

High-Fidelity Prediction of Orthologous Proteins in Bacterial Genomes

E. E. Snyder, E. K. Nordberg, O. Crasta and B. W. Sobral

Virginia Bioinformatics Institute

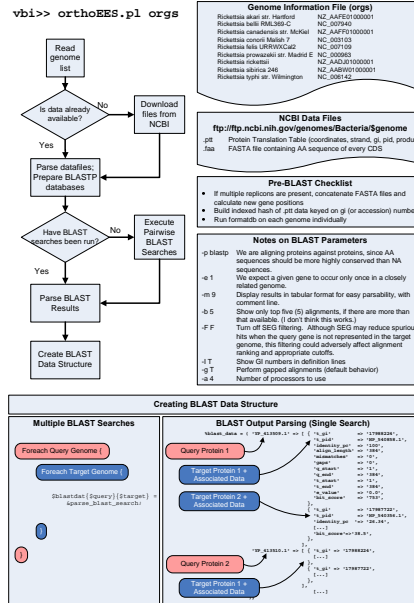
Virginia Polytechnic Institute and State University, Blacksburg, VA 24061, USA, eesnyder@vbi.vt.edu

Introduction

The PathoSystems Resource Integration Center (PATRIC) is one of eight Bioinformatics Resource Centers (BRCs) funded by the National Institute of Allergy and Infection Diseases (NIAID) to create a data and analysis resource for select NIAID priority pathogens, specifically proteobacteria of the genera *Brucella*, *Rickettsia* and *Coxiella*, and corona-, calici- and lyssaviruses and viruses associated with hepatitis A and E. The goal of the project is to provide a comprehensive bioinformatics resource for these pathogens, including consistently annotated genome, proteome and metabolic pathway data to facilitate research into countermeasures, including drugs, vaccines and diagnostics.

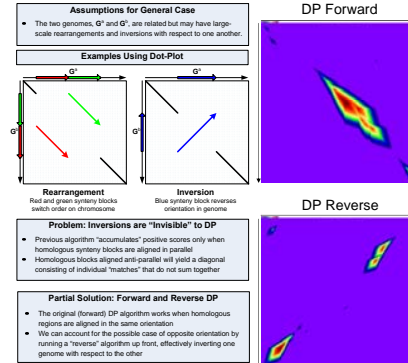
To better understand phenotypic differences between the closely related genomes in each category, it is essential to understand differential gene content and polymorphisms between orthologs, particularly when comparing virulent and avirulent strains. Naturally, this requires the construction of high-quality ortholog groups. To accomplish this, we adopted a two-stage strategy. First, similarity data was collected for each organism category by executing reciprocal BLASTP analyses between all pairwise genome combinations. Data of this type is the starting point for many comparative genomics applications. The second step involves a novel method for processing the individual pairwise BLAST searches based on dynamic programming (DP). The method resembles the classic Smith-Waterman algorithm for sequence alignment; however, instead of comparing two strings of amino acids using a substitution matrix, two strings of proteins are compared using BLAST scores. This method has the advantage of combining both similarity and conservation of synteny within a mathematically rigorous optimization framework. While local gene order is usually conserved within genomes of the same genus, there are typically large-scale segmental rearrangements and inversions which would confound the Smith-Waterman-based implementation described above. This report describes a new DP algorithm developed to accommodate these biological considerations.

BLAST Analysis



DP Allowing Inversions

General Case

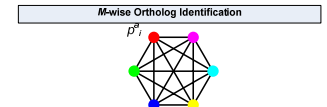


Applications

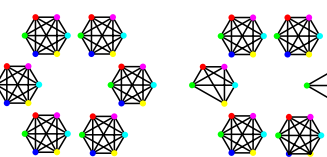
Pairwise and Multi-Genome Ortholog Prediction

Pairwise Ortholog Identification

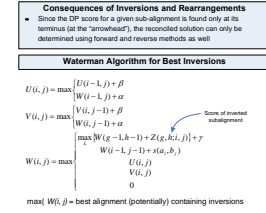
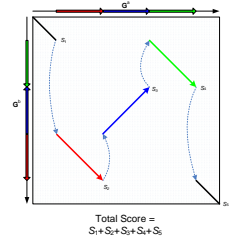
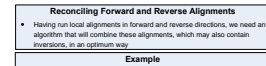
- Having established that an optimal alignment between genomes with realistic structure can be found, how does this impact our definition of orthology?
 - "Bidirectional Best-Hit" criterion is no longer required
 - The pairing between p_1 and p_2 is defined by DP



