Using string kernels to predict gene expression

Pavel P. Kuksa

NEC Laboratories America, Inc
pkuksa@nec-labs.com

In this work we focus on sequence-based gene expression prediction, i.e. inferring expression level from composition of protein sequence using expression data. In particular, we consider a task of predicting gene expression under specific environment conditions, e.g., nitrogen or sulfur deprivation, from sequence. This task is motivated by recent interest in sustainable sources of biofuels, and microalgae in particular, which is known to accumulate large quantities of oils (TAGs, triacylglycerols) when nutrient deprived [7, 2]. *Chlamydomonas reinhardtii* (*C. reinhardtii*) has been used as a model organism to study mechanisms of TAG accumulation as well as basic metabolism and physiology.

We used the complete genome of *C. reinhardtii* [1] (15497 sequences, annotation v4.0) and gene expression data under nitrogen deprivation condition [7]. We cast gene expression prediction as a classification problem: given gene/protein sequence, predict its expression (up-, down- regulation or no change); Table 1 gives an example of sequence-gene expression data.

State-of-the-art approaches to sequence classification rely on measuring sequence similarity using fixed-length representations $\Phi(X)$ of sequences as the $|\Sigma|^k$-dimensional histogram (spectra) of counts of short substrings (k-mers), contained, possibly with up to m mismatches, in a sequence, e.g.., spectrum/mismatch methods [5, 6]. These algorithms often exhibit state-of-the-art performance on tasks such as biological sequence classification, protein superfamily and fold prediction, etc. This work proposes and evaluates alternative approach for gene expression prediction using large sequence kernels [3, 4] to relate sequence composition and expression of the genes.

Experiments. We experimented with predicting changes in gene expression at (1) input level, for transcription factors (TFs) (dataset 1), and (2) whole-genome level, i.e. for all genes (dataset 2). Table 2 summarizes the two datasets used in this study.

Table 1: Gene expression as classification

#	Sequence	Expression	Class
1	MVAHGVPGLSRGLVGD..	10.7	up
2	MITGNARSRALTLCPQSLLKVTADALPAGRSVSWSQ..	1.2	no change
3	MGSSSVGTYHLLLVL..	-9.8	down

In the experiments reported below, we train our models on a subset of protein sequences for which expression level is known (train set), and use the learned model to predict the expression of other remaining protein (test set). We use cross-validation to measure prediction performance of the methods. As baselines, we use commonly used methods including PSI-BLAST and sequence composition approaches.

Table 2: Gene expression data (nitrogen deprivation)

data	# total	up	down	no-change
All proteins	15497	988	1181	13328
Transcription factors (TFs)	564	21	30	513

Table 3 compares classification performance of the mismatch string kernel and commonly used PSI-BLAST and amino acid sequence composition methods for predicting changes in gene expression for transcription factors following nitrogen deprivation. As can be seen from the results, using string kernel with k=11,m=5 results in significantly reduced error rates and increased recall rates. For example, the error rate 30% of the string kernel for predicting transcription factors with 2-fold change in gene expression (FC=2) is significantly better than the 33.7% or 37% error achieved by PSI-BLAST or by using sequence composition. Differences in prediction accuracy are even more pronounced when predicting major changes in gene expression (fold change FC=5 in Table 3). For instance, using the string kernel achieves 21% error rate compared to significantly higher error rate of 43.8% for PSI-BLAST.

For the whole-genome gene expression prediction (Table 4) we observe similar improvements using string kernels. Compared to error rates of 47.8% and 41% for PSI-BLAST and amino acid composition method, a smaller error rate of 37.4% is achieved.

Topic: data mining, pattern recognition, sequence modeling
Preference: oral/poster

Table 3: Gene expression prediction (10-fold cross-validation). Transcription factors (TFs)

Method	Balanced Error	Sensitivity
PSI-BLAST, FC[†]=2	33.68	43.33
Sequence composition, FC=2	36.94	47.00
Spectrum (k=11,m=5), FC=2	**30.15**	**61.33**
PSI-BLAST, FC=5	43.77	15.00
Sequence composition, FC=5	46.04	15.00
Spectrum (k=11,m=5), FC=5	**21.22**	**65.00**

[†] FC denotes fold change in gene expression

Table 4: Gene expression prediction. Whole genome (all genes)

Method	Balanced Error	Sensitivity
PSI-BLAST, FC[†]=2	47.81	46.57
Sequence composition, FC=2	41.03	51.27
SK, FC=2	**37.44**	**63.3**

[†] FC denotes fold change in gene expression

References

[1] Chlamydomonas reinhardtii genome assembly. http://http://genome.jgi.doe.gov/Chlre4/Chlre4.download.ftp.html.

[2] Qiang Hu, Milton Sommerfeld, Eric Jarvis, Maria Ghirardi, Matthew Posewitz, Michael Seibert, and Al Darzins. Microalgal triacylglycerols as feedstocks for biofuel production: perspectives and advances. *The Plant Journal*, 54(4):621–639, 2008.

[3] Pavel Kuksa, Pai-Hsi Huang, and Vladimir Pavlovic. Scalable algorithms for string kernels with inexact matching. In *NIPS*, 2008.

[4] Pavel Kuksa and Vladimir Pavlovic. Efficient evaluation of large sequence kernels. In *NYAS Machine Learning Symposium*, 2011.

[5] Christina Leslie and Rui Kuang. Fast string kernels using inexact matching for protein sequences. *J. Mach. Learn. Res.*, 5:1435–1455, 2004.

[6] Christina S. Leslie, Eleazar Eskin, Jason Weston, and William Stafford Noble. Mismatch string kernels for SVM protein classification. In *NIPS*, pages 1417–1424, 2002.

[7] Rachel Miller, Guangxi Wu, Rahul R. Deshpande, Astrid Vieler, Katrin Grtner, Xiaobo Li, Eric R. Moellering, Simone Zuner, Adam J. Cornish, Bensheng Liu, Blair Bullard, Barbara B. Sears, Min-Hao Kuo, Eric L. Hegg, Yair Shachar-Hill, Shin-Han Shiu, and Christoph Benning. Changes in transcript abundance in chlamydomonas reinhardtii following nitrogen deprivation predict diversion of metabolism. *Plant Physiology*, 154(4):1737–1752, 2010.