Annals of Mathematics and Computer Science ISSN: 2789-7206
Vol 35 (2026) 1-28 https://doi.org/10.56947/amcs.v35.852

# MARKOV AND HIDDEN MARKOV MODELS FOR GENOMIC SEQUENCE CLASSIFICATION

ALI SOULEYMANE DABYE[1], DOUDOU DIAKHATE[2] AND MAMADOU MOMAR FALL[2*]

Abstract. The rapid growth of genomic data generated by high-throughput sequencing technologies has created significant challenges for statistical modeling and sequence classification. In this paper, we investigate genomic sequence classification using probabilistic and machine learning approaches based on Markov chains, Hidden Markov Models (HMMs), and Support Vector Machines (SVMs). Markov and Hidden Markov models are employed to capture local nucleotide dependencies and latent biological structures associated with coding and non-coding regions. Building upon these models, we introduce a hybrid HMM–SVM framework that combines generative likelihood-based features, hidden-state representations, and biologically interpretable compositional descriptors. Experimental results on genomic data demonstrate that the proposed hybrid approach substantially improves classification performance while maintaining biological interpretability.

*Keywords.* Markov chains, Hidden Markov models, Support Vector Machines, hybrid HMM–SVM framework, genomic sequence classification, DNA sequence analysis, k-mer representations.
*2020 Mathematics Subject Classification.* Primary 60J10, 62M05, 68T10; Secondary 92D20, 62H30.

## 1. Introduction

Molecular biology has undergone a major transformation with the advent of high-throughput sequencing technologies, propelling genomics into the era of Big Data. These technologies now make it possible to rapidly generate massive volumes of genomic data, whose organization and structuring rely on reference databases such as GenBank [4] and on rigorous taxonomic classification systems [10]. This massive accumulation of data raises a fundamental challenge for modern bioinformatics: the automatic annotation of genomes. In particular, the accurate identification of functional regions—especially the discrimination between coding regions (CDS) and non-coding regions (Non-CDS)—within sequences that

*Date*: Received: May 21, 2026; Revised: Jun 17, 2026; Accepted: Jun 24, 2026.
* Corresponding author

may span millions of nucleotides constitutes a central issue in computational genomics. This problem has been extensively studied in the literature, notably by Khodaei et al. [12], as well as in the methodological works of Yandell and Ence [20] and the evolutionary approaches proposed by Mertsch and Stanke [15].

Faced with this explosion of data, one of the major challenges lies in the ability to extract relevant information from complex sequences characterized by local dependencies and rich structural organization. In this context, probabilistic and statistical approaches have emerged as fundamental tools, providing a rigorous mathematical framework for modeling nucleotide dependencies and analyzing the structural properties of genomes [3, 17].

The statistical analysis of biological sequences is rooted in a scientific tradition initiated by the pioneering works of Blaisdell [5] and Brendel et al. [6], who demonstrated that nucleotide sequences are not random but exhibit significant local dependencies. These foundational results were further developed and formalized by Churchill [7], who introduced stochastic models capable of capturing the structural heterogeneity of DNA sequences. Within this framework, Markov chains have become a central tool for modeling biological sequences, with important contributions addressing both their estimation and statistical properties [2, 19].

From a methodological perspective, Markovian modeling relies on the estimation of transition probabilities, typically carried out via maximum likelihood, as shown by Avery and Henderson [2]. This step forms the foundation of statistical inference in this framework and ensures consistent parameter estimation from observed data.

Beyond parameter estimation, model selection and evaluation play a crucial role. The goodness-of-fit of Markov models can be assessed using information-theoretic criteria such as the Kullback–Leibler divergence and perplexity, as studied by Kulkarni [13] and Abbas et al. [1]. These measures make it possible to quantify the ability of models to accurately represent sequence distributions and to capture the statistical signatures specific to different genomic regions. They thus provide a unified framework linking estimation, model selection, and predictive performance.

A major advancement in this field is the introduction of Hidden Markov Models (HMMs), which allow for the explicit representation of latent structures underlying observed sequences. Initially formalized by Rabiner [18], these models have been extensively developed in the bioinformatics context, notably by Durbin et al. [8] and Eddy [9]. HMMs are now a standard probabilistic framework for biological sequence analysis.

In particular, HMMs play a key role in tasks such as genome annotation [20] and coding region detection [15]. They have also been widely applied in other domains, especially in natural language processing, where they serve as a fundamental tool for modeling sequential data [14].

In the field of genomic sequence classification, Markovian approaches and their extensions have demonstrated their effectiveness in discriminating between different biological classes. Recent studies, such as those by Khodaei et al. [12] and

Yoon [21], confirm the relevance of probabilistic models, highlighting their ability to capture statistical signatures specific to coding and non-coding regions.

Recent contributions in the mathematical and statistical literature have also highlighted the growing importance of statistical inference, goodness-of-fit procedures, and stochastic modeling in modern applied mathematics [11].

Despite these advances, several challenges remain. The performance of Markov models strongly depends on the chosen representation and model complexity. At the same time, modern machine learning methods, such as support vector machines and decision trees, have demonstrated strong predictive capabilities in a wide range of classification problems. Recent theoretical developments have also emphasized the impact of model complexity and over-parameterization on learning performance and generalization properties [16]. Nevertheless, these approaches generally rely on vector-based representations that may overlook the intrinsic sequential structure of biological data.

This tension between probabilistic modeling and supervised learning highlights the need for hybrid approaches. In this context, it is essential to assess to what extent Hidden Markov Models can complement or outperform classical methods when applied to real genomic data. Furthermore, the choice of sequence encoding (mononucleotide versus dinucleotide) remains a key factor influencing model performance.

In this paper, we propose a unified framework for DNA sequence classification that integrates Markov chains, Hidden Markov Models, and supervised learning methods. This framework provides a principled approach for modeling sequential dependencies and is further extended to incorporate latent structures, thereby enhancing the discrimination between coding and non-coding regions.

We then conduct an empirical evaluation of the proposed models using standard performance metrics such as accuracy and F1-score.

The main contribution of this work is the development of a hybrid HMM–SVM framework that combines probabilistic sequence modeling and discriminative learning within a unified representation space.

Unlike traditional HMM-based classifiers that rely solely on likelihood comparisons, the proposed methodology integrates normalized likelihood scores, latent-state occupancy statistics, entropy-based uncertainty measures, biologically interpretable compositional descriptors, and multi-scale k-mer representations. These heterogeneous sources of information are combined through a linear Support Vector Machine, yielding a classification framework that is both highly accurate and biologically interpretable.

Beyond the empirical study, the proposed framework illustrates how generative probabilistic models and discriminative machine learning techniques can be integrated in a principled manner for genomic sequence analysis.

The remainder of the paper is organized as follows. Section 2 presents Markov chain models for genomic sequences. Section 3 introduces Hidden Markov Models and the associated inference procedures. Section 4 is devoted to the empirical study. Section 5 presents the proposed hybrid HMM–SVM framework. Section 6 discusses the methodological implications, limitations, and future research directions. Finally, Section 7 concludes the paper.

## 2. Markov Chain Models for Genomic Sequences

2.1. **DNA Sequences as Markov Chains.** Markov chains constitute a fundamental probabilistic framework for modeling local dependencies in symbolic sequences. They are widely used in bioinformatics to represent the statistical structure of genomic sequences, where successive nucleotides exhibit non-trivial dependencies related to biological constraints such as compositional biases, mutational context effects, and certain functional structures of the genome.

Within this framework, a DNA strand is represented as a sequence of random variables defined over the finite alphabet $\Sigma = \{A, C, G, T\}$.

Let $(X_n)_{n\geq 0}$ denote a stochastic process defined on the probability space $(\Omega, \mathcal{F}, \mathbb{P})$, with state space $\Sigma$. The process $(X_n)$ is a homogeneous Markov chain if, for all $n \geq 0$,

$$\mathbb{P}(X_{n+1} = j \mid X_n = i, \ldots, X_0 = i_0) = \mathbb{P}(X_{n+1} = j \mid X_n = i).$$

For all $i, j \in \Sigma$, let

$$P_{ij} = \mathbb{P}(X_{n+1} = j \mid X_n = i).$$

Then $P = (P_{ij})$ defines a stochastic matrix. Multi-step transitions satisfy the Chapman–Kolmogorov equations

$$(P^{n+m})_{ij} = \sum_{k\in\Sigma} (P^n)_{ik}(P^m)_{kj}.$$

For clarity, Figures 1 and 2 illustrate the probabilistic structures underlying the Markov and hidden Markov models considered in this work.

These schematic representations emphasize the transition structure of Markov chains and the role of latent states in Hidden Markov Models, both of which are central to the probabilistic framework developed in this study.

If the chain is irreducible and aperiodic on a finite state space, then it admits a unique stationary distribution $\pi$, characterized by

$$\pi P = \pi, \qquad \lim_{n\to\infty} (P^n)_{ij} = \pi_j,$$

where $\pi_j$ denotes the $j$-th component of the stationary distribution. This convergence shows that the distribution of $X_n$ converges to $\pi$, independently of the initial state $i$.

