SEADOG-ILP v1.0 Manual

Description

SEADOG-ILP (short for "Simultaneous Evolutionary Analysis of Domains and Genes through phylogenetic reconciliation using Integer Linear Programming") is a software package for simultaneous inference of domain-level and gene-level evolution through a joint phylogenetic reconciliation of domain, gene, and species trees. The software takes as input a rooted domain tree, rooted gene trees for the gene families in which the domains of the domain tree occur, and a rooted species tree on the species considered in the analysis. SEADOG-ILP implements an exact, ILP-based algorithm for computing optimal Domain-Gene-Species reconciliations in either a biologically restricted or unrestricted search space. SEADOG-ILP outputs all optimal Domain-Gene-Species reconciliations in its search space and often produces more optimal reconciliations that the heuristic implemented in SEADOG. However, SEADOG-ILP is not as scalable as SEADOG. A description of the domain-gene-species reconciliation model appears in the first paper listed below and a description of the exact ILP algorithm implemented by this software appears in the second. SEADOG-ILP is available open source under GPL version 3.0 and must be compiled locally by the user so it can be properly linked to the CPlex solver.

An Integrated Reconciliation Framework for Domain, Gene, and Species Level Evolution
Lei Li and Mukul S. Bansal
IEEE/ACM Transactions on Computational Biology and Bioinformatics (TCBB); 16(1): 63-76, 2019.

An Integer Linear Programming Solution for the Domain-Gene-Species Reconciliation Problem
Lei Li and Mukul S. Bansal
ACM Conference on Bioinformatics, Computational Biology, and Health Informatics (ACM-BCB) 2018: 386-397.

Installation Instructions

The software must be compiled on a Linux system using a C++ compiler implementing the C++11 standard. The software also requires access to the IBM CPLEX solver (locally installed).

CPLEX is free for academic use and can be obtained from https://www.ibm.com/products/ilog-cplex-optimization-studio

Step-by-step instructions:

1. Install CPLEX, keeping track of its install path. For example: /home/username/software/cplex
2. Inside the SEADOG-ILP folder, open the Makefile and find the following lines:

"CPLEXDIR = /home/username/softwares/cplex/cplex"
"CONCERTDIR = /home/username/softwares/cplex/concert"

Update the paths to point to your installed copy of CPLEX.

NOTE: If your makefile and source code are somehow not in the same folder, please find the line below and change accordingly.

"EXSRCCPP = ./source"

3. Inside the SEADOG-ILP folder type:

   Make clean; make

Input

The required input consists of one domain tree, a set of gene trees, and a species tree. All trees should be in Newick format, without any labels for internal nodes (may not cause any problem with the internal labels), and should be rooted and binary. Each tree must appear in a separate file. The domain tree and species tree can have arbitrary file names, but gene tree names should be of the form"geneTreeName.tree". All gene trees should be in a single directory.

Each leaf lable (domain name) in the domain tree must be of the form "domainName_GeneID_GeneTreeName", where the domainName is any label for that specific domain sequence, GeneID is the unique gene ID or name of the specific gene that contains that domain sequence, and GeneTreeName is the name of the gene tree that contains that specific gene. For example, "domainXYZ_FBgn0100324_geneTree4".

Likewise, gene names in the gene trees should be of the form “GeneID_SpeciesName”, where GeneID is the unique gene ID or name for that gene, and Species name is the label of the species from which that gene was sampled.

Command Line Arguments

The program takes the following command line arguments, among which -d, -g, and -s are required and the others are optional.

-e: Solve the exact unrestricted version of the problem with no restriction on search space. Use this option ONLY when the gene trees are small (typically less than 20 leaf nodes in total).
-d The input domain tree file.
-g Path to the directory containing the gene trees.
-s The input species tree file.
-o Output file name. By default the output file will be the input domain tree file name plus ".output".
-DD Domain duplication cost. Default value 2.
-DL Domain loss cost. Default value 1.
-DTA Domain transfer cost when the donor and recipient are in the same gene family. Default value 4.
-DTB Domain transfer cost when the donor and recipient are in different gene families. Default value

6.

-GD Gene duplication cost. Default value 2.

-GL Gene loss cost Default value 1.

Command Line Example:

./DGSILP -d domaintree1.tree -g ./ -s 12flies.stree -o output.txt -e

Output

The reconciliation output begins with a listing of the domain tree, gene trees, and species tree, with their internal nodes labeled. Each internal node is labeled by an index, followed by the indices of its two children.

The next line indicates the number of optimal reconciliations in the limited search space.

The next output block shows the reconciliation of the domain tree with the gene trees and shows the event type and mapping for each domain tree node. The reconciliation between the gene tree(s) and species tree appear next, showing the event type and mapping for each node in the gene tree(s).

If multiple optimal reconciliations are found, they are all printed one after another in arbitrary order. Each output reconciliation ends with three lines showing the final DGS reconciliation cost, the reconciliation cost between the domain tree and gene trees, and the reconciliation cost between the gene trees and species tree, respectively.

Contact

If you have any questions, suggestions, or concerns, please contact either Lei Li (lei.li@uconn.edu) or Mukul Bansal (mukul.bansal@uconn.edu).