Chapter 12
Bioinformatics for Legume Genomics Research

Vinay Kumar Singh, A.K. Singh, Arvind M. Kayastha, and B.D. Singh

Abstract Enormous legume genome sequence data are becoming available at a rapid rate through the Next-Gen Sequencing platforms. One of the biggest problems relates to management and analysis of the huge data derived from whole genome sequencing projects. To resolve this problem, researchers index their data in major biological depository systems and availability of algorithms, tools, softwares and databases and provide opportunities for analysis, annotation, and visualization of sequence data at the computational level. Different types of tools and softwares are available for the interpretation of genomes, proteomes and genes. Now researchers are using various *in-silico* techniques in *Bio-omics* (genomics, proteomics, metabolomics and transcriptomics) era for management, planning and prediction of data in cost effective and less time consuming manner. *Bio-omics* plays an important role in comparative, structural and functional biology at computational level and will play major role in different biological investigations. Identification of signal transduction pathway-associated members and gene family members will help in functional elucidation and relationship among them. In this context identification of potential candidate genes will provide an opportunity to researchers for improvement and nutritional quality enhancement of crop genomes. Based on genome blue-prints (plants, animals, fungus, microbes) one can develop potential applications to understand systems biology of legumes in fullness.

V.K. Singh, M.Sc. (✉)
Faculty of Sciences, Centre for Bioinformatics, School of Biotechnology,
Banaras Hindu University, Varanasi, Uttar Pradesh 221005, India
e-mail: vinaysingh@bhu.ac.in

A.K. Singh, Ph.D.
Department of Genetics and Plant Breeding, Institute of Agricultural Sciences, Banaras Hindu University, Varanasi, Uttar Pradesh, India

A.M. Kayastha, Ph.D. • B.D. Singh, Ph.D.
Faculty of Sciences, School of Biotechnology, Banaras Hindu University,
Varanasi, Uttar Pradesh, India

S. Gupta et al. (eds.), *Legumes in the Omic Era*, DOI 10.1007/978-1-4614-8370-0_12,

Keywords Bioinformatics • Sequence analysis • Functional annotation • Comparative mapping • Evolutionary biology • Food legume • Molecular markers • Sequence database • In silico analysis

Introduction

The complete genome has been sequenced in three legume species namely, *Medicago truncatula*, *Lotus japonicus* and soybean (*Glycine max*) (Bertioli et al. 2009; Cannon et al. 2009; Sato et al. 2008; Zhu et al. 2005; Schmutz et al. 2010). Among these, *M. truncatula* is considered as model species, and is taxonomically more related to cool-season legumes such as pea, lentil, faba bean, and chickpea (Bordat et al. 2011). Integrating the genomic and biological knowledge from model legumes to other economically important cool-season pulse crops, e.g., pea, lentil, and chickpea, warm-season food legumes, e.g., peanut and common bean, and forage legumes, e.g. alfalfa and clover, will provide a major opportunity for advancing their genomic resources (Young et al. 2005; Young and Udvardi 2009; Varshney and May 2012). For example it can foster gene identification in such species, which are less noticeable due to their large genomes (Gepts et al. 2005). Sequencing of other legumes, including common bean (Ramírez et al. 2005; David et al. 2008) is progressing rapidly and draft genome sequences of some of them like pigeonpea (Varshney et al. 2009, 2011; Singh et al. 2012) and chickpea (Garg et al. 2011; Varshney et al. 2013) are already available.

Various genome sequencing projects have produced a wealth of sequence data, which need to be properly analysed to enable prediction of the potential functional elements, genes and transcription factors. Rapid progress has been made to develop bioinformatics tools and databases for such analyses as well as for understanding of the various features of the sequenced genome (Kushwaha et al. 2008; Dutt et al. 2010; Kumari et al. 2010). Similarly, *in-silico* comparative genomics provides a great opportunity in unravelling the behaviour of genes and genomes (Udvardi 2002; Kushwaha et al. 2012). Comparative genomics uses information about signature parts at the gene level and syntenic relation at the genome level to understand the structure and function of a newly sequenced genomes, as well as to deduce its evolutionary relationships (Goffard and Weiller 2006). Gene hunting is another important application of comparative genomics to investigate coding and non-coding functional elements of the genome (Yadav et al. 2007; Kushwaha et al. 2011). It attempts to discover both similarities and differences in the genes, proteins, RNA, and regulatory regions of different organisms to infer structural and functional relationships. Comparative genomics is now focusing on discovery of regulatory regions and siRNA molecules in the genome. The available biological datasets in web repository databases allow for comparative analysis and real data validation with the existing datasets. Different databases maintained by a data model like NCBI are integrated with each other to enable their effective utilization. The experimental datasets thus give us opportunities to understand the functional and biological roles

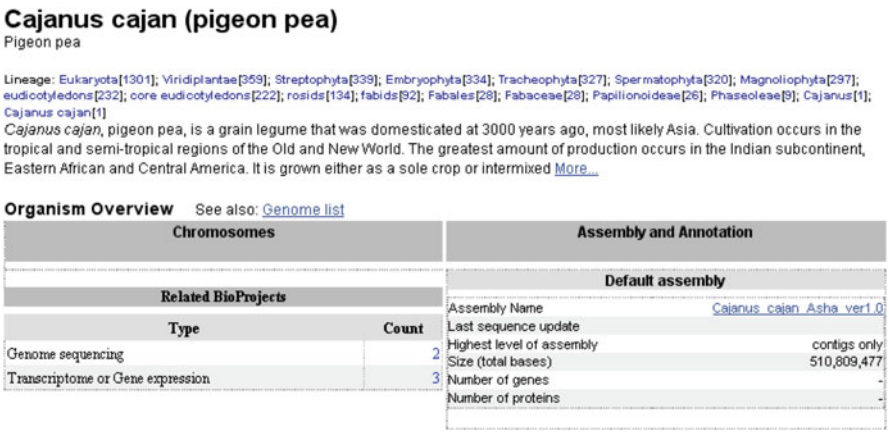
Table 12.1 Important biological databases related to legumes

Database name	URL
NCBI	http://www.ncbi.nlm.nih.gov/
DNA Data Bank of Japan DDBJ	www.ddbj.nig.ac.jp/
EMBL	www.ebi.ac.uk/em
United States Dry Bean Council (USDBC)	http://www.usdrybeans.com/
International Legume Database and Information Service (ILDIS)	http://www.ildis.org/
Legumes information System	http://www.comparative-legumes.org/
Legume "Phylo-informatics" dbase	http://www.public.asu.edu
Food Legume genome database	http://www.gabcsfl.org/
SoyBase	http://soybase.org
Medicago truncatula	http://www.medicago.org/
Illustrated Legume Genetic Resources Database	www.gene.affrc.go.jp
SSR Database of legumes	http://intranet.icrisat.org/gt1/ssr/ssrdatabase.html
Bioinformatics resources for legume researchers	http://www.legumes.org/
Chinese Legume Database and Information Service (CLDIS)	http://cldis.ibcas.ac.cn/
LegumeTFDB	http://legumetfdb.psc.riken.jp/
Lotus japonicus	http://www.kazusa.or.jp/lotus/
Phytozome v7.0	http://www.phytozome.net/
Chickpea Transcriptome Database	http://59.163.192.90:8080/ctdb/
Chickpea Root EST Database	http://www.icrisat.org/what-we-do/biotechnology/Cpest/home.asp
Gramene	http://www.gramene.org/
GmGDB	http://www.plantgdb.org/GmGDB/
Lotus japonicus genome DB	http://www.kazusa.or.jp/lotus/
Legume Information System	http://www.comparative-legumes.org/
Common Bean Database	http://jeff.ifxworks.com/Legume/common_bean.html

of unknown genes/proteins from different legumes. The availability of different biological databases related to legumes provides valuable information resource for research and analysis (Table 12.1). However, the main aim of bioinformatics is the identification of regulatory mechanisms and function of genomes and their evolution (Marla and Singh 2012).

Bioinformatics for Legume Genome Annotation

Sequencing determines the primary structure of an unbranched biopolymer. The elements with the associated function can be predicted by using DNA/protein sequences. Sequencing of a genome is a complicated and typical task that uses DNA sequencing to determine the order of nucleotides in small DNA fragments that together make up the genome. The first generation DNA sequencing was performed



Cajanus cajan (pigeon pea)
Pigeon pea

Lineage: Eukaryota[1301]; Viridiplantae[359]; Streptophyta[339]; Embryophyta[334]; Tracheophyta[327]; Spermatophyta[320]; Magnoliophyta[297]; eudicotyledons[232]; core eudicotyledons[222]; rosids[134]; fabids[92]; Fabales[28]; Fabaceae[28]; Papilionoideae[26]; Phaseoleae[9]; Cajanus[1]; Cajanus cajan[1]

Cajanus cajan, pigeon pea, is a grain legume that was domesticated at 3000 years ago, most likely Asia. Cultivation occurs in the tropical and semi-tropical regions of the Old and New World. The greatest amount of production occurs in the Indian subcontinent, Eastern African and Central America. It is grown either as a sole crop or intermixed More...