This result ensures the asymptotic stability of empirical frequencies observed in biological sequences. Moreover, the classification of states (recurrent or transient) allows for a finer characterization of the long-term behavior of the process.

2.2. **Parameter Estimation.** The likelihood of an observed trajectory $(x_0, \ldots, x_n)$ is given by

$$\mathcal{L}(P) = \mathbb{P}(X_0 = x_0) \prod_{k=0}^{n-1} P_{x_k x_{k+1}},$$

or, in aggregated form,

$$\log \mathcal{L}(P) = \sum_{i,j} N_{ij} \log P_{ij},$$

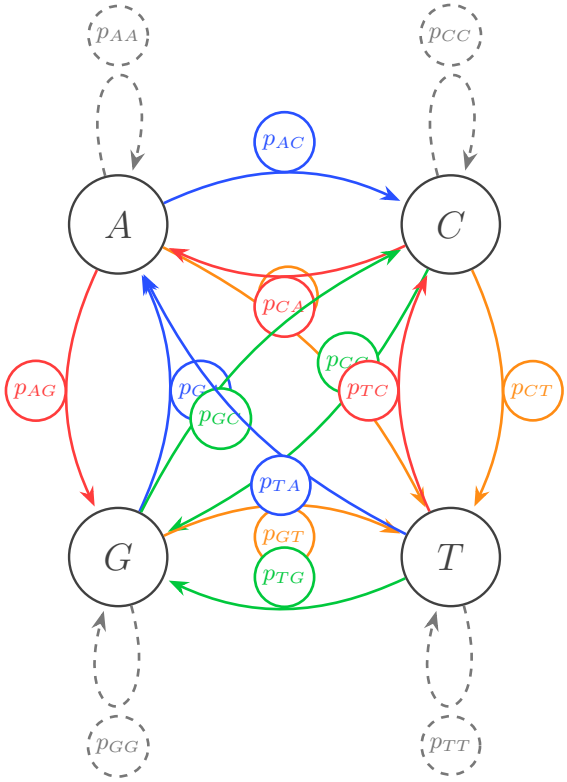


*Figure 1. Markov chain representation of nucleotide transitions on the alphabet $\Sigma = \{A, C, G, T\}$. Each edge corresponds to a transition probability $p_{ij} = \mathbb{P}(X_{n+1} = j \mid X_n = i)$.*

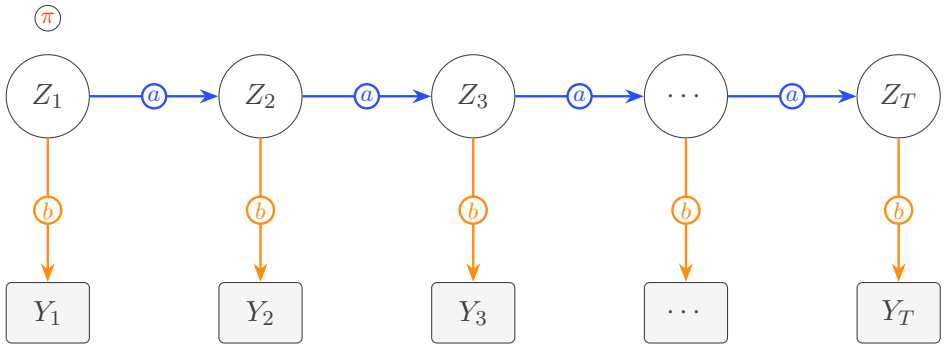


*Figure 2. Structure of a Hidden Markov Model. The hidden process $(Z_t)$ is a Markov chain with transition matrix $A = (a_{ij})$, while each observation $Y_t$ is generated conditionally on $Z_t$ according to the emission distribution $B$.*

where $N_{ij}$ is the number of transitions from $i$ to $j$. The maximum likelihood estimator is then given by

$$\widehat{P}_{ij} = \frac{N_{ij}}{N_i}, \qquad N_i = \sum_j N_{ij}.$$

Under ergodicity conditions, this estimator is consistent and asymptotically normal.

2.3. **Model Selection and Information Criteria.** In many genomic applications, local dependencies are better described by higher-order Markov chains, defined by

$$\mathbb{P}(X_n = x_n \mid X_{n-1}, \dots, X_0) = \mathbb{P}(X_n = x_n \mid X_{n-1}, \dots, X_{n-k}),$$

which makes it possible to capture complex biological patterns such as dinucleotide dependencies or GC-content biases.

To determine the appropriate model, and in particular the Markov order, one may use likelihood ratio tests together with information criteria such as

$$\mathrm{AIC} = -2\log \mathcal{L}(\widehat{P}) + 2k, \qquad \mathrm{BIC} = -2\log \mathcal{L}(\widehat{P}) + k\log n,$$

where $k$ denotes the number of free parameters and $n$ the sample size. These criteria incorporate a penalty term to balance goodness-of-fit and model complexity.

To quantify the discrepancy between probabilistic models, one may resort to information-based measures. A fundamental example is the Kullback–Leibler divergence, given by

$$D_{\mathrm{KL}}(P\|Q) = \sum_x P(x)\log\frac{P(x)}{Q(x)},$$

while the entropy of a stationary Markov chain is given by

$$H = -\sum_{i,j} \pi_i P_{ij} \log P_{ij}.$$

A practical performance measure is perplexity,

$$\text{Perplexity} = \exp\left(-\frac{1}{n}\log P(x_{1:n})\right),$$

which quantifies the predictive capacity of the model.

2.4. **Markov-Based Classification.** Beyond their descriptive capabilities, Markov chain models can be employed for classification purposes. Consider two classes of genomic sequences, for instance coding regions (CDS) and non-coding regions (Non-CDS), represented by two distinct Markov models with transition matrices $P^{(1)}$ and $P^{(2)}$.

Given an observed sequence $x_{1:n}$, the classification decision may be based on the likelihood ratio

$$\log\frac{\mathcal{L}_1(x_{1:n})}{\mathcal{L}_2(x_{1:n})} = \sum_{k=1}^{n-1}\log\frac{P^{(1)}_{x_k x_{k+1}}}{P^{(2)}_{x_k x_{k+1}}},$$

where $\mathcal{L}_1(x_{1:n})$ and $\mathcal{L}_2(x_{1:n})$ denote the likelihoods of the observed sequence under the two competing Markov models. The sequence is assigned to the class associated with the largest likelihood value.

This classification rule exploits the local transition structure of nucleotide sequences and provides a natural probabilistic framework for discriminating between different genomic regions. In particular, it allows the identification of statistical signatures associated with coding and non-coding sequences through their transition patterns.

More generally, DNA sequences may be viewed as realizations of stochastic processes taking values in the alphabet

$$\Sigma = \{A, C, G, T\}.$$

Within this framework, the objective is not only to describe local nucleotide dependencies but also to exploit them for statistical inference, prediction, and classification.

The Markovian representation captures transition preferences between nucleotides, including enrichments or depletions of specific nucleotide combinations, which often constitute characteristic molecular signatures of distinct biological regions.

The interest of this approach is twofold. First, it provides a rigorous analytical framework for studying the long-run behavior of genomic sequences through the stationary distribution $\pi$. When the chain is irreducible and aperiodic, this distribution describes the equilibrium frequencies of nucleotides and contributes to the biological interpretation of sequence composition. Second, it enables direct parameter estimation from observed data through the maximum likelihood estimator $\widehat{P}_{ij}$, thereby providing a statistically sound basis for prediction and classification.

Despite these advantages, observed Markov chains remain limited when the organization of genomic sequences depends on latent biological mechanisms that cannot be directly inferred from nucleotide transitions alone. In particular, functional regions such as coding and non-coding segments may exhibit hidden structural patterns that are not fully captured by observable states. These limitations motivate the introduction of Hidden Markov Models, which enrich the Markovian framework by incorporating an underlying latent state process. Such models provide a more flexible representation of biological sequences and play a central role in modern genomic classification and annotation.

## 3. Hidden Markov Models for Genomic Classification

3.1. **Hidden State Representation.** DNA sequences are naturally viewed as realizations of stochastic processes taking values in the alphabet $\Sigma = \{A, C, G, T\}$.

While Markov chains provide an effective framework for modeling local nucleotide dependencies, they may not fully describe the underlying biological organization of genomic sequences.

In particular, coding and non-coding regions often exhibit distinct statistical properties that cannot always be explained solely through observable nucleotide transitions. The existence of latent biological mechanisms suggests the need for a richer probabilistic representation.

Hidden Markov Models (HMMs) address this limitation by introducing an unobserved state process that governs the generation of nucleotides. In the genomic

context, these hidden states may represent functional categories such as coding and non-coding regions, while the observed symbols correspond to emitted nucleotides.

The introduction of latent states considerably extends the descriptive power of Markovian models and provides a flexible framework for segmentation, annotation, and classification of genomic sequences.

