

Equivalence of two Fourier methods for biological sequences

Eivind Coward

Department of Mathematical Sciences, Norwegian University of Science and Technology, N-7034 Trondheim, Norway
e-mail: eivindc@math.ntnu.no

Received 4 December 1995

Abstract. Two methods for defining Fourier power spectra for DNA sequences or other biological sequences are compared. The first method uses indicator sequences for each letter. The second method by Silverman and Linsker assigns to each letter a vertex of a regular tetrahedron in space, and this can be generalized to any dimension. While giving different Fourier transforms, it is shown that the power spectra of the two methods are essentially the same. This is also true if one replaces the Fourier transform in both methods with another linear transform, such as the Walsh transform.

Key words: DNA sequences – Periodicity – Fourier and Walsh spectral analysis

1 Introduction

Several authors have studied methods for detecting periodicities in DNA or protein sequences. These sequences can be written as a string of letters from a finite alphabet, corresponding to 4 bases for a DNA sequence or 20 amino acids for a protein sequence. Other alphabets are also used, for example the two-letter purine-pyrimidine alphabet. A natural tool for studying periodicities in a sequence is the discrete Fourier transform, but the conventional Fourier transform is defined only for numerical sequences. One could assign a numerical value to each letter, but the result would then depend on the chosen values. Our Fourier spectrum should be independent of the labelling, so that the DNA sequences AAGC AAGC AAGC and GGTA GGTA GGTA would give the same result, for example. This is not possible when using a simple letter-number assignment. Stoffer et al. (1993) consider all such assignments for a given sequence and choose among these certain optimal values. A simpler approach is to assign vectors instead of scalar values. Two different ways of assigning vectors, both with the required symmetry, is the use of indicator sequences (Tavaré and Giddings, 1989) and the tetrahedral

representation by Silverman and Linsker (1986) for DNA sequences (that is, a four-letter alphabet), which can be generalized to an alphabet of n letters by using a $(n - 1)$ -simplex in $(n - 1)$ -dimensional space, as pointed out by Li and Kaneko (1992). The aim of this paper is to prove that the two methods yield essentially the same result. We will first give a short review of the two methods.

2 The indicator sequence method

Consider a sequence (a_k) of length N (with k running from 0 to $N - 1$) from the n -letter alphabet $\mathcal{A} = \{\alpha^{(1)}, \alpha^{(2)}, \dots, \alpha^{(n)}\}$. For each different letter $\alpha^{(j)}$ in \mathcal{A} we form an indicator sequence $(x_{j,k})_{k=0}^{N-1}$ such that

$$x_{j,k} = \begin{cases} 1 & \text{if } a_k = \alpha^{(j)}, \\ 0 & \text{otherwise.} \end{cases}$$

Let $(\hat{x}_{j,m})_{m=0}^{N-1}$ be the ordinary discrete Fourier transform of the j th indicator sequence $(x_{j,k})_{k=0}^{N-1}$:

$$\hat{x}_{j,m} = \frac{1}{N} \sum_{k=0}^{N-1} x_{j,k} e^{-2\pi i k m / N} \quad j = 1, 2, \dots, n.$$

The power spectrum (c_m) of the sequence (a_k) is then defined by

$$c_m = \sum_{j=1}^n |\hat{x}_{j,m}|^2.$$

This can also be formulated in terms of vectors in n -space, which is convenient for comparison with the simplex method. For each $k = 0, 1, \dots, N - 1$, let

$$\mathbf{x}_k = \begin{bmatrix} x_{1,k} \\ x_{2,k} \\ \vdots \\ x_{n,k} \end{bmatrix}.$$

Thus each letter $\alpha^{(j)}$ corresponds to a standard basis vector \mathbf{e}_j in \mathbb{R}^n . The Fourier transform operates componentwise on this sequence of vectors and can be written

$$\hat{\mathbf{x}}_m = \frac{1}{N} \sum_{k=0}^{N-1} \mathbf{x}_k e^{-2\pi i k m / N}, \quad (1)$$

where each \mathbf{x}_m is a vector in \mathbb{R}^n . The power spectrum is $c_m = |\hat{\mathbf{x}}_m|^2$.

