Averaging fuzzy biopolymers [1]

J. Casasnovas, F. Rosselló

Dept. of Mathematics and Computer Science, Research Institute of Health Science (IUNICS), University of the Balearic Islands, 07122 Palma de Mallorca (Spain)
E-mail: {jaume.casasnovas,cesc.rossello}@uib.es

Abstract

Let d be a metric on the set $\mathcal{FP}(X)$ of fuzzy subsets of a finite set X. A midpoint of two fuzzy subsets $\mu, \nu \in \mathcal{FP}(X)$ is any fuzzy subset $\xi \in \mathcal{FP}(X)$ such that $d(\xi, \mu) = d(\xi, \nu) = \frac{1}{2}d(\mu, \nu)$. These midpoints can be used to represent "middle ways" or "compromises" between two situations described by the fuzzy subsets μ and ν. Now, the imprecise knowledge of a nucleic acid or protein sequence of length N can be modeled by means of a *fuzzy biopolymer*, a fuzzy subset of a kN-element set with k the number of bases, 4, in the case of nucleic acids, and of amino acids, 20, in the case of proteins. Thus, a midpoint of two fuzzy biopolymers of the same length can be understood as an average of the knowledge of the sequences represented by them. In this paper we explicitly describe the midpoints of two fuzzy biopolymers with respect to distances obtained by aggregating, through several suitable mappings, metrics on each position of the sequences represented by the fuzzy biopolymers.

Key words: Distance, metric, midpoint, fuzzy polynucleotide, profile, fuzzy polypeptide.

1 Introduction

There are many situations where it is useful to discern those fuzzy subsets that can be considered as "middle ways" or "compromises" between two given fuzzy subsets of a given finite set. For instance, a patient's symptoms profile can be described by means of a fuzzy subset of the set of attribute variables taken into account, and the ensemble of middle ways between two such descriptions

[1] This work has been partially supported by the Spanish DGES projects BFM2000-1113-C02-01 and BFM2003-00771.

of a given patient, for instance provided by two independent raters, can be used as a new representation of the patient [3].

These middle ways between two fuzzy subsets have been formalized by Nieto and Torres [16] by means of (fuzzy) midpoints. A *midpoint* of two fuzzy subsets μ, ν of a finite set X is any fuzzy subset ξ of X whose distance to μ and ν is exactly half the distance between these two fuzzy subsets. Of course, the actual content, and the range of application, of this definition depends on the chosen distance between fuzzy subsets of X [1,15].

If we consider the euclidean distance $d(\mu,\nu) = \sqrt{\sum_{x\in X}(\mu(x)-\nu(x))^2}$, then we know from euclidean geometry that any two fuzzy subsets μ and ν of X have one, and only one, midpoint: half their sum. For other distances, like the one introduced by Nieto et al. to compare polynucleotides [17], the Canberra metric [12] and Pappis-Karacapilidis metric [18], there exist pairs of fuzzy subsets without midpoints. Finally, there are well-known distances with respect to which most pairs of fuzzy subsets have infinitely many midpoints, like for instance the Hamming and maximum distances [3,16].

We propose in this paper a new application of midpoints with respect to suitable distances, as averages of fuzzy polynucleotides or fuzzy polypeptides of the same length. A *fuzzy polynucleotide*, or also a *fuzzy genome*,[2] of length N is simply a vector in Kosko's $4N$-dimensional hypercube $[0,1]^{4N}$, which can then be identified with a fuzzy subset of a $4N$-element set. Sadegh-Zadeh [21] introduced them as models of imprecisely known nucleic acid sequences of length N: the first four entries of the vector represent our knowledge of the extent to which the first base in the sequence is, respectively, an A, a C, a G or a T (U, in RNA); the next four entries represent the same for the base in the second position of the sequence; and so on. For instance, our knowledge of the composition of a *codon* (a DNA or RNA sequence of length $N = 3$) could be represented by a 12-dimensional vector

$$(1/3, 2/3, 0, 0, 0, 0, 1, 0, 1/3, 1/6, 1/3, 1/6).$$

In this vector we read that the first base in the codon cannot be a G, that the third base is C to the extent 1/6, and so on. This vector could be thought as representing the codons coding for the amino acid Arginine, as it assigns to each base and each position the relative frequency of this base at this position in these codons: CGA, CGC, CGG, CGU, AGA, AGG.

[2] Sadegh-Zadeh used the term “fuzzy genome” when he introduced this concept, but some colleagues have expressed to us that they feel its use misleading, as the word “genome” has the connotation of referring to the complete DNA sequence of an individual or a species, while the notion of fuzzy genome even includes single nucleotides.

In a similar way, a *fuzzy polypeptide* of length N would be a representation, in the sense described above, of an imprecisely known protein sequence of length N as a vector in $[0,1]^{20N}$. To simplify the language, we shall use the term *fuzzy biopolymer* to refer simultaneously to fuzzy polynucleotides and fuzzy polypeptides.

The fuzzy polynucleotide displayed above can be thought to belong to a class of fuzzy biopolymers that is often used in computational biology: the profiles. A *profile* is a representation of a group of related nucleic acid or protein sequences, usually based on a multiple alignment of these sequences [4]. Once the multiple alignment is defined, the profile is constructed by counting the occurrences of each monomer (bases in nucleic acids, amino acids in proteins) at each position along the multiple alignment and dividing these counts by the total number of sequences. Sometimes pseudocounts are introduced so that no monomer has a zero value assigned at any position, and other variations have also been used. For more information on profile derivation techniques, see [7,9]. A similar representation of multiple alignments are the *position specific scoring matrices*, *PSSMs* [8], where the numbers of monomers at each position are computed using position-based sequence weights and then they are normalized by the expected frequencies of the corresponding monomers. Profiles and PSSMs are usually given by means of matrices, with columns representing the positions in the sequence and the rows representing the monomers in some prefixed order. If we reorganize these matrices by concatenating the rows, considered as vectors of length 4 (in the nucleic acid setting) or 20 (in the protein setting), we obtain a fuzzy polynucleotide or a fuzzy polypeptide.

The comparison of profiles has been an important topic in computational biology. Some papers, like for instance [6,10,24], use iterative methods or dynamic programming algorithms for the construction of profile-profile alignments. Other papers, like [14,19], simply measure the profiles' similarity using some simple formulas like distances or correlation coefficients. Moreover, Torres and Nieto [22] have shown the interest of the comparison of fuzzy polynucleotides not necessarily arising from multiple alignments.

In this paper we consider very simple families of distances between fuzzy biopolymers, obtained by using a metric to measure the similarity at the level of each position of the content of the sequences represented by the fuzzy biopolymers, and then a suitable mapping to aggregate the values obtained in this way. Similar metrics have already been used by other authors to align profiles [10], and we believe that the asymmetrical definition of fuzzy biopolymers, where each consecutive group of a fixed number of consecutive entries represents a different position, motivates this kind of construction. In particular, this allows to assign different weights to the different positions in the sequences.

For the metrics we consider, we compute the midpoints of two fuzzy biopolymers of the same length to capture what could be considered an average of two such fuzzy biopolymers. In particular, we obtain explicit descriptions, as regions in a unit hypercube of a suitable dimension, of the sets of midpoints of two fuzzy biopolymers of the same length with respect to distances obtained by aggregating euclidean, Hamming or maximum distances on each position by means of a weighted sum, a weighted maximum mapping, or a weighted euclidean norm.

These sets of midpoints of fuzzy biopolymers can have several applications. A first, generic application of midpoints is the comparison of pairs of fuzzy subsets of a given set [3]. Given two pairs (μ_1, μ_2) and (ν_1, ν_2) of fuzzy subsets of a given set X, for instance two pairs of profiles of the same length, and a metric d on the set of fuzzy subsets of X, one can measure the similarity of (μ_1, μ_2) and (ν_1, ν_2) under d by computing the average value of d between the sets of midpoints of (μ_1, μ_2) and of (ν_1, ν_2) with respect to this metric. Another application of midpoints of fuzzy biopolymers could be the simple refinement of multiple alignments. More specifically, assume for instance that we are given two multiple alignments of the same length. Then any multiple alignment of the union of a subset of each one of these sets of sequences whose profile is a midpoint of the profiles of the original pair of alignments, could be understood as a consensus of this pair. We are currently working on this line of research, and we hope to report on it in a near future.

2 Preliminaries: midpoints and segments

Let us fix from now on a finite, n-element set

$$X = \{x_1, \ldots, x_n\};$$

we understand its elements ordered by the subscripts. Let $\mathcal{FP}(X)$ denote the set of its $[0,1]$-valued fuzzy subsets. To simplify the notations, given a fuzzy subset $\mu, \nu, \ldots$ of X, we shall write $\mu_i, \nu_j, \ldots$ instead of $\mu(x_i), \nu(x_j), \ldots$.

The mapping sending every $\mu \in \mathcal{FP}(X)$ to the vector $(\mu_1, \ldots, \mu_n) \in [0,1]^n$ is a bijection $\mathcal{FP}(X) \cong [0,1]^n$, and it allows to identify, in a one-to-one way, every fuzzy subset of X with a point of Kosko's n-dimensional hypercube [11]; hypercubical calculus has been described in [25]. To ease the language, in practice we shall systematically carry out this identification between fuzzy subsets and vectors, usually without any further notice.

