Published online March 1, 2005

1332–1336 *Nucleic Acids Research, 2005, Vol. 33, No. 4*
doi:10.1093/nar/gki271

Human pol II promoter prediction: time series descriptors and machine learning

Rajeev Gangal and Pankaj Sharma*

SciNova Technologies Pvt. Ltd, 528/43 Vishwashobha, Adjacent to Modi Ganpati, Narayan Peth, Pune 411030, Maharashtra, India

Received September 9, 2004; Revised and Accepted February 8, 2005

ABSTRACT

Although several *in silico* promoter prediction methods have been developed to date, they are still limited in predictive performance. The limitations are due to the challenge of selecting appropriate features of promoters that distinguish them from non-promoters and the generalization or predictive ability of the machine-learning algorithms. In this paper we attempt to define a novel approach by using unique descriptors and machine-learning methods for the recognition of eukaryotic polymerase II promoters. In this study, non-linear time series descriptors along with non-linear machine-learning algorithms, such as support vector machine (SVM), are used to discriminate between promoter and non-promoter regions. The basic idea here is to use descriptors that do not depend on the primary DNA sequence and provide a clear distinction between promoter and non-promoter regions. The classification model built on a set of 1000 promoter and 1500 non-promoter sequences, showed a 10-fold cross-validation accuracy of 87% and an independent test set had an accuracy >85% in both promoter and non-promoter identification. This approach correctly identified all 20 experimentally verified promoters of human chromosome 22. The high sensitivity and selectivity indicates that *n*-mer frequencies along with non-linear time series descriptors, such as Lyapunov component stability and Tsallis entropy, and supervised machine-learning methods, such as SVMs, can be useful in the identification of pol II promoters.

INTRODUCTION

One of the challenges in the field of computational biology and especially in the area of computational DNA sequence analysis is the automatic detection of promoter sites. Promoter sites typically have a complex structure consisting of multi-functional binding sites for proteins involved in the transcription initiation process. Promoters have been defined as modular DNA structures containing a complex array of *cis*-acting regulatory elements required for accurate and efficient initiation of transcription and for controlling expression of a gene.

Eukaryotic cells basically contain three different types of RNA polymerases in their nuclei, RNA polymerases I, II and III. RNA polymerase II transcribes all protein-coding sequences in eukaryotic cells, and is the most important of the three polymerases. Promoters in general contain two consensus sequences: (i) a TATA box located ~30 bp upstream from the transcriptional start site and (ii) a CCAAT box located somewhere around −75 bp, with a consensus sequence of GGCCAATCT. There are also a number of other consensus sequences that frequently occur in eukaryotic promoters, which serve as binding sites for a wide variety of protein transcription factors, such as GC box and enchancers. Since, eukaryotic promoters have highly diverse primary sequences; it has been very difficult to find generalized patterns or rules by conventional sequence analysis methods. Promoters contain vital information about gene expression and regulatory networks, including gene targets of individual cascades/signalling pathways (1). The basic aim of computer-assisted promoter recognition is the elucidation of gene transcription and associated genetic regulatory networks. Prediction of the functionality of a promoter would also be welcome for gene therapy approaches to improve the expression of newly created vector constructs.

Several algorithms for the prediction of promoters, transcriptional start points and transcription factor binding sites in eukaryotic DNA sequence now exist (2,3). Although current algorithms perform much better than the earlier attempts, it is probably fair to say that performance is still far from satisfactory.

Prometheus, a machine-learning tool, is designed to address the problem of low-prediction accuracy. It specifically deals with the application of non-linear dynamics and statistical

*To whom correspondence should be addressed. Tel: +91 20 4450282; Fax: +91 20 4450282; Email: pankaj.sharma@scinovaindia.com thermodynamics descriptors, such as Lyapunov component and Tsallis entropy along with non-linear machine-learning algorithms. Prometheus is found to perform significantly better than some other promoter finding programs, NNPP 2.2, Promoter Scan version 1.7, Promoter 2.0 Prediction Server (4), Soft Berry (5) and Dragon Promoter Finder (6).

