INFORMATION THEORY BASED ANALYSIS FOR UNDERSTANDING THE REGULATION OF HLA GENE EXPRESSION IN HUMAN LEUKEMIA

Bishwajit Das and Durjoy Majumder [*]

Department of Physiology, West Bengal State University, Berunanpukuria, Malikapur,
Barasat, North 24 Parganas, Kolkata-700126
[*] durjoy@rocketmail.com

ABSTRACT

Considering information entropy (IE), HLA surface expression (SE) regulation phenomenon is considered as information propagation channel with an amount of distortion. HLA gene SE is considered as sink regulated by the inducible transcription factors (TFs) (source). Previous work with a certain number of bin size, IEs for source and receiver is computed and computation of mutual information characterizes the dependencies of HLA gene SE on some certain TFs in different cells types of hematopoietic system under the condition of leukemia. Though in recent time information theory is utilized for different biological knowledge generation and different rules are available in those specific domains of biomedical areas; however, no such attempt is made regarding gene expression regulation, hence no such rule is available. In this work, IE calculation with varying bin size considering the number of bins is approximately half of the sample size of an attribute also confirms the previous inferences.

KEYWORDS

Information entropy, Mutual Information, HLA Surface expression, bin size.

1. INTRODUCTION

Understanding of HLA gene regulation is important particularly in malignancy and other state of disease cases. It has been reported that in cancer cells classical HLA class I (HLA class Ia) gene expression is frequently down-regulated that may enable them to escape from immune attack. It is also noted that HLA down regulation is also evident in leukemic cell both at the transcriptional and at the translational level [1]. In this connection it would be interesting to note that in cancer no mutation has been identified in HLA gene so far [2].

Regarding the mechanism of HLA down-regulation in cancer, aberrant expression or binding of transcription factor (**TF**) to Enhancer (**Enh**) A, a conserved sequence present in the HLA promoter region is already reported. However, in the promoter region of HLA has another region called Enh B region, regarded as inducible region. It has been suggested that this region has a cell/tissue specific function and plays a significant role in pathogenic transformation [3-4]. Hence, this would be interesting to find out the possible role/importance of this region in malignant cells.

Conventionally, experimental biologists test the mechanism of gene regulation through an *in vitro* experimental model system set up. With such model system, either the gene of interest (GI) i.e., TF is over-expressed within a cell line deficient to that gene or silencing the GI followed by estimation of the effect on the downstream target gene (**TG**). Such experimentation with Enh B reveals that several Enh B region binding TF like RFXB, CIITA or CREB1 alone can induce the

DOI : 10.5121/ijist.2012.2504

HLA class I expression. Therefore it could be expected that these TF could be down-regulated in primary human leukemic cells. Investigation with primary human leukemic cells reveals that there is no alteration (statistically insignificant through nonparametric statistics) of majority of the TFs except RFXB and CREB1. Contrarily, these two TFs are over expressed in primary human leukemic cells having HLA class Ia positivity [5, 6].

In recent times, information theoretic tools (like entropy analysis, mutual information) have been used in the development of sets of over- and/or under-expressed genes (through clustering) from microarray profile of different gene expression [7, 8]. Very recently, mutual information has been utilized to identify post-translator modulators of different TFs in human B-cells [9]. Along with the approach, information entropies computed for the source (TF) to receiver [TG i.e., surface expression (SE) of HLA] and computation of channel equivocation and mutual information are used to characterize the phenomenon of HLA gene regulation in different de-novo AML, CML, ALL, CLL and MDS patients.

With this approach, the TF and SE data divided into certain number of intervals and confirms the previous relationships between RFXB and CREB1 expression with HLA class Ia positivity. Moreover, mutual information analysis reveals the different cells (leukemias) types are differentially dependent on these TFs in regulating HLA expression [10]. The dependencies of HLA expression are as follows –
for RFXB: NV > AML > ALL and
for CREB1: AML > ALL > NV.

In recent time it is hypothesized that entropy function is dependent on the probability distribution of an attribute within the bins and hence, differences in the number of bins may produce in different inference [11, 12]. Here attempt has been made to validate/confirm the previous findings with varying bin size.

2. Information Theory Background

The classical concept involves a source of information that emanates certain symbols according to a probability distribution. These symbols pass through a channel and are received at the other end. The received symbol probabilities are different from the source to an extent depending upon the distortion properties of the channel.

This concept can be extended to cover data points of a TF attribute which acts as the source and an attribute surface expression (SE) as the receiver with the phenomenon of gene regulation as the underlying channel. Information entropy (IE) function provides us with this important metric. IEs computed for the source, receiver and computation of channel equivocation and mutual information could be useful to characterize the phenomenon of gene regulation. Below we provide detailed theoretical background on the concepts used in this work.