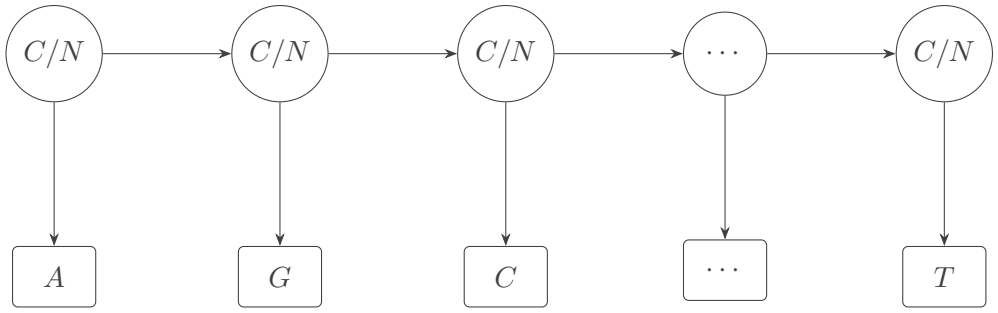


*Figure 3. Biological interpretation of a Hidden Markov Model for genomic sequences. Hidden states may represent functional classes such as coding (C) and non-coding (N) regions, while observations correspond to emitted nucleotides.*

Let

$$\mathcal{Z} = \{1, \ldots, K\}$$

denote the hidden state space.

An HMM is defined through the triplet $(A, B, \pi)$, where $A = (a_{ij})$ is the transition matrix given by

$$a_{ij} = \mathbb{P}(Z_{t+1} = j \mid Z_t = i), \qquad i, j \in \mathcal{Z}.$$

The component $B = (b_j(y))$ corresponds to the emission distribution, defined by

$$b_j(y) = \mathbb{P}(Y_t = y \mid Z_t = j), \qquad j \in \mathcal{Z},\ y \in \Sigma,$$

where $\Sigma = \{A, C, G, T\}$ denotes the observation alphabet.

Finally, the vector

$$\pi = (\pi_i)_{i \in \mathcal{Z}}$$

characterizes the initial distribution of the hidden Markov chain, where

$$\pi_i = \mathbb{P}(Z_1 = i), \qquad i \in \mathcal{Z}.$$

In genomic applications, hidden states may represent functional categories such as coding regions, non-coding regions, promoters, or regulatory elements. The observations correspond to emitted nucleotides, whose distributions vary according to the underlying biological state.

3.2. **Likelihood, Inference and Decoding.** Let

$$Y_{1:T} = (Y_1, \ldots, Y_T)$$

denote the observed nucleotide sequence and

$$Z_{1:T} = (Z_1, \ldots, Z_T)$$

the corresponding sequence of hidden states, where $Y_t \in \Sigma$ and $Z_t \in \mathcal{Z}$.

For a Hidden Markov Model characterized by the parameter triplet $\theta = (A, B, \pi)$, the joint distribution of observations and hidden states is

$$\mathbb{P}(y_{1:T}, z_{1:T}) = \pi_{z_1}\, b_{z_1}(y_1) \prod_{t=2}^{T} a_{z_{t-1} z_t}\, b_{z_t}(y_t).$$

The likelihood of the observed sequence is obtained by summing over all possible hidden-state configurations:

$$L(\theta) = \mathbb{P}_\theta(Y_{1:T} = y_{1:T}) = \sum_{z_{1:T} \in \mathcal{Z}^T} \mathbb{P}_\theta(y_{1:T}, z_{1:T}),$$

where

$$\theta = (A, B, \pi)$$

denotes the parameter vector of the Hidden Markov Model.

Direct computation of the likelihood is generally infeasible because the number of hidden-state sequences grows exponentially with $T$. Consequently, inference relies on efficient dynamic programming algorithms.

The forward probabilities are defined by

$$\alpha_t(i) = \mathbb{P}(Y_{1:t} = y_{1:t},\, Z_t = i), \qquad i \in \mathcal{Z},$$

and satisfy the recursion

$$\alpha_{t+1}(j) = \left( \sum_{i \in \mathcal{Z}} \alpha_t(i) a_{ij} \right) b_j(y_{t+1}), \qquad j \in \mathcal{Z}.$$

Similarly, the backward probabilities are defined by

$$\beta_t(i) = \mathbb{P}(Y_{t+1:T} = y_{t+1:T} \mid Z_t = i), \qquad i \in \mathcal{Z},$$

and satisfy

$$\beta_t(i) = \sum_{j \in \mathcal{Z}} a_{ij}\, b_j(y_{t+1})\, \beta_{t+1}(j), \qquad i \in \mathcal{Z}.$$

The forward–backward procedure provides an efficient computation of the likelihood and posterior state probabilities, which are central to parameter estimation and sequence analysis.

To recover the most probable sequence of hidden states, one uses the Viterbi algorithm, which computes

$$\widehat{z}_{1:T} = \arg\max_{z_{1:T}} \mathbb{P}(z_{1:T} \mid y_{1:T}).$$

In the genomic context, this decoding procedure allows the segmentation of DNA sequences into biologically meaningful regions, such as coding and non-coding segments.

3.3. **Parameter Estimation by the Baum–Welch Algorithm.** The parameters are estimated via the EM algorithm (Baum–Welch) using the quantities

$$\gamma_t(i) = \mathbb{P}(Z_t = i \mid Y_{1:T}), \qquad \xi_t(i,j) = \mathbb{P}(Z_t = i, Z_{t+1} = j \mid Y_{1:T}),$$

ensuring a monotonic increase of the likelihood.

The Baum–Welch algorithm is a particular instance of the Expectation–Maximization (EM) procedure.

During the E-step, the posterior probabilities

$$\gamma_t(i) = \mathbb{P}(Z_t = i \mid Y_{1:T})$$

and

$$\xi_t(i,j) = \mathbb{P}(Z_t = i, Z_{t+1} = j \mid Y_{1:T})$$

are computed using the forward–backward recursions.

During the M-step, the transition and emission probabilities are updated according to

$$\widehat{a}_{ij} = \frac{\sum_{t=1}^{T-1} \xi_t(i,j)}{\sum_{t=1}^{T-1} \gamma_t(i)},$$

and

$$\widehat{b}_j(y) = \frac{\sum_{t:Y_t=y} \gamma_t(j)}{\sum_{t=1}^{T} \gamma_t(j)}.$$

These updates maximize the expected complete-data log-likelihood and guarantee a monotonic increase of the observed-data likelihood.

3.4. **HMM-Based Genomic Classification.** To model not only local dependencies but also the existence of functional structures that are not directly observable, Hidden Markov Models constitute a particularly relevant generalization. In this framework, the observed sequence of nucleotides is assumed to be emitted by a latent sequence of states representing, for example, coding and non-coding regions. The joint probability of a sequence of states $z_{1:T}$ and a sequence of observations $y_{1:T}$ is given by $\mathbb{P}(y_{1:T}, z_{1:T})$.

In a classification perspective, probabilistic inference consists in comparing the compatibility of a sequence with several competing models. If, for example, one has a model learned on coding regions $M_{CDS}$ and another learned on non-coding regions $M_{Non-CDS}$, the decision can be formulated based on the maximum likelihood criterion:

Let $S = y_{1:T}$ denote an observed DNA sequence. The classification rule is based on the maximum log-likelihood criterion

$$\widehat{c}(S) = \arg \max_{c \in \{\text{CDS,Non–CDS}\}} \log P(S \mid M_c).$$

A sequence is thus assigned to the class whose model best explains the observations. This decision strategy makes it possible to transform probabilistic modeling into an operational prediction tool, while preserving a clear interpretation of the underlying statistical mechanisms.

Finally, Hidden Markov Models provide a natural probabilistic framework for representing latent biological structures underlying genomic sequences. Hidden states may be associated with functional categories such as coding and non-coding regions, while emission distributions characterize the statistical properties of observed nucleotides. Through probabilistic inference and decoding procedures, HMMs enable the segmentation and annotation of genomic sequences.

The combination of Markov chains and Hidden Markov Models therefore provides a unified framework for modeling, inference, and classification. The practical effectiveness of this framework will be assessed in the next section through an empirical study based on real genomic data, using standard performance measures such as accuracy and F1-score.

## 4. Empirical Study

In this section, we implement the theoretical concepts presented previously to address a concrete problem of genomic region classification. More specifically, we focus on distinguishing between coding and non-coding regions.

### 4.1. **Genomic Data Description.**

We used genomic data from the organism *Escherichia coli* strain K-12 substr. MG1655, a bacterial model widely studied in molecular biology and genomics. The complete genome of this strain is available through the public database **NCBI** (National Center for Biotechnology Information) [4], under the accession number `NC_000913.3`.

The file used in this study is in GenBank format, which allows the combination of the nucleotide sequence with detailed annotation describing genes, coding regions (CDS), RNAs, promoters, and other functional elements. This genome, of circular double-stranded DNA type, has a total length of 4 641 652 base pairs. It contains approximately 4 870 annotated genes, of which nearly 4 315 correspond to coding regions (CDS). In addition, more than 300 RNAs have been identified, including ribosomal RNAs (rRNA), transfer RNAs (tRNA), as well as other non-coding RNAs (ncRNA).

