**Mathematical Biology**

# Coding sequence density estimation via topological pressure

**David Koslicki · Daniel J. Thompson**

**Abstract** We give a new approach to coding sequence (CDS) density estimation in genomic analysis based on the *topological pressure*, which we develop from a well known concept in ergodic theory. Topological pressure measures the 'weighted information content' of a finite word, and incorporates 64 parameters which can be interpreted as a choice of weight for each nucleotide triplet. We train the parameters so that the topological pressure fits the observed coding sequence density on the human genome, and use this to give *ab initio* predictions of CDS density over windows of size around 66,000 bp on the genomes of Mus Musculus, Rhesus Macaque and Drososphilia Melanogaster. While the differences between these genomes are too great to expect that training on the human genome could predict, for example, the exact locations of genes, we demonstrate that our method gives reasonable estimates for the 'coarse scale' problem of predicting CDS density. Inspired again by ergodic theory, the weightings of the nucleotide triplets obtained from our training procedure are used to define a probability distribution on finite sequences, which can be used to distinguish between intron and exon sequences from the human genome of lengths between 750 and 5,000 bp. At the end of the paper, we explain the theoretical underpinning for our approach, which is the theory of Thermodynamic Formalism from the dynamical systems literature. Mathematica and MATLAB implementations of our method are available at http://sourceforge.net/projects/topologicalpres/.

D. Koslicki
Department of Mathematics, Oregon State University,
368 Kidder Hall, Corvallis, OR 97330, USA
e-mail: david.koslicki@math.oregonstate.edu

D. J. Thompson (✉)
Department of Mathematics, The Ohio State University,
100 Math Tower, 231 West 18th Avenue, Columbus, OH 43210, USA
e-mail: thompson@math.osu.edu

## 1 Introduction

We present a novel approach to genomic analysis using tools from the theory of
thermodynamic formalism. A number of recent influential works in mathematical
biology have been based on the philosophy that the methods of statistical mechanics,
and dynamical systems, can give insight into biological problems (Bialek et al. 2012;
Mora et al. 2010; Tkacik et al. 2006; Schneidman et al. 2006). In this spirit, we
adapt tools from thermodynamic formalism (which is a well established branch of
dynamical systems, developed from ideas in statistical mechanics and information
theory), to the study of bioinformatics. The principle concept that we introduce is the
*topological pressure* of a finite sequence, which is adapted from a well known concept
in ergodic theory. It is a real number which is given by counting, with weights, all
distinct subwords of an exponentially shorter length that appear in the original word,
and is interpreted as a weighted measure of complexity of a finite sequence.[1]

The structure and organization of genomes is of central concern to the study of
genome biology, and determining the distribution of coding sequences is a key com-
ponent of this pursuit (Berná et al. 2012; Mackiewics et al. 2010; Salzburger et al. 2009;
Kowalski et al. 2009). Furthermore, identification of gene-rich regions in eukaryotes
(especially in plants) is an ongoing field of research (Ksiazkiewics et al. 2013; Varshney
et al. 2006; Erayman et al. 2004). The topological pressure provides a computational
tool for predicting the distribution of coding sequences and identifying such gene-rich
regions. Our approach is particularly suitable for the study of novel genomes where
limited training data is available. This is especially useful when faced with the recent
aggregation of thousands of little-studied genomes (eg. Genome 10K Haussler et al.
2009).

The primary goals of our analysis are:

1. To use the topological pressure, trained on the human genome, to give *ab initio*
   predictions of coding sequence density on other genomes (Mus Musculus, Rhesus
   Macaque, Drososphilia Melanogaster). This establishes the key practical advantage
   of our approach, which is that we can predict CDS density using only a single
   moderately phylogenetically distant informant genome as training data.
2. To use the theory of thermodynamic formalism to turn the data encoded in the
   parameters used in (1) into a probability distribution which can measure the coding
   potential of sequences of nucleotides of lengths between 750 and 5,000 bp.

1.1 Predictions for CDS density

The *coding sequence density* (or CDS density) is the probability density function given
by the bin count of coding sequences in non-overlapping windows of a given size. We

---

[1]  See Sect. 2.1 for a precise definition, and Sect. 2.2 for biological interpretation.

focus on windows of size approximately 66,000 bp for reasons we describe later. This corresponds to dividing, for example, the autosomes of the human genome into roughly 40,000 windows. The topological pressure, which depends on 64 parameters (one for each nucleotide triplet) assigns a real number to each of these windows, and we train these parameters by maximizing the correlation with the observed CDS density on a genome.

After obtaining our parameters by training on the human genome, and cross-validating our results to check we are not overfitting, we give *ab initio* predictions of the CDS density of Mus Musculus, Rhesus Macaque and Drososphilia Melanogaster simply by computing the topological pressure along these genomes. We find that the correlation between topological pressure (trained on the human genome) and the observed CDS density on these genomes is 0.77, 0.73 and 0.60 respectively. The decrease in the correlation roughly corresponds to increasing phylogenetic distance between the human genome and the target genome.

Our predictions of CDS density can be improved by using better training data (for example, topological pressure would estimate the CDS density of Drososphilia Melanogaster very accurately if it were trained on the genome of Drososphilia Simulans), however our results emphasize that we can still make reasonable predictions of CDS density even if we are not able to train on a close relative of the target genome. This relatively low sensitivity to organism-specific genomic traits means that although our method cannot hope to predict any finer structure of a genome (for example, the exact location of genes), our technique is advantageous for the identification of regions of high CDS density for novel genomes where refined training data is unavailable. Our approach is also suitable for *ab initio* prediction on non-mammalian genomes if a suitable model genome is chosen as training data, although we do not develop this line of research here.

## 1.2 Comparison with gene-finding techniques

In the last ten years, a number of powerful and effective gene-finding software packages have been developed (e.g. Augustus, Contrast, Exoniphy, Genemark HMM, FGenesh, GenSCAN, GeneID, N-SCAN, SNAP). While these packages were not primarily designed for estimating CDS density, this information can be inferred by taking a bin count of the predicted coding sequences. These methods, which are typically based on Hidden Markov Models or conditional random fields, are often very effective at gene prediction on reasonably well understood genomes, although gene sensitivity/specificity and accuracy of predicted intron-exon structure is typically much lower (Yandell and Ence 2012, p. 333; Flicek 2007, Fig. 1).

The drawback of these gene-finding methods is that they achieve only limited success on novel genomes (Korf 2004; Yandell and Ence 2012), as they rely on parameter files which are either partially trained on the genome under study, or use detailed data from a large number of closely related informant genomes. In particular, the training procedure requires a large number of high-quality genes and error-free assemblies, and can require data that is not yet available for new genomes (Yandell and Ence 2012, p. 333; Guig et al. 2006, pp. S2.2–S2.3; Giogo and Reese 2005, p. 577; Carter and Durbin 2006, p. S6.2).

We investigate the predicted CDS density given by some of these methods for comparison. We use GeneID on each of the genomes we consider, and find the predictions to be comparably accurate to the predictions yielded by our method. While the first version of GeneID was developed over ten years ago, it remains widely used, and we found that it often outperformed more recent gene-finding software for estimating CDS density. We ran GENSCAN and GenemarkHMM on all three genomes, and they were outperformed by GeneID in all three cases.

We considered a selection of the most recent gene-finding software packages (N-SCAN, Exoniphy, CONTRAST) on the genomes where suitable data was available for their implementation. CONTRAST gave the best prediction over any method considered on Drosophilia Melongaster, yielding a correlation of 0.92. This is not surprising since CONTRAST utilizes 14 informant genomes closely related to Drosophilia Melongaster (for example, Drosophilia Simulans and Drosophilia Yakuba) to make these predictions. This amount of training data would usually be unavailable for the analysis of a novel genome. We showed that Exoniphy performed very effectively on Mus Musculus, performing as effectively as topological pressure.

Apart from these examples, we do not give a comprehensive study of the performance of these advanced gene-finding programs for estimating CDS density, but it is our expectation that they perform as well, or better, than topological pressure when good training data is available. We emphasize that the advantage of our approach is the possibility of predicting CDS density in situations where insufficient data is available to effectively train the leading gene-finding software packages.

Another advantage of our approach is its simplicity and speed: the topological pressure can predict a CDS density for a genome in a matter of seconds, while *ab initio* prediction programs typically take a few hours, and evidence-based methods can take weeks (Yandell and Ence 2012, p. 335).

### 1.3 A probability distribution on short segments of DNA sequences

Inspired once more by the techniques of ergodic theory, we demonstrate how our parameters determine a probability distribution on finite sequences, called an *equilibrium measure*. We show that this probability distribution assigns relatively large weight to sequences which are known to be exons. This property can be used to predict the coding potential of DNA sequences which are orders of magnitude shorter than those on which the topological pressure is trained.

The equilibrium measure is a Markov measure, so this construction can be interpreted as using the topological pressure (which makes no Markovian assumption at the training stage) to produce a Markov model suitable for identifying coding sequences. The theoretical basis for this construction is the Variational Principle from Sect. 5, which shows that the equilibrium measure maximizes a certain kind of entropy. While Markov models and entropy maximization are both familiar ideas in sequence modeling (Durbin et al. 1998), the new ingredients here are the method for obtaining the Markov model, and the interpretation of the Markov model via topological pressure as an equilibrium measure.

The development of robust techniques that detect the coding potential of short sequences is an important area of research (Creanza et al. 2009; Fickett and Tung

1992; Gao and Zhang 2004; Guigó and Fickett 1995; Lin et al. 2008, 2011; Saeys et al. 2007; Washietl et al. 2011) with applications to sequence annotation as well as gene prediction. We show that our equilibrium measure is reasonably effective in distinguishing between randomly selected introns and exons of length 750 bp in the human genome. While this approach is not as effective as the powerful comparative techniques developed in, for example, Washietl et al. (2011), our method could be useful on novel genomes. Furthermore, this result can be interpreted as evidence that our parameters are capturing the differences in distribution of 3-mers between coding sequences and non-coding sequences.

