Algorithmica and molecular biology
The Pisan experience

Fabrizio Luccio

Glimpses into the world born from the encounter between the machines for sequencing DNA fragments, and computers that assembly those fragments.

The Department group

Maria Federico	(federico@cli.di.unipi.it)
Claudio Felicioli	(pangon@gmail.com) *
Paolo Ferragina *	(ferragin@di.unipi.it)
Roberto Grossi	(grossi@di.unipi.it)
Fabrizio Luccio *	(luccio@di.unipi.it)
Roberto Marangoni *	(marangon@di.unipi.it)
Nadia Pisanti *	(pisanti@di.unipi.it)

* The boss (at least, the who knows everything)

* Reference person (she made most of the work)

* Trying to escape

* gmail: why? (probably paid by Google)

* ME! (parasite, but early group initiator)

Glimpses into the world etc ….

Algorithms are the winning tool.

Sorry…. good algorithms are the winning tool, especially when dealing with very large data.

Inefficient algorithms….

…. have the unpleasant property of resisting to hardware improvement:

A polynomial-time algorithm solves a problem on n data in time $t_1 = c\,n^s$

An exponential-time algorithm solves a problem on n data in time $t_2 = c\,s^n$

with c, s constants

With a computer k times faster, and same running time, we process $N > n$ data, according to the laws:

$t_1 = c\,n^s, \;\; k\,t_1 = c\,N^s \;\Rightarrow\; N = k^{1/s}\,n$

$t_2 = c\,s^n, \;\; k\,t_2 = c\,s^N \;\Rightarrow\; k\,s^n = s^N \;\Rightarrow\; N = n + \log_s k$

Publications on sequence algorithms

Mercatanti A., Rainaldi G., Mariani L., Marangoni R., Citti L. *A method for prediction of accessible sites on an mRNA sequence for target selection of hammehead ribozymes*. J. Computational Biology, 4(9) 641-653, 2002

Menconi G., Marangoni R. *A compression-based approach for coding sequences identification in prokaryotic genomes*, J. Computational Biology (to appear)

Corsi C., Ferragina P., Marangoni R. *The bioPrompt-box: an ontology-based clustering tool for searching in biological databases*. BMC bioinformatics (to appear)

Cozza A., Morandin F., Galfrè S.G., Mariotti V., Marangoni R., Pellegrini S. *TAMGeS: a Three-Array Method for Genotyping of SNPs by a dual-color approach.* BMC genomics (to appear)

Felicioli C., Marangoni R. *BpMatch: an efficient algorithm for segmenting sequences, calculating genomic distance and counting repeats*, (submitted)

Ferragina P. *String search in external memory: algorithms and data structures*. Handbook of Computational Molecular Biology, CRC Press, 2005

Publications on motifs

N. Pisanti, M. Crochemore, R. Grossi, M.-F. Sagot. *A Comparative Study of Bases for Motif Inference*. *NATO Series on String Algorithmics*, 2004.

N. Pisanti, M. Crochemore, R. Grossi, M.-F. Sagot. *Bases of Motifs for Generating Repeated Patterns with Wild Cards*. *IEEE/ACM Transactions on Computational Biology and Bioinformatics* 2(1) 40-50, 2005.

C.S.Iliopoulos, J.McHugh, P.Peterlongo, N.Pisanti, W.Rytter, M.Sagot. *A first approach to finding common motifs with gaps*, *International Journal of Foundations of Computer Science* 16(6) 1145--1155, 2005.

N. Pisanti, H. Soldano, M. Carpentier, J. Pothier. *Implicit and Explicit Representation of Approximated Motifs*. In: *Algorithms for Bioinformatics*, C. Iliopoulos et al, editors, King's College London Press, 2006.

P.Peterlongo, N.Pisanti, F.Boyer, A.Pereira do Lago, M.-F.Sagot. *Lossless filter for multiple repetitions with Hamming Distance*. *Journal of Discrete Algorithms* 2007 (to appear).

