Proceedings of the 28th IEEE
EMBS Annual International Conference
New York City, USA, Aug 30-Sept 3, 2006

DNA Binding Sites Characterization by Means of Rényi Entropy Measures on Nucleotide Transitions

Alexandre Perera, Montserrat Vallverdú, Francesc Clarià, José Manuel Soria and Pere Caminal

Abstract— **In this work, parametric information-theory measures for the characterization of binding sites in DNA are extended with the use of transitional probabilities on the sequence. We propose the use of parametric uncertainty measure such as Rényi entropies obtained from the transition probabilities for the study of the binding sites, in addition to nucleotide frequency based Rényi measures. Results are reported in this manuscript comparing transition frequencies (i.e. dinucelotides) and base frequencies for Shannon and parametric Rényi for a number of binding sites found in E. Coli, λ and T7 organisms. We observe that, for the evaluated datasets, the information provided by both approaches is not redundant, as they evolve differently under increasing Rényi orders.**

I. Introduction

Information for the gene regulatory processes lie in the so called non-coding regions of the DNA sequence. However, the determination of the pattern at the binding site for various binding elements is still a research issue. Previous attempts have been developed with a prior characterization of the patterns on the binding sites using information based measures. These measures were computed from an aligned set of sequences with empirical evidence of binding properties. A key information measure was introduced by Schneider et al. at 1986 [1]. In this contribution, authors proposed the use of a normalized Shannon Entropy measure R_S (for Redundancy). Other authors have extended this work with the use of parametric Rényi Entropy measure [2]. Authors demonstrated that low Rényi orders conveyed to sharper profiles on the binding sites for a number of different cases. On the other hand, other studies have proposed the use of generalization entropies for the characterization of complete chromosome [3] on a basis of n-mers frequencies, which could also be applied to the characterization of binding sites. Although in [1] it is suggested that dinucleotides provided similar total information (Shannon) for ribosome binding sites, the case for the rest of recognizers under a generalized entropy view still remains unclear. In this paper, we provide results for nucleotide transition based Rényi characterizing binding site profiles on E. Coli, T7 and λ organism. We observe that the total uncertainty, as computed using Rényi Redundancy (R_α), differs from the nucleotide frequency based measures. We also show that this difference depends heavily on the Rényi order (α) for most binding site cases. Results are provided using nucleotide and dinucleotide (base transition) measures of both, Shannon and Rényi Redundancy measures.

This work was supported by the the Spanish Ministerio de Ciencia y Tecnologia, Juan de la Cierva Program and by grants Ministerio de Educación y Ciencia and TEC2004-02274.

A. Perera, M. Vallverdú and P. Caminal are with Dep. ESAII, Centre for Biomedical Engineering Research, Technical University of Catalonia (UPC), Barcelona, Gargallo, 5, 08028 Barcelona, Spain (e-mail: (Corresp. Author) Alexandre.Perera@upc.edu, Montserral.Vallverdu@upc.edu and Pere.Caminal@upc.edu).

F. Clarià is Dep. d'Informàtica I Ingenyieria Industrial, Universitat de Lleida, Lleida, Spain (e-mail: Claria@eup.udll.es).

J. Manuel Soria is with the Unitat d Hemostàsia i Trombosi, Dep. d' Hematologia, Hospital de la Santa Creu i Sant Pau, Barcelona, Spain (e-mail: jsoria@santpau.es).

II. Materials and Methods

A. Database description

We present results from a set of nucleotide sequences that are recognized by a number of macromolecules (e.g. polymerases, repressors, ribosome and others). The dataset consists of a number of sequences previously aligned for E. Coli, Lambda and T7 organisms. Data is publicly available from website of Dr. Schneider's laboratory (http://www.ccrnp.ncifcrf.gov/~toms/) and is the base of the first contribution using information measures for binding site characterization by *Schneider et al.*, in [1]. A summary of the data with organism, recognizer, number of bases in the alignment and number of aligned sequences is shown in TABLE I. All calculus in this paper has been executed on data from ArgR, LexA, Ribosome, TrpR, cI/Cro, HincII and T7 recognizers.

