Università degli Studi di Palermo
Facoltà di Scienze MM.FF.NN.
Laurea Specialistica in Scienze dell'Informazione

Estensione della trasformata di Burrows-Wheeler ed applicazioni

Tesi di laurea di
Giovanna Rosone

Relatore
Prof. Antonio Restivo

Indice

- Breve introduzione della trasformata di Burrows-Wheeler (BWT)
- Estensione della BWT a k sequenze (EBWT)
- Confronto di sequenze
  - Esperimenti filogenetici
- Compressione di testi
  - Esperimenti sulla compressione

Trasformata di Burrows e Wheeler

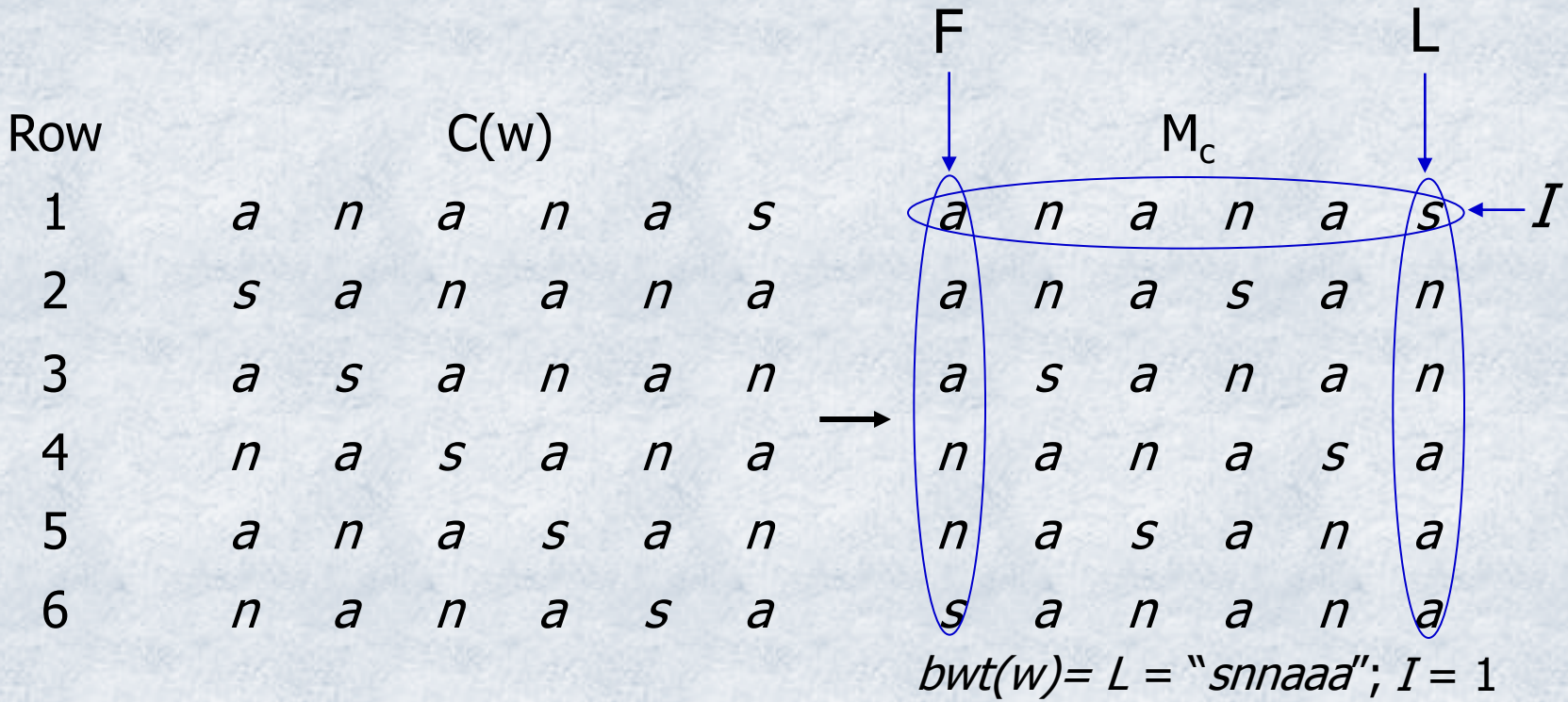
- È stata introdotta nel 1994 per migliorare la compressione dei file senza perdita d'informazione.
- È una trasformazione reversibile su testi.

