DYNAMICS, STABILITY AND CONSISTENCY IN REPRESENTATION OF GENOMIC SEQUENCES

Liming Wang and Dan Schonfeld

Department of Electrical and Computer Engineering
University of Illinois at Chicago, Chicago, IL 60607
lwang37@uic.edu, dans@uic.edu

ABSTRACT

Processing of biological data sequences by mapping into numerical signals is a commonly used technique. The operators such as de-noising filter, smoothing filter and certain algorithm could be used iteratively. Little is known about the consistency of analysis results with different mapping strategies in this situation. Meanwhile, due to the errors and noises in acquisition of data, the stability of analysis results should never be neglected. In this paper, we provide a method for analyzing the consistency between different mappings under iterations of operator. We define different concepts of mapping equivalence. We show the necessary and sufficient condition for consistency under iteration of affine operator. We present a few theoretical results on the equivalent mappings on the concept of Fatou and Julia Set. We give the definition of stability under iteration of operator and show the stability issue can be viewed as a special case of mapping equivalence. We also establish the connection of stability to Fatou and Julia set. Finally, we present experimental results on human gene AD169 sequence and rhodopsin gene sequence under one of the widely used mappings and illustrate the equivalent mapping for a smoothing filter.

1. INTRODUCTION

In order to investigate and process information in genomic sequences, one common technique is to map the symbolic data into numerical signals. Therefore a mapping method is required. Numerous mapping strategies have been proposed for different areas. For example, in DNA sequence analysis, there are many mapping methods like mapping the original nucleotide sequence into one-dimensional numerical sequence [1]; indicator sequences method; Method for non-stationary sequences [2], etc. Indeed, it is impossible to determine which mapping is preferable. Furthermore, it is conceivable that distinct mappings could lead to contradictory conclusions. There is always an issue of the consistency for the analysis results under different mappings. In previous work, the authors [3] proposed the concept of mapping equivalence and showed the equivalent transform for certain classes of operators including correlation function and Fourier transform.

However, operators as de-noising filter, smoothing filter and certain algorithm may be applied many times in processing the data. The consistency question rises again naturally in this situation. Another important question in this situation would be whether the analysis result is robust and stable to the perturbations of the data. Due to the noises and errors in acquisition of the biological data sequences, the mapping equivalence or the stability issue under the iterations of operators can not be neglected.

The use of numerical representation of biological data sequence is particularly popular in genomic sequence analysis. In this paper we provide an approach for analyzing the consistency and stability for different mappings under iterations of operator. We conduct experiment on human gene AD169 sequence and rhodopsin gene sequence under a popular mapping method [1]. We show that one-dimensional mapping method is stable for a smoothing filter and illustrate its equivalent mapping class.

2. DYNAMICS AND MAPPING CONSISTENCY UNDER ITERATIONS

Given biological data sequence $\{a_i\}_{i=0}^{n-1}$, where $a_i \in \mathcal{A}$. The set $\mathcal{A}$ could be collection of nucleotides, amino acids, etc. f is a mapping from $\mathcal{A}^n$ to $\mathbb{C}^N$, i.e. $f : \{a_i\}_{i=0}^{n-1} \mapsto z,\ z \in \mathbb{C}^N$. In particular, if we have a mapping method $\tilde{g} : \mathcal{A} \mapsto \mathbb{C}^k$, then it naturally induces the map $g : \{a_i\}_{i=0}^{n-1} \mapsto z,\ z \in \mathbb{C}^{nk}$, where $([z]_{jk+1}, [z]_{jk+2}, ..., [z]_{jk+k})^T = g(a_j),\ j = 0, 1, ..., n-1$. Therefore, for a given symbolic sequence and a mapping method f, the corresponding numerical sequence is a point in $\mathbb{C}^N$. We denote this point as z_f. Let $\mathbf{\Phi} : \mathbb{C}^N \to \mathbb{C}^N$ be a holomorphic(analytic) operator. In this paper we will assume $\mathbf{\Phi}$ is polynomial, i.e. $(\mathbf{\Phi}(z))_i = P_i(z_1, z_2, ..., z_N), i = 1, ..., N$, where P_i is a polynomial. From Taylor's theorem we know that any holomorphic map can be approximated by polynomials.

Since we need the analysis result to be consistent in some sense, we first define different concepts of mapping equivalence.

Definition 1. *Given a genomic sequence and two mapping f and g, we say f and g are asymptotically equivalent if*

$$\lim_{n\to\infty} \|\mathbf{\Phi}^{\circ n}(z_f) - \mathbf{\Phi}^{\circ n}(z_g)\| = 0 \tag{1}$$

f and g are called M-boundedly equivalent, if

$$\sup_n \|\mathbf{\Phi}^{\circ n}(z_f) - \mathbf{\Phi}^{\circ n}(z_g)\| < M. \tag{2}$$

f and g are called n-th equivalent, if

$$\|\mathbf{\Phi}^{\circ n}(z_f) - \mathbf{\Phi}^{\circ n}(z_g)\| = 0 \tag{3}$$

We say f and g are equivalent if they are k-th equivalent for any $k \in \mathbb{N}$.

