Computer annotation of plant protein sequences based on sequence similarity and orthology

Malyugin E.[1,2]*, Mustafin Z. [2,3], Pronozin A. [2,3], Genaev M. [2,3], Afonnikov D.[2,3]

[1] *Novosibirsk State University, Novosibirsk, Russia*

[2] *Institute of Cytology and Genetics, SB RAS, Novosibirsk, Russia*

[3] *Kurchatov Genomic Center of the Institute of Cytology and Genetics, SB RAS, Novosibirsk, Russia*

* *evgeny.malyugin98@gmail.com*

Introduction

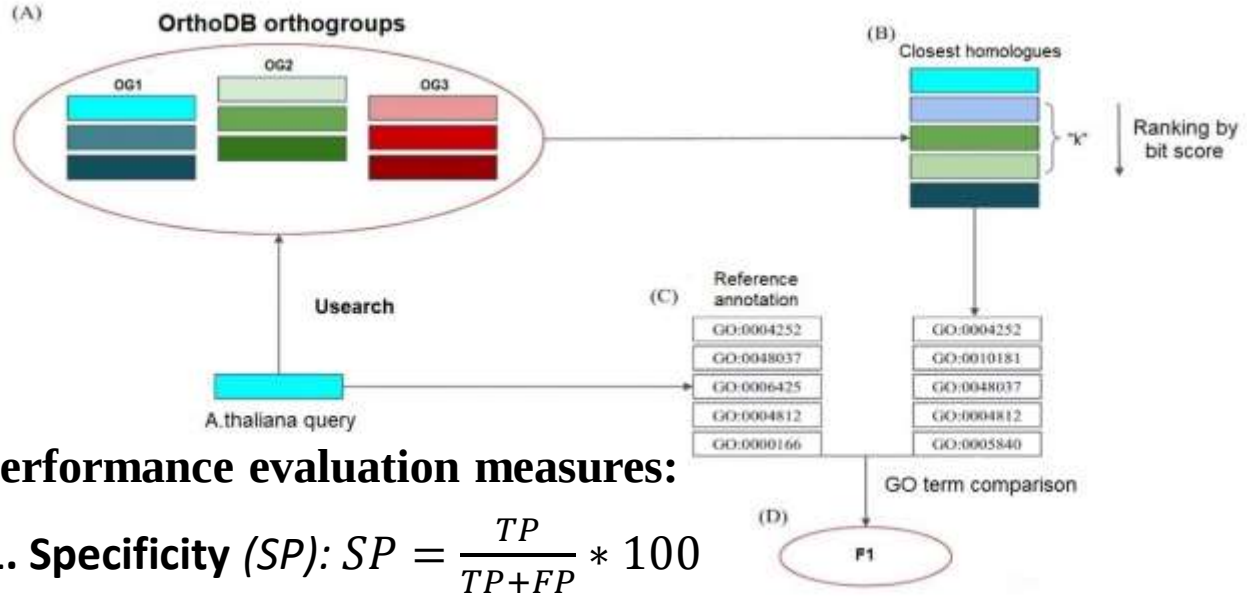
The number of protein-coding sequences is growing at a tremendous pace and requires efficient prediction of their function. Most methods use sequence homology search for functional annotation. In this case, an important factor is the gene age. It depends on the presence of gene orthologs in a large number of taxa. For genes that have appeared recently during evolution (young or taxon-specific genes), orthologs are absent or poorly represented in other organisms. These young genes are usually hard to annotate with reasonable accuracy unlike the ancient genes, for which homologs are widely represented

Our aim was to develop a method for annotation of gene functions that would work with high accuracy regardless of the age of the genes