This identification allows the translation of operations and distances on $[0,1]^n$ into operations and distances on $\mathcal{FP}(X)$. So, for instance, given $\mu, \nu \in \mathcal{FP}(X)$

and $t \in [0,1]$, by $t \cdot \mu + (1-t) \cdot \nu \in \mathcal{FP}(X)$ we denote the fuzzy subset of X defined by

$$(t \cdot \mu + (1-t) \cdot \nu)_i = t\mu_i + (1-t)\nu_i \quad \text{for every } i = 1, \ldots, n.$$

In this work we shall mainly be concerned with the following three basic distances on $\mathcal{FP}(X)$, which can be understood as translations of well known metrics on $[0,1]^n$:

- The *euclidean distance* $d_2(\mu,\nu) = \sqrt{\sum_{i=1}^n (\mu_i - \nu_i)^2}$.
- The *Hamming distance* $d_H(\mu,\nu) = \sum_{i=1}^n |\mu_i - \nu_i|$.
- The *maximum distance* $d_\infty(\mu,\nu) = \bigvee_{i=1}^n |\mu_i - \nu_i|$.

Let now d be any distance on $\mathcal{FP}(X)$. For every $\mu, \nu \in \mathcal{FP}(X)$, a fuzzy subset $\xi \in \mathcal{FP}(X)$ is a *midpoint* of μ and ν with respect to d if and only if

$$d(\xi,\mu) = d(\xi,\nu) = \frac{1}{2} d(\mu,\nu).$$

Let $\mathrm{mid}_d(\mu,\nu) \subseteq \mathcal{FP}(X)$ denote the set of all midpoints of μ and ν with respect to d.

The sets of midpoints with respect of several distances have been considered so far in the literature. In particular, they have been computed for the three distances introduced above. Specifically, for every $\mu, \nu \in \mathcal{FP}(X)$:

- It is well known from euclidean geometry that
$$\mathrm{mid}_{d_2}(\mu,\nu) = \{\frac{\mu+\nu}{2}\}.$$
- It is proved in [3,16] that, if we let
$$I_+ = \{i \mid \mu_i < \nu_i\}, \ I_- = \{i \mid \mu_i > \nu_i\},$$
then $\mathrm{mid}_{d_H}(\mu,\nu)$ consists of those $\xi \in \mathcal{FP}(X)$ that satisfy the following two conditions:
  - · $\min\{\mu_i,\nu_i\} \leq \xi_i \leq \max\{\mu_i,\nu_i\}$ for every $i = 1, \ldots, n$.
  - · $\sum_{i \in I_+} \xi_i - \sum_{i \in I_-} \xi_i = \sum_{i \in I_+} \frac{1}{2}(\mu_i + \nu_i) - \sum_{i \in I_-} \frac{1}{2}(\mu_i + \nu_i)$.

  In particular, if $\mu_i = \nu_i$ and $\xi \in \mathrm{mid}_{d_H}(\mu,\nu)$, then $\xi_i = \mu_i = \nu_i$.
- It is proved in [3] that $\mathrm{mid}_{d_\infty}(\mu,\nu)$ consists of those $\xi \in \mathcal{FP}(X)$ such that, for every $i = 1, \ldots, n$,
$$\max\{\mu_i,\nu_i\} - \frac{1}{2} d_\infty(\mu,\nu) \leq \xi_i \leq \min\{\mu_i,\nu_i\} + \frac{1}{2} d_\infty(\mu,\nu).$$
In particular, if $|\mu_i - \nu_i| = d_\infty(\mu,\nu)$ and $\xi \in \mathrm{mid}_{d_\infty}(\mu,\nu)$, then $\xi_i = (\mu_i + \nu_i)/2$.

Notice that

$$\mathrm{mid}_{d_2}(\mu,\nu) \subseteq \mathrm{mid}_{d_H}(\mu,\nu), \quad \mathrm{mid}_{d_2}(\mu,\nu) \subseteq \mathrm{mid}_{d_\infty}(\mu,\nu).$$

In general, these inclusions are strict and, furthermore, there is no relationship between $\mathrm{mid}_{d_H}(\mu,\nu)$ and $\mathrm{mid}_{d_\infty}(\mu,\nu)$.

Example 1. A vector with entries the frequencies of the four nucleotides A, C, G and T (in this order) in some specific region of a genome is a point of $[0,1]^4$ and hence it can be understood as a fuzzy subset of a 4-element set. For instance, and according to [22], these frequencies in the coding region of the *Mycobacterium tuberculosis H37Rv* are given by the vector

$$\begin{aligned}\mu_{MT} &= (\frac{702492}{3971522}, \frac{1283724}{3971522}, \frac{672608}{3971522}, \frac{1312698}{3971522}) \\ &\approx (0.1693, 0.3232, 0.3304, 0.1771),\end{aligned}$$

while these frequencies in the coding region of the *Escherichia coli K-12* are given by the vector

$$\begin{aligned}\mu_{EC} &= (\frac{985105}{4025952}, \frac{976160}{4025952}, \frac{976676}{4025952}, \frac{1088011}{4025952}) \\ &\approx (0.2425, 0.2424, 0.2704, 0.2447).\end{aligned}$$

Then (within this degree of approximation):

- $\mathrm{mid}_{d_2}(\mu_{MT},\nu_{EC})$ consists simply of

$$(0.2059, 0.2828, 0.3004, 0.2109) \in [0,1]^4.$$

- $\mathrm{mid}_{d_H}(\mu_{MT},\nu_{EC})$ is the intersection of the hyperprism

$$[0.1693, 0.2425] \times [0.2424, 0.3232] \times [0.2704, 0.3304] \times [0.1771, 0.2447]$$

  with the hyperplane defined by the equation

$$x_1 - x_2 - x_3 + x_4 = 0.2059 - 0.2828 - 0.3004 + 0.2109 = -0.1664$$

- Since $d_\infty(\mu_{MT},\nu_{EC}) = 0.0808$, we have that $\mathrm{mid}_{d_\infty}(\mu,\nu)$ is the 3-dimensional prism

$$[0.2021, 0.2097] \times \{0.2828\} \times [0.2900, 0.3108] \times [0.2043, 0.2175].$$

Thus, if we are using the maximum distance to compare these vectors of frequencies, then any vector of frequencies of nucleotides in the last prism (i.e., any vector lying in the intersection of this prism with the hyperplane $x_1+x_2+x_3+x_4=1$) can be understood as an average of the vectors corresponding to *M. tuberculosis* and *E. Coli*, while, if we use the euclidean distance, then only the usual midpoint defined as half-the-sum of the vectors can be so. Finally, the vectors of frequencies that could be understood as their average under the Hamming distance would be those in the intersection of

$$[0.1693, 0.2425] \times [0.2424, 0.3232] \times [0.2704, 0.3304] \times [0.1771, 0.2447]$$

with the plane of equations

$$x_1 + x_4 = 0.4168, \quad x_2 + x_3 = 0.5832$$

obtained by combining the equation $x_1 - x_2 - x_3 + x_4 = -0.1664$ with $x_1 + x_2 + x_3 + x_4 = 1$. □

We shall also use segments in our work. For every $\mu, \nu \in \mathcal{FP}(X)$, the *segment* defined by μ and ν with respect to a distance d on $\mathcal{FP}(X)$ is

$$\mathrm{seg}_d(\mu, \nu) = \{\xi \in \mathcal{FP}(X) \mid d(\mu, \xi) + d(\xi, \nu) = d(\mu, \nu)\}.$$

Notice that $\mathrm{mid}_d(\mu, \nu) \subseteq \mathrm{seg}_d(\mu, \nu)$. More specifically, we have the following easy result (see [16]).

Lemma 2. For every distance d on $\mathcal{FP}(X)$ and for every $\mu, \nu \in \mathcal{FP}(X)$,

$$\mathrm{mid}_d(\mu, \nu) = \{\xi \in \mathrm{seg}_d(\mu, \nu) \mid d(\xi, \mu) = d(\xi, \nu)\}.$$

□

The following segments are known:

- It is well known from euclidean geometry that, for every $\mu, \nu \in \mathcal{FP}(X)$,

$$\mathrm{seg}_{d_2}(\mu, \nu) = \{t \cdot \mu + (1 - t) \cdot \nu \mid t \in [0, 1]\}.$$

- It is proved in [16] that, for every $\mu, \nu \in \mathcal{FP}(X)$,

$$\mathrm{seg}_{d_H}(\mu, \nu) = \{\xi \mid \min\{\mu_i, \nu_i\} \leq \xi_i \leq \max\{\mu_i, \nu_i\} \text{ for every } i = 1, \ldots, n\}.$$

As far as segments with respect to d_∞ go, we have the following result.

Proposition 3. Let $\mu, \nu \in \mathcal{FP}(X)$. Then, for every $\xi \in \mathcal{FP}(X)$, the following conditions are equivalent:

(1) $\xi \in \mathrm{seg}_{d_\infty}(\mu, \nu)$.
(2) ξ satisfies the following conditions for every index $i_0 \in \{1, \ldots, n\}$ such that $d_\infty(\mu, \nu) = |\mu_{i_0} - \nu_{i_0}|$:
(i) $\min\{\mu_{i_0}, \nu_{i_0}\} \leq \xi_{i_0} \leq \max\{\mu_{i_0}, \nu_{i_0}\}$
(ii) $d_\infty(\mu, \xi) = |\mu_{i_0} - \xi_{i_0}|$ and $d_\infty(\nu, \xi) = |\nu_{i_0} - \xi_{i_0}|$.

