PROJETO DE TCC PRINCIPAIS ARTIGOS

Aluno:	Josué Crispim Vitorino
Professora Orientadora:	Maria Angélica de Oliveira Camargo Brunetto

Sumário

1 Principais Periódicos e Eventos

Periódicos

Biology Direct

BMC Bioinformatics

Bioinformatics

Computational Statistics & Data Analysis

Computational Biology and Chemistry

Current Opinion in Plant Biology

Molecular Genetics and Genomics

Plant Physiology and Biochemistry

Plant Molecular Biology

The Plant Journal

Trends in Plant Science

Eventos

International Conference on Machine Learning and Applications (ICMLA)

Genomic Signal Processing and Statistics (GENSIPS)

2 Biology Direct

Classifying transcription factor targets and discovering relevant biological features

Dustin T Holloway, Mark Kon e Charles DeLisi

Biology Direct, Vol. 3, No. 1. (30 May 2008), 22. doi:10.1186/1745-6150-3-22

Abstract

Background

An important goal in post-genomic research is discovering the network of interactions between transcription factors (TFs) and the genes they regulate. We have previously reported the development of a supervised-learning approach to TF target identification, and used it to predict targets of 104 transcription factors in yeast. We now include a new sequence conservation measure, expand our predictions to include 59 new TFs, introduce a web-server, and implement an improved ranking method to reveal the biological features contributing to regulation. The classifiers combine 8 genomic datasets covering a broad range of measurements including sequence conservation, sequence overrepresentation, gene expression, and DNA structural properties.

Principal Findings

(1) Application of the method yields an amplification of information about yeast regulators. The ratio of total targets to previously known targets is greater than 2 for 11 TFs, with several having larger gains: Ash1(4), Ino2(2.6), Yaf1(2.4), and Yap6(2.4).

(2) Many predicted targets for TFs match well with the known biology of their regulators. As a case study we discuss the regulator Swi6, presenting evidence that it may be important in the DNA damage response, and that the previously uncharacterized gene YMR279C plays a role in DNA damage response and perhaps in cell-cycle progression.

(3) A procedure based on recursive-feature-elimination is able to uncover from the large

initial data sets those features that best distinguish targets for any TF, providing clues relevant to its biology. An analysis of Swi6 suggests a possible role in lipid metabolism, and more specifically in metabolism of ceramide, a bioactive lipid currently being investigated for anti-cancer properties.

(4) An analysis of global network properties highlights the transcriptional network hubs; the factors which control the most genes and the genes which are bound by the largest set of regulators. Cell-cycle and growth related regulators dominate the former; genes involved in carbon metabolism and energy generation dominate the latter.

Conclusion

Postprocessing of regulatory-classifier results can provide high quality predictions, and feature ranking strategies can deliver insight into the regulatory functions of TFs. Predictions are available at an online web-server, including the full transcriptional network, which can be analyzed using VisAnt network analysis suite.

In silico regulatory analysis for exploring human disease progression

Dustin T Holloway, Mark Kon e Charles DeLisi

Biology Direct, Vol. 3, No. 1. (18 June 2008), 24. doi:10.1186/1745-6150-3-24

Abstract

Background

An important goal in bioinformatics is to unravel the network of transcription factors (TFs) and their targets. This is important in the human genome, where many TFs are involved in disease progression. Here, classification methods are applied to identify new targets for 152 transcriptional regulators using publicly-available targets as training examples. Three types of sequence information are used: composition, conservation, and overrepresentation.

Results

Starting with 8817 TF-target interactions we predict an additional 9333 targets for 152 TFs. Randomized classifiers make few predictions (2/18660) indicating that our predictions for many TFs are significantly enriched for true targets. An enrichment score is calculated and used to filter new predictions. Two case-studies for the TFs OCT4 and WT1 illustrate the usefulness of our predictions:

• Many predicted OCT4 targets fall into the Wnt-pathway. This is consistent with known

biology as OCT4 is developmentally related and Wnt pathway plays a role in early development.