Entropy (H) of Single attribute of TF: If an attribute value close to max., data are scattered & more is its information entropy and uncertainty. Given the n data points pertaining to a variable, the range is sub-divided into q intervals and if f_i is the number of data points occurring in the i^{th} interval, then $p_i = f_i / n$ defines a probability distribution for the variable over the chosen q intervals. Entropy $(H) = \sum_{i=1}^{q} p_i \times \log\left(\frac{1}{p_i}\right)$ gives a measure of surprise associated with this probability distribution of the variable. In general for r-based logarithm,

$$I(E) = -\log_r\left(\frac{1}{p_E}\right) \text{ r-ary units.}$$

In natural logarithms (base e) the units are nats. In our calculation we have calculated all values to 10 based logarithm. That means here r =10.

Some of the properties of entropy function is listed here that would be useful for the present work [13, 14].

i) H is symmetric and continuous. This ensures that any choice of sub-interval changes can bring out the required uncertainty measure.
ii) H_{n+1} (p_1, p_2, p_3,…, p_{n-1}, 0) = H_n (p_1, p_2, p_3,…, p_n) i.e., if an interval is empty, it does not affect entropy. This means extending the range to some global (max, min) does not affect the sample data point based calculation. Due to this property, all the different groups (normal, disease) can be governed by the same sub-interval choice without affecting the desired metric.
iii) H_n (p_1, p_2, p_3,…, p_n) ≤ H_n (1/n, 1/n, 1/n,…, 1/n). This means that if the data is uniformly distributed the entropy will be maximum while the same falls down when the data is clustered more in a certain interval. This allows an upper bound on the chosen metric and thereby facilitates comparison.

Joint Entropy (TF attribute vs. surface expression) is the amount of average information provided by the two attribute jointly. The joint entropy approaches the summation of the individual entropies when the two taken attributes are independent. This allows any arbitrary sub-ranging of the two-dimensional array involving TF and SE pair while finding the metric.

$$H = \sum_X \sum_Y p(X,Y) \times \log\left(\frac{1}{p(X,Y)}\right) \text{ or, } H(X,Y) = \sum_X \sum_Y p(X,Y) \times \log\left(\frac{1}{p(X,Y)}\right), \text{ where}$$

X and *Y* are two random variables.

Conditional Entropy [H(Y|X) or H(X|Y)] measures the uncertainty remaining about random variable X after specifying that random variable Y has taken on a particular value. The relationship between joint entropy and conditional entropy is exhibited by the fact that the entropy of a pair of random variables is the entropy of one plus the conditional entropy of the other.

Mutual Information [I(X;Y)] is the relative importance of an attribute in SE of protein. The mutual information between two random variables measures the amount of information that one conveys about the other. Equivalently, it measures the average reduction in uncertainty about X those results from learning about Y. Mutual information is always ≥ 0. In the event that the two random variables are perfectly correlated, then their mutual information is the entropy of either one alone. The mutual information of a random variable with itself is just its entropy. For this reason, the entropy H(X) of a random variable X is sometimes referred to as its self-information. The mutual information can be calculated as

$$I(X;Y) = H(X) - H(X|Y) = H(Y) - H(Y|X).$$

3. Materials and Method

3.1. Collection of Data

All gene expression data has been collected from the work of Majumder, 2006 and Majumder, 2012 [5, 6]. Primarily we have data of two attributes – HLA surface expression (HLA-ABC and HLA-DR) and transcriptional data (IRF-1, RFX5, RFXB, CIITA, CREB-1) of 10 normal volunteers (NV) and different leukemic patients [18 AML (acute myelogenous leukemia), 14 ALL (acute lymphocytic leukemia), 12 CML (chronic myelogenous leukemia) and 6 CLL (chronic lymphocytic leukemia)]. The demographic description of the patients is same as mentioned in the earlier works [1]. The TFs gene expression data and SE data were acquired through semi-quantitative reverse transcription polymerase chain reaction (RT-PCR) and by flow cytometric method respectively. The characteristics of the collected data are shown in Table 1 and 2 [1].

Table 1. Cell surface HLA-ABC and HLA-DR expression. Data are presented as mean ± SD; Mdn, Max and Min stands for median, maximum and minimum value obtained in the population.

