CSC4112

Critical Review

Registration No.: SC/2020/11730

Student Name: Navinda Hewawickrama

November 2025

Bachelor of Computer Science (Special) Degree
Department of Computer Science, University of Ruhuna

The study "A Next Generation Connectivity Map: L1000 Platform and the First 1,000,000 Profiles", was published in **Cell**, on November 30th 2017 by Aravind Subramanian, Todd R. Golub, and many others. The main purpose of this paper is to show the new scale up version of the **Connectivity Map (CMap)**, a foundational functional look up table that connects genes, drugs, and disease states by virtue of common gene-expression signatures. The L1000, a new, low cost, high throughput reduced representation expression profiling method, enables generation of over a million profiles. This publication was carried out within the NIH Library of Integrated Network-Based Cellular Signatures (LINCS) Consortium which sets the context of the viability and usefulness of developing a genuinely comprehensive, genome-scale functional resource. This analysis will give a composite evaluation of the scale-up transformation brought by the L1000 program and critically assess the methodological limitations of the program, especially genetic perturbation and target recovery rates.

1 Summary

1.1 Introduction

To truly understand Cellular functions, perturbing the system (genetically or chemically) and monitoring downstream consequences, ideally using a "functional look-up table" is required. Modern biomedical research has benefited a lot from genomic resources that provide details about genes and their associations with disease. The original CMap pilot shows this concept, but due to its small scale the usage of it is very limited. By introducing the L1000 dataset, the paper addresses this limitation — a reduced representation approach to expand CMap dramatically (Lamb et al., 2006).

1.2 Literature Review (Comparison of Approaches)

The expansion of the CMap was needed to overcome the prohibitive cost of standard gene expression profiling. The authors looked at the proposed L1000 method against three pros/cons like approaches.

- **Original CMap Approach (Affymetrix Microarrays):** The CMap concept that was used earlier, as a pilot study, used expensive affymetrix microarrays (Lamb et al., 2006). This method uses profiles of only 164 drugs and 3 cancer cells, which was a very small scale. The paper says that this scale is too small to be shown as a true genome scale resource, showing that there is a need for a cheaper alternate resource for this.

- **Standard RNA Sequencing (RNA-seq):** The standard for gene expression is noted usually as the RNA-seq, becuase of its unbiased nature. However, even in 2017, the cost of RNA-seq was too high to profile the millions of samples that was needed to create a very comprehensive CMap. Also what adds more to the cost is the fact that deep sequencing required to detect very rare transcripts. The L1000 looks at all these aspects where it helps to detect rare transcripts at a low cost and all the other problems with the CMap, due to begin hybridization based, and it shows that it is actually a superior high-throughput solution for these specific problems in CMap.

- **Alternative Reduced Representation Methods:**The method of understanding a function from a gene expression compendium started with Hughes and colleagues, while they were working in yeast. Som,e like Donner, proposed other computational methods for imputing gene expression from reduced probe sets actually exist. The L1000 changes by selecting its 1000 landmark transcripts using a dimensionality reduction approach that is unbiased and data driven (Principal Component Analysis and tight clustering of 12,031 Affymetrix profiles) that is optimized for max information recovery(82% of the full transcriptome) insted of using biological knowledge pr alternative probe selection methods, which ususally requires prior knowledge(Donner et al., 2012; Hughes et al., 2000).

  The limitations of scale and cost is overcome by the **L1000 platform**by using ligation-mediated amplification (LMA) coupled with fluorescently addressed microspheres. The reagent cost is reduced significantly, making the massive scale-up tractable.

1.3 Methodology

THe main methodology centers on the L1000 assay, which uses a reduced representation of the transcriptome.

- **Landmark Selection and Assay:** The profile of public expression was analyzed and found that 1,000 landmark transcripts were enough to replicate 82% of data contained in the entire transcriptome, and that these were selected through a data-driven clustering method. L1000 assay is based on the adaptation of ligation-mediated amplification (LMA) with capture on fluorescently addressed microspheres. Since there were only 500 colors of beads, a deconvolution method was designed to have two transcripts marked with the same bead color, which enabled determination of 978 landmark transcripts and 80 controls transcripts.

- **Data Processing and Inference:**The raw data is processed through a series of levels which include; normalization through L1000 Invariant Set Scaling (LISS) and also quantile normalization (QNORM). Ordinary least-squares (OLS) regression models are used to infer the expression of the rest of the transcripts using the measured landmark transcripts(Subramanian et al., 2017).

- **Signature Generation and Query:**Biological replicates (3 or more are most common) are reduced to a Consensus Gene Signature (CGS) that employs a moderated z score (MODZ) methodology to reduce outlier effects, which is of special relevance in shRNA experiments (Peck et al., 2006). The Weighted Connectivity Score (WTCS) is used to compare queries to the CMap database, and converted to Tau (τ) (range,100 to 100), providing a standardized score of the association between the query and a perturbagen as compared to the whole reference database.

