Predicting Functional Associations using Flanking Genes (FlaGs)

Chayan Kumar Saha, Rodrigo Sanchez Pires, Harald Brolin and Gemma Catherine Atkinson

Department of Molecular Biology and Umeå Centre for Microbial Research, Umeå University, Sweden

Abstract

Summary

Functional associations of proteins can be predicted by conservation of the genomic neighbourhood surrounding the gene encoding the protein of interest. We have developed a tool, FlaGs, for Flanking Genes, that clusters neighbourhood-encoded proteins into homologous groups and outputs the identity of the groups, a graphical visualization of the gene neighbourhood and its conservation, and – optionally – a phylogenetic tree annotated with flanking gene conservation.

Availability and implementation

The software is implemented in Python for Mac and Linux environments and is freely available along with a manual and accompanying example input data and results files from the project GitHub page:
https://github.com/GCA-VH-lab/FlaGs

Contact

gemma.atkinson@umu.se

Introduction

Conservation of gene order at long evolutionary distances is a strong indicator of a functional relationship among genes (Dandekar, et al., 1998; Overbeek, et al., 1999). Extreme examples are the tryptophan biosynthesis (Dandekar, et al., 1998), and *str* ribosomal protein operons (Lechner, et al., 1989), which are conserved from bacteria to archaea. While such functional clustering is most commonly associated with prokaryotic genomes, it can even be observed in eukaryotic genomes (Lee and Sonnhammer, 2003). The vast amount of genomic sequence data that has become available in recent decades is a treasure trove of clues about the function of uncharacterised proteins, and the pathways in which they are involved (Gabaldon and Huynen, 2004). High-throughput identification of gene order conservation in genomes is a promising approach for predicting the involvement of proteins in various pathways. Specific examples are the identification of conserved bicistronic loci seen in toxin-antitoxin (TA) systems (Sevin and Barloy-Hubler, 2007), and the prediction of novel proteins involved in adaptive immunity though association with CRISPR-cas systems (Shmakov, et al., 2018).

In addition to yielding functional predictions, the identification of conserved genomic architectures is essential for understanding the evolutionary dynamics behind the formation of gene clusters, and their restructuring, including reassembly of operons after disruption during evolution (Omelchenko, et al., 2003).

The most widely used tools for analysing flanking gene conservation include String (Szklarczyk, et al., 2015), and The Seed viewer (Overbeek, et al., 2005). String includes a number of other metrics for predicting functional associations in addition to flanking gene conservation. It takes as input one protein from a representative organism in the String database, and the results often summarise evidence from a mix of orthologues (genes related by vertical descent) and paralogues (genes related by gene duplication). The region comparison tool of The Seed viewer (Overbeek, et al., 2005) offers more control over which organisms contribute to comparative neighbourhood analyses, but similarly to String uses its own pre-determined selection of organisms.

Giving the user complete control over which genes and organisms are considered for comparative analyses expands the application of flanking gene conservation analysis to several classes of problems. It allows targeted analyses to answer questions about specific proteins over any evolutionary distance the user is interested in. For example, if the interest is a very recent horizontal transfer event, it is useful to examine all strains in one species to view how the new gene has been incorporated; if a protein has a broad distribution and no known function, it is better to expand the analysis to representatives of *all* lineages where the protein is found. If duplication of the gene of interest has been followed by functional diversification, lineage-specific interactions can be found by searching separately for flanking genes for each paralogous group, rather than pooling all homologues together. In combination with phylogenetic analysis, flanking gene conservation analysis retraces the dynamics of genomic neighbourhood architectures over time, as well as aiding in the discrimination of orthologues from paralogues.

