Quasiperiodic Biosequences and Modulo Incidence Matrices*

Honghui Wan[1†] and Enmin Song[2]

[1] *Computational Biology Branch, National Center for Biotechnology Information*
National Library of Medicine, National Institutes of Health
Building 38A, 8th Floor, 8600 Rockville Pike, Bethesda, Maryland 20894, USA

[2] *University of California at San Francisco, VA Medical Center, 116R*
4150 Clement Street, San Francisco, CA 94121, USA

Abstract

Algorithm development for finding quasiperiodic regions in sequences is at the core of many problems arising in biological sequence analysis. We solve an important problem in this area. Let $\mathbf{A}$ be an alphabet of size n and $\mathbf{A}^l$ denote the set of sequences of length l over $\mathbf{A}$. Given a sequence $S = s_1 s_2 \cdots s_l \in \mathbf{A}^l$, a positive integer p is called a *period* of S if $s_i = s_{i+p}$ for $1 \leq i \leq l-p$. S is called *p-periodic* if it has a *minimum period* p. Let $\Omega_l(p)$ denote the set of p-periodic sequences in $\mathbf{A}^l$. A natural measure of "nearness to p-periodicity" for S is the average Hamming distance to the nearest p-periodic sequence: $D(S) = min_{T \in \Omega_l(p)} D(S,T)$. If T is a sequence $\in \Omega_l(p)$ such that $D(S,T) = D(S)$, then T is called a nearest p-periodic sequence of S and S is called p-quasiperiodic associated with the score $D(S)$. This paper develops an efficient algorithm for finding a nearest p-periodic sequence of S by means of its modulo-p incidence matrix. Let $\alpha = (\alpha_1, \cdots, \alpha_n)$ and $\beta = (\underbrace{q+1, \cdots, q+1}_{r}, \underbrace{q, \cdots, q}_{p-r})$, where $l = \alpha_1 + \alpha_2 + \cdots + \alpha_n$ is a partition of l and q is the quotient and r is the remainder when l is divided by p. This paper shows that there exists a sequence in $\mathbf{A}^l$ whose modulo-p incidence matrix has row sum vector α and column sum vector β.

Keywords: Period; Nearest periodic sequence; Average Hamming distance; Modulo-p incidence matrix; Maximum flow; Network

*This work was supported by NIH grant Z01-NLM00025-11.

†Corresponding author. Tel: (301) 435-5917; Fax: (301) 435-2433; E-mail: hwan@ncbi.nlm.nih.gov

1 Introduction

The search for weak-repeated and quasiperiodic segments in sequences is an important problem in molecular biology. Actually, one of the most striking features of DNA and protein is the extent to which quasi-periodic segments occur in the genome. This is especially true of eukaryotes (higher-order organisms whose DNA is enclosed in a cell nucleus). For example, most of the human Y chromosome consists of quasi-periodic segments, and overall families of reiterated sequences account for about one third of the human genome ([2]).

It a general methodology in computational molecular biology to find biological function from some specific segments of a sequence ([2-9]), especially from those quasiperiodic segments. For example, we have presented a census of the internal quasi-periodic regions in all known proteins and drawn general conclusions about the role of quasi-periodicity in evolution of proteins. Motivated by biological sequence analysis, we in this paper investigate quasiperiodicity of sequences in combinatorial and algorithmic aspects.

In section 2, we mathematically define quasiperiodicity and the modulo incidence matrix of a sequence by means of the average Hamming distance to the nearest periodic sequence. Then we develop an efficient algorithm for finding a nearest periodic sequence of a sequence. In section 3, we give the number of M-equivalence class of a sequence and explore the relationship between a sequence in $\mathbf{A}^l$ and the modulo incidence matrix associated with a partition of l. In particular, we present a characterization of the modulo incidence matrix of a periodic sequence.

2 Quasi-periodicity

Let $\mathbf{A} = \{a_1, a_2, \cdots, a_n\}$ denote an alphabet in which a_i is called the letter of type i $(1 \leq i \leq n)$. Symbolic sequences are characterized by $\mathbf{A}$ and (usually) by a finite length l. One-dimensional strings play an important role in various fields, such as informatics, dynamical systems, biology, communication theory, linguistics, and psychology. Particularly, the digital information that underlies genetics, biochemistry, cell biology, and development can be represented by a simple string on a 4-letter alphabet (4 nucleotides: `A, C, G, T`) or a 20-letter alphabet (20 amino acids: `A, C, D, E, F, G, H, I, K, L, M, N, P, Q, R, S, T, V, W, Y`).

