Improved modelling of RNA-seq and ChIP-seq bias using multiple alternative nucleotide distributions

Software and example files: Installation and usage instructions

This document describes the installation and usage of software tools that were designed to implement the algorithms described in “Improved modelling of RNA-seq and ChIP-seq bias using multiple alternative nucleotide distributions”.

The software is a combination of:

- GUIs: the cisGenome GUI and GUIs implemented in Microsoft Excel
- Data visualisation using the cisGenome web server application and Excel spreadsheets.
- Command line applications that implement the core algorithms

The software will run on a PC running Windows XP/Vista/7 and Excel 2003 or later. The command line applications can also be built and supplied on request for other server platforms.

All of the software and some example data file are available from:

http://www2.warwick.ac.uk/fac/sci/moac/students/2007/nigel_dyer/softwareforpaper

Software Installation

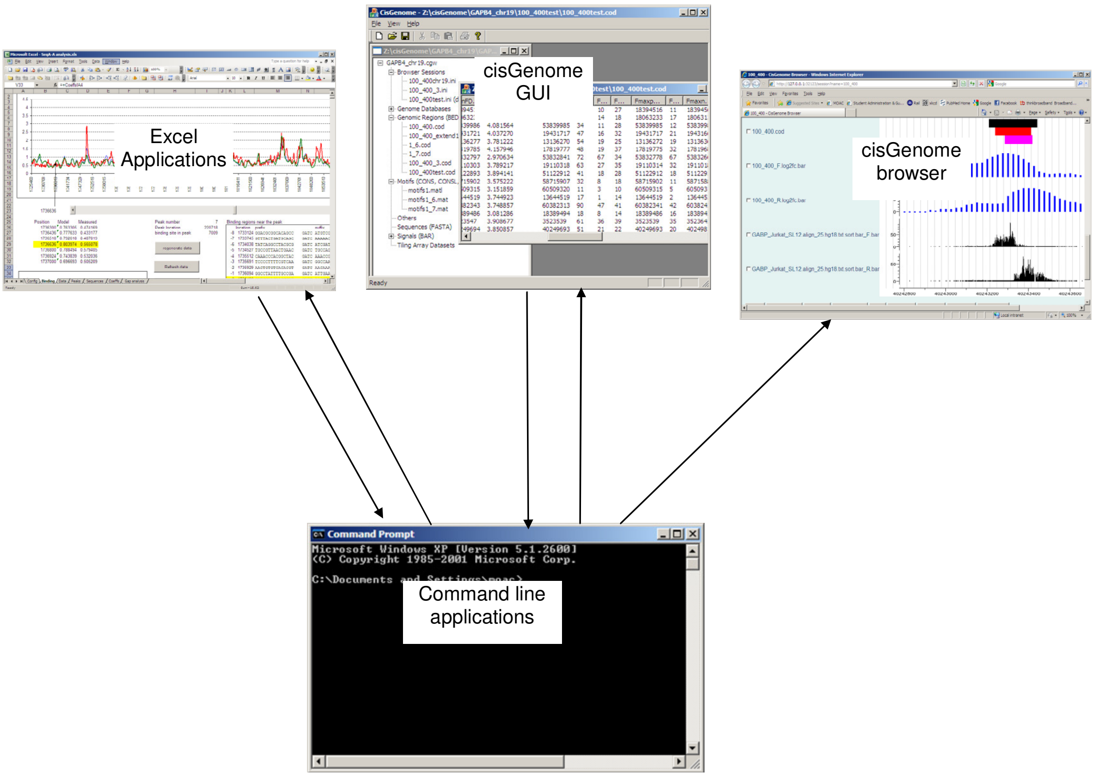
cisGenome software

The modified version of the cisGenome software was used to produce many of the plots and perform much of the data preparation described in the paper. This should be downloaded and installed.

Spreadsheets and applications

These are available from the website should be downloaded and unzipped into a single directory (<ModellingTools>) ready for use.

Software Architecture



```
Microsoft Windows XP [Version 5.1.2600]
(C) Copyright 1985-2001 Microsoft Corp.

C:\Documents and Settings\moac>
```

Figure 1 Software architecture showing interaction between different applications

The general architecture is as shown in figure 1, where both cisGenome and a series of Excel spreadsheets are used to assemble the set of command line options for running command line applications, and then they are used to display the results of the operations, or use the cisGenome browser to view the results.