Organism Overview See also: Genome list

Chromosomes	Assembly and Annotation

Related BioProjects

Type	Count
Genome sequencing	2
Transcriptome or Gene expression	3

Default assembly

Assembly Name	Cajanus_cajan_Asha_ver1.0
Last sequence update	
Highest level of assembly	contigs only
Size (total bases)	510,809,477
Number of genes	-
Number of proteins	-

Fig. 12.1 An Example of pigeonpea (*C. cajan*) genome sequence deposited in NCBI by a group of Indian scientists [Reprinted from Singh N. K., Gupta D. K., Jayaswal P. K., Mahato A.K., Dutta S., Singh S., Bhutani S., et al. (2012) The first draft of the pigeonpea genome sequence. J. Plant Biochem Biotechnol 21: 98–112 with permission from Springer Science+Business Media]

by using the chain termination method developed by Frederick Sanger and co-workers (Sanger and Coulson 1975; Sanger et al. 1977). This technique uses sequence-specific termination of a DNA synthesis reaction using modified nucleotide substrates. However, new sequencing technologies such as pyrosequencing are gaining an increasing share of the sequencing work and the next generation DNA sequencers that achieve sequencing by synthesis are based on this approach. These sequencer do not require *in vivo* library construction, are faster and much cheaper to use; they are being used for rapid genome sequencing. An example of nearly completed *C. cajan* genome sequenced by a group of Indian scientists using the second generation DNA sequencers is depicted in Fig. 12.1.

After completion of the full genome sequence, it is necessary to assemble and annotate new sequences. In fact, genome assembly is a very difficult computational task owing to large numbers of identical sequences (repeats) found in genomes. These repeats can be of thousands of nucleotides in length, and some of them may occur in a number of different locations. In a shotgun sequencing project, the entire DNA from a source (usually a single organism, ranging from a bacterium to a mammal) is first fragmented into millions of small pieces. These pieces are then "read" by automated sequencers, and each read can be up to 1,000 nucleotides long. A genome assembly algorithm works by taking all the reads and aligning them with one another, to detect all the places where two of the reads are overlapping. These overlapping reads can be merged together to form a contig and then linking information of contigs is used to create scaffolds. Subsequent to this, scaffolds are positioned along the physical map of the chromosomes.

Most of the assembler tools and packages were developed by different research groups, e.g., short oligonucleotide analysis package and *de novo* assembly tools were developed by Beijing Genomics Institute (BGI).

Table 12.2 Bioinformatics softwares available for genome annotation and *de novo* assembly

Application	Available tools
Genome annotation	TRF, Repeat Masker, Genescan, BGF, InterproScan etc.
De-novo assembly	SOAP *de-novo*, AbySS, Velvet etc.
Genome resequencing analysis	SOAPsnp\SOAPsv\SOAPInDel, SAMtools, BreakDancer, VarScan etc.

In genome annotation one can elucidate the biological information based on assembled genome sequences. In this process, called "gene prediction", one can identify functional elements in the genome and generate biological information about these elements. The genome annotation is done by the methods prescribed by Kawaji and Hayashizaki (2008). The basic level of genome annotation can be done using Basic Local Alignment Search Tool BLAST to find out similarities and differences. However, nowadays more and more additional information is added to the annotation platform. The complete annotated genome data are deposited in different biological databases, i.e., NCBI, DDBI, Phytozome, Ensembl and EMBL. These databases use genome context information, experimental datasets, and integrations of tools and resources to provide gene and genome annotations through their subsystems approach. Sequence Assembly AMOS tool can be used for manipulation with sequence files. AMOS tool is currently maintained by University of Maryland. CABOG is a tool that assembles large genomic DNA sequences produced by whole-genome shotgun sequencing. Some important annotation tools like Apollo, BLAST, Parser, MATLAB, Bioconductor package in R, Artemis and AAT tool are available. Manatee is a web-based gene evaluation and genome annotation tool for visualization, modification and storage for genomes. PASA can be used as eukaryotic genome annotation tool that exploits spliced alignments of expressed transcript sequences to gene model. Several bioinformatics tools are available for annotation, genome sequence alignment, *de novo* assembly, sequence alignments, evolution and RNA sequence analysis; some of these tools are listed in Table 12.2.

Hiremath et al. (2011) carried out a large-scale transcriptome analysis in chickpea (*C. arietinum* L.) using next generation sequencing technologies such as, Roche 454 and Illumina/Solexa. They determined a total of 103,215 tentative unique sequences (TUSs) and assigned functions for 49,437 (47.8 %) of the TUSs. Comparison of the chickpea TUSs with the *M. truncatula* genome assembly (Mt 3.5.1 build) resulted in 42,141 aligned TUSs with putative gene structures (including 39,281 predicted intron/splice junctions). These TUSs were also used to identify 728 SSR, 495 SNP, 387 conserved orthologous sequence (COS) markers, and 2,088 intron-spanning region (ISR) markers. Similarly, transcriptome assembly has been done in pigeonpea by Kudapa et al. (2012) referred to as CcTA v2, comprised 21,434 transcript assembly contigs (TACs) and 77.5 % TACs (16,622 TACs) of the total could be mapped on to the soybean genome. Based on knowledge of intron junctions, so far 10,009 primer pairs were designed from 5,033 TACs for amplifying intron spanning regions (ISRs). By using *in silico* mapping of BAC-end-derived

SSR loci of pigeonpea on the soybean genome as a reference, putative mapping positions at the chromosome level were predicted for 6,284 ISR markers, covering all the 11 pigeonpea linkage groups. The transcript assembly and markers developed will provide a useful resource for basic and applied research for genome analysis and crop improvement in chickpea and pigeonpea.

ORFs and their localization, gene structure optimization, coding region identification and location of regulatory motifs explain the complete organization of gene family with their associated functions. Identification of gene family is a better approach to investigate the various types of members related to each other and the manner in which they have evolved (Thornton and DeSalle 2000). Availability of EST datasets for a genome gives a better understanding of transcripts with tissue-specific expression. Based on bioinformatics tools and databases any one can compare biological experiment datasets with any query sequence. *In-silico* based approaches utilize information from expressed sequence tags and proteins, often derived from mass spectrometry, to improve genomic annotations. A variety of software tools have been developed to help scientists in their quest for gene and genome annotations. Identification of gene locations and the sites of other genetic control elements are often described as the biological "parts list" for the assembly of an organism. Scientists are still at an early stage of delineating this "parts list" and in understanding how all the parts fit together and work together. Gene and genetic control elements investigation can be done using publicly available biological databases and tools accessible *via* the web and other electronic means. Some statistical tools are available for the analysis of deep sequencing like ANDES Tools and DAG chainer that computes chains of syntenic genes within complete genome sequences. DNA sequence analysis tools include k-mer tool, ESTmapper, Snapper mapping reads and ATAC are available for aligning genomes. For rapid aligning of the entire genomes, a software MUMmer, can be used.

Bioinformatics for Sequence Analysis

In bioinformatics, sequence analysis refers to the process of subjecting a DNA, RNA or protein sequence using analytical methods and algorithms to understand its features, function, structure, or evolution. Methodologies used are biological database mining, comparative analysis and sequence alignment. With the development of statistical algorithm, matrices based tools for prediction of gene and protein sequences, the rate of addition of new sequences to the databases has increased exponentially. Such a collection of sequences does not, by itself, increase the scientist's understanding of the biology of organisms. However, comparing these new sequences to those with known functions is a key way of understanding the biology of an organism from which the new sequence comes. Thus, sequence analysis can be used to assign functions to genes and proteins by a study of the similarities between the compared sequences. Nowadays, there are many tools and techniques are available that provide the sequence comparisons (sequence alignment) and analyze the alignment

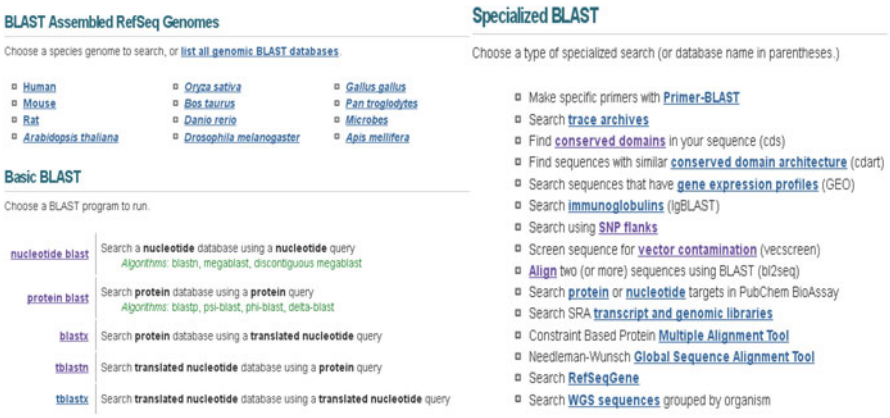


Fig. 12.2 A page showing basic local alignment search tool (BLAST; http://blast.ncbi.nlm.nih.gov/)

of a product to understand its biology. Sequence analysis in molecular biology includes a wide range of applications, some of which are listed below.

1. Comparison of different sequences in order to detect similarities among them and, often, to infer if the sequences are related (homologous).
2. Identification of intrinsic features of the different sequences, such as active sites, post-translational modification sites, gene structures, reading frames, distributions of introns and exons and the regulatory elements.
3. Identification of sequence differences and variations such as point mutations and single nucleotide polymorphisms (SNPs) in order to develop the genetic markers.
4. Unraveling the evolutionary process and assessment of genetic diversity of the sequences and the organisms.
5. Identification of molecular structure from sequence data alone.

