The Nucleus **51**, No. 1 (2014) 125-131

**The Nucleus**
A Quarterly Scientific Journal of Pakistan Atomic Energy Commission
NCLEAM, ISSN 0029-5698

# SMGCD: METRICS FOR BIOLOGICAL SEQUENCE DATA

*M. IDREES and M.U.G. KHAN[1]

Department of Computer Science and Engineering, University of Engineering and Technology, Lahore, Pakistan

[1]Bioinformatics Lab. Al-Khwarizmi Institute of Computer Science, University of Engineering and Technology, Lahore, Pakistan

(Received November 11, 2013 and accepted in revised form January 01, 2014)

In the realm of bioinformatics, the key challenges are to manage, store and retrieve the biological data efficiently. It can be classified in to structured, unstructured and semi-structured contents. Typically, the semi-structured biological data comprised of biological sequences. The complex biological sequences produce huge volume of biological data which further produce much more problems for its management, storage and retrieval. This paper proposed metrics; namely, symmetry measure, molecular weight measure, similarity or diversity measure, size base measure, size gap measure, complexity measure and size complexity diversity measure to manage the raised problems in biological data sequences. These metrics measure the sequence complexity, molecular weights, length with gaps and without gaps, its symmetry and similarity through mathematical formulations. The metrics are demonstrated and validated using the proposed hybrid technique which combines empirical evidence with theoretical formulation. This research opens new horizons for efficient management to measure the functionality and quality of metadata for single and multiple biological sequences.

**Keywords**: Semi-structured, Symmetry, Metadata, Diversities, Metrics, Hybrid

## 1. Introduction

Bioinformatics is a broad domain that encompasses richness of biological concepts, different relationships among molecular classes, hierarchies from macro molecules to micro molecules and diversity of information. Genome, transcriptome and proteome are considered to be the macromolecules in the field of molecular biology. These three macromolecules play the most essential and important roles in all sorts of organisms. They are contributing vitally in production of biological sequence data [1]. A large number of directions for research in bioinformatics domain are propagated, according to the availability of biological information. It has become necessary to assess and measure relevancy and similarity at attribute level and functional level for its cohesiveness by designing variety of its measurements in different dimensions [4].

Biological data can be classified into structured; unstructured and semi-structured contents in the domain of bioinformatics. Often, it is considered as semi-structured due to loosely-typed and diversities of its data formats. Mostly semi-structured biological data comprised of biological sequences. The rapid development of biological sequence data i.e., its volume and nature, increasing its popularity, which make its usage more complex. This growth involves Moor's law [2] for its evolutionary representation. The production yielded due to the usage of different sequencing technologies is actively playing critical role in increasing its complexity for its management due to new inventions and discoveries. The burning issues involved are due to the lack of unsuitability of data model for semi-structured data. Due to improper management in diversity of formats, dimensions and heterogeneity of databanks, it is facing a large number of uncertainties and inconsistencies among these biological sequences [3]. Biological sequence data is derived from not only in human beings; rather it is being produced in chimpanzee and gorilla also. Different techniques of sequence comparison, subsequence comparison are introduced in [5]. Biological sequences composed of long chains of alphabetical letters named nitrogenous bases that comprise various biological structures and sequences for genome, transcriptome and proteome. They also contain a large number of amino acids which make the bases for protein structures. Sequence biological data has major shortcomings in order to measure and determine its quality [6].

Metrics prove helpful to measure and assess the quantity and quality of theoretical research. The performance assessment can be made by selectivity of certain features. Basically they are used to measure and evaluate the quality of desired product by defining its threshold values and other significant roles like accuracy, consistency by means of mathematical ratings, statistical methods and analysis. Such measurement incorporates its functionality and how well it is working and fulfilling the user requirements [7]. A little attention is made by research community towards metrics identification and validation in the domain of bioinformatics.

Our contribution in this research work includes proposing SMGCD metrics for biological sequence data, which can be seen in section proposed work. The

* Corresponding author : midrees10@gmail.com

SMGCD metrics help us in understanding the sequence, sub-sequences, repeated patterns and their analysis and to determine the relationships among these biological sequences in detail.

