Alex's contributions in biology

12 June 2019

How it started

- Lunch discussion with colleagues at *Institut de Mathématiques de Luminy*, already involved in interdisciplinary collaborations (G. Rauzy, B. Mosse, B. Host,...)
- Strong incitement for mathematicians and physicists to study questions related to genomics and bio informatics
- Long term stay of C. Landes at *Centre de Physique Thorique*
- Alex eventually decided to move to the *Informatique et Génome* research unit at *Génopole, Evry*.

Alex's point of view: Signal/Data *Understanding*

- Very much in the spirit of his work on wavelets and signal processing:
  - Search for the most suitable representation for signals/data
  - Use mathematical models as starting points to draw conclusions
- *All models are wrong, but some are useful* (George Box).
  - A main question is to find the appropriate level of model sophistication, given the complexity of the problem, the lack of knowledge, and the quality of data.
  - For Alex, *appropriate* has to be understood in terms of interpretability.

Alex's approaches gradually moved towards data analysis, algorithmics and theoretical computer science.

Some contributions

- Analysis of large families of protein sequences using rate matrices (with C. Landès, A. Hénaut, M. Holschneider...)
- Rank based classification[1] (with C. Landès, A. Hénaut, A. Dress, S. Grünewald,...): start from a dissimilarity matrix, an iterative ranking procedure yields clustering and classification.
- Alignment free sequence comparison[2] (with G. Didier, E. Corel, I. Laprévotte , C. Landès,...): define and compute, an adapted variable length local decoding of a set of sequences, and compare the compositions of sequences of this decoding.

Alex also started a new project, developing very simple geometrical descriptors for protein surfaces, based upon volume estimates for tetrahedra built from consecutive atoms. These turn out to reveal very interesting structure.

[1] Devauchelle et al, *Annals of Combinatorics* 8 (2004) 441-456

[2] G. Didier et al, *Theoretical Computer Science* 462:30 (2012), 1-11

Evolution of genomic sequences

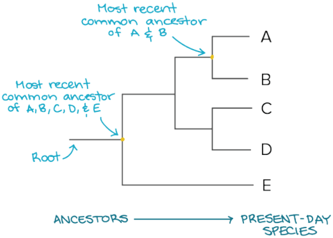
- How to compare genomic sequences (DNA, proteins) from an evolutionary perspective ?
- Phylogeny: inference of an evolutionary tree... many approaches (probability, combinatorics,...)
- Alex's approach: keep it simple ! Describe a multiple alignment using rate matrices, inspired by Markov tree models. Rate matrices provide alternative representation for data, that can be used as a starting point for further investigations.

A simple model I

- Genomic sequences at fixed time t modeled as symbolic sequences $(X_n(t))_n$, with values in a finite alphabet (4 symbols for DNA, 20 symbols for proteins)

  `QTFAVCDNVAENCMYCHCNSKGVLSHSHDIGELTHICKSSFMSIAVGNKP`

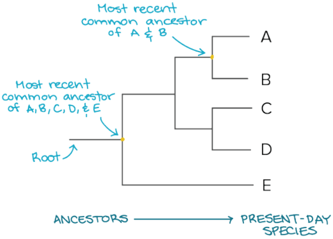
- Site independence: the X_n are modeled as *iid* random variables
- Evolution described by a tree


Most recent common ancestor of A & B
A
B
C
D
E
Most recent common ancestor of A, B, C, D, & E
Root
ANCESTORS
PRESENT-DAY SPECIES

A simple model II

- Time evolution along tree branches: stationary continuous time Markov model:

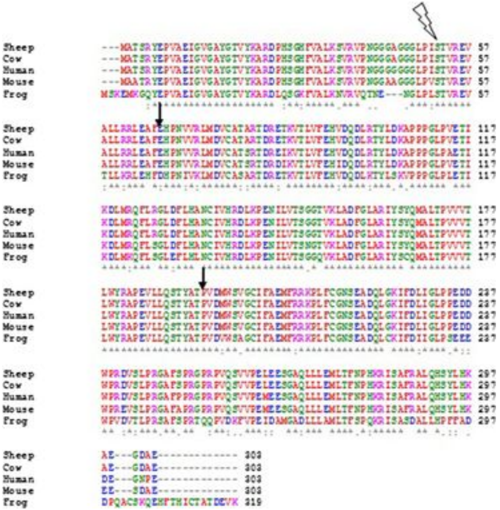
$$\mathbb{P}\{X_n(t+\tau)=a|X_n(t)=b\}=P_{ba}(\tau)\ ,\qquad P(\tau)=e^{-\tau Q}$$


Most recent common ancestor of A & B
A
B
Most recent common ancestor of A, B, C, D, & E
C
D
Root
E
ANCESTORS
PRESENT-DAY SPECIES

A simple model III

- Inference: given aligned sequences, estimate parameters (here the rate matrix) and the tree topology.
  Many sophisticated and powerful algorithms can provide estimates for parameters, sometimes uncertainty estimates
  - Maximum likelihood
  - Bayesian
  - Estimation of quad-trees followed by reconstruction
  - ...

