A benchmark of methods for horizontal transfer detection

Jennifer Becq[1], Ludovic Mallet[1], Cécile Churlaud[1] & Patrick Deschavanne[1]

[1] Équipe de Bioinformatique Génomique et Moléculaire, Inserm UMRS-726, Université Paris 7, France

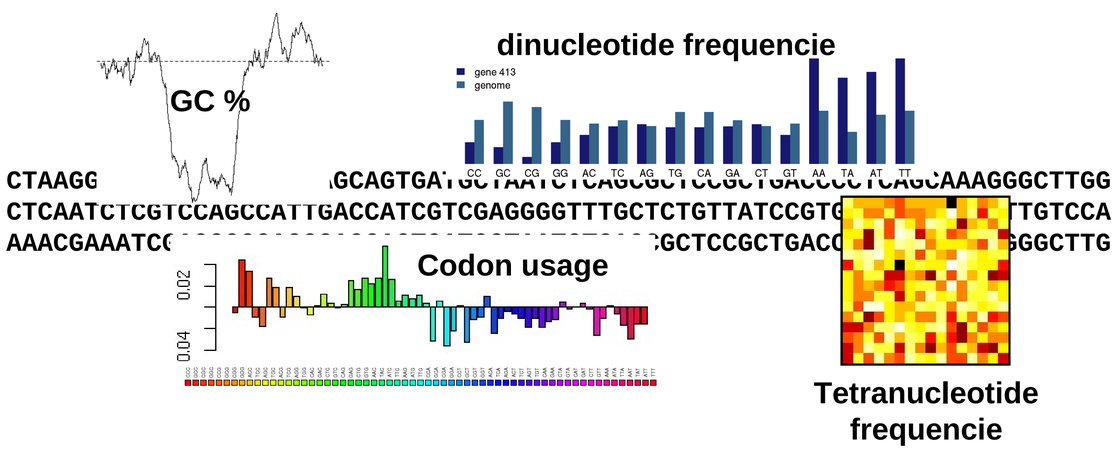
becq@ebgm.jussieu.fr

Introduction

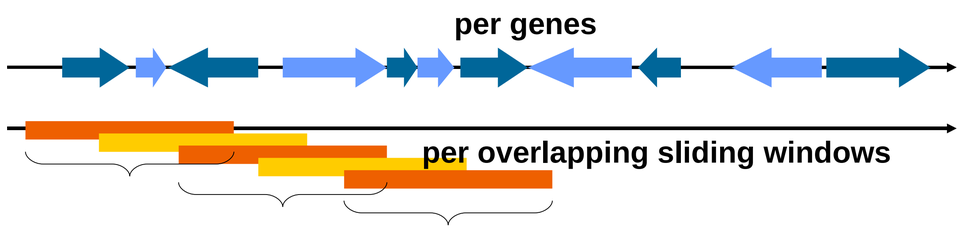
From the sequencing of complete genomes it appears that horizontal transfer (HT) was a significant factor of prokaryote evolution. Besides phylogenetic analysis that was the original method, numerous parametric methods were published in order to detect transferred genes or regions. These parametric methods often lead to different results[1,2,3] and, in absence of a gold standard reference, it is difficult to evaluate the quality of the results. It was proposed that the different methods, based on different working hypotheses, detect different types of HT. Using artificial genomes, created by a generalized hidden Markov model from genuine genomes, we performed a benchmark of almost all published methods based on compositional criteria.

I. Parametric methods

Parametric methods are based on the **compositional characteristics** of a given genome. Different **criteria** were used :



Different **genome *scanning* procedures :**



Different **scoring formulas :**

$$S(g) = \frac{1}{m}\sum_{i}^{m} \left| f_g^i - f_{WG}^i \right|$$
Absolute difference

$$S(g) = \frac{1}{m}\sum_{i}^{m} f_g^i \cdot f_{WG}^i$$
Covariance

$$S(g) = \sqrt{\sum_i \left(f_g^i - f_{WG}^i \right)^2}$$
Euclidean distance

$$S(g) = \sum_i \frac{(f_g^i - \bar{f}^i)^2}{\bar{f}^i}$$
Chi-2 test

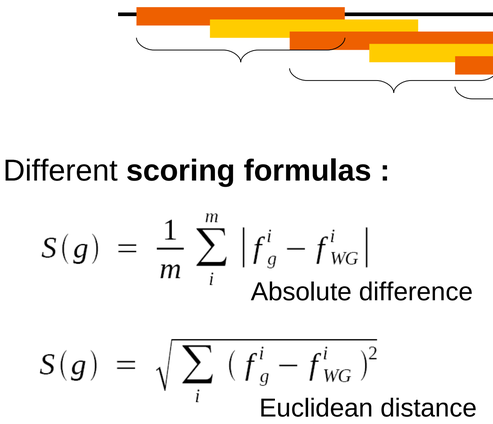
$$S(g) = \sum_i f_g^i \cdot \ln \frac{f_g^i}{f_{WG}^i}$$
Kullback-Leibler distance

$$S(g) = \left(f_g^i - f_{WG}^i \right)^T \cdot S^{-1} \cdot \left(f_g^i - f_{WG}^i \right)$$
Mahalanobis distance
(S=covariance matrix)

➡ **16 parametric methods were tested**

II. Artificial genomes

Azad *et al.*[4] has generated **11 artificial genomes** by using a generalized Markov model from genuine genomes. These homogeneous genomes were **combined** to create **different sets** of artificial **horizontal transfers**.