As an example, we have computed the power spectrum for the first $N = 360$ bases of the DNA sequence GGPRION (Gabriel et al., 1992) from

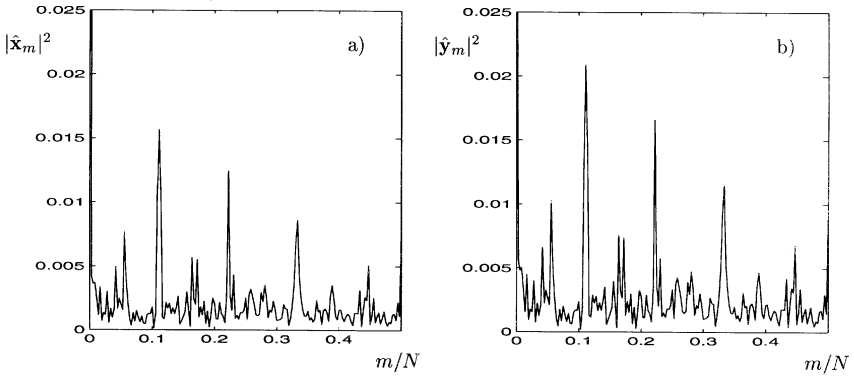


Fig. 1a,b. Fourier power spectra for the GGPRION DNA sequence, using the indicator sequence method in **a** and the simplex method in **b**

the EMBL database (release 42), see Fig. 1. Thus $n = 4$, and the alphabet is $\mathcal{A} = \{A, C, G, T\}$. This is a coding region containing nine imperfect tandem repeats of length 18 (6 amino acids), which gives peaks at $1/18 \approx 0.056$, and multiples of $1/18$. The peaks at $1/9 \approx 0.11$ and multiples of this are even more pronounced, and this indicates that the repeats are in fact almost 9-periodic on the DNA level.

3 The simplex method

Suppose for the moment that $n = 4$ (which is the case if (a_k) is a DNA sequence). Let \mathbf{r}_j , $j = 1, 2, 3, 4$, be unit vectors pointing to the vertices of a regular tetrahedron in \mathbb{R}^3 , and let each letter $\alpha^{(j)}$ (that is A, C, G and T) correspond to \mathbf{r}_j (see Fig. 2). The coordinates of \mathbf{r}_j depend of the orientation of the tetrahedron and the ordering of the vertices, but one possibility is (writing the \mathbf{r}_j as column vectors):

$$\mathbf{r}_1 = \begin{bmatrix} -\sqrt{6}/3 \\ -\sqrt{2}/3 \\ -1/3 \end{bmatrix}, \quad \mathbf{r}_2 = \begin{bmatrix} \sqrt{6}/3 \\ -\sqrt{2}/3 \\ -1/3 \end{bmatrix}, \quad \mathbf{r}_3 = \begin{bmatrix} 0 \\ 2\sqrt{2}/3 \\ -1/3 \end{bmatrix}, \quad \mathbf{r}_4 = \begin{bmatrix} 0 \\ 0 \\ 1 \end{bmatrix}.$$

The sequence (a_k) corresponds to an associated sequence (\mathbf{y}_k) of unit vectors, where $\mathbf{y}_k = \mathbf{r}_j$ when $a_k = \alpha^{(j)}$. Its Fourier transform is defined by

$$\hat{\mathbf{y}}_m = \frac{1}{N} \sum_{k=0}^{N-1} \mathbf{y}_k e^{-2\pi i k m / N},$$

exactly as in (1). The power spectrum is similarly $|\hat{\mathbf{y}}_m|^2$. It is independent of the choice of orientation and ordering of the vertices.

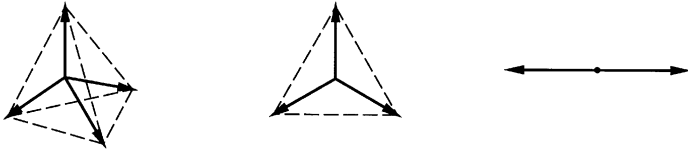


Fig. 2. Regular simplexes with unit vectors: 3-simplex (tetrahedron) to the left, 2-simplex (triangle) in the middle and 1-simplex (line segment) to the right

For a general alphabet size n we substitute our tetrahedron in \mathbb{R}^3 with a regular $(n - 1)$ -simplex in \mathbb{R}^{n-1} , which has n vertices at unit distance from the origin and all edge lengths equal. For example, a regular 2-simplex is an equilateral triangle in the plane, and a 1-simplex is simply a line segment (see Fig. 2). This gives n unit vectors \mathbf{r}_j corresponding to the n different letters. The Fourier transform and the power spectrum are defined as before, but now with each \mathbf{y}_k and $\hat{\mathbf{y}}_m$ in \mathbb{R}^{n-1} .

The power spectrum of the DNA sequence from the previous example is also computed using the simplex method, see Fig. 1. As we will see in the next section, the similarity between the two plots is no coincidence.

4 The connection between the two methods

The Fourier transform is necessarily different for the two methods (for the indicator sequence method it is a sequence of vectors in \mathbb{R}^n , while for the simplex method it is a sequence of vectors in \mathbb{R}^{n-1}). Nevertheless, it turns out that the power spectra are essentially proportional. More precisely, we have the following result.

