ICAR Consortium Research Platform on Genomics

On-line training program

on

RNAome:Profiling and Characterization of Non-coding RNAs

14th-20th March 2024

Reference Manual

Dr. Sarika Sahu, Course coordinator

Dr. Neeraj Budhlakoti, Course co-coordinator

Dr. Soumya Sharma, Course co-coordinator

Division of Agricultural Bioinformatics
ICAR-Indian Agricultural Statistics Research Institute
Library Avenue, PUSA, New Delhi - 110012
http://cabgrid.res.in/cabin/

https://iasri.icar.gov.in/

Preface

The era of the innovative world is coming with the advent of new technologies in the field of agriculture and keep enhancing the goal of sustainable development growth worldwide. The most popular and accepted theory of life's origins reveals that the first biocatalysts were made of RNA or a very similar polymer instead of protein. Experiments are beginning to confirm that the catalytic abilities of RNA are compatible with this 'RNA world' hypothesis. An RNA molecule that does not translate into a protein is known as a non-coding RNA (ncRNA). These ncRNAs have been revolutionizing the RNA world in various aspect of life. Recently, several different systematic screens have identified a surprisingly large number of new ncRNA genes. The training program on "RNAome: Profiling and characterization of non-coding RNAs" aimed to provide an insight into basic concepts of various theoretical and practical aspects of transcriptomics. This manual will help the research scholars to learn and explore the application of computational tool/techniques in their research work. The practical-oriented approach would be a big help for the new budding technologist for insight mechanisms of multicellular processes. The module contains each and every section of the program covered in the training program like 'Transcriptome Data pre-processing and Assembly', Introduction to Linux, Introduction to Ashoka, 'Differential gene expression analysis', 'Transcriptome data annotation', 'Prediction and characterization of miRNA' 'Overview of lncRNA and circular RNA', long non coding RNAs's roles in different physiological conditions in livestock, lncRNA prediction through machine learning approach and 'Gene regulatory networks to understand disease Resistance'.

The first talk on "whole transcriptome sequencing by next-generation sequencing (NGS) technologies or RNA-Seq" explained the complex landscape and dynamics of the transcriptome. The sequence reads obtained from the common NGS platforms, including Illumina, SOLiD, and 454, are often very short, ranging from 35bp to 500bp. As a result, it is necessary to reconstruct the full-length transcripts by transcriptome assembly. The theory and hand-on-session on 'Transcriptome Data pre-processing and assembly' provide the comprehensive knowledge of reconstructing entire transcriptome from raw NGS read including detailed understanding of all informatics challenges. It was followed by lectures on Differential gene expression (DGE) analysis. Differential gene expression (DGE) analysis is one of the most common applications of RNA-sequencing (RNA-seq) data. This process allows for the elucidation of differentially expressed genes across two or more conditions and is widely used in many applications of RNA-seq data analysis. Transcriptome annotation provides insight into the function and biological process of transcripts and the proteins they encode. The lectures on Transcriptome annotation explained various tools and techniques for transcriptome annotation.

Micro RNAs (miRNAs) are single stranded, small and non-coding endogenous RNA molecules, which control the gene expression at the post-transcriptional level either by suppression or degradation. Because of its highly conserved nature, *in silico* methods can be employed to predict novel miRNAs in plant species. The lecture on ‘Prediction and characterization of miRNA’ covered bioinformatics tools and techniques for miRNA prediction and functional analysis by identifying genes targeted by the miRNA.

lncRNAs are widely defined as a large and heterogeneous class of regulatory transcripts that are at least 200 nt long. circRNAs are also a subtype of endogenous ncRNAs with tissue- and cell-specific expression patterns, whose biogenesis is regulated by a particular form of alternative splicing, termed backsplicing. With the development of high-throughput technologies and extensive research reports, lncRNAs and circRNAs have gained wide attention for their roles in biological processes. The lectures on ‘Overview of lncRNA and circular RNA’ and ‘Regulatory network analysis of lncRNA’ provided detailed understanding of their roles and bioinformatics tools and techniques for analysis.

Although the manual is mainly focuses on hand-on-session but attempts are taken to explain theory of each session. The details of computational tools, commands and analysis pipeline via flow chart are mentioned for each module separately that will be helpful for the naïve bioinformatician.

Sarika Sahu

Overview of Training Programme

Sarika Sahu, Neeraj Budhlakoti, Soumya Sharma
Division of Agricultural Bioinformatics
ICAR-Indian Agricultural Statistics Research Institute, New Delhi

Introduction:

This online training "RNAome: Profiling and characterization of non-coding RNAs" organized under the aegis of CRP-Genomics project, aims to provide a comprehensive view of the main facets involved in theoretical and practical aspects of this very rapidly growing field of non-coding RNAs. An RNA molecule that does not translate into a protein is known as a non-coding RNA (ncRNA). These ncRNAs have been revolutionizing the RNA world in various aspect of life. Recently, several different systematic screens have identified a surprisingly large number of new ncRNA genes.

RNA biology is the combination of all RNAs whether coding or noncoding. The discovery of non-coding RNAs led to the revolution in RNA world (Derks *et al.* 2015). Noncoding RNAs (ncRNAs) play an important role in various biological processes and gene-disease association (Nallar and Kalvakolanu, 2013). Among the ncRNAs, the most studied ncRNAs are microRNA, which play a major role in gene expression (Hermeking, 2012). However, it has been revealed that long ncRNAs (lncRNAs) also play a very important role in various biological pathways within the cell (Huarte *et al.,* 2010). Researchers reported that several lncRNAs are expressed during stress conditions and are involved in stress-responsive regulation (Zheng *et al.* 2014, Heo *et al.* 2011, Liu *et al.* 2012). lncRNAs are non-coding RNAs whose length is more than 200 base pairs and biochemically resemble mRNAs but they do not translate into proteins. Despite noncoding RNAs, lncRNAs function as RNA genes as well as regulate distant genes. Ponting *et al.* (2009) classified lncRNAs into sense, anti-sense, bidirectional, intronic and intergenic on the basis of their chromosomal localization. In addition, the lncRNAs are normally expressed at low levels and lack sequence similarities among the plant species (Marques and Ponting, 2014). Plethora of literature is available for the identification of lncRNAs in animals while very few are reported on the presence of lncRNAs in plants (Liu *et al.,*2017). The analysis of lncRNA became very easy with the advent of state-of-art technologies like next-generation sequencing. lncRNAs were identified in model plant organisms like Arabidopsis thaliana (Wang *et al.* 2014, Lu *et al.* 2017, Sun *et al.* 2020) Two lncRNAs namely: COOLAIR (cool-assisted intronic non-coding RNA) and COLDAIR (coldassisted intronic non-coding RNA) regulates the flowering time epigenetic repression of FLC (Flowering Locus C) in Arabidopsis (Heo and Sung, 2011). Another important lncRNA: LDMAR (long-day-specific male-fertility-associated RNA) is involved in the regulation of photoperiod male sterility in rice (Ding *et al.* 2012) and participated in ripening of tomato (Zhu *et al.* 2015). These are few examples to be mentioned and suggest the importance of ncRNAs in the plant and crop systems.

Objectives of this training were

To Profile of ncRNAs through Bioinformatics approach.

To provide insight into the role of RNAs and non-coding RNA in regulatory networks.

To Develop an analytical skills through lectures and hands-on session.

Different modules covered under this training program were as following

□ Differential gene expression.

Sequencing platform and Quality Check

Assembly: *de novo* and reference based and annotation

□ **Profiling of RNA regulatory molecule and their role in the regulation of biological processes**

Prediction and characterization of miRNAs

Prediction and characterization of lncRNAs

Prediction and characterization of circRNAs

□ **Regulatory network analysis of RNAs.**

Application of machine learning in ncRNAs prediction

Different theoretical and Practical Sessions were taken during this training program. In this manual, different session taken during training are described in detail. Chapter 2 focuses over RNA-sequencing analysis. Chapter 3 mentions detailed practical procedure taught in the training for Transcriptome Data Pre-processing and Assembly while Chapter 4 given an overview of genome annotation with special focus over gene prediction. Chapter 5 gives detail about Differential Gene Expression Analysis. Chapter 6 provide detail about different tools and execution carried out for Transcriptome data annotation**.** Chapter 7 provides glimpse about world of miRNA. In chapter 8, hands on session over prediction and Characterization of miRNA is covered. Chapter 9 focuses over Circular RNA and about its basic concept and their role in various processes and also covers details of Hands-on-session for circRNA prediction. In chapter 10, aspects of RNAome in biofortification of plant and animal traits is covered.

References

1. Derks KW, Misovic B, van den Hout MC, Kockx CE, Payan Gomez C, Brouwer RW, Vrieling H, Hoeijmakers JH, van IJcken WF, Pothof J. Deciphering the RNA landscape by RNAome sequencing. RNA biology. 2015 Jan 2;12(1):30-42.
2. Ding J, Lu Q, Ouyang Y, Mao H, Zhang P, Yao J, Xu C, Li X, Xiao J, Zhang Q. A long noncoding RNA regulates photoperiod-sensitive male sterility, an essential component of hybrid rice. Proceedings of the National Academy of Sciences. 2012 Feb 14;109(7):2654-9.
3. Heo JB, Sung S. Vernalization-mediated epigenetic silencing by a long intronic noncoding RNA. Science. 2011 Jan 7;331(6013):76-9.
4. Hermeking H. MicroRNAs in the p53 network: micromanagement of tumour suppression. Nature reviews cancer. 2012 Sep;12(9):613-26.
5. Huarte M, Guttman M, Feldser D, Garber M, Koziol MJ, Kenzelmann-Broz D, Khalil AM, Zuk O, Amit I, Rabani M, Attardi LD. A large intergenic noncoding RNA induced by p53 mediates global gene repression in the p53 response. Cell. 2010 Aug 6;142(3):409-19.
6. Liu H, Wang X, Wang HD, Wu J, Ren J, Meng L, Wu Q, Dong H, Wu J, Kao TY, Ge Q. Escherichia coli noncoding RNAs can affect gene expression and physiology of *Caenorhabditis elegans*. Nature communications. 2012 Sep 25;3(1):1-1.
7. Lu Z, Xia X, Jiang B, Ma K, Zhu L, Wang L, Jin B. Identification and characterization of novel lncRNAs in *Arabidopsis thaliana*. Biochemical and biophysical research communications. 2017 Jun 24;488(2):348-54.
8. Marques AC, Ponting CP. Intergenic lncRNAs and the evolution of gene expression. Current opinion in genetics & development. 2014 Aug 1;27:48-53.
9. Nallar SC, Kalvakolanu DV. Regulation of snoRNAs in cancer: close encounters with interferon. Journal of Interferon & Cytokine Research. 2013 Apr 1;33(4):189-98.
10. Ponting CP, Oliver PL, Reik W. Evolution and functions of long noncoding RNAs. Cell. 2009 Feb 20;136(4):629-41.
11. Sun Z, Huang K, Han Z, Wang P, Fang Y. Genome-wide identification of Arabidopsis long noncoding RNAs in response to the blue light. Scientific reports. 2020 Apr 10;10(1):1-0.
12. Wang H, Chung PJ, Liu J, Jang IC, Kean MJ, Xu J, Chua NH. Genome-wide identification of long noncoding natural antisense transcripts and their responses to light in Arabidopsis. Genome research. 2014 Mar 1;24(3):444-53.
13. Zeng H, Wang G, Hu X, Wang H, Du L, Zhu Y. Role of microRNAs in plant responses to nutrient stress. Plant and Soil. 2014 Jan;374(1):1005-21.
14. Zhu B, Yang Y, Li R, Fu D, Wen L, Luo Y, Zhu H. RNA sequencing and functional

analysis implicate the regulatory role of long non-coding RNAs in tomato fruit ripening. Journal of Experimental Botany. 2015 Aug 1;66(15):4483-95.

Introduction to RNA-Sequencing

Dwijesh Chand Mishra and Neeraj Budhlakoti
Division of Agricultural Bioinformatics
ICAR-Indian Agricultural Statistics Research Institute, New Delhi

Introduction

The advent of Next-Generation Sequencing (NGS) technology has transformed genomic studies. One important application of NGS technology is the study of the *transcriptome*, which is defined as the complete collection of all the RNA molecules in a cell. Various types of RNA that have been classified so far are shown in **Fig. 1**. All of these molecules are called *transcripts* since they are produced by process of transcription.

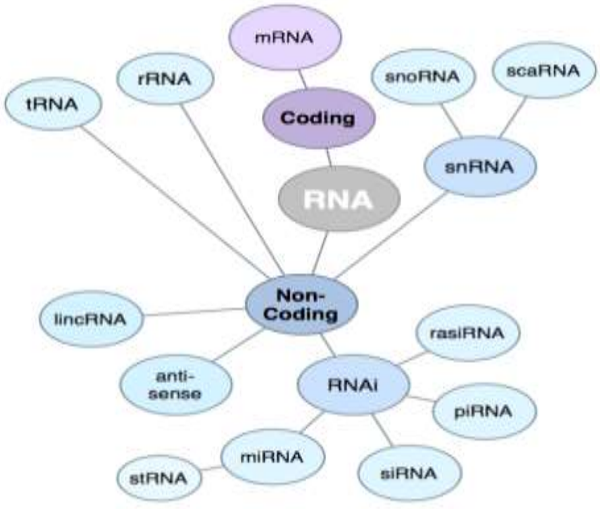


Fig. 1: Different types of RNA

(Image source: http://scienceblogs.com/digitalbio/2011/01/08/next-gene-sequencing)

Understanding the transcriptome is essential for interpreting the functional elements of the genome and revealing the molecular constituents of cells and tissues, and also for understanding development and disease [1]. The main purpose of transcriptomics are: to catalogue all species of transcript, including mRNAs, non-coding RNAs and small RNAs; to determine the transcriptional structure of genes, in terms of their start sites, 5′ and 3′ ends, splicing patterns and other post-transcriptional modifications; and to quantify the changing expression levels of each transcript during development and under different conditions.

The study of transcriptome is carried out through sequencing of RNAs. RNA sequencing *(RNA-Seq)* is a powerful method for discovering, profiling, and quantifying RNA transcripts [2].

RNA-Seq uses NGS datasets to obtain sequence reads from millions of individual RNAs. The RNA-Seq analysis is performed in several steps: First, all genes are extracted from the reference genome (using annotations of type *gene*). Other annotations on the gene sequences are preserved (e.g. CDS information about coding sequences etc). Next, all annotated transcripts (using annotations of type *mRNA*) are extracted [3]. If there are several annotated splice variants, they are all extracted. An example is shown in below **Fig. 2(a).**

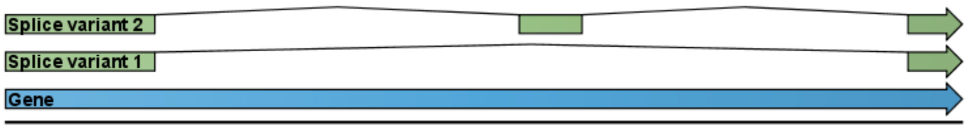


Fig. 2(a): A simple gene with three exons and two splice variants.

The given example is a simple gene with three exons and two splice variants. The transcripts are extracted as shown in **Fig. 2(b).**

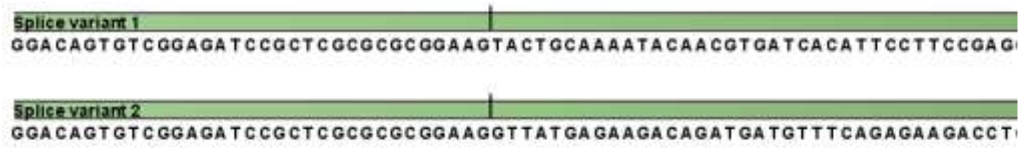

Splice variant 1
GGACAGTGTCGGAGATCCGCTCGCGCGGAAGTACTGCAAAATACAACGTGATCACATTCCTTCCGAG
Splice variant 2
GGACAGTGTCGGAGATCCGCTCGCGCGGAAGGTTATGAGAAGACAGATGATGTTTCAGAGAAGACCT

Fig. 2(b): All the exon-exon junctions are joined in the extracted transcript.

Next, the reads are mapped against all the transcripts plus the entire gene [see **Fig. 2(c)**].

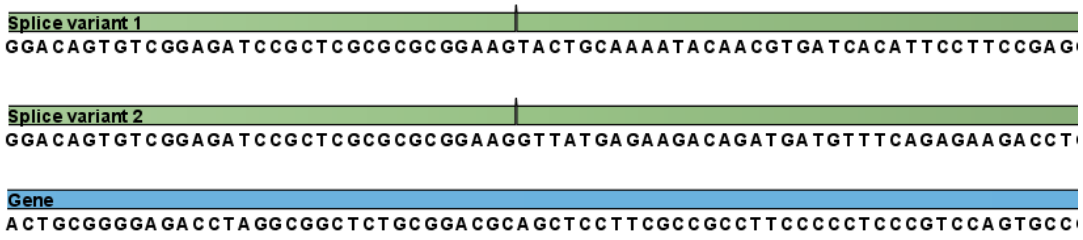

Splice variant 1
GGACAGTGTCGGAGATCCGCTCGCGCGGAAGTACTGCAAAATACAACGTGATCACATTCCTTCCGAG
Splice variant 2
GGACAGTGTCGGAGATCCGCTCGCGCGGAAGGTTATGAGAAGACAGATGATGTTTCAGAGAAGACCT
Gene
ACTGCGGGGAGACCTAGGCGGCTCTGCGGACGCAGCTCCTTCGCCGCCTTCCCCCTCCCGTCCAGTGCC

Fig. 2(c): The reference for mapping: all the exon-exon junctions and the gene.

(Image source: CLC Genomic workbench tutorials)

From this mapping, the reads are categorized and assigned to the genes and expression values for each gene and each transcript are calculated and putative exons are then identified.

RNA Sequencing Experiment

In a standard RNA-seq experiment, a sample of RNA is converted to a library of complementary DNA fragments and then sequenced on a high-throughput sequencing platform, such as Illumina's Genome Analyzer, SOLiDor Roche 454 [4]. Millions of short sequences, or reads, are obtained from this sequencing and then mapped to a reference genome (**Fig. 3**). The count of reads mapped to a given gene measures the expression level of this gene.

The unmapped reads are usually discarded and mapped reads for each sample are assembled into gene-level, exon-level or transcript-level expression summaries, depending on the objectives of the experiment. The count of reads mapped to a given gene/exon/transcript measures the expression level for this region of the genome or transcriptome.

One of the primary goals for most RNA-seq experiments is to compare the gene expression levels across various treatments. A simple and common RNA-seq study involves two treatments in a randomized complete design, for example, treated versus untreated cells, two different tissues from an organism, plants, etc. In most of the studies, researchers are particularly interested in detecting gene with differential expressions (DE). A gene is declared differentially expressed if an observed difference or change in read counts between two experimental conditions is statistically significant, i.e. if the difference is greater than what would be expected just due to random variation [5]. Detecting DE genes can also be an important pre-step for subsequent studies, such as clustering gene expression profiles or testing gene set enrichments.

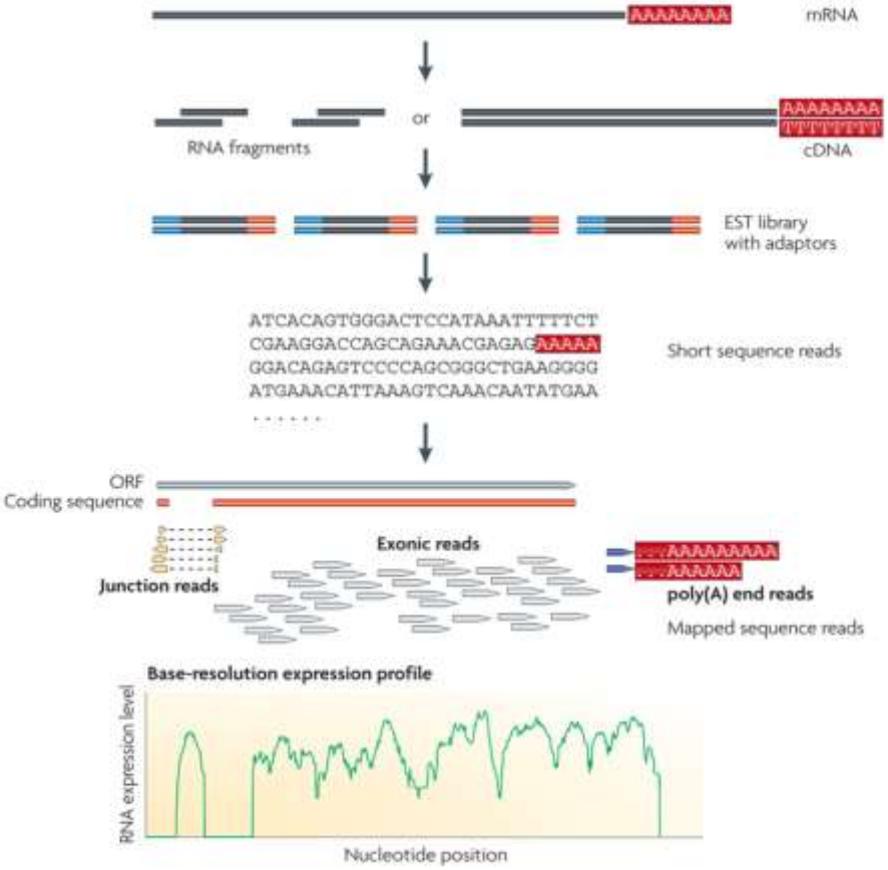


Fig. 3: General RNA-seq experiment. mRNA is converted to cDNA, and fragments from that library are used to generate short sequence reads. Those reads are assembled into contigs which may be mapped to reference sequences (Wang et al., 2009).

Analysing RNA-Seq data

RNA-seq experiments must be analyzed with robust, efficient and statistically correct algorithms. Fortunately, the bioinformatics community has been striving hard at work for incorporating mathematics, statistics and computer science for RNA-seq and building these ideas into software tools. RNA-seq analysis tools generally fall into three categories: (i) those for read alignment; (ii) those for transcript assembly or genome annotation; and (iii) those for transcript and gene quantification. Some of the open source software available for RNA-seq analysis are as follows:

- **Data preprocessing**
    - Fastx toolkit
    - Samtools
- **Short reads aligners**
    - Bowtie, TOPHAT, Stampy, BWA, Novoalign, etc
- **Expression studies**
    - Cufflinks package
    - R packages (DESeq, edgeR, *more…*)
- **Visualisation**
    - CummeRbund, IGV, Bedtools, UCSC Genome Browser, etc.

Besides there are commercially data analysis pipelines like GenomeQuest, CLCBio etc available for researchers to use. The most commonly used pipeline is to identify protein coding genes by aligning RNA-Seq data to annotate data from sources like RefSeq. After generating the alignments, the number of aligning sequences is counted for each position. Since each alignment represents a transcript, the alignments allow to count the number of RNA molecules produced from every gene.

Using NGS technology, RNA-Seq enables to count the number of reads that align to one of thousands of different cDNAs, producing results similar to those of gene expression microarrays [6]. Sequences generated from an RNA-Seq experiment are usually mapped to libraries of known exons in known transcripts. RNA-Seq can be used for discovery applications such as identifying alternative splicing events, allele-specific expression, and rare and novel transcripts [7]. The sequencing output files (compressed FASTQ files) are the input for secondary analysis. Reads are aligned to an annotated reference genome, and those aligning to exons, genes and splice junctions are counted. The final steps are data visualisation and interpretation, consisting of calculating gene- and transcript-expression and reporting differential expression. A general Bioinformatics workflow to map transcripts from RNA-seq data is shown in **Fig. 4**.

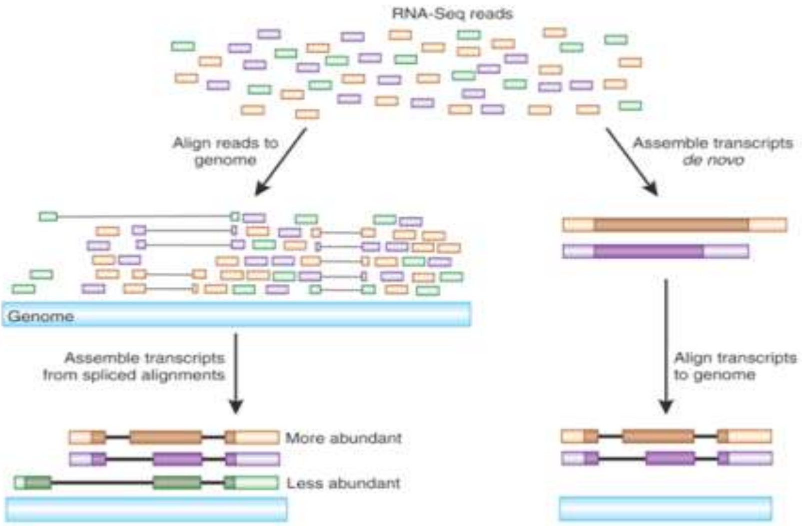

RNA-Seq reads
Align reads to genome
Assemble transcripts de novo
Genome
Assemble transcripts from spliced alignments
Align transcripts to genome
More abundant
Less abundant

Fig. 4: RNA-seq workflow (Adapted fromAdvancing RNA-Seq analysis Brian J. Haas and Michael C. Zody Nature Biotechnology 28, 421-423 (2010)

RPKM (Reads per KB per million reads)

RNA-Seq provides quantitative approximations of the abundance of target transcripts in the form of counts. However, these counts must be normalized to remove technical biases inherent in the preparation steps for RNA-Seq, in particular the length of the RNA species and the sequencing depth of a sample. The most commonly used is RPKM (Reads Per Kilobase of exon model per Million mapped reads). The RPKM measure of read density reflects the molar concentration of a transcript in the starting sample by normalizing for RNA length and for the total read number in the measurement [8]. RPKM is mathematically represented as:

$$\text{RPKM}=\frac{total\ exon\ reads}{mapped\ reads\ (millions)\text{X}\ exon\ length\ (KB)}$$

Total exon reads

This is the number of reads that have been mapped to a region in which an exon is annotated for the gene or across the boundaries of two exons or an intron and an exon for an annotated transcript of the gene. For eukaryotes, exons and their internal relationships are defined by annotations of type mRNA.

Exon length

This is calculated as the sum of the lengths of all exons annotated for the gene. Each exon is included only once in this sum, even if it is present in more annotated transcripts for the gene. Partly overlapping exons will count with their full length, even though they share the same region.

Mapped reads

The total gene reads for a gene is the total number of reads that after mapping have been mapped to the region of the gene. A gene's region is that comprised of the flanking regions, the exons, the introns and across exon-exon boundaries of all transcripts annotated for the gene. Thus, the sum of the total gene reads numbers is the number of mapped reads for the sample.

Applications of RNA-seq

This technique can be used to:

- Measure gene expression
- Transcriptome assembly, gene discovery and annotation
- Detect differential transcript abundances between tissues, developmental stages, genetic backgrounds, and environmental conditions
- Characterize alternative splicing, alternative polyadenylation, and alternative transcription.

Future Directions

Although RNA-Seq is still in the infancy stages of use, it has clear advantages over previously developed transcriptomic methods. Compared with microarray, which has been the dominant approach of studying gene expression in the last two decades, RNA-seq technology has a wider measurable range of expression levels, less noise, higher throughput, and more information to detect allele-specific expression, novel promoters, and isoforms [9]. For these reasons, RNA-seq is gradually replacing the array-based approach as the major platform in gene expression studies. The next big challenge for RNA-Seq is to target more complex transcriptomes to identify and track the expression changes of rare RNA isoforms from all genes. Technologies that will advance achievement of this goal are pair-end sequencing, strand-specific sequencing and the use of longer reads to increase coverage and depth. As the cost of sequencing continues to fall, RNA-Seq is expected to replace microarrays for many applications that involve determining the structure and dynamics of the transcriptome.

References

1. https://www.genome.gov/13014330
2. WangZ., GersteinM., SynderM. (2009). Rna-seq: a revolutionary tool for transciptomics, Nat Rev Genet 10(1): 57–63.
3. http://scienceblogs.com/digitalbio/2011/01/08/next-gene-sequencing-results-a/
4. Shendure J, Ji H (2008) Next-generation RNA sequencing. Nature Biotechnology 26: 2514-2521
5. Anders S, Huber W (2010). Differential expression analysis for sequence count data. Genome Biol. 11:R106.
   Illumina, Inc,. (2011). Getting started with RNA-Seq Data Analysis. Pub. No. 470-2011-003.
6. Illumina, Inc,. (2011). RNA-Seq Data Comparison with Gene Expression Microarrays. A cross-platform comparison of differential gene expression analysis. Pub. No. 470-2011-004
7. Yaqing Si (2012). Statistical analysis of RNA-seq data from next-generation sequencing technology. PhD thesis. Iowa State University, Ames, Iowa.
8. Mortazavi, A., Williams, B. A., McCue, K., Schaeffer, L., and Wold, B. (2008). Mapping and quantifying mammalian transcriptomes by rna-seq. Nat Methods, 5(7):621-628.
9. Wang L., Si Y., Dedow L.K., Shao Y., Liu P., Brutnell T.P. (2010). A low-cost library construction protocol and data analysis pipeline for Illumina-based strand-specific multiplex RNA-seq. PLoS One 6(10):e26426.
10. Brian J. H. and Michael C. Z. (2010). Advancing RNA-Seq analysis. Nature Biotechnology 28, 421-423.

Transcriptome Data Pre-processing and Assembly

Soumya Sharma, Sarika Sahu and Neeraj Budhlakoti
Division of Agricultural Bioinformatics
ICAR-Indian Agricultural Statistics Research Institute, New Delhi

Transcript profiling ("Transcriptomics") is a widely used technique that obtains information on the abundance of multiple mRNA transcripts within a biological sample simultaneously. Therefore, when a number of such samples are analysed, as in a scientific experiment, large and complex data sets are gene-rated. RNA-Seq technology utilizing NGS sequencing has emerged as an attractive alternative to traditional microarray platforms for conducting transcriptional profiling. Next generation sequencing (NGS) experiments generate a tremendous amount of data which can't be directly analyzed in any meaningful way. Selecting the right analytical approach along with an appropriate set of bioinformatics tools is key to extract useful information from RNA-Seq data while avoiding misinterpretation or bias. In the present section we will discuss about the assembly of short-read Illumina sequencing data, which is commonly used for RNA-Seq experiments.

Requirements for RNA-Seq Data Assembly

Hardware

- Linux environment or server
- Accessed via shell terminals, such as PuTTY or MobaXterm
- Can use a virtual machine on Windows
- 32GB RAM recommended if working with larger genomes
- 1TB storage or higher recommended for smaller projects

Software

- FastQC
  https://www.bioinformatics.babraham.ac.uk/projects/download.html
- Trimmomatic
  http://www.usadellab.org/cms/?page=trimmomatic
- Bowtie2
  https://sourceforge.net/projects/bowtie-bio/files/bowtie2/
- Tophat
  https://ccb.jhu.edu/software/tophat/index.shtml
- Cufflinks
  http://cole-trapnell-lab.github.io/cufflinks/getting_started/
- Trinity
  https://github.com/trinityrnaseq/trinityrnaseq/wiki/Installing-Trinity

Pre-processing of RNA-Seq Data

First, switch to the where the FASTQ files are stored directory. Use the cd command (i.e., change directory) followed by the path of the directory.

```
>> cd /path/to/folder_name/
```

Next, you can check the FASTQ files by using the ls command (i.e., listing), which shows the contents of the current working directory.

Data files from sequencing providers are typically compressed and have the extension “.fastq.gz”. These files contain structured information about individual NGS reads—a unique identifier, the called bases, and the associated quality scores.

Lastly, you can make an output directory using the mkdir command (i.e., make directory). Output files can be stored here.

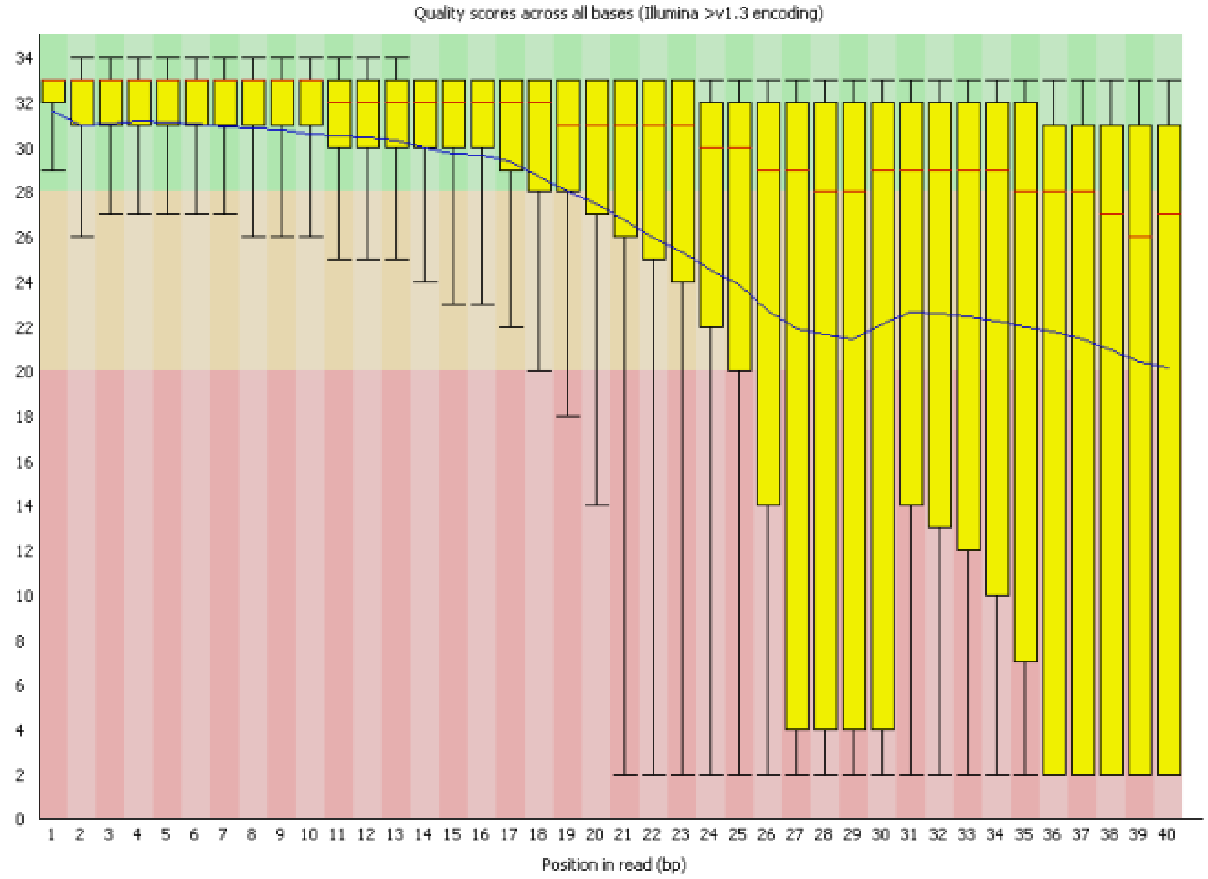
```
>> mkdir /path/to/output_folder/
```

1. Check quality with FastQC

Run FastQC to check the raw data quality.

```
>> fastqc sample_01.fastq.gz --extract -o /path/to/output_folder
```

The output contains graphs and statistics about the raw quality, including quality scores, GC content, adapter percentage, and more. Below is an examples of the output file “Per base Sequence quality”.



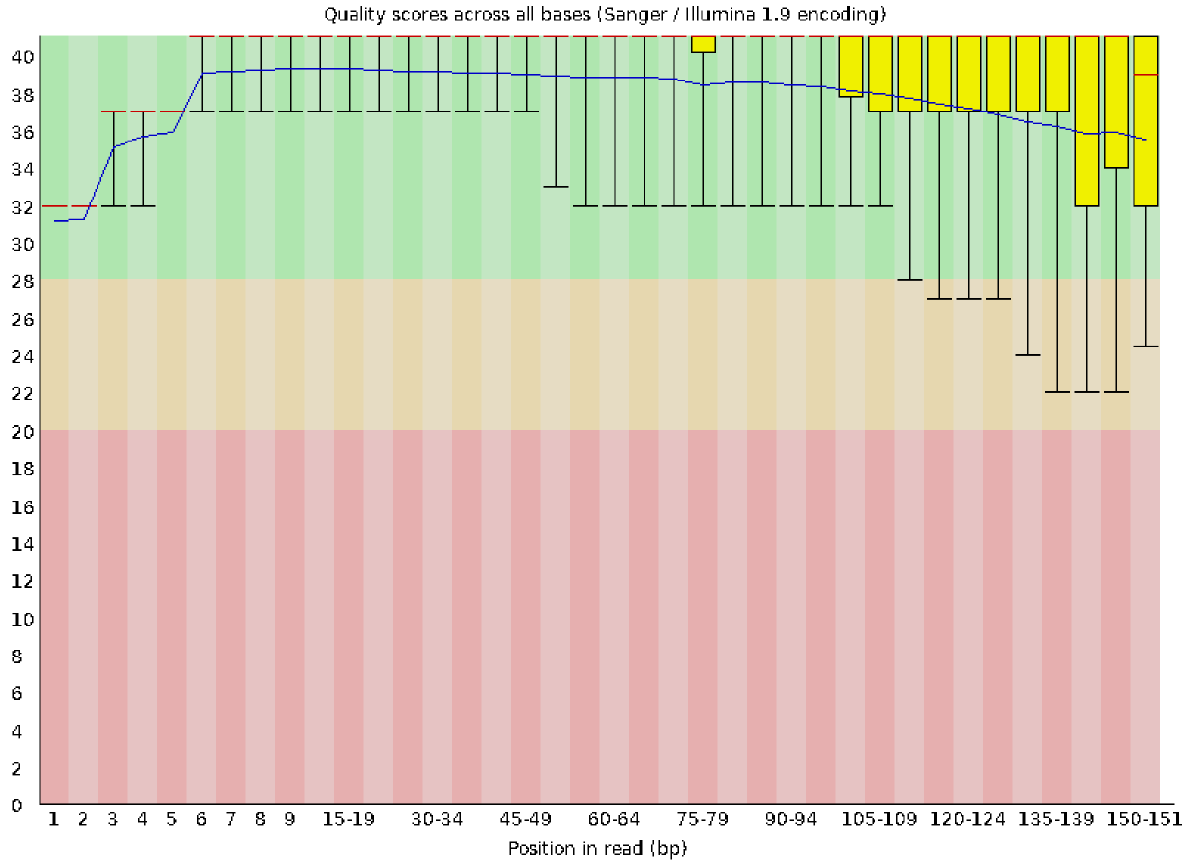
Per base sequence quality. Quality scores for each base position in the read are represented as box plots. The blue line represents the average quality score. High-quality data will typically have over 80% of bases with a quality score of 30 or higher (i.e., Q30 > 80%). Q30 represents 99.9% accuracy in the base call, or an error rate of 1 in 1000. A dip in quality is expected towards the end of the read.

2. Trim reads with Trimmomatic

Poor-quality regions and adapter sequences should be trimmed from the reads before further analysis. Trimmomatic can be used for trimming the low quality reads and adapter sequences.

```
>> trimmomatic  PE input_forward.fastq.gz  input_reverse.fastq.gz
output_forward_paired.fastq.gz  output_forward_unpaired.fastq.gz
output_reverse_paired.fastq.gz  output_reverse_unpaired.fastq.gz
ILLUMINACLIP:TruSeq3-PE.fa:2:30:10:2:keepBothReads LEADING:3 TRAILING:3
MINLEN:36
```

Run FastQC again on the trimmed treads to confirm that the new quality is acceptable.



<u>Transcriptome Assembly</u>

Refrence based Assembly

1. Indexing the reference genome

First, index the reference genome using Bowtie2 to prepare it for alignment. Adding gene annotation information to the reference genome will facilitate alignment of RNA-Seq reads across exon-intron boundaries. This indexing step is only required once; you can then use the indexed genome repeatedly in future analysis.

```
>> bowtie-build [options]* <input referencegenome fasta file> < basename of the index files >
```

It results in 6 files with extention .bt2

2. Map/Align the reads to reference Genome

Then, align the reads using Tophat.

```
>> tophat [options]* <genome_index_base> PE_reads_1.fq.gz,SE_reads.fa PE_reads_2.fq.gz
```

- or -

```
>> tophat [options]* <genome_index_base> PE_reads_1.fq.gz PE_reads_2.fq.gz,SE_reads.fa
```

Check the mapping statistics in the [sample_name]Log.final.out file to ensure the BAM file was generated properly and the reads align to the genome correctly. Uniquely mapped reads are the most useful for expression analysis, as there is high confidence in which loci they represent. In general, >60-70% for the “uniquely mapped reads %” metric is considered good; a significantly lower value warrants further investigation.

3. Assemble the mapped reads

Use Cufflinks program to assemble aligned RNA-Seq reads into transcripts, estimate their abundances, test for differential expression and regulation, and provide transcript quantification. Some of the tools part of Cufflinks can be run individually, while others are part of a larger workflow.

```
>> cufflinks [options] input_alignments.[sam|bam]
```

The program cufflinks produces number of files in its predefined output directory. Some of the generated files are:

transcripts.gtf: The GTF file contains Cufflinks’ assembled isoforms where there is one GTF record per row, and each record represents either a transcript or an exon within a transcript

isoforms.fpkm_tracking: This file contains the estimated isoform-level expression values in the generic FPKM Tracking Format

genes.fpkm_tracking: This file contains the estimated gene-level expression values in the generic FPKM Tracking Format

De novo Assembly

De novo transcriptome assembly is often the preferred method to studying non-model organisms, since reference-based methods are not possible without an existing genome. De novo assembly can be performed using Trinity assembler.

A typical Trinity command for assembling non-strand-specific RNA-seq data would be like so, running the entire process on a single high-memory server (aim for ~1G RAM per ~1M ~76 base Illumina paired reads, but often much less memory is required):

```
Trinity --seqType fq --max_memory 50G  --left reads_1.fq.gz  --right reads_2.fq.gz --CPU 6
```

If multiple sets of fastq files are available, such as corresponding to multiple tissue types or conditions, etc., indicate them to Trinity like following:

```
 Trinity --seqType fq --max_memory 50G --left condA_1.fq.gz,condB_1.fq.gz,condC_1.fq.gz
–right condA_2.fq.gz,condB_2.fq.gz,condC_2.fq.gz  --CPU 6
```

When Trinity completes, it will create a 'Trinity.fasta' output file in the 'trinity_out_dir/' output directory (or output directory specified).

Trinity groups transcripts into clusters based on shared sequence content. Such a transcript cluster is very loosely referred to as a 'gene'. This information is encoded in the Trinity fasta accession.

Genome Annotation: Gene Prediction

Sanjeev Kumar, D.C. Mishra and Jyotika Bhati
Division of Agricultural Bioinformatics
ICAR-Indian Agricultural Statistics Research Institute, New Delhi

Introduction

Until the genome revolution, genes were identified by researchers with specific interests in a particular protein or cellular process. Once identified, these genes were isolated, typically by cloning and sequencing cDNAs, usually followed by targeted sequencing of the longer genomics segments that code for the cDNAs. Once an organism's entire genome sequence becomes available, there is strong motivation for finding all the genes encoded by a genome at once rather than in a piecemeal approach. Such catalogue is immensely valuable to researchers, as they can learn much more from the whole picture than from a much more limited set of genes. For example, genes of similar sequence can be identified, evolutionary and functional relationships can be elucidated, and a global picture of how many and what types of genes are present in a genome can be seen. A significant portion of the effort in genome sequencing is devoted to the process of *annotation*, in which genes, regulatory elements, and other features of the sequence are identifies as thoroughly as possible and catalogued in a standard format in public databases so that researchers can easily use the information. Functional genomics research has expanded enormously in the last decade and particularly the plant biology research community. Functional annotation of novel DNA sequences is probably one of the top requirements in functional genomics as this holds, to a great extent, the key to the biological interpretation of experimental results.

Computational Gene Prediction

Computational gene prediction is becoming more and more essential for the automatic analysis and annotation of large uncharacterized genomic sequences. In the past two decades, many algorithms have been evolved to predict protein coding regions of the DNA sequences. They all have in common, to varying degree, the ability to differentiate between gene features like Exons, Introns, Splicing sites, Regulatory sites etc. Gene prediction methods predicts coding region in the query sequences and then annotates the sequences databases.

Gene Structure and Expression

The gene structure and the gene expression mechanism in eukaryotes are far more complicated than in prokaryotes. In typical eukaryotes, the region of the DNA coding for a protein is usually not continuous. This region is composed of alternating stretches of *exons* and *introns*. During transcription, both exons and introns are transcribed onto the RNA, in their linear order. Thereafter, a process called *splicing* takes place, in which, the intron sequences are excised and discarded from the RNA sequence. The remaining RNA segments, the ones corresponding to the exons are ligated to form the mature RNA strand. A typical multi-exon gene has the following structure (as illustrated in Fig. 1).

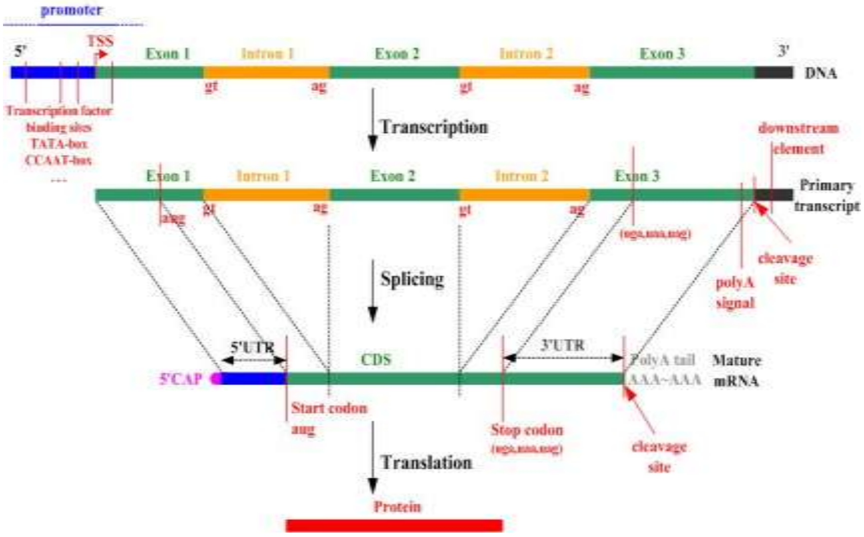


Fig. 1: Representative Diagram of Protein Coding Eukaryotic Gene

It starts with the promoter region, which is followed by a transcribed but non-coding region called *5' untranslated region (5' UTR)*. Then follows the initial exon which contains the start codon. Following the initial exon, there is an alternating series of introns and internal exons, followed by the terminating exon, which contains the stop codon. It is followed by another non-coding region called the *3' UTR*. Ending the eukaryotic gene, there is a polyadenylation (polyA) signal: the nucleotide Adenine repeating several times. The exon-intron boundaries (i.e., the splice sites) are signalled by specific short (2bp long) sequences. The 5'(3') end of an intron (exon) is called the *donor* site, and the 3'(5') end of an intron (exon) is called the *acceptor* site. The problem of gene identification is complicated in the case of eukaryotes by the vast variation that is found in gene structure.

Gene Prediction Methods

There are mainly two classes of methods for computational gene prediction (Fig. 2). One is based on sequence similarity searches, while the other is gene structure and signal-based searches, which is also referred to as *Ab initio* gene finding.

Sequence Similarity Searches

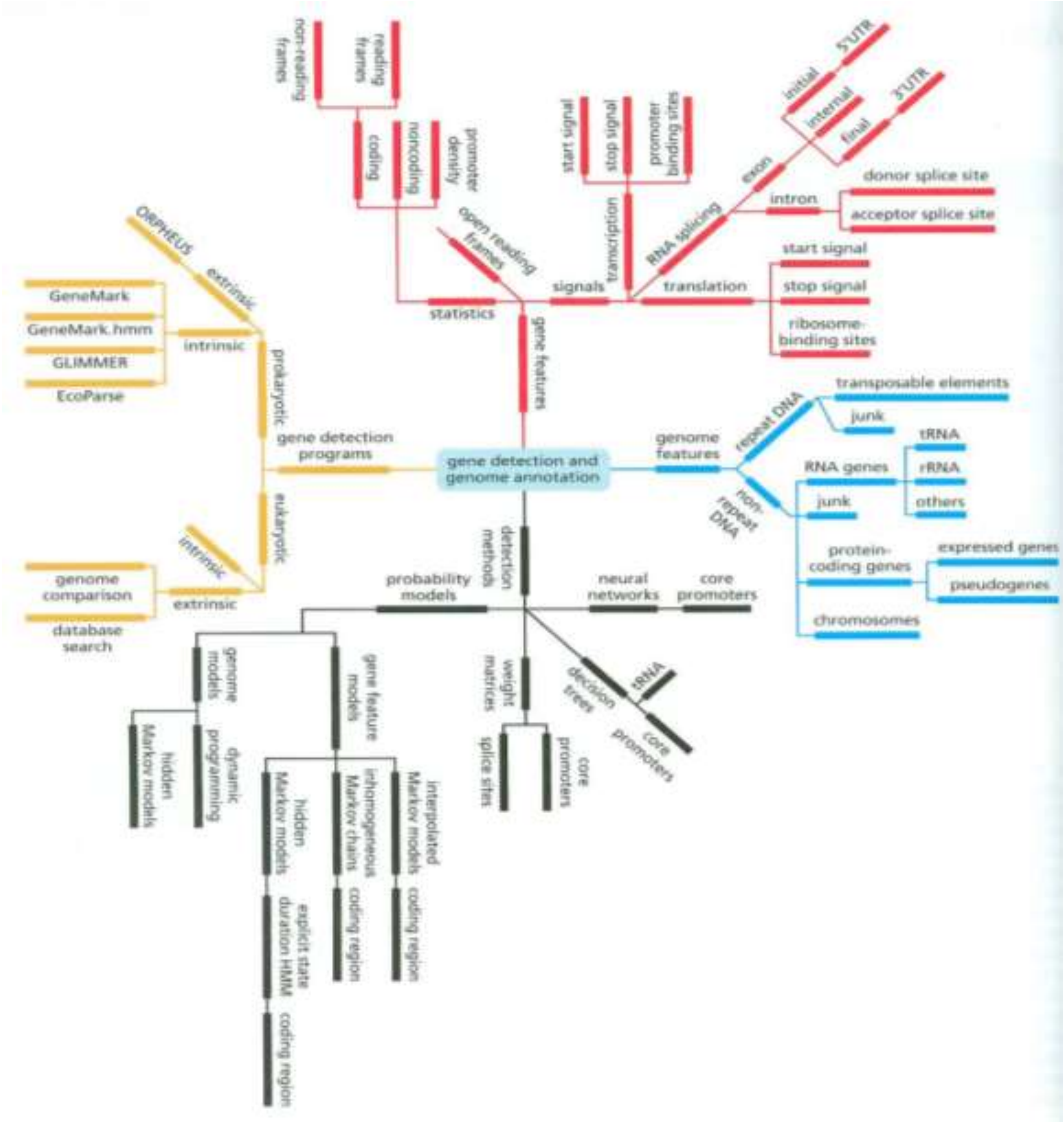
Sequence similarity search is a conceptually simple approach that is based on finding similarity in gene sequences between ESTs (expressed sequence tags), proteins, or other genomes to the input genome. This approach is based on the assumption that functional regions (exons) are more conserved evolutionarily than non-functional regions (intergenic or intronic regions). Once there is similarity between a certain genomic region and an EST, DNA, or protein, the similarity information can be used to infer gene structure or function of that region. EST-based sequence similarity usually has drawbacks in that ESTs only correspond to small portions of the gene sequence, which means that it is often difficult to predict the complete gene structure of a given region. Local alignment and global alignment are two methods based on similarity searches. The most common local alignment tool is the BLAST family of programs, which detects sequence similarity to known genes, proteins, or ESTs. The biggest limitation to this type of approaches is that only about half of the genes being discovered have significant homology to genes in the databases.

Ab initio Gene Prediction Methods

The second class of methods for the computational identification of genes is to use gene structure as a template to detect genes, which is also called *ab initio* prediction. *Ab initio* gene predictions rely on two types of sequence information: signal sensors and content sensors. Signal sensors refer to short sequence motifs, such as splice sites, branch points, poly pyrimidine tracts, start codons and stop codons. Exon detection must rely on the content sensors, which refer to the patterns of codon usage that are unique to a species, and allow coding sequences to be distinguished from the surrounding non-coding sequences by statistical detection algorithms.

Many algorithms are applied for modelling gene structure, such as Dynamic Programming, linear discriminant analysis, Linguist methods, Hidden Markov Model and Neural Network. Based on these models, a great number of *ab initio* gene prediction programs have been developed.

Fig. 2: Diagrammatic Representation of Gene Prediction and Annotation



Gene Discovery in Prokaryotic Genomes

Discovery of genes in Prokaryote is relatively easy, due to the higher gene density typical of prokaryotes and the absence of introns in their protein coding regions. DNA sequences that encode proteins are transcribed into mRNA, and the mRNA is usually translated into proteins without significant modification. The longest ORFs (open reading frames) running from the first available start codon on the mRNA to the next stop codon in the same reading frame generally provide a good, but not assured prediction of the protein coding regions. Several methods have been devised that use different types of Markov models in order to capture the compositional differences among coding regions, "shadow" coding regions (coding on the opposite DNA strand), and noncoding DNA. Such methods, including ECOPARSE, the widely used GENMARK, and Glimmer program, appear to be able to identify most protein coding genes with good performance (Fig. 3).

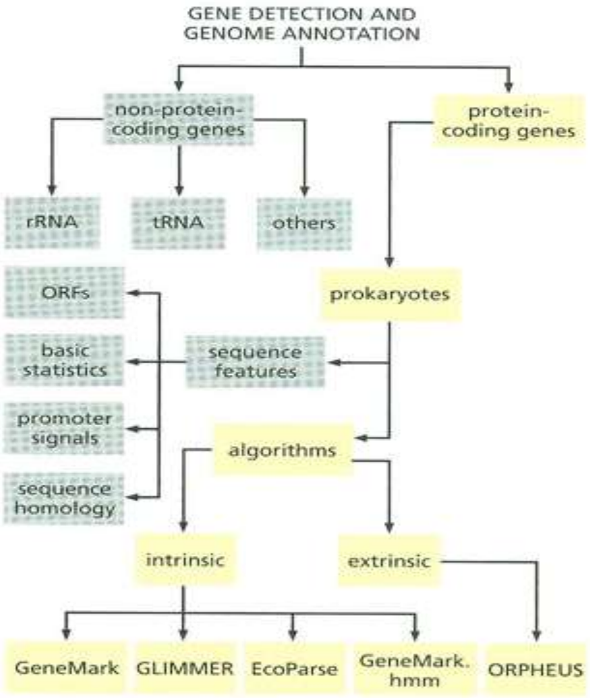


Fig. 3: Flow Diagram of Prokaryotic Gene Discovery

Gene Discovery in Eukaryotic Genome

It is a quite different problem from that encountered in prokaryotes. Transcription of protein coding regions initiated at specific promoter sequences is followed by removal of noncoding sequences (introns) from pre-mRNA by a splicing mechanism, leaving the protein encoding exons. Once the introns have been removed and certain other modifications to the mature RNA have been made, the resulting mature mRNA can be translated in the 5` to 3` direction, usually from the first start codon to the first stop codon. As a result of the presence of intron sequences in the genomic DNA sequences of eukaryotes, the ORF corresponding to an encoded gene will be interrupted by the presence of introns that usually generate stop codons (Fig.4).

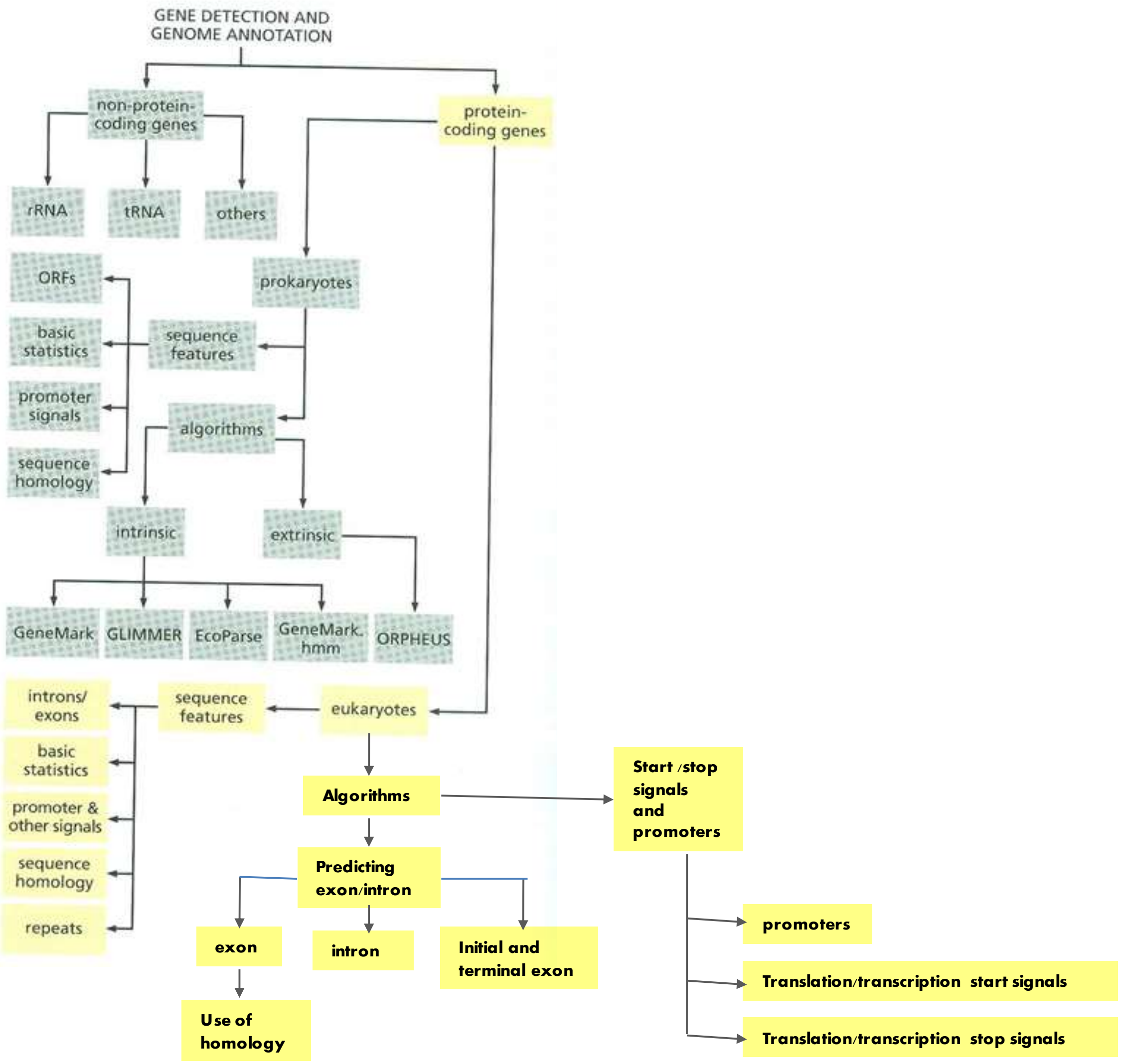


Fig. 4: Flow Diagram of Eukaryotic Gene Discovery

Gene Prediction Program

There are two basic problems in gene prediction: prediction of protein coding regions and prediction of the functional sites of genes. Gene prediction program can be classified into four generations. The first generation of programs was designed to identify approximate locations of coding regions in genomic DNA. The most widely known programs were probably TestCode and GRAIL. But they could not accurately predict precise exon locations. The second generation, such as SORFIND and Xpound, combined splice signal and coding region identification to predict potential exons, but did not attempt to assemble predicted exons into complete genes. The next generation of programs attempted the more difficult task of predicting complete gene structures. A variety of programs have been developed, including GeneID, GeneParser, GenLang, and FGENEH. However, the performance of those programs remained rather poor. Moreover, those programs were all based on the assumption that the input sequence contains exactly one complete gene, which is not often the case. To solve this problem and improve accuracy and applicability further, GENSCAN and AUGUSTUS were developed, which could be classified into the fourth generation.

GeneMark

GeneMark uses a Markov Chain model to represent the statistics of the coding and noncoding frames. The method uses the dicodon statistics to identify coding regions. Consider the analysis of a sequence x whose base at the ith position is called x_i. The Markov chains used are fifth order, and consist of a terms such as $P(a/x_1x_2x_3x_4x_5)$, which represent the probability of the sixth base of the sequence x being given a given that the previous five bases in the sequence x where $x_1x_2x_3x_4x_5$, resulting in the first dicodon of the sequence being $x_1x_2x_3x_4x_5a$. These terms must be defined for all possible pentamers with the general sequence $b_1b_2b_3b_4b_5$. The values of these terms can be obtained of analysis of data, consisting of nucleotide sequence in which the coding regions have been actually identified. When there are sufficient data, they are given by

$$P(\frac{a}{b_1b_2b_3b_4b_5}) = \frac{n_{b_{1}b_{2}b_{3}b_4b_5a}}{\sum_{a=A,C,G,T} n_{b_{1}b_{2}b_{3}b_4b_5a}}$$

where, $n_{b_{1}b_{2}b_{3}b_4b_5a}$ is the number of times the sequence $b_1b_2b_3b_4b_5a$ occurs in the training data. This is the maximum likelihood estimators of the probability from the training data.

Glimmer

The core of Glimmer is Interpolated Markov Model (IMM), which can be described as a generalized Markov chain with variable order. After GeneMark introduces the fixed-order Markov chains, Glimmer attempts to find a better approach for modeling the genome content. The motivational fact is that the bigger the order of the Markov chain, the more non-randomness can be described. However, as we move to higher order models, the number of probabilities that we must estimate from the data increases exponentially. The major limitation of the fixed-order Markov chain is that models from higher order require exponentially more training data, which are limited and usually not available for new sequences. However, there are some oligomers from higher order that occur often enough to be extremely useful predictors. For the purpose of using these higher-order statistics, whenever sufficient data is available, Glimmer IMMs.

Glimmer calculates the probabilities for all Markov chains from 0th order to 8th. If there are longer sequences (e.g. 8-mers) occurring frequently, IMM makes use of them even when there is insufficient data to train an 8-th order model. Similarly, when the statistics from the 8-th order model do not provide significant information, Glimmer refers to the lower-order models to predict genes.

Opposed to the supervised GeneMark, Glimmer uses the input sequence for training. The ORFs longer than a certain threshold are detected and used for training, because there is high probability that they are genes in prokaryotes. Another training option is to use the sequences with homology to known genes from other organisms, available in public databases. Moreover, the user can decide whether to use long ORFs for training purposes or choose any set of genes to train and build the IMM.

GeneMark.hmm

GeneMark.hmm is designed to improve GeneMark in finding exact gene starts. Therefore, the properties of GeneMark.hmm are complementary to GeneMark. GeneMark.hmm uses GeneMark models of coding and non-coding regions and incorporates them into hidden Markov model framework. In short terms, Hidden Markov Models (HMM) are used to describe the transitions from non-coding to coding regions and vice versa. GeneMark.hmm predicts the most likely structure of the genome using the Viterbi algorithm, a dynamic programming algorithm for finding the most likely sequence of hidden states. To further improve the prediction of translation start position, GeneMark.hmm derives a model of the ribosome binding site (6-7 nucleotides preceding the start codon, which are bound by the ribosome when initiating protein translation). This model is used for refinement of the results.

Both GeneMark and GeneMark.hmm detect prokaryotic genes in terms of identifying open reading frames that contain real genes. Moreover, they both use pre-computed species-specific gene models as training data, in order to determine the parameters of the protein-coding and non-coding regions.

