Modeling Biological Signals using Information-Entropy with Kullback-Leibler-Divergence

Anjali Mohapatra[†], **P.M.Mishra**[††], **S.Padhy**[†††],

IIIT, Bhubaneswar, Orissa Dept. of Energy, Govt. of Orissa Utkal University, Orissa

Summary
Biological signals are short conserved regions in DNA, RNA, or Protein sequences which correspond to some structural and/or functional feature of the bio-molecules. Finding such signals has important applications in locating regulatory sites and drug target identification. Identification of biological signals such as motifs is a challenging problem because they can exist in different sequences in mutated forms. Despite extensive studies over last few years this problem is far from being satisfactorily solved. Most existing methods formulate signal finding as an intractable optimization problem and rely either on expectation maximization (EM) or on local heuristics. Another challenge is the choice of model: simpler models such as positional weight matrices (PWM) impose biologically unrealistic assumptions where as other harder models are difficult to parameterize. In this paper a conceptually simpler and biologically relevant model based on Kullback-Leibler divergence along with information entropy framework is proposed to measure the divergence of the biological signals. Both synthetic and real data are used to test applicability of the proposed model for finding motifs in DNA sequence. Our proposed model performs better than models based on PWM or Shannon entropy.

Key words:
Information entropy, Biological signal, Motif, Kullback-Leibler divergence.

1. Introduction

Information entropy, first introduced by Claude E. Shannon in his paper "A Mathematical theory of Communications" in 1948 describes information content of a signal or event [1]. Although, it was used for modeling communications in a channel, the mathematical concepts can also be applied to other fields such as living system [2].

The information in genomic sequences can be encoded as a string over the four letter language of nucleotide bases {A, C, G, T} and genes can be viewed as substrings of a DNA sequence. These substrings are used by living cells as the blueprints for making specific proteins that carry out all the functions of cellular life. While all the cells of an organism contain the same DNA sequences, different cells make radically different sets of proteins based on different conditions or needs. Embedded in the non-coding DNA (introns) of an organism are many control sequences that influence when a gene is expressed and which coding portions of the gene (exons) are spliced together to form the gene product. Commonly, the DNA region just upstream of a particular gene is thought to contain many short substrings or signals of DNA that do not actually encode any part of a protein, but act as controlling signal for the associated gene's expression. These and other sequences that may be used to regulate gene expression are collectively known as regulatory signals [3]. A fundamental challenge in computational biology is to identify these important signals. Despite extensive studies over last few years the challenge of finding potential significant biological signals is far from being satisfactorily solved [4, 5, 6, 7, 8].

Shannon's theory deals with information over a communication channel and provide tools which are based on measuring information in terms of bits or in terms of the minimal amount of complexity of structures needed to encode a given piece of information. The biological information encoded in DNA is transmitted through the process of transcription, translation and mutation and decoded as proteins. Shannon's theory has been used to study genomic sequences by calculating the amount of information contributed by individual nucleotides during these encoding and decoding processes [9].

1.1 Entropy and biological Signals

There are close parallels between the mathematical expressions for the thermodynamic entropy S, established by Ludwig Boltzmann in the 1870s and the Shannon information theoretic entropy H [1]. Information Theory is the mathematical theory of communication to find out the speed and quantity of information transmission. It uses statistical concepts of probability to compute the extra information (redundancy) necessary to counteract the distortion and losses that may occur during transmission from one information source to another. The theory uses entropy as the "measure of the rate of transfer of information in [that] message".

Manuscript received January 5, 2009
Manuscript revised January 20, 2009

Let ∑ be an alphabet of size N, for DNA ∑ = {A, C, G, T}.

When a symbol is received from an alphabet of size N, where letters are sent with equal probability and the symbol is transmitted as a binary signal, then we must wait until all $\log_2 N$ bits of the symbol are sent before it is known with certainty which symbol has been transmitted. Thus the information entropy contained in a symbol from an alphabet of size N is $\log_2 N$, under uniform probability distribution.

Let $p_1 = \Pr[A_1],...,p_n = \Pr[A_n]$ are the probabilities of outputting characters A_i in a message, where $\sum_{i=1}^{N} p_i = 1$.

