Binary Sequence Analysis by Translated Regrouping

William F. Barnes

Abstract

A finite binary sequence may be examined for order by asking how well its visible past predicts its next member. The construction developed here begins with element regrouping. For every order g, a sequence is translated through all g alignments, regrouped into elements of $\phi_{2,g}$, and summarized by a grand population vector $\sigma_{2,g}$. The populations of the two elements that can continue the recent suffix determine a partial guess x_g. Binary weighting models combine the partial guesses, and a guess–reveal procedure measures each model against every prefix of the original sequence. With $g_{\max} = 10$, the nontrivial model space contains 1,023 weighting models. This article gives the construction, its correctness and streak statistics, its application to random and percussive sequences, and the limits of what a finite model family can establish about randomness.

1 The problem

Let

$$S_N(\phi_2) = (s_1 \; s_2 \; \cdots \; s_N) \,, \qquad \phi_2 = \{0, 1\}, \tag{1}$$

be a binary sequence with N members. The sequence might arise from a stochastic process, a deterministic rule, or a human attempt to behave randomly. The task is to distinguish sequence order that can be used for prediction from sequence order that is not visible to the present method.

A count of zeros and ones is insufficient. The sequence

$$010101010101\cdots$$

contains equal populations of its two elements, yet every digit after the first is determined. The order appears only when neighboring members are considered together. This motivates two related goals:

1. compare binary patterns at several regrouping orders; and
2. use those patterns to predict the next member of a growing sequence.

Prediction supplies the test. Given only the prefix

$$S_n(\phi_2) = (s_1 \; s_2 \; \cdots \; s_n) \,, \qquad n < N, \tag{2}$$

one calculates a guess X^{n+1}, then reveals s_{n+1} and compares the two. A rule that repeatedly guesses correctly has detected order in the sequence.

2 Element regrouping and translation

2.1 Regrouping

An order-g regrouping collects adjacent members into blocks of length g. The element vector becomes

$$\phi_{2,g} = \{00\cdots 0, 00\cdots 1, \ldots, 11\cdots 1\}, \tag{3}$$

which contains 2^g elements. In particular,

$$\phi_{2,1} = \{0, 1\}, \tag{4}$$
$$\phi_{2,2} = \{00, 01, 10, 11\}, \tag{5}$$
$$\phi_{2,3} = \{000, 001, 010, 011, 100, 101, 110, 111\}. \tag{6}$$

If g does not divide N, the final members do not complete a block. The regrouping may therefore be written

$$S_N(\phi_2) = S_{\lfloor N/g \rfloor}(\phi_{2,g}) + R_g(\phi_2), \tag{7}$$

where R_g contains N modulo g members.

Regrouping exposes patterns hidden by the original element vector. For example,

$$S_{24}(\phi_2) = 010011001110010011001110 \tag{8}$$

can be viewed as

$$S_6(\phi_{2,4}) = (e\ m\ o\ e\ m\ o), \tag{9}$$

after assigning letters to the sixteen elements of $\phi_{2,4}$. The apparent binary disorder is two repetitions of the same three-element word.

2.2 Translation

The population obtained by regrouping depends on where the first block begins. To include every alignment, define the order-t element translation

$$^{t}S_N(\phi_2) = (s_{1+t}\ s_{2+t}\ \cdots\ s_N\ s_1\ \cdots\ s_t), \tag{10}$$

where indices are interpreted cyclically. An order-g regrouping has g distinct alignments, represented by

$$^{t}S_N(\phi_2), \qquad t = 0, 1, \ldots, g-1. \tag{11}$$

Translation does not change the original binary populations, but it generally changes the populations of the regrouped elements.

3 Population vectors

For a sequence written in an element vector $\phi_{M,g}$, define the *population vector* $\rho_{M,g}$ to be the list of occurrence counts of the M^g possible elements. A translation has its own population vector, denoted

$$^{t}\rho_{M,g}. \tag{12}$$

The translation number is attached to ρ because regrouping is performed after translation.

3.1 Grand population vector

The *grand population vector* at order g is the sum over all distinct alignments:

$$\boxed{\sigma_{M,g} = \sum_{t=0}^{g-1} {}^{t}\rho_{M,g}.} \tag{13}$$

For the binary problem, $M = 2$. Each translated regrouping contributes $\lfloor N/g \rfloor$ elements, so the components of $\sigma_{2,g}$ sum to

$$g \left\lfloor \frac{N}{g} \right\rfloor . \tag{14}$$

When g divides N, the grand population counts every cyclic block of length g exactly once. When it does not, Equation (14) records the exclusion of the regrouping remainders.

