Installation

To install FragGeneScan, please follow the steps below:

1. Untar the downloaded file "FragGeneScan.tar.gz". This will automatically generate the directory "FragGeneScan".

2. Make sure that you also have a C compiler such as "gcc" and perl interpreter.

3. Run "makefile" to compile and build excutable "FragGeneScan" make clean make fgs

Running the program

1. To run FragGeneScan,

./run_FragGeneScan.pl -genome=[seq_file_name] -out=[output_file_name] -complete=[1 or 0] -train=[train_file_name] -thread=[num_thread]

[seq_file_name]: sequence file name including the full path [output_file_name]: output file name including the full path [whole_genome]: 1 if the sequence file has complete genomic sequences 0 if the sequence file has short sequence reads [train_file_name]: file name that contains model parameters; this file should be in the "train" directory. Note that four files containing model parameters already exist in the "train" directory. [complete] for complete genomic sequences or short sequence reads without sequencing error [sanger_5] for Sanger sequencing reads with about 0.5% error rate [sanger_10] for Sanger sequencing reads with about 1% error rate [454_5] for 454 pyrosequencing reads with about 0.5% error rate [454_10] for 454 pyrosequencing reads with about 1% error rate [454_30] for 454 pyrosequencing reads with about 3% error rate [illumina_5] for Illumina sequencing reads with about 0.5% error rate [illumina_10] for Illumina sequencing reads with about 1% error rate [num_thread]: number of thread used in FragGeneScan. Default 1.

2. To test FragGeneScan with a complete genomic sequence,

./run_FragGeneScan.pl -genome=./example/NC_000913.fna -out=./example/NC_000913-fgs -complete=1 -train=complete

[NC_000913.fna]: this sequence file has the complete genomic sequence of E.coli (NCBI gene predictions for this genome are available under the same folder example/)

3. To test FragGeneScan with sequencing reads,

./run_FragGeneScan.pl -genome=./example/NC_000913-454.fna -out=./example/NC_000913-454-fgs -complete=0 -train=454_10

[NC_000913-454.fna]: this sequence file has simulated reads (pyrosequencing, average length = 400 bp and sequencing error = 1%) generated using Metasim

For illumina reads, please use illumina_5 or illumina_10 as the train model.

4. To test FragGeneScan with assembly contigs, ./run_FragGeneScan.pl -genome=./example/contigs.fna -out=./example/contigs-fgs -complete=1 -train=complete

Note: -complete=1 & -train=complete are used as the parameters.

5. To test FragGeneScan with whole genome, ./run_FragGeneScan.pl -genome=./example/NC_000913.fna -out=./example/NC_000913-fgs -complete=1 -train=complete

Output

Upon completion, FragGeneScan generates four files.

1. The first file is "[output_file_name].out", which lists the coordinates of putative genes. This file consists of five columns (start position, end position, strand, frame, score). For example,

> gi|49175990|ref|NC_000913.2| Escherichia coli str. K-12 substr. MG1655, complete genome
> 108 440 - 3 1.378688 337 2799 + 1 1.303498 2801 3733 + 2 1.317386 3734 5020 + 2 1.293573 5234 5530 + 2 1.354725 5683 6459 - 1 1.290816 6529 7959 - 1 1.326412 8238 9191 + 3 1.286832 9306 9893 + 3 1.317067

2. The second file is '[output_file_name].ffn", which lists nucleotide sequences corresponding to the putative genes in "[output_file_name].out". For example,

gi|49175990|ref|NC_000913.2| Escherichia coli str. K-12 substr. MG1655, complete genome start=108 e nd=338 strand=-
GTTGTTACCTCGTTACCTTTGGTCGAAAAAAAAAGCCCGCACTGTCAGGTGCGGGCTTTTTTCTGTGT
TTCCTGTACGCGTCAGCCCGCACCGTTACCTG
TGGTAATGGTGATGGTGGTGGTAATGGTGGTGCTAATGCGTTTCATGGATGTTGTGTACTCTGTAATTT
TTATCTGTCTGTGCGCTATGCCTATATTGGT TAAAGTATTTAGTGACCTAAGTCAA
gi|49175990|ref|NC_000913.2| Escherichia coli str. K-12 substr. MG1655, complete genome start=343 e nd=2799 strand=+
TTGAAGTTCGGCGGTACATCAGTGGCAAATGCAGAACGTTTTCTGCGTGTTGCCGATATTCTGGAAA
GCAATGCCAGGCAGGGGCAGGTGGCCACCGTCC
TCTCTGCCCCCGCCAAAATCACCAACCACCTGGTGGCGATGATTGAAAAAACCATTAGCGGCCAGGA
TGCTTTACCCAATATCAGCGATGCCGAACGTAT
TTTTGCCGAACTTTTGACGGGACTCGCCGCCGCCCAGCCGGGGTTCCCGCTGGCGCAATTGAAAAC
TTTCGTCGATCAGGAATTTGCCCAAATAAAACAT
GTCCTGCATGGCATTAGTTTGTTGGGGCAGTGCCCGGATAGCATCAACGCTGCGCTGATTTGCCGTG
GCGAGAAAATGTCGATCGCCATTATGGCCGGCG

