Bioinformatics, 32(8), 2016, 1229–1231
doi: 10.1093/bioinformatics/btv726
Advance Access Publication Date: 12 December 2015
Applications Note

Sequence analysis

INSECT 2.0: a web-server for genome-wide *cis*-regulatory modules prediction

R. Gonzalo Parra[1,†], Cristian O. Rohr[1,†], Daniel Koile[2], Carolina Perez-Castro[2] and Patricio Yankilevich[2,*]

[1]Facultad De Ciencias Exactas Y Naturales, Universidad De Buenos Aires, Buenos Aires, Argentina and [2]Instituto De Investigacion En Biomedicina De Buenos Aires (IBioBA), CONICET, Partner Institute of the Max Planck Society, Godoy Cruz 2390, Buenos Aires C1425FQA, Argentina

*To whom correspondence should be addressed.
[†]The authors wish it to be known that, in their opinion, the first two authors should be regarded as Joint First Authors.

Associate Editor: John Hancock

Received on 11 September 2015; revised on 3 December 2015; accepted on 4 December 2015

Abstract

INSECT is a user-friendly web server to predict the occurrence of *Cis*-Regulatory Modules (CRMs), which control gene expression. Here, we present a new release of INSECT which includes several new features, such as whole genome analysis, nucleosome occupancy predictions, and which provides additional links to third-party functional tools that complement user capabilities, CRM analysis and hypothesis construction. Improvements in the core implementation have led to a faster and more efficient tool. In addition, this new release introduces a new interface designed for a more integrative and dynamic user experience.
Availability and implementation: http://bioinformatics.ibioba-mpsp-conicet.gov.ar/INSECT2
Contact: pyankilevich@ibioba-mpsp-conicet.gov.ar

1 Introduction

The complexity of multicellular organisms is principally a product of the intricate regulatory networks through which they control cell expression, in a range of spatiotemporal conditions. The high specificity required for gene expression regulation is acquired mainly through the physical interaction of many Transcription Factors (TFs) that bind to groups of *cis* DNA elements (Transcription Factor Binding Sites, TFBSs) in a coordinated manner. These elements are often located close to each other and follow specific and highly conserved patterns, generally known as *Cis*-Regulatory Modules (CRMs).

Mathematical models of many TFBSs have been obtained and incorporated into publicly available databases. These models, usually represented as Position Weight Matrices (PWMs), can be used to predict new instances of regulatory sites over specific sets of sequences. However, the major obstacle for prediction of TFBSs, when aligning PWMs to gene regulatory sequences, is the vast amount of artifactual findings due to random and non-functional sites with high PWM scores (i.e. false positives).

INSECT first version (Rohr *et al.*, 2013) was presented as a CRM detector that outperforms most of the available tools with similar characteristics in terms of sensitivity and specificity, by the implementation of several strategies to lower the rate of false positives. The main advantage of our method is a more realistic semantic representation of CRMs structures. INSECT offers several features that allow users to easily extract information related to the genes of interest from the Ensembl annotations and other third-party tools, such as the UCSC Genome Browser. The tool interface minimizes usage complexity, allowing a highly integrative analysis to be performed.

INSECT 2.0 is significantly faster than its previous version, allowing for genome-wide analysis. New third-party tools have been included to help in the processes of further analysis and finding of CRMs, which can maximize the likelihood of representing true regulatory elements. INSECT 2.0 is a new online tool of great value to every researcher interested to predict the occurrence of CRMs at genome level and over long lists of genes, with no requirement to install prior infrastructure.

2 New features

Increased speed

A redesigned search core and data architecture have significantly boosted search speed. While the previous version was relatively fast when analyzing several hundred genes, INSECT 2.0 can perform genome-wide analysis in the same amount of time or faster, depending on the structure complexity of the CRMs and stringency parameters.

Genome-wide analysis and improved results dissection

Genome-wide analysis can be performed for 14 genomes from Ensembl. By simply defining distance limits relative to the TSS of each gene, INSECT 2.0 automatically retrieves the corresponding sequences and performs the search, providing several genomic annotations and links to third-party tools along with the results.

It has been shown that transcriptional regulation constrains the organization of genes across and within chromosomes (Janga *et al.*, 2008). For this reason, INSECT 2.0 allows CRM searches on specific chromosomes, and the filtering of the resulting lists of genes by chromosome, when genome-wide analyses are performed.

The protein-coding genes-centric vision for gene expression regulation has until now underestimated the importance of many other gene categories present in genomes. Non-protein-coding genes are critically involved in the regulation of cellular processes, and have been proven to be as tightly regulated as protein-coding genes, as recently demonstrated for lnc-RNA genes (Alam *et al.*, 2014). In addition, the implementation of a biotype filter provides the capability to perform analyses that are focused on non-protein-coding genes.

Chromatin structure and nucleosomes positioning within DNA may play a major role in allowing TFs–DNA interaction to proceed with transcriptional regulation. While high nucleosomes occupancy in DNA mainly acts as a barrier for most TFs to target the DNA, some others, called pioneer factors, may be biased to target silent chromatin and initiate cell-fate changes (Soufi et al., 2015). We have included a new feature for nucleosomes occupancy calculations for CRMs (Kaplan *et al.*, 2009), offering more accurate estimation of whether the predicted site is a real regulatory element or not.