Orpheus

The ORPHEUS program uses homology, codon statistics and ribosome binding sites to improve the methods presented so far by using information that those programs ignored. One of the key differences is that it uses database searches to help determine putative genes, and is thus an extrinsic method. This initial set of genes is used to define the coding statistics for the organism, in this case working at the level of codon, not dicodons. These statistics are then used to define a larger set of candidate ORFs. From this set, those ORFs with an unambiguous start codon end are used to define a scoring matrix for the ribosome-binding site, which is then used to determine the 5` end of those ORFs where alternative start are present.

EcoParse

EcoParse is one of the first HMM model based gene finder, was developed for gene finding in *E.coli*. It focuses on the uses the codon structure of genes. With EcoParse a flora of HMM based gene finder, usuing dynamic programming and the viterbi algorithm to parse a sequence, emerged.

Evaluation of Gene Prediction Programs

In the field of gene prediction accuracy can be measured at three levels

a. Coding nucleotides (base level)

b. Exon structure (exon level)

c. Protein product (protein level)

At base level gene predictions can be evaluated in terms of *true positives (TP)* (predicted features that are real), *true negatives* (TN) (non-predicted features that are not real), *false positives (FP)* (predicted features that are not real), and *false negatives (FN)* (real features that were not predicted) Fig. 5. Usually the base assignment is to be in a coding or non coding segment, but this analysis can be extended to include non coding parts of genes, or any functional parts of the sequences.

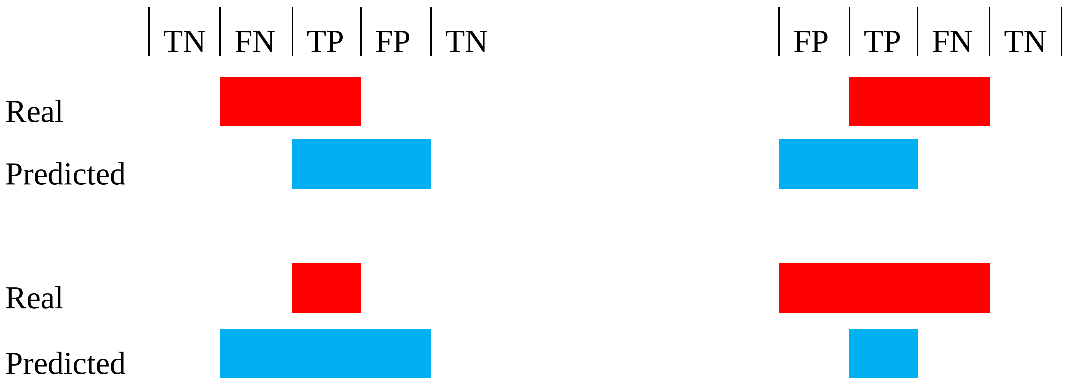


Fig. 5: Four Possible Comparisons of Real and Predicted Genes

Sensitivity (Sn): The fraction of bases in real genes that are correctly predicted to be in genes is the sensitivity and interpreted as the probability of correctly predicting a nucleotide to be in a given gene that it actually is.

$$Sn = \frac{TP}{TP + FN}$$

Specificity (Sp): The fraction of those bases which are predicted to be in genes that actually are is called the specificity and interpreted as the probability of a nucleotide actually being in a gene given that it has been predicted to be.

$$Sp = \frac{TP}{TP + FP}$$

Care has to be taken in using these two values to assess a gene prediction program because, as with the normal definition of specificity, extreme results can make them misleading.

Approximate correlation coefficient (AC) has been proposed as a single measure to circumvent these difficulties. This defined as AC=2(ACP-0.5), where

$$ACP = \frac{1}{n}\left(\frac{TP}{TP + FN} + \frac{TP}{TP + FP} + \frac{TN}{TN + FP} + \frac{TN}{TN + FN}\right),$$

At the exon level, determination of prediction accuracy depends on the exact prediction of exon start and end points. There are two measures of sensitivity and specificity used in the field, each of which measures a different but useful property.

The sensitivity measures used are

S_{n1} = CE/AE and Sn2 = ME/AE

The specificity measures used are

S_{p1}=CE/PE and S_{p2}=WE/PE

Where,

AE = No of actual exons in the data

PE = No of predicted exons in the data

CE = No of correct predicted exons

ME = No of missing exons (rarely occurs)

WE = No of wrongly predicted exons (Figure-5)

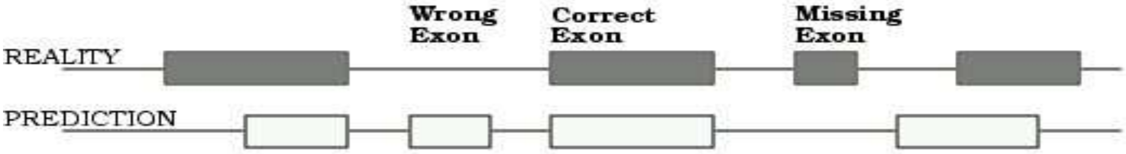


Fig. 6: Real and Predicted Exons

Gene Ontology

The gene ontology (GO, http:www.geneontology.org) is probably the most extensive scheme today for the description of gene product functions but other systems such as enzyme codes, KEGG pathways, FunCat, or COG are also widely used. Here, we describe the Blast2GO (B2G, www.blast2go.org) application for the functional annotation, management, and data mining of novel sequence data through the use of common controlled vocabulary schemas. The main application domain of the tool is the functional genomics of non-model organisms and it is primarily intended to support research in experimental labs. Blast2GO strives to be the application of choice for the annotation of novel sequences in functional genomics projects where thousands of fragments need to be characterized. Functional annotation in Blast2GO is based on homology transfer. Within this framework, the actual annotation procedure is configurable and permits the design of different annotation strategies. Blast2GO annotation parameters include the choice of search database, the strength and number of blast results, the extension of the query-hit match, the quality of the transferred annotations, and the inclusion of motif annotation. Vocabularies supported by B2G are gene ontology terms, enzyme codes (EC), InterPro IDs, and KEGG pathways.

Fig.7 shows the basic components of the Blast2GO suite. Functional assignments proceed through an elaborate annotation procedure that comprises a central strategy plus refinement functions. Next, visualization and data mining engines permit exploiting the annotation results to gain functional knowledge. GO annotations are generated through a 3-step process: blast, mapping, annotation. InterPro terms are obtained from InterProScan at EBI, converted and merged to GOs. GO annotation can be modulated from Annex, GOSlim web services and manual editing. EC and KEGG annotations are generated from GO. Visual tools include sequence color code, KEGG pathways, and GO graphs with node highlighting and filtering options. Additional annotation data-mining tools include statistical charts and gene set enrichment analysis functions.

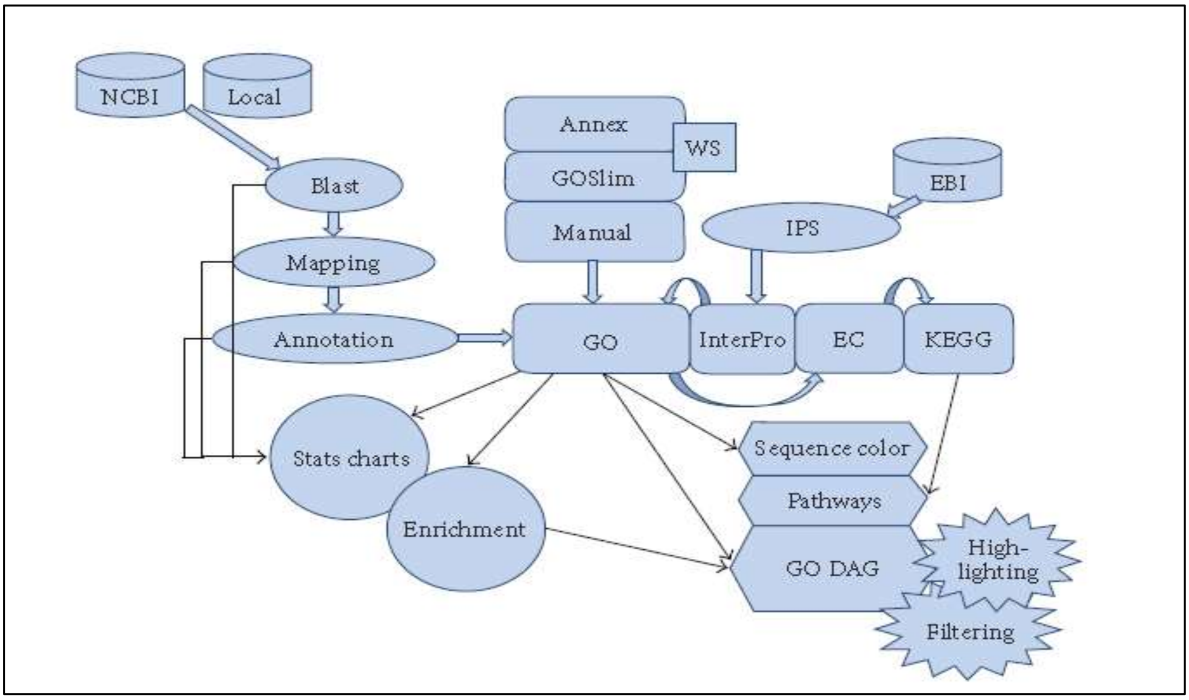


Fig. 7: Schematic Representation of Blast2GO Application.

The Blast2GO annotation procedure consists of three main steps: blast to find homologous sequences, mapping to collect GO terms associated to blast hits, and annotation to assign trustworthy information to query sequences.

Blast Step

The first step in B2G is to find sequences similar to a query set by blast. B2G accepts nucleotide and protein sequences in FASTA format and supports the four basic blast programs (blastx, blastp, blastn, and tblastx). Homology searches can be launched against public databases such as (the) NCBI nr using a query-friendly version of blast (QBlast). This is the default option and in this case, no additional installations are needed. Alternatively, blast can be run locally against a proprietary FASTA-formatted database, which requires a working www-blast installation. The Make Filtered Blast-GO-BD function in the Tools menu allows the creation of customized databases containing only GO annotated entries, which can be used in combination with the local blast option. Other configurable parameters at the blast step are the expectation value (e-value) threshold, the number of retrieved hits, and the minimal alignment length (hsp length) which permits the exclusion of hits with short, low e-value matches from the sources of functional terms. Annotation, however, will ultimately be based on sequence similarity levels as similarity percentages are independent of database size and more intuitive than e-values. Blast2GO parses blast results and presents the information for each sequence in table format.

Query sequence descriptions are obtained by applying a language processing algorithm to hit descriptions, which extracts informative names and avoids low content terms such as “hypothetical protein” or “expressed protein”.

Mapping Step

Mapping is the process of retrieving GO terms associated to the hits obtained after a blast search. B2G performs three different mappings as follows.

a. Blast result accessions are used to retrieve gene names (symbols) making use of two mapping files provided by NCBI (geneinfo, gene2accession). Identified gene names are searched in the species-specific entries of the gene product table of the GO database.

b. Blast result GI identifiers are used to retrieve UniProt IDs making use of a mapping file from PIR (Non-redundant Reference Protein database) including PSD, UniProt, Swiss-Prot, TrEMBL, RefSeq, GenPept, and PDB.

c. Blast result accessions are searched directly in the DBXRef Table of the GO database.

Annotation Step

This is the process of assigning functional terms to query sequences from the pool of GO terms gathered in the mapping step. Function assignment is based on the gene ontology vocabulary. Mapping from GO terms to enzyme codes permits the subsequent recovery of enzyme codes and KEGG pathway annotations. The B2G annotation algorithm takes into consideration the similarity between query and hit sequences, the quality of the source of GO assignments, and the structure of the GO DAG. For each query sequence and each candidate GO term, an annotation score (AS) is computed (see Figure 8). The AS is composed of two terms. The first, direct term (DT), represents the highest similarity value among the hit sequences bearing this GO term, weighted by a factor corresponding to its evidence code (EC). A GO term EC is present for every annotation in the GO database to indicate the procedure of functional assignment.

$$
\begin{aligned}
&\mathrm{DT} = \max\,(\mathrm{similarity} \times \mathrm{EC}_{\mathrm{weight}}) \\
&\mathrm{AT} = (\#\mathrm{GO} - 1) \times \mathrm{GO}_{\mathrm{weight}} \\
&\mathrm{AR} : \mathrm{lowest.node}(\mathrm{AS}(\mathrm{DT} + \mathrm{AT}) \geq \mathrm{threshold})
\end{aligned}
$$

Fig. 8: Blast2GO Annotation Rule

ECs vary from experimental evidence, such as inferred by direct assay (IDA) to unsupervised assignments such as inferred by electronic annotation (IEA). The second term (AT) of the annotation rule introduces the possibility of abstraction into the annotation algorithm. Abstraction is defined as the annotation to a parent node when several child nodes are present in the GO candidate pool. This term multiplies the number of total GOs unified at the node by a user defined factor or GO weight (GOw) that controls the possibility and strength of abstraction. When all ECw's are set to 1 (no EC control) and the GOw is set to 0 (no abstraction is possible), the annotation score of a given GO term equals the highest similarity value among the blast hits annotated with that term. If the ECw is smaller than one, the DT decreases and higher query-hit similarities are required to surpass the annotation threshold. If the GOw is not equal to zero, the AT becomes contributing and the annotation of a parent node is possible if multiple child nodes coexist that do not reach the annotation cutoff. Default values of B2G annotation parameters were chosen to optimize the ratio between annotation coverage and annotation accuracy. Finally, the AR selects the lowest terms per branch that exceed a user-defined threshold.

Blast2GO includes different functionalities to complete and modify the annotations obtained through the above-defined procedure. Enzyme codes and KEGG pathway annotations are generated from the direct mapping of GO terms to their enzyme code equivalents. Additionally, Blast2GO offers InterPro searches directly from the B2G interface. B2G launches sequence queries in batch, and recovers, parses, and uploads InterPro results. Furthermore, InterPro IDs can be mapped to GO terms and merged with blast-derived GO annotations to provide one integrated annotation result. In this process, B2G ensures that only the lowest term per branch remains in the final annotation set, removing possible parent-child relationships originating from the merging action.

References

1. Conesa, S. Gotz, J. M. Garcia-Gomez, J. Terol, M. Talon, and M. Robles, “Blast2GO: a universal tool for annotation, visualization and analysis in functional genomics research,” Bioinformatics, vol. 21, no. 18, pp. 3674–3676, 2005.
2. Conesa and S. Gotz, "Blast2GO: A Comprehensive Suite for Functional Analysis in Plant Genomics," International Journal of Plant Genomics, vol. 2008, 2008.
3. H. Ogata, S. Goto, K. Sato, W. Fujibuchi, H. Bono, and M. Kanehisa, “KEGG: Kyoto Encyclopedia of Genes and Genomes,” Nucleic Acids Research, vol. 27, no. 1, pp. 29–34, 1999.
4. J.D. Watson, R.M. Myers, A.A. Caudy and J.A. Witkowski, "Recombinant DNA: Genes and Genomes - A Short Course," 3rd Ed., 2007.
5. M. Ashburner, C. A. Ball, J. A. Blake, et al., “Gene Ontology: tool for the unification of biology. The Gene Ontology Consortium,” Nature Genetics, vol. 25, no. 1, pp. 25–29, 2000.

6. Ruepp, A. Zollner, D. Maier, et al., “The FunCat, a functional annotation scheme for systematic classification of proteins from whole genomes,” Nucleic Acids Research, vol. 32, no. 18, pp. 5539–5545, 2004.
7. R. L. Tatusov, N. D. Fedorova, J. D. Jackson, et al., “The COG database: an updated version includes eukaryotes,” BMC Bioinformatics, vol. 4, p. 41, 2003.
8. Schomburg, A. Chang, C. Ebeling, et al., “BRENDA, the enzyme database: updates and major new developments,” Nucleic Acids Research, vol. 32, Database issue, pp. D431–D433, 2004.
9. S. F. Altschul, W. Gish, W. Miller, E. W. Myers, and D. J. Lipman, “Basic local alignment search tool,” Journal of Molecular Biology, vol. 215, no. 3, pp. 403–410, 1990.
10. S. Myhre, H. Tveit, T. Mollestad, and A. Lægreid, “Additional Gene Ontology structure for improved biological reasoning,” Bioinformatics, vol. 22, no. 16, pp. 2020–2027, 2006.

Practical Hands-on:Transcriptome Data Analysis and Annotation

Sneha Murmu
Division of Agricultural Bioinformatics
ICAR-Indian Agricultural Statistics Research Institute, New Delhi

The current ecosystems of RNA-seq tools provide a varied ways analyzing RNA-seq data. Depending on the experiment goal one could align the reads to reference genome or pseduoalign to transcriptome and perform quantification and differential expression of genes or if you want to annotate your reference, assemble RNA-seq reads using a *de novo* transcriptome assembler. In this lecture, we focus on workflows that align reads to reference genomes using updated Tuxedo protocol (HISAT, StringTie, Ballgown) by Pertea et al. This updated Tuxedo protocol not only scales but is more accurate in detecting differentially expressed genes (DEGs). Lastly, we used Blast2GO for annotating the identified DEGs.

In this example, we have used the example data which is mentioned in the paper. Before starting with the actual workflow, we have briefly mentioned the steps required to set up the system.

1) Setting up the system for differential expression analysis of transcriptome data

```
#for windows system, install linux via wsl.

#install anaconda in linux (Ubuntu)

#open ubuntu terminal

$ wget https://repo.anaconda.com/archive/Anaconda3-2022.10-Linux-x86_64.sh

$ bash Anaconda3-2022.10-Linux-x86_64.sh

#set up the conda environment

$ conda env create -f environment_1.yaml

$ conda activate rnaseq_py3

# Set up complete!
```

1. **Protocol:**

```
###Align the data to the reference genome using HISAT2

##build index

(rnaseq_py3) root@DESKTOP-BJ5B6HR:/mnt/e/iasri/dabin_training/Nov2022/practical/example# mkdir index

(rnaseq_py3) root@DESKTOP-BJ5B6HR:/mnt/e/iasri/dabin_training/Nov2022/practical/example# extract_splice_sites.py resources/chrX.gtf > index/chrX.ss
```

```
(rnaseq_py3) root@DESKTOP-BJ5B6HR:/mnt/e/iasri/dabin_training/Nov2022/practical/example# extract_exons.py resources/chrX.gtf > index/chrX.exon
```

```
(rnaseq_py3) root@DESKTOP-BJ5B6HR:/mnt/e/iasri/dabin_training/Nov2022/practical/example# cd index
```

```
(rnaseq_py3) root@DESKTOP-BJ5B6HR:/mnt/e/iasri/dabin_training/Nov2022/practical/example/index# hisat2-build -p 8 --ss chrX.ss --exon chrX.exon ../resources/chrX.fa chrX_tran
```

```
(rnaseq_py3) root@DESKTOP-BJ5B6HR:/mnt/e/iasri/dabin_training/Nov2022/practical/example/index# cd ..
```

##1. mapping

```
$ fastqdir=resources/samples

mapdir=mapped

mkdir $mapdir

hisat2 -p 8 --dta -x index/chrX_tran -1 $fastqdir/ERR188044_chrX_1.fastq.gz -2 $fastqdir/ERR188044_chrX_2.fastq.gz -S $mapdir/ERR188044.sam
```

##2. sort mapped files

```
$ mapdir=mapped

samtools sort -@ 8 -o $mapdir/ERR188044.bam $mapdir/ERR188044.sam
```

##3. assembly

```
gtf=resources/chrX.gtf

assembly=assembly

mapdir=mapped

mkdir $assembly

stringtie $mapdir/ERR188044.bam -l ERR188044 -p 8 -G $gtf -o $assembly/ERR188044.gtf
```

##obtain list of each sample .gtf file in a single file (mergelist.txt)

```
$ ls assembly/*.gtf > mergelist.txt
```

##merge .gtf file of each sample

```
$ stringtie --merge -p 8 -G resources/chrX.gtf -o stringtie_merged.gtf mergelist.txt
```

##obtain sequences of transcripts

```
$ gffread -w transcripts.fa -g resources/chrX.fa stringtie_merged.gtf
```

##compare merged.gtf file with reference .gtf file

```
$ gffcompare -r resources/chrX.gtf -G -o merged stringtie_merged.gtf
```

##4. abundance estimation

```
$ abundancedir=abundance

mapdir=mapped

stringtie -e -B -p 8 -G stringtie_merged.gtf -o $abundancedir/ERR188044/ERR188044_chrX.gtf $mapdir/ERR188044.bam
```

2. Differential expression analysis

Open R console.

```
#Differential expression

#load the libraries

library(ggplot2)

library(ballgown)

library(genefilter)

library(RSkittleBrewer)

library(devtools)

library(dplyr)

library(ggrepel)

library(pheatmap)

library(gplots)

library(GenomicRanges)

library(viridis)

#lets load the sample information

pheno_data <- read.csv("resources/geuvadis_phenodata.csv")
```

```
#let's show information for first 6 samples

head(pheno_data)

#Load the expression data using ballgown

bg_chrX <- ballgown(dataDir="abundance",samplePattern="ERR",pData=pheno_data)

#Lets filter out transcripts with low variance

#This is done to remove some genes that have few counts. Filtering improves the statistical
power of differential expression analysis.

#We use variance filter to remove transcripts with low variance( 1 or less)

bg_chrX_filt<- subset(bg_chrX,"rowVars(texpr(bg_chrX))>1",genomesubset=TRUE)

#Let's test on transcripts

de_transcripts <-
stattest(bg_chrX_filt,feature="transcript",covariate="conditions",getFC=TRUE,meas="FPKM")

# the results_transcripts does not contain identifiers. We will therefore add this information

#add identifiers

de_transcripts = data.frame(geneNames=ballgown::geneNames(bg_chrX_filt),
geneIDs=ballgown::geneIDs(bg_chrX_filt), de_transcripts)

# Let's test on genes

de_genes <- stattest(bg_chrX_filt,feature="gene",covariate="conditions",getFC=TRUE,
meas="FPKM")

#lets get the gene names

bg_filt_table=texpr(bg_chrX_filt,'all')

gene_names=unique(bg_filt_table[,9:10])

features=de_genes$id

mapped_gene_names=vector()

for (i in features)

{  query=gene_names%>%filter(gene_id==i & gene_name != '.') ; n_hit=dim(query)[1]; if
(n_hit==1) {mapped_gene_names=append(mapped_gene_names,query$gene_name[[1]]) }
else

{mapped_gene_names=append(mapped_gene_names,'.') }
}

#add the mapped gene names to the de genes table

de_genes$gene_name <- mapped_gene_names

de_genes <- de_genes[, c('feature','gene_name','id','fc','pval','qval')]

de_genes[,"log2fc"] <- log2(de_genes[,"fc"])

de_transcripts[,"log2fc"] <- log2(de_transcripts[,"fc"])

#Let's arrange the results from the smallest P value to the largest

de_transcripts = arrange(de_transcripts,pval)

de_genes = arrange(de_genes,pval)

#save result in .csv

write.csv(de_genes, "de_transcripts.csv", row.names=FALSE)

write.csv(de_genes, "de_genes.csv", row.names=FALSE)

#Let's subset transcripts that are detected as differentially expressed at qval <0.05

subset_transcripts <- subset(de_transcripts,de_transcripts$qval<0.05)

#do same for the genes

subset_genes <- subset(de_genes,de_genes$qval<0.05)

#create plots

dir.create('plots')

print('generating plots')

#volcano plot

#https://biocorecrg.github.io/CRG_RIntroduction/volcano-plots.html

de_genes$diffexpressed <- "NO"

de_genes$diffexpressed[de_genes$log2fc > 1 & de_genes$pval < 0.05] <- "UP"

de_genes$diffexpressed[de_genes$log2fc < -1 & de_genes$pval < 0.05] <- "DOWN"

de_genes$delabel <- NA

de_genes$delabel[de_genes$diffexpressed != "NO"] <- de_genes$id[de_genes$diffexpressed != "NO"]
```

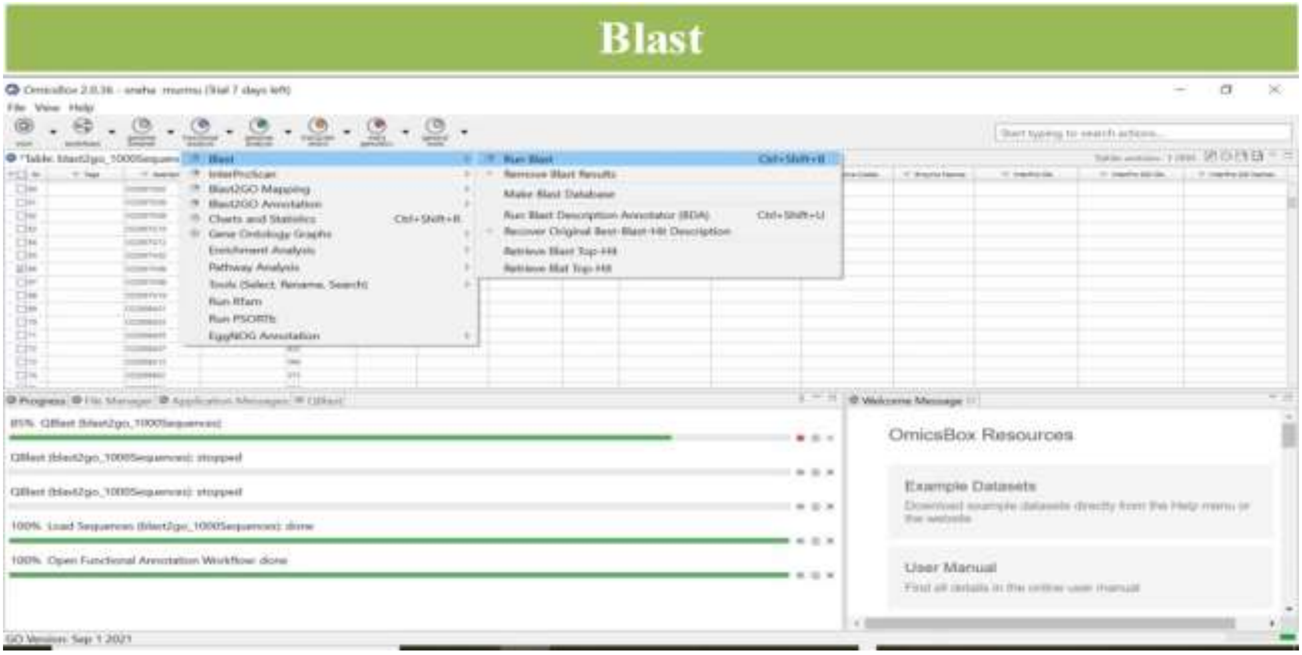
```
options(ggrepel.max.overlaps = Inf)
png('plots/volcano.png',width = 1800, height = 1000) #,width = 1800, height = 1000
volcano=ggplot(data=de_genes, aes(x=log2fc, y=-log10(pval), col=diffexpressed, label=delabel)) +
  geom_point() +
  theme_minimal() +
  geom_text_repel() +
  scale_color_manual(values=c("blue", "black", "red")) +
  geom_vline(xintercept=c(-0.8, 0.8), col="red") +
  theme(text=element_text(size=20))

  #geom_hline(yintercept=-log10(0.05), col="red")
print(volcano)
dev.off()
#DONE
#MAPLOT
#https://davetang.org/muse/2017/10/25/getting-started-hisat-stringtie-ballgown/
png('plots/maplot.png',width = 1800, height = 1000)
de_transcripts$mean <- rowMeans(texpr(bg_chrX_filt))
maplot=ggplot(de_transcripts, aes(log2(mean), log2(fc), colour = qval<0.05)) +
  scale_color_manual(values=c("#999999", "#FF0000")) +
  geom_point() +
  theme(legend.text=element_text(size=20),legend.title=element_text(size=20)) +
  theme(axis.text=element_text(size=20),axis.title=element_text(size=20)) +
  geom_hline(yintercept=0)
print(maplot)
dev.off()
```

```
#DONE

Exit R.

##extract DE transcript sequence by ID

gffread -w transcripts.fa -g chrX.fa stringtie_merged.gtf

#create index of transc.fa

cdbfasta transcripts.fa

cat up17_id_list.txt |cdbyank transcripts.fa.cidx > up17.fasta
```

3. Annotation

Functional annotation is defined as the process of collecting information about and describing a gene's biological identity—its various aliases, molecular function, biological role(s), subcellular location, and its expression domains within the plant. Blast2GO is a bioinformatics platform for high-quality functional annotation and analysis of genomic datasets. The following section mentions the four major modules involved in Blast2GO annotation.

A) Basic Local Alignment Search Tool: to search for similar (or homologous) sequences as shown in Fig 1.



85% QBlast (blast2go_1000Sequences)

QBlast (blast2go_1000Sequences): stopped

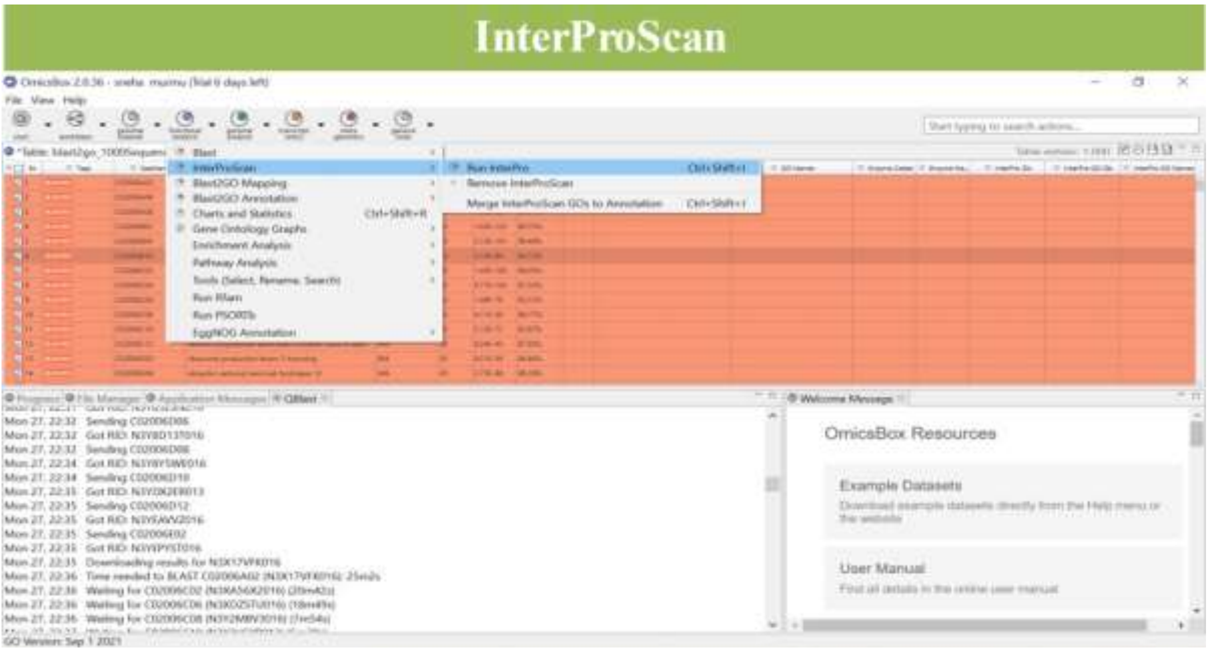
QBlast (blast2go_1000Sequences): stopped

100% Load Sequences (blast2go_1000Sequences): done

100% Open Functional Annotation Workflow: done

Figure 1: BLAST

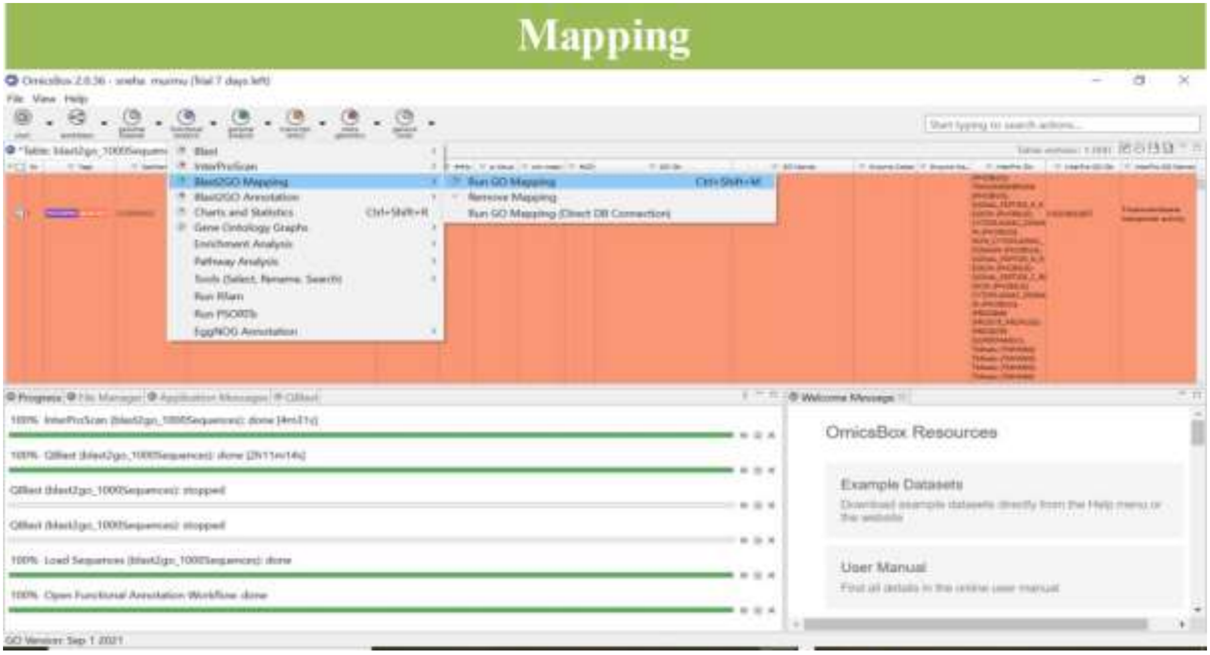
B) InterProScan: for classification of protein families as shown in Fig 2.


InterProScan
OmicsBox 2.0.36 - sneha murmu (Trial 6 days left)
File View Help
Blast
InterProScan
Blast2GO Mapping
Blast2GO Annotation
Charts and Statistics Ctrl+Shift+R
Gene Ontology Graphs
Enrichment Analysis
Pathway Analysis
Tools (Select, Rename, Search)
Run Rfam
Run PSORTb
EggNOG Annotation
Run InterPro Ctrl+Shift+I
Remove InterProScan
Merge InterProScan GOs to Annotation Ctrl+Shift+I
Progress
File Manager
Application Messages
CBlast
Welcome Message
OmicsBox Resources
Example Datasets
Download example datasets directly from the Help menu or the website
User Manual
Find all details in the online user manual

```
Mon 27, 22:32   Sending C02006D06
Mon 27, 22:32   Got RID: N3YBD137016
Mon 27, 22:32   Sending C02006D08
Mon 27, 22:34   Got RID: N3YBYSWE016
Mon 27, 22:34   Sending C02006D10
Mon 27, 22:35   Got RID: N3YDK2ER013
Mon 27, 22:35   Sending C02006D12
Mon 27, 22:35   Got RID: N3YEAWV2016
Mon 27, 22:35   Sending C02006E02
Mon 27, 22:35   Got RID: N3YEPYST016
Mon 27, 22:35   Downloading results for N3X17VFK016
Mon 27, 22:36   Time needed to BLAST C02006A02 (N3X17VFK016): 25m2s
Mon 27, 22:36   Waiting for C02006C02 (N3KA56X2016) (20m42s)
Mon 27, 22:36   Waiting for C02006C06 (N3KDZSTU016) (18m49s)
Mon 27, 22:36   Waiting for C02006C08 (N3Y2MBV3016) (7m54s)
GO Version: Sep 1 2021
```

Fig 2: InterProScan

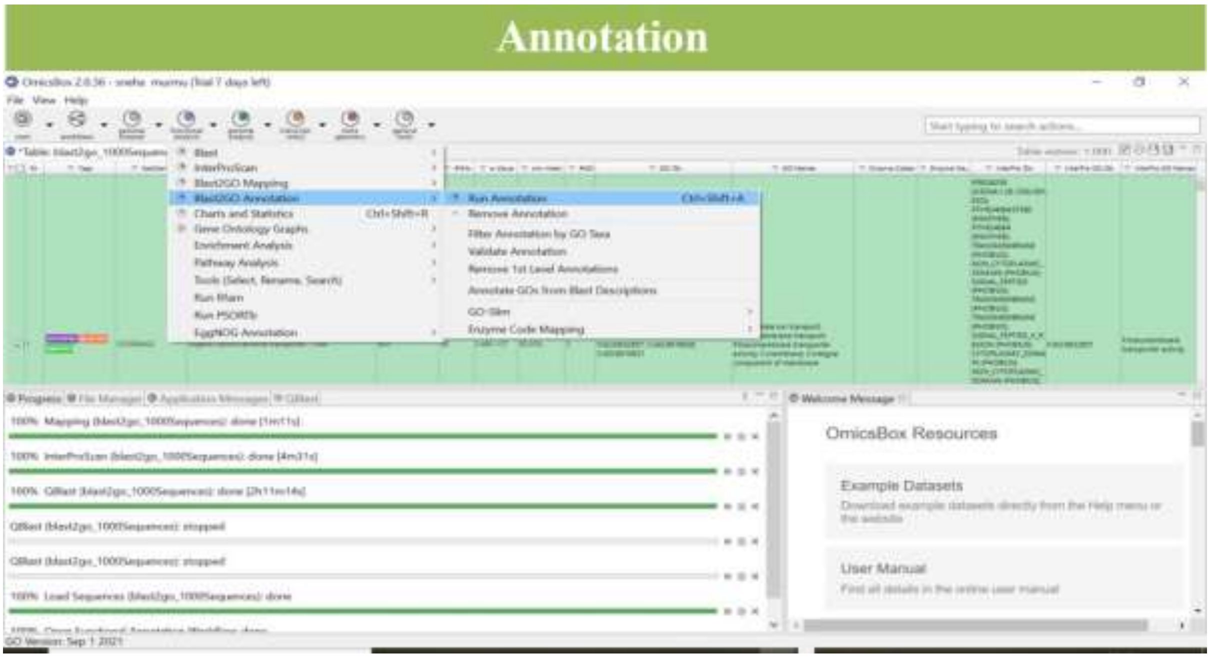
C) Blast2GO Mapping: to retrieve Gene Ontology (GO) terms as shown in Fig 3.


Mapping
OmicsBox 2.0.36 - sneha murmu (Trial 7 days left)
File View Help
Blast
InterProScan
Blast2GO Mapping
Blast2GO Annotation
Charts and Statistics Ctrl+Shift+R
Gene Ontology Graphs
Enrichment Analysis
Pathway Analysis
Tools (Select, Rename, Search)
Run Rfam
Run PSORTb
EggNOG Annotation
Run GO Mapping Ctrl+Shift+M
Remove Mapping
Run GO Mapping (Direct DB Connection)
Progress
File Manager
Application Messages
CBlast
Welcome Message
OmicsBox Resources
Example Datasets
Download example datasets directly from the Help menu or the website
User Manual
Find all details in the online user manual

```
100%  InterProScan (blast2go_1000Sequences): done [4m31s]
100%  CBlast (blast2go_1000Sequences): done [2h11m14s]
CBlast (blast2go_1000Sequences): stopped
CBlast (blast2go_1000Sequences): stopped
100%  Load Sequences (blast2go_1000Sequences): done
100%  Open Functional Annotation Workflow: done
GO Version: Sep 1 2021
```

Fig 3: Mapping

D) Blast2GO Annotation: to select reliable functions as shown in Fig 4.



Annotation

OmicsBox 2.0.36 - sneha murmu (Trial 7 days left)
File View Help
Blast
InterProScan
Blast2GO Mapping
Blast2GO Annotation
Charts and Statistics Ctrl+Shift+R
Gene Ontology Graphs
Enrichment Analysis
Pathway Analysis
Tools (Select, Rename, Search)
Run Rfam
Run PSORTb
EggNOG Annotation
Run Annotation Ctrl+Shift+A
Remove Annotation
Filter Annotation by GO Taxa
Validate Annotation
Remove 1st Level Annotations
Annotate GOs from Blast Descriptions
GO-Slim
Enzyme Code Mapping
Progress
File Manager
Application Messages
CloudBlast
Welcome Message
OmicsBox Resources
Example Datasets
Download example datasets directly from the Help menu or the website
User Manual
Find all details in the online user manual

100% Mapping (blast2go_1000Sequences): done [1m11s]

100% InterProScan (blast2go_1000Sequences): done [4m31s]

100% CBlast (blast2go_1000Sequences): done [2h11m14s]

CBlast (blast2go_1000Sequences): stopped

CBlast (blast2go_1000Sequences): stopped

100% Load Sequences (blast2go_1000Sequences): done

GO Version: Sep 1 2021

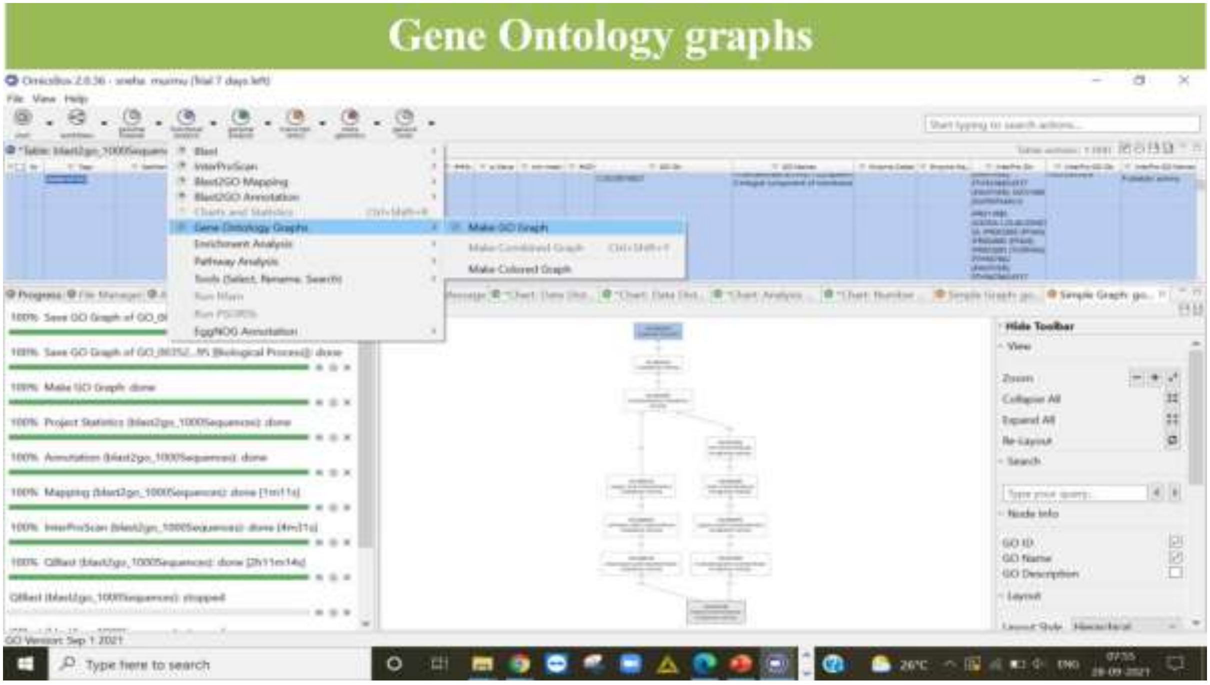
Fig 4: Annotation

Result of Blast2GO:

The result can be visualized in the following forms:

a) Gene Ontology graphs (as shown in Fig 5)

b) Pathway analysis (as shown in Fig 6)



Gene Ontology graphs

100% Save GO Graph of GO_0...

100% Save GO Graph of GO_00352...% (Biological Process): done

100% Make GO Graph: done

100% Project Statistics (blast2go_1000Sequences): done

100% Annotation (blast2go_1000Sequences): done

100% Mapping (blast2go_1000Sequences): done [1m11s]

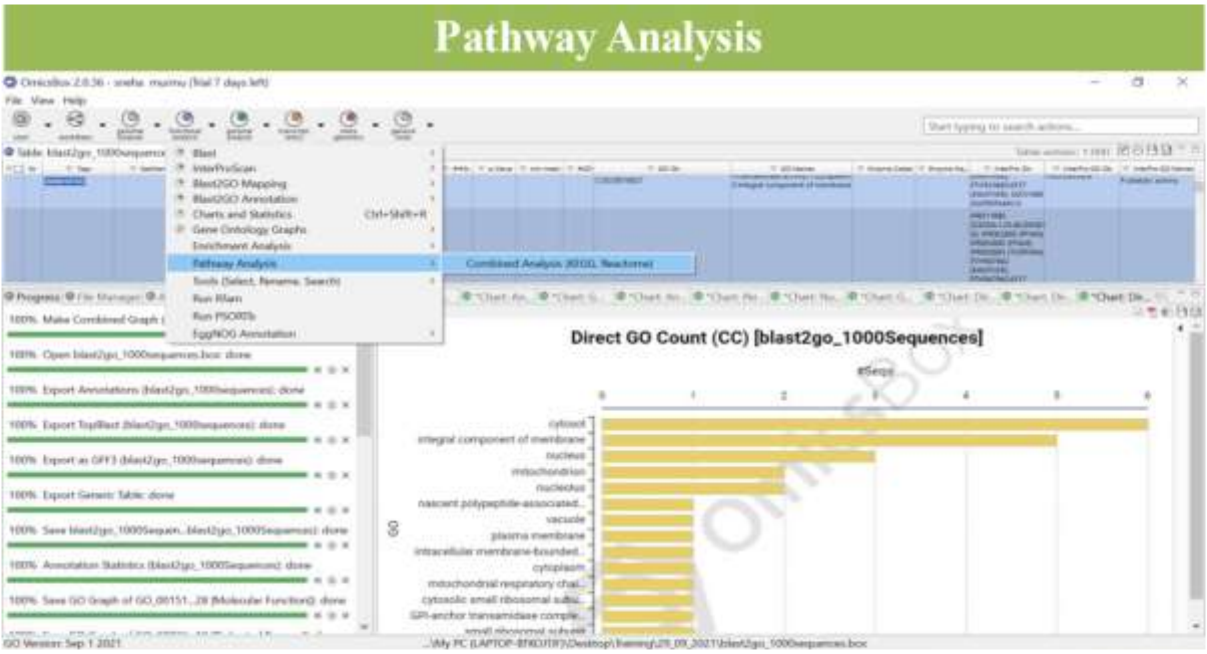
100% InterProScan (blast2go_1000Sequences): done [4m31s]

100% CBlast (blast2go_1000Sequences): done [2h11m14s]

CBlast (blast2go_1000Sequences): stopped

GO Version: Sep 1 2021

Fig 5. Gene Ontology graphs


Pathway Analysis
OmicsBox 2.0.36 - sneha murmu (trial 7 days left)
File View Help
Table: blast2go_1000sequences
Blast
InterProScan
Blast2GO Mapping
Blast2GO Annotation
Charts and Statistics Ctrl+Shift+R
Gene Ontology Graphs
Enrichment Analysis
Pathway Analysis
Combined Analysis (KEGG, Reactome)
Tools (Select, Rename, Search)
Run Rfam
Run PSORTb
EggNOG Annotation
Progress
File Manager
Direct GO Count (CC) [blast2go_1000Sequences]
#Seqs
0
1
2
3
4
5
6
GO
cytosol
integral component of membrane
nucleus
mitochondrion
nucleolus
nascent polypeptide-associated...
vacuole
plasma membrane
intracellular membrane-bounded...
cytoplasm
mitochondrial respiratory chai...
cytosolic small ribosomal subu...
GPI-anchor transamidase comple...
small ribosomal subunit
GO Version: Sep 1 2021
...\My PC (LAPTOP-8NKJ7JF)\Desktop\Training\29_09_2021\blast2go_1000sequences.box

```
100% Make Combined Graph (
100% Open blast2go_1000sequences.box: done
100% Export Annotations (blast2go_1000sequences): done
100% Export TopBlast (blast2go_1000sequences): done
100% Export as GFF3 (blast2go_1000sequences): done
100% Export Generic Table: done
100% Save blast2go_1000Sequen...blast2go_1000Sequences): done
100% Annotation Statistics (blast2go_1000Sequences): done
100% Save GO Graph of GO_00151...28 (Molecular Function): done
```

Fig 6: Pathway Analysis

References:

Pertea, M., Kim, D., Pertea, G. M., Leek, J. T., & Salzberg, S. L. (2016). Transcript-level expression analysis of RNA-seq experiments with HISAT, StringTie and Ballgown. Nature protocols, 11(9), 1650-1667.

Integrated Analysis of multiOMICS Data to Predict the Role of miRNAs

Indra Singh
Inovarion, Paris, France

MicroRNAs (miRNAs) represent a class of small non-coding RNAs that are playing diverse and pivotal roles in the post-transcriptional regulation of gene expression across various organisms. miRNAs are 18-23 nucleotide-long molecules. miRNAs are involved in various fundamental biological processes such as development, differentiation, apoptosis, and metabolism, highlighting their significance in cellular homeostasis and organismal development. Dysregulation of miRNA expression has been implicated in various diseases, including cancer, cardiovascular disorders, and neurological conditions, underscoring their potential as diagnostic biomarkers and therapeutic targets. Additionally, miRNAs exhibit evolutionary conservation, with many miRNA families being conserved across species, reflecting their essential roles in gene regulation and organismal evolution. Overall, miRNAs represent key players in the intricate regulatory networks governing gene expression and cellular function, with profound implications for both basic research and clinical applications.

1. **Computational tools for miRNA prediction and characterization**

miRBase: A comprehensive database for miRNA sequences and annotations. It serves as a valuable resource for comparing and validating predicted miRNAs.
miRDeep: A widely-used tool for the prediction of novel miRNAs from small RNA sequencing data. It integrates secondary structure analysis, sequence conservation, and machine learning algorithms for accurate prediction.
miRanda: This tool predicts miRNA target sites by examining sequence complementarity between miRNAs and potential target mRNAs. It considers both sequence complementarity and conservation across species.
TargetScan: A popular tool for miRNA target prediction, TargetScan predict miRNA target sites based on seed sequence matches, site accessibility, and evolutionary conservation.
RNAhybrid: A tool for predicting the hybridization energy and the minimum free energy (MFE) of RNA-RNA duplexes, commonly used to predict potential miRNA-mRNA interactions.
PITA (Probability of Interaction by Target Accessibility): This tool predicts miRNA target sites based on thermodynamic stability and target site accessibility, offering a probabilistic framework for target prediction.
miRDeep2: An updated version of miRDeep, miRDeep2 integrates small RNA sequencing data with genomic information to predict both known and novel miRNAs with improved accuracy.
miRPlant: Specifically designed for plant miRNA prediction, miRPlant incorporates features such as sequence conservation, secondary structure, and thermodynamic stability to identify potential miRNA candidates in plant genomes.
ShortStack: This tool integrates multiple small RNA sequencing data sets to identify and characterize miRNAs, including novel miRNAs and their targets, with a focus on plant species.
psRNATarget: A plant-specific tool for predicting miRNA targets, psRNATarget considers various factors such as target site accessibility and conservation across species to provide accurate predictions.

2. **General workflow for miRNA prediction**

i. Data retrieval:

Obtain small RNA sequencing data from the organism of interest. This data can be generated from high-throughput sequencing platforms such as Illumina or Ion Torrent.

ii. Quality Control:

Perform quality control on the raw sequencing data to remove low-quality reads, adaptors, and contaminants. Tools like FastQC can be used for this purpose.

iii. Pre-processing:

Trim adapter sequences and filter out reads of inappropriate length. Tools such as Cutadapt or Trimmomatic can be used for this step.

iv. Mapping to Reference Genome:

Map the pre-processed reads to the reference genome or transcriptome using alignment tools like Bowtie, BWA, or HISAT.

v. miRNA Identification:

Use miRNA prediction tools such as miRDeep, miRDeep2, or miRPlant to identify potential miRNA candidates. These tools integrate various features such as sequence conservation, secondary structure, and thermodynamic stability to predict miRNAs.

vi. Novel miRNA Prediction:

Identify novel miRNAs by comparing predicted miRNAs with known miRNA sequences from databases like miRBase. Tools like miRDeep2 and ShortStack often include modules for predicting novel miRNAs.

vii. Target Prediction:

Predict miRNA target genes using tools like miRanda, TargetScan, or psRNATarget. These tools analyze sequence complementarity between miRNAs and potential target mRNAs, considering factors such as seed sequence matches, site accessibility, and evolutionary conservation.

viii. Functional Annotation:

Annotate predicted target genes to elucidate their biological functions and pathways. Tools such as DAVID, GO enrichment analysis, or KEGG pathway analysis can be used for functional annotation.

ix. Experimental Validation:

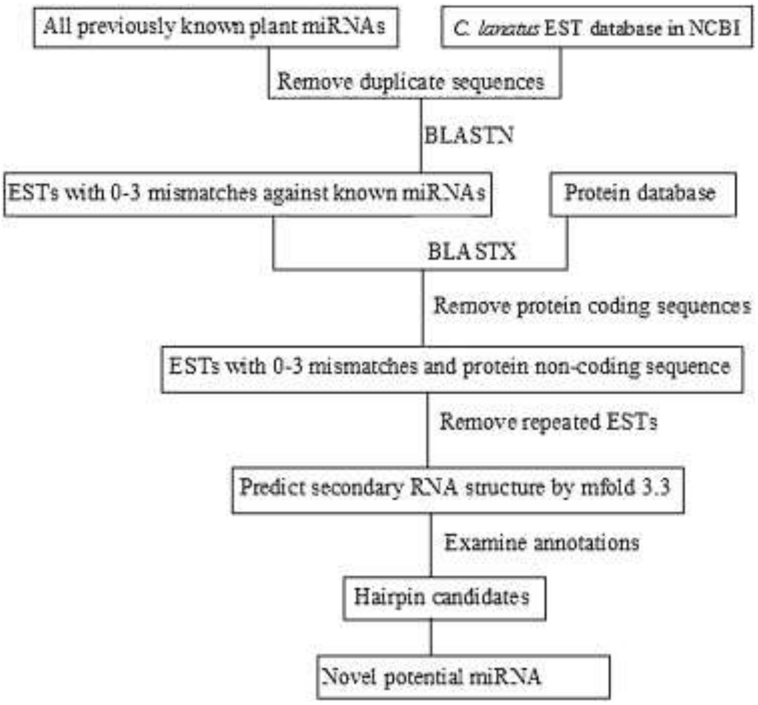
Experimentally validate predicted miRNAs and their targets using techniques such as qRT-PCR, luciferase reporter assays, or functional studies in cell lines or model organisms.

x. Integration and Visualization:

Integrate miRNA prediction results with other omics data (e.g., mRNA expression data, proteomics data) to gain a comprehensive understanding of miRNA-mediated regulatory networks. Visualization tools such as Cytoscape can be used to visualize miRNA-mRNA interaction networks.

xi. Validation and Interpretation:

Validate predicted miRNAs and their targets using experimental techniques. Interpret the results in the context of the biological system under study and generate hypotheses for further investigation.


All previously known plant miRNAs
C. lanatus EST database in NCBI
Remove duplicate sequences
BLASTN
ESTs with 0-3 mismatches against known miRNAs
Protein database
BLASTX
Remove protein coding sequences
ESTs with 0-3 mismatches and protein non-coding sequence
Remove repeated ESTs
Predict secondary RNA structure by mfold 3.3
Examine annotations
Hairpin candidates
Novel potential miRNA

Procedure of novel potential miRNA prediction by identifying homologs of previously known miRNAs in plants (Zakeel et al., 2019)

3 Integrated analysis of multiOMICS data

Integrating miRNA data into genomics, transcriptomics, proteomics, and other -omics data sets is crucial for a comprehensive understanding of gene regulation and cellular processes. These are some key applications of miRNA data integration:

Regulatory Network Reconstruction:

Integration of miRNA data allows for the reconstruction of regulatory networks encompassing miRNAs, mRNAs, and proteins. This holistic view enables researchers to unravel complex regulatory interactions governing cellular processes.Integrating miRNA data with mRNA expression profiles facilitates the identification of miRNA targets. By correlating changes in miRNA expression with alterations in mRNA abundance, putative miRNA-target interactions can be inferred.

Functional Annotation:

Integrating miRNA data with functional annotation databases (e.g., Gene Ontology, KEGG pathways) provides insights into the biological functions and pathways regulated by miRNAs. This aids in understanding the physiological implications of miRNA dysregulation.

Biomarker Discovery:

Integration of miRNA expression data with clinical outcomes or disease states can lead to the discovery of miRNA biomarkers for diagnosis, prognosis, and treatment response prediction in various diseases, including cancer and neurodegenerative disorders.

Network Dynamics Analysis:

Integrating miRNA data with dynamic modeling approaches allows for the analysis of network dynamics and the prediction of regulatory outcomes under different conditions or perturbations. This aids in elucidating the regulatory mechanisms underlying cellular responses.

Drug Discovery and Therapeutic Targeting:

Integration of miRNA data with drug response profiles and molecular pathways facilitates the identification of miRNAs as potential therapeutic targets or biomarkers for drug efficacy. This can accelerate drug discovery and personalized medicine approaches.

Evolutionary Conservation Studies:

Integrating miRNA data across species enables comparative genomics analyses to identify evolutionarily conserved miRNAs and their targets. This sheds light on the evolutionary dynamics of miRNA-mediated gene regulation and functional conservation.

Systems Biology Insights:

Integration of miRNA data into systems biology frameworks allows for the modeling and simulation of regulatory networks at a systems level. This integrative approach provides insights into emergent properties and behaviors of biological systems.

Tools commonly used for integrating omics data with miRNA data:

miRWalk: miRWalk enables the integration of miRNA-target interaction data with gene expression profiles. It allows users to input miRNA and mRNA expression data to predict potential miRNA-target interactions and perform functional enrichment analysis.

miRGator: miRGator integrates miRNA expression profiles with mRNA expression data, protein-protein interaction networks, and pathway information. It enables users to visualize miRNA-mRNA regulatory networks and identify key regulatory modules.

DIANA-miRPath: This tool integrates miRNA expression data with gene ontology (GO) terms and Kyoto Encyclopedia of Genes and Genomes (KEGG) pathways to predict the functional impact of miRNA dysregulation. It identifies enriched biological processes and pathways targeted by differentially expressed miRNAs.

miEAA: miEAA (miRNA-Enriched Annotation Analysis) integrates miRNA expression data with functional annotation databases, such as GO and KEGG, to identify miRNA-regulated biological processes and pathways. It prioritizes candidate miRNAs based on their functional relevance.

TarBase: TarBase provides a curated database of experimentally validated miRNA-target interactions. It allows users to query miRNA-target interactions based on experimental evidence and integrates miRNA-target interaction data with other omics data sets for network analysis.

miRNA Target Filter: This tool integrates miRNA expression data with target prediction algorithms, such as TargetScan and miRanda, to prioritize miRNA-target interactions based on expression correlation and target prediction scores. It facilitates the identification of high-confidence miRNA-target interactions.

miRNet: miRNet integrates miRNA expression data with protein-protein interaction networks, transcription factor-target interactions, and pathway databases. It enables users to construct and visualize miRNA-mediated regulatory networks and identify key regulatory nodes.

miRNAtap: miRNAtap integrates miRNA expression data with gene expression profiles, protein-protein interaction networks, and pathway information. It enables users to identify dysregulated miRNA-target interactions associated with specific biological processes or diseases.

These tools facilitate the integration of miRNA data with other omics data sets, enabling comprehensive analysis of miRNA-mediated regulatory networks and their functional implications.

Reference

1. Bartel, D. P. (2004). MicroRNAs: genomics, biogenesis, mechanism, and function. Cell, 116(2), 281-297. [DOI: 10.1016/S0092-8674(04)00045-5]
2. Ha, M., & Kim, V. N. (2014). Regulation of microRNA biogenesis. Nature Reviews Molecular Cell Biology, 15(8), 509-524. [DOI: 10.1038/nrm3838]
3. Jonas, S., & Izaurralde, E. (2015). Towards a molecular understanding of microRNA-mediated gene silencing. Nature Reviews Genetics, 16(7), 421-433. [DOI: 10.1038/nrg3965]
4. Mendell, J. T., & Olson, E. N. (2012). MicroRNAs in stress signaling and human disease. Cell, 148(6), 1172-1187. [DOI: 10.1016/j.cell.2012.02.005]
5. Lewis, B. P., Burge, C. B., & Bartel, D. P. (2005). Conserved seed pairing, often flanked by adenosines, indicates that thousands of human genes are microRNA targets. Cell, 120(1), 15-20. [DOI: 10.1016/j.cell.2004.12.035]
6. Bartel, D. P. (2009). MicroRNAs: target recognition and regulatory functions. Cell, 136(2), 215-233. [DOI: 10.1016/j.cell.2009.01.002]
7. Friedman, R. C., Farh, K. K., Burge, C. B., & Bartel, D. P. (2009). Most mammalian mRNAs are conserved targets of microRNAs. Genome Research, 19(1), 92-105. [DOI: 10.1101/gr.082701.108]

Circular RNA: Basic Concept and their Role in Various Processes

Sarika Sahu
Division of Agricultural Bioinformatics
ICAR-Indian Agricultural Statistics Research Institute, New Delhi

Introduction

In the eukaryotic organisms mainly two kinds of RNAs are occurred: coding, messenger RNA (mRNA), and non-coding RNA (ncRNA). With the advent of high throughput sequencing several RNAs have been discovered and are found in cells, such as microRNAs (miRNAs), long non-coding RNAs (lncRNAs), and circular RNAs (circRNAs), SnoRNA (small nucleolar), transfer RNAs (tRNAs), ribosomal RNAs (rRNAs), small interfering RNAs (siRNAs), small nuclear RNAs (snRNAs), piwi-interacting RNAs (Piwi-RNAs). ncRNA has little or no protein-coding potential but plays a vital role in various biological processes like gene regulation, chromosomal structure, genome defence, translation, splicing, DNA replication, healthy growth and development and stress responses. One of the important ncRNAs is circRNA, discovered over two decades ago as a special group of RNA transcripts featuring circular structures. The first identified circRNA was the potato spindle tuber viroid in 1976. Since, last four decades, circRNAs were often considered as by-products of splicing or aberrantly spliced products. Recent advancements in high-throughput sequencing technologies ease the unbiased deep profiling of circRNA landscape in a genome-wide manner. Subsequently, thousands of circRNAs have been reported in eukaryotes and archaea.

2. Biogenesis of circRNA

CircRNA is an endogenous single-stranded RNA molecule that is generated by the head-to-tail joining of pre-mRNA (back-splicing). There are three proposed models of circRNA biogenesis: (i) direct back-splicing, (ii) RNA-binding protein-mediated circularization, and (iii) lariat-driven circularization [Fig 1]. CircRNAs are generated when the pre-mRNA splicing machinery back splices to join a down-stream splice donor to an upstream splice acceptor. The 3′ and 5′ ends usually present in a linear mRNA molecule have been joined together covalently forming a characteristic back-splice junction (BSJ) in circRNA. Further, the U2-dependent spliceosome is account for the splicing of the vast majority of introns in both plants and animals, with GT and AG terminal dinucleotides at their 5′ and 3′ termini, respectively. However, in plants, both monocot and dicot species have different mechanism of the splice signals for circRNAs. Further, only a small portion (7.3%) of circRNAs possess canonical

GT/AG (CT/AC) splicing signals, and a large number of circRNAs share diverse non-GT/AG splicing signals, such as GC/GG, CA/GC, GG/AG, GC/CG, and CT/CC was reported in plants. CircRNAs have multiple origin sites; they can originate from multi-exonic transcripts, single exonic transcripts, uncharacterized transcripts and even fusion genes. In addition, Alternative RNA processing events have been observed in circRNAs, including exon skipping, intron retention and alternative splicing. Although most circular RNAs are lowly expressed, some of them are able to accumulate to high levels and even exceed their cognate mRNAs due to their longer half-lives. The majority of circRNAs are ecircRNAs, which are predominantly located in the cytoplasm. However, EIcircRNAs and ciRNAs are usually located in the nucleus. Once produced in the nucleus, the majority of circular RNAs are exported to the cytoplasm for their proper functions or degradation.