The FlaGs workflow

To address these challenges, we have created a Python tool for functional association prediction that gives the user complete control over which proteins and genomes are considered, and outputs figure-quality, editable vector graphics. FlaGs (standing for Flanking Genes) takes in user-determined input sequences that can come from any protein entry from any organism in the NCBI RefSeq database (around 110 million proteins as of May 2018). From an input list of accession numbers, FlaGs outputs information on the conservation of flanking genes, and their identity, in graphical and text format (Fig. 1). Such input files are easily prepared from – for example – the output from a local, or web-based NCBI BlastP search. An optional addition to the input file is the NCBI genome assembly identifier. NCBI protein accession numbers are non-redundant in that identical proteins have the same accessions and can link to multiple genomic sequences. The benefit of including assembly identifiers is that the protein is linked to a specific genome that may have its own unique architecture. Assembly identifiers are readily available from NCBI and are usually stored locally when whole genomes are downloaded for local sequence searching by tools such as Blast and HMMer. FlaGs clusters sequences using the sensitive Hidden Markov Model-based method Jackhmmer, part of the HMMER distribution (Eddy, 2011). The advantage of this approach is that it can be used for all levels of evolution – from across kingdoms or domains of life to the strain or isolate level – as the E value thresholds and numbers of iterations in the Jackhmmer searches can be modified by the user to control sensitivity.

The output of FlaGs always includes a to-scale diagram of flanking genes, number- and colour-coded by conservation groups. A "description" file is also part of the output, which includes the accession numbers and descriptions present in the flanking gene- encoded protein clusters and acts as the legend for interpreting the flanking gene diagram. An optional output is a phylogenetic tree that is annotated with flanking genes reduced to triangular pennant-like flags, also number- and colour-coded by conservation groups. The tree-building feature uses the ETE 3 Python environment (Huerta-Cepas, et al., 2016), and opens a visualisation window for user interaction with the tree. If a phylogenetic tree is not required, FlaGs can also be run without using or installing ETE 3.

Conclusion

FlaGs is a flexible tool for sensitive detection of flanking gene conservation at any evolutionary distance, and display of the results in an intuitive, figure-quality graphical format. Applications of FlaGs include the discovery of novel functional associations and the analysis of gene neighbourhood dynamics during evolution.

Acknowledgements

We wish to thank Lars Barquist for suggesting the use of Jackhmmer in our pipeline, and Vasili Hauryliuk for valuable comments on the manuscript.

Funding

The work is supported by Vetenskapsrådet (VR, the Swedish Research Council) grant 2015-04746, Jeanssons Stiftelse, and UCMR Linnaeus Program Gender Policy Support to GCA.

References

Dandekar, T., *et al.* Conservation of gene order: a fingerprint of proteins that physically interact. *Trends Biochem Sci* 1998;23(9):324-328.

Eddy, S.R. Accelerated Profile HMM Searches. *PLoS Comput Biol* 2011;7(10):e1002195.

Gabaldon, T. and Huynen, M.A. Prediction of protein function and pathways in the genome era. *Cell Mol Life Sci* 2004;61(7-8):930-944.

Huerta-Cepas, J., Serra, F. and Bork, P. ETE 3: Reconstruction, Analysis, and Visualization of Phylogenomic Data. *Mol Biol Evol* 2016;33(6):1635-1638.

Lechner, K., Heller, G. and Bock, A. Organization and nucleotide sequence of a transcriptional unit of Methanococcus vannielii comprising genes for protein synthesis elongation factors and ribosomal proteins. *J Mol Evol* 1989;29(1):20-27.

Lee, J.M. and Sonnhammer, E.L. Genomic gene clustering analysis of pathways in eukaryotes. *Genome Res* 2003;13(5):875-882.

Omelchenko, M.V., *et al.* Evolution of mosaic operons by horizontal gene transfer and gene displacement in situ. *Genome Biol* 2003;4(9):R55.

Overbeek, R., *et al.* The subsystems approach to genome annotation and its use in the project to annotate 1000 genomes. *Nucleic Acids Res* 2005;33(17):5691-5702.

Overbeek, R., *et al.* The use of gene clusters to infer functional coupling. *Proc Natl Acad Sci U S A* 1999;96(6):2896-2901.

Sevin, E.W. and Barloy-Hubler, F. RASTA-Bacteria: a web-based tool for identifying toxin-antitoxin loci in prokaryotes. *Genome Biol* 2007;8(8):R155.

