ and $\mathbb{V}(D_2)$ are the expectation and the standard deviation of D_2, respectively. Although the D_2^z similarity improves D_2, it is still dominated by the specific variation of each pattern from the background [21], [27]. To account for different distributions of the k-mers, in [21] a new statistic is defined and named D_2^*. Let $\tilde{A}_w = A_w - (n-k+1) * p_w$ and $\tilde{B}_w = B_w - (n-k+1) * p_w$ where p_w is the probability of w under the null model. Then D_2^* can be defined as follows:

$$D_2^* = \sum_{w \in \Sigma^k} \frac{\tilde{A}_w \tilde{B}_w}{(n-k+1)p_w}.$$

This latter similarity measure responds to the need of normalization of D_2. Recently Goke et al. [14] proposed N_2, a statistic based on word-neighborhood. It is closely related with D_2^*, except that it counts words with at most one mismatch and considers also the reverse-complement of words. The N_2 statistic is in practice one of the best performing for the detection CRMs [14].

All these statistics have been studied by Reinert et al. [21] and Wan et al. [27] for the detection of regulatory sequences. Both papers test the performance on synthetic and real data sets. In particular, the pattern transfer model, introduced by Reinert et al. [21], was used to simulate the exchange of genetic material between two genomes. We describe this model in Section 3 and propose a more realistic formulation.

Most of the works on sequence similarity use the word distribution to study evolutionary relationships among different organisms [13], [22], [23]. Other works, instead, compare advantages and disadvantages of alignment-free methods [12], [29]. In [12], researchers have shown that the use of k-mer frequencies can improve the construction of phylogenetic trees traditionally based on a multiple-sequence alignment, especially for distant related species. The efficiency of alignment-free measures also allows the reconstruction of phylogenies for whole genomes [8], [9], [22]. Another application is the classification of protein remotely related, which can be addressed with sophisticated word counting procedures [10], [11]. Several other applications can benefit from the use of alignment-free methods [2]; for a comprehensive review we refer the reader to [26].

In the context of alignment-free measures the selection of the best resolution, the size k of words, is a non-trivial problem [22], [29], where the performance of every method is tightly related with this parameter. For example, in [27], using synthetic data, the best performance is achieved when k is equal to the length of the implanted patterns. In a real scenario, where we do not know the length of the biological signal in sequences under examination, it may be hard to choose the best resolution k. Moreover, in case of CRMs, where multiple binding sites with different lengths are present, a fixed value of k will never capture the statistic of all binding sites. These observations motivate the use of variable-length patterns for the estimation of similarity between regulatory sequences. The result will be a parameter-free alignment-free measure.

3 Methods and Materials: $Under_2$ and the Pattern Transfer Model Revised

In this section we describe our parameter-free alignment-free similarity measure, called $Under_2$, which is based on two concepts: irredundancy and underlying positioning.

Let us consider the space of all patterns, of all lengths, that are shared by two sequences s_1 and s_2, say Σ^*. The notion of irredundancy is meant to remove the redundant patterns, i.e., those patterns that do not convey extra information for the similarity measure. The second driving principle is the fact that every position in the sequences contributes a multiple number of times to the final score. More precisely, in all scores presented in Section 2, every position except the borders is part of exactly k different k-mers. In the same way, repeats present in the sequences might alter the occurrence profile of some patterns, making them more likely to be irredundant. Here we want to limit this phenomenon so that every position is part of at most one "underlying" pattern and thus will be accounted just one time. This will permit to correlate the nucleotides of s_1 and s_2 with each other, as one would naturally think, maintaining the information on possible translocations and duplications of specific patterns at the same time.

In the following we address these two issues separately. The goal is to build a similarity measure between the sequences s_1 and s_2 using all exact patterns of all lengths, Σ^*, that are shared between the two sequences.

3.1 Removing Redundant Patterns

The notion of irredundancy has been introduced in [4] and later on modified for the problem of protein comparison [10]. In this paper we consider this latter version, also called irredundant common motifs, but restricted only to the domain of exact patterns (i.e., without allowing mismatches or gaps) in order to deal with large sets of enhancers and large sequences.

It is well known that the total number of distinct patterns of any length in a sequence of length n are $\Theta(n^2)$. Remarkably a notable family of fewer than $2n$ patterns exists such that they are maximal in the host sequence, in the sense that it is impossible to extend a word in this class by appending one or more characters to it without losing some of its occurrences [3]. The linear size set of maximal patterns can be further reduced to the set of irredundant patterns.

In [10], we extended the notion of irredundancy to the case of pairwise sequence comparison, in order to avoid overcounting common patterns that cover the same region of a sequence. Indeed, one can easily show that most sequences share an unusually large number of common patterns that do not convey extra information about the input. To keep the article self-contained, here we summarize the basic facts already proved in [10]. If the occurrence of a pattern completely overlaps with the occurrence of another longer pattern, we say that the occurrence of the first pattern is covered by the second one.

Definition 1 (Irredundant/Redundant Common Patterns). *A pattern w is* irredundant *if and only if at least an occurrence of w in s_1 or s_2 is not covered by other patterns. A pattern that does not satisfy this condition is called a* redundant common pattern.

We denote the set of irredundant common patterns as $\mathcal{I}_{s_1,s_2}$. In loose terms, all redundant common patterns can be deduced from patterns in $\mathcal{I}_{s_1,s_2}$, since they do not carry more information than the set of irredundant common patterns and therefore can be discarded.

One can show that every irredundant common pattern w is the result of some intersection of the two input sequences, where each meet in this case corresponds to a particular a set of patterns [10]. We observe again that $\mathcal{I}_{s_1,s_2}$ is a subset of the well-known linear set of maximal common patterns; therefore the number of irredundant common patterns is bounded by $|s_1| + |s_2|$.

A simple algorithm that can discover all such patterns (without gaps) can be found in [8]. It employs a generalized suffix tree of the two sequences s_1 and s_2. The construction of the generalized suffix tree and the subsequent extraction of the irredundant common patterns can be completed in time and space linear in the size of sequences, by exploiting well-known properties and implementations of suffix arrays and matching statistics [16].

