DATE: 19-03-22

WEBLEM 8

Introduction to Gene Prediction and various elements in Prokaryotes and Eukaryotes

With the rapid accumulation of genomic sequence information, there is a pressing need to use computational approaches to accurately predict gene structure. Computational gene prediction is a prerequisite for detailed functional annotation of genes and genomes. The process includes detection of the location of open reading frames (ORFs) and delineation of the structures of introns as well as exons if the genes of interest are of eukaryotic origin. The ultimate goal is to describe all the genes computationally with near 100% accuracy. The ability to accurately predict genes can significantly reduce the amount of experimental verification work required.

CATEGORIES OF GENE PREDICTION PROGRAMS

The current gene prediction methods can be classified into two major categories, ab initio–based and homology-based approaches. The ab initio–based approach predicts genes based on the given sequence alone. It does so by relying on two major features associated with genes. The first is the existence of gene signals, which include start and stop codons, intron splice signals, transcription factor binding sites, ribosomal binding sites, and polyadenylation (poly-A) sites. In addition, the triplet codon structure limits the coding frame length to multiples of three, which can be used as a condition for gene prediction. The second feature used by ab initio algorithms is gene content, which is statistical description of coding regions. It has been observed that nucleotide composition and statistical patterns of the coding regions tend to vary significantly from those of the noncoding regions. The unique features can be detected by employing probabilistic models such as Markov models or hidden Markov models to help distinguish coding from noncoding regions.

The homology-based method makes predictions based on significant matches of the query sequence with sequences of known genes. For instance, if a translated DNA sequence is found to be similar to a known protein or protein family from a database search, this can be strong evidence that the region codes for a protein. Alternatively, when possible exons of a genomic DNA region match a sequenced cDNA, this also provides experimental evidence for the existence of a coding region.

Some algorithms make use of both gene-finding strategies. There are also a number of programs that actually combine prediction results from multiple individual programs to derive a consensus prediction. This type of algorithms can therefore be considered as consensus based.

Gene Prediction Using Markov Models and Hidden Markov Models

Markov models and HMMs can be very helpful in providing finer statistical description of a gene. A Markov model describes the probability of the distribution of nucleotides in a DNA sequence, in which the conditional probability of a particular sequence position depends on k previous positions. In this case, k is the order of a Markov model. A zero-order Markov model assumes each base occurs independently with a given probability. This is often the case for noncoding sequences. A first-order Markov model assumes that the occurrence of a base depends on the base preceding it. A second-order model looks at the preceding two bases to determine which base follows, which is more characteristic of codons in a coding sequence.

The use of Markov models in gene finding exploits the fact that oligonucleotide distributions in the coding regions are different from those for the noncoding regions. These can be represented with various orders of Markov models. Since a fixed-order Markov chain describes the probability of a particular nucleotide that depends on previous k nucleotides, the longer the oligomer unit, the more non randomness can be describedfor the coding region. Therefore, the higher the order of a Markov model, the more accurately it can predict a gene.

FGENESB is a web-based program that is also based on fifth-order HMMs for detecting coding regions. The program is specifically trained for bacterial sequences. It uses the Vertibi algorithm to find an optimal match

for the query sequence with the intrinsic model. A linear discriminant analysis (LDA) is used to further distinguish coding signals from noncoding signals.

Step-by-Step Description of FGENESB annotation.

STEP 1. Finds all potential ribosomal RNA genes using BLAST against bacterial and/or archaeal rRNA databases, and masks detected rRNA genes.

STEP 2. Predicts tRNA genes using tRNAscan-SE program (Washington University) and masks detected tRNA genes.

STEP 3. Initial predictions of long ORFs that are used as a starting point for calculating parameters for gene prediction. Iterates until stabilizes. Generates parameters such as 5th-order in-frame Markov chains for coding regions, 2nd-order Markov models for region around start codon and upstream RBS site, Stop codon and probability distributions of ORF lengths.

STEP 4. Predicts operons based only on distances between predicted genes.

STEP 5. Runs BLASTP for predicted proteins against COG database, cog.pro.

STEP 6. Uses information about conservation of neighboring gene pairs in known genomes to improve operon prediction.

STEP 7. Runs BLASTP against NR for proteins having no COGs hits.

STEP 8. Predicts potential promoters (BPROM program) or terminators (BTERM) in upstream and downstream regions, correspondingly, of predicted genes. BTERM is the program predicting bacterial - independent terminators with energy scoring based on discriminant function of hairpin elements.

STEP 9. Refines operon predictions using predicted promoters and terminators as additional evidences.

Prediction Using Discriminant Analysis.

Some gene prediction algorithms rely on discriminant analysis, either LDA or quadratic discriminant analysis (QDA), to improve accuracy. LDA works by plotting a two-dimensional graph of coding signals versus all potential 3_ splice site positions and drawing a diagonal line that best separates coding signals from noncoding signals based on knowledge learned from training data sets of known gene structures. QDA draws a curved line based on a quadratic function instead of drawing a straight line to separate coding and noncoding features. This strategy is designed to be more flexible and provide a more optimal separation between the data points.

FGENES is a web-based program that uses LDA to determine whether a signal is an exon. In addition to FGENES, there are many variants of the program. Some programs, such as FGENESH, make use of HMMs. There are others, such as FGENESH C, that are similarity based. Some programs, such as FGENESH+, combine both ab initio and similarity-based approaches.

Output format:

- G - predicted gene number, starting from start of sequence
- Str - DNA strand (+ for direct or - for complementary)
- Feature - type of coding sequence: CDSf - First (Starting with Start codon), CDSi - internal (internal exon), CDSl - last coding segment, ending with stop codon)
- TSS - Position of transcription start (TATA-box position and score)
- TSS - position of transcription start
- TATA - position of TATA-box
- wTATA - Discriminant function score for TATA box
- Start and End - Position of the Feature
- Weight - Discriminant function score for the feature
- ORF - start/end positions where the first complete codon starts and the last codon ends

An issue related to gene prediction is promoter prediction. Promoters are DNA elements located in the vicinity of gene start sites (which should not be confused with the translation start sites) and serve as binding sites for the gene transcription machinery, consisting of RNA polymerases and transcription factors. Therefore, these DNA elements directly regulate gene expression. Promoters and regulatory elements are traditionally determined by experimental analysis. The process is extremely time consuming and laborious. Computational prediction of promoters and regulatory elements is especially promising because it has the potential to replace a great deal of extensive experimental analysis.

PREDICTION ALGORITHMS

Current algorithms for predicting promoters and regulatory elements can be categorized as either ab initio based, which make de novo predictions by scanning individual sequences; or similarity based, which make predictions based on alignment of homologous sequences; or expression profile based using profiles constructed from a number of co-expressed gene sequences from the same organism. The similarity type of prediction is also called phylogenetic foot-printing.

PREDICTION FOR PROKARYOTES

One of the unique aspects in prokaryotic promoter prediction is the determination of operon structures, because genes within an operon share a common promoter located upstream of the first gene of the operon. Thus, operon prediction is the key in prokaryotic promoter prediction. Once an operon structure is known, only the first gene is predicted for the presence of a promoter and regulatory elements, whereas other genes in the operon do not possess such DNA elements.

There are a number of methods available for prokaryotic operon prediction. The most accurate is a set of simple rules developed by Wang et al. This method relies on two kinds of information: gene orientation and intergenic distances of a pair of genes of interest and conserved linkage of the genes based on comparative genomic analysis. A scoring scheme is developed to assign operons with different levels of confidence. This method is claimed to produce accurate identification of an operon structure, which in turn facilitates the promoter prediction.

This newly developed scoring approach is, however, not yet available as a computer program. The prediction can be done manually using the rules, however. The few dedicated programs for prokaryotic promoter prediction do not apply the Wang et al. rule for historical reasons. The most frequently used program is BPROM.

BPROM is a web-based program for prediction of bacterial promoters. It uses a linear discriminant function combined with signal and content information such as consensus promoter sequence and oligonucleotide composition of the promoter sites. This programfirst predicts a given sequence for bacterial operon structures by using an intergenic distance of 100 bp as basis for distinguishing genes to be in an operon. This rule is more arbitrary than theWang et al. rule, leading to high rates of false positives. Once the operons are assigned, the program is able to predict putative promoter sequences. Because most bacterial promoters are located within 200 bp of the protein coding region, the program is most effectively used when about 200 bp of upstream sequence of the first gene of an operon is supplied as input to increase specificity.

Output format:

- First line - name of your sequence;
- Second and Third lines - LDF threshold and the length of presented sequence
- 4th line - The number of predicted promoters
- Next lines - positions of predicted promoters, and their scores with 'weights' of two conserved promoter boxes. Promoter position assign to the first nucleotide of the transcript (Transcription Start Site position).
- After that we present elements of Transcriptional factor binding sites for each predicted promoter (if they found).

PREDICTION FOR EUKARYOTES

The ab initio method for predicting eukaryotic promoters and regulatory elements also relies on searching the input sequences for matching of consensus patterns of known promoters and regulatory elements. The consensus patterns are derived from experimentally determined DNA binding sites which are compiled into profiles and stored in a database for scanning an unknown sequence to find similar conserved patterns. However, this approach tends to generate very high rate of false positives owing to nonspecific matches with the short sequence patterns. Furthermore, because of the high variability of transcription factor binding sites, the simple sequence matching often misses true promoter sites, creating false negatives.

To increase the specificity of prediction, a unique feature of eukaryotic promoter is employed, which is the presence of CpG islands. It is known that many vertebrate genes are characterized by a high density of CG dinucleotides near the promoter region overlapping the transcription start site. By identifying the CpG islands, promoters can be traced on the immediate upstream region from the islands. By combining CpG islands and other promoter signals, the accuracy of prediction can be improved. Several programs have been developed based on the combined features to predict the transcription start sites in particular.

The eukaryotic transcription initiation requires cooperation of a large number of transcription factors. Cooperativity means that the promoter regions tend to contain a high density of protein-binding sites. Thus, finding a cluster of transcription factor binding sites often enhances the probability of individual binding site prediction. TSSW is a web program that distinguishes promoter sequences from non-promoter sequences based on a combination of unique content information such as hexamer/trimer frequencies and signal information such the TATA box in the promoter region. The values are fed to a linear discriminant function to separate true motifs from background noise.

Output format:

- First line - name of your sequence;
- Second and Third lines - LDF threshold and the length of presented sequence
- 4th line - The number of predicted promoter regions
- Next lines - positions of predicted sites, their 'weights' and TATA box position (if found)
- Position shows the first nucleotide of the transcript (TSS position)
- After that functional motifs are given for each predicted region; (+) or (-) reflects the direct or complementary chain; S... means a particular motif identificator from the Wingender data base.
- Lower cased letters mean non-conserved nucleotides in the site consensus
- The letters except (A,T,G,C) describe ambiguous sites in a given DNA sequence motif, where a single character may represent more than one nucleotide using Standard IUPAC Nucleotide code.

ORF finder:

ORF finder searches for open reading frames (ORFs) in the DNA sequence you enter. The program returns the range of each ORF, along with its protein translation. Use ORF finder to search newly sequenced DNA for potential protein encoding segments, verify predicted protein using newly developed SMART BLAST or regular BLASTP. This web version of the ORF finder is limited to the sub range of the query sequence up to 50 kb long. Stand-alone version, which doesn't have query sequence length limitation.

Thus, TSSW and BPROM are a useful tool for the recognition promoter region and start of transcription. Understanding the regulation of gene expression is an important aspect of understanding the gene function, thus this tool will be useful in elucidating regulation and expression mechanisms of genes and may even shed light on the function of novel and uncharacterized genes with the help of knowledge of promoters. FGENESB tool is useful for prediction of bacterial operon and gene and FGENES for prediction of exons. Identifying the genes that are grouped together into operons may enhance our knowledge of gene regulation and function, and such information is an important addition to genome annotation. All this can be done with the help of FGENESB. ORF finder can used to predict open reading frames in the genome. This information of long ORFs are often used, along with other evidence, to initially identify candidate protein-coding regions or

functional RNA-coding regions in a DNA sequence. Small Open Reading Frames (small ORFs/sORFs/smORFs) are important sources of putative peptides previously dismissed as being non-functional or junk DNA.

REFERENCES:

1. Xiong, J. (2008).Gene Prediction. Essential bioinformatics. Cambridge: Cambridge University Press. 97-111.
2. Xiong, J. (2008). Promoter and Regulatory Element Prediction. Essential bioinformatics. Cambridge: Cambridge University Press. 113-119.
3. TSSW - Recognition of human PolII promoter region and start of transcription. (n.d.). Www.softberry.com. Retrieved March 18, 2022, from http://www.softberry.com/berry.phtml?topic=tssw&group=programs&subgroup=promoter
4. BPROM - Prediction of bacterial promoters. (n.d.). Www.softberry.com. Retrieved March 18, 2022, from http://www.softberry.com/berry.phtml?topic=bprom&group=programs&subgroup=gfindb
5. FGENESB - Bacterial Operon and Gene Prediction. (n.d.). Www.softberry.com. Retrieved March 18, 2022 from http://www.softberry.com/berry.phtml?topic=fgenesb&group=programs&subgroup=gfindb
6. FGENES - pattern-based gene structure prediction. (n.d.). Www.softberry.com. Retrieved March 18, 2022, from http://www.softberry.com/berry.phtml?topic=fgenes&group=programs&subgroup=gfind
7. Home - ORFfinder – NCBI. (2019). Nih.gov. Retrieved March 18, 2022, from https://www.ncbi.nlm.nih.gov/orffinder/

DATE: 19-03-22

WEBLEM 8a

TSSW

(URL: http://www.softberry.com/berry.phtml?topic=tssw&group=programs&subgroup=promoter)

AIM:

To recognize the Saccharomyces cerevisiae kinase and start of transcription using TSSW tool.

INTRODUCTION:

kinase, an enzyme that adds phosphate groups (PO_4^{3-}) to other molecules. A large number of kinases exist—the human genome contains at least 500 kinase-encoding genes. Included among these enzymes' targets for phosphate group addition (phosphorylation) are proteins, lipids, and nucleic acids. Saccharomyces cerevisiae kinase and start of transcription can to recognized using TSSW.

TSSW is a web program that distinguishes promoter sequences from non-promoter sequences based on a combination of unique content information such as hexamer/trimer frequencies and signal information such the TATA box in the promoter region. The values are fed to a linear discriminant function to separate true motifs from background noise.

METHODOLOGY:

1. Open homepage for softberry. (URL: http://www.softberry.com/)
2. Under search for promotor/functional motifs select TSSW. (URL: http://www.softberry.com/berry.phtml?topic=tssw&group=programs&subgroup=promoter)
3. Retrieve nucleotide FASTA sequence for protease from GenBank.
4. Process the FASTA sequence on TSSW.
5. Observe and interpret the results.

OBSERVATION:

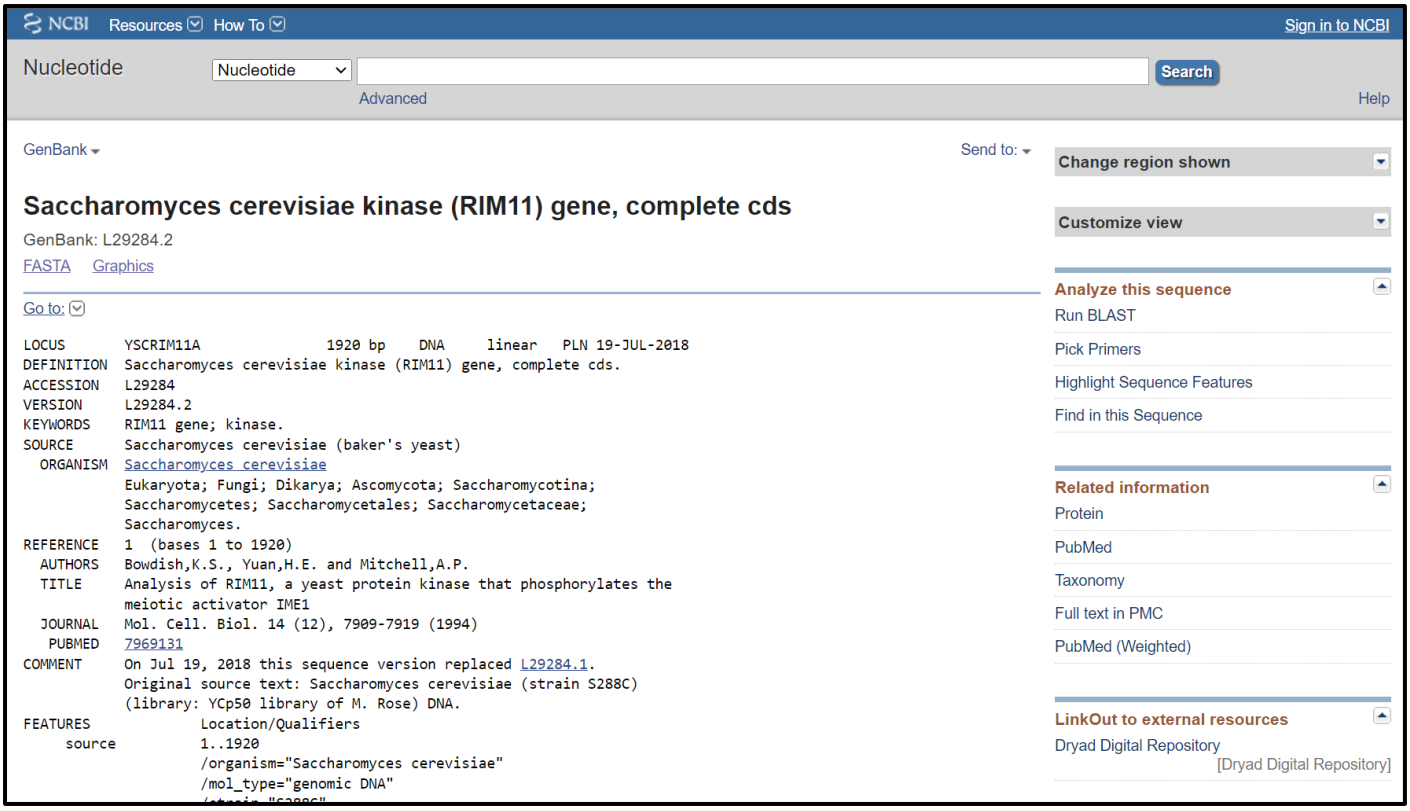

Fig1. GenBank result for Human kinase

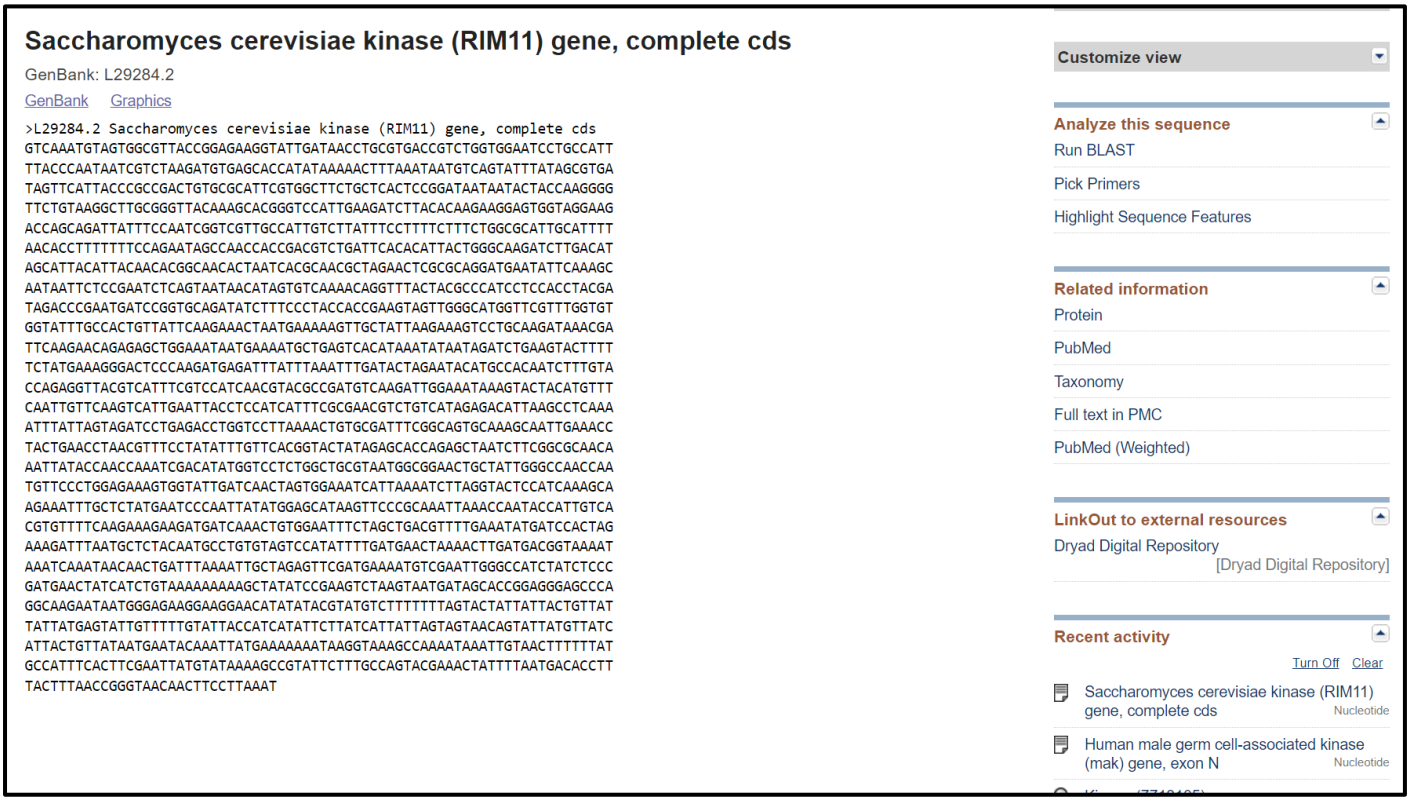

Fig2. Nucleotide FASTA sequence for Protease

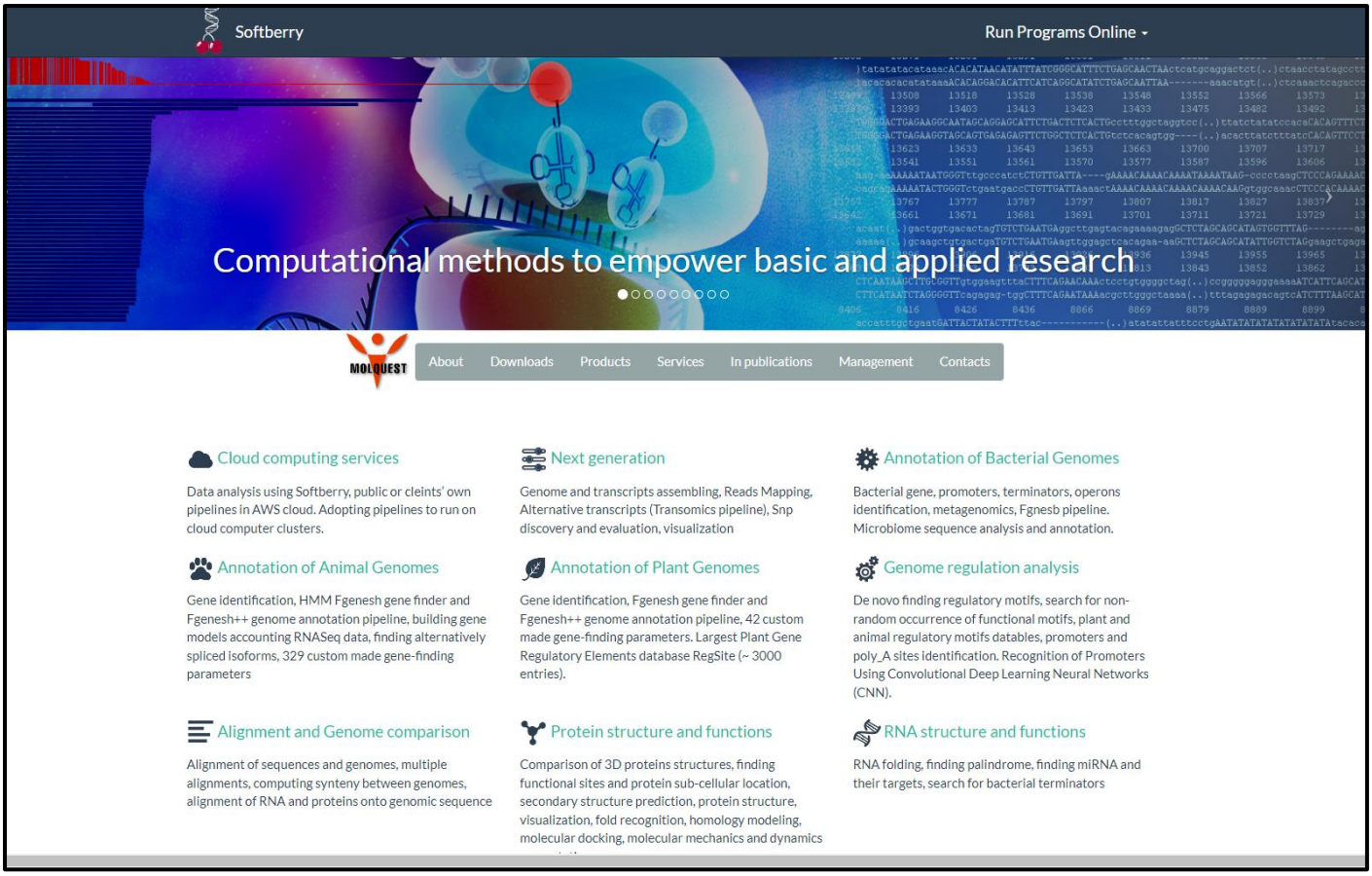

Fig3. Homepage for Softberry

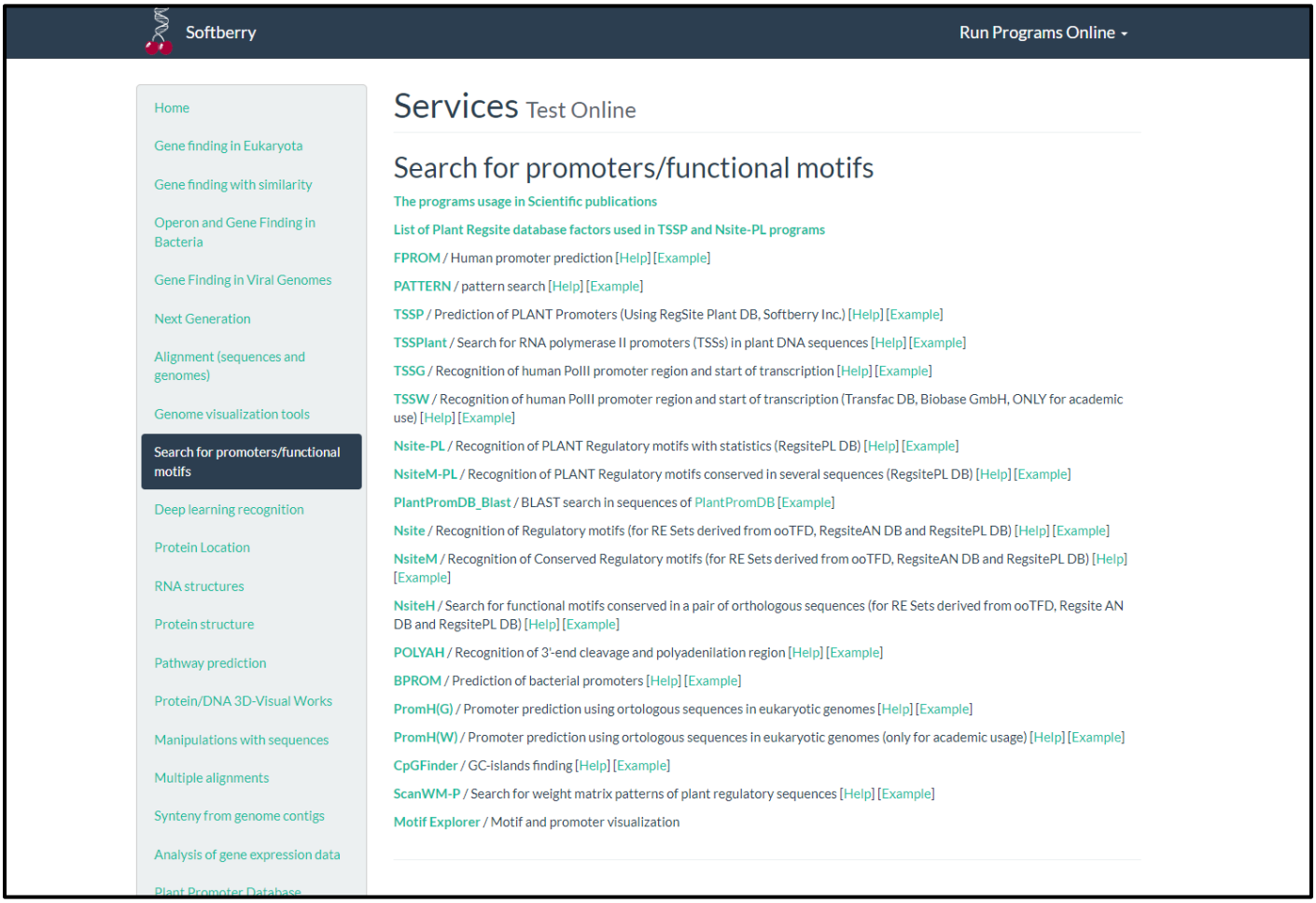

Fig4. Tools for promoters/functional motifs

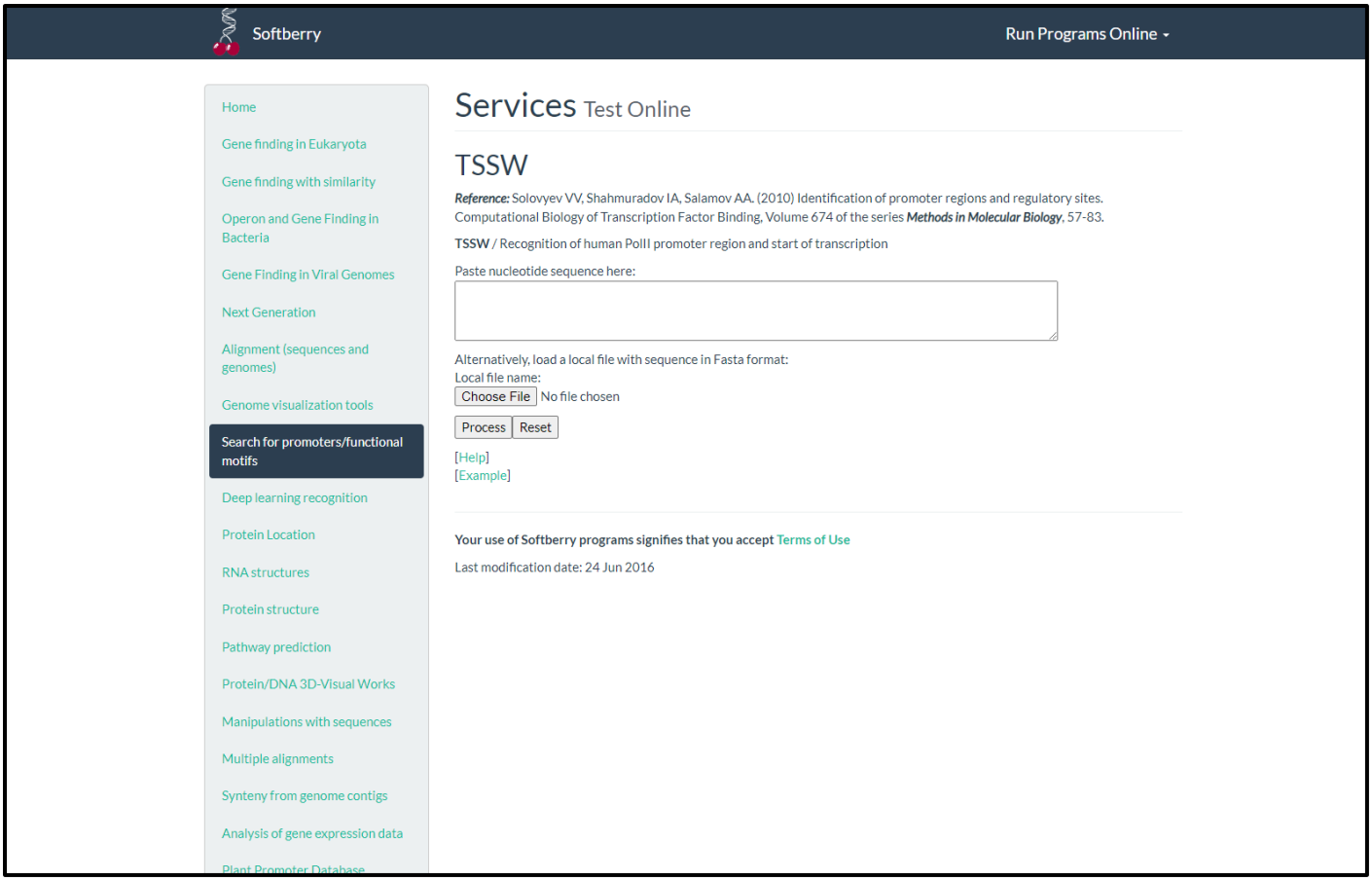

Fig5. Homepage for TSSW

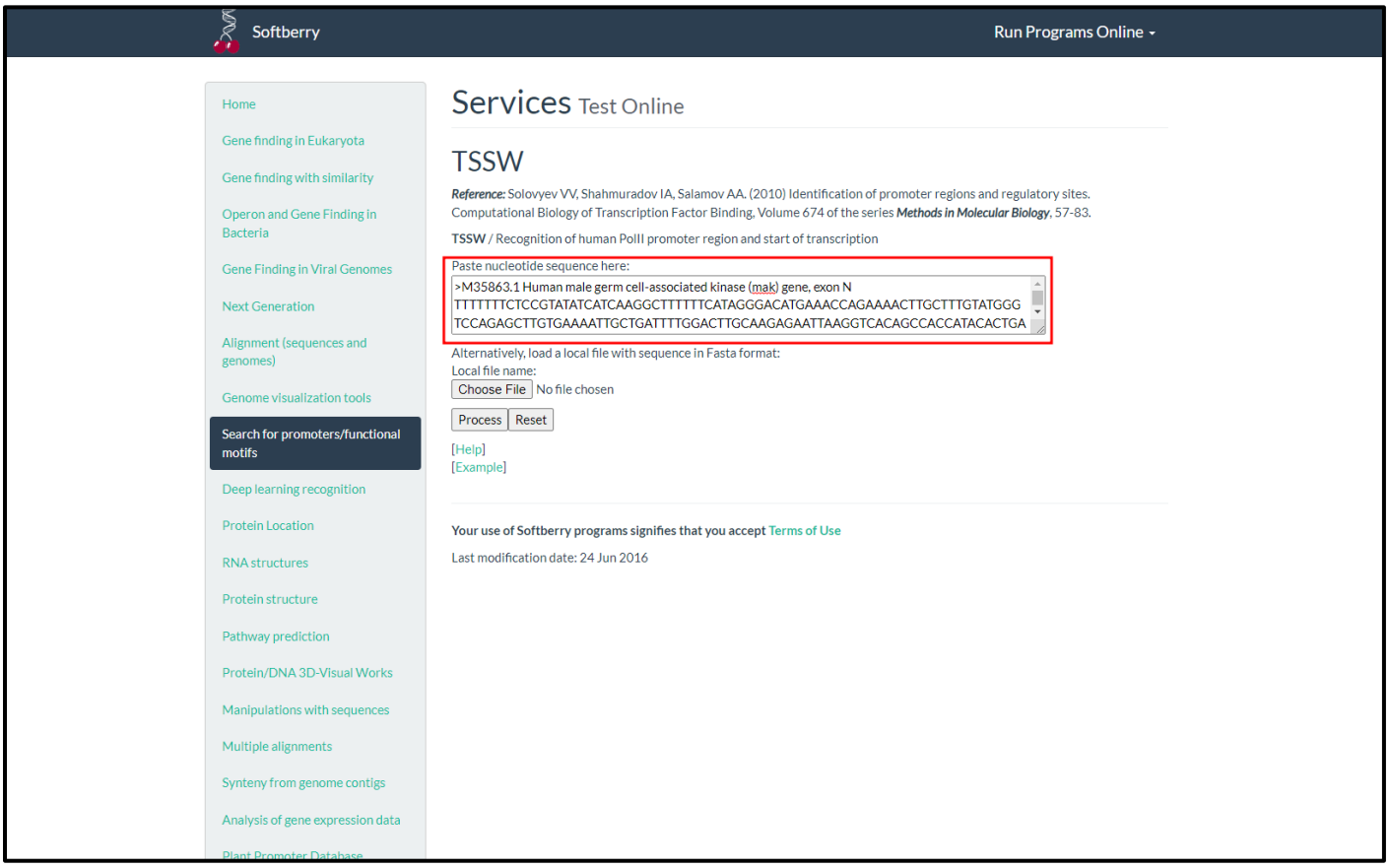

Fig6. Search for protease nucleotide FASTA sequence

```
>L29284.2 Saccharomyces cerevisiae kinase (RIM11) gene, complete cds
 Length of sequence-      1920
 Thresholds for TATA+ promoters -  0.45, for TATA-/enhancers -  3.70
    3 promoter/enhancer(s) are predicted
 Enhancer Pos:   1391 LDF-  5.79
 Promoter Pos:   1390 LDF-  5.75
 Promoter Pos:   1870 LDF-  0.84 TATA box at   1841    21.05
 Transcription factor binding sites:
for promoter at position -    1391
  1187 (+) HS$BAC_03    CCAAT
  1281 (+) HS$BAC_03    CCAAT
  1316 (+) HS$BAC_03    CCAAT
  1181 (-) HS$BAC_03    CCAAT
  1289 (-) CHICK$BAC_   TATAA
  1127 (-) CHICK$BAC_   TATAA
  1203 (-) Y$ADH2_01    TCTCC
  1185 (+) RAT$ALBU_2   AACCAAT
  1314 (+) RAT$ALBU_2   AACCAAT
  1327 (+) Y$CFES_01    GTCACGTG
  1177 (+) MOUSE$A21C   ATTGG
  1320 (-) MOUSE$A21C   ATTGG
  1285 (-) MOUSE$A21C   ATTGG
  1191 (-) MOUSE$A21C   ATTGG
  1164 (+) MOUSE$A21C   gcccagccctcccATTGGtggagacg
  1333 (-) MOUSE$A21C   gcccagccctcccATTGGtggagacg
  1298 (-) MOUSE$A21C   gcccagccctcccATTGGtggagacg
  1204 (-) MOUSE$A21C   gcccagccctcccATTGGtggagacg
  1201 (-) Y$CYC1_09    ctcatttggcgagcGTTGGt
  1146 (-) Y$CYC1_09    ctcatttggcgagcGTTGGt
  1297 (-) AD$E3_06     gggcagggTATAActcacctga
  1135 (-) AD$E3_06     gggcagggTATAActcacctga
  1380 (-) AD$E4_16     ACGTCA
  1215 (-) HS$EGFR_15   TCAAT
  1378 (-) RAT$EAI_09   GTCAG
  1283 (+) Y$GAL1_02    AcTTATAT
  1290 (-) Y$GAL1_12    ATATAA
  1110 (-) HS$BG_05     AAGATTAG
  1239 (-) MOUSE$BMG_   cagtagtTGATTgagca
  1099 (+) MOUSE$BMG_   agggccAATCTgctc
  1231 (+) MOUSE$BMG_   agggccAATCTgctc
  1187 (+) HS$GG_17     CCAATag
  1281 (+) HS$GG_17     CCAATag
  1316 (+) HS$GG_17     CCAATag
  1181 (-) HS$GG_17     CCAATag
  1286 (+) RAT$GLU_04   TATAT
  1231 (+) HS$GMCSF_0   CATTA
  1165 (-) HS$GMCSF_0   CATTA
  1389 (-) HS$GMCSF_0   TATTT
  1340 (-) HS$HO_01     ctggccCACGTGACccgc
  1231 (-) HS$HH4_01    GATTTC
  1146 (+) HS$HH4_02    GGTCC
  1335 (-) HA$HMGCR_0   gttCGTGACgtaggccg
  1137 (-) HS$IGH_04    ATTTGccgT
  1229 (-) HS$IGKL_01   TTTCCA
  1223 (+) RAT$INS_01   GTGGAAA
```

Fig7. Result for recognised promoter regions

RESULT:

Nucleotide FASTA sequence for Homo sapiens neutral protease of length 3298bps was submitted. With LDF threshold of 0.45 for promoters and 3.70 for enhancers, 3 promoters at position 809, 323 and 2884 with 9.40, 6.96 and 3.70 LDF values were recognised. 1 enhancer at position 319 with 7.38 LDF was also recognised.

CONCLUSION:

TSSW is a useful tool for the recognition of human Pol III promoter region and start of transcription. Understanding the regulation of gene expression is an important aspect of understanding the gene function, thus this tool will be useful in elucidating regulation and expression mechanisms of genes and may even shed light on the function of novel and uncharacterized genes with the help of knowledge of promoters.