Sample	HLA-ABC	HLA-DR
NV	57.23±21.97 Mdn 48.6 Max 107.37 Min 38.53	36.793±13.78 Mdn 32.48 Max 58.33 Min 21.03
AML	29.035 ± 17.325 Mdn 28.02 Max 59.81 Min 1.12	51.508±46.29 Mdn 42.96 Max 165.01 Min 1.19
ALL and CLL	32.721 ± 23.44 Mdn 25.19 Max 81.42 Min 9.2	195.909±192.43 Mdn 106.82 Max 626.36 Min 32.91

*Note: In CML cases, identification of malignant cell diagnosis is not possible through flow cytometry, hence investigation on HLA surface expression is not done. Statistical test of significance is available in Ref. 1.

Table 2. Transcriptional expression of different TFs in leukemic and normal individuals. Data are presented as mean ± SD.

	IRF1	RFX5	RFXB	CIITA	CREB1
NV	1.049 ±0.632	1.102 ±0.376	0.711 ±0.392	0.412 ±0.353	0.0
AML	1.37 ±1.451 NS	0.984 ±0.597 NS	1.83 ±0.588 P<0.005	0.801 ±0.742 NS	0.801 ±0.76 P<0.001
ALL	0.831 ±0.978 NS	1.181 ±0.5 NS	1.84 ±0.905 P≤0.02	0.735 ±0.486 P≤0.02	0.533 ±0.286 P<0.001
CML	0.842 ±1.419 P<0.02	1.15 ±0.66 NS	1.528 ±1.1 P<0.05	0.717 ±0.668 NS	0.228 ±0.218 P<0.001
CLL	1.83 ±2.16 NS	1.234 ±0.411 NS	2.253 ±1.403 P≤0.05	0.795 ±0.491 NS	0.155 ±0.158 P<0.001

NS: Not statistically significant; P means the level of statistical significance through Mann-Whitney U test.

3.2. Proposed Scheme

Five attributes (IRF1, RFX5, RFXB, CIITA AND CREB1,) pertaining to TF and two attributes of HLA surface expression (HLA-ABC, HLA-DR) (SE) have been considered (Fig. 1A). For each pair (altogether 5×2 = 10 pairs) of TF and SE, we consider the existence of an informational channel through which the concerned TF manifest into the corresponding SE. Our aim is to examine these channels. For the chosen attributes we have collected data of individuals from normal population and some individuals with different leukemic conditions. We separately examine the channels in different disease groups and compare them with normal group. This gives an insight into the phenomenon through which a TF regulates the SE. The results are not in absolute terms but based on the comparative analysis.

The TF and SE data collected is divided into a certain number of intervals. Here we choose two numbers of intervals; the intervals are 7 and 10. The number of intervals is analogous to the number of symbols in classical information theory. Calculation of the frequency distribution from data is analogous to the symbol probabilities at the source and the receiver side respectively.

Now we have considered the joint probabilities of the symbols of the source (TF) and receiver (SE). In these ways we convert the TF-SE pair into a source-receiver pair. Now we calculate the information entropies at the source H(X) i.e., TF [interval 7 and 10] and receiver H(Y) i.e., SE [interval 7 and 10]. We also calculate the joint entropy of source and receiver pair H(X,Y) by considering the joint probabilities. These provide the measure of uncertainty and from these. In a sense the mutual information I(X;Y) is the intersection between H(X) and H(Y), since it represents their statistical dependence. In the given Venn diagram we derive the channel equivocation H(X|Y) or H(Y|X) i.e., the average conditional entropy of the source given the receiver symbol or vice versa. H(Y|X) = H(X, Y) - H(X) and H(X|Y) = H(X, Y) – H(Y), Mutual information I(X;Y) can now be calculated as: I(X;Y) = H(X) – H(X|Y) = H(Y) - H(Y|X)

In this way we calculate our proposed scheme and compared our results in two intervals and also checked the results that different intervals results are same or not.

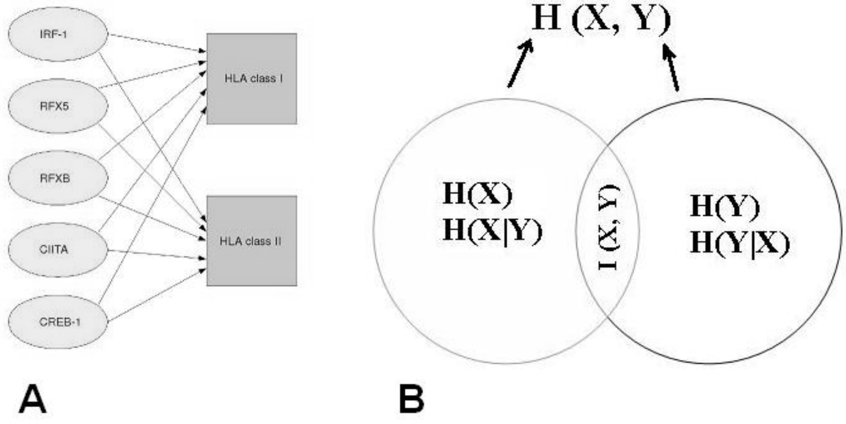

Figure.1. Analogy between transcriptional regulation and information channel (A) and Venn diagram showing relation between different entropies (B).

