An enumeration-based putative dyad predicting algorithm for promoter analysis in plants

Doctoral (Ph.D.) thesis

Mátyás Cserháti

Supervisor: Dr. Sándor Pongor and Dr. Györgyey János

Biological Research Center of the HAS, Institute of Plant Biology

University of Szeged

Szeged

2011

1. Introduction

A large number of genes play a role in abiotic stress response, since this affects a large part of the plant's physiology. Plants respond to stress in two basic ways: they either try to return to their former physiological status, or try to adapt to their changed environments. Abiotic stress is defined as certain environmental conditions, which reduce the plant's water potential, such as cold, drought, salt, or osmotic stress. Stress signals are transmitted through the plasma membrane (many times due to hormones such as ABA, cytokines, or ethylene). The signal is transmitted within the cell by secondary messengers (e.g. ROS, Ca^{2+}, or IP molecules). The complex interplay between many different kinds of transcription factors within the nucleus is responsible for changes in gene expression levels in response to a given form of stress.

Abiotic stress response in plants usually follows two basic pathways, one dependent from the plant hormone ABA, and one which is independent from it. There are significant overlaps between these two pathways, as well interactions between common transcription factors and transcription factor binding sites.

2. Objectives

Since genes which take part in abiotic stress response undergo similar regulation, we may assume that common regulatory elements can be found in their promoter regions. Since we are dealing with a complex molecular genetic phenomenon, we may also assume that a number of genes, transcription factors and transcription factor binding sites have an integrated effect on each other.

Until now a number of motif discovery programs have been developed for the analysis of DNA sequences and the prediction of DNA motifs. These are usually capable of finding short oligonucleotide motifs. According to studies done by Tompa et al., the sensitivity of a number of well-known motif discovery programs was defined to be around 0.22, therefore there is room for improving these algortihms. Since until now no such algorithm existed for the discovery of regulatory elements in co-regulated promoters, we decided to develop this kind of algorithm. The algorithm can be used in the promoter analysis of a number of agricultural crops (e.g. barley, rice, wheat) among other species.

One of the main results of the algorithm is that it is capable of predicting a number of putative optimal dyads in input promoter sets. The studied organism's promoterome can afterwards be analyzed with these optimal dyads in order to find other promoters which contain a large number of these dyads, and therefore undergo similar transcriptional regulation. The genes of these promoters can be predicted to take part in similar processes as the original genes that the analysis started out from.

3. Description of the algorithm

The algorithm searches for dyad sequences, which can be described with the formula $M_1N_nM_2$ where M_1 denotes the head motif, and M_2 the tail motif. In between them is a well-defined spacer region which is n bp long, with slight wobbling allowed. The head and tail motif are the same length, while the spacer region can be 0-52 bp long.

The algorithm is made up of a number of phases. The first phase deals with selecting the proper co-regulated genes as well as determining their promoter sequences. Afterwards we split up the promoters into different sets. The promoters are usually 2 Kbp long shorter if they overlap with upstream genes. The promoters are to be divided into a positive and negative learning set and a positive and negative test set.

The next phase is the learning phase, where the algorithm counts the total number of occurences of all possible dyads in the positive and negative learning sets. Afterwards the algorithm calculates the dyads' weight, or cdr value (cumulative difference ratio), described below, and then ranks them for further analysis and use. The cdr score is calculated as follows:

$$cdr = \frac{N_{positive} - N_{negative}}{N_{positive}}$$

Here $N_{positive}$ is the number of promoters from the positive learning set that a given dyad occurs in, and $N_{negative}$ is the number of promoters in which the dyad also occurs. The cdr score takes up a value between $-\infty$ and 1 (only those cases can be taken into account where $N_{positive}$ is greater than 1). The greater the cdr score of a dyad, the more relevant role it plays in the mechanism under study (in our case abiotic stress).

After the learning phase comes the test phase where the dyads are analyzed according to a number of different parameters in order to select the optimal set. These parameters are: occurence in the positive learning set, the possible wobbling of the spacer region (up to ±5bp), and the minimal cdr value for all selected dyads. All dyad sets were found back in the positive and negative test promoter sets. ROC analysis was used to

determine the optimal dyad set which was used in a promoterome search. Promoters found during the promoterome search were scored and ranked according to their optimal dyad content.

4. Results

The algorithm was first developed and verified in the case of Arabidopsis and then used in rice; in the promoterome study of one dicot and one monocot. In the case of Arabidopsis 125 promoters were put into the positive and negative learning sets, and 44 into the positive and negative test sets. In rice 87 promoters were put into both learning sets, while 42 were put into the positive test set and 56 into the negative test set. These genes were selected based on their invovlement in abiotic stress, which was determined based on either their annotation or data from the Genvestigator database.

