Consistent proximity between N-grams

García, Jesús E.[1]
González-López, V. A.[2]

SMTDA 2016 - Malta

[1]IMECC-UNICAMP e-mail: jg@ime.unicamp.br
[2]IMECC-UNICAMP e-mail: veronica@ime.unicamp.br

Partition Markov models

Notation

Let (X_t) be a discrete time order M Markov chain

- A the finite alphabet;

Partition Markov models

Notation

Let (X_t) be a discrete time order M Markov chain

- A the finite alphabet;
- $\mathcal{S} = A^M$ the state space;

Partition Markov models

Notation

Let (X_t) be a discrete time order M Markov chain

- A the finite alphabet;
- $\mathcal{S} = A^M$ the state space;
- $P(a|s) = \text{Prob}(X_t = a|X_{t-M}^{t-1} = s),\ a \in A,\ s \in \mathcal{S}$ the transition probabilities.

Equivalence relationship on $\mathcal{S}$

Definition

for $s, r \in \mathcal{S}$; $s \sim_p r \iff P(a|s) = P(a|r)\ \forall a \in A$.

- For any $s \in \mathcal{S}$, the equivalence class of s is given by $[s] = \{r \in \mathcal{S} | r \sim_p s\}$.

Equivalence relationship on $\mathcal{S}$

Definition

for $s, r \in \mathcal{S}$; $s \sim_p r \iff P(a|s) = P(a|r)\ \forall a \in A$.

- For any $s \in \mathcal{S}$, the equivalence class of s is given by $[s] = \{r \in \mathcal{S} | r \sim_p s\}$.
- The classes defined by $\sim_p$ are the subsets of $\mathcal{S}$ with the same transition probabilities.

Equivalence relationship on $\mathcal{S}$

- The equivalence relation defines a partition $\mathcal{L}$ of $\mathcal{S}$.

Equivalence relationship on $\mathcal{S}$

- The equivalence relation defines a partition $\mathcal{L}$ of $\mathcal{S}$.
- We have $(|A| - 1)$ transition probabilities for each "part" (element of $\mathcal{L}$), obtaining a model with $(|A| - 1)|\mathcal{L}|$ parameters.

Equivalence relationship on $\mathcal{S}$

- The equivalence relation defines a partition $\mathcal{L}$ of $\mathcal{S}$.
- We have $(|A| - 1)$ transition probabilities for each "part" (element of $\mathcal{L}$), obtaining a model with $(|A| - 1)|\mathcal{L}|$ parameters.
- The elements of $\mathcal{S}$ on the same equivalence class activate the same random mechanism to choose the next element in the Markov chain.

Markov chain with partition $\mathcal{L}$

Definition

let (X_t) be a discrete time, order M Markov chain on A and let $\mathcal{L} = \{L_1, L_2, \ldots, L_K\}$ be a partition of $\mathcal{S}$. We will say that (X_t) is a Markov chain with partition $\mathcal{L}$ if this partition is the one defined by $\sim_p$.

Example

- $A = \{0, 1\}$, $M = 2$, $\mathcal{S}(= A^M) = \{00, 01, 10, 11\}$,
  $P(0|00) = P(0|01) = 0.4$ $P(0|10) = P(0|11) = 0.2$;

Example

- $A = \{0, 1\}$, $M = 2$, $\quad \mathcal{S}(= A^M) = \{00, 01, 10, 11\}$,
  $P(0|00) = P(0|01) = 0.4 \;\; P(0|10) = P(0|11) = 0.2$;
- $P(1|s) = 1 - P(0|s) \;\; \forall s \in \mathcal{S}$;

Example

- $A = \{0, 1\}$, $M = 2$, $\mathcal{S}(= A^M) = \{00, 01, 10, 11\}$, $P(0|00) = P(0|01) = 0.4$ $P(0|10) = P(0|11) = 0.2$;
- $P(1|s) = 1 - P(0|s)$ $\forall s \in \mathcal{S}$;
- the partition for this Markov chain is $\mathcal{L} = \{\{00, 01\}, \{10, 11\}\}$ with parts $L_1 = \{00, 01\}$ and $L_2 = \{10, 11\}$;

Example

- $A = \{0, 1\}$, $M = 2$, $\quad \mathcal{S}(= A^M) = \{00, 01, 10, 11\}$, $P(0|00) = P(0|01) = 0.4 \;\; P(0|10) = P(0|11) = 0.2$;
- $P(1|s) = 1 - P(0|s) \;\; \forall s \in \mathcal{S}$;
- the partition for this Markov chain is $\mathcal{L} = \{\{00, 01\}, \{10, 11\}\}$ with parts $L_1 = \{00, 01\}$ and $L_2 = \{10, 11\}$;
- the parameters of the Markov chain with partition $\mathcal{L}$ are $P(0|L_1) = 0.4$ and $P(0|L_2) = 0.2$.