B. Information content measures

1) Rényi entropy definition

Binding site characterizations are shown with Rényi entropies[4], which can be considered as a generalization of Shannon entropy[5]. Given a discrete random variable x with N states $(x_1, x_2, ..., x_N)$ with probability for state i given by p_i, Rényi entropy is defined as,

$$H_\alpha = \frac{1}{1-\alpha}\log\left(\sum_{i=1}^{N} p_i^\alpha\right) \qquad (1)$$

where α is a real positive defined as the order of the Rényi entropy $(\alpha \neq 1)$. For equally probable values, $p_i=1/N$, Rényi takes its maximum value at $H_\alpha=log(N)$ (bits). For minimum disorder $p_i=1$, $p_k=0$ $(k \neq i)$, Rényi takes the minimum value at $H_\alpha=0$. Rényi entropy converges to Shannon entropy when $\alpha \to 1$, given by,

$$H_S = -\sum_{i=1}^{N} p_i \log p_i \qquad (2)$$

where c is taken as 1 when the logarithm base is 2.

2) Redundancy

Redundancy can be computed from previous information measures to produce a normalized index, valued between 0 and 1,

$$R = \frac{E_{\max} - E}{E_{\max}} = 1 - \frac{E}{E_{\max}} \qquad (3)$$

where E represents either Shannon or Rényi entropy measures in (1) and (2) (to result into Shannon or Rényi Redundancy, respectively). Redundancy in this case will increase as the order in the signal increases, and decrease when de disorder of the signal increases.

3) Redundancy on base frequencies: nucleotide approach

For a set of aligned binding sites sequences, Rényi and Shannon Redundancies can be computed in order to characterize sites with large degree of complexity. For this particular case, N will take 4 since we may find only four bases corresponding to the nucleotide types, A, T, G and C. The frequency for each base is taken as the estimation on the probability of appearance of that particular base on a certain site. Shannon and Rényi Redundancies (R_S and R_α, respectively) can then be computed as the schema in Fig. 1 *(a)*. This measure provides information on the complexity of the distribution of base frequencies on the conserved sequence.

4) Redundancy on the base transition frequencies: dinucleotide approach

Additional information on the transition of the bases can be obtained from the binding site sequences. This information is computed not from the frequency of the bases but from the frequencies of base transitions. For this case the number of possible symbols is now 16, since we can transit from one of the four bases to any of the four bases on the next sequence site as shown in the Fig. 1*(b)*.

III. Results

A. Comparison of Base Frequency Entropy and Base Transition Frequency Entropy measures

Rényi (orders α=*0.5*, α=*0.1*) and Shannon Redundancy for all binding sites are listed in TABLE I. Typical patterns on R convey to different shapes when computed using base frequency or base transition frequency. We include comparative results from both methods for λ cI/Cro and E. Coli ArgR in Fig. 2. Both Shannon and Rényi peak amplitudes evolve differently as is shown when comparing *a)* vs. *b)* and *c)* vs. *d)*. Furthermore, the characteristic shape of the binding site region evolves differently in both methods as the Rényi order decreases ($\alpha \rightarrow 0$). Results suggest that the base transitions contain additional information. Evidence towards this hypothesis can be obtained from both, Fig. 2 and Fig. 3. For instance, in the later figure, T7 Symmetry and LexA binding sites cases are plotted on left and right positions of Fig. 3, respectively. From top to bottom, Redundancies for base frequencies and base transition frequencies, and the transition probability maps are respectively shown. Fig. 3 *c)* shows a clear region with forbidden transitions between positions around 18 to 28, which impacts on larger values on this region. Fig. 3 *f)* shows a clear TATA sequence from position 6 to 14. On T7 Symmetry (left), sharp peaks appearing around position 20 are conserved on both algorithms, independently from the Rényi order. On the other hand, secondary peaks appearing around position 10 and 34 are eliminated in the case of base frequency, as Rényi order decreases but conserved for the base transition case (shaded areas in *a)* and *b)*).

