

From **G**lobal **A**nalysis of **S**equence **T**rees to **O**rthology **N**etworks

Going **O**nward in the **N**ew **N**eoteric
Exploration of a **T**erritory

Amos Bairoch, July 4, 2014



Swiss Institute of
Bioinformatics




UNIVERSITÉ
DE GENÈVE

Neoteric?



neoteric

Line breaks: neo|ter|ic

Pronunciation: /,ni:ə(u)'terɪk  /

ADJECTIVE

FORMAL

- 1 Belonging to recent times; recent.

MORE EXAMPLE SENTENCES

- 1.1 New or modern:

'another effort by the White House to display its neoteric wizardry went awry'

MORE EXAMPLE SENTENCES

NOUN

[Back to top](#)

A modern person; a person who advocates new ideas.

As you all know from previous talks, Gaston had already a very prestigious career in computer sciences before he embarked into his foray in the jungle of bioinformatics

It is therefore not surprising to find his name in a list of *people who have shaped technology*:

<http://www.dayintechhistory.com/technology-birthdays/>

The vagaries of alphabetic ordering place him between the inventors of the telephone and of Twitter

Noah	Glass	Twitter		
Gaston	Gonnet	Open Text Corporation		
Alexander	Graham Bell	Invented Telephone	March 3, 1847	August 2, 1922

Is Twitter the New Telephone?

A Seattle resident recently tweeted the local police department to make a records request, just one event in a string of public devotion to the platform.

BY COLIN WOOD / MAY 9, 2014

0



[Christophe Dessimoz](#)

@cdessimoz

+ Follow

Symposium for Gaston Gonnet's retirement on 4 July at ETHZ (@ETH_en). Proceedings via @thePeerJ. Sponsored by @ISBSIB
gnome2014.lnk.ch

One day in late 1991....

From Des Higgins <Des.Higgins@EMBL.BITNET>☆

Subject **Gaston Gonnet**

To bairoch@CGECMU51.BITNET☆

Message ID <D7E522554020404F@EMBL-Heidelberg.DE>

Received from JNET-DAEMON by cmu.unige.ch; Fri, 18 Oct 91 13:35 WET-DST

Received From EMBL(HIGGINS) by CGECMU51 with Jnet id 6620 for BAIROCH@CGECMU51; Fri, 18 Oct 1991 13:35 +0100

Date Fri, 18 Oct 1991 13:34 +0100

Ciao Amos: we have just had a talk from a guy called Gaston Gonnet from the ETH in Zurich. He described an experiment where he did all possible optimal alignments between all sequences in a protein database. He did this very cleverly and used the output to do some extremely interesting things. His knowledge of the algorithms and maths involved is EXTREMELY high. He wants to do this again soon and we told him to do it on Swissprot and to contact you. As a by product of his work he has a datastructure describing all significant overlaps in the database which is a MINE of information if you know what to look for. One thing he did with it was to look at the distribution of gap sizes he has made GAPS tractable in the same way that PAM matrices make substitutions tractable.

Anyway, this note is to explain that he is worth hearing if he tries to contact you he KNOWS what he is doing very well.

Des

Which lead to an email from Gaston a week later

From gonnet@inf.ethz.ch★

Subject **SwissProt and All-against-all Matching**

To bairoch@CGECMU51.BITNET★, bairoch@cmu.unige.ch★

Message ID <9110311530.AA01129@rutishauser.inf.ethz.ch>

Received from JNET-DAEMON by cmu.unige.ch; Thu, 31 Oct 91 16:31 WET-DST

Received From CERN(MAILER) by CGECMU51 with Jnet id 7556 for BAIROCH@CGECMU51; Thu, 31 Oct 1991

Received from CERN by CERN.cern.ch (Mailer R2.07B) with BSMTP id 7555; Thu, 31 Oct 91 16:31:44 GVA

Dear Dr Bairoch, I do not think we met before, but I hope we certainly do so in the future. I am a professor in Informatik at ETH Zurich. Previously I was working at the University of Waterloo in Canada. In recent times, since my coming to Zurich, I have started doing computational biochemistry together with Prof. Stephen Benner from Organic Chemistry.