This richness of annotation allows a precise distinction between coding and non-coding regions, which is essential in our classification approach. These data are automatically extracted using the `Biopython` library, which enables automated reading and processing of GenBank files.

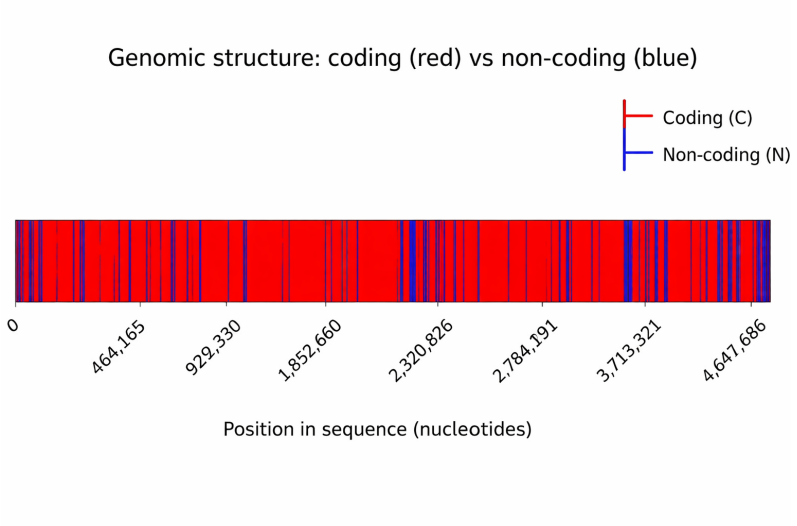


*Figure 4. Genomic structure*

*Table 1. Detailed mononucleotide statistics of CDS and Non-CDS regions*

| **Base** | **CDS (4 025 874)** | **Non-CDS (646 148)** | **Difference** |
|---|---|---|---|
| A | 24.06% | 28.26% | -4.20% |
| C | 24.55% | 21.69% | +2.86% |
| G | 27.34% | 22.08% | +5.26% |
| T | 24.04% | 27.97% | -3.93% |
| GC | 51.90% | 43.77% | +8.12% |
| AT | 48.10% | 56.23% | -8.13% |

The analysis of mononucleotide frequencies is an essential step to characterize sequence composition and highlight differences between coding and non-coding regions. In CDS regions, the base distribution is relatively balanced, with a slight predominance of G and C bases, leading to a GC content of 51.90%.

In contrast, non-coding regions show a marked enrichment in A and T bases, reaching more than 56% of the total content. This difference reflects distinct biological constraints: coding regions require a certain structural stability, favored by GC bases, whereas non-coding regions exhibit greater structural flexibility due to their AT richness.

These compositional differences constitute a discriminative molecular signature between CDS and Non-CDS regions, which can be exploited in classification models.

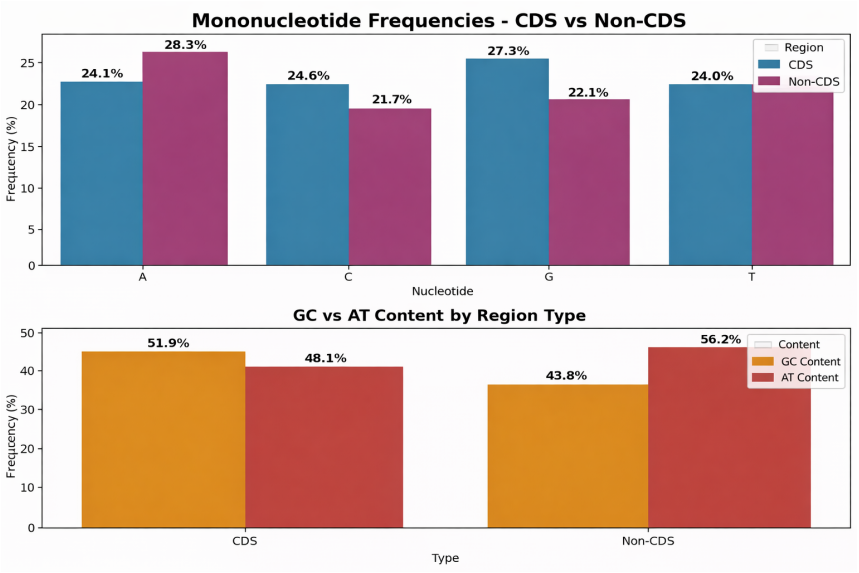


*Figure 5. Frequency distribution*

Beyond individual frequencies, the analysis of dinucleotides allows capturing local dependencies between successive nucleotides. Dinucleotides are not randomly distributed but reflect contextual constraints.

In coding regions, the dominant motifs are GC (8.65%), TG (8.14%), CG (7.85%), AA (7.44%), and GG (6.68%), confirming the GC richness. In particular, the overrepresentation of GC and CG motifs is notable.

In contrast, non-coding regions are dominated by dinucleotides AA (9.54%), TT (9.36%), AT (8.11%), TA (6.59%), and TG (6.46%), reflecting a strong dominance of A and T bases.

*Table 2. Detailed dinucleotide statistics of CDS and Non-CDS regions*

| Dinucleotide | CDS | Non-CDS | Difference |
|---|---|---|---|
| GC | 8.65% | 5.97% | +2.69% |
| CG | 7.85% | 5.18% | +2.66% |
| TT | 6.45% | 9.36% | -2.91% |
| TA | 4.23% | 6.59% | -2.37% |
| AA | 7.44% | 9.54% | -2.09% |

These differences highlight discriminative motifs, notably the enrichment of GC and CG in CDS, and TT, TA, and AA in Non-CDS regions.

These motifs constitute relevant variables for classification models. In HMMs, they allow better capture of sequential dependencies, while in supervised models (SVM), they can be incorporated as explanatory variables.

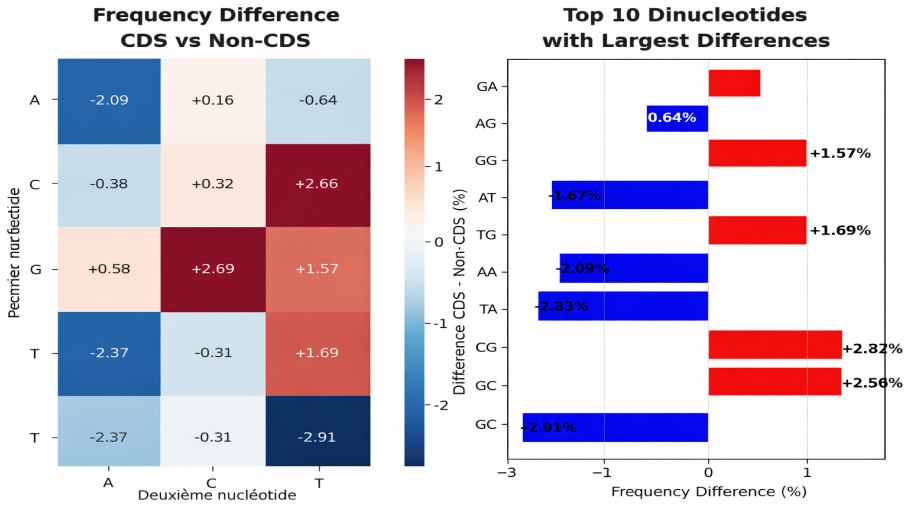


*Figure 6. Discriminative dinucleotides*

4.2. **Data Cleaning and Encoding.** The preliminary processing of DNA sequences constitutes a crucial step, as it directly affects the quality of the models developed subsequently. This phase consists in preparing raw genomic data to make it suitable for statistical analysis and machine learning.

The data cleaning process aims to retain only the information relevant to our study. More specifically, we keep coding regions (CDS) and non-coding regions (Non-CDS), while removing unnecessary elements such as redundant annotations, incomplete sequences, and ambiguous nucleotides.

Using the `Biopython` library, the following operations were performed: loading the GenBank file containing genomic annotations, extracting coding regions (CDS), inferring non-coding regions from the intervals between CDS, and filtering out sequences that are too short (length $< 30$ nucleotides), which may introduce noise.

Depending on the type of model used, specific transformations were then applied to the sequences.

In the case of Hidden Markov Models (HMMs), the sequences are encoded as discrete states. Each nucleotide from the alphabet $\Sigma = \{A, C, G, T\}$ is mapped to an integer, for example:

$$A \mapsto 0, \quad C \mapsto 1, \quad G \mapsto 2, \quad T \mapsto 3.$$

This encoding allows each sequence to be represented as a sequence of discrete observations, compatible with a categorical HMM.

For Support Vector Machines (SVMs), sequences cannot be used directly, as this type of model requires fixed-dimensional feature vectors. To address this constraint, sequences are transformed into numerical vectors using the *k-mer* approach, which is widely used in bioinformatics. This method consists of counting the occurrences of subsequences of length $k$, thereby providing a vector representation suitable for supervised learning.