A DNA sequence can be pictured as a dynamical system. It evolves continuously in the course of evolution and is thus subject to perturbation, i.e. losses and gains of single residues or fragments. It can perhaps further be characterized as a chaotic dynamical system, since a slight change in initial conditions can lead to different outcomes in terms of the final function (7).

The aim of the present study is to provide a distinct classification between promoter and non-promoter sequences. In the present study, we have used properties such as 3mer, 4mer (*n*-mer frequencies) (8) and GC% along with non-linear time series descriptors, i.e. Lyapunov exponent and Tsallis entropy (9). Non-linear time series analysis is being increasingly applied in the fields of biology and physiology, where the systems are expected to be non-linear and a simple linear stochastic description often does not account for the highly complex nature of the observed behaviour. The maximum Lyapunov exponent used here is a qualitative measure of the stability of a dynamical system. A quantitative measure of the sensitive dependence on the initial conditions is the Lyapunov exponent. It is the averaged rate of divergence (or convergence) of two neighbouring trajectories. Lyapunov exponents quantify this divergence by measuring the mean rate of exponential divergence of initially neighbouring trajectories (10). A trajectory of a system with a negative Lyapunov exponent is stable and will converge to an Attractor exponentially with time. The magnitude of the Lyapunov exponent determines how fast the attractor is approached. A trajectory of a system with a positive Lyapunov exponent is unstable and will not converge to an attractor. The magnitude of the positive Lyapunov exponent determines the rate of exponential divergence of the trajectory.

In recent years, considerable interest has been generated in the question of non-extensivity of entropy and statistics of a number of systems. Tsallis entropy, which gives the usual Shannon–Boltzmann–Gibbs entropy as a special case (11) has enjoyed considerable success in dealing with a number or non-equilibrium phenomena and hence, is a prime candidate for application to biological systems. Since, biological systems ranging from genes and proteins to cells, organisms and ecosystems are open and far from equilibrium, so Tsallis entropy might have an important role to play in chemical and biological dynamics in general (12). Tsallis entropy is given by

[:Advent3B2:customise]
[:Advent3B2:customise]
[:Advent3B2:customise]
[:Advent3B2:customise]
[:Advent3B2:customise]
[:Advent3B2:customise]
[:Advent3B2:customise]
[:Advent3B2:customise]
[:Advent3B2:customise]
[:Advent3B2:customise]

[:Advent3B2:customise]
[:Advent3B2:customise]
[:Advent3B2:customise]
[:Advent3B2:customise]
[:Advent3B2:customise]
[:Advent3B2:customise]
[:Advent3B2:customise]
[:Advent3B2:customise]
[:Advent3B2:customise]

[:Advent3B2:customise]
[:Advent3B2:customise]
[:Advent3B2:customise]
[:Advent3B2:customise]
[:Advent3B2:customise]
[:Advent3B2:customise]
[:Advent3B2:customise]
[:Advent3B2:customise]
[:Advent3B2:customise]
[:Advent3B2:customise]
[:Advent3B2:customise]
[:Advent3B2:customise]
[:Advent3B2:customise]{\color{#fff}}
[:Advent3B2:customise]
[:Advent3B2:customise]{\color{#fff}}
[:Advent3B2:customise]
[:Advent3B2:customise]
[:Advent3B2:customise]
[:Advent3B2:customise]
[:Advent3B2:customise]
[:Advent3B2:customise]{\color{#fff}}
[:Advent3B2:customise]
[:Advent3B2:customise]{\color{#fff}}
[:Advent3B2:customise]
[:Advent3B2:customise]{\it S
[:Advent3B2:customise]}
[:Advent3B2:customise]
[:Advent3B2:customise]
[:Advent3B2:customise]

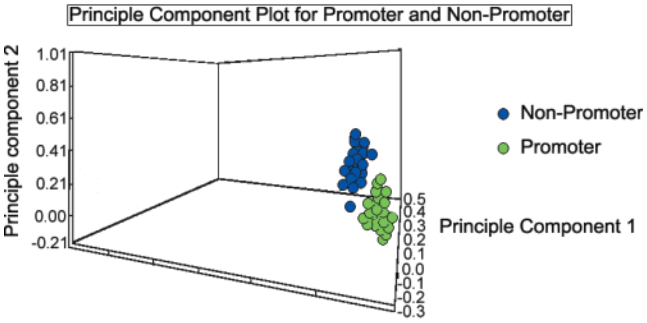

Figure 1. Principle components analysis (PCA) plot for each promoters and non-promoter. The descriptors used to discriminate between promoters and non-promoters are transformed to three orthogonal axes. A clear separation between promoter and non-promoter sequences is shown in the PCA plot.