Additional PWM libraries

For users that do not have their own PWM for their TFs of interest, having a database of known PWMs is essential. In this new release, we have added several PWMs from Hocomoco (655 PWMs) (Kulakovskiy *et al.*, 2013), SwissRegulon (189 PWMs) (Pachkov *et al.*, 2013) and FlyFactorSurvey (486 PWMs) (Zhu *et al.*, 2011). Taking into consideration those already included in our previous version, INSECT 2.0 has currently 2526 PWMs integrated in its PWM database.

Summary statistics and filters

Genomic CRMs search analysis could yield long lists of genes with several CRM hits on each of them. For this reason, we have implemented visualizations and filters that help to dissect and explore the search results.

TFs bind to specific elements at the regulatory regions of genes following certain patterns. Some TFBSs are distributed very close to the TSS of the target genes, while others show a decreased presence at those regions, shifting their occurrence to further regions (Chen et al., 2015). The distribution of distances of the predicted CRMs relative to the TSS for each gene and the PWM score for each TFBSs in the CRMs are calculated and visualized as histograms in the results page, in order to give the user additional information about the resulting binding patterns.

Dynamic filters for both TSS-relative distances and PWM scores are implemented. By adjusting these filters, results can be filtered on the fly, avoiding the need to run the job again with different parameters. This dynamism helps users to better handle the results and obtain as much information as possible by progressively changing search rules from flexible to more strict conditions, in order to identify those candidate genes with the most reliable CRM predictions.

Additional interaction with third-party tools

Prediction of CRMs is a difficult task given the amount of false positives that can arise during its search and evaluation. Despite PWMs being a useful representation of TFBSs, cells have several levels of complexity that account for transcriptional regulation beyond the occurrence of representative sites in the DNA. For this reason and to improve the working hypothesis, we have integrated several external tools to provide additional biological features about the genes that are predicted to contain potential CRMs. INSECT 2.0 is linked to GeneMANIA (Mostafavi *et al.*, 2008), enabling users to select genes with predicted CRMs to inspect whether or not they are related, according to available bibliography. Furthermore, we have included a link to the Ensembl Regulatory Build (Zerbino et al., 2015), which aims to annotate regulatory elements in human and mouse genomes, by merging both experimental and computational derived data.

3 Conclusion

INSECT 2.0 is a powerful tool for researchers requiring search and analysis capabilities for potential CRMs presence at a genomic scale. This new release has added several features to help users further explore predicted CRMs based on the biological aspects of genome structure. The implementation of search filters reduces the search space to specific sets of genes of interest, minimizing the rate of false positives. New visualizations and statistical histograms of the global results have also been included. By linking the genes with predicted CRMs to additional third-party resources such as the UCSC Genome Browser, GeneMANIA or Ensembl Regulation, the powerful functionalities of its interface make INSECT 2.0 a unique integrative platform for CRMs prediction. The combined feature set allows users to exploit CRMs information that otherwise may remain hidden in huge tables, or that would require programming, data gathering or computational skills which would be barriers for non-specialized users.

Funding

All the authors are members of the Argentine National Research Council (CONICET). This project was supported by grants from CONICET, ANPCyT and FOCEM-Mercosur.

Conflict of interest: none declared.

References

Alam,T. *et al.* (2014) Promoter analysis reveals globally differential regulation of human long non-coding RNA and protein-coding genes. *PLoS One*, **9**, e109443.

Chen,H. *et al.* (2015) An integrative analysis of TFBS-clustered regions reveals new transcriptional regulation models on the accessible chromatin landscape. *Sci. Rep.*, **5**, 8465.

Janga,S.C. *et al.* (2008) Transcriptional regulation constrains the organization of genes on eukaryotic chromosomes. *Proc. Natl Acad. Sci. U.S.A.* **105**, 15761–15766.

Kaplan,N. *et al.* (2009) The DNA-encoded nucleosome organization of a eukaryotic genome. *Nature*, **458**, 362–366.

Kulakovskiy,I.V. *et al.* (2013) HOCOMOCO: a comprehensive collection of human transcription factor binding sites models. *Nucleic Acids Res.*, **41**, 195–202.

Mostafavi,S. *et al.* (2008) GeneMANIA: a real-time multiple association network integration algorithm for predicting gene function. *Genome Biol.*, **9**(Suppl 1), S4.

Pachkov,M. *et al.* (2013) SwissRegulon, a database of genome-wide annotations of regulatory sites: recent updates. *Nucleic Acids Res.*, **41**, D214–D220.

Rohr,C.O. *et al.* (2013) INSECT: IN-silico SEarch for Co-occurring Transcription factors. *Bioinformatics*, **29**, 2852–2858.

Soufi,A. *et al.* (2015) Pioneer transcription factors target partial DNA motifs on nucleosomes to initiate reprogramming. *Cell*, **161**, 555–568.

Zerbino,D.R. *et al.* (2015) The Ensembl regulatory build. *Genome Biol.*, **16**, 56

Zhu,L. *et al.* (2011) Flyfactorsurvey: a database of drosophila transcription factor binding specificities determined using the bacterial one-hybrid system. *Nucleic Acids Res.*, **39**, D111–D117.