Model selection problem

Given a sample generated by a finite memory stationary process, how to choose a partition defining a good Markov model for the source?

Notation

Let x_1^n be a sample of the process $\left(X_t\right)$, $s \in \mathcal{S}$, $a \in A$ and $n > M$.

$$N_n(s,a) = \left|\{t : M < t \leq n, x_{t-M}^{t-1} = s, x_t = a\}\right|, \tag{1}$$

$$N_n(s) = \left|\{t : M < t \leq n, x_{t-M}^{t-1} = s\}\right|. \tag{2}$$

To simplify the notation we will omit the $_n$ on N_n.

A distance in S

Definition

We define the distance d in S,

$$\begin{aligned} d(s,r) &= \frac{2}{(|A|-1)\ln(n)} \sum_{a\in A} \left\{ N(s,a) \ln\left(\frac{N(s,a)}{N(s)}\right) \right. \\ &+ N(r,a) \ln\left(\frac{N(r,a)}{N(r)}\right) \\ &- \left. (N(s,a)+N(r,a)) \ln\left(\frac{N(s,a)+N(r,a)}{N(s)+N(r)}\right) \right) \end{aligned}$$

for any $s, r \in S$.

A distance in S

Theorem

For any $s, r, t \in S$,

i. $d(r,s) \geq 0$ *with equality if and only if* $\frac{N(s,a)}{N(s)} = \frac{N(r,a)}{N(r)} \quad \forall a \in A$,

ii. $d(r,s) = d(s,r)$,

iii. $d(r,t) \leq d(r,s) + d(s,t)$.

d can be generalized to subsets (see García, J. and González-López, V. A. (2010)).

Consistence in the case of a Markov source

Theorem

Let (X_t) be a discrete time, order M Markov chain on a finite alphabet A. Let x_1^n be a sample of the process and $0 < \alpha < \infty$, then for n large enough and for each $s, r \in S$, $d_n(r, s) < \alpha$ iff s and r belong to the same class.

Algorithm

Input: $d(s,r) \forall s \neq r \in S$; **Output**: $\hat{\mathcal{L}}_n$.
$B = S$; $\hat{\mathcal{L}}_n = \emptyset$
while $B \neq \emptyset$

select $s \in B$

define $L_s = \{s\}$

$B = B \setminus \{s\}$

for each $r \in B, r \neq s$

if $d(s,r) < \alpha$

$L_s = L_s \cup \{r\}$
$B = B \setminus \{r\}$

$\hat{\mathcal{L}}_n = \hat{\mathcal{L}}_n \cup \{L_s\}$

Return: $\hat{\mathcal{L}}_n = \{L_1, L_2, \ldots, L_K\}$

BIC criterion

Given a sample x_1^n and a partition $\mathcal{L}$ of the state space,

$$\mathrm{BIC}(x_1^n, \mathcal{L}) = \sum_{a \in A, L \in \mathcal{L}} N_n^{\mathcal{L}}(L, a) \ln\left(\frac{N_n^{\mathcal{L}}(L, a)}{N_n^{\mathcal{L}}(L)}\right) - \frac{(|A| - 1)|\mathcal{L}|}{2} \ln(n),$$

Theorem

Let (X_t) be a Markov chain of order M over a finite alphabet A, with partition $\mathcal{L}^*$, x_1^n a size n sample of X_t and let $\mathcal{P}$ be the set of partitions of $\mathcal{S} = A^M$. Define,

$$\mathcal{L}_n = \arg\max_{\mathcal{L} \in \mathcal{P}} \{BIC(\mathcal{L}, x_1^n)\}$$

then, eventually almost surely as $n \to \infty$,

$$\mathcal{L}^* = \mathcal{L}_n$$

BIC proximity

Given the set $\{X_t^i\}_{i=1,2}$ of independent stochastic processes over the same alphabet A. All of them of order $M < \infty$ with sample $(x_t^i)_1^{n_i}$ of size $n_i, i = 1, 2$, and a partition $\mathcal{L}$ of the state space A^M. And consider,

$$\mathrm{BIC}\left((x_t^1)_1^{n_1}, (x_t^2)_1^{n_2}, \mathcal{L}\right) = \sum_{a \in A, L \in \mathcal{L}} N_{n_1+n_2}^{\mathcal{L}}(L, a) \ln\left(\frac{N_{n_1+n_2}^{\mathcal{L}}(L, a)}{N_{n_1+n_2}^{\mathcal{L}}(L)}\right) - \frac{(|A|-1)}{2}|\mathcal{L}| \ln(n_1 + n_2).$$

