International Journal of Agricultural Economics and Management.
ISSN 2277-3215 Volume 7, Number 1 (2017), pp. 7-16
© Research India Publications
http://www.ripublication.com

In Silico Promter Prediction in Chromosome 10th of Sorghum Bicolor

Divya Mudgil[1]*, Bhavika Chhettri[2], Sakina Khurana[3] and Divya Singhal[4]

Department of Biotechnology Engineering, Ambala College of Engineering and Applied Research, Ambala-134003, India.

Abstract

Sorghum bicolor belong to the grass family Poaceae and it is the first plant of African origin whose genome was sequenced with genome size of 697.6 mb. Sorghum bicolor total genome consists of 10 chromosomes with their respective sizes. we have predicted and analysed the chromosome no. 10th of sorghum bicolor for research basis to find out similarity with other plants genome and use for comparison of other genomes in order to improve or modify their properties. Because bicolor grows in a wide range of temperature, high altitudes, toxic soil and recover growth after drought condition. Due to high demand of this plant we predict the chromosome 10th for research prospecting. Sorghum bicolor has 10 chromosomes out of these we selected chromosome 10th, because this chromosome has small size. Hence the chromosome cut into specific size files by using file splitter software that is used to generate software compatible files. The files of chromosome 10th (60.9 mb) were subjected to softberry software prediction(TSSP). As per the limit of software only a file with 90 kb could be entered and so each file were made of 90 kb. After undergoing computational analysis s of promoter prediction, all of them showing total length of the sequence in each subfiles, promoters and enhancer predicted, position of TATA boxes and TFBE. Their recognition by computer algorithms is fundamental for understanding gene expression pattern, cell specificity and development. This describe the advance approaches to identify promoters and we discuss an approach to identify statistically significant regulatory motifs in genomic sequence.

Keywords: Promoter Prediction; TATA Box; Transcription factors; Regulatory motif; methodology; prediction softwares; homology inference.

INTRODUCTION

Sorghum bicolor commonly called sorghum and also known as durra, jowari or milo, is a grass species cultivated for its grain, which is used for food both animals , humans and for ethanol production. Sorghum originated in northern Africa and is now cultivated widely in tropical and subtropical regions. It grows in clumps that may reach over 4 m high. The grain is small ranging from 3 to 4 mm in diameter. Sweet sorghums are cultivates that are primarily grown for foliage, syrup production and ethanol. ***[Carpita and Cann, (2008)].***

The leading producers of sorghum bicolor in 2011 were Nigeria (12.6%), India (11.2%) and the United State (10.0%). Sorghum grows in a wide range of temperature, high altitudes, toxic soils and can recover growth after some drought condition. It has four features that make it one of the most drought- resistant crops. It has a very large root-to-leaf surface area. In times of drought, its leaves to lessen water loss by transpiration. If drought continues, it will dormancy rather than dying. Its leaves are protected by a waxy cuticle. Sorghum bicolor is a multipurpose crop providing food, feed, fibre and fuel especially in those with fragile conditions. Food and Agriculture Organization data show that sorghum is currently the number five most important grain crop. Its yearly production has been stabilized at 60 million tones with a harvesting area of 44 million hectares**. [*Abeel, et al,(2008)].***

Sorghum is closely related to maize and belong to the so called C4 – plants. C4 plants have an improved photosynthesis activity at high temperature and drought. As the first fully sequenced cereal that C4 photosynthesis, analysis of the sequence provides new insights into the recruitment of C3 genes to the C4 pathways.***[Bower, et al, (2003)].***

Whole genome shotgun sequences of sorghum line ATx623. (AT/AT)n to 26.1% of all SSRs, followed by (AG/TC)n at 20.5% (AC/TG)n at 13.7% and (CG/GC)n at 11.8%. Sorghum bicolor genome sequences has promoters, enhancers, TATA Boxes, transcription factors binding sites***. [Yonemaru , et al, (2009)].***

