BIOINFORMATICS ORIGINAL PAPER

Vol. 29 no. 22 2013, pages 2852–2858
doi:10.1093/bioinformatics/btt506

Sequence analysis

Advance Access publication September 4, 2013

INSECT: IN-silico SEarch for Co-occurring Transcription factors

Cristian O. Rohr[1,†], R. Gonzalo Parra[2,†], Patricio Yankilevich[3,*] and Carolina Perez-Castro[3,*]

[1]Instituto de Ecología, Genética y Evolución (IEGEBA)-CONICET, [2]Protein Physiology Laboratory, Departamento de Química Biológica, Facultad de Ciencias Exactas y Naturales, UBA-CONICET-IQUIBICEN, C1428EGA, Buenos Aires, Argentina and [3]Instituto de Investigación en Biomedicina de Buenos Aires (IBioBA) -CONICET- Partner Institute of the Max Planck Society, C1425FQD, Buenos Aires, Argentina

Associate Editor: John Hancock

ABSTRACT

Motivation: Transcriptional regulation occurs through the concerted actions of multiple transcription factors (TFs) that bind cooperatively to *cis*-regulatory modules (CRMs) of genes. These CRMs usually contain a variable number of transcription factor-binding sites (TFBSs) involved in related cellular and physiological processes. Chromatin immunoprecipitation followed by sequencing (ChIP-seq) has been effective in detecting TFBSs and nucleosome location to identify potential CRMs in genome-wide studies. Although several attempts were previously reported to predict the potential binding of TFs at TFBSs within CRMs by comparing different ChIP-seq data, these have been hampered by excessive background, usually emerging as a consequence of experimental conditions. To understand these complex regulatory circuits, it would be helpful to have reliable and updated user-friendly tools to assist in the identification of TFBSs and CRMs for gene(s) of interest.

Results: Here we present INSECT (IN-silico SEarch for Co-occurring Transcription factors), a novel web server for identifying potential TFBSs and CRMs in gene sequences. By combining several strategies, INSECT provides flexible analysis of multiple co-occurring TFBSs, by applying differing search schemes and restriction parameters.

Availability and implementation: INSECT is freely available as a web server at http://bioinformatics.ibioba-mpsp-conicet.gov.ar/INSECT

Contact: cperezcastro@ibioba-mpsp-conicet.gov.ar or pyankilevich@ibioba-mpsp-conicet.gov.ar

Supplementary information: Supplementary data are available at *Bioinformatics* online.

Received on April 17, 2013; revised on August 14, 2013; accepted on August 26, 2013

1 INTRODUCTION

Regulation of transcription occurs through the concerted actions of multiple TFs that bind to *cis*-regulatory modules (CRMs) (Arnone and Davidson, 1997; Kirchhamer *et al.,* 1996). Experimental techniques such as chromatin immunoprecipitation followed by sequencing (ChIP-seq) have provided valuable insight into the mechanisms governing gene regulation. However, these techniques produce large amounts of data of variable quality, inherent to the complex experimental conditions used (Park, 2009). The inconvenience, together with high data analysis costs, makes this strategy unappealing without specific information to help reduce the dataset before performing experiments.

The availability of more efficient sequencing techniques, together with the growing number of genomes deposited on public databases, provides an ideal scenario for development of bioinformatics tools facilitating design of future experiments in this field. Several tools have been recently developed to search and characterize the transcription factor-binding sites (TFBSs) of CRMs by dissecting relationships among them (Van Loo and Marynen, 2009). Currently available TFBSs search tools can be grouped into two major approaches: (i) Motif discovery: TFBSs are inferred by analyzing a group of sequences that are considered to be under the regulation of a particular transcription factor (TF) and (ii) Motif search: to determine the location where a TF is more likely to bind by searching its corresponding TFBS over sequence datasets.

The constant progress in experimental technologies and the establishment and execution of initiatives such as the ENCODE project (Raney *et al*., 2011) have helped to characterize large sets of TFBSs. Thus, a considerable amount of position weight matrices (PWMs, i.e. elements that represent a group of sequences that are recognized by a specific TF) has been deposited on databases such as TRANSFAC (Matys *et al*., 2006), JASPAR (Bryne *et al*., 2008) and UniPROBE (Newburger and Bulyk, 2009). Such matrices derived from alignment and processing of target sequences can be used to search potential TFBSs over genomes and sequences as well as to establish potential sites of gene regulation. An important limitation of this methodology is its high false-positive rate (FPR). According to the futility theorem (Wasserman and Sandelin, 2004), the presence of non-functional binding sites for a given TF can be three orders of magnitude higher compared with the actual number of functional sites of genomes (Tronche *et al*., 1997). FPR reduction without compromising the sensitivity of the method is challenging, and several strategies have been designed to tackle this problem. However, many of these tools are organism specific (i-*cis*Target, DiRE), discontinued in their maintenance (MSCAN, *cis*-analyst), outdated (ModuleMiner, TFBScluster, PReMod) or lack flexibility and are difficult to use by non-computational scientists (as reviewed by Van Loo and Marynen, 2009). Importantly, most of these tools do not allow implementation of rules regarding the relationships between the individual TFBSs of a particular CRM (Frith *et al*., 2003). This caveat represents a significant limitation during CRM screening. Nevertheless, most of these tools have

*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.

proven quite valuable to the scientific community, assisting in the characterization and study of CRMs.