We denote by $\mathbf{A}^l$ the set of all sequences of length l over $\mathbf{A}$. For a sequence $S = s_1 s_2 \cdots s_l \in \mathbf{A}^l$, we denote by $S[i..j]$ the (contiguous) segment of S which starts at position i and ends at position j in S. A positive integer p $(p \leq l)$ is called a *(perfect) period* of S if $s_i = s_{i+p}$ for all i with $1 \leq i \leq l-p$. S holds *p-periodicity* or is called *p-periodic* if it has a *minimum period p*. As a simple example, `ACTACTACTACTAC` is 3-periodic.

Now we turn to investigating quasiperiodicity of sequences which has a special importance in sequence analysis. Let $\Omega_l(p)$ denote the set of all p-periodic sequences

in $\mathbf{A}^l$. Given a sequence $S \in \mathbf{A}^l$, a natural measure of "nearness to p-periodicity" for S is a distance to the nearest p-periodic sequence. The simplest distance is the *average Hamming distance*, i.e., the proportion of differences calculated simply by counting the number of residue differences divided by the length of the sequence. Hence, for two sequences S and T in $\mathbf{A}^l$, the average Hamming distance between S and T is

$$D(S,T) = \frac{k}{l},$$

where k is the number of characters that differ. If T is a sequence $\in \Omega_l(p)$ such that

$$D(S) =: min_{U \in \Omega_l(p)} D(S,U) = D(S,T),$$

then p is called a *quasi-period* of S with respect with T and $T[1..p]$ is called a *p-quasiperiodic unit* of S. In addition, T is called a *nearest p-periodic sequence* of S and $D(S)$ is called the *score* corresponding to T. For example, the DNA sequence $S =$ `ACACTACCACAC` has both a quasi-period 2 and a quasi-period 5. `AC`, `ACACT`, and `ACACC` are quasi-periodic units of S and the score corresponding to each of them is 0.25. Obviously, S has a perfect period 8.

The score corresponding to the nearest p-periodic sequence reflects the quasiness of p. In fact, the smaller is the corresponding score, the stronger is the quasi-period. Particularly, if the corresponding score is equal to zero, then the quasi-period coincides with the perfect period in the proper sense.

Let γ be a real number between 0 and 1. If there exists a sequence $R \in \Omega_r(p)$ such that $D(S[i..i+r-1], R) \leq \gamma$, then $S[i..i+r-1]$ is called a *p-quasiperiodic segment* of S associated with the score γ. For example, $S[6..11] =$`ACCACA` is 3-quasiperiodic segment of $S =$`ACACTACCACAC`. In this case, the sequence in $\Omega_6(3)$, corresponding to the segment $S[6..11]$ associated with the score $\frac{1}{6}$, is either $R =$ `ACCACC` or $R =$ `ACAACA`.

We focus on algorithm design for finding a nearest p-periodic sequence of a sequence. First, we define a function δ on $\mathbf{A} \times \mathbf{A}$ as

$$\delta(a_i, a_j) = \begin{cases} 1 & \text{if } i = j, \\ 0 & \text{otherwise.} \end{cases}$$

For a sequence $S = s_1 s_2 \cdots s_l \in \mathbf{A}^l$, we define an $n \times p$ matrix $M = (m_{ij})_{n \times p}$ in which

$$m_{ij} = \sum_{j=k \ (mod\ p)} \delta(a_i, s_k).$$

The matrix $M = (m_{ij})_{n \times p}$ is called the *modulo-p incidence matrix* of S.

Let us look at a DNA sequence: $S =$ `ACAGCTGACGTAG`. In this case, $\mathbf{A} =$ {`A`,`C`,`G`,`T`}. For $p = 5$, the modulo-5 incidence matrix is $M = (m_{ij})_{4 \times 5}$ as follows:

$$\begin{array}{c} \\ \texttt{A} \\ \texttt{C} \\ \texttt{G} \\ \texttt{T} \end{array} \begin{array}{c} \begin{array}{ccccc} 1 & 2 & 3 & 4 & 5 \end{array} \\ \begin{pmatrix} 1 & 1 & 2 & 0 & 0 \\ 0 & 1 & 0 & 1 & 1 \\ 0 & 1 & 1 & 1 & 1 \\ 2 & 0 & 0 & 0 & 0 \end{pmatrix} \end{array}. \tag{1}$$

The modulo-p incidence matrix of a sequence completely reflects complexity and symmetry of the corresponding sequence. It is a comprehensive "data structure" to represent the sequences in $\mathbf{A}^l$ and provides us with a conceptual framework for constructing a nearest p-periodic sequence of a sequence.

