Distribution on Contingency of Alignment of Two Literal Sequences under Constrains

Running title: Distribution on alignments under constrains

Lorentz JÄNTSCHI [a,b,c,d], Sorana D. BOLBOACĂ [e,c,*]

[a] Technical University of Cluj-Napoca, Department of Physics and Chemistry, 103-105 Muncii Bvd., 400641 Cluj-Napoca, Cluj. Romania. E-mail: lorentz.jantschi@gmail.com

[b] Babeş-Bolyai University, Institute for Doctoral Studies, 1st Mihail Kogalniceanu Street, 400084 Cluj-Napoca, Romania.

[c] University of Agricultural Science and Veterinary Medicine Cluj-Napoca, 3-5 Calea Mănăştur, 400372 Cluj-Napoca, Romania.

[d] The University of Oradea, Department of Chemistry, 1st Universităţii Street, 410087 Oradea, Romania.

[e] Iuliu Haţieganu University of Medicine and Pharmacy, Department of Medical Informatics and Biostatistics, 6 Louis Pasteur, 400349 Cluj-Napoca, Romania. E-mail: sbolboaca@umfcluj.ro

* Corresponding author: Phone: +4-0750-774506; Fax: +4-0364-818418

Abstract

The case of ungapped alignment of two literal sequences under constrains is considered. The analysis lead to general formulas for probability mass function and cumulative distribution function for the general case of using an alphabet with a chosen number of letters (e.g. 4 for deoxyribonucleic acid sequences) in the expression of the literal sequences. Formulas for three statistics including mean, mode, and standard deviation were obtained. Distributions are depicted for three important particular cases: alignment on binary sequences, alignment of trinomial series (such as coming from generalized Kronecker delta), and alignment of genetic sequences (with four literals in the alphabet). A particular case when sequences contain each letter of the alphabet at least once in both sequences has also been analyzed and some statistics for this restricted case are given.

Keywords: alignment; contingency matrix; probability mass function (PMF); cumulative distribution function (CDF).

1 Introduction

Researches related to sequence alignments are frequently done due to the huge amount of already identified sequence of DNA (deoxyribonucleic acid), RNA (ribonucleic acid), or proteins (Pruitt et al. 2012). Sequence alignment is defined as a way of arrange DNA, RNA (Allali et al. 2012), or amino acid (Mongiovì and Sharan 2013) sequences to identify similar regions that could reflect functional, structural or evolutionary relationships between sequences (Mount 2004). Several algorithms were developed and implemented for global (Rahrig et al. 2013; Szalkowski and Anisimova 2013) or local alignments (Phuong et al. 2006; Tabei and Asai 2009; Frith et al. 2010), each algorithm with certain advantages and disadvantages. For example, the approach proposed by Szalkowski and Anisimova (2013) detect insertions and deletions of TR (tandem repeats) units not restricted to TR unit boundaries and proved more performing (~10%) compared to other aligners for cases with divergence high TR rates. Frith et al. (2010) assessed several combinations of score parameters for alignment (495) and found that high value of X-drop parameter are not always better, when tandem repeats are masked in a non-standard way the E-values accurately indicate the rate of spurious alignment, while highly reliable subsets of aligned bases could be obtained by γ-centroid alignment.

The state-of-the art in the pairwise alignments showed that statistical significance of the alignment is directly related to scoring scheme, sequence length and number of literals in the sequence (Mott 2005). Different approaches had been developed and implemented to estimate statistical significance of scores of alignment. BLAST2.0 (Altschul et al. 1997) implement a lookup method where K and λ parameters are pre-computed for different scoring schemes using average amino acid composition of both sequences. FASTA package (PRSS program) estimate the statistical significance of the shuffled (1,000 times) score distribution (Pearson 2000) while HMMER use maximum likelihood fitting in estimation of statistical significance (Mitrophanov and Borodovsky 2006).

Different scoring functions, such as probability consistency transformation – PCT (Do et al. 2005), Burrows-Wheeler Transform – BWT (Li and Durbin 2009), PSAR (Kim and Ma 2011), gPSP (Mokaddeml and Elloumi 2013), PSAR-Align (Kim and Ma 2014), etc., are used to characterize the alignment. BWT could be seen as a scoring function since the extension of BWT introduced by Mantaci et al. (2008) led to a general method for comparing sequences.

