Unveiling combinatorial regulation through the combination of ChIP information and *in silico cis*-regulatory module detection

Hong Sun[1], Tias Guns[2], Siegfried Nijssen[2] and Kathleen Marchal[1,§]

[1]Department of Microbial and Molecular Systems, Katholieke Universiteit Leuven, Address Kasteelpark Arenberg 20, 3001 Leuven, Belgium

[2]Department of Computer Science, Katholieke Universiteit Leuven, Address Celestijnenlaan 200A, 3001 Leuven, Belgium

Introduction

Nowadays with high-throughput chromatin-immunoprecipitation technologies becoming increasingly popular for the genome-wide identification of TF binding sites, *cis*-regulatory module detection (CRM) can be used in combination with ChIP information to computationally predict with which other TFs a ChIP-assayed TF potentially interacts. In contrast to gene centered methods, ChIP information allows reducing largely the regions in which the motif of the assayed TF should be located (typically 500 bp instead of thousands of bp). However, as the binding site of the assayed TF often not coincides with the peak location, searching for CRMs in ChIP-Seq defined regions still boils down to a combinatorial search problem. In addition, as it is not known in advance with which other TF the assayed one interacts, the CRM detection approach needs to be able to search for a CRM that can include any of the known motifs.

Methodology

In this study we developed an analysis flow (Figure 1) that allows performing CRM detection on ChIP-defined regions by combining a powerful combinatorial search algorithm with a strategy to reduce the search space in a biologically motivated way. The latter is done by constraining the number of possible motif sites during the screening step using epigenetic filtering and the number of valid motif combinations during the combinatorial search. The combinatorial search is performed by CPModule, a novel approach of CRM detection with a performance that is competitive to that of other state-of-art tools but that in contrast to previous tools can handle much larger datasets (such as 100 sequences in combination with a library of 516 PWMs). The advantage of CPModule is that it builds upon a constraint based itemset mining framework CP4IM: this offers the advantage of flexibly adding relevant constraints and a straight forward application of existing itemset mining principles. This allowed us to use CPModule in a query-based setting, searching for modules only that contained our motif of interest, i.e. the motif of the assayed TF and that meet other biologically relevant constraints that help us to prioritize the most likely biologically true modules, such as encompassing a restricted region (proximity constraint) or occurring in a high number of sequences (frequency constraint (support)). Benchmarking with other state-of-the-art CRM tools shows that CPModule is competitive with other CRM tools in effectively searching CRMs in large sequence sets, even in the presence of a considerable amount of noisy motif sites.