In summary, the notion of irredundancy is useful for removing non-informative patterns, and thus for drastically reducing the number of candidates to be analyzed to estimate the sequence similarity between s_1 and s_2. Moreover, it is worth mentioning that this notion can be efficiently computed also for long sequences and large data sets [8], [24].

3.2 Selecting Underlying Patterns

A notable problem with alignment-free measures is that they are biased by the presence of repetitive words in genomes, with a major tradeoff between the number of words they can encode and the resolution, or size, and repetitiveness of these words along the sequences. For example, mammalian genomes can have very long repeats that contribute to overcounting the presence of some k-mers. At the same time different k-mers with occurrences shifted of just one or two positions with each other in s_1 can easily match different regions of the other genome s_2, leading to patterns that ambiguously represent the similarity of sequences.

Following these simple ideas, when comparing large genomes or large sets of genomic sequences, one would like to avoid that repeats or patterns with low resolution may be trivially overcounted a multiple number of times, misleading the final similarity score. To partially address this issue, repeat regions are usually masked before computing the similarity score [14]. Nevertheless, in this paper we take into account both problems, repeats and possible low resolution of patterns, in a simple and systematic manner without discarding any information on sequences.

The basic idea behind our approach is that a position on the sequences should contribute only once to the final similarity. Again, traditionally alignment-free statistics fail to comply with this simple rule. In fact, every position, apart from the borders, belongs to k different k-mers and thus contributes k times to the similarity. On the other hand every irredundant common pattern has at least one of its occurrences that is not covered by any other pattern. However, this occurrence might be partially covered by other patterns that are perhaps longer and more significant, or could be a repeat that makes the pattern more likely to be the irredundant. For example, for large genomes of similar size n we noticed a number of irredundant common patterns very close to the actual size n of each genome, and this fact can lead to account each position $O(n^2)$ times. Here we want every position to be covered by exactly one pattern so that it will be accounted only once in the similarity score, therefore correlating the nucleotides of s_1 and s_2 with each other unambiguously, even in case of translocations and duplications (which are usually not taken into account by classic alignment methods).

In previous works on whole-genome comparison, to solve this problem we used the notions of pattern priority and of underlying pattern [8]. The pattern priority rule is mainly based on the idea of selecting, for each position, those patterns that represent the largest number of matching sites between sequences, and thus that are more likely to be conserved patterns. Since it might be the case that two or more patterns with the same length overlap with each other, we exploited the fact that genomic coordinates give a total rank of all patterns to select just one "underlying " pattern for each position, in a combinatorial fashion. In practice the use of genomic coordinates proved to be valuable and efficient in whole-genome comparison, as a universal way to sort out patterns with a similar level of conservation. Note that, in principle, a score based on the probability for a pattern to appear over an entire genome would have been preferable, but we found it not feasible to be applied, because genomic regions can be characterized by highly different nucleotide profiles that bias the likelihood of appearance of a pattern (for example, in highly repeating regions).

Nevertheless, in case of enhancer sequences we can actually infer a general nucleotide profile that does not significantly vary among them. Therefore, it would be natural to include a probabilistic score for each pattern in our priority rule; we will see how to compute this score in the next section. In the following we formally define the new pattern priority rule and the underlying patterns.

Let us consider the set of irredundant common patterns $\mathcal{I}_{s_1,s_2}$ as input. Given two patterns w and w', we say that w has priority over w', denoted $w \rightarrow w'$, if and only if either $|w| > |w'|$, or $|w| = |w'|$ and w is less likely to appear in the sequences than w', or w and w' have the same length and probability to appear, but the first occurrence of w appears before the first occurrence of w'. Following the notion of pattern priority, every pattern can be defined just by its length, probability, and its starting positions in the sequences, meaning that any set of patterns is totally ordered with respect to the priority rule. We say that an occurrence l of w is *tied* to an occurrence l' of another pattern w', if these occurrences (partially) overlap to each other, $[l, l + |w| - 1] \cap [l', l' + |w'| - 1]) \neq \emptyset$, and $w' \rightarrow w$. Otherwise, we say that l is *untied* from l'.

Definition 2 (Underlying Patterns). *A set of patterns* $\mathcal{U}_{s_1,s_2} \subseteq \mathcal{I}_{s_1,s_2}$ *is said to be the* Underlying set *of* $\{s_1, s_2\}$ *if and only if:*

1. *every pattern* w *in* $\mathcal{U}_{s_1,s_2}$*, called* underlying pattern, *has at least one occurrence in both sequences that is untied from all the untied occurrences of other patterns in* $\mathcal{U}_{s_1,s_2} \setminus w$*, and*
2. *there does not exist a pattern* $w \in \mathcal{I}_{s_1,s_2} \setminus \mathcal{U}_{s_1,s_2}$ *such that* w *has at least two untied occurrences, one per sequence, from all the untied occurrences of patterns in* $\mathcal{U}_{s_1,s_2}$.

The objective of this definition is to select the most important patterns in $\mathcal{I}_{s_1,s_2}$ for each location of the sequences, according to the pattern priority rule. If a pattern w is selected, we filter out all occurrences of patterns with less priority than w that lay on the untied locations of w, in a simple combinatorial fashion.