Promoters: It is a region of DNA that initiates transcription of a particular gene. Promoters are located near the genes they transcribe on the same strand and upstream on the DNA(towards the 3' regions of the anti – sense strands also called template strand and non coding strands). They can be about 100 – 1000 base pairs long. They work as cis – regulatory element to mediate both spatial and temporal control of development by turning in specific cells and or repressing it in other cell. ***[Smale, T.; Kadonaga, T, (2003)].***

Enhancer: This is a short region of DNA that can be bound with proteins (namely, the transacting factors, much like a set of transcription factors) to enhance transcription levels of gene in a gene cluster. While enhancers are usually cis – acting, an enhancer does not need to be particularly close to the genes it acts on, and sometimes need not be located on the same chromosome***.[Roeder, (1996)].***

TATA Boxes : The TATA Boxes(also called Goldberg- Hongness box) has the core DNA sequence 5'-TATAAA-3' or a variant, which is usually followed by three or more adenine bases. It is usually located 25 base pairs upstream of the transcription site. The sequence is believed to have remained consistent throughout much of the evolutionary process, possibly originating in an ancient eukaryotic origanism. It is normally bound by the TATA binding protein (TBP) in the process of transcription, which unwind the dna and bends it through 80 degree. The AT- rich sequence facilitates easy unwinding (- stacking interactions among due to weaker base A and T rather than G and C). The TBP is an unusual protein in that it binds to the minor groove and binds with a β sheets. ***[C Yang, E Bolotin, T Jiang, FM Sladek, E Martinez, (March 2007)].***

Transcription Factors: A transcription factor (sometimes called a sequence - specific DNA - binding factors) is a protein that binds to specific DNA sequences, thereby controlling the flow (or transcription) of genetic information from DNA to m RNA. Transcription factors performs this function alone or with other proteins in complex by promoting (as an activators) or blocking (as a repressor) the recruitment of RNA polymerase (the enzyme that performs the transcripton of genetic information from DNA to RNA) to specific genes. A defining feature of transcripton factors is that they contain one or more dna- binding domain (DBDs). Which attach to specific sequences of DNA adjacent to the genes that they regulate. Additional proteins such as co-activators, chromatin remodelers, histone acetylases, deacetylases, kinases and methylase. While also playing crucial roles in gene regulation, lack DNA-binding domains and therefore are not classified as transcripton factors.

MATERIALS AND METHODS

Chromosome collection, computational promoters prediction and its tools:

Promoters predicted in silico using computational analysis or grey analysis that is the basis for using of various tools softwares for prediction. The input file or query sequence is to be entered in the FASTA formate because the algorithms formed for the softwares are programmed for the same and have a set limit of size of the query sequence to give the output result.

The tools used for promoter prediction are online as well as offline (downlodable). Every tool has its own set of algorithms responsible for the outputs.