Our interests lie on very different areas, but to avoid repeating myself, I took the liberty of sending you a copy of a technical report which describes our activities. One of such is an all-against-all matching of an entire peptide database, which we can do thanks to the use of fast algorithms (originally developed for the processing of the Oxford English Dictionary). The first time that we did the all-against-all matching, we did it with the MIPS (version 64) database (most likely a very foolish decision).

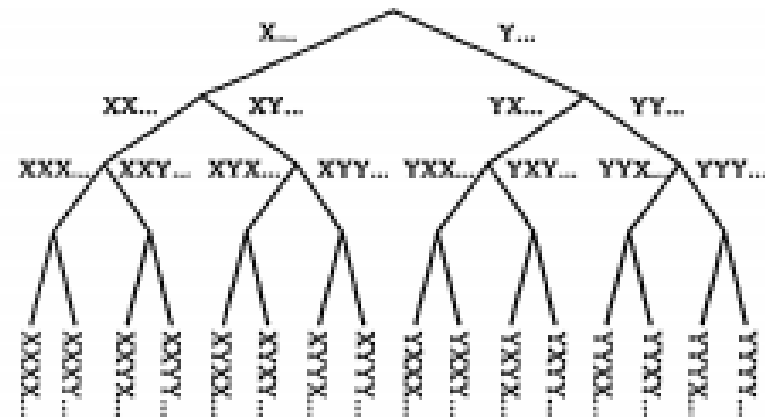
A week ago I gave a talk at EMBL, and they strongly suggested that I should use SwissProt the next time. They also encouraged me to contact you for various reasons, (1) so that you know what we are doing here

Exhaustive Matching of the Entire Protein Sequence Database

Gaston H. Gonnet, Mark A. Cohen, Steven A. Benner*

Fig. 1. Reorganization of sequences to form semi-infinite strings placed in "alphabetical order" (left to right in this diagram) on a patricia tree (13), idealized here for sequences built from just two letters. Reorganization time is almost linear with database size and requires negligible computation. Exhaustive matching is achieved by comparing patricia subtrees from the top. Time is saved because the matching of patricia subtrees is aborted when the score falls below a liberally chosen similarity limit.

Because all subsequences in the database are indexed, the fact that two similar protein sequences do not begin identically does not diminish the generality of the search.



longer improved. After refining, 1.7×10^6 matches remained, each optimally aligned, which were then used to calculate new mutation matrices and a model for scoring gaps. These new scoring parameters were then used to further refine the matches to self-consistency. The parameters provide definitive answers to the two fundamental questions concerning protein alignment: What does a mutation cost? and What does a gap cost?

**This paragraph stirred quite a controversy:
Gaston's arrival in the field of bioinformatics
was duly noticed!**

We poor people in Europe receive journals like SCIENCE with a certain delay. Having read a posting about the Gonnet et al. paper, I rushed onto the last issue to read it.

It gave me goose flesh. And I'm reassured to see that D. Davison thinks the same.

- in the summary: " Definitive mutation matrices". Later: "The parameters provide definitive answers ..." !! How can a scientist write that he got a **definitive** answer to anything ?

It is a piece of junk. (stronger word deleted). It contains **no** information that allows their work to be reproduced or checked in any way. An example of something, I'm not sure what, but it is NOT a good paper. It may be good science but you can't tell from what they wrote.

dan

We are sorry to have upset Dan so badly. Perhaps Dan will become more relaxed if he realizes: (a) that Science publishes papers that reports results of general interest to a wide range of scientists, not papers for specialists; (b) that the details of the methods needed to reproduce our results are available to Dan should he contact us, or should he read our full papers; (c) that the caption to Figure 2 already addresses for the careful reader Dan's most trenchant objections. Regardless of what the prior literature says, the data show that a linear gap penalty is not appropriate. We

A word in your protein

SIR — What is the longest word spelled out in the sequence of a protein in the protein sequence database¹ using the one-letter code for amino acids? None of the extensive literature devoted to this problem^{2–4} has taken a truly systematic approach; the longest word found to date contains only seven letters. We have matched the entire *Oxford Unabridged English Dictionary* (second edition, 20 volumes, 572,728,830 characters, with information content close to that of the human genome) against the entire SwissProt protein sequence database (version 23)⁵. Using the Patricia tree data structure¹, the matching consumed only 23 minutes of computational time. We found two words with nine characters: “hidalgism” (the manner or practice of a hidalgo) entered the English dictionary via a citation from 1887, and appears at positions 247–355 of the integrase of bacteriophage lambda (acquisition number P03700). “Ensilists” (the plural of ensilist, one who preserves

his crops by ensilage) entered the dictionary via a citation from 1883, and appears at positions 81–89 of the PRRB protein from *Escherichia coli* (acquisition number P17222).