TABLE I
SUMMARY OF THE ORGANISMS AND RECOGNIZERS ANALYZED

Organism	Recognizer	Bases	Aligned seqs.
E. Coli	ArgR	20	34
E. Coli	LexA	20	38
E. Coli	Ribosome	201	569
E. Coli	TrpR	38	6
Lambda	cI/Cro	20	38
T7	HincII	40	60
T7	Polymerase	42	17
T7	Symmetry	44	34

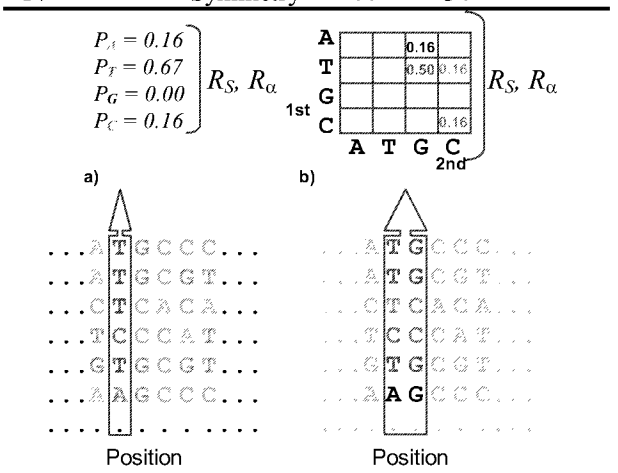

Fig. 1. Schema of the calculus of Entropy Redundancy on: (a) base frequency, and (b) base transition frequency.

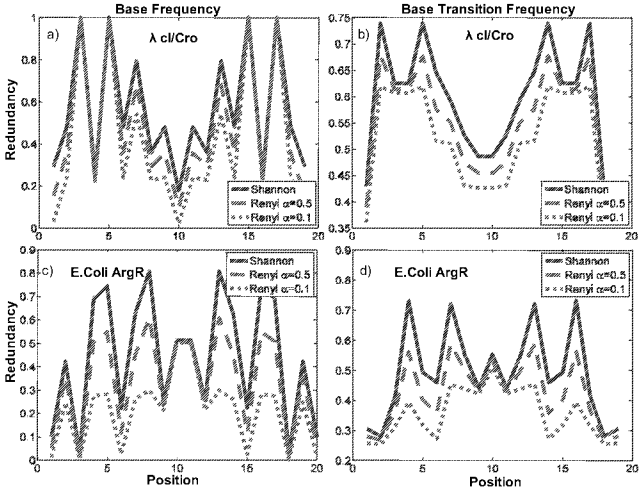

Fig. 2. Redundancy computed using base frequency (left) and base transition frequency (right) for cI/Cro (top) and and ArgR (bottom) binding sites.

