Detection of New Motifs Properties in Biodata

Nooruldeen N. Qader

University of Human Development, University of Sulaimani, Sulaimanyah, Iraq, drnnq@uhd.edu.iq, nuraddin.qadir@univsul.edu.iq

Hussein K. Al-Khafaji

Alrafidain University College, Baghdad, Iraq, dr.hkm1811@yahoo.com

Abstract— Biodata are rich of information. Knowing the properties of biological sequence can be valuable in analyzing data and making appropriate conclusions. This research applied naturalistic methodology to investigate the structural properties of biological sequences (i.e., DNA). The research implemented in the field of motif finding. Two new motifs properties were discovered named identical neighbors and adjacent neighbors. The analysis is done in different situations of background frequency and motif model, using distinctive real data set of varied data size. The analysis demonstrated the strong existence of the properties. Exploiting of these properties considers significant steps towards developing powerful algorithms in molecular biology.

Index Terms—Motif model, mining, DNA, biodata, sequence, genome, k-mers, structure, Bioinformatics, monad, composite, Background Frequency

I. Introduction

Explosion and growth of biological data in exponential rate resulted in urgent collaborative work to enable understanding and analyzing such data to be utilized in a better form in daily life, although massive efforts have been done, Bioinformatics are still infancy. There are a lot of factors that make the challenges harder; including huge information carried by a genome, lack of techniques to reveal benefit knowledge from, and difficulty of the biology laboratory test to validate correct information [1]. Data mining comes as the first techniques to design new methods and algorithms for knowledge extraction by finding patterns, classification, clustering, etc [2], [3].

The objectives are finding characteristics and properties of biosequences that make genome [4]; therefore numerous data structure and mapping have been used. Recent research motivates investigating the structural properties of biological sequences to enhance algorithms in molecular biology [1], [5], [6]. Therefore, this paper focuses on the nature of biological data to make the design more efficient following the new trends in Bioinformatics. The field of search and extraction of biological patterns is considered an active and promising field of Bioinformatics. Therefore, it has been selected for this study.

The rest of this paper is organized as follows: Biological motif model in section 2. Characteristics of biological motifs and sequences are exhibited in section 3. Section 4 & 5 concentrates on the discovery of two important properties in biodata: adjacent neighbours and identical neighbours. Finally, we ended with the research conclusions in section 6.

II. Biological Motifs Model

The motif is an abstract model for a set of sites positions with similar patterns. Motifs have multi forms of representation; this study employs string and PWM model. Motifs mining algorithms search for exact or approximate motifs utilizing motifs template, and biological motifs are generally sequence of symbols [7]. Motifs classified into monad and composite type. And composite motif has two types simple and structure motifs, simple motifs allow fixed gaps between symbols, structure motifs allows variable gaps between symbols or component [8]. Composite motifs template reveals:

- Monad motifs alphabet either in DNA alphabet or in IUPAC format,
- Motifs length, type and number of symbols in each monad motif,
- Motifs components, number of monad motifs,
- Gaps length, minimum and maximum gaps

A composite motif represents formally *M1 [l1, u1] M2 [l2, u2] M3…Mn [ln, un],* for example, motifs template shown in figure 2 consists of two monad motifs in DNA alphabets, M1 [5,17] M2 such that M1 is GGGTGGGAAGGTCGT with length of 15 base and M2 is TTAGCGGGTAT with length of 11 base and variable gap between these two monad motifs as minimum gap of 5 base and maximum gap of 17 accordingly after 5 of any base or don't care base of M1 till 17 searcher has to look for M2, where found; it is considered an event and occurrence of the pattern, and when the number of occurrences is equal or greater than minimum threshold it is regarded as frequent pattern [9].

GGGTGGGAAGGTCGT [5, 17] TTAGCGGGTAT

Fig. 1. Composite motif

III. Characteristics of Biological Sequences

Knowing the properties of biological sequence can be very valuable in analyzing data and making

appropriate conclusions. In this context, appropriate characterization of the biological sequence structures and the exploitation properties of sequences are very important steps towards the development of powerful algorithms. Biological data, or more specifically molecular biological data DNA, RNA and protein, create organism body [10]. Biodata have many properties, they are rich of information. While a detailed discussion of biological properties is beyond the scope of this research, exploring properties that affect motif search algorithms are necessary. Some of these properties have been examined previously. Some of the related properties are briefly presented as follows:

1. Small alphabet, biological sequence alphabet (DNA, RNA and protein) are generally regarded small when compared with transaction sequences (e.g. market-basket analysis). Biological sequence mining typically requires an alphabet of size less than 21; DNA and RNA consist of four alphabets and protein consists of 20 alphabets; effects of alphabet size on sequential pattern data mining have been examined in [11]–[13].
2. Long sequences, biological sequences carry full details information about organism species in genes. Biosequences are long, for example chromosome 1 of the human sized 243 megabytes and human genome sized more than 3 gigabytes. Therefore, long sequences are considered an important property of biological sequence data set, this property and its impact on sequential pattern data mining are examined in [14]–[16].
3. Mutation, it is the most outstanding property that distinguishes between biological sequences and transactional sequences. Occurrences of patterns are not always identical; some copies may be approximated. The instances of the pattern usually differ from the model in a few positions. The biological sequence pattern usually allows nontrivial numbers of insertions, deletions, and mutations. Mutation represents a real challenge of sequential pattern data mining; this issue has been referred to in [7], [14], [17], [18].
4. Adjacent neighbours (AdjN): as shown in the previous chapter, one of the vivid fields of data mining in Bioinformatics is sequential pattern mining; these patterns consist of a number of characters or symbols that determine the pattern length. Pattern length may extend to several hundred. Monad motifs are made of consequent symbols; thus they are completely constructed from AdjN. And composite motifs consist of number of monad motifs where gaps or distances are permitted between them. Therefore, AdjN represent major parts of composite motifs.
5. Identical neighbours (IdN): this research explores and exploits IdN's property as a new property; it will be explained later in details in section 3.3. IdN occurs frequently in biological sequences and patterns, using this property in designing of algorithms in molecular biology will be considered as a promising issue for better performance.

