# Fingerprint-Based Structure Retrieval Using Electron Density

Shuangye Yin and Nikolay V. Dokholyan*

Department of Biochemistry and Biophysics
University of North Carolina at Chapel Hill
Chapel Hill, NC 27599-7260
USA

## SUPPLEMENTARY MATERIALS

**Table S1.** Definition of different distance measures between two vectors $X$ ($x_i$, $i$=1..$n$) and $Y$ ($y_i$, $i$=1..$n$). $\overline{X}$ and $\overline{Y}$ are the average of X and Y, respectively.

| Distance measure name | Definition |
|---|---|
| Cosine | $1-\frac{X\cdot Y}{\lvert X\rVert Y\rvert}$ |
| Correlation | $1-\frac{(X-\overline{X})\cdot(Y-\overline{Y})}{\lvert X-\overline{X}\rVert Y-\overline{Y}\rvert}$ |
| Euclidean | $\sqrt{\sum(x_i-y_i)^2}$ |
| Manhattan | $\sum\lvert x_i-y_i\rvert$ |
| Canberra | $\sum\frac{\lvert x_i-y_i\rvert}{\lvert x_i\rvert+\lvert y_i\rvert}$ |

**Table S2**. Top 91 hits and their fingerprint similarity scores from screening GroEL electron density. The protein names are extracted from meta-data of the PDB files. Note that for several structures the name "60 KDA CHAPERONIN" is used for GroEL protein.

| PDBID | Fingerprint Difference | Protein Name |
|---|---|---|
| 1oel-G | 0.0136641 | GROEL (HSP60 CLASS) |
| 1oel-B | 0.0137425 | GROEL (HSP60 CLASS) |
| 1oel-E | 0.0138116 | GROEL (HSP60 CLASS) |
| 1oel-A | 0.0139113 | GROEL (HSP60 CLASS) |
| 1oel-F | 0.014112 | GROEL (HSP60 CLASS) |
| 1oel-C | 0.0143676 | GROEL (HSP60 CLASS) |
| 1kp8-G | 0.0144045 | GROEL PROTEIN |
| 1oel-D | 0.0144593 | GROEL (HSP60 CLASS) |

* Corresponding author, Email: dokh@med.unc.edu, Phone: 919-843-2513, Fax: 919-966-2852.

