

**DEVELOPMENT OF MACHINE-LEARNING BASED  
APPROACHES FOR THE IDENTIFICATION OF EXPANSIN  
PROTEINS**

By

Piyus Mohanty (161518)

**UNDER THE GUIDANCE OF**

Dr. Narendra Kumar



June 2020

*Submitted in partial fulfilment of the requirement*

*for the award of the degree of*

**BACHELOR OF TECHNOLOGY, BIOINFORMATICS**

**JAYPEE UNIVERSITY OF INFORMATION TECHNOLOGY,  
WAKNAGHAT**

## ACKNOWLEDGEMENT

The work done here could not have been accomplished on my own. This project is the culmination of my hardwork and the guidance received.

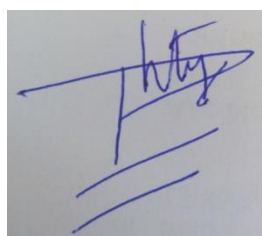
Primarily, I am grateful to my supervisor **Dr. Narendra Kumar** for providing me with an opportunity to work under him. I am highly obliged for his invested time& efforts for the success of our project as well as for playing an integral role in constantly mentoring me and guiding me throughout the process of the project. From teaching me the basics of being a good researcher to giving me life lessons, he has helped me become a better version of myself. His support is invaluable and our association is something that I will cherish for a lifetime.

I would also like to thank, **Dr. Jayashree Ramana** who introduced me to the world of machine learning in the field of bioinformatics and has played a vital role in shaping my interest & development through her constant support as well as motivation. Her teachings will definitely help me in my foray in the field of bioinformatics.

I'll be forever grateful to **Mrs. Somlata Sharma**, our lab technician for her unwavering support and teaching me the virtues of life. Also to, PhD. **Nadia Ahmad** and **Rohit Shukla** for their valuable insights throughout the project. It would be incomplete without acknowledging the contribution of my friends **Chintan, Prashant, Shubhaditya, Saishta, Shivani and Medhavi** who pushed and motivated me when things were tough.

## STUDENT DECLARATION

I hereby, declare that the research reported in the project report entitled “ Development of machine-learning-based approaches for identification of expansion proteins” submitted at the Jaypee University of Information Technology, Wagnaghat for the completion of the B.Tech, is a legitimate record of the work carried out under the supervision of Dr. Narendra Kumar. I have not submitted this work elsewhere for any other degree or diploma.



(Signature of the Student)

Piyus Mohanty (161518)

Department of Biotechnology & Bioinformatics,

Jaypee University of Information Technology,

Wagnaghat, Himachal Pradesh, India.

Date:

## CERTIFICATE

This is to certify that project report entitled “**Development of machine-learning-based approaches for identification of expansion proteins**”, submitted by Piyus Mohanty(161518) is in its partial fulfilment for the award of the degree of Bachelor of Technology in Bioinformatics to Jaypee University of Information Technology, Wagnaghat, H.P., India is an authentic record of candidate’s own work carried out by him under my supervision. This work has not been submitted partially or fully to any other university or institution in order to achieve any award or any other degree.



Dr. Narendra Kumar

Assistant Professor (Grade II),

Department of Biotechnology & Bioinformatics,

Jaypee University of Information Technology,

Wagnaghat, Himachal Pradesh, India.

## **TABLE OF CONTENTS**

<b>ABSTRACT</b> .....	6
<b>INTRODUCTION</b> .....	7
<b>MATERIALS ANDMETHODS</b> .....	13
<b>RESULTS AND DISCUSSION</b> .....	14
<b>CONCLUSION</b> .....	16
<b>REFERENCES</b> .....	17

## **ABSTRACT**

Expansin is one of the most sought out cell wall proteins in plants. It plays a crucial role in cell walls modularity processes such as, providing cell wall with plasticity along with softening certain fruits as well as root hair elongation along with several other functions. Which makes it even more imperative that sequence and structure-based knowledge of expansin be used for better understanding of expansin functions, and hence highlighting the need for correlating sequence and structure starting from identification of the said proteins.