Sequence analysis is based on sequence alignment, i.e., comparison between query and subject sequences, in which two or more sequence sets can participate. Alignment between two sequences is called pairwise alignment, and alignment between more than two sequences is called multiple sequence alignment. Two methods are used for searching for a series of identical or similar characters in the sequences to find out similarities and dissimilarities within sets of sequences; these are called global and local alignments. Global alignment finds the best alignment across the whole length of two sequences and forces alignment in such regions that show differences. Local alignment finds regions of high similarity in parts of the participating sequences, and concentrates on regions of high similarity. Basic local alignment search tool (BLAST) is an example of local alignment (Fig. 12.2). Mainly five flavors of Basic BLAST are available for comparison of the query with the subject for sequence. In case of protein query sequence, one can use BLASTp and tBLASTn. In case of nucleotide query sequence, any one of the BLASTn, BLASTx and tBLASTx can be used. Other specialized blasts are also available for conserved domain detection, SNP detection, global sequence alignment, etc.

Gene Identification and Characterization Using Comparative Genomics/Proteomics

In computational biology gene hunting or gene prediction refers to the process of identifying the regions of genomic DNA that function as genes, i.e., encode proteins or various types of RNA molecules, or as other functional elements like regulatory regions. Gene finding is one of the first and most important steps in understanding the genome of a species once it has been sequenced. Earlier "gene finding" was based on cumbersome experiments on living cells and organisms. But the availability of comprehensive genome sequences and powerful computational resources have greatly facilitated gene finding, and some of the tools and database servers dedicated to gene prediction are listed in Table 12.3.

Genome sequence of "Asha" variety of pigeonpea was obtained using GS-FLX Phase D chemistry and the GS-FLX Titanium chemistry and reads were assembled

Table 12.3 A list of some important gene prediction servers

Name	Description/function
ATGpr	Identifies translational initiation sites in cDNA sequences
AUGUSTUS	Predicts genes in eukaryotic genomic sequences
BGF	Hidden Markov model based *ab initio* gene prediction program
EUGENE	Gene hunting for *Arabidopsis thaliana*
FRAMED	Finds genes and frameshift in G + C rich prokaryotic sequences
GENIUS	For linking predicted genes in complete genomes to known protein 3D structures
GENEID	Signal, exon and gene prediction server
GENEPARSER	Detect intron and exon regions in DNA sequence
GeneMark	Family of gene prediction programs
GeneMark.hmm	A gene prediction program for prokaryotes and eukaryotes
GeneTack	Prediction of genes with frameshifts in prokaryotic genomes
NIX	Web tool gene prediction based on combining results from different programs
GLIMMER	For finding genes in microbial DNA
VEIL	Hidden Markov model for finding genes in vertebrate DNA Server
Splice Predictor	Identifies potential splice sites in (plant) pre-mRNA using Bayesian methods
GENESCAN	For finding genes using Fourier transform
FGENESH	The fastest and most accurate *ab initio* gene prediction program
NNPP	Promoter prediction by neural network
NNSPLICE	Splice site prediction using neural network method
GENOMESCAN	Predicts locations and exon-intron boundary in genomic sequences
ORF FINDER	A graphical analysis tool for open reading frame prediction
GrailEXP	Predicts exons, genes, promoters, poly-As, CpG islands and repetitive elements within DNA sequences
EuGène	Gene finder for eukaryotic system exploits probabilistic models for discriminating coding from non-coding sequences to discriminate effective splice sites from false splice sites

using "Newbler GS De Novo assembler version 2.5.3" that compares all sequence reads pairwise and reads with overlaps are joined into contigs (Singh et al. 2011). An average of all aligned reads at a specific nucleotide position is used to determine the consensus sequences for a contig, and overlapping contigs are finally merged to make scaffolds. The finished sequence was passed through fgenesh tool of Molquest software using *Arabidopsis thaliana* gene models as a reference. Predicted genes with size of >500 bp were BLAST-searched against the NCBI database, and the search output was processed using BLAST Parser software and gene annotations were manually curated and categorized based on function. Singh et al. (2012) were able to predict a total of 59,515 genes with the largest size of 11,523 bp and the smallest gene size of 501 bp of these 47,004 were protein coding genes of which 1,213 were related with plant defense and 152 were involved in abiotic stress tolerance.

Comparative phylogenetic studies within the legume family revealed high syntenic relationships between sequenced legumes and other important legumes (Wojciechowski et al. 2004), e.g. between *Medicago truncatula* and pea (Kaló et al. 2004), and common bean and soybean (Lee et al. 2001), but limited synteny is also reported to be present among other legumes, e.g., between cool-season and warm-season legumes (Zhu et al. 2005). Whole genome sequencing of some important legumes is likely to be completed in the near future, and this will facilitate a comprehensive assessment of synteny. Comparative genomics for synteny studies can accelerate exploitation of genomic resources, and facilitate more rapid progress in research efforts in an efficient and cost-effective manner. A detailed study of the syntenic relationships is a critical issue to be addressed for better allocation of genomic information from sequences of model legumes to other legumes and to other crop species. Based on conservation of synteny between pigeonpea and soybean genomes, Singh et al. (2012) found that chromosomes 1, 3, 4 and 9 of pigeonpea showed the maximum conservation with chromosomes 2, 5, 7, 8, 12, 13, 15 and 17 of soybean. Chromosome 1 of pigeonpea showed the highest number of matches with chromosomes 8 and 5 of soybean. Similarly, chromosome 2 of pigeonpea showed the maximum number of hits with chromosomes 19 and 10 of soybean. Pigeonpea chromosome 3 showed the maximum number of hits with chromosomes 13 and 15 of soybean, pigeonpea chromosome 4 showed the maximum number of hits with chromosomes 12 and 13 of soybean, chromosome 5 showed the highest number of matches with chromosomes 13, 12 and 17 of soybean, chromosome 6 showed the maximum number of matches with chromosomes 9 and 3 of soybean, chromosome 9 showed maximum number of matches with chromosomes 2, 12, 3, 11 and 16 of soybean, chromosome 10 showed the maximum number of hits with chromosomes 18, 17 and 2 of soybean, chromosome 11 showed the maximum numbers of hits with chromosomes 14 and 18 of soybean, and chromosome 7 showed maximum number of hits with chromosomes 10 and 20 of soybean, while chromosome 8 of pigeonpea showed minor synteny with chromosomes 13 and 14 of soybean. However, Singh et al. (2012) concluded that the overall synteny between the genomes of pigeonpea and soybean was only to a limited extent.

Bioinformatics for Computational Evolutionary Biology

The phylogenetic tree (phylogeny) is textual and visual representation that describes evolutionary relationships among various groups of organisms or among a family of related nucleotide or protein sequences and other entities based upon similarities and differences in their physical and genetic characteristics. In such a study, one can use morphological features (e.g., shape, size, length, etc.) and molecular data (e.g., DNA and protein sequences). The taxa/entities joined together in the tree are implied to have descended from a common ancestor. Phylogenetic trees are useful in fields of bioinformatics, systematics and comparative biology. There are rooted and unrooted types of tree inferences and main approaches for phylogeny reconstruction, i.e., distance based methods, topology search methods and Bayesian methods. Some phylogenetic tree terminologies are shown in Fig. 12.3.

A rooted phylogenetic tree defines common ancestor of all the entities at the leaves of the tree, i.e., the operational taxonomic units (OTUs). One example showing root based phylogenetic classification of Toll interleukin 1 receptor (TIR) domain among different organisms depicts the way this family might have been derived during evolution (Fig. 12.3). Phylogenetic relationships among genes can help to predict the genes that might have similar function e.g. *ortholog detection*.

TIR domain is mainly involved in plant immune responses against various pathogens. An example of Toll/interleukin-1 receptor classification is provided here TIR domain for *C. cajan* was used for find out similar homologues in different organisms using basic local alignment search tool (BLAST). Selected homologues from different species were used for multiple sequence alignment and phylogenetic

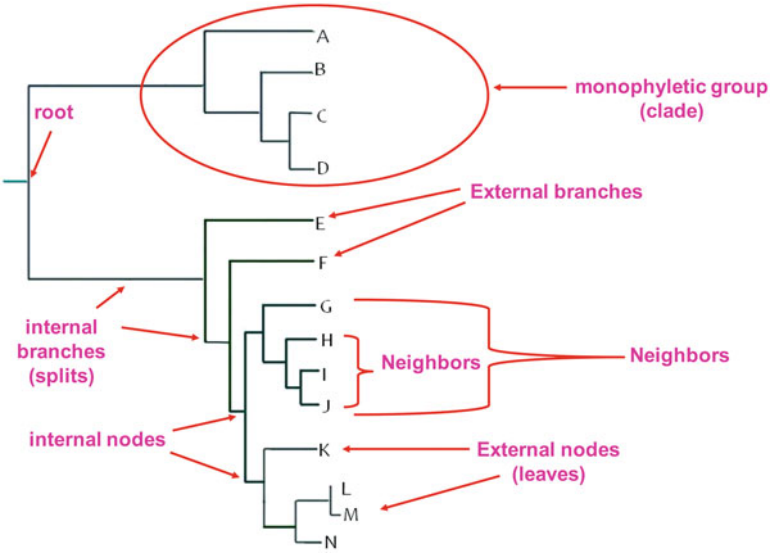


Fig. 12.3 Figure showing phylogenetic tree terminologies

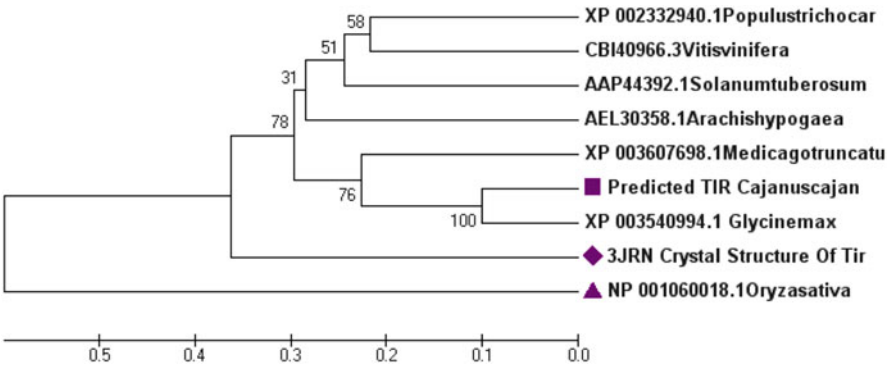


Fig. 12.4 Example of rooted tree of TIR domain homologues from *C. cajan* with six other plant species (Singh et al., unpublished data)

