Unsupervised Machine Learning and Phase Space Reduction: A Robust and Generalisable Approach for Concurrently Solving the Protein Complex Conformation Classification and Quantification Problems

Daniel Celis Garza

Science and Technology Facilities Council: Scientific Computing Department: CCP4-PDBe Collaboration

CAPRI Meeting
16 February, 2024

Science and Technology Facilities Council

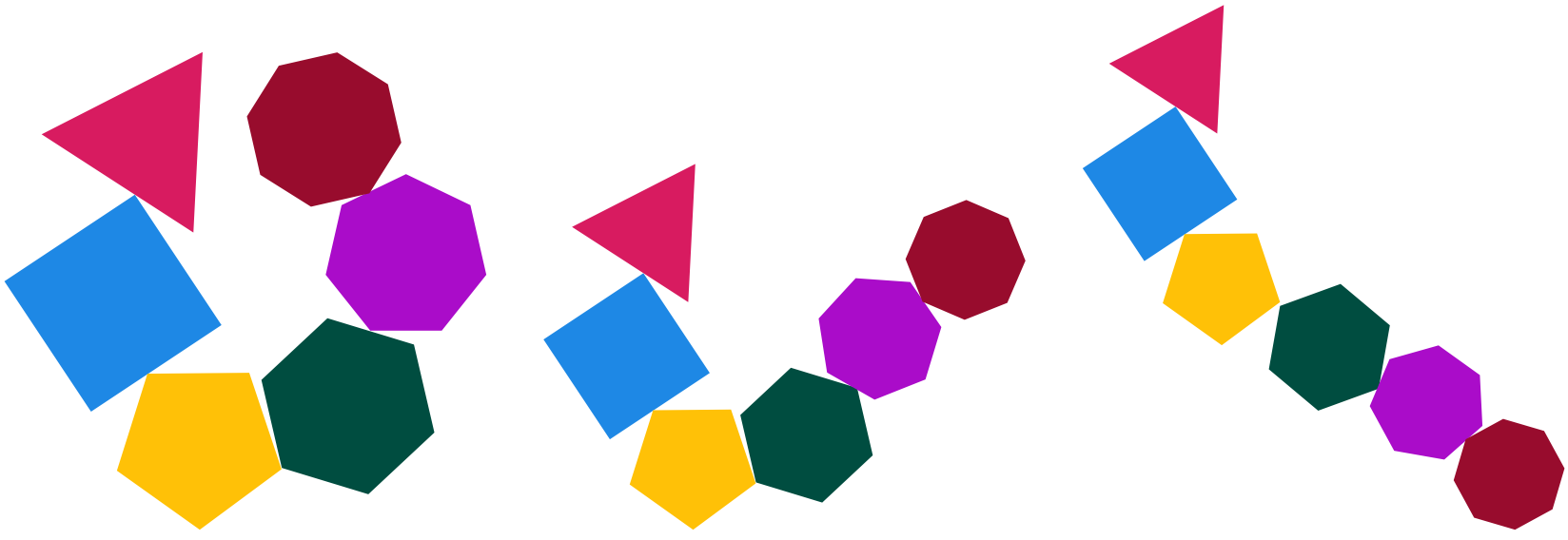
Contents

FunCLAN

Science and Technology Facilities Council

What do we want to know?

Science and Technology Facilities Council



(a) Closed. (b) Semi-open. (c) Open.

Figure 1: Three samples. Can we classify them as different conformers? Can we quantify the differences?

Science and Technology Facilities Council

- The complex's topology remains the same, so graph-theoretic approaches would *not* work without heuristics.
- It works great for classifying complex classes [1]
- For conformers, you'd have to weigh edges between *corresponding* chains according to some relational characteristic.
- How do you measure relationships? Centrality (which one), minimal spanning trees, walks, etc.

Science and Technology Facilities Council

- There are tools for working on single chains [2].
- They need some pre- and post-processing. Chain ordering and clustering respectively.
- They may be inappropriate to describe the relationships between all chains.

Assumptions

Science and Technology Facilities Council

- Sufficiently high sequence similarity between corresponding chains. In the future maybe high structural similarity like q-score [2].
- Chains are rigid bodies.

Goals

Science and Technology Facilities Council

- Identify complex-level conformers.
- "Measure" degree of similarity between them.

How do we solve the problem?

We need to synthesize some new information from the PDB file.