3. The third file is '[output_file_name].faa", which lists amino acid sequences corresponding to the putative genes in "[output_file_name].out". For example,

gi|49175990|ref|NC_000913.2| Escherichia coli str. K-12 substr. MG1655, complete genome start=108 e nd=338 strand=-
VVTSLPLVEKKSPHCQVRAFFCVSCTRQPAPLPVVMVMVVVMVVLMRFMDVVYSVIFICLCAMPILVK
VFSDLSQ gi|49175990|ref|NC_000913.2| Escherichia coli str. K-12 substr. MG1655, complete genome start=343 e nd=2799 strand=+
LKFGGTSVANAERFLRVADILESNARQGQVATVLSAPAKITNHLVAMIEKTISGQDALPNISDAERIFAELL
TGLAAAQPGFPLAQLKTFVDQEFAQIKH
VLHGISLLGQCPDSINAALICRGEKMSIAIMAGVLEARGHNVTVIDPVEKLLAVGHYLESTVDIAESTRRI
AASRIPADHMVLMAGFTAGNEKGELVVLG
RNGSDYSAAVLAACLRADCCEIWTDVDGVYTCDPRQVPDARLLKSMSYQEAMELSYFGAKVLHPRTIT
PIAQFQIPCLIKNTGNPQAPGTLIGASRDEDE
LPVKGISNLNNMAMFSVSGPGMKGMVGMAARVFAAMSRARISVVLITQSSSEYSISFCVPQSDCVRA
ERAMQEEFYLELKEGLLEPLAVTERLAIISVVG
DGMRTLRGISAKFFAALARANINIVAIAQGSSERSISVVVNNDDATTGVRVTHQMLFNTDQVIEVFVIG
VGGVGGALLEQLKRQQSWLKNKHIDLRVCGV
ANSKALLTNVHGLNLENWQEELAQAKEPFNLGRLIRLVKEYHLLNPVIVDCTSSQAVADQYADFLREGF
HVVTPNKKANTSSMDYYHQLRYAAEKSRRKF
LYDTNVGAGLPVIENLQNLLNAGDELMKFSGILSGSLSYIFGKLDEGMSFSEATTLAREMGYTEPDPRD
DLSGMDVARKLLILARETGRELELADIEIEP
VLPAEFNAEGDVAAFMANLSQLDDLFAARVAKARDEGKVLRYVGNIDEDGVCRVKIAEVDGNDPLFK
VKNGENALAFYSHYYQPLPLVLRGYGAGNDVTA AGVFADLLRTLSWKLGV

> gi|49175990|ref|NC_000913.2| Escherichia coli str. K-12 substr. MG1655, complete genome
> start=2801 end=3733 strand=+
> VKVYAPASSANMSVGFDVLGAAVTPVDGALLGDVVTVEAAETFSLNNLGRFADKLPSEPRENIVYQCW
> ERFCQELGKQIPVAMTLEKNMPIGSGLGSSAC
> SVVAALMAMNEHCGKPLNDTRLLALMGELEGRISGSIHYDNVAPCFLGGMQLMIEENDIISQQVPGF
> DEWLWVLAYPGIKVSTAEARAILPAQYRRQDCI
> AHGRHLAGFIHACYSRQPELAAKLMKDVIAEPYRERLLPGFRQARQAVAEIGAVASGISGSGPTLFALCD
> KPETAQRVADWLGKNYLQNQEGFVHICRLD TAGARVLEN

4. [output_file_name].gff gene prediction results in gff format.

License