classification. ClustalW tool was used for multiple sequence alignment and for tree classification, MEGA tool was used to find out the best tree topology. Figure 12.4 shows the rooted inferences of selected sequences of TIR domains from seven different plant species (*Populus*, *Vitis*, *Solanum*, *Arachis*, *Medicago*, *Glycine*, *Cajanus* and *Oryza*). Interestingly, it was found that TIR, *Oryza* spp. forms an outer group, while the remaining six TIR domains are much more closely related this may be expected because *Oryza* is a monocot.

The identified TIR domain from *C. cajan* was further used to determine the number of TIR loci present in the *Cajanus* genome, and a total of 148 TIR domains have been successfully identified based on the available datasets of *C. cajan* genome sequence (Taxid: 3821). Figure 12.5 shows an unrooted tree depicting the various TIR domains derived from Cajanus genome itself. Unrooted trees specify relationships but they do not depict the evolutionary path. For phylogenetic study, different online and offline softwares are available (Table 12.4). Legume diversity and evolution in a phylogenetic context has been reviewed earlier by Doyle and Luckow (2003).

In-Silico Analysis for Gene Expression Data

An expressed sequence tag (EST) is a short, ordinarily, terminal sequence of a cDNA sequence. Thus an EST results from one-shot sequencing of a cloned mRNA, i.e., several hundred base pairs of sequence starting from an end of a cDNA sequence. The cDNAs used for EST generation are typically individual clones from a cDNA library. ESTs may be used to identify gene transcripts; they are instrumental in gene discovery and gene sequence determination. The identification of ESTs has proceeded rapidly, and ~73 million ESTs are now available in the public database GenBank. The dbEST is a division of Genbank established in 1992, and the data in dbEST is directly submitted by laboratories worldwide. Based on EST

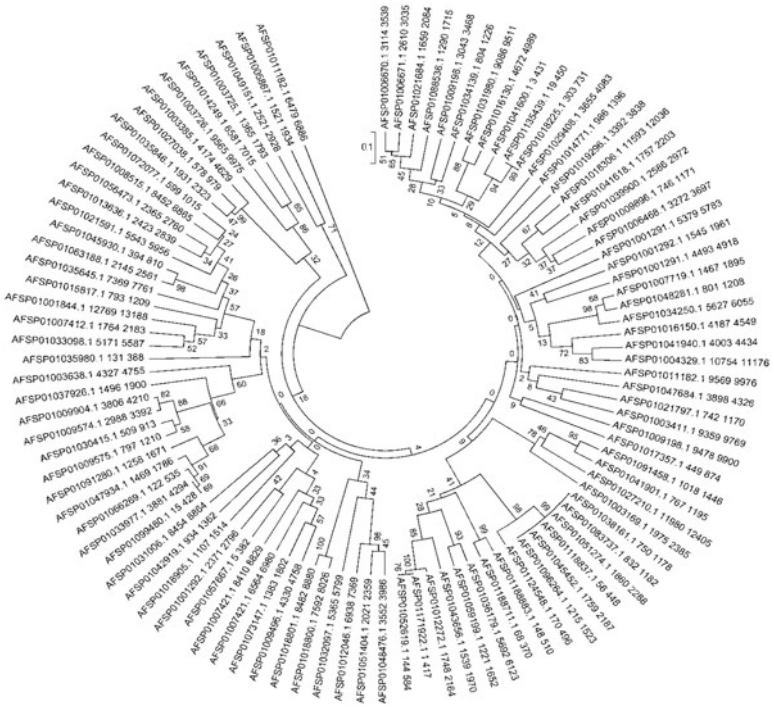


Fig. 12.5 Example of unrooted tree of identified TIR domains from *C. cajan*

Table 12.4 Tools and servers for multiple sequence alignment and phylogenetic analysis

Tools and server	URL
ClustalW2	http://www.ebi.ac.uk/Tools/msa/clustalw2/
CLUSTALW	http://www.genome.jp/tools/clustalw/
MEGA	http://www.megasoftware.net/
T-Coffee	http://www.ebi.ac.uk/Tools/msa/tcoffee/
PHYLIP	http://evolution.genetics.washington.edu/phylip.html
The PhylOgenetic Web Repeater (POWER)	http://power.nhri.org.tw/power/home.htm
BlastO	http://oxytricha.princeton.edu/BlastO/
BIONJ	http://mobyle.pasteur.fr/cgi-bin/portal.py?#forms::bionj
DendroUPGMA	http://genomes.urv.cat/UPGMA/
PhyML	http://www.atgc-montpellier.fr/phyml/binaries.php
Evolutionary Trace Server (TraceSuite II)	http://mordred.bioc.cam.ac.uk/~jiye/evoltrace/evoltrace.html
Phylogeny.fr	http://www.phylogeny.fr/
Mesquite	http://mesquiteproject.org/mesquite/mesquite.html
Winboot	http://archive.irri.org/science/software/winboot.asp

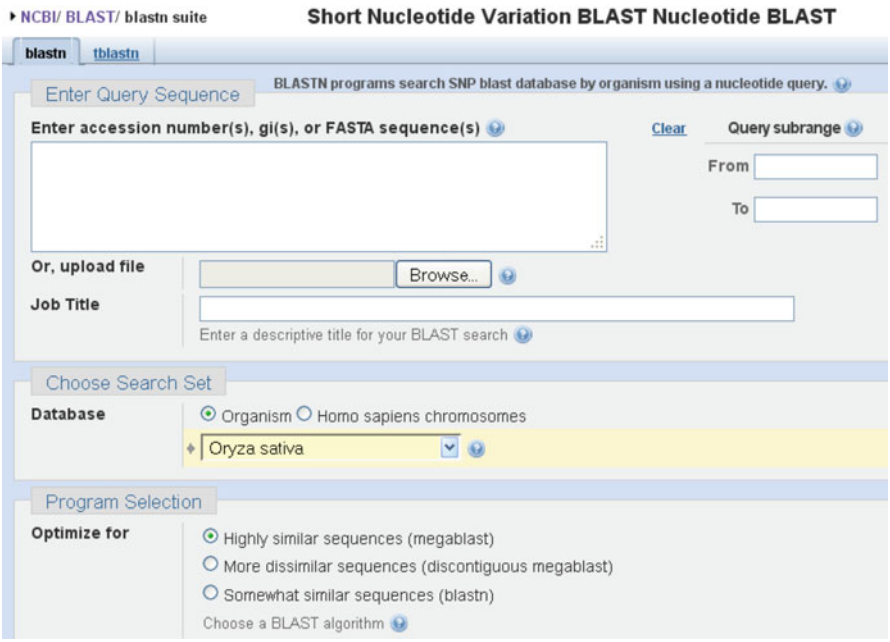


Fig. 12.6 Short nucleotide variation BLAST page

datasets any one can determine the gene function based on expression datasets. ESTs contain enough information to permit the design of precise probes for DNA microarrays that can be used to determine the gene expression. For expression microarray data analysis normalization and management, one can use Ginkgo (Comparative Genomic Hybridization package). TM4 and Magnolia packages are also designed for microarray data management for researchers who use PFGRC microarrays. The programme SNP Filter Scripts can be used to identify and detect false positive SNP calls that are present in raw data from affymetrix gene chip resequencing arrays. There are several other tools freely available, including MAGIC, CLUSFAVOUR, etc. for microarray data analysis. Short nucleotide variation analysis server is also available for this type of study (Fig. 12.6).