• Beginning with 15 known targets, 354 predictions are made for WT1. WT1 has a role in formation of Wilms' tumor. Chromosomal regions previously implicated in Wilms' tumor by cytological evidence are statistically enriched in predicted WT1 targets. These findings may shed light on Wilms' tumor progression, suggesting that the tumor progresses either by loss of WT1 or by loss of regions harbouring its targets.

• Targets of WT1 are statistically enriched for cancer related functions including metastasis and apoptosis. Among new targets are BAX and PDE4B, which may help mediate the established anti-apoptotic effects of WT1.

• Of the thirteen TFs found which co-regulate genes with WT1 ($p \leq 0.02$), 8 have been previously implicated in cancer. The regulatory-network for WT1 targets in genomic regions relevant to Wilms' tumor is provided.

Conclusion

We have assembled a set of features for the targets of human TFs and used them to develop classifiers for the determination of new regulatory targets. Many predicted targets are consistent with the known biology of their regulators, and new targets for the Wilms' tumor regulator, WT1, are proposed. We speculate that Wilms' tumor development is mediated by chromosomal rearrangements in the location of WT1 targets.

3 BMC Bioinformatics

Kernel based methods for accelerated failure time model with ultra-high dimensional data

Zhenqiu Liu, Dechang Chen, Ming Tan, Feng Jiang e Ronald B Gartenhaus

BMC Bioinformatics, Vol. 11, No. 1. (2010), 606. doi:10.1186/1471-2105-11-606

Abstract

Background

Most genomic data have ultra-high dimensions with more than 10,000 genes (probes). Regularization methods with L1 and Lp penalty have been extensively studied in survival analysis with high-dimensional genomic data. However, when the sample size n $\ll$ m (the number of genes), directly identifying a small subset of genes from ultra-high (m $>$ 10, 000) dimensional data is time-consuming and not computationally efficient. In current microarray analysis, what people really do is select a couple of thousands (or hundreds) of genes using univariate analysis or statistical tests, and then apply the LASSO-type penalty to further reduce the number of disease associated genes. This two-step procedure may introduce bias and inaccuracy and lead us to miss biologically important genes.

Results

The accelerated failure time (AFT) model is a linear regression model and a useful alternative to the Cox model for survival analysis. In this paper, we propose a nonlinear kernel based AFT model and an efficient variable selection method with adaptive kernel ridge regression. Our proposed variable selection method is based on the kernel matrix and dual problem with a much smaller n $\times$ n matrix. It is very efficient when the number of unknown variables (genes) is much larger than the number of samples. Moreover, the primal variables are explicitly updated and the sparsity in the solution is exploited.

Conclusions

Our proposed methods can simultaneously identify survival associated prognostic factors and predict survival outcomes with ultra-high dimensional genomic data. We have demonstrated the performance of our methods with both simulation and real data. The proposed method performs superbly with limited computational studies.

Learning gene regulatory networks from only positive and unlabeled data

Luigi Cerulo, Charles Elkan e Michele Ceccarelli

BMC Bioinformatics, Vol. 11, No. 1. (5 May 2010), 228. doi:10.1186/1471-2105-11-228

Abstract

Background

Recently, supervised learning methods have been exploited to reconstruct gene regulatory networks from gene expression data. The reconstruction of a network is modeled as a binary classification problem for each pair of genes. A statistical classifier is trained to recognize the relationships between the activation profiles of gene pairs. This approach has been proven to outperform previous unsupervised methods. However, the supervised approach raises open questions. In particular, although known regulatory connections can safely be assumed to be positive training examples, obtaining negative examples is not straightforward, because definite knowledge is typically not available that a given pair of genes do not interact.

Results

A recent advance in research on data mining is a method capable of learning a classifier from only positive and unlabeled examples, that does not need labeled negative examples. Applied to the reconstruction of gene regulatory networks, we show that this method significantly outperforms the current state of the art of machine learning methods. We assess the new method using both simulated and experimental data, and obtain major performance improvement.

