Robinson, Mark; **Castellano, Cristina González**; **Rezwan, Faisal**; **Adams, Rod**; **Davey, Neil**; **Rust, Alastair**; **Sun, Yi**

Combining experts in order to identify binding sites in yeast and mouse genomic data.

(English) Zbl 1254.92070
Neural Netw. 21, No. 6, 856-861 (2008).

Summary: The identification of cis-regulatory binding sites in DNA is a difficult problem in computational biology. To obtain a full understanding of the complex machinery embodied in genetic regulatory networks it is necessary to know both the identity of the regulatory transcription factors and the location of their binding sites in the genome. We show that using an SVM together with data sampling to classify the combination of the results of individual algorithms specialised for the prediction of binding site locations, can produce significant improvements upon the original algorithms. The resulting classifier produces fewer false positive predictions and so reduces the expensive experimental procedure of verifying the predictions.

MSC:

92D10 Genetics and epigenetics
92C42 Systems biology, networks
68T10 Pattern recognition, speech recognition

Keywords:

computational biology; support vector machine; imbalanced data; sampling; transcription factor binding sites

Software:

SMOTE; LIBSVM; Footprinter

Full Text: DOI

References:

[1] Abnizova, I.; Rust, A.G.; Robinson, M.; Te Boekhorst, R.; Gilks, W.R., Transcription binding site prediction using Markov models, Journal of bioinformatics and computational biology, 4, 425-441, (2006)

[2] Akbani, R., Kwek, S., & Japkowicz, N. (2004). Applying support vector machines to imbalanced datase. In \textit{15th European conference on machine learning}. 2004. pp. 39-50 · Zbl 1132.68523

[3] Apostolico, A.; Bock, M.E.; Lonardi, S.; Xu, X., Efficient detection of unusual words, Journal of computational biology, 7, 71-94, (2000)

[4] Bailey, T.L., & Elkan, C. (1994). Fitting a mixture model by expectation maximization to discover motifs in biopolymers. In \textit{Proc int conf intell syst mol biol. Vol. 2} (pp. 28-36)

[5] Blanchette, M.; Tompa, M., Footprinter: A program designed for phylogenetic footprinting, Nucleic acids research, 31, 3840-3842, (2003)

[6] Blanco, E.; Farre, D.; Alba, M.M.; Messeguer, X.; Guigo, R., ABS: A database of annotated regulatory binding sites from orthologous promoters, Nucleic acids research, 34, D63-D67, (2006)

[7] Brown, C.T.; Rust, A.G.; Clarke, P.J.; Pan, Z.; Schilstra, M.J.; De Buysscher, T., New computational approaches for analysis of cis-regulatory networks, Developmental biology, 246, 86-102, (2002)

[8] Chang, C.-C., & Lin, C.-J. (2001). LIBSVM: A library for support vector machines

[9] Chawla, N.V.; Bowyer, K.W.; Hall, L.O.; Kegelmeyer, W.P., SMOTE: synthetic minority over-sampling technique, Journal of artificial intelligence research, 16, 321-357, (2002) · Zbl 0994.68128

[10] Henery, R.J., Methods for comparison, (), 107-124 · Zbl 0424.90019

[11] Hughes, T.R.; Marton, M.J.; Jones, A.R.; Roberts, C.J.; Stoughton, R.; Armour, C.D., Functional discovery via a compendium of expression profiles, Cell, 102, 109-126, (2000)

[12] Karolchik, D.; Baertsch, R.; Diekhans, M.; Furey, T.S.; Hinrichs, A.; Lu, Y.T., The UCSC genome browser database, Nucleic acids research, 31, 51-54, (2003)

[13] Montgomery, S.B.; Griffith, O.L.; Sleumer, M.C.; Bergman, C.M.; Bilenky, M.; Pleasance, E.D., Oreganno: an open access database and curation system for literature-derived promoters, transcription factor binding sites and regulatory variation,

Bioinformatics, 22, 637-640, (2006)

[14] Radivojac, P.; Chawla, N.V.; Dunker, A.K.; Obradovic, Z., Classification and knowledge discovery in protein databases, Journal of biomedical informatics, 37, 224-239, (2004)

[15] Rajewsky, N.; Vergassola, M.; Gaul, U.; Siggia, E.D., Computational detection of genomic cis-regulatory modules applied to body patterning in the early drosophila embryo, BMC bioinformatics, 3, 30, (2002)

[16] Robinson, M.; Sharabi, O.; Sun, Y.; Adams, R.; te Boekhorst, R.; Rust, A.G., Using real-valued meta classifiers to integrate and contextualize binding site predictions, (), 822-829

[17] Robinson, M.; Sun, Y.; Boekhorst, R.T.; Kaye, P.; Adams, R.; Davey, N., Improving computational predictions of cis-regulatory binding sites, Pacific symposium on biocomputing, 391-402, (2006)

[18] Sun, Y., Robinson, M., Adams, R., Kaye, P., Rust, A.G., & Davey, N. (2005a). Using real-valued meta-classifiers to integrate binding site predictions. In \textit{IJCNN} · Zbl 1133.68411

[19] Sun, Y., Robinson, M., Adams, R., Kaye, P., Rust, A.G., & Davey, N. (2005b). Integrating binding site predictions using non-linear classification methods. In \textit{Machine learning workshop} · Zbl 1133.68411

[20] Sun Y., Robinson, M., Adams, R., te Boekhorst, R., Rust, A. G., Davey, N. (2006). Using feature selection filtering methods for binding site predictions. In \textit{5th IEEE international conference on cognitive informatics, 2006, vol. 1} (pp. 566-571)

[21] Thijs, G.; Lescot, M.; Marchal, K.; Rombauts, S.; De Moor, B.; Rouze, P., A higher-order background model improves the detection of promoter regulatory elements by Gibbs sampling, Bioinformatics, 17, 1113-1122, (2001)

[22] Tompa, M.; Li, N.; Bailey, T.L.; Church, G.M.; De Moor, B.; Eskin, E., Assessing computational tools for the discovery of transcription factor binding sites, Nature biotechnology, 23, 137-144, (2005)

[23] Veropoulos, K., Cristianini, N., & Campbell, C. (1999). Controlling the sensitivity of support vector machines. In \textit{Proceedings of the international joint conference on artificial intelligence}

[24] Wasserman, W.W.; Sandelin, A., Applied bioinformatics for the identification of regulatory elements, Nature reviews genetics, 5, 276-287, (2004)

[25] Yellaboina, S.; Seshadri, J.; Kumar, M.S.; Ranjan, A., Predictregulon: A web server for the prediction of the regulatory protein binding sites and operons in prokaryote genomes, Nucleic acids research, 32, W318-W20, (2004)

[26] Zhu, J.; Zhang, M.Q., SCPD: A promoter database of the yeast \textit{saccharomyces cerevisiae}, Bioinformatics, 15, 607-611, (1999)