Bioinformatics in Legume Nutritional Genomics

By manipulating the promoter region of seed-specific protein encoding genes one can improve the nutritional quality of any crop species. Bioinformatics tools can play a major role in the study of the promoter region of genes and for identification of *cis*-acting elements or *cis-regulatory* elements. A *cis*-acting element is a

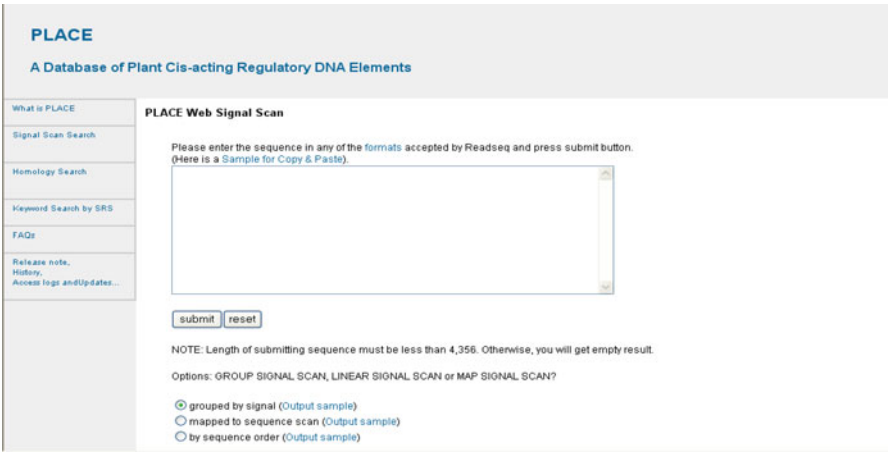


Fig. 12.7 Plant *cis*-acting elements prediction server (PLACE; http://www.dna.affrc.go.jp/PLACE/)

region of DNA or RNA that regulates the expression of genes located in the same chromosome. This term is derived from the Latin word *cis*, which means "on the same side as". The *cis*-regulatory elements are often binding sites for one or more *trans-acting* factors. These *cis*-elements may be located upstream of the coding sequences of the concerned genes, i.e., in the promoter region or even further upstream, in an intron, or downstream of the gene's coding sequence. In molecular biology and genetics, a transcription factor (sometimes called a sequence-specific DNA-binding factor) is a protein that binds to specific DNA sequences, thereby controlling the flow of genetic information (or transcription) from DNA to mRNA. Transcription factors perform this function alone or with other proteins in a complex, by promoting (as an activator)/or blocking (as a repressor) the recruitment of RNA polymerase to transcribe specific genes. Therefore, identification of potential *cis*-acting elements can help in improving the nutritional quality of seeds of plant species, and/or other traits of economic/agronomic value.

Databases of plant *cis*-acting regulatory elements like PlantCare and PLACE can be used as a portal for *in-silico* analysis of promoter sequences of plant genes (Fig. 12.7). Yadav et al. (2007) successfully identified the seed storage protein promoter specific *cis*-acting elements in cloned and sequenced promoter regions of seed storage protein genes from different cultivars of wheat, rice and oat. A database containing collection of proximal promoter sequences for RNA polymerase II with experimentally determined transcription start-sites from various plant species is available on server PlantProm DB. For retrieval and investigation of transcription factor associated genes PlnTFDB (plntfdb.bio.uni-potsdam.de/) and PlantTFDB (http://planttfdb.cbi.pku.edu.cn/) are important databases. In addition, species transcription factor databases are also available online (Fig. 12.8).

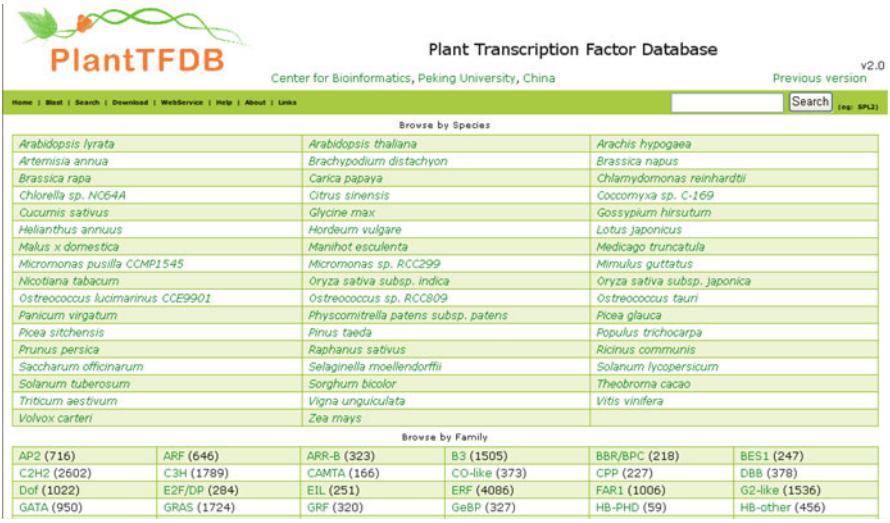


Fig. 12.8 Plant transcription factor database PlantTFDB (http://planttfdb.cbi.edu.cn/)

Prediction for Function of Protein Sequences

In the prediction of the function of a protein sequence of interest, structural visualization, 3D prediction, classification and structural alignment play important roles. In this connection homology modeling, threading and *ab-initio* prediction methods can be used for protein structure prediction. Homology modeling (comparative modeling) is a process for constructing an atomic-resolution model of the "target" protein using an experimental three-dimensional structure of a related homologous protein (the "template") derived by NMR, X-ray techniques. Homology modeling relies on the identification of one or more known protein structures likely to resemble the structure of the query sequence, and on the production of an alignment that maps residues in the query sequence to residues in the template sequence. It has been shown that protein structures are more conserved than the amino acid sequences amongst homologues, but sequences falling below 20 % sequence identity can have very different structures. For homology modeling, threading and *ab-initio* prediction several servers are available in public domain (Table 12.5). Some commercial software like MOE, Schrödinger and Discovery Studio can also be used for protein modeling and simulation. For *Ab-initio* or *de-novo* protein modeling one can use I-TASSER and ROBETTA, which are freely available. Based on different protein modeling servers, one can predict the three dimensional structure of the target protein.

Table 12.5 List of servers for homology modeling, threading and *ab-initio* based structure prediction

Server name	Description	URL
SWISS-MODEL	Automated protein structure homology-modeling server	http://swissmodel.expasy.org/
ModBase	Comparative modeling based on three-dimensional protein models. The models are derived by ModPipe, an automated modeling using PSI-BLAST and MODELLER	http://modbase.compbio.ucsf.edu/modbase-cgi/index.cgi
I-TASSER	Model is built based on multiple-threading alignments by LOMETS and iterative TASSER simulations	http://zhanglab.ccmb.med.umich.edu/I-TASSER/
LOMETS	3D model prediction by collecting high-scoring target-to-template alignments using threading programs (FUGUE, HHsearch, MUSTER, PPA, PROSPECT2, SAM-T02, SPARKS, SP3)	http://zhanglab.ccmb.med.umich.edu/LOMETS/
ESyPred3D	Homology modeling web by combining, weighting and screening the results of several multiple alignment programs using the modeling package MODELLER	http://www.fundp.ac.be/sciences/biologie/urbm/bioinfo/esypred/
3D-Jigsaw	Automated system to build three-dimensional models for proteins based on homologues of known structure	http://bmm.cancerresearchuk.org/~3djigsaw/
HMMSTR/Rosetta	Predicts the structure of proteins from the sequence: secondary, local, super secondary, and tertiary. Provided by the Depts. of Biology and Computer Science, Rensselaer Polytechnic Institute	http://www.bioinfo.rpi.edu/bystrc/hmmstr/server.php
Geno3D	Protein three-dimensional structure using comparative protein structure modeling by spatial restraints (distances and dihedral) satisfaction	http://geno3d-pbil.ibcp.fr/cgi-bin/geno3d_automat.pl?page=/GENO3D/geno3d_home.html
VADAR (Volume, Area, Dihedral Angle Reporter)	Quantitatively and qualitatively assess protein structures determined by 3D-threading or homology modelling	http://vadar.wishartlab.com/
ResProx (Resolution-by-proxy or Res(p))	A web server that predicts the atomic resolution of NMR protein structures using only PDB coordinate data as input	http://www.resprox.ca/
Robetta	*Ab initio* fragment assembly	http://robetta.bakerlab.org/

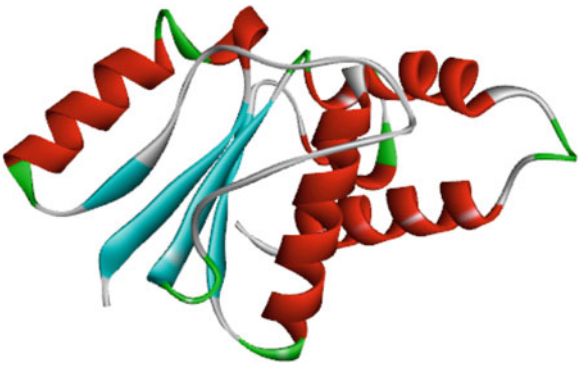


Fig. 12.9 Structure of TIR domain (PM0078097) from *C. cajan* developed using TIR domain structure from *Arabidopsis thaliana* (3JRN) based on homology modelling [Courtesy of Vinay Kumar Singh]

Qualitative and Quantitative Study of Predicted Models

Finally, predicted 3D models can be subjected to a series of tests for assessing their internal consistency and reliability. The Quality of the model can be checked with verify3D [http://nihserver.mbi.ucla.edu/Verify_3D/], Errat [http://nihserver.mbi.ucla.edu/ERRATv2/] etc. The stereochemical properties based on backbone conformation can be evaluated by inspection of Psi/Phi/Chi/Omega angle using Ramachandran plot of PDBSum database [http://www.ebi.ac.uk/pdbsum/], RAMPAGE [http://mordred.bioc.cam.ac.uk/~rapper/rampage.php] etc. Quantitative analysis can be done using accessible surface area prediction using Volume Area Dihedral Angle Reporter [VADAR; http://vadar.wishartlab.com/]. Standard bond lengths and bond angles of the model can be determined using WHAT IF [http://swift.cmbi.ru.nl/whatif/]. ResProx (Resolution-by-proxy; http://www.resprox.ca/) can be used for quality and quantity measurements at resolution level. For example, we have successfully predicted 3D model of toll-like interleukin receptor (TIR) domain of *R* genes from *C. cajan* using comparative homology modeling and the best evaluated model has been deposited to Protein Model DataBase (PMDB; http://mi.caspur.it/PMDB/) (Fig. 12.9).