Table 1. Results of models built for promoter prediction

Input: data promoter and non-promoter sequences	Correctly classified instances on training data (%)	Correctly classified instances on cross-validation data (%)	Correctly classified instances on validation data (%)[a]	Algorithm used	Correlation coefficient	Kappa statistics
Model 1[b]	100.00	87.5	85.8	SVM	0.78	0.74
Model 2[c]	100.00	87.25	86.6	SVM	0.68	0.71

[a] Twenty per cent of the training set was split and used for model validation from the training set.
[b] Model 1 includes calculation of *n*-mer frequencies, GC% and non-linear time series descriptors.
[c] Model 2 only includes calculation of *n*-mer frequencies and GC%.

Table 2. List of experimentally verified promoters on human chromosome 22

Accession number[a]	Gene name	Predicted by Prometheus
L43122	*COMT*	+
X52828	*BCR*	+
X84664	*MMP11*	+
AJ007494	*GGT1*	+
X72990	*EWSR1*	+
M63420	*LIF*	+
AF129855	*OSM*	+
AF047576	*TCN2*	+
AB016655	*LIMK2*	+
S79779	*TIMP3*	+
S58267	*HMOX1*	+
EP11091[b]	*MB*	+
X63578	*PVALB*	+
X53093	*IL2RB*	+
M87841	*H1F0*	+
AF115252	*PLA2G6*	+
EP11139[b]	*PDGFB*	+
AF106656	*ADSL*	+
D86746	*SREBF2*	+
M77378	*ACR*	+
Total 20 genes, correctly predicted instances 20 (100%)		

[a] All sequences are taken from GenBank/EMBL/EPD. See accession number for details.
[b] EPD accession number.

Table 3. Prediction done using the above models

Predicted sequences	Total no. of sequences	True positive	False positive	False negative	True negative
Model 1					
Promoter	800	707	Nil	93	Nil
Intron	1000	Nil	97	Nil	903
Human chromosome 22 experimentally verified promoters	20	20	Nil	Nil	Nil
Model 2					
Promoter	800	682	Nil	118	Nil
Intron	1000	Nil	93	Nil	907
Human chromosome 22 experimentally verified promoters	20	9	Nil	11	Nil

TP, true positives, # {correctly recognized positives}; TN, true negatives, # {correctly recognized negatives}; FN, false negatives, # {positives recognized as negatives}; and FP, false positives, # {negatives recognized as positives}.

[:Advent3B2:customise]
[:Advent3B2:customise]
[:Advent3B2:customise]
[:Advent3B2:customise]
[:Advent3B2:customise]
[:Advent3B2:customise]
[:Advent3B2:customise]
[:Advent3B2:customise]{\color{#fff}}
[:Advent3B2:customise]
[:Advent3B2:customise]{\color{#fff}}
[:Advent3B2:customise]
[:Advent3B2:customise]
[:Advent3B2:customise]
[:Advent3B2:customise]
[:Advent3B2:customise]
[:Advent3B2:customise]{\color{#fff}}
[:Advent3B2:customise]
[:Advent3B2:customise]{\color{#fff}}
[:Advent3B2:customise]
[:Advent3B2:customise]{\it q
[:Advent3B2:customise]}
[:Advent3B2:customise]

