Discovery of regulators for co-expressed human genes using large sequence search spaces

Bram Van de Sande, Zeynep Kalender Atak, and Stein Aerts

Laboratory of Computational Biology, Center for Human Genetics, University of Leuven, Belgium.

Introduction

The importance of gene expression profiles provided by microarray and RNA-Seq experiments in disease research, and more specifically in the cancer field, is universally recognized: many studies have shown the added value of this data in the prediction of the prognosis (1) and in our understanding of the pathogenesis of cancer. Knowledge of the causal biochemical mechanisms for these abnormal expression signatures, would even further improve our insights. However, elucidating the perturbed pathways and transcriptional regulation that underlie these aberrant expression profiles remains a challenge.

The gap between gene expression and biochemical pathways is bridged by transcription factors (TF). These TFs bind DNA at specific binding sites that are modeled as Position Weight Matrices (PWM)/motifs. The main challenge can thus be reformulated as the prediction of the motifs that drive a (sub-)set of differentially expressed genes. Classical approaches are based on de novo motif discovery in the proximal promotors of co-expressed genes. However, transcriptional regulation in eukaryotes, and especially in human cells, is complex and also controlled via intronic regions and more distant acting enhancers. Unfortunately, extending the search space to include these more distant regulatory active regions renders classical motif discovery unproductive because of the extended background noise. Therefore, motif discovery is only able to find regulatory signals in search spaces restricted to a region of around 2kb upstream of the TSS. Additionally, most regulatory regions are clustered in Cis-Regulatory Models (CRM) and this kind of information is usually not taken into account by motif discovery methods.

For Drosophila melanogaster, a more robust computational analysis technique named CisTargetX (2) is available for the discovery of regulatory motifs that drive a given gene signature. This technique is able to search large genomic spaces for regulatory regions and takes clustering of motifs into account. To overcome the aforementioned problems, we ported the cisTargetX algorithm to handle sets of human genes. This new computational approach will be of great value to the cancer field by enabling researchers to predict the regulatory motifs, and if known, their corresponding TFs, that underlie a given transcriptional aberrant state.

Results

Determination of the optimal search space

CisTargetX for the human genome needs an adequate delineation of the regulatory active regions of the genome and a precise mapping of these regions to the genes they regulate. To this end, we mapped ChIP-chip fragments of 127 transcription factors relative to the TSS of their nearest gene. The region upstream and downstream of the TSS was binned in 5kb intervals and the number of overlapping fragments for each bin was tallied. The resulting histogram shows that regulatory regions are concentrated around the TSS in a region extending 10kb upstream and downstream into the 5' UTR and the first intron (Figure 1.A). The same analysis was performed relative to the end of the nearest transcript, indicating an additional regulatory role for the 3'UTR region and the region downstream of the transcript.

Based on these results, we defined two regulatory search spaces. The first set encompasses the 10kb upstream region extended with the 5'UTR and first intron downstream of the TSS,

excluding coding exons, and limiting the 10kb upstream region by the transcript of the nearest upstream gene. For 21798 HGNC genes a region could be extracted, 486 were lost because of ambiguous genome locations. The second set of regulatory candidate regions was created for all unambiguously UCSC annotated genes, now extending the full transcript with the 10kb limited upstream region and the 5kb region downstream of the transcript. The latter region was also limited by the nearest downstream transcript. Again, any coding sequences overlapping with this region were excluded. This new set effectively encompasses the previous set, hopefully capturing additional regulatory signals in the 3'UTR and 5kb region downstream of the transcript.

