Core Collection Detector version 1.0 Manual

Introduction

CCD is a Java software package for the selection of core collections of diverse taxa (e.g. from germplasm collections) that are intended to capture the genetic diversity of the input dataset. The software takes as input a binary matrix which transcribes the different alleles (genetic forms) for each taxon. It then tries to find a most diverse collection of alleles under various constraints (e.g. sample size, geographic location of samples).

Availability

CCD is written in Java, and the most recent version, version 1.0, is compiled using J2SE 5.0 - the application is standalone but if you are having problems running it please ensure you have Java Virtual Machine(VM) version 5.0 or above installed, available from http://java.com/en/download/manual.jsp

Once you have a suitable Java VM installed, download the ZIP file containing CCD and its dependencies. The CCD JAR file within this ZIP file includes a manifest enabling it to be run from the command line using java -jar ccd.jar.

Using the software

Step 1 (optional): Construction of distance matrix for phylogenetic tree and network based selection methods. Distances can be exported in PHYLIP format which is then loaded into SplitsTree to build a tree or network with neighbour-joining or neighbour-net respectively.
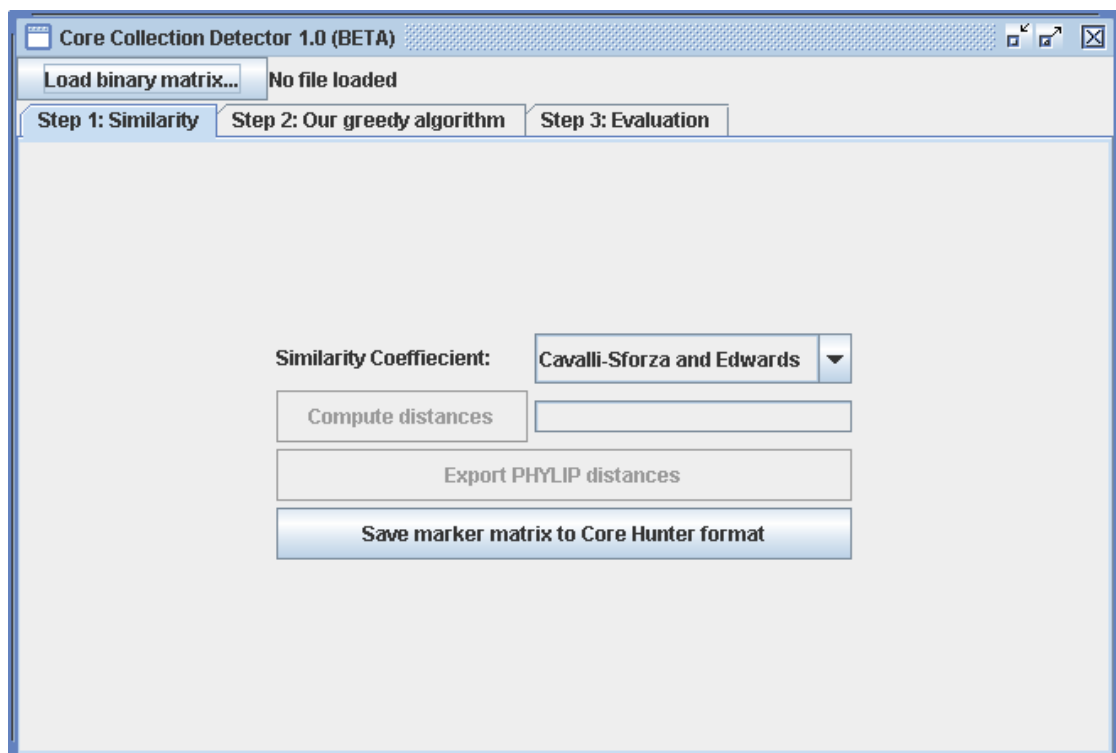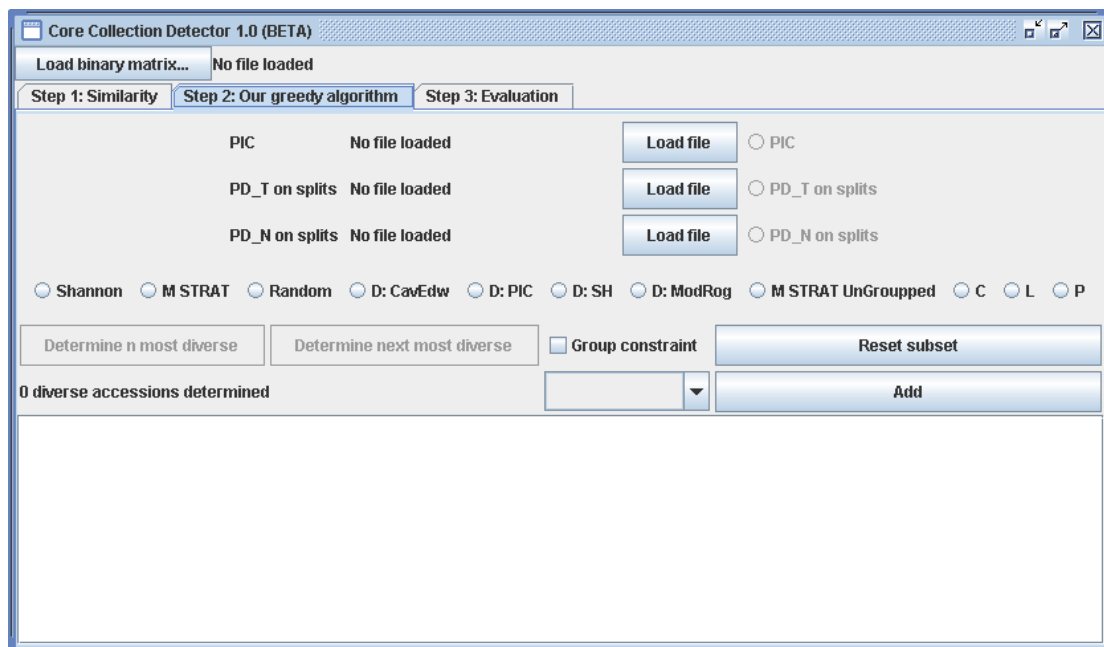
Step 2: Core collection selection.

