Predicting Cancer Phenotypes from Somatic Genomic Alterations via Genomic Impact Transformer

Yifeng Tao¹, Chunhui Cai², William W. Cohen¹,*, Xinghua Lu²,3,*

¹School of Computer Science, Carnegie Mellon University
²Department of Biomedical Informatics, University of Pittsburgh
³Department of Pharmaceutical Sciences, School of Pharmacy, University of Pittsburgh
Tumor origin and progression

- Cancers are mainly caused by somatic genomic alterations (SGAs)
  - Driver SGAs (~10s/tumor): Promote tumor progression
  - Passenger SGAs (~100s/tumor): Neutral mutations
- How to distinguish drivers from passengers?

S Nik-Zainal et al. 2017
Cancer drivers

• How to distinguish drivers from passengers?
  • Frequency: recurrent mutations more likely to be drivers

• Conserved domain: protein function significantly disturbed

• All unsupervised. But drivers are defined as mutations that promote to tumor development…
Cancer drivers

- Identify driver SGAs with supervision of downstream phenotypes
  - Change of RNA expression
  - Differentially expressed genes (DEGs)
- Candidate models
  - Bayesian model (C Cai et al. 2019)
  - Lasso/Elastic net (R Tibshirani 1994)
  - Multi-layer perceptrons (MLPs) (F Rosenblatt 1958)
- Models do prediction & driver detection?

Model (?) that predicts DEGs accurately & identifies driver SGAs
Self-attention mechanism

• Models do prediction & driver detection?
• Attention mechanism
  • Initially in CV (K Xu et al. 2015)/NLP (A Vaswani et al. 2017)
  • Better interpretability
  • Improves performance
• Self-attention mechanism (Z Yang et al. 2016)
  • Contextual deep learning framework: weights determined by all the input mutations

Model with self-attention that predicts DEGs accurately & identifies driver SGAs

\[ \alpha_0 = 1, \alpha_1, \alpha_2, \alpha_3, \alpha_4, \alpha_5, \ldots, \alpha_m \]

Cancer type

Somatic genomic alterations (SGAs)

Cancer patient: TCGA-D8-A1JJ

Over-expressed genes

Under-expressed genes
Genomic impact transformer (GIT)

- Transformer: encoder-decoder architecture
- Encoder: self-attention mechanism; Decoder: MLP
Encoder: Multi-head self-attention

- Tumor embedding is the weighted sum of gene embeddings:
  \[ e_t = 1 \cdot e_s + \sum_g \alpha_g \cdot e_g \]

- Weights determined by input gene embeddings:
  \[ \alpha_1, \alpha_2, \ldots, \alpha_m = \text{Function}_{\text{Attention}}(e_1, e_2, \ldots, e_m) \]

\[ \alpha_g = \sum_{j=1}^{h} \alpha_{g,j} = \alpha_{g,1} + \alpha_{g,2} + \ldots + \alpha_{g,h}, \ g = 1, 2, \ldots, m \]

\[ \alpha_{1,j}, \alpha_{2,j}, \ldots, \alpha_{m,j} = \text{softmax}(\beta_{1,j}, \beta_{2,j}, \ldots, \beta_{m,j}) \]

\[ \beta_{g,j} = \theta_j^T \cdot \tanh(W_0 \cdot e_g), \ g = 1, 2, \ldots, m \]
Pre-training gene embedding: Gene2Vec

- Co-occurrence pattern (e.g., mutually exclusive alterations)

\[
\Pr(c \in \text{Context}(g) \mid g) = \frac{\exp(e^T_g v_c)}{\sum_{c' \in g} \exp(e^T_g v_{c'})}
\]

Pathway 1

Pathway 2

Pathway 3

MD Leiserson et al. 2015
T Mikolov et al. 2013
Improved performance in predicting DEGs

• Predicting DEGs from SGAs
  • Conventional models
  • Ablation studies
Candidate drivers via attention mechanism
Gene embedding space

- Functionally similar genes are close in gene embedding space
  - Qualitatively and quantitatively (i.e., GO enrichment, NN accuracy)
Tumor embedding: Survival analysis

• Tumor embeddings reveal distinct survival profiles
Tumor embedding: Drug response

- Tumor embeddings are predictive of drug response
Conclusions and future work

• Biologically inspired neural network framework
  • Identifying cancer drivers with supervision of DEGs
  • Accurate prediction of DEGs from mutations

• Side products
  • Gene embedding: informative of gene functions
  • Tumor embedding: transferable to other phenotype prediction tasks