Similarly to the underlying patterns presented in [8], it is easy to see that the Underlying set exists and is unique for a pair in input $\{s_1, s_2\}$. The complete procedure to discover the set $\mathcal{U}_{s_1,s_2}$ can be found in [8]. Here below we give an overview of the algorithm.

```
Underlying Pattern Extraction (Input: $s_1$, $s_2$; Output: $\mathcal{U}_{s_1,s_2}$)
  Compute the set of Irredundant common patterns $\mathcal{I}_{s_1,s_2}$.
  Rank all patterns in $\mathcal{I}_{s_1,s_2}$ using the pattern priority rule.
  for Select the top pattern, $w$, from $\mathcal{I}_{s_1,s_2}$: do
    if Check in $\Gamma$ if $w$ has at least one untied occurrence per
    sequence that is not covered by some other patterns already
    in $\mathcal{U}_{s_1,s_2}$ then
      Add $w$ to $\mathcal{U}_{s_1,s_2}$ and update the location vector, $\Gamma$, in
      which $w$ appears as untied.
    else
      Discard $w$.
    end if
  end for
```

The algorithm requires first to order the set of irredundant common patterns by means of the pattern priority. The complexity of sorting is in general $O(l \log l)$, where $l = |s_1| + |s_2|$. However, due to the properties of the pattern priority, this step can be done in $O(l)$ time using radix sort. An auxiliary vector Γ, of length l, is used to represent all locations of s_1 and s_2. If a location i is covered by some pattern already in $\mathcal{U}_{s_1,s_2}$, then $\Gamma[i] = \mathit{false}$; otherwise, if the location is free, $\Gamma[i] = \mathit{true}$. For a pattern w in $\mathcal{I}_{s_1,s_2}$, we can check whether its occurrences are tied to other patterns by looking at the vector Γ. If some untied occurrences are found, then we can add the new underlying pattern w to $\mathcal{U}_{s_1,s_2}$, and update the vector Γ accordingly using all the untied occurrences of w. In total the extraction of all underlying patterns, using this scheme, takes $O(l^2)$ time. A more advanced algorithm with a better complexity, $O(l \log l \log \log l)$ time and $O(l)$ space, can be found in [8].

As a corollary, it can be shown that all untied occurrences of the underlying patterns in $\mathcal{U}_{s_1,s_2}$ can be mapped into the sequences s_1 and s_2 without overlaps. We will use precisely these occurrences to compute a similarity measure in which every position contributes only once.

3.3 Building the $Under_2$ Similarity Measure

Next, we want to compute a similarity measure based on the underlying patterns in $\mathcal{U}_{s_1,s_2}$. At first we can note that the set of underlying patterns U_{s_1,s_2} is not symmetric, in general $U_{s_1,s_2} \neq U_{s_2,s_1}$. Thus, in order to build a symmetric measure, we need to consider both sets. Our similarity is inspired by the Average Common Subword approach (ACS) [24], where the scores of common patterns found are averaged over the length of sequences. Here we follow the same approach, but, instead of counting all common patterns, we use just the untied occurrences of the underlying patterns, which by definition do not overlap [8].

Then, in ACS the contribution of each position is given by the length of the pattern covering that position. In our approach we use instead the ratio of the number of occurrences for an underlying pattern w, and the expected number of occurrences for that pattern. Let us define occ_w as the number of occurrences of w, and $untied^1_w$ as the number of untied occurrences of w in s_1. First we compute the score:

$$Score(s_1, s_2) = \frac{\sum_{w \in U_{s_1,s_2}} |w| * untied^1_w * \frac{occ_w}{E[occ_w]}}{|s_1|}.$$

Recalling that the untied occurrences do not overlap with each other, we notice that the term $|w| * untied^1_w$ counts the positions where w appears without overlapping any other pattern. For each such position we sum the score $\frac{occ_w}{E[occ_w]}$, where $E[occ_w]$ is the expected number of occurrences. Note that the expectation of this ratio is exactly 1. This sum is then averaged over the length of the first sequence under examination, s_1. This score is large when the two sequences are similar, therefore we take its inverse. Then, since the total number of occurrences of an underlying pattern w present in s_1 is expected to logarithmically increase with the length of s_2, we consider the measure $log_4(|s2|)/Score\ (s1, s2)$, where a base-4 logarithm is used to represent the four DNA bases.

To center the formula, such that it goes to zero when $s_1 = s_2$, we subtract the term $\log_4 |s_1|$. If $s_1 = s_2$ there will be just one underlying pattern that is equal to the sequence itself. In this case, $Score(s_1, s_1)$ will be 1 and the term $\log_4|s_1|$ makes sure that $\overline{Under_2}(s_1, s_1) = 0$. These observations are implemented in the general formula of $\overline{Under_2}(s_1, s_2)$:

$$\overline{Under_2}(s_1, s_2) = \frac{\log_4|s_2|}{Score(s_1, s_2)} - \log_4|s_1|,$$

$$Under_2(s_1, s_2) = \frac{\overline{Under_2}(s_1, s_2) + \overline{Under_2}(s_2, s_1)}{2}.$$

Finally, to correct the asymmetry, our similarity measure called $Under_2$ is the average of the two statistics $\overline{Under_2}(s_1, s_2)$ and $\overline{Under_2}(s_2, s_1)$.

An important aspect in this formula is the computation of the expected number of occurrences of a pattern w. A Bernoulli model, where each symbol is independent from each other, is usually inappropriate for genomic sequences that might be rich, for example, of CpG dinucleotides. A Markov model usually outperforms the Bernoulli model on biological sequences. In our case the length of CRMs is relatively short and thus, to avoid overfitting, we will rely on a first order Markov model as in [14]. In summary, the expectation is computed as $E[occ_w] = p_w(l - |w| + 1)$, where p_w is the probability of w using the Markov model and $l = |s_1| + |s_2|$. In our experiments we estimate the background probabilities for each pair of sequences separately. For a fair comparison also the expectations in D_2^*, D_2^z, N_2 will use the same background model. These latter statistics, with the first order Markov model, are implemented in the N_2 package [14].

Finally, we extend our approach to account for untied occurrences that are present in the reverse, complement, and reverse-complement of each sequence, in order to simulate the DNA strand and the evolution of sequences. Every underlying pattern can thus match the sequences s_1 and s_2 in different ways, but a location in the sequences will be covered at most by an untied occurrence, independently of the matching strand. For more details about this extension, we refer to [8].

3.4 The Pattern Transfer Model Revised

Reinert et al. [21] proposed two simulation models: the *common pattern model*, and the *pattern transfer model*. In the common pattern problem a pattern is implanted into two sequences at random positions. In the pattern transfer model, random fragments of the first sequence are copied into the second.