In this research we followed Naturalistic methodology [1], which enable us to discover two new properties: IdN and AdjN; they are new properties discovered in this research. Therefore, the following sections 4 & 5 are dedicated to explore and examine those two properties in more details.

IV. Adjacent Neighbours Property in Motifs Model

Motifs are monad or composite, therefore exploring property of AdjN or contiguous bases is explained according to the type of motifs template as in the following:

A. AdjN in Monad Motifs

Two nucleotide TA called AdjN if TA denotes a pair (or subsequence) of nucleotides, where A appears immediately after T. Monad motifs templates consist of numerated contiguous characters, monad motifs described by the number and types of characters involved, for example AACTG is a monad motif lengthen five characters (i.e., k=5); also called 5-mer. Positional join algorithms based on the join between the positions of continuous neighbours, therefore the number of AdjN in each monad motif equals to the number of characters in a motif minus one, in other words, number of AdjN in monad motifs is:

#AdjN= (k-mers) – 1 (1)

Figure 2 displays AdjN in monad motif AACTG, AdjN are four and length of pattern is five, i.e., 5-mer. Number of AdjN in monad motifs is always less than the length of the monad motifs by one digit because each symbol takes the role of the head once and next takes role as tail except first and last symbols (i.e., A and G). They participate only one time because positional join starts from right most (i.e., G takes only role of tail) and ends at left most (i.e., A takes only role of head).

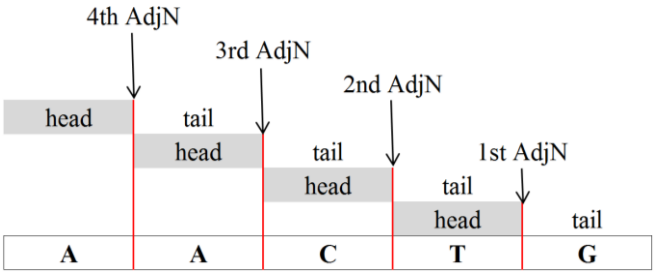

Fig. 2. AdjN in monad motif

B. *AdjN in Composite Motifs*

A composite motif can be regarded as an ordered collection of monad motifs with gap constraints between each pair of adjacent monad motifs. For example, GCAT [0, 5] AACTG is a composite motif that consists of two monad motifs and separated distance ranges from 0 to 5. An AdjN is in pairs that do not contain gaps. One may call the pairs that include gaps a gap neighbours. Figure 3 demonstrates AdjN in composite motif, most neighbours of composite motifs are adjacent (i.e., 7 neighbours are adjacent out of 8) and gap neighbours (GN) are rare (i.e., numbers of gap neighbours are the same as number of gaps), so:
The number of composite motifs components=n,

$$\text{The number of gaps} = \text{number of GN} = n - 1 \qquad (2)$$

The number of AdjN=∑number of AdjN in monad motifs= (k1-1) +(k2-1) +…+ (km-1) (3)

Where m symbolizes the total number of monad motifs in the composite motif, and k symbolizes k-mers; accordingly k1 represents the length of first monad motif and k2 denotes length of second monad motif, and so on.

To improve the likelihood of 'AdjN is greater than GN' one can imagine the lowest number of AdjN comparing with a number of the GN with small k-mers and high number of composite motif components (monad motifs). By using "Eq.2 and Eq.3" to calculate types of neighbours in the worst case where lengths of all monad motifs are equal and in lowest length (k1=k2=…km=2); let the monad motifs are composite of only two symbols and the number of composite motif components is high, i.e. n=9:

GN=9-1=8

AdjN= (2-1) + (2-1) + (2-1) + (2-1) + (2-1) + (2-1) + (2-1) + (2-1) + (2-1) =9

The presented example is impractical but it shows 'AdjN is more than GN' even in the worst case. It is known that the real monad motif length is mostly more than two; a popular example of monad motif is TFBS that has a length of 5-30 bp.

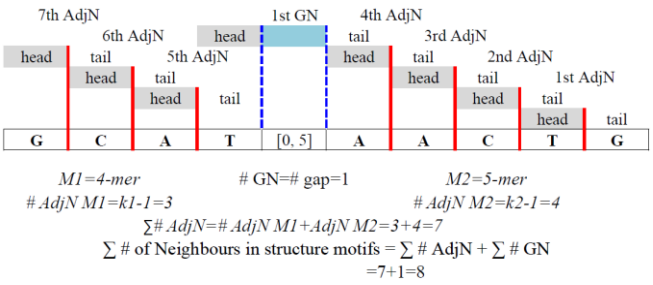

Fig. 3. AdjN in composite motif

To know the amount of AdjN in real situation somehow, let recalculate previous example and minimize each monad motifs length to minimum length of TFBS i.e. k-mers=5, and the number of composite motif components is same. n=9

GN=9-1=8

AdjN= (5-1) + (5-1) + (5-1) + (5-1) + (5-1) + (5-1) + (5-1) + (5-1) + (5-1) =36

This example shows superiority of AdjN (36) over GN (8).