### 1.4 Layout

The layout of the paper is as follows: In Sect. 2, we develop our methodology. In Sect. 3, we present the results of our analysis of topological pressure and CDS density. In Sect. 4, we demonstrate how the topological pressure defines a measure on finite sequences, and show that this measure can distinguish between coding sequences and non-coding sequences. In Sect. 5, we explain the theoretical basis for our approach, and give more general definitions suitable for use in future analyses.

## 2 Methodology

### 2.1 Topological pressure

We introduce the mathematical content of our study, and then show how it can be applied to genomic analysis. The topological pressure is a well known and well studied concept in the ergodic theory of dynamical systems. The standard version is a quantity associated to a topological dynamical system which measures the 'weighted' exponential orbit complexity of the system (Parry and Pollicott 1990; Parry and Tuncel 1982; Walters 1982). We introduce a finite implementation of topological pressure which can be interpreted as a measurement of weighted information content of a finite sequence. Topological pressure is a weighted version of topological entropy, which is a parameter free quantity introduced by Koslicki (2011). Topological entropy was shown to be effective in distinguishing between intron and exon sequences (Koslicki 2011). For ease of exposition, we state here only a special case of the definition of topological pressure, which is the one we use for our investigation of DNA sequences, and then give a series of remarks which explain why it is defined this way. We postpone the general definition of topological pressure until Sect. 5.

We consider finite sequences on the symbols $A, C, G, T$. We use the expressions 'finite sequence' and 'word' synonymously. However, 'subword' has a different meaning from 'subsequence': a subword is a subsequence whose entries are consecutive entries of the original sequence. We write a word either by using sequence notation, or juxtaposition, so the sequence $(A, G, A, T, C)$ may be written simply as $AGATC$.

We weight each word of length 3, which we think of as a nucleotide triplet, with a positive real parameter. After choosing some order for the triplets (e.g. lexicographic order), it is convenient to record these parameters in a vector

$$\mathbf{v} = (v_{AAA}, v_{AAC}, v_{AAG}, \ldots \ldots, v_{TTG}, v_{TTT}) \tag{2.1}$$

with 64 coordinates. We are free to assume that $\mathbf{v}$ is a probability vector (we explain why in Sect. 5). We define $\Phi_\mathbf{v}$ to be the real-valued function on the collection of words of length 3 that sends a word to its corresponding entry in $\mathbf{v}$. In other words, for $a_1, a_2, a_3 \in \{A, C, G, T\}$,

$$\Phi_\mathbf{v}(a_1 a_2 a_3) := v_{a_1 a_2 a_3}. \tag{2.2}$$

We can use the parameters encoded in $\mathbf{v}$ to induce a weight on a word $u = (u_1, u_2, \ldots, u_n)$ of length $n \geq 3$ by the expression

$$\text{"Weight assigned to } u\text{"} = \prod_{i=1}^{n-2} \Phi_\mathbf{v}(u_i u_{i+1} u_{i+2}). \tag{2.3}$$

The topological pressure of a word with respect to $\mathbf{v}$, whose formal definition follows, is given by counting the number of distinct subwords of an exponentially shorter length, with weights given by the expression (2.3).

**Definition 2.1** Let $m \geq n$ and let $w = (w_1, w_2, \ldots, w_m)$ be a finite sequence where each $w_i \in \{A, C, G, T\}$. We let $SW_n(w)$ denote the set of all subwords of length $n$ that appear in $w$, that is

$$SW_n(w) = \{w_i w_{i+1} \cdots w_{i+n-1} : i \in \{1, 2, \ldots, m - n + 1\}\}.$$

Suppose that $w$ has length $m = 4^n + n - 1$. Let $\mathbf{v}$ be a probability vector of the form (2.1). We define the topological pressure of $w$ with respect to the parameters $\mathbf{v}$, denoted $P(w, \mathbf{v})$, to be

$$P(w, \mathbf{v}) = \frac{1}{n} \log_4 \left( \sum_{u \in SW_n(w)} \prod_{i=1}^{n-2} \Phi_\mathbf{v}(u_i u_{i+1} u_{i+2}) \right). \tag{2.4}$$

*Remark 2.1* Since $SW_n(w)$ is defined as a set (rather than a sequence), subwords are not counted with multiplicity, so the expression inside the parentheses in (2.4) is counting the *distinct* length $n$ subwords of $w$, with weights determined by the parameters $\mathbf{v}$ via the expression (2.3).

*Remark 2.2* The definition above only applies to words whose length are of the form $4^n + n - 1$ for some $n \in \mathbb{N}$, and this is the $n$ which appears in equation (2.4). There are obvious ways to extend the definition of topological pressure to a word of arbitrary length (e.g. by truncating or averaging), but in this paper we need only consider words whose length are of this form. In this study, we set $n = 8$, so we are looking for all distinct subwords of length 8 in a window of length $4^8 + 7 = 65, 543$.

*Remark 2.3* When all entries in $\mathbf{v}$ are chosen to be equal (i.e. each entry is $\frac{1}{64}$), $P(w, \mathbf{v})$ reduces to the definition of topological entropy for finite sequences in Koslicki (2011). The reason we take the logarithm in base 4 in (2.4), and the length of form $4^n + n - 1$, rather than just $4^n$, is so that the maximum value of the topological entropy is exactly 1, and that there exist sequences on which this maximum is attained (see discussion after Definition 5.1 for details).

*Remark 2.4* It is possible to set up topological pressure so that instead of assigning a parameter value to each 3-mer, we assign a parameter value to each $k$-mer for some fixed $k \geq 1$ (we give the details in Sect. 5). We focus on $k = 3$ because of the biological importance of 3-mers in the genetic code. Furthermore, we will see that using $4^3$ parameters neither overfits nor underfits our training data. We do not expect significant improvement to the results of this paper if we considered weightings on $k$-mers with $k > 3$, and we would risk overfitting the data. Conversely, we checked that the case $k = 3$ is a better fit for the data than $k = 2$.

*Remark 2.5* In practical situations, we must also deal with the occurrence of non-*ACTG* symbols (e.g. $N$). We do this by only including the subwords composed entirely of the symbols *ACTG* in our computation of topological pressure. This is crucial for a genome like Rhesus Macaque where entries of $N$ appear throughout the genome. For a word $w$ with only a few occurrences of $N$, this has negligible effect on our computations. On the other hand, a word $w$ with many occurrences of $N$ has low topological pressure. This effect is consistent with our application to genomic analysis, because we want the topological pressure to predict low CDS density in regions with many occurrences of $N$. Alternatively, for very accurate genome assemblies such as the human genome, we can eliminate the vast majority of non-*ACTG* symbols by removing the telomeres and centromeres of each chromosome. We can then restrict our attention to sequences composed entirely of *ACTG* without difficulty.

## 2.2 High topological pressure sequences: biological interpretation

The sequences for which the topological pressure is large are those that balance high complexity against high frequency of 3-mers with relatively large parameter values. This intuition is made precise by the variational principle for topological pressure from ergodic theory which we discuss in Sect. 5.2. Regions containing a large number of coding sequences will tend to have a different distribution of 3-mers from those regions that do not, and we search for parameter values so that the topological pressure can detect this difference.

It is crucial that topological pressure maximizes complexity and frequency of strongly weighted 3-mers simultaneously: maximizing only complexity would favor random sequences, while maximizing only the frequency of strongly weighted 3-mers would favor sequences with very low complexity, neither of which we would expect to see in regions of high CDS density. On the other hand, we demonstrate that topological pressure, which balances both these effects, can be trained so that high topological pressure correlates with high CDS density.

Heuristically, we think of the 3-mers which receive a relatively large parameter value in $\mathbf{v}$ to be those which are sending a strong signal that we are in a coding region, while those with relatively small parameter value are those that are associated with non-coding regions, or do not send us a strong signal in either direction.

While this heuristic may seem simplistic given the complexity of the relationship between nucleotide composition and the structure of genes, it is supported by a number of results in this paper. In Sect. 3.7, we show that if we choose parameters based on this hueristic (by basing the parameters on the frequency of 3-mers in exons), then topological pressure correlates positively with CDS density. This correlation is significantly weaker than that obtained by our training procedure, which is consistent with our expectations. Also in keeping with this heuristic, the results of Sect. 4.2 show that the parameters obtained by our training procedure can be used to define a measure which classifies introns and exons.

## 2.3 Topological pressure and CDS density estimation

The *coding sequence density* (or CDS density) is the probability density function representing the percentage of coding sequences in non-overlapping windows of a given size. We describe our methodology for training the topological pressure to match the observed distribution of coding sequences on the human genome, and on other data sets.

We utilize the NCBI hg18 build 36.3 with coding sequences defined by NCBI RefSeq genes and accessed via the UCSC table browser (Karolchik et al. 2004). We choose a chromosome and fix an integer window size $m$ to divide the chromosome into non-overlapping windows of length $m$. The selection of the window size exhibits the typical trade-off between sensitivity and specificity: a smaller window size gives finer information on the CDS distribution, but exhibits a higher sensitivity to fluctuations in nucleotide composition. The most suitable window sizes for comparison with the topological pressure are those of the form $m = 4^n + n - 1$. We focus on a window size of 65,543 ($n = 8$), as this seems to achieve a good balance. This corresponds to dividing the autosomes of the human genome into roughly 40,000 non-overlapping windows. We remove any windows with non-$ACTG$ symbols, as the vast majority of these correspond to telomeres and centromeres. We could also carry out our analysis with different window sizes. The case $n = 7$, which gives window size $m = 16,390$, would also be a reasonable choice and could give finer results, although it would be more computationally intensive and susceptible to noise.