In study of dynamics of equivalence, the concept of Fatou and Julia set play a fundamental role. There are several different definitions of Fatou and Julia set [4, 5]. We will use the definition below. Before that, we first introduce the notion of *normality*.

Definition 2 ([6])**.** *A collection of holomorphic map $\mathcal{F}$ is called normal if every infinite sequence of maps from $\mathcal{F}$ either has a locally uniformly convergent subsequence or a subsequence diverges locally uniformly.*

Definition 3 ([4])**.** *The domain of normality F of $\mathcal{F} = \{\mathbf{\Phi}^{\circ n}\}$ is called Fatou set. Its complement*

$$J = \mathbb{C}^N \backslash F \tag{4}$$

is called Julia set.

We define the basin of infinity as set of all points which have norms go to infinity under iteration.

The connected components of Julia (Fatou) set are called Julia (Fatou) components.

We will see later the Julia set represents the chaotic behaved points and points in Fatou set show rational behavior. We can show the following propositions about the Fatou set. The proof of the following two propositions in one-dimensional case appears in [4].

Proposition 1. *A point z is in Fatou set if z is in the basin of infinity.*

Proposition 2. *Fatou (Julia) component is invariant. i.e. the operator maps one component onto another component.*

For z_f and z_g, if only one of them is in the basin of infinity, it is obviously that f and g are not boundedly or asymptotically equivalent. If both are in the basin of infinity, although theoretically we can examine the equivalence, however, from computational point of view, the point diverges very fast under polynomial iterations. After a few rounds of iterations, the numerical results will overflow. In this case, the equivalence or even analysis result turns out to be meaningless. Braverman and Yampolsky [7] showed the Julia set of certain types of polynomial can not be computed by any Oracle Turing Machine. If one of the point is in Julia set, it may not able to figure out the equivalent mapping class of the given maps. We will classify all the mappings falling in all these situations as the *computationally chaotic mapping class*. In general, from computational point of view, it is futile or meaningless to find the equivalent mappings of the element in computationally chaotic mapping class. Therefore, the only interesting case left would be if both points are in Fatou set.

As for the simplest case if $\mathbf{\Phi}$ is affine. We can show the following results for equivalent mapping.

Theorem 1. *If $\mathbf{\Phi}(z) = Az + b$, all mappings are asymptotically equivalent for any genomic sequence if and only if the spectral radius $\rho(A) < 1$.*

All mappings are boundedly equivalent for any genomic sequence if and only if either $\rho(A) < 1$ or $\rho(A) = 1$ and all the eigenvalues have index ≤ 1.

For complex manifold, one can construct a pseudo metric called *Kobayashi pseudo metric*. A complex manifold is called *hyperbolic* if the Kobayashi pseudo metric d_K is a metric. [8] is referred for the details of construction and properties of Kobayashi metric and hyperbolic manifold. One of the fundamental theorem for hyperbolic manifold would be the non-increasing principle [8].

Theorem 2 (Non-increasing Principle)**.** *If M is hyperbolic, then for any holomorphic function $\mathbf{\Phi}$ we have,*

$$d_K(\mathbf{\Phi}(x), \mathbf{\Phi}(y)) \leq d_K(x, y), x, y \in M \tag{5}$$

We can show the following theorems on certain class of operator.

Theorem 3. *If $\mathbf{\Phi}$ is non-degenerate homogenous polynomial, then its Fatou component is hyperbolic.*

Theorem 4. *If $\mathbf{\Phi}$ is non-degenerate homogenous polynomial, z_f and z_g are in Fatou set, if $d_K(z_f, z_g) < M$ then f and g will be M-boundedly equivalent under d_K metric. In particular if z_f and z_g are in the same Fatou component U where $\mathbf{\Phi}(U) = U$ and U is hyperbolic, then any two mappings z_f and z_g in this Fatou component is boundedly equivalent under Euclidean metric.*

Theorem 5. *If $\mathbf{\Phi}$ is non-degenerate homogenous polynomial, U is a Fatou component, and $\mathbf{\Phi}(U) = U$, if $d_K(\mathbf{\Phi}(x), \mathbf{\Phi}(y)) < d_K(x, y)$ for any distinct $x, y \in U$, then there exists a unique fixed point in U and any z_f and z_g in U are asymptotically equivalent.*

3. STABILITY UNDER ITERATIONS

In previous sections, we introduce the different concepts of mapping equivalence and provide some results on the equivalent mappings under iteration. Another issue in the iteration

is whether the result is stable to the perturbation of initial values. Due to the inevitable errors and noises in acquisition of data, the stability issue can not be ignored. The analysis result would be much less compelling if it is obtained under some unstable mapping strategy. Usually the biological data sequence would possess very long length, after the numerical mapping, a few acquisition error would correspond to another numerical sequence near the true sequence. Therefore the stability issue is equivalent to the question that whether small changes of the given sequence will cause a small changes in the result.

