Mini-presentation on Bowen Notebook Problem 85

Dan Thompson

August 1, 2017

Published in Math Systems Theory (1975), received Nov 1972.

Some Systems with Unique Equilibrium States

by

RUFUS BOWEN*

Department of Mathematics
University of California
Berkeley, California 94720

We shall be dealing with a homeomorphism $f: X \to X$ of a compact metric space and a continuous $\varphi: X \to R$. Let $M_f(X)$ denote the set of all f-invariant Borel probability measures on X. $\mu \in M_f(X)$ is called an *equilibrium state* (for f and φ) if

$$h_\mu(f)+\mu(\varphi) = \sup_{\nu \in M_f(X)} (h_\nu(f)+\nu(\varphi)),$$

where $h_\mu(f)$ is the entropy of μ. We want conditions on f and φ which guarantee a unique equilibrium state.

f is called *expansive* if there is an $\epsilon > 0$ such that for any two points $x \neq y$ in X there is an $n \in Z$ so that $d(f^n(x), f^n(y)) > \epsilon$. f satisfies *specification* if for each $\delta > 0$ there is an integer $p(\delta)$ for which the following is true: if $I_1, \cdots, I_n$ are intervals of integers contained in $[a, b]$ with $d(I_i, I_j) \geq p(\delta)$ for $i \neq j$ and $x_1, \cdots, x_n \in X$, then there is a point $x \in X$ with $f^{b-a+p(\delta)}(x) = x$ and $d(f^k(x), f^k(x_i)) < \delta$ for $k \in I_i$. This condition allows us to construct a lot of periodic points.

For $\varphi \in C(X)$ and $n \geq 1$ let

$$(S_n\varphi)(x) = \varphi(x)+\varphi(f(x))+\cdots+\varphi(f^{n-1}(x)).$$

Let $V(f)$ be the set of $\varphi \in C(X)$ for which an $\epsilon > 0$ and a K exist for which the following is true: $d(f^k(x), f^k(y)) \leq \epsilon$ for all $0 \leq k < n \Rightarrow |S_n\varphi(x) - S_n\varphi(y)| \leq K$.

THEOREM. *Let $f: X \to X$ be an expansive homeomorphism of a compact metric space satisfying specification. Then each $\varphi \in V(f)$ has a unique equilibrium state μ_φ.*

Remark. Let δ be any expansive constant for f. Then, if $\varphi \in V(f)$, $|\varphi|_f = \sup \{|S_n\varphi(x) - S_n\varphi(y)|: n \geq 1$ and $d(f^k(x), f^k(y)) \leq \delta \forall k \in [0, n)\}$ is finite (if ϵ, K are as in the definition of $\varphi \in V(f)$ and $d(x, y) < \epsilon$ follows from $d(f^j(x), f^j(y))$

Notebook question

Question (85.)

Codon frequencies via equilibrium states for "some potential"?

Notebook question

Question (85.)

Codon frequencies via equilibrium states for "some potential"?

Recent progress:

- Teresa Krick, Nina Verstraete, Leonardo Alonso, David Shub, Diego Ferreiro, Michael Shub and Ignacio E.Sanchez, Amino acid metabolism conflicts with protein diversity, Molecular Biology and Evolution, 2014.
- David Koslicki and Daniel Thompson, Coding sequence density estimation via topological pressure, J. Math. Biol., 2014.
- Also papers by Bruno Cessac e.g. Gibbs distribution analysis of temporal correlations structure in retina ganglion cells, Journal of Physiology, 2012: Uses topological pressure in neural networks; cites Bowen, Ruelle, etc.

Estimating coding sequence density

Can we estimate coding sequence (CDS) density in a segment of DNA by measuring its (weighted) complexity as a sequence? (coding sequences constitute about 2% of the 300,000,000 long sequence of A,T,G,C which represents human genome)

Nucleotide triplets are distributed differently in regions with low/high frequency of coding sequences. (e.g. long runs of $AAAAAA\ldots$ are associated with intergenic regions of the genome).