Table 1 displays the number of gaps and AdjN in some example of real composite motifs from species Saccharomyces cerevisiae yeast and Arabidopsis thaliana c, the last two columns show that number of AdjN are greater than the number of GN. Even the neighbour type of pairs in composite motif is generally adjacent.

Therefore, AdjN property in composite motifs is regarded a significant property and it is better to take it into consideration in designing algorithms of molecular biology.

V. IDN Property in Biodata

Trends of this research examine and analyze biological sequence properties in order to exploit them for designing motif discovery algorithms. By collecting data of neighbours in a different situation, it is possible to obtain an empirical estimate of neighbour's type rates. It cannot be known in advance the types of neighbours in the motif model, or even knowing exactly what the probability of neighbours' types is. However, it can estimate the proportion of neighbour's types from data. Certainly, it cannot predict exactly what this will be — it will vary from one motif model to another. However, collecting data from different kinds of real motif model will give real notion of the distribution of neighbour's types across the population of motif models, which in turn will provide information about the likely cost of finding motifs. To be able to trace and measure the existence and overall concept of IdN property; it is better to determine that in *background frequency, motif model, effect of identical neighbour's k-mers, and influence of data size*. Therefore, the property of IdN is explored in thees situations as in the following subsections:

Table 1 Number of GNs and AdjNs in a set of real composite motifs

Species	Structure motif	Number of components	Length of simple motif						# AdjN	# GN
			1	2	3	4	5	6		
Saccharomyces cerevisiae yeast	UASH and URS1H Binding Sites	NNDTBNGDWGDNNDH [5, 179] WBRGCSGCYVW								
		2	15	13					26	1
Arabidopsis thaliana	Copia Motif	TNGA [12,14] TWNYTNNA [19,21] TNTMYRT [4,6] WNCCNNNNRG [72,95] TGNNA [100,125] TNTANRTNRAYGA								
		6	4	8	7	10	5	13	41	5
	M1	HNGTNYDNHDNBTNNDNA [0,3] YNHTNYRHGGNBTNAR [0,2] ARDBNBH								
		3	18	16	7				38	2
	M2	TNVRNKAYNKNVVNDV [9,11] HNRR [6,8] YDNNVNNV [9,13] HB [4,5] TNNNNRBNYDBDNNRR								
		5	16	4	8	2	16		41	4
	M3	DNNNNDRYW [2,5] DS [6,7] HMM [1,2] TNDB								
		4	9	2	3	4			14	3
	M4	DBNNNND [48,102] KRRYMYNNNMRNHYNDVNYAYVH [7,10] VNNNYNNND [34,63] WD [2,8] KNNH [3,5] VNDDRNNNNNNHVNNNNNNHHH								
		6	7	23	9	2	4	23	62	5

A. *IdN Phenomenon in Background Frequency*

Background frequency is the number of occurrences of a biological character (i.e., ACGT) in the raw biological sequences [21], [22]. IdN means same character appears immediately again without any gap or any other character (i.e., in DNA alphabet AA, CC, GG or TT). Two nucleotides TT are called IdN if TT denotes a pair (or subsequence) of nucleotides, where T appears immediately after T. Generally XjYj+1 are IdN in sequence S if X=Y, X at position j and Y at position j+1. Biological sequences length may be exceed millions even billions of characters and their alphabet are small. Investigating those two primary characteristics, small alphabet and very long sequences, intuitively reveal another important characteristic that repeats the same character several times, furthermore the same character is repeated consequently (per-letter background frequencies). For example, if a segment of DNA sequence with 10000 bases long examined and if each element has the same probability, then this probability is p=1/N, where p- is the probability of each nucleotide and N- is DNA alphabet (i.e., 4). In general the elements are not likely equal; they have different probabilities pi but suppose that the previous sequence is nucleotide sequences (ACGT) and suppose each base has the same probability; thus probabilities of symbol frequency are 10000/4=2500. In other words, each symbol may occur 2500 times in just 10000 positions that increase probabilities of IdN.

Figure 4 shows a real sample of DNA sequence; starting part of chromosome number 1 of Saccharomyces cerevisiae yeast; the sample contains 800 bases. Special format (i.e., CCTT) used to indicate IdN. Although the sequence is a tiny sequence but it reflects the concentration of similar successive characters and high score of IdN (i.e., IdN in the sample is 354) that make to see it as a fact that requires no proof [3].

$$\text{Ratio of IdN bases to overall bases} = \#\text{IdN}/\Sigma \text{ Neighbours} = 354/800 = 0.4425 \quad (4)$$

The proportional relation between identical neighbour's bases and overall bases in the sample by using "Eq.4" is found 0.4425; almost half of neighbours are identical!

Characters frequencies have been used in molecular algorithms, are referred to as Position Weight Matrix (PWM) or profile and Position Frequency Matrix (PFM) [23]. However, these representations fail to capture nucleotide interdependence and it was discovered by many researchers that the nucleotides of the DNA binding site cannot be treated independently.