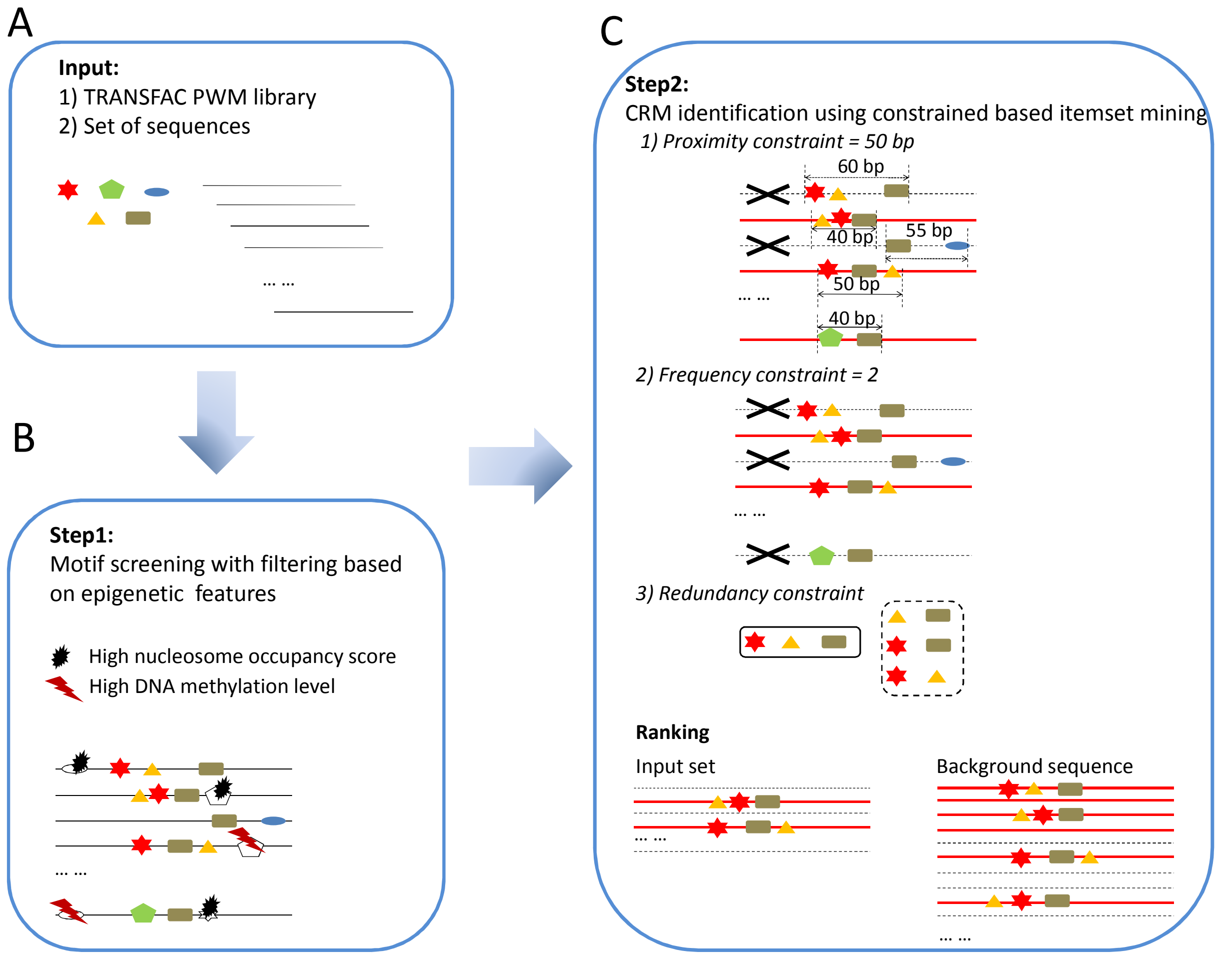
A
Input:
1) TRANSFAC PWM library
2) Set of sequences
… …
B
Step1:
Motif screening with filtering based on epigenetic features
High nucleosome occupancy score
High DNA methylation level
… …
C
Step2:
CRM identification using constrained based itemset mining
1) Proximity constraint = 50 bp
60 bp
40 bp
55 bp
50 bp
40 bp
… …
2) Frequency constraint = 2
… …
3) Redundancy constraint
Ranking
Input set
Background sequence
… …
… …

Figure 1: Analysis flow. The input consists of a library of PWMs and a set of sequences. In a first step prior to the actual CRM detection a screening with public motif databases is performed. Here we combine standard PWM screening with filtering based on epigenetic features. Only regions containing a motif site that display a low GC content and a low nucleosome occupancy are withheld. The second step consists of the actual combinatorial search. Here we use a constrained based itemset mining approach to enumerate all valid CRMs i.e. combinations of motifs that 1) of which the motif sites contributing to the CRM occur in each others proximity (user defined) 2) that occur frequent in the input set (i.e. in all sequences displayed in red) 3) that are non-redundant. Valid CRMs are finally ranked based on their specificity for the input set.

Results

We demonstrate the performance of our analysis method on real ChIP-based experiments conducted by Chen *et al.* 2008 for five key transcription factors KLF4, NANOG, OCT4, SOX2 and STAT3 involved in self-renewal of mouse embryonic stem cells. We used the previously described combinatorial interactions amongst those TFs as a benchmark. Table 1 displays which of the previously described CRMs involved in self renewal could be recovered by CPModule and also displays their rank amongst the total number of all possible CRMs or of all that contain the ChIP-assayed TF. To further validate the detected CRMs we use the ChIP-Seq of Chen *et al.* 2008 in a cross validation set up: we verified whether the motifs contributing to the predicted CRMs fell within the binding peaks of the other ChIP-Seq-assayed TFs: the reported modules were validated in at least 10% of the cases by the ChIP-Seq data of the cognate validation sets. For instance when considering the CRM composed of SOX2 and OCT4: here we could predict by performing CRM detection on the ChIP-Seq regions identified for SOX2 that it most likely interacts with OCT4. This retrieved module was ranked first amongst the 22 potential CRMs that contained OCT4. OCT4 and SOX2 co-occurred in 63% of the SOX2 ChIP-Seq identified regions within a distance of 150 bp and the identified sites for OCT4 fell within the identified OCT4 ChIP-Seq regions in 79% of the cases. Table 1 also clearly shows the added value of using ChIP-Seq data to constrain the search by querying only those CRMs that contain the motif of the assayed TF. This is illustrated by the rank of the 'true module' amongst the possible number of CRMs (so not only those containing the sites of the ChIP-Seq-assayed TF). By enumerating all possible CRMs and ranking them based on their statistical significance, CPModule allows having an insight in the position of a certain CRM amongst all possible CRMs. For this dataset it seems that of the benchmark CRMs mainly those containing STAT3 sites rank poorly. This is probably due to the low specificity of the screening results obtained with the STAT family of TFs: after screening and epigenetic filtering we still obtain on average 11 sites per sequence, indicating that STAT3 sites are frequently occurring sites in the genome. Such high genomic frequency deteriorates the specificity of CRMs containing STAT3 sites for the set of input sequences and decreases their rank. Without ChIP-Seq data these CRMs would never be considered.