The channel equivocation is an important metric that provides the information about the nature of the channel, i.e., how the channel contributes to the uncertainty propagation from source to receiver. In analogy, the metric chosen by us could provide the uncertainty with which the TFs' express themselves into the SE. In other words, it gives an idea about relative importance of the contribution of the channel in the propagation of uncertainty of TF into the uncertainty of SE. Venn diagram (Fig. 1B) shows how the different entropies are related to one another. Results indicate how differently the channel behaves from normal to disease cases. Also, the relative importance of propagation of a TF to SE would be manifested.

3.3. Grading of Independence

We sum the individual entropies of attributes already computed and compare them with their joint entropies. If sum comes close to joint entropy values then we can say that the two considered attributes are independent.

Table 3a. Entropy of different attributes with 7 intervals.

	HLA-ABC	HLA-DR	IRF1	RFX5	RFXB	CIITA	CREB1
NV	0.5159	0.4727	0.6533	0.4728	0.477	0.4579	ND
AML	0.5914	0.6866	0.72	0.7343	0.5397	0.5853	0.5491
ALL	0.6372	0.5233	0.6098	0.6239	0.6394	0.7262	0.4515
CML	ND	ND	0.3867	0.7774	0.6221	0.4265	0.2597
CLL	0.477	0.477	0.5478	0.5393	0.4191	0.4391	ND
MDS	ND	ND	0.2065	0.6016	0.4515	0.4515	0.2172
Total	0.7453	0.5595	0.692	0.7279	0.7653	0.6538	0.3963
Maximum	0.8450	0.8450	0.8450	0.8450	0.8450	0.8450	0.8450

Table 3b. Entropy of different attributes with 10 intervals.

	HLA-ABC	HLA-DR	IRF1	RFX5	RFXB	CIITA	CREB1
NV	0.5988	0.5556	0.7362	0.447	0.477	0.4556	ND
AML	0.7116	0.7839	0.7803	0.8012	0.5397	0.7153	0.8087
ALL	0.7465	0.6873	0.7626	0.6078	0.6392	0.7853	0.7522
CML	ND	ND	0.5495	0.6582	0.6221	0.4265	0.64
CLL	0.477	0.477	0.5478	0.5393	0.4191	0.4391	0.4306
MDS	ND	ND	0.6394	0.7145	0.4515	0.4515	0.5782
Total	0.7523	0.6583	0.7986	0.8582	0.8401	1.963	0.5176
Maximum	1	1	1	1	1	1	1

ND: not done due to inadequate data.

So after finding the joint entropy we have given the grading to them that denotes the degree of independence between those two attributes. Say for a joint distribution of X, Y we obtain $H_{X,Y}$ = P units of entropy and Q be the individual sum of entropies, then, $\frac{Q-P}{Q}$ × 100 is taken as a measure to find the grade which is expressed as a range of percentage. Thus grading system indicates that the lesser the percentage or grade, the two attributes are more independent to each other; whereas higher is the grade, higher is the dependency.

Similarly the calculated mutual information with respect to the entropy value of each of the variable i.e., H(X) or H(Y) is represented to percentage [$\frac{I(X;Y)}{H(X)}\times 100$ and $\frac{I(X;Y)}{H(Y)}\times 100$]. This measure denotes the relative dependencies on that variable in quantitatively.

Table 4a. Entropy of each attributes of different leukemia combinations with 7 intervals.

Type of Combination	Malignancy	HLA-ABC	HLA-DR	IRF1	RFX5	RFXB	CIITA	CREB1
Myeloid Combination	AML+CML	0.5914	0.6868	0.8339	0.8334	0.7956	0.7541	0.5395
Lymphoid Combination	ALL + CLL	0.6891	0.8412	0.6288	0.773	0.7877	0.7597	0.4398
Acute Combination	AML+ ALL	0.67	0.6259	0.7524	0.875	0.7687	0.8309	0.5593
Chronic Combination	CML+ CLL	0.477	0.477	0.6113	0.794	0.791	0.6093	0.2167
	Total	0.7453	0.5595	0.692	0.7279	0.7653	0.6538	0.3963
	Maximums	0.8450	0.8450	0.8450	0.8450	0.8450	0.8450	0.8450
Normal Volunteers	NV	0.5159	0.4727	0.6533	0.4728	0.477	0.4579	0

Table 4b. Entropy of each attributes of different leukemia combinations with 10 intervals.