| | | |
|---|---|---|
| 1ss8-G | 0.0145948 | GROEL PROTEIN |
| 1ss8-E | 0.0147399 | GROEL PROTEIN |
| 1sx3-G | 0.0147529 | GROEL PROTEIN |
| 1ss8-B | 0.014758 | GROEL PROTEIN |
| 1xck-E | 0.014776 | 60 KDA CHAPERONIN |
| 1xck-M | 0.0148479 | 60 KDA CHAPERONIN |
| 1xck-B | 0.0150037 | 60 KDA CHAPERONIN |
| 1ss8-A | 0.0150081 | GROEL PROTEIN |
| 1kp8-D | 0.0150574 | GROEL PROTEIN |
| 1ss8-F | 0.01507 | GROEL PROTEIN |
| 1xck-A | 0.0150787 | 60 KDA CHAPERONIN |
| 1xck-G | 0.0151059 | 60 KDA CHAPERONIN |
| 1xck-K | 0.0151103 | 60 KDA CHAPERONIN |
| 1kp8-C | 0.0151385 | GROEL PROTEIN |
| 1xck-D | 0.0151575 | 60 KDA CHAPERONIN |
| 1pf9-L | 0.0152614 | GROEL PROTEIN |
| 1xck-I | 0.0152616 | 60 KDA CHAPERONIN |
| 1xck-J | 0.0152663 | 60 KDA CHAPERONIN |
| 1kp8-H | 0.0152879 | GROEL PROTEIN |
| 1xck-L | 0.0153507 | 60 KDA CHAPERONIN |
| 1ss8-C | 0.0153551 | GROEL PROTEIN |
| 1we3-N | 0.0154318 | CPN60(GROEL) |
| 1kp8-N | 0.0154651 | GROEL PROTEIN |
| 1sx3-D | 0.0154789 | GROEL PROTEIN |
| 1kp8-L | 0.0154792 | GROEL PROTEIN |
| 1sx3-C | 0.0155081 | GROEL PROTEIN |
| 1pcq-L | 0.0155118 | GROEL PROTEIN |
| 1kp8-I | 0.0155339 | GROEL PROTEIN |
| 1ss8-D | 0.015566 | GROEL PROTEIN |
| 1xck-F | 0.0156018 | 60 KDA CHAPERONIN |
| 1sx3-H | 0.0157191 | GROEL PROTEIN |
| 1xck-H | 0.0157566 | 60 KDA CHAPERONIN |
| 1we3-H | 0.0157588 | CPN60(GROEL) |
| 1xck-C | 0.0158204 | 60 KDA CHAPERONIN |
| 1pf9-K | 0.0158485 | GROEL PROTEIN |
| 1sx3-L | 0.0158799 | GROEL PROTEIN |
| 1kp8-E | 0.015897 | GROEL PROTEIN |
| 1pcq-H | 0.015899 | GROEL PROTEIN |
| 1we3-I | 0.0159098 | CPN60(GROEL) |
| 1sx3-N | 0.0159351 | GROEL PROTEIN |
| 1svt-L | 0.0159482 | GROEL PROTEIN |
| 1kp8-J | 0.0159659 | GROEL PROTEIN |
| 1pcq-K | 0.0159681 | GROEL PROTEIN |
| 1pf9-H | 0.0159875 | GROEL PROTEIN |
| 1sx3-I | 0.0159942 | GROEL PROTEIN |
| 1pf9-M | 0.016006 | GROEL PROTEIN |
| 1pf9-I | 0.0161077 | GROEL PROTEIN |
| 1kp8-B | 0.0161718 | GROEL PROTEIN |
| 1kp8-M | 0.0161826 | GROEL PROTEIN |
| 1kp8-A | 0.0162605 | GROEL PROTEIN |
| 1we3-J | 0.0162749 | CPN60(GROEL) |
| 1pcq-M | 0.0162994 | GROEL PROTEIN |
| 1kp8-K | 0.016304 | GROEL PROTEIN |

| | | |
|---|---|---|
| 1sx3-E | 0.0163272 | GROEL PROTEIN |
| 1xck-N | 0.0163366 | 60 KDA CHAPERONIN |
| 1kp8-F | 0.0163975 | GROEL PROTEIN |
| 1svt-K | 0.0164025 | GROEL PROTEIN |
| 1we3-K | 0.0164082 | CPN60(GROEL) |
| 1pcq-I | 0.016422 | GROEL PROTEIN |
| 1svt-H | 0.0164316 | GROEL PROTEIN |
| 1we3-L | 0.0164931 | CPN60(GROEL) |
| 1sx3-J | 0.0165149 | GROEL PROTEIN |
| 1sx3-B | 0.0167023 | GROEL PROTEIN |
| 1pf9-J | 0.0167209 | GROEL PROTEIN |
| 1svt-M | 0.0167482 | GROEL PROTEIN |
| 1sx3-K | 0.0167914 | GROEL PROTEIN |
| 1sx3-A | 0.0168139 | GROEL PROTEIN |
| 1pf9-N | 0.0168248 | GROEL PROTEIN |
| 1sx3-M | 0.0168328 | GROEL PROTEIN |
| 1svt-I | 0.0168422 | GROEL PROTEIN |
| 1sx3-F | 0.0168654 | GROEL PROTEIN |
| 1pcq-N | 0.0170332 | GROEL PROTEIN |
| 1pcq-J | 0.0170947 | GROEL PROTEIN |
| 1svt-N | 0.0173197 | GROEL PROTEIN |
| 1svt-J | 0.0174233 | GROEL PROTEIN |
| 1we3-M | 0.0178757 | CPN60(GROEL) |
| 1grl-B | 0.0234945 | GROEL (HSP60 CLASS) |
| 1grl-F | 0.0234954 | GROEL (HSP60 CLASS) |
| 1grl-G | 0.0234956 | GROEL (HSP60 CLASS) |
| 1grl-A | 0.0234957 | GROEL (HSP60 CLASS) |
| 1grl-E | 0.0234957 | GROEL (HSP60 CLASS) |
| 1grl-C | 0.0234958 | GROEL (HSP60 CLASS) |
| 1grl-D | 0.0234958 | GROEL (HSP60 CLASS) |

