sequence data from database

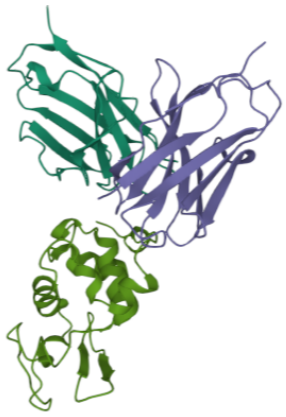
Date Created	@April 16, 2022 3:46 PM
Status	Doing
date	
category	Inno

Affinity bench v2

There are 136 set of data in database. **(chain ID provided)**

- **94** data are choose (two chains)
- Others are with multiple chains

  One of example is 1BVK_DE:F, with chain D and E are Fv Hulys11 and Chain F is HEW lysozyme. see https://www.rcsb.org/3d-view/1BVK.



Complex PDB	Functional class	Protein A	Protein B	Expe
1BVK_DE:F	Antigen-Antibody	Fv Hulys11	HEW lysozyme	10.5

data stored in

1. ./affinitybench/affinitybench.seq.txt
2. ./affinitybench/affinitybench.dg.txt

PDBbind-CN

There are 2850 complex with affinity binding (Kd, Ki or IC50)

2615 protein complex with affinity binding Kd. (remove uncertain value with '>' or '<', remove express with Ki, and IC50)

- The database does not provide temp, but mention measured in room temperature in literature. delta G is calculated using room temperature 298K

3ohm protein is replaced by 7sq2 in PDB database. https://www.rcsb.org/structure/removed/3ohm

4fqr is too large, do not have pdb file (only mmCIF) https://www.rcsb.org/structure/4FQR

The Chain ID are not provided - hard to know which chain represents the protein name on excel

select protein complex with only two chains

Using biopython select protein complex with only two chains.

Only **635** protein complex are left.

data stored in

1. ./pdbbind/pdbbind.seq.txt
2. ./pdbbind/pdbbind.dg.txt

Using Uniprot labels

select data with two uniprot labels, **1557** protein complex are selected

The uniprot sequencing data from label are obtained from:

1. unprot.csv
2. extract from website

```
def seq_uniprot_lib (label):
    cID=label

    baseUrl="http://www.uniprot.org/uniprot/"
    currentUrl=baseUrl+cID+".fasta"
    response = r.post(currentUrl)
    cData=''.join(response.text)

    Seq=StringIO(cData)
    pSeq=list(SeqIO.parse(Seq,'fasta'))
    result =str(pSeq[0].seq)
    return result

seq_uniprot_lib('Q07011')
```

data stored in;

1. ./pdbbind/pdbbind_uniprot.seq.txt
2. ./pdbbind/pdbbind_uniprot.dg.txt

SKEMPI v2

Original method is extract sequencing data from PDB model → time consuming, file is large

wild-type

225 unique wild-type protein complex with two chains

method:

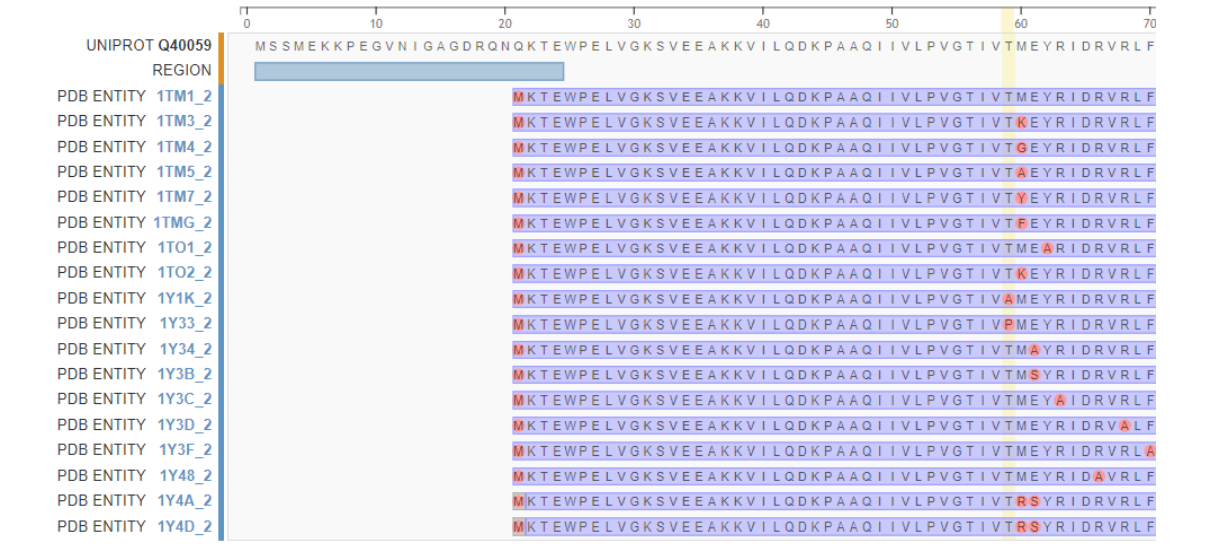
1. using corresponding uniprot label
2. using pdb files

data stored in :

1. ./skempi/wild.seq.txt
2. ./skempi/wild.dg.txt

mutated type

4165 unique mutated-type protein complex with two chains



Then find the uniprot sequencing is not always corresponding to pdb sequencing

Thus, have to use PDB file, Uniprot is not okay.

new data stored in : (all sequencing from pdb bank)

1. ./skempi/wild2.seq.txt

mutation data stored in :

1. ./skempi/mut.seq.txt
2. ./skempi/mut.dg.txt

Due to the version reason, adjust the position number of 1S1Q and 4PCA

- 1S1Q works well
- two mutations in 4PCA is not corresponding
- all the mutations with 1KBH is wrong
  - sequence length is much smaller than mutation position on csv file

overall

database	number	note	files
Affinity bench v2	94		1. ./affinitybench/affinitybench.seq.txt 2. ./affinitybench/affinitybench.dg.txt
PDBbind-CN	635 (from pdb) +1557 (from uniprot)	could be overlap using two methods	**from pdb** 1. ./pdbbind/pdbbind.seq.txt 2. ./pdbbind/pdbbind.dg.txt **from uniprot** 1. ./pdbbind/pdbbind_uniprot.seq.txt 2. ./pdbbind/pdbbind_uniprot.dg.txt
SKEMPI v2	225 wildtype + 4165 mutated type	1KBH sequencing data are lost, part of 4PCA are lost	**wild-type** 1. ./skempi/wild.dg.txt 2. ./skempi/wild2.seq.txt **mutated type** 1. ./skempi/mut.seq.txt 2. ./skempi/mut.dg.txt