Major collaborations on motifs

Lyon (group of Marie-France Sagot)

Grenoble (group of Alan Vieri)

Paris (group of Henri Soldano)

Marne-la-Valle (group of Maxime Crochemore)

Paralogy tree construction

……. via *transformation distance*

Pisanti N., Marangoni R., Ferragina P., Frangioni A., Savona A., Pisanelli C., Luccio F. *PaTre: A Method for Paralogy Trees Construction*. J. Computational Biology, 5 (10) 791-802, 2003

How does the genomic information increase?

external imports - Transfections

- Horizontal transfer

Endogenous mechanisms

(genic or genomic) duplications: Large scale

Tandem

Dispersed

Single gene

The fate of the copy

Non-functional: pseudogene

Functional: paralog

genome as a set of families of paralogs

PARALOGY TREE

- How does the genome choose the paralog to duplicate within a family?
- Is the duplication rate constant among the various families?
- Are sparse duplications correlated to sparse deletions?

Couple-comparison method
Transformation Distance (TD)

Often newest genes are the shortest ones

To insert sequences imply paying metabolic costs. To delete sequences has no metabolic cost

We need an asymmetric distance:

TD(S,T) = the cost of the minimum-length script able to

transform S into T

Elementary operations : Insertion, Copy, Inverted copy

TD: an example

f g h

S=ATCGATCAGCTGCCCAATGAATCAGATAAAGTTTC

1ÉÉÉÉÉ.ÉÉ11ÉÉ.....16ÉÉÉÉÉÉÉ..25ÉÉÉÉÉÉÉÉ35

f g h

T=ATCGATCAGCTTTCACTACGAATGAATCAGATTGGTAGCTTTGAAATAG

1ÉÉÉÉÉÉÉÉ..11ÉÉÉÉÉÉÉ...21ÉÉÉÉÉÉÉÉÉÉÉÉ.ÉÉÉ38ÉÉÉÉÉÉÉÉ48

Script transforming S into T	**Description**
1) copy f	copy (1, 1, 11)
2) insertion of TTCACTACG	insert (TTCACTACG)
3) copy g	copy (16,21,12)
4) insertion of TGGTAGC	insert (TGGTAGC)
5) inverted copy of h	copy (25,38,11,1)

PaTre

Input: TD values for each possible couple made by the genes of the family

Building-up of the directed graph of distances

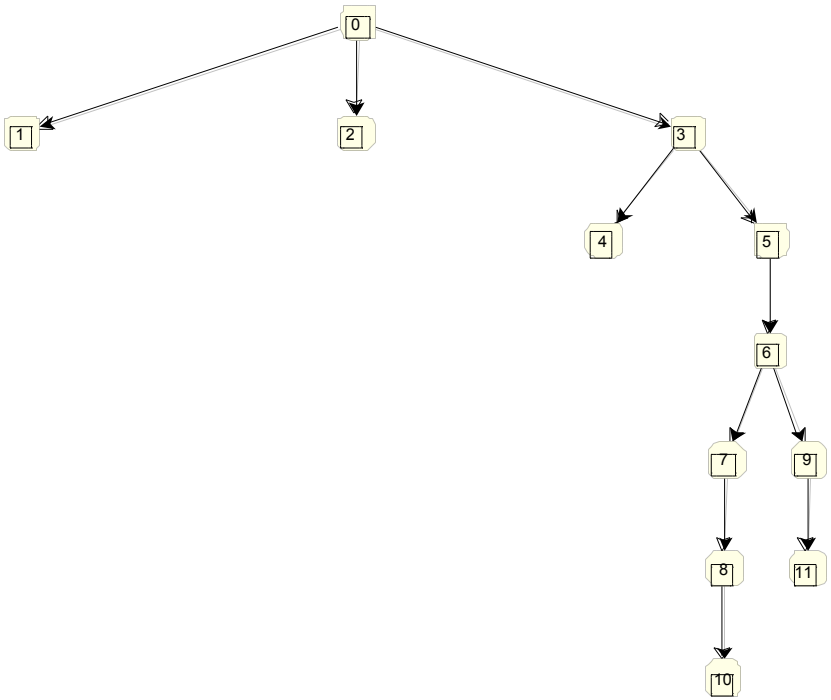
Edmonds' algorithm: extraction of the LSA (Lightest Spanning Arborescence) → optimal paralogy tree

