DNA-Inspired Time Series Encoding: A Glimpse Into The Next 4-Hour Timeframe

Bao Bui-Quang
BScCS *, MScFE †
baobuiquang@proton.me

Abstract

In this work, we introduce a bio-inspired encoding framework for forecasting the direction of financial time series. Motivated by the limitations of linear models and the opacity of many deep learning approaches, we draw an analogy to genetics: observable micro-patterns are encoded into symbolic "Financial DNA" sequences. These sequences are then analyzed using a probabilistic state-transition mechanism to estimate the likelihood of subsequent market directions. We evaluate the approach on Bitcoin hourly OHLCV data with a rolling backtest. Among the horizons considered, modeling transitions from current Financial DNA patterns to the 4-hour-ahead price direction yields the strongest results, achieving a win ratio of 0.729. The findings suggest that compact, interpretable symbolic representations can capture salient, recurring structures in noisy, non-stationary markets and support effective directional forecasts.[1]

1 Introduction

Forecasting the direction of time series remains one of the most challenging [3] and captivating problems in quantitative finance. The inherent complexity, high level of noise [12], and non-stationary nature [15] of financial time series data have rendered traditional linear models insufficient for consistent prediction [10]. This has motivated a continuous search for approaches that can better capture the intricate and dynamic patterns governing market behavior. While sophisticated machine learning and deep learning models have shown promise, many operate as "black boxes" [4], making their underlying decision-making processes difficult to interpret.

This paper introduces a novel, bio-inspired framework for forecasting financial time series direction that combines conceptual clarity with probabilistic rigor. We draw an analogy from the field of genetics, where the complex traits of a biological species are encoded within its fundamental DNA sequence [14]. By treating short-term patterns in price movements as distinct "market species", we propose a method to encode these patterns into symbolic "Financial DNA" sequences. These sequences can then be analyzed to forecast future market states, much like a genetic sequence can be used to predict an organism's characteristics and behaviors [9]. Our goal is to create a model that is not only predictive but also interpretable, allowing for a deeper understanding of the market dynamics that precede significant price movements.

*Bachelor of Science in Computer Science

†Master of Science in Financial Engineering

[1]The approach presented in this work is intended solely for academic and research purposes. Readers are strongly cautioned against using these results for actual investment or trading decisions without thorough independent validation and robust risk management.

2 Proposed Method

2.1 A Bio-Inspired Framework for Financial Time Series

In biology, the immense diversity of life can be understood by studying the underlying genetic code [1]. While it is difficult to distinguish between two closely related species based on their macroscopic appearance alone, their distinction becomes clear upon analysis of their DNA (deoxyribonucleic acid) [2]. A DNA strand is a sequence composed of four basic nucleotides: Adenine (A), Guanine (G), Cytosine (C), and Thymine (T). A representative sequence might be:

ATGAAATTTGGGCCCTGA

This sequence of simple units provides a blueprint that defines the organism's traits.

We propose a parallel concept for financial markets. The raw financial time series data is analogous to the complex, macroscopic organism, which is difficult to classify simply by observing it. Our core idea is to establish a methodology for identifying the fundamental building blocks of price movements, analogous to nucleotides. These discretized units of price action can then be formed into short sequences, which we term "Financial DNA". Each unique sequence, or "market species", represents a distinct local pattern of market behavior. The hypothesis is that these underlying encoded patterns, rather than the raw price data, hold more distinct and reliable predictive information.

2.2 State Transitions Modeling with Conditional Probability

Once the financial time series is transformed from continuous price data into a symbolic sequence of "Financial DNA", the forecasting task becomes one of modeling the stochastic process that governs the evolution of these patterns. The fundamental mathematical tool for this is **conditional probability** [6], which allows us to quantify the likelihood of a future outcome based on observed historical patterns.

The core question we seek to answer is: "Given that we have just observed a specific 'Financial DNA' pattern, what is the probability that the market will subsequently move?". This is governed by the formal definition of conditional probability. For events A and B with $P(B) > 0$, the conditional probability of A given B is:

$$P(A \mid B) = \frac{P(A \cap B)}{P(B)}$$

In our context, let F_t denote the observed "Financial DNA" pattern at time t (our feature), and let $S_{t+1} \in S = \{s_1, \ldots, s_k\}$ denote the next market state (e.g., Up, Down, Sideways). The quantity of direct interest is therefore:

$$P\big(S_{t+1} = j \mid F_t = f\big) = \frac{P\big(F_t = f,\, S_{t+1} = j\big)}{P\big(F_t = f\big)}$$

which we estimate empirically from historical data by counting how often a specific pattern f is followed by each subsequent state j. Specifically, we calculate the probability of the next state S_{t+1} being, for example, 'Up', conditioned on observing a particular "Financial DNA" sequence F_t ending at time t.