**Table S3**. Top 10 hits and their fingerprint similarity scores from screening rhodopsin electron density. The protein names are extracted from meta-data from the PDB files.

| PDBID | Fingerprint Difference | Protein Name |
|---|---|---|
| 1hzx-B | 0.023656 | RHODOPSIN |
| 1l9h-B | 0.0239033 | RHODOPSIN |
| 1f88-B | 0.0244048 | RHODOPSIN |
| 1jfp-A | 0.0289866 | RHODOPSIN |
| 3cap-B | 0.0296454 | RHODOPSIN |
| 3cap-A | 0.0298134 | RHODOPSIN |
| 2nun-A | 0.0302557 | AVIRULENCE B PROTEIN |
| 1gzm-A | 0.0309606 | RHODOPSIN |
| 3c9l-A | 0.0310502 | RHODOPSIN |
| 1gzm-B | 0.0311402 | RHODOPSIN |

**Figure S1**. The ROC curves of the structural retrieval experiments for the six SCOP families in the benchmark dataset. Five different fingerprint comparison methods are used, including cosine angle distance, correlation distance, Euclidean distance, Manhattan distance, and Canberra distance.

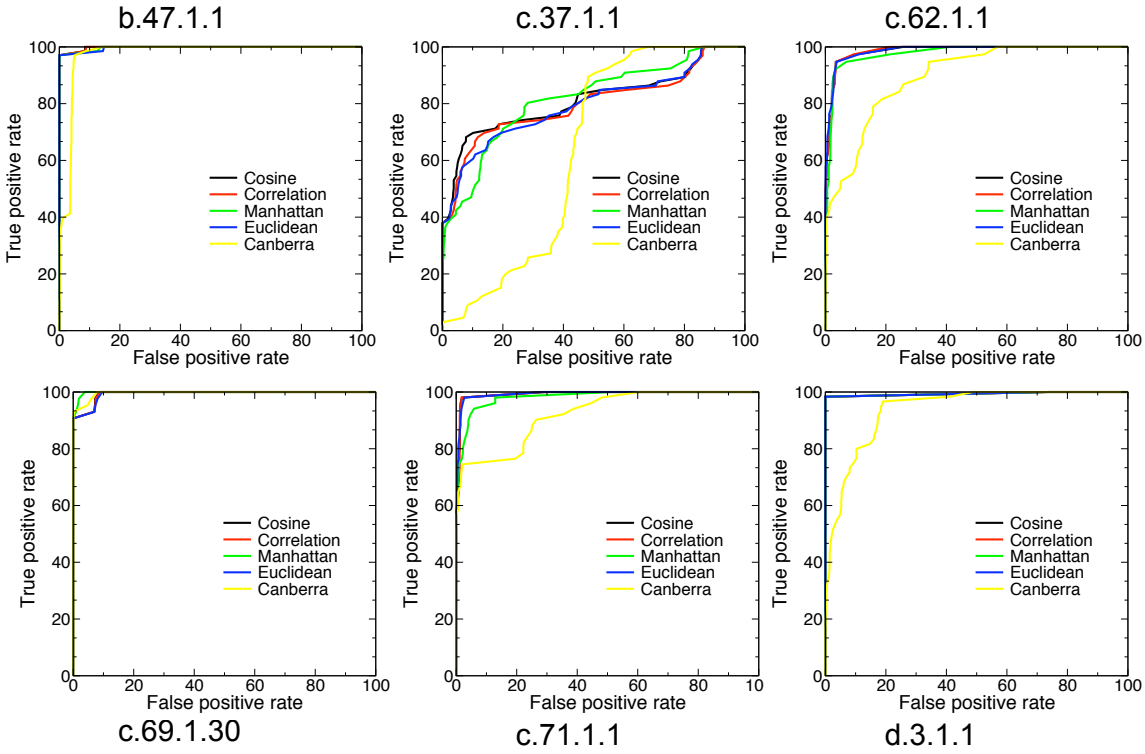


**Figure S2**. The statistics of fingerprint difference against decoys structures for the six SCOP families. The 5% statistical significance level corresponds to a fingerprint difference of approximately 0.015.