Machine learning methods like SVM are used for the purpose of classification which is carried out using an in-house python script. Several features such as amino acid composition, dipeptide composition along with several others are taken in two definitive ratios that are 1:1 and 1:2 have been used to train a model. After which the best model has been used to create a web server with an intent to help the research community and researchers working in this field to annotate sequential and structural information. The web-server created can be accessed using the link-<http://piyus22.pythonanywhere.com/predictor/>.

**Keywords:** SVM, amino acid composition, dipeptide composition, expansins

## INTRODUCTION

The cell wall is quite important as it has several decisive roles in activities such as differentiation, transportation, communication and several other roles ultimately contributing to plant growth. Primary cell wall controls the expansion of plant cells, by a coordinated process wherein it confines the shape and accordingly selectively loosens it leading to cell wall relaxation and subsequent water uptake which in turn results in enlargement of the cell <sup>[1]</sup>. The cell wall is considered to be a complex structure composed of various polysaccharides, proteins, suberin, enzymes, waxes as well as several other components because of this plants can adapt to almost all climates <sup>[2]</sup>. The cell wall is modified when the plant cells grow, develop, face environmental stresses or encounter infection by enzymatic action <sup>[3]</sup>. Structurally the cell wall is polyhedral consisting of edges and vertices <sup>[4]</sup>. The most common localized proteins are HRGPs or extensins, AGPs, GRPs and the PRPs apart from these there are several other proteins as well <sup>[5]</sup>. Even though structural components involved in cell growth and cell wall are prominent, research is still going on how these components are linked so as to form a strong and stable primary wall <sup>[6]</sup>.

Expansins were earlier exposed in cucumber hypocotyls <sup>[7]</sup>, localized in the cell wall were highly conserved. Thus, they were believed to play a significant part in the cell wall plasticity whenever plant cells expand or differentiate <sup>[8]</sup>. In an experiment when a growing cell was disrupted, it could be done by freeze/thaw cycle and after denaturation of cell wall proteins <sup>[9]</sup>, they lose the ability to extend their ability to extend which was reinstated after addition of expansin protein <sup>[10]</sup>. Expansins prefer acidic pH and hence are preferred mediators for “acid growth” in plants. This means that the expansins play a critical role in growth during acidic condition which takes place during stress condition which leads to cell elongation <sup>[11]</sup>. According to sequence-based phylogeny, it is broadly classified into two major families EXPA and EXPB. EPA's are

associated with cell wall loosening during optimum acidic pH, whereas EXPB's include several set of proteins on which research is still going on. It has four different subfamilies i.e.  $\alpha$ -expansin(EXPA),  $\beta$ -expansin(EXPB), expansin-like A(EXLA) and expansin-like B(EXLB) <sup>[12]</sup>. EXLA & EXLB are smaller families targeting the cell wall modification <sup>[13]</sup>.

Conventional method i.e. sequence comparison method, wherein we align the query sequence across a database of known sequences to find the annotation associated with it, But in the case of the diverse dataset where there is very less similarity or identity this approach fails. So we need a method or approach which would help in dealing with this flaw, this is wherein we have employed machine learning-based algorithm specifically support vector machines (SVM) which is quite popular when it comes to prediction, it could be used for prediction of structure, function or interactions along with it <sup>[14-15]</sup>, it has high accuracy and can be used to deal with high-dimensional and large diverse datasets <sup>[16-18]</sup>. With high accuracy, there is less chance of loss of function due to false prediction. It is generally used for binary classification i.e. distinguish between two categories, one could be considered positive the other might be considered negative just for the sake of convenience. It is able to do so because of the use of two key concepts which are large margin separation and kernel function. Large margin separation deals with the classification of points in two dimensions whereas kernel function deals with the similarity between two points.

