Published in final edited form as:

Proc Int Symp Symb Numer Algorithms Sci Comput. 2017 ; 19: 397–400. doi:10.1109/SYNASC.
2017.00072.

Detecting proteine coding regions using a customized multi-scales splines construction

Darian M. Onchis

Faculty of Mathematics and Computer Science, West University of Timi oara, Timi oara,
Romania; Faculty of Mathematics, University of Vienna, Vienna, Austria

Abstract

In this experimental paper, we propose a method to detect proteins coding regions using a multiresolution wavelet-like system, constructed in multi-window splines-type spaces. The method is based on considering only the generators that prominently intersect the spectrum of the DNA (Deoxyribonucleic acid) sequence and on analysing jointly three different features namely time, scale and frequency. Numerical comparison and experimental results are presented.

I Introduction

The identification of coding regions in DNA sequences is part of the main trends in bioinformatics research [1], [2], [3], [4], [5]. The precise identification of coding regions, is useful for knowing the exact structure and functionality of a protein and also for studying evolution and for constructing phylogenetic trees.

Prokaryotic and eukaryotic cells both have DNA as the basis for their genes. While in prokaryotes cells, we find long and continuous coding genes, in eukaryotes cells, the genes display coding parts (i.e. exons) disconnected by extended noncoding parts (i.e. introns).

In the scientific literature, the methods to detect the protein coding regions are either model-dependent methods [5] or model-independent methods [1], [2], [3], [4]. In this paper, we propose a model-independent method inspired by the scaling characteristics of wavelets algorithms and implemented for efficiency reasons using modulated multi-window spline-type systems. The model-independent methods, are based on a proper use of the Fast Fourier Transform and on the three-base periodicity (TBP) property. This property was observed for most of DNA sequences, and it simply means that the dominant part in coding regions of biological sequences exhibits a TBP. In order to detect these approximative repetitions of the TBP, spectral analysis techniques were efficiently used in several papers [2], [4].

DNA sequences are composed of 4 nucleotides: adenine (A), guanine (G), thymine (T), and cytosine (C) and they have a double helical structure, with the two individual strands linked by complementary bases i.e. A and T are complementary and G and C are likewise complementary. In order to obtain a numerical signal, we have considered a DNA sequence

in FASTA format and we have applied a binary mapping to it. The mapping is the characteristic function, because it is useful to indicate the relative position of each nucleotide in the genome. Therefore, we obtain 4 binary sequences associated with the nucleotides alphabet and without losing any harmonic feature. The 3-periodicity property states that the spectral energy of the sum of the 4 binary signals representing a DNA protein coding region of length N , exhibits a peak at discrete frequency $k = N/3$.

Wavelet analysis is using multiresolutions in order to cope both with small protein coding regions and large regions. The wavelet-like procedure proposed in here, makes use of the advantages of both the wavelet transform (scaling) and Gabor transform (modulations), while it is implemented for efficiency reasons using a multi-window spline-type scheme. Since TBPs present the same frequency content at distinct scales, the transform will use as dilations only the modulations that interest the spectrum. In this way, the multi-window Gabor or the traditional full scale multiresolution approaches are both optimized. Therefore, our method is suitable to be used with larger DNA sequences, taken from GenBank, EMBL etc. with the new generation of sequencing.

II Splines reformulation of multiscale wavelet-like analysis

We will work with locally compact Abelian groups and we define the left shift as the operator acting on a function or distribution f defined over the group G as

$$L_y f(x) := f(y^{-1}x), \quad y, x \in G.$$

In ST (spline-type) theory, modulation is exactly the shift in the Fourier domain over the dual group. For each $\hat{x} \in \hat{G}$ we define

$$M_{\hat{x}} f := \langle \cdot, \hat{x} \rangle f \quad \text{and} \quad \widehat{M_{\hat{x}} f} = \hat{L}_{\hat{x}} \hat{f}$$

where we have denoted with $\hat{L}_{\hat{x}}$ the left shift over the LC group $\hat{G}$. The usual frequency interpretation for the group $G = \mathbb{R}^d$ becomes straightforward once the particular form or the related characters are considered.

