Fast approximate computation of non-uniform DFTs for biological sequence analysis

Julien Epps

Periodicity is emerging as a useful method for characterising the structure within biological sequences such as DNA. For sequence data, integer periods are usually of most interest, which poses the problem of fast computation if conventional Fourier-based analyses are applied. This paper adapts an existing complex polynomial re-evaluation algorithm, and proposes a fast approximation rule applicable to any discrete Fourier transform-based analysis where the frequencies to be evaluated are not uniformly spaced. Experiments evaluating binary signals on an integer-period frequency scale show that magnitude spectrum approximations differing from the exact magnitude spectrum by less than 1-3% can be obtained with a reduction in complexity of 3-10 times.

Introduction: Despite the existence of a number of period estimation techniques, to date the biology community has relied mainly on the Fourier transform and autocorrelation approaches for the periodicity characterisation of sequence data. Estimation of non-uniformly spaced frequencies, as required by the integer periods of interest in this problem, has been studied previously, producing for example the chirp z-transform [5] and the warped DFT [4]. Neither of these methods allow evaluation at an arbitrary set of frequencies, however. Recently, an integer period discrete Fourier transform
was proposed, however due to its non-uniform frequency spacing, to date fast computation has been restricted only to symbolic data \[1\]. In this letter, we address the problem by demonstrating the applicability of a complex polynomial evaluation technique and proposing an approximation rule, based on this, that allows fast computation of the discrete Fourier transform at arbitrary non-uniformly spaced frequencies.

**Fourier transform computation as polynomial evaluation:** The discrete Fourier transform (DFT) of a discrete-time signal \(x[n]\) of length \(N\), is defined as

\[
X[k] = \sum_{n=0}^{N-1} x[n] \exp\left(-j \frac{2\pi nk}{N}\right), \quad k = 0, 1, ..., N-1, \tag{1}
\]

and evaluates the discrete-time Fourier transform (DTFT) at the digital frequencies \(\frac{2\pi k}{N}\). By contrast, the integer period DFT (IPDFT) \[1\]

\[
X_{IP}[p] = \sum_{n=0}^{N-1} x[n] \exp\left(-j \frac{2\pi n}{p}\right), \quad p = 1, ..., P < N, \tag{2}
\]

evaluates the DTFT at frequencies \(\frac{2\pi}{p}\). In both (1) and (2), the Fourier transformation can be recognised as degree \(N-1\) (complex) polynomial evaluation, where the polynomial coefficients are \(x[n]\) and the evaluation points are \(\exp\left(-j \frac{2\pi k}{N}\right)\) (DFT) and \(z_{p} = \exp\left(-j \frac{2\pi}{p}\right)\) (IPDFT). The transformations can also be expressed in terms of matrix multiplication, and it is the Vandermonde recursive-block structure of the DFT transformation matrix that leads to FFT computation in \(O(M\log N)\) time. For the IPDFT, there is no recursive-block structure, however the transformation matrix is still Vandermonde, permitting evaluation in \(O(M\log^{2} N)\) time \[3\]. If \(P << N\), which
can occur in the case of long DNA sequence analysis, then multipoint polynomial evaluation, comprising $\left\lceil N/P \right\rceil$ smaller polynomial evaluations, yields a solution of complexity $O(M \log^2 P)$ [6]. To achieve lower complexity, we propose using the DFT and then interpolating these values for use at non-uniformly spaced frequencies by polynomial re-evaluation.

**Approximate multipoint complex polynomial re-evaluation:** Given a general set of points $\{w_k \in \mathbb{Z} \mid k = 0, 1, \ldots, N-1\}$ at which a polynomial has been evaluated and values $\{y_k \in \mathbb{Z} \mid k = 0, 1, \ldots, N-1\}$ of the polynomial at those points, complex multipoint polynomial re-evaluation can be achieved by evaluating

$$R(z) = \psi(z)\sigma(z)$$  \hspace{1cm} (3)

at desired re-evaluation points $\{z_p \in \mathbb{Z} \mid p = 1, \ldots, P\}$, where

$$\psi(z) = \sum_{k=0}^{N-1} \frac{c_k}{z - w_k},$$ \hspace{1cm} (4a)

