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:

\[
\begin{align*}
t_1 &= c n^s, \quad k t_1 = c N^s & \Rightarrow & & N = k^{1/s} n \\
t_2 &= c s^n, \quad k t_2 = c s^N & \Rightarrow & & k s^n = s^N & \Rightarrow & & N = n + \log_s k
\end{align*}
\]
Publications on sequence algorithms


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


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


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

topline: genome as a set of families of paralogs

- 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:

$$TD(S,T) = \text{the cost of the minimum-length script able to transform } S \text{ into } T$$

Elementary operations : Insertion, Copy, Inverted copy
**TD: an example**

\[
S = \text{ATCGATCAGCTGCCCAATGAATCAGATAAAGTTTTC} \\
1\ldots11\ldots16\ldots25\ldots35
\]

\[
T = \text{ATCGATCAGCTTTCACTACGAATCAGATTGGTAGCTTTGAAATAG} \\
1\ldots11\ldots21\ldots38\ldots48
\]

**Script transforming S into T**

1) copy f  
2) insertion of TTCACTACG  
3) copy g  
4) insertion of TGGTAGC  
5) inverted copy of h

<table>
<thead>
<tr>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>copy (1, 1, 11)</td>
</tr>
<tr>
<td>insert (TTCACTACG)</td>
</tr>
<tr>
<td>copy (16, 21, 12)</td>
</tr>
<tr>
<td>insert (TGGTAGC)</td>
</tr>
<tr>
<td>copy (25, 38, 11, 1)</td>
</tr>
</tbody>
</table>
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.