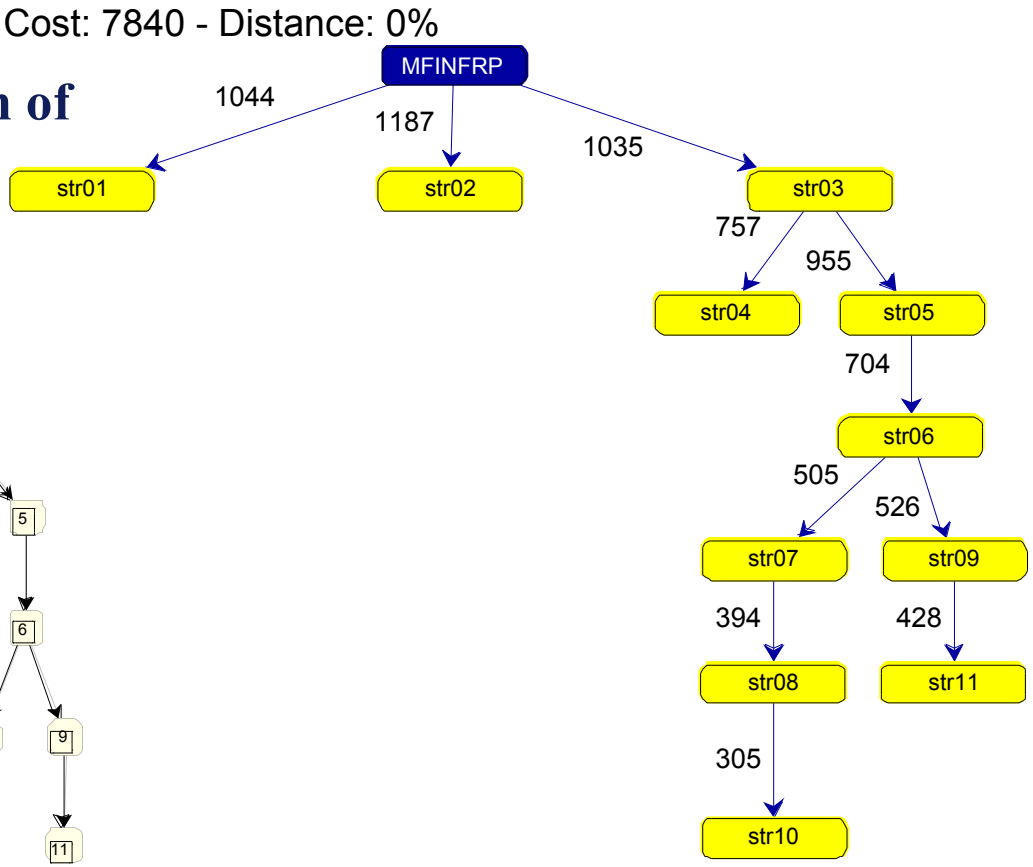
Generation of optimal and sub-optimals (space of quasi-optimal solutions)

PaTre has been tested by simulation

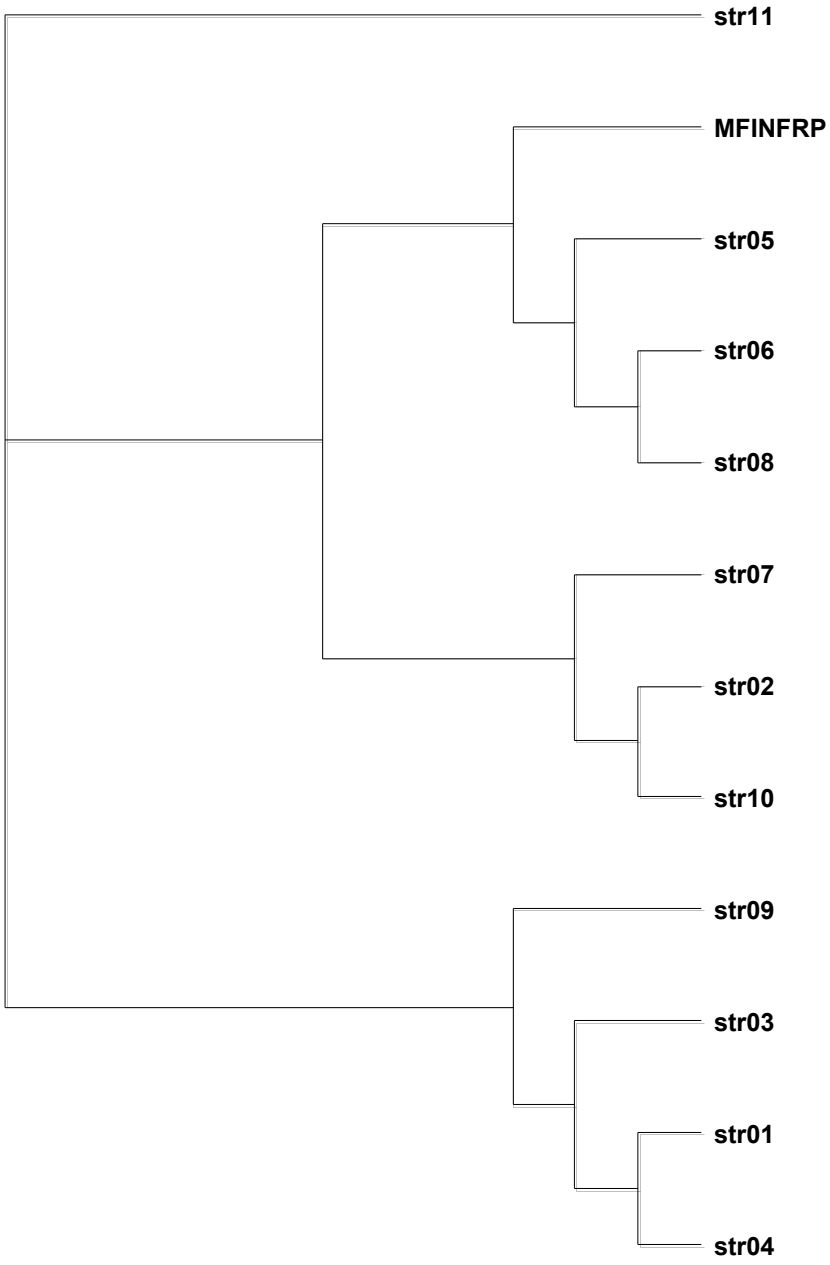
…because there are no experimental data on the history of families of genes