Conclusions

Compared to unsupervised methods for gene network inference, supervised methods are potentially more accurate, but for training they need a complete set of known regulatory connections. A supervised method that can be trained using only positive and unlabeled data, as presented in this paper, is especially beneficial for the task of inferring gene regulatory networks, because only an incomplete set of known regulatory connections is available in public databases such as RegulonDB, TRRD, KEGG, Transfac, and IPA.

Network motif-based identification of transcription factor-target gene relationships by integrating multi-source biological data

Yuji Zhang, Jianhua Xuan, Benildo G de los Reyes, Robert Clarke e Habtom W Ressom

BMC Bioinformatics, Vol. 9, No. 1. (2008), 203. doi:10.1186/1471-2105-9-203

Abstract

Background

Integrating data from multiple global assays and curated databases is essential to understand the spatio-temporal interactions within cells. Different experiments measure cellular processes at various widths and depths, while databases contain biological information based on established facts or published data. Integrating these complementary datasets helps infer a mutually consistent transcriptional regulatory network (TRN) with strong similarity to the structure of the underlying genetic regulatory modules. Decomposing the TRN into a small set of recurring regulatory patterns, called network motifs (NM), facilitates the inference. Identifying NMs defined by specific transcription factors (TF) establishes the framework structure of a TRN and allows the inference of TF-target gene relationship. This paper introduces a computational framework for utilizing data from multiple sources to infer TF-target gene relationships on the basis of NMs. The data include time course gene expression profiles, genome-wide location analysis data, binding sequence data, and gene ontology (GO) information.

Results

The proposed computational framework was tested using gene expression data associated with cell cycle progression in yeast. Among 800 cell cycle related genes, 85 were identified as candidate TFs and classified into four previously defined NMs. The NMs for a subset of TFs are obtained from literature. Support vector machine (SVM) classifiers were used to estimate NMs for the remaining TFs. The potential downstream target genes for the TFs were clustered into 34 biologically significant groups. The relationships between TFs and potential target gene clusters were examined by training recurrent neural networks whose topologies mimic the NMs to which the TFs are classified. The identified relationships between TFs and gene clusters were evaluated using the following biological validation and statistical analyses: (1) Gene set enrichment analysis (GSEA) to evaluate the clustering results; (2) Leave-one-out cross-validation (LOOCV) to ensure that the SVM classifiers assign TFs to NM categories with high confidence; (3) Binding site enrichment analysis (BSEA) to determine enrichment

of the gene clusters for the cognate binding sites of their predicted TFs; (4) Comparison with previously reported results in the literatures to confirm the inferred regulations.

Conclusion

The major contribution of this study is the development of a computational framework to assist the inference of TRN by integrating heterogeneous data from multiple sources and by decomposing a TRN into NM-based modules. The inference capability of the proposed framework is verified statistically (e.g., LOOCV) and biologically (e.g., GSEA, BSEA, and literature validation). The proposed framework is useful for inferring small NM-based modules of TF-target gene relationships that can serve as a basis for generating new testable hypotheses.

Pol II promoter prediction using characteristic 4-mer motifs: a machine learning approach

Firoz Anwar, Syed Murtuza Baker, Taskeed Jabid, Md Mehedi Hasan, Mohammad Shoyaib, Haseena Khan e Ray Walshe

BMC Bioinformatics, Vol. 9, No. 1. (04 October 2008), 414. doi:10.1186/1471-2105-9-414

Abstract

Background

Eukaryotic promoter prediction using computational analysis techniques is one of the most difficult jobs in computational genomics that is essential for constructing and understanding genetic regulatory networks. The increased availability of sequence data for various eukaryotic organisms in recent years has necessitated for better tools and techniques for the prediction and analysis of promoters in eukaryotic sequences. Many promoter prediction methods and tools have been developed to date but they have yet to provide acceptable predictive performance. One obvious criteria to improve on current methods is to devise a better system for selecting appropriate features of promoters that distinguish them from non-promoters. Secondly improved performance can be achieved by enhancing the predictive ability of the machine learning algorithms used.