The remaining paper is organized as following. In section 2, literature survey, we present related work, close to our work. Section 3, proposed work is used to discuss our proposed metrics particularly how they define the operations on biological sequence data and measure its quality. Section 4, defines the validity of our proposed metrics through different examples. We have used hybrid approach for the validation of our research work. In conclusions and future work section, we provided some suggestion and intentions for its extendibility.

## 2. Literature Survey

In this section we present existing research work available close to our proposed work briefly. Current research efforts for biological sequence metrics comprises the bioinformatics domain overall. Most of the existing work is related to similarity measure by using different techniques. These techniques get involved edit distance, distance metric, dynamic programming, integrity functions and by measuring neighboring elements for defining these biological metrics. In short a little attention is focused to address towards defining biological metrics [2]. Jason Lee et al. [8] proved in the literature that similarity searching techniques using clustering techniques for biological sequences are entirely different from traditional clustering techniques. They addressed the clustering utility metrics in their research and empirically proved that utility metrics measure strong cohesion to the quality of biological sequences. They claimed that with increasing the volume/size of the data does not affect to measure its quality. Manicassamy et al. [7] assess the mapping methodologies i.e., their cost, quality assurance, complexity, similarity, performance, optimization and scheduling from empirical to formal transformations. Consequently, they addressed that the process of assignment a symbol or a number to a certain data values of objects is used for their assessment and measurement.

In literature, designed metrics asses the similarity among transcriptome biological sequences based on diversity of techniques i.e., edit distance, base pair class based and mountain class based. By using edit distance, symmetry metric is measured through calculation of distances between the consecutive bases of biological sequences. They stated that if the distance between its all neighboring bases is same, then the bases are said to be symmetric discussed by Moulton et al. [9].

Waterman et al. [10] proposed some metrics based on biological sequences. Their research work based on distance metrics for measuring similarity, dissimilarity and evolutionary path tracking, insertion and deletion of single letter and double letters. They used edit distance formulae to prove their proposed metrics by generalizing s-metrics and t-metrics of previous work. Louie et al. [11] have discussed theoretically uncertainty metrics based on quality of biological data resource, references of these heterogeneous resources, contents containing these data resources and relationship among them. Based on these metrics, they exploited the idea of integration of biological data resources through biomediator approach. Their research work includes four uncertainty metrics consisting *Ps Measure*, *Qs Measure*, *Pr Measure* and *Qr Measure* respectively.

Shah et al. [16] proposed the similarity measures metrics for phylogenic tree for mitochondrial genomes provides the bases on effective, normalized, universal and informational distance measuring parameters and propagated general mathematical theory. Preliminary investigations reported that measuring distance is proportional to the energy. Smaller the distance between two consecutive objects, energy generated or consumed is minimum. The rate of uncertainty can be reduced to some extent and the ratio of bases addition or deletion in any sort of genome sequence is minimized. They stated that different evolutionary biological trees have been constructed for different genome available in the literature by Li et al. [12].

Raibi et al. [17] have narrated the metrics of stability and effects of stability on variations for protein sequences during motive selections and predictions. Further they have propagated their work by extending motif stability, method sensibility, conservation and ratio of motif stability for protein sequences during motif predictions as also narrated by Saidi et al. [13]. Darling et al. [14] have proposed and exemplified different metrics for high throughput genome sequencing data to assess their quality. Their works have been approved by using different assemblers to measure their quality for genome sequencings. Useful metrics proposed and measured includes Miscalled Bases, Uncalled Bases, Extra bases, Missing Bases, Extra Segments and Missing Segments outlined [14].

The detailed comparison analysis of existing metrics for this domain is presented in Table 1. The tabular structure helps the research community to understand the latest art of work. In Table 1, data metrics represents the proposed metrics, characteristics represents to proposed methods, application area, difficulty level and tools used to measure these metrics.