Sequence and structure-based information have been used to build a machine learning model consisting of various features like amino acid composition as well as dipeptide composition. AA composition refers to the number of individual residues divided by the sequence length, whereas the dipeptide composition refers to di-amino acid composition possible for all combinations (400) divided by length, using these features models are trained to predict whether the query sequence is expansin like or not along with it, if it has any associated structure that could have a similar structure which in turn would help in determining the function associated with it. This is an interesting approach as not much is done in identifying and classifying expansins into different families as well as correlating their structure with a sequence which will aid in identifying functions.



The following table explains some of the important genes and their subfamily information related to expansins and their involvement in plant cell growth along with it how they deal with stress conditions.

**Table-1**

(Expansin's effect on the plant cell development as well as their adaption to harsh conditions)

<b>Expansin name</b>	<b>Sub-family</b>	<b>Mode of expression</b>	<b>Observed phenotype</b>	<b>References</b>
AtEXPA1	$\alpha$ -Expansin	Overexpression and inhibition	Increased rate of light-induced opening of stomata and reduces the sensitivity of stomata.	[19]
AtEXPA2	$\alpha$ -Expansin	Overexpression and suppression	Overexpresses germinated faster than wild type plants while germination was delayed in mutant lines	[20]
AtEXP3	$\alpha$ -Expansin	Overexpression	Enhanced growth and larger leaves	[21]

			under normal growth conditions	
AtEXPA4	$\alpha$ -Expansin	Expression profile analyses	Thought to soften the cell wall of the stigma	[22]
AtEXPA10	$\alpha$ -Expansin	Overexpression	Large plant cells, larger leaves and longer stems	[24]
AtEXPA18	$\alpha$ -Expansin	Overexpression	Influenced root hair initiation and root growth	[26]
LeEXPA1	$\alpha$ -Expansin	Overexpression and Suppression	Overexpression of the gene resulted in softer fruits while its suppression produced firmer fruits in transgenic tomatoes	[27]
LeEXPA8	$\alpha$ -Expansin	mRNA expression analysis	Thought to influence germination since it is expressed in germinating seeds only and appears to be involved during	[28]

			the initial elongation of the radicle	
LeEXPA10	$\alpha$ -Expansin	mRNA expression analysis	Thought to influence germination as well as seed development	[29]
OsEXPA4	$\alpha$ -Expansin	Overexpression Antisense (RNAi)	Pleiotropic phenotypes in plant height, leaf number, flowering time and seed set as well as enhanced coleoptile growth Shorter plants, decreased coleoptile and mesocotyl lengths	[32]
OsEXPA8	$\alpha$ -Expansin	Overexpression	Increased root mass, number and size of leaves as well as plant height	[33]
DzEXP1	$\alpha$ -Expansin	Expression analysis	Thought to be involved in fruit/pulp softening and peel dehiscence	[35]

NtEXPA5	$\alpha$ -Expansin	Overexpression	Increased organ size especially the leaves and the stem	[36]
FaExp2	$\alpha$ -Expansin	Expression analysis	Thought to take part in cell wall polymer disassembly during fruit ripening	[37]
MaExp1	$\alpha$ -Expansin	Overexpression	Thought to affect banana ripening	[38]
PpEXP1	$\alpha$ -Expansin		Enhanced germination and abiotic stresses tolerance	[39]

## Table-2

Represents a minimum no. of genes for different families of expansin across different plant species

Species	EXPA	EXPB	EXPLA	EXPLB	Total	References
<b>Angiosperms</b>						
Arabidopsis thaliana	26	6	3	1	36	[40]
Poplar	27	3	2	4	36	[41]
Grape	20	4	1	4	29	[42]
Soybean	49	9	2	15	75	[43]

Apple	34	1	2	4	41	[44]
Chinese cabbage	39	9	2	3	53	[45]
Rice	33	18	4	1	56	[40]
Maize	36	48	4	0	88	[46]
<b>Nonflowering plants</b>						
Selaginella moellendorffii	15	2	0	0	17	[47]
Physcomitrella patens	28	7	0	0	35	[48]

## MATERIALS AND METHODS

### Data Collection

Data has been downloaded from NCBI using keyword search Expansin and “not hypothetical” and not isoform and not partial and not putative and not fragment, after which multiple sequence alignment was performed and then the corresponding percentage identity matrix was used to plot a histogram and also a dendrogram was generated.