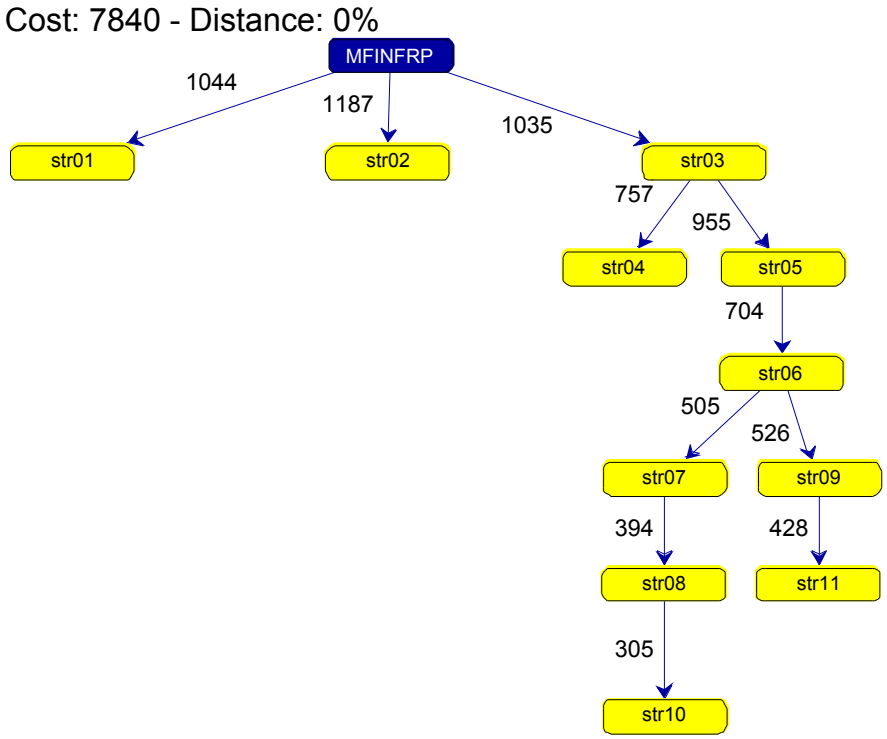
output from PaTre for the simulated Ribosomal Protein of *M. pneumoniae*





The simulated paralogy tree for the Ribosomal Proteins family of *M. pneumoniae*





The paralogy tree reconstructed by ClustalW for the Ribosomal proteins genic family of *M. pneumoniae*

After having tested PaTre on many examples,

we could conclude that PaTre is able to correctly reconstruct the simulated history of genetic families, while ClustalW and other similarity based methods fail.