Experimental protocol. For each class, the dataset was randomly divided into training and testing subsets using an 80%/20% split. Model performance was evaluated using accuracy, precision, recall, F1-score, and ROC-AUC metrics.

4.3. **Construction of HMM Models.** We constructed and trained two Hidden Markov Models (HMMs) to model coding and non-coding regions of DNA sequences, drawing in particular on the work of D. Mertsch and M. Stanke [15].

The objective is to assess the impact of sequence representation on the ability of probabilistic models to discriminate between coding (CDS) and non-coding (Non-CDS) regions.

Two encoding strategies were investigated: a mononucleotide encoding based on individual nucleotides, and a dinucleotide encoding based on successive nucleotide pairs.

In the first approach, each nucleotide

$$A,\ C,\ G,\ T$$

is mapped to an integer value. The resulting probabilistic model is a discrete HMM whose observations belong to a finite alphabet of size 4.

The hidden-state structure relies on six latent states, assumed to represent recurrent motifs or internal organizational patterns of genomic sequences. The choice of six hidden states provided a satisfactory compromise between model flexibility and estimation stability.

Training is carried out using two categories of genomic regions: coding segments (CDS) and non-coding segments (Non-CDS). Two distinct probabilistic models are therefore estimated, one for each class.

The estimated parameters consist of: the transition matrix between hidden states and the emission probability matrix associated with the observations.

Parameter estimation is performed by maximum likelihood using the Baum–Welch algorithm, which iteratively maximizes the likelihood of the observed sequences.

In the second approach, dinucleotide encoding is introduced in order to capture local dependencies between successive nucleotides. The observation alphabet then contains

$$16 = 4 \times 4$$

symbols corresponding to all possible nucleotide pairs.

The hidden-state architecture remains unchanged, with six latent states. However, the emission structure becomes richer, allowing a finer representation of short-range sequence regularities, at the cost of increased model complexity.

Training follows the same strategy as in the mononucleotide case, with separate estimation procedures for CDS and Non-CDS regions.

The partition of the data into training and test sets is adapted to the probabilistic framework. For each class, 80% of the sequences are used for training, while the remaining 20% are reserved for evaluation.

Consequently, two independent HMMs are constructed: one associated with coding regions and the other with non-coding regions.

During the testing phase, each sequence is evaluated under both competing models. The predicted class is determined according to the maximum log-likelihood criterion:

$$\widehat{c}(S) = \arg \max_{c \in \{\text{CDS}, \text{Non}-\text{CDS}\}} \log \mathbb{P}(S \mid \mathcal{M}_c),$$

where $\mathcal{M}_c$ denotes the HMM associated with class $c$.

The estimated transition structures reveal significant differences between coding and non-coding regions.

For coding regions, several transitions exhibit dominant probabilities, reflecting a highly structured sequential organization. Certain transitions appear almost systematic, which is consistent with the regularity of biological motifs observed in CDS regions.

In contrast, the transition matrices associated with non-coding regions display a more diffuse probabilistic structure. The transition probabilities are less concentrated and exhibit greater variability, indicating a more heterogeneous sequential organization and increased structural flexibility.

These differences confirm that Hidden Markov Models effectively capture the specific statistical characteristics of each type of genomic region, thereby providing a robust probabilistic framework for discriminating between coding and non-coding sequences.

## 4.4. **Baseline Models.**

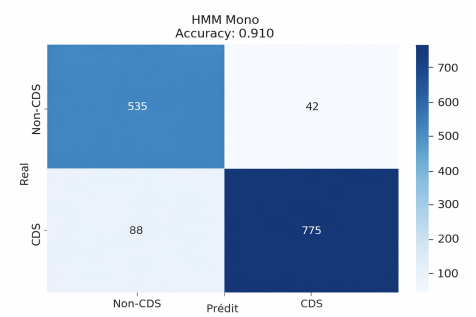


*Table 3. Classification report — Mononucleotide HMM*

| Class | Precision | Recall | F1-score | Support |
|---|---|---|---|---|
| Non-CDS | 0.86 | 0.93 | 0.89 | 577 |
| CDS | 0.95 | 0.90 | 0.92 | 863 |
| Accuracy | 0.91 (out of 1440) | | | |
| Macro avg | 0.90 | 0.91 | 0.91 | 1440 |
| Weighted avg | 0.91 | 0.91 | 0.91 | 1440 |

*Figure 7. Performance of the mononucleotide HMM model*

4.4.1. *Mononucleotide HMM Results.* The evaluation of the HMM model with mononucleotide encoding shows high performance, with an overall accuracy of 91%. This result confirms the relevance of the probabilistic approach for genomic region classification.

For non-coding regions, the model achieves a precision of 0.86 and a recall of 0.93, indicating a strong ability to correctly identify these regions (535 sequences correctly classified versus 42 errors).

For coding regions, the precision reaches 0.95 and the recall 0.90, resulting in an F1-score of 0.92. This means that 775 sequences out of 863 are correctly identified, with 88 classification errors.

The confusion matrix highlights a dominant diagonal, reflecting a good separation between classes. The errors are mainly due to CDS being classified as Non-CDS, suggesting a trade-off between sensitivity and specificity.

These results show that even with a simple encoding, the HMM effectively captures sequential dependencies.

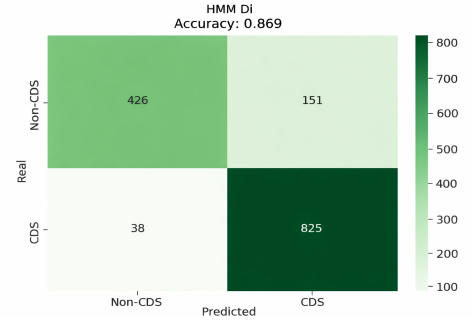


*Table 4. Classification report — Dinucleotide HMM*

| Class | Precision | Recall | F1-score | Support |
|---|---|---|---|---|
| Non-CDS | 0.92 | 0.74 | 0.82 | 577 |
| CDS | 0.85 | 0.96 | 0.90 | 863 |
| Accuracy | 0.87 (out of 1440) | | | |
| Macro avg | 0.88 | 0.85 | 0.86 | 1440 |
| Weighted avg | 0.87 | 0.87 | 0.87 | 1440 |

*Figure 8. Performance of the dinucleotide HMM model*

4.4.2. *Dinucleotide HMM Results.* The HMM model based on dinucleotide encoding allows capturing richer local dependencies. It achieves an overall accuracy of 87%, slightly lower than that of the mononucleotide model.

For Non-CDS, the precision is high (0.92), but the recall is lower (0.74), meaning that some non-coding sequences are poorly detected (426 correctly classified versus 151 errors).

For CDS, the model achieves a very high recall (0.96), but a lower precision (0.85), indicating a tendency to over-classify certain sequences as CDS.

The confusion matrix reveals an asymmetry: the errors mainly arise from Non-CDS classified as CDS (151 cases), compared to only 38 errors in the opposite direction.

These results show that the dinucleotide model better captures characteristic motifs of CDS, but at the cost of degraded performance on Non-CDS.

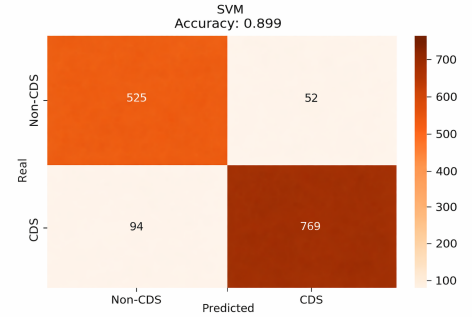


*Table 5. Classification report — SVM*

| Class | Precision | Recall | F1-score | Support |
|---|---|---|---|---|
| Non-CDS | 0.84 | 0.94 | 0.88 | 577 |
| CDS | 0.95 | 0.88 | 0.92 | 863 |
| Accuracy | 0.90 (out of 1440) | | | |
| Macro avg | 0.90 | 0.91 | 0.90 | 1440 |
| Weighted avg | 0.91 | 0.90 | 0.90 | 1440 |

*Figure 9. Performance of the SVM model*

4.4.3. *Support Vector Machine (SVM).* The SVM model, based on variables such as GC content and mono- and dinucleotide frequencies, achieves an accuracy of

90%. It provides good classification performance but relies on a vector representation of the data, without explicitly modeling sequential dependencies.

