0.1 Abstract

0.2 Introduction

0.3 Methods

0.3.1 Processing experimental maps TBD

Validation or Unlabeled see 0.3.2

Training (Labeled Sets)

for each map:

- create unlabeled dataset 0.3.2
- calculate labels
- TBD filter:
  - run Local Correlation Score and delete boxes with box center LCS < 0.8
  - For boxes labeled 0 (NONE) randmoly select an amount similar to amount of boxes labeled > 0
- TBD augment (only positive
  - only proteins (not viruses, because of symmetry)
  - Select N rotation triples
  - rotate initial map (but same grid) and pdb structure
  - run the above procedure on the rotated map

0.3.2 Dataset for Detection Problem

Create Boxes TBD

input:

1. mrc map file

oupput:

1. list of boxes (3D matrices)
2. list position coordinates for each box

procedure :

1. done Create a new grid with a required apix and interpolate, use chimera vop command
   - regions identically zero cropped out of the map
2. Done extract boxes (copy, don't interpolate)
   - with normalization
   - filter boxes with average less then average of the map
3. TBD for each box calc position of the box center

0.3.3 Databases

Simulation:Train,Test Sets

Table of simulated MRC files folders in /data/rotamersdata/mrcs

MRCs_res30apix-1	Resolution 3 Å, variable apix
MRCs_23_to_25	Variable 2.3 Å-2.5 Å, variable apix
MRCs_25_to_27	Variable 2.5 Å-2.7 Å, variable apix
MRCs_27_to_29	Variable 2.7 Å-2.9 Å, variable apix
MRCs_29_to_31	Variable 2.9 Å-3.1 Å, variable apix

1. **folder = rotamersdata/DBres30apix-1, script = create_db_res3_apix_var.py**. Simulated from Rotamers, resolution = 3 Å, sampled at 1 Å, box size = $11 \times 11 \times 11$. Box center at geometric cg of AA. No normalization
2. **folder = rotamersdata/DBres30apix-1norm01, script = create_db_res3_apix_var_norm01.py**. Simulated from Rotamers, resolution = 3 Å, sampled at 1 Å, box size = $11 \times 11 \times 11$. Box center at geometric cg of AA. **Normalization** $mean = 0, \sigma = 1$ **for each box**
3. **folders = rotamersdata/DBresXXYYnorm01**, $XX, YY \in \{(23, 25), (25, 27), (27, 29), (29, 31), (31, 33), (33, 35)\}$, **script = create_db_resXXYYnorm01.py**. Simulated from Rotamers, resolution = XXÅto YYÅ, sampled at 1 Å, box size = $11 \times 11 \times 11$. Box center at geometric cg of AA. **Normalization** $mean = 0, \sigma = 1$ **for each box**

Experimental Maps

Resolution high then 2.3 Å maps:

res Å	EMDB	pdb file	comments
1.8	emd-8194.map	5k12.pdb	glutamate dehydrogenase, needs augmentation
1.9	emd-7770.map	6cvm.pdb	beta-galactosidase, NUT USED, needs augmentation
2.2	emd-2984.map	5a1a.pdb	beta-galactosidase, VALIDATION, needs augmentation
2.3	emd-3295.map	5ftj.pdb	p97 very noisy, looks like outer loops no in the map, good for validation and analysis, needs filtering for training. There is a 2.4 map of p7 3296, so this map is moved to this category
2.26	emd-8762.pdb	5w3m	rhinovirus, PROBLEMATIC 960*960*1
2.17	7599	6csg	not yet submitted

Datasets:

1. **folder =data/cryoEM/DB0023classnoaug**, **script db0023classnoaug**. maps 8194 - train,2984 - valid,3295-train. Classification with NONEs ans ALA's, no augmentation

Resolution 2.3-2.5 Å

res Å	EMDB	pdb file	comments
2.4	emd-3296	5ftk.pdb	p97 very noisy, looks like outer loops no in the map, good for validation and analysis, needs filtering for training
2.43	7638	6cvb	not released
2.5	emd-7025	6az3	Ribosome, contains lots fo DNA