Results

In this paper, a novel approach is presented in which 128 4-mer motifs in conjunction with a non-linear machine-learning algorithm utilising a Support Vector Machine (SVM) are used to distinguish between promoter and non-promoter DNA sequences. By applying this approach to plant, Drosophila, human, mouse and rat sequences, the classification model has

showed 7-fold cross-validation percentage accuracies of 83.81%, 94.82%, 91.25%, 90.77% and 82.35% respectively. The high sensitivity and specificity value of 0.86 and 0.90 for plant; 0.96 and 0.92 for Drosophila; 0.88 and 0.92 for human; 0.78 and 0.84 for mouse and 0.82 and 0.80 for rat demonstrate that this technique is less prone to false positive results and exhibits better performance than many other tools. Moreover, this model successfully identifies location of promoter using TATA weight matrix.

Conclusion

The high sensitivity and specificity indicate that 4-mer frequencies in conjunction with supervised machine-learning methods can be beneficial in the identification of RNA pol II promoters comparative to other methods. This approach can be extended to identify promoters in sequences for other eukaryotic genomes.

Using hexamers to predict cis-regulatory motifs in Drosophila

Bob Y Chan e Dennis Kibler

BMC Bioinformatics, Vol. 6, No. 1. (27 October 2005), 262. doi:10.1186/1471-2105-6-262

Abstract

Background

Cis-regulatory modules (CRMs) are short stretches of DNA that help regulate gene expression in higher eukaryotes. They have been found up to 1 megabase away from the genes they regulate and can be located upstream, downstream, and even within their target genes. Due to the difficulty of finding CRMs using biological and computational techniques, even well-studied regulatory systems may contain CRMs that have not yet been discovered.

Results

We present a simple, efficient method (HexDiff) based only on hexamer frequencies of known CRMs and non-CRM sequence to predict novel CRMs in regulatory systems. On a data set of 16 gap and pair-rule genes containing 52 known CRMs, predictions made by HexDiff had a higher correlation with the known CRMs than several existing CRM prediction algorithms: Ahab, Cluster Buster, MSCAN, MCAST, and LWF. After combining the results of the different algorithms, 10 putative CRMs were identified and are strong candidates for future study. The hexamers used by HexDiff to distinguish between CRMs and non-CRM sequence were also analyzed and were shown to be enriched in regulatory elements.

Conclusion

HexDiff provides an efficient and effective means for finding new CRMs based on known CRMs, rather than known binding sites.

4 Bioinformatics

KIRMES: kernel-based identification of regulatory modules in euchromatic sequences

Sebastian J. Schultheiss,Wolfgang Busch,Jan U. Lohmann,Oliver Kohlbacher e Gunnar Rätsch

Bioinformatics , Vol. 25, No. 16. (15 August 2009), pp. 2126-2133.
doi:10.1093/bioinformatics/btp278

Abstract

Motivation

Understanding transcriptional regulation is one of the main challenges in computational biology. An important problem is the identification of transcription factor (TF) binding sites in promoter regions of potential TF target genes. It is typically approached by position weight matrix-based motif identification algorithms using Gibbs sampling, or heuristics to extend seed oligos. Such algorithms succeed in identifying single, relatively well-conserved binding sites, but tend to fail when it comes to the identification of combinations of several degenerate binding sites, as those often found in cis-regulatory modules.

Results

We propose a new algorithm that combines the benefits of existing motif finding with the ones of support vector machines (SVMs) to find degenerate motifs in order to improve the modeling of regulatory modules. In experiments on microarray data from Arabidopsis thaliana, we were able to show that the newly developed strategy significantly improves the recognition of TF targets.