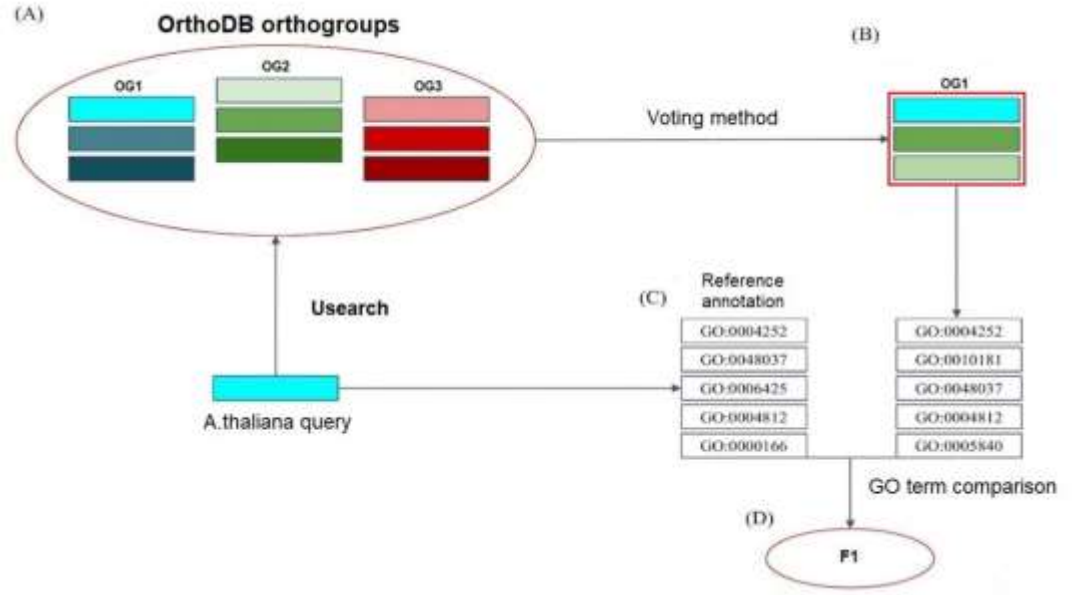
Three algorithms for gene function annotation

We developed three algorithms for annotation of proteins using search for homologs/orthologs in the OrthoDB database by Usearch algorithm. GO terms from hits were transferred to query sequence and compared with the known annotation of *A.thaliana* proteins. The performance was evaluated by *F1* measure on the basis of matching GO terms.

1) Method *k*-nearest neighbors (KNN)



2) Method using orthogroups annotation (OG)



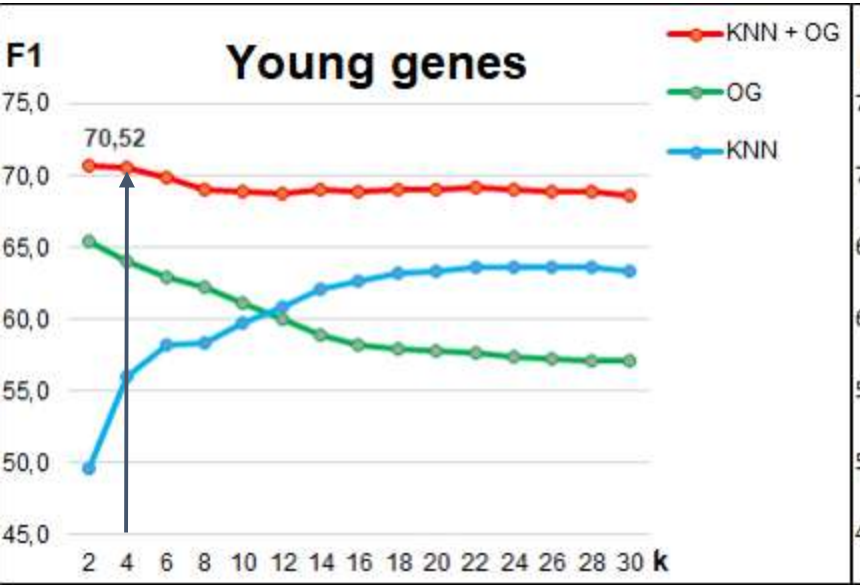
Performance evaluation measures:

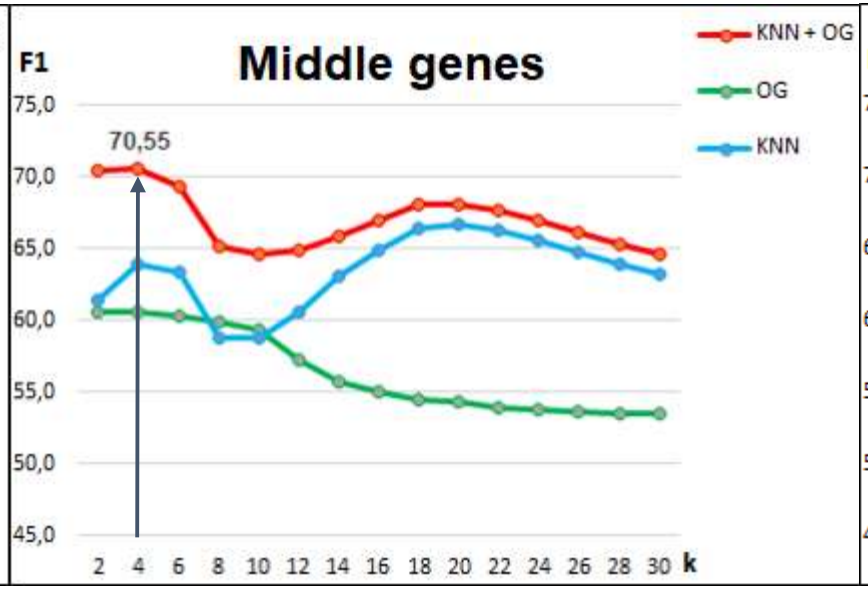
1. **Specificity** *(SP)*: $SP = \frac{TP}{TP+FP} * 100$
2. **Sensitivity** *(SN)*: $SN = \frac{TP}{TP+FN} * 100$
3. **Accuracy** *(AC)*: $AC = \frac{SN+SP}{2} * 100$
4. **F1 Score** *(F1)*: $F1 = 2\frac{SP*SN}{SP+SN} * 100$

