**[An automated method for detecting alternatively spliced protein domains](https://academic.oup.com/bioinformatics/advance-article/doi/10.1093/bioinformatics/bty425/5026666)**

Vitor Coelho Michael Sammeth

## Abstract

**Motivation**

Alternative splicing (AS) has been demonstrated to play a role in shaping eukaryotic gene diversity at the transcriptional level. However, the impact of AS on the proteome is still controversial. Studies that seek to explore the effect of AS at the proteomic level are hampered by technical difficulties in the cumbersome process of casting forth and back between genome, transcriptome and proteome space coordinates, and the naïve prediction of protein domains in the presence of AS suffers many redundant sequence scans that emerge from constitutively spliced regions that are shared between alternative products of a gene.

**Results**

We developed the AstaFunk pipeline that computes for every generic transcriptome all domains that are altered by AS events in a systematic and efficient manner. In a nutshell, our method employs Viterbi dynamic programming, which guarantees to find all score-optimal hits of the domains under consideration, while complementary optimisations at different levels avoid redundant and other irrelevant computations. We evaluate AstaFunk qualitatively and quantitatively using RNAseq in well-studied genes with AS, and on large-scale employing entire transcriptomes. Our study confirms complementary reports that the effect of most AS events on the proteome seems to be rather limited, but our results also pinpoint several cases where AS could have a major impact on the function of a protein domain.

**Availability**

The JAVA implementation of AstaFunk is available as an open source project on [http://astafunk.sammeth.net](http://astafunk.sammeth.net/).

**Contact**

micha@sammeth.net

**Supplementary information**

Supplementary data are available at Bioinformatics online.

[**DeepMirTar: a deep-learning approach for predicting human miRNA targets**](https://academic.oup.com/bioinformatics/advance-article/doi/10.1093/bioinformatics/bty424/5026656)

Ming Wen Peisheng Cong Zhimin Zhang Hongmei Lu Tonghua Li

## Abstract

**Motivation**

MicroRNAs (miRNAs) are small noncoding RNAs that function in RNA silencing and post-transcriptional regulation of gene expression by targeting messenger RNAs (mRNAs). Because the underlying mechanisms associated with miRNA binding to mRNA are not fully understood, a major challenge of miRNA studies involves the identification of miRNA-target sites on mRNA. In silico prediction of miRNA-target sites can expedite costly and time-consuming experimental work by providing the most promising miRNA-target-site candidates.

**Results**

In this study, we reported the design and implementation of DeepMirTar, a deep-learning-based approach for accurately predicting human miRNA targets at the site level. The predicted miRNA-target sites are those having canonical or non-canonical seed, and features, including high-level expert-designed, low-level expert-designed, and raw-data-level, were used to represent the miRNA-target site. Comparison with other state-of-the-art machine-learning methods and existing miRNA-target-prediction tools indicated that DeepMirTar improved overall predictive performance.

**Availability**

DeepMirTar is freely available at <https://github.com/Bjoux2/DeepMirTar_SdA>.

**Contact**

lith@tongji.edu.cn, hongmeilu@csu.edu.cn

**Supplementary information**

[Supplementary data](https://oup.silverchair-cdn.com/oup/backfile/Content_public/Journal/bioinformatics/PAP/10.1093_bioinformatics_bty424/1/bty424_supp.zip?Expires=1528223127&Signature=1ckvStobt7vh5IA11P0OpLmLCLPA9JAeGkfHJiFlzEWHDeYI8z-0H8Am6Bjl-8HYmwkhHdKvaa1yg87tB~gFMMgmUVjMaM7Uk-3c9kvyb52uOVNLE8FteZ0rIPXr5v7bI3QEUWgq2KZ4oqJ0DwDeABLm7gIb1Ztd82ew6Uj1g3hOVg52J7n9P9FktN6OQBG3loMPc-z3GzEHMplEecxXdvyp3HIO-Owmnww2WGNOYbNNL4fvbpBuTwauqrVYa7BeYZIycVcx54cnlnr-BHOazxRVBCHffet28G3nshppV1-5MOjrF4j58RGBZDrO8whwVstPY0n8u4szHI0PkD8N1w__&Key-Pair-Id=APKAIE5G5CRDK6RD3PGA) are available at Bioinformatics online.

# A graph-embedded deep feedforward network for disease outcome classification and feature selection using gene expression data

Yunchuan Kong Tianwei Yu

*Bioinformatics*, bty429, <https://doi.org/10.1093/bioinformatics/bty429>

## Abstract

**Motivation**

Gene expression data represents a unique challenge in predictive model building, because of the small number of samples (n) compared to the huge amount of features (p). This “n = p” property has hampered application of deep learning techniques for disease outcome classification. Sparse learning by incorporating external gene network information could be a potential solution to this issue. Still, the problem is very challenging because (1) there are tens of thousands of features and only hundreds of training samples, (2) the scale-free structure of the gene network is unfriendly to the setup of convolutional neural networks.

