BSSM4GSQ code package

Michael E Sparks

July 17, 2007

1 Getting Started

This software facilitates generation of splice site probabilities for use with `GeneSeqer`, and training data with which to build splice site parameters using the `gthbssmbuild` tool in the `GenomeThreader` package. `BSSM4GSQ` can process either plain text `GeneSeqer` or gthXML v1.0 (or later) output. gthXML output can be produced natively by the `gth` program of `GenomeThreader`, and can be produced from plain text `GeneSeqer` output using the `GSQ2XML.pl` script, available from either `http://www.genomethreader.org` or `http://www.public.iastate.edu/~mespar1/gthxml/`. The end user may wish to study the following reports prior to using this software:

1. Brendel V, Xing L, Zhu W. (2004) Gene structure prediction from consensus spliced alignment of multiple ESTs matching the same genomic locus. *Bioinformatics*. **20**:1157-69.

2. Sparks ME and Brendel V. (2005) Incorporation of splice site probability models for non-canonical introns improves gene structure prediction in plants. *Bioinformatics*. **21**:iii20-iii30.

2 Directions

1. Verify that the following executables are present in the bin/ directory.

   (a) `indexFasSeq`

   (b) `BSSM_build`

   (c) `BSSM_print`

If any of these are absent, cd to the src/ directory and issue "make". (This step is, however, optional, as the `Mktraindata.sh` and `Mkbssmparm.sh` scripts will build the files, when necessary.)

2. There are two subdirectories in the input directory, gsq/ and fas/. You absolutely must meet the following requirements to make the code work.

   (a) There must be a one-to-one correspondence between files in these two directories.

   (b) Files in the fas/ directory must have an extension of ".fas" and those in the gsq/ directory must have one of either ".gsq" or ".xml", for plain text `GeneSeqer` or gthXML formatted data files, respectively. You cannot mix plain text and gthXML input files.

   (c) The basename prefixes of cognate fas/gsq file pairs, i.e., the substrings prior to ".gsq" or ".fas", must be identical.

   (d) All references made to a genomic template in the spliced alignment output file must be identical to the file's "file handle" mentioned above. This essentially mandates that each fas/gsq cognate file pair correspond to one genomic sequence and its spliced alignment annotation, respectively.

   There are sample training data in the input/sample/ directory. These data will not generate any meaningful probabilities, and are only intended to demo the system.

3. Edit `Mktraindata.sh` such that the FORMAT variable is set correctly for the files placed in the input/gsq/ directory; this is described explicitly in the header of the script. Run `Mktraindata.sh`. This produces exon and intron data, sorted according to phase and placed in the output/exons_introns/ directory, and sampled, phase-sorted BSSM training data placed in the output/training_data/ directory. In each of these directories, data will be written to a subdirectory named according to the donor/acceptor dinucleotide termini trained for. (If this is unclear, inspect the contents of the output directories after unpacking this code, run the script, and look at them again.)

   `Mktraindata.sh` processes GT-AG introns by default. For other types, tune the DON and ACC variables (set these in CAPITAL letters!) and run it again. This will not overwrite any existing output in

the training_data/ or exons_introns/ directories so long as a different DON/ACC combination is used. Rerunning the script using a DON/ACC pair whose results were already recorded will cause the original data to be overwritten.

4. Run `Mkbssmparm.sh`. The script will solicit some configuration information:

   (a) Name of output file ("foo.bssm")

   (b) Root directory of training data. (If you haven't done anything non-standard up to this point, it should be safe to just say "y" here.)

   (c) Build GT model? For GT-AG parameterizations. (If you trained for these intron types, responding with "y" will put the probabilities in your *.bssm file. Else, say "n".)

   (d) Build GC model? For GC-AG parameterizations. (If you trained for these intron types, responding with "y" will put the probabilities in your *.bssm file. Else, say "n".)

   (e) File to write ascii data to ("foo.bssm.ascii")

   (Users of the `GenomeThreader` package (`http://www.genomethreader.org`) should note that these binary *.bssm files are **not compatible** with those used by that system. Code in `BSSM4GSQ` implementing the weight array matrix development routines was adapted to implement the `gthbssmbuild` program of `GenomeThreader`; given input data like that produced in steps 1-3 above, `gthbssmbuild` will generate `GenomeThreader`-compatible binary *.bssm files. Please refer to the `GenomeThreader` manual or contact Gordon Gremme at `gremme@zbh.uni-hamburg.de` for details.)

   The `BSSM_print` utility allows the user to generate an ascii representation of the trained splice site probability matrices. The *.bssm.ascii file presents the weight array matrices in the following order:

```
for TERMINAL in (0-1):
  for HYPOTHESIS in (0-6):
    print transition probabilities
```

where for TERMINAL, 0 and 1 index donor and acceptor sites, respectively; and for HYPOTHESIS, 0, 1, 2, 3, 4, 5, 6 index the T1, T2, T0, F1, F2, F0 and Fi hypotheses, respectively.

The user will, unfortunately, have to manually splice the new probabilities into the daPbm7.* header files included with the `GeneSeqer` source distribution (adjust the GU_7/GC_7 and AG_7 arrays, the name_model array, and the NMDLS macro accordingly) to incorporate the models into the software. There is a script provided in the src/plscripts/ directory, `punctuation.pl`, that will assist in adding syntactical markup to allow copying/pasting results into the header files, e.g.,

```
$ cat something.bssm.ascii | ./punctuation.pl
```

You can verify that your parameter file contains valid probability mass distributions by using the `verify_pmf.pl` script in the src/plscripts/ directory, e.g.,

```
$ cat something.bssm.ascii | ./verify_pmf.pl
```

Please see the commentary in that file for more details; output of a row of zeros is expected, and does not signal an error.

3 Contact Info

If you have questions, concerns, etc., please email me at `mespar1@iastate.edu`.