Distribution analysis found its usefulness in assessment of different natural (Bolboacă et al. 2011, Jäntschi et al. 2012a; 2011) or simulated phenomena (Jäntschi and Bolboacă 2011; Jäntschi et al. 2012b). Karlin and Altschul (1990) proved for ungapped alignment that the optimal local alignment scores used in the evaluation of sequence alignments follow an extreme-value distribution (characterized by characteristic value *K*, and scale *λ*). Computational experiments suggest that the optimal local alignment scores also apply to gapped local alignments (Smith et al. 1985; Altschul et al. 2001). The most important advantages of the island method (Olsen et al. 1999) over the direct method (Waterman 1994) in estimation of statistical parameters for gapped local sequence alignment is related with systematic errors that are easiest to be controlled (Altschul et al. 2001).

Our research started from the hypothesis that the distribution of ungapped alignments could provide useful information about the chance of their occurrences. A statistical approach based on distribution analysis able to identify the thresholds for rejecting an ungapped alignment by chance has been developed and is presented in this manuscript.

2 Methods

The global ungapped alignment of two sequences of equal length (equal number of literals, *n*) was investigated in this study. The general formulation of the problem investigated in this study along with an example of a particular case (total number of letters q = 4 – for DNA and RNA, A = adenine, C = cytosine, G = guanine, and U = uracil; respectively A, C, G, and T = thymine) is presented in Table 1.

Table 1: Alignment of two sequences of identical length: general case (left-hand) and particular case (right-hand, q=4)

j	1	2	3	4	…	n-4	n-3	n-2	n-1	n
Seq1	a_{11}	a_{12}	a_{13}	a_{14}	…	$a_{1(n-4)}$	$a_{1(n-3)}$	$a_{1(n-2)}$	$a_{1(n-1)}$	a_n
Seq2	a_{21}	a_{22}	a_{23}	a_{24}	…	$a_{2(n-4)}$	$a_{2(n-3)}$	$a_{2(n-2)}$	$a_{2(n-1)}$	a_n
Match	?	?	?	?	…	?	?	?	?	?

Seq1 = first sequence; Seq2 = second sequence;
a_{ij}: the first subscript number refers the number of sequence (1 or 2); the second subscript number refers the index of the literal in the sequence ($1 \leq j \leq n$);
n = the length of the sequence

j	1	2	3	4	…	n-4	n-3	n-2	n-1	n
Seq1	A	C	C	G	…	U	A	G	A	C
Seq2	C	A	C	U	…	A	A	G	C	A
Match	no	no	yes	no	…	no	yes	yes	no	no

A = adenine, C = cytosine, G = guanine, and U = uracil

A match is present when identical alphabet letters are present in both sequences at the same. The case presented in Table 1 could be transposed in a contingency of alignment as it is presented in Table 2. The total number of possible literals in the sequences (q) gives the size of alignment contingency (Table 2).

Table 2: Alignment contingency: general case (left-hand) and particular case (right-hand, q=4)

Seq1\Seq2	1	… …	q	Σ
1	$\mathbf{b_{11}}$	… …	b_{1q}	$b_{11}+…+b_{1q}$
…	…	… …	…	…
…	…	… …	…	…
q	b_{q1}	… …	$\mathbf{b_{qq}}$	$b_{11}+…+b_{1q}$
Σ	$b_{11}+…+b_{q1}$		$b_{1q}+…+b_{qq}$	n

Seq1 = first sequence; Seq2 = second sequence;
e.g. for the first literal in the alphabet equal A (1=A), b_{11}= no of cases when A is in the same position in both sequences

Seq1\Seq2	'A'	'C'	'G'	'U'	Σ
'A'	**Σ('A','A')**	Σ('A','C')	Σ('A','G')	Σ('A','U')	
'C'	Σ('C','A')	**Σ('C','C')**	Σ('C','G')	Σ('C','U')	
'G'	Σ('G','A')	Σ('G','C')	**Σ('G','G')**	Σ('G','U')	
'U'	Σ('U','A')	Σ('U','C')	Σ('U','G')	**Σ('U','U')**	
Σ					n

A=adenine, C=cytosine, G=guanine, and U=uracil
n = the length of the sequence from Table 1
ex. Seq1 = 'AGCUAA'; Seq2 = 'ACGUAC'
→ Σ('A','A') = 1 + 0 + 0 + 0 + 1 + 0 = 2

The number of aligned literals from two stings of n literals, *PSq(q)*, is therefore given by the main diagonal of the alignment contingency presented in Table 2 (Eq(1) for general case):

$$PSq(q) = \Sigma_{1\leq i\leq n} b_{11} + \Sigma_{1\leq i\leq n} b_{22} + \ldots + \Sigma_{1\leq i\leq n} b_{(q-1)(q-1)} + \Sigma_{1\leq i\leq n} b_{qq} \tag{1}$$

where *PSq(q)* ranges from *0* (no matches) to *n* (perfect alignment).