Prediction of regulatory networks: genome-wide identification of transcription factor targets from gene expression data

Jiang Qian, Jimmy Lin1, Nicholas M. Luscombe, Haiyuan Yu2 e Mark Gerstein

Bioinformatics, Vol. 19, No. 15. (12 October 2003), pp. 1917-1926.
doi:10.1093/bioinformatics/btg347

Abstract

Motivation

Defining regulatory networks, linking transcription factors (TFs) to their targets, is a central problem in post-genomic biology. One might imagine one could readily determine these networks through inspection of gene expression data. However, the relationship between the expression timecourse of a transcription factor and its target is not obvious (e.g. simple correlation over the timecourse), and current analysis methods, such as hierarchical clustering, have not been very successful in deciphering them.

Results

Here we introduce an approach based on support vector machines (SVMs) to predict the targets of a transcription factor by identifying subtle relationships between their expression profiles. In particular, we used SVMs to predict the regulatory targets for 36 transcription factors in the Saccharomyces cerevisiae genome based on the microarray expression data from many different physiological conditions. We trained and tested our SVM on a data set constructed to include a significant number of both positive and negative examples, directly addressing data imbalance issues. This was non-trivial given that most of the known experimental information is only for positives. Overall, we found that 63% of our TF–target relationships were confirmed through cross-validation. We further assessed the performance of our regulatory network identifications by comparing them with the results from two recent genome-wide ChIP-chip experiments. Overall, we find the agreement between our results and these experiments is comparable to the agreement (albeit low) between the two experiments. We find that this network has a delocalized structure with respect to chromosomal positioning, with a given transcription factor having targets spread fairly uniformly across the genome.

5 Computational Statistics & Data Analysis

Informative transcription factor selection using support vector machine-based generalized approximate cross validation criteria

Insuk Sohn, Jooyong Shim, Changha Hwang, Sujong Kim e Jae Won Lee

Computational Statistics & Data Analysis, Vol. 53, No. 5. (15 March 2009), pp. 1727-1735. doi:10.1016/j.csda.2008.05.001

Abstract

The genetic regulatory mechanism plays a pivotal role in many biological processes ranging from development to survival. The identification of the common transcription factor binding sites (TFBSs) from a set of known co-regulated gene promoters and the identification of genes that are regulated by the transcription factor (TF) that have important roles in a particular biological function will advance our understanding of the interaction among the co-regulated genes and intricate genetic regulatory mechanism underlying this function. To identify the common TFBSs from a set of known co-regulated gene promoters and classify genes that are regulated by TFs, the new approaches using Support Vector Machine (SVM)-based Generalized Approximate Cross Validation (GACV) criteria are proposed. Two variable selection methods are considered for Recursive Feature Elimination (RFE) and Recursive Feature Addition (RFA). Performances of the proposed methods are compared with the existing SVM-based criteria, Logistic Regression Analysis (LRA), Logic Regression (LR), and Decision Tree (DT) methods by using both two real TF target genes data and the simulated data. In terms of test error rates, the proposed methods perform better than the existing methods.

6 Computatinal Biology and Chemistry

The cross-species prediction of bacterial promoters using a support vector machine

Michael Towsey, Peter Timms, James Hogan e Sarah A. Mathews

Computational Biology and Chemistry, Vol. 32, No. 5. (October 2008), pp. 359-366. doi:10.1016/j.compbiolchem.2008.07.009

Abstract

Due to degeneracy of the observed binding sites, the in silico prediction of bacterial sigma 70-like promoters remains a challenging problem. A large number of sigma 70-like promoters has been biologically identified in only two species, Escherichia coli and Bacillus subtilis. In this paper we investigate the issues that arise when searching for promoters in other species using an ensemble of SVM classifiers trained on E. coli promoters. DNA sequences are represented using a tagged mismatch string kernel. The major benefit of our approach is that it does not require a prior definition of the typical -35 and -10 hexamers. This gives the SVM classifiers the freedom to discover other features relevant to the prediction of promoters. We use our approach to predict sigma A promoters in B. subtilis and sigma 66 promoters in Chlamydia trachomatis. We extended the analysis to identify specific regulatory features of gene sets in C. trachomatis having different expression profiles. We found a strong -35 hexamer and TGN/-10 associated with a set of early expressed genes. Our analysis highlights the advantage of using TSS-PREDICT as a starting point for predicting promoters in species where few are known.

