SYNOPSIS

% perl extract_ALL_chrs.pl

The packages DBI and DBD_mySql must be installed

This script must currently be run through an account logged onto the lexomics server

This will likely change soon

DESCRIPTION

Summary

This script accesses a database and retrieves all the different organisms on the server and gets some basic information about them

Input

The script needs access to a mysql server containing the genomes of a number of organisms and their metadata. If you have already have a mysql server the script seed.pl can be used to create and populate the database with all the micro-organisms with complete genomes in NCBI's database.

Connecting to Your Database

Once you have a database ready you will need to make a few minor edits to the script so it can connect to your database. Search this file for the mysql_dbh subroutine:

```
sub mysql_dbh {
```

Modify these four lines, replacing this generic data with your database's access info:

```
my $db         = 'test';
my $host       = 'localhost';
my $user       = 'GenomicsUser';
my $pass       = '';
```

For $db enter the name of the (MySQL) database you are using. Ex: 'GenomeDatabase'

For $host enter the name or address of the server your database is on. Ex: 'WheatonGenomics'

For $user enter your MySQL username on the database. Ex: 'wsmith'

For $pass enter the user's corresponding password (or leave it blank). Ex: 'lollipop'

Output

Every chromosome for every organism will be stored in a seperate .fna file in a subdirectory in the data_all_chr folder named after the organism. In addition in the folder the script is run in a file Chr_Stats.xls will contain for every chromosome in every organism the length, percentage coding the number of genes, the number of overlaps and the number of each type of overlap

Overlap codes

```
Complete - one gene completely within a second gene
 |-----------------|
           |----|

Partial one gene partially inside second gene
        |-------|
             |----|

(OR) the gene overlaps the origin of replication
```

AUTHORS

```
Mark D. LeBlanc
Donald W. Bass
```

Modification History

6/17/2010 (nkf)

Removed Wheaton database log-in info from the mysql_dbh subroutine. Users must replace the generic data with their own database access info.

6/15/2010 (dwb)

Made script create the folder data_all_chr if it does not already exist

6/07/2010 (dwb)

Completely rewrote code for detecting overlaps to make it more accurate and also simplified the number of types of overlaps from 5 to 2, by removing two types that were mirror images of another two types, and removing overlap type 5 which stood for no overlap, and didn't make sense with the new code which instead of running one comparision per gene, runs as many as necessary to catch all overlaps

6/03/2010 (dwb)

added additional code documentation

6/02/2010 (dwb)

added pod documentation

12/02/2008 (mdl)

worked on genic-intergenic STAT output to get a better feel for the extent of genic regions, including the types of operonds and/or overlapping regions between consecutive genes (see ChrSTAT.xls output)

11/20/2008 (mdl)

doing only CHROMOSOMES now

12/03/2007 (mdl)

REMOVED abstracted_chromosome code ... created a file for stats of plasmids

11/30/2007 (mdl)

looked again at why length of abs-chr is not equal to length of real chr + extras; NOT SOLVED ...

10/19/2007 (mdl)

start morphing to work with PLASMIDS

02/01/2007 (mdl)

dang! some genomes have a final gene that bridges the (man-made) origin (bp 1); inserting fix (and code to check if longer abstracted chromosome length is right

# 01/29/2007 (mdl) -

ABSTRACTED CHROMOSOME genes on the indirect(-) strand are stored in their reverse complement form so that all counting will be in a 5' to 3' direct all gene regions are stored independently even if two gene regions overlap;

```
5'  ...|---1---|......      3'

3' ...|---2---|....         5'
```

these two regions will be completely consecutively #1.N.#2' where 2' means the sequence in #2 is stored in reverse complement fashion;

08/22/2006 (mdl) -

modified to handle multiple chromosomes per genome

NOTE: 10/19/2007

(should also handle multiple plasmids)

07/13/2006 (mdl)

mostly done

COPYRIGHT INFORMATION

#what is 562