Suppose n, the length of a random message M (or length of nucleotide sequence) is large, let $n_i = np_i$ be the expected number of occurrences of A_i. Then the message M belongs with high probability to a set of size given by the multinomial coefficients $N_n = \frac{N!}{n_1!...n_N!}$ representing the number of ways of partitioning N into a collection of sets of sizes $n_1,...,n_N$.

The average information, $I = \frac{\log_2 N_n}{n}$

Using Sterling's formula

$$N_n \sim \frac{\sqrt{2\pi n}\, n^n e^{-n}}{\sqrt{2\pi n_1}\, n_1^{n_1} e^{-n_1} ... \sqrt{2\pi n_N}\, n_N^{n_N} e^{-n_N}},$$

$$\ln N_n \sim n\ln n - np_1 \ln(np_1) - ... - np_N \ln(n_N p_N)$$

$$= -n\sum_{i=1}^{N} p_i \ln p_i$$

Since ln and $\log_2$ are related by a constant:

$$\log_2 N_n \sim -n\sum_{i=1}^{N} p_i \log_2 p_i,$$

Thus Entropy $H(p_1,...,p_n)$ in Shannon's formula:

$$H(p_1,...,p_n) = I = \frac{\log_2 N_n}{n}$$

$$= -\sum_{i=1}^{N} p_i \log_2 p_i$$

The measure of information (*Shannon entropy*) is mathematically, similar to that of disorder (*Boltzmann entropy*).

2. Existing Modeling Approaches

2.1 Models based on Positional Weight Matrix

Positional weight matrices (PWM) are foundations to many biological problems and are commonly used to represent consensus sequences in phylogeny, structure prediction, motif finding, sequence alignment etc. It specifies the probability that a given base is observed at each index position of the sequence. Let t be a set of DNA sequences, each of which is n nucleotides long and l be the length of the unknown signal. Selecting one position in each of these t sequences forms an array $S = (S_1,S_2,....,S_t)$, with $1 \le S_i \le n - l + 1$. The l-mers starting at these positions can be compiled into an $t \times l$ alignment matrix, whose $(i,j)^{th}$ element is the nucleotide in the $(S_i + j - 1)^{th}$ element in the i^{th} sequence. Based on the alignment matrix, the $4 \times l$ profile matrix can be computed whose $(i,j)^{th}$ element holds the number of times nucleotide I appears in column j of the alignment matrix, where I varies from 1 to 4. A consensus string can be formed from the most frequent element in each column of the alignment matrix, which is the nucleotide with the largest entry in the profile matrix. [10]

Let $P(s)$ denotes the profile matrix corresponding to starting positions s, $M_{P(S)(j)}$ denotes the largest count in column j of $P(s)$. Given starting positions s, the consensus score is $Score(s,DNA) = \sum_{j=1}^{l} M_{P(S)(j)}$, which measures the strength of a profile corresponding to the given starting positions.

The Motif finding problem thus can be formulated as:

Given a set of DNA sequences, find a set of l-mers, one from each sequence that maximizes the consensus score.

Input: A $t \times n$ matrix of DNA, and l the length of the pattern to find.

Output: An array of t starting positions $S = (S_1,S_2,...S_t)$ maximizing $Score(s,DNA)$

As an example, consider the Cyclic AMP receptor proteins (CRP) as a transcription factor of E.Coli. Its binding sites are DNA sequences approximately twenty-two in length. The following sequences,

TTGTGGC	TTGTGAT	GCCTGAC
TTTTGAT	ATTTATT	TTGTGAT
AAGTGTC	ACGTGAT	TTGTGAT
ATTTGCA	ATGTGAG	GTGTGAA
CTGTGAG	TTGTGAG	CTGTGAC
ATGCAAA	CTGTAAC	ATGAGAC
GTGTTAA	CTGTGAA	TTGTGAG
ATTTGAA	TTGTGAC	

taken from Stormo and Hartzell shows the 3-9 positions of the 23 CRP binding sites [11]. The following table indicates the PWM.