- Orthologs from six genomes fully connected to one another
- Each connecting line represents a protein pair from an optimal alignment of corresponding genomes



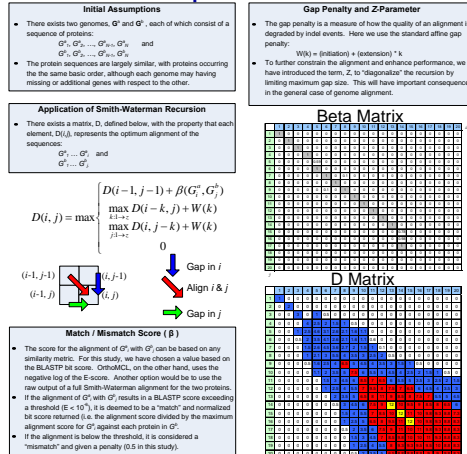
Assembling Final Solution from Forward and Reverse Runs



Bacterial Genomes of Interest

Organism Category / Organism	Taxonomic Rank / Accno	Genome Size, kb	CDS Count	Associated Diseases
<i>Brucella</i> sp.	Genus	3900	3150	Brucellosis (Undulant Fever, Malta Fever)
<i>Brucella suis</i> 1330	NC_004310	2108	3288	
<i>Brucella abortus</i> biovar 1 str. 9-941	NC_004311	1207		
<i>Brucella abortus</i> biovar 1 str. 9-941	NC_006932	2124	3085	Bovine Brucellosis (Bang's Disease, Abortive Fever)
<i>Brucella abortus</i> biovar 1 str. 9-941	NC_006933	1162		
<i>Brucella melitensis</i> 16M	NC_003317	2117	3308	
<i>Brucella melitensis</i> 16M	NC_003318	1178		
<i>Brucella melitensis</i> biovar Abortus 2308	NC_007618	2121	3034	
<i>Brucella melitensis</i> biovar Abortus 2308	NC_007624	1157		
<i>Coxiella burnetii</i>	Species	2102	2134	Q Fever
<i>Coxiella burnetii</i> RSA 493	NC_002740	1995	2102	
<i>Coxiella burnetii</i> RSA 493	NC_004704	37		
<i>Coxiella burnetii</i> Nine Mile phase I	NC_002118	37	37*	
<i>Coxiella burnetii</i> R 1140	NC_002131	33	34*	
<i>Coxiella burnetii</i> Priscilla Q177	Y15898	39	41*	
<i>Rickettsia</i> sp.	Genus	1266	850 - 1550	Epidemic Typhus, Scrub Typhus, Rocky Mountain Spotted Fever
<i>Rickettsia akari</i> str. Hartford	NZ_AAF01000001	1231	1217	
<i>Rickettsia belli</i> RML369-C	NC_007340	1522	1429	
<i>Rickettsia canadensis</i> str. McKiel	NZ_AAF01000001	1160	969	
<i>Rickettsia conorii</i> Malish 7	NC_003103	1269	1374	
<i>Rickettsia felis</i> URRWXCa2	NC_007109	1485	1512*	
<i>Rickettsia prowazekii</i> str. Madrid E	NC_000963	1112	839	
<i>Rickettsia rickettsii</i>	NZ_AADJ01000001	1258	1311	
<i>Rickettsia sibirica</i> 246	NZ_AABW01000001	1250	1234	
<i>Rickettsia typhi</i> str. Wilmington	NC_006142	1111	838	

Dynamic Programming Alignment Special Case



References

- Altschul, S.F., Gish, W., Miller, W., Myers, E.W. & Lipman, D.J. (1990) "Basic local alignment search tool." *J. Mol. Biol.* 215:403-410.
- Li L, Stockert CJ Jr, Roos DS. (2003) OrthoMCL: identification of ortholog groups for eukaryotic genomes. *Genome Res.* 9:1278-89.
- Remm, M., Storm, C.E., Sonnhammer, E.L. (2001) Automatic clustering of orthologs and in-paralogs from pairwise species comparisons. *J Mol Biol.* 314(5):1041-52.
- Waterman, M.S. (2000) *Introduction to Computational Biology: Maps, Sequences and Genomes*. Chapman & Hall/CRC Press, New York, pp. 215-219.

Acknowledgements

This project is funded by NIAID / NIH contract HHSN26620040035C to Bruno Sobral.