Tools for computational prediction are:-

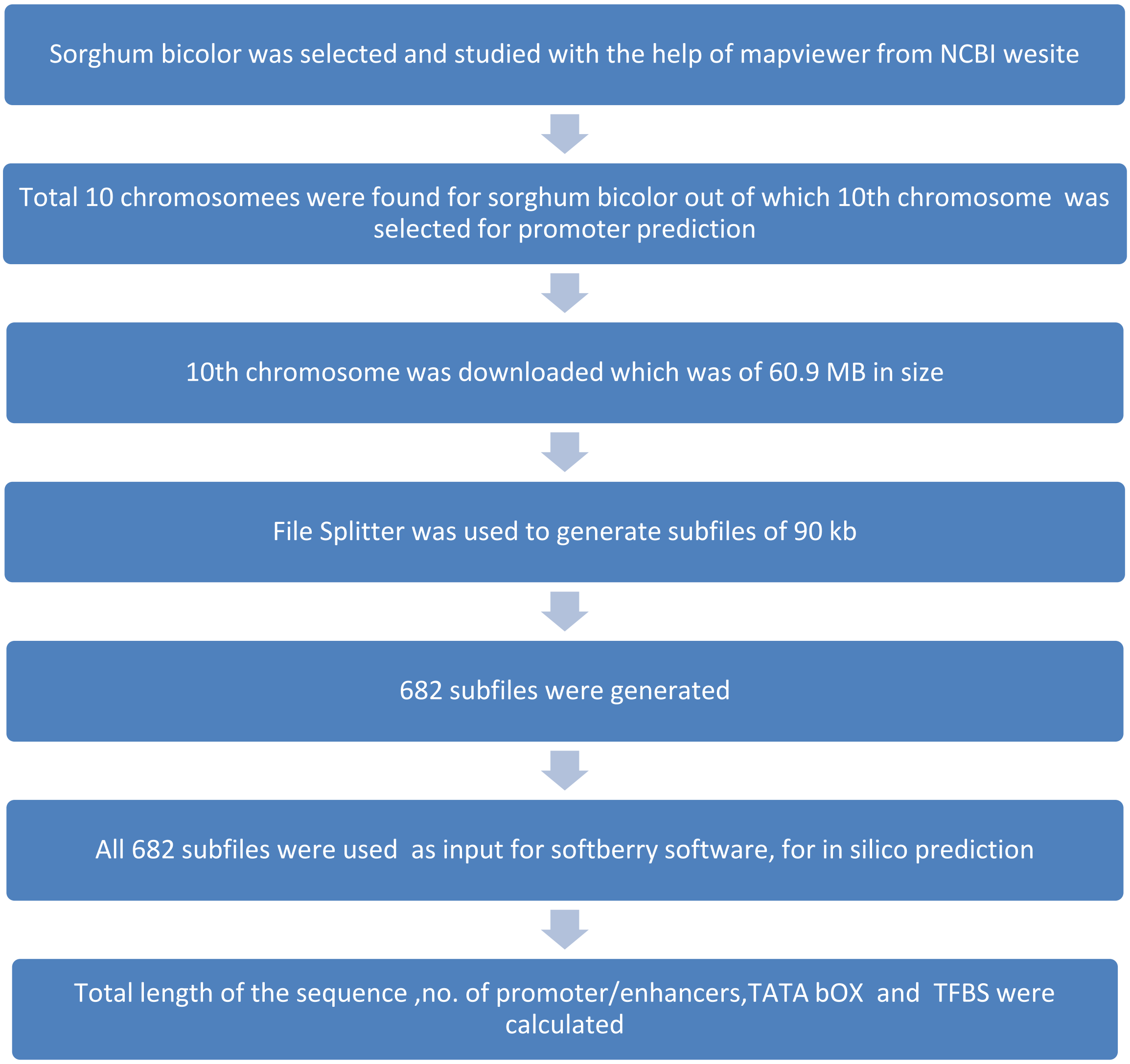
Softberry (http://linux1.softberry.com/berry.phtml)
Neural networks (http://www.fruitfly.org/seq_tools/promoter.html)
Promoter 2.0 prediction server (http://www.cbs.dtu.dk/services/promoter/)
Promoser (http://biowulf.bu.edu/zlab/promoser)
Virtual Footprint (http://www.prodoric.de/vfp)
SCOPE (http://genie.dartmouth.edu/scope)

Sorghum bicolor has 10 chromosomes in its genome. Its genome size is 738,787,382bp. Size of 1st chromosome is 73,850,631 bp, 2nd is 77,932,606 bp, 3rd is 74,441,160 bp, 4th is 68,034,345 bp 5th is 62,352,331 bp, 6th is 62,208,784 bp, 7th is 64,342,021 bp, 8th is 55,460,251 bp, 9th is 59,635,592 bp and 10th is 60,981,646 bp. Out of these 10th chromosome was selected.