Type of Combination	Malignancy	HLA-ABC	HLA-DR	IRF1	RFX5	RFXB	CIITA	CREB1
Myeloid combination	AML+CML	0.7116	0.7839	0.9341	0.9306	0.9332	0.8218	0.5402
Lymphoid combination	ALL + CLL	0.9231	0.8811	0.7979	0.814	0.7877	0.8253	0.4695
Acute combination	AML+ ALL	0.8474	0.776	0.9189	0.9465	0.9055	0.9073	0.5606
Chronic Combination	CML+ CLL	0.477	0.477	0.847	0.8334	0.791	0.7285	0.2340
	Total	0.7523	0.6583	0.7986	0.8582	0.8401	1.963	0.5176
	Maximums	1	1	1	1	1	1	1
Normal Volunteers	NV	0.5988	0.5556	0.7362	0.447	0.477	0.4556	0

4. RESULTS

4.1. Analysis of Single Attribute

Table 3a and 3b indicate the extent of deterministic behavior of different attributes of different populations in different intervals. If any attribute value is closer to the maximum value, the more is its information entropy or uncertainty. As discussed in entropy (H) function properties, uncertainty will maximize if all the intervals are equally likely i.e., data points are scattered equally over the entire range. Here deterministic behavior implies that most of the attribute values fall within a few sub-ranges while randomized behavior means the attribute values are scattered over the entire range. It is observed that with increasing in bin size the entropy values are also increased in both NV and diseased samples, as suggested by Paninski, 2004 [11]. This ensures the data points are scattered. However, the increment in bin size follow the same trend in both NV and disease sample (Table 3a and 3b). Comparing the ratio of entropy value of an attribute in between the disease to NV does not differ much more and the value does not exceed 1.43. For example, RFX5 has the difference between two bins is 0.084 while considering AML versus NV.

4.2. Analysis of Combinations of Different Leukemias

Next we have the combination of different disease conditions (Table 4a and 4b) and performed entropy analysis to find out the behavior of individual parameters in different states and types of leukemia in different intervals. Here we observed the same trend as we have observed for single attribute analysis.

4.3. Joint Entropy Analysis

Joint entropy analysis provides the amount of average information of two attribute jointly and dependency between two attributes and also a comparison between disease and normal reflects the alteration in transcriptional efficiency. The joint entropy of different combinations and its comparison with the summation of their individual entropies have been tabulated and a grade has been provided as per the grading rule mentioned in the Methods section. Example cases are tabulated in Table 5. The table shows that in normal samples, CIITA is more potent in induction of HLA-DR (more dependency) compared to HLA-ABC. Generally, in disease cases, HLA is independent of CIITA with some minor dependency in case of lymphoid leukemia. The detailed results for all TFs can be derived from mutual information analysis.

Table 5. HLA-ABC vs. CIITA (A) and HLA-DR vs. CIITA (B). DS: total disease samples.

Type	HLA-ABC vs. CIITA (A)					
	Joint Entropy		Sum of Entropies		Grading*	
	7 Vs. 7	7 Vs. 10	7 Vs. 7	7 Vs. 10	7 Vs. 7	7 Vs. 10
NV	0.6614	0.7535	0.9738	1.0544	D	D
AML	0.8398	0.9049	1.1767	1.4269	B	C
ALL	0.9866	1.0413	1.3634	1.5318	B	B
Total (DS)	1.3259	1.6281	1.3991	2.7153	B	A
Maximum	1.6901	1.8450				

Type	HLA-DR vs. CIITA (B)					
	Joint Entropy		Sum of Entropies		Grading*	
	7 Vs. 7	7 Vs. 10	7 Vs. 7	7 Vs. 10	7 Vs. 7	7 Vs. 10
NV	0.6387	0.5558	0.9306	1.0112	C	D
AML	0.7817	0.7522	1.2719	1.4992	B	B
ALL	0.8873	0.9542	1.2495	1.4726	B	C
Total (DS)	1.5670	1.3933	1.2133	2.6213	B	A
Maximum	1.6901	1.8450				

*A: 0-15%, B: 15-30%, C: 30-45%, D: above 45%.

