Mathematical models of genome-level evolution of protein domains

The distributions of genome-associated quantities, including **protein domain** families can be approximated with **power laws**. The availability of a huge number of sequenced genomes has lead to the development of mathematical models to account for the shape of these distributions and to describe evolutionary aspects of genome and domain evolution.

Birth, death and innovation (BDI) models are stochastic and stationary models [1] that account for many of the observed characteristics and allow to estimate evolutionary parameters such as the rates of domain acquisition and deletion *for each species*. Other models such as the **Chinese restaurant (CR)** are stochastic and dynamic models [2] and allow the exploration of *universal features*.

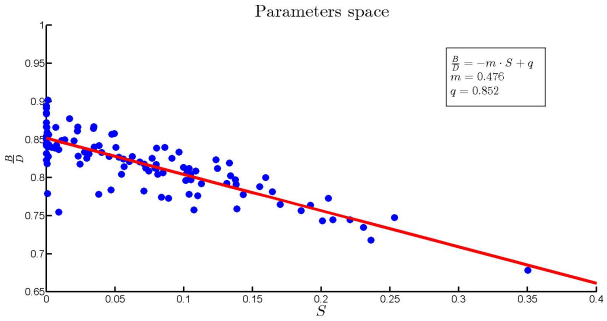

Figure 2 BDI model of genome evolution: Comparison between the ratio between the Birth and Death rates (B/D) and the Innovation rate (S). Each point represents a bacterial family (with at least 3 sequenced genomes).

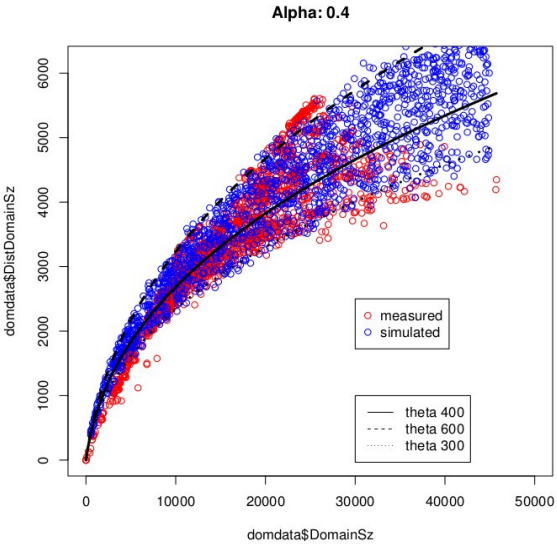

Figure 1: CR model of genome evolution. Comparison between the total number of protein domains and the total number of distinct protein domains. Each point represents a genome

Both approaches use different hypothesis and assumptions and both show a high degree of agreement between model predictions and experimental data. The *goal of this project is to connect both approaches.*

In the computational systems biology group we have analyzed the genome of more than 300 bacterial species, which provides and ideal dataset for the testing and development of this kind of models.

Requirements: The analytical analysis of these models requires the manipulation of (simple) differential equations. The numerical analysis and the simulations will be done using R or Matlab.

Thesis outline:

- Apply the BDI and CR models to the colection of bacterial genomes.
- Explore the space of parameters and the goodness of the fits.
- Relate both modelling approaches

Contact Details: **Maria Suárez Diez** **maria.suarezdiez@wur.nl**
Edoardo Saccenti **edoardo.saccenti@wur.nl**

References

1. Karev GP, Wolf YI, Koonin EV (2003) *Bioinformatics* 19: 1889–1900.

2. Cosentino Lagomarsino M, Sellerio AL, Heijning PD, Bassetti B (2009) *Genome Biol* 10: R12.