Here we present INSECT (IN-silico SEarch for Co-occurring Transcription factors), a novel motif search web tool for identification of potential TFBSs and CRMs. By combining different strategies, INSECT allows for complete and flexible analysis of multiple co-occurring TFBSs. We have compared INSECT performance with other tools analyzing experimental data related to the regulation of genes by Sox2 and Oct-4 TFs in embryonic stem cells (ESCs). The regulatory networks of mouse and human ESC properties have been studied extensively (Boyer *et al.*, 2005; Rodda *et al.*, 2005). Many target genes of the Sox2 and Oct-4 TFs have been identified using genome-wide ChIP followed by DNA microarray analysis (Boyer *et al.*, 2005; Loh *et al.*, 2006). These TFs (in addition to an additional TF, Nanog) appear to act in concert to regulate a limited repertoire of target genes, tightly regulating the ESCs' pluripotent state.

2 MATERIALS AND METHODS

2.1 Experimental datasets

To assess INSECT performance, we selected two experimental datasets.

Rodda dataset. The exact sequences and positions that are bound by Sox2 and Oct-4 in six genes were previously experimentally reported (Table 1; Rodda *et al.*, 2005). The regions from −5 to +5 kb relative to the transcription start site (TSS) were scanned for each gene.

Boyer dataset. Genome-wide ChIP assay was previously performed and reported by Boyer *et al.* (2005). We used a subset of genes reported by Sun *et al.* (2009) that bind Sox2 and Oct-4 cooperatively in their proximal promoter region (from −1 kb to the TSS) from which, after mapping to Ensembl Gene IDs, we obtained 79 genes (Supplementary Table S4).

2.2 Sox2 and Oct-4 PWMs used for CRM searches

The INSECT module to create PWMs was used to make two PWMs from the Sox2 and Oct-4 binding sequences of the Rodda dataset. In addition, the PWM for Pou5f1 (Oct-4, MA0142.1), publicly available from JASPAR, was also used.

2.3 Benchmarking to other computational approaches

INSECT performance was compared with three tools with comparable search features: CPModule (Sun *et al.*, 2012), Cluster Buster (Frith *et al.*, 2003) and MotifViz (Fu *et al.*, 2004). The performance was measured using the following parameters Equations (1–3):

$$Specificity = \frac{TN}{FP + TN} \quad (1)$$

$$Sensitivity = \frac{TP}{TP + FN} \quad (2)$$

$$MCC = \frac{TP * TN - FP * FN}{\sqrt{(TP + FP)(TP + FN)(TN + FP)(TN + FN)}} \quad (3)$$

where MCC is the Mathews Correlation Coefficient, TP are the true positives, TN are the true negatives, FP are the false positives and FN are the false negatives.

2.4 PWM construction and formats

INSECT uses the TFBS Perl module (Lenhard and Wasserman, 2002) to implement the PWMs creation from a multiple sequence alignment. Although TFBS allows for the construction of different types of scoring matrices (ICM, PFM and PWM), we chose the TFBS PWM building module. Scoring matrices in PFM and PWM format can also be uploaded. Other modules in TFBS Perl module were also used to perform PWM searches.

Table 1. Target genes for Sox2 and Oct-4

Gene	Sox2 TFBS	Oct-4 TFBS	Spacing	Orientation	Location	
Fgf4	CTTTGTT	ATGCTAAT	3	Forward	3022	3′
Utf1	CATTGTT	ATGCTAGT	0	Reverse	1838	3′
Sox2	CATTGTG	ATGCATAT	0	Forward	4041	3′
Fbx15	CATTGTT	ATGATAAA	0	Reverse	523	5′
Nanog	CATTGTA	ATGCAAAA	0	Reverse	181	5′
Pou5f1	CTTTGTT	ATGCATCT	0	Reverse	1991	5′

Note: Known target sequence, spacing, orientation and genomic location. Adapted from Rodda *et al.* (2005)

2.5 PWM scoring

The score assigned by a PWM m to a substring $S = (Sj)_{j=1}^{N}$ is defined as $\sum_{j=1}^{N} m_{s_j,j}$, where j represents position in the substring, s_j is the symbol at position j in the substring and $m_{\alpha,j}$ is the score in row α and column j of the matrix. PWM scores are calculated as the sum of position-specific scores for each symbol in a given substring. We assume a direct association between the DNA sequence variability in the binding sites and the binding affinity (or activity) for the particular protein that recognizes these sites (Berg and von Hippel, 1987; Stormo and Fields, 1998). In this context, if enough number of sequences were used to build the PWM, higher TF binding affinities would be indicated by higher PWM scores, being the maximum matrix score (MMSc), $MMSc = \sum_{j=1}^{N} max(m_j)$. Substrings will have PWM scores no greater than its MMSc. The score of every substring is divided by the MMSc to normalize the TF affinities and provide a score that is comparable among different factors without a length motif bias.

