Generative Bayesian Networks for Augmentation of Molecular Data from Commercial Genetic Panels

ZEPHYR AI

Dillon Tracy, Jeff Sherman, Maayan Baron

Zephyr AI: 1800 Tysons Blvd Suite 901, McLean VA 22102 | www.zephyrai.bio

Abstract #7373

SUMMARY

- We introduce a generative Bayesian network method for **synthesizing annotated patient feature profiles using a constrained set of genes from limited real-world molecular data**, looking specifically at somatic mutations and lung and breast cancer.
- This approach addresses challenges posed by widely clinically available, yet molecularly sparse tumor data, **enhancing the value of established real-world clinicogenomic datasets** and potentially advancing precision oncology through personalized treatment guidance, enriched data analysis and novel biomarker identification.

BACKGROUND

- Molecular data from patient tumors in real-world settings are sparse, typically limited to profiles of a few hundred genes.
- This issue of molecular sparsity is exacerbated by earlier assays, resulting in real-world clinicogenomic databases that are very rich in longitudinal clinical follow-up, but restricted in their applicability to research pursuits such as biomarker discovery.
- The number of genes on commercial NGS panels continues to increase over time, reflecting the discovery of more biomarkers in cancer research and the translation of these discoveries into clinical practice.
- We hypothesized that by modeling the joint distribution of both observed and unobserved molecular features in a large tumor cohort using a Bayesian network and Gibbs sampling, we could effectively infer and synthesize comprehensive mutational profiles for tumors with otherwise limited data from commercial NGS panels (**Fig. 1**).

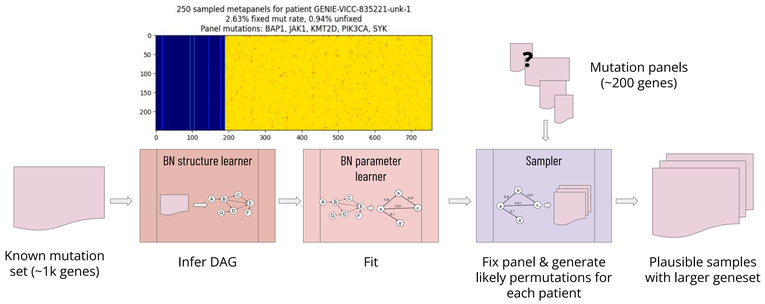

Figure 1. Generative Bayesian network approach to augment tumor mutational profiles. A directed graph is inferred on a corpus of molecularly rich training data (e.g., TCGA mutational profiles). State probabilities for the graph are learned, and Gibbs sampling is used to extract multiple likely graph states for each patient of interest. The output is a collection of profiles whose "out-of-panel" components are drawn from the joint probability distribution of all mutations in the training population. Upper left: 250 757-gene profiles generated from a 190-gene panel with 5 mutations. The blue area (the panel result) is fixed and the yellow area is generated or synthesized; red dots indicate mutations.

Modeling Mutational Profiles in TCGA Cohorts with Bayesian Networks

For any particular cohort (for instance, TCGA LUSC patients), there is a known distribution of observed features (i.e. NGS panel results) and an unknown joint distribution of unobserved features (mutation probabilities over all genes). We model this joint distribution by learning a Bayesian network over a broad feature set for which some training data is available, and generate feature profiles by applying Gibbs sampling to the learned network. In effect, a fitted Bayesian network allows sampling random variates from the joint distribution of all mutational profiles (**Fig. 2**).

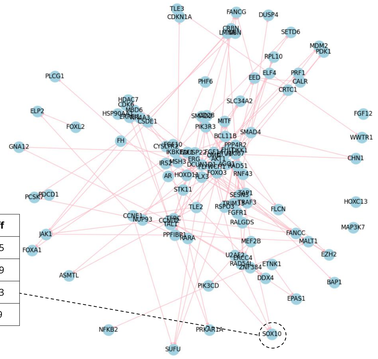

Figure 2. A directed graph inferred from TCGA LUSC mutations was generated using findr [1]. Only highly connected nodes (above degree 15) shown. Nodes are limited to 12 outgoing and 4 incoming links. The complete graph has 689 genes and 2644 edges. The fitted network model tabulates or predicts (depending on size) the mutation probability for each node, conditioned on its parents' states. The model's conditional probability table for the SOX10 vertex is shown.