Integrated Bioinformatics Tools

Some integrated tools like MEME and MAST are useful servers for motif elucidation (Fig. 12.10). For protein functional elucidation and characterization, one can use INTERPROSCAN, PROSITE, PFAM and PRODOM etc. (Fig. 12.11). SWISSPROT, DBSNP and SNP flanks tools and databases can be used for SNP/variant detection. An example of signature part of toll-like interleukin receptor domain from *C. cajan* is given in Fig. 12.12.

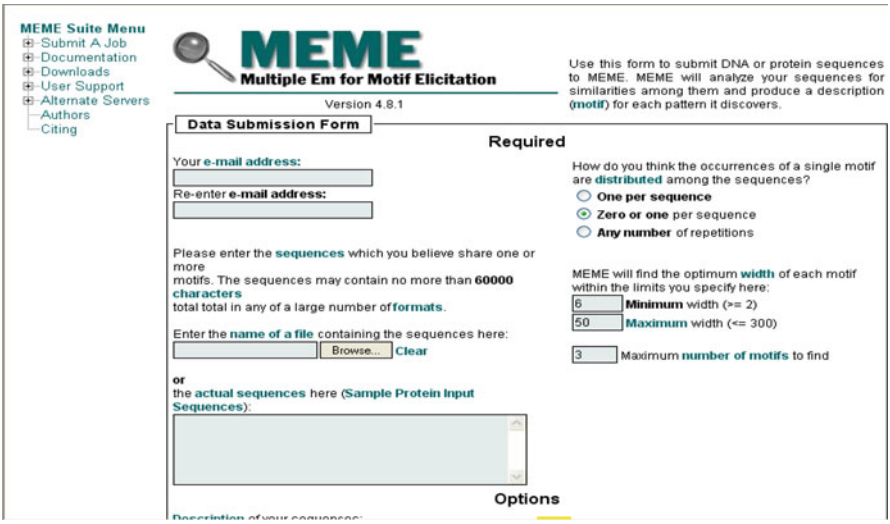


Fig. 12.10 A server to discover motifs (highly conserved regions) in groups of related DNA or protein sequences

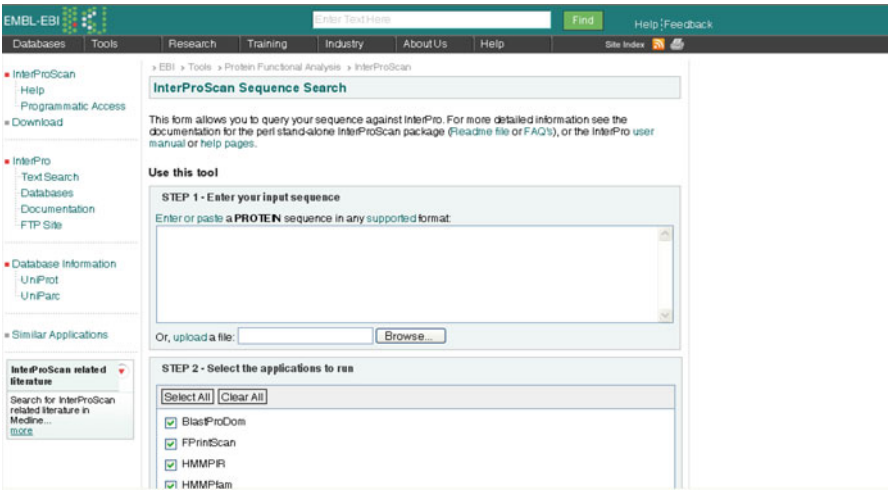


Fig. 12.11 Server for protein functional elucidation based on domain and signature motifs

Molecular Docking

In bioinformatics, molecular docking is a method that predicts the possible orientation of one molecule in relation to a second when the two are bound to each other to form a stable complex. The knowledge of the possible orientations in turn, can be

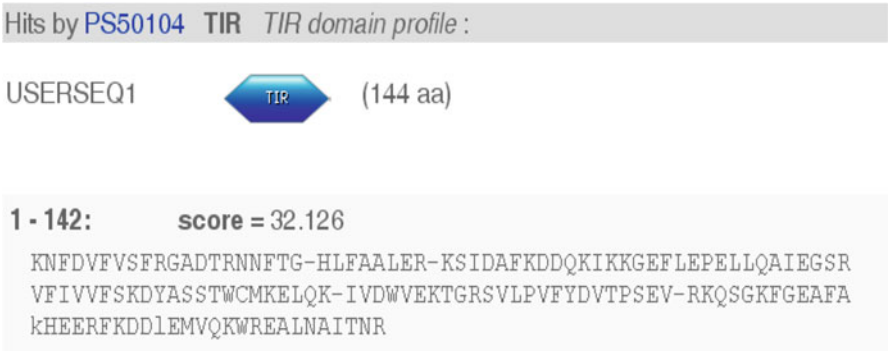


Fig. 12.12 Toll-like interleukin receptor domain form *C. cajan*

used to predict the binding affinity between the two molecules using energy scoring functions. Using molecular docking approach, one can predict the binding orientation with energy total and energy shape of a ligand (small molecule) to its protein target (receptor) to predict the affinity and activity of the small molecule. The interaction between ligand and receptor protein can result in activation or inhibition of the protein enzyme. Two main approaches are the most popular of the different molecular docking strategies. The first strategy uses a matching technique that explains protein and ligand as complementary surfaces. The second approach, however, simulates the actual docking process, in which the ligand–protein interaction energies are calculated. Molecular docking plays an important role in the rational drug designing. For a study of interaction of ligand (inhibitor and cofactor) and protein target one can use HEX, BIOSOLVEIT, DOCKING SERVER and other servers listed in Table 12.6.

Plant–Pathogen Interactions

Many microbes establish wide range of interactions with host plants. Some of these are pathogenic and some are symbiotic in nature. Such interactions involve complex recognition events between the plant and the microbe, leading to a cascade of signalling events and regulation of a number of genes is required for, or associated with, the interaction. The combined components of the transcriptomes of both plant and microorganism that are expressed during the interaction give rise to the term "interaction transcriptome". High-throughput methods to study differential gene transcription, or proteomics coupled with bioinformatics will accelerate our understanding of the molecular bases of plant–microbe interactions (Birch and Kamoun 2000; Samac and Graham 2007). For example, Soria-Guerra et al. (2010) conducted a transcriptome profiling study for soybean rust (*Phakopsora pachyrhizi*) to identify soybean rust resistance genes in *Glycine tomentella*. Among 38,400 genes

Table 12.6 List of servers related to inhibitor, cofactor and protein docking

Server	Description/function	URL
SwissDock	Predicts the molecular interactions between a target protein and a small molecule	http://swissdock.vital-it.ch/
DockingServer	Molecular docking from ligand and protein set-up	http://www.dockingserver.com/web
Blaster	Docking program developed by Pharmaceutical Chemistry Department at the California University	http://blaster.docking.og/
Docking At UTMB	Structure-based virtual screening with AutoDock Vina	http://docking.utmb.edu/
Pardock	Fully automated, all-atom energy based ligand docking	http://www.scfbio-iitd.res.in/dock/pardock.jsp
PPDock	Portal Patch Dock is a web server that can be used to dock drugs to the target proteins	http://140.112.135.49/ppdock/
iScreen	Docking and screening the small molecular database on traditional Chinese medicine (TCM) using the LEA3D genetic algorithm	http://iscreen.cmu.edu.tw/
TarFisDock	It docks small molecules into the protein targets in Potential Drug Target Database, and ranks them by the energy score, including their binding conformations	http://www.dddc.ac.cn/tarfisdock/
PLATINUM	Calculates match or mismatch in receptor–ligand complexes and hydrophobic properties of molecules	http://model.nmr.ru/platinum/

monitored using a soybean microarray, 1,342 genes exhibited significant differential expression between uninfected and *P. pachyrhizi*-infected leaves at 12, 24, 48, and 72 h post-inoculation (hpi) in both rust-susceptible and rust-resistant genotypes. Differentially expressed genes were grouped into 12 functional categories, and a large numbers of these genes relate to the basic plant metabolism. These findings provided a better insight into the mechanisms underlying resistance and general activation of plant defense mechanisms in response to rust infection in soybean.

Further, sequencing of EST libraries from pathogen-inoculated or elicitor-treated plants and microarray transcript analyses have enabled the elucidation of genome-wide gene expression changes associated with defence (Ameline-Torregrosa et al. 2006). Samac et al. (2011) used microarray analysis to identify the genes associated with disease defence responses in *M. truncatula*. They compared the genes expressed in response to three pathogens (*Colletotrichum trifolii*, *Erysiphe pisi* and *Phytophthora medicaginis*) and identified genes unique to an interaction.

Fusarium wilt, the most serious disease of pigeonpea, is a common vascular wilt fungal disease caused by *Fusarium* sp. A release draft genome assembly of six strains of different *Fusarium* sp. (Rep and Kistler 2010) gives opportunities to understand the host–pathogen interaction at computational level. In this context, bioinformatics approaches help in understanding the host–pathogen interaction at protein level, in which protein–protein interactions are used to investigate the biological process. Protein–protein interactions are interactions between two or more

Fig. 12.13 An automated protein–protein interaction server

proteins that bind together to carry out their biological function. Protein-protein docking will help understand protein–protein interactions at computational level. HEX, Z-DOCK and other tools are commonly used for protein–protein interaction studies (Fig. 12.13).