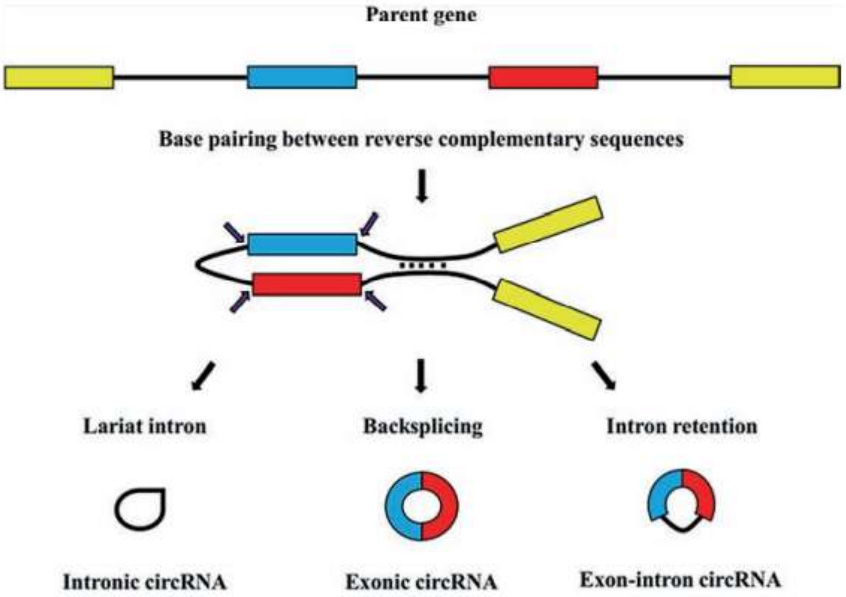


Fig1: biogenesis of different types of circRNA

3. Types of circular RNA

According to their genomic location, circRNAs are classified into exon, intron, intergenic, and exon-intron molecules. Intron circRNA mostly regulates its parental gene than exon circRNA. On the basis of origin of circRNA on the genome, circRNAs were classified into ten types (Fig. 2), at which the two back-splicing sites of a certain circRNA are located.

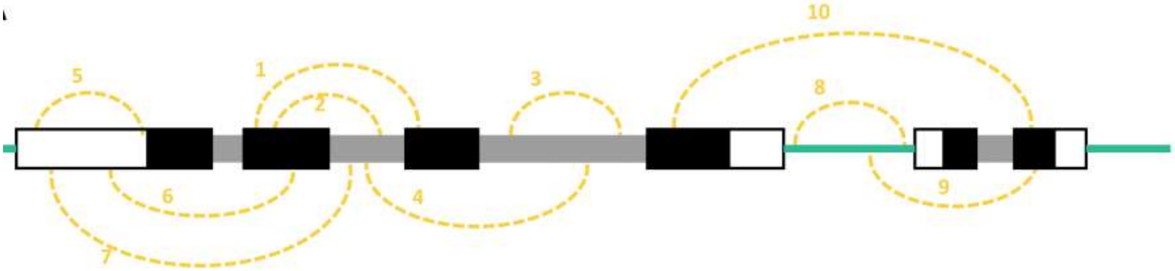


Fig2: Types of circRNAs on the basis of their generation from the parent gene. The black, grey and blank bars represent exons, introns and UTRs, respectively. The green lines represent intergenic region of the genomes

no. on fig2	Type of circRNA	Type of Origin
1	e-circRNA	two back-splicing sites of a circRNA are both at exons
2	ei-circRNA	one back-splicing site of a circRNA is at exon while the other is at intron
3	i-circRNA	two back-splicing sites of a circRNA are both at a single intron
4	ie-circRNA	two back-splicing sites of a circRNA are at two different introns across one or several exons
5	u-circRNA	two back-splicing sites of a circRNA are both at UTRs
6	ue-circRNA	one back-splicing site of a circRNA is at UTR while the other is at exon
7	ui-circRNA	one back-splicing site of a circRNA is at UTR while the other is at intron
8	ig-circRNA	two back-splicing sites of a circRNA are both at a single intergenic region
9	igg-circRNA	one back-splicing site of a circRNA is at intergenic region while the other is at genic region
10	ag-circRNA	two back-splicing sites of a circRNA are at two different genes

4. Characteristics of Plant circular RNAs

The nucleotide length of circRNAs are vary and ranges from <100 nt to >4 kb. They are conserved and have various isoforms that are generated by alternative circularization in plants. However, some circRNAs are only observed in certain plant species. The majority of plant exonic circRNAs contain 1-4 exons and large parental genes with multiple shorter exons are preferentially circularised. They are less likely to be generated from exon(s) flanked by introns containing repetitive and/or reverse complementary sequences. In Arabidopsis, out of the 13 validated plant circRNAs, only two (~15%) contain >15-bp reverse complementary sequences in their flanking introns. Similarly, in cotton (*Gossypium sp*.), despite circRNAs seem to have more repeat sequences in their flanking introns than linear genes, only ~10% of exonic circRNAs are associated with reverse complementary intronic sequences. A recent study in maize (Zea mays) found that LLERCPs (reverse complementary pairs of LINE1-like elements) are significantly enriched in the 35-kb, particularly in the 5-kb, flanking regions of circRNAs 20. The study also found that circRNAs with LLERCPs have an expression level significantly higher than those without LLERCPs nearby, indicating LLERCPs could reinforce the expression of circRNAs, although the numbers of LLERCPs seem not to be related to the expression level of circRNAs 20. Because LLERCPs were found in a relatively large flanking region of circRNAs, it is of interest to know how they are related to circRNA biogenesis. It is also of interest to know whether repeat sequences located at the flanking introns of circRNAs are associated with genome complexity so that large and polyploid genomes tend to have more repeat sequences in their flanking introns of circRNAs. In addition, multiple circRNAs can be generated from a single parental gene through alternative back splicing and circularization. Parental genes of over 700 exonic circRNAs (~15% of Arabidopsis circRNAs) are orthologs between rice and Arabidopsis. Approximately 34% and 55% of circRNA-producing soybean genes are conserved orthologs in Arabidopsis and rice, respectively. In the context of expression, they are not highly expressed while few are highly accumulated and exceed their cognate mRNAs due to their longer half-lives. Once produced in the nucleus, the majority of circular RNAs are exported to the cytoplasm for their proper functions or degradation.

5. Functional role of circRNA in plant

(i) Acting as miRNA sponges

The most extensively studied function of circRNAs is microRNA (miRNA) sponging. miRNAs are small noncoding RNAs that bind to target mRNAs and typically induce mRNA degradation or translational repression. Further, circRNAs have been found to bind miRNAs, decreasing their availability and thereby upregulating the expression of their target mRNAs. The first cases of miRNA sponging were discovered for CDR1as, with over 70 conserved target sites for miR-7, and circSry, with 16 binding sites for miR-138. circRNAs functioning as a miRNA sponge continue to be frequently documented and reported. However, studies that analysed thousands of circRNAs found that most contain a smaller number of miRNA binding sites and do not have other properties of effective miRNA sponges. These findings suggest that the majority of circRNAs do not act as miRNA sponges, and many studies have revealed other functions

(ii) Regulating transcription and translation

Further studies found that circRNAs perform many other regulatory functions, including exerting transcriptional and translational control, sequestering and translocating proteins, facilitating interactions between proteins, and translating to proteins. It was also observed that some engineered circRNAs having an internal ribosome entry site (IRES) could be translated and form small peptides in vivo.

(iii) circRNA as biomarkers

circRNAs could also be used as potential biomarkers in plants due to their unique characteristics, including resistance to degradation, long halflives, and ease the specificity of detection. Same study was reported in Arabidopsis, circRNAs used as bona fide biomarkers of functional exon-skipped AS variants, including in the homeotic MADS-box transcription factor family.

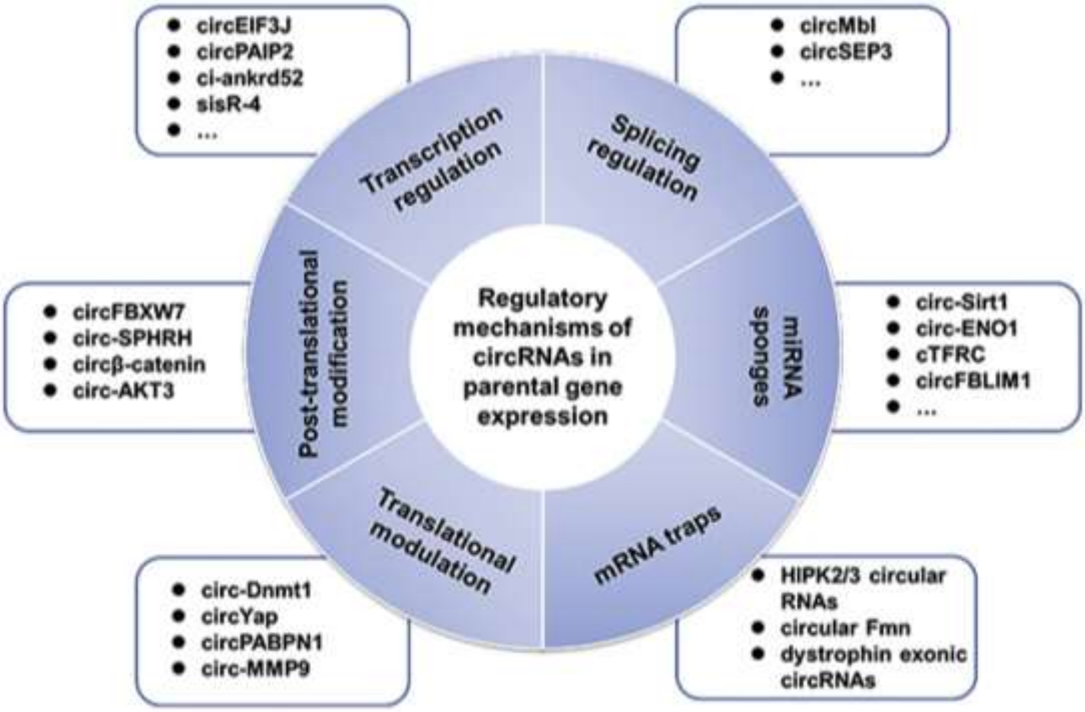


Fig3: functional role of parental gene of circRNA

(iv) Potential role of circRNAs in stress responses

circRNAs usually exhibit specific cell-type, tissue, and developmental stage expression patterns, and furthermore, the expression of circRNAs and circRNA isoforms is often induced under diverse environmental stresses, such as low- and high-light stresses, Pi-starvation conditions, low temperature stress, dehydration stress, and chewing injury stress by insects, which suggests that circRNAs might play important roles in plant development or in the response to biotic and abiotic stresses. Zhao *et al* discovered total 293 EIcircRNAs, including 183 and 175 in resistant and susceptible samples, under defoliation damage stress by cotton bollworm feeding in soybean, which indicated that EIcircRNAs might participate in the response to chewing injury resistance processes in plants. In addition, circRNAs of barley that are highly expressed in the mitochondria might be participated in micronutrient homeostasis.

(v) Role of circRNA in plant development

The overexpression of PSY1-circ1, a circRNA derived from *Phytoene Synthase 1* (*PSY1*) in tomato, resulted in a significant decrease in lycopene and β-carotene accumulation in transgenic tomato fruits, which suggests the involvement of circRNAs in plant development.

Table 1: List of tool for the prediction of circRNA

Tool	Version	Mapping tool	Address	References
circRNA finder	N/A	STAR	https://github.com/orzechoj/circRNA_finder	Westholm et al., 2014
CIRCexplorer	1.1.10	Bowtie1 and 2	https://github.com/YangLab/CIRCexplorer	Zhang et al., 2014
CIRI	1.2	Bwa	https://sourceforge.net/projects/ciri/files/	Gao et al., 2015
find circ	v2	Bowtie2	https://github.com/marvin-jens/find_circ	Memczak et al., 2013
Mapsplice	2.2.1	Bowtie1	http://www.netlab.uky.edu/p/bioinfo/MapSplice2	Wang et al., 2010
circseq-cup	1.0	STAR	http://ibi.zju.edu.cn/bioinplant/tools/circseq-cup.htm	Ye et al., 2017
KNIFE	1.4	Bowtie1, Bowtie2	https://github.com/lindaszabo/KNIFE	Szabo et al., 2015
Segemehl	0.2.0	Segemehl	http://www.bioinf.uni-leipzig.de/Software/segemehl/	Hoffmann et al., 2014
UROBORUS	0.0.2	Bowtie Bowtie2 tophat2	http://uroborus.openbioinformatics.org/en/latest/	Song et al., 2016

Table 2: List of plant database of circRNA

Database	Organisms	URL
PlantcircBase	*Oryza sativa, Arabidopsis thaliana, Zea mays, Solanum lycopersicum, Triticum aestivum, Glycine max, Gossypium hirsutum, Hordeum vulgare, Solanum tuberosum, Poncirus trifoliate, Gossypium arboretum Gossypium raimondii, Camellia sinensis, Pyrus betulifolia, Oryza sativa ssp. Indica, Nicotiana benthamiana,Brassica rapa, Cucumis sativus, Echinochloa crus-galli, Populus trichocarpa*	*http://ibi.zju.edu.cn/plantcircbase/index.php*

AtCircDB	*Arabidopsis thaliana*	*http://www.deepbiology.cn/circRNA/*
GreenCircRNA	*Ananas comosus, Amaranthus hypochondriacus, Arabidopsis lyrata, Asparagus officinalis, Arabidopsis thaliana, Botryococcus braunii, Brachypodium distachyon, Brachypodium hybridum, Brassica oleracea capitate, Brassica rapa FPsc, Brachypodium stacei, Brachypodium sylvaticum, Cicer arietinum, Citrus clementina, Capsella grandiflora, Carica papaya, Chenopodium quinoa, Chlamydomonas reinhardtii, Capsella rubella, Cucumis sativus, Citrus sinensis, Chromochloris zofingiensis, Daucus carota, Dunaliella salina, Eucalyptus grandis, Eutrema salsugineum, Fragaria vesca, Gossypium hirsutum, Glycine max, Gossypium raimondii, Helianthus annuus, Hordeum vulgare, Kalanchoe fedtschenkoi, Lactuca sativa, Linum usitatissimum, Musa acuminate, Malus domestica, Manihot esculenta, Mimulus guttatus, Marchantia polymorpha, Micromonas pusilla CCMP1545, Micromonas sp.RCC299, Medicago truncatula, Olea europaea, Oryza sativa, Oryza sativa Kitaake, Populus deltoides WV94, Panicum hallii, Physcomitrella patens, Prunus persica, Populus trichocarpa, Porphyra umbilicalis, Panicum virgatum, Phaseolus vulgaris, Ricinus communis, Sorghum bicolor, Setaria italic, Solanum lycopersicum, Spirodela polyrhiza, Salix purpurea, Solanum tuberosum, Setaria viridis, Triticum aestivum, Theobroma cacao, Trifolium pratense, Vigna unguiculata, Vitis vinifera, Zostera marina, Zea mays*	*http://greencirc.cn*

References

- Babaei, Saeid, Mohan B. Singh, and Prem L. Bhalla. "Circular RNAs repertoire and expression profile during Brassica rapa pollen development." *International journal of molecular sciences* 22, no. 19 (2021): 10297.
- Belousova, E. A., M. L. Filipenko, and N. E. Kushlinskii. "Circular RNA: new regulatory molecules." *Bulletin of Experimental Biology and Medicine* 164, no. 6 (2018): 803-815.
- Chaabane, Mohamed, Robert M. Williams, Austin T. Stephens, and Juw Won Park. "circDeep: deep learning approach for circular RNA classification from other long non-coding RNA." *Bioinformatics* 36, no. 1 (2020): 73-80.
- Cheng, Jinping, Yong Zhang, Ziwei Li, Taiyun Wang, Xiaotuo Zhang, and Binglian Zheng. "A lariat-derived circular RNA is required for plant development in Arabidopsis." *Science China Life Sciences* 61, no. 2 (2018): 204-213.
- Dong, Rui, Xu-Kai Ma, Guo-Wei Li, and Li Yang. "CIRCpedia v2: an updated database for comprehensive circular RNA annotation and expression comparison." *Genomics, proteomics & bioinformatics* 16, no. 4 (2018): 226-233.
- Gao, Yuan, Jinyang Zhang, and Fangqing Zhao. "Circular RNA identification based on multiple seed matching." *Briefings in bioinformatics* 19, no. 5 (2018): 803-810.
- Guria, Ashirbad, Kavitha Velayudha Vimala Kumar, Nagesh Srikakulam, Anakha Krishnamma, Saibal Chanda, Satyam Sharma, Xiaofeng Fan, and Gopal Pandi. "Circular RNA profiling by Illumina sequencing via template-dependent multiple displacement amplification." *BioMed research international* 2019 (2019).
- Hoffmann S, Otto C, Doose G, Tanzer A, Langenberger D, Christ S, Kunz M, Holdt L, Teupser D, Hackermueller J, Stadler PF: "A multi-split mapping algorithm for circular RNA, splicing, trans-splicing, and fusion detection", Genome Biology (2014) 15:R34.
- Jakobi, Tobias, Alexey Uvarovskii, and Christoph Dieterich. "circtools—a one-stop software solution for circular RNA research." *Bioinformatics* 35, no. 13 (2019): 2326-2328.
- Jakobi, Tobias, and Christoph Dieterich. "Computational approaches for circular RNA analysis." *Wiley Interdisciplinary Reviews: RNA* 10, no. 3 (2019): e1528.
- Jakub O. Westholm, Pedro Miura, Sara Olson, Sol Shenker, Brian Joseph, Piero Sanfilippo, Susan E. Celniker, Brenton R. Graveley, and Eric C. Lai. Genome-wide Analysis of Drosophila Circular RNAs Reveals Their Structural and Sequence Properties and Age-Dependent Neural Accumulation Westholm et al. Cell Reports, 2014.
- Kai Wang, Darshan Singh, Zheng Zeng, Stephen J. Coleman, Yan Huang, Gleb L. Savich, Xiaping He, Piotr Mieczkowski, Sara A. Grimm, Charles M. Perou, James N. MacLeod, Derek Y. Chiang, Jan F. Prins and Jinze Liu
  Nucleic Acids Research 2010; doi: 10.1093/nar/gkq622.
- Memczak, S.; Jens, M.; Elefsinioti, A.; Torti, F.; Krueger, J.; Rybak, A.; Maier, L.; Mackowiak, S.D.; Gregersen, L.H.; Munschauer, M.; et al. Circular RNAs are a large class of animal RNAs with regulatory potency. Nature 2013, 495, 333–338.
- Song X, Zhang N, Han P, Lai RK, Wang K, Lu W. Circular RNA Profile in Gliomas Revealed by Identification Tool UROBORUS. Nucleic Acids Research, 2016, 44:e87.
- Szabo L, Morey R, Palpant NJ, Wang PL, Afari N, Jiang C, Parast MM, Murry CE, Laurent LC, Salzman J. Statistically based splicing detection reveals neural enrichment and tissue-specific induction of circular RNA during human fetal development. Genome Biology. 2015, 16:126.
- Tong, Wei, Jie Yu, Yan Hou, Fangdong Li, Qiying Zhou, Chaoling Wei, and Jeffrey L. Bennetzen. "Circular RNA architecture and differentiation during leaf bud to young leaf development in tea (Camellia sinensis)." Planta 248, no. 6 (2018): 1417-1429.

- Wang, Kai, Chong Wang, Baohuan Guo, Kun Song, Chuanhong Shi, Xin Jiang, Keyi Wang, Yacong Tan, Lequn Wang, Lin Wang, Jiangjiao Li, Ying Li, Yu Cai, Hongwei Zhao and Xiaoyong Sun. “CropCircDB: a comprehensive circular RNA resource for crops in response to abiotic stress.” Database: The Journal of Biological Databases and Curation 2019 (2019): n. pag.
- Wang, Ying, Zeyang Xiong, Qian Li, Yueyang Sun, Jing Jin, Hao Chen, Yu Zou, Xingguo Huang, and Yi Ding. "Circular RNA profiling of the rice photo-thermosensitive genic male sterile line Wuxiang S reveals circRNA involved in the fertility transition." BMC plant biology 19, no. 1 (2019): 1-16.
- Yang, Zhenchao, Zhao Yang, Yingge Xie, Qi Liu, Yanhao Mei, and Yongjun Wu. "Systematic identification and analysis of light-responsive circular RNA and co-expression networks in lettuce (Lactuca sativa)." G3: Genes, Genomes, Genetics 10, no. 7 (2020): 2397-2410.
- Ye et al., Full length sequence assembly reveals circular RNAs with diverse non GT AG splicing signals in rice. RNA Biology. 2016.
- Ye, Jiazhen, Lin Wang, Shuzhang Li, Qinran Zhang, Qinglei Zhang, Wenhao Tang, Kai Wang et al. "AtCircDB: a tissue-specific database for Arabidopsis circular RNAs." Briefings in Bioinformatics 20, no. 1 (2019): 58-65.
- Yin, Shuwei, Xiao Tian, Jingjing Zhang, Peisen Sun, and Guanglin Li. "PCirc: random forest-based plant circRNA identification software." BMC bioinformatics 22, no. 1 (2021): 1-14.
- Yuan Gao†, Jinfeng Wang† and Fangqing Zhao*. CIRI: an efficient and unbiased algorithm for de novo circular RNA identification. Genome Biology (2015) 16:4.
- Zhang XO, Wang HB, Zhang Y, Lu X, Chen LL and Yang L. Complementary sequence-mediated exon circularization. Cell, 2014, 159: 134-147.
- Zhang, Pei, Yuan Fan, Xiaopeng Sun, Lu Chen, William Terzaghi, Etienne Bucher, Lin Li, and Mingqiu Dai. "A large-scale circular RNA profiling reveals universal molecular mechanisms responsive to drought stress in maize and Arabidopsis." The Plant Journal 98, no. 4 (2019): 697-713.

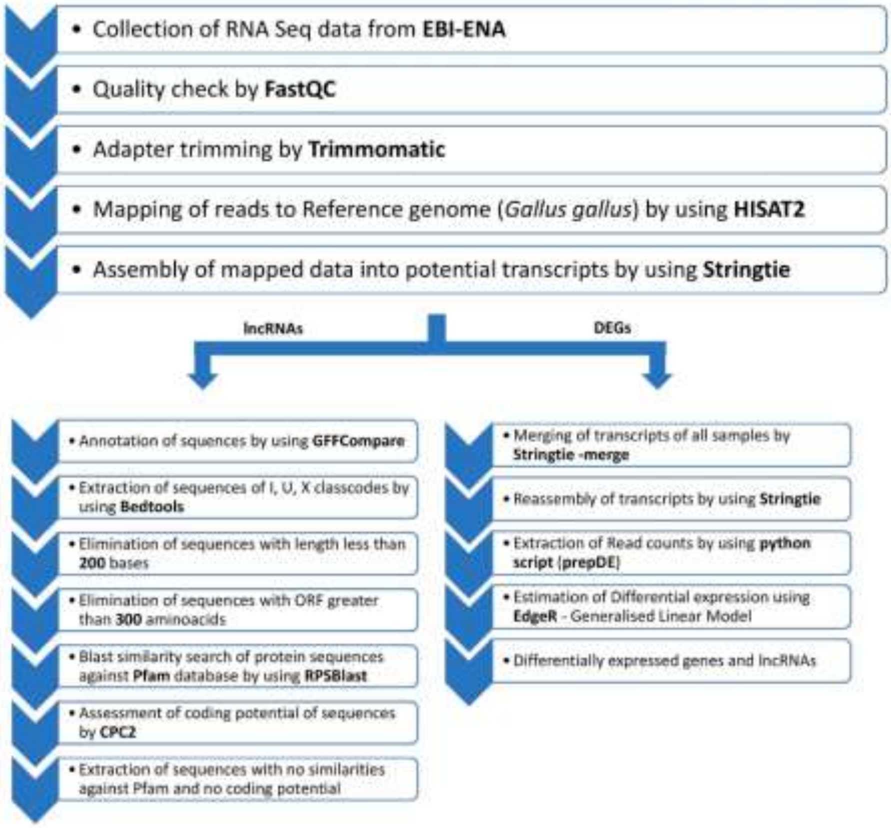
Hands-on-session for circRNA prediction

- Kindly see the manual of bwa link is given below:
- (https://bio-bwa.sourceforge.net/bwa.shtml)
- Kindly download CIRI2 from the link given below:
- https://sourceforge.net/projects/ciri/files/CIRI2/
- Step1: bwa index reference_file.fa
- Step2: bwa mem index_file fastq_file > input.sam (single end data)
- bwa mem index_file read1.fq read2.fq > input.sam (Paired-end data)
- Step3: perl CIRI2.pl --help
- perl CIRI2.pl -I input.sam -O circRNA –F reference_file.fa -T 10

In-Silico Identification of Long Non Coding RNAs Playing Key Roles during Different Physiological Conditions in Livestock

Dr. Shailesh Sharma
National Institute of Animal Biotechnology, Hyderabad

Long non-coding RNAs (long ncRNAs, lncRNA) are one among another types of RNA, generally defined as transcripts more than 200 nucleotides that are not translated into protein. Identification and analysis of the expression profiles of key molecular players specifically lncRNAs involved in host-pathogen interactions and host response against any pathogenic response like Brucellosis or NDV or during Sex differentiation will be of great value. This presentation will end up with the holistic view of the key molecular player involved in host-pathogen interactions and host response against any pathogen. This training will show insights of the interplay of key molecular players specifically lncRNAs which may play role in resistance and susceptibility pattern against pathogens. This will surely contribute in the better understanding of different physiological conditions at genomic level.



- Collection of RNA Seq data from **EBI-ENA**
- Quality check by **FastQC**
- Adapter trimming by **Trimmomatic**
- Mapping of reads to Reference genome (*Gallus gallus*) by using **HISAT2**
- Assembly of mapped data into potential transcripts by using **Stringtie**

lncRNAs

- Annotation of squences by using **GFFCompare**
- Extraction of sequences of I, U, X classcodes by using **Bedtools**
- Elimination of sequences with length less than **200** bases
- Elimination of sequences with ORF greater than **300** aminoacids
- Blast similarity search of protein sequences against **Pfam** database by using **RPSBlast**
- Assessment of coding potential of sequences by **CPC2**
- Extraction of sequences with no similarities against Pfam and no coding potential

DEGs

- Merging of transcripts of all samples by **Stringtie -merge**
- Reassembly of transcripts by using **Stringtie**
- Extraction of Read counts by using **python script** (**prepDE**)
- Estimation of Differential expression using **EdgeR** - Generalised Linear Model
- Differentially expressed genes and lncRNAs

Figure 1: Home made algorithm which will be applied for available SRA datasets. This algorithm is already applied on ~100 NDV infected *Gallus gallus* datasets and ~24 *Bos taurus* samples and relevant papers are already published.

Theme of Research:

Our team's research experience spans Genomics, Transcriptomics and Structural Bioinformatics. We focus on Identification of long non-coding RNAs and genes, deferential expression analysis, functional annotation, co-expression analysis. Apart from this, we also perform homology modelling, screening, ADMET analysis and Molecular Dynamics Simulation.

Objectives:

1. Trachea transcriptome analysis to decipher the host response during Newcastle Disease challenge in different breeds of chicken.

2. Identification and differential expression of long non-coding RNAs and their association with genes during early embryonic developmental stages of *Bos taurus* and *Sus scrofa.*

3. Deciphering the structure and function of bovine ephemeral fever virus accessory proteins.

4. Identification of lncRNAs during host response against Bovine tuberculosis in cattle.

5. Sheep breed classification on the basis of phenotypic characters by using Artificial Intelligence.

6. Identification of role of lncRNAs in Bovine uterine transcriptome response to high fertile and low fertile semen in cattle.

Recent Work:

1. Trachea transcriptome analysis to decipher the host response during Newcastle Disease challenge in different breeds of chicken.

Newcastle disease is a highly infectious economically devastating disease caused by Newcastle disease Virus (NDV) in *Gallus gallus* (Chicken). Leghorn and Fayoumi are two breeds which show differential resistance patterns towards NDV. This study aims to identify the differentially expressed genes and lncRNAs during NDV challenge which could play a potential role in this differential resistance pattern. A total of 552 genes and 1580 lncRNAs were found to be differentially expressing. Of them, 52 genes were annotated with both Immune related pathways and Gene ontologies. We found that most of these genes were upregulated in Leghorn between normal and challenged chicken but several were down regulated between different timepoints after NDV challenge, while Fayoumi showed no such downregulation. We also observed that higher number of positively correlating lncRNAs were found to be downregulated along with these genes. This shows that although Leghorn is showing higher number of differentially expressed genes in challenged than in non-challenged, most of them were downregulated during the disease between different timepoints. With this we hypothesize that the downregulation of immune related genes and co-expressing lncRNAs could play a significant role behind the Leghorn being comparatively susceptible breed than Fayoumi.

2. Identification and differential expression of long non-coding RNAs and their association with genes during early embryonic developmental stages of *Bos taurus* and *Sus scrofa.*

Porcine epiblast derived pleuripotent stem cells have application in livestock breeding. The molecular mechanism involved during pig embryo development is largely regulated by long non coding RNAs. Here we analyzed the transcriptome data of porcine scRNA-seq from four different stages; E11 epiblast cells, E14 somatic cells E14 Primordial germ cells and E31 primordial germ cells to understand the role of long non coding RNAs, their distribution across the chromosomes over time, their genomic location. The differentially expression profile of the genes between different time points shows some similarity and aslo differences in expression for certain genes as the embryo grows from E11 epiblast to E31 primordial germ cells. Further, we analyzed the differentially expressed long non coding RNAs and their co-expression. The functional annotation of the differentially expressed lncRNAs and DEGs of the pig early embryo shows important functions including anatomical structure developmental, cellular processes, metabolic processes, developmental process.

3. Deciphering the structure and function of bovine ephemeral fever virus accessory proteins.

Bovine Ephemeral Fever (BEF) virus is an arthropod-borne rhabdovirus that is enclosed in a cone- or bullet-shaped envelope and contains negative-sense single-stranded RNA. The BEF virus causes acute febrile illness in cattle and water buffalo, which results in fever, shivering, lameness, and stiff muscles in affected animals. The genome is comprised of several open reading frames (ORFs) encoding, structural (N, P, M, G & L), non-structural (GNS), and several small accessory proteins (α1, α2, α3, β, and γ). The structural proteins, namely, nucleoprotein (N, 52 kDa), phosphoprotein (P, 43 kDa), matrix protein (M, 29 kDa), glycoprotein (G, 81 kDa), and the polymerase or large protein (L, 180 kDa) constitute the virion. Since some of the accessory proteins might have the feature of viroporin. We are working on the protein-membrane complex, and we have built the protein-membrane complex for further study MDS (Fig. 2).

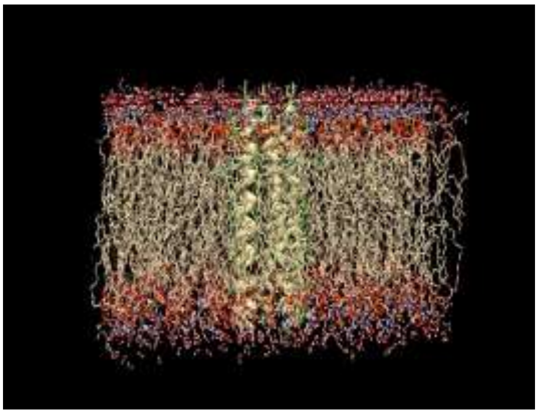


Figure 2

4. Identification of lncRNAs during host response against Bovine tuberculosis in cattle.

Long non-coding RNAs (lncRNAs) are the transcripts of length longer than 200 nucleotides. They are involved in the regulation of various biological activities. Bovine tuberculosis, caused by *Mycobacterium tuberculosis bovis* (*M. bovis*), is an important enzootic disease affecting mainly cattle, worldwide. Despite the implementation of national campaigns to eliminate the disease, bovine tuberculosis remains recalcitrant to eradication in several countries. Here, we report the analysis of the transcriptomic data of whole blood cells collected from experimentally infected calves with a virulent strain of *M. Bovis* for studying the lncRNAs involved in regulation of these genes. Using bioinformatics approaches, a total of 51,812 lncRNAs were extracted and 86 and 29 lncRNAs were deferentially expressed from infected and uninfected calf samples at each of the 8- and 20- w.p.i time points, respectively.

5. Sheep breed classification on the basis of phenotypic characters by using Artificial Intelligence.

Since a long time ago for the production of wool, meat and milk sheep are farmed by human being. Currently the worldwide population of sheep is around 1 billion and it is estimated that they come under 1000 distinct breeds. To estimate the commercial value of farming, a sheep producer need an automatic method of identification of different breeds which can be valuable for the sheep industry. An alternative method for breed identification is DNA testing but it is expensive and sometimes not affordable for a huge population. In this study we have tried to develop a CNN model and trained it using the facial images of four different breeds of sheep found in different parts of our country (India). Our aim is to classify these sheep into their respective breeds on the basis of their phenotypic characters by using artificial intelligence and deep learning algorithms. Throughout our study, we achieved training accuracy 97.68% and testing accuracy 82.66%. For more accurate and efficient classification of breeds we can use this technique in sheep farming for the welfare of both sheep and farmer.

6. Identification of role of lncRNAs in Bovine uterine transcriptome response to high fertile and low fertile semen in cattle.

Fertility is a vital factor impacting the production of *Bos taurus*, the widely recognized domestic cattle and economically significant livestock species worldwide. However, reproductive efficiency in *Bos taurus* is hindered by various fertility-related issues, which can have adverse economic implications. Recent studies have revealed the pivotal role of long non-coding RNAs (lncRNAs) in governing gene expression and cellular processes, particularly those involved in fertility. Initially, we have identified a total of 9078 lncRNAs. After differential expression analysis, in High fertile vs Low fertile groups, we have identified 128 DEGs and 1 DElncRNA. In High fertile vs Control groups, we have identified 283 DEGs and 20 DElncRNAs. In Low fertile vs Control groups, we have identified 74 DEGs and no DElncRNAs. In comparison with the previous study, in High fertile vs Low fertile groups, out of 40 DEGs identified in the previous study, 11 DEGs were found to be common with our study. In High fertile vs Control groups, out of 376 previous DEGs, 58 DEGs were found to be common. In Low fertile vs Control groups, the 1 DEG identified in the previous study was also found in our study. In Functional annotation, Cellular Process (GO:0009987) was found to be annotated to highest percentage of DEGs (21%), followed by, Metabolic Process (GO:0008152) with 17% of DEGs and Biological Regulation (GO:0065007) with 12% DEGs. About 3% of the DEGs were found to be annotated with Immune System Process (GO:0002376). In pathway annotation, under KEGG pathway categories, highest number of the annotated pathways (31%) were found to be under Human Diseases and Metabolism categories, followed by Organismal Systems category with 18% of the pathways. Under Reactome pathway categories, highest number of the annotated pathways (19%) were found to be under Signal Transduction category, 18% of the pathways under Immune System category followed by Metabolism category with 17% of the pathways.We also identified several DElncRNAs which were co-expressing with these DEGs. In conclusion, this study shows the relation of DelncRNAs corresponding to the DEGs and their functions.

References:

Vanamamalai VK, E P, T R K, **Sharma S**. Integrated analysis of genes and long non-coding RNAs in trachea transcriptome to decipher the host response during Newcastle disease challenge in different breeds of chicken. Int J Biol Macromol. 2023 Oct 2;253(Pt 5):127183. doi: 10.1016/j.ijbiomac.2023.127183. Epub ahead of print. PMID: 37793531

Jali I, Vanamamalai VK, Garg P, Navarrete P, Gutiérrez-Adán A, **Sharma S**. Identification and differential expression of long non-coding RNAs and their association with XIST gene during early embryonic developmental stages of *Bos taurus*. Int J Biol Macromol. 2022 Dec 24:S0141- 8130(22)03132-4. doi: 10.1016/j.ijbiomac.2022.12.221. Epub ahead of print. PMID: 36572076

Vanamamalai VK, Garg P, Kolluri G, Gandham RK, Jali I, **Sharma S**. Transcriptomic analysis to infer key molecular players involved during host response to NDV challenge in *Gallus gallus* (Leghorn & Fayoumi). Sci Rep. 2021 Apr 19;11(1):8486. doi: 10.1038/s41598-021-88029-6. PMID: 33875770

MiRDeep2

Priyanka Jain, Sunbul Ahmed

Amity University, Noida

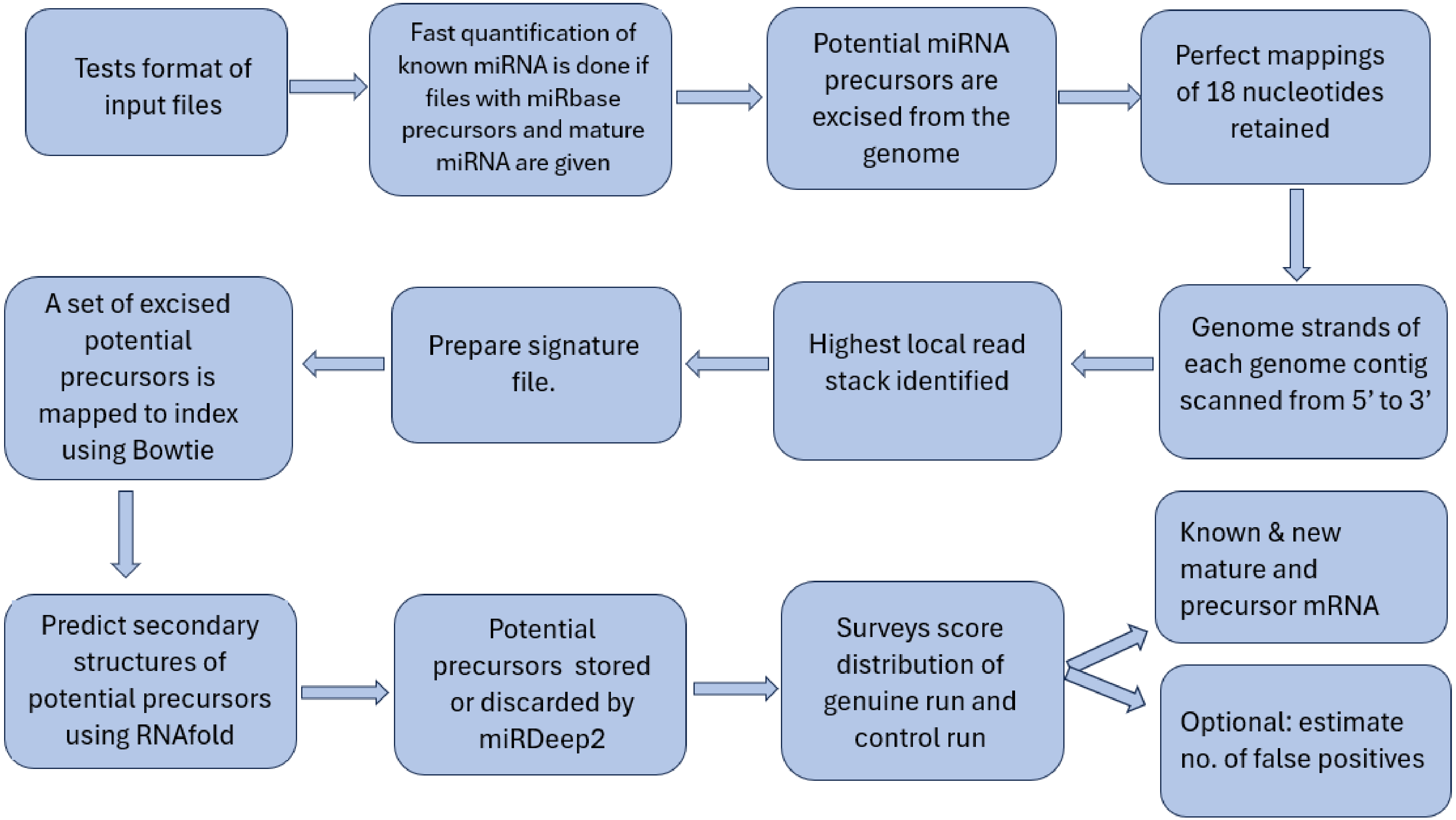
Topics to be covered in this lecture:

- microRNA
- miRDeep2
- miRDeep2 algorithm
- miRDeep2 Workflow
- MiRDeep2 script references
- Analysing and identifying miRNAs from RNA-seq data using the miRDeep2 tool in Galaxy

MiRDeep2 (Friendlar et al.)

- Developed by Sebastian Mackowiak & Marc Friedländer.
- miRDeep2 discovers active known or novel miRNAs from deep sequencing data (Solexa/Illumina, 454, ...).
- User-friendly
- Written in Perl
- Tools for read mapping, RNA folding, and calculating the significance of folding energies

Workflow of MiRDeep2 module algorithm



Tests format of input files

Fast quantification of known miRNA is done if files with miRbase precursors and mature miRNA are given

Potential miRNA precursors are excised from the genome

Perfect mappings of 18 nucleotides retained

Genome strands of each genome contig scanned from 5' to 3'

Highest local read stack identified

Prepare signature file.

A set of excised potential precursors is mapped to index using Bowtie

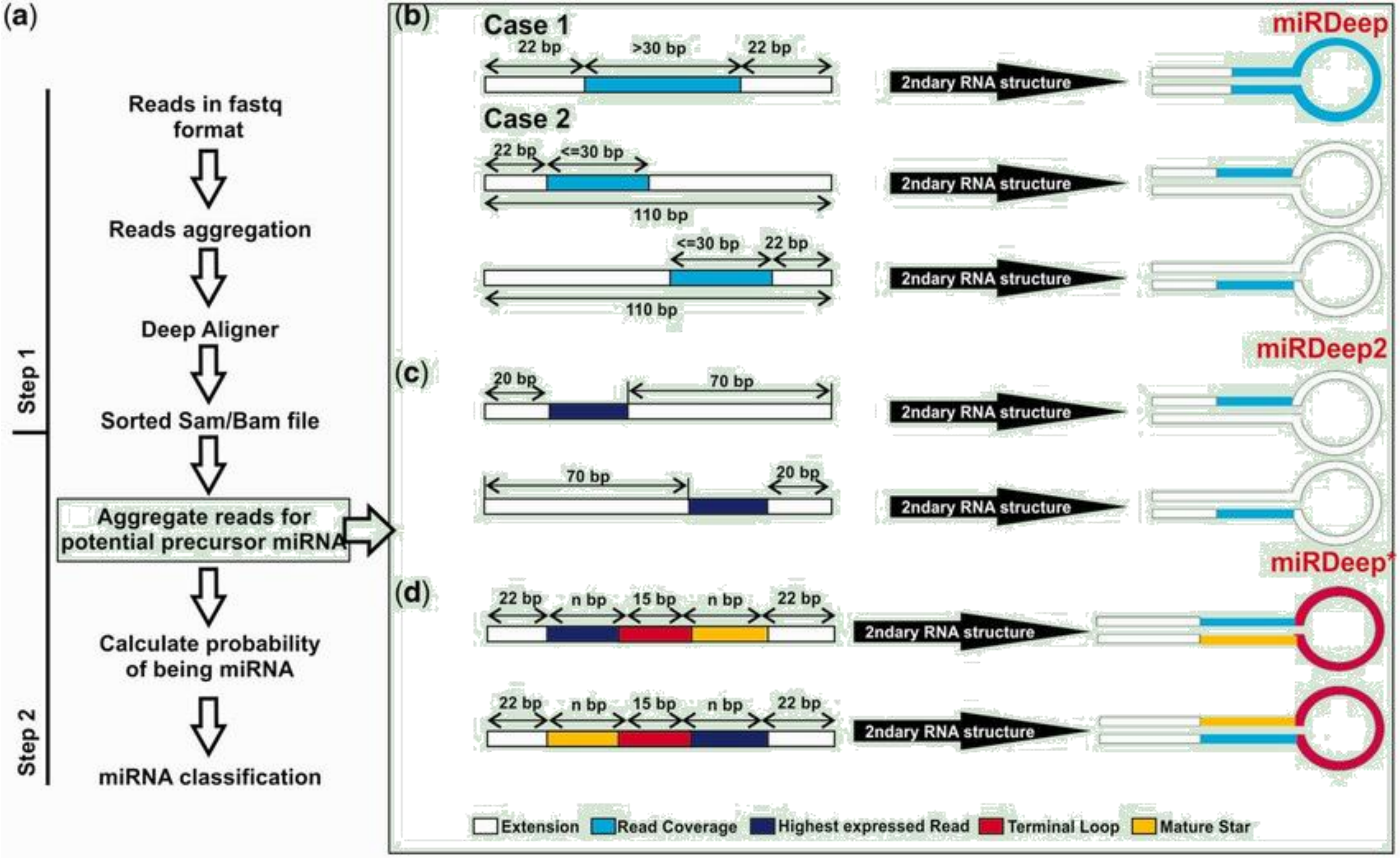
Predict secondary structures of potential precursors using RNAfold

Potential precursors stored or discarded by miRDeep2

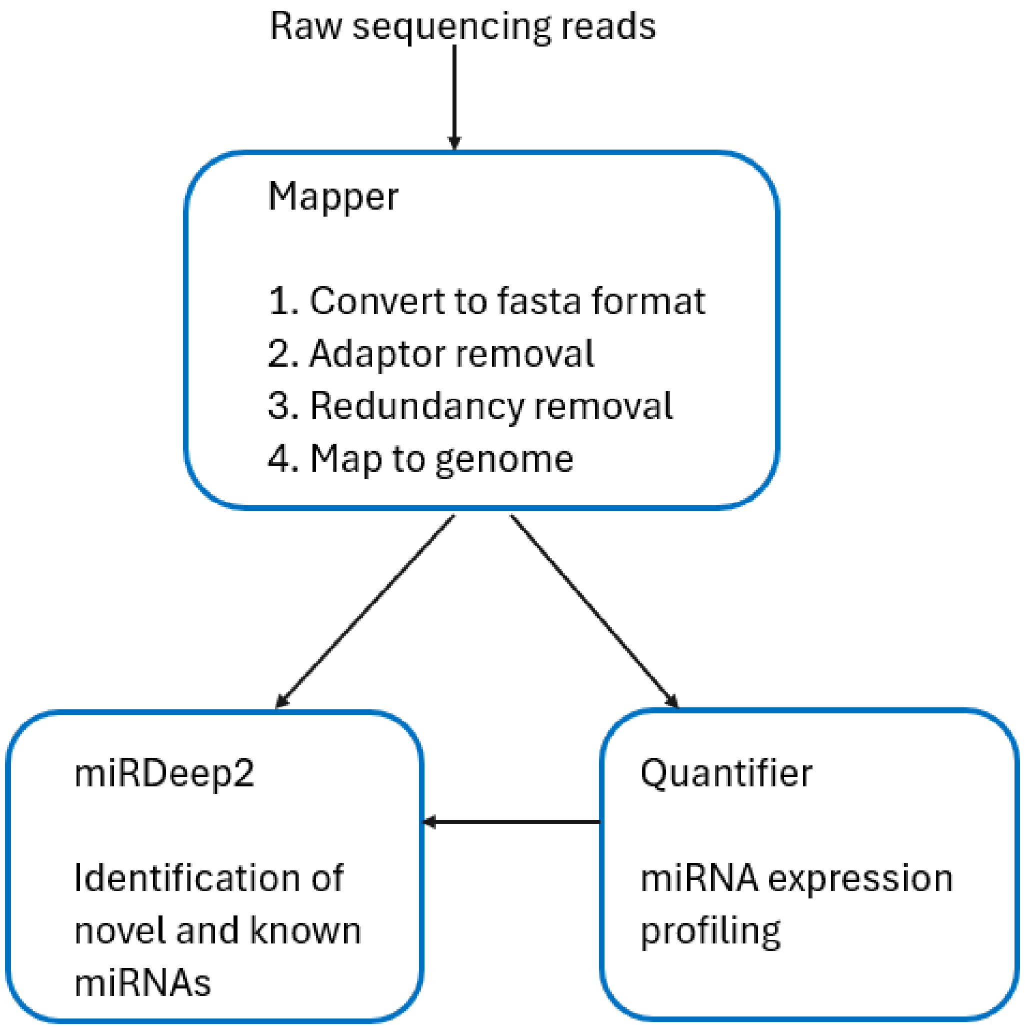
Surveys score distribution of genuine run and control run

Known & new mature and precursor mRNA

Optional: estimate no. of false positives


(a)
Step 1
Reads in fastq format
Reads aggregation
Deep Aligner
Sorted Sam/Bam file
Aggregate reads for potential precursor miRNA
Calculate probability of being miRNA
Step 2
miRNA classification
(b)
Case 1
22 bp
>30 bp
22 bp
2ndary RNA structure
miRDeep
Case 2
22 bp
<=30 bp
110 bp
2ndary RNA structure
<=30 bp
22 bp
110 bp
2ndary RNA structure
(c)
miRDeep2
20 bp
70 bp
2ndary RNA structure
70 bp
20 bp
2ndary RNA structure
(d)
miRDeep*
22 bp
n bp
15 bp
n bp
22 bp
2ndary RNA structure
22 bp
n bp
15 bp
n bp
22 bp
2ndary RNA structure
Extension
Read Coverage
Highest expressed Read
Terminal Loop
Mature Star

Analyzing and identifying miRNAs from RNA-Seq data using the miRDeep2 tool in Galaxy



Raw sequencing reads

Mapper

1. Convert to fasta format
2. Adaptor removal
3. Redundancy removal
4. Map to genome

miRDeep2

Identification of novel and known miRNAs

Quantifier

miRNA expression profiling

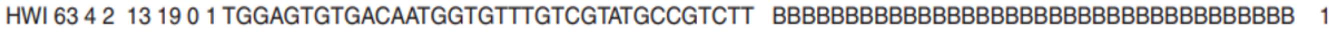
MiRDeep2 mapper

- The MiRDeep2 Mapper module is designed as a tool to process deep sequencing reads and/or map them to the reference genome

Input

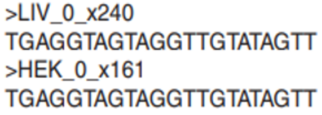
- Default input is a file in FASTA format, seq.txt or qseq.txt format. More input can be given depending on the options used.

1. qseq.txt



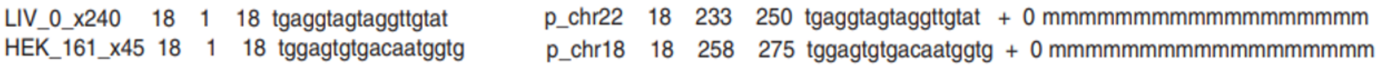
```
HWI 63 4 2  13 19 0 1 TGGAGTGTGACAATGGTGTTTGTCGTATGCCGTCTT   BBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBB   1
```

2. FASTQ



```
>LIV_0_x240
TGAGGTAGTAGGTTGTATAGTT
>HEK_0_x161
TGAGGTAGTAGGTTGTATAGTT
```

3. arf

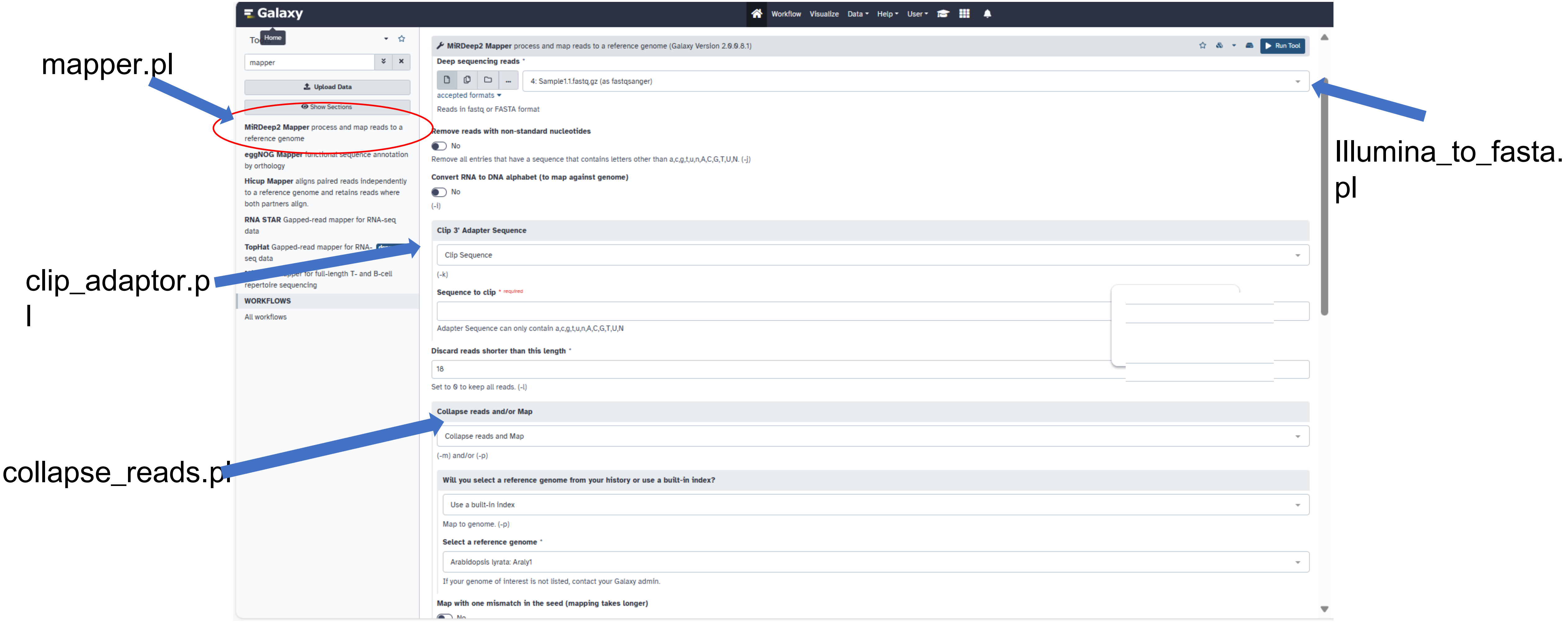


```
LIV_0_x240    18   1   18  tgaggtagtaggttgtat          p_chr22   18   233   250  tgaggtagtaggttgtat  +  0 mmmmmmmmmmmmmmmmmm
HEK_161_x45  18   1    18  tggagtgtgacaatggtg          p_chr18   18   258   275  tggagtgtgacaatggtg  +  0 mmmmmmmmmmmmmmmmmm
```

Output

- The output depends on the options used. Either a FASTA file with processed reads or an arf file with mapped reads, or both, are output.
reads, or both, are output.
- **Arf format:** This is a proprietary file format generated and processed by miRDeep2. It contains information of reads mapped to a reference genome. Each line in such a file contains 13 columns:
- read identifier
- length of read sequence
- start position in read sequence that is mapped
- end position in the read sequence that is mapped
- read sequence
- identifier of the genome part to which a read is mapped to. This is either a scaffold id or a chromosome name
- length of the genome sequence a read is mapped to
- start position in the genome where a read is mapped to
- end position in the genome where a read is mapped to
- genome sequence to which a read is mapped
- genome strand information. Plus means the read is aligned to the sense-strand of the genome. Minus means it is aligned to the antisense strand of the
is aligned to the antisense strand of the genome.
- Number of mismatches in the read mapping
- Edit string that indicates matches by lowercase 'm' and mismatches by uppercase 'M'

Summary of MiRDeep2 mapper and script commands



mapper.pl

clip_adaptor.pl

collapse_reads.pl

Illumina_to_fasta.pl

mapper

MiRDeep2 Mapper process and map reads to a reference genome

eggNOG Mapper functional sequence annotation by orthology

Hicup Mapper aligns paired reads independently to a reference genome and retains reads where both partners align.

RNA STAR Gapped-read mapper for RNA-seq data

TopHat Gapped-read mapper for RNA-seq data

...apper for full-length T- and B-cell repertoire sequencing

WORKFLOWS

All workflows

MiRDeep2 Mapper process and map reads to a reference genome (Galaxy Version 2.0.0.8.1)

Deep sequencing reads *

4: Sample1.1.fastq.gz (as fastqsanger)

accepted formats

Reads in fastq or FASTA format

Remove reads with non-standard nucleotides

No

Remove all entries that have a sequence that contains letters other than a,c,g,t,u,n,A,C,G,T,U,N. (-j)

Convert RNA to DNA alphabet (to map against genome)

No

(-i)

Clip 3' Adapter Sequence

Clip Sequence

(-k)

Sequence to clip * required

Adapter Sequence can only contain a,c,g,t,u,n,A,C,G,T,U,N

Discard reads shorter than this length *

18

Set to 0 to keep all reads. (-l)

Collapse reads and/or Map

Collapse reads and Map

(-m) and/or (-p)

Will you select a reference genome from your history or use a built-in index?

Use a built-in index

Map to genome. (-p)

Select a reference genome *

Arabidopsis lyrata: Araly1

If your genome of interest is not listed, contact your Galaxy admin.

Map with one mismatch in the seed (mapping takes longer)

No

mapper. pl

Processes reads and/or maps them to the reference genome.

Input

Default input is a file in FASTA, seq.txt, or qseq.txt format.

More input can be given depending on the options used.

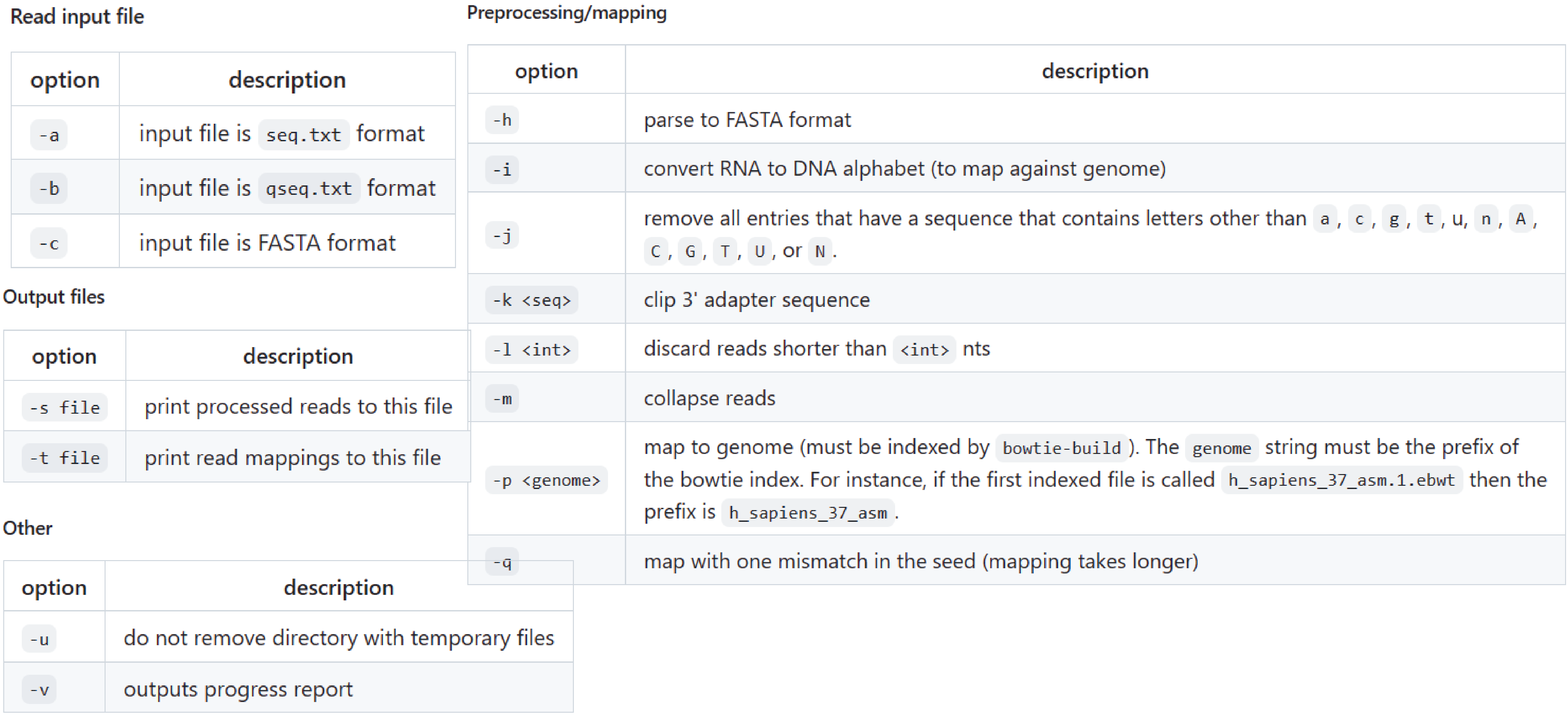
Output

The output depends on the options used (see below).

Either

- a FASTA file with processed reads, or
- an ARF file with mapped reads, or
- Both are output.

Options:



Read input file

option	description
`-a`	input file is `seq.txt` format
`-b`	input file is `qseq.txt` format
`-c`	input file is FASTA format

Output files

option	description
`-s file`	print processed reads to this file
`-t file`	print read mappings to this file

Other

option	description
`-u`	do not remove directory with temporary files
`-v`	outputs progress report

Preprocessing/mapping

option	description
`-h`	parse to FASTA format
`-i`	convert RNA to DNA alphabet (to map against genome)
`-j`	remove all entries that have a sequence that contains letters other than `a`, `c`, `g`, `t`, u, `n`, `A`, `C`, `G`, `T`, `U`, or `N`.
`-k <seq>`	clip 3' adapter sequence
`-l <int>`	discard reads shorter than `<int>` nts
`-m`	collapse reads
`-p <genome>`	map to genome (must be indexed by `bowtie-build`). The `genome` string must be the prefix of the bowtie index. For instance, if the first indexed file is called `h_sapiens_37_asm.1.ebwt` then the prefix is `h_sapiens_37_asm`.
`-q`	map with one mismatch in the seed (mapping takes longer)

clip_adapters.pl

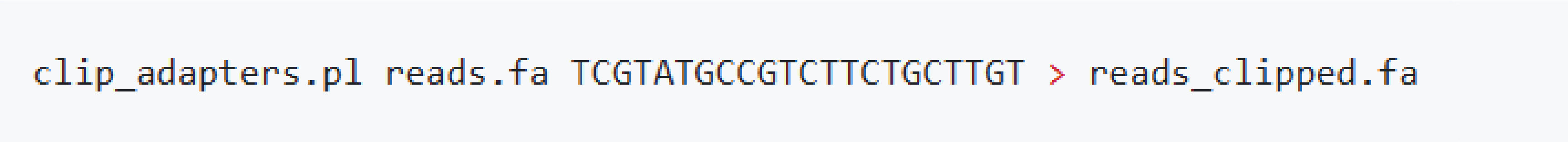
Removes 3' end adaptors from deep sequenced small RNAs.

Input

- A FASTA file with the deep sequencing reads and the adapter sequence (both in RNA or DNA alphabet).

Output

- A FASTA file with the clipped reads.
- FASTA IDs are retained. If no matches to the adapter prefixes are identified in a given read, the unclipped read is output.



```
clip_adapters.pl reads.fa TCGTATGCCGTCTTCTGCTTGT > reads_clipped.fa
```

collapse_reads.pl

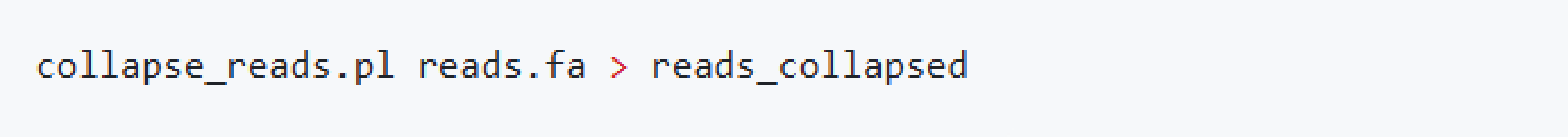
Collapses are read in the FASTA file to ensure that each sequence only occurs once. To indicate how many times reads the sequence represents, a suffix is added to each FASTA identifier. *E.g.* a sequence that represents ten reads in the data will have the _x10 suffix added to the identifier.