The algorithm to find the best matching p-periodic sequence of S is described below:

Algorithm 1. Given a sequence $S \in \mathbf{A}^l$ and a positive integer $p < l$, find a nearest p-periodic sequence T of S.
Step 1. Create the modulo-p incidence matrix of S: $M = (m_{ij})_{n\times p}$.

Step 2. Find the maximum value μ_j of elements in the jth column of M and pick up a character, say a_{i_j} in $\mathbf{A}$, corresponding to the maximum value ($j = 1, \cdots, p$).

Step 3. Make a p-quasiperiodic unit of S: $U = a_{i_1}a_{i_2}\cdots a_{i_p}$.

Step 4. Produce a nearest p-periodic sequence T of S:

$$T = \underbrace{U\cdots U}_{q}W,$$

where W is the prefix of U of length r, q is the quotient and r is the remainder when l is divided by p.

Theorem 1. *Algorithm 1 can be executed in $O(l + np)$ time.*

Proof. Step 1 takes $O(l)$ time, step 2 takes $O(np)$ time, step 3 takes $O(p)$ time, and step 4 takes $O(q+r)$ time. Thus, the time complexity of the algorithm is $O(l+np)$.

The following theorem gives the number of the nearest p-periodic sequences of a sequence of length l.

Theorem 2. *Let $M = (m_{ij})_{n\times p}$ be the modulo-p incidence matrix of $S \in \mathbf{A}^l$ and μ_j denote the maximum of elements in the jth column vector of M:*

$$\mu_j = max\{m_{ij} : 1 \le i \le n\}|.$$

Moreover, let r_j denote the number of those elements in the jth column vector of M achieving the maximum μ_j:

$$r_j = |\{i|m_{ij} = \mu_j, 1 \le i \le n\}|.$$

Then there are $\prod_{j=1}^{p} r_j$ nearest p-periodic sequences of S corresponding to the distance $1 - \frac{1}{l}\sum_{j=1}^{p}\mu_j$.

Proof. Let $I_j = \{i | m_{ij} = \mu_j, 1 \leq i \leq n\}$. We can arbitrarily choose an element i_j from I_j for each i with $1 \leq i \leq p$. There are totally $b_1 b_2 \cdots b_p$ choices for constructing p-quasiperiodic units of S. This means that we can make $b_1 b_2 \cdots b_p$ nearest p-periodic sequences of S. The number of characters that identify between S and each of the nearest p-periodic sequences is $\sum_{j=1}^{p} \mu_j$. Hence, The average Hamming distance between S and each of the nearest p-periodic sequences is $1 - \frac{1}{l} \sum_{j=1}^{p} \mu_j$.

By Theorem 2, we immediately have the following:

Corollary 3. *Let S be a sequence in $\mathbf{A}^l$ and $\mathbf{m}_j$ denote the jth column vector of M. Then the following three statements are equivalent:*

(1) S is p-periodic,

(2) $\sum_{j=1}^{p} \mu_j = l$,

(3) $\mathbf{m}_j = \mu_j \mathbf{e}_k$ for some k $(1 \leq j \leq p)$, where $\mathbf{e}_k$ denotes the k-th n-dimensional unit column vector.

As an example, let us go back to the sequence $S =$ `ACAGCTGACGTAG`. With a simple calculation, we obtain $\mu_1 = m_{41} = 2, \mu_2 = m_{12} = m_{22} = m_{32} = 1, \mu_3 = m_{13} = 2, \mu_4 = m_{24} = m_{34} = 1$, and $\mu_5 = m_{25} = m_{35} = 1$. Thus, we have $a_{i_1} =$ `T`; $a_{i_2} =$ `A`, `C`, or `G`; $a_{i_3} =$ `A`; $a_{i_4} =$ `C` or `G`; $a_{i_5} =$ `C` or `G`. There are 12 nearest p-periodic sequences of S:

`TAACCTAACCTAA`, `TAACGTAACGTAA`, `TAAGCTAAGCTAA`, `TAAGGTAAGGTAA`,
`TCACCTCACCTCA`, `TCACGTCACGTCA`, `TCAGCTCAGCTCA`, `TCAGGTCAGGTCA`,
`TGACCTGACCTGA`, `TGACGTGACGTGA`, `TGAGCTGAGCTGA`, `TGAGGTGAGGTGA`.

The number of the matching letters between the original sequence S and the nearest p-periodic sequences is 7, which is simply calculated as the sum of the maximum value in each column. The average Hamming distance between S and the nearest p-periodic sequences is 6/13.