4.4. Mutual Information Analysis

Mutual information analysis may reveal the relative importance of an attribute (TF) on the regulation of SE (HLA-ABC or HLA-DR) in normal and leukemic state (Table 6 and Table 7). If mutual information decreases, it indicates that the association becomes more independent and the channel (gene regulation) distorts the passage of information from TF to SE. With respect to particular TF, percentage of dependency (as mentioned in section 3.3) is also calculated and compared with NV. If the calculated value is increased in disease sample, dependency to that TF is more for SE of HLA in disease cases.

Table 6. Mutual information analysis with bin size combination of 7 vs. 7. Values in parentheses indicates % dependency (out of mutual information analysis) with respect to entropy value of X i.e., TF* [H(X)] (in **A**) and Y i.e., SE† [H(Y)] (in **B**).

(A) HLA Gene	Type	$I(X;Y) = H(X) - H(X \mid Y)^{*}$				
		CIITA	RFX5	RFXB	IRF1	CREB1
HLA-ABC	NV	0.3124 (68.22)	0.5158 (109.09)	0.2148 (45.03)	0.215 (32.90)	-0.0117 (100)
	AML	0.3369 (57.56)	1.2068 (164.34)	0.3998 (74.49)	0.2156 (29.94)	0.1863 (33.92)
	ALL	0.3768 (51.88)	0.3292 (52.76)	0.4985 (77.96)	0.3675 (60.26)	0.3106 (68.79)
HLA-DR	NV	0.2919 (63.74)	0.2693 (56.95)	0.2719 (57.00)	0.0339 (5.18)	0.2075 (100)
	AML	0.4902 (83.75)	0.5008 (68.20)	0.4029 (75.06)	0.4464 (62.00)	0.5605 (102.07)
	ALL	0.3622 (49.87)	0.352 (56.41)	0.3846 (60.15)	0.3053 (50.06)	0.6588 (145.47)

(B) HLA Gene	Type	$I(X;Y) = H(Y) - H(X \mid Y)^{\dagger}$				
		CIITA	RFX5	RFXB	IRF1	CREB1
HLA-ABC	NV	0.3704 (71.79)	0.5589 (108.33)	0.2537 (49.17)	0.0776 (15.04)	0.5042 (97.73)
	AML	0.343 (57.99)	1.0639 (179.89)	0.4545 (76.85)	0.087 (14.71)	0.2286 (38.65)
	ALL	0.2878 (45.16)	0.3425 (53.75)	0.4963 (77.88)	0.3949 (61.97)	0.4963 (77.88)
HLA-DR	NV	0.3067 (64.88)	0.2692 (56.94)	0.2676 (56.61)	-0.1467 (-31.03)	0.6802 (143.89)
	AML	0.5915 (86.14)	0.4531 (65.99)	0.5528 (80.51)	0.413 (60.15)	0.698 (101.66)
	ALL	0.1593 (30.44)	0.2514 (48.04)	0.2685 (51.30)	0.2188 (41.81)	0.7286 (139.23)

From Table 6 and Table 7, we observed that mutual information for RFXB is high in general in both intervals. This means that both HLA class I and II (HLA-DR) are dependent on this TF with an indication that dependency of HLA class I is more than HLA-DR. This dependency becomes more pronounced in leukemic condition. Though under normal condition CIITA dependency of

HLA-DR is more but under the condition of malignancy this dependency decreases for AML but increased in ALL cases in both intervals. However, HLA-ABC is less dependent on RFX5 and CIITA in normal and myeloid leukemic cases. For lymphoid leukemia dependency is almost unaltered.

Results imply that in induction of HLA-ABC, CREB1 has no role both in normal and leukemia; however, in lymphoid leukemia it plays a role. Similarly, in HLA-DR expression CREB1 has no role in normal and leukemia in general excepting lymphoid leukemia. The overall observation is that different TF plays different role in different type of leukemia (cell) and also observed that if we change the intervals then it not affect the roles of different types of leukemia.

Table 7. Mutual information analysis with bin size combination of 7 vs. 10. Values in parentheses indicates % dependency (out of mutual information analysis) with respect to entropy value of X i.e., TF[*] [H(X)] (in **A**) and Y i.e., SE[†] [H(Y)] (in **B**).