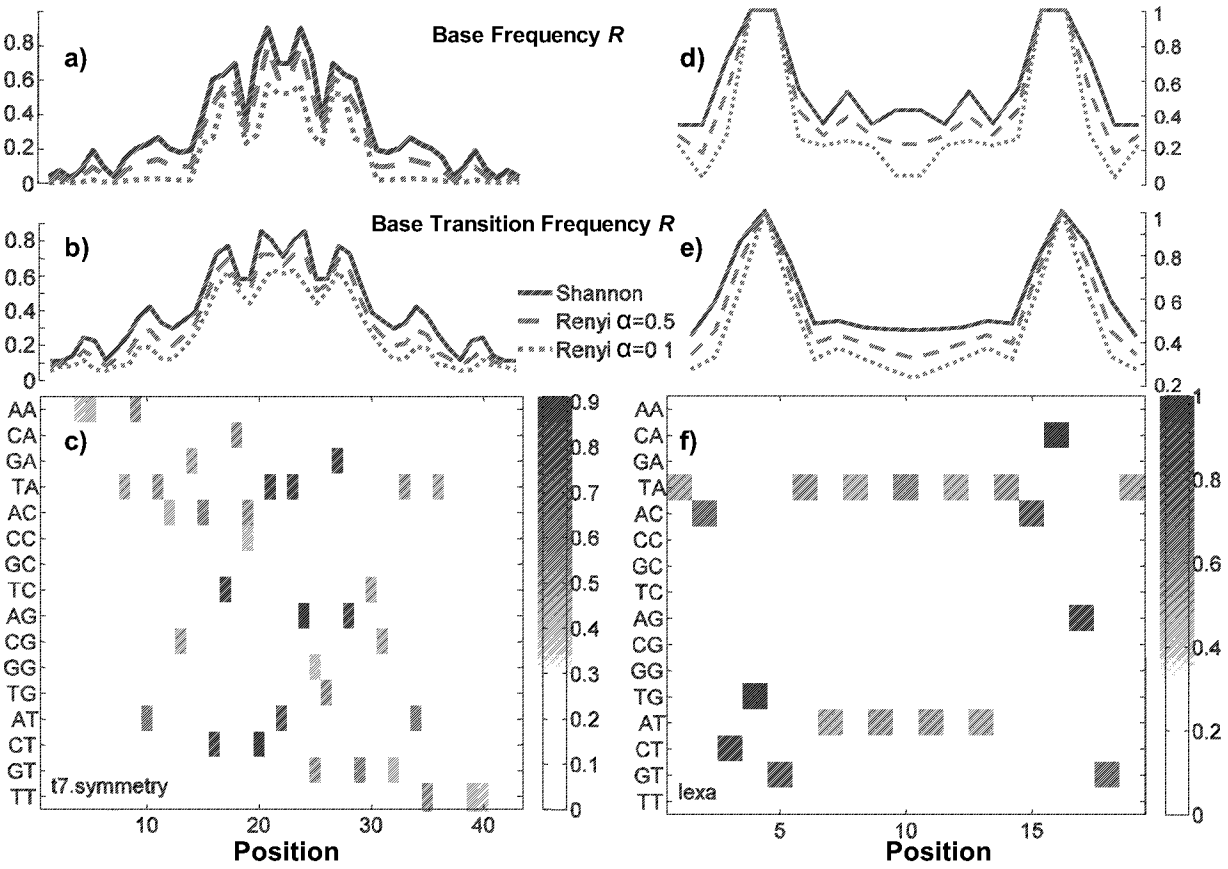

Fig. 3. Base Frequency Redundancy (*a)* and *d)*) and Base Transition Redundancy (*b)* and *e)*) for T7 symmetry (*left*) and E. Coli LexA (*right*) binding sites. Transition probabilities can be graphically observed on *c)* and *f)* figures.

On the other hand, for LexA binding sites, an opposite process occurs when peaks appearing on the base frequency Redundancy are conserved with low Rényi orders (shaded areas in *d)*). However, the same position is down-played by base transition Redundancy measure at low Rényi orders (shaded areas in *e)*). Low Rényi orders trend to penalize those combinations with low p_i in (1). Given that the number of symbols for base frequency is 4 and for base transition frequency the number of words is 16, the structural shape for the binding site evolves differently for $\alpha \rightarrow 0$. Also it should be noted that, for a fixed sample set, 4 symbols provides with larger resolution than 16 on the calculus of Entropy. However, from our observations, it is suggested that both measures do not provide the same information. Information captured by base transition entropy measures complements the information measured with base frequency on each position of the binding site. It is therefore reasonable to argue that results from both methods could be used for the characterization of binding sites in DNA.