E. coli was used as host genome

Donor genomes where clustered into three groups according to proximity to *E. coli* :

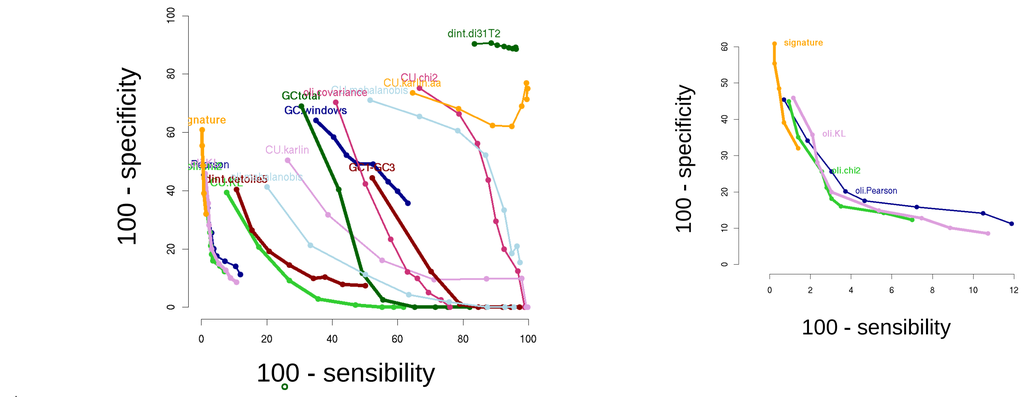
- **close** (in shades of blue)
- **Intermediate** (green colors)
- **Far** (red, orange and pink)

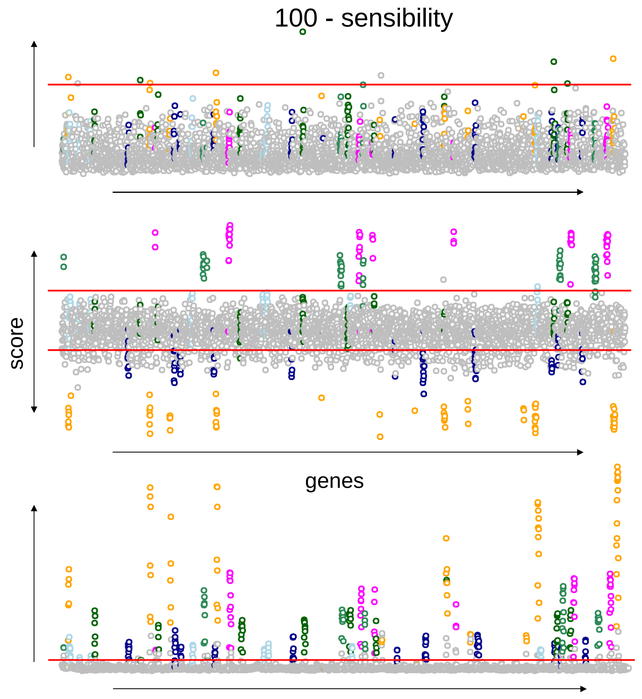
The methods were assessed according to

- the **origin** of the HT, *i.e.* the proximity of the donor species
- the overall **percentage** of HT in the genome, from 1 to 20 %
- the **size** of the HT islands in terms of number of genes per transfer

III. Threshold determination

The best threshold for each HT detection method was determined by **ROC curves**





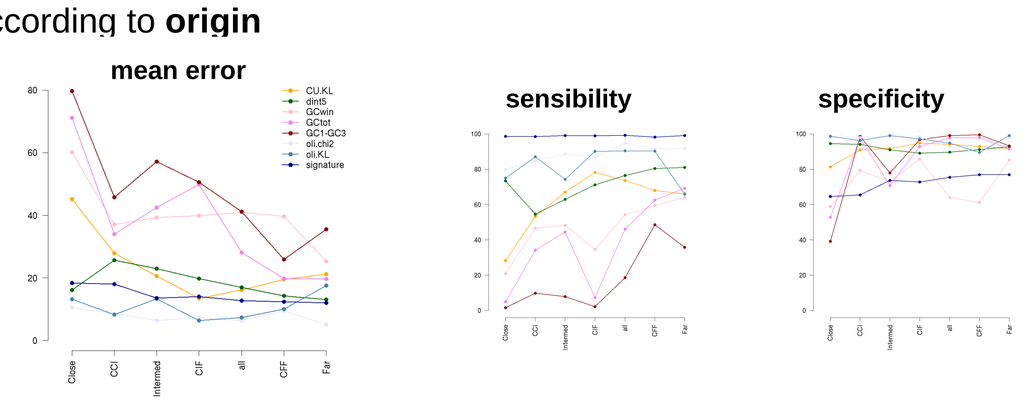
Codon usage using **Mahalanobis** (CU.mahalanobis) : no specificity, HT genes can not be distinguished from host genes.