7 Current Opinion in Plant Biology

cis-Regulatory elements in plant cell signaling

Henry D Priest, Sergei A Filichkin e Todd C Mockler

Current Opinion in Plant Biology, Vol. 12, No. 5. (01 October 2009), pp. 643-649. doi:10.1016/j.pbi.2009.07.016

Abstract

Plant cell signaling pathways are in part dependent on transcriptional regulatory networks comprising circuits of transcription factors (TFs) and regulatory DNA elements that control the expression of target genes. Here, we describe experimental and bioinformatic approaches for identifying potential cis-regulatory elements. We also discuss recent integrative genomics studies aimed at elucidating the functions of cis-regulatory elements in aspects of plant biology, including the circadian clock, interactions with the environment, stress responses, and regulation of growth and development by phytohormones. Finally, we discuss emerging technologies and approaches that offer great potential for accelerating the discovery and functional characterization of cis-elements and interacting TFs–which will help realize the promise of systems biology.

'Omics' analyses of regulatory networks in plant abiotic stress responses

Kaoru Urano, Yukio Kurihara, Motoaki Seki, Kazuo Shinozaki

Current Opinion in Plant Biology, Vol. 13, No. 2. (April 2010), pp. 132-138. doi:10.1016/j.pbi.2009.1

Abstract

Plants must respond and adapt to abiotic stresses to survive in various environmental conditions. Plants have acquired various stress tolerance mechanisms, which are different processes involving physiological and biochemical changes that result in adaptive or morphological changes. Recent advances in genome-wide analyses have revealed complex regulatory networks

that control global gene expression, protein modification, and metabolite composition. Genetic regulation and epigenetic regulation, including changes in nucleosome distribution, histone modification, DNA methylation, and npcRNAs (non-protein-coding RNA) play important roles in abiotic stress gene networks. Transcriptomics, metabolomics, bioinformatics, and high-through-put DNA sequencing have enabled active analyses of regulatory networks that control abiotic stress responses. Such analyses have markedly increased our understanding of global plant systems in responses and adaptation to stress conditions.

8 Molecular Genetics and Genomics

Characterization of the TaAIDFa gene encoding a CRT/DRE-binding factor responsive to drought, high-salt, and cold stress in wheat

Zhao-Shi Xu, Zhi-Yong Ni, Li Liu, Li-Na Nie, Lian-Cheng Li, Ming Chen, You-Zhi Ma

Molecular Genetics and Genomics, Vol. 280, No. 6. (1 December 2008), pp. 497-508. doi:10.1007/s00438-008-0382-x

Abstract

Dehydration responsive element-binding factors (DBFs) belong to the AP2/ERF superfamily and play vital regulatory roles in abiotic stress responses in plants. In this study, we isolated three novel homologs of the DBF gene family in wheat (Triticum aestivum L.) by screening a drought-induced cDNA library and designated them as TaAIDFs (T. aestivum abiotic stress-induced DBFs). Compared to TaAIDFb and TaAIDFc, TaAIDFa lacks a short Ser/Thr-rich region, a putative phosphorylation site, following the AP2/ERF domain. The TaAIDFa gene, located on chromosome 3BS, is interrupted by a single intron at the 17th Arg (R) in the N-terminal domain. The N-terminal region of the TaAIDFa protein modulates nuclear localization. The TaAIDFa protein is capable of binding to CRT/DRE elements in vitro and in vivo, and of trans-activating reporter gene expression in yeast cells. The TaAIDFa promoter, with various stress-related cis-acting elements, drives expression of the GUS reporter gene in wheat calli under stress conditions. This was further confirmed by responses of TaAIDFa transcripts to drought, salinity, low-temperature, and exogenous ABA. Furthermore, overexpression of TaAIDFa activated CRT/DRE-containing genes under normal growth conditions, and improved drought and osmotic stress tolerances in transgenic Arabidopsis plants. These results suggested that TaAIDFa encodes a CRT/DRE element-binding factor that might be involved in multiple abiotic stress signal transduction pathways.