BIC proximity

Theorem

Let (X_t^i) be a Markov chain of order M over a finite alphabet A, $\{x^i\}_1^n$ a size n sample of X_t^i. This for $i = 1, 2$. Let $\mathcal{P}$ be the set of partitions of $\mathcal{S} = A^M$. Define,

$$\mathcal{L}^1{}_n = \arg\max_{\mathcal{L}\in\mathcal{P}}\{BIC(\mathcal{L}, \{x^1\}_1^n)\}$$

$$\mathcal{L}^2{}_n = \arg\max_{\mathcal{L}\in\mathcal{P}}\{BIC(\mathcal{L}, \{x^2\}_1^n)\}$$

$$\mathcal{L}^{12}{}_n = \arg\max_{\mathcal{L}\in\mathcal{P}}\{BIC(\mathcal{L}, \{x^1\}_1^n, \{x^2\}_1^n)\}$$

then, eventually almost surely as $n \to \infty$, $X_t^1 \neq^d X_t^2$ if, and only if, $BIC\left((x_t^1)_1^{n_1}, (x_t^2)_1^{n_2}, \mathcal{L}^{12}\right) < \sum_{k=1,2} BIC((x_t^k)_1^{n_k}, \mathcal{L}^k)$

BIC proximity

A corolary of this theorem is that eventually almost surely as $n \to \infty$, $X_t^1 \neq^d X_t^2$ if, and only if, $d(\{x^1\}_1^n, \{x^2\}_1^n) > 1$, with $d(\{x^1\}_1^n, \{x^2\}_1^n)$ defined by

Definition

$$d(\{x^1\}_1^{n_1}, \{x^2\}_1^{n_2}) = \frac{\sum_{a \in A} B(\mathcal{L}^1, n_1, a) + B(\mathcal{L}^2, n_2, a) - B(\mathcal{L}^{12}, n_1 + n_2, a)}{(|A| - 1)\{|\mathcal{L}^1| \ln(n_1) + |\mathcal{L}^2| \ln(n_2) - |\mathcal{L}^{12}| \ln(n_1 + n_2)\}}$$

and $B(\mathcal{L}, n, a) = \sum_{L \in \mathcal{L}} N_n^{\mathcal{L}}(L, a) \ln\left(\frac{N_n^{\mathcal{L}}(L,a)}{N_n^{\mathcal{L}}(L)}\right)$.

The data

The five DNA sequences registered so far:

- (i) B95-8-type I, the reference sequence (accession number NC_007605), see Baer *et al.* (1984)[3], named "EBV.WT" according to Kwok *et al.* (2012)[2];
- (ii) GD1-type I (accession number AY961628) denoted by "GD1", see Zeng *et al.* (2005)[4];
- (iii) AG876-type II (accession number DQ279927) denoted by "AG876", see Dolan *et al.* (2006)[5];
- (iv) GD2-type 1 (accession number HQ020558) denoted by "GD2", see Liu *et al.* (2011)[6] and
- (v) the new sequence reported in 2012, an EBV strain type I denoted by HKNPC1 (accession number JQ009376), see Kwok *et al.* (2012)[2].

- The minimal Markov model for each sequence was obtained by applying the algorithm introduced in García and González-López (2011)[8].
- The 20 amino acid alphabet plus the stop codon was used with IUPAC notation.
- The concatenation of amino acids observed in the code was the string of realizations x_1^n.

The size of each sequence is shown in the next table,

Table: Total number of amino acids of each DNA sequence: EBV.WT (accession number NC_007605), GD1 (accession number AY961628), AG876 (accession number DQ279927), GD2 (accession number HQ020558) and HKNPC (accession number JQ009376).

EBV.WT	GD1	AG876	GD2	HKNPC1
54373	57219	54670	52074	54913

For the incomplete sequence HKNPC1, the occurrences of each string were computed separately from the beginning of each stretch of sequence.