A	.35	.043	0	.043	.13	.83	.26
C	.17	.087	.043	.043	0	.043	.3
G	.13	0	.78	0	.83	.043	.17
T	.35	.87	.17	.91	.043	.087	.26

Once a PWM is created, it can be used to score new sequences. An important assumption made here is that the base identity at any position does not depend on the base identity of any other position in the sequence. We can calculate the joint probability by multiplying all individual probabilities at each position. Molecular biologists determine the most frequent base at each position called a 'consensus'. The position 3 of the above PWM has 0% A, 4.3%C, 78%G, and 17%T. If we make a consensus model of this position, we would call it 'G'. This in turn means, when one looks into this site, 22% of the time it will not be correctly recognized. Yet, this method is extremely widespread in computational molecular biology. Models based on only consensus sequences as foundation is thus expected to make significant error in real biological predictions.

Limitations: Existing algorithms based on PWM consider the most popular element in each column of the alignment matrix, and completely ignore the rest which are comparatively weak (probability of occurrence is less). From information entropy point of view all sequences (strong or weak) can contain different amount of information about the biological signal in various forms and thus can be a potential candidate for replacing the PWM model.

2.2 Models based on Information Entropy

Information Entropy describes the communication of symbols through a channel [12].

The entropy $H(p_1,..., p_n)$ in Shannon's formula has been obtained as:

$$= -\sum_{i=1}^{N} p_i \log_2 p_i$$

= The measure of information (*Shannon entropy*)

Given a set of genomic sequences, the above formula can be used to rank each member sequence according to its information content. The information content is represented in bits which provides a universal scale and allows information from independent substrings to be summed together. This helps in evaluating each input sequence and takes care of weak signals which were completely ignored by earlier PWM model. For example a new sequence GGAGCCG is completely ignored since the probability of each position is lowest in the PWM. However, the information content of this sequence is equal to the information content of an existing sequence TTGTGAT that differs from the consensus sequence ATGTGAC at only two positions and occurs twice in the list of 23 sequences.

Limitations: The model based on information entropy measures the information quantity; it says nothing about the quality of information. It may so happen that given two substrings in a sequence with equal quantity of information, one may contain biologically relevant information (e.g. motif) while the other does not. When the goal is to find signal of interest (e.g. motif) from voluminous biological data, accurate identification is difficult because they possess same quantity of information and are short size in the midst of noise. For example the sequences TTGTGAT and GGAGCCG have the same information content in spite of their wide distance from the consensus. (The sequence TTGTGAT differs at 2 positions where as GGAGCAG differs at all 7 positions from the consensus). Thus we need a model that besides measuring the quantity of information can also distinguishes between relevant and irrelevant information.

3. Proposed Model Using Shannon Entropy Along With Kullback-Leibler Divergence.

3.1 Motivation from biological observation.

When individual nucleotides are linked by condensation reactions, Nucleic - acid chains such as DNA/RNA sequences are formed. In condensation reaction a molecule of H_2O is liberated when two nucleotides are joined. The $5'$ phosphate of the incoming nucleotide is linked to the $3'$ hydroxyl (-OH) group of the growing chain. The nucleic acid chain is thus extended in the $5' \rightarrow 3'$ directions. In living organisms, such reactions are catalyzed by polymerase enzymes and an incoming nucleotide building block has a chain of three phosphate groups that is cleaved to provide energy for the chain-building reaction [13]. Hence, individual nucleotides are

more “pure” in the form of nucleotide composition. When they join the resulting chain of nucleotides become more stable than individual nucleotides by liberating H_2O in the process of condensation reactions, there by moving to a lower entropy state. The biological motif, carry itself repeatedly in each of the sequences; can be viewed as a much more stable structure and is expected to have entropy different from other stretches of the sequence.

3.2 Motivation from physical observation.

If the energy distribution of a system is unbalanced, the laws of thermodynamics governs the direction of all physical changes taking place and with time, the energy within a system tends to become balanced and distributed in the most probable pattern. This "most probable pattern" is actually a state of equal energy among particles, as collisions cause bodies to exchange heat.