### *2.1 Analysis of Existing Metrics*

Table 1. Comparative analysis of existing metrics for biological sequence data

| Data Metrics/ Characteristics | Proposed Methods | Applications Area | Difficulty Level | Tools Used |
|---|---|---|---|---|
| Sequence similarity, local similarity, global similarity, scoring functions etc. [1] | Geometrically | Generic biological sequences | High | Empirical |
| Similarity, evolutionary distance etc. [2] | Statistically | Generic | Low | BLAST |
| Quality Metrics [4] | Mathematically | | Medium | Quality Augmented Model |
| Secondary structure, polarity, codon diversity score, electrostatic diversity etc. [6] | Multivariate analysis | Amino acids | Medium | Numerical assessment |
| Cluster utility etc. [8] | Graphically | Generic | Low | Clustering approach |
| Mountain metrics, Global, Local etc. [9] | Mathematically | RNA | Medium | Ensemble |
| Deletion, insertion, mutation etc. [10] | Mathematically | Nucleotide sequences | Medium | Distance Calculations |
| Uncertainty measures etc. [11] | Probabilistically | Protein structures, Genes | Low | COG and Bio-Mediator System |
| Similarity measure etc. [12] | Empirically | Mammalian DNA | High | Evolutionary and language tree |
| [14] | Empirically | DNA, genome | High | Scoring methods |

## 3. Proposed Work

In this section we propose SMGCD novel metrics for biological sequence data. The vast usage and fragmented accumulation of biological sequence data in diverse repositories has generated enormous challenges for bioinformatics research community to measure its quality [16]. The SMGCD metrics are helpful in defining the attributes and functions from biological sequence data especially for genome, transcriptome and proteome. Biological sequences and their structures have a close relationship due to their base compositions and hierarchical structures [17]. The SMGCD metrics are examined by exemplifying based on local and global studies of biological sequence data. The measuring aspects are represented by using different notations, formulas and functions respectively as follows:

### *3.1. Symmetry Measure (SM)*

This metric used to measure the similarity among different genome, transcriptome and proteome. Let B = {A, C, G, T, U} provides a set of nitrogenous basis for different biological sequences. We can define and measure the symmetry of biological sequences by using the following defined function

SM = f (A, C) = f (T, G) = { 1 iff A = T $\wedge$ C= G} +

f (T, G) = f (A, C) = {1 iff T = A $\wedge$ G = C}

The base A makes hydrogen bond with base T and base C makes hydrogen bond with base G, so they are said to be symmetric if in both sequences, the corresponding bases are satisfying the above conditions for DNA and RNA biological sequences. In RNA, the property remains same because only base T is replaced by U. This metric (symmetric property) helps for the determination of subsequence, similarity measure and pattern recognition and gene identifications in different biological sequences. By using comparison mechanism for this metric, we can predict and identify new patterns, functions and biological structures.

### *3.2. Molecular Weight Measure (MWtM)*

This metric is used to measure the molecular weight of biological sequences. It determines the molecular weights of all biological sequences by taking the sum of molecular weights of all the bases existing in a biological sequence.

M wt. M = Sum of ($\sum_{i=0}^{\tau}$(Bi) )

Where Bi denotes the molecular weights of all biological bases contained in a specific sequence. It helps us to determine the molecular weights of DNA, RNA and protein structures and their derived biological

sub-sequences. The proposed metric facilitates for sequence analysis, for the identifications of motifs and subsequences, structures classifications and for predictions of various biological objects [7].

### *3.3. Similarity/Diversity Measure (S/DM)*

Similarity metric used to determine/measure the distances between various consecutive neighboring biological objects. The measured distance can be effective during normalization and its measuring characteristics depend on different parameters [12]. Our proposed method of measuring similarity is simple than other existing methods used in the literature. Similarity measure is a metric which determines the similarity between two biological sequences by comparing their corresponding bases of different sequences. Let us consider two biological sequences i.e., A∧ B; and they have their lengths L1 ∧ L2 respectively. These lengths i.e., L1 ∧ L2 can be characterized as L1 = $\prod_{i=0}^{j} B_i$∧ L2 =$\prod_{m=0}^{r} B_m$ respectively. Then the similarity measure s/d M can be defined as:

s/d M= $\prod_{i=0}^{j} B_i$ / $\prod_{m=0}^{r} B_m$= { 1 if L1 = L2 ∧ m=n }

= {0 if A≠ B ‖ m≠n }

In case 1 similarity is achieved to be 100% if and only if, both of the biological sequences are equal in length and their corresponding letters quantify similar bases which denote to exact matching.