Theorem 1 *Let $(a_k)_{k=0}^{N-1}$ be a sequence from the n -letter alphabet \mathcal{A} , and let (\mathbf{x}_k) and (\mathbf{y}_k) be the associated vector sequences in the indicator sequence method and the simplex method, respectively. Then the power spectra are related by*

$$|\hat{\mathbf{y}}_m|^2 = \frac{n}{n-1} |\hat{\mathbf{x}}_m|^2 \quad \text{for } m \neq 0, \quad (2)$$

$$|\hat{\mathbf{y}}_0|^2 = \frac{n}{n-1} |\hat{\mathbf{x}}_0|^2 - \frac{1}{n-1}. \quad (3)$$

Proof. The basis vectors $\mathbf{e}_1, \mathbf{e}_2, \dots, \mathbf{e}_n$ in \mathbb{R}^n are the vertices of a $(n - 1)$ -simplex, lying in a hyperplane Π given by $\sum x_j = 1$ (Coxeter (1973), pp. 120–122). The centroid of the simplex is $\mathbf{t} = \frac{1}{n}[1, 1, \dots, 1]$. When this simplex is translated by $-\mathbf{t}$ (moving the centroid to the origin), and then rotated and scaled appropriately, each vertex coincides with the corresponding \mathbf{r}_j (when \mathbb{R}^{n-1} is embedded naturally in \mathbb{R}^n).

Thus the correspondence between \mathbf{x}_k and \mathbf{y}_k can be expressed as a translation by $-\mathbf{t}$, followed by a rotation R (a $(n-1) \times n$ matrix), followed by a scaling by a factor a :

$$\mathbf{y}_k = aR(\mathbf{x}_k - \mathbf{t}). \quad (4)$$

If (\mathbf{x}_k) is any sequence of vectors in Π and \mathbf{y}_k is given by (4) for every k , then

$$\hat{\mathbf{y}}_m = \frac{1}{N} \sum_{k=0}^{N-1} aR(\mathbf{x}_k - \mathbf{t})e^{-2\pi imk/N} = \begin{cases} aR\hat{\mathbf{x}}_m & \text{if } m \neq 0 \\ aR(\hat{\mathbf{x}}_0 - \mathbf{t}) & \text{if } m = 0 \end{cases} \quad (5)$$

Since a rotation preserves the euclidean length of a vector,

$$|\hat{\mathbf{y}}_m|^2 = a^2|\hat{\mathbf{x}}_m|^2 \quad \text{for every } m \neq 0. \quad (6)$$

For $m = 0$ we have

$$|\hat{\mathbf{y}}_0|^2 = a^2|\hat{\mathbf{x}}_0 - \mathbf{t}|^2 = a^2(|\hat{\mathbf{x}}_0|^2 - 2\text{Re}\hat{\mathbf{x}}_0^*\mathbf{t} + |\mathbf{t}|^2).$$

Furthermore, we observe that

$$\begin{aligned} \sum_{j=1}^n \hat{x}_{j,m} &= \frac{1}{N} \sum_{k=0}^{N-1} e^{-2\pi imk/N} \sum_{j=1}^n x_{j,k} = \frac{1}{N} \sum_{k=0}^{N-1} e^{-2\pi imk/N} \\ &= \begin{cases} 0 & \text{if } m \neq 0, \\ 1 & \text{if } m = 0. \end{cases} \end{aligned} \quad (7)$$

This implies that

$$\hat{\mathbf{x}}_0^*\mathbf{t} = \frac{1}{n} \sum_{j=1}^n \hat{x}_{j,0} = \frac{1}{n},$$

whence,

$$|\hat{\mathbf{y}}_0|^2 = a^2 \left(|\hat{\mathbf{x}}_0|^2 - \frac{1}{n} \right). \quad (8)$$

The scaling constant a is the ratio between the side length $\sqrt{2n/(n-1)}$ of the normalized simplex in \mathbb{R}^{n-1} (see Coxeter (1973), pp. 294–295) and the corresponding side length of the simplex defined by the basis vectors \mathbf{e}_j in \mathbb{R}^n , which is $\sqrt{2}$. Thus $a = \sqrt{n/(n-1)}$, which gives equations (2) and (3) when substituted into (6) and (8). \square

Remarks:

1. The extra term $1/(n-1)$ for $m = 0$ can be interpreted as a bias in the sequence (\mathbf{x}_k) due to the fact that the basis vectors \mathbf{e}_j are not symmetrically distributed around the origin, in contrast to the simplex vertices \mathbf{r}_j .
2. Only the first $n-1$ components of $\hat{\mathbf{x}}_m$ have to be computed by the Fourier transform, then the last component is given by equation (7). (This was pointed out by Tavaré and Giddings (1989) for the Walsh transform.)