Exploiting alignment data I



```
Sheep     ---MATSRYEPVAEIGVGAYGTVYKARDPHSGHFVALKSVRVPNGGGAGGGLPISTVREV 57
Cow       ---MATSRYEPVAEIGVGAYGTVYKARDPHSGHFVALKSVRVPNGGGAGGGLPISTVREV 57
Human     ---MATSRYEPVAEIGVGAYGTVYKARDPHSGHFVALKSVRVPNGGGGGGGLPISTVREV 57
Mouse     ---MAATRYEPVAEIGVGAYGTVYKARDPHSGHFVALKSVRVPNGGAAGGGLPVSTVREV 57
Frog      MSKEMKGQYEPVAEIGVGAYGTVYKARDLQSGKFVALKNVRVQTNE---NGLPLSTVREV 57
                  :*******************  :**:*****.***  ..     .***:******

Sheep     ALLRRLEAFEHPNVVRLMDVCATARTDRETKVTLVFEHVDQDLRTYLDKAPPPGLPVETI 117
Cow       ALLRRLEAFEHPNVVRLMDVCATARTDRETKVTLVFEHVDQDLRTYLDKAPPPGLPVETI 117
Human     ALLRRLEAFEHPNVVRLMDVCATSRTDREIKVTLVFEHVDQDLRTYLDKAPPPGLPAETI 117
Mouse     ALLRRLEAFEHPNVVRLMDVCATSRTDRDIKVTLVFEHIDQDLRTYLDKAPPPGLPVETI 117
Frog      TLLKRLEHFDHPNIVKLMDVCASARTDRETKVTLVFEHVDQDLKTYLSKVPPPGLPLETI 117
          :**:*** *:***:*:******::*****: ********:****:***.*.****** ***

Sheep     KDLMRQFLRGLDFLHANCIVHRDLKPENILVTSGGTVKLADFGLARIYSYQMALTPVVVT 177
Cow       KDLMRQFLRGLDFLHANCIVHRDLKPENILVTSGGTVKLADFGLARIYSYQMALTPVVVT 177
Human     KDLMRQFLRGLDFLHANCIVHRDLKPENILVTSGGTVKLADFGLARIYSYQMALTPVVVT 177
Mouse     KDLMRQFLSGLDFLHANCIVHRDLKPENILVTSNGTVKLADFGLARIYSYQMALTPVVVT 177
Frog      KDLMKQFLSGLEFLHLNCIVHRDLKPENILVTSGGQVKLADFGLARIYSCQMALTPVVVT 177
          ****:*** **:*** ****************.* ************* **********

Sheep     LWYRAPEVLLQSTYATPVDMWSVGCIFAEMFRRKPLFCGNSEADQLGKIFDLIGLPPEDD 237
Cow       LWYRAPEVLLQSTYATPVDMWSVGCIFAEMFRRKPLFCGNSEADQLGKIFDLIGLPPEDD 237
Human     LWYRAPEVLLQSTYATPVDMWSVGCIFAEMFRRKPLFCGNSEADQLGKIFDLIGLPPEDD 237
Mouse     LWYRAPEVLLQSTYATPVDMWSVGCIFAEMFRRKPLFCGNSEADQLGKIFDLIGLPPEDD 237
Frog      LWYRAPEVLLQSTYATPVDVWSAGCIFAEMFKRKPLFCGNSEADQLCKIFDIIGLPSEEE 237
          *******************:**.********:************** ****:****.*::

Sheep     WPRDVSLPRGAFSPRGPRPVQSVVPELEESGAQLLLEMLTFNPHKRISAFRALQHSYLHK 297
Cow       WPRDVSLPRGAFSPRGPRPVQSVVPELEESGAQLLLEMLTFNPHKRISAFRALQHSYLHK 297
Human     WPRDVSLPRGAFPPRGPRPVQSVVPEMEESGAQLLLEMLTFNPHKRISAFRALQHSYLHK 297
Mouse     WPREVSLPRGAFAPRGPRPVQSVVPEMEESGAQLLLEMLTFNPHKRISAFRALQHSYLHK 297
Frog      WPVDVTLPRSAFSPRTQQPVDHFVPEIDAMGADLLLAMLTFSPQKRISASDALLHPFFAD 297
          ** :*:***.**.**  :**:..***::  **:*** ****.*:*****  ** *.::  .

Sheep     AE---GDAE------------- 303
Cow       AE---GDAE------------- 303
Human     DE---GNPE------------- 303
Mouse     EE---SDAE------------- 303
Frog      DPQACSKQEHFTHICTATDEVK 319
```

Exploiting alignment data II

Alex's contribution[3][4]: limit the analysis to simple ideas, find convenient representations that can be analyzed:

- consider pairwise alignments: two sequences x, y
- estimate counts matrices $\Pi_{x,y}$, corresponding Markov matrices $P_{x,y}$ and *observed rate matrices* $L_{x,y} = \log P_{x,y}$ when possible.
- observed rate matrices $L_{x,y}$ provide a nice representation of data, from which relevant questions can be answered; their trace $d_{x,y} = \mathrm{Tr}(L_{x,y})$ provide dissimilarity estimates.
- multivariate analysis: are observed rate matrices (close to) multiple of a generic rate matrix Q ? taxon dependent ?...
- Even simpler representations: diagonals of estimated rate matrices
- See C. Landes' talk

[3] Devauchelle et al, *J. Comp. Biol.* 8:4 (2001) 381-399

[4] Weyer-Menkhoff et al, *Computer Biology and Chemistry*, 29 (2005) 196-203

Summary

Besides his scientific contributions and publications, Alex brought a very original point of view to the field, based on

- the construction of appropriate data representations
- the will of finding the right level of modeling.

His scientific approach was multi/inter/trans-disciplinary... long before this became fashionable.

Many thanks to Alain Hénaut for his help in the preparation of this presentation.