$\{c_k \in \mathbb{Z} \mid k = 0, 1, \ldots, N-1\}$ are constant weights given by

$$c_k = \frac{y_k}{d\sigma \bigg|_{z=w_k}}$$ \hspace{1cm} (4b)

and

$$\sigma(z) = \prod_{k=0}^{N-1} (z - w_k).$$ \hspace{1cm} (5)

A theorem due to Reif [6] states that given a degree $N-1$ polynomial $R(z)$ evaluated at points $w_k$, $P$ re-evaluation points $z_p$ on or inside the unit circle, and $\varepsilon > 0$, then an $\varepsilon$-approximate polynomial evaluation can be computed within $O(P \log^2 r + M \log N)$, where $r = O \left( \log \left( \frac{\sum_{k=0}^{N-1} c_k}{\varepsilon} \right) \right)$. 


**Fast approximate non-uniform DFT evaluation:** In this paper, we evaluate the DTFT at non-uniformly spaced frequencies, based on an initial DFT computation. In this case, \( y_k = X[k] \) as given by equation (1), and \( w_k = \exp\left(-j \frac{2\pi k}{N}\right) \). Since \( w_k \) are the \( N \) complex roots of unity, \( \sigma(z) = z^N - 1 \), so that \( \frac{d\sigma}{dz}\big|_{z=w_k} = N \exp\left(j \frac{2\pi k}{N}\right) \), and equation (3) can be stated as

\[
R(z) = \sigma(z)\psi(z) = \frac{(z^N - 1)}{N} \sum_{k=0}^{N-1} \left( \frac{X[k]\exp\left(-j \frac{2\pi k}{N}\right)}{z - \exp\left(-j \frac{2\pi k}{N}\right)} \right)
\]  
(6)

The proposed approximate evaluation method operates by dropping terms from \( \psi(z) \) for which \( |X[k]| \) is small, or for which \( |z - \exp\left(-j \frac{2\pi k}{N}\right)| \) is large. Since in general the frequencies to be evaluated would not change from one analysis frame to the next, while \( |X[k]| \) does, we take the latter approach, for each \( k \) retaining the \( \gamma \times 100\% \) smallest terms of \( |z - \exp\left(-j \frac{2\pi k}{N}\right)| \), i.e. on a frequency-dependent basis. Denoting the sets of retained terms as \( \Psi_p \), where the index \( p \) represents the \( p \)'th non-uniform frequency to be evaluated, an approximate \( \psi(z) \) is given by

\[
\tilde{\psi}(z) = \sum_{k \in \Psi_p} \left( \frac{X[k]\exp\left(-j \frac{2\pi k}{N}\right)}{z - \exp\left(-j \frac{2\pi k}{N}\right)} \right),
\]  
(7)
while the corresponding approximated version of the desired non-uniform
discrete Fourier transform $R(z)$ is $\tilde{R}(z_p)$.

Summarising, the proposed algorithm for approximation at the complex values
$z_p = e^{j\theta_p}$ representing frequencies $\theta_p$, $p = 1, 2, ..., P$, is as follows:

(i) pre-compute $z^N - 1$ for each $z_p$ and 
$$\frac{\exp(-j\frac{2\pi k}{N})}{z - \exp(-j\frac{2\pi k}{N})} \forall z_p, k;$$

(ii) pre-compute $\left| z - \exp\left(-j\frac{2\pi k}{N}\right) \right|$ for each combination of $k$ and $z_p$, and

hence produce the sets $\Psi_p$ by retaining the $\gamma \times 100\%$ smallest terms for
each value of $z_p$;

(iii) for each analysis frame

(a) compute $X[k]$ using the FFT

(b) for each $z_p$, calculate $\sigma(z_p)$, $\tilde{\psi}(z_p)$ using (7) and hence $\tilde{R}(z_p)$.

This represents the computation, in addition to the FFT, of $(\gamma + 1)N$
multiplications and $\gamma N$ additions per evaluation point.