Input

- A FASTA file, either in standard format or in the collapsed suffix format.

Output

- A FASTA file in the collapsed suffix format.



```
collapse_reads.pl reads.fa > reads_collapsed
```

illumina_to_fasta.pl

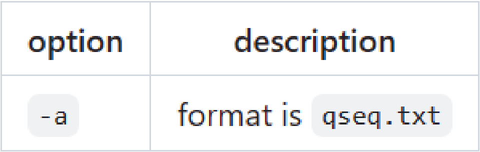
- parses seq.txt or qseq.txt output from the Solexa/Illumina platform to FASTA format.

Input

- A seq.txt or
- qseq.txt file.
- By default seq.txt.

Output

- A FASTA file, one entry for each line of seq.txt.
- The entries are named seq plus a running number that is incremented by one for each entry. Any . characters in the seq.txt file is substituted with an N.

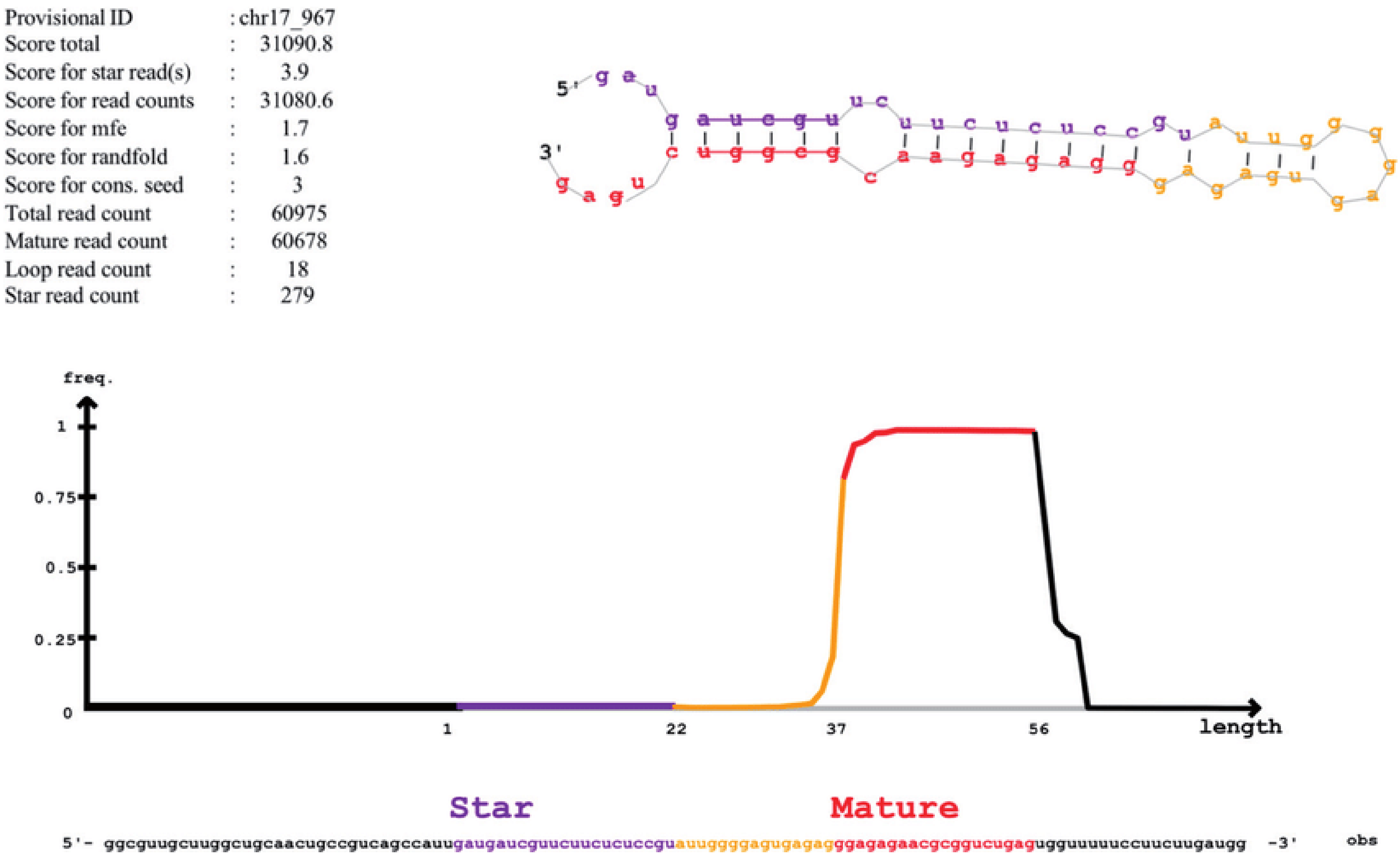


option	description
`-a`	format is `qseq.txt`

MiRDeep2 quantifier

- The module maps the deep sequencing reads to predefined miRNA precursors and determines the expression of the corresponding miRNAs. First, the predefined mature miRNA sequences are mapped to the predefined precursors. Optionally, predefined star sequences can be mapped to the precursors too. By that, the mature and star sequence in the precursors are determined. Second, the deep sequencing reads are mapped to the precursors. The number of reads falling into an interval 2nt upstream and 5nt downstream of the mature/star sequence is determined.
- **Input**
- A FASTA file with precursor sequences, a FASTA file with mature miRNA sequences, a FASTA file with deep sequencing reads, and optionally a FASTA file with star sequences and the 3-letter code of the species of interest.
- **Output**
- A tab separated file with miRNA identifiers and their rated read count, a signature file, an HTML file that gives an overview of all miRNAs the input data, and a pdf that contains for each miRNA a pdf file showing its signature and structure.

MiRDeep2 quantifier example output



```
Provisional ID           : chr17_967
Score total              :   31090.8
Score for star read(s)   :       3.9
Score for read counts    :   31080.6
Score for mfe            :       1.7
Score for randfold       :       1.6
Score for cons. seed     :         3
Total read count         :     60975
Mature read count        :     60678
Loop read count          :        18
Star read count          :       279
```

```
5'- ggcguugcuuggcugcaacugccgucagccauugaugaucguucuucucuccguauuggggagugagagggagagaacgcggucugagugguuuuuccuucuugaugg  -3'     obs
```

MiRDeep2 quantifier script reference

quantifier.pl

The module maps the deep sequencing reads to predefined miRNA precursors and determines by that the expression of the corresponding miRNAs.

Input

•A FASTA file with precursor sequences,

•a FASTA file with mature miRNA sequences,

•a FASTA file with deep sequencing reads, and

•optionally a FASTA file with star sequences and the 3 letter code of the species of interest.

Output

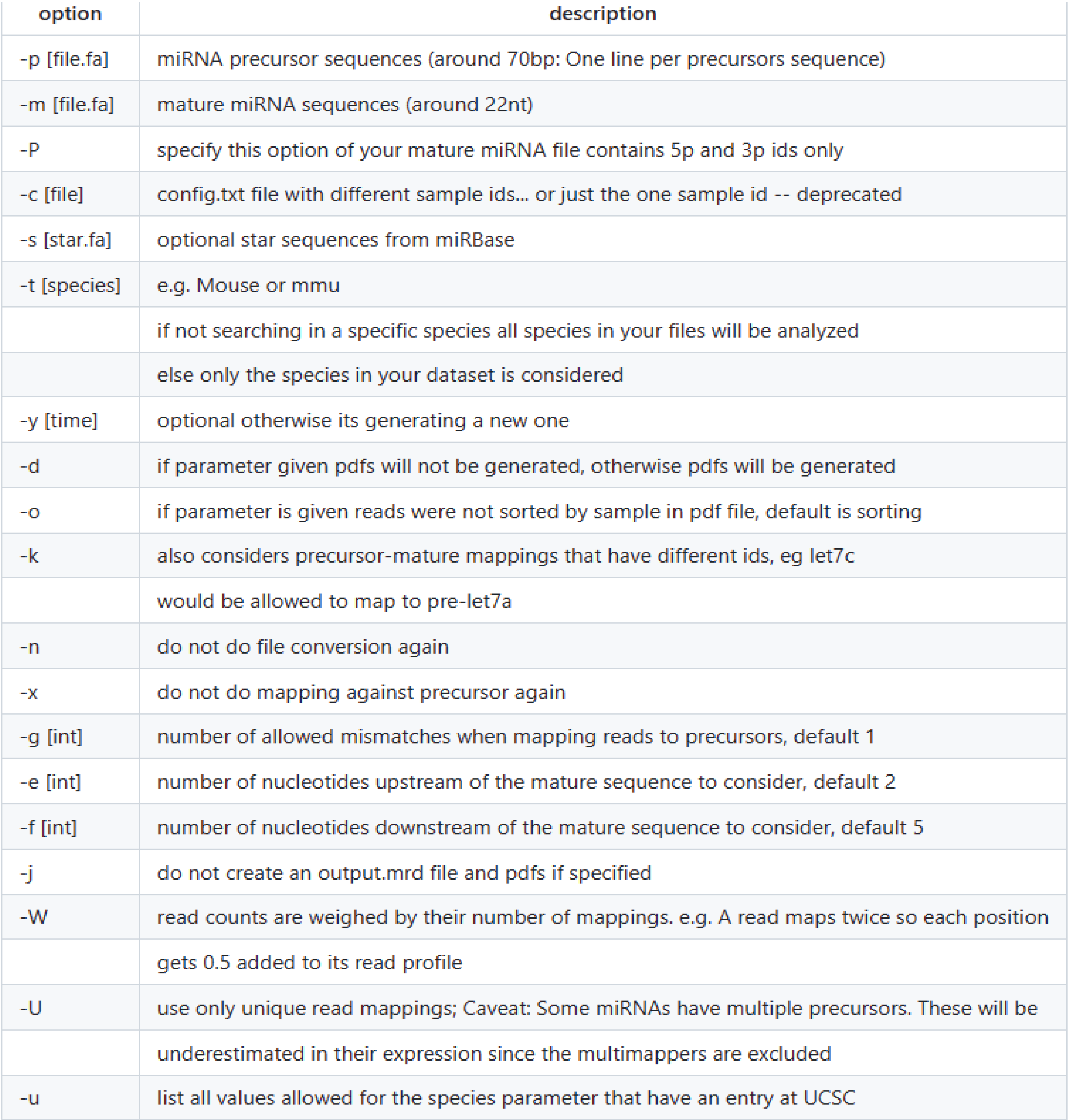
•A 2 column table file called miRNA_expressed.csv with miRNA identifiers and its read count,

•a file called miRNA_not_expressed.csv with all miRNAs having 0 read counts,

•a signature file called miRBase.mrd,

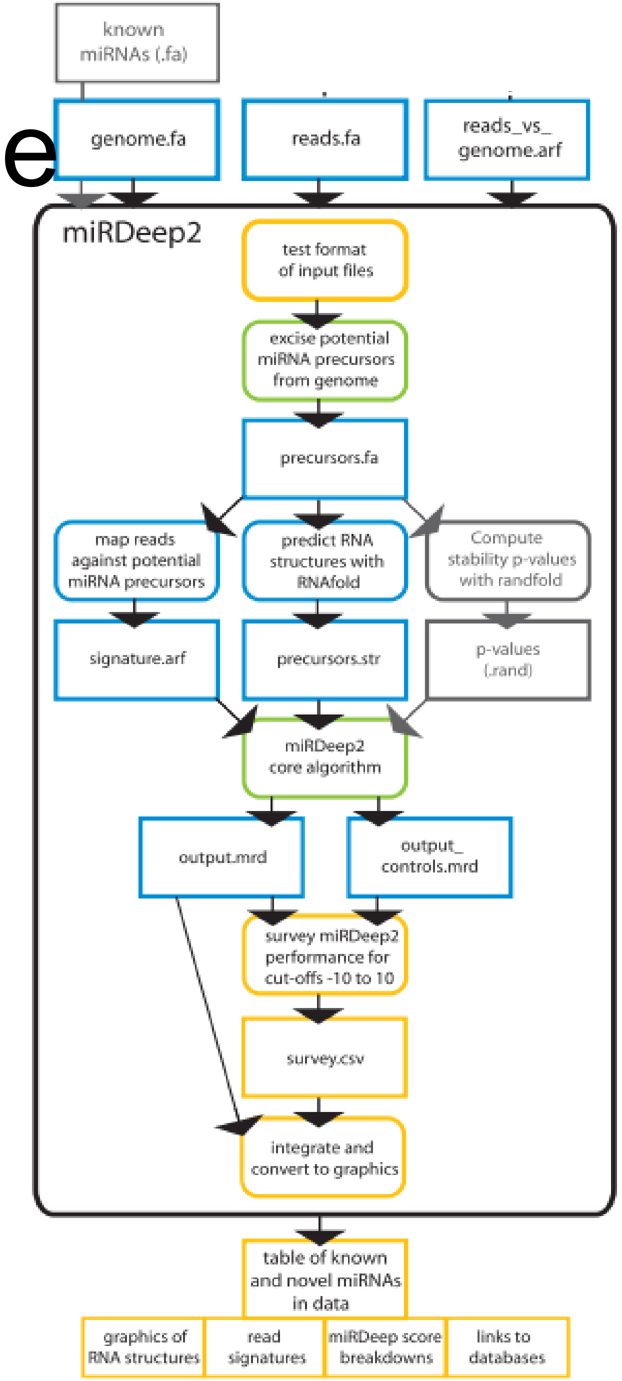
•a file called expression.html that gives an overview of all miRNAs the input data, and

•a directory called pdfs that contains for each miRNA a PDF file showing its signature and structure.



option	description
-p [file.fa]	miRNA precursor sequences (around 70bp: One line per precursors sequence)
-m [file.fa]	mature miRNA sequences (around 22nt)
-P	specify this option of your mature miRNA file contains 5p and 3p ids only
-c [file]	config.txt file with different sample ids... or just the one sample id -- deprecated
-s [star.fa]	optional star sequences from miRBase
-t [species]	e.g. Mouse or mmu
	if not searching in a specific species all species in your files will be analyzed
	else only the species in your dataset is considered
-y [time]	optional otherwise its generating a new one
-d	if parameter given pdfs will not be generated, otherwise pdfs will be generated
-o	if parameter is given reads were not sorted by sample in pdf file, default is sorting
-k	also considers precursor-mature mappings that have different ids, eg let7c
	would be allowed to map to pre-let7a
-n	do not do file conversion again
-x	do not do mapping against precursor again
-g [int]	number of allowed mismatches when mapping reads to precursors, default 1
-e [int]	number of nucleotides upstream of the mature sequence to consider, default 2
-f [int]	number of nucleotides downstream of the mature sequence to consider, default 5
-j	do not create an output.mrd file and pdfs if specified
-W	read counts are weighed by their number of mappings. e.g. A read maps twice so each position
	gets 0.5 added to its read profile
-U	use only unique read mappings; Caveat: Some miRNAs have multiple precursors. These will be
	underestimated in their expression since the multimappers are excluded
-u	list all values allowed for the species parameter that have an entry at UCSC

Flowchart for miRDeep2 module


known miRNAs (.fa)
genome.fa
reads.fa
reads_vs_genome.arf
miRDeep2
test format of input files
excise potential miRNA precursors from genome
precursors.fa
map reads against potential miRNA precursors
predict RNA structures with RNAfold
Compute stability p-values with randfold
signature.arf
precursors.str
p-values (.rand)
miRDeep2 core algorithm
output.mrd
output_controls.mrd
survey miRDeep2 performance for cut-offs -10 to 10
survey.csv
integrate and convert to graphics
table of known and novel miRNAs in data
graphics of RNA structures
read signatures
miRDeep score breakdowns
links to databases

MiRDeep2 script reference

miRDeep2 analyses can be performed using the three scripts miRDeep2.pl, mapper.pl and quantifier.pl.

miRDeep2.pl : Wrapper function for the miRDeep2.pl program package. The script runs all necessary scripts of the miRDeep2 package to perform a microRNA detection deep sequencing data analysis.

Input

•A FASTA file with deep sequencing reads,

•a FASTA file of the corresponding genome,

•a file of mapped reads to the genome in miRDeep2 ARF format,

•an optional FASTA file with known miRNAs of the analyzed species, and

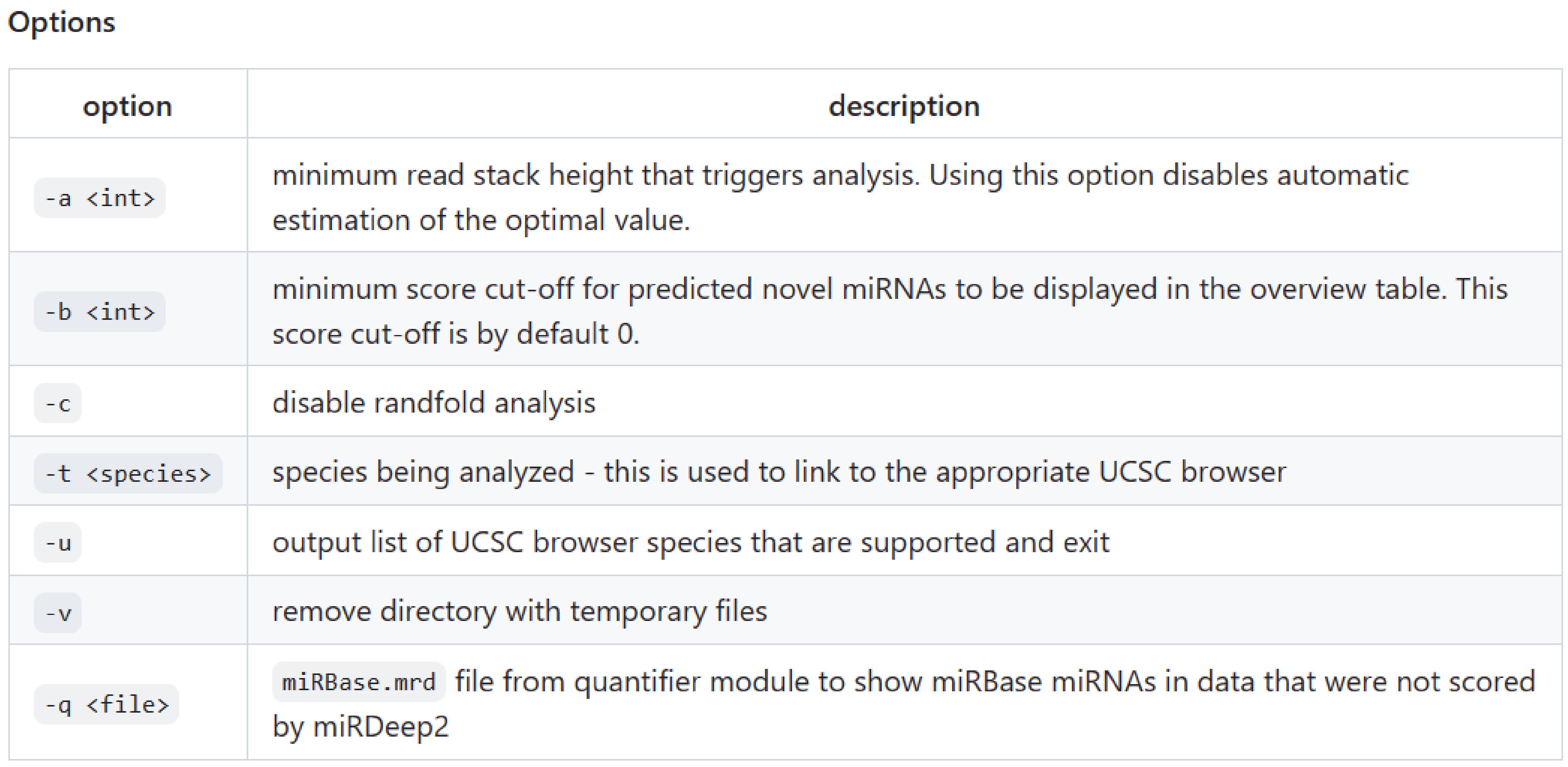
•an optional FASTA file of known miRNAs of related species.

Output

•A spreadsheet and

•an HTML file

with an overview of all detected miRNAs in the deep sequencing input data.

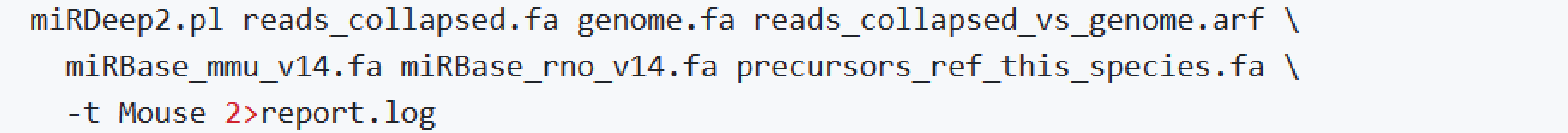


Options

option	description
`-a <int>`	minimum read stack height that triggers analysis. Using this option disables automatic estimation of the optimal value.
`-b <int>`	minimum score cut-off for predicted novel miRNAs to be displayed in the overview table. This score cut-off is by default 0.
`-c`	disable randfold analysis
`-t <species>`	species being analyzed - this is used to link to the appropriate UCSC browser
`-u`	output list of UCSC browser species that are supported and exit
`-v`	remove directory with temporary files
`-q <file>`	`miRBase.mrd` file from quantifier module to show miRBase miRNAs in data that were not scored by miRDeep2

Examples

- For example: The user wishes to identify miRNAs in mouse deep sequencing data, using default options. The miRBase_mmu_v14.fa file contains all miRBase mature mouse miRNAs, while the miRBase_rno_v14.fa file contains all the miRBase mature rat miRNAs. The 2> will pipe all progress output to the report.log file.



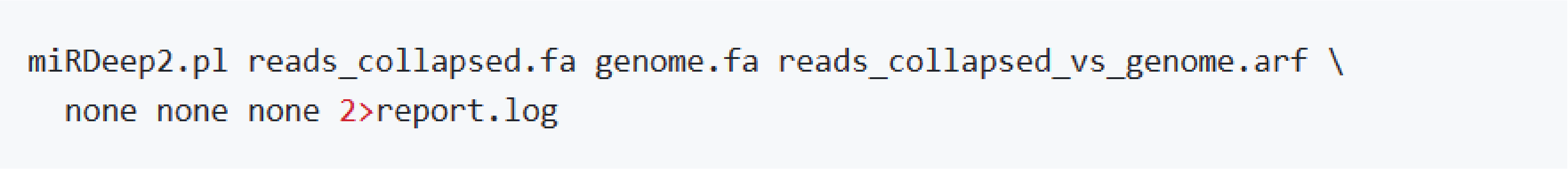
```
miRDeep2.pl reads_collapsed.fa genome.fa reads_collapsed_vs_genome.arf \
  miRBase_mmu_v14.fa miRBase_rno_v14.fa precursors_ref_this_species.fa \
  -t Mouse 2>report.log
```

This command will generate

- a directory with PDFs showing the structures, read signatures, and score breakdowns of novel and known miRNAs in the data,
- an HTML webpage that links to all results generated (result.html),
- a copy of the novel and known miRNAs contained in the webpage but in text format which allows easy parsing (result.csv),
- a copy of the performance survey contained in the webpage but in text format (survey.csv), and
- a copy of the miRNA read signatures contained in the PDFs but in text format (output.mrd).

Example 2

The user wishes to identify miRNAs in deep sequencing data from an animal with no related species in

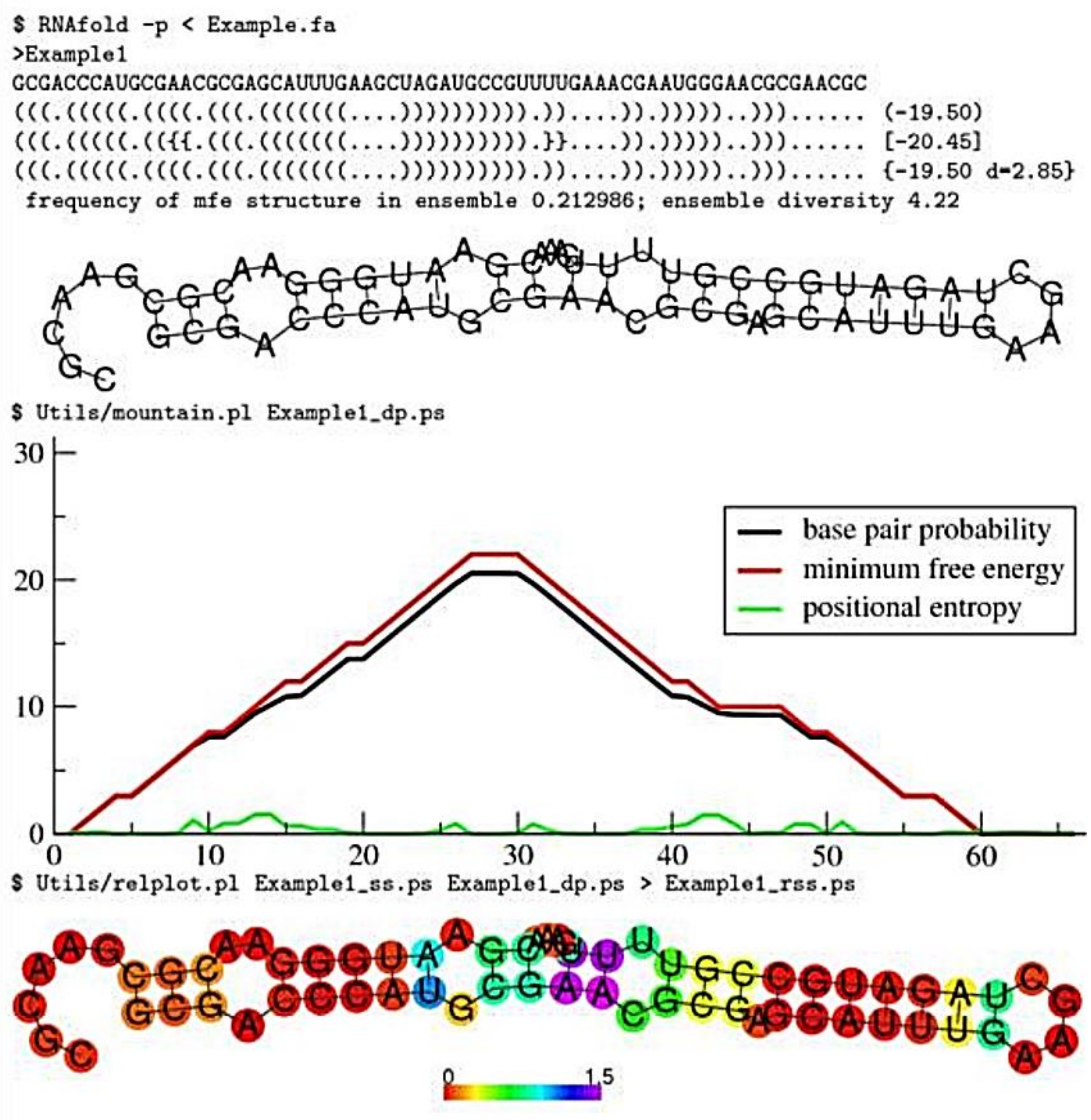


```
miRDeep2.pl reads_collapsed.fa genome.fa reads_collapsed_vs_genome.arf \
  none none none 2>report.log
```

- This command will generate the same type of files as in the example before. Note that there it will in practice always improve miRDeep2 performance if miRNAs from some related species is input, even if it is not closely related.

RNAfold

- Main secondary structure prediction tool
- Computes the minimum free energy (MFE) and backtraces an optimal secondary structure.



```
$ RNAfold -p < Example.fa
>Example1
GCGACCCAUGCGAACGCGAGCAUUUGAAGCUAGAUGCCGUUUUGAAACGAAUGGGAACGCGAACGC
(((.(((((.((((.(((.(((((((....)))))))))).))....)).)))))..)))...... (-19.50)
(((.(((((.(({{.(((.(((((((....)))))))))).}}....)).)))))..)))...... [-20.45]
(((.(((((.((((.(((.(((((((....)))))))))).))....)).)))))..)))...... {-19.50 d=2.85}
 frequency of mfe structure in ensemble 0.212986; ensemble diversity 4.22
```

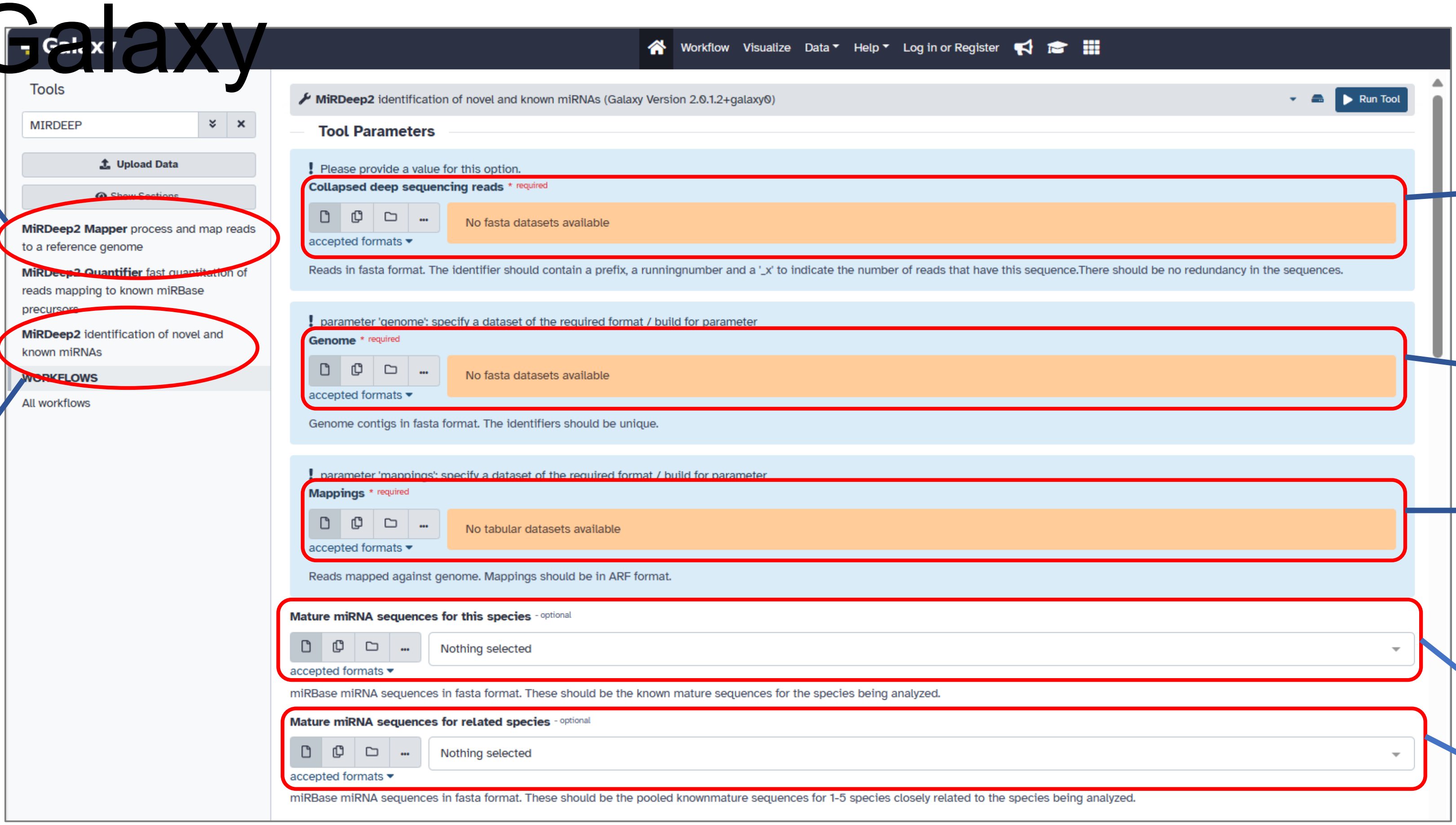
```
$ Utils/mountain.pl Example1_dp.ps
```

```
$ Utils/relplot.pl Example1_ss.ps Example1_dp.ps > Example1_rss.ps
```

MiRDeep2 module homepage in Galaxy

The mapper module is used to transform the raw Illumina sequencing output files (qseq.txt) to the widely used fasta (.fa) files. And maps processed reads to a reference genome

The miRDeep2 module identifies known and novel miRNAs high-throughput sequencing data.



MiRDeep2 Mapper process and map reads to a reference genome

MiRDeep2 Quantifier fast quantitation of reads mapping to known miRBase precursors

MiRDeep2 identification of novel and known miRNAs

MiRDeep2 identification of novel and known miRNAs (Galaxy Version 2.0.1.2+galaxy0)

Tool Parameters

! Please provide a value for this option.

Collapsed deep sequencing reads * required

No fasta datasets available

accepted formats

Reads in fasta format. The identifier should contain a prefix, a runningnumber and a '_x' to indicate the number of reads that have this sequence.There should be no redundancy in the sequences.

! parameter 'genome': specify a dataset of the required format / build for parameter

Genome * required

No fasta datasets available

accepted formats

Genome contigs in fasta format. The identifiers should be unique.

! parameter 'mappings': specify a dataset of the required format / build for parameter

Mappings * required

No tabular datasets available

accepted formats

Reads mapped against genome. Mappings should be in ARF format.

Mature miRNA sequences for this species - optional

Nothing selected

accepted formats

miRBase miRNA sequences in fasta format. These should be the known mature sequences for the species being analyzed.

Mature miRNA sequences for related species - optional

Nothing selected

accepted formats

miRBase miRNA sequences in fasta format. These should be the pooled knownmature sequences for 1-5 species closely related to the species being analyzed.

A set of high-throughput sequencing reads

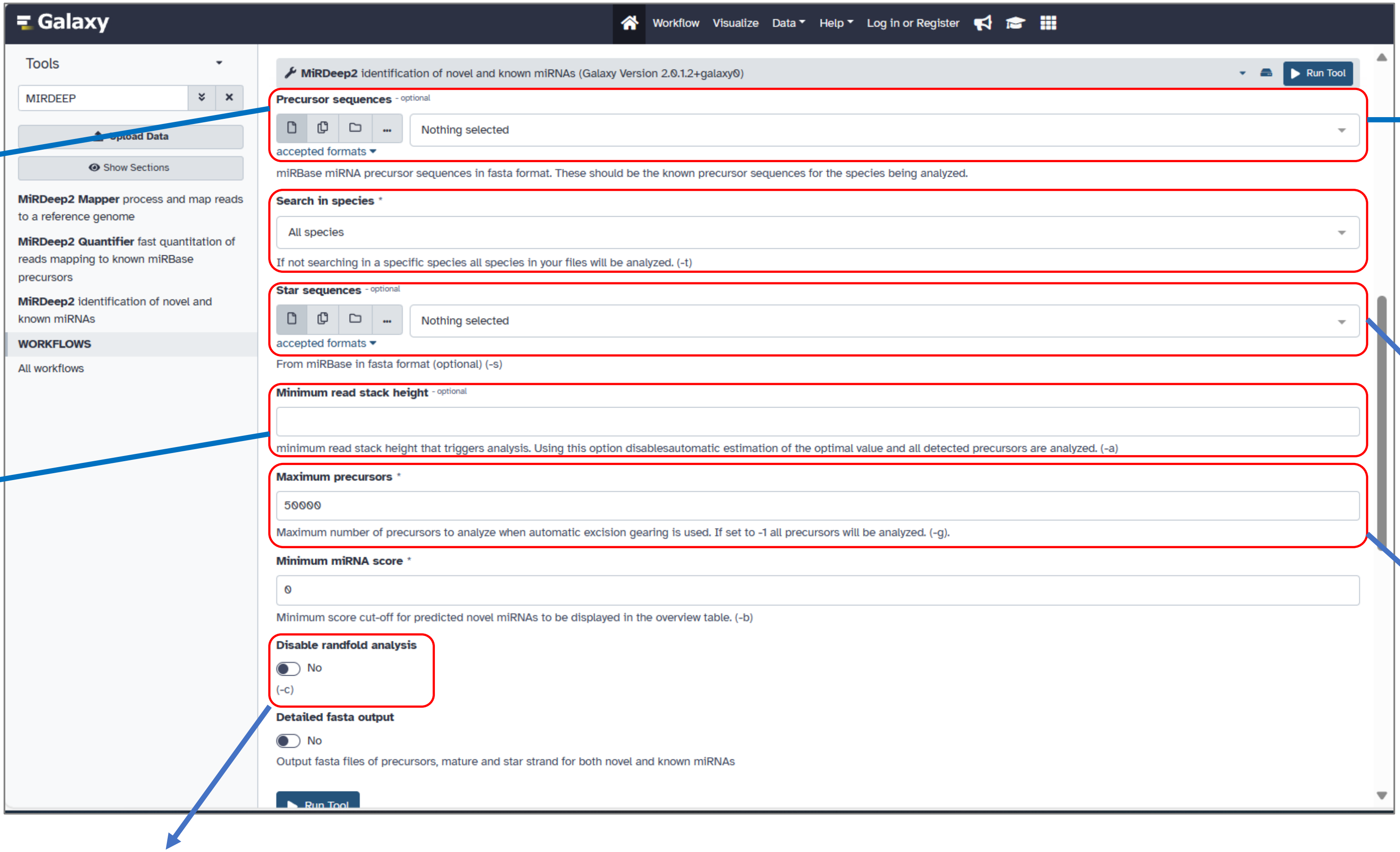
Reference genome ex. Homo sapiens hg38

File with positions of reads mapped against the genome

Optionally, known mature, star, and precursor miRNAs from related species can be input

Potential miRNA precursor sequences excised from the genome using mapped repeats as guidelines

Excision is initiated when the highest stack of reads is encountered within 70nt.



MIRDEEP

MiRDeep2 Mapper process and map reads to a reference genome

MiRDeep2 Quantifier fast quantitation of reads mapping to known miRBase precursors

MiRDeep2 identification of novel and known miRNAs

WORKFLOWS

All workflows

MiRDeep2 identification of novel and known miRNAs (Galaxy Version 2.0.1.2+galaxy0)

Precursor sequences - optional

Nothing selected

accepted formats

miRBase miRNA precursor sequences in fasta format. These should be the known precursor sequences for the species being analyzed.

Search in species *

All species

If not searching in a specific species all species in your files will be analyzed. (-t)

Star sequences - optional

Nothing selected

accepted formats

From miRBase in fasta format (optional) (-s)

Minimum read stack height - optional

minimum read stack height that triggers analysis. Using this option disablesautomatic estimation of the optimal value and all detected precursors are analyzed. (-a)

Maximum precursors *

50000

Maximum number of precursors to analyze when automatic excision gearing is used. If set to -1 all precursors will be analyzed. (-g).

Minimum miRNA score *

0

Minimum score cut-off for predicted novel miRNAs to be displayed in the overview table. (-b)

Disable randfold analysis

No

(-c)

Detailed fasta output

No

Output fasta files of precursors, mature and star strand for both novel and known miRNAs

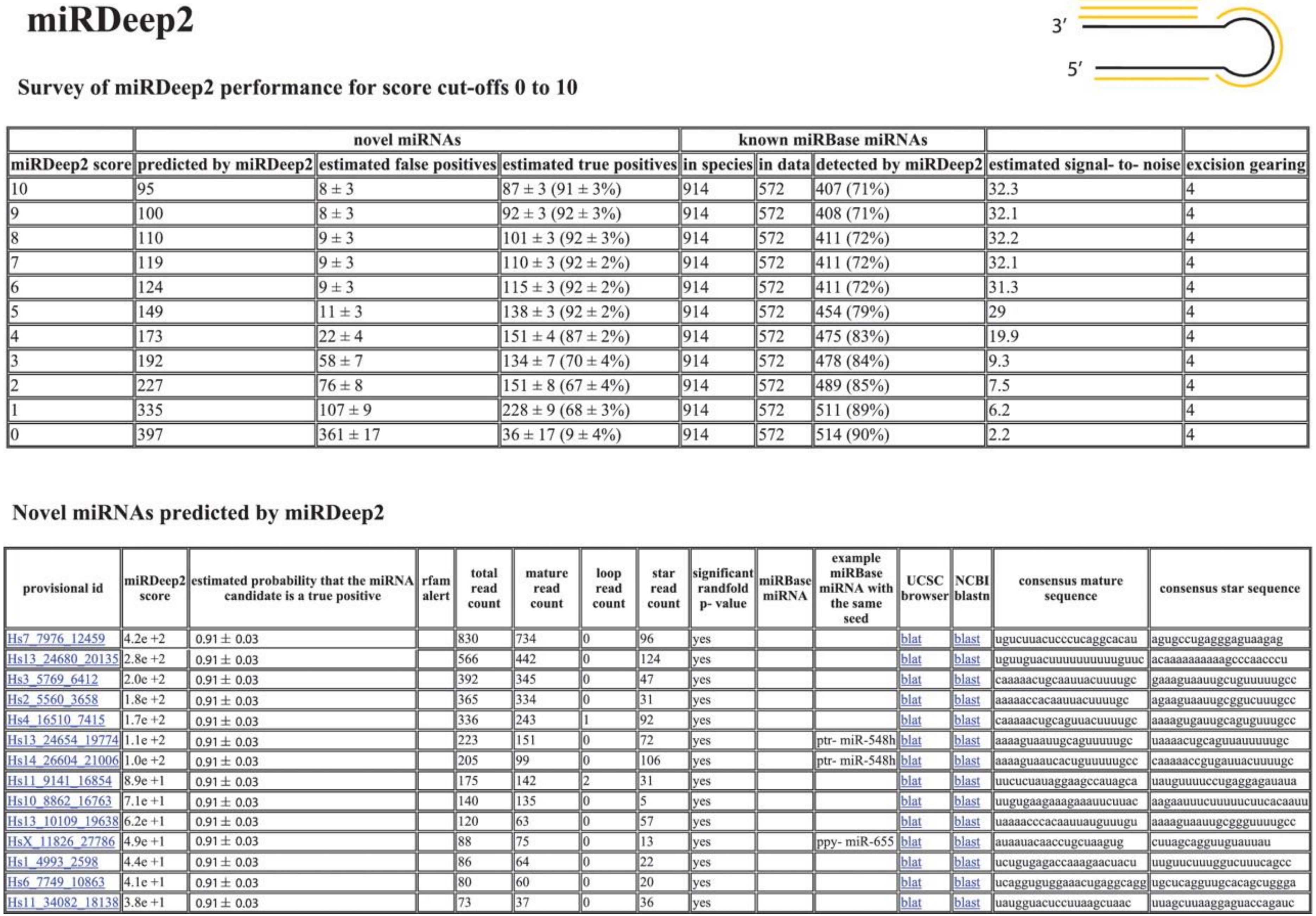
Select species in which the precursor sequences can be searched.
Default: All species

Optionally, input star sequences.

The total number of potential precursor sequences excised should be less than 50,000 (two precursors per genomic locus) for downstream analysis to take place.

randfold P-values are calculated for a subset of potential precursors.

Output



miRDeep2

Survey of miRDeep2 performance for score cut-offs 0 to 10

	novel miRNAs			known miRBase miRNAs				
miRDeep2 score	predicted by miRDeep2	estimated false positives	estimated true positives	in species	in data	detected by miRDeep2	estimated signal- to- noise	excision gearing
10	95	8 ± 3	87 ± 3 (91 ± 3%)	914	572	407 (71%)	32.3	4
9	100	8 ± 3	92 ± 3 (92 ± 3%)	914	572	408 (71%)	32.1	4
8	110	9 ± 3	101 ± 3 (92 ± 3%)	914	572	411 (72%)	32.2	4
7	119	9 ± 3	110 ± 3 (92 ± 2%)	914	572	411 (72%)	32.1	4
6	124	9 ± 3	115 ± 3 (92 ± 2%)	914	572	411 (72%)	31.3	4
5	149	11 ± 3	138 ± 3 (92 ± 2%)	914	572	454 (79%)	29	4
4	173	22 ± 4	151 ± 4 (87 ± 2%)	914	572	475 (83%)	19.9	4
3	192	58 ± 7	134 ± 7 (70 ± 4%)	914	572	478 (84%)	9.3	4
2	227	76 ± 8	151 ± 8 (67 ± 4%)	914	572	489 (85%)	7.5	4
1	335	107 ± 9	228 ± 9 (68 ± 3%)	914	572	511 (89%)	6.2	4
0	397	361 ± 17	36 ± 17 (9 ± 4%)	914	572	514 (90%)	2.2	4

Novel miRNAs predicted by miRDeep2

provisional id	miRDeep2 score	estimated probability that the miRNA candidate is a true positive	rfam alert	total read count	mature read count	loop read count	star read count	significant randfold p- value	miRBase miRNA	example miRBase miRNA with the same seed	UCSC browser	NCBI blastn	consensus mature sequence	consensus star sequence
Hs7_7976_12459	4.2e +2	0.91 ± 0.03		830	734	0	96	yes			blat	blast	ugucuuacucccucaggcacau	agugccugagggaguaagag
Hs13_24680_20135	2.8e +2	0.91 ± 0.03		566	442	0	124	yes			blat	blast	uguuguacuuuuuuuuuuguuc	acaaaaaaaaaagcccaacccu
Hs3_5769_6412	2.0e +2	0.91 ± 0.03		392	345	0	47	yes			blat	blast	caaaaacugcaauuacuuuugc	gaaaguaauugcuguuuuugcc
Hs2_5560_3658	1.8e +2	0.91 ± 0.03		365	334	0	31	yes			blat	blast	aaaaaccacaauuacuuuugc	agaaguaauugcggucuuugcc
Hs4_16510_7415	1.7e +2	0.91 ± 0.03		336	243	1	92	yes			blat	blast	caaaaacugcaguuacuuuugc	aaaagugauugcaguguuugcc
Hs13_24654_19774	1.1e +2	0.91 ± 0.03		223	151	0	72	yes		ptr- miR-548h	blat	blast	aaaaguaauugcaguuuuugc	uaaaacugcaguuauuuuugc
Hs14_26604_21006	1.0e +2	0.91 ± 0.03		205	99	0	106	yes		ptr- miR-548h	blat	blast	aaaaguaaucacuguuuuugcc	caaaaaccgugauuacuuuugc
Hs11_9141_16854	8.9e +1	0.91 ± 0.03		175	142	2	31	yes			blat	blast	uucucuauaggaagccauagca	uauguuuuccugaggagauaua
Hs10_8862_16763	7.1e +1	0.91 ± 0.03		140	135	0	5	yes			blat	blast	uugugaagaaagaaauucuuac	aagaauuucuuuuucuucacaauu
Hs13_10109_19638	6.2e +1	0.91 ± 0.03		120	63	0	57	yes			blat	blast	uaaaacccacaauuauguuugu	aaaaguaauugcggguuuugcc
HsX_11826_27786	4.9e +1	0.91 ± 0.03		88	75	0	13	yes		ppy- miR-655	blat	blast	auaauacaaccugcuaagug	cuuagcagguuguauuau
Hs1_4993_2598	4.4e +1	0.91 ± 0.03		86	64	0	22	yes			blat	blast	ucugugagaccaaagaacuacu	uuguucuuuggucuuucagcc
Hs6_7749_10863	4.1e +1	0.91 ± 0.03		80	60	0	20	yes			blat	blast	ucaggugugga aacugaggcagg	ugcucagguugcacagcuggga
Hs11_34082_18138	3.8e +1	0.91 ± 0.03		73	37	0	36	yes			blat	blast	uaugguacuccuuaagcuaac	uuagcuuaaggaguaccagauc

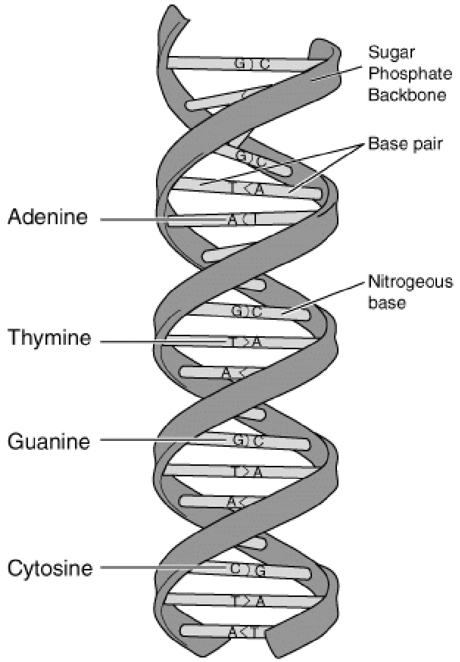
Differential Gene Expression Analysis

Sudhir Srivastava
Division of Agricultural Bioinformatics
ICAR-Indian Agricultural Statistics Research Institute

Introduction

DNA

- double stranded, helical structure
- sequences of nucleotides (A, T, G & C)
- base pairs (A with T and G with C)

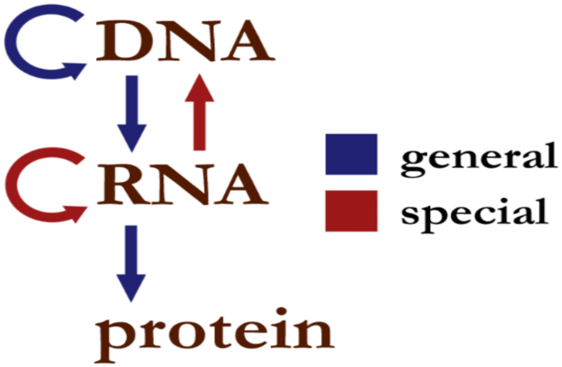


Introduction…

Central Dogma of Molecular Biology

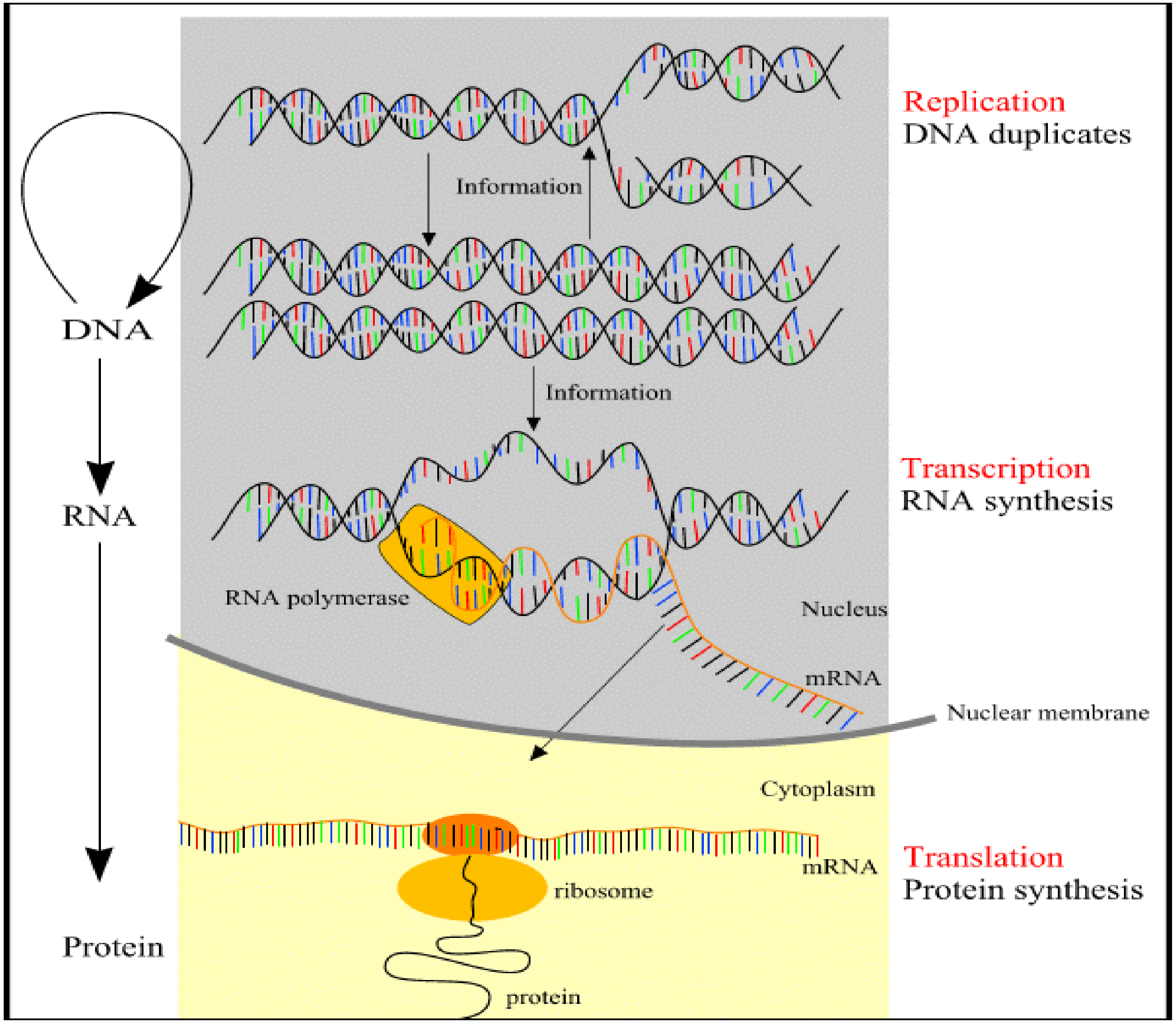
The Central Dogma. This states that once 'information' has passed into protein it cannot get out again. In more detail, the transfer of information from nucleic acid to nucleic acid, or from nucleic acid to protein may be possible, but transfer from protein to protein, or from protein to nucleic acid is impossible. Information means here the precise determination of sequence, either of bases in the nucleic acid or of amino acid residues in the protein. [Francis Crick,1958]

The central dogma of molecular biology deals with the detailed residue-by-residue transfer of sequential information. It states that such information cannot be transferred back from protein to either protein or nucleic acid. [Francis Crick, re-stated in a Nature paper, 1970]



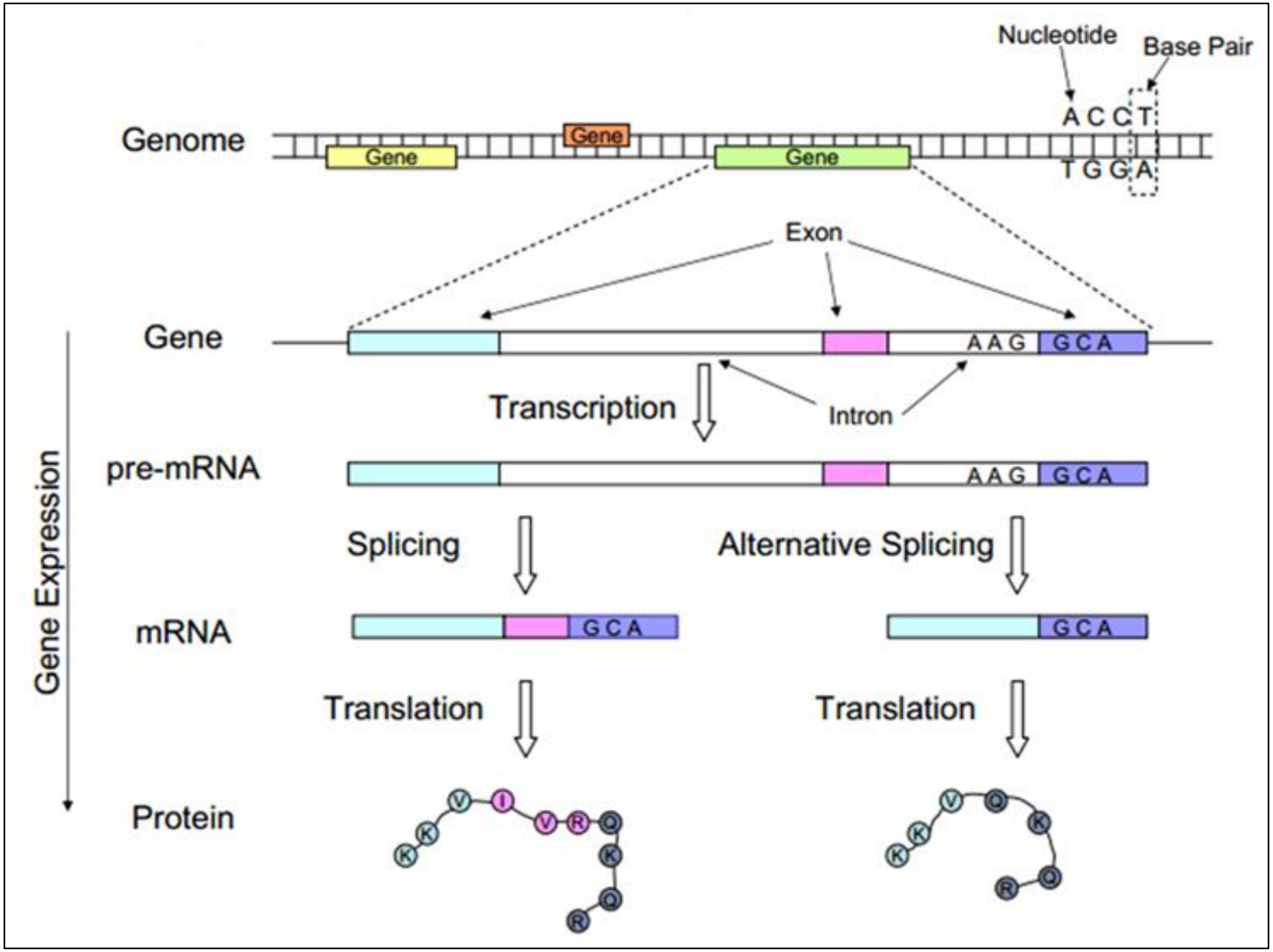
Introduction…

Central Dogma of Molecular Biology



Introduction…

Central Dogma of Molecular Biology

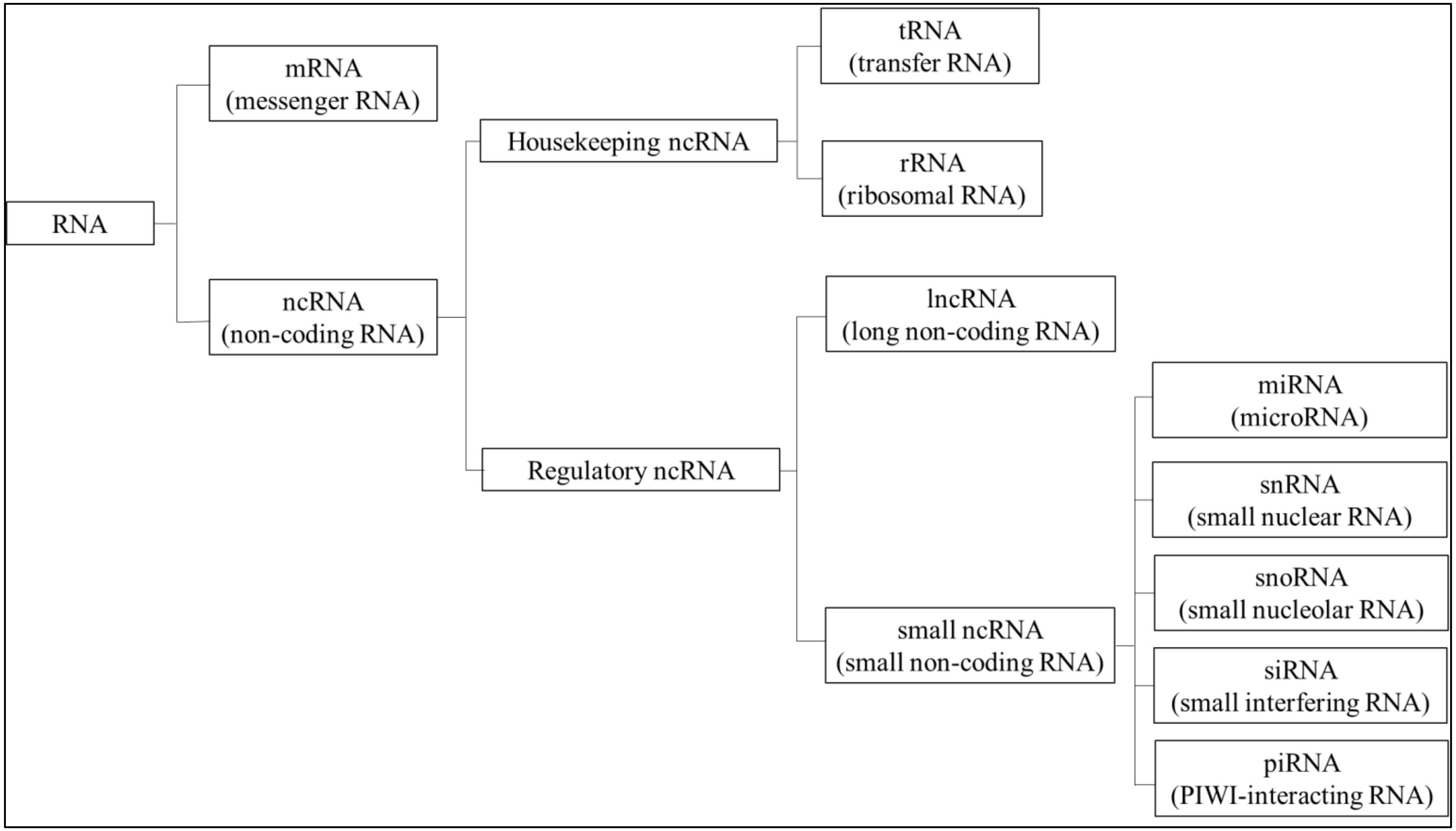


Introduction…

- The advent of Next-Generation Sequencing (NGS) technology has transformed genomic studies.
- One important application of NGS technology is the study of the transcriptome.
- Transcriptome is defined as the complete collection of all the RNA molecules in a cell.

Introduction…

Different types of RNA


RNA
mRNA (messenger RNA)
ncRNA (non-coding RNA)
Housekeeping ncRNA
Regulatory ncRNA
tRNA (transfer RNA)
rRNA (ribosomal RNA)
lncRNA (long non-coding RNA)
small ncRNA (small non-coding RNA)
miRNA (microRNA)
snRNA (small nuclear RNA)
snoRNA (small nucleolar RNA)
siRNA (small interfering RNA)
piRNA (PIWI-interacting RNA)

- All of these molecules are called transcripts since they are produced by process of transcription.
- ~ 2% mRNA

Introduction…

- RNA-Sequencing uses NGS technology to reveal the presence and quantity of RNA in a biological sample at a given moment.
- It allows transcript quantification and differential gene expression analysis.
- Several machines/ protocols are available for generating RNA-Seq data:
  - Illumina (MiSeq, NextSeq, HiSeq, NovaSeq)
  - Ion Torrent (Proton, Personal Genome Machine)
  - SOLiD
  - Roche 454

Introduction…

- Important steps of RNA-Seq experiments:
  - Data generation (experimental design, sample collection, sequencing design, quality control)
  - Quantification of reads to estimate the expression values
  - Normalization
  - Differential expression analysis

Introduction…

- **Applications of RNA-Seq experiments**
  - Quantification of transcriptome/RNA-Seq expression levels to study gene expression in complex experiments
  - Novel gene discovery
  - Gene annotation
  - Detection of differentially expressed features (genes/ transcripts/ exons) between different conditions
  - Detection of splicing events
  - Identification of introns and exon boundaries

Bioinformatics Tools for NGS data preprocessing

Tools for quality check/ filtering/ trimming

- **FASTQC** - A quality control tool for high throughput sequence data

  (https://www.bioinformatics.babraham.ac.uk/projects/fastqc/)

- **NGS QC** - Quality Control

- **FastqCleaner** – A shiny app for Quality Control, Filtering and Trimming of FASTQ Files

- **Trimmomatic** – Trimming of FASTQ files

Bioinformatics Tools for NGS data preprocessing…

- **FASTX toolkit** – A collection of command line tools for Short-Reads FASTA/FASTQ files preprocessing

  (http://hannonlab.cshl.edu/fastx_toolkit/)

- **ShortRead** – R package for filtering and trimming reads, and for generating a quality assessment report

Bioinformatics Tools for NGS data preprocessing…

Samtools: A suite of programs for interacting with high-throughput sequencing data (http://www.htslib.org/)

Three separate repositories:

- Samtools - Reading/writing/editing/indexing/viewing SAM/BAM format
- BCFtools - Reading/writing BCF2/VCF/gVCF files and calling/filtering/summarising SNP and short indel sequence variants
- HTSlib - A C library for reading/writing high-throughput sequencing data

Bioinformatics Tools for NGS data preprocessing…

Short read aligners

- Bowtie
- TOPHAT
- BWA
- Novoalign
- STAR

Bioinformatics Tools for NGS data preprocessing…

de novo assemblers

- SOAPdenovo-Trans
- Trans-AbySS
- Trinity
- SPAdes

Tools for Visualization

- CummeRbund
- IGV
- Bedtools
- UCSC Genome Browser

Experimental design and heterogeneity issues

- The purpose of experimental design is to plan experiment in an effective way so that it can answer the biological question under consideration.