In Arabidopsis we found 81 putative optimal dyads with a minimum occurence of 14 in the positive learning set, a minimum cdr value of 0.9, and a ±2bp wobbling. In rice 38 optimal dyads were found with a minimum occurence of 9 in the positive learning set, a minimum cdr value of 0.89, with no wobbling. The 81 optimal Arabidopsis dyads were clustered into 11 groups, based on how similar two given dyads were to each other. A Hamming distance was calculated for each dyad pair, with a maximum similarity of 10 (since we studied pentamer pairs). The maximum Hamming distance was 3.

According to the individual promoterome searches, we studied the top 3100 Arabidopsis promoters and the top 4600 rice promoters. The reason these numbers of top promoters were selected for both organisms was because here the ratio of non-stress promoters to all promoters was the lowest. According to the promoterome search in Arabidopsis, 78.6% of the found promoters were shown to be involved in abiotic stress response. This means that 49 hypothetical genes and 1224 genes (1273 in total) without any chip data were newly predicted to be involved in abiotic stress. In rice, 98.7% of the genes were shown to be involved in abiotic stress, meaning that 1245 hypothetical genes and 1437 genes without an Affymetrix id were also predicted to be involved in abiotic stress (2682 in total).

38 of the 81 dyads found in Arabidopsis were clustered into 11 groups. 7 clusters and 5 individual dyads were found to be play an important role in the network-type regulation of 5 cor and 4 erd genes. 1224 tentative REPs (Regulatory Element Pairs) (with a modified cdr score above 0.5) play a role in abiotic stress response.

In Arabidopsis we performed a promoterome search in order to find promoters with significant REP content. We calculated the Jacquard coefficient between each gene and each *cor* gene. Based on this we calculated the difference in REP content between each gene and each of the *cor* genes. We selected those 25 promoters whose difference in REP content was below 0.5. In this way we found 1 hypothetic gene and 5 genes of unknown function which could then be newly annotated to have a function similar to *cor* genes.

In the case of the 30 rice aldo-keto reductase (AKR) genes we found 28 new putative dyads which occured in at least 7 of the input promoters, with a minimum cdr value of 0.9. We studied three of the 30 AKR genes in detail, and found that one of them (AKR1, or Os01g0847600) contained more dyads than the other ones (AKR2, and AKR3), which were shown experimentally to be induced by osmotic stress to a lesser degree. In this way also we were capable of independently verifying our algorithm.

We studied tetrad dyads in the promoter region of 91 genes belonging to six rice gene familes (glucanases, chtinases, PR1, PR4, PR5, and PR9 genes). These genes were homologous to certain wheat genes which play a role in biotic stress response. We found many motifs in these promoter regions which have a match in the PlantCARE database (e.g. the W1-box, the EIRE, and a WUN-motif). Because of homology we assumed that we would be able to find similar regulatory elements in the promoter regions of the rice genes homologous to the wheat genes. We studied the promoters of the wheat genes and found that half of the predicted dyads match those in rice with slight modifications.

We ran the algorithm on these 91 rice homolog biotic stress promoters. 13 of these were used as a positive learning set since they were the best scoring homologs, while the rest were put into the negative learning set. Overall 263 dyads were found

which had a minimum cdr score of 0.9. 28 dyads were found which occured in promoters of at least 4 of the 6 gene families, therefore their occurence was taken to be statistically significant. Motifs matching these dyads were also found in the PLACE database.

We compared our algorithm with two well-known motif finding algorithms, YMF and dyad-analysis. We ran these two programs on the 125 stress learning promoters from Arabidopsis. With YMF we found 283 promoters which contained a substantial amount of putative regulatory elements. Out of these, only 3 belonged to the original 125 positive learning promoters. 3.1% were shown to be stress-inducible based on data from the Genevestigator database. The dyad-analysis program found 149 promoters, which contained a substantial amount of regulatory elements, amongst which only 1 of them belonged to the original set of 125 positive learning promoters. 3.6% of these were shown to be involved in abiotic stress according to data in the Genevestigator database. These results show that our algorithm is much more capable of finding putative regulatory elements which are involved in abiotic stress response, as well as discovering the involvement of further genes in physiological processes in which the original co-regulated genes take part in, and whose promoters contain a large number of such regulatory elements.

The algorithm was run in a 64 bit IRIX64 programming environment using a combination of awk (GNU Awk 3.1.5), C shell, and C (GCC 3.4.6) scripts. The algorithm can be dowloaded from its own website (which a short description) in the form of a desktop application: http://bhd.szbk.u-szeged.hu/dyadscan/. Input parameters include a positive and negative learning promoter set, length of motifs, maximum length of the spacer region, minimum occurence of dyads in the positive learning set, and minimum cdr value. The program's output is a list of dyads which meet the input criteria. In the output the dyads' sequence, occurence in the positive and negative learning sets, and the cdr score is given.