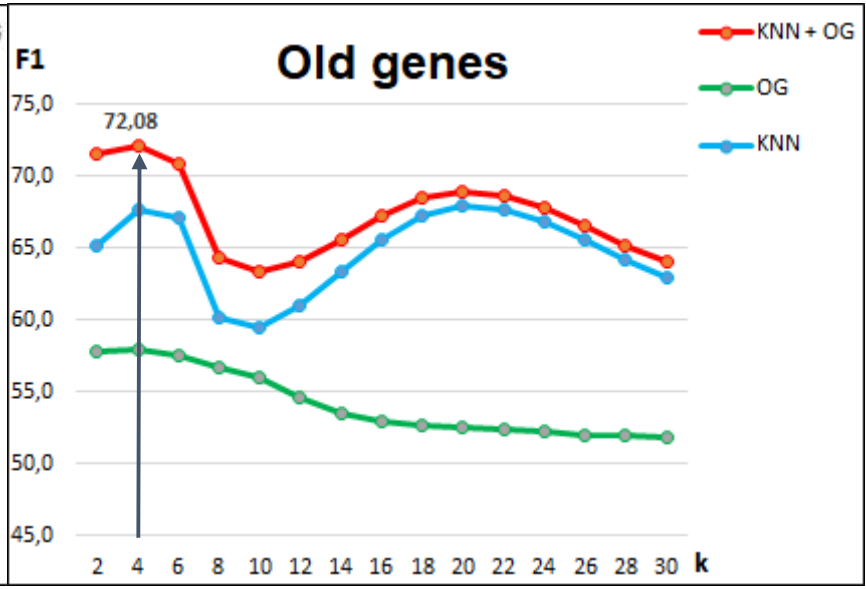
3) In addition, we used the combinations of the two previous methods (KNN+OG**): GO terms obtained by methods (1) and (2) were merged.**

Comparison of prediction methods

The age of genes was calculated by Orthoscape tool (Mustafin et al., 2019). Genes were split into three categories by age: old (14528), middle (6699) and young (4671). We changed the number of closest homologs (*k*) searching for its optimal values. The optimal values were achieved at hit similarity > 30%, coverage > 60%.







The combination of the results of KNN and OG methods (method KNN + OG) gives a fundamental improvement in the accuracy of function recognition for genes of all ages. At $k = 4$ it allows to decrease the effect of the gene age on the performance of their annotation. The performance becomes approximately the same and approaches 70,5-72%.

Comparison with Blast2GO

	KNN+OG				Blast2GO			
	SN	**SP**	**AC**	**F1**	**SN**	**SP**	**AC**	**F1**
Young genes	61.49	82.66	72.08	**70.52**	42.55	66.41	54.48	**51.86**
Middle genes	58.22	89.50	73.86	**70.55**	46.02	72.97	59.50	**56.44**
Old genes	58.94	92.75	75.85	**72.08**	46.39	76.86	61.62	**57.85**

In general, KNN+OG method is more accurate than Blast2GO by more than 14% using *A. thaliana* proteins as a benchmark.

Conclusion

An OG method is proposed for predicting gene functions, taking into account information about query sequence homologs and orthological group. In combination with KNN results, this gives a better performance and decrease the effect of gene age on the annotation.

Comparison of the proposed method with the Blast2GO method showed that the *F1* measure of our approach exceeds that of Blast2GO by 14-18%.

Acknowledgements

Development of algorithms was funded by the Kurchatov Genome Center of the Institute of Cytology and Genetics of Siberian Branch of the Russian Academy of Sciences, agreement with the Ministry of Education and Science of the Russian Federation, no. 075-15-2019-1662

Thank you for your attention!