• Code and pretrained gene embedding:
  https://github.com/yifengtao/genome-transformer

• Future work
  • Fine-grained embedding representation in codon level
  • Tumor evolutionary features, e.g., hypermutability, intra-tumor heterogeneity
Acknowledgments

• Dr. Xinghua Lu
• Dr. William W. Cohen
• Dr. Chunhui Cai
• Michael Q. Ding
• Yifan Xue
Quantitative measurement of gene embeddings

- Functional similar genes → closer in embedding space
  - Go enrichment:
    \[
    \text{enrichment} = \frac{\mathbb{E}_{\text{Clust}(e_g) = \text{Clust}(e_c)} [\mathbb{1}(\text{GO}(g) \cap \text{GO}(c) \neq \emptyset)]}{\mathbb{E}_{g,c \in G} [\mathbb{1}(\text{GO}(g) \cap \text{GO}(c) \neq \emptyset)]}
    \]
  - NN accuracy:
    \[
    \text{NN accuracy} = \mathbb{E}_{e_c \in \text{NN}(e_g)} [\mathbb{1} (\text{GO}(g) \cap \text{GO}(c) \neq \emptyset)]
    \]
Tumor embedding space
Gene2Vec algorithm

Data: Genomic alterations in each tumor: $T = \{ T_i = \{ g_{i1}, g_{i2}, \ldots, g_{im(i)} \} \}_{i=1,2,\ldots,N}$.

Result: Pretrained gene embedding of each gene:
$E = \{ e_g \in \mathbb{R}^n \}_{g \in G}$.

Context gene embeddings:
$V = \{ v_g \in \mathbb{R}^n \}_{g \in G}$.

\begin{align*}
    f(g) & \leftarrow \frac{1}{N} \sum_{i=1}^{N} \mathbb{1}(g \in T_i), \quad g \in G; \quad \text{\quad // Gene frequency} \\
    f_n(g) & \leftarrow \frac{1}{Z_n} f(g)^{3/4}, \quad g \in G; \quad \text{\quad // Normalized frequency} \\
    e_g & \sim U \left( -\frac{0.5}{n}, \frac{0.5}{n} \right)^n, \quad v_g \leftarrow 0^n, \quad g \in G; \quad \text{\quad // Initialize gene embeddings and context embeddings} \\
\end{align*}

while not converges do
\begin{align*}
    l & \leftarrow 0; \quad \text{\quad // Total loss of a mini-batch samples} \\
    \text{for } b = 1, 2, \ldots, \text{batch\_size} \text{ do} \\
    \quad g & \sim f; \quad \text{\quad // Sample a gene} \\
    \quad g_c & \sim \text{Context}(g; T); \quad \text{\quad // Sample a context gene} \\
    \quad g_{nr} & \sim f_n, \quad r = 1, 2, \ldots, R; \quad \text{\quad // Sample negative context genes} \\
    \quad l & \leftarrow l + \text{NSLoss} \left( g, g_c, \{ g_{nr} \}_{r=1}^R ; E, V \right); \quad \text{\quad // Update} \\
    \end{align*}
\begin{align*}
    (E, V) & \leftarrow (E, V) - \eta \cdot \frac{\partial l}{\partial (E, V)}; \quad \text{\quad // Gradient descent} \\
\end{align*}
end

Function Context ($g; T$)
\begin{align*}
    P_c & \leftarrow U \left( \{ g_c \mid g_c \in T_i, g \in T_i \}_{i=1,2,\ldots,N} \right); \quad \text{\quad // Uniform distribution on sequence of adjacent mutations} \\
    \text{return } P_c
\end{align*}

Function NSLoss ($g, g_c, \{ g_{nr} \}_{r=1}^R; E, V$)
\begin{align*}
    l & \leftarrow \log \sigma (e_g^T v_{g_c}) + \sum_{r=1}^R \log \sigma (-e_g^T v_{g_{nr}}); \quad \text{\quad // Negative sampling loss of one sample} \\
    \text{return } l
\end{align*}
Gene2Vec: Co-occurrence patterns

• Co-occurrence does not necessarily mean similar embeddings
  • Ex 1: two cats sit there.
  • Ex 2: two cats stand there.
  • Ex 3: two dogs sit there.

Pathway 1: number
  one, two, several

Pathway 2: noun
  cat, dog

Pathway 3: verb
  sit, stand, lie

MD Leiserson et al. 2015
T Mikolov et al. 2013