Trasformata

La BWT prende in input un testo *w* di lunghezza *n* e produce:

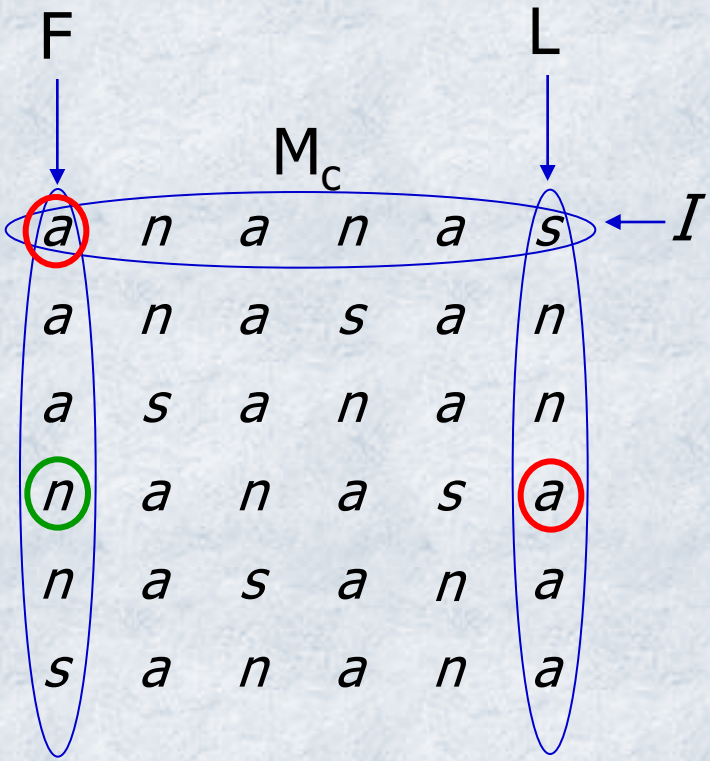
- Il testo permutato *bwt(w)*
- Un indice *I*

Esempio: *w="ananas", n=6*



Proprietà

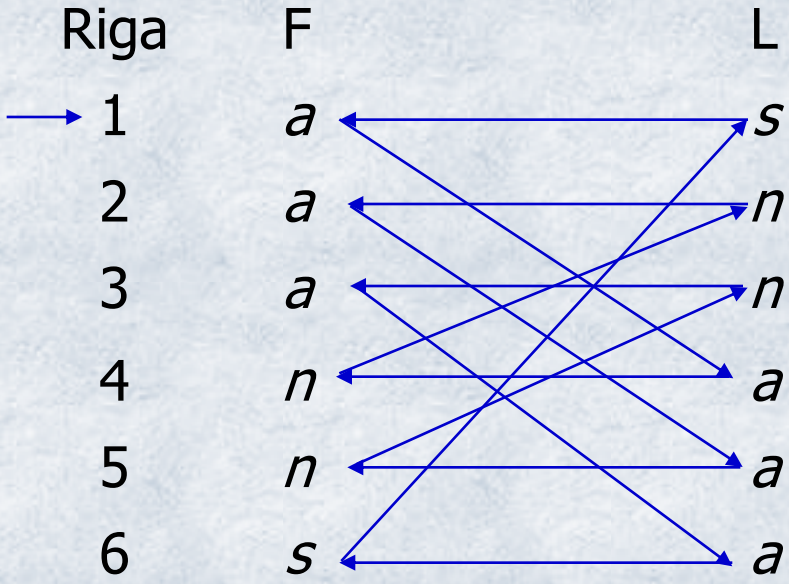
- Ogni colonna di M_c è una permutazione di *w*;
- Il primo carattere di *w* è F[*I*];
- Per ogni carattere α, l'i-esima occorrenza di α in F corrisponde all'i-esima occorrenza di α in L;
- Per ogni j = *1…n*, j≠*I*, il simbolo L[j] è seguito in *w* dal simbolo F[j].



Anti-trasformata

- Costruisce F ordinando lessicograficamente i caratteri di L;
- Costruisce la permutazione τ tale che per ogni j=1…*n*, se F[j] è la k-esima istanza di α in F, allora L[τ[j]] è la k-esima istanza di α in L.

bwt(w)= L = "snnaaa"; I = 1



- w[1] = F[I]=F[**1**]=a
- w[2] = F[τ[I]]=F[**4**]=n
- w[3] = F[τ^2[I]]=F[τ[4]]=F[**2**]=a
- w[4] = F[τ^3[I]]=F[τ[2]]=F[**5**]=n
- w[5] = F[τ^4[I]]=F[τ[5]]=F[**3**]=a
- w[6] = F[τ^5[I]]=F[τ[3]]=F[**6**]=s

$$\tau = \begin{pmatrix} 1 & 2 & 3 & 4 & 5 & 6 \\ 4 & 5 & 6 & 2 & 3 & 1 \end{pmatrix} = \begin{pmatrix} 1 & 4 & 2 & 5 & 3 & 6 \end{pmatrix}$$