3.2 A binary example

Consider the sequence

$$S_{48}(\phi_2) = 101101001010101101100110 \tag{15}$$

$$101100010001110101011010. \tag{16}$$

At order one, the population vector is

$$\sigma_{2,1} = \rho_2 = \{23_0, 25_1\}. \tag{17}$$

At order two, summing the two translated populations gives

$$\sigma_{2,2} = \{6_{00}, 17_{01}, 17_{10}, 8_{11}\}. \tag{18}$$

At order three, the corrected grand population is

$$\sigma_{2,3} = \{2_{000}, 4_{001}, 10_{010}, 7_{011}, \\ 4_{100}, 13_{101}, 7_{110}, 1_{111}\}. \tag{19}$$

The components in each case sum to 48. Motifs such as 101 and 010 are far more common than 000 and 111.

4 Partial and final guesses

4.1 Partial guess

Suppose $S_N(\phi_2)$ is to be extended by one member. At order g, take the final $g-1$ members and call the resulting suffix u_g. Only two elements of $\phi_{2,g}$ can continue this suffix:

$$u_g0 \qquad \text{and} \qquad u_g1. \tag{20}$$

Let $\sigma_{2,g}[b]$ denote the component of the grand population belonging to the block b. The order-g *partial guess* is

$$x_g = \begin{cases} 0, & \sigma_{2,g}[u_g0] > \sigma_{2,g}[u_g1], \\ 1, & \sigma_{2,g}[u_g1] > \sigma_{2,g}[u_g0], \\ \frac{1}{2}, & \sigma_{2,g}[u_g0] = \sigma_{2,g}[u_g1]. \end{cases} \tag{21}$$

The value 1/2 represents an undecided population rather than an arbitrary binary choice.

The working sequence in Equation (16) ends in 0, 10, and 010. Equations (17)–(19), together with the order-four population, give

$$x_1 = 1 \qquad (25 > 23), \tag{22}$$
$$x_2 = 1 \qquad (\sigma_{2,2}[01] = 17 > 6), \tag{23}$$
$$x_3 = 1 \qquad (\sigma_{2,3}[101] = 13 > 4), \tag{24}$$
$$x_4 = 1 \qquad (\sigma_{2,4}[0101] = 8 > 2). \tag{25}$$

The same construction continues to a chosen cutoff $g_{\max}$.

4.2 Weighting models

A set of partial guesses is combined by weights w_g. A weighting model is the vector

$$W_q = (w_1, w_2, \ldots, w_{g_{\max}}). \tag{26}$$

For the present model space, each weight is binary:

$$w_g \in \{0, 1\}. \tag{27}$$

Thus an order either contributes to the final guess or does not. The model number is

$$q = \sum_{g=1}^{g_{\max}} w_g 2^{g-1}. \tag{28}$$

The first component w_1 is the least significant binary place. For example,

$$W_1 = (1, 0, 0, 0, \ldots), \tag{29}$$
$$W_2 = (0, 1, 0, 0, \ldots), \tag{30}$$
$$W_3 = (1, 1, 0, 0, \ldots). \tag{31}$$

Let $G_N = \min(N, g_{\max})$ and define the number of active weights by

$$A_q(N) = \sum_{g=1}^{G_N} w_g. \tag{32}$$

When $A_q(N) > 0$, the weighted mean is

$$Y_q(N) = \frac{1}{A_q(N)} \sum_{g=1}^{G_N} w_g x_g. \tag{33}$$

The final binary guess is

$$\boxed{X_q^{N+1} = \begin{cases} 1, & Y_q(N) > \frac{1}{2}, \\ 0, & Y_q(N) \le \frac{1}{2}. \end{cases}} \tag{34}$$

An exact tie is therefore resolved as 0. If $A_q(N) = 0$ during the first few members of a sequence, the same zero convention is used until one of the model's selected orders becomes available.