In the particular case (q=4), the Eq(1) became:

$$PSq(4) = \Sigma_{v\in\{A,C,G,U\}} \Sigma_{1\leq i\leq n} (Seq1_i = v, Seq2_i = v) \tag{1P}$$

Based on the perfect alignment (the same literal exists at the same position in both sequences), the alignment ratio (AR) could be obtained for the general case using the formulas presented in Eq(2):

$$AR = PSq(q)/n \tag{2}$$

and for the particular case (q=4) using the formula presented in Eq(2P):

$$AR = PSq(4)/n \qquad (2P)$$

Two cases of alignment of two sequences of equal length were investigated in this study:

- Unrestricted case: no restriction is imposed in regards of appearance of letters in each sequence.
- Restricted case: all letters of the alphabet (genetic sequence alignment: A, C, G, T or A, C, G, U) appear at least once in each of both sequences.

A full enumeration study was conducted on ungapped alignment of two sequences with identical length from 2 to 10, with 2 to 4 literals in the alphabet (0≤i≤10, where i = number of matches). The full enumeration study was conducted in accordance with conventional procedures to generate all numbers whose representation in base q (number of literals in the alphabet, 2≤n≤4) has n (the length of the sequence) digits; two by two such sequences were generated and then aligned. All frequencies of alignments in the generated sequences were counted for unrestricted case and for restricted case (where were counted only if accomplished the criterion of all letters appearance).

The results of the full enumeration analysis were used to identify (whenever possible) the general formulas for:

- Number of possibilities of arrangements $Sq(i;n,q)$;
- Number of perfect alignments $Sq(i=n;n,q)$ and number of no matches $Sq(i=0;n,q)$;
- Statistical parameters of alignment: mode, mean, and variance;
- Probability mass function (PMF) and cumulative distribution function (CDF);
- Thresholds for alignment by chance (q=4 and 4≤n≤40) at a significance level of 5% by Monte-Carlo experiment.

3 Results and Discussion

The results and associated discussion presented in this section refer to the unrestricted case as well as, in certain case, to the restricted case (the restriction referred to apparition at least once of any letters from alphabet in each sequence, equations referred with 'R' along the manuscript). The results obtained on full enumeration study are given in Tables 3-5.

Table 3: Total number of possible arrangements: 2≤*q*≤4 and *n*≤10

	Unrestricted case			**Restricted case**		
n	***q*=2**	***q*=3**	***q*=4**	***q*=2**	***q*=3**	***q*=4**
2	16			4		
3	64	729		36	36	
4	256	6561	65536	196	1296	576
5	1024	59049	1048576	900	22500	57600
6	4096	531441	16777216	3844	291600	2433600
7	16384	4782969	268435456	15876	3261636	70560000
8	65536	43046721	4294967296	64516	33593616	1666598976
9	262144	387420489	68719476736	260100	329422500	34774790400
10	1048576	3486784401	1099511627776	1044484	3133760400	669974990400

Table 4: Full enumeration results on unrestricted case: *q*=2 and *q*=4, *q*≤*n*≤10

					***q*=2 and 2≤*n*≤10**						
n	***i = 0***	***i = 1***	***i = 2***	***i = 3***	***i = 4***	***i = 5***	***i = 6***	***i = 7***	***i = 8***	***i = 9***	***i = 10***
2	4	8	4								
3	8	24	24	8							
4	16	64	96	64	16						
5	32	160	320	320	160	32					
6	64	384	960	1280	960	384	64				
7	128	896	2688	4480	4480	2688	896	128			
8	256	2048	7168	14336	17920	14336	7168	2048	256		
9	512	4608	18432	43008	64512	64512	43008	18432	4608	512	
10	1024	10240	46080	122880	215040	258048	215040	122880	46080	10240	1024
					***q*=4 and 4≤*n*≤10**						
4	20736	27648	13824	3072	256						
5	248832	414720	276480	92160	15360	1024					
6	2985984	5971968	4976640	2211840	552960	73728	4096				
7	35831808	83607552	83607552	46448640	15482880	3096576	344064	16384			
8	429981696	1146617856	1337720832	891813888	371589120	99090432	16515072	1572864	65536		
9	5159780352	15479341056	20639121408	16052649984	8026324992	2675441664	594542592	84934656	7077888	262144	
10	61917364224	206391214080	309586821120	275188285440	160526499840	64210599936	17836277760	3397386240	424673280	31457280	1048576