REFERENCE:

1. Xiong, J. (2008). Promoter and Regulatory Element Prediction. Essential bioinformatics. Cambridge Cambridge University Press. 113-119.
2. PROTEOLYTIC ENZYMES (PROTEASES): Overview, Uses, Side Effects, Precautions, Interactions Dosing and Reviews. (n.d.). Www.webmd.com. Retrieved March 18, 2022, from https://www.webmd.com/vitamins/ai/ingredientmono-1623/proteolytic-enzymes-proteases#:~:text=Proteolytic%20enzymes%20(proteases)%20are%20enzymes
3. Homo sapiens neutral protease alpha subunit gene, complete cds. (2016). NCBI Nucleotide. Retrieved March 18, 2022, from https://www.ncbi.nlm.nih.gov/nuccore/AH001431.2?report=genbank
4. Softberry Home Page. (n.d.). Www.softberry.com. Retrieved March 18, 2022, from http://www.softberry.com/
5. TSSW - Recognition of human PolII promoter region and start of transcription. (n.d.). Www.softberry.com. Retrieved March 18, 2022, from http://www.softberry.com/berry.phtml?topic=tssw&group=programs&subgroup=promoter
6. Softberry - TSSW result. (n.d.). Www.softberry.com. Retrieved March 18, 2022, from http://www.softberry.com/cgi-bin/programs/promoter/tssw.pl

DATE: 19-03-22

WEBLEM 8b

BPROM

(URL: http://www.softberry.com/berry.phtml?topic=bprom&group=programs&subgroup=gfindb)

AIM:

To predict bacterial promoter for Enterococcus Alcedinis using BPROM tool.

INTRODUCTION:

Two Gram-positive, catalase-negative bacterial strains were isolated from the cloaca of common kingfishers (Alcedo atthis). Repetitive sequence-based PCR fingerprinting using the (GTG)5 primer grouped these isolates into a single cluster separated from all known enterococcal species. The two strains revealed identical 16S rRNA gene sequences placing them within the genus Enterococcus with Enterococcus aquimarinus LMG 16607(T) as the closest relative (97.14 % similarity).

BPROM is a web-based program for prediction of bacterial promoters. It uses a linear discriminant function combined with signal and content information such as consensus promoter sequence and oligonucleotide composition of the promoter sites. This program first predicts a given sequence for bacterial operon structures by using an intergenic distance of 100 bp as basis for distinguishing genes to be in an operon. This rule is more arbitrary than the Wang et al. rule, leading to high rates of false positives. Once the operons are assigned, the program is able to predict putative promoter sequences. Because most bacterial promoters are located within 200 bp of the protein coding region, the program is most effectively used when about 200 bp of upstream sequence of the first gene of an operon is supplied as input to increase specificity.

METHODOLOGY:

1. Open homepage for softberry. (URL: http://www.softberry.com/)
2. Under operon and gene finding select BPROM. (URL:http://www.softberry.com/berry.phtml?topic=bprom&group=programs&subgroup=gfindb)
3. Retrieve bacterial nucleotide FASTA sequence from GenBank.
4. Process the FASTA sequence on BPROM.
5. Observe and interpret the results.

OBSERVATION:

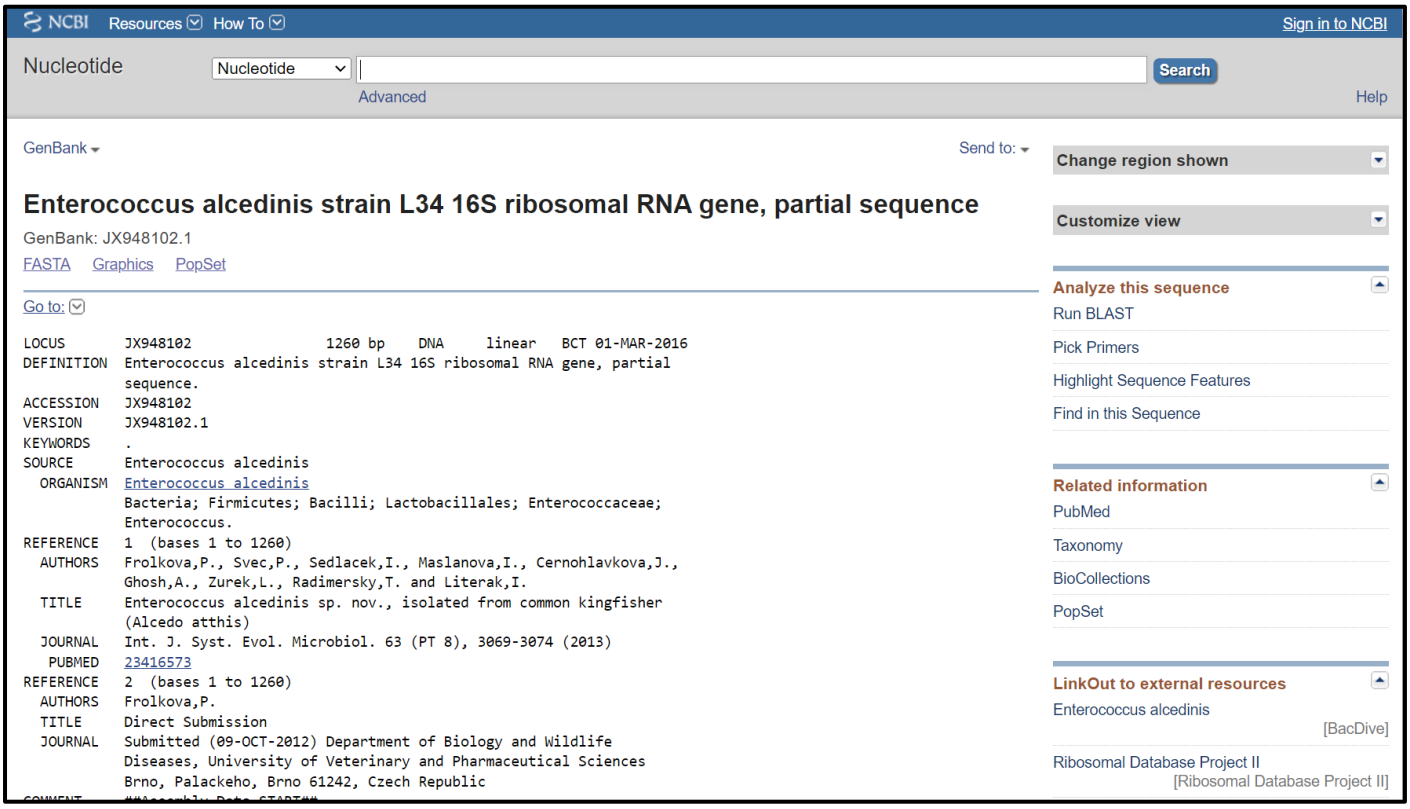

Fig1. GenBank result for Neisseria gonorrhoeae

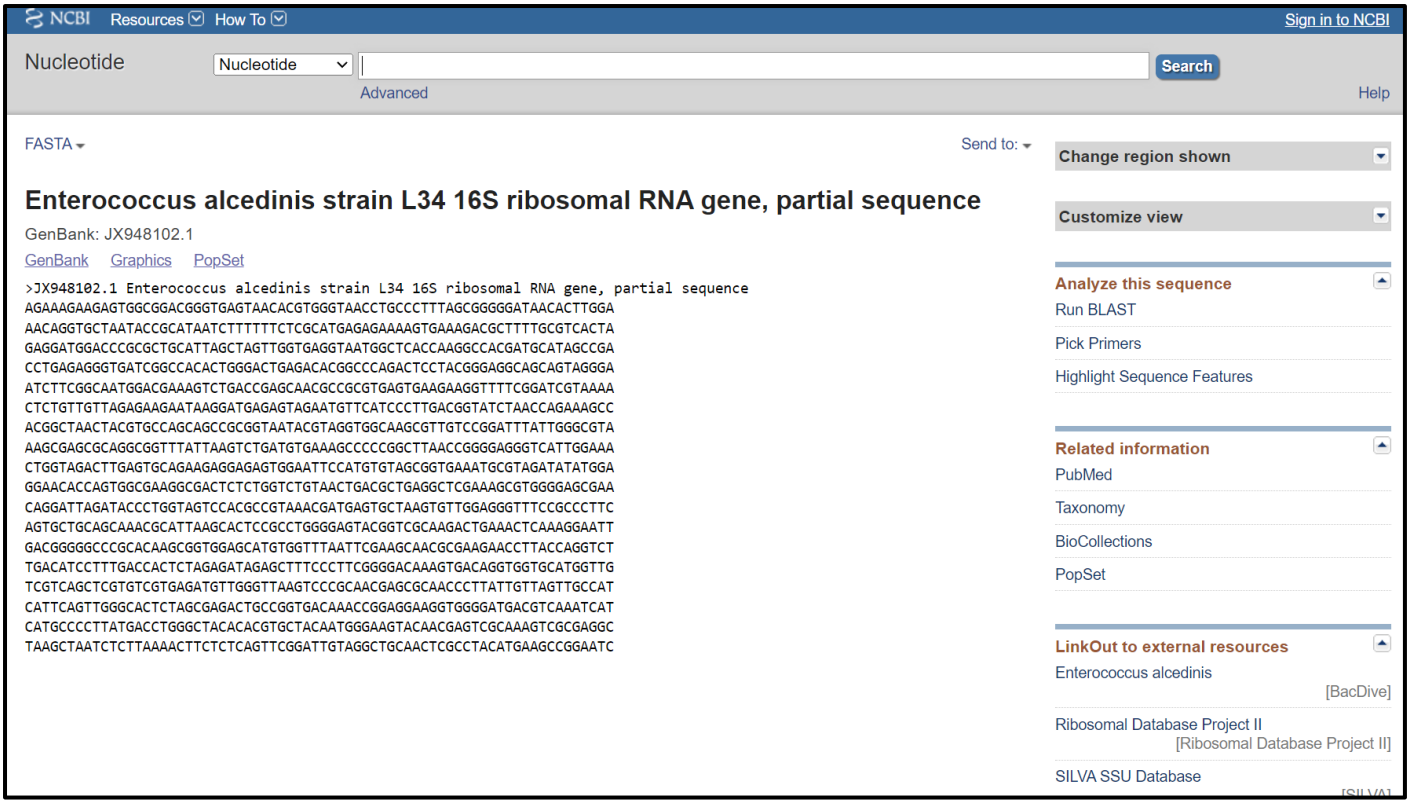

Fig2. Nucleotide FASTA sequence

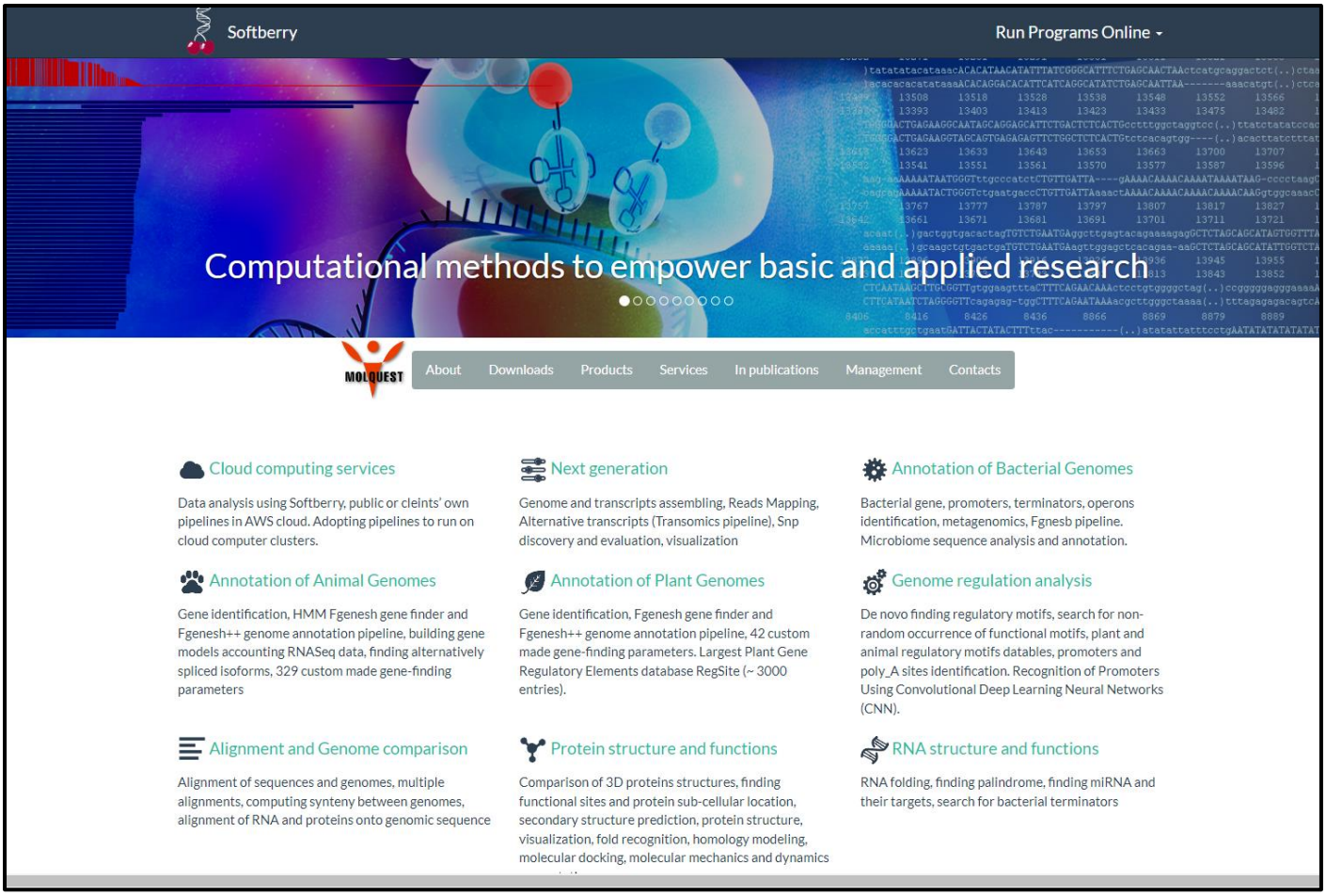

Fig3. Homepage for softberry

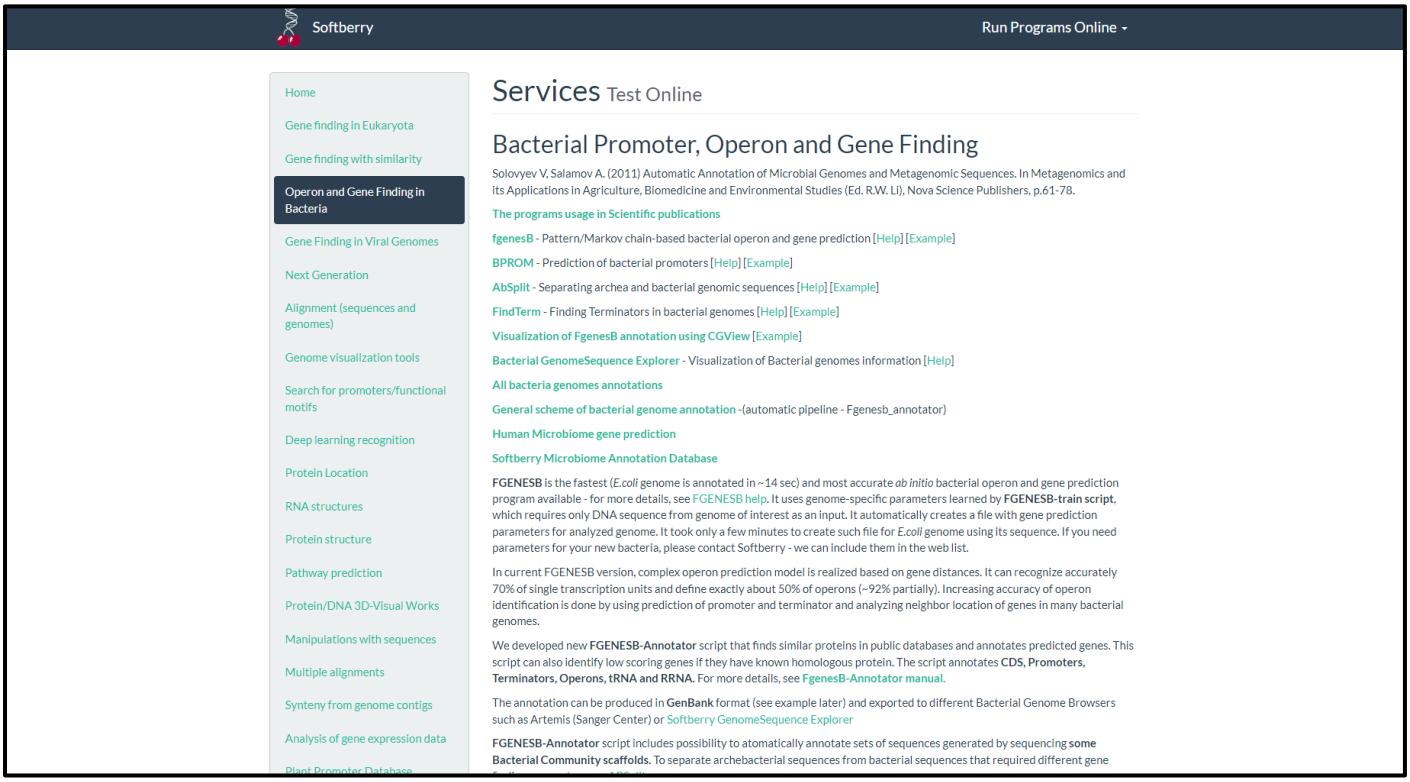

Fig4. Tools for bacterial promoter, operon and gene finding

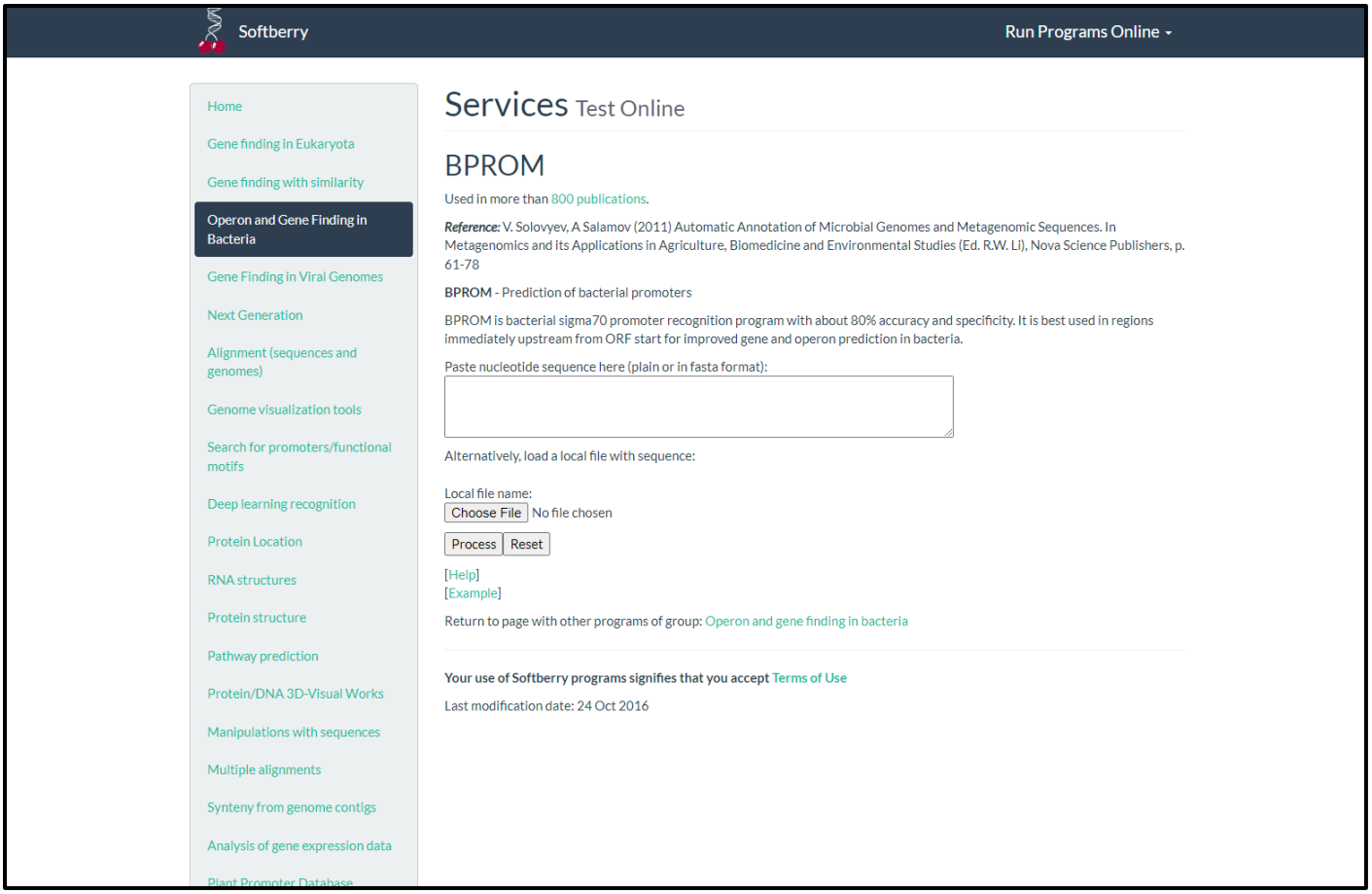

Fig5. Homepage for BPROM

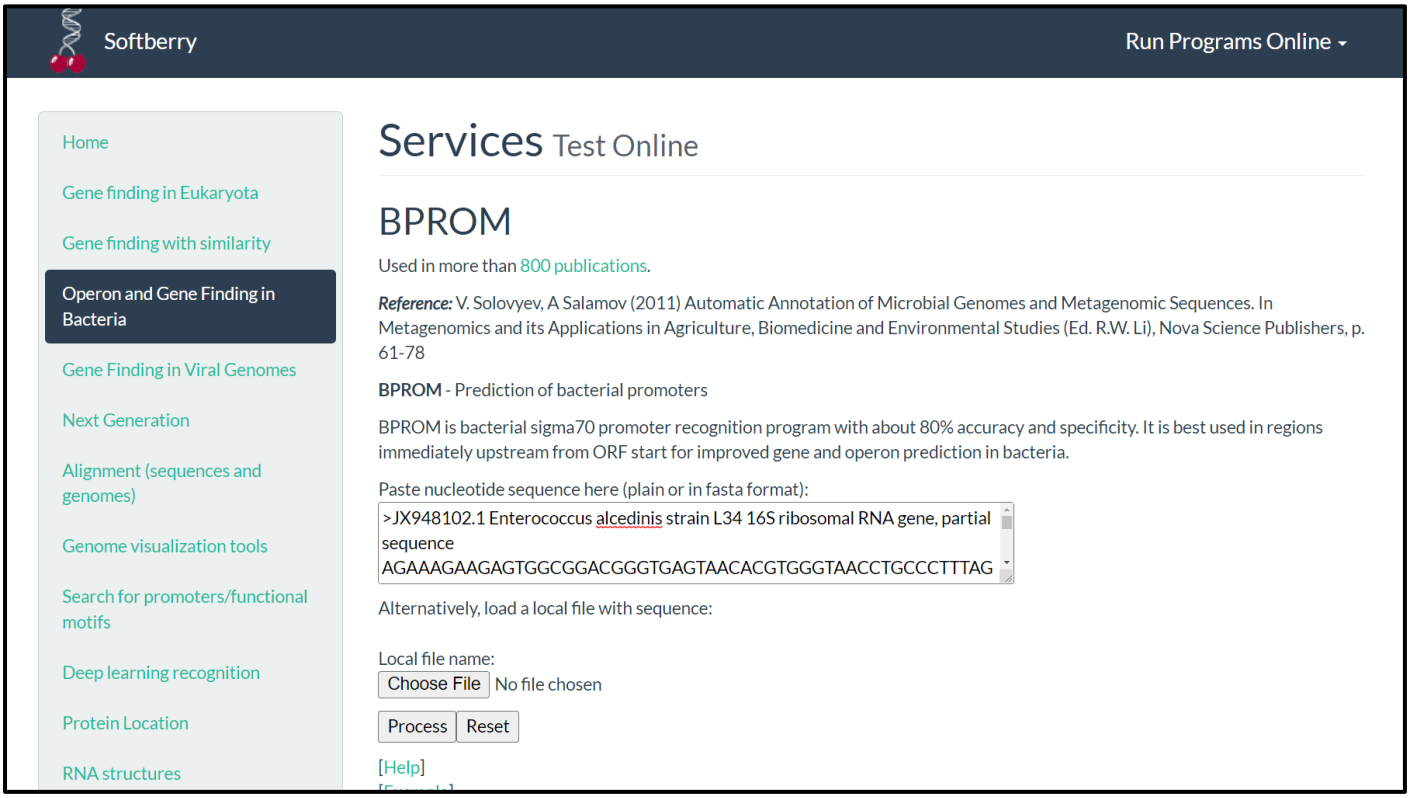

Fig6. Homepage for nucleotide FASTA sequence

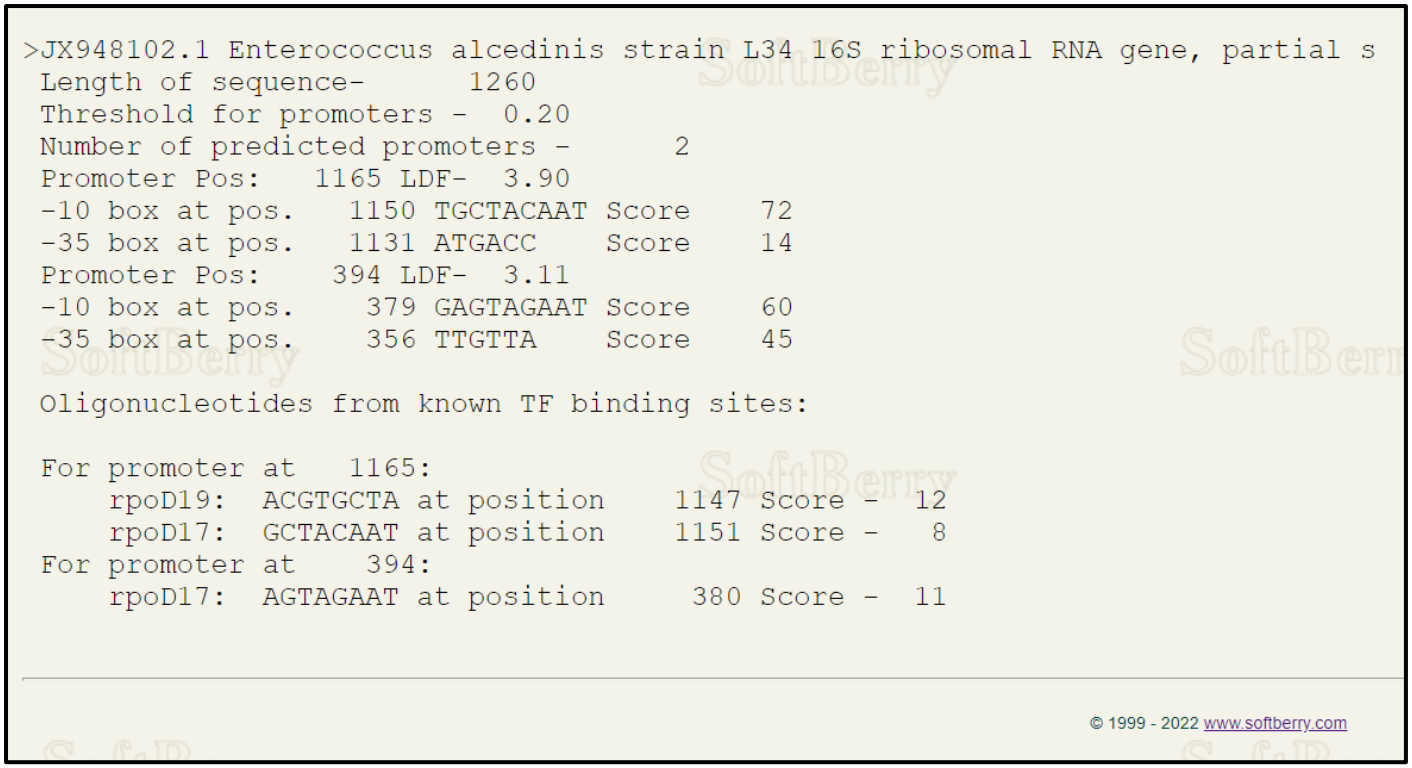

Fig7. Result for predicted promoters and known TF binding sites

RESULT:

Nucleotide FASTA sequence for Enterococcus alcedinis strain L34 16S ribosomal RNA gene, partial sequence with length 1260 was submitted. With threshold LDF value of 0.20, 9 promoters were predicted using BPROM with their TF binding sites.

CONCLUSION:

BPROM is a useful tool for the recognition of bacterial promoter region. Understanding the regulation of gene expression is an important aspect of understanding the gene function, thus this tool will be useful in elucidating regulation and expression mechanisms of genes and may even shed light on the function of novel and uncharacterized genes with the help of knowledge of promoters.

REFERENCES:

1. Xiong, J. (2008). Promoter and Regulatory Element Prediction. Essential bioinformatics. Cambridge: Cambridge University Press. 113-119.