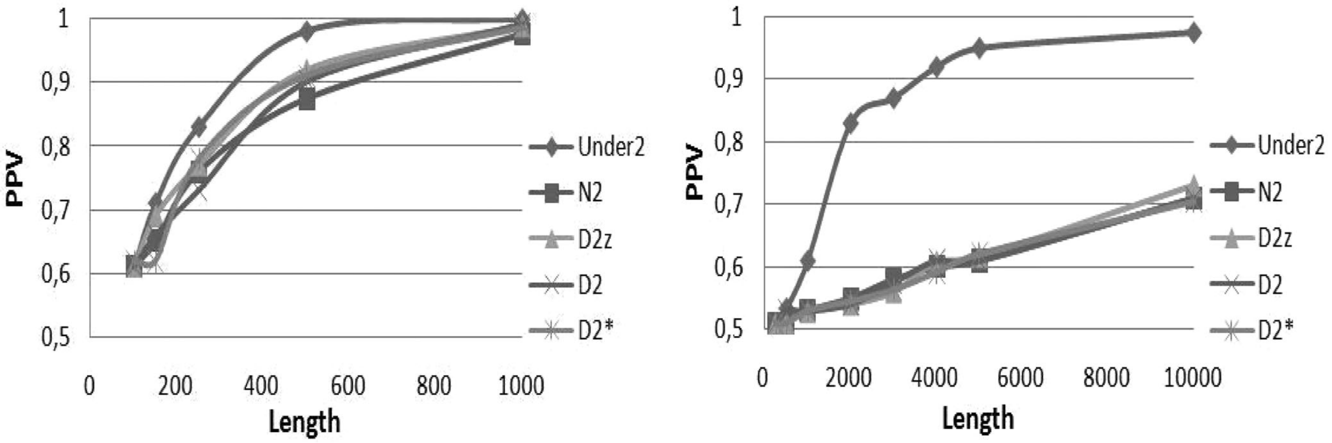
The latter model simulates the exchange of genetic material, e.g., horizontal gene transfer, and can model sequences from distant related organisms. This model is of particular interest for the analysis of CRMs. Let us consider two input strings $S = s_1s_2s_3 \dots s_n$ and $W = w_1w_2w_3 \dots w_n$. In the pattern transfer model the string S is associated with a set of Bernoulli random variables Z_1, $Z_2 \dots$, Z_{n-k+1} where $P(Z_i = 1) = \lambda$. Every time $Z_i = 1$, a word of length l is transferred from $s_is_{i+1} \dots s_{i+l-1}$ to $w_iw_{i+1} \dots w_{i+l-1}$. If a word is transferred from position i of S to position i of W, we ignore the values of Z_j for $i < j \leq i + l - 1$ and the process restarts from position $i + l$, therefore we do not allow overlaps. We say that the resulting sequences W and S are related according to the pattern transfer model. This model has been used for different studies, see for example [21], [27].

There are two basic observations that we can make regarding the pattern transfer model. First the model always transfers patterns of fixed-length l. In general the binding sites of a CRMs can have different lengths, but also their degrees of preservation can be different, thus the use of a fixed-length model does not often fit CRMs. Moreover, since almost all alignment-free methods use words of fixed-length k, the performance of these methods is tightly related with the choice of k and its relation with l. A second observation is the fact that often binding sites in similar CRMs appear with a different orientation. More precisely the same binding site can appear in another regulatory sequence as reverse, complement or as reverse-complement. To accommodate these observations we formulate the *pattern transfer model revised* as follows.

Similarly to the original model, if $Z_i = 1$, a word is transferred from position i of S to position i of W. However, this word will be of variable size, between l to $l + \delta$. The length of the word to be transferred is chosen uniformly at random, i.e., with probability $1/\delta$. This word is then copied from position i of W as it is, or as reverse, complement, or reverse-complement again with equal probability. The resulting string W is now related to S through the pattern transfer model revised. As we will see later in Section 4, in general the detection of correlations between sequences related to each other through the pattern transfer model revised is harder than with the original model.

4 Experimental Results on Synthetic and Real Data

To assess the performance of $Under_2$ and compare it to the other statistics presented above, we devised a series of tests on real and simulated data.



$\lambda = 0.01$
(a)

$\lambda = 0.001$
(b)

Fig. 1. PPV/accuracy scores for various methods on the original pattern transfer model.

4.1 Pattern Transfer on Simulated Data

We follow the experimental setup of [18] and of [20]. Let us first generate a set of random sequences using an i.i.d. distribution, as background model (or negative set). We implant patterns using the pattern transfer model into a copy of the same set of sequences to simulate CRMs; this will be our positive set. We then compute the pairwise scores between all pairs in the negative set, and between all pairs in the positive set. For all methods, we assess if the pairs from the positive set score higher than the pairs from the negative set. This is done by sorting all pairwise scores in one combined list. We consider as positive predictive value (PPV) the percentage of pairs from the positive set that are in the top half of this list. This setup was proposed in [18] and, in this case, PPV, or precision, is equivalent to both sensitivity and accuracy. A score of 1 means a perfect separation between negative and positive sets, while values close to 0.5 imply no statistical power.

In our experiments we generate the random sequences using two different i.i.d. models. In the first model all symbols have the same probability $p_A = p_C = p_G = p_T = 1/4$, whereas in the second model we simulate GC-rich sequences with $p_A = p_T = 1/6$ and $p_C = p_G = 1/3$. We assess the performance of the methods presented in Section 2 for two values of λ, 0.01 and 0.001, while varying the length of the input sequences. All results show the average scores over 20 simulations, every time drawing 50 new random sequences.

We start following the original pattern transfer model, by inserting words of length 5, similarly to [20]. With this experimental setup it has been shown that all others statistics give the best performance with $k = 5$. In all our experiments we use the most performing setup also for N_2, which is using k-mers (again $k = 5$) with at most one mismatch and considering the reverse-complement. All statistics are computed using the package ALF from the SeqAn library (http://www.seqan.de).