Given two sequences $\{x^1\}_1^{n_1}$ and $\{x^2\}_1^{n_2}$, we used as a measure of dissimilarity,

$$d(\{x^1\}_1^{n_1}, \{x^2\}_1^{n_2}) = \frac{\sum_{a\in A} B(\mathcal{L}^1, n_1, a) + B(\mathcal{L}^2, n_2, a) - B(\mathcal{L}^{12}, n_1+n_2, a)}{(|A|-1)\{|\mathcal{L}^1|\ln(n_1) + |\mathcal{L}^2|\ln(n_2) - |\mathcal{L}^{12}|\ln(n_1+n_2)\}}$$

where $B(\mathcal{L}, n, a) = \sum_{L\in\mathcal{L}} N_n^{\mathcal{L}}(L, a) \ln\left(\frac{N_n^{\mathcal{L}}(L,a)}{N_n^{\mathcal{L}}(L)}\right)$.

Results

Table: 2. Minimal Good Partition for each sequence. We show each part L, a member of $\hat{\mathcal{L}}_n$, as a collection of amino acids.

EBV.WT	GD1	AG876
{stop,K,M,T,V}	{stop,D,H,K,L,M,V}	{stop,E,K,Q}
{A,C,D,H,L,S}	{A,G,P,R,W}	{A,G,P,R,W}
{E,Q}	{C,F,I,N,Y}	{C,D,H,L,M,S,T,V}
{F,I,N,Y}	{E,Q}	{F,I,N,Y}
{G,P}	{S,T}	
{R,W}		

GD2	HKNPC1	
{stop,E,K,Q}	{stop,E,K,Q}	
{A,D,G,H,L,M,S,T,V,W}	{A,D,L,M,S,T,V,W}	
{C,F,I,N,Y}	{C,F,H,I,N,Y}	
{P,R}	{G,P,R}	

For instance, if L is composed of amino acids F, I, N and Y, then the part L, denoted as the set $L = \{\text{F, I, N, Y}\}$ means that

- F, I, N and Y have the same probability of choosing the next symbol in the DNA sequence, or
- for any element a in the alphabet of amino acids A, $P(a|\text{ F}) = P(a|\text{I}) = P(a|\text{N}) = P(a|\text{Y})$, or
- F, I, N and Y are equivalent.

- We note that the minimal good partitions of two nasopharyngeal carcinoma-related EBV strains, GD2 and HKNPC1, are very similar, except for the positions of the amino acids G and H. This result agrees with the findings in Kwok *et al.* (2012)[2].

- We note that the minimal good partitions of two nasopharyngeal carcinoma-related EBV strains, GD2 and HKNPC1, are very similar, except for the positions of the amino acids G and H. This result agrees with the findings in Kwok *et al.* (2012)[2].
- In Kwok *et al.* (2012)[2] the authors show that HKNPC1 has closer phylogenetic relationship to GD1 and GD2 than EBV.WT and AG876. We emphasize that both sequences, GD2 and HKNPC1, were obtained from epithelial cells. GD1 (the remaining nasopharyngeal carcinoma EBV strain) was not directly harvested from epithelial tissue, but from saliva.

To confirm the proximity between GD2 and HKNPC1, we built a dendrogram using our measure $d()$.

Table: 3. $d(.,.)$

	EBV.WT	GD1	AG876	GD2	HKNPC1
EBV.WT	0.0000	0.0018	0.0017	0.0017	0.0018
GD1	-	0.0000	0.0021	0.0026	0.0023
AG876	-	-	0.0000	**0.0009**	**0.0007**
GD2	-	-	-	0.0000	**0.0006**
HKNPC1	-	-	-	-	0.0000