File Splitter: This program is C language based program in which according to command, loop continue and hence the chromosome is cut into specific size files. In this case loop was set in which 90 kb sized files were generated.

Softberry: Based on fastest and most accurate ab initio gene prediction program, FGENESH. Softberry fully automatic genome annotation pipeline, FGENESH++C, is the best available. FPROM, TSSG and TSSW are used for the promoter prediction of human genome, TSSP, NSITE-PL AND NSITEM-PL are used for the promoter prediction of plant genome, NSITEM, NSITE are used for the recognition of regulatory motifs, BPROM is used for bacterial promoter prediction. From this, TSSP was selected.

Flow chart showing steps of methodology:



Sorghum bicolor was selected and studied with the help of mapviewer from NCBI wesite

Total 10 chromosomees were found for sorghum bicolor out of which 10th chromosome was selected for promoter prediction

10th chromosome was downloaded which was of 60.9 MB in size

File Splitter was used to generate subfiles of 90 kb

682 subfiles were generated

All 682 subfiles were used as input for softberry software, for in silico prediction

Total length of the sequence ,no. of promoter/enhancers,TATA bOX and TFBS were calculated

Figure 1

RESULTS

The 682 subfiles of chromosomes 10th (60.9 MB) were subjected to softberry software prediction. As per the limit of software only a file with 90 kb size could be entered and so each file was made of 90 kb. After undergoing computational analysis of promoter prediction, all of them gave respective results showing total length of the sequence in each subfile, promoters and enhancers predicted, position of TATA boxes and TFBS as shown in table 1. It was found that out of 682 files, 27 files did not gave any output with no promoter/enhancer, no TATA boxes and no TFBS. First 28 files out of 682 files were of variable sizes ranging upto 60-90 kb.

As shown table 1, the summarized results of 682 files of chromosome 10th in Sorghum bicolor are mentioned i.e total length of the chromosome no. 10th is 60.9 mb, where total number of promoters/enhancers predicted was 27417, total numer of TATA boxes predicted was 21647 and toatal number of TFBS predicted was 943341.

Table 1: Summarized results of promoter prediction in chromosome no. 10th.

Total length of all subfiles (682)	60.9 mb
Total number of enhancers and promoterS	27417
Total number of TATA boxes	21647
Total numbers of TFBS	943341

DISCUSSION

According to Plantagora a whole genome data resource of plant assembly , most of the assemblies for the plantagora project were produced from read based on chromosome one of the rice genome, because very high demand made on computer resources by the assembly of larger eukaryotic genomes. The choice to use sorghum as a grasses model is based on its relatives small genome for a crop plant (third the size of maize genome, quarter the size of the human genome), its low level of duplication and his high number of repetitive element. Drought tolerance makes it a staple for human populations in arid environments. It is also a good source of feed, fiber and fuel in the global agronomics and economics. We used this single chromosome (45 mb) as a smaller model genome for testing a wide range of sequencing and assembly approaches. In order to compare our results from this smaller model to those from the larger plant genomes that currently challenge researchers, we performed a focused series of assemblies with the Arabdopsis thaliana (119 mb) genome, the whole Oryza sativa genome (382 mb) ,and with the Sorghum bicolor genome (697.6 mb). A single assembly was performed for each genome using each of the different assembler/platform combinations used for the Plantagora project *(http://www.plantagora.org/).*Nitrogen (N) fertilizers are a major agricultural input where more than 100 million tons are supplied annually. Cereals are particularly inefficient at soil N uptake, where the unrecovered nitrogen causes serious environmental damage. Sorghum bicolor (sorghum) is an important cereal crop, particularly in resource-poor semi-arid regions, and is known to have a high NUE in comparison to other major cereals under limited N conditions. ***[Massel K, Campbell BC, Mace ES, Tai S, Tao Y, Worland BG, Jordan DR, Botella JR, Godwin ID].***

Softberry software used for annotation of plant genome (Gene, promoter prediction, functional motif and protein subcellular localization).**Plant Molecular Biology (2005), 57, 3, 445-460:** Evaluated the five ab initio programs (FGENESH, GeneMark.hmm, GENSCAN, GlimmerR and Grail) for their accuracy in predicting maize genes. FGENESH yielded the most accurate and GeneMark.hmm the second most accurate predictions" (FGENESH identified 11% more correct gene models than GeneMark on a set of 1353 test genes).