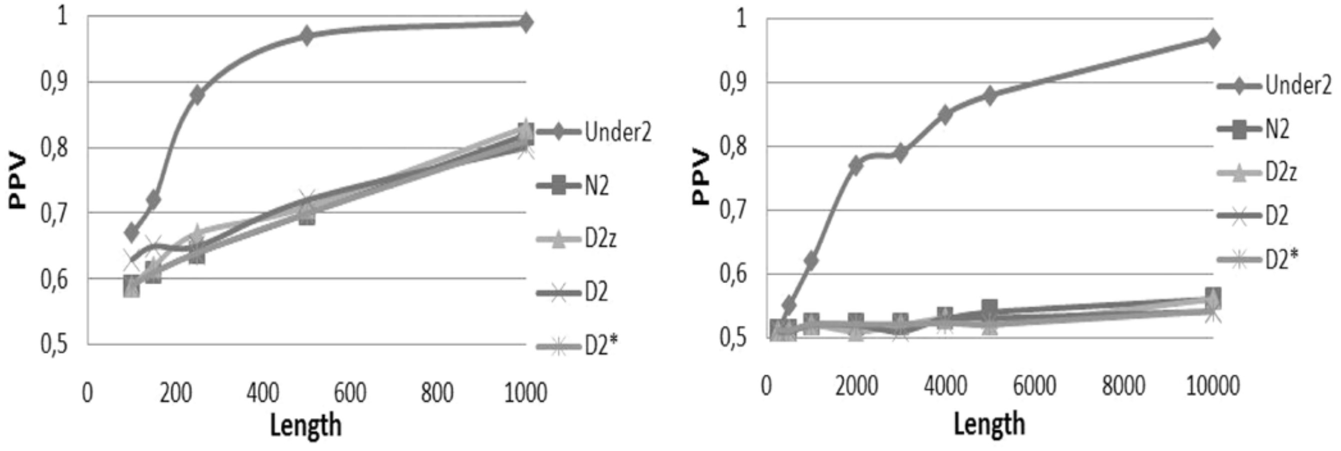
Fig. 1 presents the performance of all statistics when using the original pattern transfer model, a background with equally probable symbols, and the two tested values of λ. In general, the performance of all methods increases with the length of the sequences, where the number of implanted patterns also increases. For large values of λ the number of implanted patterns is high enough to get all methods performing very well. With $\lambda = 0.001$ the instance becomes difficult and all the other methods need longer sequences to reach a significant score, usually no greater than 0.75, whereas $Under_2$ classifies correctly the two sets of sequences with an accuracy of 0.9 even when the length of sequences is just $4,000$ bp.

By applying the same experimental setup, we test all similarity measures using the pattern transfer model revised, with the length of implanted patterns in the range [4-6]. Results are presented in Fig. 2. We can observe that the performance of all methods degrades and, as we expected, with this revised version it is more difficult to detect the implanted biological signals. For all methods except $Under_2$ the performance reduction is clear, especially for $\lambda = 0.01$. Similarly with the previous figure, the behavior of all other measures is almost indistinguishable.

In Fig. 3 the setup is similar to Fig. 2 except that the background sequences are generated with the proposed GC-rich model. Our method, $Under_2$, remains the best performing, whereas now we can observe more clearly the differences between the other measures. As expected D_2 is the worse, as it does not correct for sequence biases. The performance of D_2^z slightly degrades, while N_2 and D_2^*, which are very similar measures, take advantage of the nucleotide distribution.

4.2 Pattern Transfer on Drosophila Genome

In the above simulations we study the power of all statistics to detect relationships between random sequences under the pattern transfer model revised for different settings. However, random sequences are not a good background model for genomes. Here we test the ability to cluster related sequences when the background is a real genomic sequence. We follow the experimental setup of [20] and use the same real genomic data as a background. We first download all the intergenic sequences of the Drosophila genome



$\lambda = 0.01$
(a)

$\lambda = 0.001$
(b)

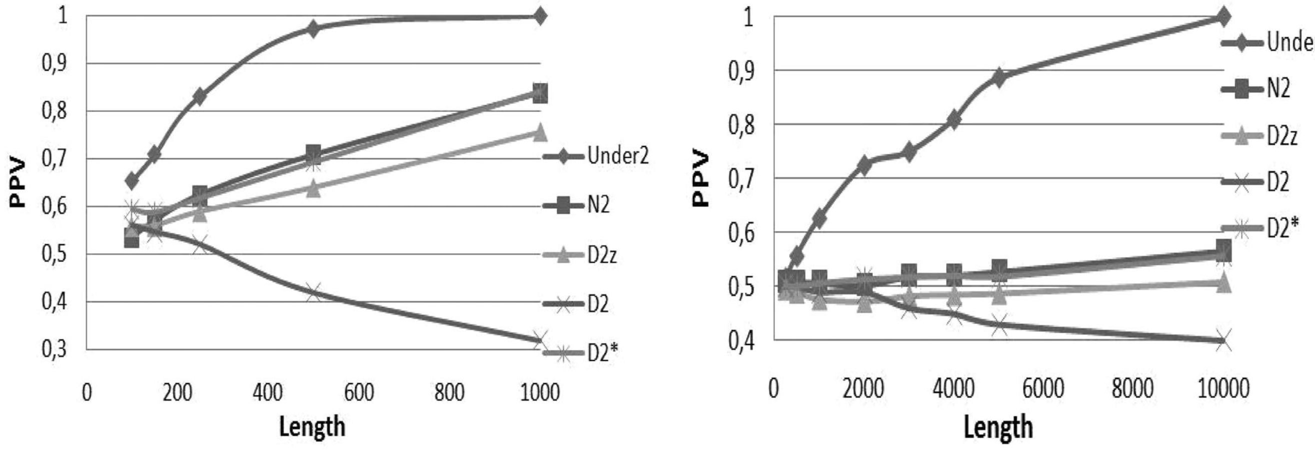
Fig. 2. PPV/accuracy scores for various similarity measures after applying the pattern transfer model revised.

from FlyBase (http://flybase.org, dmel-all-intergenic-r5.49.fasta). Then, we pick at random 50 sequences of the same length as a negative set, and create the positive set using the pattern transfer model revised as above. We repeat this process 20 times and report the average scores in Fig. 4 for all statistics, for different lengths in background, and different values of λ.