Science and Technology Facilities Council

- Corresponding chain pairs between samples.
- Find how the relationships between a complex's chains change from sample to sample.
- ???
- Profit.

Chain pairs

Threshold

Science and Technology Facilities Council

- Sequence match two chains.
- Chains must be sufficiently similar to be matched with one another.

$$I_a \times I_b > t^2 \,. \tag{1}$$

- Where I_i is the percentage of the sequence of chain i, found in the matched sequence, and t is the threshold value.
- The pairing is not in general bijective. We need to make it so later.

Chain pairs

1-to-1 matches

Science and Technology Facilities Council

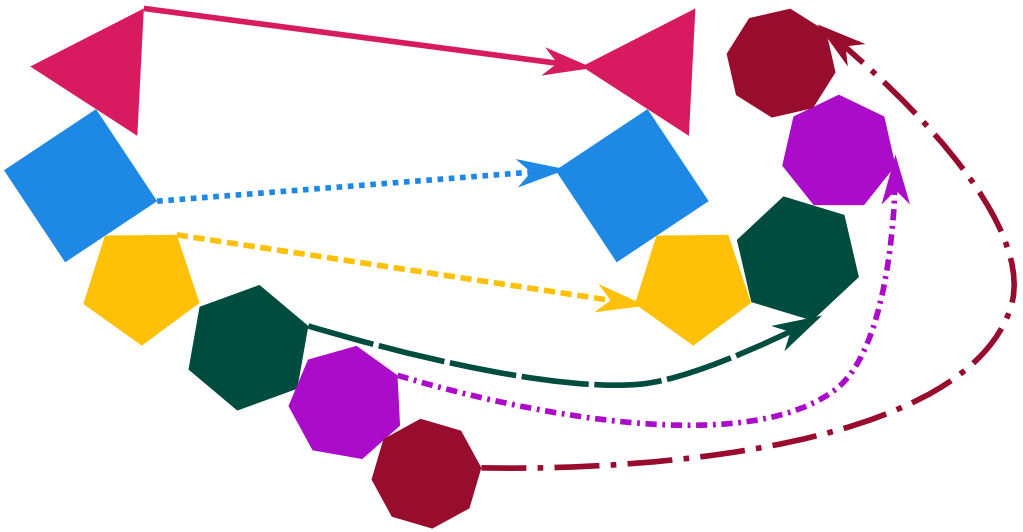


Figure 2: In the ideal case, chains can only be matched 1-to-1. In non-ideal cases a single chain may have multiple matches (homo N-mers).

Chain pairs

N^2 matches

Science and Technology Facilities Council

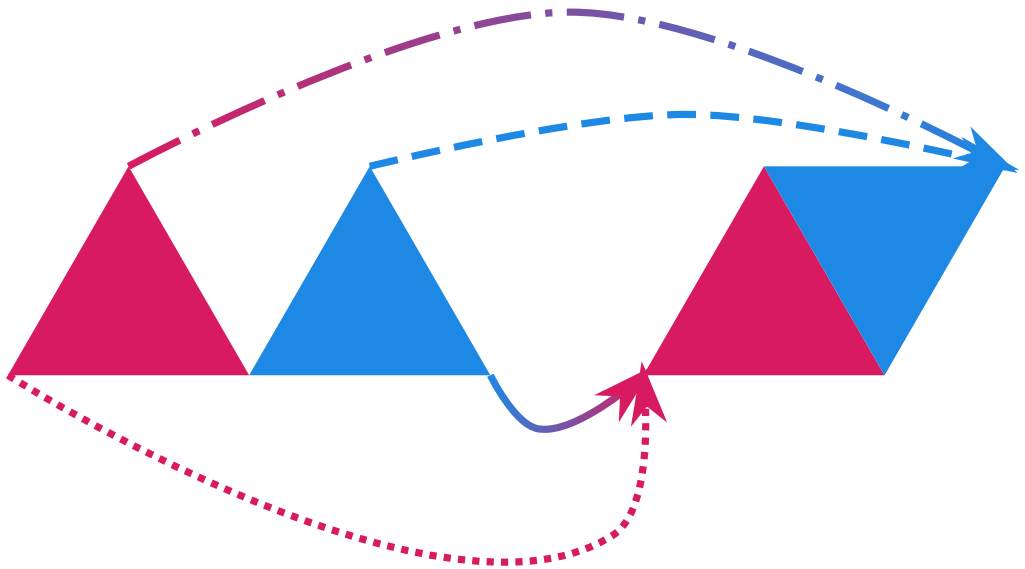


Figure 3: Two chains of the same type (colour only for clarity). In non-ideal cases, a single chain may have multiple matches. Homo N-mers may have up to N^2 matching pairs.