3 Modulo-p incidence matrices

For $S \in \mathbf{A}^l$, we denote by α_i the frequencies of occurrences of a_i in S for $1 \leq i \leq n$. Obviously, $\alpha_1 + \alpha_2 + \cdots + \alpha_n = l$. The vector $\alpha = (\alpha_1, \cdots, \alpha_n)$ is called the *frequency vector* of S. Let $\beta = (\underbrace{q+1, \cdots, q+1}_{r}, \underbrace{q, \cdots, q}_{p-r})$, where q is the quotient and r is the remainder when l is divided by p. Then the modulo-p incidence matrix $M = (m_{ij})_{n \times p}$ of S has row sum vector α and column sum vector β:

$$\sum_{j=1}^{p} m_{ij} = \alpha_i, i = 1, \cdots, n;$$

$$\sum_{i=1}^{n} m_{ij} = \begin{cases} q+1 & \text{for } 1 \le j \le r, \\ q & \text{for } r < j \le p. \end{cases}$$

For the sequence $S =$ ACAGCTGACGTAG, we have the frequency vector $\alpha = (4,3,4,2)$ as well as $\beta = (3,3,3,2,2)$. Its modulo-5 incidence matrix in (1) has row sum vector $\alpha = (4,3,4,2)$ and column sum vector $\beta = (3,3,3,2,2)$.

The set of sequences with the modulo-p incidence matrix $M = (m_{ij})_{n\times p}$ is called *M-equivalence class*, denoted by $\mathcal{S}(M)$. For any sequence $S \in \mathcal{S}(M)$, there are $q+1$ characters at positions $j, p+j, \cdots$, and $qp+j$: m_{1j} of type 1, m_{2j} of type 2, $\cdots$, and m_{nj} of type n ($j = 1,2,\cdots,r$). The number of arrangements of these $q+1$ characters is $\binom{q+1}{m_{1j},\cdots,m_{nj}}$. Similarly, there are q characters at positions $j, p+j, \cdots$, and $(q-1)p+k$: m_{1k} of type 1, m_{2k} of type 2, $\cdots$, and m_{nk} of type n ($k = r+1,\cdots,p$). The number of arrangements of these q characters is $\binom{q}{m_{1k},\cdots,m_{nk}}$. Therefore, the number of sequences in $\mathcal{S}(M)$ is given by

$$|\mathcal{S}(M)| = \prod_{j=1}^{r} \binom{q+1}{m_{1j},\cdots,m_{nj}} \prod_{k=r+1}^{p} \binom{q}{m_{1k},\cdots,m_{nk}}. \tag{2}$$

For the matrix $M = (m_{ij})_{4\times 5}$ in (1), the size of $\mathcal{S}(M)$ is:

$$\binom{3}{1,0,0,2,}\binom{3}{1,1,1,0,}\binom{3}{2,0,1,0,}\binom{2}{0,1,1,0,}\binom{2}{0,1,1,0,} = 216.$$

A partition of a positive integer l is a representation

$$l = \alpha_1 + \alpha_2 + \cdots + \alpha_n. \tag{3}$$

The numbers $\alpha_1,\cdots,\alpha_n$ are the *parts* of the partition. Hence (3) is a partition of l into n parts.

Theorem 4. *Let $l = \alpha_1+\alpha_2+\cdots+\alpha_n$ be a partition of l. Let $\alpha = (\alpha_1,\cdots,\alpha_n)$ and $\beta = (\underbrace{q+1,\cdots,q+1}_{r},\underbrace{q,\cdots,q}_{p-r})$, where q is the quotient and r is the remainder when l is divided by p. Then there exists a sequence in $\mathbf{A}^l$ whose modulo-p incidence matrix has row sum vector α and column sum vector β.*

Proof. According to (2), it is enough to prove that there exists an $n\times p$ nonnegative integral matrix with row sum vector α and column sum vector β. We use the network flow approach to proving the theorem. The basics of network flow theory can be found in [1].

