

Module: Control Generation (CoGent)

Summary:

CoGent (**Control Generator**) creates data sets for controlling informatic analyses. These data sets consist of genomic loci and/or genomic sequences. The data is taken from a database of actual genomic sequence and annotations, as opposed to ad-hoc generation, sequence scrambling, or the like. This produces biologically relevant and accurate results which allow for stronger controls. The controls are matched against a user provided data set via a number of parameters.

User defined parameters include:

- The species, assembly, and annotation to utilize (e.g. Human – NCBI_b35 - RefSeq)
- The locus type to retrieve from (gene, exon, UTR, etc)
- Min/max or matching polynucleotide lengths
- Sequence concatenation
- Matching GC content

Algorithm:

CoGent utilizes the Hocus Locus database structure and access manager (as described in “Database Schema”) to provide the user with a list of available species, assemblies, and annotations to choose from. It then retrieves random samples and filters this data based on user defined parameters. These parameters can be contextual to the annotation (CDS only, 5' UTRs, etc) and they can be matched to the user's data set for greater control accuracy.

I - Procedure:

See Fig 1

A data set is loaded into CoGent in the form of a LocusSet. This represents the loci/sequences to be controlled for. Each record in the data set will produce a matched control, and each evaluated criteria is contextual to the current user record being examined. First the user must choose which species/assembly database (as described in “Database Schema”) they wish to use. Once selected, the user can then be presented with a list of annotation tables, and again a selection must be made. Examples of annotation tables are: RefSeq, KnownGene, miRNAs, Transcription Factor Binding sites, Methylation, etc.

The user then sets parameters which will act as filters on the data. The first level of filtering happens during data retrieval. A random sample is selected from the user defined table and only the specified loci are returned. The possible loci are contextual to the annotation table selected. For example, miRNAs would just have a single locus per record, while KnownGene could return whole gene regions, CDS, UTR, etc. This sample size is configurable and is used to maintain a pool of data, thus minimizing database lookups. CoGent then uses this pool of data and applies the second set of filtering criteria.

The algorithm branches depending on whether the user requested sequences, or loci only. For the latter, the algorithm iterates over the loci in the pool and attempts to apply any length criteria (matching length, minimum, etc). If the locus, or a subset, can meet the criteria it is saved to the control set and the next user record is examined. Otherwise it is discarded.

If the user requested sequences, then the actual nucleotide sequence is retrieved for the loci in the pool. The user can decide if they would like their control sequences to come from a single concatemerized sequence. This avoids creating any 'center selection' bias when randomly selecting regions from within a given locus. If this is the case, then an appropriate length sequence is selected with a random starting point, continuing across one or more sequences as needed to complete the length. If concatemerization is not required, then the algorithm iterates over the loci in the pool and attempts to apply any length criteria (as before). Once an appropriate length sequence is found, it is checked for matching GC content. GC content can be set to match a given percentage threshold from $\pm 100\%$ (GC does not need to be matched) up to $\pm 5\%$. If the locus matches required GC content, it is saved to the control set and the next user record is examined. Otherwise it is discarded.

Once all the records in the user defined data set have a matched control, the algorithm exits and the control set is returned to the user.

Potentially Novel Items:

1. Generation of informatic controls based on actual biological data from a user-defined source.
2. Ability to define locus-based subsets of data to use (genes, exons, etc).
3. Ability to virtually concatemerize sequence to avoid 'center selection' bias when randomly selecting sub-regions.

Figure 2

CoGent – User Definable Parameters...