With a real background we can observe that all statistics are no longer monotonic. In the simplest instance, with $\lambda = 0.01$, the evolutionary signal can be easily recovered by N_2 and $Under_2$, while in general the relative performance of all statistics does not mutate with respect to the previous experiment. When the evolutionary signal becomes more subtle, $\lambda = 0.001$, only $Under_2$ can detect it, whereas all other statistics have no discrimination power (accuracy close to 0.5).

4.3 The Effect of Evolutionary Time after Pattern Transfer with Real Data

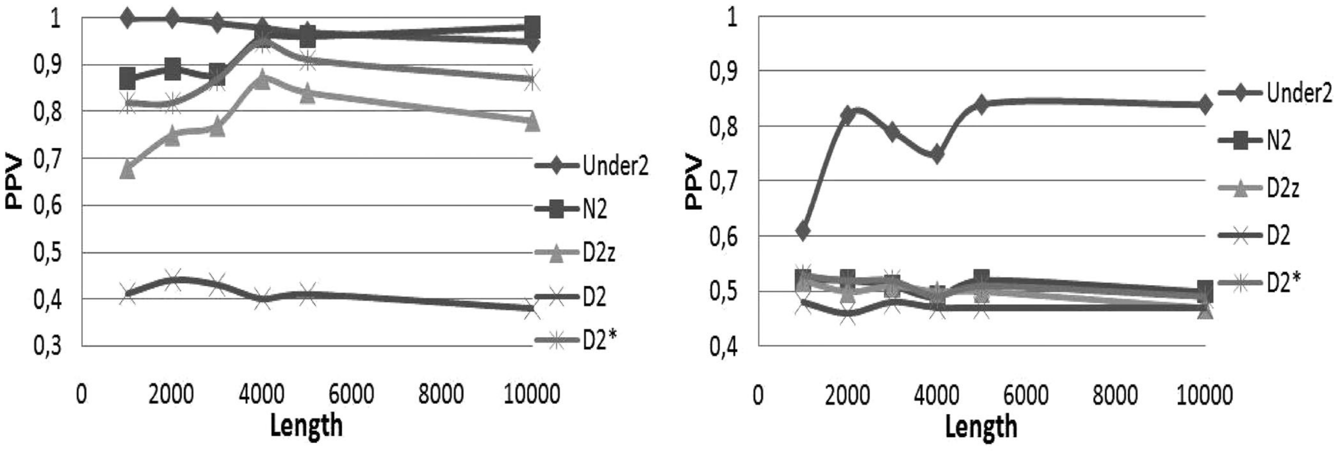
In the above simulations we studied the ability of the statistics to detect related sequences immediately after the pattern transfer model revised. Although this model might mimic the exchange of genetic material, in many real situations the two sequences continue to evolve afterward. Thus, it is of interest to understand how evolution affects the performance of the different statistics. To devise a suitable test we use as input Drosophila background sequences with length of 2,000 taken from the previous simulation and $\lambda = 0.01$. We then evolve the sequences using the HKY model presented in [17], with ratio equal to 2.0. We write $\theta = t\eta$, where t is the time and η is the rate of mutation. In our experiments we consider the average human mutation rate of $\eta = 10^{-8}$. As a result, the evolutionary time depends now only on θ. We took as input the Drosophila sequences after the pattern transfer and then evolve this set for various values of $\theta = 0.01 - 0.1$. In Fig. 5 we report the performance of the different methods with respect to the evolutionary time. As expected, the power of all methods decreases over time as a function of θ. We continue to see that the relative performance of all methods remains unchanged. When $t = 10^6$ generations, which implies $\theta = 0.01$, the power of $Under_2$ remains close to 1; however, after $t = 10^7$



$\lambda = 0.01$
(a)

$\lambda = 0.001$
(b)

Fig. 3. PPV/accuracy scores for various methods after applying the pattern transfer model revised with a GC-rich background.



$\lambda = 0.01$

(a)

$\lambda = 0.001$

(b)

Fig. 4. PPV/accuracy scores for various methods on the pattern transfer model revised using real DNA sequences as background (Drosophila).

generations and $\theta = 0.1$ it goes down to 0.6, where all others statistics have already lost their ability to discriminate longer before that time.

4.4 Comparison of Mouse Regulatory Sequences

The above simulations deal with artificial CRMs from unrelated sequences, in this section we use real ChIP-seq data to discriminate in vivo identified enhancers. Tissue-specific enhancers in mouse embryos have been discovered by Visel et al. [25] and Blow et al. [6]. We use enhancers active in forebrain, midbrain, limb, and heart as positive set (as in [14]). To correct for length differences, we select those sequences with length of about $1,000$ bases.

In a first test, for the negative set we randomly sampled sequences of the same length from the mouse genome, and compute all pairwise scores and the accuracy as above. We maintain the parameters of previous experiments to have a consistent and comparable setup. The results are presented in Table 1. Again, the results are averaged over 20 runs, where in every run we select 50 sequences from both sets. Across all tissues, $Under_2$ gives the best results, demonstrating that it is the most suitable to detect tissue-specific activities of regulatory sequences. The results confirm the relative power of the other statistics. In particular, with this real data set we can observe that N_2 performs slightly better than D_2^*, confirming that the use of reverse-complement, as implemented in N_2, plays an important role.

The previous test indicates that tissue-specific enhancers can have a similar word content. However, the comparison with random genomic sequences can be biased by the technology, e.g. when it more likely extracts sequences with high or similar GC-content. To avoid this bias introduced by the technology, we also compare different ChIP-seq sequences between each other. This is a much more challenging test, that can be used by biologists to select enhancers that drive a similar expression pattern. We use as positive set the enhancers active in one tissue, and as negative set the enhancers active in all others tissues. Table 2 reports the results of this test. Although the accuracy decreases compared to Table 2, these later experiments confirm that similar tissue-specific enhancers have a higher sequence similarity, and thus they can be detected with alignment-free methods. In this difficult test $Under_2$ obtains an average accuracy of 0.7, while the advantage with respect to all others methods remains substantial, further proving to be a valuable statistic for the identification and analysis of regulatory sequences.

