

# ASSESSING THE SIGNIFICANCE OF CONSERVED GENOMIC REGIONS

Simona Grusea

INSA de Toulouse, Institut Mathématique de Toulouse

Journées Statistiques du Sud 2010  
Mèze, 23 juin 2010

- 1 BIOLOGICAL CONTEXT
- 2 COMPOUND POISSON APPROXIMATION
- 3 MEASURES FOR THE EXCEPTIONALITY OF GENE ORDER

1 BIOLOGICAL CONTEXT

2 COMPOUND POISSON APPROXIMATION

3 MEASURES FOR THE EXCEPTIONALITY OF GENE ORDER

## CONSERVED GENOMIC REGION

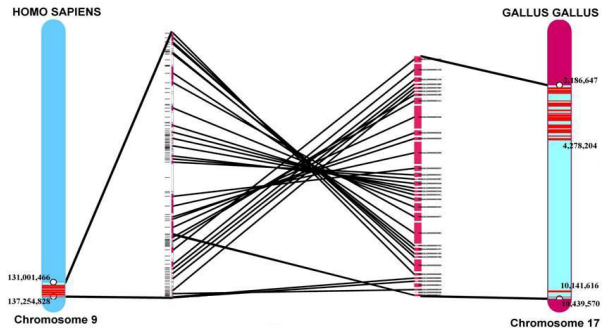
*Two chromosomic regions, in two different species, sharing a certain number of orthologous genes.*

(No restriction on the gaps between the orthologs or on their order).

*Significant, if very improbable to have appeared by chance.*

↪ a *sign* for

- evolutionary relationships between species ;
- functional selective pressures.



- *reference-region* approach.

# STATISTICAL TEST FOR SIGNIFICANCE

$H_0$  : random gene order in the genome B.

Statistical test :

- adapted to the reference-region approach ;
- taking into account the existence of multigene families (co-orthologs) ;
- taking into account gene order and gene orientation.

## THE DATA

- $m$  : the # of genes in the reference region having orthologs in B ;
- $\phi_j$  : the # of orthologs in B for the gene  $j$  from the reference-region,  $j = 1, \dots, m$  ;
- $n = \phi_1 + \dots + \phi_m$  : the # of genes in B, orthologous of genes from the reference region ;
- $N$  : the size of the genome B.

## NOTATION

- $\phi_{min} := \min\{\phi_j : j = 1, \dots, m\}$
- $n_{min} := \#\{j : \phi_j = \phi_{min}\}$ .

## SIMPLIFIED MATHEMATICAL MODEL

- The genome  $B$  as the interval  $[0,1]$  ;
- Under  $H_0$ , the positions in  $B$  of the  $n$  orthologs as i.i.d. r.v.'s  $\sim \mathcal{U}([0, 1])$ .

## TAKING INTO ACCOUNT THE MULTIGENE FAMILIES

The *weight* of an interval  $I \subset [0, 1]$  :

$$\mu_m(I) = \sum_{i=1}^n L_i \mathbf{1}_{\{U_{(i)} \in I\}}$$

- $U_{(1)} < \dots < U_{(n)}$  : order statistics of  $n$  i.i.d. r.v.'s  $\sim \mathcal{U}([0, 1])$  ;
- $L_i = \frac{1}{\phi_j}$ , if the gene in position  $U_{(i)}$  is orthologous to the gene  $j$  from the reference region.

## CLUSTER SIGNIFICANCE

Observed cluster, of weight  $h$  and length  $r \in (0, 1)$ .

Let  $h_* := \lceil h\phi_{min} \rceil$ .

- For every  $i = 1, \dots, n - h_* + 1$ ,

$$A_i := \{\mu_m([U_{(i)}, U_{(i)} + r]) \geq h\}$$

(the event of having a cluster of weight  $\geq h$  and of length  $\leq r$  starting with the  $i$ -th ortholog).

- The # of clusters of type  $(h, r)$  in B :

$$W_m := \sum_{i=1}^{n-h_*+1} \mathbf{1}_{A_i}.$$

$\rightsquigarrow$  the  $p$ -value of the cluster =  $\mathbb{P}(\bigcup_i A_i) = \mathbb{P}(W_m \geq 1)$ .

- 1 BIOLOGICAL CONTEXT
- 2 COMPOUND POISSON APPROXIMATION
- 3 MEASURES FOR THE EXCEPTIONALITY OF GENE ORDER

# COMPOUND POISSON APPROXIMATION FOR $W_m$

The events  $A_i$  are *locally strongly dependent*.

