Explainable Substructure Partition Fingerprint for Protein, Drug, and More

Kexin Huang
Health Data Science Program
Harvard University
Boston, MA 02115

Cao Xiao
Analytics Center of Excellence
IQVIA
Cambridge, MA 02139

Lucas Glass
Analytics Center of Excellence
IQVIA
Cambridge, MA 02139

Jimeng Sun
CSE
Georgia Institute of Technology
Atlanta, GA 30313

Interpretable fingerprint that links sub-structures' relevance to the molecular properties is highly desirable to guide decision making in drug discovery. For example, it is useful to understand which functional groups of the drugs lead to specific property. To our best knowledge, existing fingerprinting methods, as categorized below, all lack this level of explainability.

- Molecule-level fingerprints such as composition-transition-distribution (Govindan and Nair (2011)) for the protein generate numerical values to measure the molecule-level features. However, these fingerprints only provide molecule-level information.
- Hashing-based fingerprints such as Morgan fingerprint (Rogers and Hahn (2010)) and Daylight-type fingerprint (James (2004)) for drugs hash circular or path sub-structures to numerical values. However, it is difficult to trace back to the sub-structures due to the hashing nature. For latent fingerprints such as mol2vec (Jaeger *et al.* (2018)), it is unclear how the dense numerical values map back to sub-structures.
- Expert-curated fingerprints such as PubChem (Bolton *et al.* (2008)), MACCS (Cereto-Massagué *et al.* (2015)) for drugs and amino acid composition (Cao *et al.* (2013)) for protein represent each sub-structure with a bit of a bit vector. However, the output vector usually assigns numerous granular sub-structures (~100 for PubChem) to even a small molecule, where many sub-structures are a subset of other ones, making it intractable to know which specific sub-structures lead to the outcome.

To fill the gap, we postulate that ideally, an interpretable fingerprint should be able to cleverly partition the input to discrete pieces of moderate-sized sub-structures and provide a tractable path to trace the partitioned sub-structural signals to the prediction outcome.

To tackle the above challenges, we propose an Explainable Substructure Partition Fingerprint (ESPF) that decomposes drugs and proteins into a discrete set of moderate-sized sub-structures that are customized to the data at hand and have strong predictive values. ESPF is inspired by the subword units (Sennrich *et al.* (2015)) in the natural language processing domain and is based on the Byte Pair Encoding algorithm (Gage (1994)). The input of ESPF is a database of sequences of entities (e.g. SMILES for drug, the amino acid sequence for protein). Given this database, it discovers frequent recurring subsequences (subword units) in a database and replaces the original sequence (word) with the most plausible combination of subsequences. The output is the subsequences vocabulary set and their frequencies. With these two pieces of information, it can decompose any new unseen sequence to a sequence of frequent subsequences. This sequence can then be turned to a bit vector where each bit corresponds to one item in the discovered subsequences set (See Algo. 1).

Empirically, we find ESPF outputs suitable sized sub-structure ordered partition. It successfully identifies important functional groups for drugs and motifs for proteins. The suitable sized non-

Learning Meaningful Representation of Life Workshop at NeurIPS 2019

Algorithm 1: ESPF

Input: $\mathbb{V}$ as the set of all initial amino acids/SMILES tokens; $\mathbb{W}$ as the set of tokenized proteins/drugs; θ as the specified frequency threshold; ℓ as the maximum size of $\mathbb{V}$.

for $t = 1 \dots \ell$ **do**

  (A, B), FREQ $\leftarrow$ scan $\mathbb{W}$

  // (A, B) is the frequentest consecutive tokens.

  if FREQ $< \theta$ **then**

    break // (A, B) 's frequency lower than threshold

  $\mathbb{W} \leftarrow \mathrm{find}(\mathrm{A}, \mathrm{B}) \in \mathbb{W}, \mathrm{replace\ with}\ (\mathrm{AB})$

  // update $\mathbb{W}$ with the combined token (AB)

  $\mathbb{V} \leftarrow \mathbb{V} \cup (\mathrm{AB})$ // add (AB) to the token vocabulary set $\mathbb{V}$

Output: $\mathbb{W}$, the updated tokenized proteins/drugs; $\mathbb{V}$, the updated token vocabulary set.

overlapping outputs provide a tractable path to see which sub-structures contribute to machine learning predictive outcomes. For example, it identifies contributing factors in the interaction between Sildenafil and Isosorbide Mononitrate. In addition, it has competitive predictive performance against state-of-the-art fingerprinting methods. e.g., a 2% increase in PR-AUC over ECFP on the drug-drug interaction prediction task. The ESPF method can also be used for any other sequential biomedical entities such as DNA sequences.

References

Bolton, E. E., Wang, Y., Thiessen, P. A., and Bryant, S. H. (2008). Pubchem: integrated platform of small molecules and biological activities. In *Annual reports in computational chemistry*, volume 4, pages 217–241. Elsevier.

Cao, D.-S., Xu, Q.-S., and Liang, Y.-Z. (2013). propy: a tool to generate various modes of chou's pseaac. *Bioinformatics*, **29**(7), 960–962.

Cereto-Massagué, A., Ojeda, M. J., Valls, C., Mulero, M., Garcia-Vallvé, S., and Pujadas, G. (2015). Molecular fingerprint similarity search in virtual screening. *Methods*, **71**, 58–63.

Gage, P. (1994). A new algorithm for data compression. *The C Users Journal*, **12**(2), 23–38.

Govindan, G. and Nair, A. S. (2011). Composition, transition and distribution (ctd)—a dynamic feature for predictions based on hierarchical structure of cellular sorting. In *2011 Annual IEEE India Conference*, pages 1–6. IEEE.

Jaeger, S., Fulle, S., and Turk, S. (2018). Mol2vec: unsupervised machine learning approach with chemical intuition. *Journal of chemical information and modeling*, **58**(1), 27–35.

James, C. A. (2004). Daylight theory manual.

Rogers, D. and Hahn, M. (2010). Extended-connectivity fingerprints. *Journal of chemical information and modeling*, **50**(5), 742–754.

Sennrich, R., Haddow, B., and Birch, A. (2015). Neural machine translation of rare words with subword units. *ACL 2015*.