parent	child	coeff
FLI1	SOX10	0.645
IRF2	SOX10	0.779
NFE2	SOX10	0.473
PTPRO	SOX10	0.779

Validation of Model-Generated Mutation Profiles in Lung Cancer

By way of validation, we examined the marginal mutation rates of profiles generated by the model to see how well they recapitulate rates seen in the training data. Results show close agreement across a decade of range in the mutation rate (**Fig. 3A**). We further compared mutation coincidence rates for nine lung cancer-pertinent genes with rich coincidence data [9], again showing agreement between training ground truth and the generated profiles, with the possible exception of EGFR-TP53 in LUAD (**Fig. 3A**).

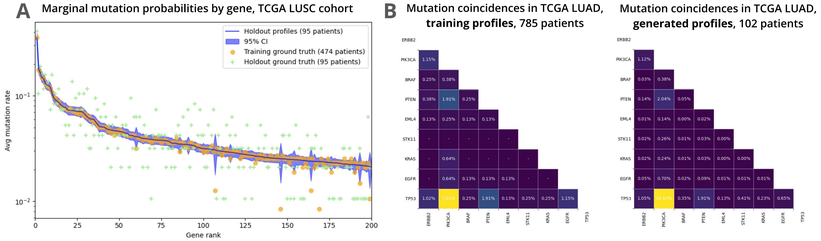

Figure 3. Analysis of Training and Predicted Mutation Probabilities and Coincidences in TCGA Lung Cancer Cohorts. A) Marginal mutation probabilities by gene in the TCGA LUSC cohort, for the 200 most frequently mutated genes in the generated profiles. Training rates, shown in orange, are averages over patients in the training split (80%). Holdout rates, shown in blue, are averages of averages of profiles generated over patients (5k per patient) in the holdout split (20%). Confidence limits are Wilson scores on the per-patient averages. Holdout patient ground truth is shown in green. **B)** Mutation coincidences in the TCGA LUAD cohort, for 9 genes of interest in NSCLC.

Characterizing Drug Response Using Generated Patient Mutational Data

One application of this type of generative model is in downstream modeling and biomarker discovery. Using an internal drug response prediction model we found variations in augmented profiles typically induced small perturbations to modeled drug response (**Fig. 4**). Interestingly, when outputs were discordant between limited and expanded actual gene panel inputs, synthetic data were more concordant with results from expanded panels (**Fig. 5**). Moreover, synthetic data enables exploration of numerous genomic possibilities simultaneously, supporting identification of off-panel genes that may significantly influence sensitivity, thus aiding biomarker discovery.

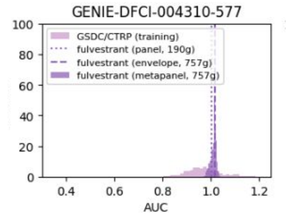

Figure 4. Consistency in drug sensitivity predictions using real and synthesized mutational profiles for fulvestrant in a BRCA patient. Fulvestrant drug response prediction scores were generated for a single BRCA patient using 5000 profiles (metapanel, 757g) generated from an actual 190-gene set. Predictive performance for this drug response model was assessed using actual data as input from a 190-gene (panel, 190g) and 757-gene panel (envelope, 757g). The outcome demonstrates high consistency between real and synthesized profiles (higher AUC = increased resistance).

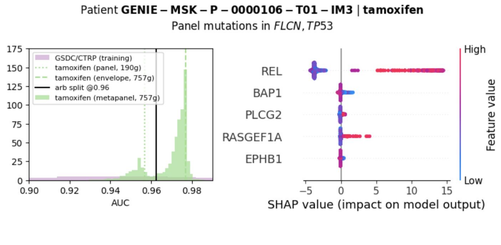

Figure 5. Tamoxifen response predictions between synthetic and expanded gene panels are concordant. Tamoxifen response scores were obtained for an additional BRCA patient using a synthetic 757-gene (metapanel, 757g) generated from an actual 190 gene mutation panel result for this patient. Predictions were discordant between the panel (panel, 190g; predicted sensitive) and the generated profile (envelope, 757g; predicted resistant), highlighting the impact in predictive power of larger gene panels. Remarkably, the insensitive peak in the predicted response distribution (metapanel, 757g) closely matched the actual 757g panel, demonstrating the robustness of our method. Solid line demarcates the sensitive/insensitive boundary for the binary classifier whose feature importance appears in the SHAP analysis (right). Positive and red SHAP values indicate REL mutations are linked to tamoxifen sensitivity.