2. Neisseria gonorrhoeae - an overview | ScienceDirect Topics. (n.d.). Www.sciencedirect.com. Retrieved March 18, 2022, from https://www.sciencedirect.com/topics/medicine-and-dentistry/neisseria-gonorrhoeae#:~:text=Neisseria%20gonorrhoeae%20is%20a%20bacterial
3. Neisseria gonorrhoeae strain NG_869 plasmid pNG869_3, whole genome shotgun sequence. (2022).NCBI Nucleotide. Retrieved March 18, 2022, fromhttps://www.ncbi.nlm.nih.gov/nuccore/NZ_CM003348.1
4. Softberry Home Page. (n.d.). Www.softberry.com. Retrieved March 18, 2022, fromhttp://www.softberry.com/
5. BPROM - Prediction of bacterial promoters. (n.d.). Www.softberry.com. Retrieved March 18, 2022,from http://www.softberry.com/berry.phtml?topic=bprom&group=programs&subgroup=gfindb
6. Softberry - BPROM result. (n.d.). Www.softberry.com. Retrieved March 18, 2022, from http://www.softberry.com/cgi-bin/programs/gfindb/bprom.pl

DATE: 19-03-22

WEBLEM 8c

FGENESB

(URL: http://www.softberry.com/berry.phtml?topic=fgenesb&group=programs&subgroup=gfindb)

AIM:

To predict bacterial operon and gene for *Enterococcus alcedinis* using FGENESB tool.

INTRODUCTION:

Two Gram-positive, catalase-negative bacterial strains were isolated from the cloaca of common kingfishers (Alcedo atthis). Repetitive sequence-based PCR fingerprinting using the (GTG)5 primer grouped these isolates into a single cluster separated from all known enterococcal species. The two strains revealed identical 16S rRNA gene sequences placing them within the genus Enterococcus with Enterococcus aquimarinus LMG 16607(T) as the closest relative (97.14 % similarity).