1.4 Results and Findings

CMap-L1000v1 compendium has 1,319,138 L1000 profiles, which consists of 42,080 perturbagens, of which almost 20,000 are small molecules and thousands of genetic perturbations. This unparalleled magnitude was confirmed to be strong, and with high technical reproducibility (median pairwise correlation ¿0.9).

- **Inference Accuracy:**The existence of inference models accurately predicted 9,196 of 11350 inferred genes (81%). Inference was also found to be critical to connections, with 20 percent of all the anticipated connections being lost in case of landmarks only.
- **ShRNA Analysis:**ShRNA signatures comparison indicated that the scale of the off-target effects (similarity of shared seed sequences) were significantly large compared to that of the on-target effectsJackson et al. (2003). The further evolution of the Consensus Gene Signature (CGS) protocol was proved to increase the on-target signal through the averaging of independent replicas of shRNAs.
- **MOA Recovery and Prediction:**The CMap was able to restore 63% of known cell-line compound-target relationships when the scores of cell lines were summed up. Using Perturbagen Classes (PCLs) optimized classification of compounds (to obtain 171 high-confidence classes). CMap was also able to foretell new mechanisms of action of unannotated compounds, including BRD-2751 as a strong ROCK1 inhibitor (KD 56 nM) and BRD-1868 as a selective CSNK1A1 inhibitor (KD 2.2 500 μM).
- **Clinical and Genetic Utility:**The CMap was proven to be a functional annotator of genetic variants, where wild-type FBXW7 and those involving a loss-of-function (LoF) allele were differentiated. In addition, clinical trial biopsy analysis revealed that CMap query outcomes might be indicative of target activity in vivo and potentially clinical activity (e.g.positive linkage of connectivity to cell cycle inhibition signatures to longer PHA-793887 treatment)

2 Critical Evaluation

2.1 Strengths and Contributions

The work described represents a significant advancement in functional genomics, achieving what was previously deemed cost-prohibitive

Technological Innovation (L1000): L1000 technique offers the information rich gene expression readout at a minute scale (approximately two dollars) which significantly reduces the entry cost to large scale perturbation experiments. The depth and breadth is made possible by this innovation to form a genome-scale resource that is needed.**Development of a Resource with Comprehensive View:** The development of CMap-L1000v1 (1.3 million profiles) instantaneously created the largest public resource of cellular perturbation data. The resource encompasses the variety of compounds, genetic perturbation, and cell lines.**Evidence of Use in Discovery:** The experiment was able to confirm the essence of CMap, which is discovery facilitation. It showed a high recovery rate of known mechanisms (63% success rate) and, more to the point, it allowed to discover novel small-molecule inhibitors (e.g.ROCK1 and CSNK1A1 inhibitors) purely by computational analysis. **Addressing Complex Data Problems:** The authors tackled the methodological issues that are critical to the high-throughput data-collection. This involves the design of the CGS procedure to reduce potent shRNA off-target effects, and the inference model that is able to efficiently maximize the information obtained using the reduced transcriptome set. **Data Accessibility:** Distribution of all 1.3 million profiles publicly over the CLUE cloud platform (and GEO) would be a way of making the data as useful as it can be to the research community.

2.2 Weakness and Limitations

The authors openly address the drawbacks of the L1000 platform and the compendium that came out of it.

Incomplete Mechanism Recovery: 37 percent of the compounds tested were completely not detected to be linked to their targets even after being highly enriched. The failure is indicative of possible problems, such as the incomplete inhibition of the target, robust off-target effects, or inherent variations between the phenocopy of pharmacological inhibition and genetic loss-of-function (LoF)(Subramanian et al., 2017).

Simplicity of the Inference Model: Inference of the levels of non-measured transcripts is based on Ordinary Least Squares (OLS) regression. As the OLS did a good job (81% accuracy), it is reported that the model can be less effective in the cell types to which the training was made. OLS can provide only a limited measure of the nonlinear, complex relationships that biological systems comprise, particularly in comparison to more sophisticated machine learning tools(Donner et al., 2012).

Noise persistence even after 2 hours: CGS shows significant signal improvement but the original observation that shRNA off-target effects significantly outweigh on-target effects is alarming. The CGS process has been credited with imperfection, and this comingles with the fact that there may still be some spurious associations.

Cellular Context Selectivity: The results of the analysis revealed that 43% of the compounds had cell-type selective gene expression signature. This brings to light the fact that a universal CMap query can be missing context-dependent biology, which implies a wider range of cell types, such as specialized or patient-derived cells, are required in the future(Subramanian et al., 2017; Lamb et al., 2006).