In addition to being the longest strings appearing simultaneously in the English and protein languages, these are also candidates for the most unusable pieces of information simultaneously in lexicography and in biochemistry. Their discovery does, however, demonstrate the power of these data structures in handling large amounts of information.

Gaston H. Gonnet

Steven A. Benner

*Institute for Scientific Computation, and
Institute for Organic Chemistry,
ETH, Zurich, CH-8092 Switzerland*

1. Gonnet, G. H., Cohen, M. A. & Benner, S. A. *Science* **256**, 1443–1445 (1992).
2. Price, N. C. *Trends biochem. Sci.* **12**, 349 (1987).
3. Purton, M. *Trends biochem. Sci.* **13**, 48 (1988).
4. Jimenez, A. *Trends biochem. Sci.* **14**, 14 (1989).
5. Bairoch, A. & Boeckmann, B. *Nucleic Acids Res.* **20**, 2019–2022 (1992).

More protein talk

SIR — Gonnet and Benner¹ have searched for the longest English word in the protein sequence databank. But what of other languages? Given that the ownership of the longest peptide-word will undoubtedly become a source of intense national pride, I thought it wise to investigate.

I performed a similar analysis to Gonnet and Benner using a standard hashing algorithm to search the SwissProt databank² with a multilingual word list of 1.3 million words from Danish,

Dutch, English, Finnish, German, Italian, Norwegian, Swedish and some Estonian. I considered only if a word contained no accented letters or special characters.

Apart from English, four of the other languages provided nine-letter words: *ansvarlig* (a Danish word meaning 'liable') in entry HX_YEAST at position 85; *haletante* (French for 'breathless') in K1C0_XENLA at 145; *saltsilda* (Norwegian for 'salted herring') in PA11_BOVIN at 271, and *stillassi* (the perfect subjunctive of the Italian word *stillare* — 'to drip') in STE2_YEAST at 207. The most apt nine-letter word was *salasivat*, PEHX_ERWCH at 10, the past tense of *salata*, Finnish for 'to keep hidden' or 'to encode'.

Although I did not find any other English words of nine letters or more in SwissProt, the search did turn up a 10-letter Italian word, *annidavate*, at position 45 in databank entry PHEA_FREDI (C-phycoerythrin α -chain from *Fremyella diplosiphon*). The word is from the past imperfect tense of the word *annidare*, meaning 'to nest'.

An honourable mention must also go to the 10-letter American-English word *Wallawalla* (the language of the Shahap-tian people of southeast Washington or Oregon), and the Dutch word *tariefklas* (literally 'tariff class') which are found reversed in the databank at position 18 of BVGB_BORPE and position 938 of ITA4_HUMAN, respectively.

The race is now on for the next longest word. How long will we have to wait before Germany finally scoops the honours with the possible 27-letter peptide-word for 'social sciences':

Gesellschaftswissenschaften

David Jones

*Biomolecular Structure and Modelling
Unit,
Department of Biochemistry and
Molecular Biology,
University College,
London WC1E 6BT, UK*

1. Gonnet, G. H. & Benner, S. A. *Nature* **361**, 121 (1993).
2. Bairoch, A. & Boeckmann, B. *Nucleic Acids Res.* **19**, 2247–2249 (1991).

So, is
Gesellschaftswissenschaften
now present in one of the
70 million UniProt entries?

SSENACHVFAEL
SSENSCHAFTEN
***** : ** . * : *

Submitted name:

CBN-TAG-163 protein (EMBL EGT39212.1)

Name: **Cbn-tag-163** (EMBL EGT39212.1)

ORF Names: CAEBREN_23701 (EMBL EGT39212.1)

Caenorhabditis brenneri (Nematode worm) [Complete proteome]

GOMNL7
GONNET

It is maybe not so well known that Gaston not only worked on sequence analysis and orthology but also applied his skills to proteomics.



