Do It Yourself Comparative Genomics

Almeida, João Manuel Feio de

Centro de Recursos Microbiológicos (CREM), Departamento de Ciências da Vida,
Faculdade de Ciências e Tecnologia, Universidade Nova de Lisboa, 2829-516, Caparica, Portugal

Available from: http://sourceforge.net/projects/bidiblast/ or http://moodle.fct.unl.pt/course/view.php?id=2079

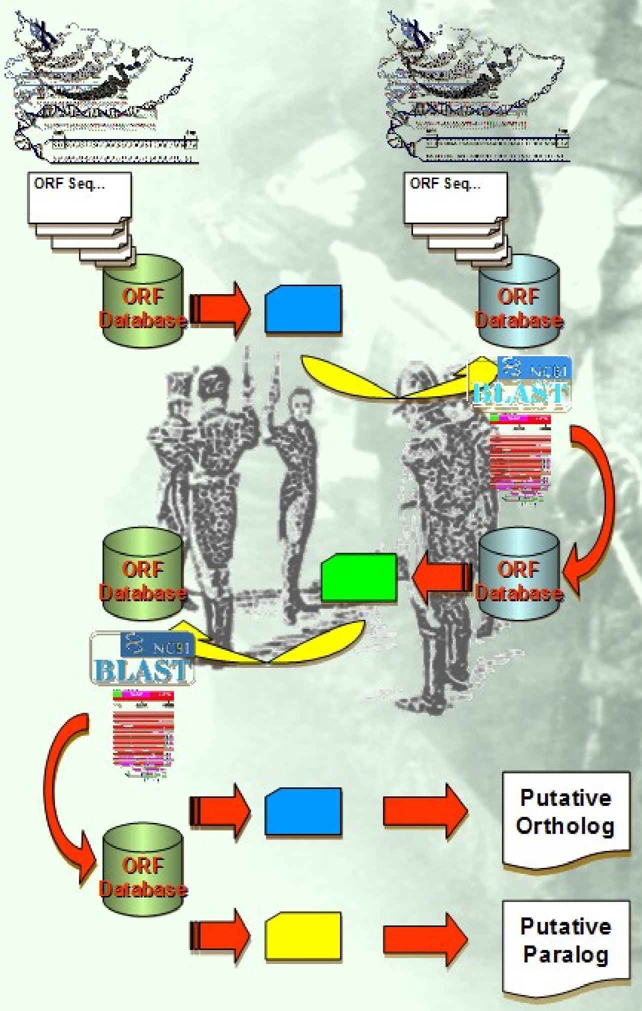
Driving Motivation

The amount of genome sequences available in public databases increases steadily. But the full exploitation of this bonanza of data is hampered by the limitations in sequence annotation. These limitations result from an imbalance between the rate of accumulation of new sequences, and the throughput of wet-bench researchers. The gap is usually filled by in silico analysis, mostly done through data pipeline software (e.g. EMBL Bank to TrEMBL). The results are more often than not stored in secondary databases after a most scant quality control assessment due to limitations in staff. This state of affairs result in the need to enforce a most strict set of parameters during the in silico analysis in order to avoid or limit the emergence of artifacts (e.g. annotation transfer from analogs).

The ability to repeat those analysis according to one's own parameter values is of paramount importance to independently check annotation made available by the genome sequencing centers.

Here this capability is awarded even to less performant personal computers with a minimum burden to the user.

Bi-Directional Procedure



ORF Seq...

ORF Database

BLAST

Putative Ortholog

Putative Paralog

Implementation Details

Collection of 15 JAVA classes – 3 Packages

General routines - **bidiblastsup**
Data structures – **bidiblastsup.objects**
User interface – **bidiblastsup.ui**

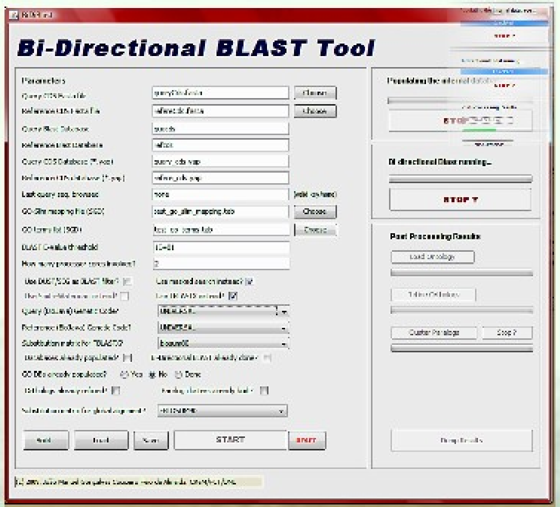
Uses 3 third-party libraries

BioJava 1.4 – mainly trasnlation tasks
DB4O 5.0 – data management and ...
NeoBio – scoring schemes including ambiguity codes

Integrates 4 command line tools

NCBI BLAST (blastall –p blastn)
align0 (FASTA) – ORF alignment
stretcher (EMBOSS) – protein alignment
yn00 (PAML) – dN/dS calculation

Graphic User Interface



Bi-Directional BLAST Tool

Parameters

Last query seq. browsed: none

START

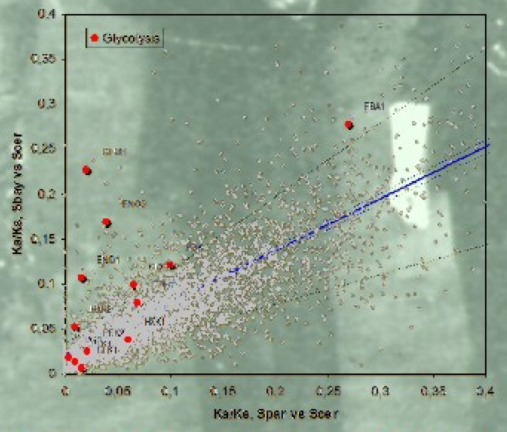
STOP ?