Common principles

Many of the applications work on the principle of having a working directory (-o option) for the parameter file ("weights.csv") and the output files. Once a model fitting process has been completed, the weights.csv file, and any other files might be subsequently needed should be copied from the working directory to some other location.

An introduction to the cisGenome GUI

The cisGenome GUI is a way of holding information about a set of files in a '.cgw' project file. It provides sets of dialogs for running programs where the files can be selected from those included in the project file.

It also allows results to be displayed in a web browser interface, typically by selecting and double clicking the item to be viewed. These include:

- A PSSM set described in a matl file
- A PSSM in a mat file

ChIP-seq data analysis

Data preparation

The starting point is ChIP-seq data that has been aligned to a genome with an application such as Bowtie.

The ChIP-seq sequence tag data is created in a wide variety of different formats and must first be converted to the cisGenome .bar file format.

This is done using the cisGenome ‘sequencing/Alignment->bar’ function which can convert a wide variety of different the tag data format files into three files:

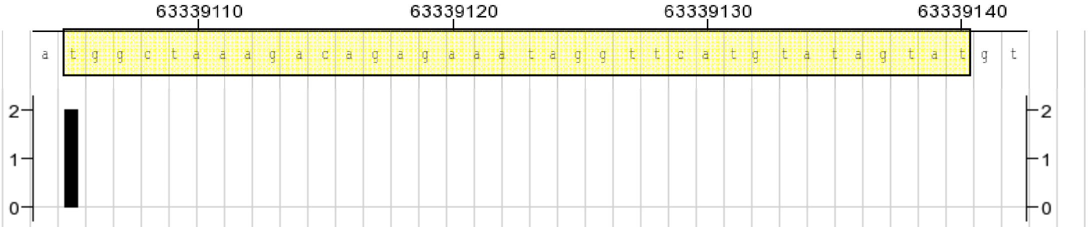
- <filename>.bar
- <filename>.bar_F.bar
- <filename>.bar_R.bar

These files contain all of the tag data, the tag data that maps to the forward strand and the tags that map to the reverse strand respectively.

There are some formats which this function cannot handle. However the ‘Format Conversion/<XX>->ALN’ function is able to convert some of these into a form that the ‘Alignment to bar’ function can handle.

There is not a universal convention as to either the numbering of nucleotides or the nucleotide numbering that should be attached to an aligned tag, so it is necessary to confirm that any aligned data confirms to the convention that has been adopted within the bias analysis and modelling software.

In the case of forward reads the nucleotide that should be identified is the first nucleotide of the tag or fragment, which is the first nucleotide after the point where the DNA fractures, as in this example from the SL523 data.

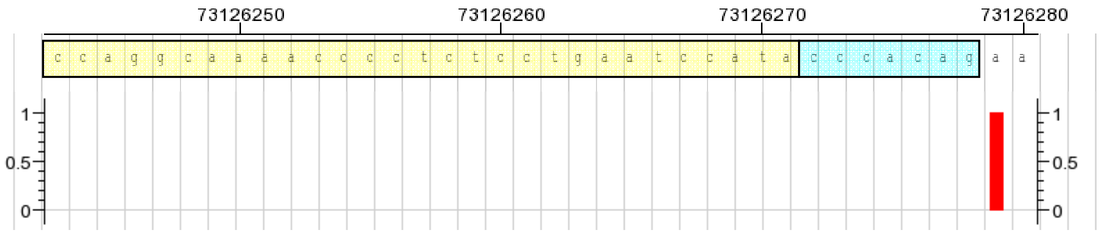

63339110
63339120
63339130
63339140
a t g g c t a a a g a c a g a g a a a t a g g t t c a t g t a t a g t a t g t
2
1
0
2
1
0

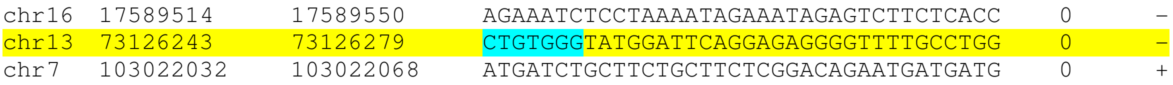
```
chr11  22194284      22194320      TTTTTAGAGATGTATGTTTTCATTAATCTTGGTTAA    1       +
chr11  63339105      63339141      TGGCTAAAGACAGAGAAATAGGTTCATGTATAGTAT    0       +
chr18  20223355      20223391      GGAAAGAAAATTTATATTTTGTTGCCTTATAAATAT    0       +
```

In the case of the reverse reads the sequence provided in the alignment file may be either the original sequence that was read or the reverse complement of the sequence that was originally found in order to male it easier to see the alignment to the forward strand.

