dna2vec

Dna2vec is an open-source library to train distributed representations of variable-length k-mers.

For more information, please refer to the paper: dna2vec: Consistent vector representations of variable-length k-mers

Installation

Note that this implementation has only been tested on Python 3.5.3, but we welcome any contributions or bug reporting to make it more accessible.

1. Clone the dna2vec repository: `git clone https://github.com/pnpnpn/dna2vec`
2. Install Python dependencies: `pip3 install -r requirements.txt`
3. Test the installation: `python3 ./scripts/train_dna2vec.py -c configs/small_example.yml`

Training dna2vec embeddings

Note: I had file chr21.fa after downloading data since I was training model on my computer.

1. Download hg38 from http://hgdownload.cse.ucsc.edu/goldenPath/hg38/bigZips/hg38.chromFa.tar.gz. This will take a while as it's 938MB.
2. Untar with `tar -zxvf hg38.chromFa.tar.gz`. You should see FASTA files for chromosome 1 to 22: chr1.fa, chr2.fa, ..., chr22.fa.
3. Move the 22 FASTA files to folder inputs/hg38/
4. Start the training with: `python3 ./scripts/train_dna2vec.py -c configs/hg38-20161219-0153.yml`
5. Wait for a couple of days ...
6. Once the training is done, there should be a `dna2vec-<ID>.w2v` and a corresponding `dna2vec-<ID>.txt` file in your results/directory.

Reading pretrained dna2vec

You can read pretrained dna2vec vectors pretrained/dna2vec-*.w2v using the class MultiKModel in dna2vec/multi_k_model.py. For example:

Note: Please open Python terminal on Command Line before running below command

Note:Updated version of MultiKModel compatible with latest version of Gensim is being added to folder.

```
from dna2vec.multi_k_model import MultiKModel
filepath = 'pretrained/dna2vec-20161219-0153-k3to8-100d-10c-29320Mbp-sliding-Xat.w2v'
mk_model = MultiKModel(filepath)
```

You can fetch the vector representation of AAA with:

```
>>> mk_model.vector('AAA')
array([ 0.023137  ,  0.156295 , ...
```

Compute the cosine distance between two k-mers via dna2vec:

```
>>> mk_model.cosine_distance('AAA', 'GCT')
0.14546435594464155
>>> mk_model.cosine_distance('AAA', 'AAAA')
0.89000147450211231
```

FAQ

Does the pre-trained dna2vec data (w2vfile) cover all k-mers?

The pre-trained data should cover all k-mers for $3 \leq k \leq 8$

```
>>> [len(mk_model.model(k)) for k in range(3,9)]
[64, 256, 1024, 4096, 16384, 65536]
```

```
>>> [4**k for k in range(3,9)]
[64, 256, 1024, 4096, 16384, 65536]
```