(3) ξ satisfies conditions (i) and (ii) above for some index $i_0 \in \{1, \ldots, n\}$ such that $d_\infty(\mu, \nu) = |\mu_{i_0} - \nu_{i_0}|$

Proof. To prove the implication (1)$\Longrightarrow$(2), let us assume that $\xi \in \mathrm{seg}_{d_\infty}(\mu, \nu)$, i.e., that

$$\bigvee_{i=1}^{n} |\mu_i - \xi_i| + \bigvee_{i=1}^{n} |\nu_i - \xi_i| = \bigvee_{i=1}^{n} |\mu_i - \nu_i|,$$

and let $i_0 \in \{1, \ldots, n\}$ be any index such that $d_\infty(\mu, \nu) = |\mu_{i_0} - \nu_{i_0}|$. Then

$$\begin{aligned} |\mu_{i_0} - \nu_{i_0}| &\leq |\mu_{i_0} - \xi_{i_0}| + |\xi_{i_0} - \nu_{i_0}| \\ &\leq \textstyle\bigvee_{i=1}^{n} |\mu_i - \xi_i| + \bigvee_{i=1}^{n} |\nu_i - \xi_i| = \bigvee_{i=1}^{n} |\mu_i - \nu_i| = |\mu_{i_0} - \nu_{i_0}|, \end{aligned}$$

which implies that

$$|\mu_{i_0} - \nu_{i_0}| = |\mu_{i_0} - \xi_{i_0}| + |\xi_{i_0} - \nu_{i_0}|$$

and

$$|\mu_{i_0} - \xi_{i_0}| + |\xi_{i_0} - \nu_{i_0}| = \bigvee_{i=1}^{n} |\mu_i - \xi_i| + \bigvee_{i=1}^{n} |\nu_i - \xi_i|.$$

Now, the first equality is equivalent to

$$\min\{\mu_{i_0}, \nu_{i_0}\} \leq \xi_{i_0} \leq \max\{\mu_{i_0}, \nu_{i_0}\},$$

and, since

$$|\mu_{i_0} - \xi_{i_0}| \leq \bigvee_{i=1}^{n} |\mu_i - \xi_i|, \qquad |\xi_{i_0} - \nu_{i_0}| \leq \bigvee_{i=1}^{n} |\nu_i - \xi_i|,$$

the second equality is equivalent to

$$\bigvee_{i=1}^{n} |\mu_i - \xi_i| = |\mu_{i_0} - \xi_{i_0}| \text{ and } \bigvee_{i=1}^{n} |\nu_i - \xi_i| = |\xi_{i_0} - \nu_{i_0}|,$$

which completes the proof of this implication.

The implication (2)$\Longrightarrow$(3) is straightforward. Finally, as far as the implication (3)$\Longrightarrow$(1) goes, simply notice that if $i_0 \in \{1, \ldots, n\}$ is such that $d_\infty(\mu, \nu) = |\mu_{i_0} - \nu_{i_0}|$ and ξ satisfies conditions (i) and (ii) for this index, then

$$\begin{aligned} d_\infty(\mu, \xi) + d_\infty(\nu, \xi) &= |\mu_{i_0} - \xi_{i_0}| + |\nu_{i_0} - \xi_{i_0}| \qquad \text{(by (ii))} \\ &= |\mu_{i_0} - \nu_{i_0}| \qquad \text{(by (i))} \\ &= d_\infty(\mu, \nu). \end{aligned}$$

□

3 Midpoints for aggregations of distances

Let

$$X = X_1 \sqcup X_2 \sqcup \cdots \sqcup X_k$$

be a partition of the finite set X and, for every $j = 1, \ldots, k$, let d_j be a distance on $\mathcal{FP}(X_j)$. For every $\mu \in \mathcal{FP}(X)$, let $\mu^{(j)} \in \mathcal{FP}(X_j)$ denote henceforth the restriction $\mu|_{X_j}$.

Let $\Phi : (\mathbb{R}^+)^k \to [0, +\infty]$ be a non-decreasing mapping such that:

(A1) $\Phi(0, \ldots, 0) = 0$.
(A2) Φ is subadditive: if $c_i \leq a_i + b_i$ for every $i = 1, \ldots, k$, then

$$\Phi(c_1, \ldots, c_k) \leq \Phi(a_1, \ldots, a_k) + \Phi(b_1, \ldots, b_k).$$

(A3) $\Phi(c_1, \ldots, c_k) = 0$ implies $c_1 = \cdots = c_k = 0$.

Then the mapping

$$d : \mathcal{FP}(X) \times \mathcal{FP}(X) \to \mathbb{R}^+$$

defined, for every $\mu, \nu \in \mathcal{FP}(X)$, by

$$d(\mu, \nu) = \Phi(d_1(\mu^{(1)}, \nu^{(1)}), \ldots, d_k(\mu^{(k)}, \nu^{(k)}))$$

is a distance on $\mathcal{FP}(X)$, called the *aggregation* of $d_1, \ldots, d_k$ through Φ and which we shall denote from now on by $\Phi(d_1, \ldots, d_k)$; see [2,20].

Let $\omega = (\omega_1, \ldots, \omega_k) \in (\mathbb{R}^+)^k$ be any vector of positive weights. It is easy to check that conditions (A1) to (A3) above are satisfied, among others, by the following three mappings:

$$\Phi_{2,\omega}(c_1, \ldots, c_k) = \sqrt{\textstyle\sum_{i=1}^k \omega_i c_i^2}, \qquad \Phi_{H,\omega}(c_1, \ldots, c_k) = \textstyle\sum_{i=1}^k \omega_i c_i,$$
$$\Phi_{\infty,\omega}(c_1, \ldots, c_k) = \textstyle\bigvee_{i=1}^k \omega_i c_i.$$

We have now the following results on midpoints with respect to aggregations of distances on parts of X through these three mappings.

Proposition 4. Let D be a distance on $\mathcal{FP}(X)$ of the form

$$\Phi_{2,\omega}(d_1, \ldots, d_k)$$

for some $\omega = (\omega_1, \ldots, \omega_k) \in (\mathbb{R}^+)^k$ and some distances d_j on the subsets X_j of X, $j = 1, \ldots, k$. Then, for every $\mu, \nu \in \mathcal{FP}(X)$, the set $\mathrm{mid}_D(\mu, \nu)$ consists of those fuzzy subsets $\xi \in \mathcal{FP}(X)$ such that $\xi^{(j)} \in \mathrm{mid}_{d_j}(\mu^{(j)}, \nu^{(j)})$ for every $j = 1, \ldots, k$.

Proof. Let $\xi \in \mathrm{mid}_D(\mu,\nu)$. Then $\xi \in \mathrm{seg}_D(\mu,\nu)$, i.e.,

$$\sqrt{\sum_{j=1}^{k} \omega_j d_j(\mu^{(j)},\xi^{(j)})^2} + \sqrt{\sum_{j=1}^{k} \omega_j d_j(\nu^{(j)},\xi^{(j)})^2} = \sqrt{\sum_{j=1}^{k} \omega_j d_j(\mu^{(j)},\nu^{(j)})^2}.$$

Since

$$d_j(\mu^{(j)},\nu^{(j)}) \leq d_j(\mu^{(j)},\xi^{(j)}) + d_j(\nu^{(j)},\xi^{(j)}) \quad \text{for every } j=1,\ldots,k,$$

arguing as in usual proof of the triangular inequality for the euclidean norm and the characterization of the cases when it becomes an equality explained in any first undergraduate linear algebra course (cf., for instance, [13, I, §4] or [23, §7.1]), we obtain that

$$d_j(\mu^{(j)},\nu^{(j)}) = d_j(\mu^{(j)},\xi^{(j)}) + d_j(\nu^{(j)},\xi^{(j)}) \quad \text{for every } j=1,\ldots,k,$$

and that if $\xi \neq \mu,\nu$, then there exists some $t>0$ such that

$$d_j(\nu^{(j)},\xi^{(j)}) = t \cdot d_j(\mu^{(j)},\xi^{(j)}) \quad \text{for every } j=1,\ldots,k.$$

When we impose moreover that $D(\mu,\xi)=D(\nu,\xi)$, i.e.,

$$\sqrt{\sum_{j=1}^{k} \omega_j d_j(\mu^{(j)},\xi^{(j)})^2} = \sqrt{\sum_{j=1}^{k} \omega_j d_j(\nu^{(j)},\xi^{(j)})^2},$$

we obtain that either $\xi=\mu=\nu$ or $t=1$. In all, ξ satisfies that

$$\left.\begin{array}{l} d_j(\mu^{(j)},\nu^{(j)}) = d_j(\mu^{(j)},\xi^{(j)}) + d_j(\nu^{(j)},\xi^{(j)}) \\ d_j(\nu^{(j)},\xi^{(j)}) = d_j(\mu^{(j)},\xi^{(j)}) \end{array}\right\} \quad \text{for every } j=1,\ldots,k,$$

i.e., that $\xi^{(j)} \in \mathrm{mid}_{d_j}(\mu^{(j)},\nu^{(j)})$ for every $j=1,\ldots,k$. This proves the "only if" implication. The "if" implication is straightforward. □

Thus, roughly speaking, all midpoints of two fuzzy subsets μ and ν of X with respect to a distance of the form $\Phi_{2,\omega}(d_1,\ldots,d_k)$ are obtained by concatenating midpoints of the restrictions of μ and ν to each X_j with respect to the corresponding distance d_j.