CisTargetX: a method to identify over-represented motifs in millions of basepairs

We have developed a methodology for motif discovery in human co-expressed genes that is similar to CisTargetX for Drosophila melanogaster (2). Briefly, CisTargetX starts by scoring the above-defined search space for the presence of homotypic clusters of binding sites of a known PWM via Cluster-Buster (3). This is performed for a collection of 3731 PWMs, resulting in a score-based ranking of the whole genome for each motif in the collection. Next, corresponding regulatory regions in multiple related species are scored for the presence of these clusters. Combining these rankings via order statistics results in a single whole genome raking for each motif. Finally, given a set of co-expressed genes, the recovery of these genes based on the ranking associated with each motif in the collection is assessed by calculating the cumulative recovery of these genes with increasing gene rank. Of special interest is the early retrieval of genes in the "query" gene set, therefore the area under the curve (AUC) for the top ranked genes is used as a metric to quantify the enrichment of these genes at the top of a ranking. The distribution of this metric for all motifs provides a method to define exceptionally good recovery, i.e. a z-score ($=(AUC-\mu_{AUC})/\sigma_{AUC}$) is computed for each motif and only motifs associated with a recovery above a certain threshold are considered possible regulatory drivers. Note that the calculation of the whole-genome gene ranking is performed only once and reused for multiple recovery analyses on different co-expressed gene sets. This effectively reduces the computational burden for cisTargetX to the calculation of recovery curves, making cisTargetX an on-the-fly analysis tool.

Benchmark analysis

To validate this technique, we used sets of known or candidate target genes for 17 human transcription factors from the AMADEUS benchmark (4). An additional prerequisite for these sets was the availability of a known motif modeling the TF bindings site. By using the z-score of a motif, CisTargetX provides a ranked list of motifs for each TF in our benchmark. To account for the inherent similarity in the motif collection, STAMP (5) was used to cluster the database of 3731 motifs in 914 distinct clusters. By equating the best motif rank in a cluster to the rank of the whole cluster, this ranked list of motifs was converted into a ranked list of clusters. The cumulative recovery of the clusters that contain the known motifs associated with the benchmark sets with increasing cluster rank resulted in a motif cluster recovery curve for the whole benchmark. As a control we shuffled the union of all target genes in the benchmark and compiled 25 new random sets for each transcription factor, keeping the cardinality of the set for each TF constant. The cluster recovery for these control sets was also calculated and averaged (Figure 1.B).

For the 10kb upstream-5' UTR- 1st intron rankings, 6 out of 17 benchmark sets (35% of sets) have there known motif ranked at position one, and 9 sets (53% of sets) are ranked among the top 10 motifs, i.e. in the top 1% of motif clusters. To quantify the early retrieval of known benchmark motifs, we determined the area under the curve for the first 5% ranked motif clusters. As expected, our benchmark had a higher AUC (0.61) value compared to the control case (0.17). Additionally, a wilcoxon ranksum test on the cluster ranks was performed, yielding an average p-value of 0.0040. This highlights the significantly better discovery of known motifs by cisTargetX compared to the control. The recovery for the limited 10kb upstream, 5'UTR and 1st intron regions is slightly better than that for the full transcript

combined with upstream and downstream regions. However this slight improvement in recovery is not statistically significant (p-value of Wilcoxon rank-sum test on cluster ranks is 0.55). In conclusion, our method allows the correct identification of the motif underlying the regulation of a set of co-expressed human genes using large regulatory search spaces.

Determining the optimal combination of vertebrate species

Frequently, the expression of a gene is conserved across related species implicating a possible conservation of the underlying CRM. To benefit from this conservation of CRMs, orthologuous regulatory regions in 10 related species were deduced using the LiftOver standalone program from the Kent software suite available from the UCSC web site and, as previously mentioned, integrated via order statistics in an overall genome-wide ranking. The 10 species used in our analysis are Chimpanzee (Pan troglodytes), Macaque (Macaca mulatta), Mouse (Mus musculus), Rat (Rattus norvegicus), Dog (Canis familiaris), Cow (Bos taurus), Opossum (Monodelphis domestica), Chicken (Gallus gallus), Pufferfish (Tetraodon nigroviridis) and Zebrafish (Danio rerio). Using the same benchmark data set for different combinations of species, we have determined the optimal combination of related species for CRM discovery in human.