Thermodynamic entropy is the measure of this disorder ness in a system. Lower the entropy more is the stability. Consequently, chaotic systems have a higher entropy value and the stable systems have lower entropy value. Motifs are short conserved regions in DNA, RNA, or Protein sequences which correspond to some structural and/or functional feature of the bio-molecules. Hence we may hope that, these “important” stretches of DNA will have entropy different from other stretches. These interesting simple repetitions are quite stable and are expected to have lower entropies. [14]

3.3 Motivation from Kullback-Leibler divergence.

In probability theory the Kullback–Leibler divergence is a non-commutative measure of the difference between two probability distributions [15]. KL divergence or Relative-Entropy is an asymmetric dissimilarity measure between two probability distributions, p and q . It measures the added number of bits required for encoding events sampled from p using q as a reference. Typically p represents the "true" distribution of data, observations, or a precisely calculated theoretical distribution. The measure q typically represents a theory, model, description, or approximation of p.

Let a discrete distribution have probability function p_k, let a second discrete distribution have probability function q_k . Then the Relative-Entropy of p with respect to q, also called the Kullback-Leibler distance or divergence, defined by.

$$d(p||q) = \sum_{k=1}^{m} p_k \ln (p_k/q_k)$$

One of the properties of the relative entropy is that it is non-negative and is 0 if both distributions are equivalent ($p=q$).The smaller the relative entropy, the more similar is the distribution of the two variables [16]. This motivates that the measure based on KL divergence can be used to model the important aspects of interesting signals or significant patterns (e.g. motifs) in DNA sequences.

D. Expectation from the proposed model:

Our proposed model is able to measure the quantity of information entropy associated with a sequence and predicts a qualitative interpretation. This interpretation can be used to distinguish between relevant and irrelevant information.

4. Method.

This paper uses and extends the idea of using KL-Divergence by incorporating information- entropy guided KL-Divergence based modeling. Our hypothesis is that short conserved regions of the bio-molecules are “important” stretches and can have entropy different from other stretches.

4.1 Algorithm

Step1. Given a set of sequences and length of an unknown signal, divide each sequence to get substrings of length equal to the given length of the signal.

Step2. Compute the information entropy of all substrings obtained in Step 1.

Step3. Choose those substrings that fall within a selected range of lower information entropies computed in step2.

Step4. Find a consensus sequence of the set of substrings obtained in step3.

Step5.Calculate the Kullback Leibler divergence for each substrings of step3 with respect to the consensus string obtained in step4.

Step6. Keep those substrings that fall within a selected range of lower Kullback-Leibler divergence computed in step5.

Step7. The consensus of the set of strings described in Step6 is the required signal.

5. Implementation

As a proof of concept, we tested our method on various synthetic datasets. As the results on the synthetic data set are promising, we applied our method to transgenic data regions of E.Coli ECRDB70 database.

The model was implemented using C++ in Windows environment. The analysis was carried out using MATLAB 7.1 and Bioinformatics Toolbox 2.1.1. Sample inputs from both synthetic data with implanted motif and using large datasets generated from Escherichia coli RegulonDB for testing the efficiency of the proposed model [17][18]. The synthetic data comprised of randomly generated DNA sequences of 81 characters length each with an implanted motif of length 15. The program was executed for finding the motifs from the random samples using the PWM model, information entropy model and the proposed entropy with Kullback-Leibler divergence model.

6. Results and Discussion

The three models were evaluated using the synthetic data with implanted motifs. For the first model, the PWM was computed from the 10 sample sequences. The consensus sequence from the PWM indicated "AAAAAAAAAAAAAAA" as the motif whereas the implanted motif was "AAAAAAAGGGGGGG".

In the second model, the information entropy was computed for each substring equal to the length (length=15) of the implanted motif as indicated in step1 of the algorithm. From the plot of Information entropy and Signal position in Fig.1, it is observed that the lower entropy range contains the implanted motif. There were 42 probable motifs out of 806 signals within the lower entropy range of 2.97 to 3.24 bits. The consensus of these 42 candidate motifs resulted in correctly predicting the implanted motif "AAAAAAAGGGGGGG".