Bioinformatics in Molecular Marker Development

For trait analysis using association mapping approaches, and for various other studies on populations including pattern of evolution, population structure, genetic diversity a number of software are available in public domain (Table 12.7). Bioinformatics plays very important role in molecular marker developments, for which several bioinformatics tools and servers are available (Table 12.8). Best optimized primers are essential for good specificity and efficiency. Anyone can design the primer pairs using genomics, mRNA, cDNA, SNP-based sequences. One can design degenerate, expression and universal primers using bioinformatics tools based on servers listed in Table 12.8. For example, Jayashree et al. (2006) have developed a database for EST based simple sequence repeats from cereals and

Table 12.7 Statistical analysis tool and software details with uniform resource locator

Tools/software's name	Description/function	URL (uniform resource locator)
TASSEL	Trait Analysis by Association, Evolution and Linkage; implements general linear model and mixed linear model approaches for association mapping; takes into account population and family structures	http://www.maizegenetics.net/
STRUCTURE	A software package uses multi-locus genotype data to investigate population structure to infer the presence of distinct populations; assigns individuals to populations, detects hybrid zones, identifies migrants and admixed individuals	http://pritch.bsd.uchicago.edu/structure.html
SPAGeDi (Spatial Pattern Analysis of Genetic Diversity)	A computer package primarily designed to characterize the spatial genetic structure of mapped individuals and/or mapped populations using genotype data of any ploidy level	http://ebe.ulb.ac.be/ebe/Software.html
EIGENSTRAT	Uses principal components analysis to explicitly model ancestry differences between cases and controls along continuous axes of variation; the resulting correction is specific to a candidate marker's variation in frequency across ancestral populations; minimizes spurious associations and maximizes power to detect true associations	http://genepath.med.harvard.edu/~reich/Software.htm
MTDFREML	Multiple Trait Diversity Analysis and analysis of variance components	http://aipl.arsusda.gov/curtvt/mtdfreml.html
ASERML	A statistical software package for fitting linear mixed models using restricted maximum likelihood, which is commonly used in plant and animal breeding, and quantitative genetics, and other fields; fits very large and complex data sets efficiently, due to its use of the average information algorithm and sparse matrix methods	http://www.vsni.co.uk/software/asreml
R	A free software environment for statistical computing and graphics; provides a wide variety of statistical (linear and nonlinear modelling, classical statistical tests, time-series analysis, classification, clustering etc.) and graphical techniques, and is highly extendible	http://www.r-project.org/
LDMAP	A program for constructing linkage disequilibrium (LD) maps	http://cedar.genetics.soton.ac.uk/pub/PROGRAMS/LDMAP
SAS	Standard statistical package for traditional statistical analysis	http://www.sas.com/software/sas9/
SPSS	Data mining, statistical analysis and data management softwares	http://www.spss.co.in
NTSys	Discovers patterns and structures in multivariate data	http://www.exetersoftware.com
SigmaPlot	Scientific data and graphing software	http://www.sigmaplot.com

Table 12.8 List of servers used in molecular marker development

Tool/servers name	Description/function	Designated website
Primer3	Widely used program for designing PCR primers	http://frodo.wi.mit.edu/
Gene Fisher	Primer designing based on multiple sequence alignment	http://bibiserv.techfak.uni-bielefeld.de/genefisher/
Web Primer	PCR primer design	http://www.yeastgenome.org/cgi-bin/web-primer
CODEHOP	COnsensus-DEgenerate Hybrid Oligonucleotide Primer	http://blocks.fhcrc.org/codehop.html
PCR Designer	PCR Designer for Restriction Analysis of Sequence Mutations	http://cedar.genetics.soton.ac.uk/public_html/primer.html
Primo Multiplex 3.4	Multiplex PCR Primer Design	http://www.changbioscience.com/primo/primoml.html
Primer Quest	PCR Primers with Probe	http://eu.idtdna.com/scitools/applications/primerquest/
Primo Pro 3.4	PCR Primer Design	http://www.changbioscience.com/primo/primo.html
Primo Degenerate 3.4	Degenerate PCR Primer Design	http://www.changbioscience.com/primo/primod.html
MethPrimer	Design Primers for Methylation PCRs	http://www.urogene.org/methprimer/index1.html
Primaclade	Identifies a set of PCR primers that will bind across the alignment	http://www.umsl.edu/services/kellogg/primaclade.html
Primer3Plus	Pick primers from a DNA sequence	http://www.bioinformatics.nl/cgi-bin/primer3plus/primer3plus.cgi
PrimerBLAST	Finding primers specific to PCR template	http://www.ncbi.nlm.nih.gov/tools/primer-blast/index.cgi
SNP Primers	Creating primers around SNPs in genomic DNA	http://pcrsuite.cse.ucsc.edu/SNP_Primers.html
SSRLocator	Simple Sequence Repeat based primer designing	http://www.ufpel.tche.br/faem/fitotecnia/fitomelhoramento/faleconosco.html
MISA	MIcroSAtellite identification based primer designing	http://pgrc.ipk-gatersleben.de/misa/

legumes. Based on the available resources any one can design EST SSR-based markers for wet-lab experimentation. Large-scale transcriptome assembly using next generation sequencing technologies such as, Roche/454 and Illumina/Solexa, are now used for development of molecular markers, which will serve as a useful resource to accelerate genetic research and breeding applications in legumes. For example, Hiremath et al. (2011) developed 728 SSR, 495 SNP, 387 conserved orthologous sequence (COS) markers, and 2,088 intron-spanning region (ISR) markers in chickpea. Kudapa et al. (2012) predicted for 6,284 intron spanning regions (ISR) covering all the 11 pigeonpea linkage groups.

Mishra et al. (2012) retrieved a total of 18,552 EST sequences (equivalent to 11.3 MB) from the EST database available in the NCBI public domain and analysed for repeat patterns using the tandem repeat finder program at http://c3.biomath.

mssm.edu/trf.html, followed by their assembly using the CAP3 software program (Huang and Madan 1999). After pre-processing, they identified SSR-containing sequences by a perl script-based program, MISA software (MICROSATELLITE identification tool, http://pgrc.ipk-gatersleben.de/misa/). They detected 10,800 unigenes from 18,522 pea EST sequences and screening of 10,800 unigenes by MISA revealed 2,612 (14.1 %) eSSRs in 2,395 (12.9 %) SSR-containing ESTs, from which 577 (24.1 %) primer pairs were designed. Out of these, 68 randomly selected primer pairs showed high rate (48–85 %) of transferability in leguminous species with high level of polymorphism, reproducibility and presence of 3.8 alleles/locus. Similarly, De Caire et al. (2012) retrieved a total of 6,327 mRNA sequences and screened them through a JAVA based programme to design gene-based SSR markers. They successfully identified 45 new polymorphic eSSR markers. e-SSRs identified in these two studies will be used in linkage mapping analyses and provide a good scaffold for comparative mapping in pea and other sequenced legumes.

The molecular markers can be used for linkage mapping using mapping populations developed from biparental crosses. Software like MAPMAKER, QTL-ALL, QTLNETWORK, QUANTO, QU-GENE, QUTIE etc. are used for mapping of markers and oligogenes, while QTL cartographer, QGENE, QTL CAFE, QTL EXPRESS etc. are available for mapping of quantitative trait loci (QTLs). The genes/QTLs detected for target traits need to be confirmed in other replicate studies. Further the marker found linked to the genes/QTLs have to be validated in unrelated germpasm/materials before they can be used for markers-assisted selection (MAS) in plant breeding programmes. Alternatively, marker trait associations can be detected by linkage disequilibrium (LD) based association mapping that uses germplasm collections/breeding lines in the place of biparental mapping populations.

Conclusion and Perspectives

Omics era in the twenty-first century provides us opportunities to understand the legume genome at sequence-structural-functional levels. While legume omics is still in its infancy, it holds great promise, and is expected to yield insights into many aspects of evolution and regulatory mechanisms of legume species. The rapid development of various molecular tools and techniques including large scale analysis of genome organization, gene expression, protein–protein interaction and protein–ligand interaction etc. are generating enormous amount of data, which need to be analyzed and interpreted to develop a biologically meaningful concepts. The need for handling such large amounts of data as forced rapid development of bioinformatics techniques to create, manage and utilize databases of biological information and development of tools and software packages to make efficient and meaningful use of these tools and databases. A variety of software packages are now available to serve various needs of the researchers. However, there is need to develop user friendly bioinformatics tools to decipher functional features of legume genome sequences.