9 Plant Physiology and Biochemistry

Prediction of regulatory interactions in Arabidopsis using gene-expression data and support vector machines

Xiaoqing Yu, Taigang Liu, Xiaoqi Zheng, Zhongnan Yang e Jun Wang

Plant Physiology and Biochemistry (12 January 2011) doi:10.1016/j.plaphy.2011.01.002

Abstract

Identification of regulatory relationships between transcription factors (TFs) and their targets is a central problem in post-genomic biology. In this paper, we apply an approach based on the support vector machine (SVM) and gene-expression data to predict the regulatory interactions in Arabidopsis. A set of 125 experimentally validated TF-target interactions and 750 negative regulatory gene pairs are collected as the training data. Their expression profiles data at 79 experimental conditions are fed to the SVM to perform the prediction. Through the jackknife cross-validation test, we find that the overall prediction accuracy of our approach achieves 88.68%. Our approach could help to widen the understanding of Arabidopsis gene regulatory scheme and may offer a cost-effective alternative to construct the gene regulatory network.

Research highlights

* Gene-expression data and SVMs are explored to predict regulatory interactions in Arabidopsis. * Experimentally validated regulatory relationships were collected as the positive example. * Negative training examples were randomly selected TF-target pairs under some strategies. * Expression data of 79 experimental conditions were extracted as the sequence feature. * Through the jackknife test, the overall accuracy of our prediction achieved 88.68%.

10 Plant Molecular Biology

An in silico strategy identified the target gene candidates regulated by dehydration responsive element binding proteins (DREBs) in Arabidopsis genome

Shichen Wang, Shuo Yang, Yuejia Yin, Xiaosen Guo, Shan Wang e Dongyun Hao

Plant Molecular Biology, Vol. 69, No. 1. (1 January 2009), pp. 167-178. doi:10.1007/s11103-008-9414-5

Abstract

Identification of downstream target genes of stress-relating transcription factors (TFs) is desirable in understanding cellular responses to various environmental stimuli. However, this has long been a difficult work for both experimental and computational practices. In this research, we presented a novel computational strategy which combined the analysis of the transcription factor binding site (TFBS) contexts and machine learning approach. Using this strategy, we conducted a genome-wide investigation into novel direct target genes of dehydration responsive element binding proteins (DREBs), the members of AP2-EREBPs transcription factor super family which is reported to be responsive to various abiotic stresses in Arabidopsis. The genome-wide searching yielded in total 474 target gene candidates. With reference to the microarray data for abiotic stresses-inducible gene expression profile, 268 target gene candidates out of the total 474 genes predicted, were induced during the 24-h exposure to abiotic stresses. This takes about 57% of total predicted targets. Furthermore, GO annotations revealed that these target genes are likely involved in protein amino acid phosphorylation, protein binding and Endomembrane sorting system. The results suggested that the predicted target gene candidates were adequate to meet the essential biological principle of stress-resistance in plants.

11 The Plant Journal

Research on plant abiotic stress responses in the post-genome era: past, present and future

Takashi Hirayama e Kazuo Shinozaki

The Plant Journal, Vol. 61, No. 6. (March 2010), pp. 1041-1052. doi:10.1111/j.1365-313X.2010.04124.x