Two generalizations should be mentioned:

1. We assumed that (\mathbf{x}_k) was a sequence of indicator vectors, that is, one component is 1 and the others are 0. But in the proof we only need that all \mathbf{x}_k lie in the hyperplane Π , where the sum of the components of each \mathbf{x}_k is 1.

This could be used, for example, to transform sequences containing ambiguous or unknown letters, which could be represented by vectors whose components indicate the probability of each letter.

2. For concreteness, we have assumed that we used the ordinary discrete Fourier transform. Other transforms suitable for detecting periodicity may be used in the same way, for example the Walsh transform (Tavaré and Giddings, 1989). The same result still holds, precise conditions are formulated below.

Theorem 2 Let $(\mathbf{x}_k)_{k=0}^{N-1}$ be a sequence of vectors in \mathbb{R}^n such that the sum of the components of each \mathbf{x}_k is 1, and let \mathbf{y}_k be defined by (4). Let \mathcal{F} be a linear transform for scalar sequences, with the property that

$$\mathcal{F}(1, 1, \dots, 1) = (c, 0, 0, \dots, 0) \quad (9)$$

for some number c . Denote by $(\mathbf{X}_m) = \mathcal{F}(\mathbf{x}_k)$ the sequence of vectors obtained by applying \mathcal{F} to each component separately. Similarly, write $(\mathbf{Y}_m) = \mathcal{F}(\mathbf{y}_k)$. Then

$$|\mathbf{Y}_m|^2 = \frac{n}{n-1} |\mathbf{X}_m|^2 \quad \text{for } m \neq 0,$$

$$|\mathbf{Y}_0|^2 = \frac{n}{n-1} |\mathbf{X}_0|^2 - \frac{c}{n-1}.$$

Proof. For a general linear transform applied componentwise, equation (5) still holds, with the exponential (and the factor $1/N$) substituted by some coefficient $f_{m,k}$. The derivation in equation (7) becomes

$$\begin{aligned} \sum_{j=1}^n X_{j,m} &= \sum_{k=0}^{N-1} f_{m,k} \sum_{j=1}^n x_{j,k} = \sum_{k=0}^{N-1} f_{m,k} \\ &= (\mathcal{F}(1, 1, \dots, 1))_m = \begin{cases} c & \text{if } m = 0, \\ 0 & \text{if } m \neq 0. \end{cases} \end{aligned}$$

The rest of the proof is unchanged. \square

Remarks:

1. The requirement (9) on the transform \mathcal{F} seems to be reasonable for a transform which detects periodicity. For the Fourier and Walsh transforms one may consider it as a consequence of the orthogonality of the associated function bases.
2. The theorem can be generalized even further to infinite sequences and functions with continuous domains. However, this is probably not relevant for biological applications.

5 Conclusion

The power spectra provided by the indicator sequence method and the simplex method for Fourier transforms of biological sequences contain

exactly the same information. In fact, they are directly proportional, except for the 0th spectral coefficient. Therefore it seems reasonable to use the indicator sequence method, which is the simplest of the two methods.

Acknowledgements. I wish to thank Kristian Seip, Finn Drabløs and the referee for useful suggestions. This work is supported by a Ph.D. grant from the Norwegian Research Council.

References

1. Coxeter, H. S. M.: *Regular Polytopes*, 3rd ed. New York: Dover 1973
2. Gabriel, J. M., Oesch, B., Kretzschmar, H., Scott, M., Prusiner, S. B.: Molecular cloning of a candidate chicken prion protein. *Proc. Natl. Acad. Sci. U.S.A.* **89**, 9097–9101 (1992)
3. Li, W., Kaneko, K.: Long-range correlation and partial $1/f^z$ spectrum in a noncoding DNA sequence. *Europhys. Lett.* **17**(7), 655–660 (1992)
4. Silverman, B. D., Linsker, R.: A measure of DNA periodicity. *J. theor. Biol.* **118**, 295–300 (1986)
5. Stoffer, D. S., Tyler, D. E., McDougall, A. J.: Spectral analysis for categorical time series: Scaling and the spectral envelope. *Biometrika* **80**(3), 611–622 (1993)
6. Tavaré, S., Giddings, B. W.: Some statistical aspects of the primary structure of nucleotide sequences. In: M. S. Waterman (ed.): *Mathematical Methods for DNA Sequences* (pp. 117–131), Boca Raton, Florida: CRC Press (1989)