In the following example, the selected tag starts with the sequence CTGTGGG..., which is the start (right hand side) of the sequence on the reverse strand. The maps to the sequence tag that ends ...CCCACAG

The nucleotide that is then used to identify the sequence is the nucleotide immediately after the end of the tag (shown in red), which is also the first nucleotide after the point where the DNA fractures, as in this example from the SL523 data





```
chr16  17589514     17589550     AGAAATCTCCTAAAATAGAAATAGAGTCTTCTCACC     0       -
chr13  73126243     73126279     CTGTGGGTATGGATTCAGGAGAGGGGTTTTGCCTGG     0       -
chr7   103022032    103022068    ATGATCTGCTTCTGCTTCTCGGACAGAATGATGATG     0       +
```

In some cases, as is the case with the SL523 data, the original alignments conform to this convention. The next most frequent example is that the identities need to be shifted by one nucleotide for both forward and reverse strands.

The Alignment->bar dialog box includes the options of adding an offset to both strands, and an additional offset for the reverse strand.

The aligned SL523 data is available for download

Analysing sequence bias

The 'Get Biases' button within 'ChIP-seq sequence bias analysis.xls' converts the tag information in a set of bar files into a csv file that contains sequence bias for all of the N-mers in each chromosome. Options are available for selecting some or all of the chromosomes, and only selecting the forward or reverse strand.

The output filename is of the format:

```
SL523.txt.TagAlign.sort_ana.bar_o-1_l8_c0_all.csv
```

Which identifies the offset, N-mer length, number of chromosomes (0 for all chromosomes) and whether the forward, reverse of both strands were used to generate the data.

As the tag data will not include tags for unmappable regions, these should not be included when calculating the statistics for sequence distribution. This is done with the exclude unmappable regions options. The filename contain this information must have a filename of the similar to matches36F.cod, where 36 is the sequence length for which the unmappable regions were calculated, and the F indicates that it is for the forward strand. This file is also used to create the mappability information for the reverse strand mapping, which is the same position data shifted by the sequence read length.

The matches36F.cod file for the human genome hg18 is available for download

A pre-generated bias summary file is available for download

Model fitting

ChIP-seq model fit.xls provides a simple GUI for running the optimiseBreaks.exe executable in order to find a set of PCMs that best match the best match to the sequence bias data previously generated. A set of working and output files are generated in the working directory (e.g. 'C:\ModellingTools\chipSeq analysis'), in particular the weights.csv file which contains that generated PCM papermeters. The process is outlined in figure 2, with the identity of the spreadsheet buttons highlighted