**Results**

To address these issues and build a robust classification model, we propose the Graph-Embedded Deep Feedforward Networks (GEDFN), to integrate external relational information of features into the deep neural network architecture. The method is able to achieve sparse connection between network layers to prevent overfitting. To validate the method’s capability, we conducted both simulation experiments and real data analysis using a Breast Invasive Carcinoma (BRCA) RNA-seq dataset and a Kidney Renal Clear Cell Carcinoma (KIRC) RNA-seq dataset from The Cancer Genome Atlas (TCGA). The resulting high classification accuracy and easily interpretable feature selection results suggest the method is a useful addition to the current graph-guided classification models and feature selection procedures.

**Availability**

The method is available at <https://github.com/yunchuankong/GEDFN>.

**Contact**

tianwei.yu@emory.edu.

**Pathway-Structured Predictive Modeling for Multi-Level Drug Response in Multiple Myeloma**

Xinyan Zhang Bingzong Li Huiying Han Sha Song Hongxia Xu Zixuan Yi Yating HongWenzhuo Zhuang Nengjun Yi

Author Notes

*Bioinformatics*, bty436, <https://doi.org/10.1093/bioinformatics/bty436>

## Abstract

**Motivation**

Molecular analyses suggest that myeloma is composed of distinct subtypes that have different molecular pathologies and various response rates to certain treatments. Drug responses in multiple myeloma (MM) are usually recorded as a multi-level ordinal outcome. One of the goals of drug response studies is to predict which response category any patients belong to with high probability based on their clinical and molecular features. However, as most of genes have small effects, gene-based models may provide limited predictive accuracy. In that case, methods for predicting multi-level ordinal drug responses by incorporating biological pathways are desired but have not been developed yet.

**Results**

We propose a pathway-structured method for predicting multi-level ordinal responses using a two-stage approach. We first develop hierarchical ordinal logistic models and an efficient quasi-Newton algorithm for jointly analyzing numerous correlated variables. Our two-stage approach first obtains the linear predictor (called the pathway score) for each pathway by fitting all predictors within each pathway using the hierarchical ordinal logistic approach, and then combines the pathway scores as new predictors to build a predictive model. We applied the proposed method to two publicly available datasets for predicting multi-level ordinal drug responses in MM using large-scale gene expression data and pathway information. Our results show that our approach not only significantly improved the predictive performance compared with the corresponding gene-based model but also allowed us to identify biologically relevant pathways.

**Availability**

The proposed approach has been implemented in our R package BhGLM, which is freely available from the public GitHub repository <https://github.com/abbyyan3/BhGLM>.

**Contact**

nyi@uab.edu; zhuangwenzhuo@suda.edu.cn

**VSClust: feature-based variance-sensitive clustering of omics data**

Veit Schwämmle Ole N Jensen

*Bioinformatics*, bty224, <https://doi.org/10.1093/bioinformatics/bty224>

## Abstract

**Motivation**

Data clustering is indispensable for identifying biologically relevant molecular features in large-scale omics experiments with thousands of measurements at multiple conditions. Optimal clustering results yield groups of functionally related features that may include genes, proteins and metabolites in biological processes and molecular networks. Omics experiments typically include replicated measurements of each feature within a given condition to statistically assess feature-specific variation. Current clustering approaches ignore this variation by averaging, which often leads to incorrect cluster assignments.

**Results**

We present VSClust that accounts for feature-specific variance. Based on an algorithm derived from fuzzy clustering, VSClust unifies statistical testing with pattern recognition to cluster the data into feature groups that more accurately reflect the underlying molecular and functional behavior. We apply VSClust to artificial and experimental datasets comprising hundreds to >80 000 features across 6–20 different conditions including genomics, transcriptomics, proteomics and metabolomics experiments. VSClust avoids arbitrary averaging methods, outperforms standard fuzzy c-means clustering and simplifies the data analysis workflow in large-scale omics studies.

**Availability and implementation**

Download VSClust at <https://bitbucket.org/veitveit/vsclust> or access it through computproteomics.bmb.sdu.dk/Apps/VSClust.

**Contact**

veits@bmb.sdu.dk

**Supplementary information**

[Supplementary data](https://oup.silverchair-cdn.com/oup/backfile/Content_public/Journal/bioinformatics/PAP/10.1093_bioinformatics_bty224/2/bty224_supp.zip?Expires=1528313265&Signature=K1ybCOPGN2tPAg2k2InEbZbBrNeeLRNEbLURNPnnYXiJ3E4bSLQraTO5AEercOdtLcSg-k3Rkty~0Peb9Ai3RinJCDsCcP3RI-~D8D6GHnVFFzDNjTE9cr5hmpiBm1sgsMdDuErjggwDB8R62lIDEfLFOeSnVfYqs-zLZmE0~wmZ2P8~yXjJzgCOba5PBohGfaGrbpMJR7R1uuCbszTJtN4PmKqzToTYMCRImeWWwG6FPzaXVRA2w296eCNfYGXcpNErH~SpitzsSosNE0~kETpt-KGBVbkandRsK7F454TmLKP9JklXHcP5o9DkLU82mZJQB~GjzRthEYLMW0pBIQ__&Key-Pair-Id=APKAIE5G5CRDK6RD3PGA) are available at Bioinformatics online.