Resolution 2.5-2.7 Å

res Å	EMDB	pdb file	comments
2.5	emd-7025	6az3	Ribosome, contains lots fo DNA, NOT USED
2.53	emd-8754	5w3e	rhinovirus B14, problematic, needs TEMPy
2.54	emd-8361	5t5h	Trypanosoma cruzi 60S ribosomal subunit contains lots fo DNA, NOT USED
2.6	emd-6272	3j9s_all.pdb	rotavirus VP6 at, VALIDATION, good but small, 100 PRO, 20 TRP,
2.6	emd-8743	5vy5.pdb	muscle aldolase using 200keV, good for train, only 13000 residues
2.7	3528	5mm2	nora virus structure, GOOD, but not fitted,Paper not publoshed
2.7	7024	6az1	Ribosome, contains lots fo DNA, NOT USED

Resolution 2.7-2.9 Å

used

res Å	EMDB	pdb file	comments
2.7	emd-6741	5xnl.pdb	C2S2M2-type PSII-LHCII, GOOD, requires filtering
2.71	8761	5w3l_all.pdb	rhinovirus B14 in complex w Nedd resampling with TEMPy
2.78	emd-7452	6cbe.pdb	rationally engineered gene delivery vector GOOD
2.79	emd-8189	5k0u_all.pdb	human rhinovirus C GOOD
2.8	emd-8604	5us7.pdb	bocavirus 3,GOOD
2.8	emd-7442	6caj.pdb	eukaryotic translation initiation factor 2B ,GOOD
2.8	emd-3246	5foj_all.pdb	Grapevine Fanleaf Virus complex with Nanobody GOOD
2.8	emd-8574	5uf6_all.pdb	chimeric adeno-associated virus-DJ GOOD
2.84	emd-7300	6bwx.pdb	f human bufavirus 1 GOOD
2.89	8314	5l35_all.pdb	NEEDS resampling with TEMPy.headful DNA-packaging bacterial virus at 2.89
2.9	8598	5urf.pdb	bocavirus 1 GOOD, same as 8604
2.9	6224	3j9c_all.pdb	toxin protective antigen pore , parts unfitted
2.9	6374	3jb0.pdb	RNA transcription and capping in a dsRNA virus, REquires TEMPy for resamplimg

unused

res Å	EMDB	pdb file	comments
2.7	3528	5mm2	nora virus structure, GOOD, but not fitted,Paper not publoshed
2.7	7024	6az1	Ribosome, contains lots of DNA, NOT USED
2.73	7589	6cs4	Not Published
2.76	7632	6cv1	Not Published
2.79	7636	6cv5	Not Published
2.8	emd-8191	5k0z.pdb	lactate dehydrogenase (LDH) in complex with GSK2837808A, NOT SO GOOD, looks like only one domain is available
2.86	7633	6cv2	Not Published
2.88	7083	6bco	TRPM4 in ATP bound state NOT GOOD, contains some unfitted regions
2.87	3951	6et5	NOT GOOD , has unufitted loops
2.9	3713	5nwy	NOT USED, lots of RNA
2.9	8875	5wpc	NOT Published
2.9	7048	6b44.pdb	Contains RNAType I-F CRISPR crRNA-guided Csy surveillance complex
2.9	8878	5wpf	Not Published
2.9	8343	5t2a	NOT USED, RNA,donovani 80S ribosome at 2.9
2.9	6555	3jci	Close-packed PCV2 Virus-like Particles GOOD
2.9	2847	5afi	Problems with Chimera
2.9	8872	5wp7	Not Published
2.9	8873	5wp8	8873
2.9	3883	6ek0	Ribosome Lots of RNA
2.9	8876	5wpd	Not Published
2.9	8877	5wpe	Not Published
2.9	8879	5wpg	Not Published
2.9	7600	6csh	Not Published
2.9	3640	5ngm	Ribosome, lots of RNA
2.9	7022	6ayp	Not Published

0.4 Results