Abstract

Understanding abiotic stress responses in plants is an important and challenging topic in plant research. Physiological and molecular biological analyses have allowed us to draw a picture of abiotic stress responses in various plants, and determination of the Arabidopsis genome sequence has had a great impact on this research field. The availability of the complete genome sequence has facilitated access to essential information for all genes, e.g. gene products and their function, transcript levels, putative cis-regulatory elements, and alternative splicing patterns. These data have been obtained from comprehensive transcriptome analyses and studies using full-length cDNA collections and T-DNA- or transposon-tagged mutant lines, which were also enhanced by genome sequence information. Moreover, studies on novel regulatory mechanisms involving use of small RNA molecules, chromatin modulation and genomic DNA modification have enabled us to recognize that plants have evolved complicated and sophisticated systems in response to complex abiotic stresses. Integrated data obtained with various 'omics' approaches have provided a more comprehensive picture of abiotic stress responses. In addition, research on stress responses in various plant species other than Arabidopsis has increased our knowledge regarding the mechanisms of plant stress tolerance in nature. Based on this progress, improvements in crop stress tolerance have been attempted by means of gene transfer and marker-assisted breeding. In this review, we summarize recent progress in abiotic stress studies, especially in the post-genomic era, and offer new perspectives on research directions for the next decade.

12 Trends in Plant Science

Organization of -acting regulatory elements in osmotic- and cold-stress-responsive promoters

K. Yamaguchishinozaki, K. Shinozaki

Trends in Plant Science, Vol. 10, No. 2. (February 2005), pp. 88-94. doi:10.1016/j.tplants.2004.12.01

Abstract

cis-Acting regulatory elements are important molecular switches involved in the transcriptional regulation of a dynamic network of gene activities controlling various biological processes, including abiotic stress responses, hormone responses and developmental processes. In particular, understanding regulatory gene networks in stress response cascades depends on successful functional analyses of cis-acting elements. The ever-improving accuracy of transcriptome expression profiling has led to the identification of various combinations of cis-acting elements in the promoter regions of stress-inducible genes involved in stress and hormone responses. Here we discuss major cis-acting elements, such as the ABA-responsive element (ABRE) and the dehydration-responsive element/C-repeat (DRE/CRT), that are a vital part of ABA-dependent and ABA-independent gene expression in osmotic and cold stress responses.

13 Genomic Signal Processing and Statistics (GENSIPS)

Transcription Factor Discovery using Support Vector Machines and Heterogeneous Data

Barbe, J.F.; Tewfik, A.H.; Khodursky, A.B.;

(June 2007), pp. 1-4. doi:10.1109/GENSIPS.2007.4365812 Key: Barbe2007

Abstract

In this work we analyze the suitability of expression and sequence data for discovery of co-regulatory relationships using Support Vector Machines. In addition, we try to assess the possibility of improving such results by heterogeneous data fusion and by estimating a probability of a correct classification. As shown in other studies, we have found that transcription co-expression is a good estimator for genetic co-regulation. We also have found some evidence that operator site sequence motifs can be used to estimate co-regulation, but the kernels used for feature extraction did not achieve classification rates comparable to expression data. Finally, the additional information provided by combining sequence and expression data can be exploited to estimate the probability of correct classification.

14 International Conference on Machine Learning and Applications (ICMLA)

SVMotif: A Machine Learning Motif Algorithm

Mark Kon, Dustin T Holloway e Charles DeLisi

Machine Learning and Applications, 2007. ICMLA 2007. Sixth International Conference on (15 December 2007), pp. 573-580. doi:10.1109/ICMLA.2007.105

Abstract

We describe SVMotif, a support vector machine-based learning algorithm for identification of cellular DNA transcription factor (TF) motifs extrapolated from known TF-gene interactions. An important aspect of this procedure is its ability to utilize negative target information (examples of likely non-targets) as well as positive information. Applications involve situations where clusters of genes are distinguished in experiments with known transcription factors without known binding locations. We apply this to yeast TF data with target identifications from ChIP-chip and other sources, and compare performance with Gibbs sampling methods such as BioProspector. We verify that in yeast this method implies well-defined and cross-validated statistical correlations between TF binding and secondary motifs whose binding properties (either with the primary TF or other possible promoters) are not certain, and discuss some implications of this. SVMotif can be a useful standalone method or a complement to existing techniques, and it will be made publicly available.

Referências