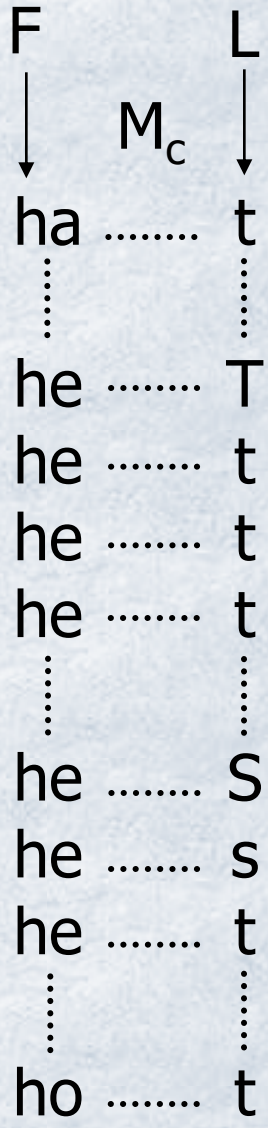
w = a n a n a s

1 4 2 5 3 6

Perché BWT migliora la compressione?

- La stringa prodotta dalla BWT è localmente omogenea: consiste nella concatenazione di diverse sottostringhe contenenti solo un piccolo numero di simboli distinti.

w="…She…the…The… he…the…that…the…she…the…those…"



Alcuni articoli sulla BWT

P. Fenwick, *The Burrows-Wheeler Transform for Block Sorting Text Compression - Principles and Improvements*, The Computer Journal, Vol. 39(9), pp. 731-740, 1996.

G. Manzini, *An Analysis of the Burrows-Wheeler Transform*, Journal of the ACM. Vol. 48, n. 3, (2001), pp. 407-430.

M. Effros, *Universal lossless source coding with the Burrows-Wheeler transform*, in: Proc. IEEE Data Compression Conference, IEEE Computer Society, 1999, pp. 178–187.

K. Sadakane, *On optimality of variants of the block sorting compression*, in: Proc. IEEE Data Compression Conference, IEEE Computer Society, 1998, p. 570.

Z. Arnavut, S. S. Magliveras, *Block sorting and compression*, in: Proc. IEEE Data Compression Conference, IEEE Computer Society, 1997, pp. 181–190.

A. I. Wirth, A. Moffat, *Can we do without ranks in Burrows Wheeler transform compression?*, in: Proc. IEEE Data Compression Conference, IEEE Computer Society, 2001, pp. 419–428.

P. Ferragina, G. Manzini, *Indexing Compressed Text*, Journal of the ACM, Vol. 52 (2005), pp. 552-581.

P. Ferragina, R. Giancarlo, G. Manzini, M. Sciortino, *Boosting Textual Compression in Optimal Linear Time*, Journal of the ACM, Vol. 52 (2005), pp. 688-713.

S. Mantaci, A. Restivo, M. Sciortino, *Burrows-Wheeler Transform and Sturmian Words*, Information Processing Letters, Vol. 86(5), pp. 241-246 (2003).

Trasformata estesa (EBWT)

- [M. Crochemore, J. Désarménien, D. Perrin, “A note on the Burrows-Wheeler Transform”, Theoretical Computer Science, 2005];
- [M. Gessel, C. Reutenauer, “Counting permutations with given cycle structure and descent set”, J. Comb. Theory, 1993].

Una nuova relazione d'ordine

- Una parola v in A^* è detta primitiva se $v=u^n$ implica v=u e n=1.
- Per ogni parola v in A^*, esiste un'unica parola primitiva w e un unico intero k tale che $v=w^k$, cioè w=root(v) e k=exp(v).
- Sia u una parola in A^*, si denoti con u^ω la parola infinita $u^\omega=uuu\ldots$.
- La nuova relazione d'ordine, denotata con $\leq_\omega$, è una relazione d'ordine totale definita:

$$u \leq_\omega v \Leftrightarrow \begin{cases} \exp(u) \leq \exp(v) \text{ se } root(u) = root(v) \\ u^\omega <_{lex} v^\omega \quad \text{altrimenti} \end{cases}$$

Relazione d'ordine $\leq_\omega$ e Teorema di Fine e Wilf

Proposizione:

Date due parole primitive u e v,

$$u \leq_\omega v \Leftrightarrow pref_k(u^\omega) <_{lex} pref_k(v^\omega)$$

dove k=|u|+|v|-mcd(|u|,|v|).

Esempio: u=aabaac, v=aabaacaab $\Rightarrow$ v $\leq_\omega$u, poiché u^ω e v^ω differiscono per il carattere in posizione k=6+9-3=12:

u= aabaacaabaac…

v= aabaacaabaab…

La trasformata estesa

Input:

$S=\{u_1, u_2, \ldots, u_k\}$, dove $u_i \in A^*$ e u_i primitiva.

Output:

- La sequenza *ebwt(S);*
- Un insieme di indici  .

Sia C(S) l'insieme delle coniugate delle parole in S.