Approximate the distribution of  $W_m$  by a compound Poisson distribution.

## COMPOUND POISSON DISTRIBUTION

$$CP(\boldsymbol{\lambda}) = \mathcal{L}\left(\sum_{j=1}^M X_j\right),$$

- $\boldsymbol{\lambda} = \sum_{i=1}^{\infty} \lambda_i \delta_i$ ,  $\Lambda := \sum_{i=1}^{\infty} \lambda_i < \infty$ ;
- $X_j \sim \frac{1}{\Lambda} \boldsymbol{\lambda}$  are i.i.d., independent of  $M \sim \text{Poisson}(\Lambda)$ .

# STEIN-CHEN METHOD, THE COUPLING APPROACH

BARBOUR, CHEN, LOH (1992), ROOS (1993)

The # of events which occur in the neighborhood of *strong dependence* of  $A_k$  :

$$Z_k := \sum_{j=k-(h_*-2)}^{k+h_*-2} \mathbf{1}_{A_j}.$$

## THE CHOICE OF $\lambda$

$$\hat{\lambda}_i := \frac{1}{i} \sum_{k=1}^{n-h_*+1} \mathbb{P}(A_k \cap \{Z_k = i\}), \quad i = 2, \dots, h_* - 1,$$

$$\hat{\lambda}_1 = \mathbb{E}(W_m) - \sum_{i=2}^{h_*-1} i \hat{\lambda}_i.$$

$$\mathbb{P}(W_m \geq 1) \simeq 1 - \exp \left\{ - \sum_{i=1}^{h_*-1} \hat{\lambda}_i \right\}.$$

## THEOREM

If  $n \rightarrow \infty, r \rightarrow 0$  s.t.  $nr \rightarrow 0$ ,  $n_{min} \asymp n$  and  $\mathbb{E}(W_m)$  is constant, then the error of the approximation is  $\mathcal{O}(\frac{1}{n})$ .

## COMPUTATION OF THE PARAMETERS

For  $i = 2, \dots, h_* - 1$  :

$$\hat{\lambda}_i \simeq (n - h_* + 1) \mathbb{P}(A_1^C \cap A_2 \cap A_3 \cap \dots \cap A_{i+1} \cap A_{i+2}^C).$$

“Markovian” approximation :

$$\hat{\lambda}_i \simeq (n - h_* + 1) \pi q^{i-1} (1 - q)^2,$$

where  $\pi = \mathbb{P}(A_1)$  and  $q = \mathbb{P}(A_2|A_1)$ .

- 1 BIOLOGICAL CONTEXT
- 2 COMPOUND POISSON APPROXIMATION
- 3 MEASURES FOR THE EXCEPTIONALITY OF GENE ORDER

# THE CASE OF NO MULTIGENE FAMILIES

*Observed* conserved genomic region  $\mathcal{R}$  in the genome B :

- containing  $h$  orthologs ;
- starting with the  $i$ -th ortholog, in position  $U_{(i)}$ .

*Goal : measure the exceptionality of the order of the orthologs in  $\mathcal{R}$  / the reference region.*

*Problem : find a “good” measure*

(biologically relevant + computationally accessible).

$Id_n$  : the order of the orthologs in the reference region ;

$\pi$  : the order of the orthologs in the genome B ;

$\pi_{i,h} = [\pi(i), \dots, \pi(i + h - 1)]$  : the order of the  $h$  orthologs in  $\mathcal{R}$ .

Under  $H_0$  :

- $\pi$  is a random uniform permutation in  $S_n$  ;
- $\pi_{i,h}$  is a random uniform permutation of  $h$  elements out of  $n$ .

# THE TRANSPOSITION DISTANCE

- $d_{trp}(\sigma, Id_n) = n - c(\sigma)$ ;
- # of permutations of  $S_n$  having  $k$  cycles =  $s(n, k)$   
(signless Stirling number of first kind).