### Redundancy removal and PSI-BsLAST

CD-Hit suite was used to remove redundancy from the collected dataset. CD-Hit with sequences identity cut off 0.6, 0.5, 0.4 were used, it is a tool used for clustering and classifying proteins <sup>[49]</sup>. After which PSI Blast was performed on the results from 0.6 cut-off from CD-HIT wherein individual sequences from the dataset were treated as query sequences and aligning it against the database created of remaining sequences. PSI-BLAST helps in determining distant relationship among proteins in the corresponding database, which BLAST fails in doing so <sup>[50]</sup>. To check the diversity of non-redundant sequences, multiple sequence alignment was carried out after which percentage identity matrix was extracted and a histogram was plotted.

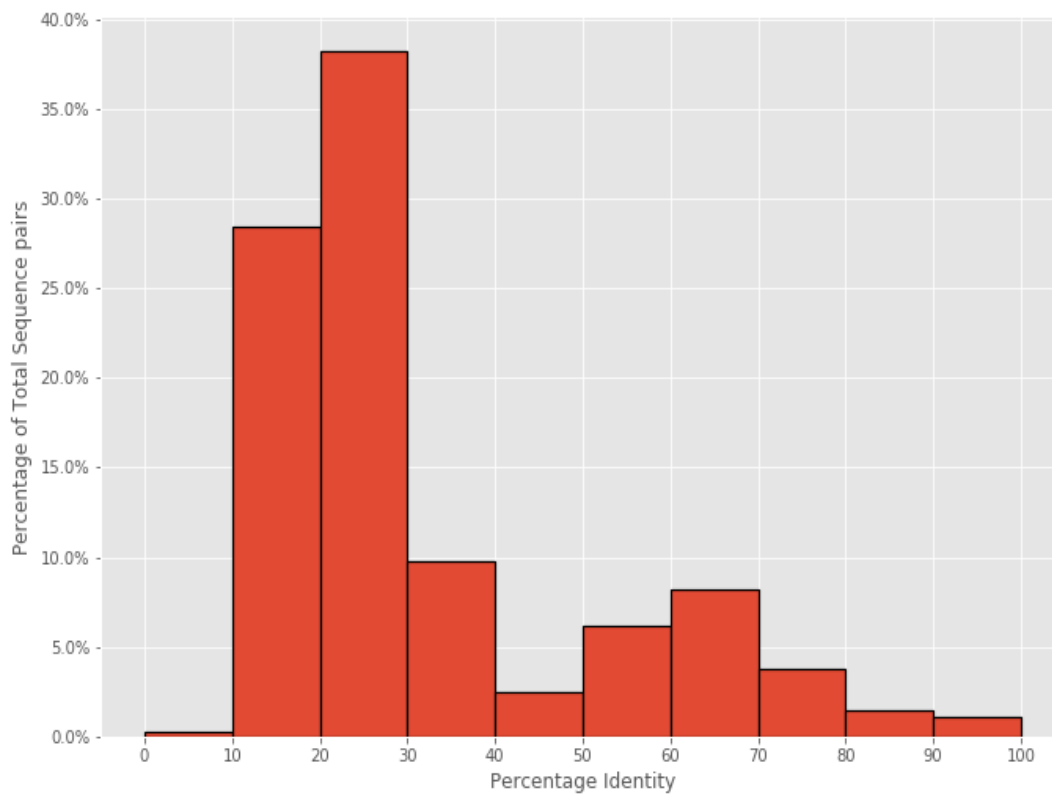
## **Model Generation**

Using in-house python scripts, feature calculation was carried out for features such as amino acid composition, dipeptide composition as well as several other features in 1:1 and 1:2 definitive ratios were calculated. After acquiring features different models were trained with SVM as a classifier, To check the efficiency of the models,<sup>2</sup> Accuracy refers to total no of correct prediction across total samples.