Si associ ad ogni $w \in C(S)$, una tripla

$(pref_H(w^\omega), L(w), \chi_s(w))$ dove

$$\chi_S(w) = \begin{cases} 1 & \text{se } w \in S \\ 0 & \text{altrimenti} \end{cases}$$

e

$$H = \max\left\{ |u_i| + |u_j| - mcd\left(|u_i|, |u_j|\right) \text{ tale che } i, j = 1, \ldots, k \right\}$$

Esempio

S={*abac*, *bca*, *cbab*, *cba*}

H=6

ebwt(S)=*ccbbbcacaaabba*

✋ = {1, 9, 13, 14}

	M_C	$Pref_H$	L	χ	F
→ 1	a b a c	a b a c a b	c	1	a
2	a b c	a b c a b c	c	0	a
3	a b c b	a b c b a b	b	0	a
4	a c a b	a c a b a c	b	0	a
5	a c b	a c b a c b	b	0	a
6	b a b c	b a b c b a	c	0	b
7	b a c a	b a c a b a	a	0	b
8	b a c	b a c b a c	c	0	b
→ 9	b c a	b c a b c a	a	1	b
10	b c b a	b c b a b c	a	0	b
11	c a b a	c a b a c a	a	0	c
12	c a b	c a b c a b	b	0	c
→ 13	c b a b	c b a b c b	b	1	c
→ 14	c b a	c b a c b a	a	1	c

Proprietà della trasformata estesa

S={*abac, bca, cbab, cba*}, k=4, m=14

→	1	a b a c
	2	a b c
	3	a b c b
	4	a c a b
	5	a c b
	6	b a b c
	7	b a c a
	8	b a c
→	9	b c a
	10	b c b a
	11	c a b a
	12	c a b
→	13	c b a b
→	14	c b a

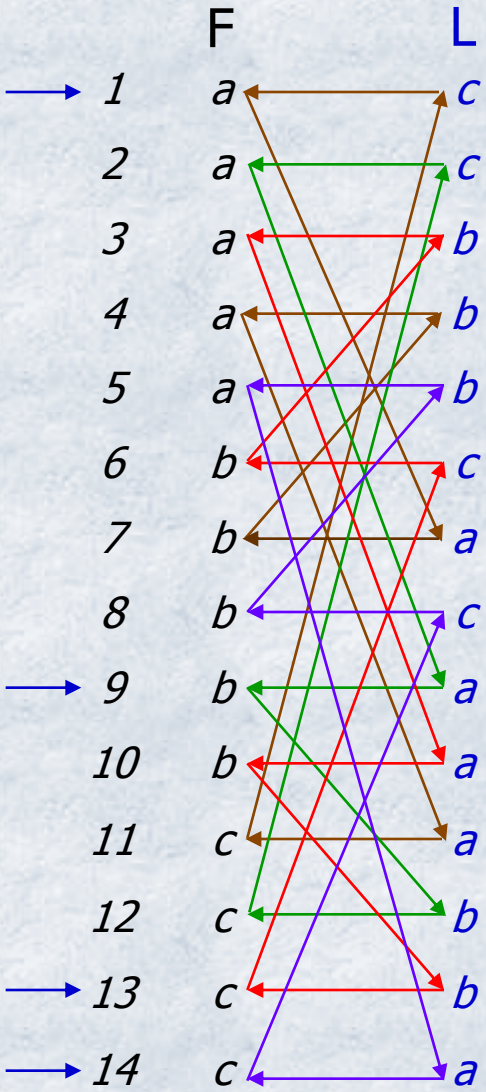
Sia $S=\{u_1, u_2, \ldots, u_k\}$, dove $u_j \in A^*$ $\forall j=1,\ldots,k$

Sia

$$m = \sum_{i=1}^{k} |u_i|$$

- Per un dato carattere α, l'i-esima occorrenza di α in F corrisponde all'i-esima occorrenza di α in L
- Per ogni i=1,…,m, $i \notin$ ✋, F[i] segue L[i] in una delle parole in S.
- Il primo carattere della parola u_j coincide con il carattere F[✋$_j$], per ogni j=1, …, k

Trasformata estesa inversa



ebwt(S)=L=*ccbbbcacaaabba*

✋={1, 9, 13, 14}

F=	a	a	a	a	a	b	b	b	b	b	c	c	c	c
θ =	1	2	3	4	5	6	7	8	9	10	11	12	13	14
	7	9	10	11	14	3	4	5	12	13	1	2	6	8
L=	c	c	b	b	b	c	a	c	a	a	a	b	b	a

θ = (1 7 4 11) (2 9 12) (3 10 13 6) (5 14 8)

a b a c a b c a b c b a c b

S={*abac*, *bca*, *cbab*, *cba*}

Reversibilità e ordinamento

La trasformata estesa non è reversibile nel caso si usi l'ordinamento lessicografico.