  (i) Biological aspects:
  - Any biological experimental plan starts with a biological question or hypothesis.
  - The experimenter might have some prior knowledge of the question under study before conducting the experiments, e.g., expression levels of some known genes, proteins, etc.

  (ii) Technical aspects:
  - These include the choice of platform and avoiding systematic errors.
  - If the experiment has systematics errors, then the result obtained for comparative analysis will be biased, irrespective of the precision of measurement and the number of experimental units.

  (iii) Economic aspects:
  - Cost of experiment and its analysis
  - Budget available
  - Time required to complete the experiment and its analysis
  - Whether pilot study is required or not, etc.

Experimental design and heterogeneity issues…

Other points to be considered:

- Availability of enough samples for experiment;
- Availability of enough RNA, DNA or proteins from samples;
- Biopsies collected from same part of tissue or other tissues;
- Number of replicates required;
- Effect size, *etc*.

Heterogeneity

- A heterogeneous sample or population means that every observed data has different value for the corresponding characteristic of interest.
- There may be various factors responsible for influencing expression in any feature.
- The major sources of variations are due to technical, genetic, demographic and environmental factors.

Experimental design and heterogeneity issues…

- There are two important points to be considered while designing RNA-Seq experiments which are namely, the sequencing depth and the number of replicates (biological and technical) required to observe significant changes in expression.
- The cost can be reduced by optimizing the designing process of these experiments.
- Tools and software for sample size estimation and power analysis:
  - RNASeqPowerCalculator
  - RNASeqPower
  - Scotty
  - PROPER

RNA-Seq Experiments

- The basic steps for summarizing a typical RNA-Seq experiment:
  - Purified RNA is converted to cDNA, fractionated, ligated with technology specific adapters and sequencing is done.
  - Millions of short read sequences are generated from one end (single-end) or both ends (paired-end) of the cDNA fragments.
  - These sequences are aligned to a reference genome.
  - The number of reads mapped to known features are recorded and summarized in a table.
- The features can be either genes, transcripts (alternative transcripts) or exon level expression.

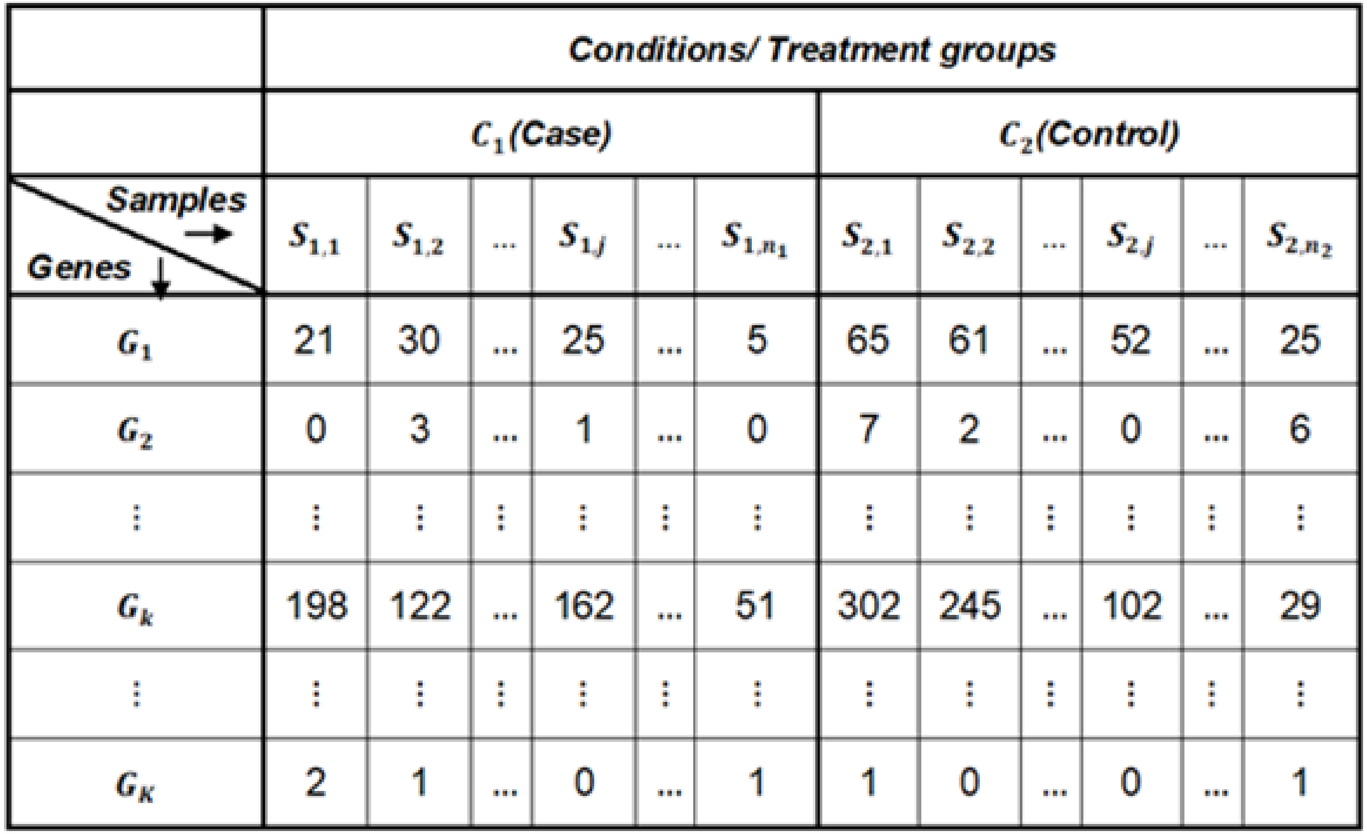
RNA-Seq Experiments…

Example of a biological experiment with I conditions/groups denoted by $G_i(i = 1, 2, \dots, I)$ having N_i individuals/samples denoted by $S_{i,j}$ $(j =$

	$\boldsymbol{G_1}$					…	$\boldsymbol{G_i}$					…	$\boldsymbol{G_I}$				
	$\boldsymbol{S_{1,1}}$	…	$\boldsymbol{S_{1,j}}$	…	$\boldsymbol{S_{1,N_1}}$		$\boldsymbol{S_{i,1}}$	…	$\boldsymbol{S_{i,j}}$	…	$\boldsymbol{S_{i,N_i}}$		$\boldsymbol{S_{I,1}}$	…	$\boldsymbol{S_{I,j}}$	…	$\boldsymbol{S_{I,N_I}}$
$\boldsymbol{F_1}$	$y_{1,1,1}$	…	$y_{1,j,1}$	…	$y_{1,N_1,1}$		$y_{i,1,1}$	…	$y_{i,j,1}$	…	$y_{i,N_i,1}$		$y_{I,1,1}$	…	$y_{I,j,1}$	…	$y_{I,N_I,1}$
⋮	⋮	⋮	⋮	⋮	⋮		⋮	⋮	⋮	⋮	⋮		⋮	⋮	⋮	⋮	⋮
$\boldsymbol{F_k}$	$y_{1,1,k}$	…	$y_{1,j,k}$	…	$y_{1,N_1,k}$		$y_{i,1,k}$	…	$\boldsymbol{y_{i,j,k}}$	…	$y_{i,N_i,k}$		$y_{I,1,k}$	…	$y_{I,j,k}$	…	$y_{I,N_I,k}$
⋮	⋮	⋮	⋮	⋮	⋮		⋮	⋮	⋮	⋮	⋮		⋮	⋮	⋮	⋮	⋮
$\boldsymbol{F_K}$	$y_{1,1,K}$	…	$y_{1,j,K}$	…	$y_{1,N_1,K}$		$y_{i,1,K}$	…	$y_{i,j,K}$	…	$y_{i,N_i,K}$		$y_{I,1,K}$	…	$y_{I,j,K}$	…	$y_{I,N_I,K}$

RNA-Seq Experiments…

A table of read counts for a hypothetical case-control study



	Conditions/ Treatment groups											
	C_1(Case)						C_2(Control)					
Samples → Genes ↓	$S_{1,1}$	$S_{1,2}$	...	$S_{1,j}$	...	S_{1,n_1}	$S_{2,1}$	$S_{2,2}$	...	$S_{2,j}$	...	S_{2,n_2}
G_1	21	30	...	25	...	5	65	61	...	52	...	25
G_2	0	3	...	1	...	0	7	2	...	0	...	6
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
G_k	198	122	...	162	...	51	302	245	...	102	...	29
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
G_K	2	1	...	0	...	1	1	0	...	0	...	1

Transcript quantification

- The most common application of RNA-seq is to estimate gene and transcript expression.
- This application is primarily based on the number of reads that map to each transcript sequence.
- The simplest approach to quantification is to aggregate raw counts of mapped reads using programs such as HTSeq-count or featureCounts.
- Metrics to normalize data considering the gene length and sequencing depth
  - RPKM (reads aligned per kilobase of exon per million reads mapped)
  - FPKM (fragments per kilobase of exon per million fragments mapped)
  - TPM (transcripts per kilobase million)
- Normalization is required before performing the differential expression analysis.

Transcript quantification…

- htseq-count
- featureCounts
- Cufflinks
- Stringtie
- RSEM
- Sailfish

Differential Expression Analysis

- One of the primary goals for RNA-seq experiments is to compare the gene expression levels across various experimental conditions, treatments, tissues, or time points.
- The researchers are particularly interested in detecting gene with differential expressions.
- The study of determining which genes have changed significantly in terms of their expression across two or more conditions is referred to as differential expression analysis.
- Identification of differentially expressed genes helps researchers to understand the functions of genes in response to a given condition.

Differential Expression Analysis…

- A large number of statistical models and tools have been developed to perform differential expression analysis for RNA-Seq data.
- Differential expression analysis methods for RNA-Seq can be grouped into two broad categories:

- **Parametric method**
  - It captures all information about the data within the parameters.
  - Each expression value for a given gene is mapped into a particular distribution, such as Poisson or negative binomial.
- **Non-parametric method**
  - A non-parametric model uses a flexible number of parameters.
  - The number of parameters often grows as it learns from more data.
  - A non-parametric model is computationally slower, but makes fewer assumptions about the data.

RNA-Seq Experiments…

Estimation of parameters based on NB distribution

- The estimation of parameters is an essential step for design, sample size calculation and differential expression analysis.

- The parameter estimation can be done by using various methods such as method of moments estimation (MME), maximum likelihood estimation (MLE), maximum quasi-likelihood estimation (MQLE).

- Besides these methods, there are various methods/models for estimation of parameters such as pseudo-likelihood, quasi-likelihood, conditional maximum likelihood (CML), conditional inference, quantile-adjusted CML, conditional weighted likelihood.

RNA-Seq Experiments…

Estimation of parameters based on NB distribution without scaling factor

- Let Y_{ij} be a NB random variable with mean μ_i and dispersion parameter ϕ_i, i.e., $Y_{ij}\sim NB(\mu_i,\phi_i)$, then its probability mass function is given by

$$p(Y_{ij}=y_{ij})=\frac{\Gamma\left(y_{ij}+\frac{1}{\phi_i}\right)}{\Gamma\left(\frac{1}{\phi_i}\right)\Gamma(y_{ij}+1)}\frac{(\mu_i\phi_i)^{y_{ij}}}{(1+\mu_i\phi_i)^{y_{ij}+\frac{1}{\phi_i}}};y=0,1,2,\dots$$

- The likelihood function is given by

$$L(\mu_i,\phi_i|y_{ij};\ j=1,2,\dots,N_i)=\prod_{j=1}^{N_i}\frac{\Gamma\left(y_{ij}+\frac{1}{\phi_i}\right)}{\Gamma\left(\frac{1}{\phi_i}\right)\Gamma(y_{ij}+1)}\frac{(\mu_i\phi_i)^{y_{ij}}}{(1+\mu_i\phi_i)^{y_{ij}+\frac{1}{\phi_i}}}$$

- The log-likelihood function is given by

$$\begin{aligned}
&l(\mu_i,\phi_i|y_{ij};\ j=1,2,\dots,N_i)\\
&=\sum_{j=1}^{N_i}\ln\Gamma\left(y_{ij}+\frac{1}{\phi_i}\right)-\sum_{j=1}^{N_i}\Gamma\left(\frac{1}{\phi_i}\right)-\sum_{j=1}^{N_i}\ln\Gamma(y_{ij}+1)\\
&+\sum_{j=1}^{N_i}y_{ij}\ln(\mu_i\phi_i)-\sum_{j=1}^{N_i}\left(y_{ij}+\frac{1}{\phi_i}\right)\ln(1+\mu_i\phi_i)
\end{aligned}$$

Differential Expression Analysis…

Method	Read count distribution assumption/model	Normalization	Differential analysis test
edgeR	Negative binomial distribution	TMM/ Upper quartile / RLE / None (all scaling factors are set to be one)	Exact test analogous to Fisher's exact test or likelihood ratio test
DESeq	Negative binomial distribution	DESeq size factors	Exact test analogous to Fisher's exact test
DESeq2	Negative binomial distribution	DESeq size factors	Wald test
baySeq	Negative binomial distribution	Scaling factors (quantile/ TMM/ total)	Posterior probability through Bayesian approach
EBSeq	Negative binomial-beta empirical Bayes model	DESeq median normalization	
SAMseq	Non-parametric method	Based on the read count mean over the null features of data set.	Wilcoxon rank statistics based permutation test
NOIseq	Non-parametric method	RPKM / TMM / Upper quartile	Corresponding logarithm of fold change and absolute expression differences have a high probability than noise values
limma+voom	Similar to t-distribution with empirical Bayes approach	TMM	Moderated t-test

Differential Expression Analysis…

Tools for Differential Expression Analysis

- Cufflinks package
- R packages: DESeq, DESeq2, edgeR

edgeR for RNA-Seq Data Analysis

1. Download and Install R

https://cran.r-project.org/bin/windows/base/

2. Download and Install RStudio

https://www.rstudio.com/products/rstudio/download/#download

3. Open RStudio

4. Install the required R packages: Here, we will install edgeR.

```
if (!requireNamespace("BiocManager", quietly = TRUE))

    install.packages("BiocManager")

BiocManager::install("edgeR")
```

edgeR for RNA-Seq Data Analysis…

https://bioconductor.org/packages/release/bioc/html/edgeR.html

Example: A paired design RNA-seq experiment of oral squamous cell carcinomas and matched normal tissue from three patients

- The aim of the analysis is to detect genes differentially expressed between tumor and normal tissue, adjusting for any differences between the patients.
- RNA was sequenced on an Applied Biosystems SOLiD System 3.0 and reads mapped to the UCSC hg18 reference genome.
- Read counts, summarised at the level of refSeq transcripts are available in Table S1 of Tuch *et al*. (https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2824832/).

Online Tool for RNA-Seq Data Analysis

http://bioinformatics.sdstate.edu/idep/

https://kcvi.shinyapps.io/START/

References

Conesa, A., Madrigal, P., Tarazona, S., Gomez-Cabrero, D., Cervera, A., McPherson, A., Szcześniak, M. W., Gaffney, D. J., Elo, L. L., Zhang, X., & Mortazavi, A. (2016). A survey of best practices for RNA-seq data analysis. *Genome biology*, **17**, 13. https://doi.org/10.1186/s13059-016-0881-8

Costa-Silva, J., Domingues, D., & Lopes, F. M. (2017). RNA-Seq differential expression analysis: An extended review and a software tool. *PloS one*, **12(12)**, e0190152. https://doi.org/10.1371/journal.pone.0190152

Li D. (2019). Statistical Methods for RNA Sequencing Data Analysis. In: Husi H, editor. Computational Biology [Internet]. Brisbane (AU): Codon Publications; Chapter 6. Available from: https://www.ncbi.nlm.nih.gov/books/NBK550334/; doi:10.15586/computationalbiology.2019.ch6

References…

Ge, S.X., Son, E.W. & Yao, R. iDEP: an integrated web application for differential expression and pathway analysis of RNA-Seq data. BMC Bioinformatics 19, 534 (2018). https://doi.org/10.1186/s12859-018-2486-6

Nelson, JW, Sklenar J, Barnes AP, Minnier J. (2016) “The START App: A Web-Based RNAseq Analysis and Visualization Resource.” Bioinformatics. doi: 10.1093/bioinformatics/btw624.

McCarthy DJ, Chen Y, Smyth GK (2012). “Differential expression analysis of multifactor RNA-Seq experiments with respect to biological variation.” Nucleic Acids Research, 40(10), 4288-4297. doi: 10.1093/nar/gks042.

Robinson MD, McCarthy DJ, Smyth GK (2010). “edgeR: a Bioconductor package for differential expression analysis of digital gene expression data.” Bioinformatics, 26(1), 139-140. doi: 10.1093/bioinformatics/btp616.

Protein Structure Prediction

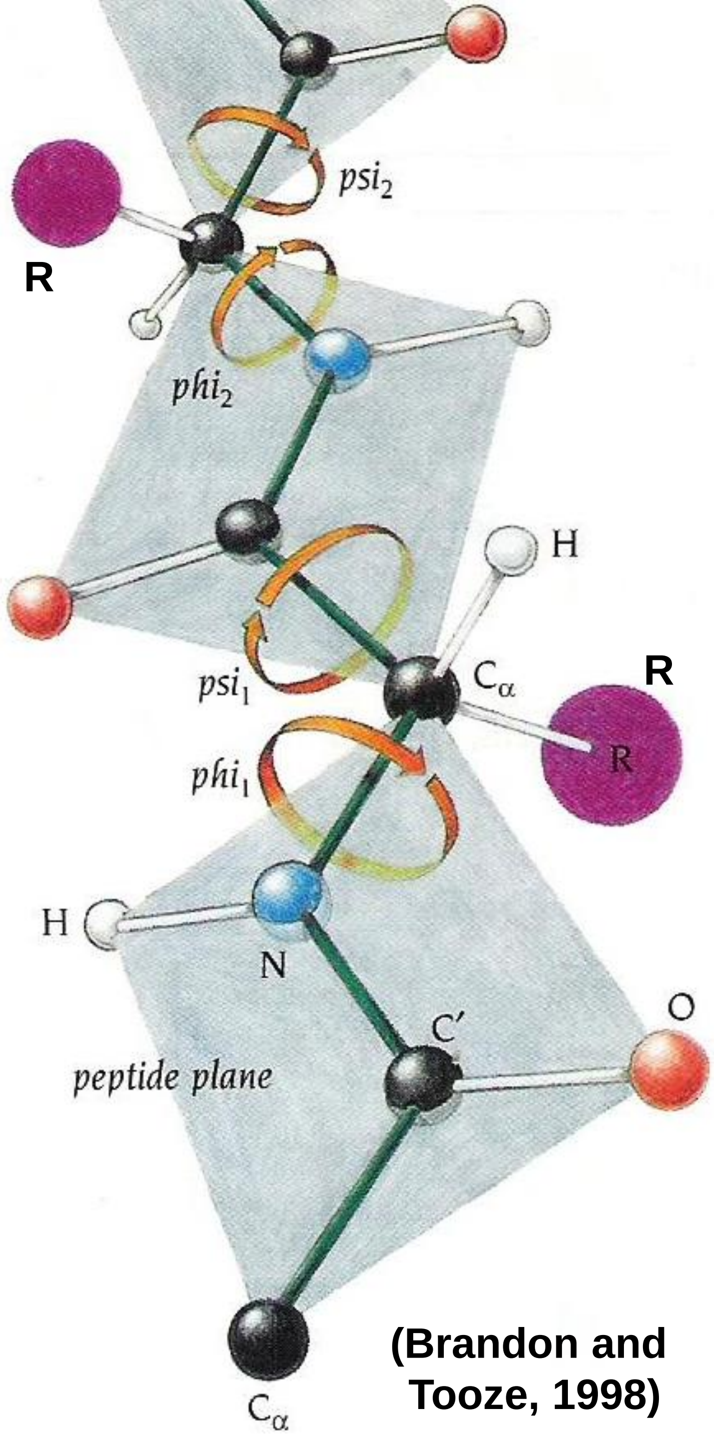
Dr. Sunil Kumar

Principal scientist
Division of Agricultural Bioinformatics
E-mail: skybiotech@gmail.com

ICAR-Indian Agricultural Statistics Research Institute
Library Avenue, Pusa, New Delhi 110012

Protein Primary Structures

- Amino acid sequence of a polypeptide chain.
- 20 amino acids, each with a different side chain (R).
- Peptide units are building blocks of protein structures.
- The angle of rotation around the N–C_α bond is called phi (ϕ), and the angle around the C_α–C' bond from the same C_α atom is called psi (ψ).



(Brandon and Tooze, 1998)

Protein Secondary Structures

- **Local substructures as a result of hydrogen bond formation between neighboring amino acids (backbone interactions).**

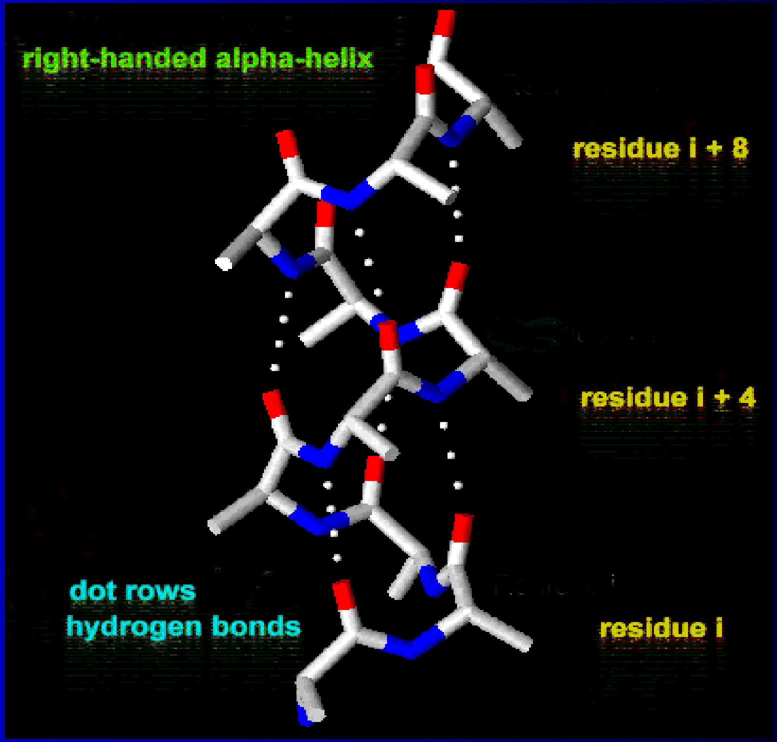
- **The amino acid side chains affect secondary structure formation.**

- **Types of secondary structures:**
  - α helix,
  - β sheet,
  - Loop or random coil.

α Helix

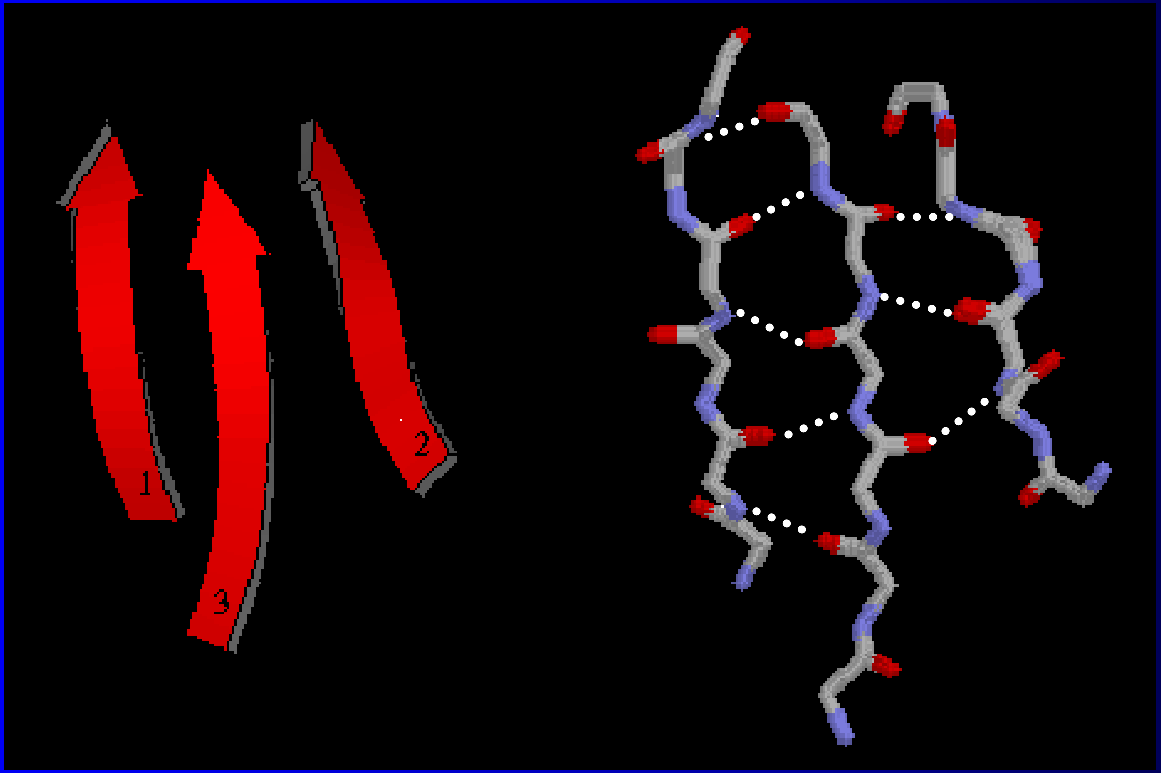
- Most abundant secondary structure.
- 3.6 amino acids per turn, and hydrogen bond formed between every fourth residue.
- Often found on the surface of proteins.

Michael Summers, HHMI at UMBC



β Sheet

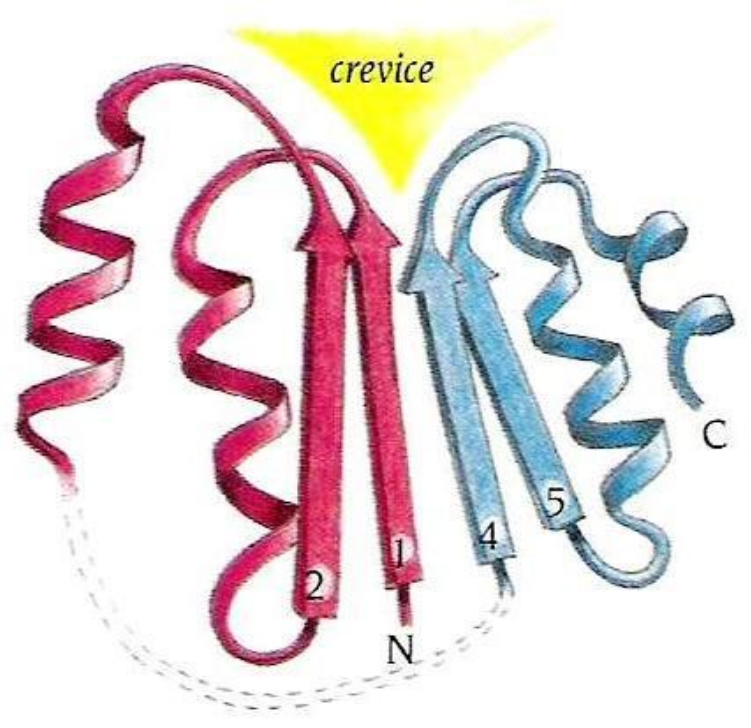
- Hydrogen bonds formed between adjacent polypeptide chains.
- The chain directions can be same (parallel sheet), opposite (anti-parallel), or mixed.



Loop or Coil

- Regions between α helices and β sheets.
- Various lengths and 3-D configurations.
- Often functionally significant (*e.g.*, part of an active site).

The active site of open α/β-barrel structures is in a crevice outside the carboxy ends of the β strands.

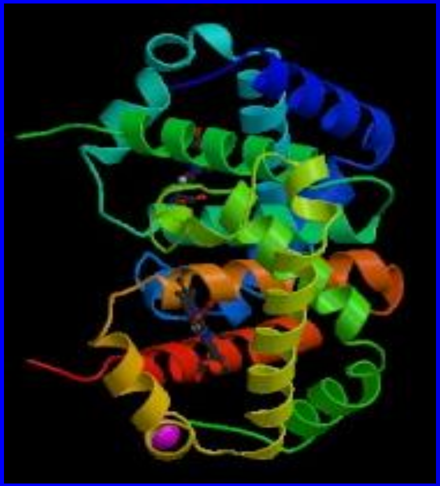


(Brandon and Tooze, 1998)

Protein Tertiary Structure

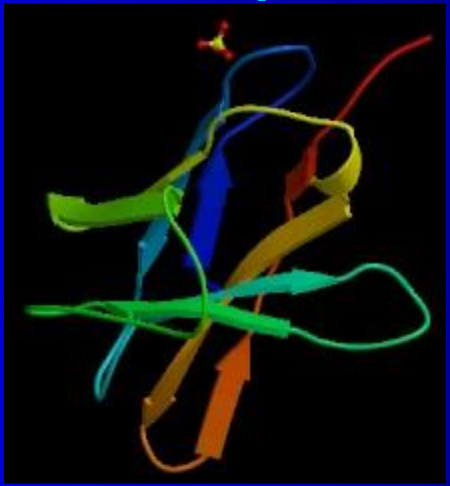
- The 3-D structure of a protein is assembled from different secondary structure components.
- Tertiary structure is determined primarily by hydrophobic interactions between side chains.
- Different classes of protein structures:

All α



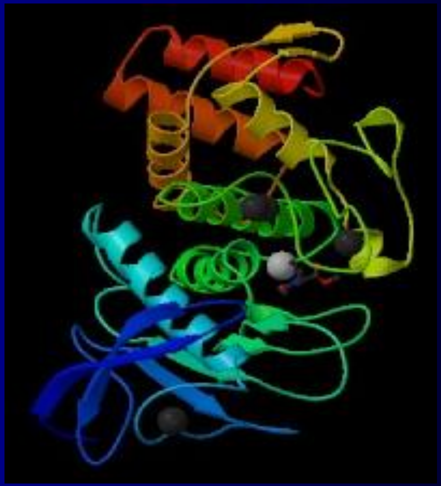
Hemoglobin (3HHB)

All β



T cell CD8 (1CD8)

Mixed



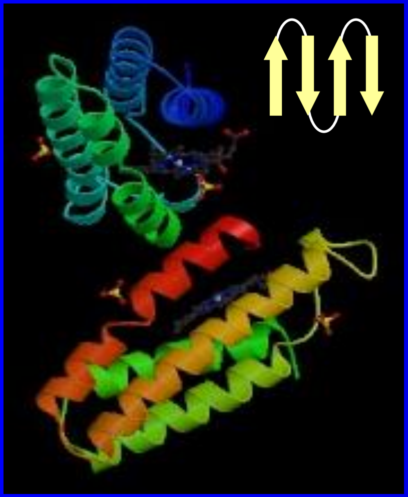
Thermolysin (7TLN)

Protein Tertiary Structure (Cont'd)

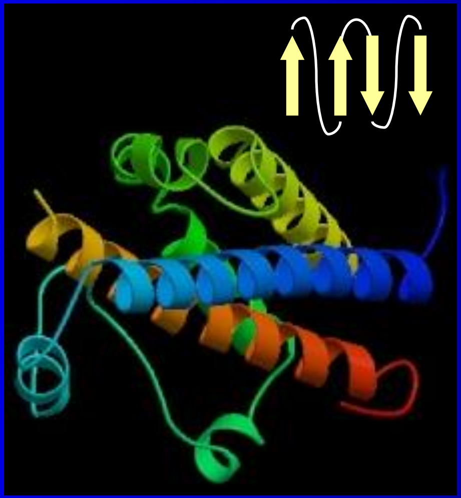
- Fold: a certain type of 3-D arrangement of secondary structures.

- Protein structures evolves more slowly than primary amino acid sequences.

Four-helix bundles

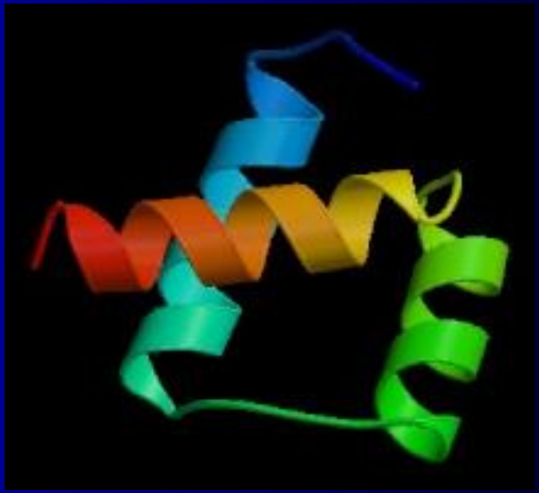


E. coli cytochrome b562 (256B)



Human growth hormone (1HUW)

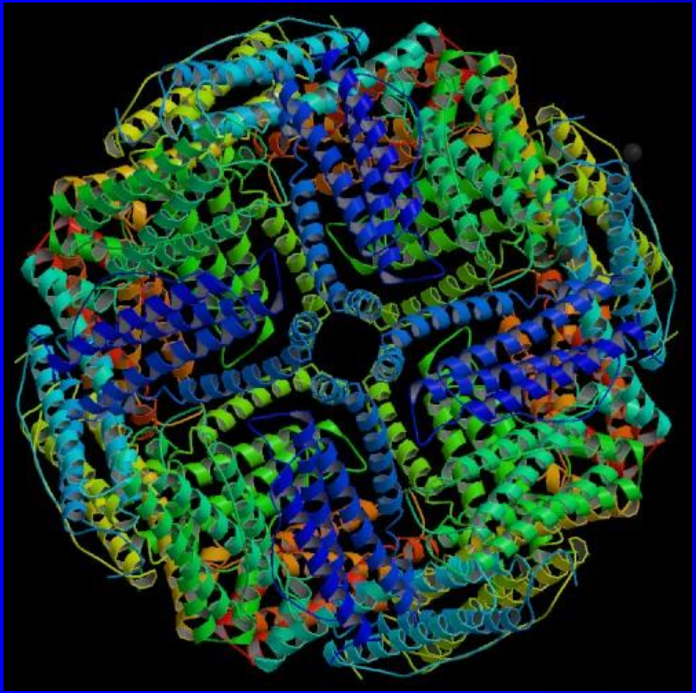
Three-helix bundle



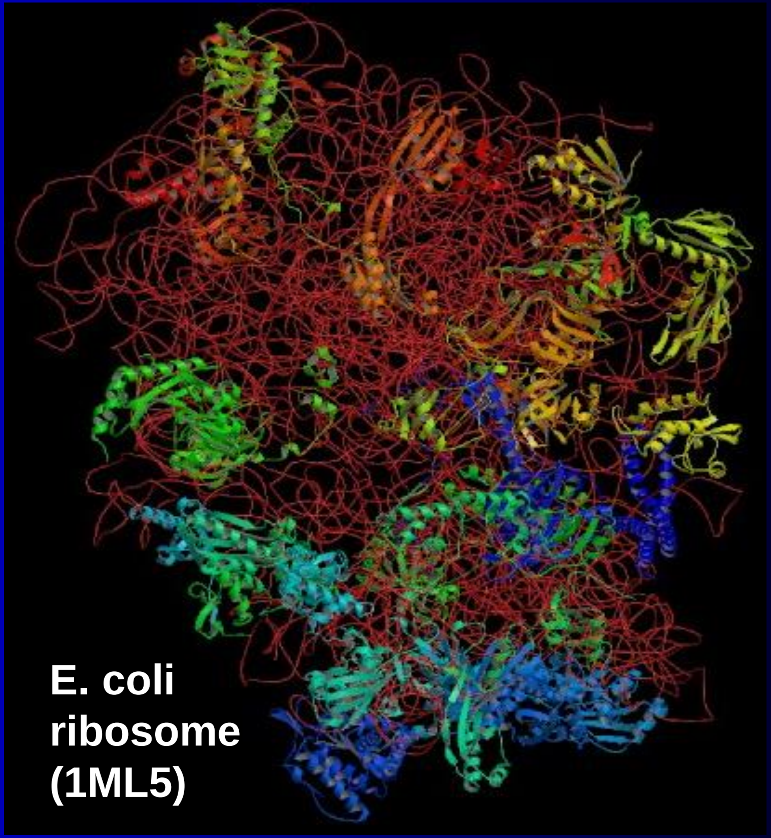
Drosophila engrailed homeodomain (1ENH)

Protein Quaternary Structure

- Two or more independent tertiary structures are assembled into a larger protein complex.
- Important for understanding protein-protein interactions.

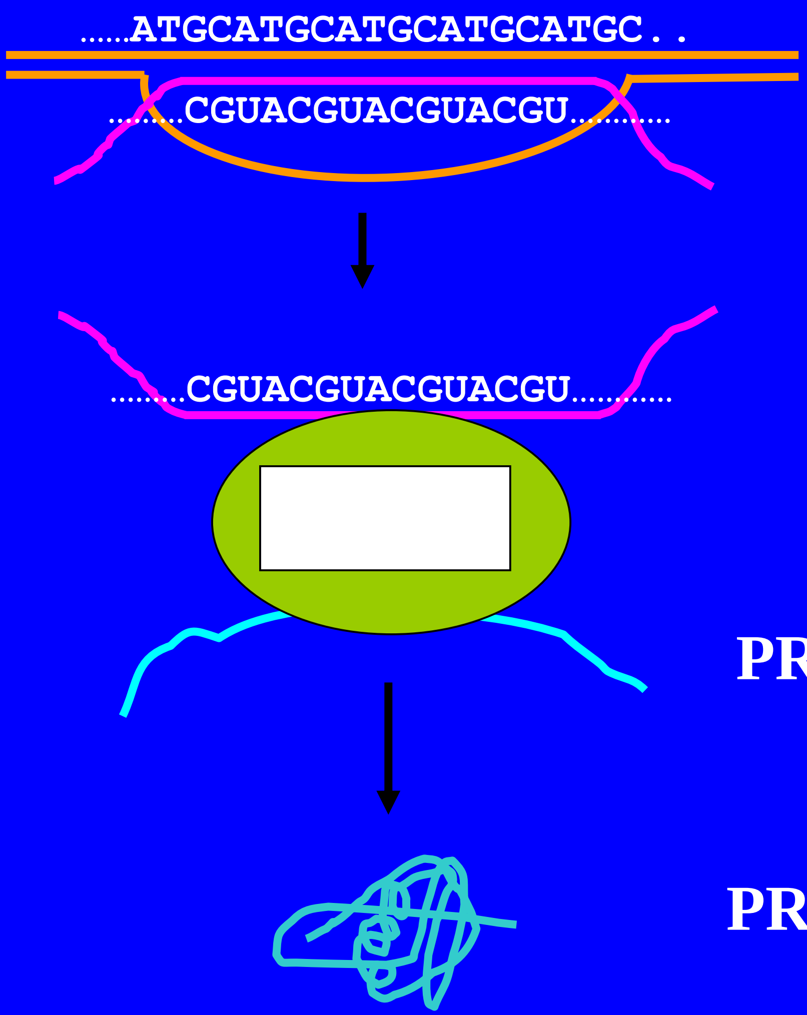


Horse spleen ferritin (1IES)



E. coli ribosome (1ML5)

Information Transfer pathway within the cell



......ATGCATGCATGCATGCATGC . .
.........CGUACGUACGUACGU............
.........CGUACGUACGUACGU............

DNA

RNA

PROTEIN Sequence

PROTEIN Structure

Biological function

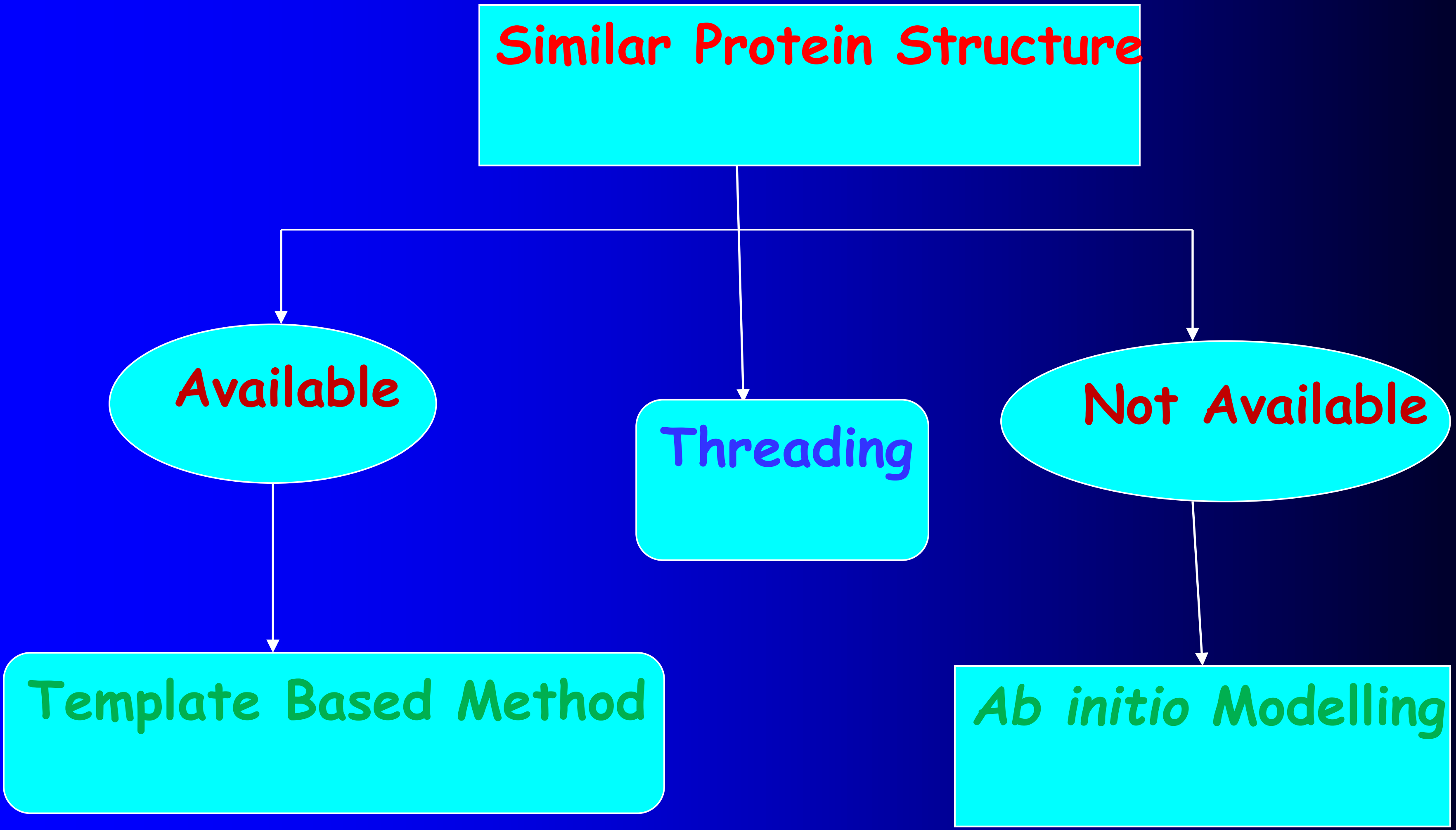
From Sequence to Structure

Protein structure is hierarchic:

- Primary – sequence of covalently attached amino acid
- Secondary – local 3D patterns (helices, sheets, loops)
- Tertiary – overall 3D fold
- Quaternary – two or more protein chains

Motivation to Acquire a Structure

- Identifying active and binding sites
- Characterization of the protein's mechanism (catalysis & interactions)
- Searching for ligand of a given binding site
- Understanding the molecular basis of diseases
- Designing mutants
- Drug design
- And more...



Similar Protein Structure

Available

Threading

Not Available

Template Based Method

Ab initio Modelling

General Scheme

1. Searching for structures related to the query sequence
2. Selecting templates
3. Aligning query sequence with template structures
4. Building a model for the query using information from the template structures
5. Evaluating the model

What is Homology Modeling?

An approach to predict a model of the three-dimensional structure of a given protein sequence (TARGET) based on an alignment to one or more known protein structures (TEMPLATES)

The homology modeling method is based on the assumption that the structure of an unknown protein is similar to known structures of reference proteins

Why a Model?

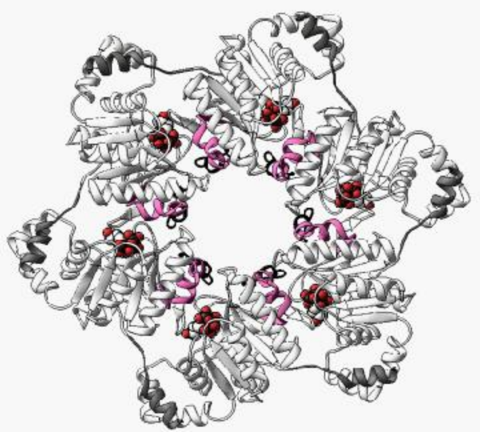
A model is desirable when either X-ray crystallography or NMR spectroscopy can not determine the structure of a protein in time or at all.

While the 3-D structure of proteins can be determined by x-ray crystallography and NMR spectroscopy. These experimental techniques are time consuming and not possible if a sufficient quantity and quality of a proteins is not available.

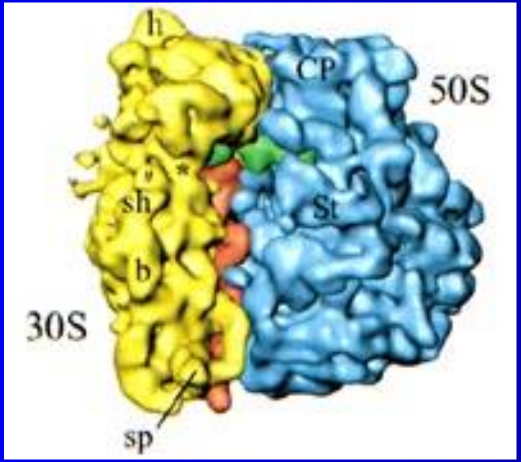
The built model provides a wealth of information of how the protein functions with information at residue property level. This information can than be used for mutational studies or for drug design..

Protein Structure Determination

- High-resolution structure determination
  - X-ray crystallography (~1Å)
  - Nuclear magnetic resonance (NMR) (~1-2.5Å)



- Low-resolution structure determination
  - Cryo-EM (electron-microscropy) ~10-15Å



X-ray crystallography

- most accurate
- An extremely pure protein sample is needed.
- The protein sample must form crystals that are relatively large without flaws. **Generally the biggest problem.**
- Many proteins aren't amenable to crystallization at all (i.e., proteins that do their work inside of a cell membrane).
- ~$100K per structure

Nuclear Magnetic Resonance

- Fairly accurate
- No need for crystals
- limited to small, soluble proteins only.

Steps in homology modelling

Target's sequence

↓

1. Identification of structures that will form the template for modelling

2. Sequence Alignment of the target with template

3. Transfer of the coordinates from the template(s) to the target of structurally conserved regions (SCR's)

4. Modelling the missing regions

5. Refinement and validation of the model

↓

Target's structure

Template search

- Homology modeling is based on using similar structures
  i.e. **no Similar structures = No Model**

- 40% amino acid identity or higher is best; below that is not advisable but examples of success do exist

- Need sequence similarity across the whole sequence, not just in one part

Sequence Alignment

```
GGTGGATCTA
   III II
   GGA–CT - GTAC
```

Key Step:

Sequence alignment of the target with the basis structures

Good Alignment

↓

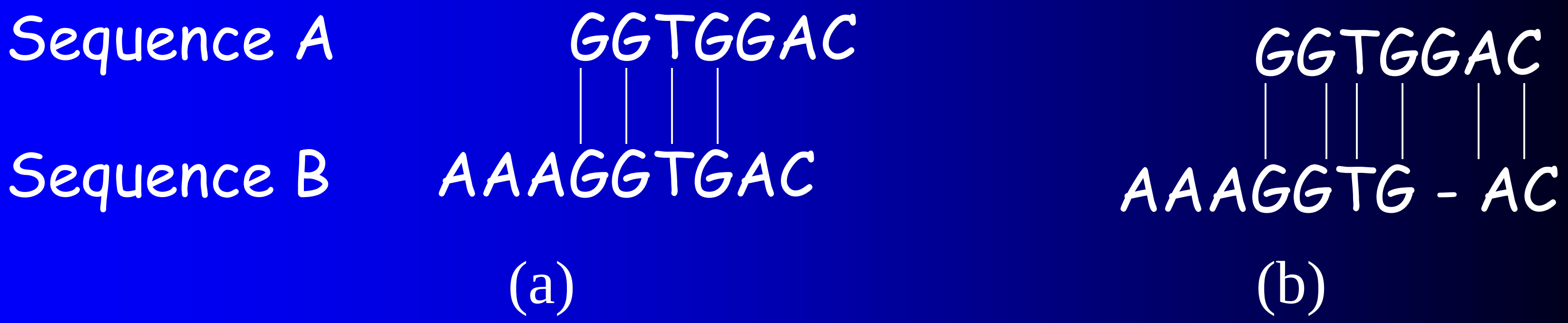
Good Model

- Sequence alignment is a basic technique in homology modeling.

- It is used to establish a one-to-one correspondence between the amino acids of the reference protein (template) and those of the unknown protein (target) in the structurally conserved regions.

- The correspondence is the basis for transferring coordinates from the reference to the model protein

- **What is sequence alignment ?**
  - To find out the conserved residues the residues of one sequence are directly mapped on to the residues of other sequence. The process of mapping is called sequence alignment.

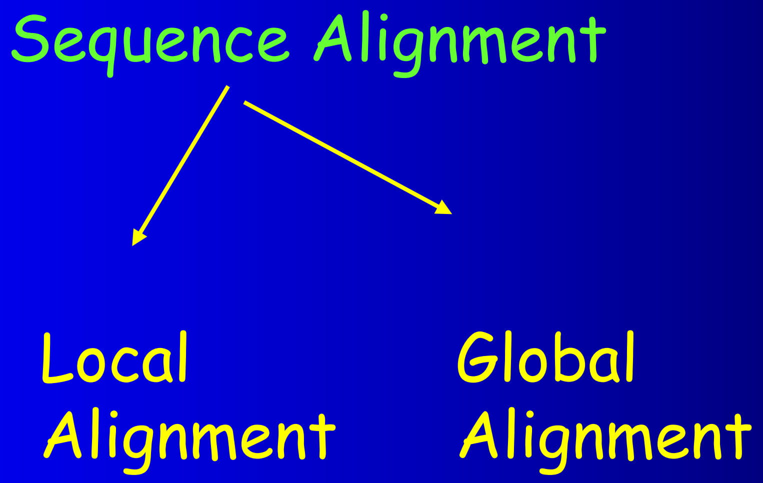


```
Sequence A           GGTGGAC                    GGTGGAC
                     ||||                       ||||  ||
Sequence B        AAAGGTGAC                  AAAGGTG-AC
                     (a)                        (b)
```

A Sample alignment of two DNA sequences

(a) Un-gapped alignment

(b) Gapped alignment. The “I” indicates matching nucleotides



Sequence Alignment

Local Alignment

Global Alignment

- In global alignment whole sequences are consider where as in local alignment only parts of sequences are consider.

- **Basic Goal:** To achieve an alignment which gives rise to maximum number of matches. (i.e. high sequence similarity)

Applications:

Global alignment : essential for comparative modeling.

Local alignment : sufficient for functional domains.

N.B: Global alignment is computationally more time consuming than the local alignment.

Computational Methods for Alignment

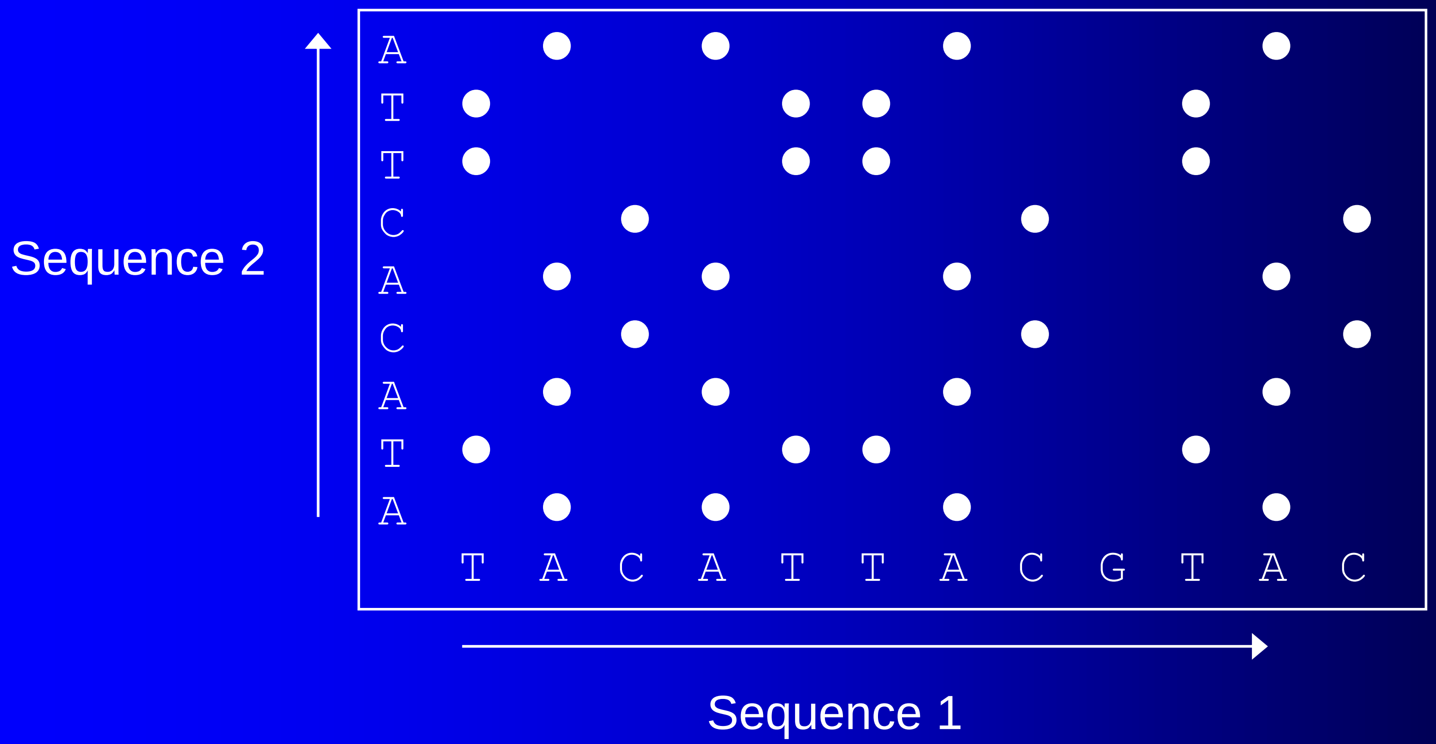
- Dot - matrix analysis
- Dynamic programming (DP) algorithms
- Heuristic methods

Dot matrix analysis

- simple graphical method
- used for finding regions of local matches between two sequences.
- The two sequences to be compared are placed as row and column of matrix.
- All the residues of the first sequence placed column wise are compared against all the residues of the second sequence placed row wise.
- Whenever a match is found a dot is placed on the corresponding position in the matrix.

Dotplot:

A dotplot gives an overview of all possible alignments



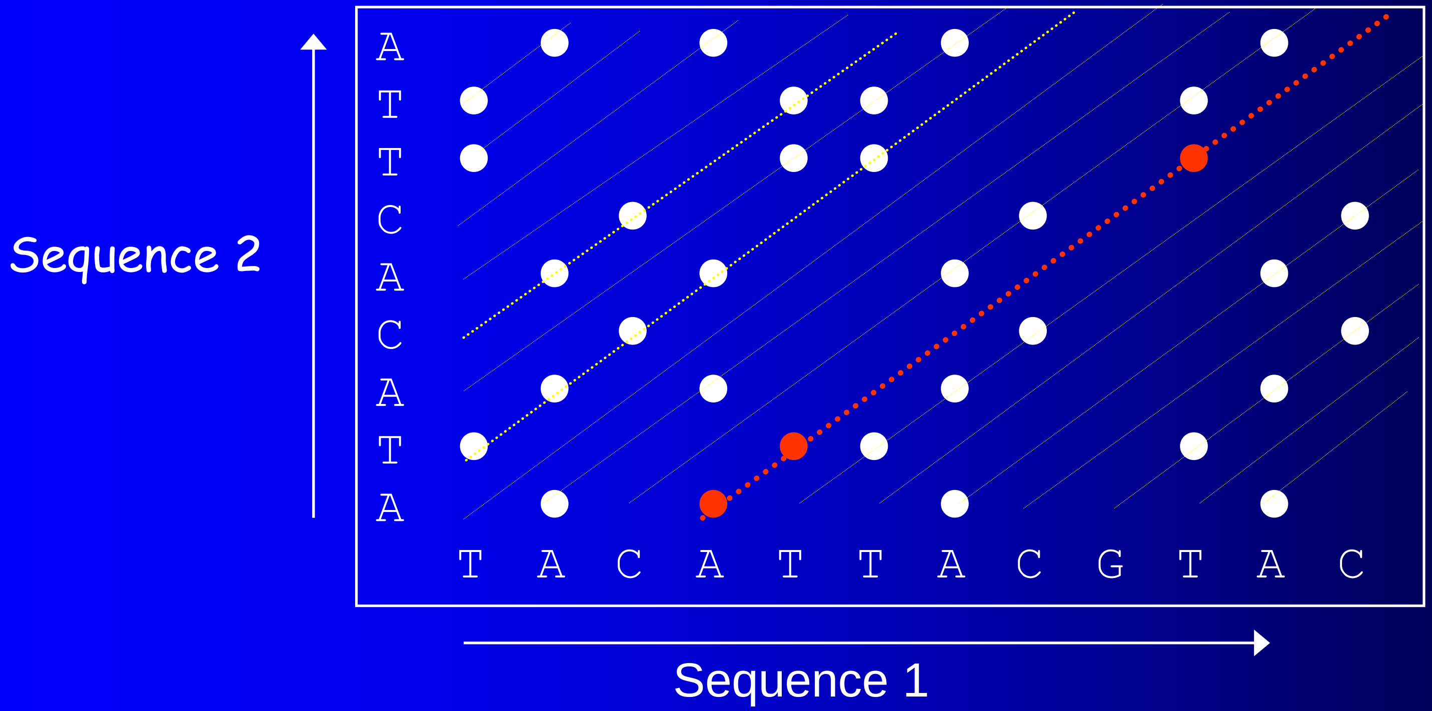
Sequence 2

A		●		●			●				●	
T	●				●	●				●		
T	●				●	●				●		
C			●					●				●
A		●		●			●				●	
C			●					●				●
A		●		●			●				●	
T	●				●	●				●		
A		●		●			●				●	
	T	A	C	A	T	T	A	C	G	T	A	C

Sequence 1

Dotplot:

In a dotplot each diagonal corresponds to a possible (ungapped) alignment



One possible alignment:

```
T A C A T T A C G T A C
      | |       |
      A T A C A C T T A
```

Dynamic Programming

Automatic procedure that finds the **best alignment** with an **optimal score** depending on the chosen parameters.

- Needleman and Wunsch Algorithm

  - **Global** Alignment -

- Smith and Waterman Algorithm

  - **Local** Alignment -

Needleman and Wunsch

(global alignment)

Sequence 1:	HEAGAWGHEE
Sequence 2:	PAWHEAE
Scoring parameters:	BLOSUM50
Gap penalty:	Linear gap penalty of 8

Basic principles of dynamic programming

- Creation of an **alignment path matrix**
- **Stepwise** calculation of score values
- **Backtracking** (evaluation of the optimal path)

Creation of ………(..contd..)

Idea:

Build up an optimal alignment using previous solutions for optimal alignments of smaller subsequences

- Construct matrix *F* indexed by *i* and *j* (one index for each sequence)
- *F*(*i*,*j*) is the score of the best alignment between the initial segment $x_{1...i}$ of *x* up to x_i and the initial segment $y_{1...j}$ of *y* up to y_j
- Build *F*(*i*,*j*) recursively beginning with *F*(0,0) = 0

Creation of ……..(..contd..)

- If *F*(*i*-1,*j*-1), *F*(*i*-1,*j*) and *F*(*i*,*j*-1) are known we can calculate *F*(*i*,*j*)

- Three possibilities:
  - x_i and y_j are aligned, $F(i,j) = F(i-1,j-1) + s(x_i, y_j)$ ↘
  - x_i is aligned to a gap, $F(i,j) = F(i-1,j) - d$ →
  - y_j is aligned to a gap, $F(i,j) = F(i,j-1) - d$ ↓

- The best score up to (*i*,*j*) will be the **largest** of the three options

Smith-Waterman Algorithm

- compares segments of all possible lengths (LOCAL alignments) and chooses whichever maximises the similarity measure.
- calculates ALL possible paths leading to each cell
- paths can be of any length and can contain insertions and deletions

Smith and Waterman

(local alignment)

Two differences:

$$1.\ F(i, j) = \max \begin{cases} 0 & \\ F(i, j) = F(i\text{-}1, j\text{-}1) + s(x_i, y_j) & \searrow \\ F(i, j) = F(i\text{-}1, j) - d & \rightarrow \\ F(i, j) = F(i, j\text{-}1) - d & \downarrow \end{cases}$$

2. An alignment can now end anywhere in the matrix

Example:

Sequence 1	H E A G A W G H E E
Sequence 2	P A W H E A E
Scoring parameters:	BLOSUM
Gap penalty:	Linear gap penalty of 8

Heuristic Methods:

- BLAST
- FASTA

Comparative Modelling Methods

-Assembly of rigid fragments

  -COMPOSER

  (Sutcliffe et al 1987 Protein Eng. 1 377)

  -Segment matching modelling (SMM)

  (Levitt, J.Mol. Biol. 226 507-533)

-Restrained based methods

  -MODELLER

  (Sali and Blundell, 1993)

MODELLER is a program for comparative modeling written by Prof. Šali's group at Rockefeller University.

- The program uses a scripting language.
- The user provides an alignment of a sequence to be modeled with known related structures.
- MODELLER automatically calculates a model with all non-hydrogen atoms.

The input are:

- Protein Data Bank (PDB) atom files of known protein structures;
- their alignment with the target sequence to be modeled.

The output is a model for the target that includes all non-hydrogen atoms.

- MODELLER can calculate sequence and structure alignments, however, it is better to prepare the alignment carefully by other means.