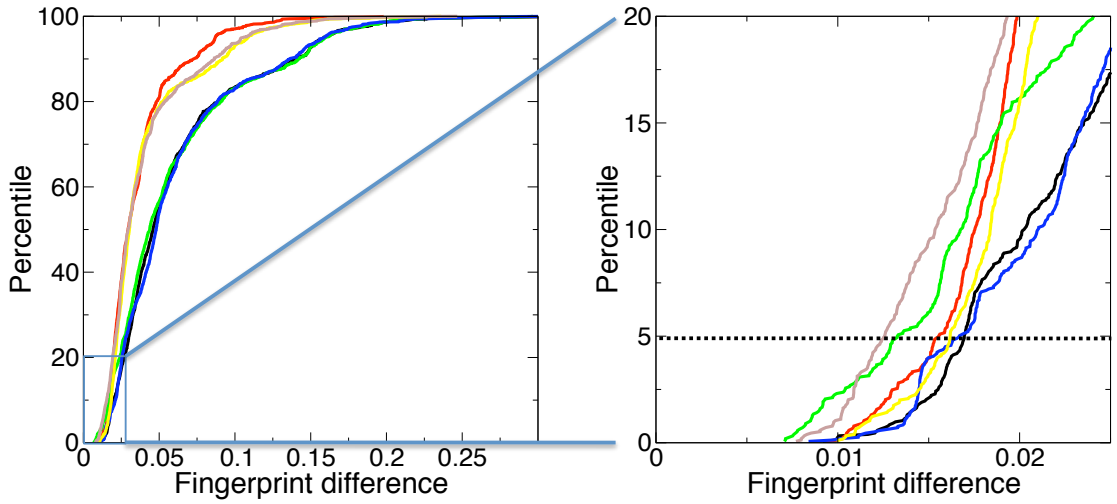

**Figure S3**. The AUC of structure retrieval for the six SCOP families with various Zernike order cutoffs.

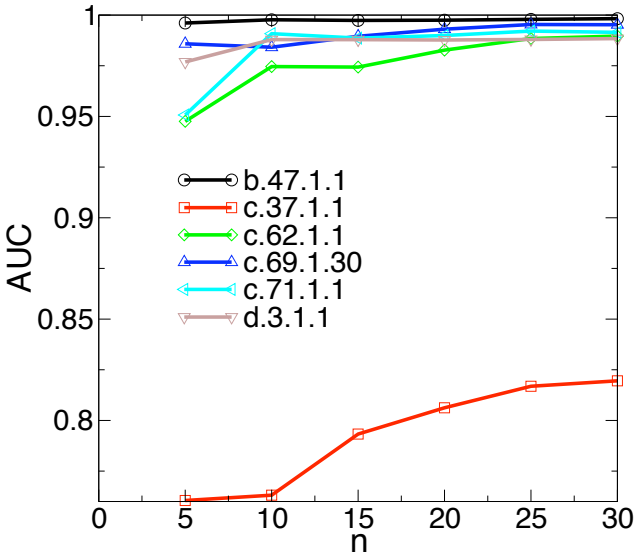


**Figure S4**. The ROC curve of structure retrieval for the six SCOP families (left panel), and the fingerprint differences as a function of RMSD from the query structures (right panel).