**Evaluation:** In this section, we evaluate the processing speed and accuracy of
the approximate computed spectrum relative to direct computation methods,
using MATLAB. The accuracy is measured in terms of the mean-square error
(MSE) between the approximate magnitude spectrum and the exact (directly
computed) magnitude spectrum, and all processing speeds are shown for a
Core 2 Duo 2.8GHz CPU with 4GB RAM.
Firstly, the MSE was examined as a function of frequency, for non-uniform frequencies $\frac{2\pi(k + \frac{m}{p})}{N}, k = 0,1,.., N - 1, m = 1,2,.., M - 1$, between the DFT evaluation frequencies $\frac{2\pi k}{N}$, to obtain a general understanding of the spectral characteristics of the approximation error. This was repeated for 500 instances of unity variance white Gaussian noise, and for each frequency (with $N = 512$, $M = 10$) and value of the approximation constant ($\gamma = 0.9$, 0.5, 0.2 and 0.05), the maximum (i.e. worst-case) MSE was recorded. Although results are not shown due to space limitations, the MSE was found to attain its maximum value halfway between DFT evaluation points (where the approximation is furthest from the exactly evaluated spectrum), and the log maximum MSE between adjacent pairs of DFT points was reasonably consistent across the spectrum for a given value of $\gamma$.

Secondly, the MSE was investigated as a function of the computation time, by varying the approximation constant $\gamma$ to produce different trade-offs between accuracy and processing time, for non-uniform integer-period frequencies $\frac{2\pi}{p}, p = 2,3,.., N - 1, \frac{N}{p} \in \mathbb{Z}$. This was done for four different example synthetic signal classes: (i) a single sinusoid, of randomly selected frequency; (ii) five sinusoids, of randomly selected frequencies and amplitudes; (iii) white Gaussian noise; and (iv) a random binary sequence. Other numerical representations exist for symbolic biological sequences, such as complex numbers, however due to the linearity property of the DFT, we consider only the binary indicator sequence representation, with minimal loss of generality [2]. The experiment was repeated for 100 instances of each signal type, the
approximate IPDFT magnitude spectrum was calculated based on a 512-point FFT, and then the maximum (i.e. worst-case) MSE was recorded. Although the subject of this paper is fast computation, note that the IPDFT has previously also been applied to real DNA sequence data [2].

Figure 1 shows that significant complexity savings can be achieved using the proposed method to approximate the IPDFT instead of direct computation, with minimal impact on the magnitude spectrum accuracy. For example, to compute the IPDFT to within 1% error (in terms of signal power), i.e. $\gamma = 0.2$, requires at most 0.011 seconds for white noise or random binary sequences or 0.022 seconds for sinusoidal signals, while exact computation ($\gamma = 1$, zero MSE) requires 0.026s using the proposed approach and direct computation requires 0.034 seconds. Poorer approximations for sinusoidal signals occur mainly for signal frequencies in the range $\left(\frac{\pi}{p}, \pi\right)$, over which $\sigma(z)$ and $\psi(z)$ are not computed even once for integer-period frequencies $\frac{2\pi}{p}$, $p = 1, \ldots, p < N$, however use of the IPDFT for such signals is practically unlikely. For the more typical application of analysing binary signals from biological sequence data, however, the proposed technique can achieve approximation errors in the range 1-3% within 0.0043s, a relative improvement of nearly ten times compared with direct computation.

Conclusion: A technique for fast approximate computation of non-uniform discrete Fourier transforms has been proposed. The technique has been shown to be capable of reducing the computational complexity of calculating integer-period frequencies by one-third to two-thirds of the computation time.
required by direct evaluation, depending on the signal type and allowable error in the resulting magnitude spectrum. There is thus considerable promise for bioinformatics applications, in which fast periodicity analysis of large sequences is a problem of interest. Future work will consider other types of non-uniform discrete Fourier transforms, and applications of the approximate computation in areas such as DNA correlation analysis.

Acknowledgement: This research was supported by a University of New South Wales Faculty of Engineering Early Career Research Grant for genomic signal processing, 2009.

References


Author affiliations:

J. Epps (School of Electrical Engineering, The University of New South Wales, Sydney NSW 2052 Australia, (j.epps@unsw.edu.au) and National ICT Australia, Eveleigh, NSW 1430 Australia)

Figure 1. Maximum mean-square error between magnitude spectral estimates calculated using the proposed approximation approach and calculated exactly, as a function of computation time for single sinusoids (dashed), five sinusoids (dotted), white Gaussian noise (dashed-dotted) and random binary sequences (solid) of length $N = 512$. 