(A) HLA Gene	Type	I (X;Y) = H(X) – H(X│Y)[*]				
		CIITA	RFX5	RFXB	IRF1	CREB1
HLA-ABC	NV	0.3009 (66.04)	0.3175 (71.02)	0.2977 (62.41)	0.3808 (51.72)	-0.0209 (100)
	AML	0.522 (72.97)	0.3368 (42.03)	0.4531 (84.42)	0.3961 (50.76)	0.5661 (70.00)
	ALL	0.4905 (62.46)	0.3677 (60.49)	0.6076 (95.05)	0.6296 (82.55)	0.7206 (95.79)
HLA-DR	NV	0.4554 (99.95)	0.3264 (73.02)	0.2545 (53.35)	0.5781 (78.52)	0.2634 (100)
	AML	0.747 (104.43)	0.4893 (61.07)	0.5002 (93.19)	0.4684 (60.02)	0.7053 (87.21)
	ALL	0.5184 (66.01)	0.4078 (67.09)	0.5484 (85.79)	0.6221 (81.57)	0.7617 (101.26)

(B) HLA Gene	Type	I (X;Y) = H(Y) – H(X│Y)[†]				
		CIITA	RFX5	RFXB	IRF1	CREB1
HLA-ABC	NV	0.4441 (74.16)	0.4693 (78.37)	0.4195 (70.05)	0.2434 (40.64)	0.5779 (96.50)
	AML	0.5183 (72.83)	0.2472 (34.73)	0.628 (88.25)	0.3274 (46.00)	0.469 (65.90)
	ALL	0.4517 (60.50)	0.5064 (67.83)	0.7149 (95.76)	0.6135 (82.18)	0.7149 (95.76)
HLA-DR	NV	0.5554 (99.96)	0.435 (78.29)	0.3331 (59.95)	0.3975 (71.54)	0.819 (147.40)
	AML	0.8156 (104.04)	0.472 (60.21)	0.7474 (95.34)	0.472 (60.21)	0.6805 (86.80)
	ALL	0.4204 (61.16)	0.4873 (70.90)	0.5965 (86.78)	0.5468 (79.55)	0.6968 (101.38)

5. CONCLUSIONS

The present trend for the understanding of gene expression/regulation made between disease and normal by microarray method followed by analysis through different heuristic approaches. However microarray technology provides a range and heuristic based approaches are not truely mechanistic [15]. For understanding of HLA gene regulation there is no microarrray chip is available.So we depend on PCR (polymerase chain reaction) base gene expression data. Previous work suggest that relative importance of propagation of individual TF to SE [10]. In other words this establishes dependency of HLA SE on different individual TF in different cell types of heamatopoietic system under the condition of leukemia.

The parametric variation in population (signal) is a major concern of biological investigation. However the measurement uncertainty (noise) may shadow the signal. So conventionally experimental biologist look for much more instrumental sophistication like simple polymerase chain reaction based method to real time PCR based method. Information theory based analysis in previous work reveal that some of the attributes are dependent on each other. This signifies the measurement noise is less with a certain number of intervals (bin size) for all attributes. Increasing in bin size increases entropy value as suggested by Paninsky, 2004 [11]; however, the incremental difference in entropy of an attribute in disease sample do not differ much compared to the same attribute of NV while differing in bin sizes.

It is to be mentioned here that we haven't made our analysis with a bin size exceeding the sample size [11] and the maximum number of bins selected were approximately half compare to the sample size. Information theory based analysis is being utilized for generation of biological knowledge in several cases; however, there is no rule based method has yet been established for analysis of gene regulation [16]. It is worthwhile mentioned here that the source data [5, 6, 10] was collected with enough procedural validation and a lot of criticisms are also present regarding real time PCR based methodology [17]. So we can different statistical analytical tools (non parametric statistics together with information theory) can reveal gene regulatory mechanism from population data without much more experimental dimensionality and investigation cost. From our previous analysis [10] and this analysis, it can be inferred that previous association analysis (by χ^2) [5, 6] between HLA class Ia expression and CREB1 in leukemia may be due to the effect of malignancy; however, RFXB may play a significant role in HLA regulation. The present work suggests that with varying in bin size, the inference doesn't differ from the previous inferences.

6. REFERENCES

[1] Majumder D, Bandyopadhyay D, Chandra S, Mukhopadhayay A, Mukherjee N, Bandyopadhyay SK, Banerjee S (2005) "Analysis of HLA class Ia transcripts in human leukaemias", *Immunogenet* Vol. 57, pp. 579-589.

[2] Demanet C, Mulder A, Deneys V, Worsham MJ, Maes P, Claas FH, Ferrone S (2004) "Down-regulation of HLA-A and HLA-Bw6, but not HLA-Bw4, allospecificities in leukemic cells: an escape mechanism from CTL and NK attack?" *Blood* Vol. 103, pp. 3122-3130.

[3] Girdlestone J, Transcriptional regulation of MHC class I genes, *Eur J Immunogenet* 23: 395-413, 1996.