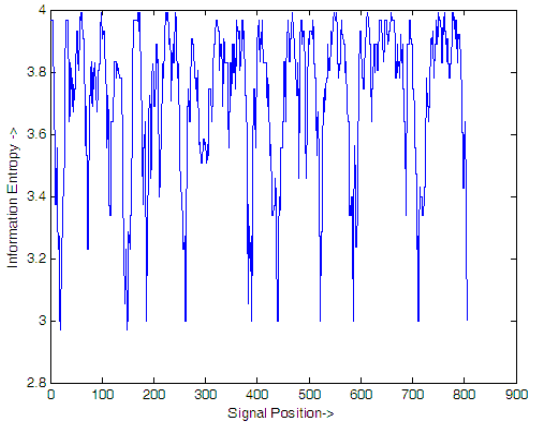

Figure-1

Another set of synthetic data with randomly implanted motif "AGAGAGAGAGAGAGA" was tested against the same entropy model which resulted in 63 candidate motifs within the range of 2.97bits to 3.29bits out of the 806 signals. The consensus of these candidate motifs resulted in correctly predicting the implanted motif "AGAGAGAGAGAGAGA". The plot of entropy and sequence position is indicated in Fig-2.

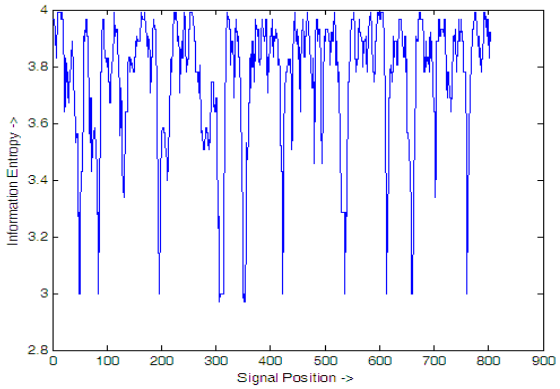

Figure-2

However, the entropy model failed to predict the motif when a randomly generated motif "TAGCTTCATCGTTGA" was implanted in the sample sequences. As depicted in Fig-3, the lower entropy ranges 3.23bits to 3.7bits contained 91 out of 806 sequences and the consensus generated was "AAAAAAGAAAAAGAA".

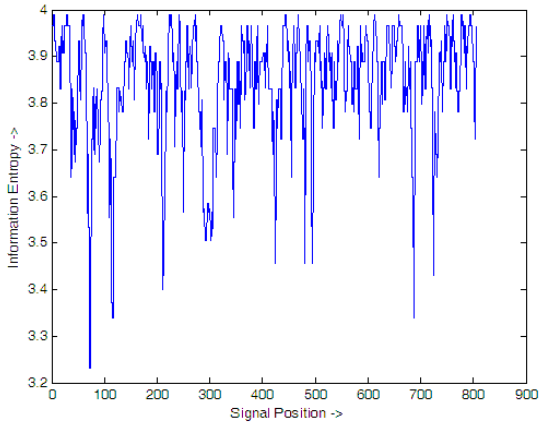

Figure-3

In the third model, the Kullback-Leibler divergence was computed for finding the implanted motif. The implanted motif "AGAGAGAGAGAGAGA" could be found as a consensus of the 34 candidate motifs out of 806 signals within the divergence range of 0.16 to 0.30 as indicated in Fig-4.

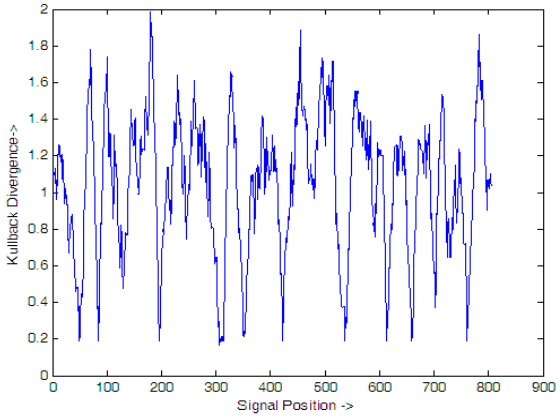

Figure-4

The Kullback-Leibler divergence could predict the random motif "TAGCTTCATCGTTGA" as "AATTTTCTTCATTTA" with matching at nine places as compared to only three matching in the entropy model. 67 candidate motifs out of 806 signals were used for generating the consensus within the divergence range of 0.44 to 0.46 as indicated in Fig-5. Selecting another divergence range of 0.45 to 0.46 could predict the motif as "TAGTTTCATCATTGA" with matching at thirteen places.