Ad esempio, se S={u=ab, v=aba} usando $<_{lex}$ si avrebbe: ebwt(S)=bbaaa e ✋={2, 3}.

	M_c	L	M_χ	F
1	a a b	b	0	a
2	a b	b	1	a
3	a b a	a	1	a
4	b a	a	0	b
5	b a a	a	0	b

$$\theta = \begin{pmatrix} 1 & 2 & 3 & 4 & 5 \\ 3 & 4 & 5 & 1 & 2 \end{pmatrix} = (1 \quad 3 \quad 5 \quad 2 \quad 4)$$

L'anti-trasformata produrrebbe: S={abaab, ababa}

Si osservi che $ab <_{lex} aba$, ma $aba \leq_\omega ab$, infatti u^ω e v^ω differiscono per il carattere in posizione k=2+3-1=4:

abab...

abaa...

Esempio

	F	L
→ 1	a	c
2	a	c
3	a	b
→ 4	a	b
5	a	b
6	b	c
→ 7	b	a
8	b	c
9	b	a
10	b	a
→ 11	c	a
12	c	b
13	c	b
14	c	a

ebwt(S)=L=*ccbbbcacaaabba*

=={1, 9, 13, 14}

→ S={*abac*, *bca*, *cbab*, *cba*}

F= a a a a a b b b b b c c c c

θ = (1) 2 3 (4) 5 6 (7) 8 9 10 (11) 12 13 14

7 9 10 11 14 3 4 5 12 13 1 2 6 8

L= c c b b b c a c a a a b b a

θ = (1 7 4 11) (2 9 12) (3 10 13 6) (5 14 8)

a b a c *a b c* *a b c b* *a c b*

abac
baca
acab
caba

abc
bca
cab

abcb
bcba
cbab
babc

acb
cba
bac

S={*abac*, *abc*, *abcb*, *acb*}

Biezione

Proposizione.

Per ogni parola L, esiste un multinsieme S di parole tale che ebwt(S)=L.

Esempio: L=babacab, non esiste nessuna parola u, tale che bwt(u)=L, ma esiste un multinsieme S={aab, abcb}, tale che ebwt(S)=L.

Teorema.

La trasformata estesa definisce una biezione tra le parole in A^* e la famiglia dei multinsiemi finiti delle classi di coniugio delle parole primitive in A^*.

Applicazioni

- Confronto di sequenze
- Compressione di testi

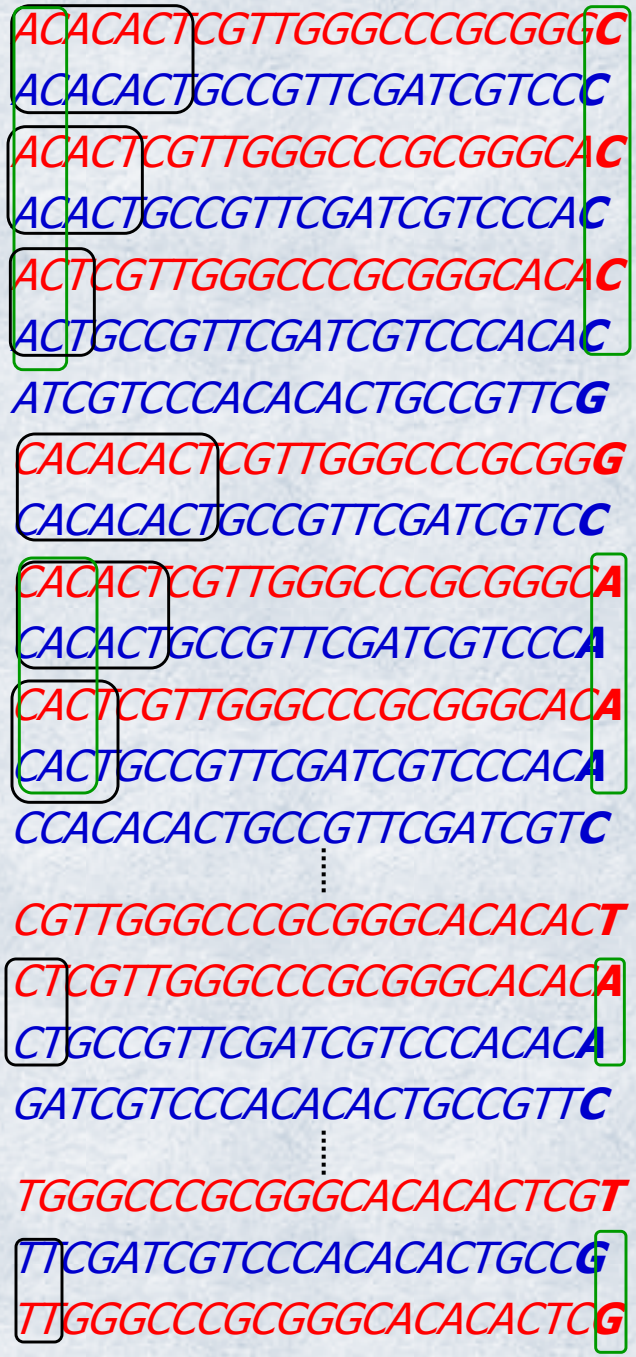
Confronto di sequenze

Esempio:

u=GTTGGGCCCGCGGGCACACACTC,

v=TTCGATCGTCCCACACACTGCCG

$$\rho(u,v) = \sum_{i=1}^{k} \left| c_i(u) - c_i(v) \right|$$



Misura basata sui blocchi monotoni Esempio

Esempio: S={*u = aaabbbb*, *v = abaabbb*}

	Mc	*ebwt*	$\lvert c_i(u) - c_j(v) \rvert$
1	*aaabbbb*	*b*	*0*
2	*aabbbab*	*b*	
3	*aabbbba*	*a*	*1*
4	*abaabbb*	*b*	*1*
5	*abbbaba*	*a*	*0*
6	*abbbbaa*	*a*	
7	*baaabbb*	*b*	*1*
8	*baabbba*	*a*	*1*
9	*babaabb*	*b*	*0*
10	*bbaaabb*	*b*	
11	*bbabaab*	*b*	
12	*bbbaaab*	*b*	
13	*bbbabaa*	*a*	*0*
14	*bbbbaaa*	*a*	

$$\rho(u,v) = \sum_{i=1}^{k} \left| c_i(u) - c_i(v) \right|$$

$\rho(u,v)=4$

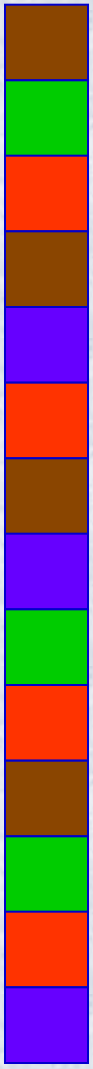
Proprietà della misura ρ

1. $\rho(u,v)= \rho(v,u)$, cioè la misura ρ è simmetrica.
2. Prese due parole u e v, allora $\rho(u,v)=0$ se e solo se u e v sono coniugate.
3. Se u' è una coniugata di u e v' è una coniugata di v, allora $\rho(u,v)= \rho(u',v')$.
4. $S=\{u_1, u_2, \ldots, u_k\}$, con $k\geq 2$ è possibile calcolare la misura ρ fra tutte le coppie di sequenze in S applicando la trasformata estesa un'unica volta all'intero multinsieme S.

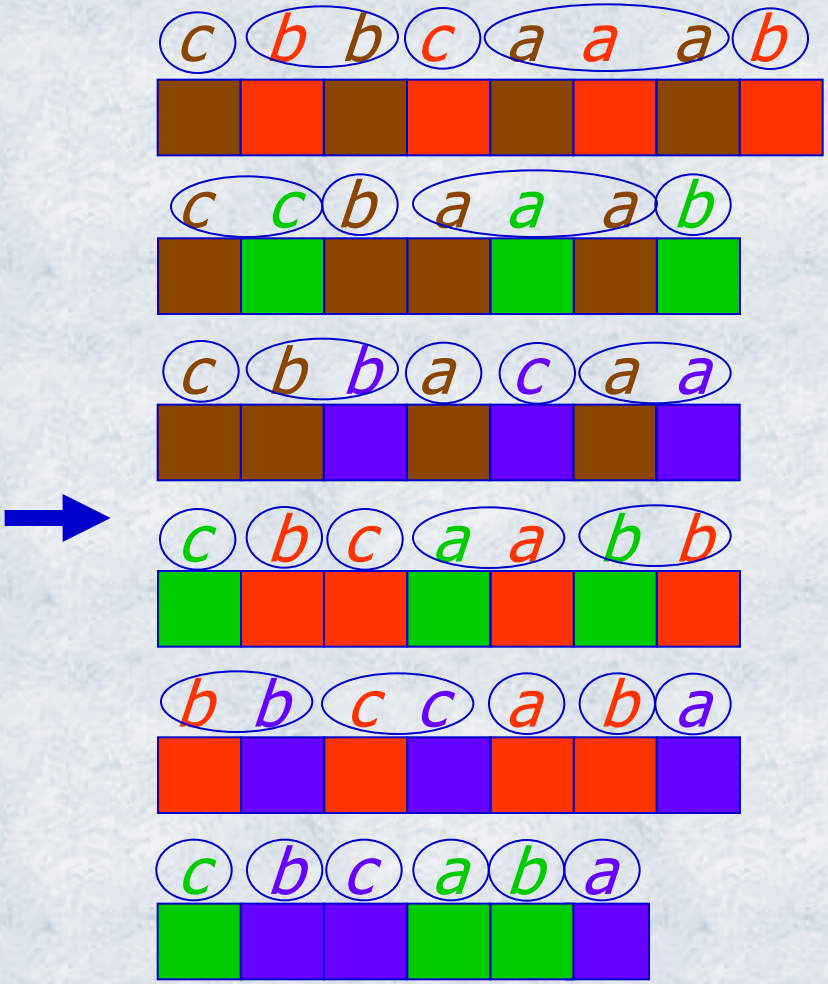
- ρ non soddisfa la disuguaglianza triangolare. Per esempio se u=aaabbbb, v=abaabbb, z=abababb, si ha $\rho(u,v)=4$, $\rho(v,z)=6$ e $\rho(u,z)=12$.