2.6 Score threshold setting

TFBSs search using a PWM requires defining a score threshold to determine whether a sequence is defined as a potential TFBS or a false positive. A low-score threshold usually leads to the appearance of spurious matches, making the prediction of potential TFBSs difficult owing to a high FPR. Therefore, users need to define and set an optimal threshold for each search. For Sox2 and Oct-4, score threshold estimation refers to the Results (section 4.1).

3 INSECT WEB SERVER

INSECT was designed as a user-friendly tool for non-computational scientists to perform motif search complex analysis. Here we briefly describe the main features in the INSECT web server.

3.1 Organisms and genes

Genes and putative regulatory regions from the genomes of 14 organisms can be defined and sequences automatically retrieved from the local copy of the latest Ensembl release (Flicek *et al.*, 2013). These genomes correspond to the most represented organisms of the TF PWMs stored in JASPAR, TRANSFAC and UniPROBE. Users need only provide a list of Ensembl gene IDs or gene symbols and define the upstream and downstream sequence limits relative to the TSS. Additionally, a multi-fasta file containing up to 500 sequences can be uploaded to the server to perform the search. In this case, the positions informed by

INSECT are numbered from 0 to $L-1$, which corresponds to the length of the sequences.

3.2 Phylogenetic footprinting

To reduce the FPR, INSECT can apply a phylogenetic footprinting search through analysis of a conservation window between orthologous genes. Orthologous sequences are automatically retrieved from the local copy of the Ensembl genomes or uploaded in a multi-fasta file. For every potential CRM identified in a gene, the INSECT search algorithm requires the orthologous gene sequence to have an equivalent CRM. A CRM is considered equivalent by INSECT if positions relative to the TSS are conserved within a range defined by the user (default set to 1000 bp). Users are encouraged to apply this option carefully, as it has been shown that 50% of regulatory elements are conserved between human and mouse. It has also been postulated that most of the conserved sites are those of high binding specificity (Dermitzakis and Clark, 2002; Wray *et al*., 2003). Note that filtering out a specific site by applying phylogenetic footprinting does not necessarily mean that it is a false positive; instead, it increases the remaining sites' chances of being true positives owing to conservation. Therefore, phylogenetic footprinting is an effective method of filtering out false positives, resulting in more robust CRM identification.

3.3 TFs and TFBSs

One major benefit of INSECT resides in the flexibility of defining search parameters. Users can search TFBSs by selecting PWMs from the JASPAR (version 4.0), TRANSFAC (release 7.0) and UniPROBE public databases. Owing to the low number of matrices present in publicly (TRANSFAC 398, JASPAR 457 and UniPROBE 418) versus privately (TRANSFAC Professional 1665) curated databases, INSECT allows the user to build their own matrices from a multiple sequence alignment as described in the TFBS Perl module (Lenhard and Wasserman, 2002), or to upload their own PWMs.

3.4 TFBSs search modes

Because CRMs are defined by a set of TFBSs and by specific relationships among them, a correct CRM model definition and build is essential. INSECT performs TFBSs search by two different methods.

Sliding window search. Target sequences are scanned using a user-defined window of fixed length (Fig. 1A). Analyzed subsequences must contain matches to all the entered TFs to report a CRM. Additionally, users can define specific TFs as not necessarily required to report a CRM by checking the Allow missing option located in the TFs and TFBSs section of the selected TFs.

Master-driven search. A master TF is selected to drive the search of co-occurring TFs, which are determined by maximum spacing restrictions relative to the master (Fig. 1B). For both search options, TFBSs can be searched in the direct, reverse or both strand orientations. This represents a key feature of INSECT, as TFs usually co-localize in very specific ways when binding to DNA owing to protein–protein interactions and complex spatial arrangements. The master-driven search mode is more restrictive than the sliding window search mode, as it requires different factors to strictly satisfy imposed restrictions with respect to a reference master TF. INSECT allows a maximum of five TFs, including the master.

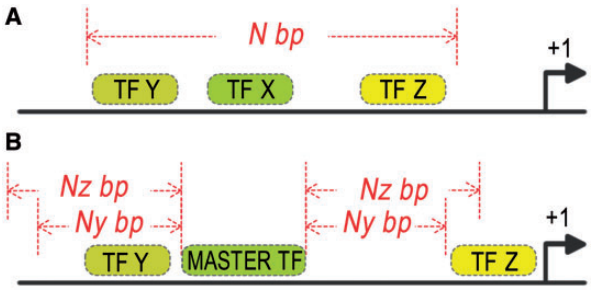

Fig. 1. (**A**) Sliding window search for three co-occurring TFBSs with spacing restriction set by a window size of N bp. (**B**) Master-driven search for a Master TF and two co-occurring TFs with differing maximum spacing restrictions

3.5 INSECT results and visualization

User-friendly interfaces to run the analysis (Supplementary Fig. S1) and generate clear results (Supplementary Fig. S2) are important issues for a bioinformatics web server. INSECT integrates various features to achieve this goal. For genes in which valid CRMs were identified by INSECT, Gene Ontology mapping (Ashburner *et al*., 2000), biological process, molecular functions and cellular components can be visualized, increasing the confidence of the results. Moreover, diagrams corresponding to CRMs along with the gene exon/intron structure for every annotated transcript on Ensembl are drawn. INSECT dynamically generates tracks that can be submitted and automatically opened with the UCSC Genome Browser (http://genome.ucsc.edu/, only available for Ensembl automatically retrieved sequences) to analyze the detected CRMs, their relationship to surrounding genomic regions and other UCSC Genome Browser-annotated tracks. Finally, spreadsheets with the TFBSs sequences and PWM scores as calculated in Lenhard and Wasserman (2002) along with the Gene Ontology table, CRMs with exons/introns structure diagrams and GFF files (that can be submitted as UCSC Genome Browser tracks) can be downloaded for each gene.