In case 2, it measures the diversity/dissimilarity for given biological sequences, if the numbers of given biological sequences are dissimilar, or they do not have exact match of corresponding alphabets, then we say both sequence are dissimilar or diverse.

This is very powerful operator, which determines the many sorts of functions i.e., similarity, diversity, mismatching, comparison and many others among biological sequences. Similarity of two biological sequences is measured by comparing their lengths and concurrent bases of both biological sequences.

### *3.4. Size Base Measure (SBM)*

This metric is used to determine the size (count the numbers) of biological bases in a given biological sequences when the given sequences do not consist of any gaps or any missing values among them. Size base measure is useful metric for the determination of length of biological bases of these sequences. Let B = {$X_1$… $X_T$} denotes the number of bases in a sequence, then SBM metric is used to determine its size as following

sbM = $\sum_{i=0}^{r}(Bi)$ when $B_i \neq$ gaps

It is to remember that this useful metric is used to measure the fine grained purified and curated biological sequences only which are gap free among them. Its output determines certain number that tells the actual length of the specific biological sequence. It helps to facilitate for measuring numbers and the orders of genes and for the identification and prediction of new gens among biological sequences. By using this function we can determine the actual length of biological sequences.

### *3.5. Size Gap Measure (SGM)*

This useful metric is used to find out the overall length (count the number) of bases, missing values and gaps of sequences as well. SGM metric helps us for measuring the general length of a given biological sequence. Let S is a sequence, which consist of a number of bases containing irregular gaps among these bases. The SGM metric is used to calculate its size with gaps by using the function as follows:

SGM = $\sum_{i=0}^{r}(Si)$ where $S_i$ = {{A, C, G, T, U}∧{gaps}}

Its output determines certain number that tells the overall length of the specific biological sequence and total number of genes. This metric helps to facilitate for measuring overall length of biological sequences, which might include the regular and irregular number of gaps and other missing values.

### *3.6. Derivation Gaps Measure (DGM)*

As the name indicates, that derived gaps measure metric is used to calculate the number of gaps in any desired biological sequences. This metric is derived from the above proposed two metrics like size base measure and size gaps measure, in general. The results of the above proposed metrics i.e., SBM and SGM can be used to calculate the numeric values, so the DGM is calculated as following:

DGM = output (SGM) - output (SBM)

The outputs of both mentioned metrics i.e. size gaps measure and size base measure must be numeric values, so their negation is also a numeric. The proposed metric shows, if we know the sequence lengths of only bases and with gaps, then we might be in a position to determine the number of gaps in a specific sequence. DGM is beneficial metric, by using this metric we can predict number of genes and the organization these genes in sequence bases.

### *3.7. Complexity Measure (CM)*

Complexity measure for a linear sequence is determined by O (CM). Biological sequences are linear which may contain gaps and without gaps, so their complexity might be determined as for linearly sequences. It must be O (CM) = n, for linear sequences.

If we measure separately for linear sequences without gaps and with gaps, then it may be n+n = 2n, by ignoring co-efficient 2, it remains n only. Here it is important that, this metric measures complexity for addition of biological sequences only. It may be different in case of multiplication and division of two or more biological sequences.

In case of single and linear biological sequences, the complexity is simply determined by using the mentioned metric i.e., "n". In case of multiplication or union of two sequences, the complexity may be denoted as $n^2$.

### *3.8. Size Complexity Diversity Measure (SCDM)*

SCDM metric basically the description of theoretical inter-related concepts. These three terms are highly interconnected to each other. SCDM represents the monotonic relationship among biological sequences. Now, we consider size, complexity and diversity, three independent terms and use different notations for their description i.e. S, C and D respectively. Here notation S measures its volume, notation C measures heterogeneity and D measures different formats of biological sequence data. Then there exists a monotonic relationship i.e. if S increases then D also increases and if D increases then C also increases among them, which is represented as follows:

S α D ∧ D α C => S α C, because S, D and D, C are directly proportional to each other, which can be written as follows:

S > = C >= D, it means
SCDM = increasing monotonically

## 4. Results and Discussion

The proposed metrics for biological sequence information are presented in section 3. For validation of our designed metrics, we have used hybrid approach. The proposed hybrid approach can be used both for the theoretical and empirical validation of research work. Due its usage for both theoretical as well as for empirical, we named it hybrid. Hybrid approach is the extension of existing work discussed in [15]. The proposed framework for our hybrid approach can be described in Figure 1.