From motif discovery to target genes and regulatory networks

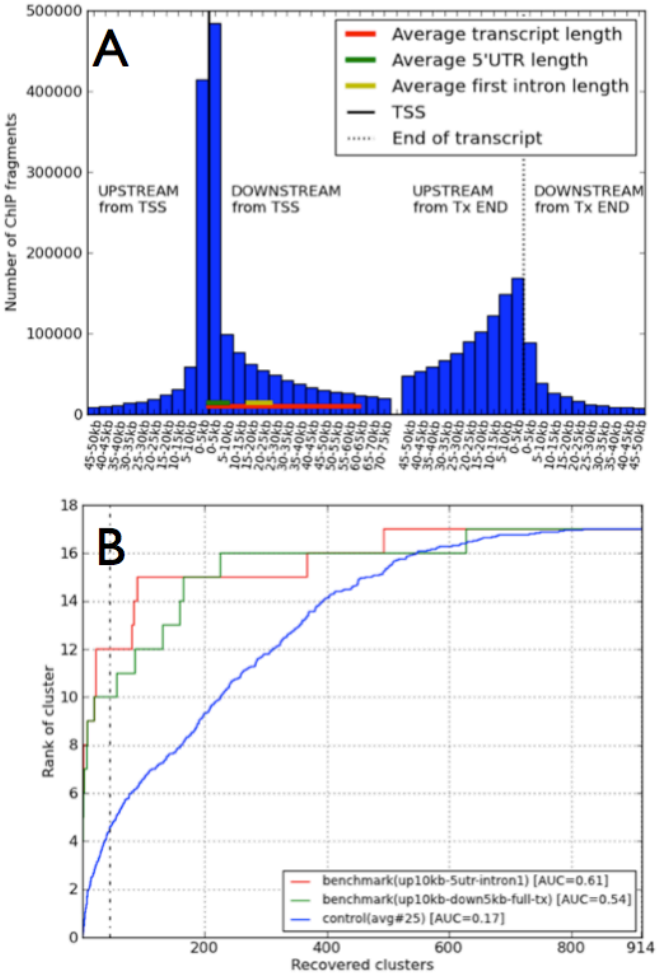
Not only does cisTargetX predict the motifs that regulate a given set of related genes, based on the recovery curve associated with each motif a list of candidate target genes for the predicted motif is also provided. For an enriched motif, the maximum deviation of its recovery curve from the average recovery over the whole motif collection serves as a cutoff that effectively filters the original set of genes to the ones most plausible controlled by the motif under investigation. For the motifs in our collection that were taken from TRANSFAC (6) and JASPAR (http://jaspar.genereg.net/) the TF for which they model the binding site is known. Additionally, extending these mappings to similar motifs based on the STAMP clustering significantly extends this TF-motif association. In this way, motif to candidate target genes associations predicted by cisTargetX can be translated to small regulatory networks connecting every TF with its multiple candidate target genes. TF-target relationships are the building blocks of Gene Regulatory Networks (GRN) in which transcriptionally interacting genes are connected. Thus, CisTargetX takes us from motif discovery to gene regulatory networks.

Application to cancer gene signatures

The vast amount of publicly available gene expression profiles are being curated and catalogued as sets of up- or downregulated genes, i.e. gene signatures. One of such efforts, GeneSigDB (7), contains 1865 human signatures of which many are cancer related. This collection enables us to extract cancer-related TF-target genes relationships by applying cisTargetX on each signature, culminating in the construction of large cancer involved GRNs which will reveal hotspots relevant in the pathogenesis of cancer. For example, we found TP53 (1st motif, MA0106 from JASPAR4) in signature 18366635-Table1 (8) together with RUNX (2nd position, motif M00769 from TRANSFAC) and ETS (8th position, motif M00971 from TRANSFAC). This signature of upregulated genes in breast cancer was derived from a microarray experiment and the resulting subnetwork, created via Cytoscape (9), is shown in Figure 1.C.