**Notation 2.1** We divide each chromosome of the human genome into non-overlapping windows of length $m = 65,543$, assuming the chromosome is read in the $p$ to $q$ direction.

Let $\text{Chr}(i)$ denote the word which represents the $i$th chromosome of the human genome, and $\text{Chr}(i, [n, m])$ denote the subword which starts at position $n$ and ends at position $m$. Let $w(i; n)$ denote the sequence which represents the $n$th such window along the ith chromosome of the human genome.[2] In other words,

---

[2] We are left with a shorter window at the end of each chromosome, and we omit these from our study.

$$w(i; n) = \text{Chr}(i, [(n - 1)m + 1, (n - 1)m + m]). \tag{2.5}$$

**Definition 2.2** We define the bin count for coding sequences in each window as follows:

$$\#CS(i; n) := \#\{\text{RefSeq coding sequences with initial nucleotide}$$
$$\text{contained in } w(i; n)\}.$$

The coding sequence density on chromosome $i$ is defined to be

$$\text{CDS}(i, n) := \#CS(i, n)/\#CS(i),$$

where $\#CS(i) := \#\{\text{Known coding sequences in Chr}(i)\}$.

For fixed $i$, $\text{CDS}(i, n)$ is a probability density function of $n$. Note that our notation suppresses our choice of window size, as this stays fixed at $m = 65,543 = 4^8 + 7$ throughout this work.

**Notation 2.2** Given a probability vector $\mathbf{v}$ with 64 entries, as described at (2.1), we consider the topological pressure with respect to $\mathbf{v}$ of each of the sequences $w(i; n)$ using the following notation:

$$P(i, n, \mathbf{v}) := P(w(i; n), \mathbf{v}),$$

where $P(\cdot, \cdot)$ is the topological pressure given by (2.4). Thus, $P(i, n, \mathbf{v})$ is the topological pressure with respect to $\mathbf{v}$ of the sequence which arises as the $n$th non-overlapping window of length 65,543 along the $i$th chromosome of the human genome.

On each chromosome, i.e. for each fixed $i$, we can consider $\text{CDS}(i, n)$ and $P(i, n, \mathbf{v})$ as functions in $n$. In fact, we want to consider these functions as $i$ ranges over a specified collection of chromosomes, most often the collection of all autosomes of the human genome. That is, the indices $i$ and $n$ are replaced with a new index $t = t(i, n)$ which tells us which window of this data set is under consideration. We modify the normalization of the coding sequence density so that $\text{CDS}(t)$ is a probability density function of $t$, and we consider $\text{CDS}(t)$ and $P(t, \mathbf{v})$ as functions in $t$. This is essentially equivalent to considering the concatenation of all the autosomes as a single sequence. Similarly, we can consider $\text{CDS}(t)$ and $P(t, \mathbf{v})$ ranging over even larger data sets, for example by concatenating all the autosomes from a number of different model species into a single sequence.

After fixing our data set, we train the parameters $\mathbf{v}$ for maximum positive correlation between $\text{CDS}(t)$ and $P(t, \mathbf{v})$. Our focus is mainly on the case when the data set is all autosomes of the human genome, although other data sets, both larger and smaller, are investigated where appropriate in this study. We demonstrate that our training procedure neither underfits nor overfits this training data.

2.4 Details of training procedure

For a fixed collection of chromosomes as described above, we use the method of Nelder and Mead (1965) to maximize the correlation between $P(t, \mathbf{v})$ and CDS$(t)$ with respect to probability vectors $\mathbf{v}$ with 64 entries.

Considered as functions in $t$, both CDS$(t)$ and $P(t, \mathbf{v})$ are inherently noisy due to random fluctuations in nucleotide composition in a given chromosome as well as due to incomplete knowledge regarding coding sequences (eg. incorrectly annotated sequences). The noise in both functions is suppressed by utilizing a Gaussian filter. The radius of the Gaussian filter is chosen so that it coincides at each $t$ with the Gaussian kernel density estimation of CDS$(t)$.

We checked that other standard smoothing techniques (moving medians, exponential moving averages, convolution with a smoothing kernel) lead to similar results, and chose the Gaussian filter for our analysis due to its simplicity and speed of implementation.

We utilize the method of Nelder and Mead (1965) in MATLAB (2012) to maximize the correlation between $P(t, \mathbf{v})$ and CDS$(t)$ with respect to $\mathbf{v}$. The precision threshold for the convergence of this heuristic maximization technique was set to $10^{-6}$ and convergence was typically achieved in 10,000 steps of the algorithm.

We focus on the case where the training data is the collection of all human autosomes. We did not include the sex chromosomes due to the well-known differences in mutation rate, selection, gene death and gene survival between the autosomes and the sex chromosomes (Wilson and Makova 2009a, b; Kvikstad et al. 2007; Graves 2006; Makova et al. 2004). We denote the parameters trained on all human autosomes as $\mathbf{v}_{max}$.

## 3 Results

Using the methodology above, we present our results on CDS prediction using the topological pressure.

### 3.1 Training on the human genome

Our training procedure yields parameters $\mathbf{v} = \mathbf{v}_{max}$ so that $P(t, \mathbf{v})$ and CDS$(t)$ have correlation above 0.9 across all autosomes of the human genome. It is not at all obvious that our training procedure should work this effectively, as we are training 64 parameters to maximize correlation over approximately 40,000 data points. That our training procedure even works gives evidence that topological pressure can detect structure in the training data.

### 3.2 Cross-validation

Since our method yields a very high correlation between $P(t, \mathbf{v})$ and CDS$(t)$, we must check if we are overfitting the 64 parameters in $\mathbf{v}$. We performed a traditional 7-fold cross-validation on chromosomes 1 through 21 (Picard and Cook 1984). We randomly

partitioned the chromosomes into 7 equal-size samples. Of these 7, a single sample of three chromosomes was retained as a test sample. We performed the maximization procedure outlined in Sect. 2.4 on the remaining 6 samples and used the resulting parameters to obtain a correlation value between the topological pressure and the test sample CDS density. An average is then taken over the 7 possible choices of test sample. We repeated this procedure 50 times. The resulting mean correlation was 0.8049 with a variance of 0.0003232. This demonstrates that the maximization procedure outlined in Sect. 2.4 is not overfitting.

### 3.3 Training on multiple genomes

We can also train topological pressure on multiple informant genomes. Using the methodology of Sect. 2.4, we obtained parameters by training on the data set given by concatenating all autosomes of the human, mouse (mm9 ) and rat (rn4) genomes. These are the parameters we use when we refer to 'topological pressure (trained on 3 genomes)' in the following sections. This is intended simply to demonstrate that topological pressure can incorporate information from multiple genomes, and a thorough investigation of the effectiveness of this idea is beyond the scope of this paper.

### 3.4 CDS density estimation on Rhesus Macaque

We used the parameters $\mathbf{v}_{\max}$ obtained from training on the human genome and showed that the correlation of the topological pressure with the coding sequence density given by RefSeq genes over all the autosomes of the Rhesus Macaque build rheMac3 was 0.726. We repeated the experiment using the parameters trained on 3 genomes, and obtained a very slightly improved correlation of 0.738. We compare this with the predictions given by GeneMarkHMM (Lukashin and Borodovsky 1998), GeneID (Blanco et al. 2002), GENSCAN (Burge and Karlin 1997), and N-SCAN (van Baren et al. 2007).

We used the GeneID and GeneMarkHMM software to obtain predicted coding sequences for the Rhesus Macaque autosomes. For GENSCAN and NSCAN, we obtained this information from the corresponding track on the UCSC table browser. For each program, we then took the bin counts of predicted coding sequences over all autosomes in the non-overlapping windows described at (2.5). Table 1 summarizes the correlation with the known coding sequence bin counts (obtained from RefSeq genes) and the bin counts predicted by each method. Figure 1 demonstrates how well GeneID and topological pressure reconstruct the coding sequence density on chromosome 2.

We see that topological pressure yields the highest correlation of all the methods we looked at on this genome, and N-SCAN gave the best prediction yielded by the gene-finding programs we considered.

### 3.5 CDS density estimation on Mus Musculus

The correlation of the topological pressure, trained on the human genome, with the coding sequence density of the autosomes from Mus Musculus build mm9 was 0.765.

**Table 1** Comparison of predictions of CDS density on rheMac3

| Method | Correlation over all autosomes |
| --- | --- |
| Topological pressure (trained on human) | 0.726 |
| Topological pressure (trained on 3 genomes) | 0.738 |
| GeneMarkHMM | 0.624 |
| GENSCAN | 0.402 |
| GeneID | 0.660 |
| N-SCAN | 0.684 |



- · - Pressure   - - - GeneID   — Actual CDS Density
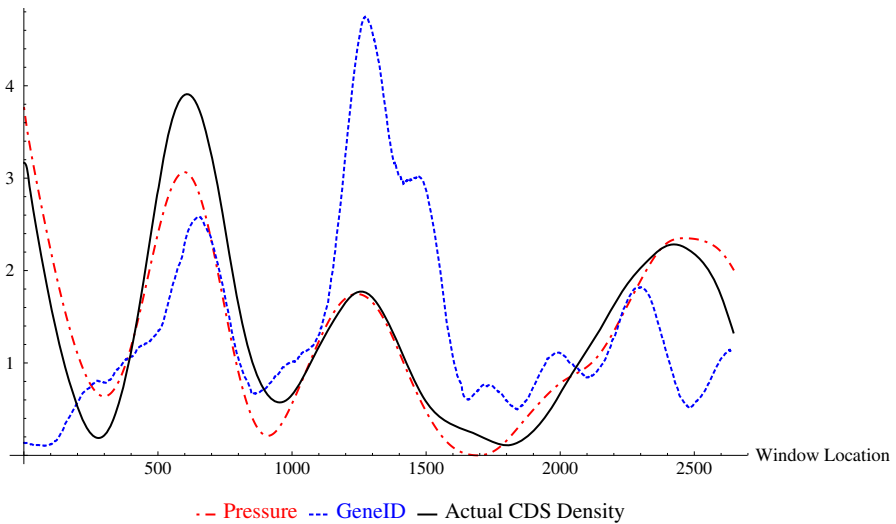
**Fig. 1** Topological pressure (trained on the human genome), CDS density predicted by GeneID, and known CDS density on chromosome 2 of rheMac3

We compare the topological pressure with predictions yielded by gene-finding techniques using the same methodology described in the previous section.