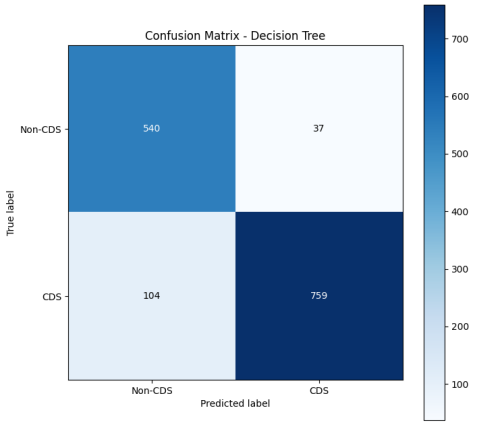


*Table 6. Classification report — Decision Tree*

| Class | Precision | Recall | F1-score | Support |
|---|---|---|---|---|
| Non-CDS | 0.84 | 0.94 | 0.88 | 577 |
| CDS | 0.95 | 0.88 | 0.92 | 863 |
| Accuracy | 0.90 (out of 1440) | | | |
| Macro avg | 0.90 | 0.91 | 0.90 | 1440 |
| Weighted avg | 0.91 | 0.90 | 0.90 | 1440 |

*Figure 10. Performance of the Decision Tree model*

4.4.4. *Decision Tree.*

4.4.5. *Decision Tree.* The Decision Tree classifier achieves competitive performance, with an overall accuracy of approximately 90%. Although its predictive performance remains slightly below that of the Hidden Markov Model and substantially below that of the proposed hybrid HMM–SVM framework, it offers an important advantage in terms of interpretability.

Unlike probabilistic sequence models, decision trees produce explicit classification rules that can be directly analyzed and visualized. The learned decision paths highlight biologically meaningful descriptors, including GC content, nucleotide composition, and specific dinucleotide motifs. These variables are known to be associated with functional differences between coding and non-coding genomic regions.

From a biological perspective, the resulting tree provides insight into the mechanisms driving classification decisions and facilitates the identification of sequence characteristics that contribute most strongly to class discrimination. This interpretability makes decision trees particularly attractive for exploratory genomic analyses, where understanding the relationship between sequence composition and biological function is often as important as predictive accuracy.

However, because decision trees rely on fixed-dimensional feature representations and hierarchical splitting rules, they do not explicitly model sequential dependencies or latent biological structures. Consequently, their ability to capture the complex organization of genomic sequences remains more limited than that of Hidden Markov Models or the proposed hybrid HMM–SVM framework.

## 5. Hybrid HMM–SVM Framework

In this section, we propose a hybrid framework for genomic sequence classification that explicitly combines sequential probabilistic modeling based on Hidden Markov Models (HMMs) with a discriminative decision mechanism provided by

a Support Vector Machine (SVM). The objective is to overcome the classical opposition between generative and discriminative models by constructing a unified representation in which HMMs extract latent structural information from genomic sequences, while the SVM learns an optimal decision boundary in the resulting feature space.

5.1. **Hybrid Feature Representation.** Let

$$x = (x_1, \dots, x_T)$$

be a genomic sequence of length $T$, and let

$$y \in \{-1, +1\}$$

denote its class label, corresponding respectively to non-coding and coding regions.

Within the standard generative framework, two class-specific Hidden Markov Models are estimated: a model $M_C$ associated with coding regions and a model $M_N$ associated with non-coding regions. Classification is then based on the comparison of the likelihoods

$$p(x \mid M_C) \qquad \text{and} \qquad p(x \mid M_N).$$

Although this approach is biologically meaningful, it remains limited from a discriminative perspective, since the final decision is often reduced to a scalar likelihood contrast.

In contrast, SVM-based approaches operate on fixed-dimensional feature vectors and generally provide a stronger separation between classes. However, they do not explicitly account for the sequential dependence structure or the latent organization of nucleotides within genomic sequences.

The proposed framework therefore seeks to combine the strengths of both approaches. Hidden Markov Models provide a representation that captures the sequential and latent structure of genomic data, while the Support Vector Machine exploits this representation to perform maximum-margin classification. In this way, the resulting hybrid model benefits simultaneously from the descriptive power of probabilistic sequence modeling and the discriminative efficiency of modern machine learning methods.

For a genomic sequence $x$, we first compute the normalized log-likelihoods associated with the two class-specific models:

$$\ell_C(x) = \frac{1}{T} \log p(x \mid M_C), \qquad \ell_N(x) = \frac{1}{T} \log p(x \mid M_N).$$

Normalization by sequence length ensures comparability across sequences of different sizes. We then define the generative contrast

$$\Delta\ell(x) = \ell_C(x) - \ell_N(x),$$

which quantifies the extent to which the sequence is better explained by the coding model than by the non-coding model.

However, in order to avoid reducing the sequential information to these global likelihood scores alone, we also exploit the posterior quantities produced by the

forward–backward algorithm. For each model, we consider the average hidden-state occupancies together with average entropy measures summarizing the dispersion of the posterior distributions along the sequence.

The former describe how a sequence activates the latent states of the Hidden Markov Model, whereas the latter quantify the degree of uncertainty associated with this activation. These quantities are particularly relevant in a biological context, since coding and non-coding regions may differ not only in their overall nucleotide composition but also in the stability of their internal sequential organization.

Consequently, the extracted HMM features provide a richer representation of genomic sequences, combining global likelihood information with latent structural characteristics inferred from the posterior state distributions.

In addition to the generative features described above, we incorporate a set of biologically interpretable compositional descriptors, denoted by

$$z(x).$$

This feature vector includes mononucleotide and dinucleotide frequencies, GC and AT contents, as well as other simple sequence composition indicators. The inclusion of these variables is motivated by their well-established biological relevance, particularly in distinguishing GC-rich regions from AT-rich regions.

The final representation of a sequence is defined through the hybrid embedding

$$\Phi(x) = \Big[\ell_C(x), \ell_N(x), \Delta\ell(x), \bar{\gamma}^{(C)}(x), \bar{\gamma}^{(N)}(x), H_C(x), H_N(x), z(x)\Big].$$

This representation possesses an important methodological advantage: it combines within a single feature space global generative evidence, latent sequential structure, and explicit compositional information. Consequently, it is neither a simple handcrafted projection of the sequence nor a purely probabilistic reduction based on a single likelihood score. Rather, it defines a mixed representation space that is simultaneously biologically interpretable and statistically discriminative.

From a machine learning perspective, this hybrid embedding allows the classifier to exploit complementary sources of information. The likelihood-based features summarize the global compatibility of a sequence with each probabilistic model, the posterior-state statistics characterize its latent sequential organization, and the compositional descriptors capture biologically meaningful patterns related to nucleotide content and local dependencies.

5.2. **Multi-Scale k-mer Extension.** The proposed hybrid representation is formulated within a multi-scale framework. More precisely, we consider two levels of granularity,

$$k = 1 \qquad \text{and} \qquad k = 2,$$

corresponding respectively to mononucleotide and dinucleotide representations.

For each value of $k$, a hybrid feature vector

$$\Phi_k(x)$$

is constructed according to the procedure described above. The resulting representations are then concatenated into a common feature space.

This strategy makes it possible to simultaneously integrate information captured at different local resolutions of the sequence. It also strengthens the methodological scope of the proposed framework, since the study of k-mer effects is no longer treated as a secondary experimental analysis but becomes an intrinsic component of the classification model itself.

From a biological perspective, different values of $k$ capture complementary aspects of genomic organization. Mononucleotide representations characterize global compositional properties, whereas dinucleotide representations provide information about short-range dependencies and local sequence motifs. Their combination therefore yields a richer and more informative description of genomic sequences.

The resulting multi-scale embedding can be written schematically as

$$\Phi_{\mathrm{MS}}(x) = \Big[\Phi_1(x), \Phi_2(x)\Big],$$

where $\Phi_1(x)$ and $\Phi_2(x)$ denote the hybrid representations associated with the mononucleotide and dinucleotide levels, respectively.

5.3. **Hybrid HMM–SVM Classification Rule.** The final classification step is performed using a linear Support Vector Machine defined by

$$f(x) = \operatorname{sign}\big(w^\top \Phi(x) + b\big),$$

where $w$ denotes the weight vector and $b$ the intercept term, both estimated from the labeled training data.

The choice of a linear SVM is not motivated solely by its computational efficiency and numerical robustness. It also provides a direct interpretation of the learned coefficients, which is particularly desirable in an explainable bioinformatics framework. Indeed, the components of the vector $w$ quantify the contribution of each feature to the classification decision and therefore make it possible to identify the variables that play the most important role in discriminating between coding and non-coding regions.

This interpretability is especially valuable in genomic applications, where understanding the biological significance of predictive features is often as important as achieving high classification accuracy. In particular, the analysis of the learned coefficients allows one to assess the relative importance of likelihood-based scores, hidden-state occupancy statistics, entropy measures, and compositional descriptors in the final decision process.

Consequently, the proposed HMM–SVM framework combines the representational power of probabilistic sequence models with the discriminative strength and interpretability of modern supervised learning techniques.

5.4. **Hybrid HMM–SVM Results.** The proposed hybrid HMM–SVM framework was evaluated on the genomic dataset described in Section 4 in order to assess its ability to discriminate between coding (CDS) and non-coding (Non-CDS) regions. The objective of this experiment is twofold. First, we investigate whether the integration of probabilistic sequence modeling and discriminative learning improves classification performance. Second, we analyze the relative

importance of the extracted hybrid features in order to better understand the biological and statistical mechanisms underlying the classification process.