Table 5: Full enumeration results on restricted case: *q*=2 and *q*=3, *q*≤*n*≤10

						q = 2					
n	***i = 0***	***i = 1***	***i = 2***	***i = 3***	***i = 4***	***i = 5***	***i = 6***	***i = 7***	***i = 8***	***i = 9***	***i = 10***
2	2	0	2								
3	6	12	12	6							
4	14	48	72	48	14						
5	30	140	280	280	140	30					
6	62	360	900	1200	900	360	62				
7	126	868	2604	4340	4340	2604	868	126			
8	254	2016	7056	14112	17640	14112	7056	2016	254		
9	510	4572	18288	42672	64008	64008	42672	18288	4572	510	
10	1022	10200	45900	122400	214200	257040	214200	122400	45900	10200	1022
						q = 3					
3	12	18	0	6							
4	288	504	324	144	36						
5	3180	7410	7020	3660	1080	150					
6	26640	77220	94230	63000	24570	5400	540				
7	195132	671622	996030	828030	417690	128646	22680	1806			
8	1326528	5262768	9159192	9139536	5721660	2305296	586152	86688	5796		
9	8624460	38653578	77098392	89828424	67375476	33753132	11308248	2449656	312984	18150	
10	54532080	272115900	611443890	814654800	712684980	427812840	178525620	51181200	9664110	1089000	55980

The general formula for the total number of possibilities of arrangements between two sequences of identical length (*Sq(i;n,q)*) in unrestricted case was identified and is given by Eq(3):

$$\sum_{i=0}^{n} \mathrm{Sq}(i;n,q) = \left(q^{n}\right)^{2} \qquad (3)$$

where *i* = number of matches, *n* = length of sequence, *q* = number of letters in the alphabet.

The total number of possibilities of arrangements between two sequences of identical length (*n*) in restricted case is given by Eq(3R) along with formula for q=2:

$$\sum_{i=0}^{n} \mathrm{Sq}(i;n,q) = \left(\sum_{k=0}^{q-1}(-1)^k \cdot \binom{q}{k} \cdot (q-k)^n\right)^2 = \left(q! \cdot \left\{ {n \atop q} \right\}\right)^2 \qquad (3R)$$

where *i* = number of matches, *n* = length of sequences, *q* = number of letters in the alphabet, *k* = integer, $\binom{q}{k}$ = the number of *q*-combinations from a given set *k* of *n* (k!/((k-q)!·q!)), and $\left\{ {n \atop q} \right\}$ = the number of ways to partition a set of n objects into k non-empty subsets, or Stirling number of the second kind (Sharp 1968).

The formulas on three restricted particular cases of Eq(3R) for 2≤*q*≤4 are given bellow:

$$\sum_{i=0}^{n} \mathrm{Sq}(i;n,q) = \begin{cases} (q^n - q)^2 & \text{for} \quad q=2 \\ (q^n - q\cdot 2^n + q)^2 & \text{for} \quad q=3 \\ (q^n - q\cdot 3^n + (q+2)\cdot 2^n - q)^2 & \text{for} \quad q=4 \end{cases}$$

The formula is more complex for the restricted case compared to unrestricted case, as could be observed when Eq(3) is compared with Eq(3R). As expected, the total number of possible arrangements is smaller for restricted case compared to unrestricted case and the difference increased with *q* (total number of letters in the alphabet). Furthermore, the difference for the same *q* decreases with the increasing of sample size (*n*), while the decrease is faded for large *n*.

Formula presented in Eq(3R) could be checked for given *n* and *q* with the full enumeration results presented in Table 3.

General formula for the total number of *i* matches out of *n* for unrestricted case (Eq(4)) and some particular formulas for restricted case (2≤q≤4) (coming from q^{2n} = q^n·$((q-1)+1)^n$, Eq(4R)) were obtained as follows:

$$\mathrm{Sq}(i;n,q) = q^n \cdot (q-1)^{n-i} \cdot 1^i \cdot \binom{n}{i} = q^n \cdot \mathrm{NewtonBin}(i, q-1, 1, n)\text{, where}$$

$$\mathrm{NewtonBin}(i,x,y,n) = \binom{n}{i} \cdot x^{n-i} \cdot y^i \qquad (4)$$

and

$$\mathrm{Sq}(i;n,q=2)=\begin{cases} q^n-q, i=0 \\ (q^n-q^2)\cdot\begin{pmatrix} n \\ i \end{pmatrix}, 0<i<n \\ q^n-q, i=n \end{cases}$$