## **RESULTS AND DISCUSSION**

### **Data Visualization and Initial Analysis**

502 sequences were downloaded from NCBI. After which MSA was carried out and the corresponding percentage identity matrix was used to plot histogram which yielded following results which conveyed that 37.5% of sequence pair lie between 20-30 per cent identity whereas 27.5% data lie between 10-20 per cent identity which invariably shows diversity across dataset as well as warrants the use of ML-based methods so as to give more accurate results. Thereafter CD-Hit was performed resulting in 126 sequences below 0.6 sequence cut-off, 88 sequences below 0.5 cut-offs and 57 sequences below 0.4 cut-offs.



**Fig-1:** Histogram depicting data variability across the dataset

**Table-3**

(Tabular description of the entire dataset)

<b>Initial count</b>	<b>CD-Hit60</b>	<b>CD-Hit50</b>	<b>CD-Hit40</b>
502	126	88	47

## Generated Models

The features were used to train the model using an in-house python script which yielded the following results, the analysis of the result shows that this model could be used for prediction of expansin proteins. This shows much better results when compared to conventional methods and highlights the importance or rather the need for using machine learning-based techniques.

## Table-4

(Represents results generated after model training done)

Feature	Accuracy	Precision	Recall	ROC Area
Aa_comp(1:1)	93.8571 %	0.931	0.929	0.929
Aa_comp(1:2)	90.3852 %	0.907	0.914	0.909
Dip_comp(1:1)	91.6667 %	0.917	0.907	0.917
Dip_comp(1:2)	82.0635 %	0.821	0.851	0.911

## CONCLUSION

In this current study, we have carried out a machine learning-based approach using in-house scripts to develop a predictor which will ease the identification and annotation of expansin proteins by correlating sequence and structural information. This predictor has been built using sequential data collected from NCBI protein and structural information from PDB. After which data cleaning and filtration was carried out. Thereafter feature calculation for features such as amino acid composition, dipeptide composition and several others was carried out. The final SVM model has been trained with 1:2 data of amino acid composition and 1:1 data of dipeptide composition. With the advancement in expansin based study in future, the predictor will help in



correlating sequence-structure based studies as well as make the predictor even more robust and relevant with increased information pertaining to sequence and structure. Making it a go-to site for prediction pertaining to it.

## REFERENCES

1. Cosgrove DJ: Growth of the plant cell wall. *Nat Rev Mol Cell Biol* 2005, 6:850-861.
2. Foster, Adriance Sherwood, and Ernest Milton Gifford. *Morphology and evolution of vascular plants*. San Francisco, USA: WH Freeman and Company, 1989.
3. Cassab, Gladys I., and Joseph E. Varner. "Cell wall proteins." *Annual Review of Plant Physiology and Plant Molecular Biology* 39.1 (1988): 321-353.
4. Korn, Robert W. "Positional specificity within plant cells." *Journal of Theoretical Biology* 95.3 (1982): 543-568.
5. Jun, Tang, et al. "Extracellular calmodulin-binding proteins in plants: purification of a 21-kDa calmodulin-binding protein." *Planta* 198.4 (1996): 510-516.
6. Cosgrove, Daniel J. "Assembly and enlargement of the primary cell wall in plants." *Annual review of cell and developmental biology* 13.1 (1997): 171-201.
7. McQueen-Mason, Simon, Daniel M. Durachko, and Daniel J. Cosgrove. "Two endogenous proteins that induce cell wall extension in plants." *The Plant Cell* 4.11 (1992): 1425-1433.
8. Cosgrove, Daniel J. "Re-constructing our models of cellulose and primary cell wall assembly." *Current opinion in plant biology* 22 (2014): 122-131.
9. Cosgrove, Daniel J. "Characterization of long-term extension of isolated cell walls from growing cucumber hypocotyls." *Planta* 177.1 (1989): 121-130.
10. McQueen-Mason, Simon, and Daniel J. Cosgrove. "Disruption of hydrogen bonding between plant cell wall polymers by proteins that induce wall extension." *Proceedings of the National Academy of Sciences* 91.14 (1994): 6574-6578.
11. Rayle, David L., and Robert E. Cleland. "The Acid Growth Theory of auxin-induced cell elongation is alive and well." *Plant physiology* 99.4 (1992): 1271-1274.
12. Marowa, Prince, Anming Ding, and Yingzhen Kong. "Expansins: roles in plant growth and potential applications in crop improvement." *Plant cell reports* 35.5 (2016): 949-965.