Proposition 5. Let D be a distance on $\mathcal{FP}(X)$ of the form

$$\Phi_{H,\omega}(d_1,\ldots,d_k)$$

for some $\omega=(\omega_1,\ldots,\omega_k)\in(\mathbb{R}^+)^k$ and some distances d_j on the subsets X_j of X, $j=1,\ldots,k$. Then, for every $\mu,\nu\in\mathcal{FP}(X)$, the set $\mathrm{mid}_D(\mu,\nu)$ consists of those fuzzy subsets $\xi\in\mathcal{FP}(X)$ that satisfy the following two conditions:

(i) $\xi^{(j)} \in \operatorname{seg}_{d_j}(\mu^{(j)}, \nu^{(j)})$, for every $j = 1, \ldots, k$.

(ii) $\sum_{j=1}^{k} \omega_j d_j(\mu^{(j)}, \xi^{(j)}) = \sum_{j=1}^{k} \omega_j d_j(\nu^{(j)}, \xi^{(j)})$.

Proof. By Lemma 2, we have that $\xi \in \operatorname{mid}_D(\mu, \nu)$ if and only if $D(\mu, \xi) + D(\nu, \xi) = D(\mu, \nu)$ and $D(\mu, \xi) = D(\nu, \xi)$. This second condition is directly equivalent to (ii). As far as the first condition goes, it says

$$\sum_{j=1}^{k} \omega_j d_j(\mu^{(j)}, \xi^{(j)}) + \sum_{j=1}^{k} \omega_j d_j(\nu^{(j)}, \xi^{(j)}) = \sum_{j=1}^{k} \omega_j d_j(\mu^{(j)}, \nu^{(j)}).$$

Since

$$d_j(\mu^{(j)}, \nu^{(j)}) \leq d_j(\mu^{(j)}, \xi^{(j)}) + d_j(\nu^{(j)}, \xi^{(j)}) \qquad \text{for every } j = 1, \ldots, k,$$

and $\omega_j > 0$ for every j, this equality is equivalent to

$$d_j(\mu^{(j)}, \xi^{(j)}) + d_j(\nu^{(j)}, \xi^{(j)}) = d_j(\mu^{(j)}, \nu^{(j)}),$$

i.e., to $\xi^{(j)} \in \operatorname{seg}_{d_j}(\mu^{(j)}, \nu^{(j)})$, for every $j = 1, \ldots, k$. □

Proposition 6. Let D be a distance on $\mathcal{FP}(X)$ of the form

$$\Phi_{\infty,\omega}(d_1, \ldots, d_k)$$

for some $\omega = (\omega_1, \ldots, \omega_k) \in (\mathbb{R}^+)^k$ and some distances d_j on the subsets X_j of X, $j = 1, \ldots, k$. Then, for every $\mu, \nu \in \mathcal{FP}(X)$, the set $\operatorname{mid}_D(\mu, \nu)$ consists of those fuzzy subsets $\xi \in \mathcal{FP}(X)$ that satisfy the following condition: if $j_0 \in \{1, \ldots, k\}$ is such that

$$\omega_{j_0} d_{j_0}(\mu^{(j_0)}, \nu^{(j_0)}) \geq \omega_j d_j(\mu^{(j)}, \nu^{(j)}) \qquad \text{for every } j = 1, \ldots, k,$$

then

(i) $\xi^{(j_0)} \in \operatorname{mid}_{d_{j_0}}(\mu^{(j_0)}, \nu^{(j_0)})$.

(ii) $\omega_j d_j(\mu^{(j)}, \xi^{(j)}) \leq \omega_{j_0} d_{j_0}(\mu^{(j_0)}, \xi^{(j_0)})$ and $\omega_j d_j(\nu^{(j)}, \xi^{(j)}) \leq \omega_{j_0} d_{j_0}(\nu^{(j_0)}, \xi^{(j_0)})$ for every $j = 1, \ldots, k$.

Proof. Assume that $\xi \in \mathcal{FP}(X)$ is such that

$$(*) \qquad \begin{aligned} \bigvee_{j=1}^{k} \omega_j d_j(\xi^{(j)}, \mu^{(j)}) &= \bigvee_{j=1}^{k} \omega_j d_j(\xi^{(j)}, \nu^{(j)}) \\ &= \tfrac{1}{2} \bigvee_{j=1}^{k} \omega_j d_j(\mu^{(j)}, \nu^{(j)}) = \tfrac{1}{2} \omega_{j_0} d_{j_0}(\mu^{(j_0)}, \nu^{(j_0)}). \end{aligned}$$

In particular,

$$\begin{aligned} \omega_{j_0} d_{j_0}(\mu^{(j_0)}, \xi^{(j_0)}) &\leq \tfrac{1}{2} \omega_{j_0} d_{j_0}(\mu^{(j_0)}, \nu^{(j_0)}), \\ \omega_{j_0} d_{j_0}(\nu^{(j_0)}, \xi^{(j_0)}) &\leq \tfrac{1}{2} \omega_{j_0} d_{j_0}(\mu^{(j_0)}, \nu^{(j_0)}). \end{aligned}$$

Now,

$$\begin{aligned}\omega_{j_0}d_{j_0}(\mu^{(j_0)},\nu^{(j_0)}) &\leq \omega_{j_0}d_{j_0}(\mu^{(j_0)},\xi^{(j_0)})+\omega_{j_0}d_{j_0}(\nu^{(j_0)},\xi^{(j_0)})\\ &\leq \tfrac{1}{2}\omega_{j_0}d_{j_0}(\mu^{(j_0)},\nu^{(j_0)})+\tfrac{1}{2}\omega_{j_0}d_{j_0}(\mu^{(j_0)},\nu^{(j_0)})=\omega_{j_0}d_{j_0}(\mu^{(j_0)},\nu^{(j_0)})\end{aligned}$$

and, since $\omega_{j_0}>0$, this entails that

$$d_{j_0}(\mu^{(j_0)},\xi^{(j_0)})=d_{j_0}(\nu^{(j_0)},\xi^{(j_0)})=\frac{1}{2}d_{j_0}(\mu^{(j_0)},\nu^{(j_0)}),$$

which is point (i) in the statement. Moreover, by (*), this implies that

$$\omega_{j_0}d_{j_0}(\mu^{(j_0)},\xi^{(j_0)})=\bigvee_{j=1}^{k}\omega_j d_j(\xi^{(j)},\mu^{(j)}),\ \omega_{j_0}d_{j_0}(\nu^{(j_0)},\xi^{(j_0)})=\bigvee_{j=1}^{k}\omega_j d_j(\xi^{(j)},\nu^{(j)}),$$

which is equivalent to point (ii).

This proves the "only if" implication. As far as the converse implication goes, condition (ii) says that

$$D(\mu,\xi)=\omega_{j_0}d_{j_0}(\mu^{(j_0)},\xi^{(j_0)})\qquad\text{and}\qquad D(\nu,\xi)=\omega_{j_0}d_{j_0}(\nu^{(j_0)},\xi^{(j_0)}),$$

and then condition (i) entails that

$$\omega_{j_0}d_{j_0}(\mu^{(j_0)},\xi^{(j_0)})=\omega_{j_0}d_{j_0}(\nu^{(j_0)},\xi^{(j_0)})=\frac{1}{2}\omega_{j_0}d_{j_0}(\mu^{(j_0)},\nu^{(j_0)})=D(\mu,\nu).$$

□

Actually, the proof of the last proposition proves that if $\xi\in\mathrm{mid}_D(\mu,\nu)$, then it satisfies conditions (i) and (ii) *for every* j_0 such that $\omega_{j_0}d_{j_0}(\mu^{(j_0)},\nu^{(j_0)})=D(\mu,\nu)$, while if ξ satisfies these conditions *for some* j_0 such that $D(\mu,\nu)=\omega_{j_0}d_{j_0}(\mu^{(j_0)},\nu^{(j_0)})$, then $\xi\in\mathrm{mid}_D(\mu,\nu)$.

If we take $k=n$, $X_i=\{x_i\}$, $\omega=(1,\ldots,1)$ and the distances d_i defined by the absolute value of the difference of the images, then these propositions entail the descriptions of the sets of midpoints with respect to the euclidean, Hamming and maximum distances recalled in Section 2. On the other hand, taking k, each X_i and each d_i as before but now $\omega\in(\mathbb{R}^+)^k$ arbitrary, we obtain descriptions of the sets of midpoints with respect to the *weighted* euclidean, Hamming and maximum distances: the weighted euclidean case is well-known from elementary algebra (still half their sum), while the weighted Hamming and maximum cases were already described in [3].