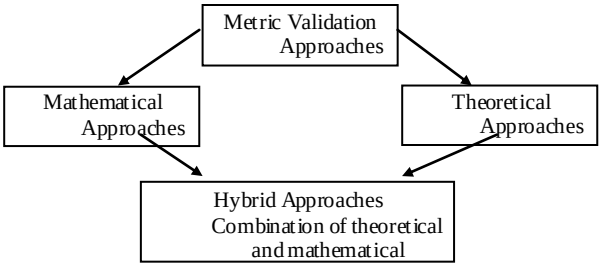


Figure 1. Metrics validation strategy.

### *4.1. Hybrid Validation Approach*

The Figure 1 shows the proposed hybrid approach for metric validation strategy. The proposed metrics are assessed by using above hybrid strategy. To assess the validity of proposed symmetric metric, let us take an example.

Y = {5´-CCAGTAA-3´}∧

Z = {3´-GGTCATT-5´}

Two fragments of biological sequences which are denoted by different set notations i.e., Y and Z respectively. The directions i.e. clockwise and anti-clockwise of both biological sequences i.e., Y and Z can be denoted by 5´- 3´ and 3´- 5´notations respectively. The elements of both sets i.e., Y and Z, consist of different nitrogenous bases which present the actual contents of biological sequences. The nitrogenous base A forms double covalent hydrogen bond with nitrogenous base T and similarly base C makes triple covalent hydrogen bond with base G. By using our proposed symmetric metric, both sequences given in set Y and Z are symmetric, because they fulfill all requirements and conditions of symmetry. Both sets i.e., Y and Z contain nitrogenous bases which are complement and having opposite directions to each other. Hence these two biological sequences belongs to two different sequence structures interconnecting the bases. This verifies the requirements of symmetry metric.

The second proposed metric is used to determine the molecular weight of the complete biological sequence. Let us consider a set of sequence such that

X = {ACGT}

The set X consists of sequence of only four bases as given. We propose here empirical values as molecular weights against these bases, such that A = 10.00, C = 12.00, G = 13.00 and T = 25.00 corresponding their molecular weights in a sequence. The proposed metric will assess the value by taking sum as following

M wt. M = Sum of $(\sum_{i=0}^{\tau}(Bi))$ = 50.00

Hence it measures X = 50.00 empirical value for the proposed sequence X.

Our next proposed metric measures the similarity and diversity between two or more biological sequences. The proposed metric is illustrated by taking three different cases with different examples.

**Case 1**

In Case 1, we consider two sets of biological sequences, such that

Y = {CCAGTAA} ∧

Z = {GGTCATT}

Both of the biological sequences represented in sets i.e., Y and Z can be entirely dissimilar. Both sequences have equal length i.e. number of bases are 7, which is common. When we compare the corresponding bases of these sequences, they represent different behavior.

**Case 2**

For Case 2, let us consider again two sets Y and Z, as following

Y = {CCAGTAA} ∧

Z = {CCAGTAA}

Both sets of sequences have equal length and the corresponding bases are also similar. The threshold value for these biological sequences is 1 i.e., it measures the similarity 100%.

**Case 3**

For Case 3, again here we consider

Y = {CCAGTAA}∧

Z = {CCAGTAAAGGA}

If we analyze the two sets of biological sequences i.e., in set Y and Z, it is clear that set Y is a subset of Z, but according to our metric, it measures dissimilarity. Though it determines subsequence, but our proposed metric measures it dissimilar, due to not having equal length of both sequences.

size base measure metric sbM = $\sum_{i=0}^{\tau}(Bi)$,

This metric simply determine the length of bases in biological sequences. In case of set Z as given below:

Z = {CCAGTAAAGGA}

It determines its length as 11 numeric characters. The point to note here, it does not contain any gap among bases in set Z. It consists of only contents of bases which count 11.

In size gap measure, let us consider a set of sequences such that

Z = {A---A----CG-G-CCT---T--CC}

The set Z contains some missing/null values in this biological sequence. Missing or null values are uncertain values, about which we do not establish any kind of fact. The size gap measure metric calculates the entire length including gaps of the sequence.