13. Sampedro, Javier, et al. "Evolutionary divergence of  $\beta$ -expansin structure and function in grasses parallels emergence of distinctive primary cell wall traits." *The Plant Journal* 81.1 (2015): 108-120.
14. Boser, Bernhard E., Isabelle M. Guyon, and Vladimir N. Vapnik. "A training algorithm for optimal margin classifiers." *Proceedings of the fifth annual workshop on Computational learning theory*. ACM, 1992.
15. Schölkopf, Bernhard, Alexander J. Smola, and Francis Bach. *Learning with kernels: support vector machines, regularization, optimization, and beyond*. MIT press, 2002.
16. Vapnik, Vladimir. *The nature of statistical learning theory*. Springer science & business media, 2013.
17. Muller, K-R., et al. "An introduction to kernel-based learning algorithms." *IEEE transactions on neural networks* 12.2 (2001): 181-201.
18. Schölkopf, Bernhard, Koji Tsuda, and Jean-Philippe Vert. *Support vector machine applications in computational biology*. MIT press, 2004.
19. Wei, Peng-Cheng, et al. "Regulation of stomatal opening by the guard cell expansin AtEXPA1." *Plant signaling & behavior* 6.5 (2011): 740-742.
20. Yan, An, et al. "AtEXP2 is involved in seed germination and abiotic stress response in Arabidopsis." *PloS one* 9.1 (2014): e85208.
21. Kwon, Ye Rim, et al. "Ectopic expression of Expansin3 or Expansin $\beta$ 1 causes enhanced hormone and salt stress sensitivity in Arabidopsis." *Biotechnology letters* 30.7 (2008): 1281-1288.
22. Mollet, Jean-Claude, et al. "Cell wall composition, biosynthesis and remodeling during pollen tube growth." *Plants* 2.1 (2013): 107-147.
23. Cosgrove, Daniel J., et al. "The growing world of expansins." *Plant and Cell Physiology* 43.12 (2002): 1436-1444.
24. Rose, Jocelyn KC, et al. "Detection of expansin proteins and activity during tomato fruit ontogeny." *Plant physiology* 123.4 (2000): 1583-1592.

25. Brummell, David A., et al. "Modification of expansin protein abundance in tomato fruit alters softening and cell wall polymer metabolism during ripening." *The Plant Cell* 11.11 (1999): 2203-2216.
26. Chen, Feng, PeetambarDahal, and Kent J. Bradford. "Two tomato expansin genes show divergent expression and localization in embryos during seed development and germination." *Plant Physiology* 127.3 (2001): 928-936.
27. Minoia, Silvia, et al. "Induced mutations in tomato SExp1 alter cell wall metabolism and delay fruit softening." *Plant science* 242 (2016): 195-202.
28. Cho, Hyung-Taeg, and Hans Kende. "Expression of expansin genes is correlated with growth in deepwater rice." *The Plant Cell* 9.9 (1997): 1661-1671.
29. Choi, Dongsu, et al. "Regulation of expansin gene expression affects growth and development in transgenic rice plants." *The Plant Cell* 15.6 (2003): 1386-1398.
30. Marowa, Prince, Anming Ding, and Yingzhen Kong. "Expansins: roles in plant growth and potential applications in crop improvement." *Plant cell reports* 35.5 (2016): 949-965.
31. Zou, Hanyan, et al. "OsEXPB2, a  $\beta$ -expansin gene, is involved in rice root system architecture." *Molecular breeding* 35.1 (2015): 41.
32. Ma, Nana, et al. "Overexpression of OsEXPA8, a root-specific gene, improves rice growth and root system architecture by facilitating cell extension." *PLoS One* 8.10 (2013): e75997.
33. YU ZM, KANG B. "Root hair-specific expansins modulate root hair elongation in rice." *The Plant Journal* 66.5 (2011): 725-734.
34. Palapol, Yossapol, et al. "Expression of expansin genes in the pulp and the dehiscence zone of ripening durian (*Duriozibethinus*) fruit." *Journal of plant physiology* 182 (2015): 33-39.
35. Kuluev, B. R., et al. "Effect of ectopic expression of NtEXPA5 gene on cell size and growth of organs of transgenic tobacco plants." *Russian journal of developmental biology* 44.1 (2013): 28-34.
36. Civello, Pedro Marcos, et al. "An expansin gene expressed in ripening strawberry fruit." *Plant Physiology* 121.4 (1999): 1273-1279.