4 Midpoints of fuzzy words

Let Σ be any alphabet. A *fuzzy letter* in Σ is a fuzzy subset of it, i.e. a mapping $\mu : \Sigma \to [0,1]$. If Σ has m elements, then, after fixing an ordering of these elements, every fuzzy letter in it can be understood as a point in the unit hypercube $[0,1]^m$: if $\Sigma = \{L_1, \ldots, L_m\}$, with the order given by the subscripts, then we can identify a fuzzy letter $\mu : \Sigma \to [0,1]$ with the point

$$(\mu(L_1), \ldots, \mu(L_m)) \in [0,1]^m.$$

Now, a *fuzzy word* of length N over the alphabet Σ is an element of $\mathcal{FP}(\Sigma)^N$, i.e., a sequence $\underline{\mu} = (\mu^{(1)}, \ldots, \mu^{(N)})$ of fuzzy letters in Σ.

If we concatenate the fuzzy letters in Σ, or rather their representations as vectors in $[0,1]^m$, of a fuzzy word of length N in the order as they appear in it, we obtain a representation of this fuzzy word as an element of $[0,1]^{mN}$: namely,

$$(\mu^{(1)}(L_1), \ldots, \mu^{(1)}(L_m), \mu^{(2)}(L_1), \ldots, \mu^{(N)}(L_m)).$$

A fuzzy word $\underline{\mu}$ of length N can be identified with a fuzzy subset of an mN-element set

$$X = \{x_{i,j} \mid i = 1, \ldots, N,\ j = 1, \ldots, m\},$$

which we shall still denote by $\underline{\mu}$: if $\underline{\mu} = (\mu^{(1)}, \ldots, \mu^{(N)})$, then we simply define $\underline{\mu}(x_{i,j}) = \mu^{(i)}(L_j)$ for every $i = 1, \ldots, N$ and $j = 1, \ldots, m$. To simplify the notations, we shall denote henceforth each $\mu^{(i)}(L_j)$ by $\mu_j^{(i)}$.

Special cases of fuzzy words have already made their appearance in computational biology. *Profiles* [4] and *position specific scoring matrices* [8] of ensembles of nucleic acid or protein sequences, based on a multiple alignment of them, can be understood as fuzzy words over the corresponding alphabet of monomers. More in general, Sadegh-Zadeh introduced in [21] the fuzzy words over the alphabet $\Sigma = \{A, C, G, T\}$ of nucleotides, and he called them *fuzzy genomes* or also *fuzzy polynucleotides*, as a way to represent imprecisely known DNA sequences beyond frequencies or probability distributions. In a similar way, a *fuzzy polypeptide* can be defined as a fuzzy word over the 20-letter alphabet of amino acids. In general, we define a *fuzzy biopolymer* as a fuzzy word over some biologically relevant alphabet of monomers: fuzzy polynucleotides and fuzzy polypeptides are special cases of fuzzy biopolymers.

Torres and Nieto have shown in [22] the interest of the comparison of fuzzy polynucleotides, and they used with this purpose a distance they introduced in [17]. Since fuzzy words are defined by concatenating fuzzy letters, we consider that to measure the dissimilarity of fuzzy words of a fixed length N over Σ it makes sense to compute first the dissimilarity at the level of the fuzzy letters in the same places in both words, and then to aggregate the values obtained in this way: to sum them, to take their maximum, ... Metrics defined in a similar way have already been used in profile alignment [10]. Thus, to compare fuzzy words we can use distances obtained by aggregating, in the sense of the previous section, distances on each fuzzy letter. We do it here for simple metrics and simple aggregation mappings.

For every $j = 1, \ldots, N$, let $X_j = \{x_{j,1}, \ldots, x_{j,m}\}$, so that $X = X_1 \sqcup \cdots \sqcup X_N$. In the sequel we shall consider aggregations through $\Phi_{2,\omega}$, $\Phi_{H,\omega}$ or $\Phi_{\infty,\omega}$ of euclidean, Hamming or maximum distances on these sets X_j. More specifically, for every $\omega = (\omega_1, \ldots, \omega_N) \in (\mathbb{R}^+)^N$ and for every $a, b \in \{2, H, \infty\}$, let $D_{a,b,\omega}$ denote the distance on X obtained by aggregating distances of type d_b on each X_j through the mapping $\Phi_{a,\omega}$. For instance,

$$\begin{aligned} D_{H,\infty,\omega}(\underline{\mu}, \underline{\nu}) &= \textstyle\sum_{j=1}^N \omega_j (\bigvee_{i=1}^m |\mu_i^{(j)} - \nu_i^{(j)}|) \\ D_{\infty,H,\omega}(\underline{\mu}, \underline{\nu}) &= \textstyle\bigvee_{j=1}^N \omega_j (\sum_{i=1}^m |\mu_i^{(j)} - \nu_i^{(j)}|) \end{aligned}$$

When $\omega = (1, \ldots, 1)$, we shall simply write $D_{a,b}$ instead of $D_{a,b,\omega}$. Notice that if $a = b$ and $\omega = (1, \ldots, 1)$, then $D_{a,b}$ is nothing but the metric d_a defined on the whole X: for instance,

$$D_{H,H}(\underline{\mu}, \underline{\nu}) = \sum_{j=1}^N \Big(\sum_{i=1}^m |\mu_i^{(j)} - \nu_i^{(j)}| \Big) = \sum_{\substack{j=1,\ldots,N \\ i=1,\ldots,m}} |\mu_i^{(j)} - \nu_i^{(j)}|.$$

More in general, distances of the form $D_{a,a,\omega}$ are simply weighted versions of the distance d_a. Therefore, in the sequel we shall only consider the case $a \neq b$. One could also consider aggregations of weighted euclidean, Hamming or maximum distances on the sets X_j. For simplicity we do not consider this case here, but it should be clear to the reader that the corresponding results are straightforward generalizations of the results given below for aggregations of non-weighted distances.

Now, as direct consequences of the propositions established in the last section and the descriptions of segments and sets of midpoints provided in Section 2, we obtain the following results. We begin with the distances of the form $D_{2,b,\omega}$, for $b = H, \infty$. In this case, and as we already pointed out after Proposition 4, the midpoints of two fuzzy words $\underline{\mu}$ and $\underline{\nu}$ of length N with respect to a distance $D_{2,b,\omega}$ are obtained by taking, for every $j = 1, \ldots, N$, a midpoint of the fuzzy letters $\mu^{(j)}$ and $\nu^{(j)}$ with respect to the corresponding distance d_b, and then concatenating them in their order.

Corollary 7. For every $\underline{\mu}, \underline{\nu} \in \mathcal{FP}(\Sigma)^N$ and for every $j = 1, \ldots, N$, let

$$I_+^{(j)} = \{i \mid \mu_i^{(j)} < \nu_i^{(j)}\}, \ I_-^{(j)} = \{i \mid \mu_i^{(j)} > \nu_i^{(j)}\}.$$

Then, for every $\underline{\xi} \in \mathcal{FP}(\Sigma)^N$, $\underline{\xi} \in \mathrm{mid}_{D_{2,H,\omega}}(\underline{\mu},\underline{\nu})$ if and only if it satisfies the following two conditions:

(i) For every $j = 1,\ldots,N$ and $i = 1,\ldots,m$,

$$\min\{\mu_i^{(j)},\nu_i^{(j)}\} \le \xi_i^{(j)} \le \max\{\mu_i^{(j)},\nu_i^{(j)}\}.$$

(ii) For every $j = 1,\ldots,N$,

$$\sum_{i\in I_+^{(j)}} \xi_i^{(j)} - \sum_{i\in I_-^{(j)}} \xi_i^{(j)} = \sum_{i\in I_+^{(j)}} \frac{1}{2}(\mu_i^{(j)}+\nu_i^{(j)}) - \sum_{i\in I_-^{(j)}} \frac{1}{2}(\mu_i^{(j)}+\nu_i^{(j)}).$$

□

Corollary 8. For every $\underline{\mu},\underline{\nu},\underline{\xi} \in \mathcal{FP}(\Sigma)^N$, $\underline{\xi} \in \mathrm{mid}_{D_{2,\infty,\omega}}(\underline{\mu},\underline{\nu})$ if and only if it satisfies the following condition: for every $j = 1,\ldots,N$ and for every $i = 1,\ldots,m$,

$$\max\{\mu_i^{(j)},\nu_i^{(j)}\} - \frac{1}{2}d_\infty(\mu^{(j)},\nu^{(j)}) \le \xi_i^{(j)} \le \min\{\mu_i^{(j)},\nu_i^{(j)}\} + \frac{1}{2}d_\infty(\mu^{(j)},\nu^{(j)}).$$

□

It is straightforward to check that $\mathrm{mid}_{D_{2,b,\omega}}(\underline{\mu},\underline{\nu}) \subseteq \mathrm{mid}_{d_b}(\underline{\mu},\underline{\nu})$, for $b = H,\infty$. These inclusions are, in general, strict if $N \ge 2$, as the following simple example shows.