A **Markov chain** [11, 13] is a structured way to organize such conditional probabilities under a simplifying assumption about memory. Let S_t be a discrete market state at time t. The first-order

Markov property assumes:

$$P\big(S_{t+1} \mid S_t, S_{t-1}, \dots\big) = P\big(S_{t+1} \mid S_t\big)$$

so that the model is fully specified by the transition matrix with entries:

$$P_{ij} = P\big(S_{t+1} = j \mid S_t = i\big)$$

In this view, a Markov chain does not supersede conditional probability; rather, it constrains which conditional relationships are retained (only those on the most recent state), trading model flexibility for parsimony and interpretability.

Our framework retains the clarity of conditional probability while allowing pattern information to inform transitions. Specifically, we condition on the observed "Financial DNA":

$$P\big(S_{t+1} = j \mid F_t = f\big)$$

or, when beneficial, augment the Markov state with pattern context:

$$P\big(S_{t+1} = j \mid S_t = i,\, F_t = f\big)$$

The former yields a feature-conditioned forecasting rule driven purely by patterns; the latter can be interpreted as a Markov model on an augmented state space (S_t, F_t), restoring the memoryless property in that expanded representation.

Instead of defining the "current state" as a single, isolated symbol, we define it as the entire observed **Financial DNA sequence**. This approach is analogous to a higher-order Markov chain, where the condition for the next transition is not just the last state but a longer sequence of prior states. By using the entire DNA sequence as our condition, we incorporate a richer history into our probability calculation, allowing for a more nuanced and potentially more accurate forecast.

3 Workflow

This research develops and tests a complete methodology based on the principles outlined above. The workflow implemented in this work is as follows:

1. **Data Discretization**: We first process historical time series data. We define a method to convert short-term, continuous price movements between consecutive data points into a finite alphabet of symbols, our "financial nucleotides".
2. **Sequence Encoding**: We then slide a window across the time series of these symbols to construct "Financial DNA" sequences. Each sequence represents a specific "market species" or local pattern.
3. **Probabilistic Forecasting**: For each identified "Financial DNA" sequence, we analyze the historical data to compute the conditional probability distribution of the subsequent market state (e.g., price moving up, down, or sideways).
4. **Signal Generation**: When a known "Financial DNA" pattern is recognized in new data, the model generates a forecast based on the pre-calculated probability of the most likely subsequent state.

4 Experiments

4.1 Dataset and Preprocessing

We preprocessed on the Bitcoin hourly OHLCV time series dataset [5] consisting of about 96 thousand raw data points, and finalized with more than 65 thousand data points for training and backtesting.

Figure 1: Preview of a sample in the Bitcoin Hourly OHLCV Time Series Dataset.

We prepared a set of 48 train/test dataframe pairs (backtesting sets) in a Python environment. These backtesting sets are sliding windows of the time series, 30 days apart. Each sliding window was divided into a training window and a testing window.

4.2 Encoding Scheme

A "Financial DNA" sequence, encoded from a consecutive data points sequence, is a set of DNA units. Each DNA unit is a combination of the single data point self-relation (at t) and the duo consecutive data points relation (at t and $t-1$).

Values Relation	$C_t > O_t$	$C_t < O_t$	$C_t \approx O_t$
Encoding Symbol	**U**	**D**	**S**
Interpretation	Up	Down	Sideways

Table 1: Single data point self-relation, with t is the timestamp of the data point

Values Relation	$x_t > x_{t-1}$	$x_t < x_{t-1}$	$x_t \approx x_{t-1}$
Encoding Symbol	**A**	**B**	**E**
Interpretation	Above	Below	Equal

Table 2: Duo consecutive data points relation, with $x_t \in X = \{O_t, H_t, L_t, C_t\}$

A DNA unit u_t contains a set of symbols n_i, where $n_i \in \{U, D, S, A, B, E\}$.