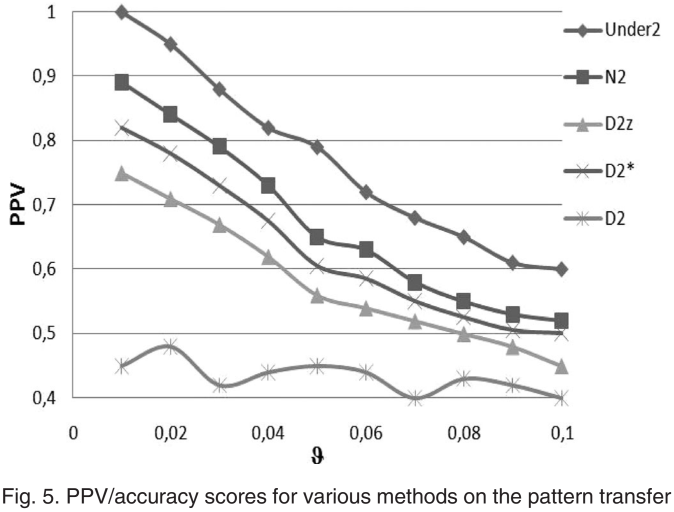


Fig. 5. PPV/accuracy scores for various methods on the pattern transfer model revised using real DNA sequences as background, and evolving the sequences.

TABLE 1
Comparison of ChIP-seq Data of Mouse Tissue-Specific Enhancers versus Random Mouse Genomic Sequences

Tissue	U_2	N_2	D_2^z	D_2^*	D_2
ForeBrain	**0.85**	0.78	0.65	0.77	0.59
MidBrain	**0.81**	0.76	0.65	0.74	0.55
Heart	**0.75**	0.70	0.62	0.69	0.51
Limb	**0.80**	0.71	0.63	0.72	0.49
Average	**0.80**	0.74	0.63	0.73	0.535

Values in the table represent the accuracy of each method for different tissues. The best scores are in bold.

TABLE 2
Comparison of ChIP-Seq Data of Mouse Tissue-Specific Enhancers versus Others Tissue-Specific Enhancers

Tissue	U_2	N_2	D_2^z	D_2^*	D_2
ForeBrain	**0.73**	0.66	0.53	0.65	0.51
MidBrain	**0.69**	0.64	0.55	0.63	0.45
Heart	**0.67**	0.59	0.54	0.58	0.44
Limb	**0.71**	0.6	0.54	0.6	0.47
Average	**0.70**	0.62	0.54	0.61	0.47

Values in the table represent the accuracy of each method for different tissues. The best scores are in bold.

5 Conclusion and Future Work

In this paper we studied the use of alignment-free measures to detect functional and/or evolutionary similarities among regulatory sequences. We introduced a parameter-free alignment-free method called $Under_2$ that is designed around the use of variable-length words combined with specific statistical and syntactical properties. A new model to simulate the exchange of genetic material has been introduced and studied. To evaluate the performance of several alignment-free methods, we devised an extensive series of tests on both synthetic and real data. In almost all simulations our method $Under_2$ outperforms all other statistics. The performance gain becomes more evident when the pattern transfer model revised is applied, or when the evolutionary signal becomes more subtle. Importantly, $Under_2$ is also able to detect similarities between *in vivo* identified enhancer sequences, e.g., of mouse. This will allow to verify and study the architecture of regulatory elements. As shown in the article, a similarity measure based on variable-length word count can successfully detect tissue-specific enhancers. This suggests that different binding site contents, captured by variable-length words, may play an important role to the tissue-specificity of enhancers. This will help to understand sequence-dependent code within CRMs, which is responsible for the large diversity of cell types. As a future direction of investigation, we will consider the use of $Under_2$ for the assembly-free comparison of genomes based only on short reads.

Acknowledgments

Matteo Comin was partially supported by the Ateneo Project CPDA110239. This work was developed within the scope of the P.R.I.N. Project 20122F87B2.

References

[1] S. Altschul, W. Gish, W. Miller, E.W. Myers, and D. Lipman, "Basic Local Alignment Search Tool," *J. Molecular Biology*, vol. 215, no. 3, pp. 403-410, 1990.

[2] M. Antonello and M. Comin, "Fast Computation of Entropic Profiles for the Detection of Conservation in Genomes," *Proc. Eighth IAPR Int'l Conf. Pattern Recognition in Bioinformatics*, pp. 277-288, 2013.

[3] A. Apostolico, "Maximal Words in Sequence Comparisons Based on Subword Composition," *Algorithms and Applications*, T. Elomaa, H. Mannila, and P. Orponen, eds., pp. 34-44, Springer-Verlag, 2010.

[4] A. Apostolico and L. Parida, "Incremental Paradigms of Motif Discovery," *J. Computational Biology*, vol. 11, no. 1, pp. 15-25, 2004.

[5] B.E. Blaisdell, "A Measure of the Similarity of Sets of Sequences not Requiring Sequence Alignment," *Proc. Nat'l Academy of Sciences USA*, vol. 83, no. 14, pp. 5155-5159, 1986.

[6] M.J. Blow et al., "ChIP-Seq Identification of Weakly Conserved Heart Enhancers," *Nature Genetics*, vol. 42, no. 9, pp. 806-810, 2010.

[7] S. Bonn et al., "Tissue-Specific Analysis of Chromatin State Identifies Temporal Signatures of Enhancer Activity during Embryonic Development," *Nature Genetics*, vol. 44, no. 2, pp. 148-156, 2012.

[8] M. Comin and D. Verzotto, "Alignment-Free Phylogeny of Whole Genomes Using Underlying Subwords," *BMC Algorithms for Molecular Biology*, vol. 7, article 34, 2012.