4 RESULTS

INSECT was tested by searching two experimental datasets for TFBSs of Sox2 and Oct-4, two TFs involved in maintaining pluripotency in ESCs. In the Rodda dataset analysis, CRMs composed by Sox2 and Oct-4 binding sites were searched over potential regulatory sequences corresponding to six genes that were experimentally demonstrated to have these TFBS by Rodda *et al*. (2005). We compared INSECT results and performance with three existing motif search tools. We show that INSECT outperformed the other tools in almost every analysis. In the Boyer dataset analysis, Sox2 and Oct-4 co-occurrences were analyzed in gene sequences from a ChIP experiment reported by Boyer *et al*. (2005).

4.1 Performance comparison using experimental data (Rodda dataset)

To perform this analysis, we selected DNA regions from −5 to +5 kb relative to the TSS of the six genes listed in Table 1. The sensitivity, specificity and MCC as described in Equations (1–3) in the Materials and Methods section were used to compare tools' performances in terms of TP, FP, TN and FN values.

Although there are two available matrices for Sox2 and Oct-4 in JASPAR (MA0143.1, MA0142.1), both matrices have the Sox2 and Oct-4 motifs concatenated in a single matrix. Thus, the Oct-4 and Sox2 binding sites cannot be separated and a spacing parameter cannot be specified. Therefore, these matrices are incapable of detecting the Sox2/Oct-4 CRM in the Fgf4 gene, where spacing between the single motifs is >0 bp (see Table 1). As both matrices are equivalent (Fig. 2), we selected the Oct-4 matrix (named Pou5f1 in JASPAR) for performance comparison.

INSECT was used to create two separate PWMs for Sox2 and Oct-4 using the sequences from sites reported in Table 1. The sequence logos for the created PWM matrices show that the motifs are consistent with the Pou5f1 matrix (Fig. 2).

Optimal threshold estimation. We evaluated a range of thresholds to define the optimal score value for the Rodda dataset analysis. Decreasing thresholds were applied from 100 to 75% (Supplementary Tables S1 and S2), and the MCC value was calculated for both Sox2 and Oct-4 motifs (Fig. 3). Maximum MCC values were obtained from 86 to 89% score thresholds.

Comparative performance analysis. We selected CPModule, MotifViz and ClusterBuster to compare their performances against INSECT. These three tools share similar features with INSECT, are updated and provide many options to enhance CRMs searching. For this comparison, regulatory regions from −5 to +5 kb relative to the TSS were automatically retrieved from Ensembl using INSECT. The Sox2 and Oct-4 matrices created with INSECT PWM construction functionality were used for Sox2/Oct-4 CRMs search in all tools. Distance restrictions between the TFBS were set when possible.

INSECT Master Search mode was applied, with the following search parameters. Sox2 was set as Master TF and Oct-4 as a co-occurring (note that when <3 factors are analyzed, the factor–cofactor order is considered not relevant), with a maximum spacing set at 3 bp. Sox2 strand restriction was set to both strand and same strand for Oct-4. A score threshold of 86% was used in the analysis for both factors. CPModule was used with eight additional randomly selected matrices from TRANSFAC (Supplementary Table S3). The miner-proximity parameter was set to 18 bp, as it measures distance from the beginning of the first motif to the end of the second. A raw score of 7 was set, and mouse chromosome 19 used as a background sequence. MotifViz was used initially with default parameters, but no sites were detected. Consequently, the *P*-value cutoff parameter was varied until all true positives, except for Fbxo15, were detected ($P = 0.18$). Further *P*-value increase did not improve results (data not shown). The overall raw score threshold parameter was set to 6.16 corresponding to the lowest true-positive score (data not shown). ClusterBuster was executed with default parameters and gap parameter set to 3 bp. INSECT was the more robust tool related to both FP and FN rates, and also with regard to MCC values. Results are summarized in Table 2.

The results of the Rodda dataset analysis for the Pou5f1 JASPAR PWM are shown in Table 3. With the exception of INSECT, the tools rely on *P*-value analysis to perform motif searches. The tools behavior is variable when matrices change, in terms of TP and TN, reflecting the heterogeneity of the strategies, and the dependence on the PWM used. CPModule was the most affected tool in terms of FP by the matrices replacement, whereas ClusterBuster was the most affected in terms of FN. Contrastingly, MotifViz improved its performance by drastically lowering its FPR.