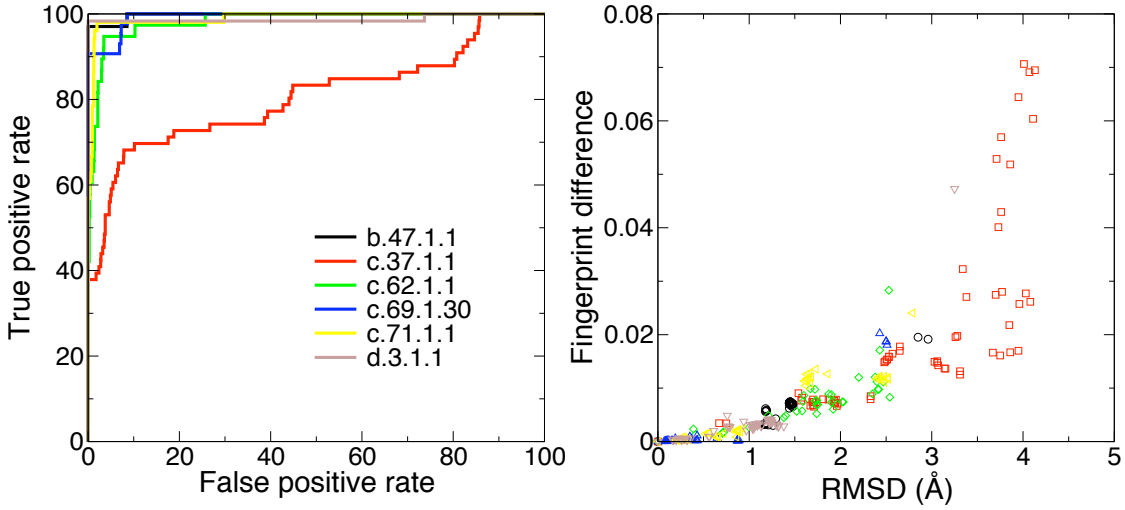

**Figure S5**. The fingerprints of six representative structures, and the standard deviations of the fingerprints on random rotations. All fingerprints have been normalized so that the first component (corresponding to total density) has magnitude 1.