Format for Modeller:

```
INCLUDE
SET ATOM_FILES_DIRECTORY = './:../'
SET PDB_EXT = '.atm'
SET STARTING_MODEL = 1
SET ENDING_MODEL = 20
SET MD_LEVEL = 'refine1'
SET DEVIATION = 4.0
SET KNOWNS ='1JKE'
SET HETATM_IO = off
SET WATER_IO = off
SET ALIGNMENT_FORMAT = 'PIR'
SET SEQUENCE = 'target1'
SET ALNFILE = 'multiple1.ali
CALL ROUTINE = 'model'
```

Steps for homology Modelling

VDLEKIPIEEVFQQLKCSREGLTTQEGEDRIQIFGPNKLEEKKESKLLKFLGFMWNPLSW
VMEMAAIMAIALANGDGRPPDWQDFVGIICLLVINSTISFIEENNAGNAAAALMAGLAP
K
TKVLRDGKWSEQEAAILVPGDIVSIKLGDIIPADARLLEGDPLKVDQSALTGESLPVTKH
PGQEVFSGSTCKQGEIEAVVIATGVHTFFGKAAHLVDSTNQVGHFQKVLTAIGNFCICSI

Target sequence

Perform BLAST search

Select PDB in BLAST database

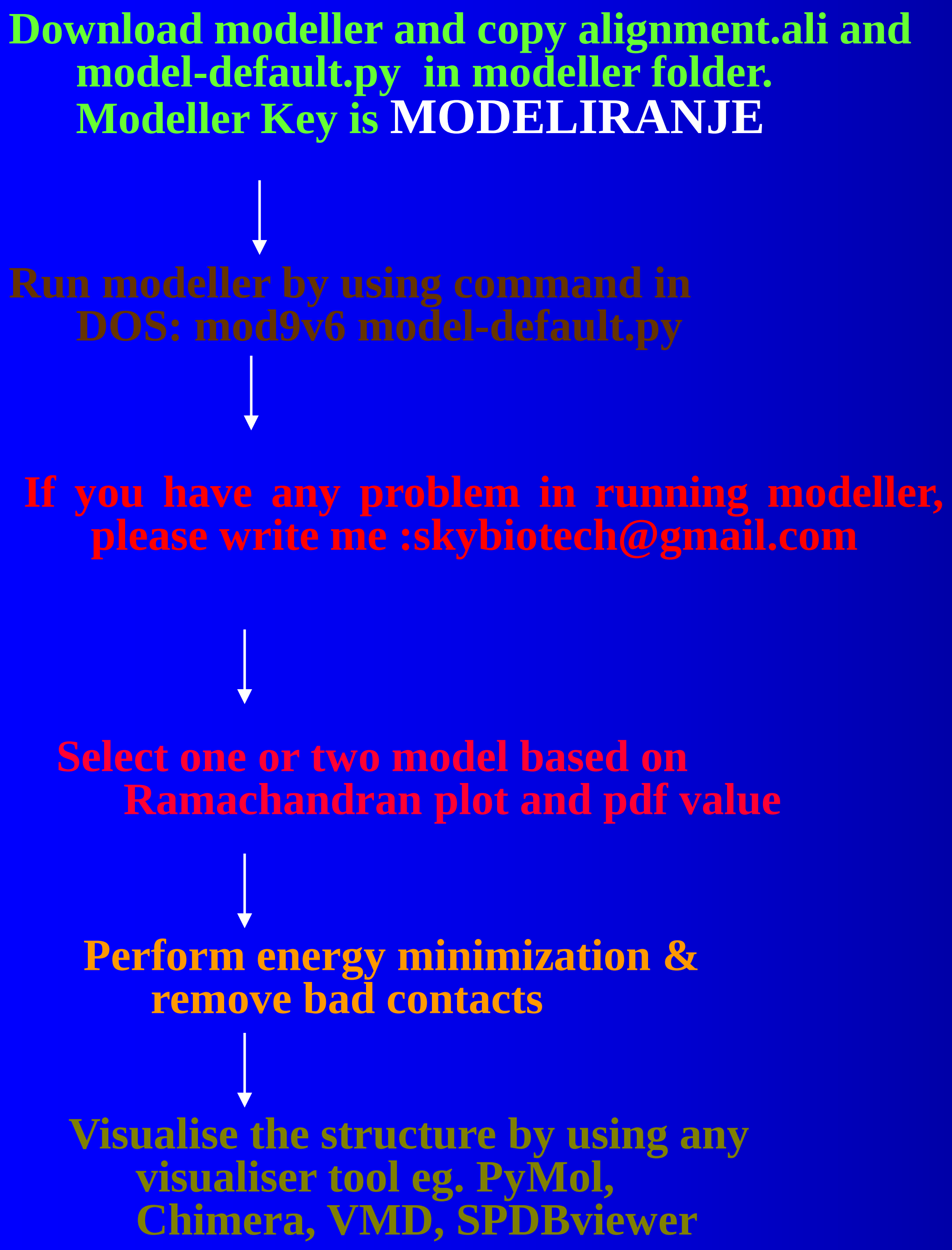
Select template from the BLAST hit

May select more than one template, if required

Do the Sequence alignment between target & template

Perform alignment in PIR format, modeller accept only PIR format

Use the alignment file for modeller



Download modeller and copy alignment.ali and model-default.py in modeller folder. Modeller Key is MODELIRANJE

Run modeller by using command in DOS: mod9v6 model-default.py

If you have any problem in running modeller, please write me :skybiotech@gmail.com

Select one or two model based on Ramachandran plot and pdf value

Perform energy minimization & remove bad contacts

Visualise the structure by using any visualiser tool eg. PyMol, Chimera, VMD, SPDBviewer

Saves server can be run for Ramachandran plot

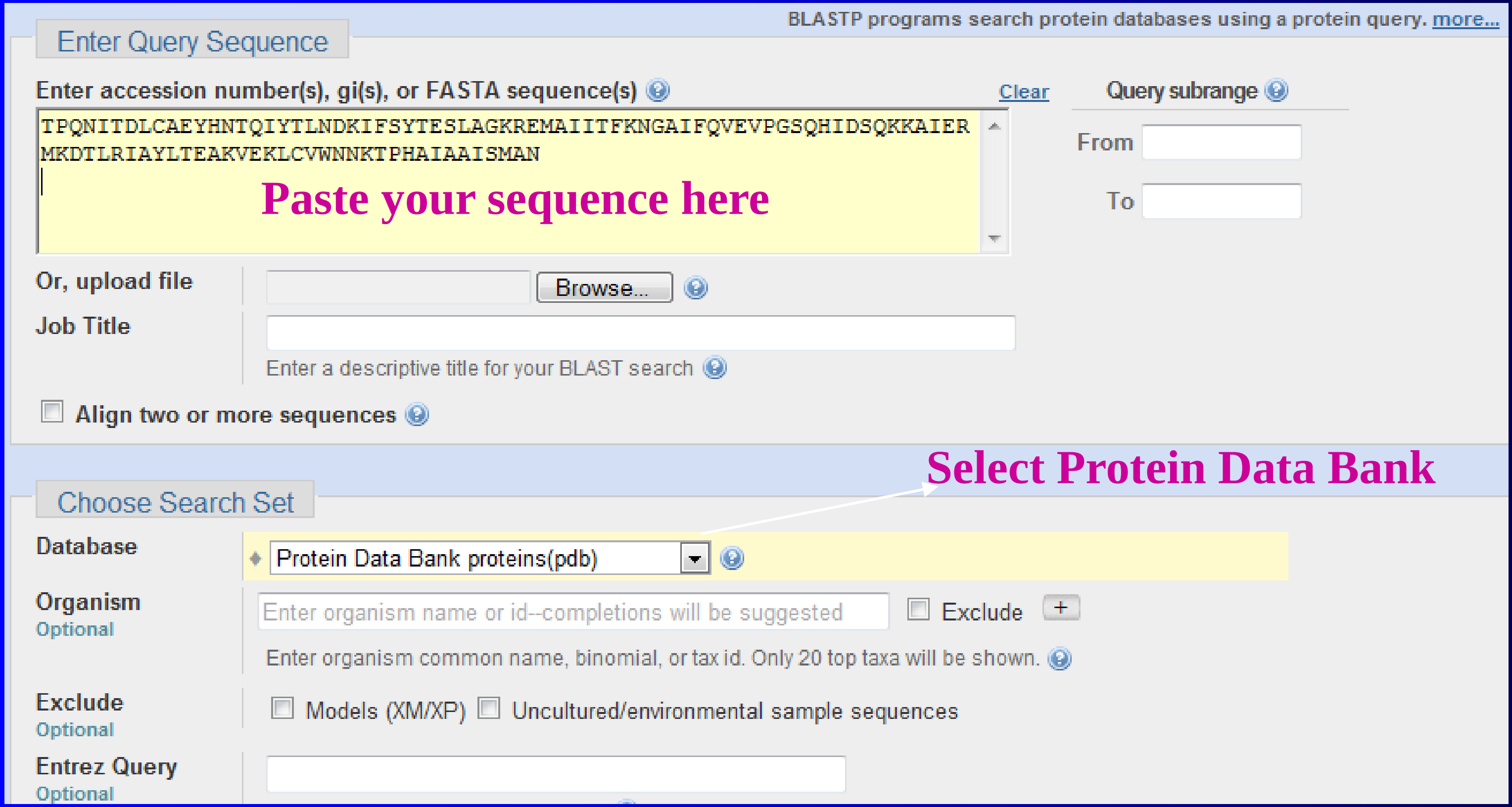
Example

TPQNITDLCAEYHNTQIYTLNDKIFSYTESLAGKREMAIITFKNGAIFQVEVP
GSQHIDSQKKAIERMKDTLRIAYLTEAKVEKLCVWNNKTPHAIAAISMAN

↓

Perform BLAST search

↓



BLASTP programs search protein databases using a protein query. more...

Enter Query Sequence

Enter accession number(s), gi(s), or FASTA sequence(s)

Clear

Query subrange

```
TPQNITDLCAEYHNTQIYTLNDKIFSYTESLAGKREMAIITFKNGAIFQVEVPGSQHIDSQKKAIER
MKDTLRIAYLTEAKVEKLCVWNNKTPHAIAAISMAN
```

Paste your sequence here

From

To

Or, upload file | Browse...

Job Title

Enter a descriptive title for your BLAST search

□ Align two or more sequences

Select Protein Data Bank

Choose Search Set

Database: Protein Data Bank proteins(pdb)

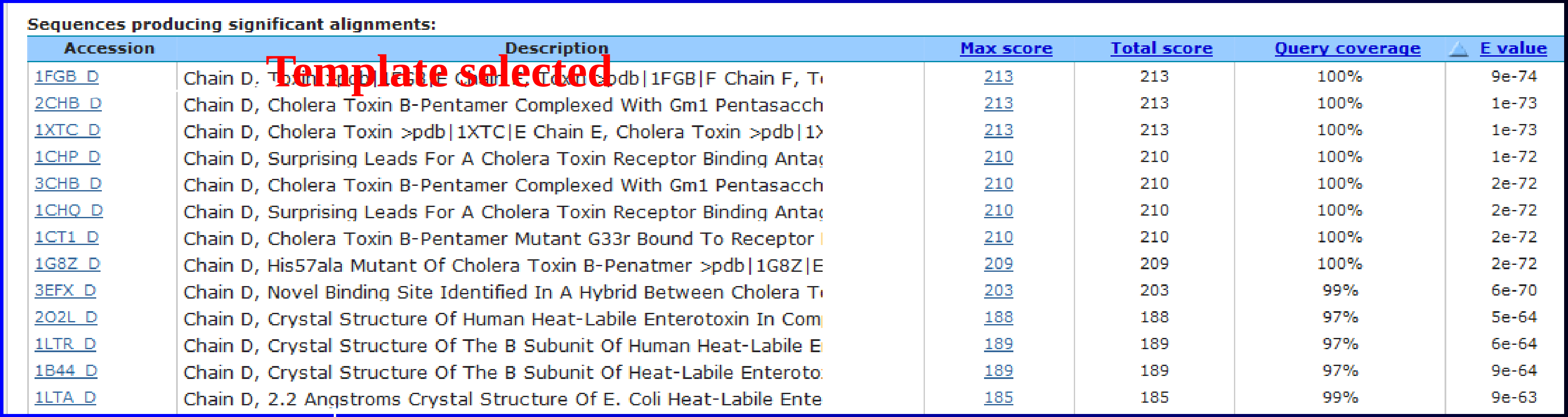
Organism Optional: Enter organism name or id--completions will be suggested □ Exclude +

Enter organism common name, binomial, or tax id. Only 20 top taxa will be shown.

Exclude Optional: □ Models (XM/XP) □ Uncultured/environmental sample sequences

Entrez Query Optional

BLAST Result



Template selected

Sequences producing significant alignments:

Accession	Description	Max score	Total score	Query coverage	E value
1FGB_D	Chain D, Toxin >pdb|1FGB|E Chain E, Toxin >pdb|1FGB|F Chain F, T	213	213	100%	9e-74
2CHB_D	Chain D, Cholera Toxin B-Pentamer Complexed With Gm1 Pentasacch	213	213	100%	1e-73
1XTC_D	Chain D, Cholera Toxin >pdb|1XTC|E Chain E, Cholera Toxin >pdb|1X	213	213	100%	1e-73
1CHP_D	Chain D, Surprising Leads For A Cholera Toxin Receptor Binding Anta	210	210	100%	1e-72
3CHB_D	Chain D, Cholera Toxin B-Pentamer Complexed With Gm1 Pentasacch	210	210	100%	2e-72
1CHQ_D	Chain D, Surprising Leads For A Cholera Toxin Receptor Binding Anta	210	210	100%	2e-72
1CT1_D	Chain D, Cholera Toxin B-Pentamer Mutant G33r Bound To Receptor	210	210	100%	2e-72
1G8Z_D	Chain D, His57ala Mutant Of Cholera Toxin B-Penatmer >pdb|1G8Z|E	209	209	100%	2e-72
3EFX_D	Chain D, Novel Binding Site Identified In A Hybrid Between Cholera T	203	203	99%	6e-70
2O2L_D	Chain D, Crystal Structure Of Human Heat-Labile Enterotoxin In Com	188	188	97%	5e-64
1LTR_D	Chain D, Crystal Structure Of The B Subunit Of Human Heat-Labile E	189	189	97%	6e-64
1B44_D	Chain D, Crystal Structure Of The B Subunit Of Heat-Labile Enteroto	189	189	97%	9e-64
1LTA_D	Chain D, 2.2 Angstroms Crystal Structure Of E. Coli Heat-Labile Ente	185	185	99%	9e-63

Do the Sequence alignment between target & template

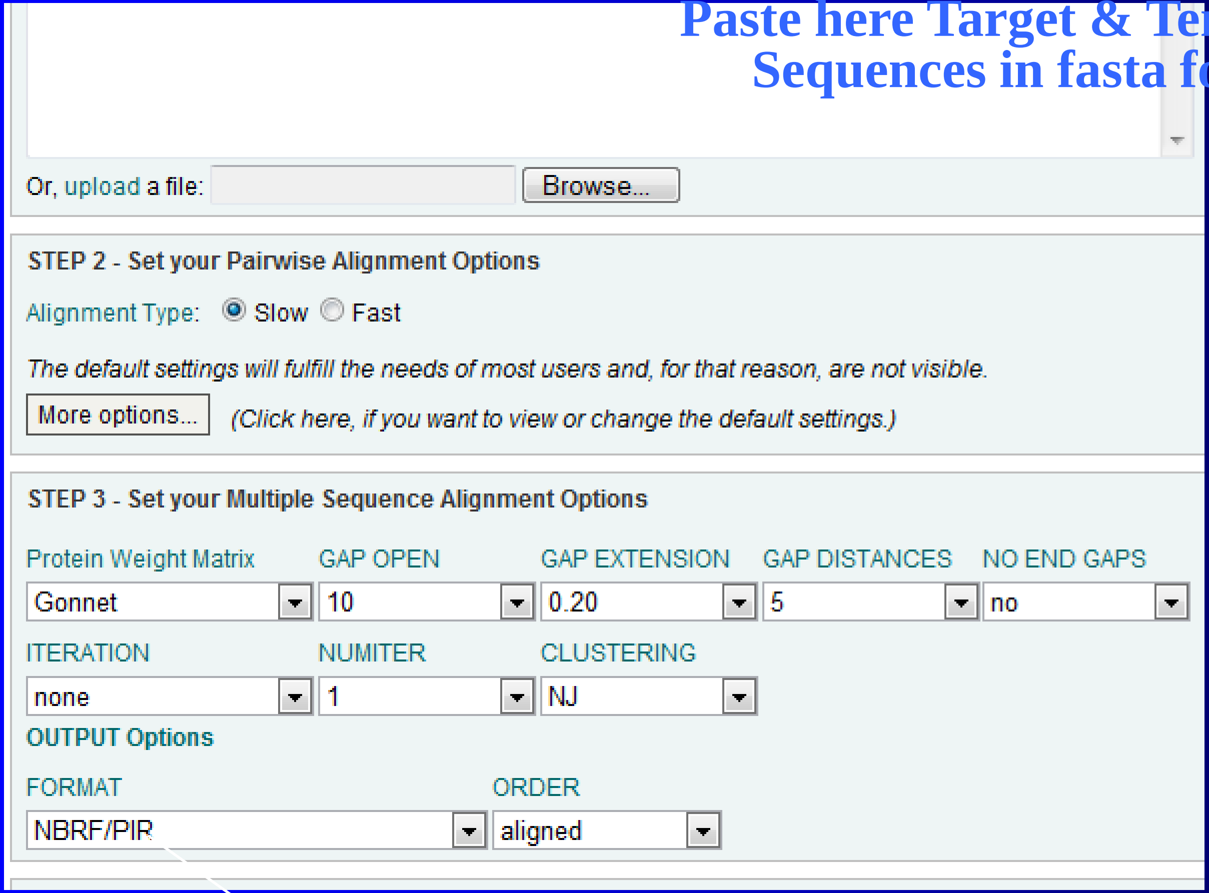
Download pdb file from PDB (www.rcsb.org/pdb)

Target sequence

```
>ctx
TPQNITDLCAEYHNTQIYTLNDKIFSYTESLAGKREMAIITFKNGAIFQVEVPGSQHID
SQKKAIERMKDTLRIAYLTEAKVEKLCVWNNKTPHAIAAISMAN
>2CHB
TPQNITDLCAEYHNTQIHTLNDKIFSYTESLAGKREMAIITFKNGATFQVEVPGSQHID
SQKKAIERMKDTLRIAYLTEAKVEKLCVWNNKTPHAIAAISMAN
```

Template sequence

Use ClustalW for Alignment



Paste here Target & Template Sequences in fasta format

Or, upload a file:

STEP 2 - Set your Pairwise Alignment Options

Alignment Type: ● Slow ○ Fast

The default settings will fulfill the needs of most users and, for that reason, are not visible.

(Click here, if you want to view or change the default settings.)

STEP 3 - Set your Multiple Sequence Alignment Options

Protein Weight Matrix	GAP OPEN	GAP EXTENSION	GAP DISTANCES	NO END GAPS
Gonnet	10	0.20	5	no

ITERATION	NUMITER	CLUSTERING
none	1	NJ

OUTPUT Options

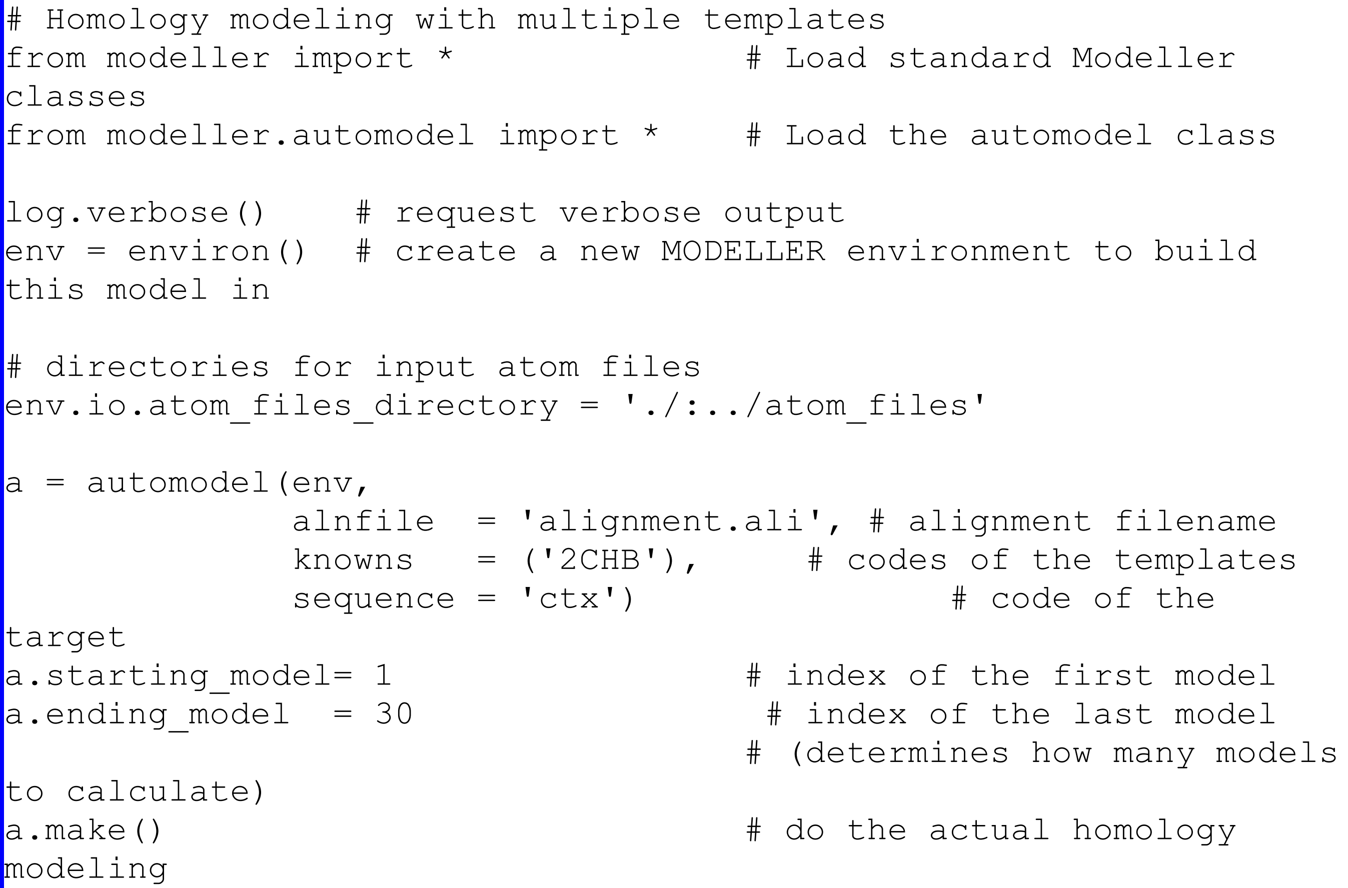
FORMAT	ORDER
NBRF/PIR	aligned

Select here PIR

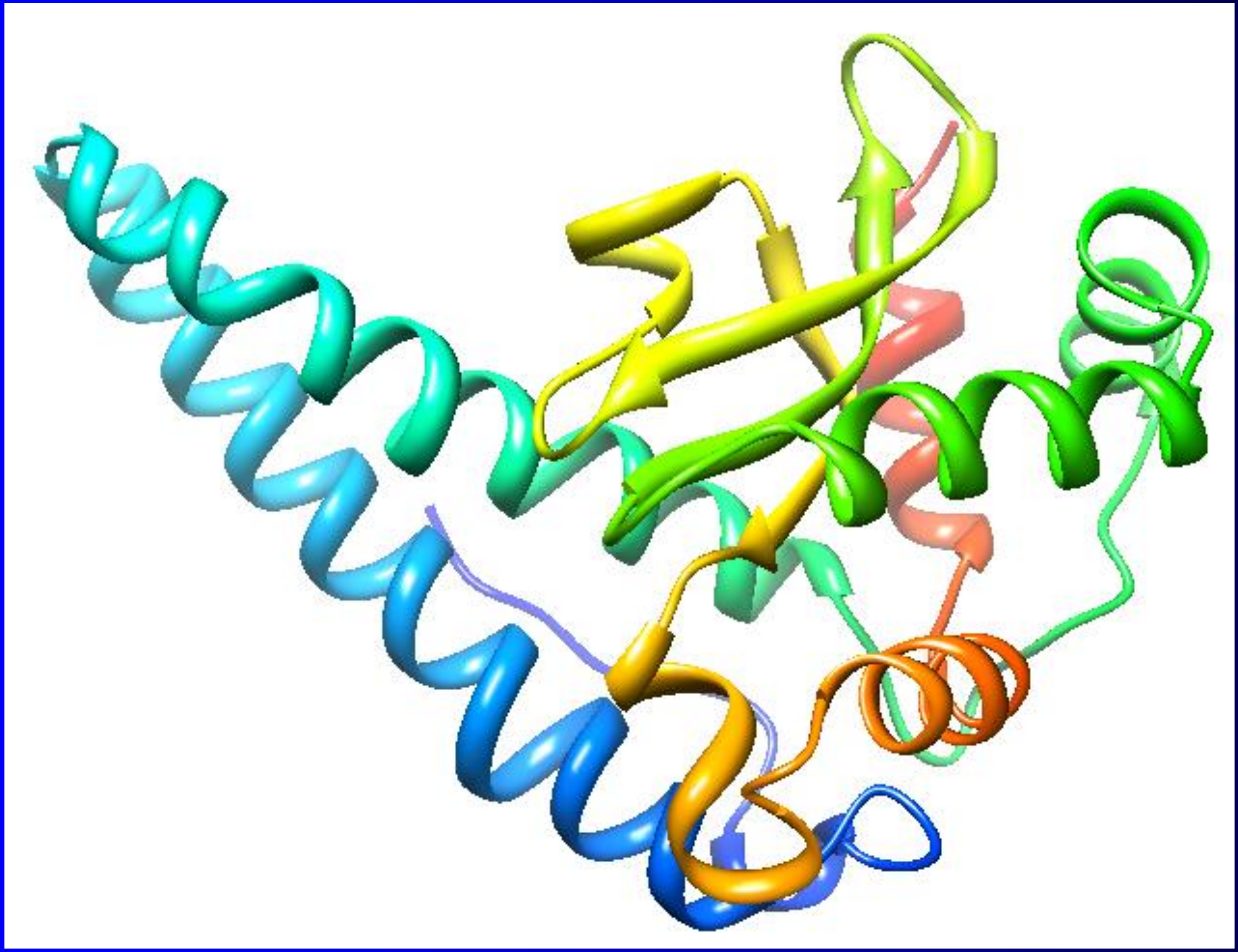
Alignment.ali

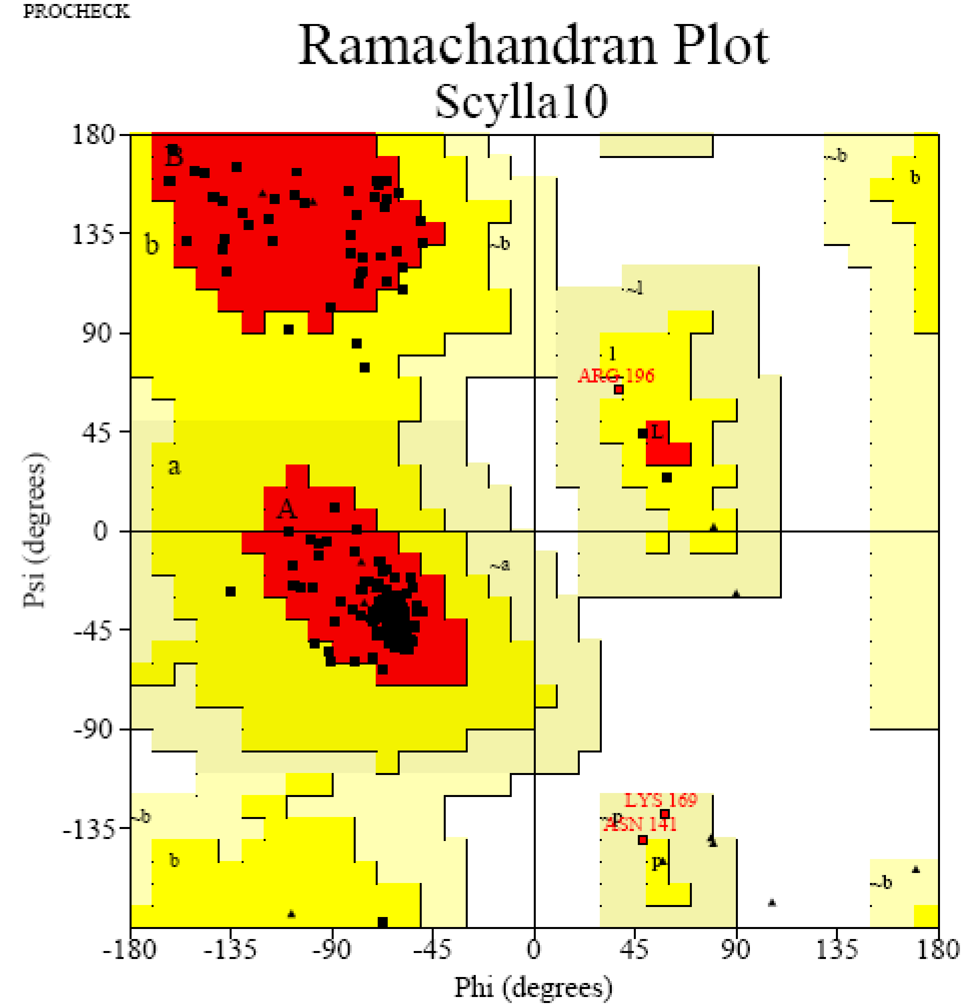
```
>P1;ctx
sequence:ctx::::::::
TPQNITDLCAEYHNTQIYTLNDKIFSYTESLAGKREMAIITFKNGAIFQVEVPGSQHID
SQKKAIERMKDTLRIAYLTEAKVEKLCVWNNKTPHAIAAISMAN
*
>P1;2CHB
structureX:2CHB:1:D:103:D::::
TPQNITDLCAEYHNTQIHTLNDKIFSYTESLAGKREMAIITFKNGATFQVEVPGSQHID
SQKKAIERMKDTLRIAYLTEAKVEKLCVWNNKTPHAIAAISMAN
*
```

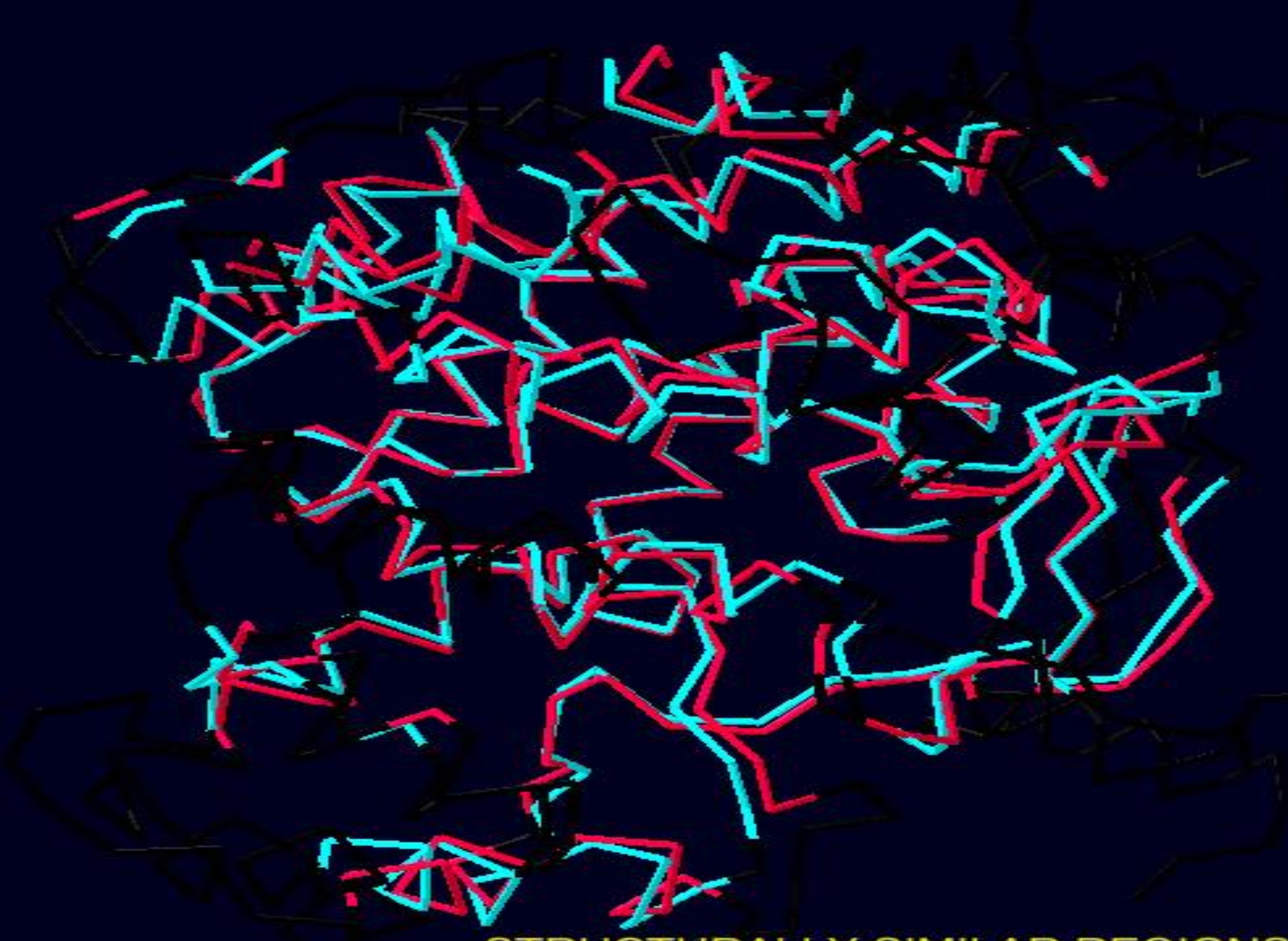
Model-default. py



```
# Homology modeling with multiple templates
from modeller import *              # Load standard Modeller
classes
from modeller.automodel import *    # Load the automodel class

log.verbose()    # request verbose output
env = environ()  # create a new MODELLER environment to build
this model in

# directories for input atom files
env.io.atom_files_directory = './:../atom_files'

a = automodel(env,
              alnfile  = 'alignment.ali', # alignment filename
              knowns   = ('2CHB'),     # codes of the templates
              sequence = 'ctx')                # code of the
target
a.starting_model= 1                 # index of the first model
a.ending_model  = 30                 # index of the last model
                                    # (determines how many models
to calculate)
a.make()                            # do the actual homology
modeling
```



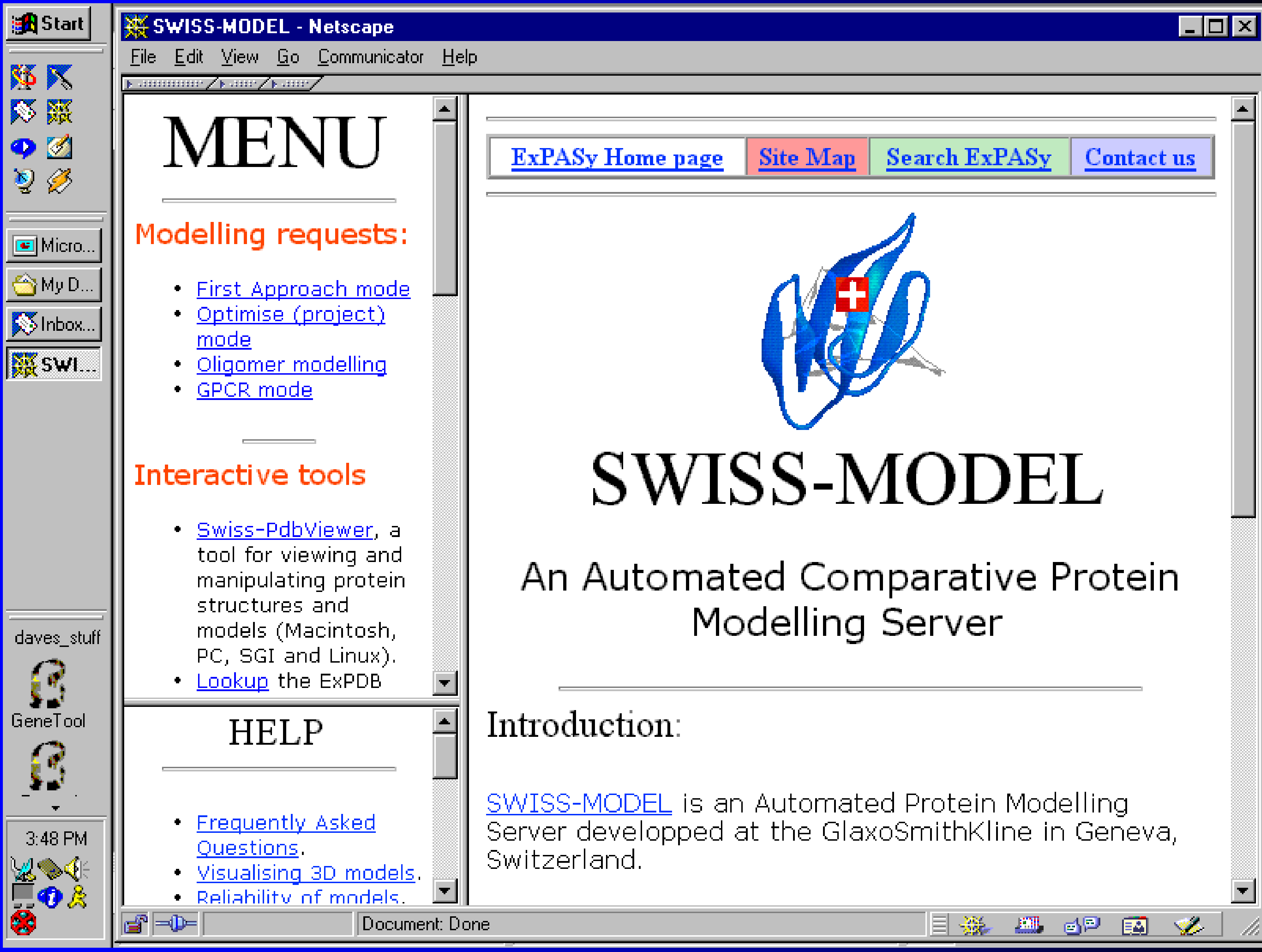

PROCHECK
Ramachandran Plot
Scylla10
180
135
90
45
0
-45
-90
-135
Psi (degrees)
-180
-135
-90
-45
0
45
90
135
180
Phi (degrees)
B
b
~b
b
~b
~l
l
ARG 196
L
a
A
~a
LYS 169
~p
ASN 141
p
~b
b
~b



STRUCTURALLY SIMILAR REGIONS

Modelling on the Web

- Prior to 1998 homology modelling could only be done with commercial software or command-line freeware
- The process was time-consuming and labor-intensive
- The past few years has seen an explosion in automated web-based homology modelling servers
- Now anyone can homology model!



MENU

Modelling requests:

- First Approach mode
- Optimise (project) mode
- Oligomer modelling
- GPCR mode

Interactive tools

- Swiss-PdbViewer, a tool for viewing and manipulating protein structures and models (Macintosh, PC, SGI and Linux).
- Lookup the ExPDB

HELP

- Frequently Asked Questions.
- Visualising 3D models.
- Reliability of models.

SWISS-MODEL

An Automated Comparative Protein Modelling Server

Introduction:

SWISS-MODEL is an Automated Protein Modelling Server developped at the GlaxoSmithKline in Geneva, Switzerland.

Application of Comparative Modeling

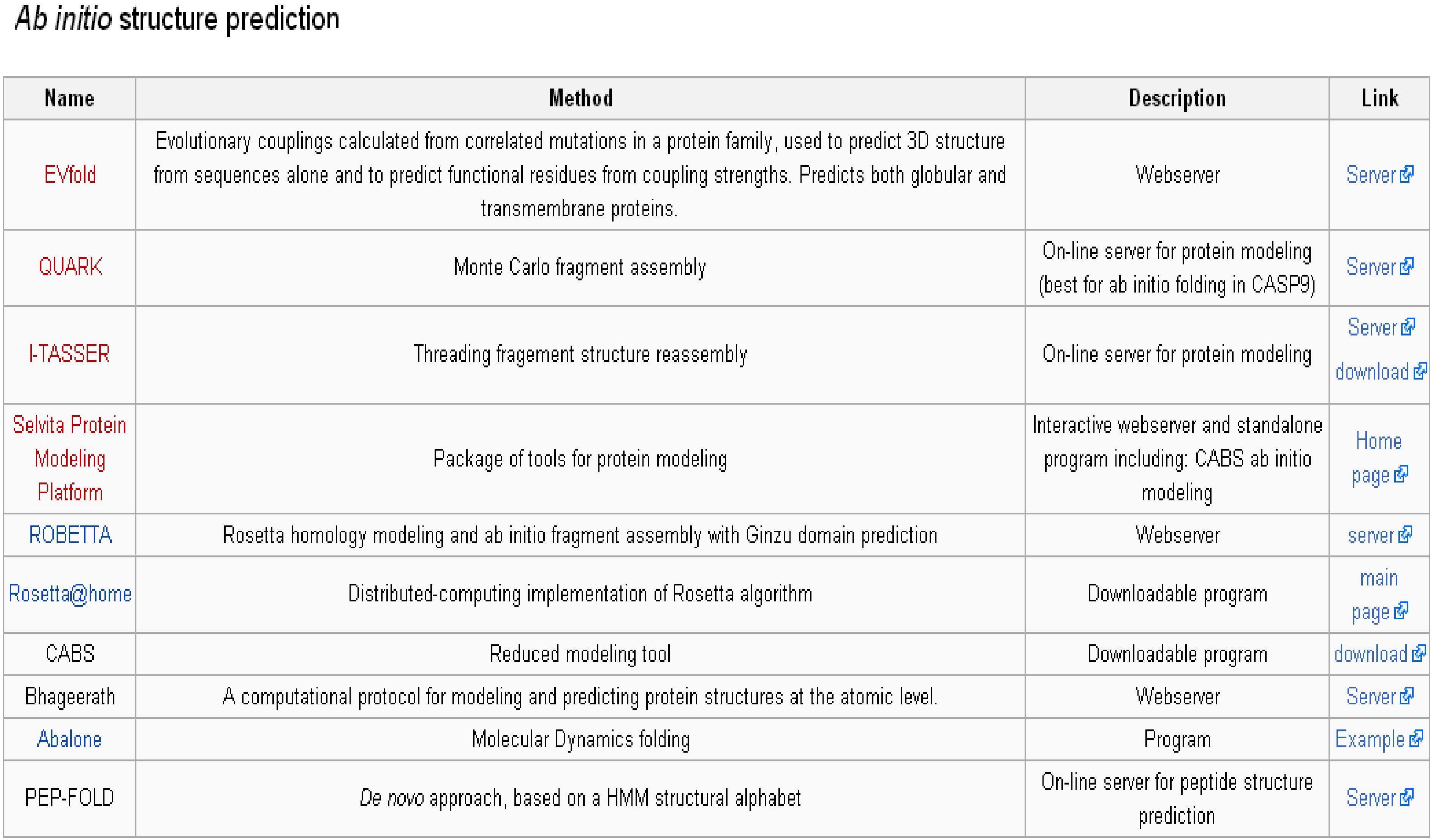
- Comparative modeling is an efficient way to obtain useful information about the proteins of interest. For example – comparative modeling can be helpful in
  - Designing mutants to text hypothesis about the proteins function.
  - Identifying active and binding sites.
  - Searching for designing and improving.
  - Modeling substrate specificity.
  - predicting antigenic epitopes.
  - Simulating protein – protein docking.
  - Confirming a remote structural relationship.

ab initio method of Modelling

Ab initio protein structure prediction is a method to determine the tertiary structure of protein in the absence of experimentally solved structure of a similar/homologous protein. This method builds protein structure guided by energy FUNCTION.

***ab Initio* modelling conducts a conformational search under the guidance of A designed energy function.**

This procedure usually generates a number of possible conformations (structure decoys) and final models are selected from them.



Ab initio structure prediction

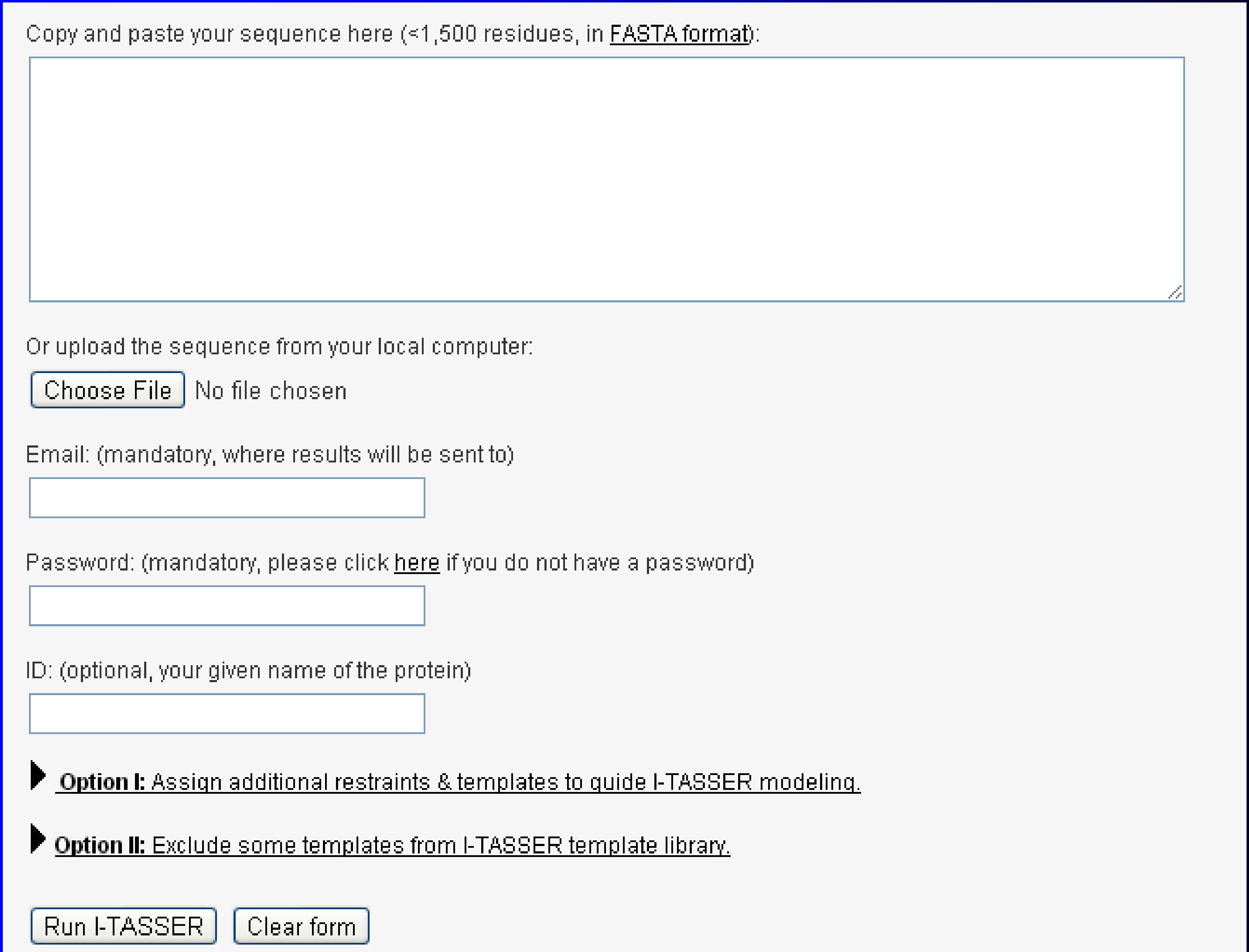
Name	Method	Description	Link
EVfold	Evolutionary couplings calculated from correlated mutations in a protein family, used to predict 3D structure from sequences alone and to predict functional residues from coupling strengths. Predicts both globular and transmembrane proteins.	Webserver	Server
QUARK	Monte Carlo fragment assembly	On-line server for protein modeling (best for ab initio folding in CASP9)	Server
I-TASSER	Threading fragement structure reassembly	On-line server for protein modeling	Server download
Selvita Protein Modeling Platform	Package of tools for protein modeling	Interactive webserver and standalone program including: CABS ab initio modeling	Home page
ROBETTA	Rosetta homology modeling and ab initio fragment assembly with Ginzu domain prediction	Webserver	server
Rosetta@home	Distributed-computing implementation of Rosetta algorithm	Downloadable program	main page
CABS	Reduced modeling tool	Downloadable program	download
Bhageerath	A computational protocol for modeling and predicting protein structures at the atomic level.	Webserver	Server
Abalone	Molecular Dynamics folding	Program	Example
PEP-FOLD	*De novo* approach, based on a HMM structural alphabet	On-line server for peptide structure prediction	Server

I-TASSER

- I-TASSER server is an on-line platform for protein structure and function predictions. 3D models are built based on multiple-threading alignments by LOMETS and iterative template fragment assembly simulations; function inslights are derived by matching the 3D models with BioLiP protein function database.

- I-TASSER (as 'Zhang-Server') was ranked as the No 1 server for protein structure prediction in recent CASP7, CASP8, CASP9, and CASP10 experiments.

- It was also ranked as the best for function prediction in CASP9. The server is in active development with the goal to provide the most accurate structural and function predictions using state-of-the-art algorithms. The server is only for non-commercial use



Copy and paste your sequence here (<1,500 residues, in FASTA format):

Or upload the sequence from your local computer:

Email: (mandatory, where results will be sent to)

Password: (mandatory, please click here if you do not have a password)

ID: (optional, your given name of the protein)

▶ **Option I:** Assign additional restraints & templates to guide I-TASSER modeling.

▶ **Option II:** Exclude some templates from I-TASSER template library.

Protein-Protein & Protein-Ligand Interactions

What is docking?

Prediction of the optimal physical configuration and energy between two molecules

The docking problem optimizes:

- Binding between two molecules such that their orientation maximizes the interaction
- Evaluates the total energy of interaction such that for the best binding configuration the binding energy is the minimum
- The resultant structural changes brought about by the interaction

Molecular Docking

- In the process of "docking" a ligand to a binding site mimics the natural course of interaction of the ligand and its receptor via a lowest energy pathway.

- Put a compound in the approximate area where binding occurs and evaluate the following:

  - ❖Do the molecules bind to each other?
  - ❖If yes, how strong is the binding?
  - ❖How does the molecule (or) the protein-ligand complex look like. (understand the intermolecular interactions)
  - ❖Quantify the extent of binding.

Few terms

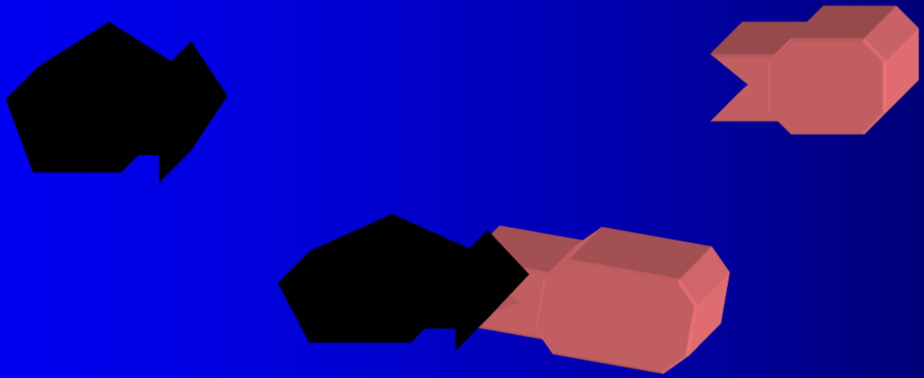
- **Receptor:** The receiving molecule, most commonly a protein or other biopolymer.
- **Ligand:** The complementary partener molecule which binds to the receptor. Ligands are most often small molecules but could also be another biopolymer.
- **Docking:** Computational simulation of a candidate ligand binding to a receptor.
- **Binding mode:** The orientation of the ligand relative to the receptor as well as the conformation of the ligand and receptor when bound to each other.
- **Pose:** A candidate binding mode.
- **Scoring:** The process of evaluating a particular pose by counting the number of favorable intermolecular interactions such as hydrogen bonds and hydrophobic contacts.
- **Ranking:** The process of classifying which ligands are most likely to interact favorably to a particular receptor based on the predicted free-energy of binding.

Classes of Docking

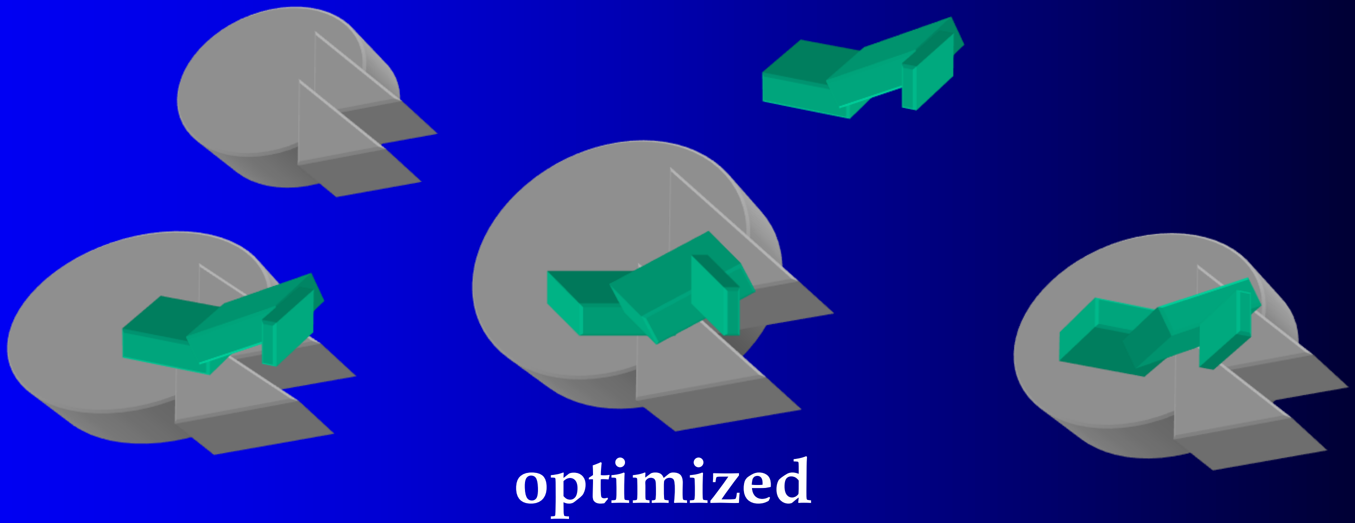
- **Protein-Protein docking**
- Both molecules usually considered rigid
- 6 degree of freedom, 3 for rotation, 3 for translation
- First apply only steric constraints to limit search space.
- Then examine energetics of possible binding confirmations
- The first approximation is to allow the substrate to do a random walk in the space around the protein to find the lowest energy.