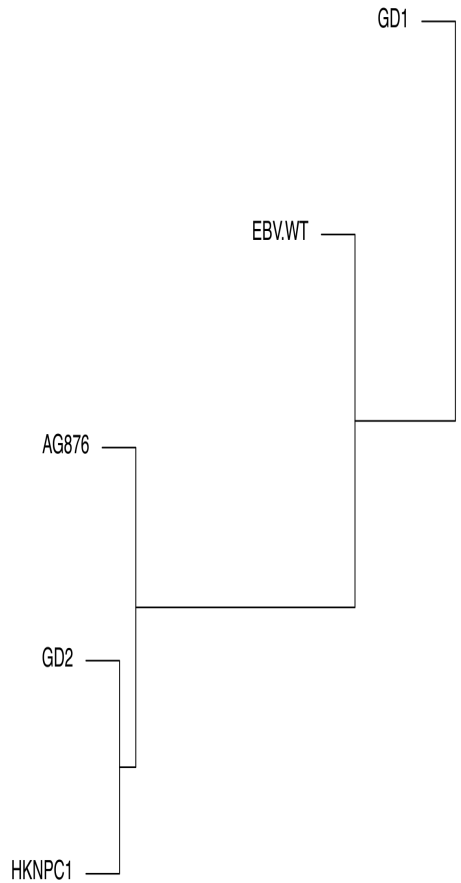

Figure: Dendrogram for our dataset.

The dendrogram exposes a proximity also between the group {GD2, HKNPC1} and the strain AG876; this is explained by the low values of d(GD2,AG876), and d(HKNPC1, AG876) (see table 3).

Conclusion

- The members of some part of a minimal good partition from some sequences can be considered as natural units of genome architecture and can also reveal stochastic proximity between sequences as shown in this paper.

Conclusion

- The members of some part of a minimal good partition from some sequences can be considered as natural units of genome architecture and can also reveal stochastic proximity between sequences as shown in this paper.
- We show that the nasopharyngeal carcinoma-related EBV strains GD2 and HKNPC1 are closer, according to symmetrized relative entropy. This finding is in accordance with the results of Kwok *et al.* (2012)[2].

Conclusion

- The members of some part of a minimal good partition from some sequences can be considered as natural units of genome architecture and can also reveal stochastic proximity between sequences as shown in this paper.
- We show that the nasopharyngeal carcinoma-related EBV strains GD2 and HKNPC1 are closer, according to symmetrized relative entropy. This finding is in accordance with the results of Kwok *et al.* (2012)[2].
- Also we obtained results consistent with McGeoch and Gatherer (2007)[1] in relation to the distance between EBV.WT, AG876 and GD1.


D.J. McGeoch and D. Gatherer. Lineage structures in the genome sequences of three Epstein-Barr virus strains. *Virology* 359 (1) p. 1, 2007.


H. Kwok, A.H. Tong, C.H. Lin, S. Lok, P.J. Farrel, D.L. Kwong and A. K. Chiang. Genomic sequencing and comparative analysis of Epstein-Barr Virus genome isolated from primary nasopharyngeal carcinoma biopsy. *PLoS ONE* 7 (5): e36939, 2012.


R. Baer, A.T. Bankier, M.D. Biggin, P.L. Deininger, P.J. Farrell, T.J. Gibson, G. Hatfull, G.S. Hudson, S.C. Satchwell, C. Séguin, P.S. Tuffnell and B.G. Barrell, B.G. DNA sequence and expression of the B95-8 Epstein-Barr virus genome, *Nature* 310 (5974) p. 207, 1984.


M.S. Zeng, D.J. Li, Q.L. Liu, L.B. Song, M.Z. Li, R.H. Zhang, X.J. Yu, H.M. Wang, I. Emberg and Y.X. Zeng. Genomic sequence analysis of Epstein-Barr Virus strain GD1 from a nasopharyngeal carcinoma patient. *Journal of Virology* 79 (24) p. 15323, 2005.


A. Dolan, C. Addison, D. Gatherer, A.J. Davison and D.J. McGeoch. The genome of Epstein-Barr virus type 2 strain AG876. *Virology* 350 (1) p. 164, 2006.


P. Liu, X. Fang, Z. Feng, Y.M. Guo, R.J. Peng, T. Liu, Z. Huang, Y. Feng, X. Sun, Z. Xiong, X. Guo, S.S. Pang, B. Wang, X. Lv, F.T. Feng, D.J. Li, L.Z. Chen, Q.S. Feng, W.L. Huang, M.S. Zeng, J.X. Bei, Y. Zhang, Y.X. Zeng. Direct sequencing and characterization of a clinical isolate of Epstein-Barr virus from nasopharyngeal carcinoma tissue by using next-generation sequencing technology. *Journal of Virology* 85 (21) p.11291, 2011.


J. Garcia and V.A. Gonzalez-Lopez. Minimal markov models. arXiv preprint arXiv:1002.0729, 2010.


J.E. García and V.A. González-López. Minimal Markov Models. In *Proceedings of the Fourth Workshop on Information Theoretic Methods in Science and Engineering. Helsinki.* v.1. p.25 - 28, 2011.

Thanks for your attention!