How Predictable Are Biological Sequences?

Izydor Apostol
Amgen Inc.
One Amgen Center Drive
Thousand Oaks, CA 91320
U.S.A.
iapostol@amgen.com

Philippe Jacquet
INRIA
Rocquencourt
78153 Le Chesnay Cedex
France
Philippe.Jacquet@inria.fr

Wojciech Szpankowski*
Dept. of Computer Science
Purdue University
W. Lafayette, IN 47907
U.S.A.
spa@cs.purdue.edu

Keywords: Molecular sequence analysis, molecular structures, strings algorithms and prediction, combinatorics, stochastic modeling.

1 Introduction

We apply our novel prediction algorithm called *Sampled Pattern Matching* (SPM) predictor that was recently developed by us in [1] to analyze molecular sequences. It is assumed that a (biological) sequence $x_1^n := x_1, \ldots, x_n$ is given. Each symbol x_i belongs to a finite alphabet $\mathcal{A}$. For a fixed integer $K \geq 1$, the algorithm will predict the next K symbols, that is, $(\hat{x}_{n+1}, \ldots, \hat{x}_{n+K})$.

Let us fix $0 < \alpha < 1$. The SPM prediction algorithm works as follows:

1. Find the largest suffix of x_1^n whose copy appears somewhere in the string x_1^n. We call this suffix the *maximal suffix* and denote its length by D_n. More precisely, $D_n := l$ where l is the largest integer such that $(x_{n-l+1}, \ldots, x_n) = (x_{n-i-l+1}, \ldots, x_{n-i})$ for some $1 \leq i \leq n$.
2. Take an α fraction of the maximal suffix of length $k_n := \lceil \alpha D_n \rceil$, that is, the suffix $x_{n-k_n+1}, \ldots, x_n$. Such a fractional suffix occurs more than twice in the original string. Each such a occurrence defines a *marker* (i.e., a substring), and the K positions after markers are called the *marked positions*. Finally, by the *sampled sequence* we mean a sequence composed of all symbols from the K-tuple marked positions.
3. Let now $N(x_1, \ldots, x_K)$ be the number of non-overlapping K-tuple $(x_1, \ldots, x_K)$ occurrences in the sampled sequence. The SPM predictor assigns
$$(\hat{x}_{n+1}, \ldots, \hat{x}_{n+K}) = \arg\max N(x_1, \ldots, x_K) \tag{1}$$
with a tie broken in an arbitrary manner (e.g., by a random selection). In words, $(\hat{x}_{n+1}, \ldots, \hat{x}_{n+K})$ is assigned to the most frequent K-tuple occurring in the sampled sequence.

Example 2. *SPM Predictor for $K = 1$*

Below is presented a text with the largest suffix and its copy framed (defined in Step 1 of the above algorithm):

SLJZGGDL $\boxed{\text{YGSJSLJZ}}$ KGSSLJZIDSLJZJGZ $\boxed{\text{YGSJSLJZ}}$

In fact, $D_{40} = 8$. Let $\alpha = 0.5$. Then the fractional suffix SLJZ is used to find all markers. They are shown below:

$\boxed{\text{SLJZ}}$ GGDLYGSJ $\boxed{\text{SLJZ}}$ KGS $\boxed{\text{SLJZ}}$ KLJZJGZYGSJ $\boxed{\text{SLJZ}}$

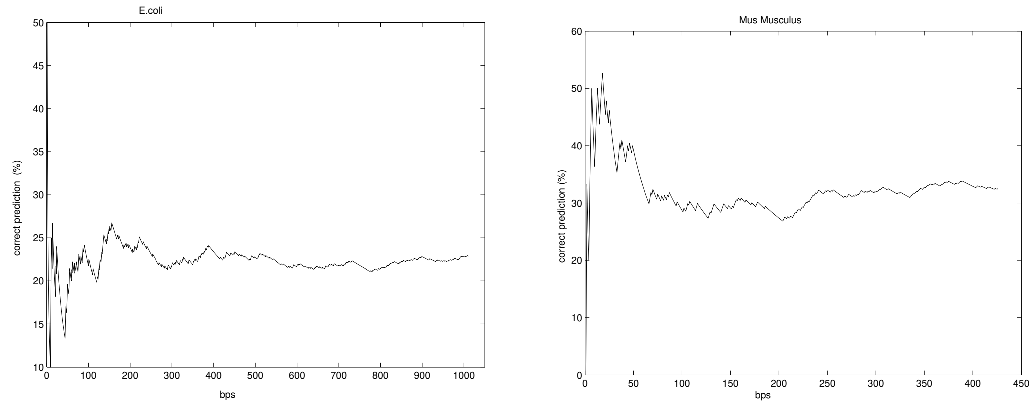
The sampled sequence is GKK, thus the SPM predicts $\hat{x}_{41} = K$.

