Ig-Indel-Identifier README

The program identifies legitimate insertions/deletions (indels), vague indels and artifact indels, as explained in our paper:
Michaeli, M., Noga, H., Tabibian-Keissar, H., Barshack, I. and Mehr, R., (2012). Automated cleaning and pre-processing of immunoglobulin gene sequences from high-throughput sequencing. Frontiers in Immunology 3:386.

Preparations:

- Please make sure your computer has a Java Runtime Environment installed. You can freely download it at: http://www.java.com/en/download/.
- Please save the attached JAR file in the same directory where all input files are saved.
- Running the program is done by double clicking on the JAR file, after it is saved in the same directory with all input files.
- Please remember to change the name of the example input file into "input.txt".

Input:
Ig-Indel-Identifier works on *.txts files that are generated during a ClustalW2 run using the "pir" option (please see: http://www.ebi.ac.uk/Tools/msa/clustalw2/). Each *.txts file should contain a clone of sequences and their germline. The header of the germline should be precisely: ">G.L".
Please put the FASTA formatted file containing all the sequences in the same directory. This file extension should be changed to *.input (instead of *.fasta or *.txt for example), when '*' means the original name of the file.
In addition, the user should prepare an input file exactly named "input.txt" and include it in the same directory with the pir files and the *.input file containing the sequences, as mentioned above. The "input.txt" file includes all information for the run: The minimum number of sequences in a clone that must share the same indel or a low quality score point mutation for this indel or mutation to be considered legitimate, HPT length, and the minimal quality score for a point mutation to be considered legitimate. If the user is not interested in identifying such mismatches, the value of the minimal quality score given in the input.txt file should be set to −1. Please find below an example of an input.txt file and follow the instructions using this example.

----------EXAMPLE START----------

```
Number of sequences
2
Homopolymer tract length
3
Minimal quality score
20
```

----------EXAMPLE END----------

Each Ig-Indel-Identifier run is performed on a single sample.
For more clarifications regarding input files please see the paper.

Output:
Ig-Indel-Identifier generates three files containing the sequences, as follows.

- OriginalNameOfFile-WithoutIndels.txt – contains all sequences in which no indels were found.
- OriginalNameOfFile-IllegitimateIndels.txt – contains all sequences in which artifact indels were found.
- OriginalNameOfFile-CloneOfSize1WithIndels.txt – contains all sequences in which vague indels were found.

For more clarifications regarding the definitions of indels, please see the paper.
In addition, the program generates a log file (named OriginalNameOfFile-Ig-Indel-Identifier.log), in which all output is written to. At the end of this file you can find some statistics on the numbers of indels found, for example:

```
The current sample had 44116 suspected indels and 330 artifact indels.
Out of that, 11385 were insertions and 32731 were deletions.
```

```
Number of 5' indels: 10984. Number of 3' indels: 13543. Number of indels in the middle
of a homopolymer tract: 2427
Number of OK sequences: 838
```

The last line denotes the number of sequences without any indels, i.e., the number of sequences written in the OriginalNameOfFile-WithoutIndels.txt file.