Basic component in molecular biology nucleotides and amino acids are arranged in sequences of DNA, RNA and proteins. These components construct structures of all organisms and perform related biological functions [10]. Therefore, the interdependencies, correlations and interrelated

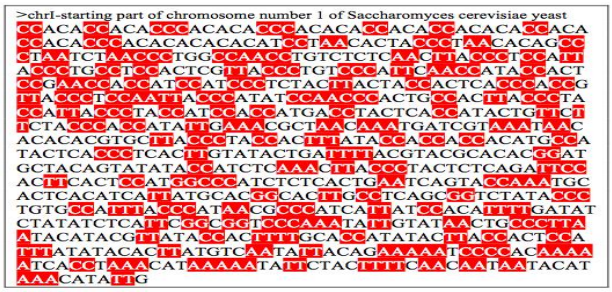

Fig. 4 Random sample of DNA sequence

relations between the basic components are expected, and even certain [24], [25]. Other studies and research tried to find the interdependencies, and model them. But efforts based on assumptions and failed to capture the reality of the biosequences data. Thus the naturalistic methodology is the attempt to correct the direction.

The string and the matrix representations share an important common weakness. They assume the occurrence of each nucleotide at a particular position of a binding site that is independent of the occurrence of nucleotides at other positions. This assumption does not represent the true picture as discussed in [24]. While PWM and PFM deal with single base frequency but here we revisit symbols frequency to consider frequency of IdN or the dependent occurrences of neighboring bases. Such direction promises to find new ways of data mining in Bioinformatics and understand genome coding better. Motif discovery is an example of role of IdN in Bioinformatics, and determine the role in genome coding is left as a suggestion for future work (State of art in genome en/coding based on single base A, C, G or T. What will be the effect of enlarging basis i.e., TT, TTT, TTTT?).

Biological motif is a piece of biological sequence. Anyway due to the functional associations of biological motifs, they are embedded in background sequences. Therefore, a collection of k-mers is likely to be a true motif. This understanding motivated and forms the direction of this research.

B. *IdN Score in Motif Model*

Scine the motif model is a sample of biological sequence; therefore all biological sequence properties must be reflected in motif model. Thus motif model includes IdN, similar adjacent bases or repeating same characters consequently. In order to find the rate of IdN in biological sequential patterns, set of real motifs have been investigated, for example, transcription factors, URS1H and UASH that are known to cooperatively regulate 10 genes of S. cerevisiae yeast. These 10 genes are listed in SCPD as the promoter DB of Saccharomyces cerevisiae yeast [24], as shown in Table 2. The first gene of them named ZIP1 which regulated by the TFBS shown in Figure 5.

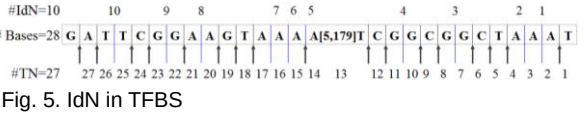

Fig. 5. IdN in TFBS

TFBS shown in Figure 5 is a composite motif of gene ZIP1 in S. cerevisiae yeast, the figure displayed included IdN; it is an example of composite motif model, it consists of two monad motifs and distance allowed between them ranged from minimum bases of 5 to maximum bases of 179. Composite motif GATTCGGAAGTAAAA [5, 179] GATTCGGAAGTAAAA is a combination of 28 characters.

Positional join based algorithms join contiguous bases [9], [26]. Contiguous bases either are identical or not. IdN denotes to contiguous bases that are identical (i.e., AA, CC, GG or TT) and total neighbours (TN) denote to all contiguous bases regardless of identicalness of them (i.e., TT or TC). The composite motif compounded of 10 IdN out of 27 TN.

The following example shows the analysis of two sets of real composite motifs goals to evaluate the weight and attendances of IdN. The actual number of IdN in each motif is calculated to determine the average of IdN of the set, and then estimating the rate of IdN in each set:

1- The first set consists of 10 composite motifs that have been shown in Table 2 and Table 3, the motif model consists of 28 nucleotides in two components of monad motifs. The number of TN in Table 3 is 27.

TABLE 2 UASH AND URS1H BINDING SITES

Genes	UASH		URS1H		Distance
	Site	Pos	Site	Pos	
ZIP1	GATTCGGAAGTAAAA	-42	==TCGGCGGCTAAAT	-22	20
MEI4	TCTTTCGGAGTCATA	-121	==TGGGCGGCTAAAT	-98	23
DMC1	TTGTGTGGAGAGATA	-175	AAATAGCCGCCCA==	-143	32
SPO13	TAATTAGGAGTATAT	-119	AAATAGCCGCCGA==	-100	19
MER1	GGTTTTGTAGTTCTA	-152	TTTTAGCCGCCGA==	-115	37
SPO16	CATTGTGATGTATTT	-201	==TGGGCGGCTAAAA	-90	111
REC104	CAATTTGGAGTAGGC	-182	==TTGGCGGCTATTT	-93	89
RED1	ATTTCTGGAGATATC	-355	==TCAGCGGCTAAAT	-167	188
REC114	GATTTTGTAGGAATA	-288	==TGGGCGGCTAACT	-94	194
MEK1	TCATTTGTAGTTTAT	-233	==ATGGCGGCTAAAT	-150	83
Motif	**NNDTBNGDWGDNNDH**		**==WBRGCSGCYVW==**		**[5, 179]**