We ran GeneMarkHMM on Mus Musculus genome build mm9 and obtained the GENSCAN, GeneID, and Exoniphy tracks from the UCSC table browser for this genome. Table 2 summarizes the correlation of each method with the known coding sequences density (obtained from RefSeq Genes).

Topological pressure was outperformed on this genome by GeneID and Exoniphy, but performed better than GeneMarkHMM and GENSCAN.

### 3.6 CDS density estimation on Drosophila Melanogaster

The correlation of the topological pressure, trained on the human genome, with the coding sequence density of the autosomes from Drosophila Melanogaster build dm3 was 0.601. This improved to 0.674 when we used the parameters trained on 3

**Table 2** Comparison of predictions of CDS density on mm9

| Method | Correlation over all autosomes |
|---|---|
| Topological pressure (trained on human) | 0.765 |
| GeneMarkHMM | −0.440 |
| GENSCAN | 0.695 |
| GeneID | 0.817 |
| Exoniphy | 0.861 |

**Table 3** Comparison of predictions of CDS density on dm3

| Method | Correlation over all autosomes |
|---|---|
| Topological pressure (trained on human) | 0.601 |
| Topological pressure (trained on 3 genomes) | 0.674 |
| GeneMarkHMM | 0.368 |
| GENSCAN | 0.608 |
| GeneID | 0.871 |
| CONTRAST | 0.918 |

genomes, and we expect that the correlation would improve significantly if we trained on a genome which was more closely related to Drosophila Melanogaster. We do not do this precisely because we want to demonstrate that we can still make reasonable predictions even when a close relative of the target genome is not available for training.

In Table 3, we compare the CDS prediction via topological pressure to those given by the following gene-finding techniques: GeneMarkHMM, GENSCAN, GeneID, and CONTRAST. The best performing method is CONTRAST. This may not be surprising since it uses 14 informant genomes closely related to Drosophila Melanogaster (for example, Drosophila Simulans and Drosophila Yakuba).

## 3.7 Other approaches to parameter selection

The topological pressure can be considered using parameters selected by means other than training against known data. To detect CDS density, we can select the parameters **v** according to the heuristic rule that '3-mers which we believe to be associated to coding sequences are assigned greater weight'. We give an example.

Many single sequence techniques for measuring the coding potential of DNA sequences are based upon frequencies of $n$-mers in known intronic and exonic regions (Akashi 2001; Comeron and Aguadé 1998; Creanza et al. 2009; Karlin et al. 1998). We can use this principle to write down parameters $\mathbf{v}_{exon}$ which are based simply on the frequency of codons in the exon sequences. More precisely, for a codon $w$, the corresponding parameter value in $\mathbf{v}_{exon}$ is assigned by the following procedure: for a segment of an autosome that corresponds to a known exon region, we count the number of times (counting overlaps) that $w$ appears, and then we sum this over all such segments. We normalize by the total number of codons (counting overlaps) that
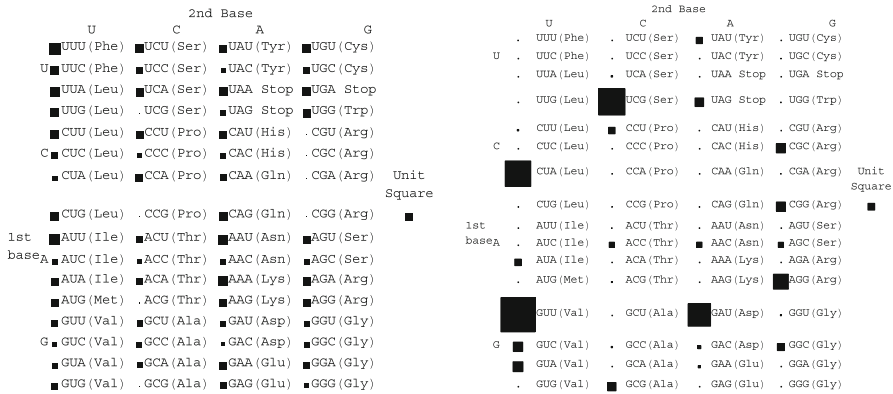
```
                2nd Base                                          2nd Base
       U        C        A        G                      U        C        A        G
   ■UUU(Phe) ■UCU(Ser) ■UAU(Tyr) ■UGU(Cys)        . UUU(Phe)  . UCU(Ser)  ■ UAU(Tyr)  . UGU(Cys)
 U ■UUC(Phe) ■UCC(Ser) .UAC(Tyr) ■UGC(Cys)      U  . UUC(Phe)  . UCC(Ser)  . UAC(Tyr)  . UGC(Cys)
   ■UUA(Leu) ■UCA(Ser) ■UAA Stop ■UGA Stop         . UUA(Leu)  . UCA(Ser)  . UAA Stop  . UGA Stop
   ■UUG(Leu) .UCG(Ser) ■UAG Stop ■UGG(Trp)         . UUG(Leu)  ■ UCG(Ser)  ■ UAG Stop  . UGG(Trp)
   ■CUU(Leu) ■CCU(Pro) ■CAU(His) .CGU(Arg)         . CUU(Leu)  ■ CCU(Pro)  . CAU(His)  . CGU(Arg)
 C ■CUC(Leu) .CCC(Pro) ■CAC(His) ■CGC(Arg)      C  . CUC(Leu)  . CCC(Pro)  . CAC(His)  ■ CGC(Arg)
   .CUA(Leu) ■CCA(Pro) ■CAA(Gln) .CGA(Arg)   Unit  ■ CUA(Leu)  . CCA(Pro)  . CAA(Gln)  . CGA(Arg)  Unit
                                            Square                                               Square
   ■CUG(Leu) .CCG(Pro) ■CAG(Gln) .CGG(Arg)  ■      . CUG(Leu)  . CCG(Pro)  . CAG(Gln)  ■ CGG(Arg)  ■
1st ■AUU(Ile) ■ACU(Thr) ■AAU(Asn) ■AGU(Ser) 1st    . AUU(Ile)  . ACU(Thr)  . AAU(Asn)  . AGU(Ser)
base ■AUC(Ile) ■ACC(Thr) ■AAC(Asn) ■AGC(Ser) base  . AUC(Ile)  ■ ACC(Thr)  ■ AAC(Asn)  ■ AGC(Ser)
   ■AUA(Ile) ■ACA(Thr) ■AAA(Lys) ■AGA(Arg)  A      ■ AUA(Ile)  . ACA(Thr)  . AAA(Lys)  . AGA(Arg)
   ■AUG(Met) .ACG(Thr) ■AAG(Lys) ■AGG(Arg)         . AUG(Met)  . ACG(Thr)  ■ AAG(Lys)  ■ AGG(Arg)
   ■GUU(Val) ■GCU(Ala) ■GAU(Asp) ■GGU(Gly)         ■ GUU(Val)  . GCU(Ala)  ■ GAU(Asp)  . GGU(Gly)
 G ■GUC(Val) ■GCC(Ala) .GAC(Asp) ■GGC(Gly)      G  ■ GUC(Val)  . GCC(Ala)  . GAC(Asp)  ■ GGC(Gly)
   ■GUA(Val) ■GCA(Ala) ■GAA(Glu) ■GGA(Gly)         ■ GUA(Val)  . GCA(Ala)  . GAA(Glu)  . GGA(Gly)
   ■GUG(Val) .GCG(Ala) ■GAG(Glu) ■GGG(Gly)         . GUG(Val)  ■ GCG(Ala)  . GAG(Glu)  . GGG(Gly)
```

**Fig. 2** Values of $50 \times \mathbf{v}_{exon}$ and $35 \times \mathbf{v}_{max}$ overlaid on the genetic code

appear in the collection of segments considered, and this yields the entry in $\mathbf{v}_{exon}$ for the codon $w$. See Fig. 2.

The correlation between $P(t, \mathbf{v}_{exon})$ and CDS($t$) is 0.4886. The positive correlation matches our expectations, but it is much weaker than the correlation obtained using $\mathbf{v}_{max}$.

### 3.8 Analysis of parameter values

The biology enters our machinery via our choice of parameters. Since we train against known CDS density, the parameters reflect the relationship between the distribution of 3-mers and the distribution of coding sequences along the genome. Although our method is entirely combinatorial, it would be desirable to give biological interpretation to the values assigned to 3-mers by $\mathbf{v}_{max}$. Obvious questions include:

1. What relationship between 3-mers and coding sequences does topological pressure really detect? We are not simply detecting the average frequency of appearance of 3-mers in coding sequences, since the values associated to the 3-mers by $\mathbf{v}_{max}$ have a different, and much less uniform, distribution than average frequencies would suggest (see Fig. 3). The parameters are detecting a more sophisticated relationship between the appearance of 3-mers, and their role in coding sequence formation than simply calculating frequencies, and it would be desirable to identify what biological mechanisms explain our parameter values.
2. Do the values of $\mathbf{v}_{max}$ tell us anything about codon usage in the human genome? If we train on different genomes, what are the differences between the parameters obtained? Can this help us understand differences in codon usage between species?