37. Asif, Mehar Hasan, et al. "Transcriptome analysis of ripe and unripe fruit tissue of banana identifies major metabolic networks involved in fruit ripening process." *BMC plant biology* 14.1 (2014): 316.
38. Xu, Qian, et al. "Transgenic tobacco plants overexpressing a grass PpEXP1 gene exhibit enhanced tolerance to heat stress." *PLoS One* 9.7 (2014): e100792.
39. Lü, Peitao, et al. "RhEXPA4, a rose expansin gene, modulates leaf growth and confers drought and salt tolerance to Arabidopsis." *Planta* 237.6 (2013): 1547-1559.
40. Sampedro, Javier, et al. "Use of genomic history to improve phylogeny and understanding of births and deaths in a gene family." *The Plant Journal* 44.3 (2005): 409-419.
41. Sampedro, Javier, Robert E. Carey, and Daniel J. Cosgrove. "Genome histories clarify evolution of the expansin superfamily: new insights from the poplar genome and pine ESTs." *Journal of plant research* 119.1 (2006): 11-21.
42. Dal Santo, Silvia, et al. "Genome-wide analysis of the expansin gene superfamily reveals grapevine-specific structural and functional characteristics." *PLoS One* 8.4 (2013): e62206.
43. Zhu, Yan, et al. "Soybean (*Glycine max*) expansin gene superfamily origins: segmental and tandem duplication events followed by divergent selection among subfamilies." *BMC plant biology* 14.1 (2014): 93.
44. Zhang, Shizhong, et al. "A genome-wide analysis of the expansin genes in *Malus domestica*." *Molecular genetics and genomics* 289.2 (2014): 225-236.
45. Krishnamurthy, Panneerselvam, et al. "Genome-wide analysis of the expansin gene superfamily reveals Brassica rapa-specific evolutionary dynamics upon whole genome triplication." *Molecular Genetics and Genomics* 290.2 (2015): 521-530.
46. Zhang, Wei, et al. "Genome-wide identification and characterization of maize expansin genes expressed in endosperm." *Molecular genetics and genomics* 289.6 (2014): 1061-1074.

47. Carey, Robert E., Nathan K. Hepler, and Daniel J. Cosgrove. "Selaginella moellendorffii has a reduced and highly conserved expansin superfamily with genes more closely related to angiosperms than to bryophytes." *BMC plant biology* 13.1 (2013): 4.
48. Carey, Robert E., and Daniel J. Cosgrove. "Portrait of the expansin superfamily in *Physcomitrella patens*: comparisons with angiosperm expansins." *Annals of Botany* 99.6 (2007): 1131-1141.
49. Huang, Ying, et al. "CD-HIT Suite: a web server for clustering and comparing biological sequences." *Bioinformatics* 26.5 (2010): 680-682.
50. Altschul, Stephen F., et al. "Gapped BLAST and PSI-BLAST: a new generation of protein database search programs." *Nucleic acids research* 25.17 (1997): 3389-3402.
51. Holmes, Geoffrey, Andrew Donkin, and Ian H. Witten. "Weka: A machine learning workbench." (1994).