References

Ameline-Torregrosa C et al (2006) Transcriptomic approaches to unravel plant–pathogen interactions in legumes. Euphytica 147: 25–36

Bertioli DJ, Moretzsohn MC, Madsen LH, Sandal N, Leal-Bertioli SC, Guimarães PM, Hougaard BK, Fredslund J, Schauser L, Nielsen AM, Sato S, Tabata S, Cannon SB, Stougaard J (2009) An analysis of synteny of *Arachis* with *Lotus* and *Medicago* sheds new light on the structure, stability and evolution of legume genomes. BMC Genomics 23: 45

Birch PRJ, Kamoun S (2000) Studying interaction transcriptome: coordinated analyses of gene expression during plant–microorganism interactions. In: New technologies for life sciences: a trends guide. Elsevier, London, UK, pp 77–82

Bordat A, Savois V, Nicolas M, Salse J, Chauveau A, Bourgeois M, Potier J, Houtin H, Rond C, Murat F, Marget P, Aubert G, Burstin J (2011) Translational genomics in legumes allowed placing *in silico* 5460 unigenes on the pea functional map and identified candidate genes in *Pisum sativum* L. *G3 (Bethesda)*. Genes Genomes Genet 1(2): 93–103

Cannon SB, May GD, Jackson SA (2009) Three sequenced legume genomes and many crop species: rich opportunities for translational genomics. Plant Physiol 151: 970–977

David P, Sévignac M, Thareau V, Catillon Y, Kami J, Gepts P, Langin T, Geffroy V (2008) BAC end sequences corresponding to the B4 resistance gene cluster in common bean: a resource for markers and synteny analyses. Mol Genet Genomics 280:521–533

De Caire J, Coyne CJ, Brumett S, Shultz JL (2012) Additional pea EST-SSR markers for comparative mapping in pea (*Pisum sativum* L.). Plant Breed 131:222–226

Doyle JJ, Luckow MA (2003) The rest of the iceberg. Legume diversity and evolution in a phylogenetic context. Plant Physiol 131(3):900–910

Dutt S, Singh VK, Marla SS, Kumar A (2010) *In silico* analysis of sequential, structural and functional diversity of wheat cystatins and its implication in plant defense. Genomics Proteomics Bioinformatics 8: 42–56

Garg R, Patel RK, Jhanwar S, Priya P, Bhattacharjee A, Yadav G, Bhatia S, Chattopadhyay D, Tyagi AK, Jain M (2011) Gene discovery and tissue-specific transcriptome analysis in chickpea with massively parallel pyrosequencing and web resource development. Plant Physiol 156(4):1661–1678

Gepts P, Beavis WD, Brummer EC, Shoemaker RC, Stalker HT, Weeden NF, Young ND (2005) Legumes as a model plant family. Genomics for food and feed report of the cross-legume advances through genomics conference. Plant Physiol 137:1228–1235

Goffard N, Weiller G (2006) Extending MapMan: application to legume genome arrays. Bioinformatics J 22: 2958–2959

Hiremath PJ, Farmer A, Cannon SB, Woodward J, Kudapa H, Tuteja R et al (2011) Large-scale transcriptome analysis in chickpea (*Cicer arietinum* L.), an orphan legume crop of the semi-arid tropics of Asia and Africa. Plant Biotechnol J 9:922–931

Huang X, Madan A (1999) CAP3: A DNA sequence assembly program. Genome Res 9: 868-877

Jayashree B, Punna R, Prasad P, Bantte K, Hash CT, Chandra S, Hoisington DA, Varshney RK (2006) A database of simple sequence repeats from cereal and legume expressed sequence tags mined in silico: survey and evaluation. In Silico Biol 6:607–620

Kaló P, Seres A, Taylor SA, Jakab J, Kevei Z, Kereszt A, Endre G, Ellis TH, Kiss GB (2004) Comparative mapping between *Medicago sativa* and *Pisum sativum*. Mol Genet Genomics 272:235–246

Kawaji H, Hayashizaki Y (2008) Genome annotation. Methods Mol Biol 452:125–139

Kudapa H, Bharti AK, Cannon SB, Farmer AD, Mulaosmanovic B, Kramer R et al (2012) A comprehensive transcriptome assembly of pigeonpea (*Cajanus cajan* L.) using Sanger and second-generation sequencing platforms. Mol Plant 5:1020–1028

Kumari A, Singh VK, Fitter J, Polen T, Kayastha AM (2010) Alpha-amylase from germinating soybean (Glycine max) seeds—purification, characterization and sequential similarity of conserved and catalytic amino acid residues. Phytochemistry 71:1657–1666

Kushwaha H, Gupta N, Singh VK, Kumar A, Yadav D (2008) In silico analysis of PCR amplified DOF (DNA binding with one finger) transcription factor domain and cloned genes from cereals and millets. Online J Bioinformatics 9:130–143

Kushwaha H, Gupta S, Singh VK, Rastogi S, Yadav D (2011) Genome wide identification of Dof transcription factor gene family in sorghum and its comparative phylogenetic analysis with rice and *Arabidopsis*. Mol Biol Rep 38: 5037–5053

Kushwaha H, Gupta S, Singh VK, Bisht NC, Sarangi BK, Yadav D (2012) Cloning, in silico characterization and prediction of three dimensional structure of Sbdof1, Sbdof19, Sbdof23 and Sbdof24 proteins from Sorghum [*Sorghum bicolor* (L.) Moench]. Mol Biotechnol 1:12

Lee JM, Grant D, Vallejos CE, Shoemaker RC (2001) Genome organization in dicots. II. Arabidopsis as a "bridging species" to resolve genome evolution events among legumes. Theor Appl Genet 103:765–773

Marla SS, Singh VK (2012) LOX genes in blast fungus (*Magnaporthe grisea*) resistance in rice. Funct Integr Genomics 12: 265–275

Mishra RK, Gangadhar BH, Nookaraju A, Kumar S, Park SW (2012) Development of EST-derived SSR markers in pea (*Pisum sativum*) and their potential utility for genetic mapping and transferability. Plant Breed 131:118–124

Ramírez M, Graham MA, Blanco-López L, Silvente S, Medrano-Soto A, Blair MW et al (2005) Sequencing and analysis of common bean ESTs. Building a foundation for functional genomics. Plant Physiol 137:1211–1227

Rep M, Kistler HC (2010) The genomic organization of plant pathogenicity in *Fusarium* species. Curr Opin Plant Biol 13: 420–426

Samac DA, Graham MA (2007) Recent advances in legume–microbe interactions: recognition, defense response, and symbiosis from a genomic perspective. Plant Physiol 144:582–587

Samac DA et al (2011) Expression of coordinately regulated defence response genes and analysis of their role in disease resistance in *Medicago truncatula*. Mol Plant Pathol 12:786–798

Sanger F, Coulson AR (1975) A rapid method for determining sequences in DNA by primed synthesis with DNA polymerase. J Mol Biol 94:441–448

Sanger F, Nicklen S, Coulson AR (1977) DNA sequencing with chain-terminating inhibitors. Proc Natl Acad Sci U S A 74:5463–5467

Sato S, Nakamura Y, Kaneko T, Asamizu E, Kato T, Nakao M, Sasamoto S, Watanabe A, Ono A, Kawashima K et al (2008) Genome structure of the legume, *Lotus japonicus*. DNA Res 15:227–239

Schmutz J, Cannon JB, Schlueter J, Ma J, Mitros T et al (2010) Genome sequence of the palaeopolyploid soybean. Nature 463:178–183

Singh VK, Singh AK, Chand R, Kushwaha C (2011) Role of bioinformatics in agriculture and sustainable development. Intern J Bioinformatics Res 3: 221–226

Singh NK, Gupta DK, Jayaswal PK, Mahato AK, Dutta S, Singh S, Bhutani S et al (2012) The first draft of the pigeonpea genome sequence. J Plant Biochem Biotechnol 21:98–112

Soria-Guerra RE et al (2010) Transcriptome analysis of resistant and susceptible genotypes of Glycine tomentella during Phakopsora pachyrhizi infection reveals novel rust resistance genes. Theor Appl Genet 120:1315–1333

Thornton JW, DeSalle R (2000) Gene family evolution and homology: genomics meets phylogenetics. Annu Rev Genomics Hum Genet 1: 41–73

Udvardi MK (2002) Legume genomes and discoveries in symbiosis research. Genome Biol 3:reports 4028

Varshney RK, May GD (2012) Next-generation sequencing technologies: opportunities and obligations in plant genomics. Brief Funct Genomics 11:1–2

Varshney RK, Close TJ, Singh NK, Hoisington DA, Cook DR (2009) Orphan legume crops enter the genomics era! Curr Opin Plant Biol 12:202–210

Varshney RK, Chen W, Li Y, Bharti AK, Saxena RK, Schlueter JA, Donoghue MT, Azam S et al (2011) Draft genome sequence of pigeonpea (*Cajanus cajan*), an orphan legume crop of resource-poor farmers. Nat Biotechnol 30:83–89

Varshney RK, Chi S, Saxena RK, Azam S, Sheng Y, Shapre AG, Steven C, Jongmin B, Rosen BD et al (2013) Draft genome sequence of chickpea provides a resource for trait improvement. Nat Biotechnol 31:240–246

Wojciechowski MF, Lavin M, Sanderson MJ (2004) A phylogeny of legumes (Leguminosae) based on analyses of the plastid matK gene resolves many well-supported subclades within the family. Am J Bot 91:1846–1862

Yadav D, Singh VK, Singh NK (2007) *In silico cis*-regulatory elements analysis of seed storage protein promoters cloned from different cultivars of wheat, rice and oat. Online J Bioinformatics 8(2):1–9

Young ND, Udvardi M (2009) Translating *Medicago truncatula* genomics to crop legumes. Curr Opin Plant Biol 12:193–201

Young ND, Cannon SB, Sato S, Kim D, Cook DR et al (2005) Sequencing the gene spaces of *Medicago truncatula* and *Lotus japonicus*. Plant Physiol 137:1174–1181

Zhu H, Choi HK, Cook DR, Shoemaker RC (2005) Bridging model and crop legumes through comparative genomics. Plant Physiol 137:1189–1196