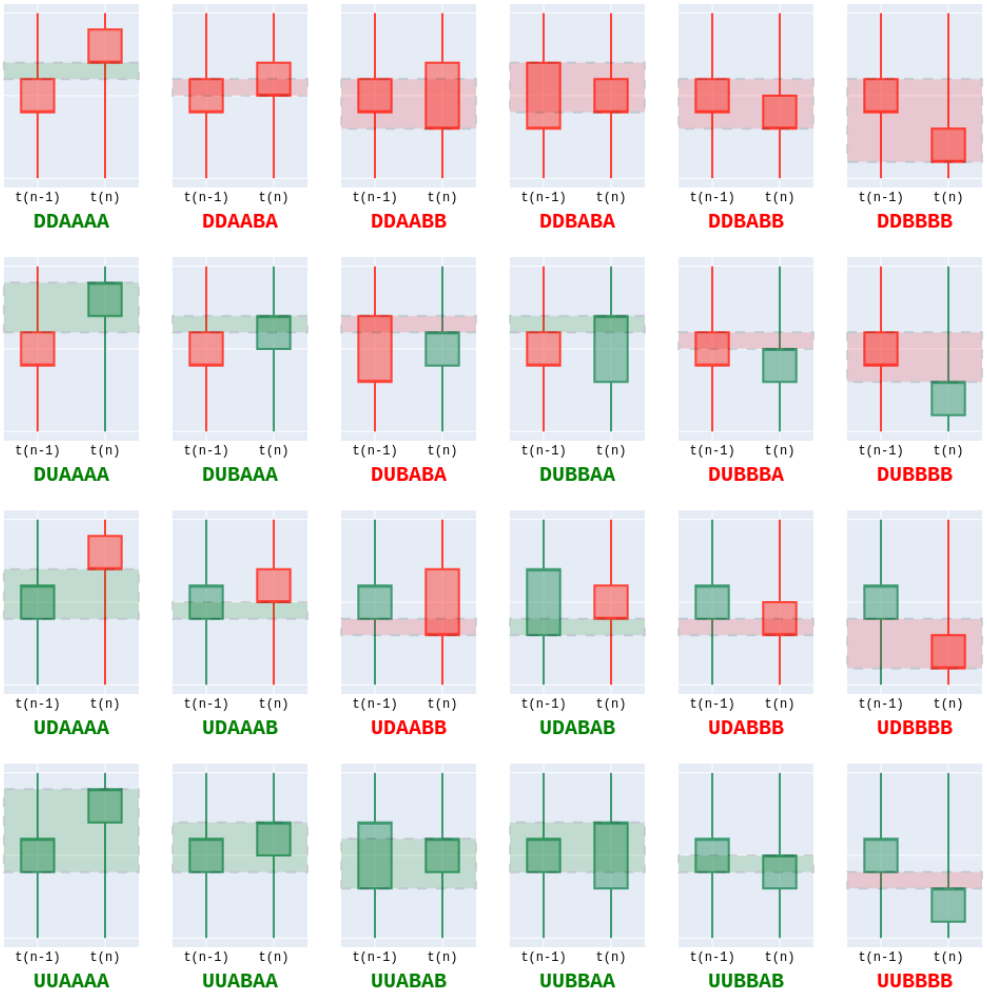

Figure 2: Examples of “Financial DNA” units.

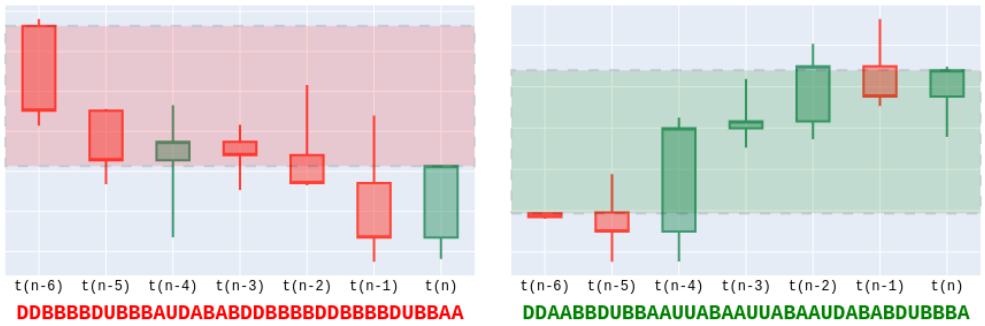

Figure 3: Examples of “Financial DNA” sequences.

4.3 Training and Backtesting Results

A backtesting set is considered "win" when:

$$\frac{\text{number of win entries} - \text{number of lose entries}}{\text{total testing entries}} \geq \epsilon$$

where ϵ is a small float ($\epsilon > 0$) to prevent the case when the number of win entries is too close or equal to the number of lose entries.