Yu et al. (2002) : A draft sequence of the rice genome (Oryza sativa L. ssp. indica). Science 296:79-92. As part of rice genome sequencing project, the team led by Beijing Genomics Institute has compared several wellknown ab initio gene prediction programs and shown that FGENESH is by far the most accurate .

Galagan et al. (2003) : The genome sequence of the filamentous fungus Neurospora crassa. Nature 422:859- 868. Neurospora genome annotation based on FGENESH and FGENESH+. result, their rice genome annotation was based almost exclusively on FGENESH .

Analysis of TSS-motifs revealed that their composition is different in dicots and monocots, as well as for TATA and TATA-less promoters. The database serves as learning set in developing plant promoter prediction programs. One such program (TSSP) based on discriminant analysis has been created by Softberry **[*Ilham A. Shahmuradov, Alex J. Gammerman, John M. Hancock, Peter M. Bramley1 and Victor V. Solovyev2,*]***

promoters search on genes with known mRNAs by different promoter-finding programs. Reproduced with changes ***from Liu and States (2002) Genome Research 12:462-469***. Promoter Prediction Programs TSSW, TSSG and PromH. Two of the most accurate eukaryotic PolII promoter prediction programs, TSSW and TSSG, are based on discriminant analysis combining characteristics of functional elements of regulatory sequence with the database of regulatory motifs. TSSG is the most accurate stand-alone promoter prediction program available: it correctly predicts 50-60% of promoters, and 80-85% of promoters predicted by TSSG are true promoters. Accuracy of TSSW is only slightly lower (see Table 4). PromH is an enhancement of TSSW that adds information on syntenic regions of mouse and human genomes into prediction algorithm. That results in additional 20% accuracy improvement, especially pronounced on TATA+ promoters. TSSW and PromH programs contain elements of older version of Transfac database (www.biobase.de), and their use may require Transfac license.

CONCLUSION

In this study of “In silico promoter prediction”, Sorghum bicolor genome was used which was of 697.6 mb in size. It has total 10 chromosomes out of 10 chromosome we used chromosome no. 10th for promoter identification. The 60.9 mb file of chromosome 10th was downloaded and splitted into total 682 files of 90 kb each using a C- program File Splitter. All the subfiles were subjected to promoter prediction using softberry softwares (online). Input files were entered in FASTA formate and results were achieved in four different categories.

In result we achieved total length of each subfiles, total number of promoter/enhancers, total number of TATA boxes and total number of TFBS. It was found out that in chromosome 10th of sorghum bicolor genome promoters found 27417, TATA boxes found were 21647 and TFBS found were 943341. Moreover 27 files were found which did not gave ay results i.e no promoters, no enhancers, no TATA boxes and no TFBS were found in these. Subfiles number 362-389 did not show any promoter region, either they may have non-codind region or repetitive sequences.

These result can be used for comparision with other genome in order to improve and modified their properties and also be use as research basis to find out similarity with genomes. Future prospecting of this it is used for insertion of drought resistant gene in to the other crops which enhance the drought resistant condition of crops.