Example 9. Let $N = 2$ and $m = 1$, so that we can identify every fuzzy word $\underline{\mu}$ of length 2 with the 2-dimensional vector $(\mu_1^{(1)},\mu_1^{(2)})$. Let $\underline{\mu} = (1,0.6)$ and $\underline{\nu} = (0,0.4)$. Then

$$\begin{aligned}
&\mathrm{mid}_{D_{2,H}}(\underline{\mu},\underline{\nu}) = \mathrm{mid}_{D_{2,\infty}}(\underline{\mu},\underline{\nu}) = \{(0.5,0.5)\}\\
&\mathrm{mid}_{d_H}(\underline{\mu},\underline{\nu}) = \{(s,t)\in[0,1]^2 \mid s+t=1\}\\
&\mathrm{mid}_{d_\infty}(\underline{\mu},\underline{\nu}) = \{0.5\}\times[0.1,0.9]
\end{aligned}$$

□

As far as the midpoints with respect to distances of the form $D_{H,b,\omega}$ go, for $b = 2,\infty$, we have the following two results.

Corollary 10. For every $\underline{\mu},\underline{\nu},\underline{\xi} \in \mathcal{FP}(\Sigma)^N$, $\underline{\xi} \in \mathrm{mid}_{D_{H,2,\omega}}(\mu,\nu)$ if and only if

$$\xi^{(j)} = t_j\cdot\mu^{(j)} + (1-t_j)\cdot\nu^{(j)} \qquad \text{for every } j = 1,\ldots,N,$$

for some $(t_j)_{j=1,\dots,N} \in [0,1]^N$ such that

$$\sum_{j=1}^{N}(t_j - \frac{1}{2})\omega_j d_2(\mu^{(j)}, \nu^{(j)}) = 0.$$

Proof. As a direct application of Proposition 5 and the description of segments with respect of the euclidean distance given in Section 2, we obtain that $\underline{\xi} \in \mathrm{mid}_{D_{H,2,\omega}}(\mu, \nu)$ if and only if there exists, for every $j = 1, \dots, N$, some $t_j \in [0,1]$ such that $\xi^{(j)} = t_j \cdot \mu^{(j)} + (1 - t_j) \cdot \nu^{(j)}$, and

$$(*) \qquad \textstyle\sum_{j=1}^{N} \omega_j d_2(\mu^{(j)}, \xi^{(j)}) = \sum_{j=1}^{N} \omega_j d_2(\nu^{(j)}, \xi^{(j)}).$$

Now, if $\xi^{(j)} = t_j \cdot \mu^{(j)} + (1 - t_j) \cdot \nu^{(j)}$, then $d_2(\mu^{(j)}, \xi^{(j)}) = (1 - t_j) d_2(\mu^{(j)}, \nu^{(j)})$ and $d_2(\nu^{(j)}, \xi^{(j)}) = t_j d_2(\mu^{(j)}, \nu^{(j)})$, and hence (*) is equivalent to the equality $\sum_{j=1}^{N}(t_j - \frac{1}{2})\omega_j d_2(\mu^{(j)}, \nu^{(j)}) = 0$ given in the statement. □

Corollary 11. For every $\underline{\mu}, \underline{\nu} \in \mathcal{FP}(\Sigma)^N$ and for every $j = 1, \dots, N$, let $\ell_j \in \{1, \dots, m\}$ be such that $d_\infty(\mu^{(j)}, \nu^{(j)}) = |\mu^{(j)}_{\ell_j} - \nu^{(j)}_{\ell_j}|$, and let

$$J_+ = \{j \mid \mu^{(j)}_{\ell_j} < \nu^{(j)}_{\ell_j}\},\ J_- = \{j \mid \mu^{(j)}_{\ell_j} > \nu^{(j)}_{\ell_j}\}.$$

Then, for every $\underline{\xi} \in \mathcal{FP}(\Sigma)^N$, $\underline{\xi} \in \mathrm{mid}_{D_{H,\infty,\omega}}(\mu, \nu)$ if and only if it satisfies the following three conditions:

(i) $\min\{\mu^{(j)}_{\ell_j}, \nu^{(j)}_{\ell_j}\} \leq \xi^{(j)}_{\ell_j} \leq \max\{\mu^{(j)}_{\ell_j}, \nu^{(j)}_{\ell_j}\}$ for every $j = 1, \dots, N$.

(ii) For every $j = 1, \dots, N$ and for every $i = 1, \dots, m$,

$$|\mu^{(j)}_i - \xi^{(j)}_i| \leq |\mu^{(j)}_{\ell_j} - \xi^{(j)}_{\ell_j}| \text{ and } |\nu^{(j)}_i - \xi^{(j)}_i| \leq |\nu^{(j)}_{\ell_j} - \xi^{(j)}_{\ell_j}|.$$

(iii) $\sum_{j\in J_+} \omega_j \xi^{(j)}_{\ell_j} - \sum_{j\in J_-} \omega_j \xi^{(j)}_{\ell_j} = \sum_{j\in J_+} \frac{\omega_j}{2}(\mu^{(j)}_{\ell_j} + \nu^{(j)}_{\ell_j}) - \sum_{j\in J_-} \frac{\omega_j}{2}(\mu^{(j)}_{\ell_j} + \nu^{(j)}_{\ell_j})$.

Proof. It is a direct application of Proposition 5, the description of segments with respect to the maximum distance given in Proposition 3, and the fact that, if $(\xi^{(1)}_{\ell_1}, \dots, \xi^{(N)}_{\ell_N})$ satisfies condition (i) in the statement, then

$$\begin{aligned}
&\textstyle\sum_{j=1}^{N} \omega_j |\mu^{(j)}_{\ell_j} - \xi^{(j)}_{\ell_j}| - \sum_{j=1}^{N} \omega_j |\nu^{(j)}_{\ell_j} - \xi^{(j)}_{\ell_j}| \\
&\quad = 2(\textstyle\sum_{j\in J_+} \omega_j \xi^{(j)}_{\ell_j} - \sum_{j\in J_-} \omega_j \xi^{(j)}_{\ell_j}) \\
&\qquad - (\textstyle\sum_{j\in J_+} \omega_j(\mu^{(j)}_{\ell_j} + \nu^{(j)}_{\ell_j}) - \sum_{j\in J_-} \omega_j(\mu^{(j)}_{\ell_j} + \nu^{(j)}_{\ell_j})).
\end{aligned}$$

□

Finally, as far as the midpoints with respect to distances of the form $D_{\infty,b}$ go, for $b = 2, H$, the following two results are direct consequences of Proposition 6 and the descriptions of sets of midpoints in Section 2.

Corollary 12. For every $\mu, \nu \in \mathcal{FP}(\Sigma)^N$, let $j_0 \in \{1, \dots, N\}$ be such that

$$\omega_{j_0} d_2(\mu^{(j_0)}, \nu^{(j_0)}) \geq \omega_j d_2(\mu^{(j)}, \nu^{(j)}) \qquad \text{for every } j = 1, \dots, N.$$

Then, for every $\underline{\xi} \in \mathcal{FP}(\Sigma)^N$, $\underline{\xi} \in \mathrm{mid}_{D_{\infty,2,\omega}}(\mu, \nu)$ if and only if it satisfies the following two conditions:

(i) $\xi^{(j_0)} = (\mu^{(j_0)} + \nu^{(j_0)})/2$.

(ii) For every $j = 1, \dots, N$,

$$\omega_j d_2(\mu^{(j)}, \xi^{(j)}) \leq \omega_{j_0} d_2(\mu^{(j_0)}, \xi^{(j_0)}) \text{ and } \omega_j d_2(\nu^{(j)}, \xi^{(j)}) \leq \omega_{j_0} d_2(\nu^{(j_0)}, \xi^{(j_0)}).$$

□

Corollary 13. For every $\mu, \nu \in \mathcal{FP}(\Sigma)^N$, let $j_0 \in \{1, \dots, N\}$ be such that

$$\omega_{j_0} d_H(\mu^{(j_0)}, \nu^{(j_0)}) \geq \omega_j d_H(\mu^{(j)}, \nu^{(j)}) \qquad \text{for every } j = 1, \dots, N,$$

and set

$$I_+^{(j_0)} = \{i \mid \mu_i^{(j_0)} < \nu_i^{(j_0)}\},\ I_-^{(j_0)} = \{i \mid \mu_i^{(j_0)} > \nu_i^{(j_0)}\}.$$

Then, for every $\underline{\xi} \in \mathcal{FP}(\Sigma)^N$, $\underline{\xi} \in \mathrm{mid}_{D_{\infty,H,\omega}}(\mu, \nu)$ if and only if it satisfies the following three conditions:

(i) $\min\{\mu_i^{(j_0)}, \nu_i^{(j_0)}\} \leq \xi_i^{(j_0)} \leq \max\{\mu_i^{(j_0)}, \nu_i^{(j_0)}\}$ for every $i = 1, \dots, m$.

(ii) $\displaystyle\sum_{i \in I_+^{(j_0)}} \xi_i^{(j_0)} - \sum_{i \in I_-^{(j_0)}} \xi_i^{(j_0)} = \sum_{i \in I_+^{(j_0)}} \frac{1}{2}(\mu_i^{(j_0)} + \nu_i^{(j_0)}) - \sum_{i \in I_-^{(j_0)}} \frac{1}{2}(\mu_i^{(j_0)} + \nu_i^{(j_0)})$.

(iii) For every $j = 1, \dots, N$,

$$\omega_j d_H(\mu^{(j)}, \xi^{(j)}) \leq \omega_{j_0} d_H(\mu^{(j_0)}, \xi^{(j_0)}) \text{ and } \omega_j d_H(\nu^{(j)}, \xi^{(j)}) \leq \omega_{j_0} d_H(\nu^{(j_0)}, \xi^{(j_0)}).$$

□

Let us finish the main body of this paper with some specific examples of computations of midpoints of fuzzy biopolymers.

