Identitag documentation

Using Identitag for tag-to-gene mapping

1 Introduction

Serial Analysis of Gene Expression (SAGE) is a method of large-scale gene expression analysis that has the potential to generate the full list of mRNAs present whithin a cell population at a given time and their frequency. This method is based on the assumption that short cDNA sequences (tags) are in most cases sufficient to identify transcripts. Thus the tags count provides a numerical measurement of the quantity of each transcript. An essential step in SAGE library analysis is the unambiguous assignment of each tag to the transcript from which it is derived, which is called tag-to-gene mapping.

We designed and implemented a tool called Identitag for tag-to-gene mapping. This tool is based on a relational database which structure can be depicted as three interconnected modules represented in Identitag relational schema (available on Identitag website : http ://pbil.univ-lyon1.fr/software/identitag.html). The first one stores virtual tags extracted from transcript sequences belonging to the species considered, the second stores experimental tags observed in SAGE experiments, and the third allows the annotation of the transcript sequences used for virtual tag extraction. Identitag therefore connects an observed tag to a virtual tag and to the transcript sequence from which it is derived, and to its functional annotation when available. For a complete description of Identitag tables see the data dictionary (available on Identitag website).

Identitag website provides Identitag sources which can be used to build a database for tag identification in any species from which transcript sequences are available. For questions and comments on Identitag please contact Céline Keime : keime@ cgmc.univ-lyon1.fr.

2 What is necessary to build an Identitag database ?

2.1 Software requirements

- A Bourne Shell (/bin/sh) and Perl interpreter (/usr/bin/perl)

- A MySQL server and client
  (LOAD DATA LOCAL must be active on MySQL client)

2.2 Data requirements

- A file containing transcript sequences from the species considered, in Fasta format : the word immediately following the greater than symbol (>) is the identifier of the sequence, all characters following this identifier (separated by one ore more blanks) correspond to the description of this sequence, all lines not beginning with the symbol > correpond to the sequence whose characteristics are written above and preceded by >.

- A file containing SAGE data, i.e. tag sequences and their occurence number in the library considered (one file for each library you want to analyze). This file must contain an header line (not taken into account during this analysis). After the header each line must contain a different tag sequence following by its occurence number, separated by a tabulation (this file could for example be generated using SAGE 2000 software to extract tags from ditag concatemers). In this file, tag sequences must not contain the restriction site of the anchoring enzyme used : i.e. for SAGE using BsmfI as tagging enzyme, the length of this sequence must be 10 bp, and for long SAGE using MmeI as tagging enzyme, the length of this sequence must be 17 bp.

- A BLASTX file result obtaining with the -m 8 option of blastall. This file must contain the results of sequence comparisons between transcript sequences from the species considered (either all sequences from the fasta file described in first item or a part of these sequences) and proteins.

- The three files previously described are required to create and load Identitag tables. By using these files, Identitag scripts load all fields from Identitag tables excepted the last four fields of Protein table. You can choose to load them by answering yes to the corresponding question during Identitag building procedure. For this you need a file containing information about proteins used for BLASTX comparison, e.g. sprot.dat or/and trembl.dat (these files could be found on EBI ftp site : ftp.ebi.ac.uk/pub/databases/sp_tr_nrdb/). Be aware this could spend relatively long time.

3 Building an Identitag database

- Create a database and a client with privileges to drop, create table and insert into tables belonging to this database, on MySQL DBMS.

- Uncompress the file identitag.tgz : this produces a directory called "identitag" which contains all scripts needed to create and load all Identitag tables.

- In identitag directory, run sh Identitag.sh.

- Answer to all questions, this will create and load all Identitag tables.