A parameter sensitivity analysis will be a crucial first step in the investigation and interpretation of $\mathbf{v}_{max}$, and we hope to address these questions in future work.

We mention a feature of $\mathbf{v}_{max}$ which does match with biological intuition: 3-mers made up of a single repeating nucleotide are assigned a low value by $\mathbf{v}_{max}$. Thus, the topological pressure will assign a low value to a long sequence of single repeated
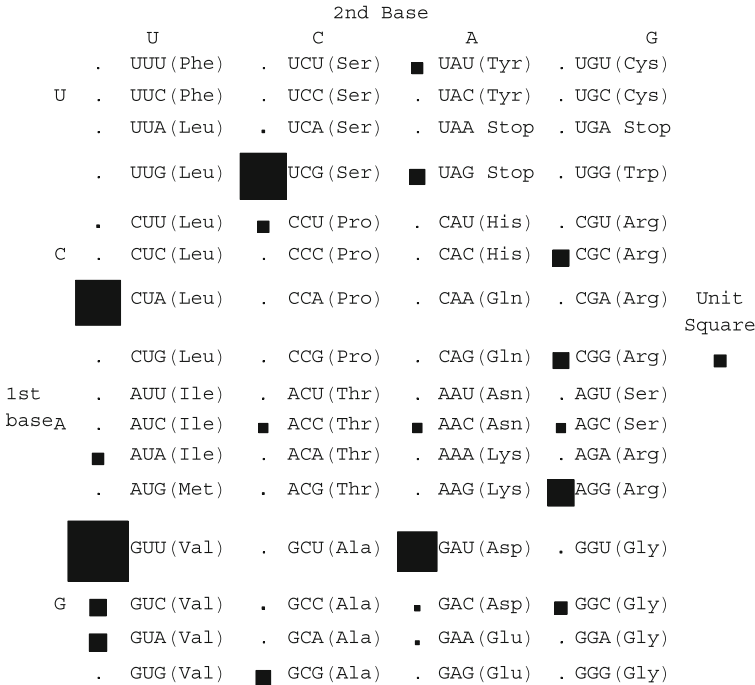
|     |     | 2nd Base |     |     |
| --- | --- | --- | --- | --- |
|     | U | C | A | G |
| U | UUU(Phe) | UCU(Ser) | UAU(Tyr) | UGU(Cys) |
|   | UUC(Phe) | UCC(Ser) | UAC(Tyr) | UGC(Cys) |
|   | UUA(Leu) | UCA(Ser) | UAA Stop | UGA Stop |
|   | UUG(Leu) | UCG(Ser) | UAG Stop | UGG(Trp) |
| C | CUU(Leu) | CCU(Pro) | CAU(His) | CGU(Arg) |
|   | CUC(Leu) | CCC(Pro) | CAC(His) | CGC(Arg) |
|   | CUA(Leu) | CCA(Pro) | CAA(Gln) | CGA(Arg) |
|   | CUG(Leu) | CCG(Pro) | CAG(Gln) | CGG(Arg) |
| A | AUU(Ile) | ACU(Thr) | AAU(Asn) | AGU(Ser) |
|   | AUC(Ile) | ACC(Thr) | AAC(Asn) | AGC(Ser) |
|   | AUA(Ile) | ACA(Thr) | AAA(Lys) | AGA(Arg) |
|   | AUG(Met) | ACG(Thr) | AAG(Lys) | AGG(Arg) |
| G | GUU(Val) | GCU(Ala) | GAU(Asp) | GGU(Gly) |
|   | GUC(Val) | GCC(Ala) | GAC(Asp) | GGC(Gly) |
|   | GUA(Val) | GCA(Ala) | GAA(Glu) | GGA(Gly) |
|   | GUG(Val) | GCG(Ala) | GAG(Glu) | GGG(Gly) |

1st base — A

Unit Square ■

**Fig. 3** Values of $35 \times \mathbf{v}_{max}$ overlaid on the genetic code

nucleotides. This is consistent with the presence of repetitive elements in intergenic regions of the genome.

## 4 A probability measure for detection of coding potential

An important area of research is to develop single sequence measures that effectively distinguish between short coding sequences and short non-coding sequences (Creanza et al. 2009; Fickett and Tung 1992; Gao and Zhang 2004; Guigó and Fickett 1995; Lin et al. 2008, 2011; Saeys et al. 2007; Washietl et al. 2011). The theory of thermodynamic formalism gives us a means of selecting a Markov measure $\mu_{\mathbf{v}}$, which reflects the properties of the topological pressure with respect to the parameters $\mathbf{v}$. We carry out this procedure for our parameters $\mathbf{v} = \mathbf{v}_{max}$ and obtain a measure that is effective for the analysis of relatively short segments of DNA sequences. We explain the theoretical underpinning for our methodology, and generalize this construction, in Sect. 5. We demonstrate that $\mu_{\mathbf{v}}$ can distinguish between coding and non-coding sequences with a reasonably high probability of success. The advantage of using the measure $\mu_{\mathbf{v}}$ rather than the topological pressure associated to $\mathbf{v}$ is that the measure is effective in analyzing relatively short DNA sequences (750–5,000 bp).

This represents a strategy in which large scale information (parameters obtained by considering windows of 66,000 bp along the whole human genome) can be utilized to extract information at a much smaller scale (measure of a sequence of length 750–5,000 bp).

4.1 Construction of $\mu_{\mathbf{v}}$ from $\mathbf{v}$

We use the parameters

$$\mathbf{v} = (v_{AAA}, v_{AAC}, v_{AAG}, \ldots\ldots, v_{TTG}, v_{TTT})$$

to define $\mu_{\mathbf{v}}$ as a stationary Markov measure of memory 2. In other words, our construction gives a Markov chain whose state space is the collection of all sequences of length 2 in the DNA alphabet, and whose transition probabilities are obtained from the parameters $\mathbf{v}$ by the rule (4.1) below. The measure $\mu_v$ is then given by the standard rule for probability of a finite path of a Markov chain. See, for example, Durbin et al. (1998) for a standard reference for these ideas in the context of biological sequence analysis.

More precisely, let $\mathcal{B} = \{A, C, G, T\}^2$, and enumerate $\mathcal{B}$ by

$$w_1 = AA, w_2 = AC, w_3 = AG, w_4 = AT, w_5 = CA, \ldots, w_{16} = TT.$$

We now use $\mathbf{v}$ to define a non-negative matrix $M$ of dimension 16 as follows. Let $M_{ij} = \mathbf{v}_w$, where if $w_i = IJ$, and $w_j = JK$, then $w = IJK$. Let $M_{ij} = 0$ if the second letter in $w_i$ is not the same as the first letter in $w_j$. The Perron-Frobenius theorem guarantees that there is a maximal eigenvalue $\lambda > 0$ and a strictly positive vector $r$ such that

$$Mr = \lambda r.$$

Now define the matrix $P$ by the equation

$$P_{ij} = \frac{M_{ij}r_j}{\lambda r_i}. \tag{4.1}$$

It is a standard exercise to check that $P_{ij}$ is a stochastic matrix and that there is a unique probability vector $p$ so that $pP = p$. More explicitly, $p_i$ is given by normalizing the vector $l_i r_i$, where $l$ is a strictly positive left eigenvector for $M$. For $a, b, c \in \{A, C, T, G\}$, let $p(ab) = p_i$ when $ab = w_i$, and let $P(ab, bc) = P_{ij}$ when $ab = w_i$ and $bc = w_j$.

**Definition 4.1** We define a stationary probability measure $\mu_{\mathbf{v}}$ on $\mathcal{A}^n$ for any fixed $n \geq 3$, by the formula

$$\mu_{\mathbf{v}}(x_1 \cdots x_n) = p(x_1 x_2) P(x_1 x_2, x_2 x_3) P(x_2 x_3, x_3 x_4) \ldots P(x_{n-2} x_{n-1}, x_{n-1} x_n)$$

for each $x_1 \cdots x_n \in \mathcal{A}^n$.

As an illustrative example, to compute $\mu_{\mathbf{v}}$ for the word $GCTAC$, we use the formula

$$\mu_{\mathbf{v}}(GTCAC) = p(GT) P(GT,TC) P(TC,CA) P(CA, AC),$$

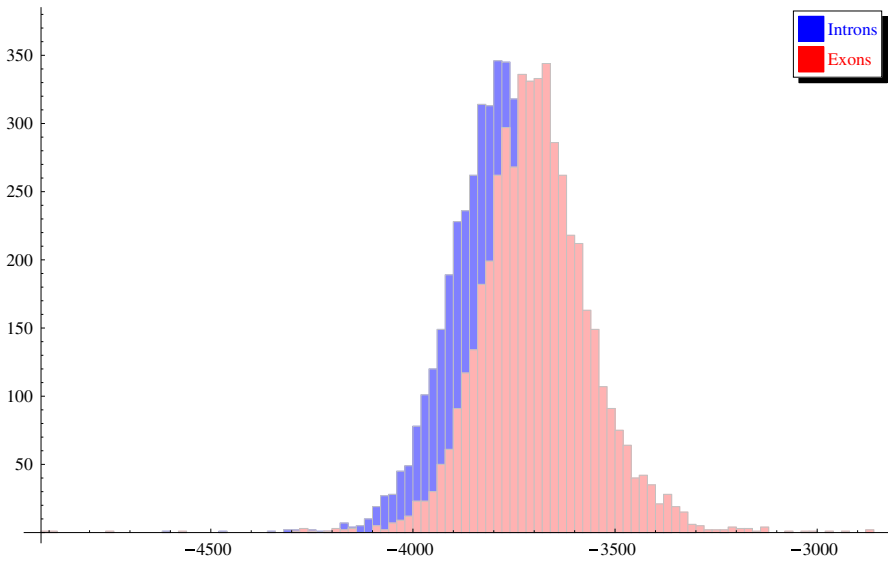and read off the appropriate values for the right hand side of the equation.