Example 14. Torres and Nieto computed in [22] the frequencies of the nucleotides A, C, G and T at the three base sites of a codon in the coding section of two bacteria, *Mycobacterium tuberculosis H37Rv* and *Escherichia coli K-12.* In this way they associated to each organism a fuzzy word of length 3 over the alphabet of nucleotides $\Sigma = \{A, C, G, T\}$. These words (presented as vectors in $[0,1]^{12}$ and approximated to only 4 digits for simplicity) are

$$\begin{aligned}
\underline{\mu}_{MT} &= (0.1724, 0.3089, 0.3556, 0.1632, 0.1763, 0.3145, 0.3056, 0.2036,\\
&\qquad 0.1593, 0.3461, 0.3302, 0.1645)\\
\underline{\mu}_{EC} &= (0.2600, 0.2420, 0.3374, 0.1605, 0.2846, 0.2286, 0.1752, 0.3116,\\
&\qquad 0.1831, 0.2568, 0.2981, 0.2619)
\end{aligned}$$

What could be understood as an average of these descriptions? The answer would depend on the metric used to compare fuzzy words of this kind: once fixed a metric, we can compute the sets of midpoints of these two fuzzy polynucleotides with respect to this metric and any fuzzy word in this set could be considered as such an average within this range of approximation. If, moreover, this vector is such that its entries 1 to 4 add up 1, as well as its entries 5 to 8 and 9 to 12, then it would also be a vector of frequencies of A, C, G and T at the three base sites of a codon.

As a way of example, we shall only consider the distances $D_{H,\infty}$ (the sum of maximum differences) and $D_{\infty,H}$ (the maximum of the sums of differences), in both cases, for simplicity, with vector of weights $\omega = (1,1,1)$, although, as already Torres and Nieto mention in the Conclusions section of [22], it could be interesting to introduce weights that represent the different roles played by the first two positions in a codon and the last one [5].

- A simple computation gives that $D_{H,\infty}(\underline{\mu}_{MT}, \underline{\mu}_{EC}) = 0.293$ and, using Corollary 11, that $\mathrm{mid}_{D_{H,\infty}}(\underline{\mu}_{MT}, \underline{\mu}_{EC})$ consists of those vectors $(x_1, \ldots, x_{12}) \in [0,1]^{12}$ such that, on the one hand,

$$0.1724 \leq x_1 \leq 0.26, \quad 0.1752 \leq x_7 \leq 0.3056, \quad 0.1645 \leq x_{12} \leq 0.2619$$
$$x_1 - x_7 + x_{12} = 0.1465$$

and, on the other hand,

$$\begin{aligned}
&|0.3089 - x_2| \leq x_1 - 0.1724 \text{ and } |0.2420 - x_2| \leq 0.26 - x_1\\
&|0.3556 - x_3| \leq x_1 - 0.1724 \text{ and } |0.3374 - x_3| \leq 0.26 - x_1\\
&|0.1632 - x_4| \leq x_1 - 0.1724 \text{ and } |0.1605 - x_4| \leq 0.26 - x_1\\
&|0.1763 - x_5| \leq 0.3056 - x_7 \text{ and } |0.2846 - x_5| \leq x_7 - 0.1752\\
&|0.3145 - x_6| \leq 0.3056 - x_7 \text{ and } |0.2286 - x_6| \leq x_7 - 0.1752\\
&|0.2036 - x_8| \leq 0.3056 - x_7 \text{ and } |0.3116 - x_8| \leq x_7 - 0.1752\\
&|0.1593 - x_9| \leq x_{12} - 0.1645 \text{ and } |0.1831 - x_9| \leq 0.2619 - x_{12}
\end{aligned}$$

$$|0.3461 - x_{10}| \leq x_{12} - 0.1645 \text{ and } |0.2568 - x_{10}| \leq 0.2619 - x_{12}$$
$$|0.3302 - x_{11}| \leq x_{12} - 0.1645 \text{ and } |0.2981 - x_{11}| \leq 0.2619 - x_{12}$$

For instance,

$$(0.1745, 0.3095, 0.3545, 0.1615, 0.2835, 0.2285, 0.1765, 0.3115,$$
$$0.1585, 0.3455, 0.3295, 0.1665)$$

satisfies these conditions and therefore it is a midpoint of $\underline{\mu}_{MT}$ and $\underline{\mu}_{EC}$ with respect to $D_{H,\infty}$. Solving the corresponding linear system one founds that this is the vector of frequencies of the bases A, C, G, and T at the three base sites of a codon in, for instance, a 4000-codon long DNA sequence containing 599 AAT codons, 99 ATA, 318 CCG, 67 CCT, 853 CTC, 418 GCC, 706 GGG, 294 GTG, 535 TAA, and 111 TCC. Such a DNA sequence could be understood as a middle way between the original DNA sequences as far as frequencies of bases in base sites of a codon concern.

Of course, any other combination of codons yielding this or another vector of frequencies in $\mathrm{mid}_{D_{H,\infty}}(\underline{\mu}_{MT}, \underline{\mu}_{EC})$ could also be understood so.

- Another simple computation gives that $D_{\infty,H}(\underline{\mu}_{MT}, \underline{\mu}_{EC}) = 0.4326$ and, using Corollary 13, that $\mathrm{mid}_{D_{\infty,H}}(\underline{\mu}_{MT}, \underline{\mu}_{EC})$ consists of those vectors $(x_1, \ldots, x_{12}) \in [0,1]^{12}$ such that, on the one hand,

$$0.1763 \leq x_5 \leq 0.2846, \qquad 0.2286 \leq x_6 \leq 0.3145$$
$$0.1752 \leq x_7 \leq 0.3056, \qquad 0.2036 \leq x_8 \leq 0.3116$$
$$x_5 - x_6 - x_7 + x_8 = -0.0239$$

and, on the other hand,

$$|0.1724 - x_1| + |0.3089 - x_2| + |0.3556 - x_3| + |0.1632 - x_4| \leq 0.2163$$
$$|0.26 - x_1| + |0.242 - x_2| + |0.3374 - x_3| + |0.1605 - x_4| \leq 0.2163$$
$$|0.1593 - x_9| + |0.3461 - x_{10}| + |0.3302 - x_{11}| + |0.1645 - x_{12}| \leq 0.2163$$
$$|0.1831 - x_9| + |0.2568 - x_{10}| + |0.2981 - x_{11}| + |0.2619 - x_{12}| \leq 0.2163$$

So, for instance,

$$(0.2106, 0.2816, 0.3467, 0.1611, 0.23415, 0.28645, 0.2255, 0.2539,$$
$$0.1656, 0.3205, 0.3176, 0.1963)$$

satisfies these conditions and therefore it is a midpoint of $\underline{\mu}_{MT}$ and $\underline{\mu}_{EC}$ with respect to $D_{\infty,H}$. □

Example 15. Figure 1 displays the alignments, defined by their position in the 3'-accept stem, of the first six bases of the tRNA-Met molecules of (a) ten eubacterias and (b) the mitochondria of ten single cells or fungi.

(a)	Mycoplasma Capric.	`GGCGGG`	(b)	Chlamydomon Reinh.	`AGACAC`
	Mycoplasma Gen.	`GGAUCU`		Penicillium Urtic.	`AGCGAA`
	Mycoplasma Mycoid.	`GGCGGG`		Pichia Canad.	`CGCACU`
	Mycoplasma Pneumo.	`GGCUGG`		Aspergillus Nidul.	`GCCAAA`
	Spiroplasma Melif.	`GGCGGG`		Saccharomyces Cer.	`GCUUGU`
	Staphylococ. Aure.	`GGCGGU`		Williopsis Mrakii	`GCUUAU`
	Helicobacter Pylo.	`GGAUUC`		Hansenula Wingei	`CGCACU`
	Bacillus Subtilis	`GGACCU`		Torulopsis Glab.	`ACUUGU`
	E. Coli	`GGCUAC`		Pichia Jad.	`GCUUGU`
	Haemophilus Influ.	`CGCGGG`		Trichophyton Rubr.	`GCCCGA`

Fig. 1. Two alignments of pieces of tRNA-Met.

The matrices of frequencies of the ribonucleotides A, C, G and U corresponding to these two alignments are, respectively,

$$\underline{\mu}_{Eub} = \begin{pmatrix} 0 & 0 & 0.3 & 0 & 0.1 & 0 \\ 0.1 & 0 & 0.7 & 0.1 & 0.2 & 0.2 \\ 0.9 & 1 & 0 & 0.5 & 0.6 & 0.5 \\ 0 & 0 & 0 & 0.4 & 0.1 & 0.3 \end{pmatrix} \qquad \underline{\mu}_{Mit} = \begin{pmatrix} 0.3 & 0 & 0.1 & 0.3 & 0.4 & 0.3 \\ 0.2 & 0.6 & 0.5 & 0.2 & 0.2 & 0.1 \\ 0.5 & 0.4 & 0 & 0.1 & 0.4 & 0 \\ 0 & 0 & 0.4 & 0.4 & 0 & 0.6 \end{pmatrix}$$

Concatenating the rows of these matrices we would obtain two fuzzy words of length 6 over the alphabet $\Sigma = \{A, C, G, U\}$, which we would still denote by $\underline{\mu}_{Eub}$ and $\underline{\mu}_{Mit}$; for the sake of clarity we shall use their matrix representation.