For the Pou5f1 PWM INSECT analysis, we used an 86% score threshold in sliding window mode, without strand restriction. CPModule was used using the same eight random matrices as before, the raw score was set in 7 and the background sequence was mouse chromosome 19. Cluster Buster was executed with the default parameters. For MotifViz, the motif score threshold was set in 7.62, and mouse chromosome 19 was used as background.

Note that by using the JASPAR matrix, a single motif search instead of a CRM search was performed, as the two motifs for Sox2 and Oct-4 were included in the same matrix. By applying this approach, the master-driven search is not necessary any longer, as the distance restriction is implicitly imposed in the matrix definition itself (spacing 0 between Sox2 and Oct-4). Additionally, a relative orientation between the two sub motifs (Sox2/Oct-4 is allowed, whereas Oct-4/Sox2 is not) is implicitly applied.

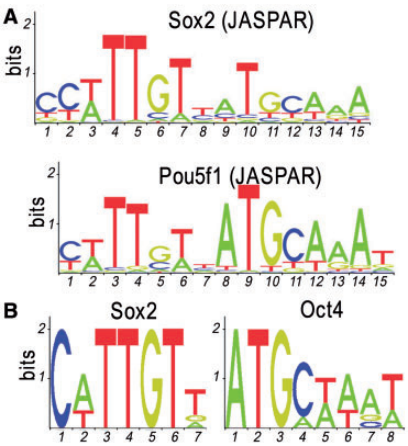

Fig. 2. (**A**) Pou5f1 and Sox2 JASPAR PWMs logos. (**B**) Rodda-derived PWM logos for Sox2 and Oct-4

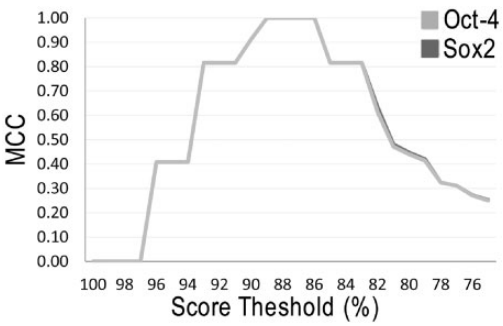

Fig. 3. Sox2 and Oct-4 optimal score threshold analysis. MCC was calculated according to Equation (3) described in Materials and Methods

4.2 Comparative analysis over ChIP data (Boyer dataset)

Matrices derived from the Rodda dataset and the threshold that maximized the MCC for detecting the true positive over those genes were initially used to perform the Sox2/Oct-4 CRM search over 79 genes from the Boyer dataset. As with the Rodda dataset comparison, the Pou5f1 matrix from JASPAR was also used. The spacing between Sox2 and Oct-4 binding sites is known to be >0 for genes like *Fgf4*, but the Pou5f1 matrix contains both Sox2 and Oct-4 motifs with spacing 0 among them. We overcame this co-occurrence limitation in the analysis by truncating the Pou5f1 matrix into the two independent Sox2 and Oct-4 matrices to perform the CRM search. This methodology is valid due to independence among positions in a PWM, where the score calculation at each position is independent from other positions. These two matrices were used to search Sox2/Oct-4 binding sites with a spacing distance up to 3 bp, the same as for Rodda-derived matrices. Additionally, the relative orientation between Sox2 and Oct-4 was also suppressed when splitting Pou5f1 matrix into two independent matrices.

Table 2. Rodda dataset motif search results (Results for Sox2- and Oct-4-derived PWM matrices.)

	Sox2						
Tool	TP	FP	FN	TN	SI	SP	MCC
INSECT	6	0	0	59 958	1.00	1.00	1.00
CPModule	5	1	1	59 957	0.83	1.00	0.83
MotifViz	6	38	0	59 920	1.00	1.00	0.37
Cluster buster	6	0	0	59 958	1.00	1.00	1.00
	Oct-4						
Tool	TP	FP	FN	TN	SI	SP	MCC
INSECT	6	0	0	59 952	1.00	1.00	1.00
CPModule	5	0	1	59 952	0.83	1.00	0.91
MotifViz	5	31	1	59 921	0.83	1.00	0.34
Cluster buster	5	4	1	59 948	0.83	1.00	0.68

Note: TP, FP, FN and TN refer to the number of true positives, false positives, false negatives and true negatives, respectively. SI is the sensitivity, SP is the specificity and MCC is the Mathews Correlation Coefficient.

Table 3. Rodda dataset motif search results [Results for Pou5f1 PWM matrix (JASPAR)]

	Pou5f1 JASPAR						
Tool	TP	FP	FN	TN	SI	SP	MCC
INSECT	5	3	1	59 907	0.83	1.00	0.72
CPModule	5	12	1	59 898	0.83	1.00	0.50
MotifViz	5	7	1	59 903	0.83	1.00	0.59
Cluster buster	4	3	2	59 907	0.67	1.00	0.62

Note: TP, FP, FN and TN refer to the number of true positives, false positives, false negatives and true negatives, respectively. SI is the sensitivity, SP is the specificity and MCC is the Mathews Correlation Coefficient.