To define the contraction (or delay) operator is more complicated, since it doesn't fit as well as the modulation operator with the group structure. To mimic the case of the real group, we would need to give the LC group an external product with scalar e.g. a vectorial structure. Given $a > 0$ we could define

$$D_a f := f(a \cdot)$$

in such a way that the natural framework where to build our generating set is the class of Radial Basis Functions (RBFs). We remember that a RBF is a function defined between Banach spaces, $\phi: \mathbb{B}^1 \rightarrow \mathbb{B}^2$, such that there exist a function $\varphi: \mathbb{R}_+ \rightarrow \mathbb{B}^2$ for which

$$\phi(x) = \varphi\left(\|x\|\right) \quad \forall x \in \mathbb{B}^1$$

For our experimental purpose, it is convenient to use generalized spline type systems (GST) [22], allowing the construction of ST space as infinite unions of ST spaces, each of them generated by a unique function (in the wavelet case a different contraction) and a different subgroup.

Wavelet and Gabor systems can be reformulated as ST systems generated by one window function and the shifts over a subgroup of the affine and Heisenberg group respectively. The first group leads to a time-scale shift, while the representation of the latter is interpreted as the classical time-frequency shift. With the GST construction, we unfortunately lose this behaviour for both Gabor and wavelet, but this is not such a big drawback since the group-representation construction leads to lack of flexibility in the inversion of the related convolutions. More interesting is the reformulation of wavelet and Gabor systems as multiwindows spline type spaces. In [19] we have shown that the Gabor case can be exactly reformulated in a biorthogonal construction problem; this construction becomes a powerful and easy to use tool in signal analysis. This is due to the favourable case for which it is possible to periodize both time and frequency domain, from the ST perspective, the original LC group and its dual. The same is not possible for wavelet system since the collection of generating functions is infinite, not periodizable, and defined over different lattices. Wavelet can be indeed defined as a collection of function $\{\phi_i\}_{i \in I}$ which are characterized by being the contraction of a unique mother function ϕ and their shifts over a related lattice H_i such that

$$H_i \subset H_{i+1}$$

The wavelet system becomes

$$\left\{ L_x(D^{(i)}\psi) : i \in I, x \in H_i \right\}$$

in the same way a Gabor systems can be defined as the collection of shifts over the same lattice of different modulations

$$\left\{ L_x(M_\omega\phi) : x \in H, \omega \in K \right\}$$

being K a discrete subset of $\hat{G}$. For any arbitrary set of functions $X = \{x_n\}$, not only for the particular case of spline type systems, we can define the so-called system operator.

$$Sf := \sum_n \langle f, x_n \rangle x_n = U_X V_X f$$

This is called frame operator elsewhere, but we always prefer to maintain the word *system* to approach the general case in which the collection is not a frame.

For the aforementioned ST spaces

$$Sf := \sum_i U_{\phi_i, H_i} V_{\phi_i, H_i} f$$

being ϕ_i different modulations and H_i equal to a unique lattice, or contractions and a monotone sequence of lattices, in the case of Gabor and Wavelet system respectively. The Gabor system operator commutes with both shift in time and frequency. This is not the case for the wavelet for which only the shift in time commutes, and not the shift in scale.

In both cases, if the system operator is invertible, it is possible to build the functions $S^{-1}\phi_i$, called the canonical duals, the related ST system behaves in each lattice as biorthogonal.

The inversion of the system operator in the Gabor case is possible once it is a frame; then it is computationally feasible only for the rational sampling case, where the Walnut representation or other analogue representation such as the one obtained through the Zak transform are employed. This leads to a numerical scheme that is the analogue to the one obtained through the spline type reformulation, since the system operator is partitioned in a similar fashion to obtain a fast code. Under this perspective biorthogonal and dual window are highly correlated, but the first can achieve the same fast algorithm for a much wider class of ST systems.