The following results illustrate both the predictive performance of the proposed framework and the contribution of the different categories of features included in the hybrid representation.

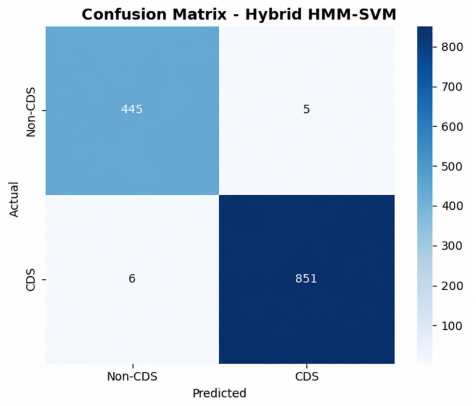


*Table 7. Classification Report — Hybrid HMM–SVM*

| Class | Precision | Recall | F1-score | Support |
|---|---|---|---|---|
| Non-CDS | 0.99 | 0.99 | 0.99 | 450 |
| CDS | 0.99 | 0.99 | 0.99 | 857 |
| Accuracy | 0.99 (out of 1307) | | | |
| Macro avg | 0.99 | 0.99 | 0.99 | 1307 |
| Weighted avg | 0.99 | 0.99 | 0.99 | 1307 |

*Figure 11. Performance of the Hybrid HMM–SVM Model*

*Table 8. Top 20 Most Discriminative Hybrid Features*

| Rank | Feature | Coef. | \|Coef.\| |
|---|---|---|---|
| 1 | k1_ll_cds | 4.126 | 4.126 |
| 2 | k1_delta_ll | 3.654 | 3.654 |
| 3 | k2_ll_cds | 1.801 | 1.801 |
| 4 | k2_delta_ll | 1.720 | 1.720 |
| 5 | k1_occ_cds_0 | -1.633 | 1.633 |
| 6 | k2_di_AT | -1.472 | 1.472 |
| 7 | k1_di_AT | -1.472 | 1.472 |
| 8 | k2_occ_non_cds_0 | 1.427 | 1.427 |
| 9 | k2_di_TA | 1.417 | 1.417 |
| 10 | k1_di_TA | 1.417 | 1.417 |
| 11 | k2_di_CA | 1.113 | 1.113 |
| 12 | k1_di_CA | 1.113 | 1.113 |
| 13 | k2_di_AC | -1.102 | 1.102 |
| 14 | k1_di_AC | -1.102 | 1.102 |
| 15 | k1_occ_cds_2 | -1.042 | 1.042 |
| 16 | k1_entropy_cds | 0.968 | 0.968 |
| 17 | k1_occ_non_cds_3 | 0.921 | 0.921 |
| 18 | k2_occ_cds_0 | 0.913 | 0.913 |
| 19 | k1_di_AG | -0.841 | 0.841 |
| 20 | k2_di_AG | -0.841 | 0.841 |

Performance Analysis. As shown in Figure 11, the proposed hybrid model achieves an overall accuracy of 0.99 on the test set composed of 1307 genomic sequences. The corresponding precision, recall, and F1-scores are all close to 0.99 for both

classes. Such results indicate an excellent balance between sensitivity and specificity and demonstrate the ability of the model to accurately discriminate between coding and non-coding regions.

The classification report further confirms the robustness of the proposed approach. The performance remains remarkably stable across both classes, suggesting that the classifier does not improve global accuracy at the expense of one category of genomic regions. Moreover, the confusion matrix indicates that only 11 sequences are misclassified, providing strong evidence that the hybrid representation captures the essential statistical differences between coding and non-coding DNA sequences.

These results substantially improve upon the performance obtained with the individual HMM, SVM, and Decision Tree models considered previously. The gain in predictive accuracy suggests that the different sources of information incorporated into the hybrid representation are highly complementary.

5.5. **Feature Interpretation.** A major advantage of the proposed methodology lies in the interpretability of the learned classifier. Because a linear SVM is employed, the estimated coefficients directly quantify the contribution of each feature to the classification decision. Consequently, the learned model provides not only accurate predictions but also valuable insight into the variables driving genomic discrimination.

The ranking displayed in Table 8 reveals that the most influential variables are primarily likelihood-based descriptors derived from the Hidden Markov Models, particularly `k1_ll_cds`, `k1_delta_ll`, `k2_ll_cds`, and `k2_delta_ll`. These quantities represent the global probabilistic evidence supporting either the coding or non-coding nature of a sequence. Their predominance demonstrates that the generative component of the framework provides the strongest discriminatory signal.

The analysis also highlights the importance of latent-state descriptors. Several hidden-state occupancy variables appear among the most influential predictors, indicating that the internal organization of the hidden states contains biologically meaningful information beyond the global likelihood scores. Furthermore, the presence of entropy-based descriptors among the most discriminative variables suggests that the uncertainty associated with the posterior state distributions itself contributes to the classification process. This result emphasizes the importance of considering not only the inferred latent structure but also the confidence associated with that structure.

In addition to the HMM-derived variables, several compositional descriptors associated with dinucleotide frequencies exhibit substantial coefficients. In particular, motifs involving `AT`, `TA`, `CA`, `AC`, and `AG` repeatedly appear among the most discriminative features. Their simultaneous presence across multiple scales confirms that local nucleotide composition remains an important source of biological information. More importantly, the multi-scale representation demonstrates that such compositional patterns complement rather than replace the information extracted from probabilistic sequence models.

Biological and Methodological Interpretation. Taken together, these findings provide strong evidence that effective genomic classification emerges from the integration of multiple complementary information sources. The proposed hybrid representation combines global likelihood evidence, latent-state organization, uncertainty measures, and biologically interpretable compositional descriptors within a unified feature space.

From a methodological perspective, the results validate the central hypothesis underlying the proposed framework: Hidden Markov Models and Support Vector Machines should not be viewed as competing paradigms, but rather as complementary tools whose integration yields a more powerful and informative classification methodology. The resulting framework therefore achieves not only very high predictive performance but also a meaningful level of interpretability, making it particularly attractive for bioinformatics applications where both accuracy and biological understanding are essential.

## 6. Discussion

The results obtained in this study highlight both the strengths and limitations of classical probabilistic and machine learning approaches for genomic sequence classification. More importantly, they demonstrate the benefits of integrating these complementary paradigms within the proposed hybrid HMM–SVM framework.

The experiments first confirm the relevance of Hidden Markov Models for genomic sequence analysis. Owing to their ability to explicitly represent latent biological structures and sequential dependencies, HMMs provide a natural probabilistic framework for distinguishing coding and non-coding regions. In particular, the mononucleotide HMM achieves an accuracy of approximately 91%, confirming that local nucleotide dependencies contain substantial information for genomic classification.

Support Vector Machines and Decision Trees also exhibit competitive performance, with accuracies close to 90%. These methods benefit from their discriminative nature and their ability to exploit compositional descriptors such as GC content and nucleotide frequencies. However, because they operate on fixed-dimensional feature vectors, they do not directly model the sequential organization of genomic data and therefore cannot fully exploit latent structural information contained in DNA sequences.

The main contribution of this work lies in the development of a hybrid HMM–SVM framework that combines the complementary strengths of both approaches. Rather than treating generative and discriminative models as competing paradigms, the proposed methodology integrates them within a unified representation space. Hidden Markov Models provide likelihood-based features, hidden-state occupancy statistics, and entropy measures that capture latent sequential organization, while the SVM learns an optimal discriminative boundary from these probabilistic descriptors together with biologically meaningful compositional variables.

The empirical results strongly support the effectiveness of this strategy. The hybrid model achieves an accuracy, F1-score, and ROC-AUC close to 0.99, substantially outperforming the individual HMM, SVM, and Decision Tree models. Such an improvement suggests that the different sources of information exploited by the hybrid representation are highly complementary.

6.1. **Methodological Discussion and Scope of the Model.** Beyond its predictive performance, the proposed framework provides several methodological contributions. First, it establishes a principled mechanism for combining probabilistic sequence modeling and discriminative learning within a unified classification architecture. The HMM component captures both the sequential dependencies and the latent biological organization of genomic sequences, whereas the SVM component transforms this information into an optimal classification rule.

The analysis of the learned coefficients reveals that the most influential variables are dominated by the likelihood-based quantities associated with the Hidden Markov Models, particularly the coding likelihood scores and likelihood contrasts. This observation demonstrates that the probabilistic component constitutes the primary source of discriminative information. At the same time, hidden-state occupancies, entropy measures, and compositional descriptors contribute significantly to the classification process, indicating that the hybrid model effectively combines multiple complementary sources of information.

From a broader perspective, the proposed representation integrates three fundamental aspects of genomic sequence analysis:

(1) probabilistic evidence through HMM likelihoods;
(2) latent biological organization through hidden-state representations;
(3) explicit sequence composition through nucleotide and dinucleotide descriptors.