TABLE 3 AVIDN IN BINDING SITES OF 10 GENES

Gene	Simple Motif1	Simple Motif2	# of IdN
ZIP1	GATTCGGAAGTAAAA	TCGGCGGCTAAAT	10
MEI4	TCTTTCGGAGTCATA	TGGGCGGCTAAAT	8
DMC1	TTGTGTGGAGAGATA	AAATAGCCGCCCA	7
SPO13	TAATTAGGAGTATAT	AAATAGCCGCCCA	8
MER1	GGTTTTGTAGTTCTA	TTTTAGCCGCCGA	10
SPO16	CATTGTGATGTATTT	TGGGCGGCTAAAA	9
REC104	CAATTTGGAGTAGGC	TTGGCGGCTATTT	10
RED1	ATTTCTGGAGATATC	TCAGCGGCTAAAT	6
REC114	GATTTTGTAGGAATA	TGGGCGGCTAACT	8
MEK1	TCATTTGTAGTTTAT	ATGGCGGCTAAAT	8

2- Second set is found by searching first chromosome of A. thaliana for composite motif reported in [9], [27] as motif1, that is, HNGTNYDNHDNBTNNDNA [0, 3] YNHTNYRHGGNBTNAR [0, 2] ARDBNBH that results 8 occurrences as shown in Figure 3, each one consists of 41 nucleotides in 3 components of monad motifs.

The mean is the sum of all values in a set, divided by the number of values. To find Estimate Identical Neighbours (EIdN), follow the following steps:

1. Compute the average of actual IdN (AvIdN), i.e., AvIdN in the first data set is calculated by searching 10 composite motifs, those composite motifs known as binding sites regulated by transcription factors UASH and URSH1 in S.

$$AvIdN = \frac{\sum_1^n IdN}{n} \quad (5)$$

Where n is the number of composite motifs under test. For example, AvIdN of the first data sets that is shown in Table 3 is calculated by using "Eq.5":

$$AvIdN = (10 + 8 + 7 + 8 + 10 + 9 + 10 + 6 + 8 + 8)/10 = 8.4$$

Table 4 shows AvIdN of the second data set that calculated by using "Eq.5":

$$AvIdN = (9 + 7 + 7 + 11 + 10 + 10 + 7 + 11)/8 = 9$$

2. Total neighbours of the composite motif are calculated by using "Eq.6":

$$TN = \text{Number of composite motif bases} - 1 \quad (6)$$

3. Using the following equation to find estimated percentage of IdN:

$$\text{Estimated of Identical Neighbours} = EIdN = AvIdN/TN \quad (7)$$

cerevisiae yeast genes, in this computation IdN between monad motifs had been excluded, because

they are separated by variable gap and they were not contiguous.

Simple Motif1	Simple Motif2	Simple Motif3	# of IdNs
CTGTTTGCAGATTAAAGA	CTATCTGTGGTCTGAA	AAGCTTC	9
AAGTGTTGAGACTACGCA	CTATCTATGGTCTTAA	AATCTCC	7
AAGTGTTGAGACTACGCA	CTATCTATGGTCTTAA	AATCTCC	7
ACGTTTGCCGATTGAAGA	CAATATATGGGCTTAA	AAGCACA	11
AGGTCTACAAGCTTAAGA	CTTTATATGGTCTAAG	AAGCTCC	10
CTGTTTGTCGTTTGCATA	CTTTATATGGTCTAAG	AAGCTTC	10
ATGTGTGTCTTCTCCACA	CCTTGTATGGTCTGAA	AGTCTCC	7
CTGTTTGTCGGTTGAAGA	CCTTATATGGACTCAA	AAGCCTC	11

Where AvIdN is the average of actual identical neighbours and TN is total neighbours in the composite motifs.

For example; EIdN of the first data sets that is shown in Table 3 is calculated by using "Eq.7" as follows:

AvIdN=8.4

TN =28-1=27

EIdN=8.4/27≈0.32

And EIdN of the second data set that shown in Table 4 is calculated by using "Eq.7" as follows:

AvIdN=9

TN =41-1=40

EIdN=9/40≈0.23

Table 4 Motif1 occurrences in first chromosome of A. thaliana

Composite motifs analyzed in Table 3 and Table 4 proved high rate of IdN to total neighbours; they are 32% and 23%; these rates reflect the richness of the property under investigation, while these rates is not constant for all motifs and may be increased or decreased, however it demonstrated new and promised property of motif that could be utilized in algorithms for molecular biology. The number of IdN varies from motifs model to another. Distribution of bases in real motifs models and background sequences is not uniform; nevertheless evidence provided by analyzed data sets of real motifs and sequences were presented abundantly of this property.

C. Frequency and k-mers of IdN [18]

The sample distribution of IdN is approximately normal due to the smallness alphabet and non-uniform distribution of biological bases in real biological sequences. Table 5 shows lengths or k-mers of IdN in upstream of S. cerevisiae yeast, for example for 5-mer, there are more than 30,000 of TTTTT and AAAAA. Consider the high frequency of long patterns of IdN making a heuristic approach to estimate the IdN slightly raises and overestimates, the table shows occurrences of a long pattern of IdN reaching up to 30 in relatively small data set under test, viz. 2.8 megabytes. For example, where pattern length is 30, IdN of nucleotide T occur 30 times and for nucleotide A occurs 13 times.

Moreover, the data in Table 6 and Table 7 demonstrated that the frequency of C and G is less than the frequency of A and T in upstream sequences of S. cerevisiae yeast. The C in a CG pair is often modified by a process known as methylation (where the C is replaced by methyl-C, which tends to mutate to T) [28]. As a result, CG pairs occur infrequently; therefore it is referred in literature CG Island [29]–[34]. And k-mers of nucleotide T is greater than others, e.g., in the sample data upstream of S. cerevisiae yeast it exceeds 40-mer.