Problem occurs for the wavelet case since the frame operator does not commutes with the scale shift operator, so the usual dual construction $S^{-1}(L_x D_a \phi) = L_x D_a S^{-1} \phi$ is not possible and a dual has to be computed at each level. On the other hand the reformulation in GST theory leads to good result for theoretical purpose [22] but displays a huge raise in the complexity of the related algorithm [22]. For this reason we decided to use the wavelet scheme in its typical multiresolution reformulation. This approach was used since the early develop of wavelet analysis, especially in image analysis. Starting from the coarsest lattice the image is analysed through a low frequency window and approximated by a linear combination of the shifts of this atom, usually through a collocation scheme; then at each subsampled lattice, the same procedure is performed on the reminder of the preceding level through a related contraction.

For the approximation of the signal at each scale we can use the biorthogonal construction for a singular window, and the reproduction formula to express the projection of the signal over the ST space $\mathcal{S}(D_{a_i}, H_i)$.

But our interest is focussed on the construction of a system that can analyse different features of a signal, and not only the scale and frequency taken alone.

Using multiwindow spline type spaces, we can improve the accuracy with a small drawback in complexity. Fixed a scale, we chose to modify the window obtaining different atoms, having the same scale.

This positions our work as a numerical scheme that can handle 3 different features analysis:

- Time = through the ST structure
- Scale = through the multiresolution iteration
- Frequency = through the multiwindow nature.

A Multiresolution Analysis

Wavelet transform was, since its early development in the 50' a powerful tool to counterpoint the Fourier transform. In 1987 the field had a huge development through the seminal paper of Mallat [27]. In there, the author gave the mathematical justification of multiresolution analysis (MRA).

MRA is based on the construction of an ascending chain of ST spaces in $L^2(\mathbb{R}^d)$

$$V_{-\infty} \subset \dots \subset V_{-1} \subset V_0 \subset V_1 \subset \dots \subset V_{\infty} = L^2(\mathbb{R}^d)$$

each of them generated by a scaling function $D_a^j \phi$ over the lattice $a^j \mathbb{Z}^d$.

At each level j is then considered W_j the orthogonal complement of V_j in V_{j+1} . Therefore, it is possible to consider any element of $L^2(\mathbb{R}^d)$ as expansion

$$L^2(\mathbb{R}^d) = V_{j_0} \oplus W_{j_0} \oplus W_{j_0+1} \oplus \dots$$

for an arbitrary fixed index j_0

B The ST multiresolution algorithm

As shown in [19], it is possible to approximate a signal in ST spaces considering only that generators that better intersect its spectrum, since the characterization of ideal on ST space relies in the corresponding Fourier domain. The first step to employ is to consider the low frequency content of the signal. In traditional wavelet construction this role is covered by the scaling function. We use $\Phi := \{M_{\hat{x}_i} \hat{\phi}\}$, some modulations of a low frequency Gaussian, is such a way that only the 3 ÷ 4% of the Fourier domain is covered.

To handle the three different behaviours that we stated above, we have to modify the multiresolution algorithm building at each scale a consistent generating set for which the corresponding ST space has minimal distance with the given signal. Because the characterization of a ST space relies in its spectrum, we cover the whole spectrum through a modified $\bar{M}$ modulation operator which will be exploited in Section III. After the whole set of generating atoms are built, we compute the correlation through the semi-product

compatible to the selected Banach space, e.g. we simply rank the outcome of the multiwindow-analysis operator.

We can then select the best atoms according to a fixed threshold or fixing the numbers of element of the generating set. We prefer the latter since it would be possible to compute the complexity of the algorithm on the fly.

After the atoms are selected, we check whether they are linearly independent and we discard the ones that give us inconsistency and replace them with other atoms having slightly smaller correlation.

Finally, we can build the corresponding biorthogonal system and store the analysis coefficient to compute the reproduction formula.

It is important to notice that we have to check the linear independence of the Fourier transforms of the atoms only for a singular element of $\hat{G}/H^\perp$, due to periodicity: by definition $\dot{x} = \dot{y}$ iff $xH = yH$.

III Numerical experiments

For comparison reasons with existing work in the scientific literature, we considered the analysis of DNA sequences of eukaryotics that has been processed for detecting coding regions. In particular, we have used the sequence F56F11.4 of *C. elegans* of 8,000 bp, which contains five coding exons in positions 928-1039, 2528-2857, 4114-4377, 5465-5644, and 7255-7605 (GenBank access number AF099922 and positions 7021-15020). This sequences has been also analyzed in [2].