The CRM searches of the Boyer dataset were performed using INSECT on the proximal promoter regions corresponding to the −1 kb/TSS for each gene because the ChIP peaks were detected experimentally within this regions. INSECT analysis of the Boyer dataset, with Rodda-derived matrices and threshold, found two genes only. For the case of the Pou5f1 matrix, 19 genes were hit. The number of hit genes increased to 31 when the partitioned Pou5f1 matrices were used (Table 4). The difference in the values obtained is most probably due to the Pou5f1 matrix possibly allowing for higher variability in the represented TFBS (Fig. 2), and higher variability of the TFBSs within the regulatory regions of genes from Boyer versus Rodda datasets. Because the true-positive binding sites are not known, MCC values cannot be calculated to optimize the threshold parameter, or to compare performance with other tools. Alternatively, the number of genes that had at least one valid match (positive genes), the total detected TFBS sites and the average number of sites per gene were reported. A score threshold analysis found that an 80% threshold improves the number of hit genes without compromising the number of detected sites per gene (data not shown). These results are summarized in Table 5.

4.3 Integration of INSECT results with other resources

In recent years, the number of databases containing experimental-derived information has proliferated. Databases such as Gene Ontology and the ENCODE project have provided researchers with powerful resources to assist in hypothesis construction and further validation (Raney *et al.*, 2011). INSECT allows the visualization of hits for positive genes as tracks in the UCSC Genome Browser, which permits examination of the detected CRMs

Table 4. Boyer dataset motif search results using INSECT (Results using the score thresholds and parameters derived from the analysis of the Rodda dataset)

Matrices	Thr	MaxS	HG	TFBS	APG	SR
Rodda PWMs	86	3	2	4	2.00	Yes
Pou5f1 JASPAR	86	–	19	24	1.26	–
Pou5f1 JASPAR truncated	86	3	31	87	2.81	Yes

Note: Thr, MaxS, HG, TFBS, APG and SR refer to threshold, maximum spacing allowed between the Sox2 and Oct-4 TFBS, total number of hit genes with detected sites, total number of TFBSs detected among all the genes, average TFBSs per gene and strand restriction, respectively.

Table 5. Boyer dataset motif search results using INSECT (Results using optimized parameters for this dataset)

Matrices	Thr	MaxS	HG	TFBS	APG	SR
Rodda PWMs	80	3	30	89	2.97	Yes
Pou5f1 JASPAR	80	–	52	107	2.06	–
Pou5f1 JASPAR truncated	80	3	70	610	8.71	Yes

Note: Thr, MaxS, HG, TFBS, APG and SR refer to threshold, maximum spacing allowed between the Sox2 and Oct-4 TFBS, total number of hit genes with detected sites, total number of TFBSs detected among all the genes, average TFBSs per gene and strand restriction, respectively.

genomic environment in the context of experimental information present with many available tracks. This INSECT feature is important because the uncovering of false positives is one of the most troublesome aspects of *in silico* CRMs search tools.

As a case study, we analyzed the INSECT results from the Boyer dataset using the Rodda-derived matrices (Table 5) for the *HOXB1* gene (ENSG000001200094) and its corresponding UCSC Genome Browser tracks (Fig. 4). We paid particular attention to the ENCODE TF ChIP-seq data tracks. As shown in Figure 4, histone marker tracks were added given more confidence to the CRM identification.

INSECT permits several analyses to be performed within the web server, making the analysis more focused and understandable in the context of a network. The main INSECT features are summarized in Table 6.

5 DISCUSSION

INSECT is designed as a user-friendly interface to minimize tool usage complexity along with integration of several methods that aim to assist in visualization and analysis of results. INSECT dynamically generates tracks that can be submitted and automatically opened within the UCSC Genome Browser to analyze the detected CRMs and their relationship to surrounding genomic regions. In this context, INSECT tracks can be analyzed along with tracks, as the ChIP-Seq, DNase I Hypersensitivity, RNA-Seq, DNA methylation information and many other features provided by the ENCODE project on the UCSC Genome Browser.

While INSECT only provides *in silico* estimation, the information generated using this tool may be a valuable guide for experimentation design and hypothesis construction. For example, it was previously suggested that the relative orientation between Sox2 and Oct-4 binding sites is critical, based on structural analysis of ternary complex among these two factors and DNA (Remenyi *et al.*, 2003). INSECT is the only solution among the tools evaluated here that can apply the strand restriction and restrict relative orientations between TFBSs within CRMs.

We have demonstrated how INSECT facilitates CRM search analysis on experimental data derived from different biochemical techniques in the Rodda dataset. We found that INSECT outperforms other tools in terms of sensitivity and specificity in cases where the true positive is known *a priori*. We also showed a more real exploratory analysis with ChIP data from the Boyer dataset. Finally, we analyzed the genomic environment of different CRMs detected by displaying them in the UCSC Genome Browser as a custom track. To detect those genes with more promising TFBS hits, analysis accounted for whether CRMs were located on regions corresponding to the ENCODE TFs, ChIP-Seq data, histone marks and regions of high transcriptional activity.