Generally, in biological sequence data sequential pattern usually allows a nontrivial number of insertions, deletions, and mutations. Thus some bases (i.e., A and T) occur more frequently than others. However, experimental results data in the table indicate IdN of all nucleotides from 2-mer to 30-mer.

A shorter minimum motif length may yield many motifs, some of which might be sub-sequences of other longer motifs. Data from Table 5 also demonstrate that when k-mers increased the number of occurrences reduced; however, there exists identical neighbour's pattern of lengths more than 30-mer of A and T. Therefore, IdN could be considered a potential property.

D. Correlation between DB Size and IdN Property [3]

To find the influence of DB size on IdN, k-mers of IdN's patterns must be specified, thus 8-mer and 20-mer have been selected. 8-mer IdN pattern of T means TTTTTTTT. Table 6 shows an IdN pattern for A. thaliana genome (five chromosomes) and some chromosomes of Homo sapience (human) [35], [36]. To

k- mers	A	C	G	T
2	342,132	99,946	95,786	344,599
3	141,800	19,965	18,579	145,782
4	63,728	4,048	3,640	67,145
5	30,781	865	734	33,268
6	16,635	229	196	18,361
7	**10,197**	**75**	**72**	**11,415**
8	6,605	37	40	7,412
9	4,578	20	27	5,196
10	3,225	13	19	3,746
11	2,300	19	14	2,744
12	1,659	8	11	2,058
13	1,224	7	8	1,563
14	922	6	5	1,218
15	724	5	2	972
16	573	4	1	785
17	459	3	0	637
18	368	2		515
19	295	1		416
20	231	0		332
21	183			266
22	145			213
23	107			167
24	76			131
25	53			99
26	39			77
27	31			59
28	24			47
29	18			37
30	13			30

examine IdN patters in more length and larger data set Table 7 demonstrated IdN patterns of 20-mer in Homo sapience (human) genome. Investigated IdN patterns were basic nucleotides of DNA (A, C, G, and T).

Table 5 k-mers of IdN in S. cerevisiae yeast

Evidence from observed data demonstrates two important features; the first is that high concentration of IdN, and the second is that the number of IdN linearly increases with data set size. This increase presents another side of potentiality of IdN's property. As can be seen in the Figure 6, number of IdN grows faster with data set size.

Table 6 IdN of 8-mers in A. thaliana and human genome

Species Name	Chromosome No.	Size (MB)	AAAAAAAA	CCCCCCCC	GGGGGGGG	TTTTTTTT
A. thaliana	1	29.4	32,736	309	294	32,534
	2	19.1	21,851	208	216	21,635
	3	22.7	25,153	273	191	23,533
	4	18	19,680	218	134	19,272
	5	26.1	29,227	307	221	29,594
Homo sapiens (human)	1	242.5	375,804	2,833	2,705	377,033
	2	231.4	358,673	2,751	2,759	360,064
	3	192.6	289,640	2,017	1,782	295,157
	4	185.9	257,954	1,560	1,634	255,945
	6	166.5	249,566	1,824	1,714	250,399
	7	153.2	227,766	1,679	1,665	234,262
	8	142.4	206,737	1,426	1,365	208,730
	20	58	96,302	868	990	103,706

Table 7 demonstrated human genomes and contains IdN of length 20 characters of all DNA symbols. The table also supports previous conclusions regarding the relation between data set size and IdN.

E. Concluding Remarks

Estimation obtained from alternative models and their competing assumptions are often believable. Rather than making an estimate based on a single model, several models can be considered and make the results more confident.

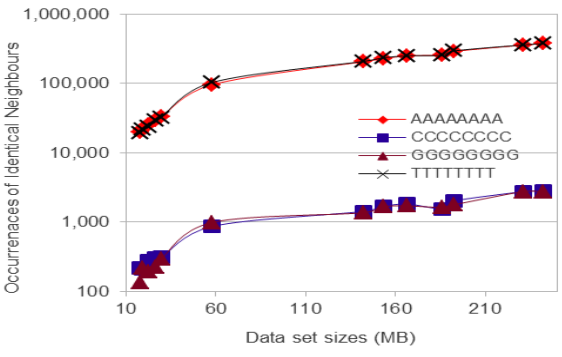

FIG. 6. OCCURRENCES OF IDN WITH DATA SET SIZE. NOTE: IDN OCCURRENCES ARE PLOTTED AGAINST THE DATA SET SIZES ON A LOGARITHMIC SCALE

For these reasons, previous subsections analyzed IdN's property in different models. To empower the analysis results, the following cases are tried:

- The analysis is accomplished at different situations of background frequency and motif model.
- The analysis is done using a distinctive real data set of S. cerevisiae yeast, A. thaliana, and homo sapience.
- The analysis is performed using various measurement of k-mers; from 2-mer to 30-mer.
- The analysis is completed using a variety of data set sizes.