For the detection of coding regions in the DNA sequence, we have first computed the transform coefficients for all the 4 binary signals corresponding to the 4 nucleotides. Then, we have computed the squared spectral sum in order to have a common signal S . For the identification of prospective coding regions, S is projected onto the position axis and local maxima region of the projections are delimited. In order to exclude low-value coefficients, a threshold depending on the maximum value of the data has been set for the projection.

A function that is often used in wavelet scheme is the following radial function (centered in $N/3$)

$$\varphi(x) = \left(1 - \left\|x\right\|^2\right)e^{-\frac{\left\|x\right\|^2}{2}} \quad x \in \mathbb{R} \quad (1)$$

Because we are working with digital signals, we consider a sampled and periodized time domain having the length equal with the one of the DNA sequence.

The ST scheme has the great flexibility to work with modulation as different generating atoms, so we can build a generating set that does not rely on the modulation operator. For

this reason, we decided to modify the shift in frequency in the following way: we shift the positive and negative parts of the Fourier transform in opposite direction.

We build a modified modulation operator

$$\bar{M}_{\hat{x}} = IFT\left(\hat{L}_{\hat{x}}^{FT}(f)\left(0, \dots, \frac{L}{2}\right) + \hat{L}_{\hat{x}}^{FT}(f)\left(\frac{L}{2} + 1, \dots, L\right)\right)$$

This will be done at each wavelet level of the multiresolution analysis, not on the scale level.

In this way, we can consider different frequency features of a signal. By simple algebraic computation we can check that this modified modulation gives us a new window that possess the same decay of the original window, so same scale parameter. To overcome the problem of badly localization in low frequency, we decided to complete the scheme using as scaling function some low frequency Gaussians, e.g. small traditional modulation of a slow decay normalized Gaussian. The results of applying the procedure for the spectrum of the DNA sequence is presented in Figure 1.

Let Γ be a set of windows having different scale features suitable for the analysis of signal in $\mathbb{C}^L$. The multiresolution ST procedure has the time complexity:

$$\mathcal{O}(\text{card}(\Gamma) \cdot L)$$

IV Conclusions

We presented in this paper a procedure to identify protein coding regions based on wavelet-like systems constructed in multiwindow spline-type spaces. The method is model-independent and it has the advantages of using a smaller number of modulation than the classical Gabor transform and also scaling flexibility through the multiresolution construction. The results for the DNA sequences tested, were accurate and the detected regions are in the range of the ones presented in the literature. The computational advantage of our method is easy observable since we are not using all the modulations needed but only a sequence/signal dependent selection.

Acknowledgment

The author gratefully acknowledge the support of the Austrian Science Fund (FWF): project number P27516.

References

- [1]. Hassani Saadi, Hamed; Sameni, Reza; Zollanvari, Amin. Interpretive time-frequency analysis of genomic sequences. BMC Bioinformatics-BMC series - open, inclusive and trusted. 2017; (Suppl 4):154.
- [2]. Mena-Chalco JP, Carrer H, Zana Y, Cesar RM Jr. Identification of protein coding regions using the modified Gabor-wavelet transform. IEEE/ACM Trans Comput Biol Bioinform. 2008 Apr-Jun; 5(2):198–207. [PubMed: 18451429]
- [3]. Subbanna N, Zeevi Y. Macromolecular sequence analysis using multiwindow Gabor representations. Signal Processing. 2008; 88(4):877–890.