Bi directional Blast running...

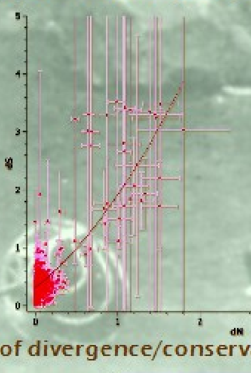
Post Processing Results

Sample Results

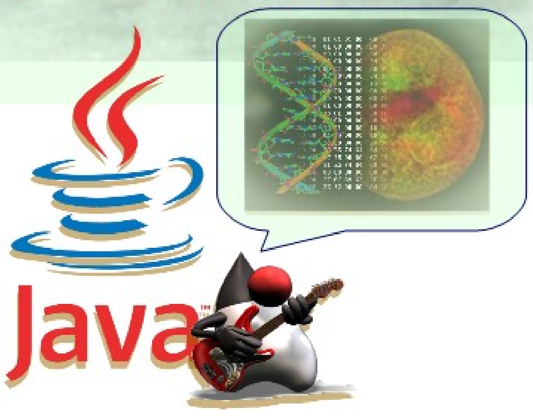
(Plots were made with other applications)



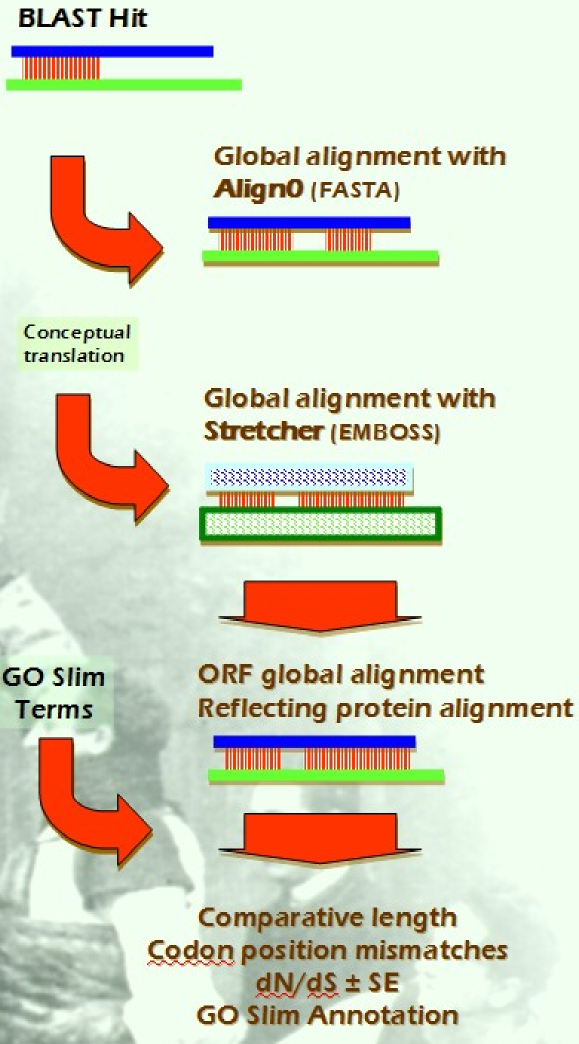
Genomewide comparison of evolution rates highlighting specific ontological groups



Distribution of divergence/conservation rates



Refining Hits Procedure



BLAST Hit

Global alignment with **Align0** (FASTA)

Conceptual translation

Global alignment with **Stretcher** (EMBOSS)

ORF global alignment
Reflecting protein alignment

GO Slim Terms

Comparative length
Codon position mismatches
dN/dS ± SE
GO Slim Annotation

Usage scope

- Costumized comparative genomics
- Annotation of ORF from newly sequenced genomes
- Estimation of evolution rates for sets sequence
- *etc*

References

Altschul S. F., et al. 1990. J. Mol. Biol 215:403-410.

Camon E., et al. 2003. Comp Funct Genomics. 4:71–74.

Myers E. W., and W. Miller. 1988. Comput. Appl. Biosci. 4:11-17.

Rice P., et al. 2000. Trends in Genetics 16:276-277.

Rivera M. C., et al. 1998. PNAS 95:6239-6244.

SGD project 2009. ftp://ftp.yeastgenome.org/yeast/.

Yang Z. 2007. Mol Biol Evol 24:1586-1591.

Technical Requirements

Hardware

Processor - Pentium 4 or newer
RAM – 1 GB (preferably 3 GB)
Hard Disk Space – (depends on data)

Software

Operating System – Windows 32 byte version (XP or Vista)
(Windows 7 was not tested)
Java Run Time Environment (SUN) – 1.4 through 1.6
Relational Database System – advisable but not required

End User License Agreement

Freeware according to GNU-GPL (see www.gnu.org for details)