Definition 4. *A mapping f is stable, if for any $\delta > 0$, there exists $\epsilon > 0$ such that for any point z_g in the ball of radius ϵ, centered at z_f we have*

$$\|\mathbf{\Phi}^{\circ n}(z_f) - \mathbf{\Phi}^{\circ n}(z_g)\| < \delta, \forall n \in \mathbb{N} \tag{6}$$

In another word, all the mappings in the ball are δ-boundedly equivalent.

We can show the following results about the stability.

Theorem 6. *If f is not in basin of infinity, then a mapping f is stable if and only if z_f is in Fatou set.*

From theorem 6, we can see that Fatou set represents the good-behaved mappings. Any mapping close enough will be a boundedly equivalent mapping. On the contrary, for the mapping in the Julia set, no matter how close the mapping is, it may not even be a boundedly equivalent mapping.

4. GENOMIC-SEQUENCE ANALYSIS

We conduct the experiments on human gene AD169 sequences (GenBank accession no. X17403) and rhodopsin gene sequence (GenBank accession no. U49742). We consider the operator $\mathbf{\Phi}$ as a non-linear smoothing filter defined as follow,

$$\mathbf{\Phi}(z_1, z_2, ..., z_N) = (\frac{{z_1}^2 + {z_2}^2}{2}, ..., \frac{{z_i}^2 + {z_{i+1}}^2}{2}, ..., \frac{{z_N}^2}{2}) \tag{7}$$

We consider the mapping $\tilde{f}$ as in [1],

$$\tilde{f}(a) = \begin{cases} 1 & \text{if } a = \mathtt{A} \\ -1 & \text{if } a = \mathtt{T} \\ i & \text{if } a = \mathtt{G} \\ -i & \text{if } a = \mathtt{C} \end{cases} \tag{8}$$

This is one of the widely used mappings. We denote the induced mapping point as z_f.

In Fig. 1, we show the slices of Julia and Fatou set of $\mathbf{\Phi}$ at $(z, 1, 1, ..., 1)$, $(z, i, i, ..., i)$ and $(z, 0.25 + 0.75i, ..., 0.25 + 0.75i)$. Julia set commonly possesses a fractal shape and could be connected or disconnected.

It can be shown the Fatou component U containing origin satisfies all the assumptions in theorem 5. Therefore any two mappings in U will be asymptotically equivalent and $z_f \in U$ for both human gene AD169 sequence and rhodopsin gene sequence. We consider the following perturbed mapping $\tilde{f}'$,

$$\tilde{f}' = \tilde{f} + \Delta z \tag{9}$$

where $\Delta z \in \mathbb{C}$. In Fig. 2, we show the slice of the Fatou component U with z_f at origin and Δz is varying in the ball of radius 0.1, centered at 0. The white area is in the Fatou component.

In Fig. 3, we show how Euclidean distance for two arbitrarily chosen mappings which are in the previous Fatou component U changes with the number of iterations for human gene AD169 sequence. As we can see the distance converges to 0 with the increase of number of iterations.

In Fig. 4, we show the case how Euclidean distance for two mappings changes with the number of iterations. One is in the previous Fatou component U and the other is not. As we can see the distance diverges with the increase of number of iterations.

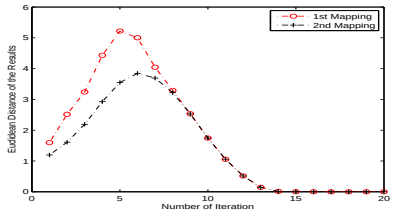


Fig. 3. Consistency case: The illustration of how Euclidean distance for two mappings which are in the previous Fatou component U changes with the number of iterations for human gene AD169 sequence.

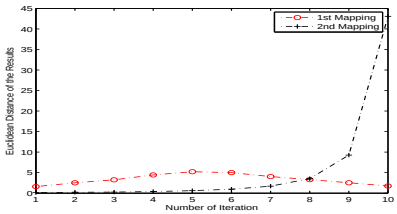
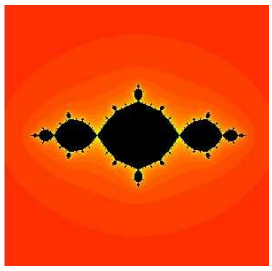


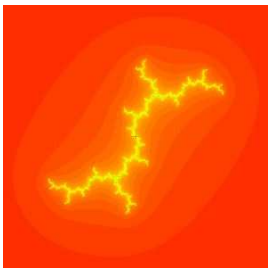
Fig. 4. Inconsistency case: The illustration of how Euclidean distance for two mappings, for which one is in the previous Fatou component U and the other is not, changes with the number of iterations for human gene AD169 sequence.