FGENESB is a web-based program that is also based on fifth-order HMMs for detecting coding regions. The program is specifically trained for bacterial sequences. It uses the Vertibi algorithm to find an optimal match for the query sequence with the intrinsic model. A linear discriminant analysis (LDA) is used to further distinguish coding signals from noncoding signals.

METHODOLOGY:

1. Open homepage for softberry. (URL: http://www.softberry.com/)
2. Under operon and gene finding in bacteria select FGENESB. (URL:http://www.softberry.com/berry.phtml?topic=fgenesb&group=programs&subgroup=gfindb)
3. Retrieve bacterial nucleotide FASTA sequence from GenBank.
4. Process the FASTA sequence on FGENESB.
5. Observe and interpret the results.

OBSERVATION:

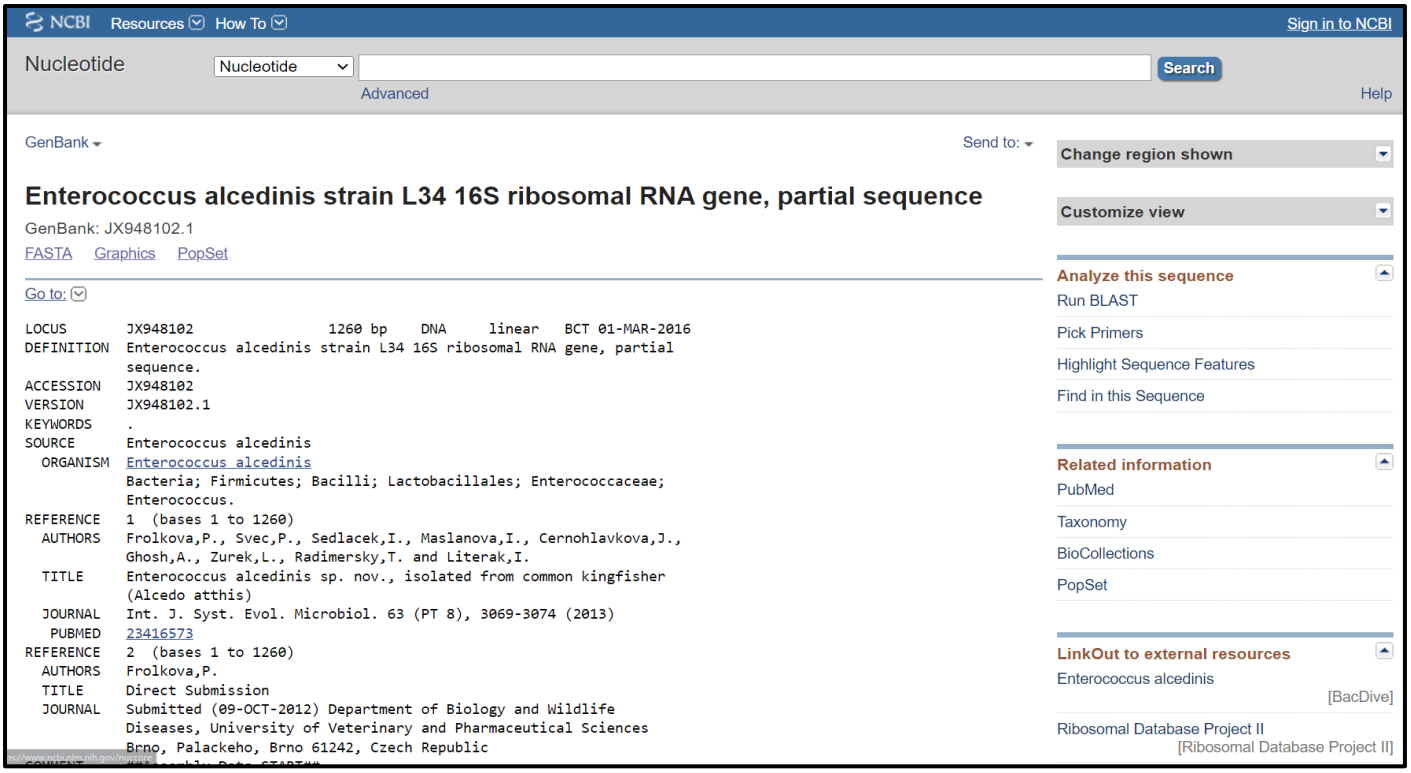

Fig1. GenBank result for *Enterococcus alcedinis*

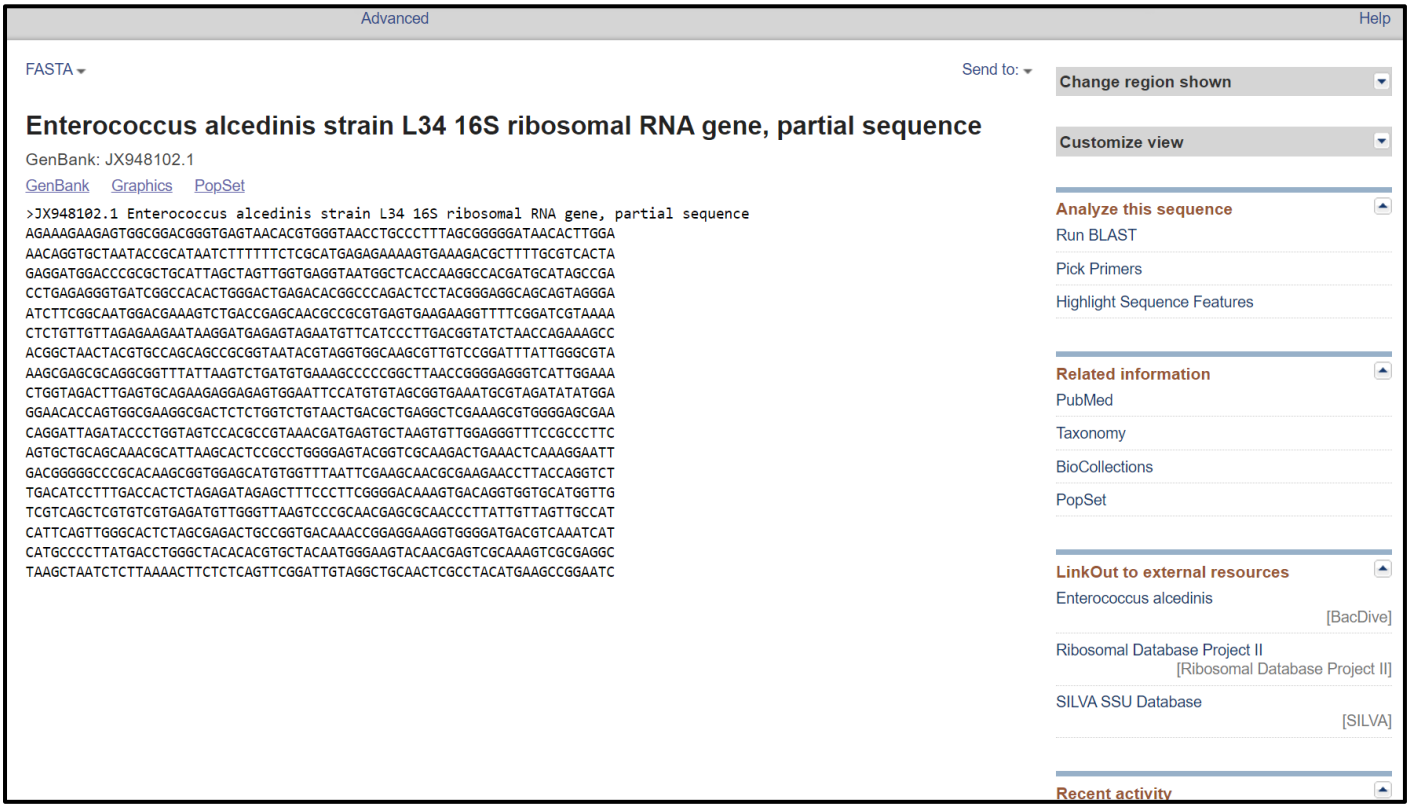

Fig2. Nucleotide FASTA sequence

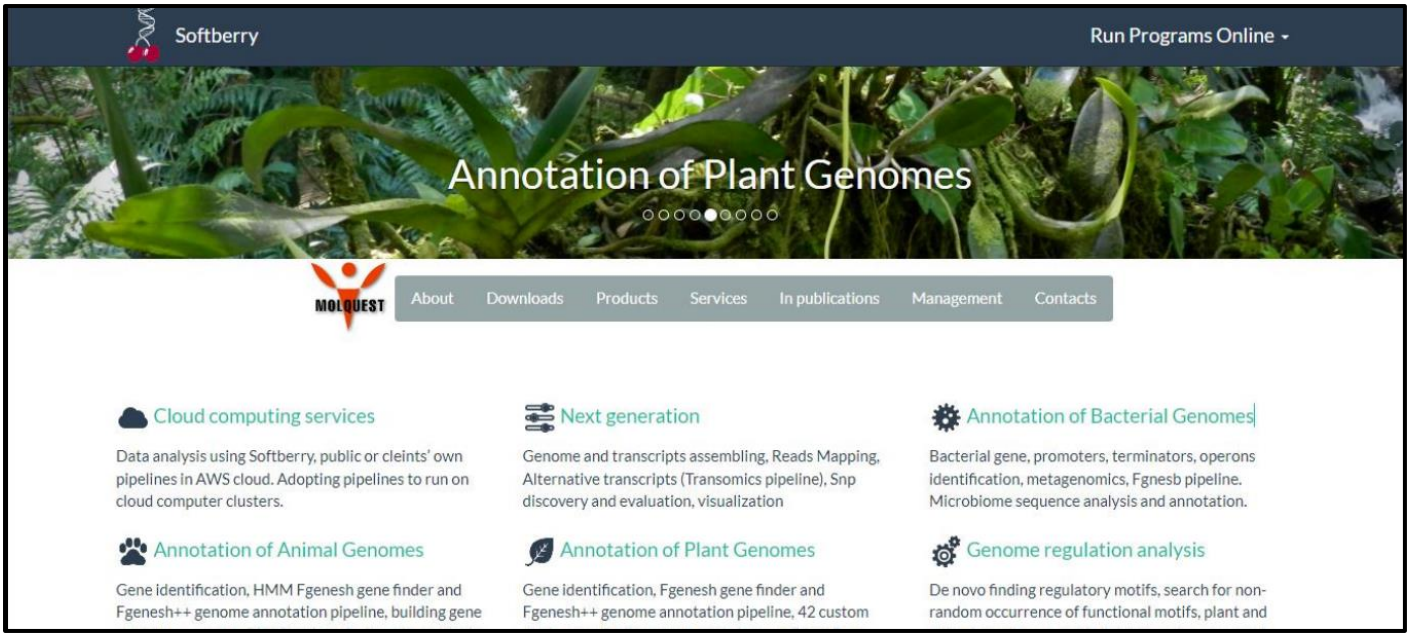

Fig3. Homepage for softberry

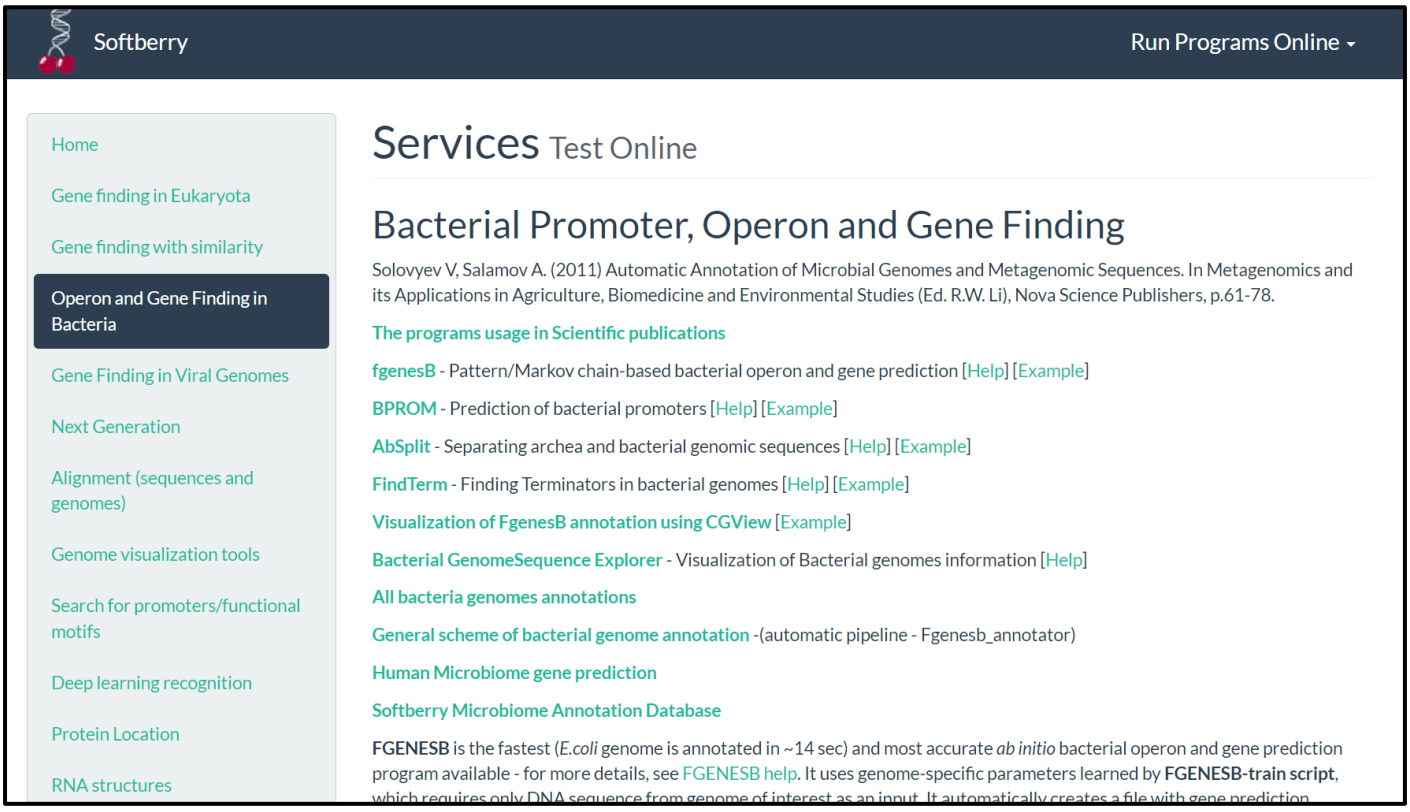

Fig4. Tools for bacterial promoter, operon and gene finding

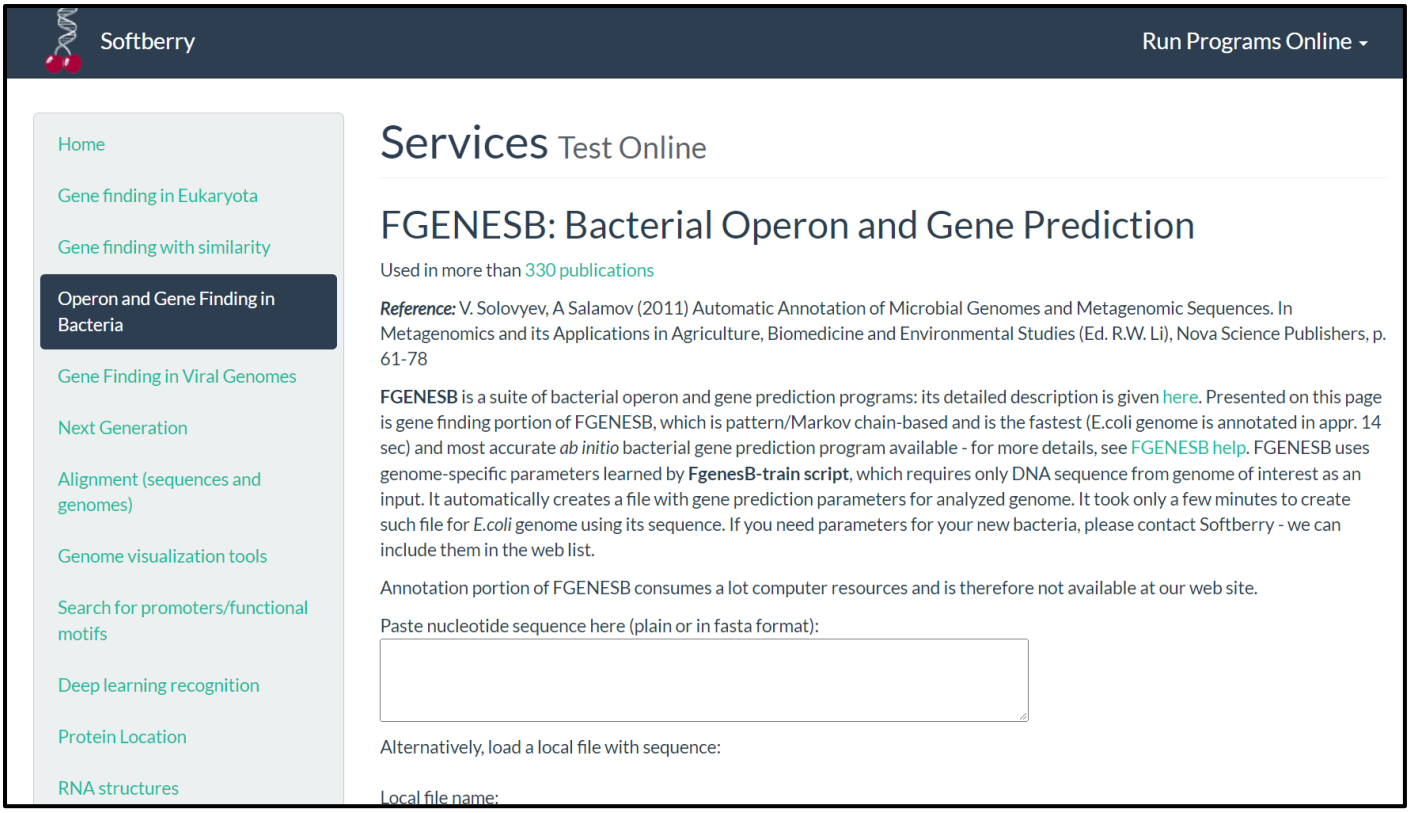

Fig5. Homepage for FGENESB

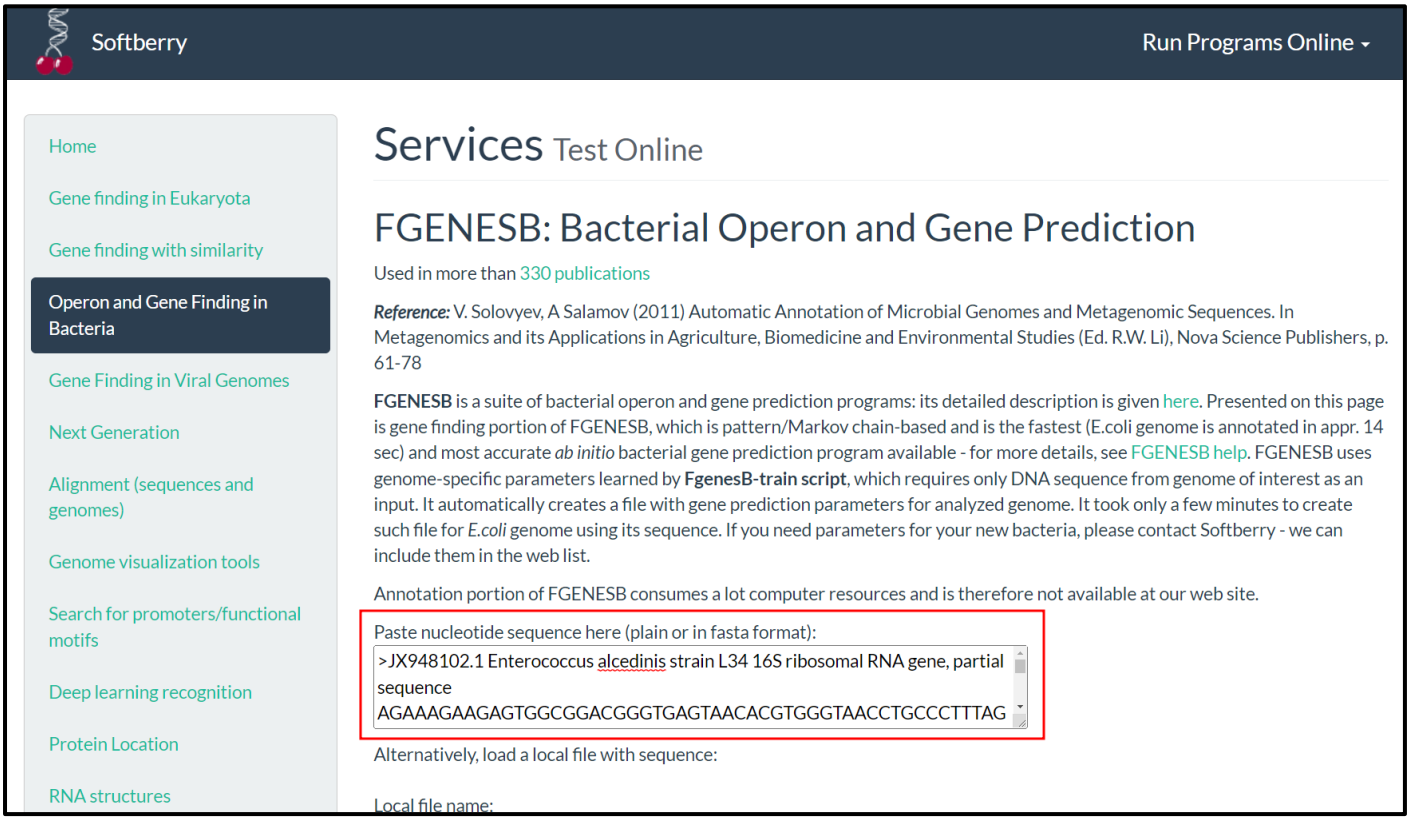

Fig6. Search for nucleotide FASTA sequence

Fig7. Result for predicted bacterial operon and genes

RESULT:

Nucleotide FASTA sequence for Enterococcus alcedinis strain L34 16S ribosomal RNA gene, partial sequence with length of 1260 was submitted and 9 genes, 5 transcriptional units and 2 operons were predicted using FGENESB.

CONCLUSION:

FGENESB tool is useful for prediction of bacterial operon and gene. Gene prediction information is a prerequisite for detailed functional annotation of genes and genomes. Identifying the genes that are grouped together into operons may enhance our knowledge of gene regulation and function, and such information is an important addition to genome annotation. All this can be done with the help of FGENESB.

REFERENCES:

1. Xiong, J. (2008).Gene Prediction. Essential bioinformatics. Cambridge: Cambridge University Press.97111.
2. Neisseria gonorrhoeae - an overview | ScienceDirect Topics. (n.d.). Www.sciencedirect.com.Retrieved March 18, 2022, from https://www.sciencedirect.com/topics/medicine-and-dentistry/neisseria-gonorrhoeae#:~:text=Neisseria%20gonorrhoeae%20is%20a%20bacterial
3. Neisseria gonorrhoeae strain NG_869 plasmid pNG869_3, whole genome shotgun sequence. (2022).NCBI Nucleotide. Retrieved March 18, 2022, from https://www.ncbi.nlm.nih.gov/nuccore/NZ_CM003348.1
4. Softberry Home Page. (n.d.). Www.softberry.com. Retrieved March 18, 2022, from http://www.softberry.com/

5. FGENESB - Bacterial Operon and Gene Prediction. (n.d.). Www.softberry.com. Retrieved March 18, 2022, from http://www.softberry.com/berry.phtml?topic=fgenesb&group=programs&subgroup=gfindb
6. Softberry - fgenesB results. (n.d.). Www.softberry.com. Retrieved March 18, 2022, from http://www.softberry.com/cgi-bin/programs/gfindb/fgenesb.pl

DATE: 19-03-22

WEBLEM 8d

FGENES

(URL: http://www.softberry.com/berry.phtml?topic=fgenes&group=programs&subgroup=gfind)

AIM:

To predict exon signals in Kinase using FGENES tool.

INTRODUCTION:

Kinase, an enzyme that adds phosphate groups (PO_4^{3-}) to other molecules. A large number of kinases exist—the human genome contains at least 500 kinase-encoding genes. Included among these enzymes' targets for phosphate group addition (phosphorylation) are proteins, lipids, and nucleic acids. Saccharomyces cerevisiae kinase and start of transcription can to recognized using FGENES.

FGENES is a web-based program that uses LDA to determine whether a signal is an exon. In addition to FGENES, there are many variants of the program. Some programs, such as FGENESH, make use of HMMs. There are others, such as FGENESH C, that are similarity based. Some programs, such as FGENESH+, combine both ab initio and similarity-based approaches.

METHODOLOGY:

1. Open homepage for softberry. (URL: http://www.softberry.com/)
2. Under Gene for Eukaryotes select FGENES. (URL:http://www.softberry.com/berry.phtml?topic=fgenes&group=programs&subgroup=gfind)
3. Retrieve nucleotide FASTA sequence for protease from GenBank.
4. Process the FASTA sequence on FGENES.
5. Observe and interpret the results.

OBSERVATION:

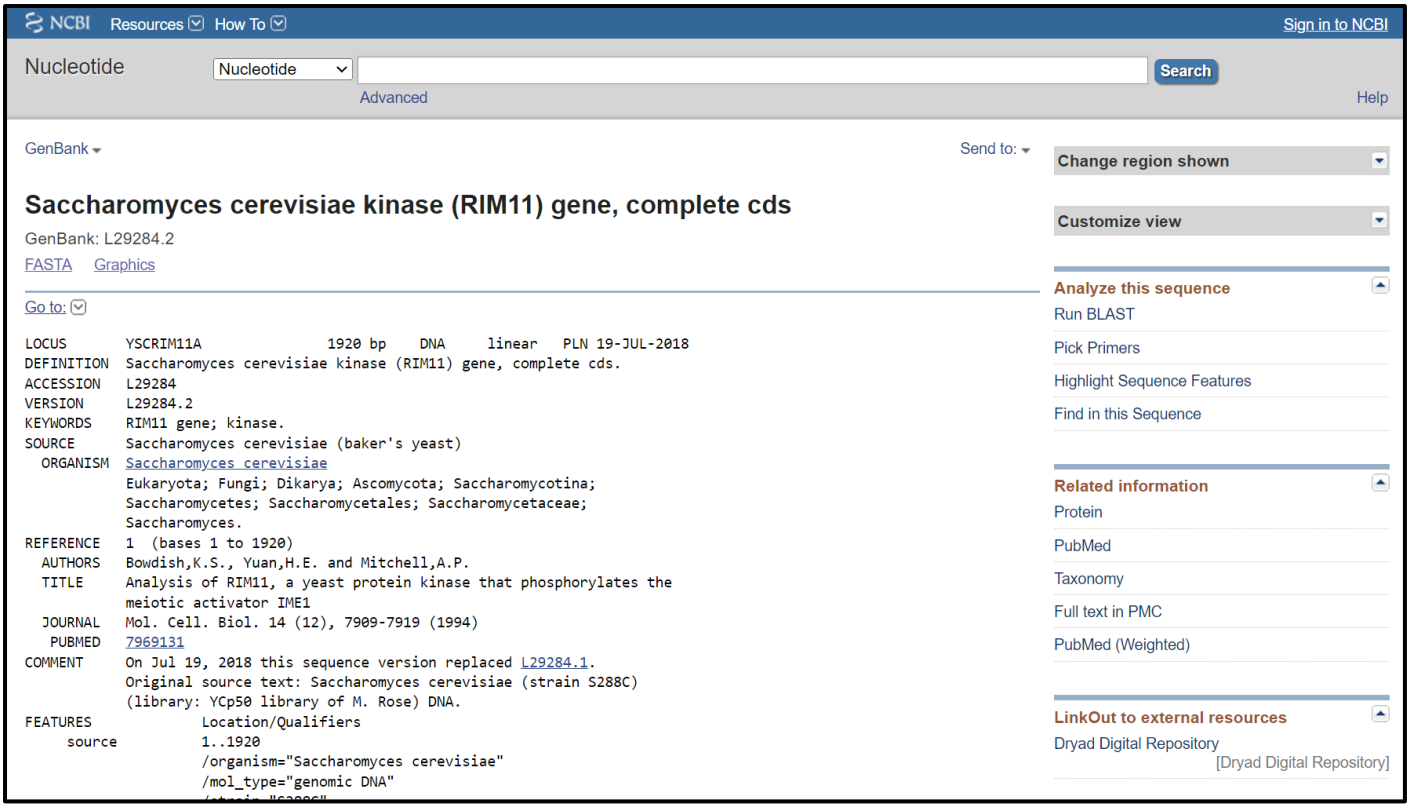

Fig1. GenBank result for Saccharomyces cerevisiae kinase

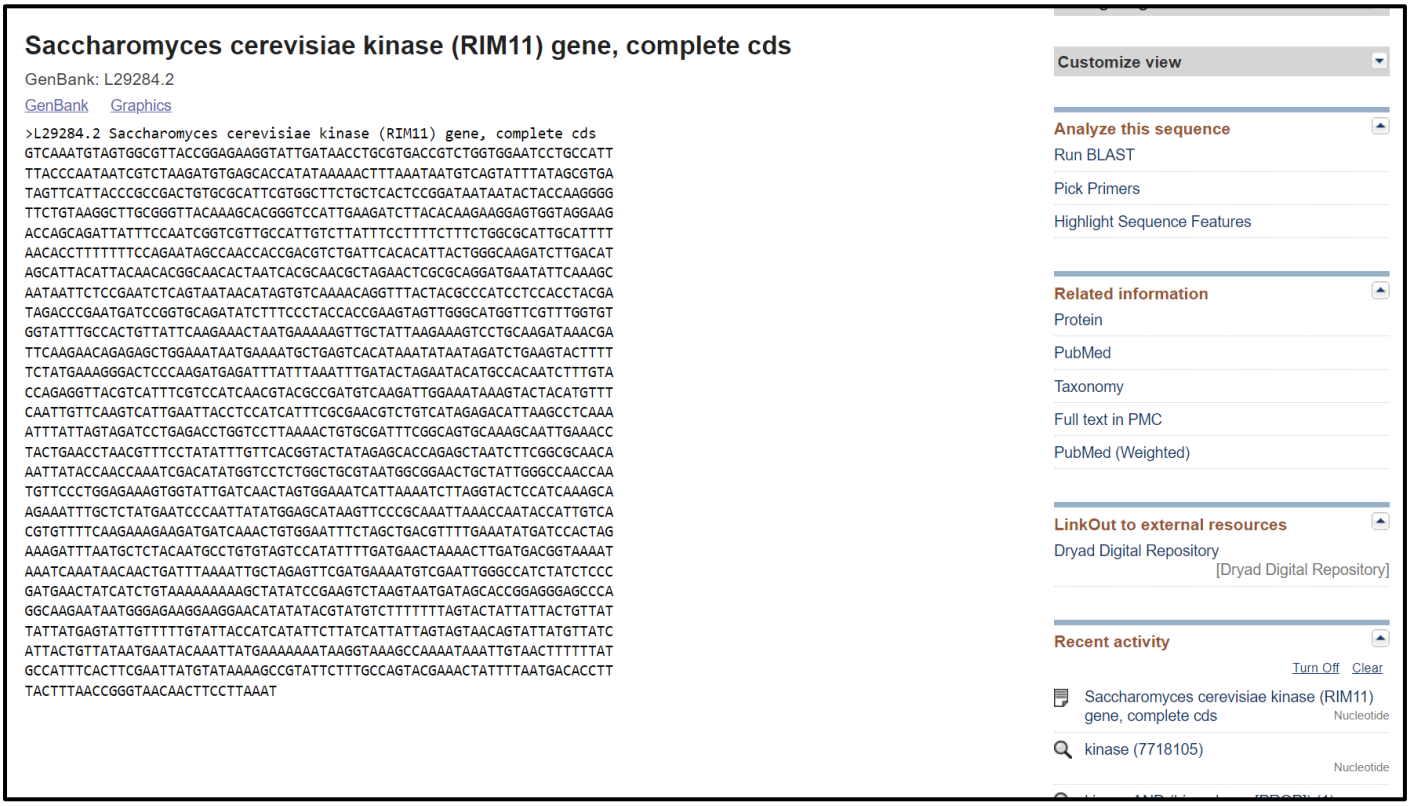

Fig2. Nucleotide FASTA sequence for Saccharomyces cerevisiae kinase

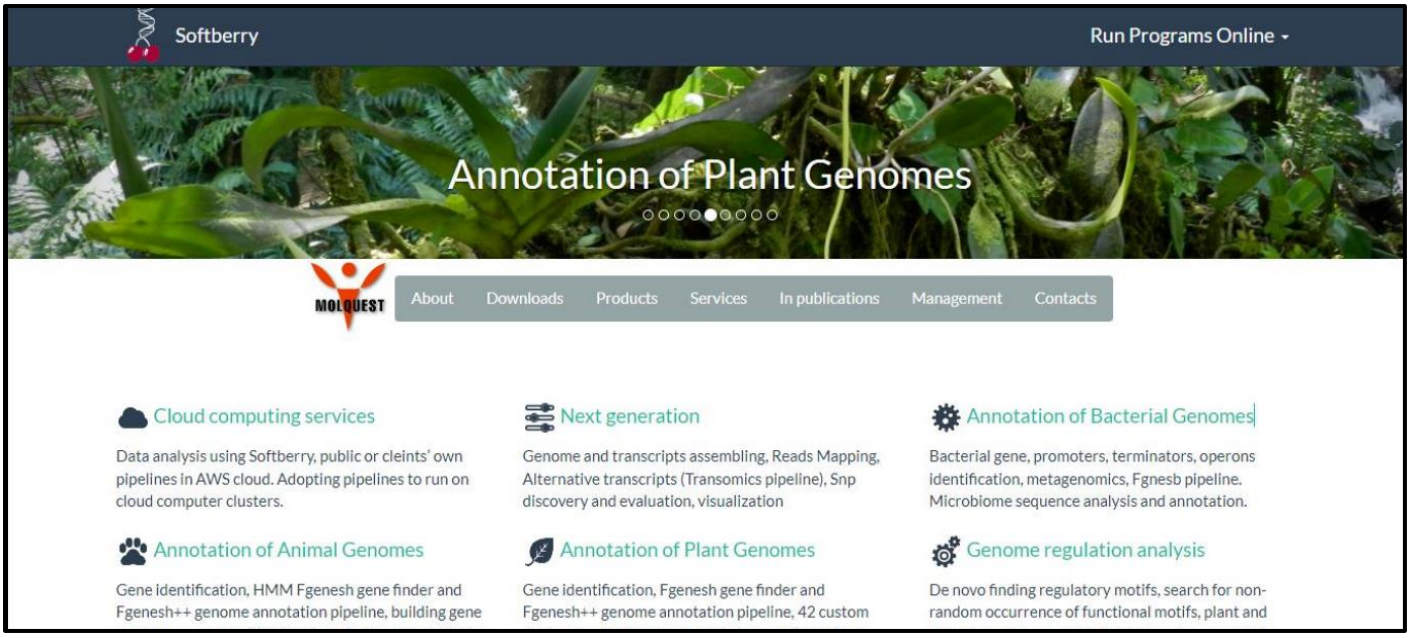

Fig3. Homepage for softberry

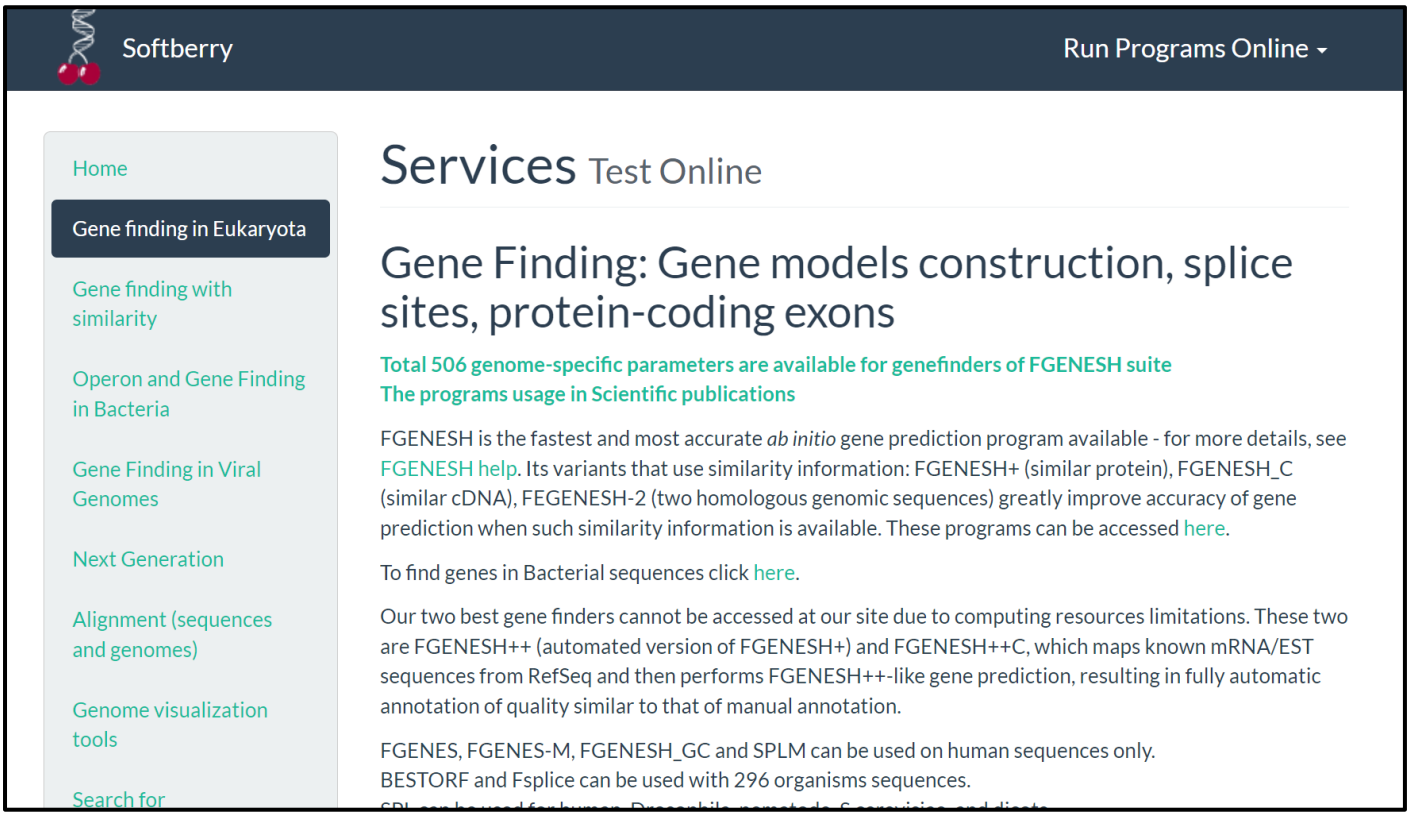

Fig4. Tools for gene finding in Eukaryota

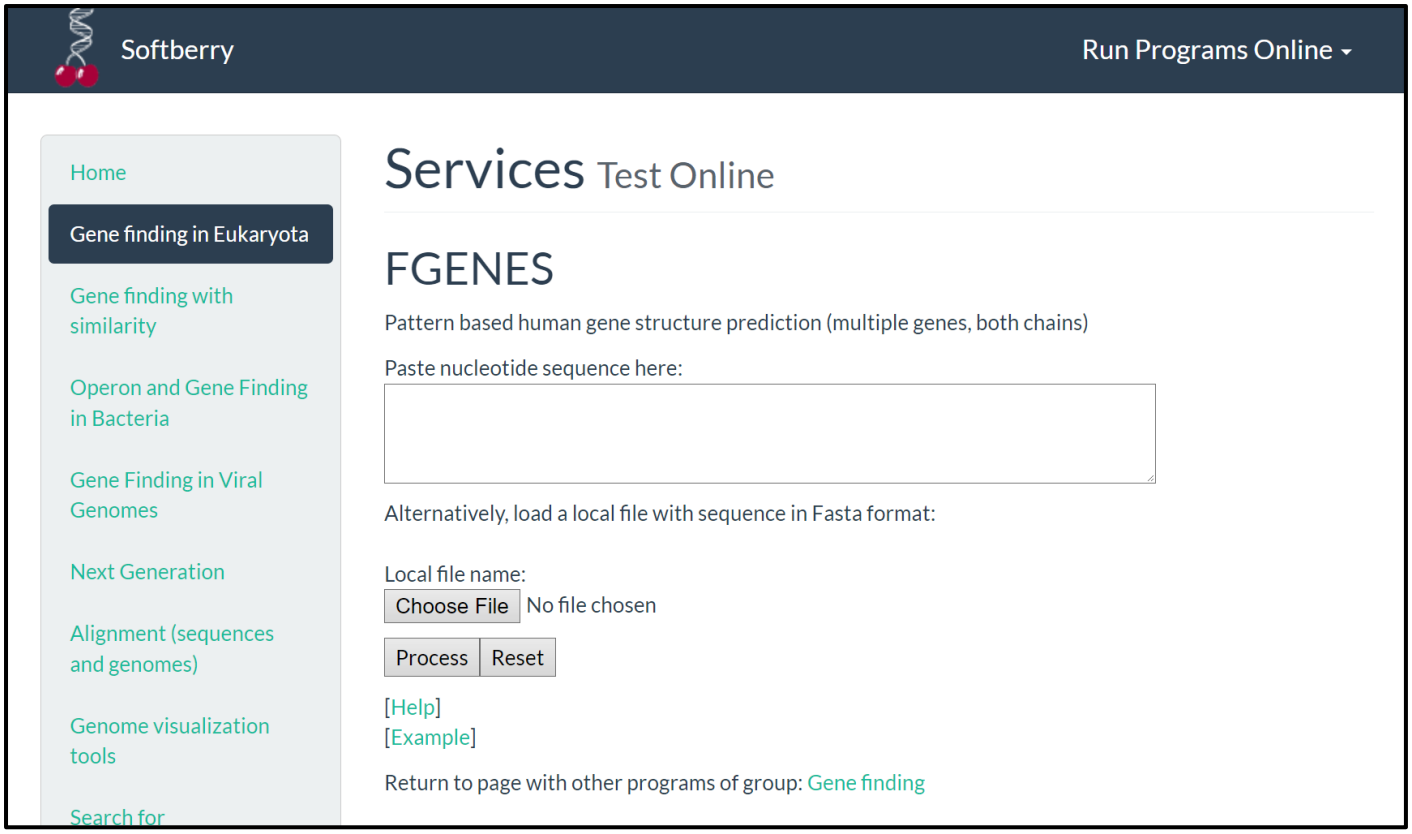

Fig5. Homepage for FGENES

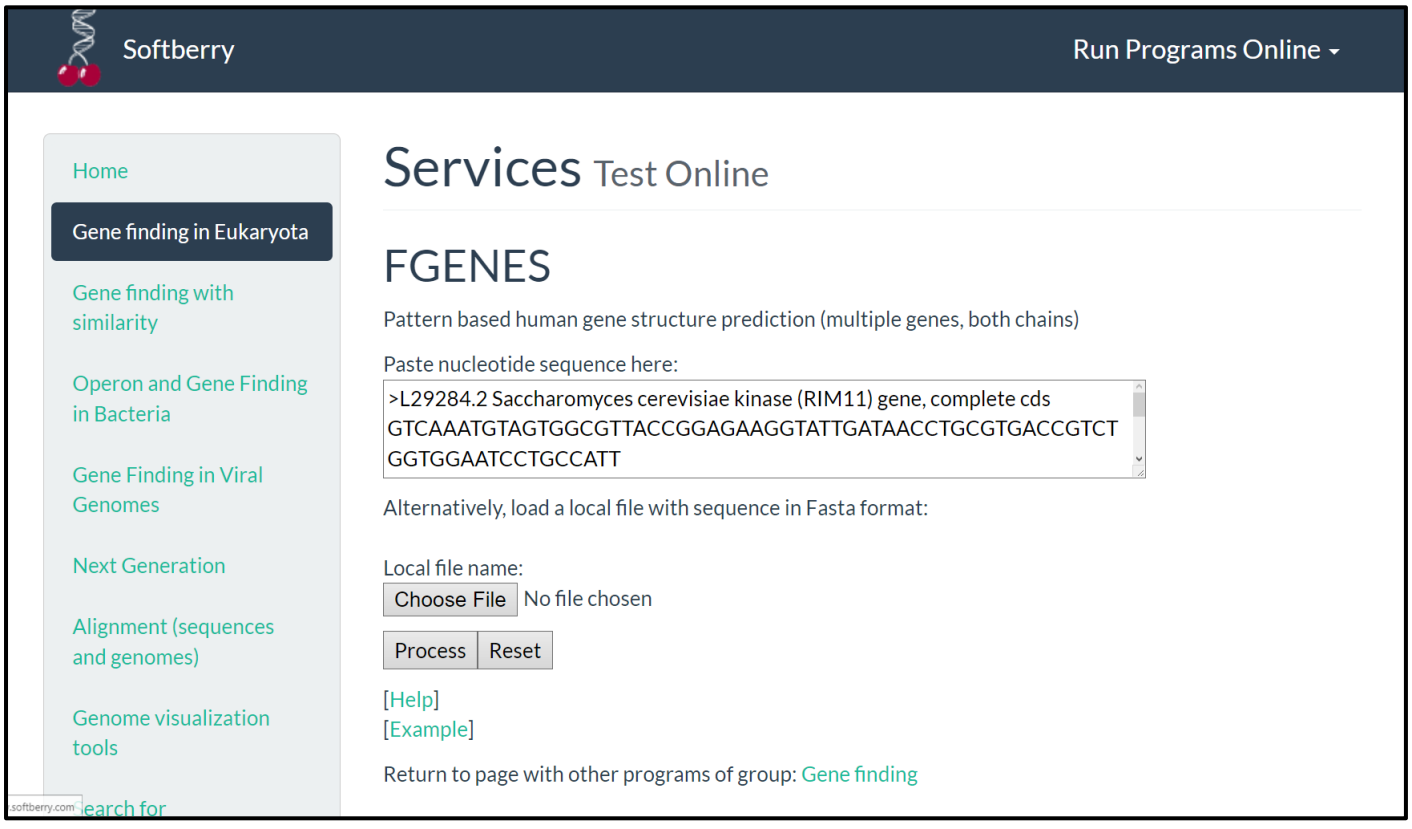

Fig6. Search for kinase nucleotide FASTA sequence

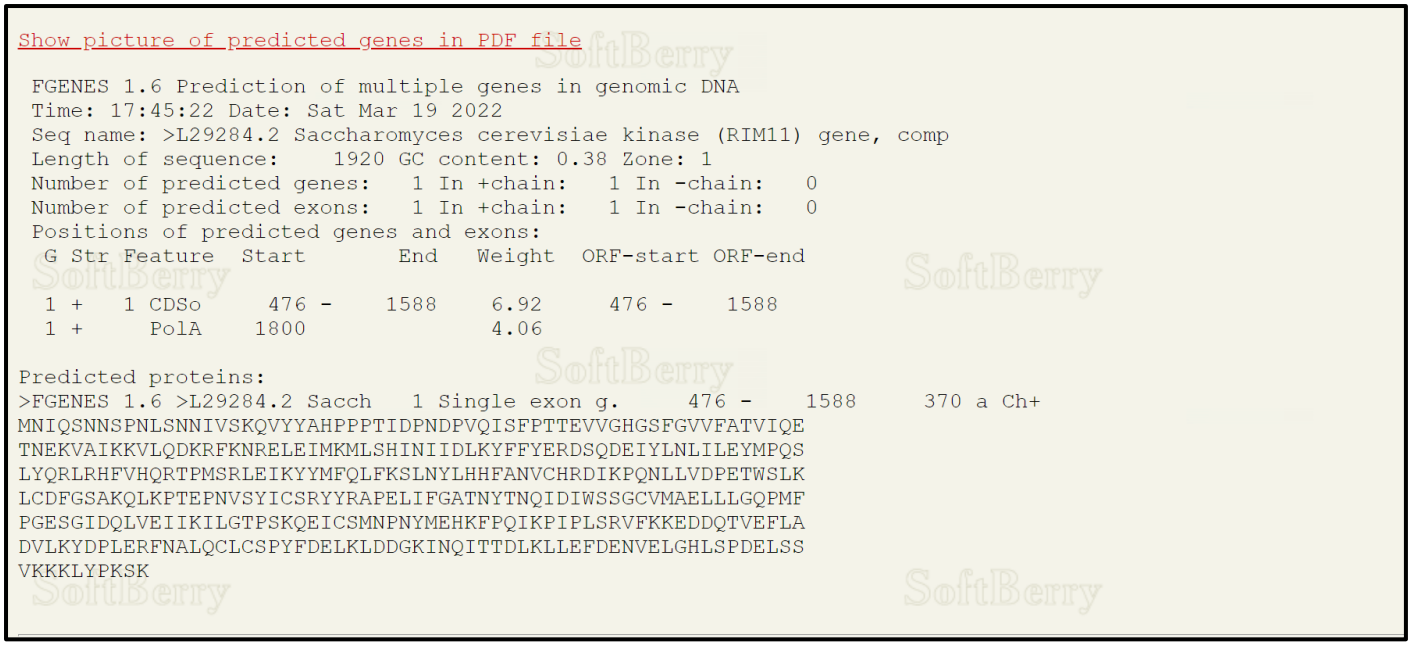

Fig7. Result for predicted exon signals

RESULT:

Nucleotide FASTA sequence Saccharomyces cerevisiae kinase of length 1260 was submitted. With GC content of 0.57, 1 gene and 10 exons were predicted on the + chain. ORF start and end of each exon was predicted.

CONCLUSION:

FGENES tool is useful for prediction of exon signals. Exon prediction information can be used to predict genes and annotate genes and genomes. It can help in undering the gene function.

REFERENCES:

1. Xiong, J. (2008).Gene Prediction. Essential bioinformatics. Cambridge: Cambridge University Press. 97-111.
2. PROTEOLYTIC ENZYMES (PROTEASES): Overview, Uses, Side Effects, Precautions, Interactions, Dosing and Reviews. (n.d.). Www.webmd.com. Retrieved March 18, 2022, from https://www.webmd.com/vitamins/ai/ingredientmono-1623/proteolytic-enzymes-proteases#:~:text=Proteolytic%20enzymes%20(proteases)%20are%20enzymes
3. Homo sapiens neutral protease alpha subunit gene, complete cds. (2016). NCBI Nucleotide. Retrieved March 18, 2022, from https://www.ncbi.nlm.nih.gov/nuccore/AH001431.2?report=genbank
4. Softberry Home Page. (n.d.). Www.softberry.com. Retrieved March 18, 2022, from http://www.softberry.com/
5. FGENES - pattern-based gene structure prediction. (n.d.). Www.softberry.com. Retrieved March 18, 2022, from http://www.softberry.com/berry.phtml?topic=fgenes&group=programs&subgroup=gfind
6. Softberry - FGENES result. (n.d.). Www.softberry.com. Retrieved March 18, 2022, from http://www.softberry.com/cgi-bin/programs/gfind/fgenes.pl

DATE: 19-03-22

WEBLEM 8e

ORF finder-NCBI

(URL: https://www.ncbi.nlm.nih.gov/orffinder/)

AIM:

To search for ORF region in Neisseria gonorrhoeae using ORF finder tool.

INTRODUCTION:

Two Gram-positive, catalase-negative bacterial strains were isolated from the cloaca of common kingfishers (Alcedo atthis). Repetitive sequence-based PCR fingerprinting using the (GTG)5 primer grouped these isolates into a single cluster separated from all known enterococcal species. The two strains revealed identical 16S rRNA gene sequences placing them within the genus Enterococcus with Enterococcus aquimarinus LMG 16607(T) as the closest relative (97.14 % similarity).

ORF finder searches for open reading frames (ORFs) in the DNA sequence you enter. The program returns the range of each ORF, along with its protein translation. Use ORF finder to search newly sequenced DNA for potential protein encoding segments, verify predicted protein using newly developed SMART BLAST or regular BLASTP. This web version of the ORF finder is limited to the subrange of the query sequence up to 50 kb long. Stand-alone version, which doesn't have query sequence length limitation.

METHODOLOGY:

1. Open homepage for ORF finder in NCBI. (URL: https://www.ncbi.nlm.nih.gov/orffinder/)
2. Retrieve nucleotide FASTA sequence for Neisseria gonorrhoeae from GenBank.
3. Submit the FASTA sequence.
4. Observe and interpret the results.