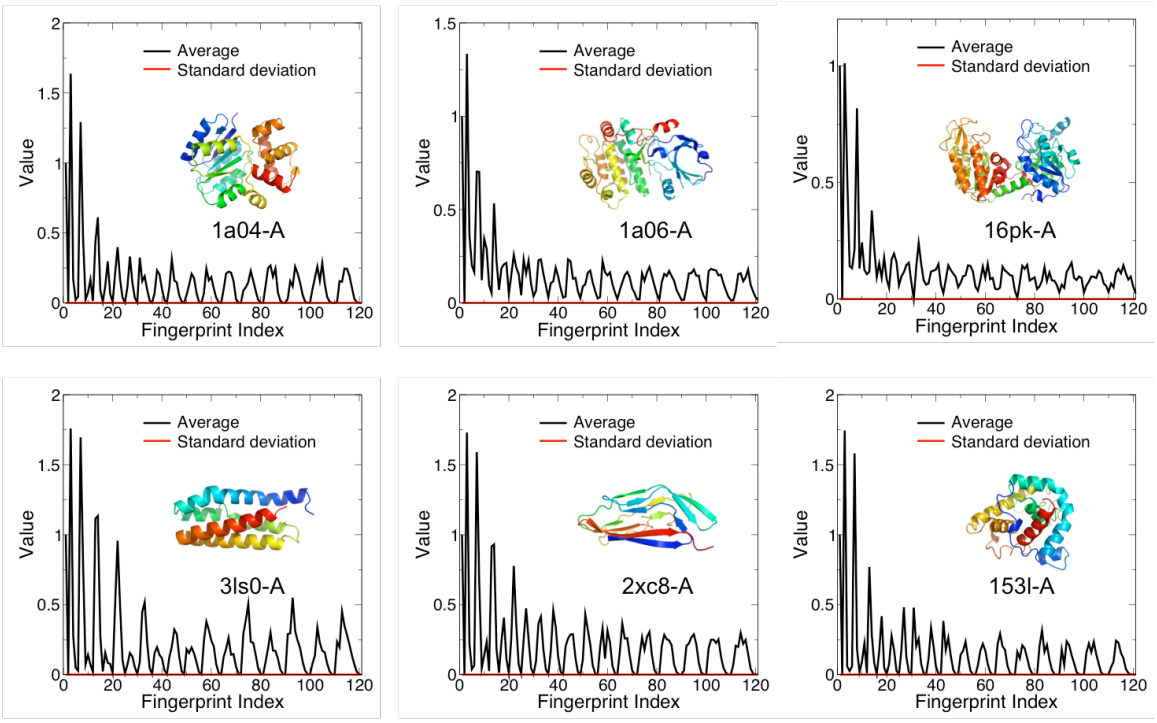


**Figure S6**. (a) The fingerprint differences of six representative structures as a function of center shift. (b) The histograms of the observed center shifts for the six SCOP families in the benchmark set. Using a fingerprint difference cutoff of 0.015 will allow deviation of density center of approximately 2-3.5Å, which is beyond the observed density center shift for the six SCOP families in the benchmark set.

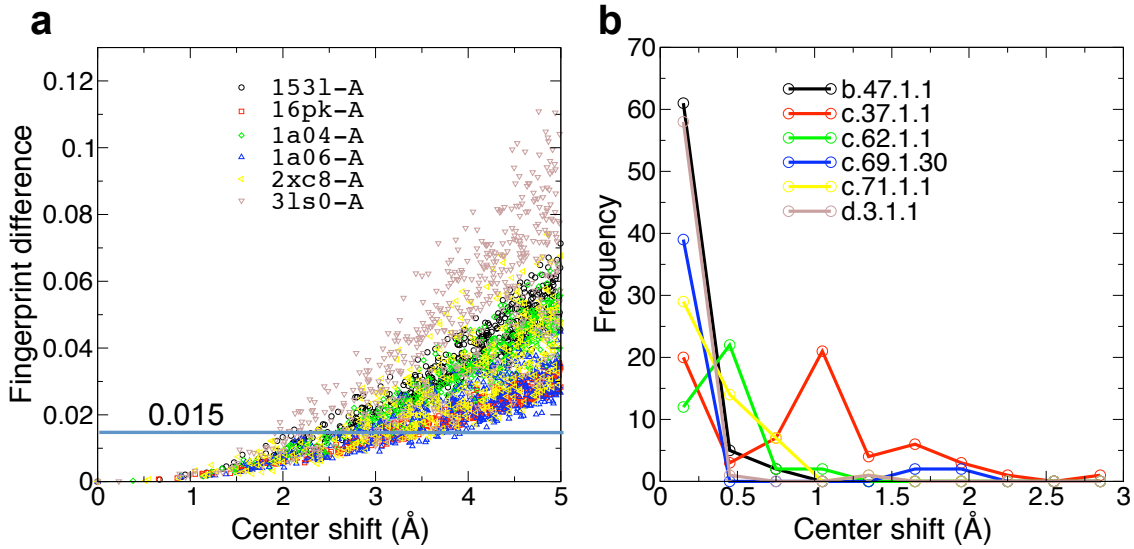

**Figure S7**. The top 4 distinct false positive structures and their rankings (shown in brackets following the PDB code) compared with the true positive hits (PDB code 1hzx-B, ranked No. 1).