Conclusion

In this abstract we present a new computational technique to predict the transcription factors that explain a set of human co-expressed genes. This new approach is more robust than classical de novo motif discovery in that it can search large genomic search spaces for regulatory signals. Moreover, this approach also allows the construction of GRNs and enables the venture from regulatory genomics into systems biology.



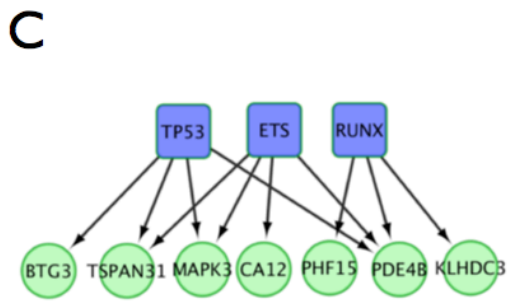


Figure 1
A. Distribution of ChIP-chip fragments of 127 transcription factors relative to the start and end of the nearest transcript. The average length of a transcript and its 5'UTR and first intron are shown inside the histogram. TSS = Transcription Start Site, 5'UTR = 5' UnTranslated Region

B. The cumulative recovery for each benchmark TF motif. The AUC (=Area Under the Curve) is calculated for the clusters ranked in the top 5% (N=45).

C. Small Gene Regulatory Network (GRN) extracted from a set of upregulated genes in breast cancer (8) via cisTargetX. The 3 transcription factors TP53, ETS and RUNX are displayed at the top and connected to their candidate target genes at the bottom.

References

1. Kretschmer C, Sterner-Kock A, Siedentopf F, Schoenegg W, Schlag P, Kemmner W. Identification of early molecular markers for breast cancer. Mol Cancer 2011 Feb;10(1):15.
2. Aerts, Quan, Claeys, Naval Sanchez, Tate, Yan, Hassan. Robust target gene discovery through transcriptome perturbations and genome-wide enhancer predictions in Drosophila uncovers a regulatory basis for sensory specification. PLoS Biol 2010 Jan;8(7):e1000435.
3. Frith, Li, Weng. Cluster-Buster: Finding dense clusters of motifs in DNA sequences. Nucleic acids research 2003 Jul;31(13):3666-8.
4. Linhart C, Halperin Y, Shamir R. Transcription factor and microRNA motif discovery: the Amadeus platform and a compendium of metazoan target sets. Genome Research 2008 Jul;18(7):1180-9.
5. Mahony S, Benos P. STAMP: a web tool for exploring DNA-binding motif similarities. Nucleic acids research 2007 Jul;35(Web Server issue):W253-8.
6. Matys V, Kel-Margoulis O, Fricke E, Liebich I, Land S, Barre-Dirrie A, Reuter I, Chekmenev D, Krull M, Hornischer K, Voss N, Stegmaier P, Lewicki-Potapov B, Saxel H, Kel A, Wingender E. TRANSFAC and its module TRANSCompel: transcriptional gene regulation in eukaryotes. Nucleic acids research 2006 Jan;34(Database issue):D108-10.
7. Culhane, Schwarzl, Sultana, Picard, Picard, Lu, Franklin, French, Papenhausen, Correll, Quackenbush. GeneSigDB--a curated database of gene expression signatures. Nucleic Acids Research 2010 Jan;38(Database issue):D716-25.
8. Natowicz R, Incitti R, Horta E, Charles B, Guinot P, Yan K, Coutant C, Andre F, Pusztai L, Rouzier R. Prediction of the outcome of preoperative chemotherapy in breast cancer using DNA probes that provide information on both complete and incomplete responses. BMC Bioinformatics 2008 Jan;9:149.
9. Smoot M, Ono K, Ruscheinski J, Wang P, Ideker T. Cytoscape 2.8: new features for data integration and network visualization. Bioinformatics 2011 Feb;27(3):431-2.