Many tools were developed to search for potential TFBSs within CRMs by implementing PWMs. However, there are differing criteria for scoring and determining whether a given subsequence is a potential TFBS or not. A group of tools statistically decides whether a substring constitutes a possible TFBS comparing its enrichment within a promoter region with respect to a background set and the subsequent calculation of *P*-values. There are several ways of achieving this comparison, as seen in the implementation of tools such as CPModule, MotifViz and Cluster Buster. As *P*-value-based algorithms require a background set for comparison, there are many variables that can affect results that are not inherent to the PWM definition itself, but the background. Some parameters, such as the length of the analyzed sequences, number of random PWM used to measure enrichment or even the motif lengths represented by the PWMs, can affect the results obtained by *P*-value-based methods. We considered that normalization of PWM scores by *MMSc* eliminates this problem, given that determination of quality for a site does not rely on comparison with random sequences, but the PWM itself. The latter relies on the hypothesis that, if the PWM is representative of the real motif, the *MMSc* value potentially has a more relevant biological meaning, as it relates directly to the affinity between TFs and their DNA target sites.

We have built a flexible method that allows researchers to define parameters to find potential CRMs instead of isolated

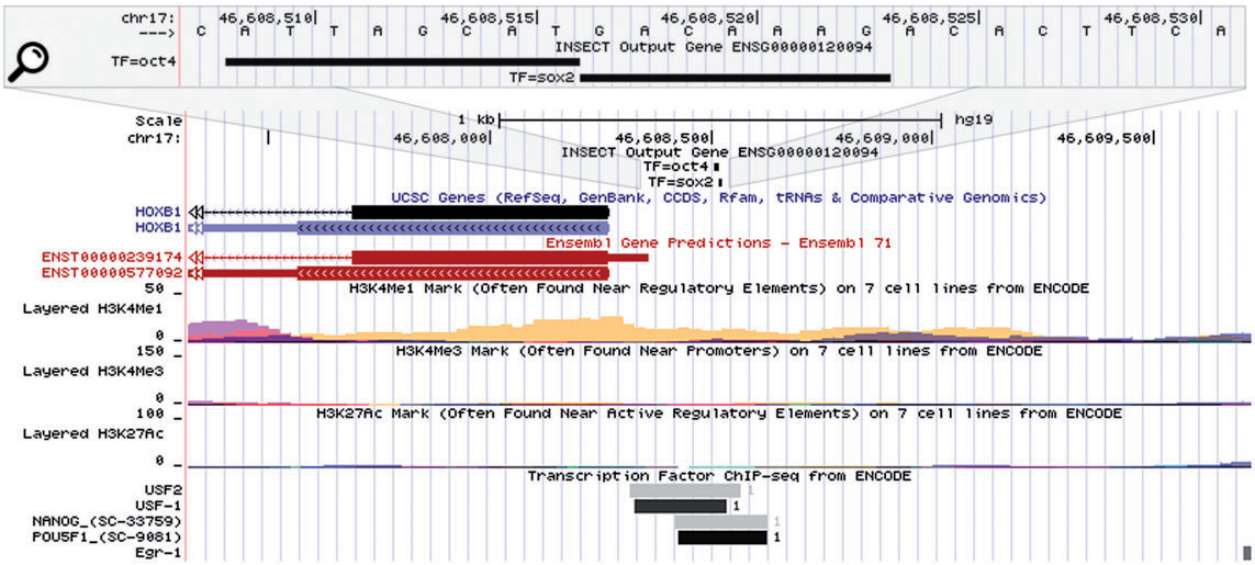

Fig. 4. INSECT visualization in UCSC Genome Browser for HOXB1 along with histone markers and TFs ChIP-seq tracks from ENCODE

Table 6. Feature comparison of INSECT with the three other motif search tools used in the performance analysis

		INSECT	CPModule	MotifViz	Cluster Buster
Inputs	TF Model				
	User PWM	Yes	Yes	Yes	Yes
	Create PWM from multiple sequence alignment	Yes	No	No	No
	TFs collections	JASPAR, Transfac and UniPROBE	No	JASPAR	JASPAR
	Sequence				
	Format	Fasta	Fasta	Fasta/GenBank	Fasta/GenBank
	Automatic retrieval	Yes (14 genomes)	No	Yes (GenBank ID)	Yes (GenBank ID)
	Strand restrictions	Yes	No	No	No
	Distance restrictions	Yes (window/master mode)	Yes (window)	No	Yes (average)
Outputs	TFBS positions/Sequences	GFF/Custom CSV	Custom output	Custom output	Custom output
	Local visualization	Yes, TFBS and transcripts	No	Yes, marked in sequence	Yes, with GenBank input
	UCSC Browser link	Yes	No	No	No
	GO information	Yes	No	No	No
Others	TFBS Filtering	Normalized Max. PWM value	*P*-value	*P*-value	Cluster and motif thresholds
	Phylogenetic footprinting	Yes	No	No	No
	Availability	Web server	Comand line	Web server/Command line	Web server

sets of TFBSs. By adding the phylogenetic footprinting search option, INSECT makes CRM identification more robust. We confirmed the search and result display capabilities of INSECT in two different datasets. The ability to perform analyses that are usually difficult, or not possible, with most of the available tools, because they are not available fully as web servers, command line programs are difficult to use or they do not offer the INSECT web server functionality, demonstrates the advantages of having an integrated motif search tool.