Experimental results from what has been stated above, without any previous assumptions, demonstrate:

1. Strong existences of the property exceeding a quarter of the motif model that motivates to describe it as an axiom. High scoring of the property may correlate with structurally or functionally important genes.
2. Active and vivid presentation of IdN property leading to its description as a potential and significant property
3. Due to GC Island, insertion, deletion and mutation in biological sequences, the frequency of A and T are higher than C and G, therefore IdN of A and T is also higher than C and G. The situation is vice versa in GC Island.
4. IdN pattern is a special kind of patterns that consists of identical symbols with different length, these patterns form part of biological molecular structure. Patterns have been regarded a good task of data mining that reveal information.
5. This property is a novel biological structural property, discovered in this research for the first time.
6. Discovered properties; above remarks guarantee that IdN is important structure property of biological sequence. They motivate to exploit it in molecular biology algorithms generally and in en/decoding genome in particular.
7. Existing motif models suffer from common weakness as discussed in [24], [25]. They assume the occurrence of each base at a particular position of motifs is independent of the occurrence of bases at other positions. The assumption is not true since sequences are biologically related. Thus the naturalistic method of research is necessary, especially in Bioinformatics, because this method concentration on the real data (assumption free methodology) in order to reveal properties.

TABLE 7 IDN IN HUMAN GENOME

Species Name	Homo sapiens (human)				
Motifs template	M1=20 AA;M2=20CC;M3=20GG;M4=20TT				
Sequence set	Size (MB)	Number of Occurrence			
		M1	M2	M3	M4
Chromosome 1	242.458	36372	8	16	37017
Chromosome 2	236.572	34620	8	191	35112
Chromosome 3	192.626	28126	16	2	29062
Chromosome 4	185.945	24513	2	16	23572
Chromosome 5	175.985	24765	0	8	25567
Chromosome 6	166.452	23822	0	4	24820
Chromosome 7	154.803	25104	10	3	26254
Chromosome 8	142.376	20397	2	10	21160
Chromosome 9	137.365	19610	9	16	19170
Chromosome 10	131.842	21362	17	16	20795
Chromosome 11	131.328	19019	7	11	18492
Chromosome 12	130.205	21219	7	0	22340
Chromosome 13	112.032	12531	23	2	12486
Chromosome 14	104.425	14334	14	12	13999
Chromosome 15	99.737	14493	6	0	14496
Chromosome 16	87.893	17523	8	2	17420
Chromosome 17	78.983	18117	3	3	17682
Chromosome 18	75.95	10274	1	0	9599
Chromosome 19	57.519	15625	4	10	15791
Chromosome 20	61.309	9247	5	0	10691
Chromosome 21	46.818	4854	0	0	5376
Chromosome 22	49.907	7294	0	0	7675
Chromosome X	151.04	19956	18	14	18731
Chromosome Y	57.756	4192	0	2	3971
Total	**3011.326**	**447369**	**168**	**338**	**451278**

IdN property is not based on position independency; in contrast it exists arbitrary in biodata. Therefore, the importance of discovered property could be presented in:

1. Reflecting the nature of bio sequences,
2. Expressing biological related functions,
3. Promising to enhance motif representation.

VI. CONCLUSION AND FUTURE WORK

In this paper, a new direction of research method was implimented called 'Naturalistic' which has been conducted in the field of motifs finding, accordingly new biological structure properties have been discovered named IdN and AdjN. The analysis is done in different situations of background frequency and motif model, using distinctive real data set of varied data size. The analysis demonstrated the strong existence of the properties. In the next paper, the new properties we will use them to develop positional join algorithm for motif discovery.

BIBLIOGRAPHY

[1] N. N. Qader and H. K. Al-khafaji, "Motivation and Justification of Naturalistic Method for Bioinformatics Research," *J. Emerg. Trends Comput. Inf. Sci.*, vol. 5, no. 2, pp. 80–87, 2014.

[2] N. N. Qader and H. K. Al-khafaji, "Motif Discovery and Data Mining in Bioinformatics," *Int. J. Comput. Technol.*, vol. 13, no. 1, pp. 4082–4095, 2014.

[3] S. Bandyopadhyay, S. Mallik, and A. Mukhopadhyay, "A Survey and Comparative Study of Statistical Tests for Identifying Differential Expression from Microarray Data.," *IEEE/ACM Trans. Comput. Biol. Bioinform.*, vol. 11, no. 1, pp. 95–115, Nov. 2013.

[4] D.-J. Yu, J. Hu, J. Yang, H.-B. Shen, J. Tang, and J.-Y. Yang, "Designing template-free predictor for targeting protein-ligand binding sites with classifier ensemble and spatial clustering.," *IEEE/ACM Trans. Comput. Biol. Bioinform.*, vol. 10, no. 4, pp. 994–1008, 2013.

[5] M. Friberg, P. Von Rohr, and G. Gonnet, "Scoring Functions for Transcription Factor Binding Site Prediction," *BMC Bioinformatics*, vol. 6, no. 1, pp. 1–11, 2005.

[6] E. Milotti, V. Vyshemirsky, M. Sega, S. Stella, F. Dogo, and R. Chignola, "Computer-aided biophysical modeling: a quantitative approach to complex biological systems.," *IEEE/ACM Trans. Comput. Biol. Bioinform.*, vol. 10, no. 3, pp. 805–10, 2013.

[7] H. Chen-Ming, C. Chien-Yu, and L. Baw-Jhiune, "WildSpan: mining structured motifs from protein sequences," *Algorithms Mol. Biol.*, vol. 6, no. 1, p. 6, 2011.

[8] A. M. Carvalho, A. T. Freitas, A. L. Oliveira, and M.-F. Sagot, "An Efficient Algorithm for the Identification of Structured Motifs in DNA Promoter Sequences," *IEEE/ACM Trans. Comput. Biol. Bioinforma.*, vol. 3, no. 2, pp. 126–140, 2006.