Comparing the outcome obtained on the same dataset with different screening strategies also showed that the quality of the screening input largely affects the outcome of the combinatorial search. A too dense screening obtained by a non-stringent screening threshold results in too many motif combinations that make the problem intractable or in case an output is obtained decreases the prediction power (too many false positive valid combinations are possible). Just increasing the stringency of the screening seems not to be an option as then many true sites and thus also true CRMs seem to be missing. With the availability of ChIP-Seq and ChIP-chip data for eukaryotic TFs, it indeed becomes increasingly clear that only in few cases the physically bound sites correspond to the 'best conserved or highest scoring' sites obtained with a PWM screening. This is probably partially due to the fact that PWMs stored in public database are biased towards sites discovered by their resemblance to the already stored motif model (circular reasoning) but also because other physical factors such as chromatin positioning determine the accessibility of a site. Using a lower screening threshold in combination with a filtering procedure based on epigenetic features seemed to provide a good trade off between recovering true sites while still keeping the number of false positives within a reasonable range.

Conclusion

Our results illustrate that using ChIP-Seq information together with combinatorial CRM detection is able to prioritize true combinatorial interactions between the assayed TF and any other TF. The success of our approach stems from combining ChIP-Seq information to not only determine a set of coregulated genes, but to also delineate the region in which at least the assayed TF binds with a powerful combinatorial approach that allows detecting combinations of the binding site of the assayed TF with any other known TF for which a PWM have been reported.

Table 1: CRMs obtained with CPModule in combination with epigenetic filtering (non-stringent screening with filtering for all TFs except the assayed one). The set of sequences corresponding to the

100 top scoring peak region of the assayed TF were screened with a set of 516 non-redundant TRANSFAC motifs using a non-stringent screening threshold. Epigenetic filtering was applied on all motif sites except on the ones of the assayed TF.

ChIP-Seq-assayed TF	CRM	Rank	Support	Cross validation	Proximity constraint (bp)	Total number of solutions/Number of solutions containing the ChIP-Seq-assayed TF	Percentile of rank
KLF4	KLF4, STAT4	143/2	60%	40.00%	300	147/3	97.28%/66.67%
NANOG	NANOG, OCT1	6846/4	61%	70.49%	300	6868/17	99.68%/23.53%
	NANOG, STAT3	14017/10	60%	25.00%	350	14033/26	99.89%/38.46%
OCT4	OCT4, STAT1, [XFD2, STAT4, STAT6]	5/5	63%	11.10%	150	5068/613	0.99%/0.82%
SOX2	SOX2, OCT4	430/1	63%	79.40%	150	14180/22	3.03%/4.55%
	SOX2, STAT3, [CDXA, PAX2, STAT5A]	61807/24	60%	23.33%	250	117006/189	52.82%/12.70%
STAT3	STAT3, OCT4, [STAT1, STAT5A, STAT6]	1/1	61%	24.59%	150	1366/20	0.07%/5.00%

ChIP-Seq-assayed TF: TF from which the top 100 binding peaks were used to perform the analysis. **CRM:** obtained CRMs that correspond to previously well described modules for the assayed TF**; [**between brackets are indicated other TFs that were predicted to belong to the same CRM, but that have not previously been described to interact with the assayed TF**]. Rank:** rank this CRM received (ranks were assigned by taking into account all the found CRMs/CRMs that contained the assayed TF)**. Support:** the percentage of sequences from the input set in which this CRM occurs (should be higher than the frequency constraint). **Cross validation:** we started from the ChIP-Seq data of one TF and tried to predict using CRM detection with which other TFs the assayed TF interacts. We verified whether the motif sites contributing to the predicted CRMs fell within the binding peaks of the other ChIP-Seq-assayed TFs. **Proximity constraint (bp):** the proximity constraint at which the displayed CRM was found. **Total number of solutions/Number of solutions containing the ChIP-Seq-assayed TF:** the total number of valid CRMs/the number of solutions containing the motif for the ChIP-Seq-assayed TF. **Percentile of rank:** the percentile of the rank comparing with the total number of solutions/total number of solutions containing the motifs for the ChIP-Seq-assayed TFs.