Can we use these differences to detect/predict coding sequences when viewing the genome simply as a long string of data?

Topological pressure for finite sequences

We introduce notion of topological pressure for finite sequences.

The topological pressure of a finite sequence is given by counting the number of distinct subwords at an exponentially shorter length, with weights determined by a locally constant function.

Topological pressure for finite sequences

We introduce notion of topological pressure for finite sequences.

The topological pressure of a finite sequence is given by counting the number of distinct subwords at an exponentially shorter length, with weights determined by a locally constant function.

We consider potential functions depending on 3 symbols. That is, depending on nucleotide triples aka "codons".

Topological pressure for finite sequences

We introduce notion of topological pressure for finite sequences.

The topological pressure of a finite sequence is given by counting the number of distinct subwords at an exponentially shorter length, with weights determined by a locally constant function.

We consider potential functions depending on 3 symbols. That is, depending on nucleotide triples aka "codons".

Sequences with high topological pressure balance high complexity and high frequency of words which are weighted strongly

Potential can be selected based on some underlying principle; e.g. GC content, or by training computationally against a data set. Potential determines a Markov measure as its equilibrium state, which we use to determine 'coding potential' (intron or exon)

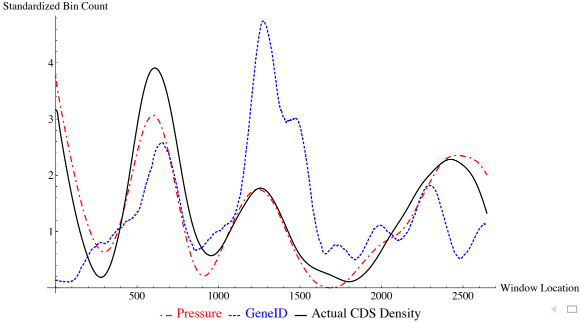
Which potential?

We use a window of length around $66,000$ bp on human genome, and obtain our parameters by optimizing the correlation between CDS density and topological pressure across these roughly $40,000$ windows

Which potential?

We use a window of length around 66,000 bp on human genome, and obtain our parameters by optimizing the correlation between CDS density and topological pressure across these roughly 40,000 windows

Using these parameters, we compute the topological pressure along the genomes of fruit fly, monkey, etc... On Rhesus Macaque, correlation was 0.73:



Which potential?

	2nd Base				
	U	C	A	G	
U	UUU (Phe)	UCU (Ser)	UAU (Tyr)	UGU (Cys)	
	UUC (Phe)	UCC (Ser)	UAC (Tyr)	UGC (Cys)	
	UUA (Leu)	UCA (Ser)	UAA Stop	UGA Stop	
	UUG (Leu)	UCG (Ser)	UAG Stop	UGG (Trp)	
C	CUU (Leu)	CCU (Pro)	CAU (His)	CGU (Arg)	
	CUC (Leu)	CCC (Pro)	CAC (His)	CGC (Arg)	
	CUA (Leu)	CCA (Pro)	CAA (Gln)	CGA (Arg)	Unit Square
	CUG (Leu)	CCG (Pro)	CAG (Gln)	CGG (Arg)	
1st base A	AUU (Ile)	ACU (Thr)	AAU (Asn)	AGU (Ser)	
	AUC (Ile)	ACC (Thr)	AAC (Asn)	AGC (Ser)	
	AUA (Ile)	ACA (Thr)	AAA (Lys)	AGA (Arg)	
	AUG (Met)	ACG (Thr)	AAG (Lys)	AGG (Arg)	
G	GUU (Val)	GCU (Ala)	GAU (Asp)	GGU (Gly)	
	GUC (Val)	GCC (Ala)	GAC (Asp)	GGC (Gly)	
	GUA (Val)	GCA (Ala)	GAA (Glu)	GGA (Gly)	
	GUG (Val)	GCG (Ala)	GAG (Glu)	GGG (Gly)	