Matrices with row sum vector α and column sum vector β can be considered as maximum integral flows in the following network. The nodes consist of a source s, a sink t, and $a_1, a_2, \cdots, a_n, b_1, b_2, \cdots, b_p$. There is an arc from s to a_i with capacity α_i

for $i = 1, 2, \cdots, n$. There is an arc from b_j to t with capacity β_j for $j = 1, 2, \cdots, p$, where

$$\beta_j = \begin{cases} q+1 & \text{for } 1 \leq j \leq r, \\ q & \text{for } r < j \leq p. \end{cases}$$

Finally, there are arcs from a_i to b_j with capacity β_j for $i = 1, 2, \cdots, n$ and $j = 1, 2, \cdots, p$. Let f_{ij} be the flow from a_i to b_j and let $F = (f_{ij})$. We immediately deduce that the arc from s to a_i is saturated and thus α_i units of flow leave a_i. Hence $f_{i1}+f_{i2}+\cdots+f_{in} = \alpha_i$ and so the ith row sum of F is α_i. We deduce that $F = (f_{ij})$ has row sum vector α and column sum vector β. Conversely, from a matrix $F = (f_{ij})$ with row sum vector α and column sum vector β we can construct an integral maximum flow of size $l = \alpha_1 + \alpha_2 + \cdots + \alpha_n$. Thus there exists a matrix with row sum vector α and column sum vector β if and only if the network has a maximum flow of size l. We are using the well-known result that one can find a maximum integral flow in a network with integral capacities.

An arbitrary *cut* in the network is formed from a set X of nodes with $s \in X$ and $t \in \bar{X}$ (where the bar refers to the complement of the set). The cut is the set of arcs whose tail is in X and whose head is in $\bar{X}$. let our arbitrary set X consist of s with a subset of the nodes in $\{a_1, a_2, \cdots, a_n\}$ indexed by I and a subset of the nodes in $\{b_1, b_2, \cdots, b_p\}$ indexed by $\bar{J}$. The cut would consist of the arcs (s, a_i) for $i \in \bar{I}$, (a_i, b_j) for $i \in I, j \in J$ and (b_j, t) for $j \in \bar{J}$. The capacity of a cut is the sum of the capacities of the arcs in the cut. The max flow-min cut theorem ensures ([1]) that there exists a matrix with row sum vector α and column sum vector β if and only if no cut has capacity less than l:

$$\sum_{i \in \bar{I}} \alpha_i + \sum_{j \in \bar{J}} \beta_j + \sum_{i \in I, j \in J} \beta_j \geq l, \tag{4}$$

for all index sets $I \subset \{1, 2, \cdots, n\}$, $J \subset \{1, 2, \cdots, p\}$.

It is easy to know that the inequality always holds. Thus, there exists a matrix with row sum vector α and column sum vector β.

Combining Corollary 3 and Theorem 4, we obtain a characterization of modulo-p incidence matrix of a p-periodic sequence in $\mathbf{A}^l$.

Corollary 5. *Let* $\alpha = (\alpha_1, \cdots, \alpha_n)$ *and* $\beta = (\underbrace{q+1, \cdots, q+1}_{r}, \underbrace{q, \cdots, q}_{p-r})$, *where* $l = \alpha_1 + \alpha_2 + \cdots + \alpha_n$ *is a partition of* l *and* q *is the quotient and* r *is the remainder when* l *is divided by* p. *Then there exists a* p*-periodic sequence in* $\mathbf{A}^l$ *whose modulo-*p *incidence matrix has row sum vector* α *and column sum vector* β *if and only if each* α_i *can be expressed a linear integral combination of* $q+1$ *and* q*:*

$$\alpha_i = \lambda_i(q+1) + \kappa_i q, i = 1, 2, \cdots, n.$$

References

[1] L. R. Ford, Jr. and D. R. Fulkerson, *Flows in Networks*, Princeton University Press, Princeton, N.J., 1962

[2] E. McConkey, *Human Genetics: The Molecular Revolution*, Jones and Bartlett, Boston, MA, 1993.

[3] H. Wan, Weak-periods in biological sequences, *First SIAM Conference on Computational Science and Engineering (CSE00)*, 2000.

[4] H. Wan and J. C. Wootton, Axiomatic foundations of complexity functions of biological sequences. *Ann. Comb.*, 3 (1999), 105-127.

[5] H. Wan and J. C. Wootton, A global compositional complexity measure for biological sequences: *AT*-rich and *GC*-rich genomes encode less complex proteins, *Comput. Chem.*, 24 (2000), 71 - 94.

[6] H. Wan and J. C. Wootton, Algorithms for computing lengths of chains in integral partition lattices, *Theoretical Comput. Sci.*, (2001).

[7] H. Wan and J. C. Wootton, Detecting simple biological sequences by a new complexity measure. *J. Computational Biology* (2002).

[8] H. Wan and J. C. Wootton, Quasiperiodicity analysis of biological sequences and sequence databases, *Proc. Nat. Acad. Sci. USA*, accepted.

[9] J. C. Wootton, Simple Sequences of Protein and DNA, In: M. J. Bishop and C. J. Rawlings (eds), *DNA and Protein Sequence Analysis.*, Oxford University Press, pp. 169-83, 1997.