Reciprocal BLAST for *x* Genomes: A Tool to Identify Best Reciprocal BLAST Hits from Multiple Microbial Genomes

Fahed Rafati[1,2,3], Mary Makarious[1,2,4], Ariane Quenum[1], and Anthony Volchek[1,2]

[1]Department of Bioinformatics; Loyola University of Chicago; Chicago, IL
[2]Department of Biology; Loyola University of Chicago; Chicago, IL
[3]Statistics Program; Department of Mathematics; Loyola University of Chicago; Chicago, IL
[4]Behavioral/Cognitive Neuroscience Program; Department of Psychology; Loyola University of Chicago; Chicago, IL

Abstract

Reciprocal BLAST is a useful, widespread method that identifies orthologs between different genomes. This tool was designed to meet the needs of those in the research community looking to compare 10s to 100s of bacterial genomes via their FASTA files and identify candidate orthologs. The purpose of this work is to design a tool to identify orthologs between a multitude of microbial genomes, in addition to identifying global similarities and differences that exist among bacterial genomes. This tool can be used to assess a gene or protein evolution and comparative genomics. Given *x* genomes, this program will generate the best reciprocal blast hit between pairs of bacterial genomes.

Availability: ReciprocalBLAST is accessible at https://github.com/m-makarious/ReciprocalBLAST

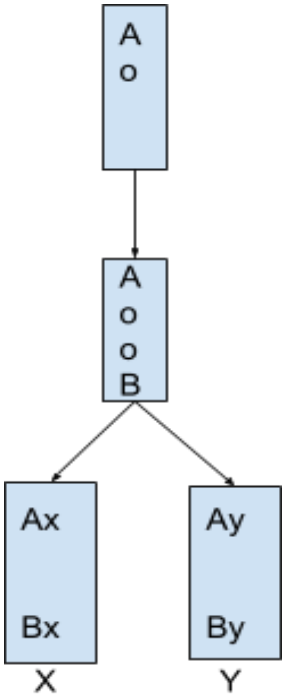
Contact: mmakarious@luc.edu or falrafati@luc.edu

Introduction

Sequence analysis, especially comparative genomics, often starts with a sequence similarity search using standard tools such as Basic Local Alignment Search Tool (BLAST) (Altschul et al., 1990). BLAST-based similarity searches are commonly used in several applications involving both nucleotide and protein sequences. Although Blast is more efficient in its method of comparison when compared to direct methods, the introduction of longer queries can be problematic for the software in its current state, potentially leading to slowdowns and mismatches. (Camacho et al., 2009)

Reciprocal BLAST is an extension of BLAST aimed at finding orthologous sequences between two species and also a common computational method used to predict orthologues. Reciprocal BLAST is done by taking a gene and BLAST-ing it to a database of the gene sequences from the organism of interest. The gene with a higher score is BLASTed to a database of the gene sequences. If the BLAST returns the original gene used as the highest scorer, in that case the genes are considered candidate orthologs; however, only experimental evidence can prove orthology.

Homologous genes diverge after a speciation event. Orthologs are genes in different species that evolved from a common ancestral gene and tend to retain the same function throughout evolution. Paralogs, defined as homologous genes are genes related by duplication within a genome. Since they have been proposed as a source of functional innovation, they are therefore less expected to have similar functions.



Ortholog vs Paralog. Two duplicated copies (A and B) of a gene from a single species diverge as evolution occurs over time. These genes are related and in this case are paralogs. In the case of orthologs, species that come from a common ancestor contain descendants of the two duplicated genes. When comparing two related species (X and Y) Ax and Ay, and Bx and By are said to by orthologous.

Implementation

The FASTA files were obtained from the NCBI Gene Database (https://www.ncbi.nlm.nih.gov/gene).

The program functions via performing reciprocal BLAST on user inputted FASTA files from bacterial origin. Initially, prompts direct the user to choose whether protein or nucleotide files will be used and to input the directory of the FASTA populated folder as well as an output folder for a comma delimited file (CSV) containing the best orthologous groups for said input. Based on user input, either protein or nucleotide Local BLAST Databases are created from the FASTA files and a reciprocal BLAST is performed via BLASTing each FASTA file against each database file. This results in comma delimited output files, for example:

***Salmonella_enterica_serovarTyphiTy2_vs_EcoliK12_db*.csv**

These output files are then parsed according to a user defined e-value in order to ascertain the "Best Hit" and populate a new CSV file with it.

The nature of the Reciprocal Blast Program as well as the parser means that the programs requirement for memory allocation and run time is linear (O(n)) and grows with increased user input, for example, reciprocal BLAST and CSV parsing on 10 FASTAs takes 13.6 seconds on 40 CPUs at 3000 MHz each. In terms memory allocation, 20 nucleotide FASTA files generate 160 Mb of output.

Schematic of RecipBlast. A folder of FASTA files is inputted, this will allow the generation of corresponding databases and subsequent BLASTing against the databases. Text files with results will be produced and the user has the option to parse the results to a .csv

Results

Now that the basis of reciprocal blast has been established, we can delve into the output. Following this, the output is generated, which is a folder containing all of the FASTA files blasted against each respective database. Each generated file reveals the length, query score, E-value, percent identity, and bit score. The number of output files generated is dependent on the number of FASTA files inputted into the pipeline.

In the case of a sample run with 3 FASTAs, reciprocal BLAST output was created and placed in the target folder. The results were then parsed into a .csv for future use in other programs.

Bit Scores			
	Lactoccocus_lactis_subsp_IL1403	Lactoccocus_lactis_KF147	EcoliK12
Lactoccocus_lactis_subsp_IL1403	0	205600	1101
Lactoccocus_lactis_KF147	205600	0	1092
EcoliK12	1101	1092	0

Sample Run Results. .CSV output of 3 FASTA run with bitscores from reciprocal BLAST results present.

In sum, we provide a means to perform a reciprocal BLAST on *x* genomes. This tool is built to identify orthologs between a multitude of microbial genomes. In addition to this, global similarities and differences that exist among microbial genomes are also identified and .csv file is made for both organization and further use in other programs.

References

1. Altschul, Stephen F., et al. "Basic local alignment search tool." Journal of molecular biology 215.3 (1990): 403-410.
2. Camacho, Christiam, et al. "BLAST+: architecture and applications." BMC bioinformatics 10.1 (2009): 421.
3. Ward, Natalie, and Gabriel Moreno-Hagelsieb. "Quickly finding orthologs as reciprocal best hits with BLAST, LAST, and UBLAST: how much do we miss?." PloS one 9.7 (2014): e101850.
4. Dalquen, Daniel A., and Christophe Dessimoz. "Bidirectional best hits miss many orthologs in duplication-rich clades such as plants and animals." Genome biology and evolution 5.10 (2013): 1800-1806.