[9] Y. Zhang and M. Zaki, "SMOTIF: efficient structured pattern and profile motif search," *Algorithms Mol. Biol.*, vol. 1, no. 1, p. 22, Jan. 2006.

[10] P. Boyen, F. Neven, D. Van Dyck, F. L. Valentim, and A. D. J. Van Dijk, "Maximally Covering Interactions in a Protein-Protein Interaction Network," vol. 10, no. 1, pp. 73–86, 2013.

[11] E. Loekito, J. Bailey, and J. Pei, "A Binary Decision Diagram Based Approach for Mining Frequent Subsequences," *Knowl. Inf. Syst.*, vol. 24, no. 2, pp. 235–268, Sep. 2010.

[12] P. P. Kuksa and V. Pavlovic, "Efficient motif finding algorithms for large-alphabet inputs.," *BMC Bioinformatics*, vol. 11 Suppl 8, no. 1471–2105, p. S1, Jan. 2010.

[13] F. Masseglia, P. Poncelet, and M. Teisseire, *Successes and New Directions in Data Mining*. Information Science Reference, 2008, p. 386.

[14] M. Piipari, T. Down, and T. Hubbard, "Large-Scale Gene Regulatory Motif Discovery with NestedMICA," *Sci. Eng. Biol. Informatics*, vol. 7, p. 1, 2011.

[15] K. Gouda, M. Hassaan, and M. Zaki, "Prism: An effective approach for frequent sequence mining via prime-block encoding," *J. Comput. Syst. Sci.*, vol. 1, pp. 1–15, 2010.

[16] F. Hadzic, T. Dillon, and H. Tan, *Mining of Data with Complex Structures*. 2011, p. 348.

[17] G. Chen and Q. Zhou, "Heterogeneity in DNA multiple alignments: modeling, inference, and applications in motif finding," *Biometrics*, vol. 66, no. 3, pp. 694–704, 2010.

[18] W. Li, B. Ma, and K. Zhang, "Optimizing Spaced k -mer Neighbors for Efficient Filtration in Protein Similarity Search," vol. 11, no. 2, pp. 398–406, 2014.

[19] "Index of /~zaki/software/sMotif." [Online]. Available: http://www.cs.rpi.edu/~zaki/software/sMotif/. [Accessed: 31-May-2014].

[20] "Arabidopsis thaliana (ID 4) - Genome - NCBI." [Online]. Available: http://www.ncbi.nlm.nih.gov/genome/4. [Accessed: 31-May-2014].

[21] L. Mao and W. J. Zheng, "Combining comparative genomics with de novo motif discovery to identify human transcription factor DNA-binding motifs," *BMC Bioinformatics*, vol. 7, no. Suppl 4, p. S21, 2006.

[22] H. M. Lodhi and S. H. Muggleton, *Elements of computational systems biology*, vol. 8. John Wiley & Sons Inc, 2009.

[23] I. Kulakovskiy and A. Favorov, "Motif discovery and motif finding from genome-mapped DNase footprint data," *Bioinformatics*, 2009.

[24] F. Chin and H. C. M. Leung, "DNA Motif Representation with Nucleotide Dependency," *IEEE/ACM Trans. Comput. Biol. Bioinforma.*, vol. 5, no. 1, pp. 110–9, 2008.

[25] H. Ji and W. H. Wong, "Computational biology: toward deciphering gene regulatory information in mammalian genomes," *Biometrics*, vol. 62, no. 3, pp. 645–663, 2006.

[26] Y. Zhang and M. Zaki, "EXMOTIF: efficient structured motif extraction," *Algorithms Mol. Biol.*, vol. 18, 2006.

[27] M. Halachev and N. Shiri, "Fast Structured Motif Search in DNA Sequences," *Bioinforma. Res. Dev.*, pp. 58–73, 2008.

[28] K. Jensen and M. Styczynski, "A Generic Motif Discovery Algorithm for Sequential Data," *Bioinforma.*, vol. 22, no. 1, pp. 21–28, 2006.

[29] M. Hasan, Q. Liu, H. Wang, and J. Fazekas, "GIST: Genomic island suite of tools for predicting genomic islands in genomic sequences," vol. 8, no. 4, pp. 203–205, 2012.

[30] W. K. Ching and M. K. Ng, *Markov chains: models, algorithms and applications*, vol. 83. Springer-Verlag New York Inc, 2006.

[31] T. Jiang, Y. Xu, and M. Q. Zhang, *Current topics in computational molecular biology*. The MIT Press, 2002.

[32] V. S. Mathura and P. Kangueane, *Bioinformatics: a Concept-Based Introduction*. Springer Verlag, 2009, p. 196.

[33] T. T. Nguyen and I. P. Androulakis, "Recent Advances in the Computational Discovery of Transcription Factor Binding Sites," *Algorithms*, vol. 2, no. 1, pp. 582–605, Mar. 2009.

[34] E. Lander, "Initial impact of the sequencing of the human genome," *Nature*, vol. 470, no. 7333, pp. 187–97, Feb. 2011.

[35] "Genomes - Genome - NCBI." [Online]. Available: http://www.ncbi.nlm.nih.gov/genome/genomes/4. [Accessed: 31-May-2014].

[36] "GoldenPath of currentGenomes -Homo_sapiens chromosomes." [Online]. Available: ftp://hgdownload.cse.ucsc.edu/goldenPath/currentGenomes/Homo_sapiens/chromosomes/. [Accessed: 05-Apr-2011].