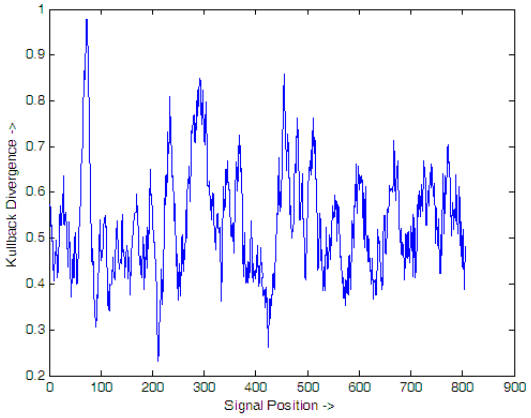

Figure-5

As the results on the various synthetic data sets are encouraging, the proposed model was tested on a real data set. The proposed model was tested using the transgenic data regions of the E.coli ECRDB70 database to predict the motifs. The motif is of length 9 and is a repressor signal present in the transgenic regions occurring at 12 locations between the base pair positions 2342723 and 3598553 [19]. The model could predict 9 out of the 12 DnaA motifs using 533 candidate motifs out of 2852 signals within the divergence range of 0.49 to 0.62 for generating the consensus (Fig-6). None of the motifs were observed in the higher ranges of divergence.

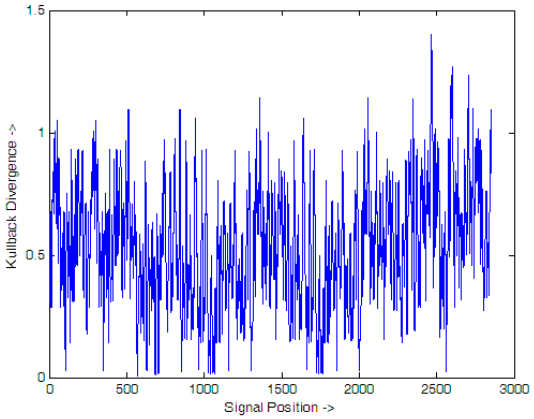

Figure-6

7. Performance Evaluation

The performance of our model on the real data set was computed using the following standard parameters.

(a) Sensitivity: It is the percent of true motifs correctly predicted as true motifs. It measures how good our model in hitting the real motifs is.

$$Sensitivity(S_n) = \frac{TP}{TP + FN} * 100$$

(b) Specificity: It is the percent of non-motifs correctly predicted as non-motifs. It measures how good our model in hitting the non motifs.

$$Specificity(S_p) = \frac{TN}{TN + FP} * 100$$

(c) Accuracy: It is percentage of correctly predicted motifs. It measures how much accurate is the model in correctly predicting the motifs.

$$Accuracy(A_{cc}) = \frac{TP + TN}{TP + TN + FP + FN} * 100$$

The KL-Divergence model generated the consensus with an improved accuracy of 74% as compared to 57% in the model based only on Information entropy. The sensitivity and specificity of the KL-Divergence model were 75% and 72% respectively as compared to 33% and 59% in the model based on Information entropy.

8. Summary and conclusion

Comparison of models for Prediction of motif

Dataset	Model based on PWM	Model based on Info. Entropy	Model based on Proposed approach
Synthetic data with implanted motif	**No**	**Yes**	**Yes**
Synthetic data with new implanted motif	**No**	**Yes**	**Yes**
Synthetic data with random motif	**No**	**~33% match**	**~77% match**
Real data with known motif	**No**	**~52% match**	**~74% match**

Yes: Model could predict the motif
No: Model could not predict the motif

When PWM is applied to all datasets, it failed to predict the motif. Information entropy and KL divergence could predict the motif accurately from the first two datasets as indicated in the above table. Better identification was observed from the third and fourth datasets because:
The PWM based model ignores the weak signals since the model completely discards the bases with lower probability of occurrence. The model based on information entropy overcomes the deficiencies associated with the PWM model but fails to distinguish between signal and noise when the information contents are same.

The proposed model based on information entropy along with Kullback-Leibler divergence successfully addresses the problem of distinguishing between the relevant and irrelevant information. We observe from the above test runs of the proposed algorithm that besides predicting the motifs which are correctly predicted by other methods it also predicts more effectively the motifs which are otherwise not correctly predicted by other models.