As with step 1, 'Load Binary Matrix' can be used to load in the matrix of marker values in which, for each accession there are present/absent marker values.
A number of selection methods are implemented including:-
- Shannon Entropy
- M Strategy either using or ignoring groupings
- D Strategy (using Cavalli Sforza-Edwards distance)
- D Strategy (using Polymorphism Information Content)
- D Strategy (using Shannon Entropy)
- D Strategy (using Modified Rodgers distance)
- C Strategy
- L Strategy
- P Strategy

The `group constraint` option enforces picking the same number of accessions from each group.

`Determine n most diverse` prompts for a number and builds a core collection of that size.

`Determine next most diverse` adds one more diverse accession to the current core collection, if there is no core collection this adds the two most diverse accessions.
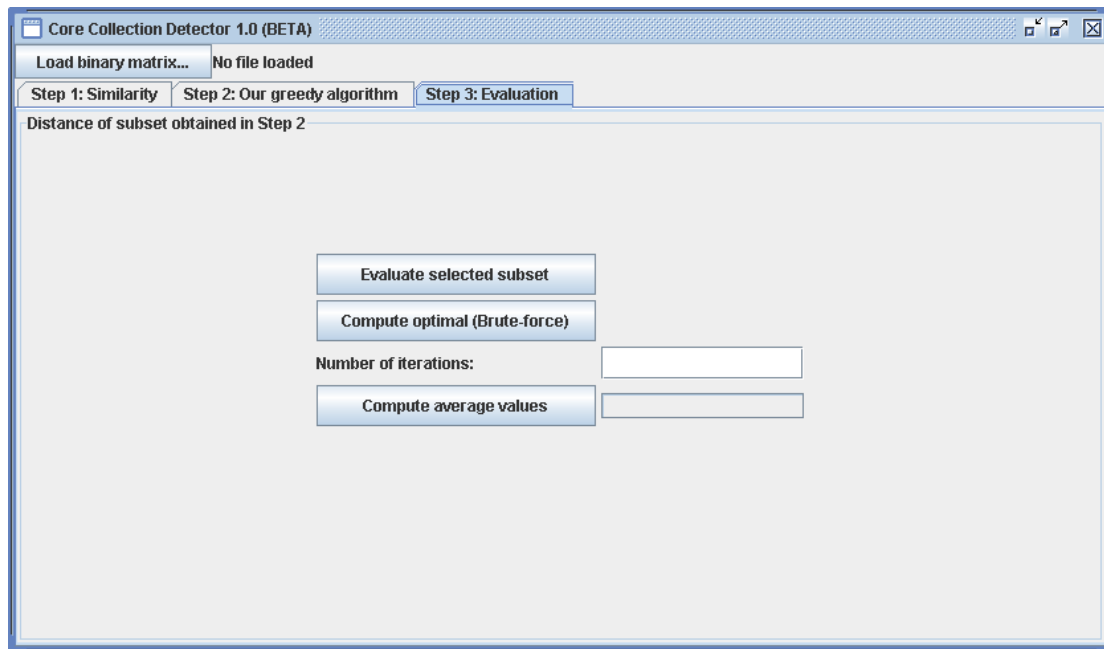
Evaluate selected subset will measure the current core collection for polymorphism information content, Shannon entropy, number of alleles, type of alleles, and if a phylogenetic tree or network is loaded for the current dataset the distance on those will also be given.

Compute optimal (Brute-Force) can be used to check all possible subsets of the same size as the current core collection from Step 2, the best subset will then be returned.

Number of iterations/Compute average values can be used to get a simple mean average for randomly generated core collections of the same size as the current core collection from Step 2.

Acknowledgement:

The work was funded in part by BBSRC grant BB/E004105/1.

Disclaimer

This software is supplied as-is, with no warranty of any kind expressed or implied. We have made every effort to avoid errors in design and execution of this software, but we will not be liable for its use or misuse. The user is solely responsible for the validity and consequences of any results generated.

Core Collection detector is © Copyright 2010 Martin Lott and Vincent Moulton.