GC1-GC3 % : good sensibility for distant HTs but close HT genes resemble host genes.

Tetranucleotide relative frequency using **chi2** measure (oli.chi2) : good sensibility and specificity.

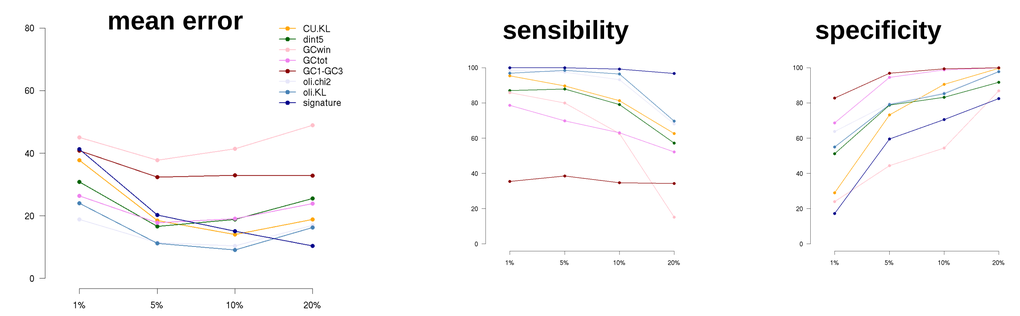
IV. Method evaluation

According to **origin**



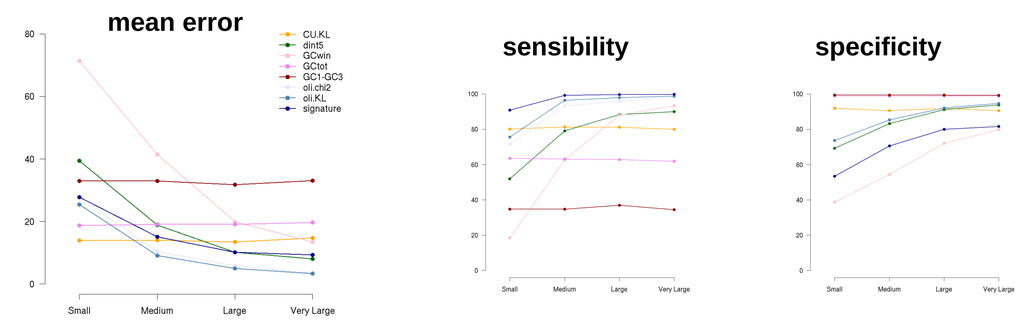
For the majority of the methods, the farther the donor species is the better the results are. Most methods improve in sensibility, *i.e.* farther HT are better detected.

According to **overall percentage of HT**



Most methods worsen when there are too few (1%) or too many (20%) HTs.

According to **HT island size**



Gene based methods (CU.KL – codon usage using Kullback-Leibler – and GCtot) present better results than window based methods for small islands (1 to 5 genes).

Conclusion & Perspectives

Even though the methods are tested in “perfect” conditions (intra-genomic variability is reduced and HTs are not ameliorated), the methods present very variable results. Some methods only detect one type of HT. Improvements are needed concerning threshold determination and the size of the windows for genome scanning. For the best results, we suggest in a preliminary step, to measure the intra-genomic variability of the genome under consideration in order to determine the best method to use and then to combine a window based method with a gene based method.

References

[1] M.A. Ragan, **On surrogate methods for detection lateral gene transfer**. *FEMS Microbiol. Lett.*, 201(2):187-191, 2001.

[2] J.G. Lawrence and H. Ochman, **Reconciling the many faces of lateral gene transfer**. *Trends Microbiol.* 10(1):1-4, 2002.

[3] C. Dufraigne, B. Fertil, S. Lespinat, A. Giron and P. Deschavanne, **Detection and characterization of horizontal transfers in prokaryotes using genomic signature**. *Nucleic Acids Res.*, 33(1):e6, 2005.

[4] R.K. Azad and J.G. Lawrence, **Use of artificial genomes in assessing methods for atypical gene detection.** *PloS Comput. Biol.*, 1(6):e56, 2005.