Signal	Count sets win	Count sets lose	Win ratio
SIGNAL(t+1)	30	18	0.625
SIGNAL(t+2)	33	15	0.688
SIGNAL(t+3)	29	19	0.604
SIGNAL(t+4)	**35**	**13**	**0.729**
SIGNAL(t+5)	26	22	0.542
SIGNAL(t+6)	27	21	0.563
SIGNAL(t+7)	24	24	0.500
SIGNAL(t+8)	27	21	0.563
SIGNAL(t+9)	19	29	0.396
SIGNAL(t+10)	19	29	0.396
SIGNAL(t+11)	19	29	0.396
SIGNAL(t+12)	20	28	0.417

Table 3: SIGNAL(t+n) is the direction of price in the following n-hour

Testing on 12 signals, SIGNAL(t+4) produced the best state transitions model with 35 sets win and 13 sets lose on the total of 48 backtesting sets. SIGNAL(t+2) also produced a high win ratio with 33 sets win and 15 sets lose. From t+9, the win ratio decreases quickly and under 0.5, telling us about the weakness of the patterns formed by "Financial DNA" with too-far direction information.

5 Conclusion

In this work, we presented a DNA-inspired framework for forecasting financial time series data. We proposed a method to encode short-term patterns in price movements into symbolic "Financial DNA" sequences. These sequences can then be analyzed to predict the next market directions.

For the Bitcoin hourly OHLCV time series dataset, the backtesting results show that the patterns of current "Financial DNA" with the following 4-hour price direction lead to the best state transitions model with the highest win ratio (0.729), giving us the result of 35 sets win out of the total of 48 backtesting sets.

Despite the promising results, the approach presented in this work is intended solely for academic and research purposes. Financial markets are inherently noisy, non-stationary, and subject to regime shifts [7], unforeseen events, and microstructure effects [8] (slippage, transaction costs, liquidity constraints) that are not accounted for in this study. Readers are strongly cautioned against using these results for actual investment or trading decisions without thorough independent validation and robust risk management.

Training Window	Testing Window	W	L	N	Set win
23/06/16 - 25/06/15	25/06/16 - 25/08/15	93	97	1249	False
23/05/17 - 25/05/16	25/05/17 - 25/07/16	96	72	1271	True
23/04/17 - 25/04/16	25/04/17 - 25/06/16	102	80	1257	True
23/03/18 - 25/03/17	25/03/18 - 25/05/17	93	115	1231	False
23/02/16 - 25/02/15	25/02/16 - 25/04/17	91	110	1238	False
23/01/17 - 25/01/16	25/01/17 - 25/03/18	114	111	1214	False
22/12/18 - 24/12/17	24/12/18 - 25/02/16	115	114	1210	False
22/11/18 - 24/11/17	24/11/18 - 25/01/17	120	95	1224	True
22/10/19 - 24/10/18	24/10/19 - 24/12/18	132	89	1218	True
22/09/19 - 24/09/18	24/09/19 - 24/11/18	116	95	1228	True
22/08/20 - 24/08/19	24/08/20 - 24/10/19	118	110	1211	True
22/07/21 - 24/07/20	24/07/21 - 24/09/19	127	108	1204	True
22/06/21 - 24/06/20	24/06/21 - 24/08/20	126	115	1198	True
22/05/22 - 24/05/21	24/05/22 - 24/07/21	125	105	1209	True
22/04/22 - 24/04/21	24/04/22 - 24/06/21	120	93	1226	True
22/03/23 - 24/03/22	24/03/23 - 24/05/22	129	113	1197	True
22/02/21 - 24/02/21	24/02/22 - 24/04/22	124	126	1189	False
22/01/22 - 24/01/22	24/01/23 - 24/03/23	149	129	1161	True
21/12/23 - 23/12/23	23/12/24 - 24/02/22	147	98	1194	True
21/11/23 - 23/11/23	23/11/24 - 24/01/23	130	90	1219	True
21/10/24 - 23/10/24	23/10/25 - 23/12/24	123	84	1232	True
21/09/24 - 23/09/24	23/09/25 - 23/11/24	117	81	1241	True
21/08/25 - 23/08/25	23/08/26 - 23/10/25	112	104	1223	True
21/07/26 - 23/07/26	23/07/27 - 23/09/25	129	109	1201	True
21/06/26 - 23/06/26	23/06/27 - 23/08/26	147	95	1197	True
21/05/27 - 23/05/27	23/05/28 - 23/07/27	142	94	1203	True
21/04/27 - 23/04/27	23/04/28 - 23/06/27	113	85	1241	True
21/03/28 - 23/03/28	23/03/29 - 23/05/28	105	95	1239	True
21/02/26 - 23/02/26	23/02/27 - 23/04/28	101	92	1246	True
21/01/27 - 23/01/27	23/01/28 - 23/03/29	113	91	1235	True
20/12/28 - 22/12/28	22/12/29 - 23/02/27	108	83	1248	True
20/11/28 - 22/11/28	22/11/29 - 23/01/28	117	77	1245	True
20/10/29 - 22/10/29	22/10/30 - 22/12/29	94	88	1257	False
20/09/29 - 22/09/29	22/09/30 - 22/11/29	81	89	1269	False
20/08/30 - 22/08/30	22/08/31 - 22/10/30	102	80	1257	True
20/07/31 - 22/07/31	22/08/01 - 22/09/30	97	77	1265	True
20/07/01 - 22/07/01	22/07/02 - 22/08/31	87	96	1256	False
20/06/01 - 22/06/01	22/06/02 - 22/08/01	104	102	1233	False
20/05/02 - 22/05/02	22/05/03 - 22/07/02	107	98	1234	True
20/04/02 - 22/04/02	22/04/03 - 22/06/02	126	110	1203	True
20/03/03 - 22/03/03	22/03/04 - 22/05/03	105	103	1231	False
20/02/02 - 22/02/01	22/02/02 - 22/04/03	101	93	1245	True
20/01/03 - 22/01/02	22/01/03 - 22/03/04	111	108	1220	False
19/12/04 - 21/12/03	21/12/04 - 22/02/02	111	92	1236	True
19/11/04 - 21/11/03	21/11/04 - 22/01/03	119	103	1217	True
19/10/05 - 21/10/04	21/10/05 - 21/12/04	120	89	1230	True
19/09/05 - 21/09/04	21/09/05 - 21/11/04	106	92	1241	True
19/08/06 - 21/08/05	21/08/06 - 21/10/05	89	90	1260	False
					35 sets win 13 sets lose