B. Effect of the Rényi order on information measures

In this section, the comparative effect of the Rényi order α with respect to Shannon case (α=1) is shown for both methods. Although entropies for different binding sites classes are not comparable, as its value will depend on the length of the region under study, it can be compared how the entropy varies with α in both methods for the same binding site. In TABLE II, mean values and standard deviation of Redundancy are shown for Rényi ($\alpha=0.5, \alpha=0.1$) and Shannon measures. Mean Redundancy captures the overall information measured in the binding site region whereas standard deviation captures how smooth these values are for the different base positions in the region. Interestingly, the trend on Redundancy values for both methods (Base Frequency and Base Transition Frequency) shows a similar evolution with decreasing Rényi orders in different recognizers. Generally, Base Transition Frequency retains higher Redundancy values for all Rényi orders. However, standard deviation follows a different evolution with α. In order to better expose this effect a normalized Redundancy is computed in order to compare the relative evolution from Shannon to low order Rényi measures (this is $\alpha=1$ *to* $\alpha=0.01$). This value is normalized to the Shannon value for both cases:

$$\overline{R}_{\alpha}^{BF} = \frac{R_{\alpha}^{BF}}{R_{S}^{BF}}, \quad \overline{R}_{\alpha}^{BTF} = \frac{R_{\alpha}^{BTF}}{R_{S}^{BTF}}, \tag{4}$$

where BF stands for Base Frequency and BTF for Base Transition Frequency. Obtained mean values are shown in Fig. 4. As suggested before, information retained by BTF is generally better conserved than BF as the Rényi order decreases, except for the case of HincII binding site. This translates in lower slope for the Base Transition case. However, sharpness of the peaks for different positions is overall not as well conserved as in the BF method.

TABLE II
COMPARATIVE REDUNDANCY MEASURES

Recognizer	Base Frequencies*			Base Transition Frequencies		
	Shannon	Rényi(0.5)	Rényi(0.1)	Shannon	Rényi(0.5)	Rényi(0.1)
ArgR	0.440 ±0.27	0.339 ±0.21	0.211 ±0.15	0.492 ±0.15	0.423 ±0.11	0.423 ±0.11
LexA	0.557 ±0.26	0.456 ±0.30	0.352 ±0.34	0.604 ±0.20	0.522 ±0.23	0.522 ±0.23
Ribosome	0.005 ±0.01	0.003 ±0.00	0.001 ±0.00	0.013 ±0.02	0.006 ±0.01	0.006 ±0.01
TrpR	0.528 ±0.31	0.489 ±0.32	0.454 ±0.33	0.630 ±0.18	0.614 ±0.18	0.614 ±0.18
cI/Cro	0.549 ±0.29	0.471 ±0.32	0.393 ±0.35	0.601 ±0.10	0.559 ±0.10	0.559 ±0.10
HincII	0.153 ±0.31	0.139 ±0.31	0.128 ±0.31	0.185 ±0.28	0.164 ±0.28	0.164 ±0.28
Polymerase	0.489 ±0.38	0.439 ±0.38	0.385 ±0.39	0.570 ±0.31	0.531 ±0.31	0.531 ±0.31
Symmetry	0.319 ±0.27	0.238 ±0.24	0.152 ±0.20	0.409 ±0.24	0.331 ±0.23	0.331 ±0.23