Esempio

M_c	*ebwt*
a b a c	c
a b c	c
a b c b	b
a c a b	b
a c b	b
b a b c	c
b a c a	a
b a c	c
b c a	a
b c b a	a
c a b a	a
c a b	b
c b a b	b
c b a	a



S={u=abac, v=cbab, w=bca, z=cba}

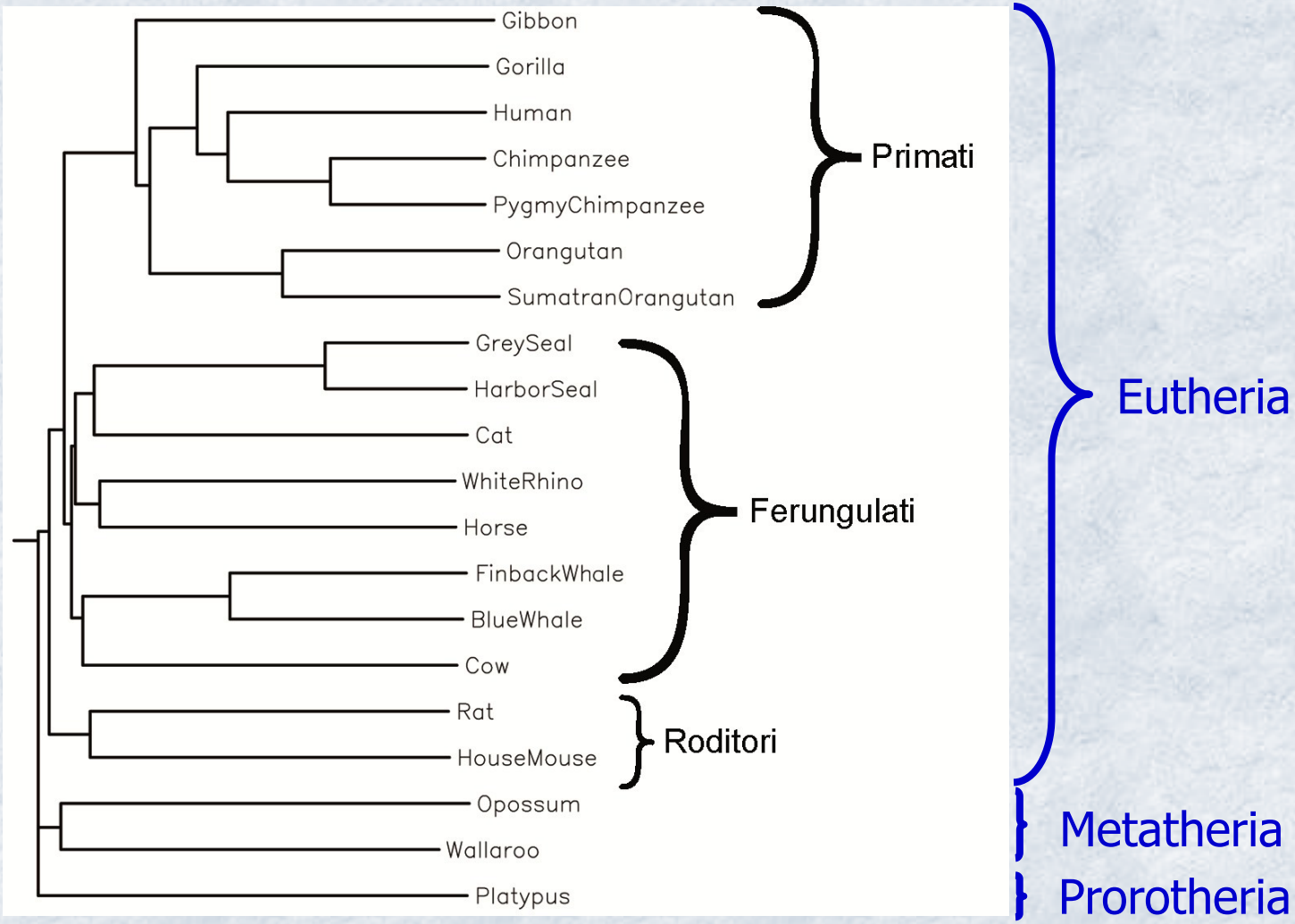


Esperimenti sulla filogenesi dei mammiferi

- La classe dei Mammiferi è suddivisa nelle Sottoclassi:
  - Prototheria (monotremi: mammiferi che si riproducono con le uova): *Platypus*.
  - Metatheria (marsupiali: mammiferi che si riproducono usando il marsupio): *Walloroo* e *Opossum*.
  - Eutheria (mammiferi placentali: mammiferi che si riproducono mediante la placenta).