$$\mathrm{Sq}(i;n,q=3)=\begin{cases} (2\cdot q)^n-2\cdot q\cdot 4^n+2\cdot q\cdot 3^n+q\cdot 2^n-2\cdot q, i=0 \\ \text{to be determined}, 0<i<n \\ q^n-q\cdot 2^n+q, i=n \end{cases}$$

$$\mathrm{Sq}(i;n,q=4)=\begin{cases} \text{to be determined}, i=0 \\ \text{to be determined}, 0<i<n \\ q^n-q\cdot 3^n+(q+2)\cdot 2^n-q, i=n \end{cases} \qquad (4R)$$

where i = number of matches, q = number of letters in the alphabet, n = length of the sequences.

The number of perfect matches (i=n) in both unrestricted and restricted case is the square root of the total number of possibilities of arrangements as could be verified with the results presented in Table 4 and 5:

$$\mathrm{Sq}(i=n;n,q)=\sqrt{\sum_{i=0}^{n}\mathrm{Sq}(i;n,q)} \qquad (5\&5R)$$

It should be noted that (see 5&5R) the problem of counting the number of all full alignments (when i = n) is equivalent with the problem of counting the number of partitions of an n-set into k non-empty but distinguishable boxes (ordered non-empty subsets). The explicit formula for them can be obtained by applying the principle of inclusion-exclusion, when using the universal set consisting of all partitions of the n-set into k (possibly empty) distinguishable boxes, and the exclusion property that the partition has the associated box empty, the principle of inclusion-exclusion gives the answer for the related result (Brualdi 2010). Each arrangement (from the ones of which number is given by 5&5R) can be seen as an individual sequence obeying the imposed rule to contain all literals and here is a bijective function (if exists a sequence containing all letters and following the imposed rule then exists a perfect arrangement having the second sequence identical with the first one and vice versa). Therefore, the number of paired sequences is the square of this number and the proof for the (3R) formula is completed too.

One important property that can be observed by analyzing the results presented in Table 4 is that, for q=2 and 0≤i≤10, the distribution of total number of matches is symmetric and had one pick for even n and two equal values at the pick of the distribution for odd n. The distribution of total number of matches become asymmetrical and always have just one pick for q=4 and 0≤i≤10. The second property that could be observed by analyzing the results presented in Table 4 refers the total number of matches for i=n which verify the formula q^n for both q=2 and q=4. When two letters are in the alphabet (q=2), the symmetry of the distribution associated to the total number of matches observed for unrestricted case (Table 4) has also been observed for the restricted case (Table 5). Moreover, in the restricted case for q=2, one pick is observed for even n and two equal values at the pick of the distribution for odd n. Similar with the non-restricted case, the distribution of the total number of matches become asymmetrical for q=3 and with just one value to the pick of the distribution and with a tendency to symmetry for large sample sizes.

The probability mass function (PMF) associated with the frequency of apparition of matches using freely q literals (unrestricted case) is given by Eq(6) while the cumulative distribution function (CDF) is given by Eq(7).

$$\mathrm{PMF}_{\mathrm{Sq(i;n,q)}} = \frac{q^{n}\cdot(q-1)^{n-i}\cdot 1^{i}}{q^{2n}}\cdot\binom{n}{i} = \frac{(q-1)^{n-i}}{q^{n}}\cdot\binom{n}{i} \qquad (6)$$

$$\mathrm{CDF}_{\mathrm{Sq(i;n,q)}} = \sum_{k=0}^{i}\mathrm{PDF}_{\mathrm{Sq(i;n,q)}} = \sum_{k=0}^{i}\frac{(q-1)^{n-k}}{q^{n}}\cdot\binom{n}{k} = \frac{(q-1)^{n}}{q^{n}}\sum_{k=0}^{i}(q-1)^{-k}\cdot\binom{n}{k} \qquad (7)$$

Graphical representation of PMF for 2≤q≤4 is showed in Figure 1 (Eq(6)), while Figure 2 showed the CDF (Eq(7)).

[Place Figure 1 here]

Figure 1. Probability mass function for 2≤q≤4 (0.0-0.1-0.2-0.3-0.4-0.5 with red, green, blue, cyan and magenta) with PMF on vertical axis and number of aligned literals (i) on frontal axis (unrestricted case)

[Place Figure 2 here]

Figure 2. Cumulative distribution function for 2≤*q*≤4 (0.0-0.1-0.2-0.3-0.4-0.5 with red, green, blue, cyan and magenta) with CDF on vertical axis and number of aligned literals (*i*) on frontal axis (unrestricted case)

Figure 1 shows the symmetrical distribution for the unrestricted case as already observed and the remoteness from the symmetry with the increase of the number of letters in the alphabet. The PMF and CDF associated with the frequency of apparition of matches for restricted case is under investigation in our laboratory and the absence of the results is due to our available power of computation.