Shmakov, S.A., *et al.* Systematic prediction of genes functionally linked to CRISPR-Cas systems by gene neighborhood analysis. *Proc Natl Acad Sci U S A* 2018.

Szklarczyk, D., *et al.* STRING v10: protein-protein interaction networks, integrated over the tree of life. *Nucleic Acids Res* 2015;43(Database issue):D447-452.

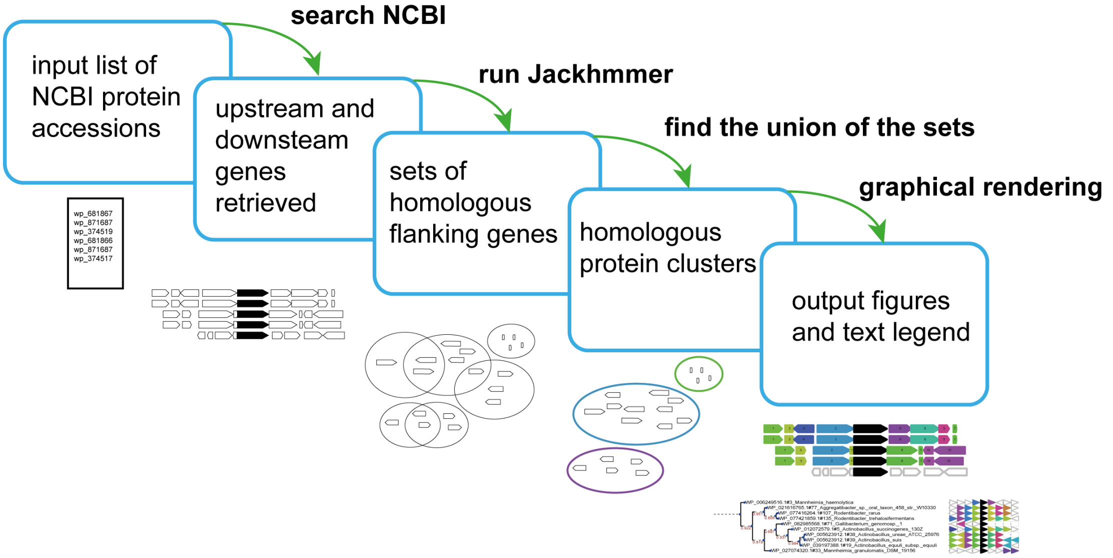
search NCBI
run Jackhmmer
find the union of the sets
graphical rendering
input list of NCBI protein accessions
upstream and downsteam genes retrieved
sets of homologous flanking genes
homologous protein clusters
output figures and text legend
wp_681867
wp_871687
wp_374519
wp_681866
wp_871687
wp_374517
WP_006249516.1#3_Mannheimia_haemolytica
WP_021616765.1#77_Aggregatibacter_sp._oral_taxon_458_str._W10330
WP_077416264.1#107_Rodentibacter_rarus
WP_077421859.1#135_Rodentibacter_trehalosifermentans
WP_082985568.1#71_Gallibacterium_genomosp._1
WP_012072579.1#5_Actinobacillus_succinogenes_130Z
WP_005623912.1#38_Actinobacillus_ureae_ATCC_25976
WP_005623912.1#39_Actinobacillus_suis
WP_039197388.1#19_Actinobacillus_equuli_subsp._equuli
WP_027074320.1#33_Mannheimia_granulomatis_DSM_19156
0.951
0.805
0.922
0.687
0.937
0.819
0.984
1

Figure 1. The FlaGs workflow. The user inputs a list of protein accession numbers – optionally with GCF assembly IDs – and can specify the number of adjacent flanking genes to consider and the sensitivity of the Jackhmmer search though changing the E value cut-off and number of iterations. The output always includes a to-scale figure of flanking genes, a description of the flanking gene identities as a legend, and optionally, a phylogenetic tree annotated with colour- and number-coded pennant flags.