ISMARA: Completely automated inference of gene regulatory networks from high-throughput data

Mikhail Pachkov[1,2*], Piotr J. Balwierz[3], Phil Arnold[4], Andreas J. Gruber[1,2], Mihaela Zavolan[1,2] and Erik van Nimwegen[1,2]

[1]Biozentrum, University of Basel, Basel, Switzerland
[2]Swiss Institute of Bioinformatics, Basel, Switzerland
[3]MRC Clinical Sciences Centre, London, United Kingdom
[4]Novartis Institutes for BioMedical Research, Basel, Switzerland

* Corresponding author: Mikhail Pachkov <pachkov@gmail.com>

Introduction

As the costs of high-throughput measurement technologies continue to fall, experimental approaches in biomedicine are increasingly data intensive and the advent of big data is justifiably seen as holding the promise to transform medicine. However, as data volumes mount, researchers increasingly realize that extracting concrete, reliable, and actionable biological predictions from high-throughput data can be very challenging. Our laboratory has pioneered a number of methods for inferring key gene regulatory interactions from high-throughput data. For example, we developed motif activity response analysis (MARA)[1], which models genome-wide gene expression (RNA-Seq, or microarray) and chromatin state (ChIP-Seq) data in terms of comprehensive predictions of regulatory sites for hundreds of mammalian regulators (TFs and micro-RNAs). Using these models, MARA identifies the key regulators driving gene expression and chromatin state changes, the activities of these regulators across the input samples, their target genes, and the sites on the genome through which these regulators act. We recently completely automated MARA in an integrated webserver (ismara.unibas.ch)[2] that allows researchers to analyze their own data by simply uploading RNA-Seq or ChIP-Seq datasets, and provides results in an integrated web interface as well as in downloadable flat form.

Methods

The main aim of the analysis is to identify which regulatory motifs play an important role and how these motifs contribute into explaining expression changes across the samples. ISMARA infers the motif activities according to a linear model using a Bayesian procedure described in details in Balwierz et al.[2]. Briefly, the log-expression (or ChIP-Seq signal) value E_{ps} of a promoter p in sample s is modeled as a linear function of the site-counts N_{pm} for all motifs m associated with the promoter p: $E_{ps} = c_s + c_p + \Sigma_m N_{pm} A_{ms}$, where c_s *and* c_p are sample and promoter dependent constants, and A_{ms} is an unknown activity of a motif m . By fitting the model above we compute activities for each motif in each sample and corresponding uncertainties that allows us to quantify significance of a motif for explaining the expression variation across the samples. As well we predict which promoters are targets of a motif and quantify their significance.

To perform the ISMARA analysis we first need to define a set of promoters for a given organism and second we need to predict transcription factor binding sites (TFBS) for the given promoter set. The promoter set is defined on the basis of existing transcript annotation and experimentally verified transcription start sites (TSS) if available. The annotated transcript starts and measured TSS are clustered together forming a promoterome, which is used for the expression quantification and for prediction of the regulatory sites. Each promoter sequence from the set is extended by 500bp upstream and 500bp downstream. For a given extended sequence we extract orthologous sequences from other organisms, e.g. for human and mouse we extract orthologous sequences from rhesus macaque, cow, dog, horse and opossum. For TFBS prediction we use Motevo algorithm[3], which integrates a suite of Bayesian probabilistic methods for the prediction of regulatory sites on multiple alignments of phylogenetically related sequences. Motevo use the set of multi-aligned orthologous sequence and our own curated set of weight matrices[4] for accurate

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.3328v1 | CC BY 4.0 Open Access | rec: 6 Oct 2017, publ: 6 Oct 2017

prediction of regulatory sites. For details on promoterome construction and TFBS predictions please see supplementary materials in Balwierz et al.[2].

Results

The ISMARA webserver provides web-interface for uploading expression or chromatin dynamics datasets in variety of formats and without any limits for dataset size. Currently ISMARA support: *a*) microarray data for human, mouse and yeast, accepting files in .CEL format, currently 23 affymetrix chips in total are supported, *b*) unmapped RNA-Seq and ChIP-Seq data for human, mouse and rat, accepting files in FASTQ format, *c*) mapped RNA-Seq and ChIP-Seq data for human, mouse and rat, accepting files in .BAM or .BED format. There is also “expert mode” option where user can upload self-prepared expression and site-count tables for the ISMARA analysis. One of our aims was to simplify user interaction with the tool as much as possible, so there is only a necessary minimum of options for user to set.

The ISMARA results are presented as a collection of interactive webpages. The results pages could be downloaded for offline browsing. The main report page shows motif list sorted according to their significance so a user can immediately see what are the inferred key regulators in the current dataset. Each motif is annotated with list of associated transcription factors, activity graph and a motif logo. For every motif there is a corresponding page containing the motif annotation, correlation between the associated transcription factors expression and the activity of the motif, and a list of the motif targets sorted by their significance. The predicted motif targets allow us to perform additional analysis such as enrichment of target genes in gene ontology categories[5] and categories from MSigDB[6], constructing a direct interaction network between transcription factors, clustering of gene targets according to StringDB[7] on the basis of known protein-protein interactions. ISMARA also provide per sample overview showing, which motifs are most significant in the given sample and how the motif targets are expressed in the given sample. It is also possible to see target pages where in interactive manner user can see how well the model fit the observed promoter expression and what is a contribution of each motif in the explaining of the experimental data.

The ISMARA webserver is free for non-commercial use.

References

1. Suzuki H, Forrest ARR, van Nimwegen E, *et al.* (2009) The transcriptional network that controls growth arrest and differentiation in a human myeloid leukemia cell line. *Nature Genetics*, 41, 553-562
2. Balwierz PJ, Pachkov M, Arnold P, *et* al. (2014) ISMARA: automated modeling of genomic signals as a democracy of regulatory motifs. *Genome Research*, 24(5), 869-884
3. Arnold P, Erb I, Pachkov M, *et al.* (2012) MotEvo: integrated Bayesian probabilistic methods for inferring regulatory sites and motifs on multiple alignments of DNA sequences. *Bioinformatics*, 28(4), 487-494
4. Pachkov M, Balwierz, Arnold P, *et al, (2013)* SwissRegulon: a database of genome-wide annotations of regulatory sites. *Nucleic Acids Research*, 41, D214-D220
5. The Gene Ontology Consortium (2015) Gene Ontology Consortium: going forward. *Nucleic Acids Research*, 43(D1), D1049–D1056
6. Subramaniana A, Tamayoa P, Mootha VK, *et al.* (2005) Gene set enrichment analysis: A knowledge-based approach for interpreting genome-wide expression profiles. *PNAS*, 102(43), 15545–15550
7. Szklarczyk D, Morris JH, Cook H, *et al.* (2017) The STRING database in 2017: quality-controlled protein-protein association networks, made broadly accessible. *Nucleic Acids Research*, 45, D362-D368