General formulas of three alignment statistical parameters (named mode, mean, and variance) associated to *Sq(i;n,q)* have been identified for the unrestricted case and are given by Eq(8) - Eq(10).

$$\hat{\mu} = \begin{cases} \{k, k+1\}, n = q \cdot k + q - 1 \\ k, n \neq q \cdot k + q - 1 \end{cases} \qquad (8)$$

$$\mu = \frac{n}{q} \qquad (9)$$

$$\sigma^2 = \frac{n \cdot (q-1)}{q^2} \qquad (10)$$

where $\hat{\mu}$ = mode, μ = arithmetic mean, q = number of letters in the alphabet, n = length of the sequences, k = integer, σ^2 = variance.

Formulas presented in Eq(8)-Eq(10) become the well known binomial formulas for q=1/p (p ∈ [0, 1], probability of success). Therefore, the unrestricted case could be seen as a binomial experiment.

The mode – Eq(8) – defines the alignment with highest probability to be observed by chance. The distribution of the mode in unrestricted case for 2≤*q*≤4 and *q*≤*n*≤10 is presented in Figure 3.

[Place Figure 3 here]

Figure 3. Mode in distribution of the alignment by chance vs. length of sequences: unrestricted case

Figure 3 showed different number of modes for different length of the sequence and $2 \le q \le 4$. The alignment by chance proved systematically bimodal to every odd n when q=2 ($n=q \cdot k+(q-1)$, k being any positive integer). For q=3, first bimodal alignment appeared when n=5 and occurred with a step equal with q ($n=q \cdot k+(q-1)$). For q=4, first bimodal alignment appeared when n=7 and occurred with a step equal with q ($n=q \cdot k+(q-1)$).

In the restricted case, the formula for the mean proved the same as for unrestricted while the mode in the distribution of alignment by chance is:

- q=2: bimodal distribution for odd n at $(n-1)/2$ and $(n+1)/2$ and unimodal for even n (excepting n=2) at $n/2$.
- q=3: unimodal distribution when n ranges from $q \cdot k$ to $q \cdot k+q-1$ at k.
- q=4 (and higher): the expression of mode for the distribution of alignment by chance has not yet been identified.

The threshold of the alignment by chance CDF$_{95}$ (k, CDF$_{\mathrm{Sq}(k;n,q)\ge 0.95}$)) for restricted case when $2 \le q \le 4$ had been estimated from a Monte-Carlo experiment and is approximated by Eq(11)-Eq(13).

$$q\text{=2: } \mathrm{k}_{\mathrm{CDF}\ge 95\%} = (0.84 \pm 0.05) \cdot (1+\mathrm{n})^{0.91\pm 0.02};\ k \to 0.56 \cdot n,\ \text{where } n \to \infty \qquad (11)$$

$$q\text{=3: } \mathrm{k}_{\mathrm{CDF}\ge 95\%} = (0.68 \pm 0.05) \cdot (1+\mathrm{n})^{0.89\pm 0.02};\ k \to 0.39 \cdot n,\ \text{where } n \to \infty \qquad (12)$$

$$q\text{=4: } \mathrm{k}_{\mathrm{CDF}\ge 95\%} = (0.66 \pm 0.07) \cdot (1+\mathrm{n})^{0.83\pm 0.03};\ k \to 0.30 \cdot n,\ \text{where } n \to \infty \qquad (13)$$

where q = number of letters in the alphabet, n = length of the sequences.

The formulas presented in Eq(11)-Eq(13) provide the threshold (CDF$_{95}$) at which, with a risk smaller than the significance level (in this case a significance level of 5% was used), the obtained matches did not appear by chance.

The distribution of CDF$_{95}$ for restricted case, q=4 and $q \le n \le 40$ obtained through simulations is presented in Figure 4.

[Place Figure 4 here]

Figure 4. Thresholds to reject matches by chance for two equally length sequences and q=4

The plot presented in Figure 4 showed that for example when 4≤n≤6 if there are observed more than three matches, with a 5% risk to be in error, these matches are not by chance. With two exceptions (when 10≤n≤11 and 30≤n≤33), the thresholds to reject matches by chance in ungapped alignment of two equally length sequences repeatedly are the same for three sample sizes (e.g. 4≤n≤6).