* Results in the form *mean* ± *std*

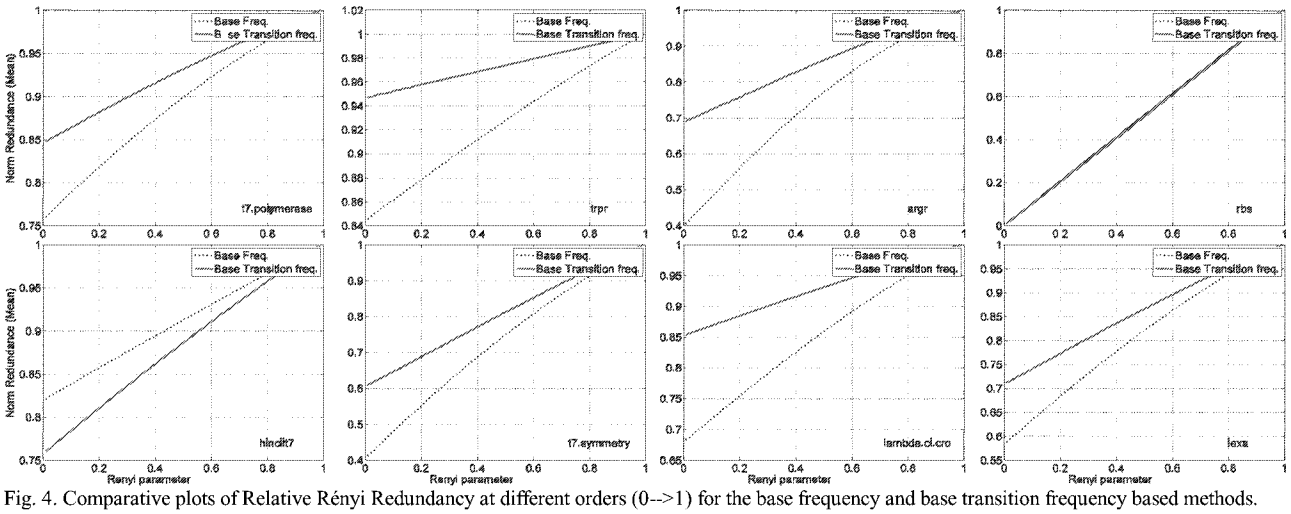

Fig. 4. Comparative plots of Relative Rényi Redundancy at different orders (0-->1) for the base frequency and base transition frequency based methods.

Overall, these results strengthen the idea than both measures are complementary. It is also clear that the amount of information measured by Rényi at low orders is different for the dinucleotide approach than using the nucleotide frequency approach.

IV. Conclusion

In this paper it is shown that additional information can be found focusing to the frequency of transition of the bases entropy complementing base frequency entropy. Redundancy results are provided using Shannon and Rényi entropies for a number of binding site classes in E. Coli, T7 and λ organisms. It is observed that, in the sense of binding site characterization, the amount of information captured by transition based entropies is overall larger than the information captured with base frequency entropies. Also, the information captured is better conserved for base transition method on lower Rényi orders when compared to the base frequency method. Results on the evolution of the uncertainty measures with Rényi orders also suggest a degree of complementary information provided on both approaches. Therefore, Rényi measures on transition base frequency are suggested for studying and characterization of DNA binding sites.

This information could complement several approaches that tackle individual sequence analysis for binding site discovery. In [6], authors suggest the construction of a weight matrix from the frequencies of each nucleotide at each position of the aligned sequence. This matrix is then applied to individual sequences themselves for the determination of the possible sequence conservation of the individual sequence. As we have shown in the present work, weight matrix could also be built from base transitions frequencies in order to be applied to individual sequences, under Rényi based metrics.

References

[1] T. D. Schneirder, G. D. Stormo, L. Gold and A. Ehrenfeuch, "The Information Content of Binding Sites on Nucleotide Sequences," in *J. Mol. Biol.* vol. 188, pp. 415-431, 1986.

[2] A. Krishnamachari, V. moy Mandal and Karmeshu, "Study of DNA binding sites using Rényi parametric entropy measure," in *J. Theor. Biol.*, vol. 227, pp 429—436, 2004.

[3] D. Holste, I. Grosse and H. Herzel, "Statistical Analysis of the DNA sequence of human chromosome 22," in *Phys. Review E,.* vol. 64, 041917, 2001.

[4] A. Rényi, "On measures of information and entropy," in Proceedings *of the 4th Berkeley Symposium on Mathematics, Statistics and Probability,* 1961, pp. 547-561.

[5] C. E. Shannon, "A mathematical theory of communication," in *Bell System Technical Journal,* vol. 27, pp. 379-423 and 623-656, 1948.

[6] T. D. Schneider, "Information Content of Individual Genetic Sequences," in *J. Theor. Biol.* vol.189, pp. 427-441, 1997.