jD2Stat User Manual

Updated 15 January 2015

jD2Stat (previously known as **JIWA***) is a Java package that utilises a series of D_2 statistics to extract *k*-mers (subsequences at defined *k* length) from a set of biological sequences, and generate pairwise distances for each possible pair. This distance can be used directly for phylogenetic inference using neighbour-joining.

*: This program is not to be confused with, and has no relation to, the proprietary trademark and software of Jiwa Financials, North Sydney, Australia.

The D_2 methods included are:

- D_2 based on exact k-mer counts (**D2**)
- D_2^S - similar to D_2, normalised based on probability of occurrences of specific *k*-mers (**D2S**)
- D_2^* - similar to D_2, normalised based on means and variance (**D2St**)
- D_2^n - extension of D_2 to allow for *n* number of wildcards (*n* neighbourhood)

Requirements

jD2Stat runs on all platforms that support the Java Runtime Environment JRE 1.7.x or higher. jD2Stat comes in a .jar file that can be invoked on command line directly under JRE, so no compilation or installation is necessary. In the rare occasion that you have no Java installed on your computer, Java is freely available from http://www.java.com/.

Memory requirement varies depending on the number and length of the sequences, overall sequence similarity and the *k*-mer length, for obvious reasons. RAM of 1GB would be sufficient for most simple analyses. For instance, in our analysis of 5000 16S rRNA sequences, the memory usage is about 2.5GB.

Input

jD2Stat accepts DNA and protein sequences in FASTA format. A file should contain a set of two or more sequences, and **MUST** have any one of these extensions: .faa, .fas, .fasta, .ffn, .fna or .frn. Both single- and multi-line FASTA format are accepted.

IMPORTANT: Note that jD2Stat considers only conventional nucleotide and amino acid characters in the analysis: A, C, G and T for DNA, and the standard 20 amino

acids (single-letter code) for proteins. The presence of any non-conventional characters in the sequences will result in an error.

Output

For each D_2 analysis, jD2Stat generates a pairwise distance matrix in PHYLIP format, which can be used as input for **`neighbor`** in PHYLIP to infer a phylogenetic tree using neighbour-joining.

Implementation

jD2Stat can be invoked directly on command line:

```
java –Xmx<vmem> –jar jD2Stat_1.0.jar <options> -i <input> -o <output>
```

Quick starts

In following examples: input sequences in **`myseq.fas`**, output matrix file named **`myseq.matrix`**.

```
java –Xmx4g –jar jD2Stat_1.0.jar –d D2 –k 8 -i myseq.fas -o myseq.matrix
```

This runs D_2 statistic on DNA sequences at *k*-mer length **8** with max memory set at 4GB.

```
java –Xmx4g –jar jD2Stat_1.0.jar –a aa –d D2St –k 4 -i myseq.fas -o myseq.matrix
```

This runs D_2^* statistic on protein sequences at *k*-mer length **4** with max memory set at 4GB.

```
java –Xmx8g –jar jD2Stat_1.0.jar –n 1 –k 8 -i myseq.fas -o myseq.matrix
```

This runs D_2^n statistic (with neighbourhood $n = 1$) on DNA sequences at *k*-mer length **8** with max memory set at 8GB.

```
java –Xmx12g –jar jD2Stat_1.0.jar –a aa –d all –k 4 -i myseq.fas
```

This runs all three D_2, D_2^S and D_2^* statistics, independently on protein sequences at *k*-mer length **4** with max memory set at 12GB. Three distance matrices in separate files will be generated.

Parameters

`-XmX<vmem>`	`<vmem>` is the maximum amount of RAM that JAVA can use (this value by default on most computers is **`4g`**).
`-i <input>`	`<input>` is the input file in FASTA format (extensions .faa, .fas, .fasta, .ffn, .fna or .frn)

`-o <output>`	`<output>` is the output file name for the pairwise distance matrix. You must specify this to get a matrix. Default: no output unless in **`-d all`** mode (see options below).

Options

`-a <a>`	`<a>` is the alphabet/character type of the sequences: `dna` (default) or `aa` (amino acid).
`-d <d>`	`<d>` is the choice of D_2 method: **`D2`** (default), **`D2S`**, **`D2St`** or **`all`** to run all three methods of `D2`, `D2S` and `D2St`. **`-d all`** yields all three matrices in separate files even without `-o` specification.
`-h`	Display all available options on the screen
`-k <k>`	`<k>` is the length of *k*-mers to be used in the analysis. Default: **`8`**
`-n <n>`	`<n>` is the neighbourhood value of n in D_2^n. For instance, **`-n 1`** specifies $D_2^{n=1}$. When `-n` is specified, only D_2^n is run regardless of specification in `-d`.
`-t <t>`	`<t>` is the number of threads to use in the analysis. Default: **`4`**

Citation

If you use jD2Stat or D_2^n in your work, please cite:

Chan CX, Bernard G, Poirion O, Hogan JM and Ragan MA (2014). Inferring phylogenies without multiple sequence alignment. *Scientific Reports*, **4:** 6504. DOI:10.1038/srep06504.

Contact

For more information about jD2Stat and if you have technical issues with the program, please contact:

Guillaume Bernard: guillaume.bernard@imb.uq.edu.au

Cheong Xin Chan: c.chan@imb.uq.edu.au