- **Protein-ligand docking**
- Flexible ligand, rigid receptor
- Search space much larger

1. Protein-Protein Docking

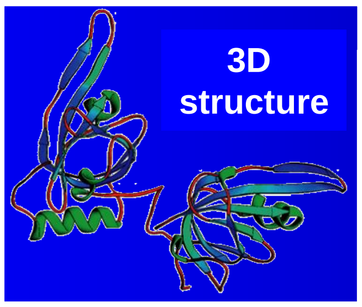


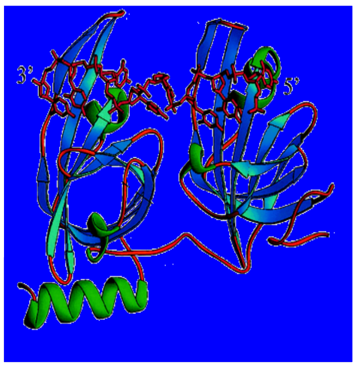
2. Protein-Ligand Docking



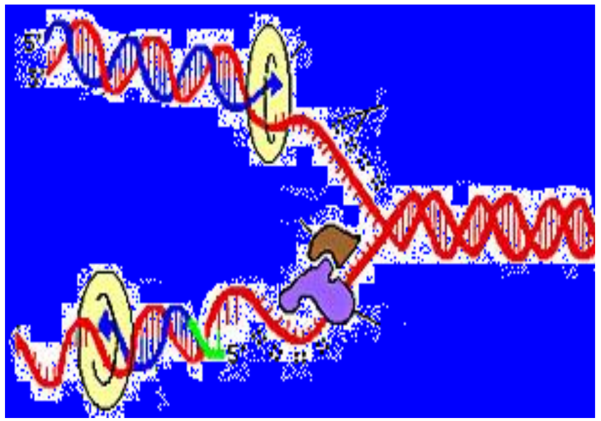
Biological Structure

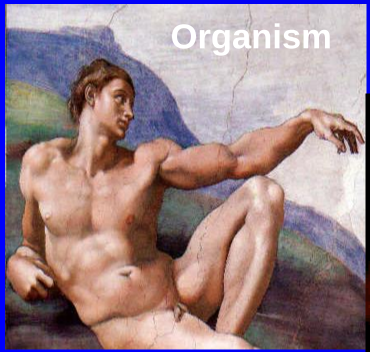
MESDAMESETMESSRSMYN
AMEISWALTERYALLKINCAL
LMEWALLYIPREFERDREVIL
MYSELFIMACENTERDIRATV
ANDYINTENNESSEEILIKENM
RANDDYNAMICSRPADNAPRI
MASERADCALCYCLINNDRKI
NASEMRPCALTRACTINKAR
KICIPCDPKIQDENVSDETAVS
WILLWINITALL

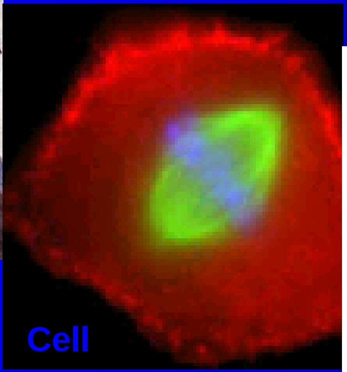


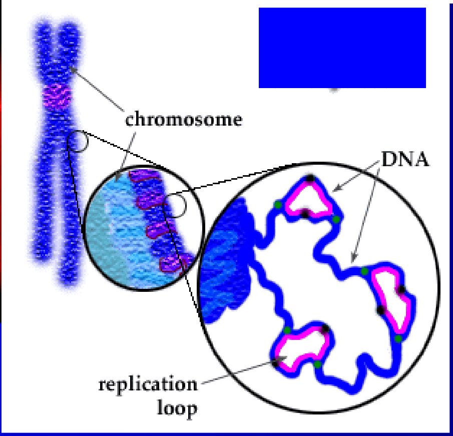


Structural Scales





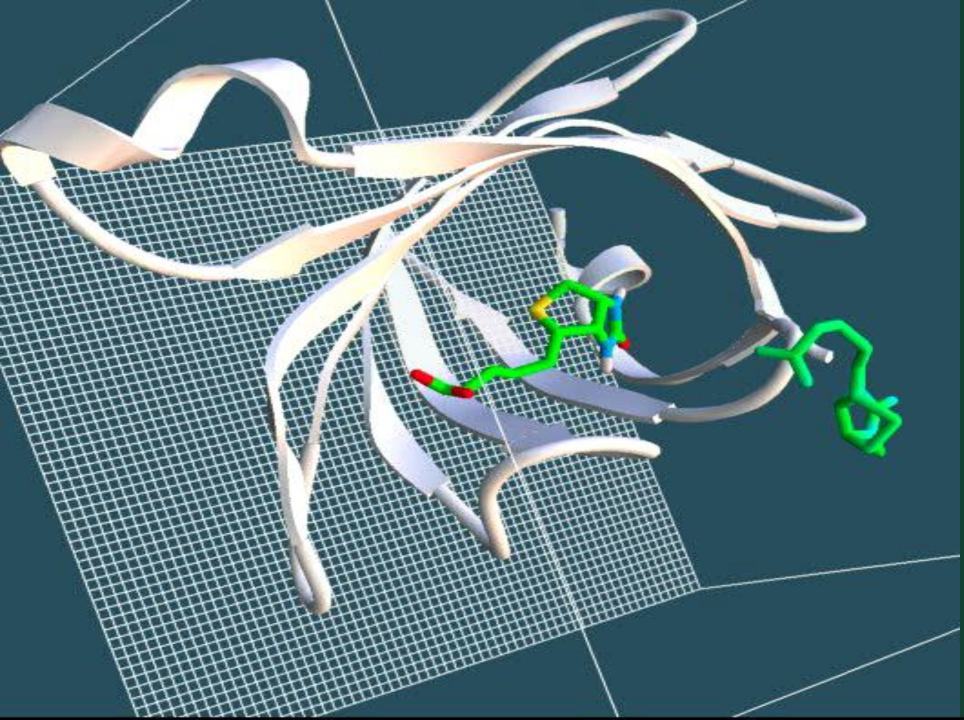


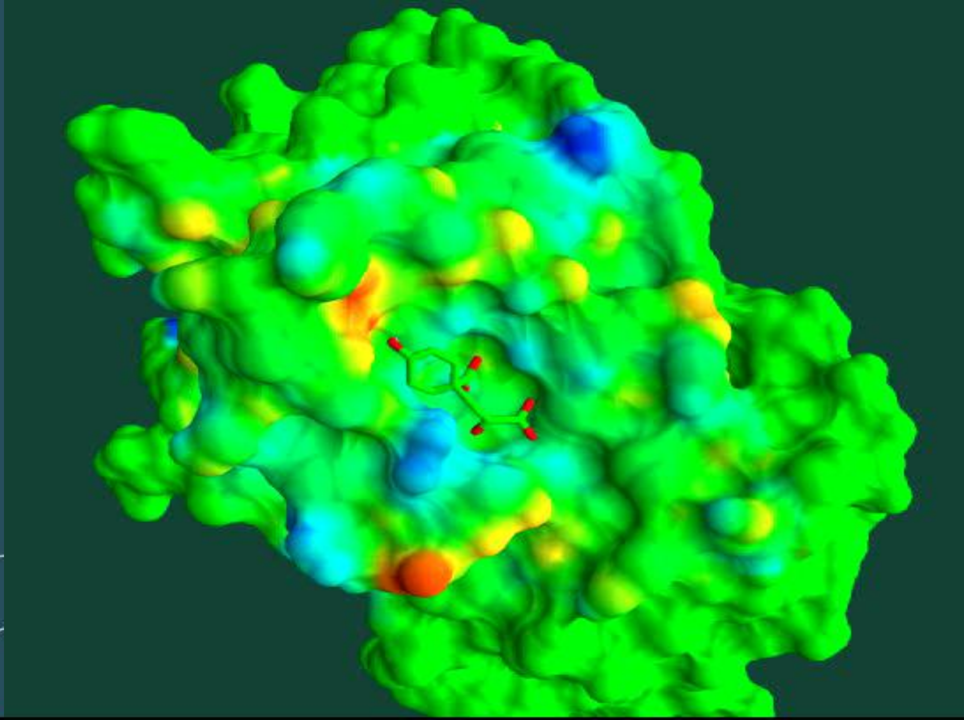


System Dynamics

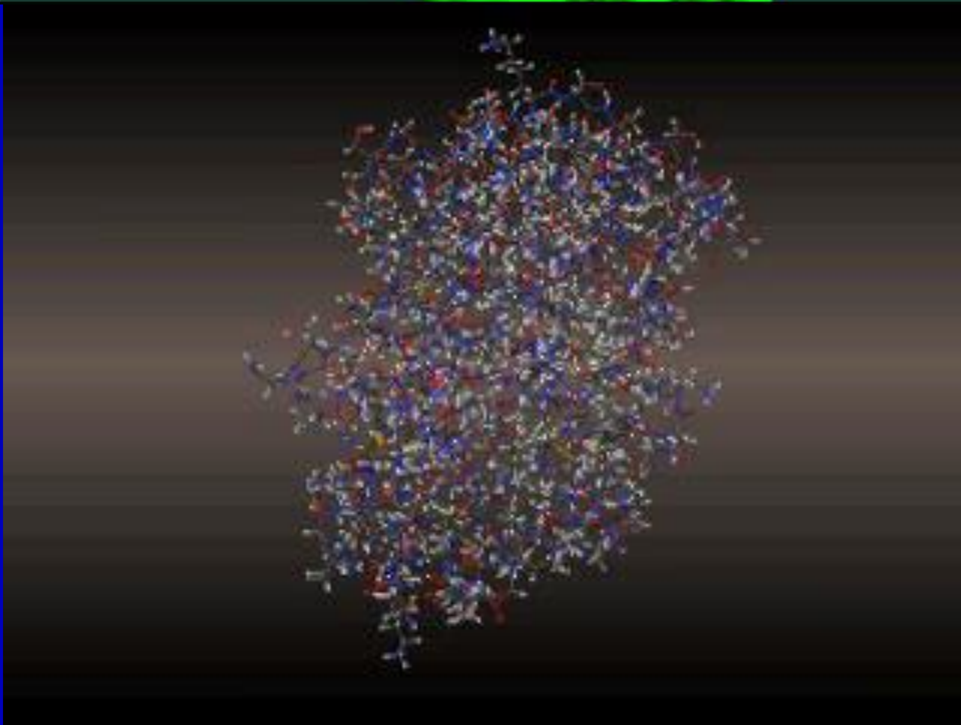
Some Available Programs to Perform Docking

- Affinity
- AutoDock
- BioMedCAChe
- CAChe for Medicinal Chemists
- DOCK
- DockVision
- FlexX
- Glide
- GOLD
- Hammerhead
- PRO_LEADS
- SLIDE
- VRDD

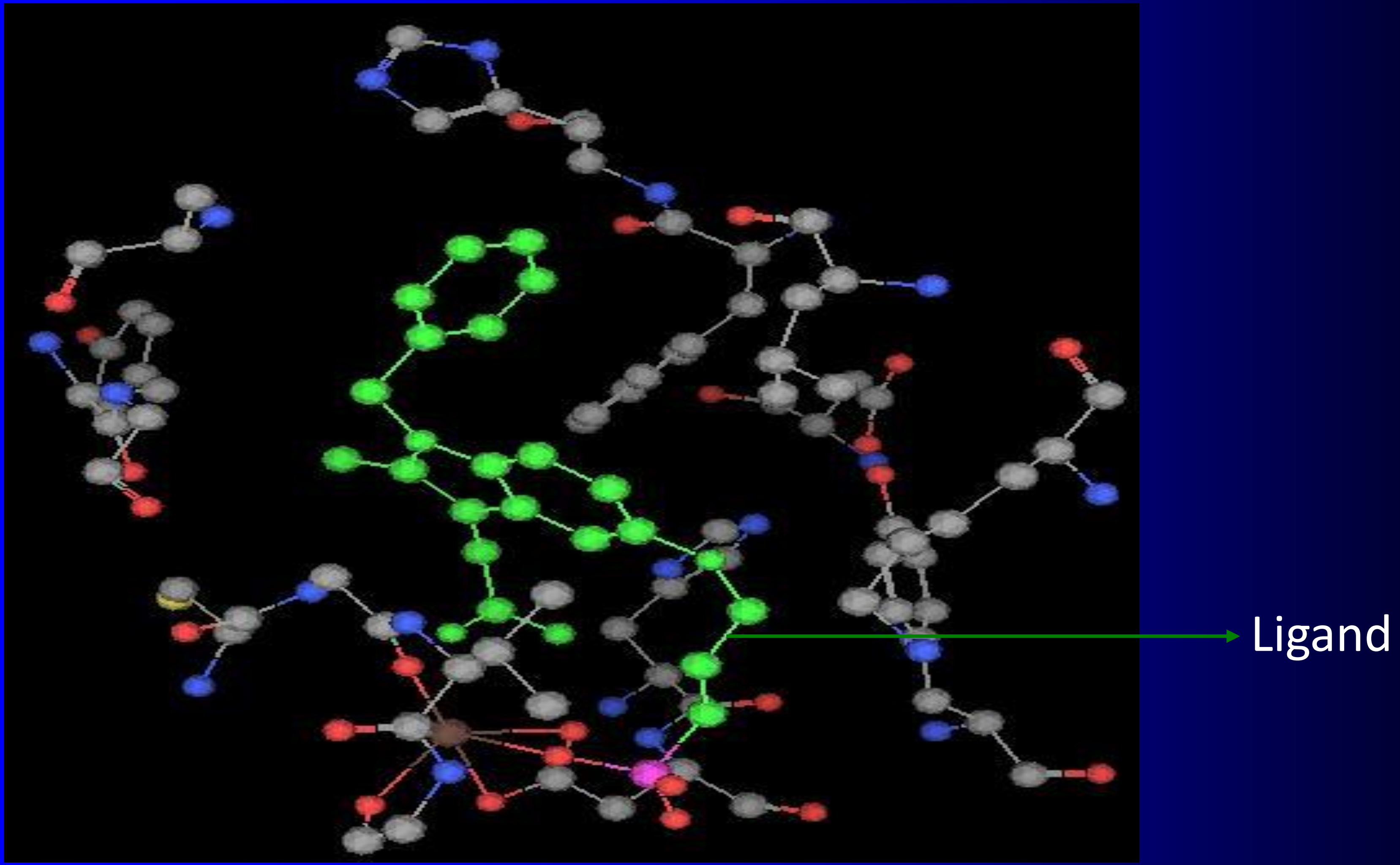




Different views of Docking



Ligand in Active Site Region

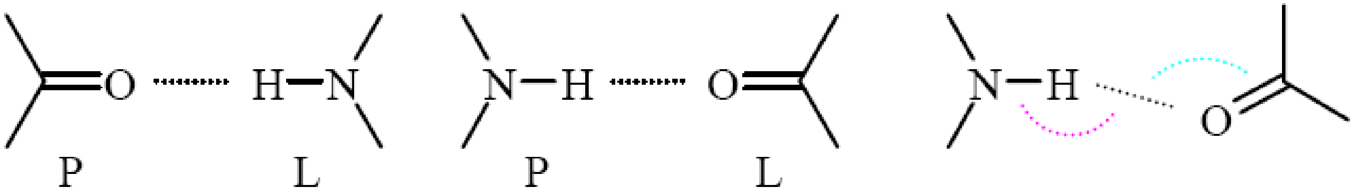


Active site residues

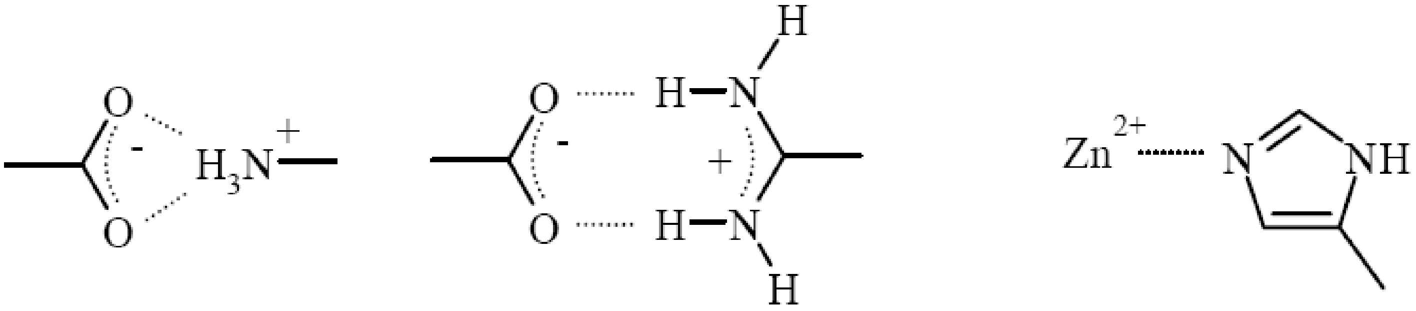
Histidine 6; Phenylalanine 5; Tyrosine 21; Aspartic acid 91; Aspartic acid 48; Tyrosine 51; Histidine 47; Glycine 29; Leucine 2; Glycine 31; Glycine 22; Alanine 18; Cysteine 28; Valine 20; Lysine 62

Types of Protein-Ligand interactions

- Hydrogen bonds
  - Electrostatic interaction with distance (N...O: 2.8-3.2 Å) and angle dependency (N-H...O: >150°, C=O...H: 100-180°)

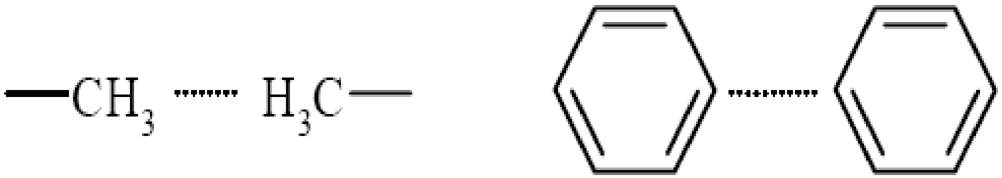


- Ionic interactions (salt bridges)
  - Strong coulomb interactions (2.7-3.0 Å)

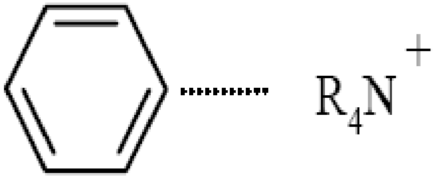


Types of Protein-Ligand interactions

- Hydrophobic interacions
  - Non-directional interactions of lipophilic regions of protein and ligand (everything not forming polar or hydrogen bond interactions) Aromatic interactions are directional!
  - Direct contribution to the binding affinity is small ⇒ acts via displacement of water molecules



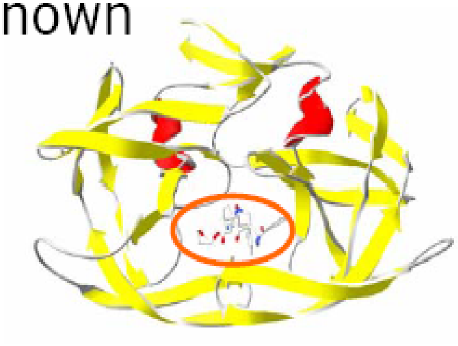
  - Cation-π interactions
    mainly polarization effect, often with quarternary nitrogens



Possible docking scenarios

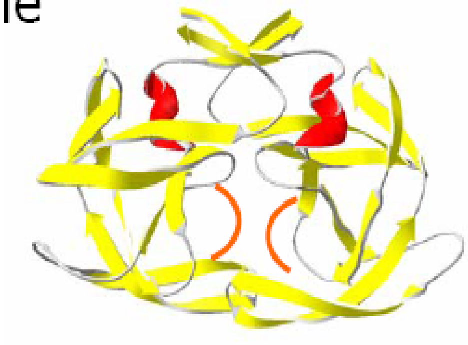
- Structure of a ligand in the binding pocket is known
  - How do other ligands interact with the protein?
  - How large is the binding affinity?

  ⇒ Docking programs (mainly flexible ligands only)



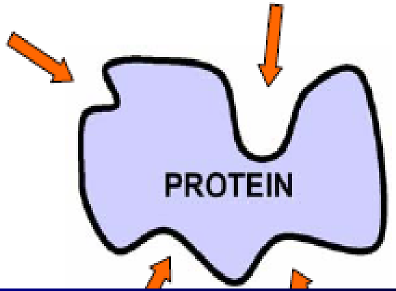
- Only structure of empty binding pocket available
  - How does a ligand change the pocket?

  ⇒ Protein modelling and/or flexible docking



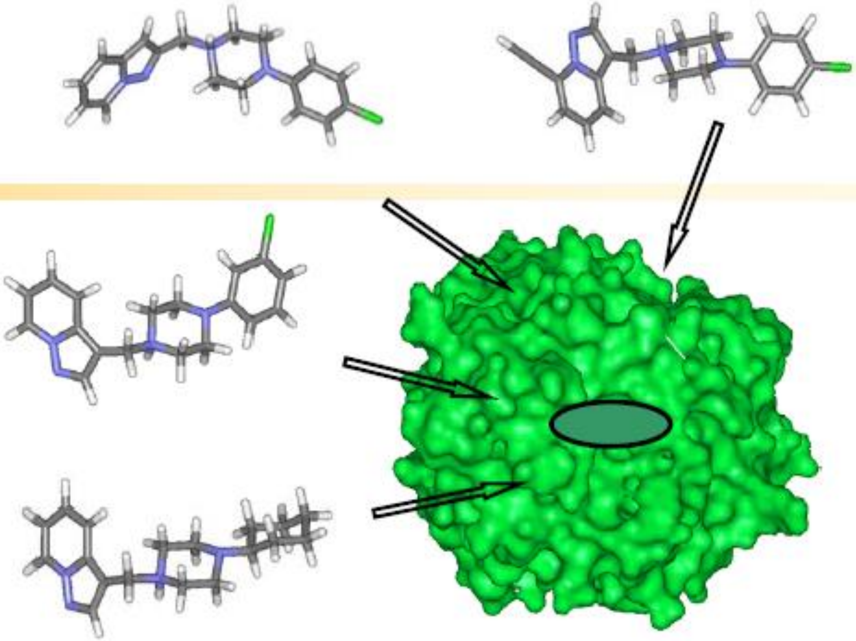
- Binding pocket is not known

  ⇒ try to find it by preceding search algorithms or extensive docking

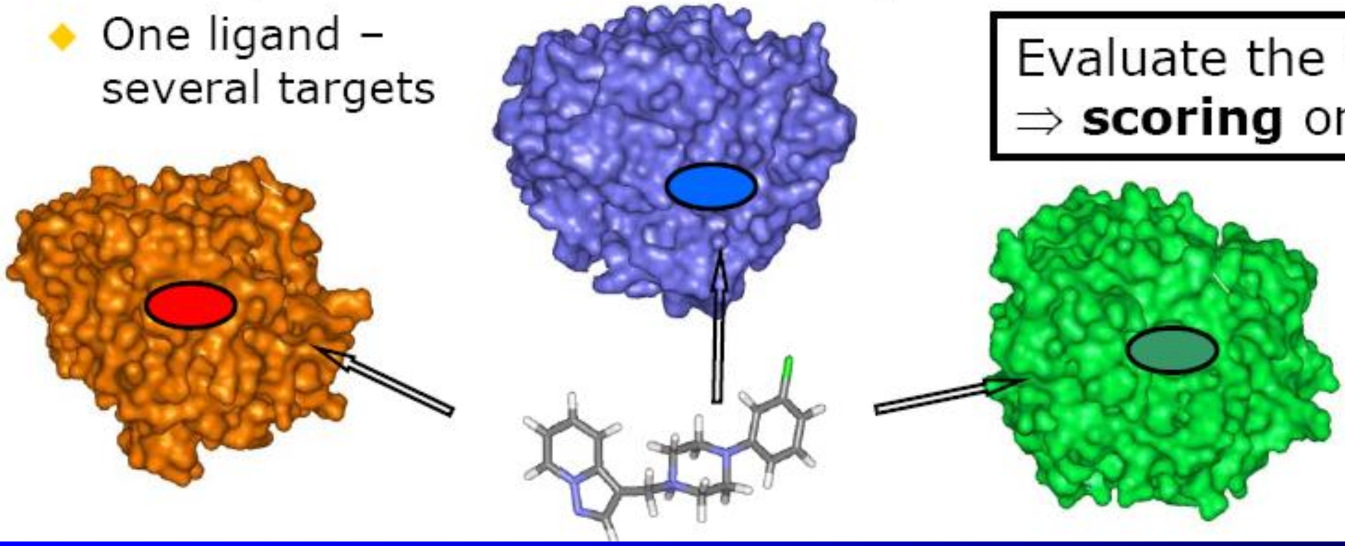


More scenarios

- Virtual screening
  - One target protein – many possible ligands



- Selectivity
  - One ligand – several targets



Evaluate the interaction?
⇒ **scoring** or **ranking**

Protein – Ligand Docking Programs

AutoDock http://www.scripps.edu/mb/olson/doc/autodock/

GOLD http://www.ccdc.cam.ac.uk/products/life_sciences/gold/

FLEXX

http://www.biosolveit.de/FlexX/

GLIDE

http://www.schrodinger.com/

ICM

http://www.molsoft.com/docking.html

Dock

http://www.cmpharm.ucsf.edu/kuntz/dock.html

Protein protein Docking Programs

ZDOCK : http://zlab.bu.edu/zdock/

HEX : http://www.csd.abdn.ac.uk/hex/

GRAMM : http://vakser.bioinformatics.ku.edu/resources/gramm

ICM : http://www.molsoft.com/docking.html

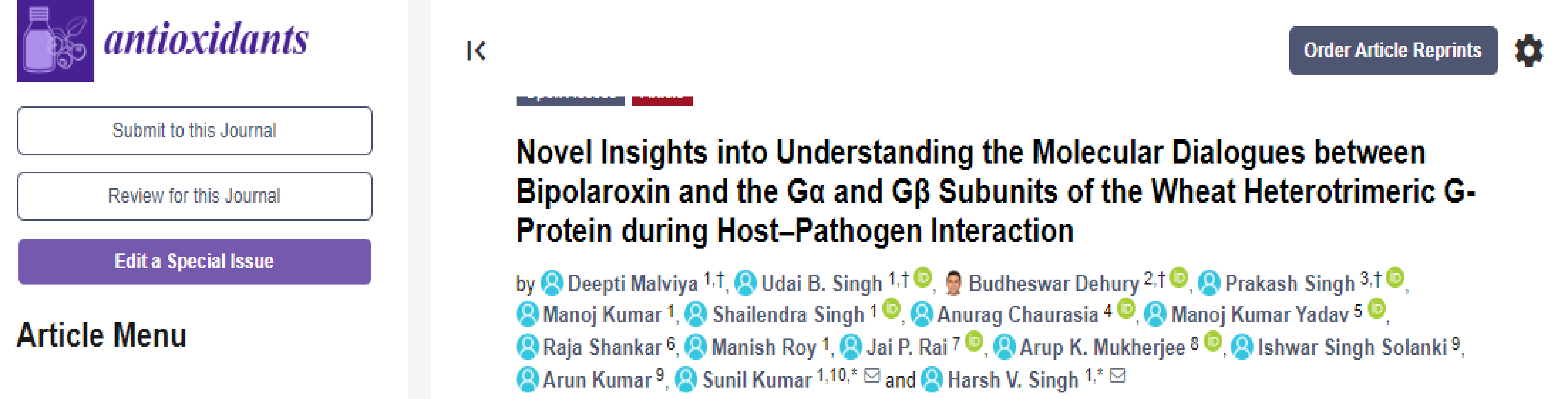
CLUSPRO : http://nrc.bu.edu/cluster/clusdoc.html

KORDO : http://www.bioinfo.de/isb/gcb99/poster/zimmermann/

MOLFIT : http://www.weizmann.ac.il/Chemical_Research_Support//molfit/

PATCHDOCK:

Case Studies



Novel Insights into Understanding the Molecular Dialogues between Bipolaroxin and the Gα and Gβ Subunits of the Wheat Heterotrimeric G-Protein during Host–Pathogen Interaction

by Deepti Malviya [1,†], Udai B. Singh [1,†], Budheswar Dehury [2,†], Prakash Singh [3,†], Manoj Kumar [1], Shailendra Singh [1], Anurag Chaurasia [4], Manoj Kumar Yadav [5], Raja Shankar [6], Manish Roy [1], Jai P. Rai [7], Arup K. Mukherjee [8], Ishwar Singh Solanki [9], Arun Kumar [9], Sunil Kumar [1,10,*] and Harsh V. Singh [1,*]

- Spot blotch disease of wheat, caused by the fungus *Bipolaris sorokiniana* (Sacc.) Shoem., produces several toxins which interact with the plants and thereby increase the blightening of the wheat leaves, and Bipolaroxin is one of them.
- There is an urgent need to decipher the molecular interaction between wheat and the toxin Bipolaroxin for in-depth understanding of host-pathogen interactions.
- we have developed the three-dimensional structure of G-protein alpha subunit from *Triticum aestivum*. Molecular docking studies were performed using the active site of the modelled G-protein alpha and cryo-EM structure of beta subunit from T. aestivum and 'Bipolaroxin'
- All-atoms molecular dynamics (MD) simulation studies were conducted for G-alpha and -beta subunit and Bipolaroxin complexes to explore the stability, conformational flexibility, and dynamic behavior of the complex system.
- In planta studies clearly indicated that application of Bipolaroxin significantly impacted the physio-biochemical pathways in wheat and led to the blightening of leaves in susceptible cultivars as compared to resistant ones. Further, it interacted with the Gα and Gβ subunits of G-protein, phenylpropanoid, and MAPK pathways, which is clearly supported by the qPCR results.
- This study gives deeper insights into understanding the molecular dialogues between Bipolaroxin and the Gα and Gβ subunits of the wheat heterotrimeric G-protein during host-pathogen interaction.

Antioxidant; 2022;

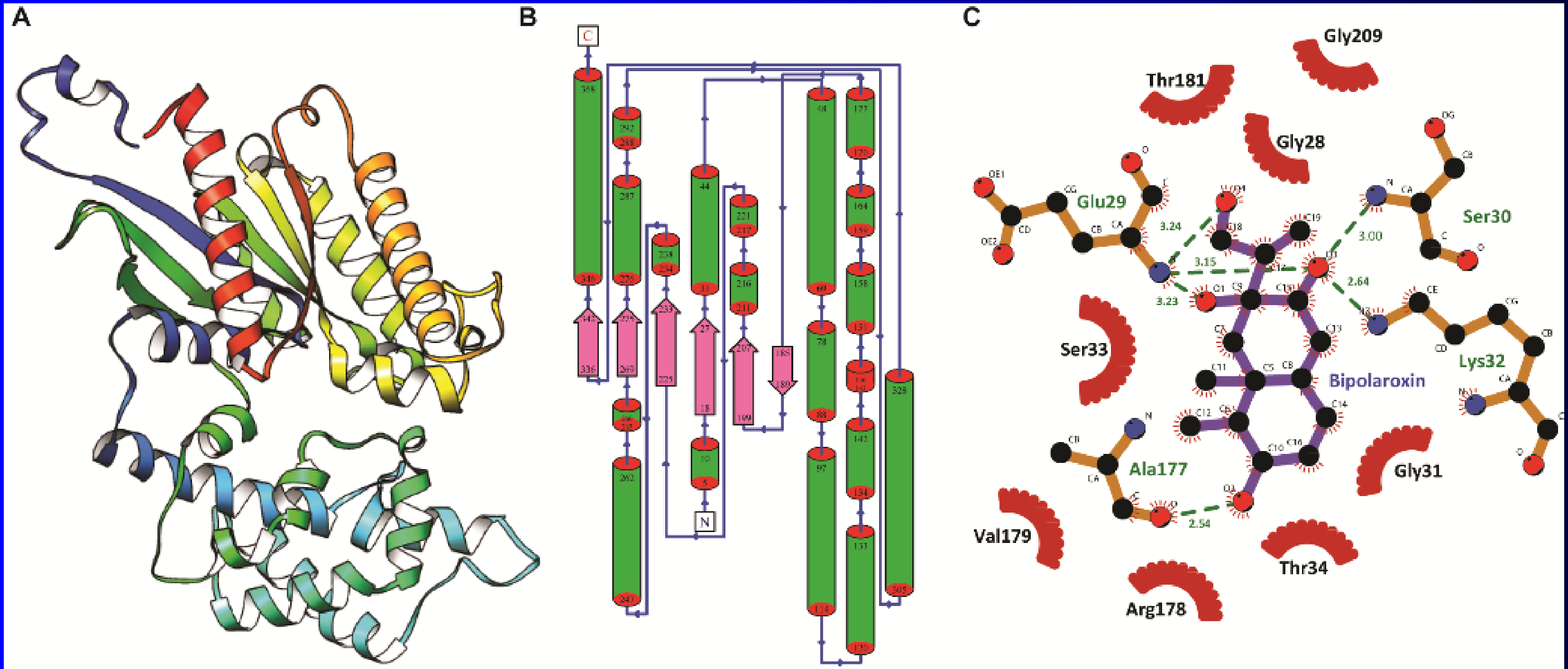


Figure 1. Three-dimensional model and 2D representation of the interaction between *Biopolaroxin* and G-Alpha subunit. (**A**) The solid ribbon representation of three-dimensional architecture of modeled G alpha subunit of wheat with domains (helical: lower half and Ras domain: upper half). (**B**) The topology of architecture of the modeled structure where the position of each helix and strand has been labeled from N-terminal end to c-terminal end. (**C**) Molecular representation of the top-ranked docked complex obtained from molecular docking of G-alpha subunit with Bipolaroxin rendered using LigPlot$^+$ tool. The green dotted lines show the hydrogen bonds, residues with dark-red semicircles forming hydrophobic contacts, and residues labeled in green portray the H-bond forming amino acids.

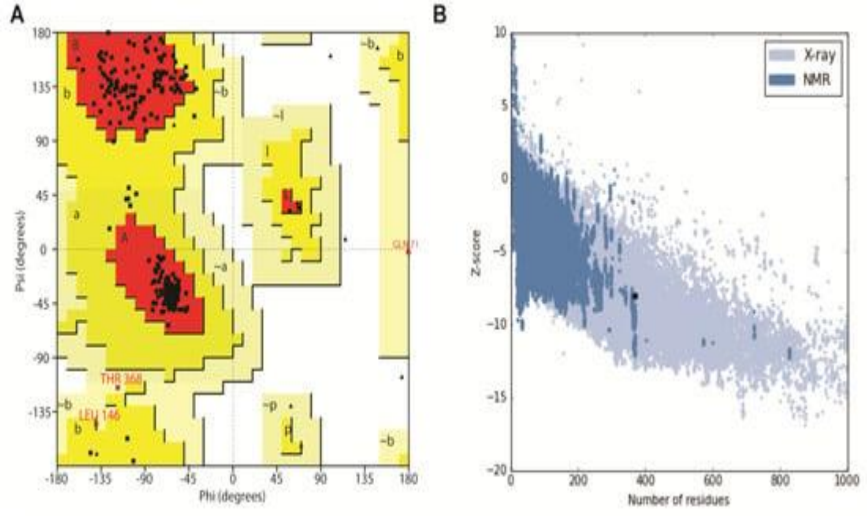
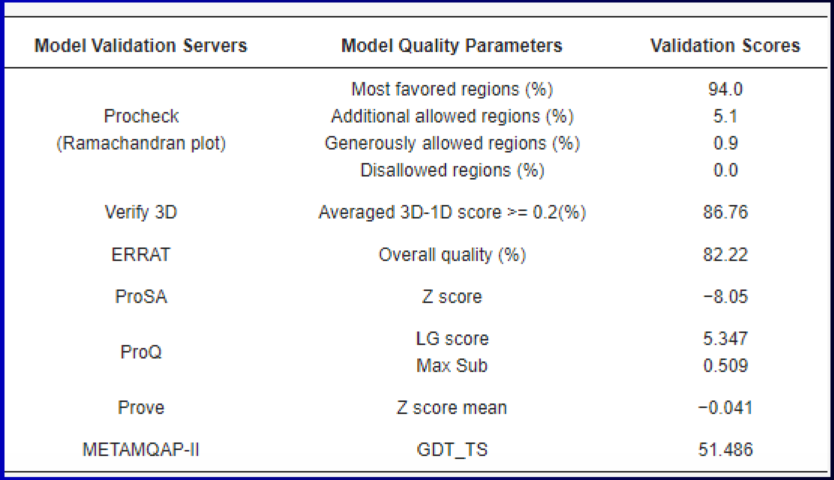


Figure 2. Ramachandran plot and ProSA z-score analysis of the modeled G-alpha subunit of *Triticum aestivum*. (**A**) The Ramachandran plot was generated using Procheck tool embedded in SAVES and the z-score plot was plotted using ProSA-Web (**B**).



Model Validation Servers	Model Quality Parameters	Validation Scores
Procheck (Ramachandran plot)	Most favored regions (%) Additional allowed regions (%) Generously allowed regions (%) Disallowed regions (%)	94.0 5.1 0.9 0.0
Verify 3D	Averaged 3D-1D score >= 0.2(%)	86.76
ERRAT	Overall quality (%)	82.22
ProSA	Z score	−8.05
ProQ	LG score Max Sub	5.347 0.509
Prove	Z score mean	−0.041
METAMQAP-II	GDT_TS	51.486

Table 1. Model validation statistics of G-protein alpha subunit of *Triticum aestivum* using various structural evaluation servers.

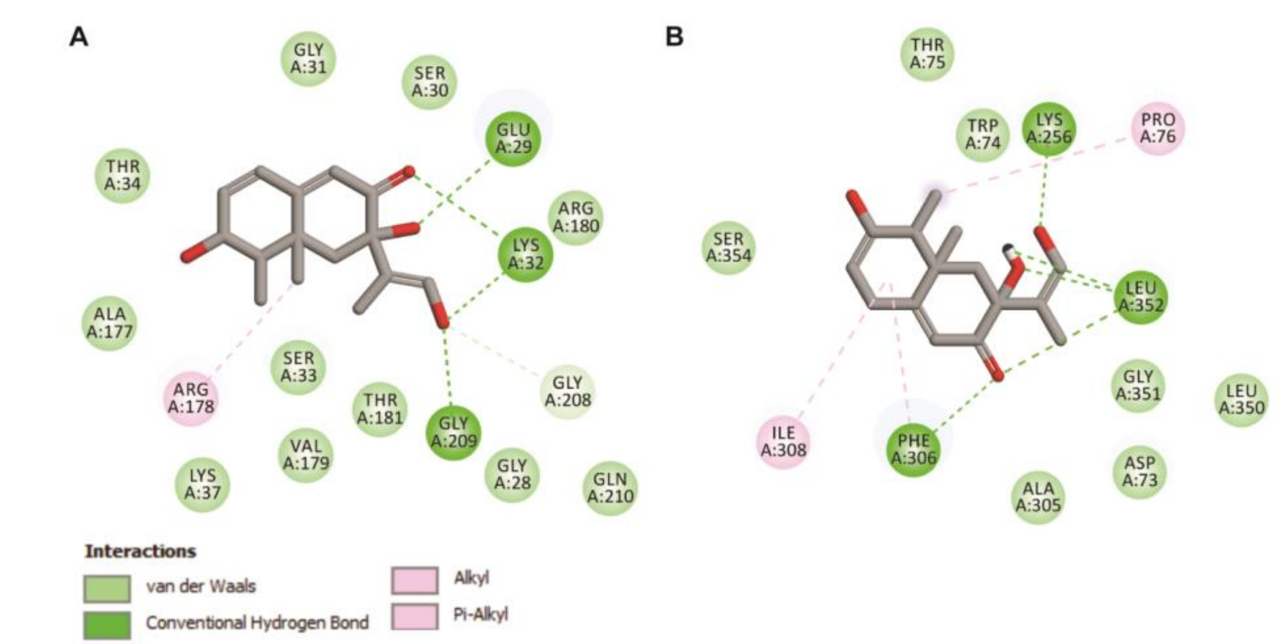
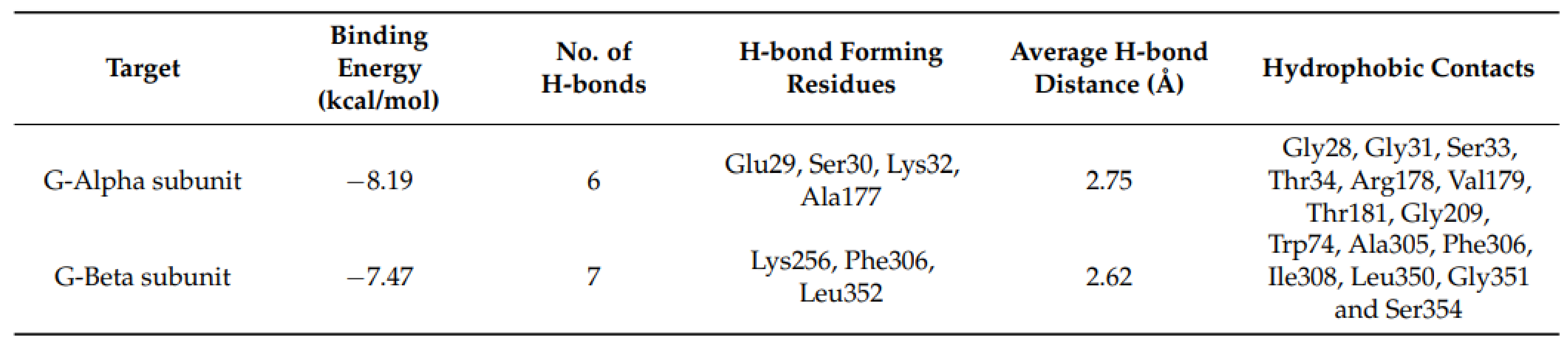


Figure 3. Intermolecular contact analyses of the top-ranked poses of Bipolaroxin with G-protein alpha (A) subunit and G-beta subunit of Triticum aestivum (B). The image was generated using BIOVIA DSV.



Target	Binding Energy (kcal/mol)	No. of H-bonds	H-bond Forming Residues	Average H-bond Distance (Å)	Hydrophobic Contacts
G-Alpha subunit	−8.19	6	Glu29, Ser30, Lys32, Ala177	2.75	Gly28, Gly31, Ser33, Thr34, Arg178, Val179, Thr181, Gly209,
G-Beta subunit	−7.47	7	Lys256, Phe306, Leu352	2.62	Trp74, Ala305, Phe306, Ile308, Leu350, Gly351 and Ser354

Table 2. Molecular docking results of Bipolaroxin with G-protein alpha and beta subunit of Triticum aestivum using AutoDock.

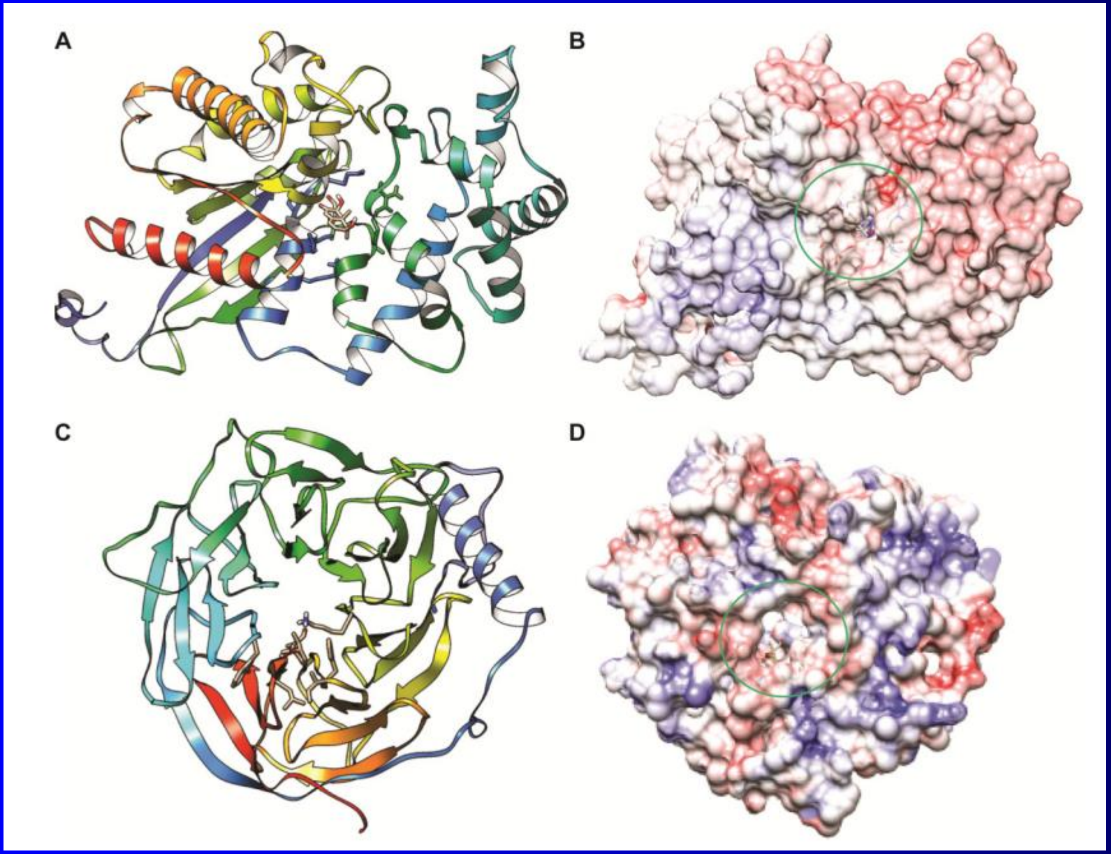


Figure 4. Docked conformational states and electrostatic surface representation of Bipolaroxin with G-protein alpha and G beta subunit. (A) Solid ribbon representation of the G-protein alpha subunit with Bipolaroxin (stick format) with the binding pocket residues. (B) Electrostatic surface potential map of G-protein alpha subunit with Bipolaroxin (ligand binding pocket has been marked in circle). (C) Solid ribbon representation of the G protein beta subunit with Bipolaroxin (stick format). (D) Electrostatic surface potential map of G protein beta subunit with Bipolaroxin (ligand binding pocket has been marked in circle). The electrostatic surface potential maps were generated using APBS and rendered using Chimera.

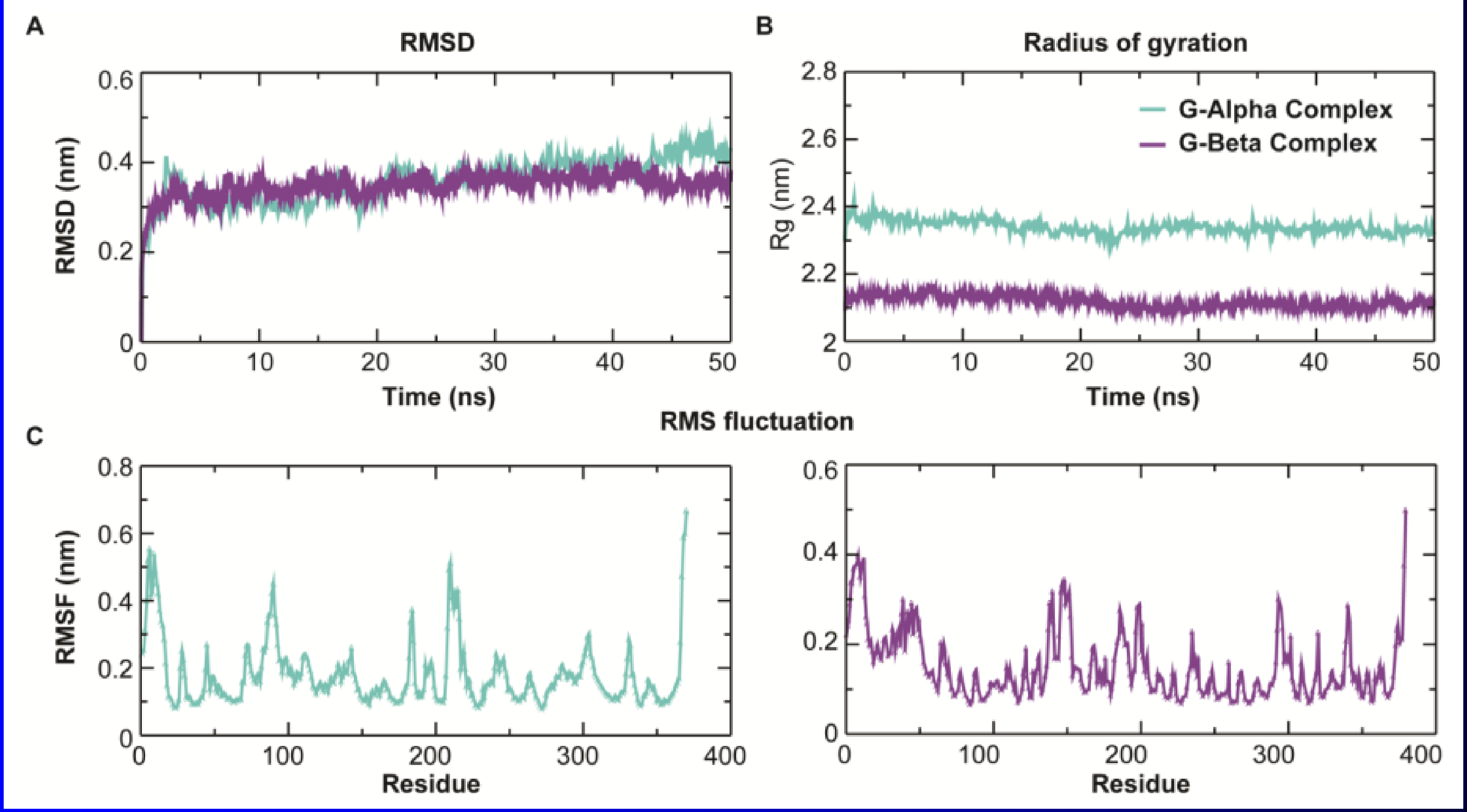


Figure 5. Intrinsic dynamics stability parameters of the G-alpha and G-beta subunit Bipolaroxin complexes during 50 ns MD simulation. (**A**) Root mean square deviation (RMSD) of backbone atoms of the modeled G-alpha subunit and experimental beta subunit in complexes with Bipolaroxin during 50 ns MD. (**B**) The compactness of the trajectory by calculating the radius of gyration (R_g) of the proteins during 50 ns MD in aqueous solution. (**C**) The root mean square fluctuation (RMSF) for Cα atoms of the G-alpha (left) and G-beta (right) complex systems.

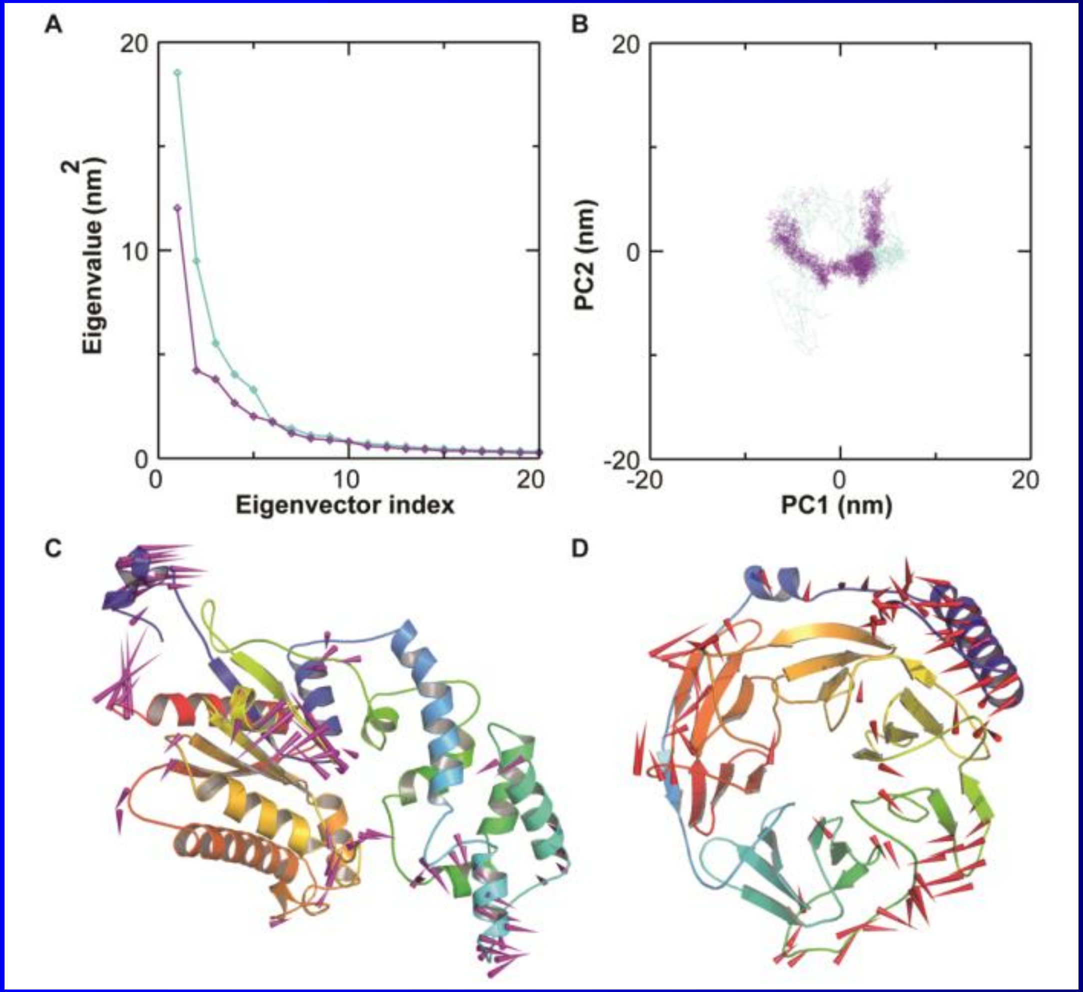


Figure 6. PCA of the protein–ligand (Gα and Gβ) systems using the resultant MD trajectories. (**A**) Eigenvalues for the complex as a function of the first 20 eigenvectors. (**B**) The cloud epitomizes the 50 ns trajectories projected onto the first two PCs where the x-axis and y-axis show the projection of the structures of the main-chain atoms in the MD trajectories onto the phase space defined by first two sets PCs (PC1 vs. PC2). (**C**) Porcupine plot depicting the movement of main-chain atoms of the first PC (PC1) of the G-α- Bipolaroxin complex. (**D**) Porcupine plot depicting the movement of main-chain atoms of the first PC of the G-β- Bipolaroxin complex. The direction of arrows indicates the motion and thickness of the arrow indicates the strength of motion. The image was generated using modevector.py script in PyMOL.

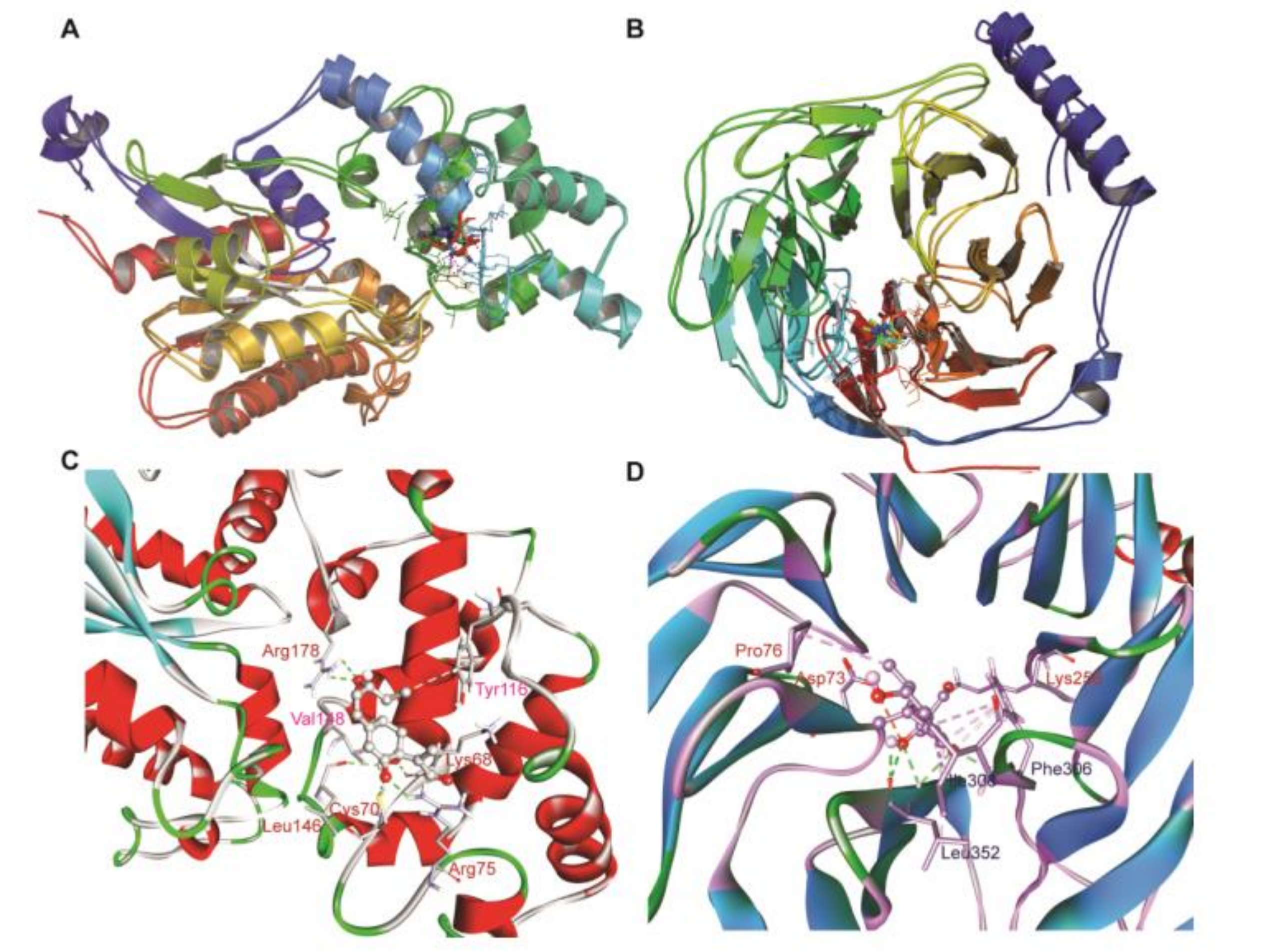
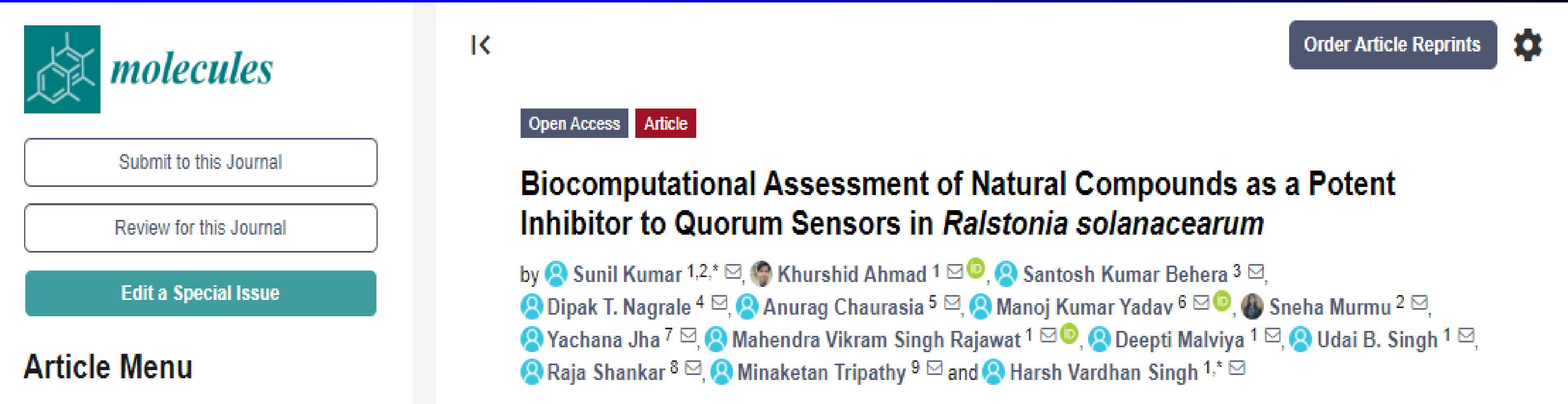


Figure 7. Structural superimposed view of the top two structural ensemble (top-ranked two clusters obtained from clustering). (**A**) Superimposed architecture of top-ranked two clusters from G-alpha subunit complex. (**B**) Superimposed architecture of top-ranked two clusters from G-beta subunit complex. Both the images were rendered using PyMOL. (**C**) Intermolecular contact of top-ranked cluster of Bipolaroxin with Gα subunit where the H-bond forming residues are marked in red while other nonbonded contacts are in pink. (**D**) Protein–ligand interaction analysis of the top-ranked cluster of Bipolaroxin with Gβ subunit where the H-bond forming residues are marked in blue while other nonbonded contacts are in red. The image was generated using BIOVIA DSV.



Biocomputational Assessment of Natural Compounds as a Potent Inhibitor to Quorum Sensors in *Ralstonia solanacearum*

by Sunil Kumar [1,2,*], Khurshid Ahmad [1], Santosh Kumar Behera [3], Dipak T. Nagrale [4], Anurag Chaurasia [5], Manoj Kumar Yadav [6], Sneha Murmu [2], Yachana Jha [7], Mahendra Vikram Singh Rajawat [1], Deepti Malviya [1], Udai B. Singh [1], Raja Shankar [8], Minaketan Tripathy [9] and Harsh Vardhan Singh [1,*]

- *Ralstonia solanacearum* is among the most damaging bacterial phytopathogens with a wide number of hosts and a broad geographic distribution worldwide. The pathway of phenotype conversion (Phc) is operated by quorum-sensing signals and modulated through the (R)-methyl 3-hydroxypalmitate (3-OH PAME) in R. solanacearum. However, the molecular structures of the Phc pathway components are not yet established, and the structural consequences of 3-OH PAME on quorum sensing are not well studied.
- In this study, 3D structures of quorum-sensing proteins of the Phc pathway (PhcA and PhcR) were computationally modeled, followed by the virtual screening of the natural compounds library against the predicted active site residues of PhcA and PhcR proteins that could be employed in limiting signaling through 3-OH PAME.
- Two of the best scoring common ligands ZINC000014762512 and ZINC000011865192 for PhcA and PhcR were further analyzed utilizing orbital energies such as HOMO and LUMO, followed by molecular dynamics simulations of the complexes for 100 ns to determine the ligands binding stability.
- The findings indicate that ZINC000014762512 and ZINC000011865192 may be capable of inhibiting both PhcA and PhcR. We believe that, after further validation, these compounds may have the potential to disrupt bacterial quorum sensing and thus control this devastating phytopathogenic bacterial pathogen.

Molecules: 2022: IF: 4.9

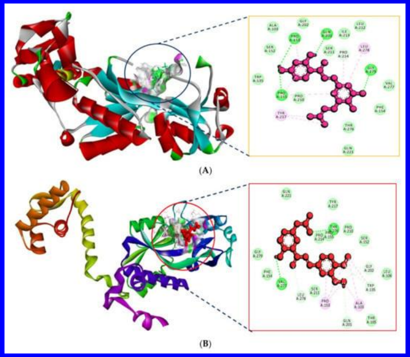


Figure Interaction of PhcA with ZINC000014762512. (**B**) Interaction of PhcA with ZINC000011865192.

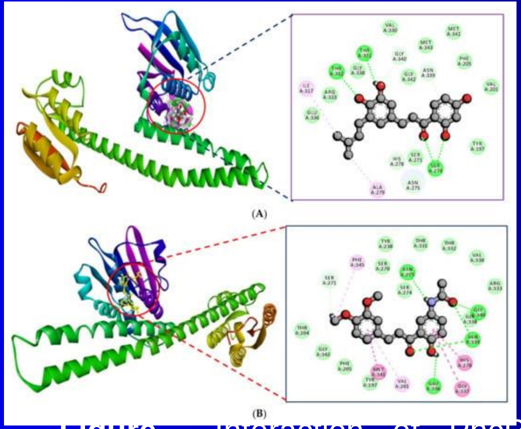


Figure Interaction of PhcR with ZINC000014762512. (**B**) Interaction of PhcR with ZINC000011865192.

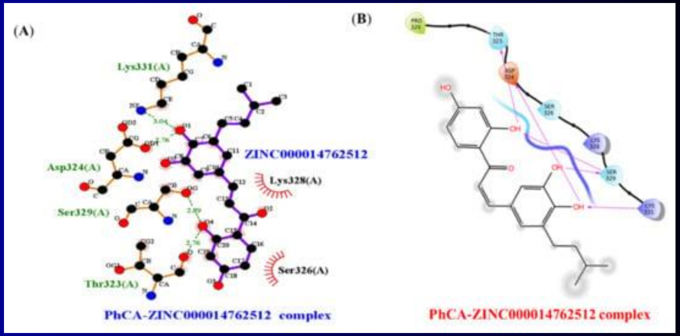


Figure. Intermolecular H-bonding, electrostatic, and hydrophobic interactions formed between PhccA–ZINC000014762512 complexes. The image (**A**) is drawn by the LigPlot+ tool and (**B**) ligand interaction module of Schrödinger.

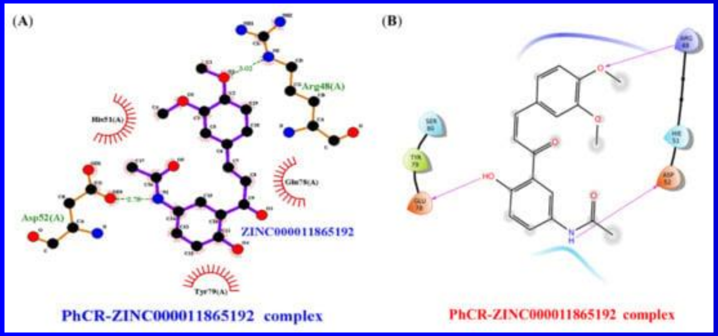


Figure 5. Intermolecular H-bonding, electrostatic, and hydrophobic interactions formed between PhccR–ZINC000011865192 complexes. The image (**A**) is drawn by the LigPlot+ tool and (**B**) ligand interaction module of Schrödinger.

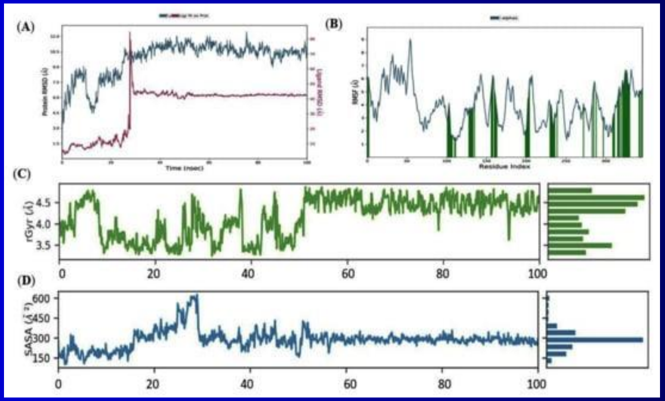
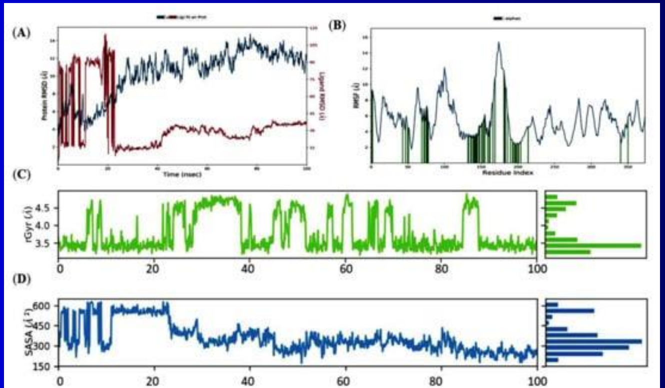
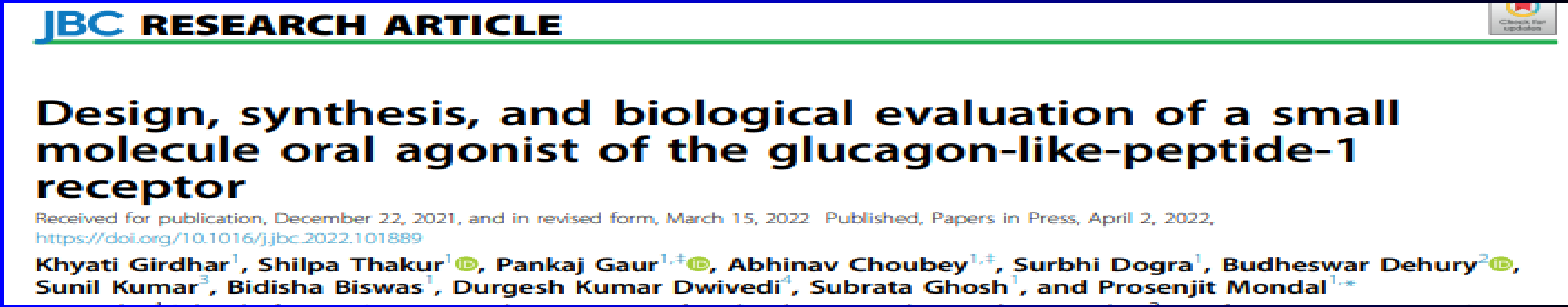


Figure Conformational constancy of ‘Apo’ and ‘Holo’ states of PhcA protein simulation study. (**A**) Backbone-RMSD of PhcA. (**B**) Cα-RMSF profile of PhcA. (**C**) Rg profile of PhcA. (**D**) SASA analysis of Apo and Holo states of PhcA protein throughout the simulations.




JBC RESEARCH ARTICLE

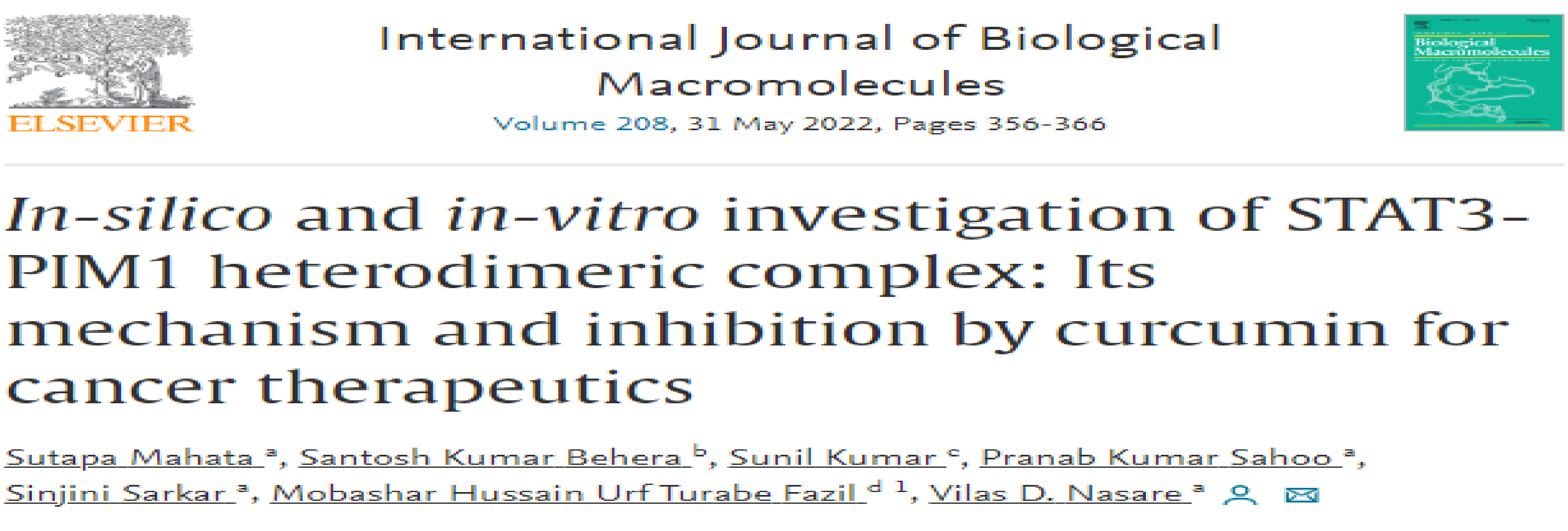
Design, synthesis, and biological evaluation of a small molecule oral agonist of the glucagon-like-peptide-1 receptor

Received for publication, December 22, 2021, and in revised form, March 15, 2022 Published, Papers in Press, April 2, 2022,
https://doi.org/10.1016/j.jbc.2022.101889

Khyati Girdhar[1], Shilpa Thakur[1], Pankaj Gaur[1,‡], Abhinav Choubey[1,‡], Surbhi Dogra[1], Budheswar Dehury[2], Sunil Kumar[3], Bidisha Biswas[1], Durgesh Kumar Dwivedi[4], Subrata Ghosh[1], and Prosenjit Mondal[1,*]

- An absolute or relative deficiency of pancreatic β-cells mass and functionality is a crucial pathological feature common to type 1 diabetes mellitus and type 2 diabetes mellitus. Glucagon-like-peptide-1 receptor (GLP1R) agonists have been the focus of considerable research attention for their ability to protect β-cell mass and augment insulin secretion with no risk of hypoglycemia.
- Presently commercially available GLP1R agonists are peptides that limit their use due to cost, stability, and mode of administration.
- To address this drawback, strategically designed distinct sets of small molecules were docked on GLP1R ectodomain and compared with previously known small molecule GLP1R agonists. One of the small molecule PK2 (6-((1-(4-nitrobenzyl)-1H-1,2,3-triazol-4-yl)methyl)-6Hindolo[2,3-b]quinoxaline) displays stable binding with GLP1R ectodomain and induces GLP1R internalization and increasing cAMP levels.
- PK2 also increases insulin secretion in the INS1 cells. The oral administration of PK2 protects against diabetes induced by multiple low-dose streptozotocin administration by lowering high blood glucose levels.
- Similar to GLP1R peptidic agonists, treatment of PK2 induces β-cell replication and attenuate β-cell apoptosis in STZ-treated mice. Mechanistically, this protection was associated with decreased thioredoxin-interacting protein expression, a potent inducer of diabetic β-cell apoptosis and dysfunction. Together, this report describes a small molecule, PK2, as an orally active nonpeptidic GLP1R agonist that has efficacy to preserve or restore functional β-cell mass

JBC; 2022; IF: 5.3



ELSEVIER

International Journal of Biological Macromolecules

Volume 208, 31 May 2022, Pages 356-366

In-silico and *in-vitro* investigation of STAT3-PIM1 heterodimeric complex: Its mechanism and inhibition by curcumin for cancer therapeutics

Sutapa Mahata [a], Santosh Kumar Behera [b], Sunil Kumar [c], Pranab Kumar Sahoo [a], Sinjini Sarkar [a], Mobashar Hussain Urf Turabe Fazil [d] [1], Vilas D. Nasare [a]

- The functional activity among STAT3 and PIM1, are key signaling events for cancer cell function. Curcumin, a diarylheptanoid isolated from turmeric, effectively inhibits STAT3 signaling.
- Selectively, we attempted to address interactions of STAT3, PIM1 and Curcumin for therapeutic intervention using *in silico* and *in vitro* experimental approaches.
- Firstly, protein-protein interactions (PPI) between STAT3-PIM1 by molecular docking studies reflected salt bridges among Arg279 (STAT3)-Glu140 (PIM1) and Arg282 (STAT3)-Asp100 (PIM1), with a binding affinity of −38.6 kcal/mol.
- Secondly, molecular dynamics simulations of heterodimeric STAT3-PIM1 complex with curcumin revealed binding of curcumin on PIM-1 interface of the complex through hydrogen bonds (Asp155) and hydrophobic interactions (Leu13, Phe18, Val21, *etc.*) with a binding energy of −7.3 kcal/mol.
- These PPIs were confirmed *in vitro* by immunoprecipitation assays in MDA-MB-231 cells. Corroborating our results, expression levels of STAT3 and PIM1 decreased after curcumin treatment.
- We observed that PIM1 interacts with STAT3 and these functional interactions are disrupted by curcumin. The calculated band energy gap of heterodimeric STAT3-PIM1-Curcumin complex was of 9.621 kcal/mol.
- The present study revealed the role of curcumin in STAT3/PIM1 signaling and its binding affinity to the complex for design of advanced cancer therapeutics.