ABBREVATIONS

BRE	:	B –Recognition element
DB	:	Database
DBD	:	DNA Binding Domain
DNA	:	Deoxyribonucleic acid
RNA	:	Ribonucleic acid
NCBI	:	National Center for Biotechnological Information
SCOPE	:	Suite for computational Identification of promoter Element
SSR	:	Single Nucleotide Polymorphism
TBP	:	TATA Binding Protein
TFBS	:	Transcription Factor Binding Sites
TSSP	:	Transcripton Start Site Programme

ACKNOWLEDGEMENT

ACE, Kurukshetra University, Haryana (India) are thankfully acknowledged for assistance with promoter prediction analysis. The authors also sincerely appreciate laboratory facilities received from ACE. We are delighted to express my sincere thanks to preoject supervisor Ms. Divya Singhal and Dr.R.K Pundir, Head,

Department of Biotechnology for kind help, encouraging extending helping hand in all possible way. The authors declare that they have no conflict of interest.

REFERENCES

[1] Abeel T,et al (2008) Core promoter prediction based on unsupervised clustering of DNA physical profiles.Bioinformatics,24:24-31.

[2] Bowers, J.E. et al (2003). A high-density genetic recombination map of sequence-tagged sites for Sorghum, as a framework for comparative structural and evolutionary genomics of tropical grains and grasses. Genetics ,165:367-386.

[3] Carpita NC, McCann MC (2008). Maize and sorghum : genetic resources for bioenergy grasses. Trends plant science, 13:415-420.

[4] Hampson, S et al (2003). LineUp: Statistical detection of chromosomal homology wiyh application to plant comparative genomics. Genome Res, 13:1-12.

[5] Harlan JR (1972). A simplified classification of sorghum, Crop Science, 12:172-176.

[6] Latchman DS (1997). “Transcription factors: an overview”. Int. J. Biochem. Cell Biology , 29(12): 1305-12.

[7] Miller, J.T. et al (1998). Cloning and characterization of a centromere-sepecific repetitive DNA element from Sorghum bicolour. Theor. Appl. Genet. 96:832-839.

[8] Roeder RG (1996). The role of general initiation factors in transcription by RNA polymerase II. Trends Biochemistry Science, 21 (9): 327-335.

[9] Rooney WL, Blumenthal J, Bean B, Mullet JE (2007). Designing sorghum as a dedicated bioenergy feedstock. Biofuels Bioproducts Biorefining-Biofpr, 1:147-157.

[10] Salse, J ., Piegu, B.,Cooke, R., and Delseny, M (2004). New in silico insight into the synteny between rice(Oryza sativa L.). and maize (Zea mays L.) highlights reshuffling and identifies new duplications in the rice genome. Plant journal, 38:396-409.

[11] Shiraki T, et al (2003). Cap analysis gene expression for high-throughput analysis of transcriptional starting point and identification of promoter usage. Proc. Natl Acad. Sci. USA, 100:15776-15781.

[12] Vermerris W (2011). Survey of genomics approaches to improve bioenergy traits in maize, Sorghum and sugarcane, J Integr Plant Biol, 53:105-119.

[13] Waibel, A.H et al (1989). Phoneme Recognition Using Time-Delay Neural Networks. IEEE Transactionss on Acoustic, Speech, and Signal Processing, 37 (3):328-339.

[14] Wakaguri H, et al (2008). Dtabase of transcription start sites, progress report. Nucleic Acids Res, 36:97-101.

[15] Wu S, et al (2007). Eukaryotic promoter prediction based on relative entropy and positional information. Phys.Rev. Stat.Nonlin. Soft Matter Phys,75:41-90.

[16] Xie X, et al (2006). An effective promoter identification method based on the AdaBoost algorithm. Bioinformatics, 22:2722-2728.

[17] Yamaguchi, D., Li, G.D.,and Nagai, M.(2007). Verification of effectiveness for grey relational analysis models. Journal of Grey System, 10(3):169-182.

[18] Young, W.E., Teetes, G.L.,(1977). Sorghum Entomology . Ann. Rev. Entomol.22:193-218.

[19] Zhao, L.Y., et al, (2009). Biomass yield and changes in chemical composition of sweet sorghum grown for biofuel. Field Crop Res., 111:55-64.