CONCLUSION

- The Bayesian network method for synthesizing patient genetic profiles tackles the issue of limited molecular data in real-world clinical settings, significantly enhancing real-world clinicogenomic datasets that typically lack molecular detail but have extensive clinical follow-up.
- This augmented data opens avenues for downstream analysis, supports the discovery of potential biomarkers for tumor classification, prognosis, or therapy response, and may improve diagnostic tool accuracy by including a wider array of genetic information.
- Additionally, these enhanced datasets hold the potential to facilitate development of machine learning models, utilizing vast amounts of real-world data to address diverse questions that support the advancement of precision medicine.

REFERENCES

1. Wang L, Audenaert P, Michoel T. High-Dimensional Bayesian Network Inference From Systems Genetics Data Using Genetic Node Ordering. Front Genet. 2019 Dec 20;10:1196. doi: 10.3389/fgene.2019.01196. PMID: 31921278; PMCID: PMC6933017.
2. Lee J, Choi MK, Song IS. Recent Advances in Doxorubicin Formulation to Enhance Pharmacokinetics and Tumor Targeting. Pharmaceuticals (Basel). 2023 May 29;16(6):802. doi: 10.3390/ph16060802. PMID: 37375753; PMCID: PMC10301446.
3. Chiba N, Ozawa Y, Hikita K, Okihara M, Sano T, Tomita K, Takano K, Kawachi S. Increased expression of HOXB9 in hepatocellular carcinoma predicts poor overall survival but a beneficial response to sorafenib. Oncol Rep. 2017 Apr;37(4):2270-2276. doi: 10.3892/or.2017.5474. Epub 2017 Feb 23. PMID: 28260092.
4. Malash, I., Mansour, O., Gaafar, R. *et al. Her2/EGFR-PDGFR* pathway aberrations associated with tamoxifen response in metastatic breast cancer patients. *J Egypt Natl Canc Inst* 34, 31 (2022). https://doi.org/10.1186/s43046-022-00132-5
5. Chouhan S, Singh S, Ashvale D, Ramteke P, Vanuopadath M, Nair BG, Nair SS, Bhat MK. Sensitization of hepatocellular carcinoma cells towards doxorubicin and sorafenib is facilitated by glucose dependent alterations in reactive oxygen species, P-glycoprotein and DKK4. J Biosci. 2020;45:97. PMID: 32713860.
6. Williams MM, Cook RS. Bcl-2 family proteins in breast development and cancer: could Mcl-1 targeting overcome therapeutic resistance? Oncotarget. 2015 Feb 28;6(6):3519-30. doi: 10.18632/oncotarget.2792. PMID: 25784482; PMCID: PMC4414133.
7. Bertucci A, Bertucci F, Gonçalves A. Phosphoinositide 3-Kinase (PI3K) Inhibitors and Breast Cancer: An Overview of Current Achievements. Cancers (Basel). 2023 Feb 23;15(5):1416. doi: 10.3390/cancers15051416. PMID: 36900211; PMCID: PMC10001361.
8. Burkitt, M., Williams, J., Townsend, T. *et al.* Mice lacking NF-κB1 exhibit marked DNA damage responses and more severe gastric pathology in response to intraperitoneal tamoxifen administration. *Cell Death Dis* 8, e2939 (2017). https://doi.org/10.1038/cddis.2017.332
9. Dearden S, Stevens J, Wu YL, Blowers D. Mutation incidence and coincidence in non small-cell lung cancer: meta-analyses by ethnicity and histology (mutMap). Ann Oncol. 2013 Sep;24(9):2371-6. doi: 10.1093/annonc/mdt205. Epub 2013 May 30. PMID: 23723294; PMCID: PMC3755331.
10. Zhang J, Zhou ZZ, Chen K, Kim S, Cho IS, Varadkar T, Baker H, Cho JH, Zhou L, Liu XM. A CD276-Targeted Antibody-Drug Conjugate to Treat Non-Small Lung Cancer (NSCLC). Cells. 2023 Sep 30;12(19):2393. doi: 10.3390/cells12192393. PMID: 37830607; PMCID: PMC10572050.