Biochemical and Biophysical Research
Communications

Volume 195, Issue 1, 31 August 1993, Pages 58–64

Regular Article

Protein Identification by Mass Profile Fingerprinting

P. James, M. Quadroni, E. Carafoli, G. Gonnet

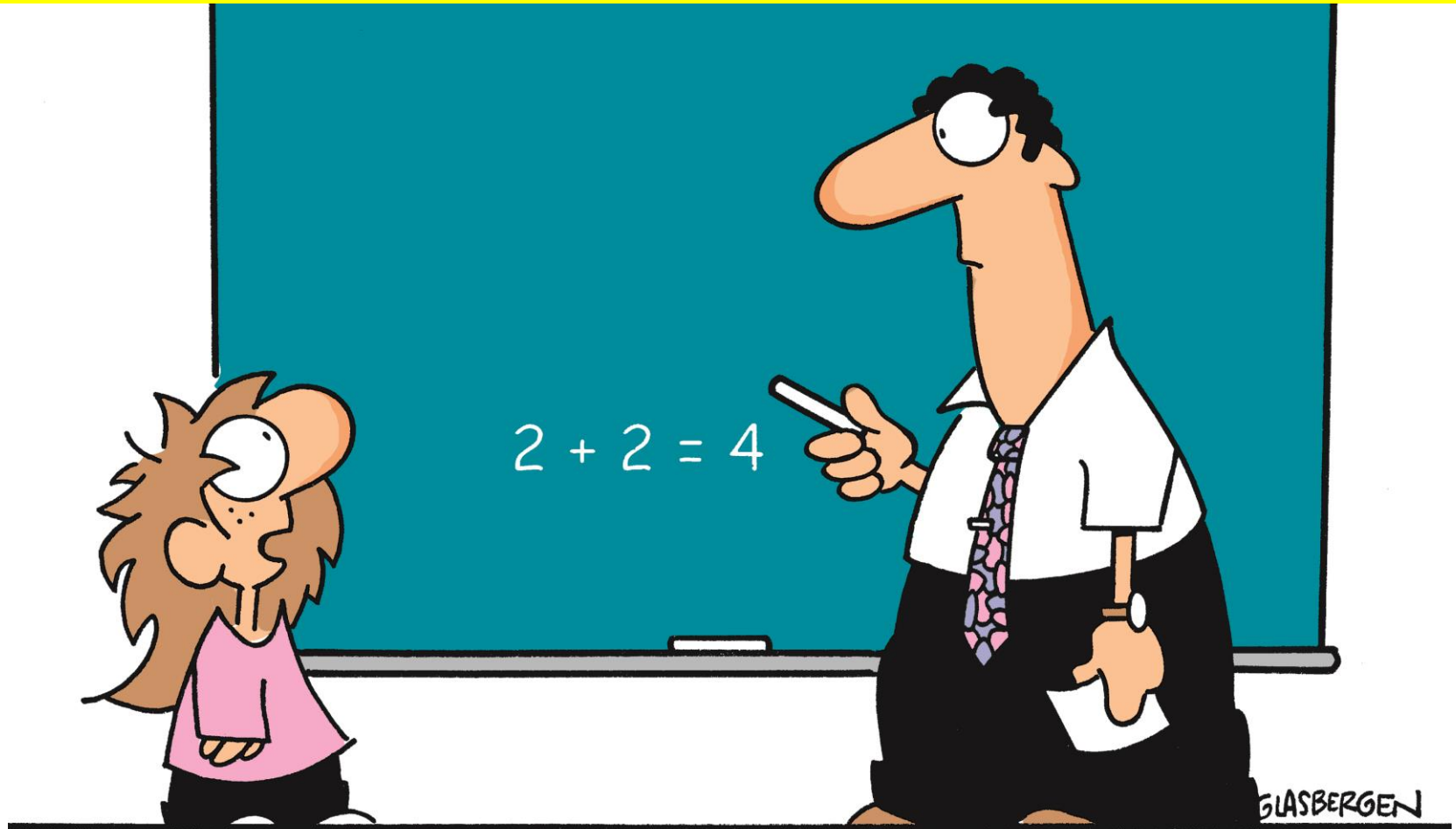


FROM GENOME TO PROTEOME

2nd Siena 2D electrophoresis meeting
Siena, Italy, Sept. 16 to 18, 1996

**He was therefore
invited to give a talk at
the 2nd Siena meeting in
1996**

He started by stating that biologists were using slides (yes real ones), bioinformaticians used transparencies and he, as a mathematician was going to use the blackboard and chalks