[4] Girdlestone J (2001) "Regulation of HLA class I loci by CIITA", *Blood* Vol. 97, pp. 1520.

[5] Majumder D (2006) "Transcriptional regulation of immune recognition in hematological malignancies", Ph. D. Thesis, Jadavpur University, India.

[6] Majumder D (2012) *HLA Expression in Leukemia: Status, Regulation & Therapeutic Implications of HLA Expression in Leukemia*, LAP LAMBERT Academic Publishing, Saarbrucken, Berlin, Leipzig, UK, ISBN: 978-3-8484-3247-9

[7] Margolin AM, Nemenman I, Basso K, Wiggins C, Stolovitzky G, Favera RD, Califano A (2006) "ARACNE: an algorithm for the reconstruction of gene regulatory networks in a mammalian cellular context", *BMC Bioinfor* Vol. 7, Suppl 1, pp. S7 [doi: 10.1186/1471-2105-7-S1-S7].

[8] Priness I, Maimon O, Ben-Gal I (2007) "Evaluation of gene-expression clustering via mutual information distance measure", *BMC Bioinfor* Vol. 8, pp. 111 [doi: 10.1186/1471-2105-8-111]

[9] Wang K, Saito M, Bisikirska BC, Alvarez MJ, Lim WK, Rajbhandari PR, Shen Q, Nemenman I, Basso K, Margolin AA, Klein U, Dalla-Favera R, Califano A (2009) "Genome-wide identification of

post-translational modulators of transcription factor activity in human B cells", *Nature Biotech* Vol. 27, No. 9, pp. 830-837.

[10] Majumder D (2012) "Application of information theory for understanding of HLA gene regulation in leukemia", In: Advances in Computing and Information Technology, Vol. 177, pp. 161-173, Meghanathan N, Nagamalai D, Chaki N (Eds.), Kacprzyk J (Ed- in-Chief), Springer-Verleg, Berlin, Heidelberg, ISSN: 2194-5357.

[11] Paninski L (2004) "Estimating entropy on m bins given fewer than m samples", *IEEE Transactions on Information Theory*, Vol. 50, No. 9, pp. 2200-2203.

[12] Margolin AA, Wang K, Califano A, Nemenman I (2010) "Multivariate dependence and genetic networks inference", *IET Syst Biol*, Vol. 4, No. 6, pp. 428-440.

[13] Abramson N (1963) *Information theory and coding*, McGraw Hill, New York, San Francisco, Toronto, London.

[14] Hamming RW (1980) *Coding and Information theory*, Prentice Hall Inc., EngleWood Cliffs, New Jersey, London, Sydney, Toronto, New Delhi, Tokyo, Singapore.

[15] Weston AD and Hood L (2004) "Systems biology, proteomics, and the future of health care: toward predictive, preventative, and personalized medicine", *J Proteome Res* Vol. 3, pp. 179-196.

[16] Kalet IJ (2009) "Probabilistic Biomedical Knowledge", Part I, Ch 3, *Principle of Biomedical Informatics*, Elsevier, Amsterdam, Boston, London, Paris, Newyork, Singapore, Sydney, Tokyo, pp.185-221.

[17] Bustin SA and Nolan T (2004) "Pitfalls of quantitative real-time reverse-transcription polymerase chain reaction", *J Biomol Techq* Vol. 15, pp. 155-166.

Authors

Bishwajit Das completed his M.Sc (Tech) from West Bengal University of Technology, Kolkata. Currently he is pursuing his Ph.D. from West Bengal State University. His research interest includes Bioinformatics and statistical analysis of biological data.

Durjoy Majumder is working as an Assistant Professor of Physiology, West Bengal State University. He received his Ph.D. from Jadavpur University in 2006. He gained his research expertise by working in different premier institutes of India like Calcutta School of Tropical Medicine, Sanjay Gandhi Post Graduate Institute of Medical Sciences, Lucknow and Saha Institute of Nuclear Physics, Kolkata. Before joining the present institution, he held a faculty position of Bioinformatics at the School of Information Technology, Bengal Engineering & Science University, Shibpur. His cross-disciplinary research experience led him to be interested in the area of Systems Biology & Systems Medicine. In particular his research interest is focussed on the quantitative aspect of spatio-temopral dynamics of the Physiological Systems that leads towards the implementation of Systems Medicine. His research interest is not only with the theoreticial development in biology rather it includes the practical aspects of the clinical scenario. He has supervised several M.Tech. Thesis, acted as a peer reviewer of more than 10 international journals and invited speakers of different international conferences.