**Fig. 4** Histogram of $\log(\mu_{\mathbf{v}})$ on 5,000 introns and exons of length 750 bp

### 4.2 Detection of coding potential using $\mu_{\mathbf{v}}$

We take the measure $\mu_{\mathbf{v}}$ corresponding to the parameters $\mathbf{v} = \mathbf{v}_{\max}$ from Sect. 3. The construction of the measure is designed so that $\mu_{\mathbf{v}}$ reflects the properties of the topological pressure with respect to $\mathbf{v}$ (see Sect. 5.3 for details). Thus, we expect that the sequences with relatively large measure are those with higher coding potential.

We demonstrate this phenomena by showing that $\mu_{\mathbf{v}}$ can partially distinguish between a randomly selected assortment of intron and exon sequences of length 750 bp. Sequences of this length are produced by some next-generation sequencing platforms (e.g. PacBio RS II, Roche GS FLX+). We randomly select 5,000 intron sequences and 5,000 exon sequences from human chromosome 1, and truncate to a length of 750 bp. These sequences are completely un-preprocessed: no information such as ORF's, stop/start codons or repeat masking is utilized.

As expected, $\mu_{\mathbf{v}}$ typically weights exon sequences more heavily than intron sequences. This is demonstrated by Fig. 4, which shows the histogram of $\log(\mu_{\mathbf{v}})$ evaluated on the test sequences. The area under the ROC (true positive rate vs. false positive rate) curve is 0.701.

We repeated the experiment for a randomly selected assortment of introns and exons of length 5,000 bp, and include in Fig. 5 the ROC curve associated to the resulting $\mu_{\max}$. The area under the ROC curve increased to 0.826.

We expect that this classification could be improved, particularly for shorter sequences, by the following strategies:

1. considering more parameters in the topological pressure, which would yield a Markov measure of higher order (as described in Sect. 5.1);
2. training on windows of much smaller length than the $\sim 66,000$ bp used previously.

**Fig. 5** ROC curve for $\log(\mu_\mathbf{v})$ on 5,000 introns and exons of length 5,000 bp
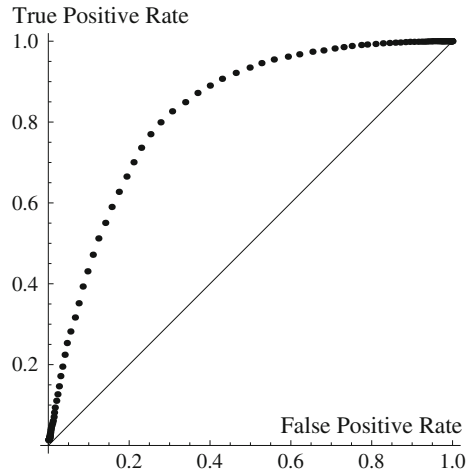


We do not pursue this here, and as it stands, the comparative techniques already available on the human genome (Washietl et al. 2011; Creanza et al. 2009) are more accurate classifiers of introns and exons than $\mu_\mathbf{v}$. Nevertheless, the equilibrium measure could potentially be a useful classifier of introns and exons on less well understood genomes. Furthermore, these results demonstrate how the parameter values for topological pressure can be used to construct a Markovian model, which captures the biological information incorporated into our machinery via the training data.

## 5 Theoretical underpinnings

Topological pressure and equilibrium measures are the principle object of study of thermodynamic formalism, which is a well established branch of ergodic theory and dynamical systems. Standard references are Baladi (2000); Bowen (1975); Parry and Pollicott (1990); Parry and Tuncel (1982); Walters (1982). In this section, we explain the connections between the present work and the classical theory.

First, we extend the definition of topological pressure for finite sequences to full generality. Let $\mathcal{A}$ be an alphabet, that is, a finite collection of symbols, and $|\mathcal{A}|$ denote the number of elements in $\mathcal{A}$. We denote the space of sequences of length $n$ by $\mathcal{A}^n$, the space of finite sequences (of any length) $\mathcal{A}^{<\mathbb{N}}$, the space of finite sequences of length at least $n$ by $\mathcal{A}^{\geq n}$ and the space of infinite sequences by $\Sigma = \mathcal{A}^{\mathbb{N}}$. For a suitable choice of $k$, we select a weight for each word in $\mathcal{A}^k$. The weights can be encoded by a vector $\mathbf{v}$, as in Sect. 2, or by a function $\psi : \mathcal{A}^k \mapsto \mathbb{R}$ so that $\psi(w)$ is the weight assigned to $w$. We use the latter notation here, because it is consistent with the conventions of the dynamical systems literature. In ergodic theory, $\psi$ is customarily called the 'potential function'. We avoid this terminology as the word 'potential' has other meanings in biology. Often, for a function $\Phi > 0$, we are interested in the weights corresponding to $\psi = \log \Phi$.

For $n \geq k$, we assign a weight to each word $u \in \mathcal{A}^n$ by the rule

$$\text{"Weight assigned to } u\text{"} = \exp\left\{\sum_{i=1}^{n-k+1} \psi(u_i u_{i+1} \cdots u_{i+k-1})\right\}$$

$$= \prod_{i=1}^{n-k+1} \Phi(u_i u_{i+1} \cdots u_{i+k-1}), \text{ if } \Phi > 0, \psi = \log \Phi.$$

**Definition 5.1** Let $\psi : \mathcal{A}^k \mapsto \mathbb{R}, m \geq n \geq k$ and let $w = (w_1, w_2, \ldots, w_m)$ be a finite sequence where each $w_i \in \mathcal{A}$. We let $SW_n(w)$ denote the set of all subwords of length $n$ that appear in $w$, that is

$$SW_n(w) = \{w_i w_{i+1} \cdots w_{i+n-1} : i \in \{1, 2, \ldots m - n + 1\}\}.$$

Now suppose that $w$ has length $4^n + n - 1$, i.e. suppose $m = 4^n + n - 1$. Then we can define the topological pressure of $w$ with respect to $\psi$, denoted $P(w, \psi)$, to be

$$P(w, \psi) = \frac{1}{n} \log_{|\mathcal{A}|} \left( \sum_{u \in SW_n(w)} \exp\left\{ \sum_{i=1}^{n-k+1} \psi(u_i u_{i+1} \cdots u_{i+k-1}) \right\} \right). \quad (5.1)$$

If $\Phi > 0$, and $\psi = \log \Phi$, where log denotes natural logarithm, then

$$P(w, \log \Phi) = \frac{1}{n} \log_{|\mathcal{A}|} \left( \sum_{u \in SW_n(w)} \prod_{i=1}^{n-k+1} \Phi(u_i u_{i+1} \cdots u_{i+k-1}) \right). \quad (5.2)$$

For a word $w$ with $|\mathcal{A}|^n + n - 1 \leq |w| < |\mathcal{A}|^{n+1} + n$, we define the topological pressure of $\psi$ on $w$ to be the topological pressure of $\psi$ on the first $|\mathcal{A}|^n + n - 1$ symbols of $w$.

Definition 5.1 generalizes Definition 2.1 because $P(w, \mathbf{v}) = P(w, \log \Phi_{\mathbf{v}})$, where $\Phi_{\mathbf{v}}$ is the function defined at (2.2). When $\psi = 0$, (5.1) reduces to the definition of topological entropy for finite sequences in Koslicki (2011). We denote the greatest topological pressure for words of length $4^n + n - 1$ by

$$P_{\max}(n, \psi) = \max\{P(w, \psi) : |w| = 4^n + n - 1\}. \quad (5.3)$$

For each $n$, there exists a word $w_{\max}^n$ of length $4^n + n - 1$ which has every word of length $n$ as a subword. This follows easily from the fact that the De Brujn graph is a Hamiltonian graph (Gheorghiciuc and Ward 2008, obs. 1.6). It follows that $P_{\max}(\psi, n) = P(w_{\max}^n, \psi)$, and thus $P_{\max}(n, 0) = 1$.

Taking a multiple of $\Phi$ (equivalently adding a constant to $\psi$) does not affect the quantities associated to the topological pressure that we study in this paper, particularly

correlation with the CDS density developed in Sect. 2.3. For any $t > 0$, and word $w$ of length $4^n + n - 1$ we have the formula

$$P(w, \log t\Phi) = \frac{n - k}{n} \log_{|\mathcal{A}|} t + P(w, \log \Phi). \tag{5.4}$$

Since the difference between $P(w, \log t\Phi)$ and $P(w, \log \Phi)$ is a constant independent of $w$, the correlations studied in Sect. 2 will remain unchanged when normalizing $\Phi$. Hence we are free to assume that $\mathbf{v}$ is a probability vector in Sect. 2.

### 5.1 Equilibrium measures

Given a function $\psi : \mathcal{A}^k \mapsto \mathbb{R}$, there is a unique probability measure $\mu_\psi$, called the *equilibrium measure* for $\psi$, whose properties reflect those of the topological pressure with respect to $\psi$. The measure $\mu_{\mathbf{v}}$ constructed in Sect. 4.1 is an equilibrium measure. In this section, we describe how to construct equilibrium measures and explain the theoretical basis for their useful properties.

The construction is a generalization of the construction of $\mu_{\mathbf{v}}$, and a special case of more general expositions given by Baladi (2000); Bowen (1975); Parry and Pollicott (1990); Parry and Tuncel (1982); Walters (1982). We take our finite alphabet $\mathcal{A}$, and a function $\psi : \mathcal{A}^k \mapsto \mathbb{R}$.