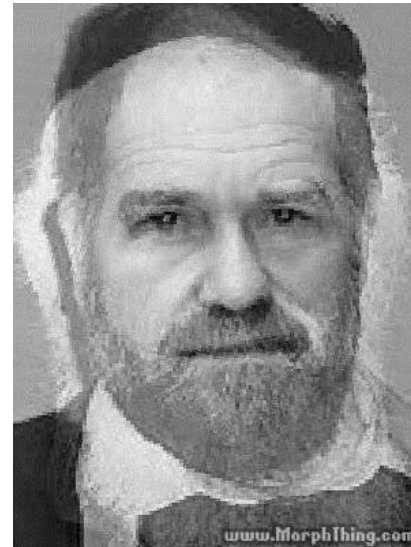
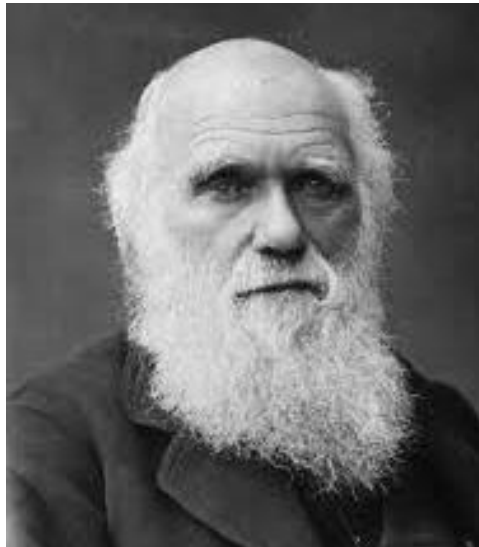
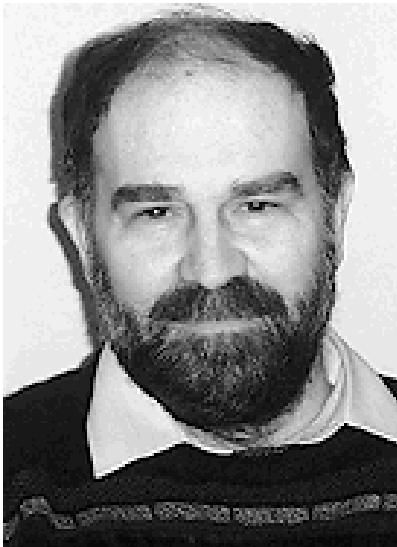
“How can I trust your information when you’re using such outdated technology?”

Darwin

In Gaston own words: Darwin is a programming language especially tailored to the wants and needs of the bioinformatics researcher.

I would summarize it as: Maple for bioinformaticians

But is there hidden agenda in the use of Darwin name for Gaston's sequence analysis language?



In early 1998, the SIB was being created

Subject **Re: Relational Databases**

To Gaston Gonnet <gonnet@inf.ethz.ch>★

Cc Amos.Bairoch@medecine.unige.ch★

Message ID <01IUA7RMJHJW0005OK@cmu.unige.ch>

In reply to ▶ [1](#), [2](#), [3](#), [4](#), [5](#), [6](#), [7](#), [8](#), [9](#), <199802261321.OAA14583@vinci.inf.ethz.ch>

Received from cmu.unige.ch by cmu.unige.ch (PMDF V5.1-10 #10385) id <01IU9T004RI80005OK@cmu.unige.ch>
BAIROCH@cmu.unige.ch; Wed, 4 Mar 1998 21:46:15 MET

Yep we need to talk about a lot of things.

Can you send me a list of possible dates for you and i will get back to you with a proposition.

sorry about tel. style., but we are in the middle of many important changes (creation of the swiss Institute of bioinformatics, of a company (GeneBio), etc. we need to discuss all this as i think you will be interested by these developments.