Publications, which form a basis of the Ph.D. thesis:

Turóczy, Z., Kis, P., Török, K., **Cserháti, M.**, Lendvai, Á., Dudits, D., and Horváth, G: Overproduction of a rice aldo-keto reductase increases oxidative and heat stress tolerance by malondialdehyde and methylglyoxal detoxification, *Plant Molecular Biology*, 2011

Cserháti M., Turóczy, Z., Zombori, Z., Cserző, M., Dudits, D., Pongor, S., Györgyey, J.: Prediction of new abiotic stress genes in Arabidopsis thaliana and Oryza sativa according to enumeration-based statistical analysis, *Molecular Genetics and Genomics*, 2011.

Cserháti, M., Pongor, S. and Györgyey, J: Statistical methods for finding biologically relevant motifs in promoter regions and a few of its implementations, In: 5th International Conference of PhD Students, University of Miskolc, Hungary, 14-20 August 2005, (Eds L. Lehoczky and L. Kalmár) Published by University of Miskolc, Innovation and Technology Transfer Centre, pp. 41-46, 2005

Cserháti, M., Pongor, S., Dudits, D., and Györgyey, J: (2006). „Enumerációs módszereken alapuló algoritmusok használata promóter motívumok keresésére." Tavaszi Szél 2006 conference. Kaposvár. ISBN 963 229 773 3

Cserháti M.: Usage of enumeration method based algorithms for finding promoter motifs in plant genomes. Acta Biol Szeged 2006, 50(3-4):145.

Veronika Pós, Klára Manninger, Krisztián Halász, Éva Hunyadi-Gulyás, Emília Szájli, **Mátyás Cserháti**, Huijun Duan, Katalin Medzihradszky, János Györgyey, Noémi Lukács: Proteomic changes of the wheat apoplast associated with resistance against leaf rust. 15th International Congress of the Hungarian Society for Microbiology: July 18-20, 2007, Eötvös Loránd University (Budapest, Hungary)

Pós Veronika, **Cserháti Mátyás**, Hunyadi-Gulyás Éva, Manninger Sándorné, Györgyey János, Medzihradszky Katalin, Lukács Noémi: KÖZÖS CISZ-REGULÁLÓ elemek LEVÉLROZSDA FERTŐZÉSSEL ASSZOCIÁLT BÚZA APOPLASZTFEHÉRJÉK GÉNEXPRESSZIÓJÁBAN. A Magyar Biokémiai Egyesület 2007. évi Vándorgyûlése 2007. augusztus 26-29. Debreceni Egyetem (Debrecen, Magyarország)

Other publications:

Cserháti, M. and Györgyey J. 2006. „Génkutatás i*n silico*", könyvfejezet: „Korszakváltás a molekuláris biológiában" c. könyvben. Szerkesztő: Dudits Dénes.

Dudits, D., **Cserháti, M.**, Miskolczi, P., Horváth, G. The growing family of plant cyclin-dependant kinases with multiple functions in cellular and developmental regulation. 2006. Cell cycle control and plant development. Editor Dirk Inzé. Blackwell Publishing, Oxford.

Cserháti, M., Turóczy, Z., Dudits, D., Horváth, G., and Györgyey, J: Bioinformatic analysis of heptamer palindromes in rice stress promoters. 3rd EPSO vonference, Visegrád, poster.

Turóczy, Z., Kis, P., **Cserháti, M.** Dare to bet? –from the *in silico* predictions to the demonstration of stress induced gene expression. 7th Biologist Days, Cluj Napoca, Romania

Turóczy, Z., Kis, P., **Cserháti, M.**, Dudits, D., Horváth, G. Response of rice AKR genes to abiotic stresses: expression profiling and enzyme activity characterization. 3rd EPSO conference, Visegrád, poster.

Dénes, D., **Cserháti, M.**, Miskolczi, P., Fehér, A., Ayaydin, F. and Horváth, G. V.: Use of Alfalfa In Vitro Cultures in Studies on Regulation of Cyclin-Dependent Kinase (CDK) Functions. 2006. Proceedings of the 11th IAPTC&B Congress, Beijing. Editors: Z. Xu, J. Li, I.K. Vasil, Y. Xue and W. Yang.

Cserháti, M., Turóczy, Z., Sečenji, M., Pongor, S., Cserző, M., Dudits, D., Horváth V., G., Györgyey, J. Növényi promóterek analízise abiotikus stressz folyamatok megértésében. 2006. Straub napok előadás, November 15-17.

András Cseri, András Palágyi, **Mátyás Cserháti**, János Pauk, Dénes Dudits, Ottó Törjék: EcoTILLING analysis of drought related candidate genes in barley. Plant Abiotic Stress - from signaling to development, 2nd meeting of INPAS(Interrnational Network of Plant Abiotic Stress), 14-17 May 2009, Tartu, Estonia