[:Advent3B2:customise]
[:Advent3B2:customise]
[:Advent3B2:customise]
[:Advent3B2:customise]
[:Advent3B2:customise]
[:Advent3B2:customise]
[:Advent3B2:customise]
[:Advent3B2:customise]{\color{#fff}}
[:Advent3B2:customise]
[:Advent3B2:customise]{\color{#fff}}
[:Advent3B2:customise]
[:Advent3B2:customise]
[:Advent3B2:customise]
[:Advent3B2:customise]
[:Advent3B2:customise]
[:Advent3B2:customise]
[:Advent3B2:customise]
[:Advent3B2:customise]
[:Advent3B2:customise]
[:Advent3B2:customise]{\color{#fff}}
[:Advent3B2:customise]
[:Advent3B2:customise]{\color{#fff}}
[:Advent3B2:customise]
[:Advent3B2:customise]
[:Advent3B2:customise]
[:Advent3B2:customise]
[:Advent3B2:customise]

Table 4. Program accuracy

Program name	NNPP (threshold 0.8)	Soft Berry (TSSW)	Promoter Scan version 1.7	Dragon Promoter Finder version 1.4	Promoter 2.0 Prediction Server	Prometheus
Sensitivity (%)[a]	32	60	40	38	50	86
Specificity (%)[b]	34	65	56	64	54	88
Correlation coefficient[c]	0.34	0.27	0.11	0.18	0.20	0.74

Line missing

[:Advent3B2:customise]
[:Advent3B2:customise]
[:Advent3B2:customise]
[:Advent3B2:customise]
[:Advent3B2:customise]
[:Advent3B2:customise]
[:Advent3B2:customise]
[:Advent3B2:customise]
[:Advent3B2:customise]
[:Advent3B2:customise]
[:Advent3B2:customise]
[:Advent3B2:customise]
[:Advent3B2:customise]
[:Advent3B2:customise]
[:Advent3B2:customise]
[:Advent3B2:customise]
[:Advent3B2:customise]
[:Advent3B2:customise]
[:Advent3B2:customise]
[:Advent3B2:customise]
[:Advent3B2:customise]=
[:Advent3B2:customise]
[:Advent3B2:customise]
[:Advent3B2:customise]
[:Advent3B2:customise]{
[:Advent3B2:customise]
[:Advent3B2:customise]
[:Advent3B2:customise]
[:Advent3B2:customise]
[:Advent3B2:customise]
[:Advent3B2:customise]
[:Advent3B2:customise]
[:Advent3B2:customise]
[:Advent3B2:customise]
[:Advent3B2:customise]{\rm
[:Advent3B2:customise]
[:Advent3B2:customise]
[:Advent3B2:customise]
[:Advent3B2:customise]
[:Advent3B2:customise]
[:Advent3B2:customise]
[:Advent3B2:customise]
[:Advent3B2:customise]

[:Advent3B2:customise]{\color{#fff}}
[:Advent3B2:customise]
[:Advent3B2:customise]{\color {#fff}}
[:Advent3B2:customise]
[:Advent3B2:customise]
[:Advent3B2:customise]
[:Advent3B2:customise]
[:Advent3B2:customise]
[:Advent3B2:customise]{\color{#fff}}
[:Advent3B2:customise]
[:Advent3B2:customise]{\color{#fff}}
[:Advent3B2:customise]
[:Advent3B2:customise]1
[:Advent3B2:customise]}
[:Advent3B2:customise]
[:Advent3B2:customise]\over
[:Advent3B2:customise]
[:Advent3B2:customise]
[:Advent3B2:customise]
[:Advent3B2:customise]
[:Advent3B2:customise]
[:Advent3B2:customise]
[:Advent3B2:customise]
[:Advent3B2:customise]
[:Advent3B2:customise]
[:Advent3B2:customise]
[:Advent3B2:customise]{\color{#fff}}
[:Advent3B2:customise]
[:Advent3B2:customise]{\color{#fff}}
[:Advent3B2:customise]
[:Advent3B2:customise]
[:Advent3B2:customise]
[:Advent3B2:customise]
[:Advent3B2:customise]
[:Advent3B2:customise]{\color{#fff}}
[:Advent3B2:customise]
[:Advent3B2:customise]{\color{#fff}}
[:Advent3B2:customise]
[:Advent3B2:customise]{\it q
[:Advent3B2:customise]}
[:Advent3B2:customise]
[:Advent3B2:customise]
[:Advent3B2:customise]