5. CONCLUSION

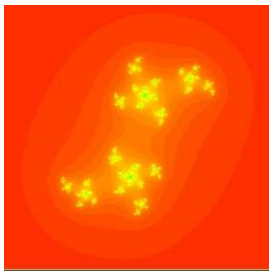
In this paper, we provide a method for analyzing the consistency between different mappings under iterations of operator. We define different mapping equivalence concepts including asymptotical, bounded and n-th iteration equivalence. We provide the necessary and sufficient condition for consistency



(a) Julia Set and Fatou Set

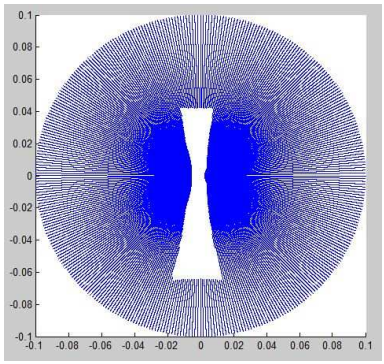


(b) Julia Set and Fatou Set

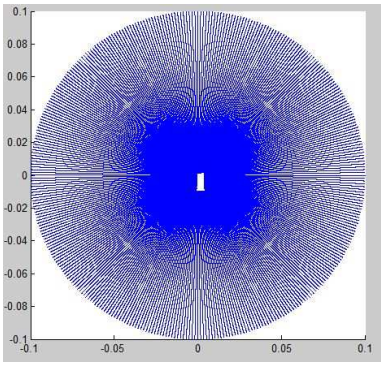


(c) Julia Set and Fatou Set

Fig. 1. Slices of the Fatou and Julia Set of at $(z, 1, 1, ..., 1)$, $(z, i, i, ..., i)$ and $(z, 0.25 + 0.75i, ..., 0.25 + 0.75i)$ respectively. The Julia set is represented as the golden color. The red and black color represent the Fatou Set.



(a) Illustration of slice of the Fatou component U for human gene AD169 sequence.



(b) Illustration of slice of the Fatou component U for rhodopsin gene sequence.

Fig. 2. DNA sequence analysis: (a), (b) show the slice of Fatou component U containing z_f for human gene AD169 sequence and rhodopsin gene sequence respectively. The origin represents z_f and the central white area is in U.

under iteration of affine operator. We present a few results on the equivalent mappings based on the concept of Fatou and Julia Set. We give the definition of stability under iteration of operator and show the stability issue can be viewed as a special case of mapping equivalence. We also establish the connection of stability to Fatou and Julia set. Finally, we conduct experiment on human gene AD169 sequence and rhodopsin gene sequence under a popular mapping method. We show that it is stable for a smoothing filter and illustrate its equivalent mapping class. In the future, we will study the dynamics and consistency problem where there is a composition of different operators.

6. ACKNOWLEDGEMENT

The authors would like to thank Prof. Laura DeMarco from department of mathematics for helpful discussion.

7. REFERENCES

[1] S. V. Buldyrev et al., "Long-range correlation properties of coding and noncoding dna sequences: Genbank analysis," *Phys. Rev. E*, vol. 51, pp. 5084–5091, 1995.

[2] W. Wang and D. H. Johnson, "Computing linear transforms of symbolic signals," *IEEE Trans. on Signal Processing*, vol. 50, no. 3, pp. 628–635, March 2002.

[3] L. Wang and D. Schonfeld, "Mapping equivalence for symbolic sequences: Theory and applications," *IEEE Trans. on Signal Processing,*, vol. 57, no. 12, pp. 4895 –4905, Dec. 2009.

[4] J. Milnor, *Dynamics in One Complex Variable*, vol. 160 of *Annals of Mathematics Studies*, Princeton University Press, 3rd edition, 2006.

[5] L. Carleson and T. W. Gamelin, *Complex Dynamics*, Springer-Verlag, 1993.

[6] L. Ahlfors, *Complex Analysis*, McGraw-Hill, 3rd edition, 1979.

[7] M. Braverman and M. Yampolsky, "Constructing non-computable julia sets," in *STOC '07*, New York, NY, USA, 2007, pp. 709–716, ACM.

[8] S. Kobayashi, *Hyperbolic Manifolds and Holomorphic Mappings*, Marcel Dekker, New York, 1970.