Table 4: Backtesting results of SIGNAL(t+4) with 0.729 win ratio

References

[1] Lewis J Alberts B, Johnson A. The Diversity of Genomes and the Tree of Life. In *Molecular Biology of the Cell. 4th edition.* NIH, 2002.

[2] Lewis J Alberts B, Johnson A. The Structure and Function of DNA. In *Molecular Biology of the Cell. 4th edition.* NIH, 2002.

[3] Adam Bouland, Wim van Dam, Hamed Joorati, Iordanis Kerenidis, and Anupam Prakash. Prospects and challenges of quantum finance, 2020.

[4] Samuel N. Cohen, Derek Snow, and Lukasz Szpruch. Black-box model risk in finance, 2021.

[5] CryptoCompare API collected by Mouad Jaouhari. Kaggle Dataset - Bitcoin Hourly OHLCV Dataset (kaggle.com/datasets/mouadjaouhari/bitcoin-hourly-ohclv-dataset).

[6] Steven Cruickshank. Conditional probability and Bayes' theorem. In *Mathematics and Statistics in Anaesthesia*. Oxford University Press, 08 1998.

[7] Marina Dolfin, George Kapetanios, Leone Leonida, and Jose De Leon Miranda. Investor behavior and multiscale cross-correlations: Unveiling regime shifts in global financial markets, 2024.

[8] Andreas Krause. Inventory effects on daily returns in financial markets. *International Journal of Theoretical and Applied Finance*, 06(07):739–765, November 2003.

[9] Benjamin Kuznets-Speck, Buduka K. Ogonor, Thomas P. Wytock, and Adilson E. Motter. Generative prediction of causal gene sets responsible for complex traits. *bioRxiv*, 2025.

[10] Haochun Ma, Davide Prosperino, Alexander Haluszczynski, and Christoph Räth. Linear and nonlinear causality in financial markets, 2023.

[11] S. Meyn and R.L. Tweedie. *Markov Chains and Stochastic Stability*. Cambridge Mathematical Library. Cambridge University Press, 2009.

[12] Omkar Nabar and Gautam Shroff. Conservative Predictions on Noisy Financial Data. In *4th ACM International Conference on AI in Finance*, ICAIF '23, page 427–435. ACM, November 2023.

[13] J. R. Norris. *Markov Chains*. Cambridge University Press, Cambridge, UK, 1997.

[14] Suvam Roy and Supratim Sengupta. The RNA-DNA world and the emergence of DNA-encoded heritable traits. *bioRxiv*, 2024.

[15] Thilo A. Schmitt, Desislava Chetalova, Rudi Schäfer, and Thomas Guhr. Non-stationarity in financial time series: Generic features and tail behavior. *EPL (Europhysics Letters)*, 103(5):58003, September 2013.