4.3 Model space

Choosing $g_{\max} = 10$ gives $2^{10} = 1{,}024$ binary weighting vectors. The all-zero vector W_0 contains no active order and is excluded from comparison. The nontrivial model space is therefore

$$\mathcal{W}_{10} = \{W_1, W_2, \ldots, W_{1023}\}. \tag{35}$$

The last model,

$$W_{1023} = (1, 1, 1, 1, 1, 1, 1, 1, 1, 1), \tag{36}$$

gives every regrouping order an equal vote.

5 The guess–reveal analysis

Fix a model W_q. Beginning with the one-member prefix $S_1(\phi_2)$, calculate X_q^2, reveal s_2, and record whether the guess was correct. Next calculate X_q^3 from $S_2(\phi_2)$, reveal s_3, and continue until the sequence is exhausted. The full sequence is used only as a record of what is eventually revealed; no guess may depend on a later member.

Define

$$I_q(n) = \begin{cases} 1, & X_q^{n+1} = s_{n+1}, \\ 0, & X_q^{n+1} \neq s_{n+1}, \end{cases} \qquad n = 1, \ldots, N-1. \tag{37}$$

There are $T = N - 1$ prediction trials. The correct count and correctness rating are

$$K_q = \sum_{n=1}^{N-1} I_q(n), \tag{38}$$

$$C_q = \frac{K_q}{N-1}. \tag{39}$$

Let R_q denote the longest consecutive run of indices for which $I_q(n) = 1$.

Every model in $\mathcal{W}_{10}$ is subjected to the same sequence of guess–reveal trials. A winning model W_{q^*} is selected by

1. the greatest correctness rating C_q;
2. if correctness is tied, the greatest longest streak R_q; and
3. if both are tied, the smaller model number q.

After the winner is selected, one further application to $S_N(\phi_2)$ gives the unscored forward prediction

$$X_{q^*}^{N+1}. \tag{40}$$

5.1 Complete procedure

The analysis may be summarized as follows:

1. For each prefix $S_n(\phi_2)$, form translated regroupings for $1 \leq g \leq \min(n, g_{\max})$.
2. Add the translated population vectors to obtain $\sigma_{2,g}$.
3. Compare the populations of u_g0 and u_g1 to obtain every x_g.

4. Use each W_q to calculate $Y_q(n)$ and X_q^{n+1}.
5. Reveal s_{n+1} and update K_q, the current streak, and R_q.
6. After $N-1$ trials, rank the models and retain W_{q^*}.
7. Apply W_{q^*} to the full sequence for the next-member prediction.

6 Randomness diagnostics

6.1 The fair-guess baseline

For T independent fair binary guesses, the expected correct count is $T/2$ and the standard deviation is

$$\sigma_T = \frac{1}{2}\sqrt{T}. \tag{41}$$

Equivalently, the standard deviation of the correctness fraction is

$$\sigma_C = \frac{1}{2\sqrt{T}}. \tag{42}$$

For $T = 100$, this is 5%; for $T = 1{,}000$, it is about 1.6%. A convenient standardized deviation for model W_q is

$$Z_q = \frac{K_q - T/2}{\sqrt{T}/2}. \tag{43}$$

The familiar 45%–55% random zone is a useful descriptive band near $T = 100$, not a sample-size-independent significance test. Moreover, W_{q^*} is selected from 1,023 competitors. The distribution of the winning correctness is therefore not the binomial distribution of one model chosen in advance. Equation (43) remains a common scale for comparison, but it should not be read as an uncorrected p-value for the winner.

6.2 Longest streak

Correctness alone can conceal local order. A model may score near 50% over the full sequence while predicting one long interval without error. This is why the longest streak R_q is retained separately.

For a fair binary guess sequence of length T, a simple approximation for the probability of observing a run at least as long as r begins with

$$\lambda(T, r) = \frac{T - r + 1}{2^{r+1}}, \tag{44}$$

the approximate expected number of new length-r runs. Treating those starts as rare events gives

$$P(R \geq r) \approx 1 - \exp[-\lambda(T, r)]. \tag{45}$$

A value below 1% marks an abnormally long streak for descriptive purposes. As with Z_{q^*}, selection of the winning model means this is not an exact family-wise probability.

6.3 Model spread

Let

$$C_{\min} = \min_{W_q \in \mathcal{W}_{10}} C_q. \tag{46}$$

The model spread is

$$\Delta C = C_{q^*} - C_{\min}. \tag{47}$$

A narrow spread means that most regrouping mixtures behave similarly. A wide spread means that the choice of pattern orders matters strongly. Some orders follow the sequence while others are uninformative or effectively inverted.