OBSERVATION:

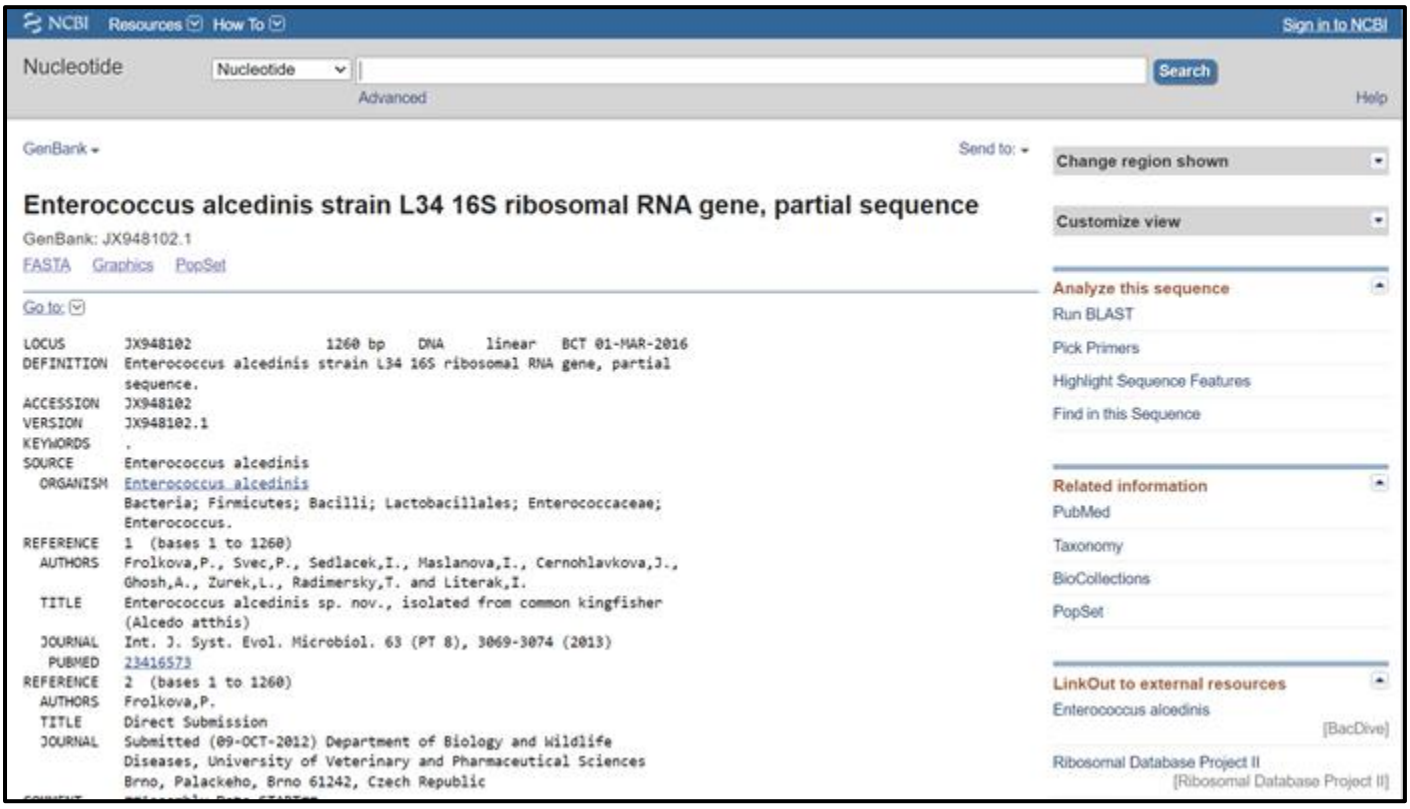

Fig1. Result for Enterococcus alcedinis

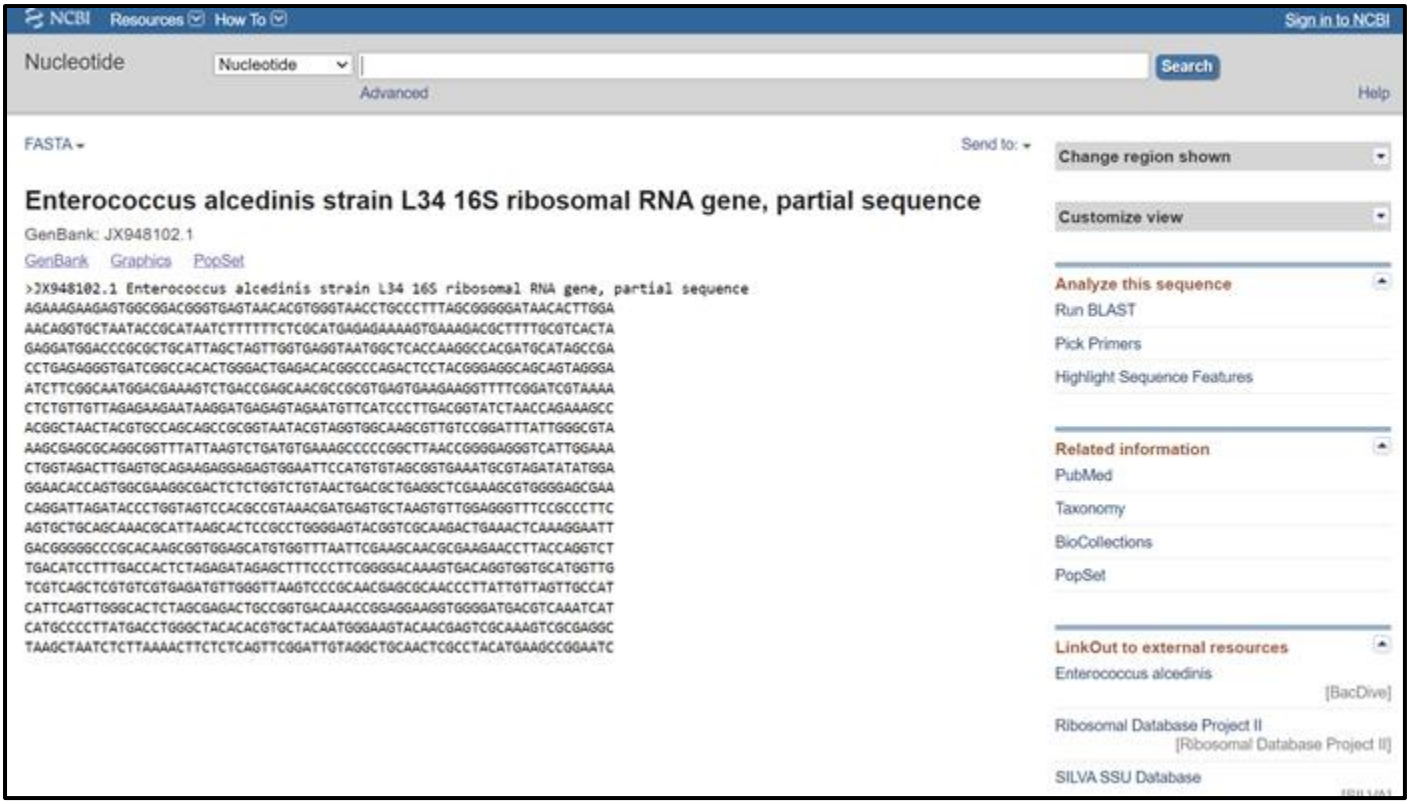

Fig2. Nucleotide FASTA sequence for Enterococcus alcedinis

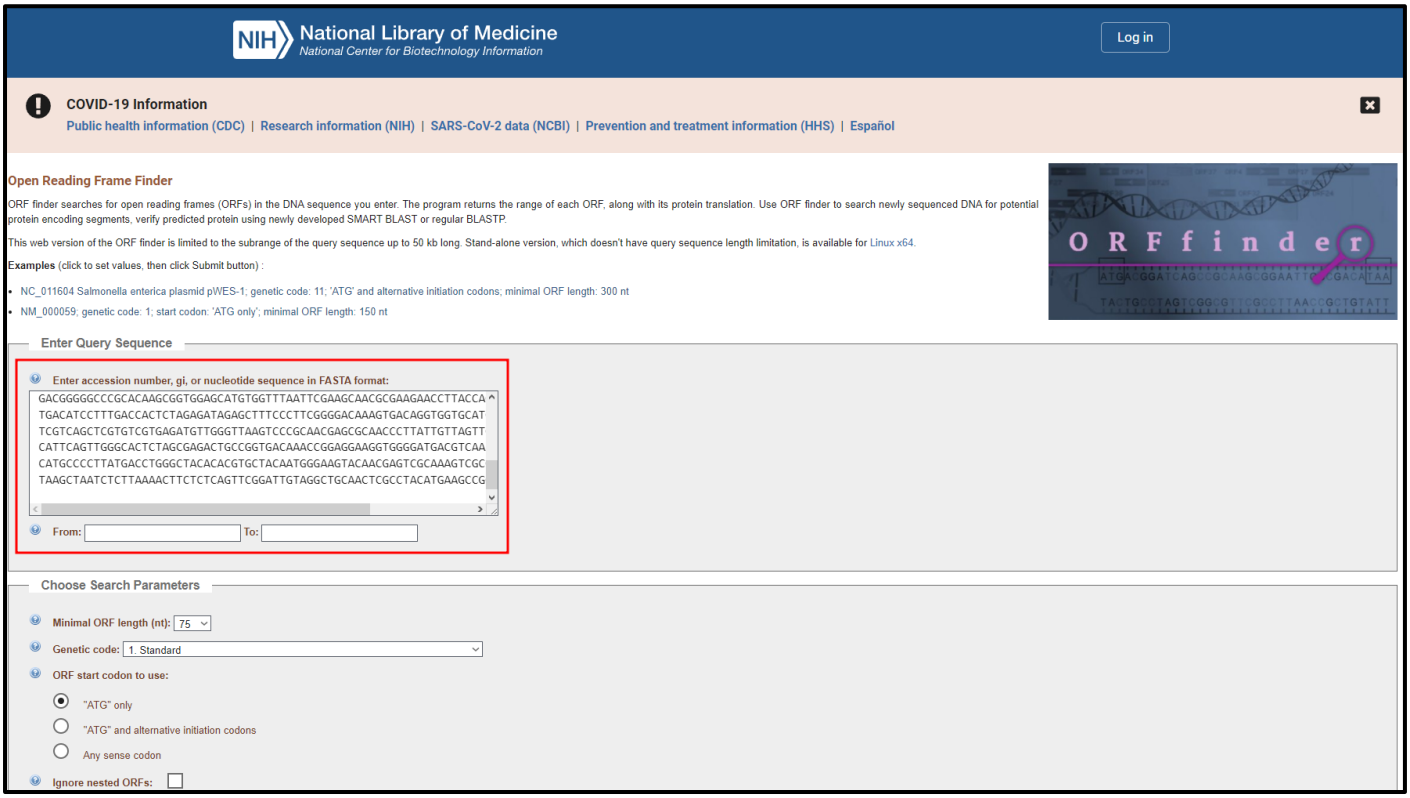

Fig3. Search for Nucleotide FASTA sequence

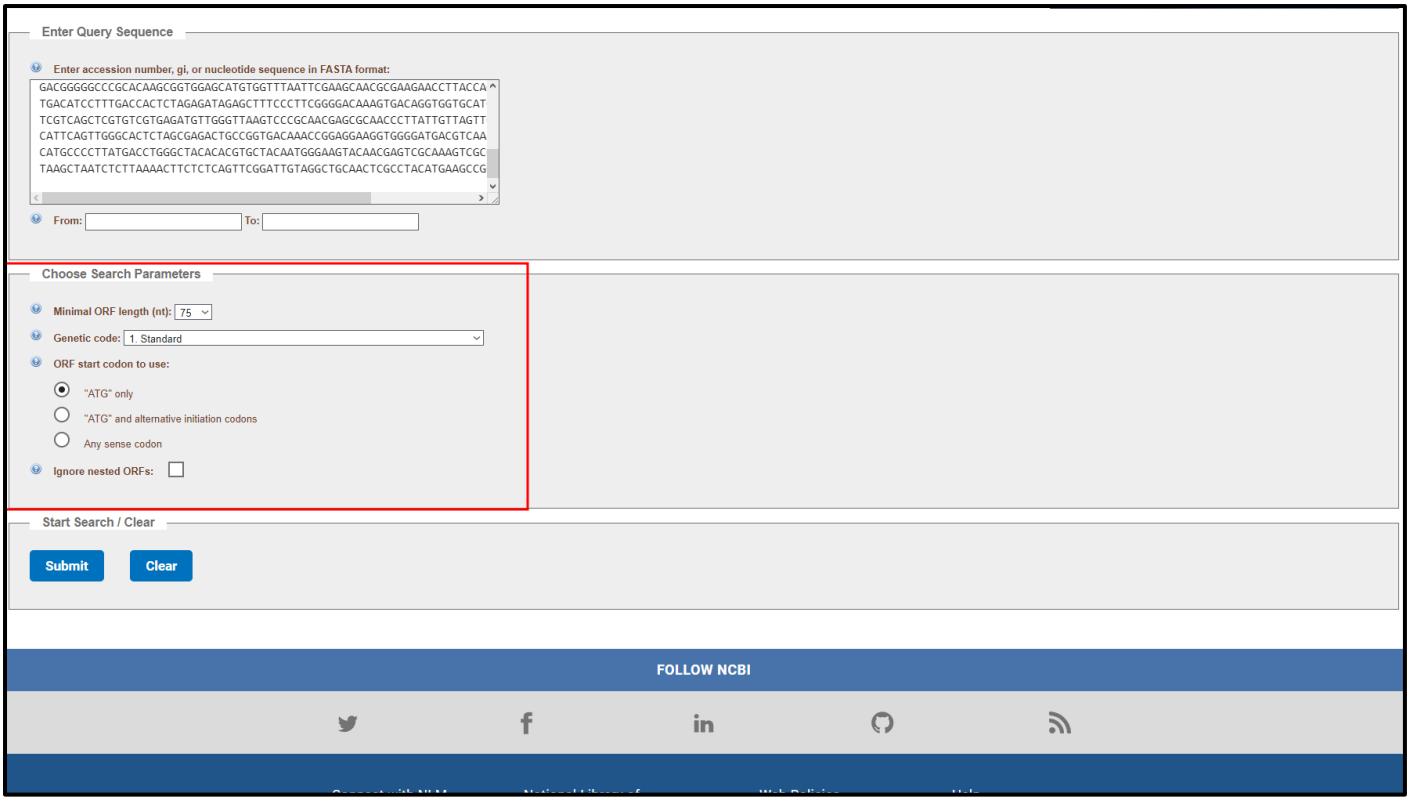

Fig4. Search parameters

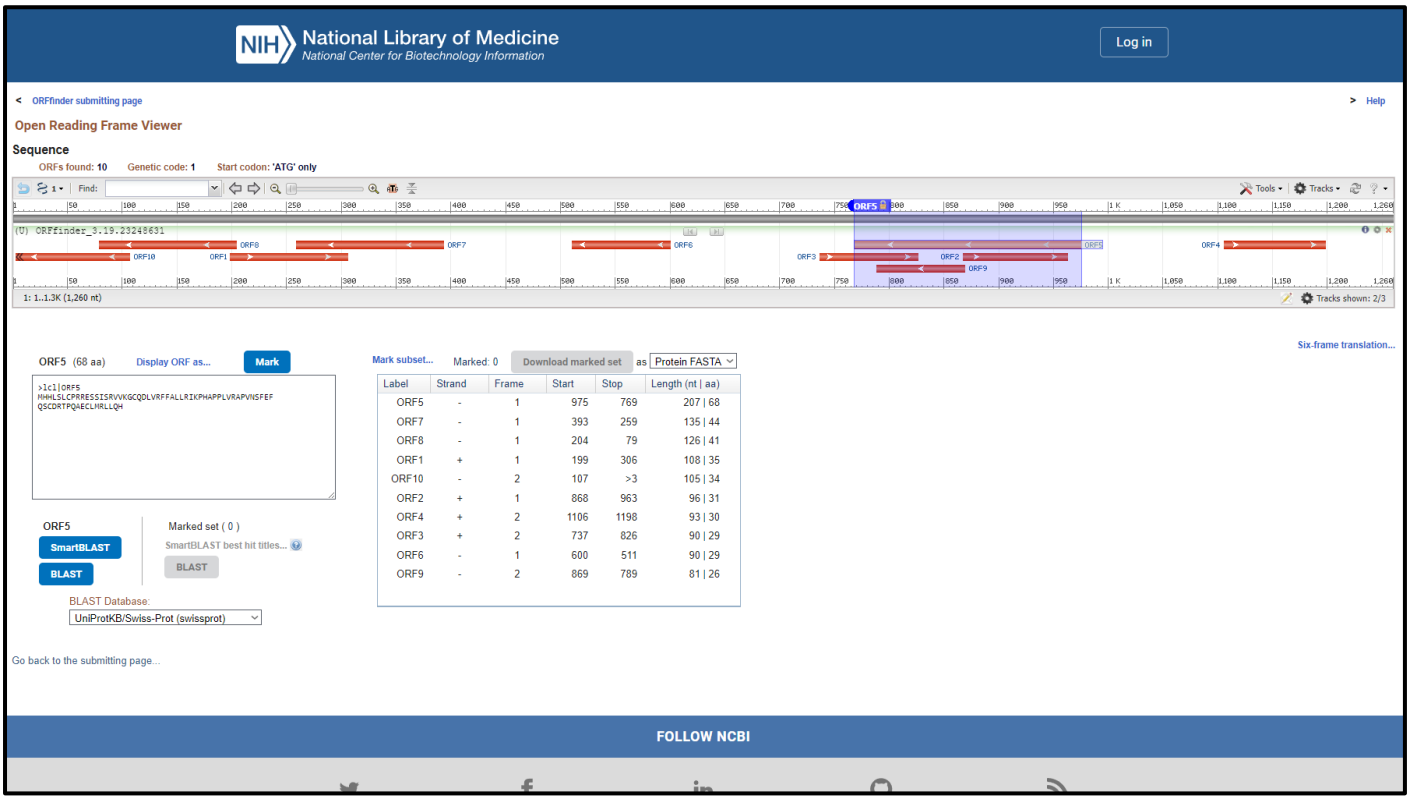

Fig5. Result for recognised ORFs and tracks information

Results:

After submitting nucleotide FASTA sequence for complete genome of Enterococcus alcedinis on NCBI's ORF finder, 360 ORFs were predicted.

CONCLUSION:

ORF finder can used to predict open reading frames in the genome. This information of long ORFs are often used, along with other evidence, to initially identify candidate protein-coding regions or functional RNA-coding regions in a DNA sequence. Small Open Reading Frames (small ORFs/sORFs/smORFs) are important sources of putative peptides previously dismissed as being non-functional or junk DNA.

REFERENCES:

1. Neisseria gonorrhoeae - an overview | ScienceDirect Topics. (n.d.). Www.sciencedirect.com.Retrieved March 18, 2022, from https://www.sciencedirect.com/topics/medicine-and-dentistry/neisseria-gonorrhoeae#:~:text=Neisseria%20gonorrhoeae%20is%20a%20bacterial
2. Neisseria gonorrhoeae strain NJ189125 chromosome, complete genome. (2022). NCBI Nucleotide. Retrieved March 18, 2022, from https://www.ncbi.nlm.nih.gov/nuccore/NZ_CP041586.1
3. Home - ORFfinder - NCBI. (2019). Nih.gov. Retrieved March 18, 2022, from https://www.ncbi.nlm.nih.gov/orffinder/