So, sgM = $\sum_{i=0}^{\tau}(Si)$

Such that

$S_i$ = {{A, C, G, T, U} ∧ {gaps}}

It determines its overall length i.e., 25, out of which only 11 are known bases and 14 are the gaps or unknown values. If we take difference of 25 – 11 = 14, our next metric measures these gaps in the given biological sequences.

## 5. Conclusions

In this paper, we proposed different metrics for biological sequence data at fundamental level. They can be proved helpful in identifying and determining the relationships between different biological sequences of genome, transcriptome and proteome. These metrics are also useful for the determination and extraction of various functions for biological sequences. It is supposed to help us in understanding the primary, secondary and tertiary structures and sequences of DNA, RNA and Protein structures, and for the prediction of various genes based on size of biological sequences. In the last but not the least, the research work measures and assesses the quality of the given biological sequences in different perspectives and dimensions with the help of examples. The current proposed work falls at micro-level and will be extended for different biological domain concepts at macro-level for its quality assessments and measures.

## References

[1] A. Stojmirovic and Y.K. Yu, Journal of Computational Biology **16**, No. 4 (2009) 579.

[2] D. P. Miranker, Metric-Space Search in Bioinformatics. National Science Foundation, Institute of National Health **2**, No. 2 (2010) 32.

[3] W. J. MacMullen and S.O. Denn, Journal of the American Society for Information Science and Technology **56**, No. 5 (2005) 447.

[4] A. Martinez and J. Hammer, Making Quality Count in Biological Data Sources, Proceedings of the 2nd International Workshop on Information Quality in Information Systems (June 2005) pp. 16-27.

[5] M. Schoniger and M.S. Waterman, Bulletin of Mathematical Biology, Elsevier **54**, No. 4 (1992) 521.

[6] W. R. Atchley, S. J. Zhao, A. D. Fernandes and T. Druke, Proceedings of the National Academy of Sciences of United States of America **102**, No. 18 (2005) 6395.

[7] J. Manicassamy and P. Dhavchelvan, International Journal of Recent Trends in Engineering **1**, No. 1 (2009) 550.

[8] J. Lee and S. Kim, Cluster Utility: A New Metric for Clustering Biological Sequences Proceedings of the 2005 IEEE Computational Systems Bioinformatics Conference Workshops and Poster Abstracts, IEEE Computer Society (August, 2005) pp.45-46.

[9] V. Moulton, M. Zuker, M. Steel, R. Pointon and D. Penny, Journal of Computational Biology **7**, No. ½, (2004) 277.

[10] M. S. Waterman, T. F. Smith and W.A. Beyer, Journal of Advances in Mathematics **20**, No. 3 (1976) 367.

[11] B. Louie, L. Detwiler, N. Dalvi, R. Shaker, P.T. Hornoch and D. Suciu, Incorporating Uncertainty Metrics into a General-Purpose Data Integration System, 19th International Conference on Scientific and Statistical Database Management, IEEE Computer Society (July 2007) pp. 19.

[12] M. Li, X. Chen, X. Li and B. Maw, IEEE Transactions on Information Theory **50**, No. 12 (2004) 3250.

[13] R. Saidi, A. Saber, M. Mondher and M.N. Engelbert, Novel Metrics for Feature Extraction Stability in Protein Sequence Classification. LIMOS: Blasé Pascal University **1** (November 2011) pp. 1-7.

[14] A.E. Darling, A. Tritt, J.A. Eisen and M.T. Faccoitti, Journal of Bioinformatics **27**, No. 19 (August 2011) 2756.

[15] Shazia, M. Shoaib, Iqra, K. Kalsoom, S. Majid and F. Majeed, Pakistan Journal of Science **63**, No. 1 (2011) 26.

[16] S. Shah, Applied Mathematics Corner: DNA Computation and Algorithm Design, Harvard University, Cambridge, MA 02138 (2009) pp. 83-89.

[17] R. Saidi and S. Aridhi, Feature Extraction in Protein Sequences Classification: A New Stability Measure, Proceedings of the ACM Conference on Bioinformatics, Computational Biology and Bio-Medicine (2012) pp. 683-689.