2.3 Conclusion

2.3.1 Overall Conclusion of the Work

The experiment is effective to prove the possibility of generating a massive library of perturbations with functional effects on a large scale with the L1000 platform. Through this effort, the NIH LINCS Consortium developed CMap-L1000v1 - a huge public dataset that included 1.3 million profiles by, dramatically, cutting the costs. It is confirmed that this resource is a potent aid to explore small-molecule mechanism of action, functionally annotate genetic variants, and produce therapeutic hypotheses by relating biological states signatures to perturbagens. L1000 platform is scalable and highly reproducible and has a comparable performance with RNA-seq in perturbation profiling.

2.3.2 Presenting a Novel Idea to Extend/Improve the Method

The current method of L1000 has a strong reliance on Ordinary Least Squares (OLS) regression to deduce the expression of 81% of non-measured genes. Although OLS is an easy and powerful linear model, relationships between measured landmark genes and inferred non-measured genes of a biological process are probably complicated and highly non-linear.

One way of expanding or improving the technique is to substitute the OLS inference framework with a deep learning framework to enhance prediction accuracy, especially in the 17 percent of inferred genes with low performance.

2.3.3 Proposed Extension: Deep Learning inference model

In place of OLS, a deep learning model, including a Deep Fully Connected Neural Network (DFNN), may be adopted. Deep learning algorithms are superior at representation learning as well as offering a universal approximation model that can model complex and nonlinear functions of high-dimensional biological data.

2.3.4 Justification using sources

Dealing with Complexity: Machine learning models are required in dealing with nonlinear effects which are complex. The deep learning algorithms of the study of crop yield prediction as the CNN-RNN architecture showed better results than the old machine learning models (such as the Random Forest and LASSO) of the study since they can catch nonlinear functions. A comparable DFNN model with multiple stacked nonlinear layers might be trained on the CMap data and yields the highly complex mappings of the 978 landmark genes (input features) and the 11350 non-measured genes (target outputs)(Consortium et al., 2015).

Learning feature: The deep learning models do not need hand-crafted features and instead learn the most appropriate representations directly out of the data. This capacity may improve the accuracy of inferred gene expression over the 81% of OLS, and, therefore, the noise in the final signatures will be reduced, and some of the 37% of viable compound-target interactions that are currently not seen by the CMap may be restored.

Strength in Contexts: The CMap signatures might be stronger inference mechanisms that would enable application to cell types not dissimilar to those used in training the model, which is a weakness of the existing OLS methodology. This would involve training the DFNN with available RNA-seq data (such as the data set GTEx consortium used earlier) with the landmark measurements being the inputs and the entire transcriptome as the outputs.

References

GTEx Consortium, Kristin G Ardlie, David S Deluca, Ayellet V Segrè, Timothy J Sullivan, Taylor R Young, Ellen T Gelfand, Casandra A Trowbridge, Julian B Maller, Taru Tukiainen, et al. The genotype-tissue expression (gtex) pilot analysis: multitissue gene regulation in humans. *Science*, 348(6235):648–660, 2015.

Yoni Donner, Ting Feng, Christophe Benoist, and Daphne Koller. Imputing gene expression from selectively reduced probe sets. *Nature methods*, 9(11):1120–1125, 2012.

Timothy R Hughes, Matthew J Marton, Allan R Jones, Christopher J Roberts, Roland Stoughton, Christopher D Armour, Holly A Bennett, Ernest Coffey, Hongyue Dai, Yudong D He, et al. Functional discovery via a compendium of expression profiles. *Cell*, 102(1):109–126, 2000.

Aimee L Jackson, Steven R Bartz, Janell Schelter, Sumire V Kobayashi, Julja Burchard, Mao Mao, Bin Li, Guy Cavet, and Peter S Linsley. Expression profiling reveals off-target gene regulation by rnai. *Nature biotechnology*, 21(6):635–637, 2003.

Justin Lamb, Emily D Crawford, David Peck, Joshua W Modell, Irene C Blat, Matthew J Wrobel, Jim Lerner, Jean-Philippe Brunet, Aravind Subramanian, Kenneth N Ross, et al. The connectivity map: using gene-expression signatures to connect small molecules, genes, and disease. *science*, 313(5795):1929–1935, 2006.

David Peck, Emily D Crawford, Kenneth N Ross, Kimberly Stegmaier, Todd R Golub, and Justin Lamb. A method for high-throughput gene expression signature analysis. *Genome biology*, 7(7):R61, 2006.

Aravind Subramanian, Rajiv Narayan, Steven M Corsello, David D Peck, Ted E Natoli, Xiaodong Lu, Joshua Gould, John F Davis, Andrew A Tubelli, Jacob K Asiedu, et al. A next generation connectivity map: L1000 platform and the first 1,000,000 profiles. *Cell*, 171(6):1437–1452, 2017.