The global alignment of both sets of sequences, preserving the nucleotides' positions, has a matrix of frequencies that is exactly half the sum of this pair of matrices, and therefore it represents a midpoint of $\underline{\mu}_{Eub}$ and $\underline{\mu}_{Mit}$ with respect to all metrics considered here. Now the question arises: given an specific metric, are there subalignments of this global alignment whose matrix of frequencies is a midpoint of $\underline{\mu}_{Eub}$ and $\underline{\mu}_{Mit}$ with respect to this metric? Such a midpoint-subalignment could represent a consensus or an average subalignment of the global alignment.

Consider for instance the metric $D_{H,\infty}$. The subalignment given in Figure 2

Mycoplasma Capric.	`GGCGGG`
Mycoplasma Gen	`GGAUCU`
Mycoplasma Pneumo.	`GGCUGG`
Pichia Canad.	`CGCACU`
Staphylococ. Aure.	`GGCGGU`
E. Coli	`GGCUAC`
Penicillium Urtic.	`AGCGAA`
Saccharomyces Cer.	`GCUUGU`
Williopsis Mrakii	`GCUUAU`
Torulopsis Glab.	`ACUUGU`

Fig. 2. A subalignment of the join of the previous two alignments

has matrix of frequencies

$$\begin{pmatrix} 0.2 & 0 & 0.2 & 0.1 & 0.3 & 0.1 \\ 0.1 & 0.2 & 0.6 & 0 & 0.2 & 0.2 \\ 0.7 & 0.8 & 0 & 0.3 & 0.4 & 0.2 \\ 0 & 0 & 0.2 & 0.6 & 0.1 & 0.5 \end{pmatrix}$$

and it easy to check that the corresponding fuzzy word satisfies the equations that describe $\mathrm{mid}_{D_{H,\infty}}(\underline{\mu}_{Eub}, \underline{\mu}_{Mit})$ given in Corollary 11 applied to $\underline{\mu}_{Eub}$ and $\underline{\mu}_{Mit}$. Furthermore, it is easy to deduce from these equations that no subalignment with less than 10 sequences can have a matrix of frequencies that is a midpoint of $\underline{\mu}_{Eub}$ and $\underline{\mu}_{Mit}$ with respect to $D_{H,\infty}$. Therefore, and as far as this metric concerns, this subalignment is a minimal average of the pair of original alignments.

We have used a toy alignment, with a small and easy to use number of sequences of small length, just to simplify the presentation of this example, but the same problem can be attacked for alignments greater in size and in length. We are currently working on it. □

5 Conclusion

The concept of midpoint of two fuzzy subsets of a given finite set with respect to a metric, as a formalization of the middle ways between the situations described by the fuzzy subsets, was invented by Nieto and Torres in 2003, and several applications to the comparison of medical data were soon proposed [3,16]. The main goal of this paper is to introduce this concept in the field of the comparison of biological sequences.

An imprecisely known biopolymer can be described as a *fuzzy biopolymer*, a vector in a unit hypercube representing a fuzzy set that assigns to each position and each possible monomer (bases in nucleic acids, amino acids in proteins) the extent to which this monomer appears in this position. This kind of descriptions also include, for instance, profiles derived from multiple alignments. But, although in our main examples we have only considered fuzzy biopolymers that come from matrices of frequencies, and hence profiles, it should be clear that the concept of fuzzy biopolymer is more general. For instance, a fuzzy polymer could assign to each position and each monomer the *possibility* that this monomer appears in this position, deduced from a poorly designed sequencing experiment [2].

We have considered in this paper metrics on fuzzy biopolymers that are obtained aggregating simple metrics (euclidean, Hamming, maximum distances) defined on each piece of the fuzzy biopolymer that corresponds to a position. For these metrics we have described the sets of midpoints of two fuzzy biopolymers of the same length as sets of vectors in the unit hypercube of a suitable dimension defined by an explicit set of equations and inequations.

Several possible applications of midpoints of fuzzy biopolymers have been outlined in the introduction and we have given some examples at the end of the last section. We hope other applications will arise. This will probably involve the generalization of this work to other metrics used in profile alignment; we hope to report on them elsewhere. To widen the range of application of midpoints in computational biology as well as in other fields of research, it is also necessary to overcome a major drawback. The very definition of midpoint of two fuzzy subsets *of a given set* makes its generalization to fuzzy words to make sense only for fuzzy words *of the same length*. One possibility to overcome this drawback could be to develop a general theory of midpoints of fuzzy subsets of different sets. Another possibility is to use a metric to align the given fuzzy words, with suitable gap costs, and then to use the same metric to compute the midpoints. This increases the interest of the study of midpoints with respect of metrics used in profile alignment.

Acknowledgements. We acknowledge with thanks the comments and advices of the anonymous referees, J. Miró, and G. Valiente on the first version of this paper, which have led to a great improvement of it.

References

[1] I. Bloch, “On Fuzzy Spatial Distances.” In: *Advances in Imaging and Electronic Pysics* vol. 128, Elsevier (2003), 51–122.

[2] J. Casasnovas, J. Miró, F. Rosselló, “Distances between possibilistic descriptions of RNA structures.” In: *Artificial Intelligence Research and Development*,

Frontiers in Artificial Intelligence and Applications vol. 100, IOS Press (2003), 15–26.

[3] J. Casasnovas, F. Rosselló, "Midpoints as average representations of pairs of descriptions by means of fuzzy subsets." To appear in: *Proceedings of the IPMU04 Conference.*

[4] R. Durbin, S. Eddy, A. Krogh, G. Mitchinson, *Biological sequence analysis.* Cambridge Univ. Press (1998).

[5] E. Garner, P. Cannon, P. Romero, Z. Obradovic, A. Dunker, "Predicting disordered regions from amino acid sequence: common themes despite differing structural characterization." Genome Inf. 9 (1998), 201–214.

[6] O. Gotoh, "Optimal alignment between groups of sequences and its application to multiple sequence alignment." Comput. Appl. Biosc. 9 (1993), 361–370.

[7] M. Gribskov, R. Lüthy, D. Eisenberg, "Profile analysis." Methods in Enzym. 183 (1990), 146-159.

[8] M. Gribskov, A. D. McLachlan, D. Eisenberg, "Profile analysis: Detection of distantly related proteins." Proc. Nat. Acad. Sci. USA 84 (1987), 4355–4358.

[9] M. Gribskov, S. Veretnik, "Identification of sequence patterns with profile analysis." Methods in Enzym. 266 (1996), 343–367.

[10] J. D. Kececioglu, W. Zhang, "Aligning alignments." In: *Proc. 9th Annual Symposium Combinatorial Pattern Matching, CPM 98*, Lecture Notes in Computer Science vol. 1448 (Springer-Verlag, 1998), 189–208

[11] B. Kosko, *Neural networks and fuzzy systems.* Prentice-Hall (1992).

[12] G. N. Lance, W. T. Williams, "A general theory of classificatory sorting strategies I: Hierarchical systems." The Computer Journal 9 (1967), 373–380.

[13] S. Lang, *Introduction to linear algebra.* Undergraduate Texts in Mathematics, Springer-Verlag (2nd edition, 1986).

[14] R. B. Lyngso, C. N. Pedersen, H. de Nielsen, "Metrics and similarity measures for Hidden Markov Models." In: *Proceedings of ISMB 1999* (AAAI Press, 1999), 178–186.

[15] S. Montes, I. Couso, P. Gil, C. Bertolouzza, "Divergence measure between fuzzy sets." Int. J. Approx. Reason. 30 (2002), 91–105.

[16] J. Nieto, A. Torres, "Midpoints for fuzzy sets and their application in medicine." Artif. Intell. Med. 27 (2003), 81–101.

[17] J. Nieto, A. Torres, M, Vàzquez, "A metric space to study differences between polynucleotides." To appear in Applied Mathematical Letters.

[18] C. P. Pappis, N. I. Karacapilidis, "A comparative assessment of masures of similarity of fuzzy values." Fuzzy Sets and Systems 56 (1993), 171–174.

[19] S. Pietrokovski, “Searching databases of conserved sequence regions by aligning protein multiple-alignments.” Mucl. Acid Res. 24 (1996), 3836–3845.

[20] A. Pradera, E. Trillas, E. Castiñeira, On Distances Aggregation. *Proceedings of IPMU2000* (Madrid, July 3-7), 693–700

[21] K. Sadegh-Zadeh, “Fuzzy genomes.” Artif. Intell. Med. 18 (2000), 1–28.

[22] A. Torres, J. Nieto, “The fuzzy polynucleotide space: basic properties.” Bioinformatics 19 (2003), 587–592.

[23] R. J. Valenza, *Linear Algebra*. Undergraduate Texts in Mathematics, Springer-Verlag (1993).

[24] G. Yona, M. Levitt, “Within the twilight zone: A sensitive profile-profile comparison tool based on information theory.” J. Mol. Biol. 315 (2002), 1257–1275.

[25] M. Zaus, *Crisp and soft computing with hypercubical calculus*. Physica-Verlag (1999).