IJBM; 2022; IF: 8.05

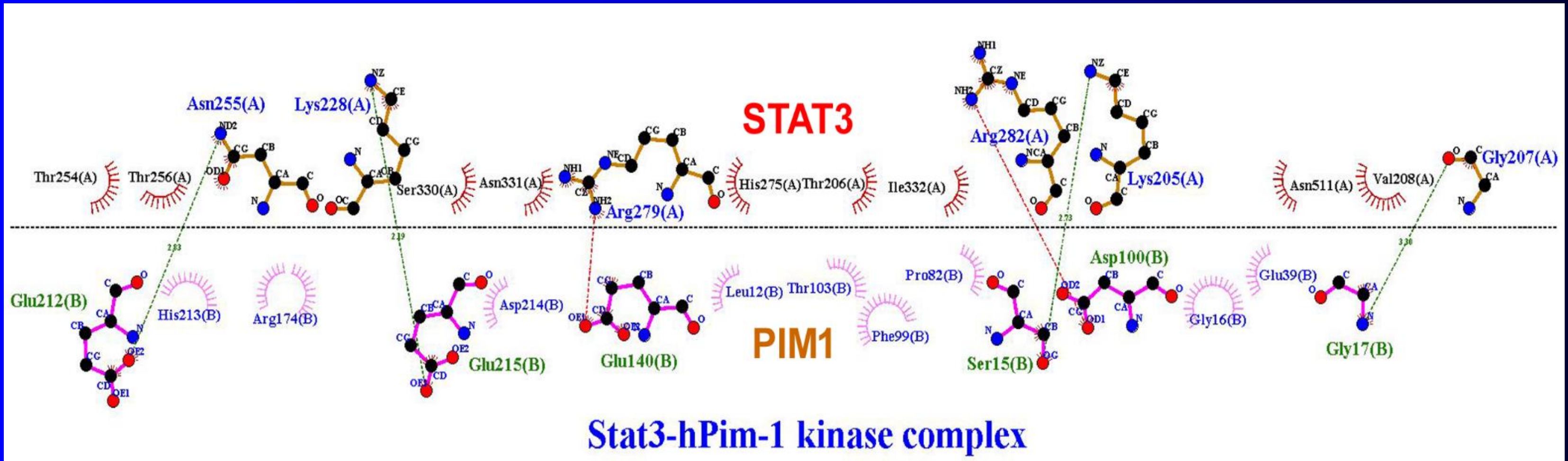
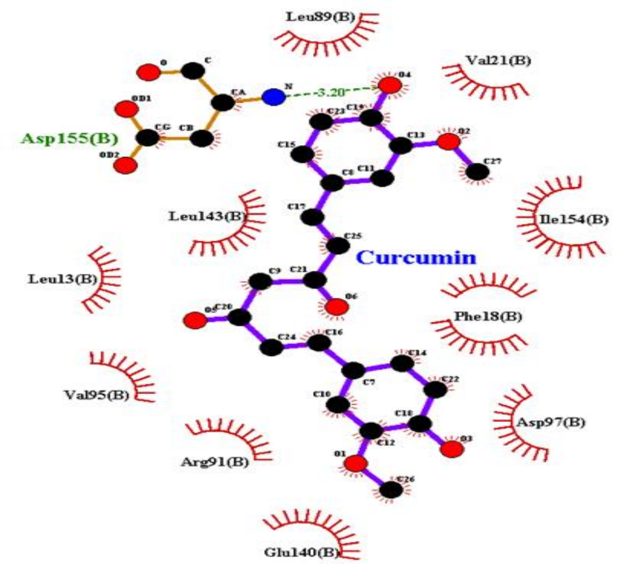
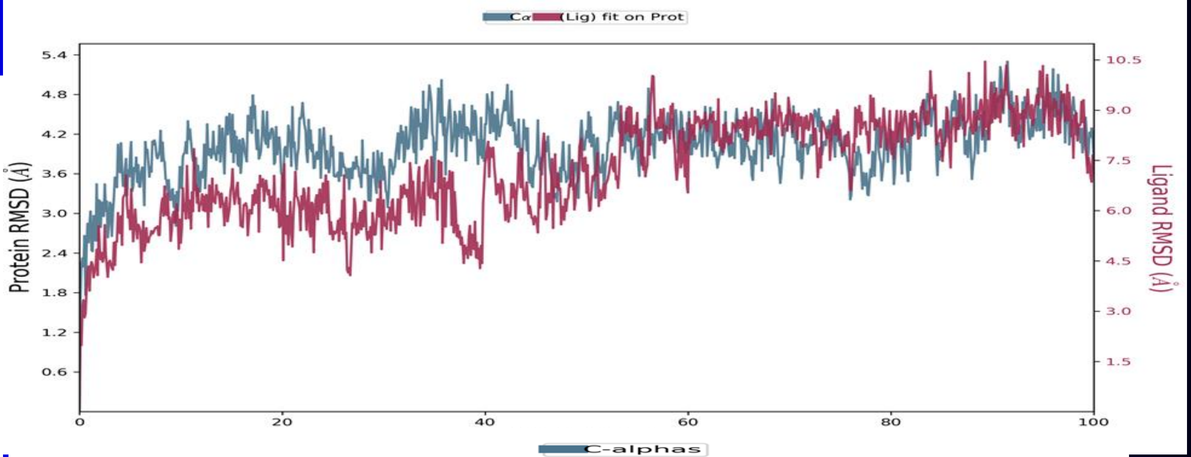
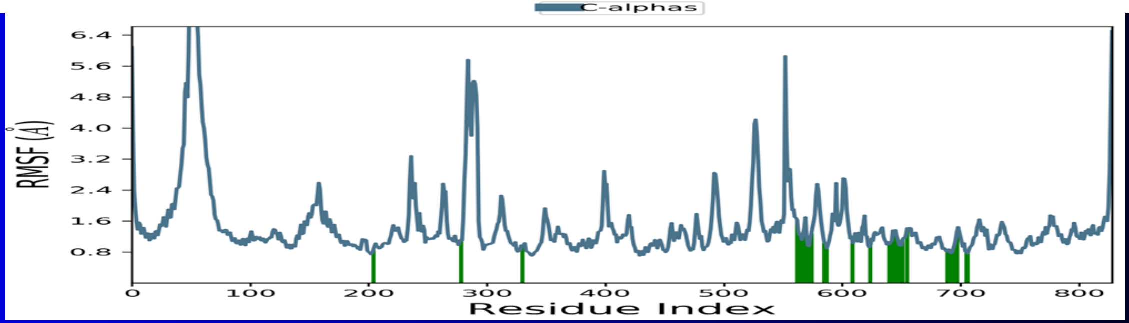


Figure Protein- protein interaction and three-dimensional structure analysis of hPim-1 kinase (PIM1) with Signal transducer and activator of transcription 3 (STAT3)







A Drug Repurposing Approach to Identify Therapeutics by Screening Pathogen Box Exploiting SARS-CoV-2 Main Protease

- (COVID-19) is caused by severe acute respiratory syndrome coronavirus -2 (SARS-CoV2) and is responsible for a higher degree of morbidity and mortality worldwide. There is a smaller number of approved therapeutics available to target the SARS-CoV-2 virus. The main protease (Mpro) enzyme of SARS-CoV-2 is essential for replication and transcription of the viral genome, thus could be a potent target for the treatment of COVID-19.

- We performed an in-silico screening analysis of 400 diverse bioactive inhibitors with proven antibacterial and antiviral properties against Mpro drug target. Ten compounds showed a higher binding affinity for Mpro than the reference compound (N3), with desired physicochemical properties.

- in-depth docking and superimposition revealed that three compounds (MMV1782211, MMV1782220, and MMV1578574) are actively interacting with the catalytic domain of Mpro. In addition, the molecular dynamics simulation study showed a solid and stable interaction of MMV178221-Mpro complex compared to the other two molecules (MMV1782220, and MMV1578574).

- In conclusion, the present in silico analysis revealed MMV1782211 as a possible and potent molecule to target the Mpro and must be explored *in vitro* and *in vivo* to combat the COVID-19

JBSD; 2023; IF: 5.25

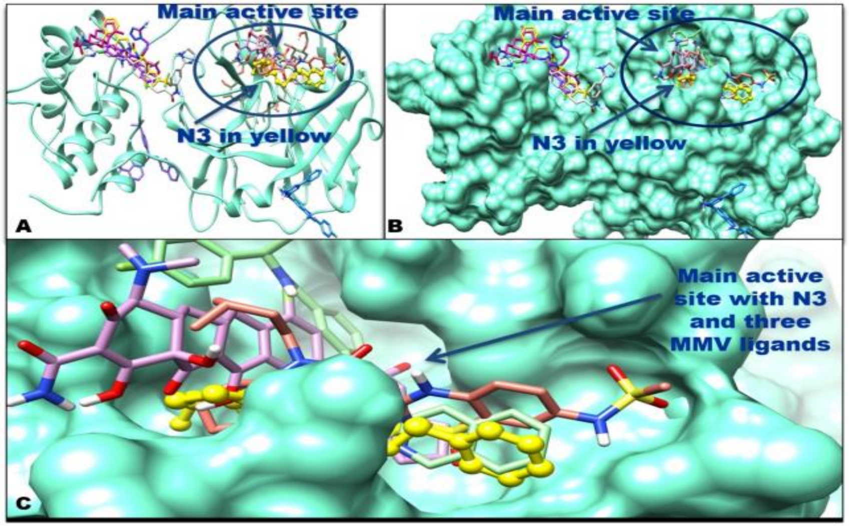


Figure 3. Binding of ligands at the M^{pro} receptor protein. (A) Secondary structure representation B) Surface view; Only three ligan (MMV1782220 MMV1782211 and MMV1578574) out of 10 binds at the main active site of the receptor (C) enlarged surface view the main active site showing three ligands at main active site pocket along with N3 (yellow color). Chimera software is used f visualization of complexes.

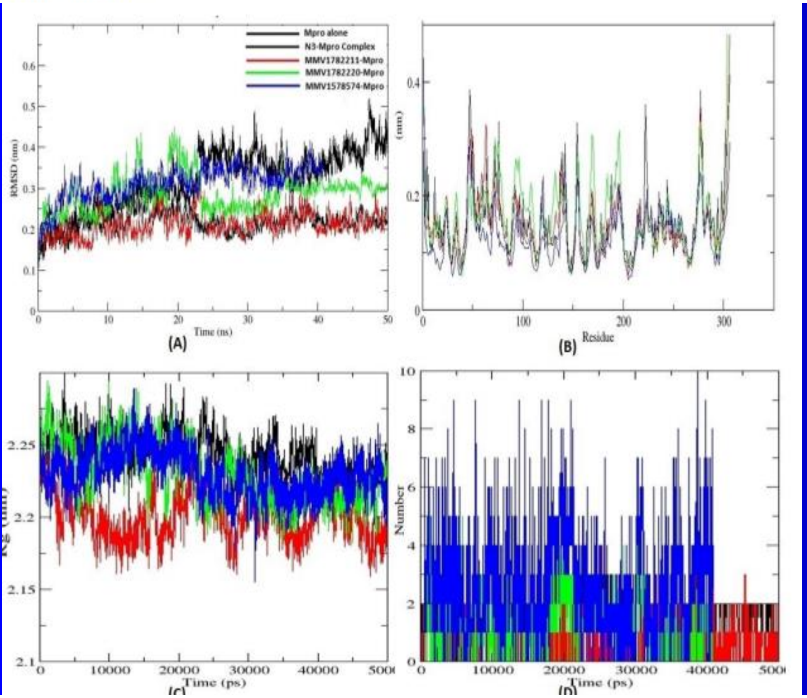


Figure 5. Comparison of molecular dynamics simulation trajectories (A) Root mean square deviation, (B) Root mean square fluctuations, (C) Radius of gyration, and (D) Number of hydrogen bond formation, for M^{pro} protein docked with the reference ligand N3 (black), MMV1782211 (red), MMV1782220 (green), and MMV1578574 (blue) over the 50 ns simulations. The trajectory graphs are developed using XMGRACE tool.

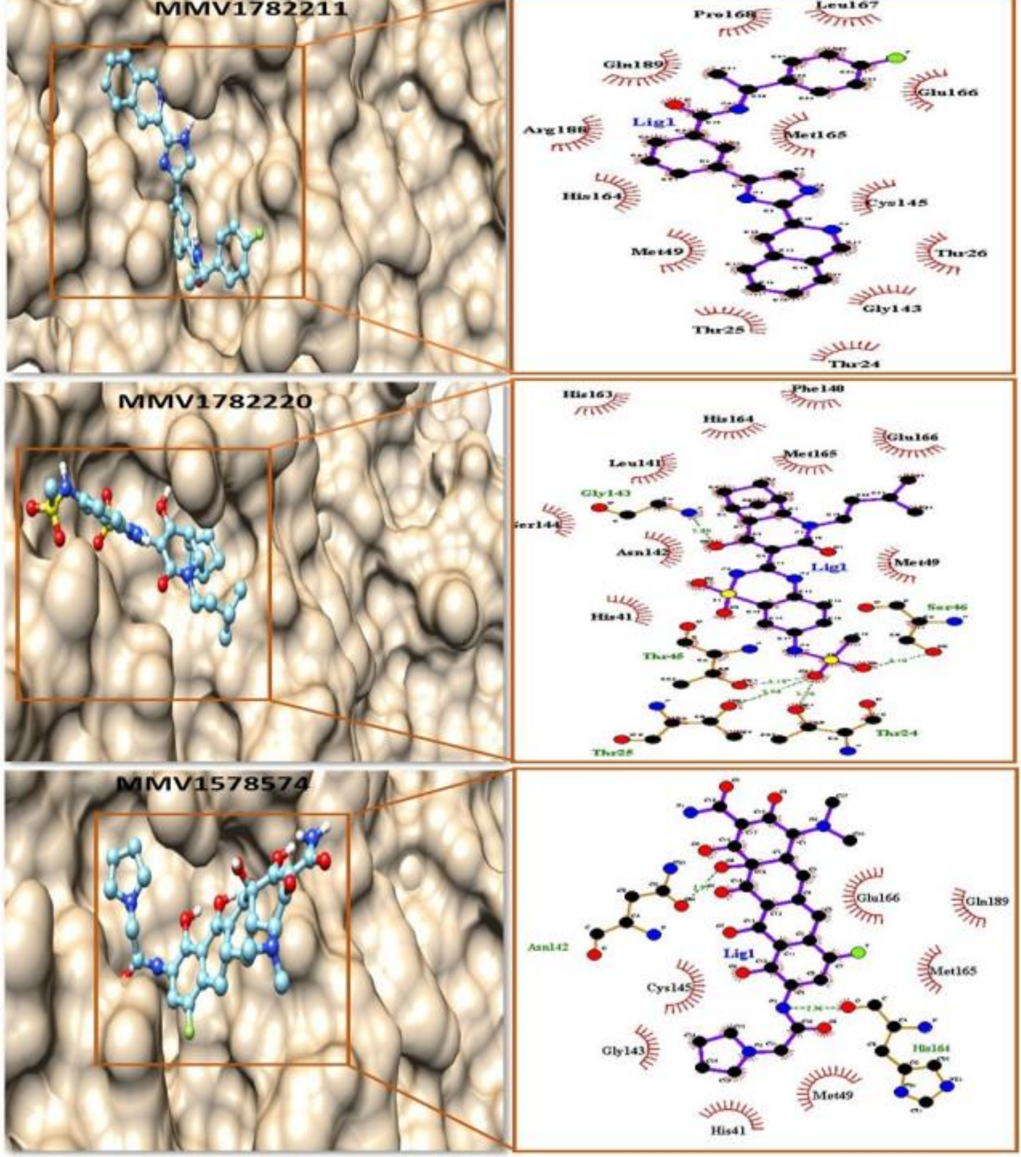
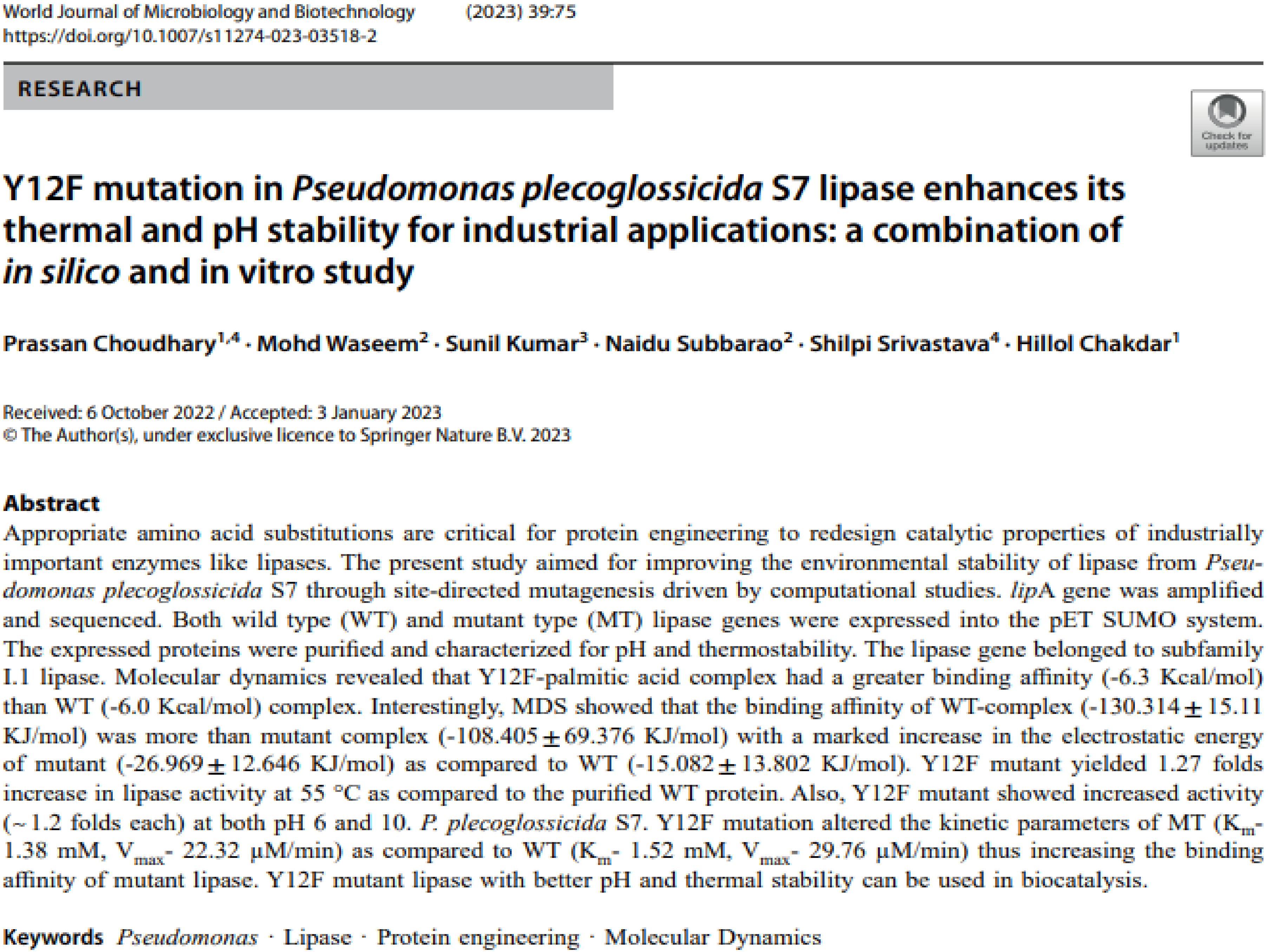


Figure 4. Binding modes of minimum energy conformers after docking experiments of MMV compounds: 3D structure of M^{pro} protein is shown as molecular surface models in Tan color and ligands are represented as ball and stick models on the left-hand side using Chimera software while ligand-receptor interactions and their close contact residues are visible on the right-hand side pane using LigPlot program where hydrogen bonds are labeled in green color.



World Journal of Microbiology and Biotechnology (2023) 39:75
https://doi.org/10.1007/s11274-023-03518-2

RESEARCH

Y12F mutation in *Pseudomonas plecoglossicida* S7 lipase enhances its thermal and pH stability for industrial applications: a combination of *in silico* and in vitro study

Prassan Choudhary[1,4] · Mohd Waseem[2] · Sunil Kumar[3] · Naidu Subbarao[2] · Shilpi Srivastava[4] · Hillol Chakdar[1]

Received: 6 October 2022 / Accepted: 3 January 2023

Abstract

Appropriate amino acid substitutions are critical for protein engineering to redesign catalytic properties of industrially important enzymes like lipases. The present study aimed for improving the environmental stability of lipase from *Pseudomonas plecoglossicida* S7 through site-directed mutagenesis driven by computational studies. *lip*A gene was amplified and sequenced. Both wild type (WT) and mutant type (MT) lipase genes were expressed into the pET SUMO system. The expressed proteins were purified and characterized for pH and thermostability. The lipase gene belonged to subfamily I.1 lipase. Molecular dynamics revealed that Y12F-palmitic acid complex had a greater binding affinity (-6.3 Kcal/mol) than WT (-6.0 Kcal/mol) complex. Interestingly, MDS showed that the binding affinity of WT-complex (-130.314 ± 15.11 KJ/mol) was more than mutant complex (-108.405 ± 69.376 KJ/mol) with a marked increase in the electrostatic energy of mutant (-26.969 ± 12.646 KJ/mol) as compared to WT (-15.082 ± 13.802 KJ/mol). Y12F mutant yielded 1.27 folds increase in lipase activity at 55 °C as compared to the purified WT protein. Also, Y12F mutant showed increased activity (~ 1.2 folds each) at both pH 6 and 10. *P. plecoglossicida* S7. Y12F mutation altered the kinetic parameters of MT (K_m- 1.38 mM, V_{max}- 22.32 µM/min) as compared to WT (K_m- 1.52 mM, V_{max}- 29.76 µM/min) thus increasing the binding affinity of mutant lipase. Y12F mutant lipase with better pH and thermal stability can be used in biocatalysis.

Keywords *Pseudomonas* · Lipase · Protein engineering · Molecular Dynamics

WJMB; 2023; IF: 4.25

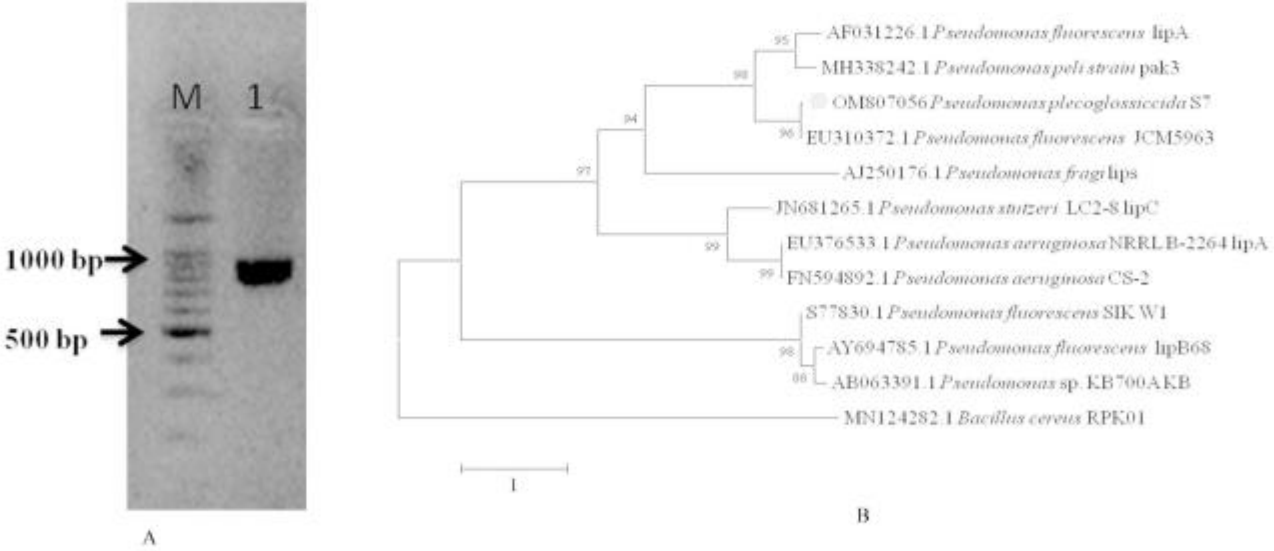


Fig. 1 Molecular and phylogenetic analyses of *lip*A gene from *P. plecoglossicida* S7 (A) PCR amplification of *lip*A gene. (B) Phylogram showing the evolutionary relatedness of *lipA* with lipases from other *Pseudomonas* spp

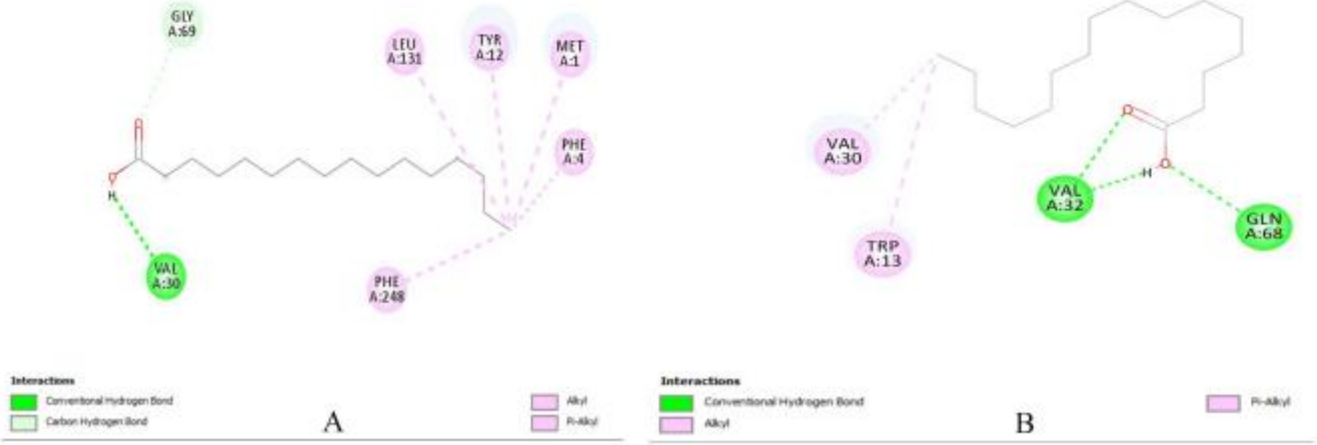


Fig. 2 Cloning of *lip*A gene from *P. plecoglossicida* S7 showing (A) positive transformants on Luria Bertani plate supplemented with Kanamycin (50 µg/mL) and (B) molecular confirmation of positive transformants though colony PCR.

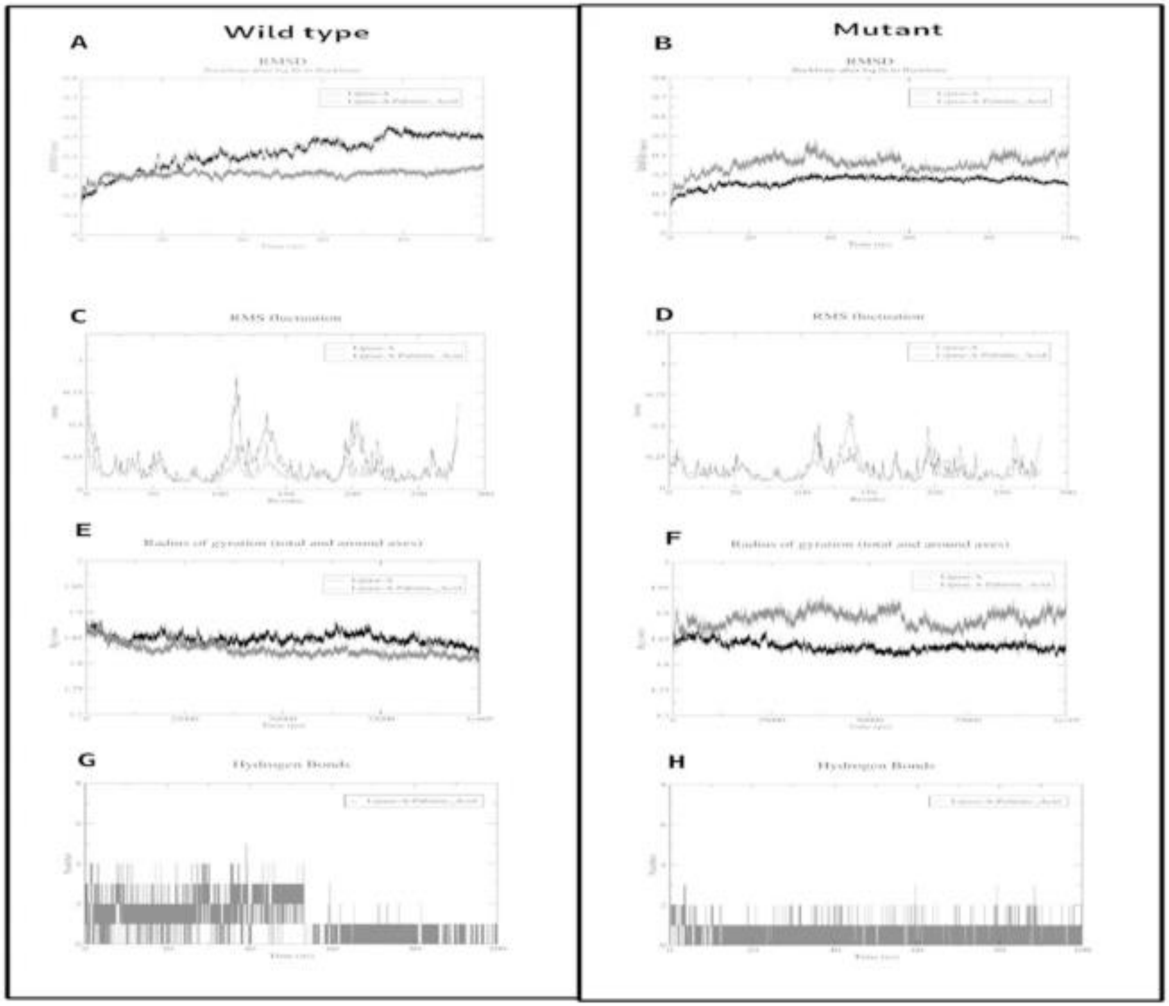


Fig. 3 *in silico* prediction and validation of lipase (LipA) from *P. plecoglossicida* S7 showing (A) 3-D protein model and (B) contour plot of amino acid residues obtained from Verify3D tool of Structure validation and analysis (SAVES) server

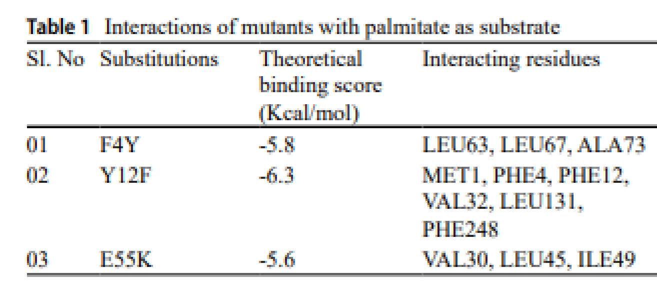


Table 1 Interactions of mutants with palmitate as substrate

Sl. No	Substitutions	Theoretical binding score (Kcal/mol)	Interacting residues
01	F4Y	-5.8	LEU63, LEU67, ALA73
02	Y12F	-6.3	MET1, PHE4, PHE12, VAL32, LEU131, PHE248
03	E55K	-5.6	VAL30, LEU45, ILE49

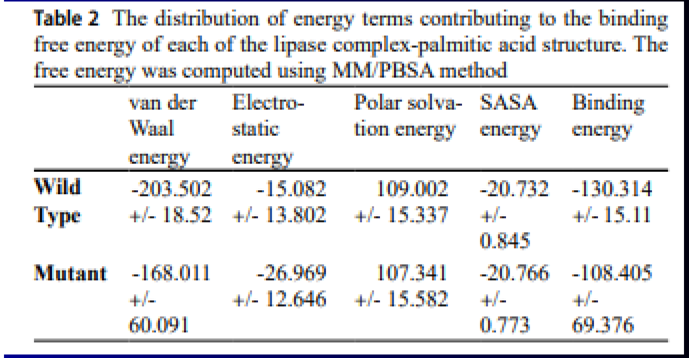
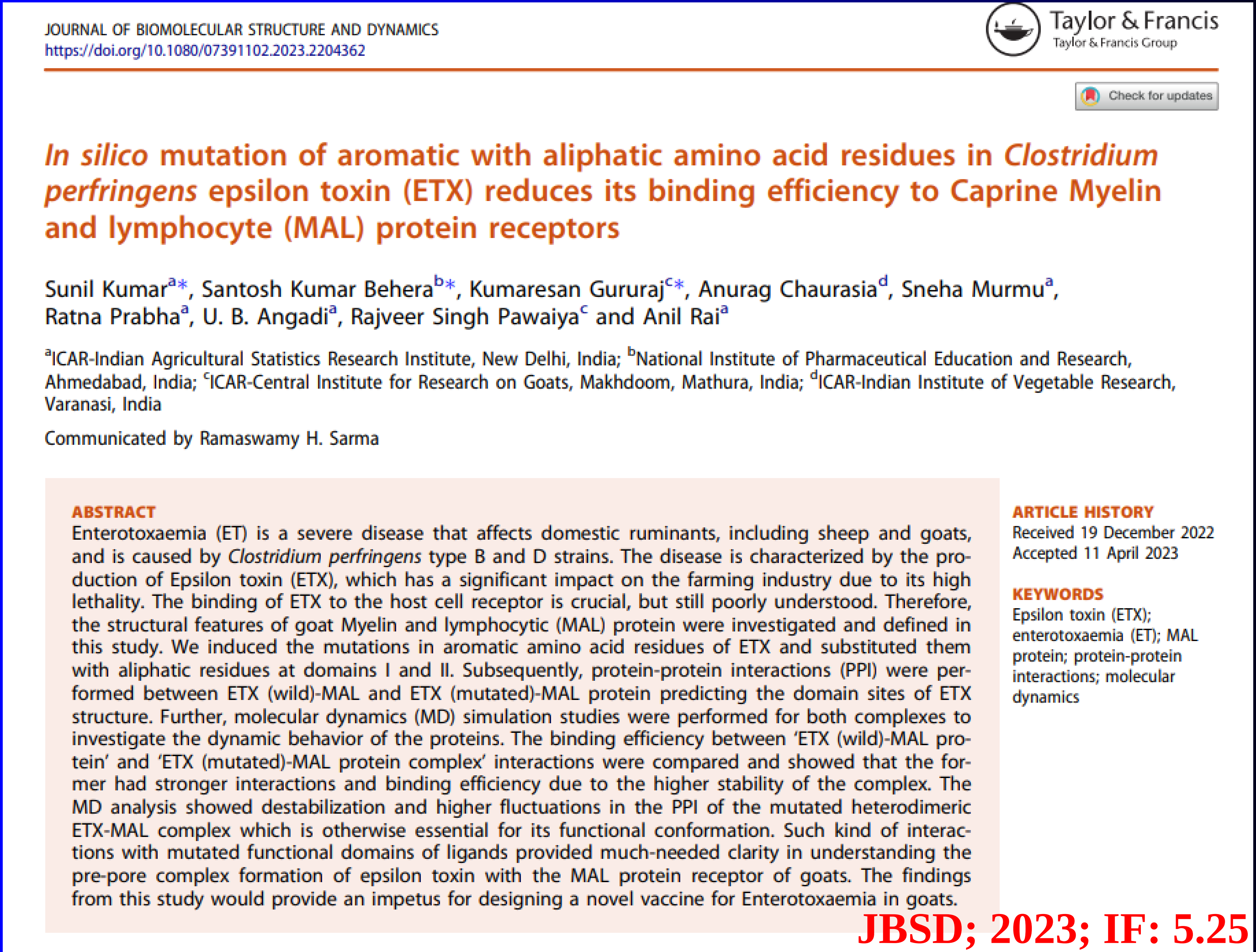


Table 2 The distribution of energy terms contributing to the binding free energy of each of the lipase complex-palmitic acid structure. The free energy was computed using MM/PBSA method

	van der Waal energy	Electro-static energy	Polar solva-tion energy	SASA energy	Binding energy
Wild Type	-203.502 +/- 18.52	-15.082 +/- 13.802	109.002 +/- 15.337	-20.732 +/- 0.845	-130.314 +/- 15.11
Mutant	-168.011 +/- 60.091	-26.969 +/- 12.646	107.341 +/- 15.582	-20.766 +/- 0.773	-108.405 +/- 69.376



JOURNAL OF BIOMOLECULAR STRUCTURE AND DYNAMICS
https://doi.org/10.1080/07391102.2023.2204362

Taylor & Francis
Taylor & Francis Group

In silico mutation of aromatic with aliphatic amino acid residues in *Clostridium perfringens* epsilon toxin (ETX) reduces its binding efficiency to Caprine Myelin and lymphocyte (MAL) protein receptors

Sunil Kumar[a*], Santosh Kumar Behera[b*], Kumaresan Gururaj[c*], Anurag Chaurasia[d], Sneha Murmu[a], Ratna Prabha[a], U. B. Angadi[a], Rajveer Singh Pawaiya[c] and Anil Rai[a]

[a]ICAR-Indian Agricultural Statistics Research Institute, New Delhi, India; [b]National Institute of Pharmaceutical Education and Research, Ahmedabad, India; [c]ICAR-Central Institute for Research on Goats, Makhdoom, Mathura, India; [d]ICAR-Indian Institute of Vegetable Research, Varanasi, India

Communicated by Ramaswamy H. Sarma

ABSTRACT

Enterotoxaemia (ET) is a severe disease that affects domestic ruminants, including sheep and goats, and is caused by *Clostridium perfringens* type B and D strains. The disease is characterized by the production of Epsilon toxin (ETX), which has a significant impact on the farming industry due to its high lethality. The binding of ETX to the host cell receptor is crucial, but still poorly understood. Therefore, the structural features of goat Myelin and lymphocytic (MAL) protein were investigated and defined in this study. We induced the mutations in aromatic amino acid residues of ETX and substituted them with aliphatic residues at domains I and II. Subsequently, protein-protein interactions (PPI) were performed between ETX (wild)-MAL and ETX (mutated)-MAL protein predicting the domain sites of ETX structure. Further, molecular dynamics (MD) simulation studies were performed for both complexes to investigate the dynamic behavior of the proteins. The binding efficiency between 'ETX (wild)-MAL protein' and 'ETX (mutated)-MAL protein complex' interactions were compared and showed that the former had stronger interactions and binding efficiency due to the higher stability of the complex. The MD analysis showed destabilization and higher fluctuations in the PPI of the mutated heterodimeric ETX-MAL complex which is otherwise essential for its functional conformation. Such kind of interactions with mutated functional domains of ligands provided much-needed clarity in understanding the pre-pore complex formation of epsilon toxin with the MAL protein receptor of goats. The findings from this study would provide an impetus for designing a novel vaccine for Enterotoxaemia in goats.

ARTICLE HISTORY
Received 19 December 2022
Accepted 11 April 2023

KEYWORDS
Epsilon toxin (ETX); enterotoxaemia (ET); MAL protein; protein-protein interactions; molecular dynamics

JBSD; 2023; IF: 5.25

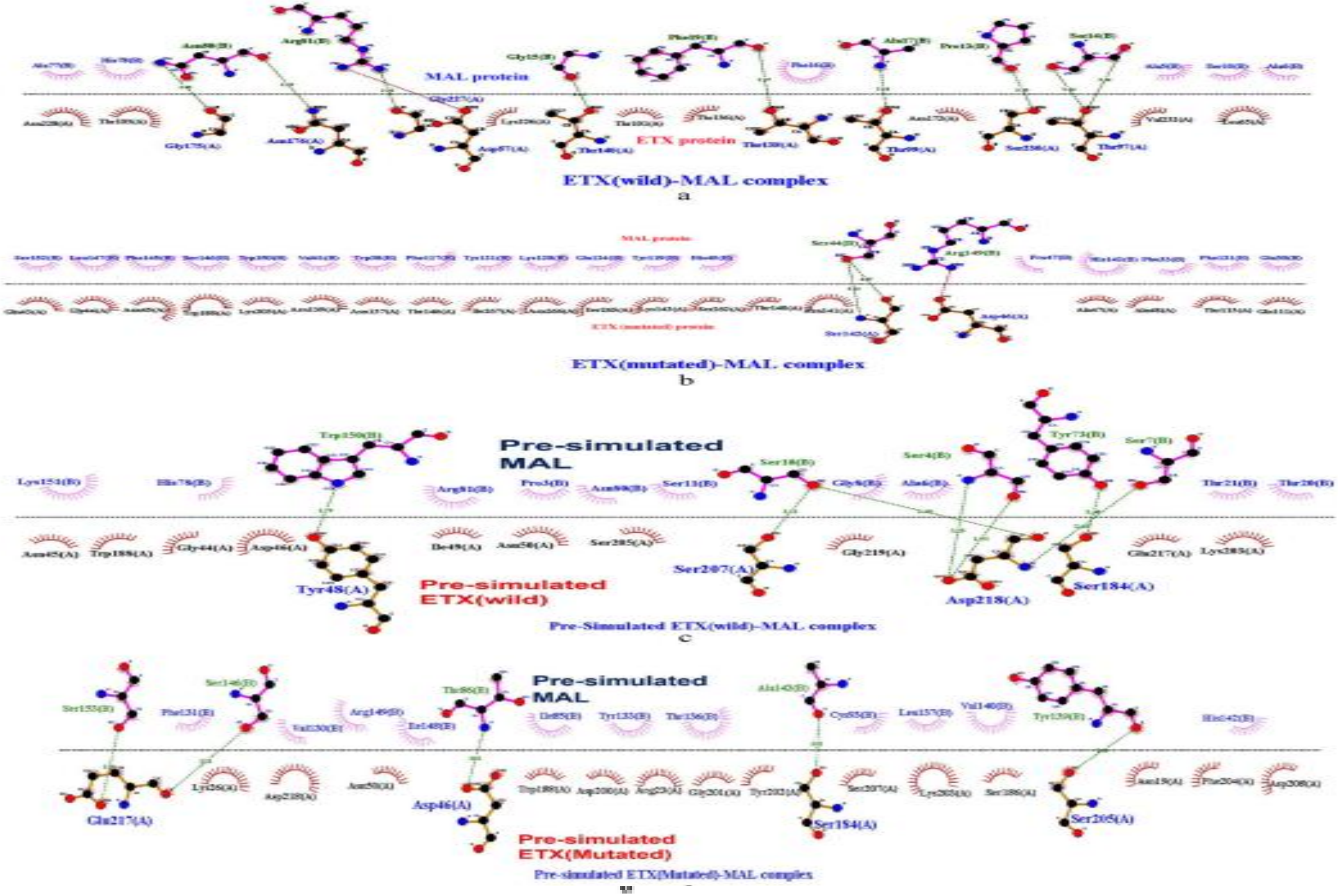
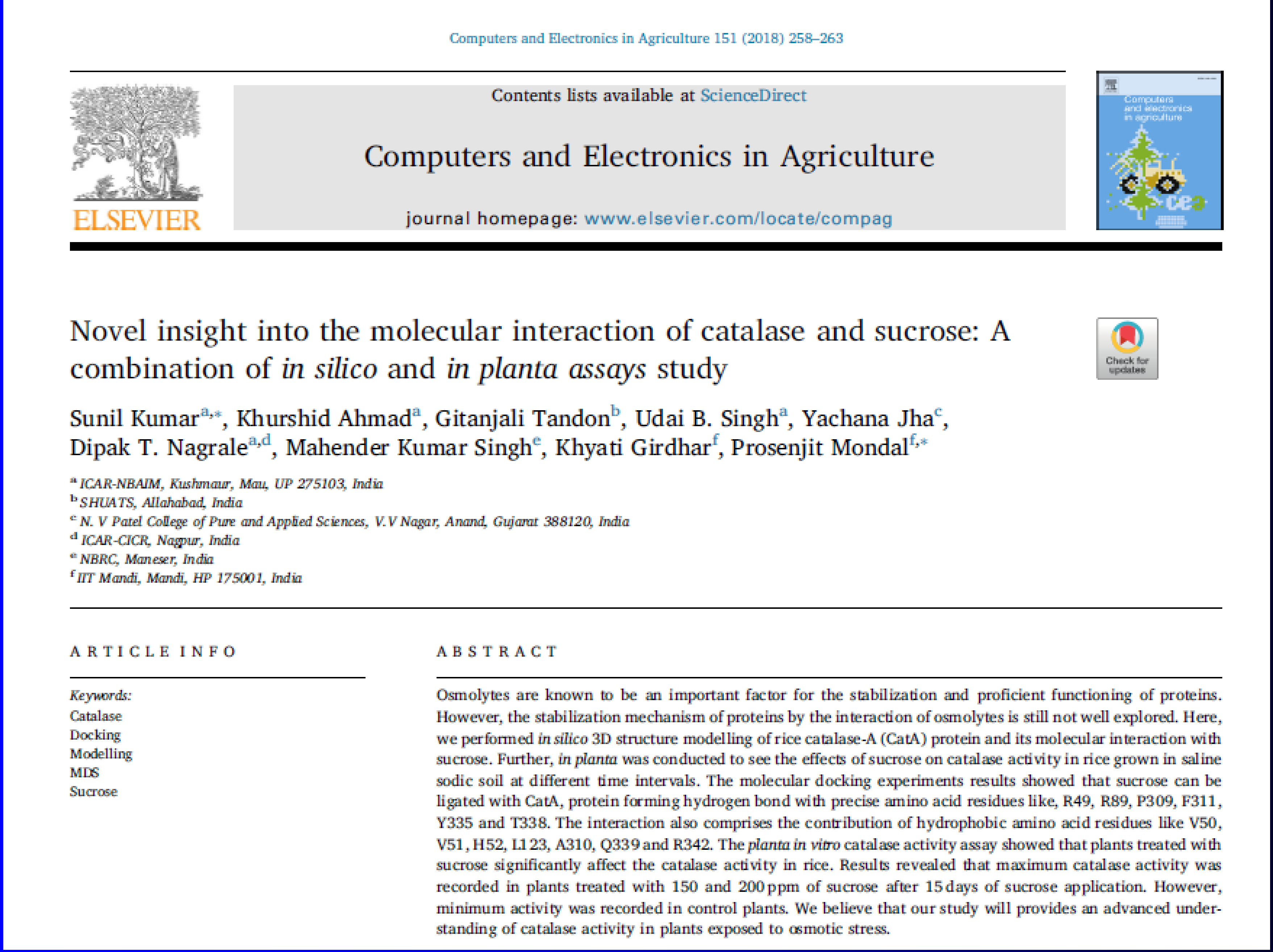


Figure 1. (a) The protein–protein interaction of non-simulated Epsilon toxin (ETX-wild) with non-simulated Myelin and lymphocyte (MAL) protein. (b) The protein–protein interaction of non-simulated Epsilon toxin (ETX-mutated) with non-simulated Myelin and lymphocyte (MAL) protein. (c) The protein–protein interaction of pre-simulated Epsilon toxin (ETX-wild) with pre-simulated Myelin and lymphocyte (MAL) protein. (d) The protein–protein interaction of pre-simulated Epsilon toxin (ETX-mutated) with pre-simulated Myelin and lymphocyte (MAL) protein.



Computers and Electronics in Agriculture 151 (2018) 258–263

Contents lists available at ScienceDirect

Computers and Electronics in Agriculture

journal homepage: www.elsevier.com/locate/compag

Novel insight into the molecular interaction of catalase and sucrose: A combination of *in silico* and *in planta assays* study

Sunil Kumar[a,*], Khurshid Ahmad[a], Gitanjali Tandon[b], Udai B. Singh[a], Yachana Jha[c], Dipak T. Nagrale[a,d], Mahender Kumar Singh[e], Khyati Girdhar[f], Prosenjit Mondal[f,*]

[a] ICAR-NBAIM, Kushmaur, Mau, UP 275103, India
[b] SHUATS, Allahabad, India
[c] N. V Patel College of Pure and Applied Sciences, V.V Nagar, Anand, Gujarat 388120, India
[d] ICAR-CICR, Nagpur, India
[e] NBRC, Maneser, India
[f] IIT Mandi, Mandi, HP 175001, India

ARTICLE INFO

Keywords:
Catalase
Docking
Modelling
MDS
Sucrose

ABSTRACT

Osmolytes are known to be an important factor for the stabilization and proficient functioning of proteins. However, the stabilization mechanism of proteins by the interaction of osmolytes is still not well explored. Here, we performed *in silico* 3D structure modelling of rice catalase-A (CatA) protein and its molecular interaction with sucrose. Further, *in planta* was conducted to see the effects of sucrose on catalase activity in rice grown in saline sodic soil at different time intervals. The molecular docking experiments results showed that sucrose can be ligated with CatA, protein forming hydrogen bond with precise amino acid residues like, R49, R89, P309, F311, Y335 and T338. The interaction also comprises the contribution of hydrophobic amino acid residues like V50, V51, H52, L123, A310, Q339 and R342. The *planta in vitro* catalase activity assay showed that plants treated with sucrose significantly affect the catalase activity in rice. Results revealed that maximum catalase activity was recorded in plants treated with 150 and 200 ppm of sucrose after 15 days of sucrose application. However, minimum activity was recorded in control plants. We believe that our study will provides an advanced understanding of catalase activity in plants exposed to osmotic stress.

IF 8.

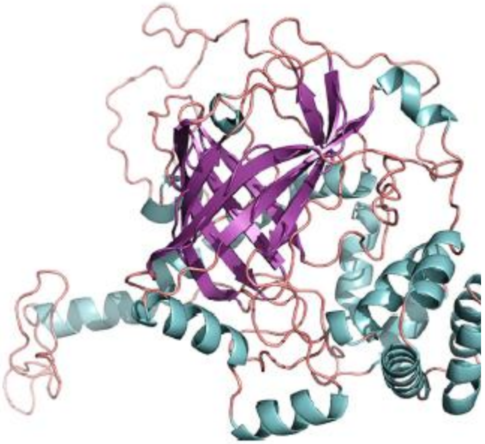


Fig. 1. 3D structure of CatA protein (Cartoon view).

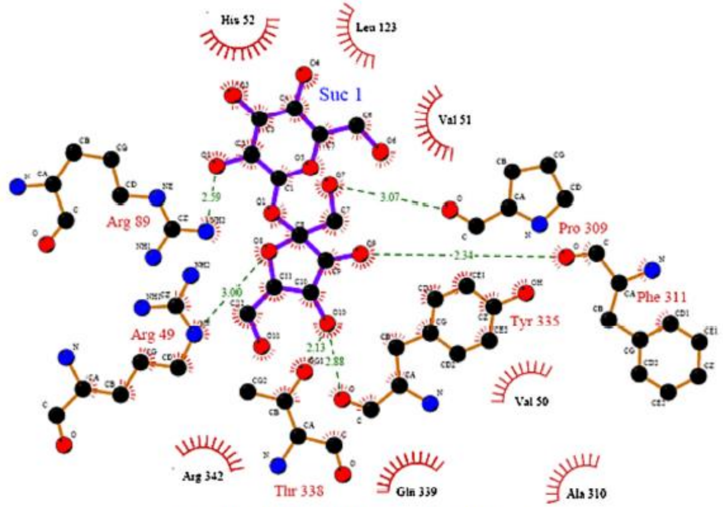


Fig. 6. Molecular interaction between catalase and sucrose.

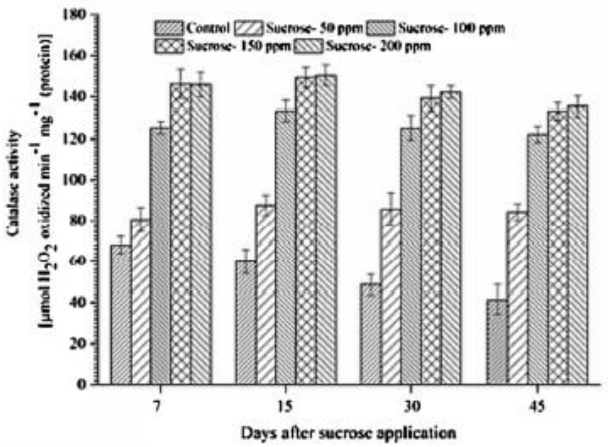


Fig. 7. Quantitative estimation of catalase activity.

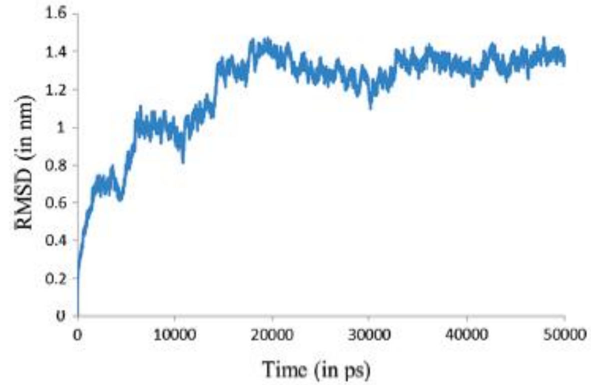


Fig. 2. Conformational stability of catalase sucrose complex at 50 ns time duration.

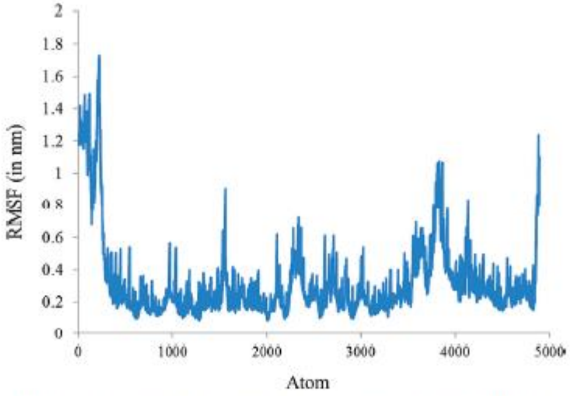


Fig. 3. Cα RMSF profile of catalase sucrose complex during 50 ns MD.

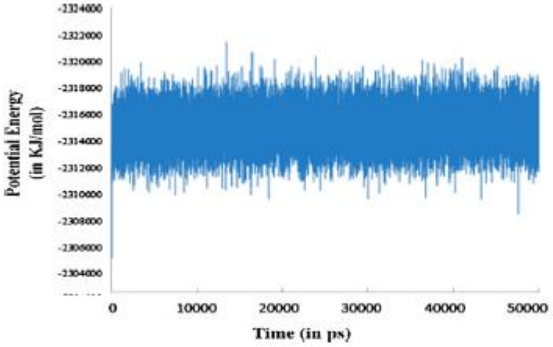


Fig. 4. Energy calculation of catalase sucrose complex.

RNA-Protein Interactions

RNA-binding protein

RNA-binding proteins (often abbreviated as **RBPs**)

- are proteins that bind to the double or single stranded RNA in cells
- and participate in forming ribonucleoprotein complexes.
- RBPs contain various structural motifs, such as
  - ✓ RNA recognition motif (RRM),
  - ✓ dsRNA binding domain,
  - ✓ zinc finger and others.
- They are cytoplasmic and nuclear proteins
- since most mature RNA is exported from the nucleus relatively quickly, most RBPs in the nucleus exist as complexes of protein and pre-mRNA called heterogeneous ribonucleoprotein particles (hnRNPs).

➢ RBPs have crucial roles in various cellular processes such as:

- ▪ cellular function, transport and localization.
- ▪ They especially play a major role in post-transcriptional control of RNAs
  - ✓ (splicing, polyadenylation, mRNA stabilization, mRNA localization and translation.

- Eukaryotic cells encode diverse RBPs, approximately 500 genes, with unique RNA-binding activity and protein–protein interaction.
- During evolution, the diversity of RBPs greatly increased with the increase in the number of introns.
- Diversity enabled eukaryotic cells to utilize RNA exons in various arrangements, giving rise to a unique RNP (ribonucleoprotein) for each RNA.

Structure

- Many RBPs have modular structures and are composed of multiple repeats of just a few specific basic domains that often have limited sequences.
- These sequences are then arranged in varying combinations to fulfill the need for diversity. A specific protein's recognition of a specific RNA has evolved through the rearrangement of these few basic domains.
- Each basic domain recognizes RNA, but many of these proteins require multiple copies of one of the many common domains to function.

Diversity

- As nuclear RNA emerges from RNA polymerase, RNA transcripts are immediately covered with RNA-binding proteins and function in RNA biogenesis, maturation, transport, cellular localization and stability.
- All RBPs bind RNA, however they do so with different RNA-sequence specificities and affinities, which allows the RBPs to be as diverse as their targets and functions.
- These targets include
  - mRNA, which codes for proteins, as well as a number of functional non-coding RNAs.
  - NcRNAs almost always function as ribonucleoprotein complexes and not as naked RNAs.

  ✓ non-coding RNAs include

  - ✓ microRNAs,
  - ✓ small interfering RNAs (siRNA), as well as
  - ✓ splicesomal small nuclear RNAs (snRNA).

Protein–RNA interactions

- RNA-binding proteins exhibit highly specific recognition of their RNA targets by recognizing their sequences and structures.
- Specific binding of the RNA-binding proteins allow them to distinguish their targets and regulate a variety of cellular functions via control of the generation, maturation, and lifespan of the RNA transcript.
- This interaction begins during transcription as some RBPs remain bound to RNA until degradation whereas others only transiently bind to RNA to regulate RNA splicing, processing, transport, and localization. In this section, three classes of the most widely studied RNA-binding domains will be discussed
  - ✓ **RNA-recognition motif,**
  - ✓ **double-stranded RNA-binding motif,**
  - ✓ **zinc-finger motif).**

RNA-recognition motif (RRM)

- The RNA recognition motif, which is the most common RNA-binding motif, is a small protein domain of 75–85 amino acids that forms a four-stranded β-sheet against the two α-helices. This recognition motif exerts its role in numerous cellular functions, especially
  - ✓ **in mRNA/rRNA processing, splicing, translation regulation, RNA export, and RNA stability.**
- Ten structures of an RRM have been identified through NMR spectroscopy and X-ray crystallography. These structures illustrate the intricacy of protein–RNA recognition of RRM as it entails RNA–RNA and protein–protein interactions in addition to protein–RNA interactions.
- Despite their complexity, all ten structures have some common features. All RRMs' main protein surfaces' four-stranded β-sheet was found to interact with the RNA, which usually contacts two or three nucleotides in a specific manner. In addition, strong RNA binding affinity and specificity towards variation are achieved through an interaction between the inter-domain linker and the RNA and between RRMs themselves. This plasticity of the RRM explains why RRM is the most abundant domain and why it plays an important role in various biological functions

Double-stranded RNA-binding motif

- The double-stranded RNA-binding motif (dsRM, dsRBD), a 70–75 amino-acid domain, plays a critical role in RNA processing, RNA localization, RNA interference, RNA editing, and translational repression.
- All three structures of the domain solved as of 2005 possess uniting features that explain how dsRMs only bind to dsRNA instead of dsDNA.
- The dsRMs were found to interact along the RNA duplex via both α-helices and β1-β2 loop. Moreover, all three dsRBM structures make contact with the sugar-phosphate backbone of the major groove and of one minor groove, which is mediated by the β1-β2 loop along with the N-terminus region of the alpha helix 2.
- This interaction is a unique adaptation for the shape of an RNA double helix as it involves 2'-hydroxyls and phosphate oxygen.
- Despite the common structural features among dsRBMs, they exhibit distinct chemical frameworks, which permits specificity for a variety for RNA structures including stem-loops, internal loops, bulges or helices containing mismatches.

Zinc fingers

- CCHH-type zinc-finger domains are the most common DNA-binding domain within the eukaryotic genome. In order to attain high sequence-specific recognition of DNA, several zinc fingers are utilized in a modular fashion. Zinc fingers exhibit ββα protein fold in which a β-hairpin and a α-helix are joined via a Zn^{2+} ion.
- Furthermore, the interaction between protein side-chains of the α-helix with the DNA bases in the major groove allows for the DNA-sequence-specific recognition. Despite its wide recognition of DNA, there has been recent discoveries that zinc fingers also have the ability to recognize RNA.
- In addition to CCHH zinc fingers, CCCH zinc fingers were recently discovered to employ sequence-specific recognition of single-stranded RNA through an interaction between intermolecular hydrogen bonds and Watson-Crick edges of the RNA bases. CCHH-type zinc fingers employ two methods of RNA binding.
- First, the zinc fingers exert non-specific interaction with the backbone of a double helix whereas the second mode allows zinc fingers to specifically recognize the individual bases that bulge out. Differing from the CCHH-type, the CCCH-type zinc finger displays another mode of RNA binding, in which single-stranded RNA is identified in a sequence-specific manner. Overall, zinc fingers can directly recognize DNA via binding to dsDNA sequence and RNA via binding to ssRNA sequence.

Thank you!
ARTIE.COM