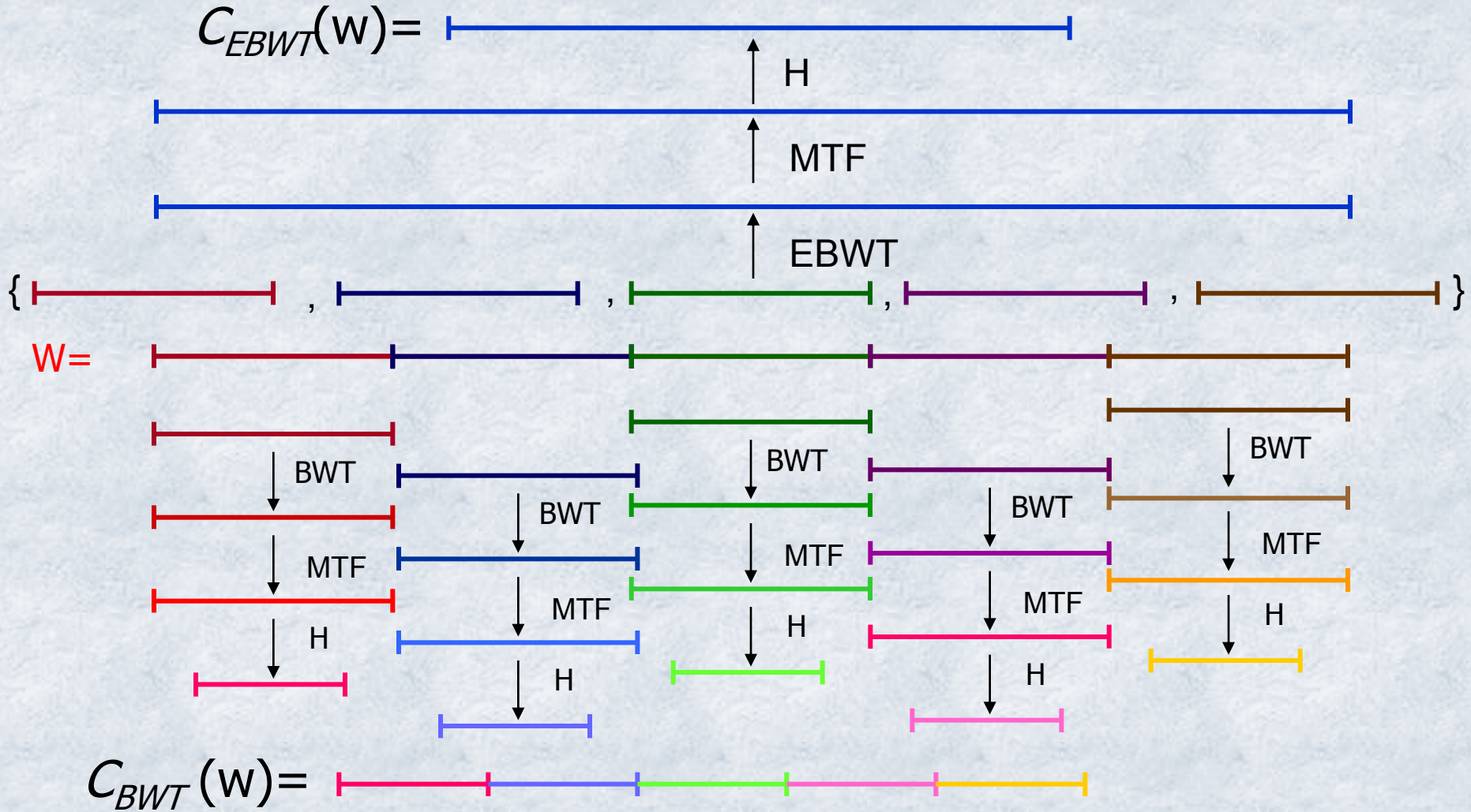
- Quali sono i gruppi dei mammiferi placentali (Primati, Ferungulati e Roditori) più vicini?

Albero filogenetico con la misura ρ mediante il metodo neighbor-joining



(Prototheria, Metatheria, (Roditori, (Primati, Ferungulati)))

Compressione



Grazie per l’attenzione