## NOTATION

$$p(n, k) := \frac{1}{n!} \sum_{j=k}^n s(n, j),$$

*the probability that a random permutation in  $S_n$  has at least  $k$  cycles.*

## A SIMPLE STRATEGY...

**Example :**

$$n = 10, h = 5, i = 3, \pi_{i,h} = [1, 7, 5, 2, 4].$$

Compare the order of the  $h$  orthologs in  $\mathcal{R}$  vs. the reference region.

$$\begin{pmatrix} 1 & 2 & 4 & 5 & 7 \\ 1 & 7 & 4 & 2 & 5 \end{pmatrix}$$

$$\rightsquigarrow d_{trp} = 5 - 3 = 2.$$

## A SIMPLE STRATEGY...

**Example :**

$$n = 10, h = 5, i = 3, \pi_{i,h} = [1, 7, 5, 2, 4].$$

Compare the order of the  $h$  orthologs in  $\mathcal{R}$  vs. the reference region.

$$\begin{pmatrix} 1 & 2 & 4 & 5 & 7 \\ 1 & 7 & 4 & 2 & 5 \end{pmatrix}$$

$$\rightsquigarrow d_{trp} = 5 - 3 = 2.$$

Under  $H_0$ , random permutation in  $S_h$ .

### PROPOSITION

For  $0 \leq d \leq h - 1$

$$\mathbb{P}(d_{trp}(\pi_{i,h}, Id_n|_{\{\pi(i), \dots, \pi(i+h-1)\}}) \leq d) = p(h, h - d).$$

## ...A MORE COMPLICATED STRATEGY...

$$d(\pi_{i,h}, Id_n) := \min\{d_{trp}(\sigma, Id_n) : \sigma \in S_n, \sigma_{i,h} = \pi_{i,h}\}.$$

**Example :** Here  $i = 3$ .

$$\begin{pmatrix} 1 & 2 & 3 & 4 & 5 & 6 & 7 & 8 & 9 & 10 \\ \cdot & \cdot & 1 & 7 & 4 & 2 & 5 & \cdot & \cdot & \cdot \end{pmatrix}$$

## ...A MORE COMPLICATED STRATEGY...

$$d(\pi_{i,h}, Id_n) := \min\{d_{trp}(\sigma, Id_n) : \sigma \in S_n, \sigma_{i,h} = \pi_{i,h}\}.$$

**Example :** Here  $i = 3$ .

$$\begin{pmatrix} 1 & 2 & 3 & 4 & 5 & 6 & 7 & 8 & 9 & 10 \\ 3 & . & 1 & 7 & 4 & 2 & 5 & . & . & . \end{pmatrix}$$

## ...A MORE COMPLICATED STRATEGY...

$$d(\pi_{i,h}, Id_n) := \min\{d_{trp}(\sigma, Id_n) : \sigma \in S_n, \sigma_{i,h} = \pi_{i,h}\}.$$

**Example :** Here  $i = 3$ .

$$\begin{pmatrix} 1 & 2 & 3 & 4 & 5 & 6 & 7 & 8 & 9 & 10 \\ 3 & 6 & 1 & 7 & 4 & 2 & 5 & . & . & . \end{pmatrix}$$

## ...A MORE COMPLICATED STRATEGY...

$$d(\pi_{i,h}, Id_n) := \min\{d_{trp}(\sigma, Id_n) : \sigma \in S_n, \sigma_{i,h} = \pi_{i,h}\}.$$

**Example :** Here  $i = 3$ .

$$\begin{pmatrix} 1 & 2 & 3 & 4 & 5 & 6 & 7 & 8 & 9 & 10 \\ 3 & 6 & 1 & 7 & 4 & 2 & 5 & 8 & 9 & 10 \end{pmatrix}$$

- $cc(\pi_{i,h}) := \#$  of *closed cycles* of  $\pi_{i,h}$ .
- We have

$$d(\pi_{i,h}, Id_n) = n - \{cc(\pi_{i,h}) + (n - h)\} = h - cc(\pi_{i,h}).$$

**Example :**

$$\begin{pmatrix} 1 & 2 & 3 & 4 & 5 & 6 & 7 & 8 & 9 & 10 \\ 3 & 6 & 1 & 7 & 4 & 2 & 5 & 8 & 9 & 10 \end{pmatrix}$$

Here

- $cc(\pi_{i,h}) = 1$  (the cycle  $(4, 7, 5)$ );
- $d(\pi_{i,h}, Id_n) = 5 - 1 = 4$ .