To assess quality of our SPS *universal* predictor (designed for *unknown distributions*) we compare it to the *optimal* predictor designed for *known distributions* (that always predicts the most likely symbol). We define the *predictability* $\hat{\pi}_n(\mathcal{M})$ of our predictor $\hat{X}_{n+1}$ as the average prediction error, that is, $\hat{\pi}_n(\mathcal{M}) = \frac{1}{n}\sum_{i=1}^n \Pr\{\hat{X}_i \neq X_i\}$ where $\mathcal{M}$ is class of sources. The primary goal of universal prediction is to find predictors that minimize (asymptotically) the predictability $\hat{\pi}_n(\mathcal{M})$ (i.e., they match asymptotically the optimal predictability that we denote as $\pi_n^*(\mathcal{M})$). However, among such predictors one looks for those that minimize the *redundancy*, r_n, defined as the difference between the average prediction error of our SPM predictor and the *optimal* predictor that always guesses the most likely symbol $r_n := \hat{\pi}_n(\mathcal{M}) - \pi_n^*(\mathcal{M})$.

In [1] we proved the following theorem.

Theorem 1 *Let $\alpha > \frac{1}{2}$ and K be fixed. Consider the Sampled Pattern Matching algorithm that predicts the next K outcomes of a sequence $X_1, \ldots, X_n$ drawn from a a Markov source $\mathcal{M}$. Then there exists $0 < \nu < \frac{1}{2}$ such that $r_n = \hat{\pi}_n(\mathcal{M}) - \pi_n^*(\mathcal{M}) = O(n^{-\nu})$ as $n \to \infty$.*

†The work of this author was supported by the NSF Grant CCR-0208709 and by NIH Grant R01 GM068959-01.



(a) (b)

Figure 1. The percentage of correct prediction using SPS for (a) *E.coli* **genome with noncoding and coding segments; (b)** *Mus musculus* **genome with introns and exons.**

2 Prediction of Some Molecular Sequences

We now briefly report on our experimental results. We did perform some experiments on DNA genome of *Mus musculus* and *E.coli* as well as protein *human adenovirus* sequences using the SPM predictor. In general, the amino acids are distributed quite non-uniformly in proteins while the four bases are almost equally likely to occur in a DNA sequence.

Our first experiment deals with the protein *human adenovirus* of a relatively short length of 807. The relative frequencies of amino acids $\{T, Y, \ldots, M\}$ vary from 0.7% to 11%. The first-order empirical entropy turns out to be $h \approx 2.66$, thus based on theoretical results [1] we expect the longest suffix to be of length $D_n \sim \frac{1}{h}\ln(807) \approx 2.5$ (cf. [2]). Indeed, in all 100 predictions we perform the length of the longest suffix is equal to 3. We choose $\alpha = 0.5$, and all markers are of length 2. We found the relative prediction success equal to 22% that is, we predicted correctly in one out of five times. If this seems to be an unimpressive number, we must point out that our Theorem 1 tells us this is the best one can expect. Indeed, having rigorously proved Theorem 1 we do not need to search for non-existing "philosophers' stone" of biology.

We did some additional experiments with the protein human adenovirus. We reduce the alphabet to ten (by lumping similar amino acids). Then the relative frequencies vary from 0.7% to 17% and the relative prediction success is 23%. Finally, we reduced the alphabet to a binary alphabet (i.e., polar/nonpolar alphabet) with relative frequencies 43% and 57%. This time the relative prediction success is 54%.

The second experiment is new and more interesting. We consider genomes of *E.coli* and *Mus musculus*. For the former we apply our SPS algorithm to a segment of about 1600 bps length that runs across noncoding ("junk or repetitive DNA) and coding segments. We expect different predictability for these segments since it is known that distributions of bases in coding and noncoding regions are different. In Figure 2(a) we report our results. We observe a change in predictability around 150th base which agree with the transition from noncoding to coding regions.

In Figure 2(b) we run our algorithm on a coding segment of *Mus musculus* with introns and exons regions. As before we expect some change in predictability, although exons might be too short to observe these changes. Nevertheless we see in Figure 2(b) some fluctuation of prediction between exons and intron regions. We conclude that prediction algorithms, including our SPM predictor, may find applications in finding change points of molecular sequences, however, more experiments are desirable.

References

[1] P. Jacquet, W. Szpankowski, and I. Apostol, A Universal Predictor Based on Pattern Matching, *IEEE Trans. Information Theory*, 48, 1462-1472, 2002.

[2] W. Szpankowski, *Average Case Analysis of Algorithms on Sequences*, Wiley, New York, 2001.