[9] M. Comin and D. Verzotto, "Whole-Genome Phylogeny by Virtue of Unic Subwords," *Proc. 23rd Int'l Workshop Database and Expert Systems Applications (DEXA-BIOKDD '12)*, pp. 190-194, 2012.

[10] M. Comin and D. Verzotto, "The Irredundant Class Method for Remote Homology Detection of Protein Sequences," *J. Computational Biology*, vol. 18, no. 12, pp. 1819-1829, 2011.

[11] M. Comin and D. Verzotto, "Classification of Protein Sequences by Means of Irredundant Patterns," *BMC Bioinformatics*, vol. 11, Suppl. 1, article S16, 2010.

[12] Q. Dai and T. Wang, "Comparison Study on K-Word Statistical Measures for Protein: From Sequence to Sequence Space," *BMC Bioinformatics*, vol. 9, article 394, 2008.

[13] L. Gao and J. Qi, "Whole Genome Molecular Phylogeny of Large dsDNA Viruses Using Composition Vector Method," *BMC Evolutionary Biology*, vol. 7, article 41, 2007.

[14] J. Goke, M.H. Schulz, J. Lasserre, and M. Vingron, "Estimation of Pairwise Sequence Similarity of Mammalian Enhancers with Word Neighbourhood Counts," *Bioinformatics*, vol. 28, no. 5, pp. 656-663, 2012.

[15] T. Goto et al., "Early and Late Periodic Patterns of Even Skipped Expression Are Controlled by Distinct Regulatory Elements that Respond to Different Spatial Cues," *Cell*, vol. 57, no. 3, pp. 413-422, 1989.

[16] D. Gusfield, *Algorithms on Strings, Trees, and Sequences: Computer Science and Computational Biology*. Cambridge Univ. Press, 1997.

[17] M. Hasegawa, H. Kishino, and T. Yano, "Dating of the Human-Ape Splitting by a Molecular Clock of Mitochondrial DNA," *J. Molecular Evolution*, vol. 22, no. 2, pp. 160-174, 1985.

[18] M.R. Kantorovitz, G.E. Robinson, and S. Sinha, "A Statistical Method for Alignment-Free Comparison of Regulatory Sequences," *Bioinformatics*, vol. 23, no. 13, pp. i249-i255, 2007.

[19] R.A. Lippert, H.Y. Huang, and M.S. Waterman, "Distributional Regimes for the Number of K-Word Matches between Two Random Sequences," *Proc. Nat'l Academy of Sciences USA*, vol. 100, no. 13, pp. 13980-13989, 2002.

[20] X. Liu, L. Wan, J. Li, G. Reinert, M.S. Waterman, and F. Sun, "New Powerful Statistics for Alignment-Free Sequence Comparison under a Pattern Transfer Model," *J. Theoretical Biology*, vol. 284, no. 1, pp. 106-116, 2011.

[21] G. Reinert, D. Chew, F. Sun, and M.S. Waterman, "Alignment-Free Sequence Comparison (I): Statistics and Power," *J. Computational Biology*, vol. 16, no. 12, pp. 1615-1634, 2009.

[22] G.E. Sims, S.-R. Jun, G.A. Wu, and S.-H. Kim, "Alignment-Free Genome Comparison with Feature Frequency Profiles (FFP) and Optimal Resolutions," *Proc. Nat'l Academy of Sciences USA*, vol. 106, no. 8, pp. 2677-2682, Feb. 2009.

[23] J. Qi, H. Luo, and B. Hao, "CVTree: A Phylogenetic Tree Reconstruction Tool Based on Whole Genomes," *Nucleic Acids Research*, vol. 32, pp. W45-W47, (WebServerIssue), 2004.

[24] I. Ulitsky, D. Burstein, T. Tuller, and B. Chor, "The Average Common Substring Approach to Phylogenomic Reconstruction," *J. Computational Biology*, vol. 13, no. 2, pp. 336-350, 2006.

[25] A. Visel et al., "ChIP-Seq Accurately Predicts Tissue-Specific Activity of Enhancers," *Nature*, vol. 457, no. 7231, pp. 854-858, 2009.

[26] S. Vinga and J. Almeida, "Alignment-Free Sequence Comparison—A Review," *Bioinformatics*, vol. 19, no. 4, pp. 513-523, 2003.

[27] L. Wan, G. Reinert, D. Chew, F. Sun, and M.S. Waterman, "Alignment-Free Sequence Comparison (II): Theoretical Power of Comparison Statistics," *J. Computational Biology*, vol. 17, no. 11, pp. 1467-1490, 2010.

[28] M.D. Wilson et al., "Species-Specific Transcription in Mice Carrying Human Chromosome 21," *Science*, vol. 322, no. 5900, pp. 434-438, 2008.

[29] T. Wu, Y. Huang, and L. Li, "Optimal Word Sizes for Dissimilarity Measures and Estimation of the Degree of Dissimilarity between DNA Sequences," *Bioinformatics*, vol. 21, no. 22, pp. 4125-4132, 2005.

Matteo Comin received the MS and PhD degrees in computer science from the University of Padova, Italy, in 2003 and 2007, respectively. Since 2007, he has been an assistant professor at the University of Padova. His research interests focus on the area of algorithms for computational biology. During his activity, he has been a research intern at IBM T.J. Watson Research Center twice where he developed motif discovery systems for biological sequences. His research interests also include computational methods for protein structural comparison and protein-protein docking prediction, and also in developing algorithms for next-generation sequencing. He has been a visiting researcher at the Purdue University (US) and at the Universitat Politècnica de Catalunya, Barcelona (Spain) three times. He is a co-inventor of three US patent and author of more than 30 publications. In 2007, he received the C. Offelli Award for best young researcher from the University of Padova.

Davide Verzotto received the MS and PhD degrees in computer science from the University of Padova, Italy, in 2008 and 2012, respectively. He is currently a postdoctoral fellow at the Genome Institute of Singapore (A*STAR). His research interests focus on the area of pattern discovery and data mining in computational biology. During his PhD, he was a visiting student in the Department of Computer Science and Engineering at the University of California, Riverside.