## THEOREM

For  $0 \leq d \leq h - 1$

$$\mathbb{P}(d(\pi_{i,h}, Id_n) \leq d) = \frac{1}{\binom{n}{h}} \sum_{m=h-d}^h \binom{n-m-1}{n-h-1} p(m, h-d).$$

## SKETCH OF THE PROOF

- Condition on

$M := \#$  of elements of  $\{i, \dots, i + h - 1\}$  in closed cycles of  $\pi_{i,h}$ ;

## SKETCH OF THE PROOF

- Condition on

$M := \#$  of elements of  $\{i, \dots, i + h - 1\}$  in closed cycles of  $\pi_{i,h}$ ;

- $\mathbb{P}(d(\pi_{i,h}, Id_n) \leq d | M = m) = p(m, h - d)$ ;

## SKETCH OF THE PROOF

- Condition on

$M := \#$  of elements of  $\{i, \dots, i + h - 1\}$  in closed cycles of  $\pi_{i,h}$ ;

- $\mathbb{P}(d(\pi_{i,h}, Id_n) \leq d | M = m) = p(m, h - d)$ ;

$$\begin{aligned}\mathbb{P}(M = m) &= \frac{\#\{\pi_{i,h} : M = m\}}{\binom{n}{h} h!} \\ &= \frac{1}{\binom{n}{h} h!} \sum_{\substack{k_1, \dots, k_{n-h} \geq 0 \\ k_1 + \dots + k_{n-h} = h - m}} \binom{h}{k_1, \dots, k_{n-h}, m} m! \prod_{j=1}^{n-h} k_j! \\ &= \frac{1}{\binom{n}{h}} \binom{n - m - 1}{n - h - 1}.\end{aligned}$$

## ...A TOO COMPLICATED STRATEGY...

$$d(\pi_{i,h}, Id_n) := \min_{1 \leq k \leq n-h+1} \min\{d_{trp}(\sigma, Id_n) : \sigma \in S_n, \sigma_{k,h} = \pi_{i,h}\}.$$

## ...AND A STRATEGY OF COMPROMISE

$$d(\pi_{i,h}, Id_n) := \min\{d_{trp}(\sigma, Id_n) : \sigma \in S_n, \sigma_{i^*,h} = \pi_{i,h}\},$$

where  $i^* \in \{1, \dots, n - h + 1\}$  maximizes

$$L^* := \#\{\pi(i), \dots, \pi(i + h - 1)\} \cap \{i^*, \dots, i^* + h - 1\}.$$

**Example :** Here  $i^* = 1$  and  $L^* = 4$ .

$$\begin{pmatrix} 1 & 2 & 3 & 4 & 5 & 6 & 7 & 8 & 9 & 10 \\ 1 & 7 & 4 & 2 & 5 & . & . & . & . & . \end{pmatrix}$$

## ...AND A STRATEGY OF COMPROMISE

$$d(\pi_{i,h}, Id_n) := \min\{d_{trp}(\sigma, Id_n) : \sigma \in S_n, \sigma_{i^*,h} = \pi_{i,h}\},$$

where  $i^* \in \{1, \dots, n - h + 1\}$  maximizes

$$L^* := \#\{\pi(i), \dots, \pi(i + h - 1)\} \cap \{i^*, \dots, i^* + h - 1\}.$$

**Example :** Here  $i^* = 1$  and  $L^* = 4$ .

$$\begin{pmatrix} 1 & 2 & 3 & 4 & 5 & 6 & 7 & 8 & 9 & 10 \\ 1 & 7 & 4 & 2 & 5 & 6 & 3 & 8 & 9 & 10 \end{pmatrix}$$

- $cc(\pi_{i,h}, i^*) = 2$  (the cycles (1), (5));
- $d^*(\pi_{i,h}, Id_n) = 5 - 2 = 3$ .

# THE DISTRIBUTION OF $d^*$

Condition on

$$L^* = \#\{\pi(i), \dots, \pi(i+h-1)\} \cap \{i^*, \dots, i^*+h-1\}.$$

The distribution of  $L^*$  is known (*conditional discrete scan statistic*).

## PROPOSITION

For  $0 \leq d \leq h-1$  and  $h-d \leq \ell \leq h$

$$\mathbb{P}(d^*(\pi_{i,h}, Id_n) \leq d | L^* = \ell) = \frac{1}{\binom{h}{\ell}} \sum_{m=h-d}^{\ell} \binom{h-m-1}{h-\ell-1} p(m, h-d).$$

# PERSPECTIVES

- try other ways of taking into account the multigene families ;
- take into account the gene order in the case of multigene families (distances between multipermutations) ;
- generalize to other distances between permutations ;
- take into account gene orientation also (distances between signed permutations) ;
- proximity + order = ?
- multiple genomes comparison.

*MERCI*

*à Etienne Pardoux, Pierre Pontarotti et  
l'Equipe Evolution Biologique et Modélisation,  
L.A.T.P, Université de Provence,*

*aux organisateurs pour l'invitation,*

*à vous pour votre attention!*