Let $\mathcal{B} = \mathcal{A}^{k-1}$ and enumerate $\mathcal{B}$ by some natural ordering. Define a $1 - 0$ square matrix $S$ of dimension $|\mathcal{A}|^{k-1}$ as follows. Let $S_{ij} = 1$ if and only if the word obtained by omitting the first symbol of $w_i$ is the same as the word obtained by omitting the last symbol in $w_j$. In this case, define $\pi(w_i, w_j) \in \mathcal{A}^k$ as the word $w_i b$, where $b \in \mathcal{A}$ is the last symbol in $w_j$. Equivalently, $\pi(w_i, w_j) = aw_j$, where $a \in \mathcal{A}$ is the first symbol of $w_i$.

We now use $\psi$ to define a non-negative matrix $M$ of dimension $|\mathcal{B}|$ as follows. If $S_{ij} = 1$, then let

$$M_{ij} = e^{\psi(\pi(w_i, w_j))}, \tag{5.5}$$

and if $S_{ij} = 0$, then let $M_{ij} = 0$. The Perron-Frobenius theorem gives a maximal eigenvalue $\lambda > 0$ and a strictly positive vector $r$ such that

$$Mr = \lambda r.$$

Now define a matrix $P$ of dimension $|\mathcal{B}|$ by

$$P_{ij} = \frac{M_{ij} r_j}{\lambda r_i}. \tag{5.6}$$

It is easy to check that $P_{ij}$ is a stochastic matrix and that there is a unique probability vector $p$ so that $pP = p$. More explicitly, $p_i$ is given by normalizing the vector $l_i r_i$, where $l$ is a strictly positive left eigenvector for $M$.

To define a measure on $\mathcal{A}^{\mathbb{N}}$, it suffices to define the measure on the cylinder sets

$$[x_1 \cdots x_n] := \{y \in \mathcal{A}^{\mathbb{N}} \mid y_1 = x_1, y_2 = x_2, \ldots, y_n = x_n\}, \tag{5.7}$$

since these are open sets which generate the natural topology on $\mathcal{A}^{\mathbb{N}}$ (Walters 1982).

**Definition 5.2** We define a probability measure $\mu_\psi$ on $\mathcal{A}^{\mathbb{N}}$ by the formula

$$\mu_\psi([x_1 \cdots x_n]) = p_{i_1} P_{i_1 i_2} P_{i_2 i_3} \cdots P_{i_{n-k} i_{n-k+1}}, \tag{5.8}$$

for any $x_1 \cdots x_n \in \mathcal{A}^n$ with $n \geq k$, where $w_{i_1} = x_1 \cdots x_{k-1}$, $w_{i_2} = x_2 \cdots x_k, \ldots,$ $w_{i_{n-k+1}} = x_{n-k+1} \cdots x_n$. We call the measure $\mu_\psi$ the equilibrium measure for $\psi$ on $\mathcal{A}^{\mathbb{N}}$.

For any fixed $n \geq k$, we can take the value assigned to each $x_1 \cdots x_n$ by the formula (5.8) to define a probability measure on $\mathcal{A}^n$, which we refer to as the equilibrium measure for $\psi$ on $\mathcal{A}^n$. Thus, the probability measure $\mu_{\mathbf{v}}$ from Definition 4.1 is the equilibrium measure for $\log \Phi_{\mathbf{v}}$ on $\{A, C, G, T\}^n$.

## 5.2 Relation to theory of dynamical systems: the full shift and the Variational Principle

In the next few sections, we recall the classical theory from dynamical systems which explains the importance of $\mu_\psi$. We demonstrate the relationship between the concepts introduced in this paper and the dynamics of the full shift (defined below).

**Definition 5.3** The *full shift* over an alphabet $\mathcal{A}$ is the dynamical system $(\Sigma, \sigma)$, where $\Sigma = \mathcal{A}^{\mathbb{N}}$ is the space of infinite sequences on $\mathcal{A}$, and $\sigma$ is the *shift map* $\sigma : \Sigma \to \Sigma$, which is the map defined by 'shifting' a sequence one position to the left. That is, for $(x_1, x_2, x_3, \ldots) \in \Sigma$,

$$\sigma((x_1, x_2, x_3, \ldots)) := (x_2, x_3, x_4, \ldots).$$

**Definition 5.4** Given a continuous function $\psi : \Sigma \to \mathbb{R}$, the *topological pressure* of $\psi$ on $\Sigma$ is defined to be:

$$P(\Sigma, \psi) = \lim_{n \to \infty} \frac{1}{n} \log \left( \sum_{u \in \mathcal{A}^n} \exp \sum_{i=0}^{n-1} \psi(\sigma^i u) \right).$$

The following result (Parry and Tuncel 1982; Walters 1982) gives the fundamental relationship between the topological pressure and $\sigma$-invariant probability measures[3] on $\Sigma$.

---

[3] That is, probability measures which satisfy $\mu(\sigma^{-1}A) = \mu(A)$ for all Borel sets $A \subset \Sigma$.

**Theorem 5.1** (Variational Principle) *The topological pressure of $\psi$ on $\Sigma$ satisfies:*

$$P(\Sigma, \psi) = \sup_{m} \left\{ h_m + \int \psi \, dm \right\}, \tag{5.9}$$

*where the supremum is taken over all $\sigma$-invariant probability measures on $\Sigma$, and $h_m$ denotes the measure theoretic entropy, given by*

$$h_m = \lim_{n \to \infty} -\frac{1}{n} \sum_{w \in \mathcal{A}_n} m([w]) \log m([w]).$$

*A measure achieving the supremum in* (5.9) *is called an* equilibrium measure *for $\psi$.*

The following result, proved by Parry and Tuncel (1982), tells us that the measure constructed in the previous section is indeed an equilibrium measure in this sense.

**Theorem 5.2** *The measure $\mu = \mu_\psi$ defined in Definition 5.2 is the unique equilibrium measure for $\psi$ (in the sense of Theorem 5.1), and*

$$P(\Sigma, \psi) = h_\mu + \int \psi \, d\mu = \log \lambda,$$

*where $\lambda$ is the Perron-Frobenius eigenvalue of the matrix* (5.5).

The Variational Principle illustrates the trade-off between structure and complexity which is detected by the topological pressure, simultaneously maximizing entropy (which is itself maximized by the uniform measure) and the integral of $\psi$ (which is itself maximized by a Dirac measure).

### 5.3 The Gibbs property

The relationship between $\psi$ and $\mu_\psi$ is captured by the *Gibbs property*, see Bowen (1975); Parry and Pollicott (1990). To simplify notation, we return to the case of $\psi : \mathcal{A}^3 \mapsto \mathbb{R}$, which is the important case for this paper.

**Theorem 5.3** (Gibbs property) *For $\psi : \mathcal{A}^3 \mapsto \mathbb{R}$ and any $w \in \mathcal{A}^n$,*

$$\mu_\psi([w]) \asymp \exp \left\{ -n P(\Sigma, \psi) + \sum_{i=1}^{n-2} \psi(w_i w_{i+1} w_{i+2}) \right\},$$

*where $[w]$ is the cylinder set defined at* (5.7), *and $a_n \asymp b_n$ means there exists a constant $C > 1$ so that $C^{-1} \le a_n/b_n \le C$ for all $n$.*

Thus, if $\psi = \log \Phi$ and we normalize $\psi$ so that $P(\Sigma, \psi) = 0$ (which is done by taking a suitable multiple of $\Phi$), then

$$\mu_\psi([w]) \asymp \prod_{i=1}^{n-2} \Phi(w_i w_{i+1} w_{i+2}). \tag{5.10}$$

In the context of Sect. 4.2, this formula provides the intuition that sequences which have a relatively high frequency of words $w \in \mathcal{A}^3$ where $v_w$ is large, and a relatively small frequency of words $w \in \mathcal{A}^3$ where $v_w$ is small, will be assigned relatively large measure by $\mu_{\mathbf{v}}$. This gives a theoretical underpinning for using $\mu_{\mathbf{v}}$ to predict the coding potential of short sequences.

## 5.4 Relationship between topological pressure for finite sequences and topological pressure on the full shift

We continue to focus on the case when $\psi : \{A, C, T, G\}^3 \to \mathbb{R}$ for simplicity, and we write $\Sigma$ for the full shift on $\{A, C, T, G\}$. The following result is essentially that of Walters (1982, Theorem 7.30). Let $M$ be the matrix constructed in Sect. 4.1, and recall that $\lambda$ is its Perron-Frobenius eigenvalue. We consider the matrix norm of $M$ given by $\|M\| = \sum_{i,j} |m_{ij}|$.

**Theorem 5.4** *We have*

$$P_{max}(\psi, n) = \frac{1}{n} \log_4 \left( \sum_{u \in \mathcal{A}^n} \exp \left\{ \sum_{i=1}^{n-2} \psi(u_i u_{i+1} u_{i+2}) \right\} \right) = \log_4 \|M^{n-2}\|^{1/n},$$

*The sequence $\|M^{n-2}\|^{1/n}$ converges to $\lambda$ exponentially fast as $n \to \infty$.*

This theorem tells us that for large $n$, $P_{max}(\psi, n)$ is very close to $\log_4 \lambda$. Since $P(\Sigma, \psi) = \log \lambda$, this describes the relationship between topological pressure for finite sequences and topological pressure on the full shift.

## 6 Conclusion

We demonstrated that the topological pressure can train on the human genome to fit the observed bin count of coding sequences on windows of size approximately 66,000 bp. We showed that topological pressure, trained on the human genome, gave effective estimates of CDS density on Rhesus Macaque, Mus Musculus and Drosophilia Melanogaster, despite the phylogenetic distance between these target genomes and the informant genome. We compared these results with predictions of CDS density yielded by a selection of current gene-finding packages. These often performed extremely well, but required detailed organism-specific training data that is not required to train the topological pressure, and is not typically available for novel genomes.

We showed that the topological pressure defines a probability measure which can distinguish between segments of human intron and exon sequences of length between 750 and 5,000 bp. Finally, we established the theoretical basis for our results, adapting ideas and results from ergodic theory.