The genetic sequence responsible for alpha hemoglobin stabilizing protein (AHSP) on *Homo sapiens* (HS, chromosome 16) and *Mustela putorius furo* (MP) were used to exemplify the usefulness of CDF on restricted case. One hundred and eight pairs of strings of 8 bp were obtained, and 26 of them proved to belong to ungapped alignment restricted case (in both string all letters appears at least one time). The number of matches varied from 0 (15% [3.99; 34.47], where the lower and upper bound of 95% confidence interval calculated using an exact approach (Jäntschi and Bolboacă 2010) are provided in square brackets) to 5 (0.15; 26.78) with the highest frequency at two matches (38% [19.38; 57.54]). According to full enumeration results, for n=8 and q=4, the matches are not by chance if more than 3 are observed. More than three matches were observed in four out of twenty-six cases (15% [3.99; 34.47]) on AHSP experiment, leading to the conclusions that these ungapped alignments between HS and MP are not by chance.

Although there is much remains to be done, the work presented in this manuscript generates findings in the field of distribution analysis on equal length sequence alignment. The case with and without restriction were investigated for the number of letters in the alphabet that varied from 2 to 4. Although the present study provides full results for the unrestricted case of ungapped alignment and yielded some finding for the restricted case, its design is not without deficiencies. One of the main limitations of our study is relatively small size of the studies samples (n≤10) but this is linked with the applied method, full enumeration. Other main limitation is the lack of full characterization of the restricted case. Even if most formulas were identified and can be verified using the results obtained by full enumeration, due to the complexity of the calculations, we did not succeed yet to identify the general formulas for probability mass function and cumulative distribution function for restricted case, these two statistics being under investigation in our lab. Although the formulas for some alignment statistical parameters were identified (mean and mode) for both investigated cases (unrestricted and restricted case), the variance formula has been identified just for the unrestricted case. Investigation of other statistical parameters such as skewness and kurtosis could also bring valuable information regarding the distribution of alignments. Thus, these statistics could be used as approximation method (e.g. Fisher-Tippet (1928) with the same skewness and kurtosis) of the distribution functions in limit cases. The approximate formulas for the alignment by chance had been obtained for restricted case. Furthermore, it could also be interesting to extend the research regarding the alignment by chance for the particular case of amino-acids sequences. Despite its limitations, this study can be seen as the first step in assessment of the distribution analysis of real ungapped sequences alignments.

4 Conclusions

General formula for total number of possibilities of arrangements between two sequences of identical length $\Sigma_{0\leq i\leq n}Sq(i;n,q)$, total number of matches $Sq(i;n,q)$, and perfect alignment $Sq(i=n;n,q)$ had been identified for both unrestricted case and restricted case, for number of letters in the alphabet from 2 to 4. Formulas of mean and mode associated to $Sq(i;n,q)$ had been identified for both cases while the formula of variance had been identified just for the unrestricted case. Furthermore, the probability mass function and cumulative distribution function had also been identified for the unrestricted case, while approximate formulas for restricted alignment by chance are presented for $2\leq q\leq 4$ (where q = number of letters in the alphabet).

References

Allali J, Saule C, Chauve C et al. (2012) BRASERO: A resource for benchmarking RNA secondary structure comparison algorithms. Adv. Bioinformatics art. no. 893048.

Altschul SF, Madden TL, Schäffer AA et al. (1997) Gapped BLAST and PSI-BLAST: A New Generation of Protein Database Search Programs. Nucleic Acids Res. 25:3389–3402.

Karlin S, Altschul SF (1990) Methods for assessing the statistical significance of molecular sequence features by using general scoring schemas. Proc. Natl. Acad. Sci. USA 87:2264-2268.

Altschul SF, Bundschuh R, Olsen R, et al. (2001) The estimation of statistical parameters for local alignment score distribution. Nucleic Acids Res. 29:351–361.

Bolboacă SD, Jäntschi L, Sestraş RE (2011) Distribution fitting 12. Sampling distribution of compounds abundance from plant species measured by instrumentation. Application to plants metabolism classification. Bulletin UASVM Horticulture 68:54–61.

Brualdi RA (2010) Introductory Combinatorics (5th ed.), Prentice-Hall.

Do CB, Mahabhashyam MS, Brudno M et al. (2005) ProbCons: Probabilistic consistency-based multiple sequence alignment. Genome Res. 15:330–340.

Fisher RA, Tippett LHC. (1928) Limiting forms of the frequency distribution of the largest and smallest member of a sample. Proc. Camb. Phil. Soc. 24:180–190.