[:Advent3B2:customise]
[:Advent3B2:customise]
[:Advent3B2:customise]
[:Advent3B2:customise]{\color{#fff}}
[:Advent3B2:customise]
[:Advent3B2:customise]{\color{#fff}}
[:Advent3B2:customise]
[:Advent3B2:customise]
[:Advent3B2:customise]
[:Advent3B2:customise]
[:Advent3B2:customise]
[:Advent3B2:customise]
[:Advent3B2:customise]
[:Advent3B2:customise]
[:Advent3B2:customise]
[:Advent3B2:customise]{\color{#fff}}
[:Advent3B2:customise]
[:Advent3B2:customise]{\color{#fff}}
[:Advent3B2:customise]
[:Advent3B2:customise]
[:Advent3B2:customise]
[:Advent3B2:customise]
[:Advent3B2:customise]
[:Advent3B2:customise]
[:Advent3B2:customise]
[:Advent3B2:customise]
[:Advent3B2:customise]
[:Advent3B2:customise]
[:Advent3B2:customise]
[:Advent3B2:customise]
[:Advent3B2:customise]
[:Advent3B2:customise]
[:Advent3B2:customise]
[:Advent3B2:customise]
[:Advent3B2:customise]
[:Advent3B2:customise]
[:Advent3B2:customise]
[:Advent3B2:customise]
[:Advent3B2:customise]
[:Advent3B2:customise]
[:Advent3B2:customise]
[:Advent3B2:customise]
[:Advent3B2:customise]
[:Advent3B2:customise]{\minus}
[:Advent3B2:customise]
[:Advent3B2:customise]
[:Advent3B2:customise]
[:Advent3B2:customise]{\rm
[:Advent3B2:customise]
[:Advent3B2:customise]
[:Advent3B2:customise]
[:Advent3B2:customise]
[:Advent3B2:customise]
[:Advent3B2:customise]
[:Advent3B2:customise]

[:Advent3B2:customise]
[:Advent3B2:customise]{\color{#fff}}
[:Advent3B2:customise]
[:Advent3B2:customise]{\color {#fff}}
[:Advent3B2:customise]
[:Advent3B2:customise]
[:Advent3B2:customise]
[:Advent3B2:customise]
[:Advent3B2:customise]
[:Advent3B2:customise]{\color{#fff}}
[:Advent3B2:customise]
[:Advent3B2:customise]{\color{#fff}}
[:Advent3B2:customise]
[:Advent3B2:customise]1
[:Advent3B2:customise]}
[:Advent3B2:customise]
[:Advent3B2:customise]
[:Advent3B2:customise]
[:Advent3B2:customise]}
[:Advent3B2:customise]
[:Advent3B2:customise]
[:Advent3B2:customise]
[:Advent3B2:customise]
[:Advent3B2:customise]
[:Advent3B2:customise]
[:Advent3B2:customise]
[:Advent3B2:customise]
[:Advent3B2:customise]{\color{#fff}}
[:Advent3B2:customise]
[:Advent3B2:customise]{\color{#fff}}
[:Advent3B2:customise]
[:Advent3B2:customise]
[:Advent3B2:customise]
[:Advent3B2:customise]
[:Advent3B2:customise]
[:Advent3B2:customise]
[:Advent3B2:customise]
[:Advent3B2:customise]
[:Advent3B2:customise]
[:Advent3B2:customise]{\color{#fff}}
[:Advent3B2:customise]
[:Advent3B2:customise]{\color{#fff}}
[:Advent3B2:customise]
[:Advent3B2:customise]
[:Advent3B2:customise]
[:Advent3B2:customise]
[:Advent3B2:customise]
[:Advent3B2:customise]
[:Advent3B2:customise]
[:Advent3B2:customise]
[:Advent3B2:customise]
[:Advent3B2:customise]
[:Advent3B2:customise]
[:Advent3B2:customise]
[:Advent3B2:customise]