## References

Akashi H (2001) Gene expression and molecular evolution. Curr Opin Genet Dev 11(6):660–666

Baladi V (2000) Positive transfer operators and decay of correlations, vol 16. World Scientific, Singapore

Berná L, Chaurasia A, Angelini C, Federico C, Saccone S, D'Onofrio G (2012) The footpring of metabolism in the organization of mammalian genomes. BMC Bioinform 13(174):1–13

Bialek W, Cavagna A, Giardina I, Mora T, Silvestri E, Viale M, Walczak A (2012) Statistical mechanics for natural flocks of birds. PNAS 109:4786–4791

Blanco E, Parra G, Guigó R (2002) Using geneid to identify genes, current protocols in bioinformatics, vol 1. John Wiley & Sons Inc., New York

Bowen R (1975) Equilibrium states and the ergodic theory of Anosov diffeomorphisms, lecture notes in mathematics, vol 470. Springer-Verlag, Berlin

Burge C, Karlin S (1997) Prediction of complete gene structures in human genomic DNA. J Mol Biol 268:78–94

Carter D, Durbin R (2006) Vertebrate gene finding from multiple-species alignments using a two-level strategy. Genome Biol 7(1):S6.1–12

Comeron JM, Aguadé M (1998) An evaluation of measures of synonymous codon usage bias. J Mol Evol 47(3):268–274

Creanza TM, Horner DS, D'Addabbo A, Maglietta R, Mignone F, Ancona N, Pesole G (2009) Statistical assessment of discriminative features for protein-coding and non coding cross-species conserved sequence elements. BMC Bioinform 10(Suppl 6):S2. doi:10.1186/1471-2105-10-S6-S2

Durbin R, Eddy S, Krogh A, Mithcison G (1998) Biological sequence analysis: probabilistic models of proteins and nucleic acids. Cambridge University Press, Cambridge, MA

Erayman M, Sandhu D, Sidhu D, Dilbirligi M, Baenziger PS, Gill KS (2004) Demarcating the gene-rich regions of the wheat genome. Nucleic Acids Res 32(12):3546–3565

Fickett JW, Tung CS (1992) Assessment of protein coding measures. Nucleic Acids Res 20(24):6441–6450

Flicek P (2007) Gene prediction: compare and CONTRAST. Genome Biol 8(233):233.1–233.3. doi:10.1186/gb-2007-8-12-233

Gao F, Zhang CT (2004) Comparison of various algorithms for recognizing short coding sequences of human genes. Bioinformatics 20(5):673–681. doi:10.1093/bioinformatics/btg467

Gheorghiciuc I, Ward M (2008) On Correlation polynomials and subword complexity. DMTCS proceedings, pp 1–18

Giogo R, Reese M (2005) EGASP: collaboration through competition to find human genes. Nature Methods 2:575–577

Graves J (2006) Sex chromosome specialization and degeneration in mammals. Cell 124(5):901–914

Guigó R, Fickett JW (1995) Distinctive sequence features in protein coding genic non-coding, and intergenic human DNA. J Mol Biol 253(1):51–60. doi:10.1006/jmbi.1995.0535

Guig R, Flicek P, Abril J, Reymond A, Lagarde J, Denoeud F, Antonarakis S, Ashburner M, Bajic V, Birney E, Castelo R, Eyras E, Ucla C, Gingeras T, Hubbard T, Lewis S, Reese M (2006) EGASP: the human ENCODE genome annotation assessment project. Genome Biol 7(Suppl 1):S2.1–S2.31

Haussler D, O'Brien S, Ryder O, Barker F, Clamp M, Crawford A, Hanner R, Hanotte O, Johnson W, McGuire J, Miller W, Murphy R, Murphy W, Sheldon F, Sinervo B, Venkatesh B, Wiley E, Allendorf F, Baker S, Bernardi G, Brenner S, Cracraft J, Diekhans M, Edwards S, Estes J, Gaubert P, Graphodatsky A, Green E, Hebert P, Helgen K, Kessing B, Kingsley D, Lewin H, Luikart G, Martelli P, Nguyen N, Orti G, Pike B, Rawson D, Schuster S, Seunez H, Shaffer H, Springer M, Stuart J, Teeling E, Vrijenhoek R, Ward R, Wayne R, Williams T, Wolfe N, Zhang YP (2009) Genome10K: a proposal to obtian whole-genome sequence for 10,000 vertebrate species. J Hered 100(6):659–674

Karlin S, Mrázek J, Campbell A (1998) Codon usages in different gene classes of the Escherichia coli genome. Mol Microbiol 29(6):1341–1355

Karolchik D, Hinrichs A, Furey T, Roskin K, Sugnet C, Haussler D, Kent W (2004) The UCSC table browser data retrieval tool. Nucleic Acids Res 164:D493–D496

Korf I (2004) Gene finding in novel genomes. BMC Bioinform 5(59):9

Koslicki D (2011) Topological entropy of DNA sequences. Bioinformatics 27(8):1061–1067. doi:10.1093/bioinformatics/btr077

Kowalski J, Waga W, Zawierta M, Cebrat S (2009) Phase transition in the genome evolution favors non-random distribution of genes on chromosomes. Int J Mod Phys C 20(08):1299–1309

Ksiazkiewics M, Wyrwa K, Szxzepaniak A, Rychel S, Majcherkiewics K, Przysiecka L, Karlowski W, B W, Naganowska B (2013) Comparative genomics of lupinus angustifolius gene-righ regions: BAC library exploration, genetic mapping and cytogenetics. BMC Genomics 14(79):1–16

Kvikstad E, Tyekucheva S, Chiaromonte F, Makova K (2007) A macaque's-eye view of human insertions and deletions: differences in mechanisms. PLoS Comput Biol 3(9):1772–1782. doi:10.1371/journal.pcbi.0030176

Lin MF, Deoras AN, Rasmussen MD, Kellis M (2008), Performance and scalability of discriminative metrics for comparative gene identification in 12 Drosophila genomes. PLoS Comp Biol 4(4):e1000067, Doi:10.1371/journal.pcbi.1000067

Lin MF, Jungreis I, Kellis M (2011) PhyloCSF: a comparative genomics method to distinguish protein coding and non-coding regions. Bioinformatics 27(13):i275–i282. doi:10.1093/bioinformatics/btr209

Lukashin AV, Borodovsky M (1998) GeneMark.hmm: new solutions for gene finding. Nucleic Acids Res 26:1107–1115

Mackiewics D, Zawierta M, Waga W, Cebrat S (2010) Genome analyses and modelling the relationships between coding density, recombination rate and chromosome length. J Theor Biol 267(2):186–192

Makova K, Yang S, Chiaromonte F (2004) Insertions and deletions are male biased too: a whole-genome analysis in rodents. Genome Res 14(4):567–573. doi:10.1101/gr.1971104.autosome

MATLAB (2012) version 8.0 (R2012b) The MathWorks Inc., Natick, Massachusetts

Mora T, Walczak A, Bialek W, Callan CJ (2010) Maximum entropy models for antibody diversity. PNAS 107(12):5405–5410

Nelder J, Mead R (1965) A simplex method for function minimization. Comput J 7(4):308

Parry W, Pollicott M (1990) Zeta functions and the periodic orbit structure of hyperbolic dynamics. No. 187–188 in Astérisque, Soc. Math. France

Parry W, Tuncel S (1982) Classification problems in ergodic theory, London Mathematical Society lecture note series, vol 67. Cambridge University Press, Cambridge, statistics: textbooks and monographs, 41

Picard R, Cook D (1984) Cross-validation of regression models. J Am Stat Assoc 79:575–583

Saeys Y, Inza I, Larrañaga P (2007) A review of feature selection techniques in bioinformatics. Bioinformatics 23(19):2507–2517. doi:10.1093/bioinformatics/btm344

Salzburger W, Steinke D, Braasch I, Meyer A (2009) Genome desertification in Eutherians: can gene deserts explain the uneven distribution of genes in placental mammalian genomes? J Mol Evol 69:207–216

Schneidman E, Berry JI, Segev R, Bialek W (2006) Weak pairwise correlations imply strongly correlated network states in a neural population. Nature 440:1007–1012

Tkacik G, Schneidman E, Berry MJI, Bialek W (2006) Ising models for networks of real neurons. eprint arXiv:q-bio/0611072 arXiv:q-bio/0611072

van Baren MJ, Koebbe BC, Brent MR (2007) Using n-scan or twinscan to predict gene structures in genomic dna sequences. In: Boyle A (ed) Current protocols in bioinformatics. John Wiley & Sons, New York

Varshney R, Gross I, U H, Siefken R, Prasad M, Stein N, Langridge P, Altschmied L, Graner A (2006) Genetic mapping and BAC assignment of EST-derived SSR markers shows non-uniform distribution of genes in the barley genome. Theor Appl Genet 113:239–250

Walters P (1982) An introduction to ergodic theory, graduate texts in mathematics, vol 79. Springer, New York

Washietl S, Findeiss S, Müller SA, Kalkhof S, von Bergen M, Hofacker IL, Stadler PF, Goldman N (2011) RNAcode: robust discrimination of coding and noncoding regions in comparative sequence data. RNA 17(4):578–594. doi:10.1261/rna.2536111

Wilson M, Makova K (2009a) Evolution and survival on eutherian sex chromosomes. PLoS Genet 5(7):11. doi:10.1371/journal.pgen.1000568

Wilson M, Makova K (2009b) Genomic analyses of sex chromosome evolution. Annu Rev Genomics Hum Genet 10:333–354. doi:10.1146/annurev-genom-082908-150105

Yandell M, Ence D (2012) A beginner's guide to eukaryotic genome annotation. Nat Rev 13:329–342