No single diagnostic is decisive in every case. A strong correctness rating, an abnormal streak, and a wide model spread are complementary evidence of sequence order.

7 Case studies

7.1 Random-like sequences

A sequence of $N = 2{,}000$ members generated as a binary random control produced a best correctness near 52%, a worst correctness near 48%, and a longest winning streak of 13. The best model was W_{66}. Both the correctness band and the streak length are compatible with a random-like sample at this length; translated regrouping found no persistent order.

The central column of the Rule 30 cellular automaton supplies a second benchmark. For a sample of $N = 2{,}000$, the best model was W_2 with about 51% correctness and a longest streak of 11, while the least successful models were near 46%. Again the model family found no useful predictive order.

These conclusions are deliberately one-sided. Successful prediction is evidence of detected structure. Unsuccessful prediction means only that no structure was found within the tested regrouping orders and model family.

7.2 Percussive sequences

A *percussive sequence* is periodic, as if generated by repeating a rhythmic seed. Write

$$P_{T,N}(\phi_2) \tag{48}$$

for a binary sequence of length N and period T. The minimum block required to reproduce it is the seed D_T. Translations of the same seed describe the same cyclic pattern.

For example,

$$D_2 = 01, \tag{49}$$
$$D_5 = 01011, \tag{50}$$
$$D_{10} = 0001110011, \tag{51}$$
$$D_{30} = 010011000111000011110000011111. \tag{52}$$

Low-period percussive sequences are learned quickly because their repeated blocks appear prominently in one or more grand population vectors. Periods larger than $g_{\max}$ can still leave shorter motifs that are strongly predictive.

The following measurements use $N = 240$, $g_{\max} = 10$, exact correctness before rounding, and the ranking rules stated above.

Seed	Winner	Correct	Streak	Spread	$N - R_{q^*}$
D_5	W_{17}	97.9%	231	39.3%	9
D_{10}	W_{162}	97.1%	221	66.9%	19
D_{30}	W_{770}	89.5%	175	54.8%	65

The period-five and period-ten sequences approach perfect prediction after a short learning interval. Even the period-thirty sequence, whose seed is three times longer than $g_{\max}$, is predicted correctly almost 90% of the time. The shorter-order motifs do not describe the full period directly, but they still constrain what can follow.

7.3 Convergence time

For a model that eventually stops making errors, define its convergence time by

$$\tau_q = \min \{n_0 : I_q(n) = 1 \text{ for every } n \geq n_0\}. \tag{53}$$

When the longest winning streak reaches the end of the sequence,

$$\tau_q \approx N - R_q. \tag{54}$$

The final column of the table gives this estimate for the winning models. The quantity separates early learning errors from later stable prediction, which a single correctness percentage does not.

8 Scope of the method

The analysis tests finite-memory pattern rules. Its conclusions are bounded by the chosen maximum regrouping order, the binary weighting family, and the length of the sample.

First, a deterministic sequence may depend on motifs longer than $g_{\max}$ or on a rule that changes with position. Failure of $\mathcal{W}_{10}$ does not prove the absence of such order. Second, the same data are used to select and score the winning model. Short samples are particularly vulnerable to an accidentally strong winner among 1,023 candidates. Independent holdout data or repeated samples from the same source provide a stricter test. Third, the analysis concerns predictability of the observed symbols, not the physical cause of the sequence.

The asymmetry is important: reliable prediction supplies positive evidence of structure, whereas a near-50% result supplies only an absence of detected structure. The proper description in the latter case is *random-like under the tested models*, not a proof of randomness.

9 Conclusion

Element regrouping turns a binary sequence into pattern sequences over the vectors $\phi_{2,g}$. Translation removes the arbitrary choice of where a regrouping begins, and the grand population vector $\sigma_{2,g}$ collects the resulting motif counts. Comparing the two motifs that can continue the recent suffix produces the partial guess x_g. Binary weighting models combine the partial guesses into final predictions.

The guess–reveal analysis measures those predictions without look-ahead. Correctness, longest streak, standardized deviation, and model spread describe different aspects of the result, while the winning model identifies which regrouping orders carry predictive value. Percussive sequences are learned rapidly; random controls remain close to the fair-guess baseline. The final application of the winning model gives a concrete prediction for the next member, but the method does not certify randomness.