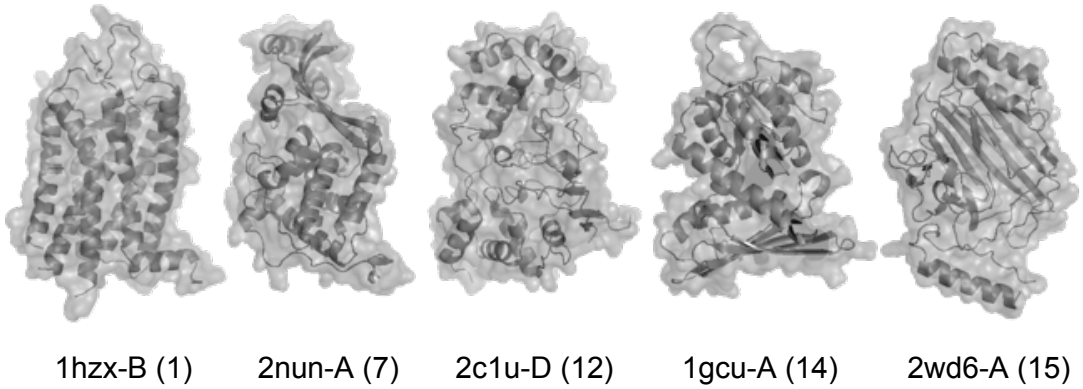


**Figure S8**. The histogram of the $r_{max}$ to $r_g$ ratio of all high-resolution structures in the PDB snapshot. The ratio is between 1.5 and 3.5 for the majority of structures.

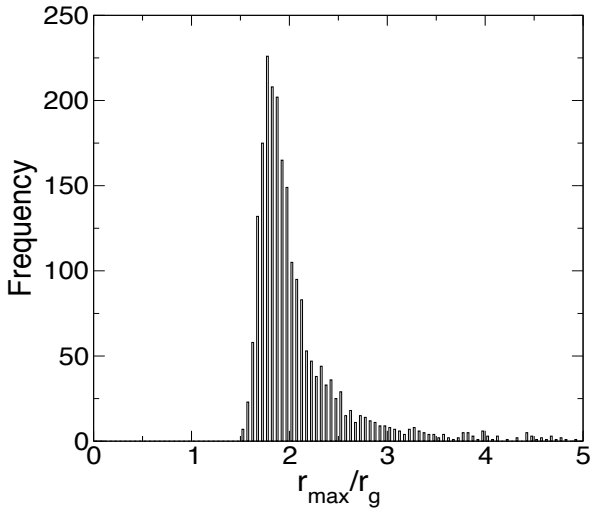

**Figure S9.** (a) The number of residues and conformational variation of all protein structures in the PDB that are classified into the rhodopsin family by SCOP or CATH. The structural difference is measured using RMSD from the top hit (PDB code 1hzx-B). Three structures are identified as incorrectly classified by SCOP. (b) Structural alignment of the representative structures of the two clusters.

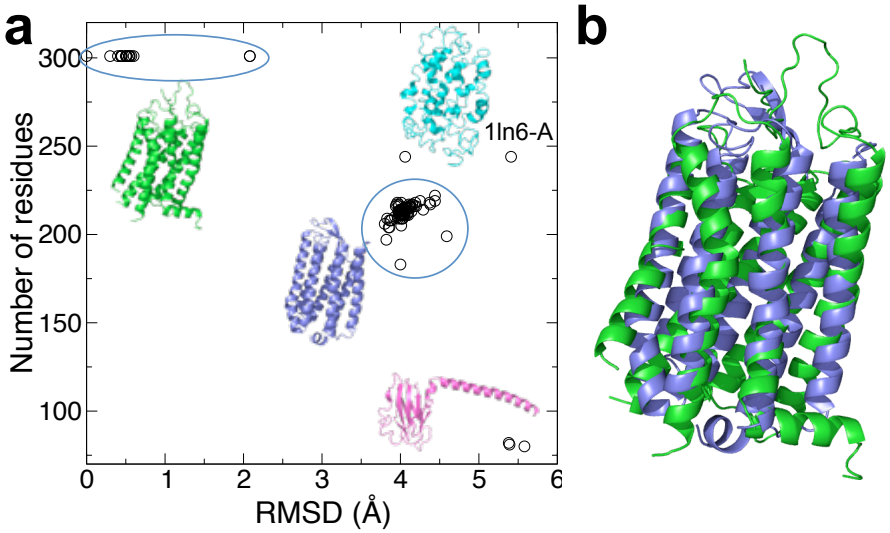


**Figure S10**. (a) The number of residues and conformational variations of all protein structures in the PDB that are classified into the GroEL family by SCOP or CATH. The structural difference is measured using RMSD from the top hit (PDB code 1oel-G). (b) Structural alignment of the representative structures of the two clusters.