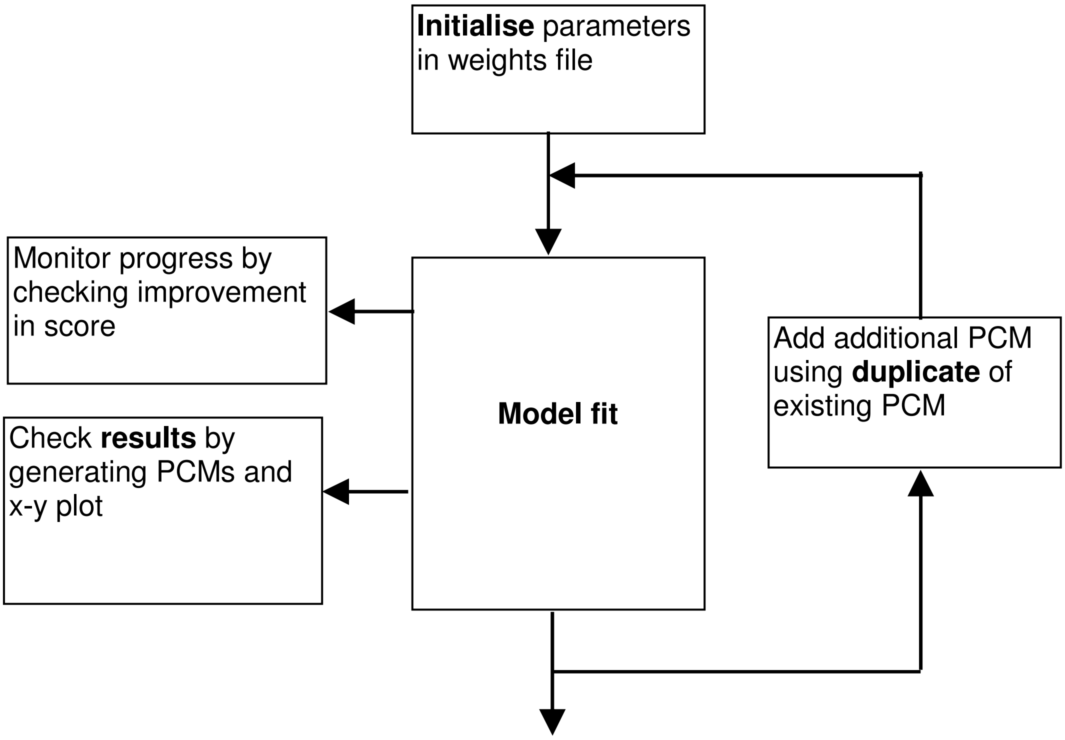

Initialise parameters in weights file
Monitor progress by checking improvement in score
Check **results** by generating PCMs and x-y plot
Model fit
Add additional PCM using **duplicate** of existing PCM

Figure 2 Process of model fitting

An initial set of 'null' weights is created using **Initialise**

Model fitting starts the model fitting process in a separate thread, allowing the **results** button to be used to view the current match between the model and the observed bias as the model fitting proceeds.

Viewing the **results** also causes a set of PCMs to be generated. To view these using cisGenome, open cisGenome, and then open the 'breakPSSM.cgw' project file that has automatically been generated in the working directory, and then double click the 'breakPSSM.matl' entry within the motifs section to view the set of PCMs using a web browser. If the results are regenerated, it is only necessary to refresh the web browser to see the new PCMs.

Results can be viewed in this way while model fitting is proceeding in order to monitor the progress of the fitting.

Progress can be monitored using Progress.xls browsing for **progress.csv** in the current working directory, and refreshing to view progress, or waiting for the automatic refresh. The contents of the progress file can be **Reset**, which should be done in order to see the current progress in more detail.

If the model-fitting is showing that is unable to proceed any further with the current number of PCMs, the process can be interrupted.

Duplicate adds an additional PCM, with the same weights as the PCM that is currently associated with most sequences. **Model fitting** can then proceed again.

At the end of the model fitting the weights.csv file should be copied/renamed as a record of the outcome of the process. The **copy weights from** allows a set of weights to be obtained from the location where a previous set of weights had been retained.

RNA-seq data analysis

RNA-seq model fit.xls provides a simple GUI for running the optimiseStartSeq.exe executable which is used for model fitting RNA-seq data

In the case of RNA-seq data it is more likely that the starting point is raw data that will need to be aligned prior to starting the model fitting.

The approach used for the datasets in the paper is to align the tag data to the representative cDNA sequence and then map the unaligned data to a set of cDNA sequences that cater for some of the transcript variants that exist for each gene.

The model fitting software requires access to a fasta file containing the sequences and identity for the full set of transcripts (which include the representative sequences as a subset.

The TAIR10 sequences used is available for download

The optimiseStartSeqs software that is called to perform the model fitting or to calculate the degree of fitting automatically converts the Bowtie output into a compressed form with a .alnmap suffix if one does not already exist or uses the compressed version if it already exists.

The compressed version for the Arabidopsis 24 hr replicate 1 data is available for download

The process of model fitting follows the same approach as for the ChIP-seq data (Figure 2), with the results and working files contained in a working directory (e.g. 'C:\ModellingTools\rnaSeq analysis'). The PCMs can be viewed by opening the rnaPSSM.cgw file in the working directory with the cisGenome GUI and double clicking rnaPSSM.matl within the motifs section of the tree.

The weights file that is generated follows the topology that is described in the paper, with multiple PCMs for the first 6 nucleotides (and one before the start of the fragment) and a single PCM for nucleotide 7 onwards. The topology can be changed by manually editing the layout of the weights.csv file.

The **setref** button allows a particular model prediction to be saved as a reference line so that the improvements that occur during subsequent model fitting can be monitored.

Nigel Dyer

21/5/13