- [4]. Machado JA, Costa AC, Quelhas MD. Wavelet analysis of human DNA. *Genomics*. 2011 Sep; 98(3):155–63. [PubMed: 21672622]
- [5]. Guigo, R. DNA composition, codon usage and exon prediction *Nucleic acid databases*. Bishop, M, editor San Diego, CA: Academic Press; 1999. 53–80.
- [6]. Aldroubi, Akram; Feichtinger, Hans. Complete iterative reconstruction algorithms for irregularly sampled data in spline-like spaces. *Acoustics, Speech, and Signal Processing*, 1997. ICASSP-97., 1997 IEEE International Conference on; IEEE; 1997. 1857–1860.
- [7]. Aldroubi, Akram; Feichtinger, Hans. Exact iterative reconstruction algorithm for multivariate irregularly sampled functions in spline-like spaces: the L^p -theory. *Proceedings of the American Mathematical Society*. 1998; 126(9):2677–2686.
- [8]. Bownik, Marcin. The structure of shift-invariant subspaces of $L^2(\mathbb{R}^n)$. *Journal of Functional Analysis*. 2000; 177(2):282–309.
- [9]. Deboor, Carl; DeVore, Ronald A; Ron, Amos. The structure of finitely generated shift-invariant spaces in $L^2(\mathbb{R}^d)$. *Journal of Functional Analysis*. 1994; 119(1):37–78.
- [10]. Eldar, Yonina C. Sampling without input constraints: Consistent reconstruction in arbitrary spaces *Sampling, Wavelets, and Tomography*. Birkhäuser; Boston: 2004. 33–60.
- [11]. Feichtinger, Hans G. Spline-type spaces in Gabor analysis *Wavelet analysis*. Vol. 1. Hong Kong: 2002. 100–122. 2001
- [12]. Feichtinger, Hans G, Strohmer, Thomas, editors *Gabor analysis and algorithms: Theory and applications*. Springer Science & Business Media; 2012.
- [13]. Gabor, Dennis. Theory of communication. Part 1: The analysis of information. *Journal of the Institution of Electrical Engineers-Part III: Radio and Communication Engineering*. 1946; 93(26):429–441.
- [14]. Golub, Gene H, Van Loan, Charles F. *Matrix computations*. Vol. 3. JHU Press; 2012.
- [15]. Gröchenig, Karlheinz. *Foundations of time-frequency analysis*. Springer Science & Business Media; 2013.
- [16]. Loomis, Lynn H. *Introduction to abstract harmonic analysis*. Courier Corporation; 2011.
- [17]. Jia, Rong-Qing; Micchelli, Charles A. On linear independence for integer translates of a finite number of functions. *Proceedings of the Edinburgh Mathematical Society (Series 2)*. 1993; 36(01):69–85.
- [18]. Feichtinger, Hans G; Onchis, Darian M. Constructive realization of dual systems for generators of multi-window spline-type spaces. *Journal of computational and applied mathematics*. 2010; 234(12):3467–3479.
- [19]. Onchis, Darian M; Zappala, Simone. Approximate dual Gabor-like frames based on realizable multi-window spline-type constructions. *SYNASC 2016, 18th International Symposium on Symbolic and Numeric Algorithms for Scientific Computing*; 2016.
- [20]. Reiter, Hans; Stegeman, Jan Derk. *Classical harmonic analysis and locally compact groups*. Courier Corporation; 2000.
- [21]. Ron, Amos. Introduction to shift-invariant spaces: Linear independence *Multivariate Approximation and Applications*. Pinkus, A, Leviatan, D, Dyn, N, Levin, D, editors Cambridge University Press; Cambridge: 2001. 112–151.
- [22]. Ron, Amos; Shen, Zuowei. Generalized shift-invariant systems. *Constructive approximation*. 2005; 22(1):1–45.
- [23]. Søndergaard, Peter L. Gabor frames by sampling and periodization. *Advances in Computational Mathematics*. 2007; 27(4):355–373.
- [24]. Unser, Michael; Aldroubi, Akram. A general sampling theory for nonideal acquisition devices. *Signal Processing, IEEE Transactions on*. 1994; 42(11):2915–2925.
- [25]. Zibulski, Meir; Zeevi, Yehoshua Y. Analysis of multiwindow Gabor-type schemes by frame methods. *Applied and Computational Harmonic Analysis*. 1997; 4(2):188–221.
- [26]. [Accessed: 2016-03-01] Various, LTFAT: The Large Time-Frequency Analysis Toolbox. <http://ltfat.sourceforge.net/>
- [27]. Mallat, SG. A compact multiresolution representation: the wavelet model. *Moore School of Electrical Engineering, Department of Computer and Information Science*; 1987.

Figure 1. Coefficients of the DNA sequence