The resulting framework therefore goes beyond a simple juxtaposition of a generative classifier and a discriminative classifier. Instead, it provides a coherent methodology in which probabilistic modeling and supervised learning reinforce one another. This integration is particularly relevant in genomics, where both local sequential dependencies and global compositional properties play a crucial role in determining biological function.

Another important feature of the proposed methodology is its interpretability. Unlike many modern black-box classification systems, the hybrid HMM–SVM framework allows direct identification of the variables driving the classification decision. This property is especially valuable in bioinformatics applications, where biological understanding and predictive performance are equally important.

More generally, the proposed framework should be viewed as a flexible methodological platform rather than a model restricted to a particular genomic dataset. The hybrid representation can naturally accommodate additional probabilistic descriptors, alternative hidden-state architectures, higher-order sequence models, or different discriminative classifiers. Consequently, the methodology introduced in this work may be applicable to a broader range of sequence analysis problems beyond the coding versus non-coding classification task considered here.

6.2. **Limitations and Future Research Directions.** Despite these encouraging results, several limitations should be acknowledged. First, the empirical study is restricted to the genome of *Escherichia coli*, a well-annotated prokaryotic organism whose genomic organization is relatively simple compared with that of higher eukaryotes. Although this dataset constitutes a standard benchmark for methodological evaluation, additional experiments on more complex genomes would be necessary to assess the generalizability of the proposed framework.

Second, the current implementation relies on first-order Hidden Markov Models together with mononucleotide and dinucleotide representations. While these models successfully capture short-range dependencies, they may not adequately represent longer-range biological interactions. Extensions based on higher-order Markov models or more sophisticated latent-state architectures could potentially provide additional improvements.

Several promising research directions emerge from this work. Future investigations may include the development of higher-order HMM representations, the integration of deep neural architectures within the hybrid framework, and the application of the proposed methodology to eukaryotic genomes exhibiting complex exon–intron structures. Another important direction concerns the theoretical analysis of the hybrid representation itself, including its statistical properties, robustness, consistency, and generalization capabilities.

Overall, the present study demonstrates that combining probabilistic sequential modeling with discriminative machine learning yields a powerful and interpretable framework for genomic sequence classification. Beyond its practical performance, the proposed HMM–SVM methodology provides a principled bridge between stochastic modeling and modern supervised learning, thereby opening new perspectives for the analysis, annotation, and understanding of biological sequences.

## 7. Conclusion

In this paper, we developed a unified framework for genomic sequence classification based on Markov chains, Hidden Markov Models, and supervised learning techniques.

The study first revisited the role of Markovian models in the probabilistic representation of DNA sequences and highlighted the importance of local nucleotide dependencies for distinguishing coding and non-coding regions. Hidden Markov Models were then employed to incorporate latent biological structures and to provide a richer representation of genomic organization.

The principal contribution of this work is the introduction of a hybrid HMM–SVM framework that combines probabilistic sequential modeling, latent-state information, and biologically interpretable compositional descriptors within a common feature space. Experimental results obtained on the *Escherichia coli* genome demonstrate that this hybrid representation substantially improves classification performance, achieving an accuracy close to 99% and outperforming the individual HMM, SVM, and Decision Tree models.

Beyond its predictive performance, the proposed methodology provides an interpretable framework that bridges generative probabilistic modeling and discriminative machine learning. The analysis of the learned features confirms that likelihood-based descriptors, hidden-state statistics, and compositional variables all contribute significantly to genomic discrimination.

Future work will focus on the extension of the framework to higher-order Markov models, more complex latent-state architectures, and large-scale genomic datasets, particularly eukaryotic genomes exhibiting rich exon–intron structures. Another promising direction concerns the theoretical analysis of the hybrid representation and its statistical properties.

Overall, the proposed HMM–SVM methodology provides a flexible and effective framework for genomic sequence classification and opens new perspectives for the integration of stochastic modeling and modern machine learning in computational genomics.

## References


1. Abbas, A. E., Cadenbach, A. H., & Salimi, E. (2017). A Kullback–Leibler view of maximum entropy and maximum log-probability methods. *Entropy*, 19(5), 232. https://doi.org/10.3390/e19050232
2. Avery, P. J., & Henderson, D. A. (1999). Fitting Markov chain models to discrete state series such as DNA sequences. *Journal of the Royal Statistical Society: Series C*, 48(1), 53–61. DOI: 10.1111/1467-9876.00139
3. Baldi, P., & Brunak, S. (2001). *Bioinformatics: The Machine Learning Approach* (2nd ed.). MIT Press. ISBN-10 : 0-262-02506-X
4. Benson, D. A., Boguski, M. S., Lipman, D. J., Ostell, J., Ouellette, B. F., Rapp, B. A., & Wheeler, D. L. (1999). GenBank. *Nucleic Acids Research*, 27(1), 12–17. https://doi.org/10.1093/nar/27.1.12
5. Blaisdell, B. E. (1985). Markov chain analysis in DNA sequences. *Journal of Molecular Evolution*, 21(3), 278–288. https://doi.org/10.1007/BF02102360
6. Brendel, V., Beckmann, J. S., & Trifonov, E. N. (1986). Linguistics of nucleotide sequences. *Journal of Biomolecular Structure and Dynamics*, 4(1), 11–21. https://doi.org/10.1080/07391102.1986.10507643
7. Churchill, G. A. (1989). Stochastic models for heterogeneous DNA sequences. *Bulletin of Mathematical Biology*, 51(1), 79–94. https://doi.org/10.1007/BF02458837
8. Durbin, R., Eddy, S. R., Krogh, A., & Mitchison, G. (1998). *Biological Sequence Analysis: Probabilistic Models of Proteins and Nucleic Acids*. Cambridge University Press. https://doi.org/10.1017/CBO9780511790492
9. Eddy, S. R. (1996). Hidden Markov models. *Current Opinion in Structural Biology*, 6(3), 361–365. https://doi.org/10.1016/S0959-440X(96)80056-X
10. Federhen, S. (2012). The NCBI taxonomy database. *Nucleic Acids Research*, 40(D1), D136–D143. https://doi.org/10.1093/nar/gkr1178
11. Kamalov, F., Choutri, S. E., & Atiya, A. F. (2025). Analytical formulation of synthetic minority oversampling technique (SMOTE) for imbalanced learning. *Gulf Journal of Mathematics*, 19(1), 1–15. https://doi.org/10.56947/gjom.v19i1.2639
12. Khodaei, A., et al. (2021). Markov chain-based classification of DNA sequences. *BioImpacts*, 11(2), 87–95. https://doi.org/10.34172/bi.2021.16

13. Kulkarni, V. G. (2016). *Modeling and Analysis of Stochastic Systems*. Chapman & Hall/CRC. https://doi.org/10.1201/9781315367910
14. Manning, C. D., & Schütze, H. (1999). *Foundations of Statistical Natural Language Processing*. MIT Press. ISBN: 9780262133609
15. Mertsch, D., & Stanke, M. (2022). Evolutionary models for coding region detection. *Bioinformatics*, 38(7), 1857–1862. https://doi.org/10.1093/bioinformatics/btac028
16. Morlando, F. (2026). The paradox of over-parameterization in learning algorithm. *Annals of Mathematics and Computer Science*, 34, 1–15. https://doi.org/10.56947/amcs.v34.824
17. Privault, N. (2013). *Understanding Markov Chains*. Springer. https://doi.org/10.1007/978-981-4451-51-2
18. Rabiner, L. R. (1989). A tutorial on hidden Markov models and selected applications in speech recognition. *Proceedings of the IEEE*, 77(2), 257–286. https://doi.org/10.1109/5.18626
19. Wu, T.-J., et al. (2001). Statistical measures of DNA sequences. *Biometrics*, 57(2), 441–448. https://doi.org/10.1111/j.0006-341X.2001.00441.x
20. Yandell, M., & Ence, D. (2012). A beginner's guide to eukaryotic genome annotation. *Nature Reviews Genetics*, 13(5), 329–342. https://doi.org/10.1038/nrg3174
21. Yoon, B.-J. (2009). Hidden Markov models and their applications in biological sequence analysis. *Current Genomics*, 10, 402–415. https://doi.org/10.2174/138920209789177575
22. Haitami, A.E. (2025). A Markovian switching diffusion for an Sirs epidemic model on complex network. Gulf Journal of Mathematics, 20(2), 253-274. https://doi.org/10.56947/gjom.v20i2.3079
23. Haitami, A.E., Boulmakoul, H., & Myr, A.E. (2026). Analysis of SIR epidemic model with relapse incorporating Ornstein-Uhlenbeck process. Gulf Journal of Mathematics, 22(2). https://doi.org/10.56947/gjom.v22i2.3987

[1] Laboratoire d'Etude et de Recherche en Statistique, Theorie Aleatoire et Applications (LERSTAD), Universite Gaston Berger, Saint-Louis, Senegal.
*Email address*: ali.dabye@ugb.edu.sn

[2] Departement de Mathematiques, Universite Gaston Berger, Saint-Louis, Senegal
*Email address*: douds7@yahoo.fr
*Email address*: mamadoumomar@gmail.com