ACKNOWLEDGEMENTS

The authors thank James Smyth, Sergio Baranzini and Ken Kobayashi for critically reading the manuscript.

Funding: All the authors are members of the Consejo Nacional de Investigaciones Científicas y Técnicas (Argentina) (CONICET). Grants from CONICET and Agencia Nacional de Promoción Cientifica y Tecnológica (ANPCyT) PICT-2010-1587, Argentina.

Conflict of Interest: none declared.

REFERENCES

Arnone,M.I. and Davidson,E.H. (1997) The hardwiring of development: organization and function of genomic regulatory systems. *Development*, **124**, 1851–1864.

Ashburner,M. *et al.* (2000) Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nat. Genet.*, **25**, 25–29.

Berg,O.G. and von Hippel,P.H. (1987) Selection of DNA binding sites by regulatory proteins. Statistical-mechanical theory and application to operators and promoters. *J. Mol. Biol.*, **193**, 723–750.

Boyer,L.A. *et al.* (2005) Core transcriptional regulatory circuitry in human embryonic stem cells. *Cell*, **122**, 947–956.

Bryne,J.C. *et al.* (2008) JASPAR, the open access database of transcription factor-binding profiles: new content and tools in the 2008 update. *Nucleic Acids Res.*, **36**, D102–D106.

Dermitzakis,E.T. and Clark,A.G. (2002) Evolution of transcription factor binding sites in Mammalian gene regulatory regions: conservation and turnover. *Mol. Biol. Evol.*, **19**, 1114–1121.

Flicek,P. *et al.* (2013) Ensembl 2013. *Nucleic Acids Res.*, **41**, 48–55.

Frith,M.C. *et al.* (2003) Cluster-Buster: finding dense clusters of motifs in DNA sequences. *Nucleic Acids Res.*, **31**, 3666–3668.

Fu,Y. *et al.* (2004) MotifViz: an analysis and visualization tool for motif discovery. *Nucleic Acids Res.*, **32**, W420–W423.

Kirchhamer,C.V. *et al.* (1996) Modular *cis*-regulatory organization of developmentally expressed genes: two genes transcribed territorially in the sea urchin embryo, and additional examples. *Proc. Natl Acad. Sci. USA*, **93**, 9322–9328.

Lenhard,B. and Wasserman,W.W. (2002) TFBS: computational framework for transcription factor binding site analysis. *Bioinformatics*, **18**, 1135–1136.

Loh,Y.H. *et al.* (2006) The Oct4 and Nanog transcription network regulates pluripotency in mouse embryonic stem cells. *Nat. Genet.*, **38**, 431–440.

Matys,V. *et al.* (2006) TRANSFAC and its module TRANSCompel: transcriptional gene regulation in eukaryotes. *Nucleic Acids Res.*, **34**, D108–D110.

Newburger,D.E. and Bulyk,M.L. (2009) UniPROBE: an online database of protein binding microarray data on protein-DNA interactions. *Nucleic Acids Res.*, **37**, 77–82.

Park,P.J. (2009) ChIP-seq: advantages and challenges of a maturing technology. *Nat. Rev. Genet.*, **10**, 669–680.

Raney,B.J. *et al.* (2011) ENCODE whole-genome data in the UCSC genome browser (2011 update). *Nucleic Acids Res.*, **39**, D871–D875.

Remenyi,A. *et al.* (2003) Crystal structure of a POU/HMG/DNA ternary complex suggests differential assembly of Oct4 and Sox2 on two enhancers. *Genes Dev.*, **17**, 2048–2059.

Rodda,D.J. *et al.* (2005) Transcriptional regulation of nanog by OCT4 and SOX2. *J. Biol. Chem.*, **280**, 24731–24737.

Stormo,G.D. and Fields,D.S. (1998) Specificity, free energy and information content in protein-DNA interactions. *Trends Biochem. Sci.*, **23**, 109–113.

Sun,H. *et al.* (2009) ModuleDigger: an itemset mining framework for the detection of *cis*-regulatory modules. *BMC Bioinformatics*, **10(Suppl. 1)**, S30.

Sun,H. *et al.* (2012) Unveiling combinatorial regulation through the combination of ChIP information and in silico *cis*-regulatory module detection. *Nucleic Acids Res.*, **40**, e90.

Tronche,F. *et al.* (1997) Analysis of the distribution of binding sites for a tissue-specific transcription factor in the vertebrate genome. *J. Mol. Biol.*, **266**, 231–245.

Van Loo,P. and Marynen,P. (2009) Computational methods for the detection of *cis*-regulatory modules. *Brief. Bioinform.*, **10**, 509–524.

Wasserman,W.W. and Sandelin,A. (2004) Applied bioinformatics for the identification of regulatory elements. *Nat. Rev. Genet.*, **5**, 276–287.

Wray,G.A. *et al.* (2003) The evolution of transcriptional regulation in eukaryotes. *Mol. Biol. Evol.*, **20**, 1377–1419.