Frith MC, Hamada M, Horton P (2010) Parameters for accurate genome alignment. BMC Bioinformatics 11:80.

Jäntschi L, Bolboacă SD (2010) Exact Probabilities and Confidence Limits for Binomial Samples: Applied to the Difference between Two Proportions. TheScientificWorldJOURNAL 10:865–78.

Jäntschi L, Bolboacă SD (2011) Distributing Correlation Coefficients of Linear Structure-Activity/Property Model. Leonardo J. Sci. 19:27–48.

Jäntschi L, Bolboacă SD, Bălan M et al. (2011) Distribution fitting 13. Analysis of independent, multiplicative effect of factors. Application to the effect of essential oils extracts from plant species on bacterial species. Application to the factors of antibacterial activity of plant species. Bulletin UASVM Animal Science and Biotechnologies 68:323–331.

Jäntschi L, Sobolu RS, Bolboacă SD (2012a) An Analysis of the Distribution of Seed Size: A Case Study of the Gymnosperms. Not Bot Horti Agrobo 40:46–52.

Jäntschi L, Bolboacă SD, Sestraş RE (2012b). A simulation study for the distribution law of relative moments of evolution. Complexity 17:52–63.

Kim J, Ma J (2011) PSAR: measuring multiple sequence alignment reliability by probabilistic sampling. Nucleic Acids Res. 39:6359–6368.

Kim J, Ma J (2014) PSAR-Align: improving multiple sequence alignment using probabilistic sampling. Bioinformatics 30(7):1010-1012.

Li H, Durbin R (2009) Fast and accurate short read alignment with Burrows-Wheeler transform. Bioinformatics 25:1754–1760.

Mantaci S, Restivo A, Rosone G, Sciortino M (2008) A New Combinatorial Approach to Sequence Comparison. Theor. Comput. Syst. 42:411–429.

Mitrophanov AY, Borodovsky M (2006) Statistical Significance in Biological Sequence Analysis. Brief Bioinform 7:2–24.

Mokaddeml A, Elloumi M (2013) Motalign: A multiple sequence alignment algorithm based on a new distance and a new score function. Proceedings - International Workshop on Database and Expert Systems Applications, DEXA, Article number 6621350, pp. 81–84.

Mongiovì M, Sharan R (2013) Global alignment of protein-protein interaction networks. Methods Mol Biol 939:21–34.

Mott R (2005) Alignment: Statistical Significance. Encyclopedia of Life Sciences. Available at: http://mrw.interscience.wiley.com/emrw/9780470015902/els/article/a0005264/current/abstract. Accessed February 24, 2014.

Mount DM (2004) Bioinformatics: Sequence and Genome Analysis. (2nd Ed.). Cold Spring Harbor Laboratory Press, Cold Spring Harbor, NY.

Olsen R, Bundschuh R, Hwa T (1999) Rapid assessment of extremal statistics for gapped local alignment. In: T. Lengauer, R. Schneider, P. Bork, D. Brutlag, J. Glasgow, H.-W. Mewes, R. Zimmer (Eds). Proceedings of the Seventh International Conference on Intelligent Systems for Molecular Biology, AAAI Press, Menlo Park, C.A., pp. 211–222.

Pearson WR (2000) Flexible Sequence Similarity Searching with the FASTA3 Program Package. Methods Mol Biol 132:185–219.

Phuong TM, Do CB, Edgar RC et al. (2006) Multiple alignment of protein sequences with repeats and rearrangements. Nucleic Acids Res 34:5932–5942.

Pruitt KD, Tatusova T, Brown GR et al. (2012). NCBI Reference Sequences (RefSeq): current status, new features and genome annotation policy. Nucleic Acids Res 40:D130–D135.

Rahrig RR, Petrov AI, Leontis NB et al. (2013) R3D Align web server for global nucleotide to nucleotide alignments of RNA 3D structures. Nucleic Acids Res 41:W15–W21.

Sharp H (1968) Cardinality of finite topologies. J Combin Theory 5:82–86.

Smith TF, Waterman MS, Burks C (1985) The statistical distribution of nucleic acid similarities. Nucleic Acids Res 13:645–656.

Szalkowski AM, Anisimova M (2013) Graph-based modeling of tandem repeats improves global multiple sequence alignment. Nucleic Acids Res 41:e162.

Tabei Y, Asai K (2009). A local multiple alignment method for detection of non-coding RNA sequences. Bioinformatics 25:1498–1505.

Waterman M (1994) Estimating statistical significance of sequence alignments. Phil Trans R Soc Lond B 344:383–390.