We have presented a conceptually simple, and a very general approach based on information entropy along with Kullback-Leibler divergence for modeling biological signal. The efficacy of the proposed method was discussed on motif finding problem. One of the major advantages of this method is that it does not ignore any sequence whether weak or strong and takes into account all input sequences while finding the interesting regions. We plan to extend this work by increasing the real case examples and analyzing the relationship between the entropy of the target motif and the effectiveness of our approach. Also, other divergence measures can be explored and applied to this problem.

References

[1] C.E.Shannon, A Mathematical Theory of Communication, *Bell System Tech.j*, 1948, 379-423.
[2] T.Dewey and H.Herzel, Application of Information Theory to Biology, Pacific Symposium on Biocomputing, 2000, 597-598.
[3] E.Eskin and P.Pevzner, Finding composite regulatory patterns in DNA sequences, Bioinformatics, 2002, 354-367.
[4] C. Lawrence and A. Reilly, An Expectation Maximization (EM) Algorithm for the identification and characterization of common sites in unaligned sequences, Proteins: Structure, Function, Genetics, 1990, 41-51.
[5] M.Sagot, Spelling approximate or common Motifs using a Suffix Tree, Springer-verlag LNCS, 1998, 111-127.
[6] S.Sinha and M.Tompa, A statistical method for finding Transcription factor Binding sites, Proceedings. Eighth International Conference on Intelligent Systems for Molecular Biology, 2000, 344-354.
[7] Lawrence et.al, Detecting Subtle sequence signals: A GIBBS sampling strategy for multiple alignments, Science, 1993, 208-214.
[8] P.Pevzner and S.Sze, Combinatorial approaches to finding subtle signals in DNA sequences, Intelligent systems for Molecular biology, 2000, 269-278
[9] T.D.Schneider, Information content of Individual Genetic Sequences, Journal of Theoretical Biology, 1997, 427-441.
[10] Neil C Jones, Pavel A. Pevzner, An Introduction to Bioinformatics Algorithms, MIT Press, 2004
[11] G.Hertzell and G.Stormo, Identifying Protein-binding sites from unaligned DNA fragments. Proc.Nat.Acad.Science, 1989, 1183-1187.
[12] P.Clote and R.Dackofen, Computational Molecular Biology, (John weily publication, 2000)
[13] Phillip Sheeler, Donald E. Bianchi, Cell and Molecular Biology, Third edition, John Wiley & Sons
[14] P.K.Naq, Engineering Thermodynamics – Mechanical Engineering Series, The McGraw-Hill.
[15] S. Kullback (1987) The Kullback-Leibler distance, The American Statistician 41:340-341.
[16] T.Koski, Hidden Markov Models for Bioinformatics (Kluwer Academic Publishers, 2001)

[17] http://www.bioalgorithms.info
[18] Jianjun Hu, Bin Li and Daisuke Kihara, Nucleic Acids Res. 2005; 33(15): 4899–4913.
[19] Jianjun Hu, Yifeng David Yang, and Daisuke Kihara., EMD: an ensemble algorithm for discovering regulatory motifs in DNA sequences, BMC Bioinformatics 2006.

Anjali Mohapatra has contributed as faculty during her assignments in leading institutions of the State. She has published and presented many research papers in the national and international forum. She is continuing her research in novel application areas of computer science such as Computational Molecular Biology, Soft Computing and Algorithms. She is M.Tech in Computer Science from Utkal University and is currently working as a Faculty in the International Institute of Information Technology (IIIT), Bhubaneswar.

P.M.Mishra is an Electrical Engineer with M.Tech in Computer Science. He has vast experience in the field of Information technology and has served the State Government and different public sector companies. He has active interest in Computational biology, ERP, ISMS, IT infrastructure, Power system studies, Energy conservation and Renewable energy. He is presently working as an Executive Engineer (Electrical) in the O/O EIC (Elect) – cum-PCEI, Orissa, Department of Energy, Govt. of Orissa.

Sudarsan Padhy is a Professor of Mathematics at Utkal University, India. He obtained his Ph.D. degree in Mathematics from Utkal University in 1979 and Postdoctoral research at University of Freiburg, Germany during 1980-81. He has over fifty published research papers and five books to his credit extending over Fluid dynamics, Finite difference and Finite element method for solving partial differential equations, Operation research, Parallel algorithms, Computational finance and Computational biology.