Best regards

Amos



Swiss Institute of
Bioinformatics

Gaston's SIB career

- President of the SIB scientific advisory board (SAB) from 1999 to 2003;
- SIB foundation council member from 1999 to 2008;
- SIB group leader from 2006 onward;
- And first group leader to retire 😊

A meeting of the SIB SAB in 2002



A discussion during a meeting of the SIB foundation council in 2008

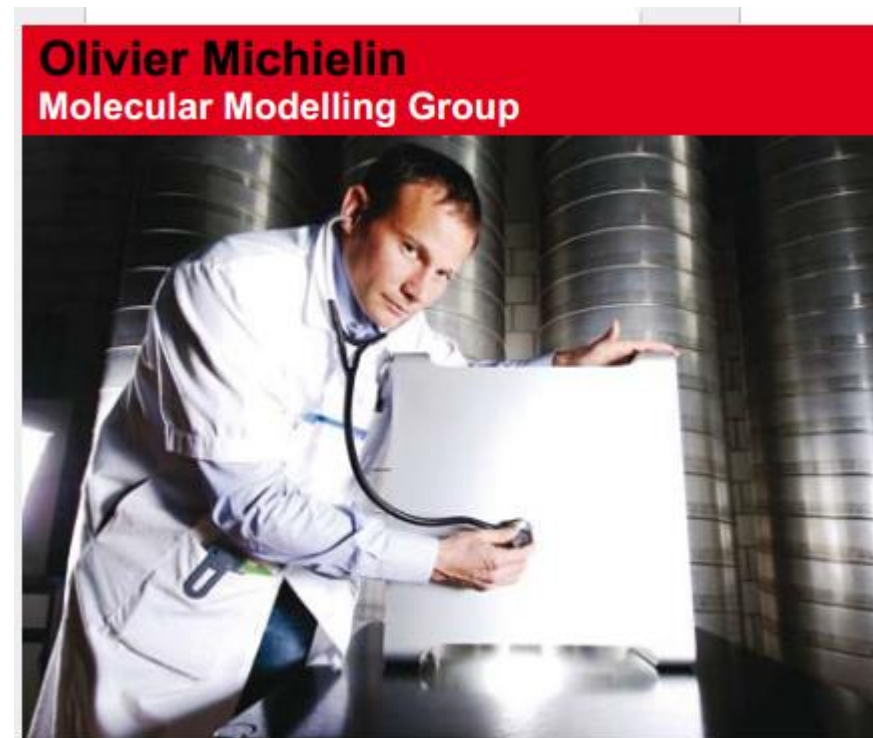


«Personne haute en couleur, il a, surtout dans les premières années, fait en sorte que nous ayons des débats animés. C'est un collègue sympathique et engagé que nous voyons partir à regret..» Ron Appel 3/7/2014

A few years ago the SIB commissioned Nicolas Righetti, a famous Swiss photographer to take original pictures of group leaders



Each GL had to answer a questionnaire that Nicolas drafted and which had to do with the GLs hobbies, passions, etc.



This is the official pictures of Gaston in the SIB GL portrait series



But in fact
Gaston had
jokingly
indicated in his
questionnaire's
answers, his
admiration for
Hannibal Lecter,
the famous
serial killer in
the Silence of
the Lambs





Retirement? If yes, where should Gaston



Rue Gaston Gonnet 84170 Monteux, France

Prix de l'immobilier

[www.lacoteimmo.com/...](http://www.lacoteimmo.com/)

D'une longueur de 9

située à Monteux, da

Alpes-Côte d'Azur).



South of France seems a reasonable idea: sun, good food and Gaston is already well known there!

Yesterday I was in Lyon for a bioinformatics HDR thesis: on the way from the IBCP to the University my way forward was interrupted by a moving van....



GONNET

DEMENAGE BIEN



04 78 00 86 43

PARTICULIERS ET TRANSFERTS DE BUREAUX

Some almost 'Final' words

- For me it has been a privilege and a pleasure to interact with Gaston for the last 23 years!
- I am also thankful for him having pushed forward the career of many young bioinformaticians that matured in his group
- And specifically Christophe Dessimoz who collaborated closely with members of my former group

A very nice memory

In 2006, Gaston sent Christophe to Fortaleza in Brazil to attend the conference for the 20th anniversary of Swiss-Prot



Gaston asked him to bring me a gift that I cherish: an original printing plate from the Oxford English Dictionary



**Gaston, I wish
you a smooth
sailing in your
next endeavor**