Comparing Transformations

How do transformations of sets of chains change between samples?

Science and Technology Facilities Council

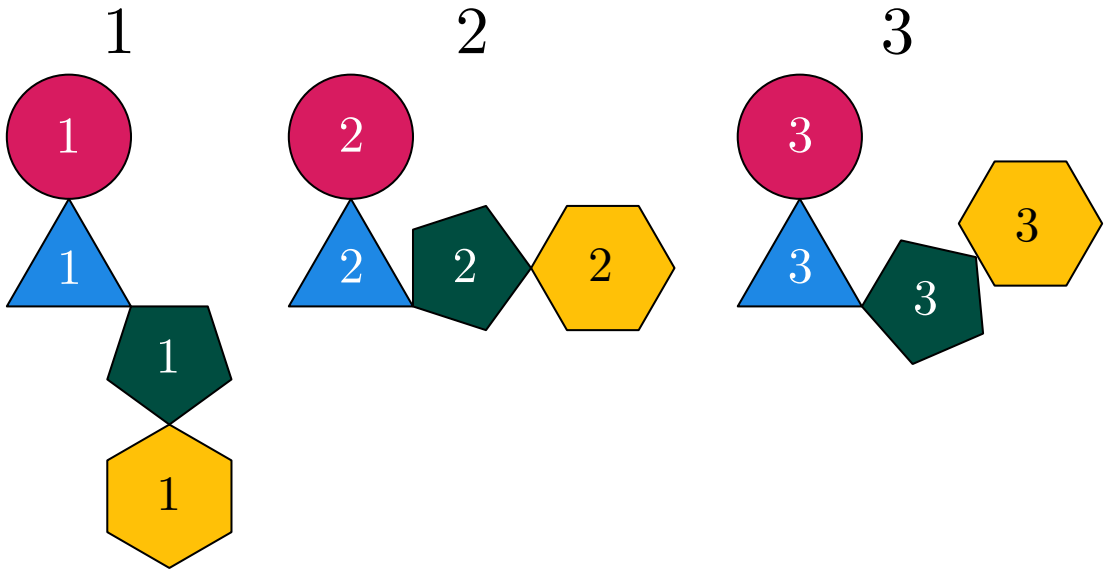


Figure 4: Complex conformations are defined by *"concerted" transformations of sets of chains*, when going from one conformation to another.

Scoring system

Desirable properties

Science and Technology Facilities Council

- Comparing transformation a to b must be the same as comparing b to a.
  - $s_{a,\,b} = s_{b,\,a}$
- Comparing equivalent transformations (ie their inverse) must yield the same result.
  - $s_{a,\,b} = s_{a^{-1},\,b}$
- Comparing a transformation to itself must have a score of zero, $s_{a,\,a} = 0$.
- The larger the difference in the transformation, the larger the score.

Scoring system

Definitions

Science and Technology Facilities Council

- Let $\mathcal{A}$ and $\mathcal{B}$ be two superposition transformations (rotation matrix + translation vector).
- Let $\hat{\boldsymbol{v}}_i$ be a basis column vector in 3D Euclidean space for dimension $i \in \{1, 2, 3\}$.

Scoring system

Science and Technology Facilities Council

$$\boldsymbol{a}_1 = \mathcal{A}(\hat{\boldsymbol{v}}_1), \quad \boldsymbol{a}_2 = \mathcal{A}(\hat{\boldsymbol{v}}_2), \quad \boldsymbol{a}_3 = \mathcal{A}(\hat{\boldsymbol{v}}_3) \tag{2}$$

$$\boldsymbol{b}_1 = \mathcal{B}(\hat{\boldsymbol{v}}_1), \quad \boldsymbol{b}_2 = \mathcal{B}(\hat{\boldsymbol{v}}_2), \quad \boldsymbol{b}_3 = \mathcal{B}(\hat{\boldsymbol{v}}_3) \tag{3}$$

$$s_{\mathcal{A},\mathcal{B}} = \sqrt{\frac{1}{3}\left(\|\boldsymbol{a}_1 - \boldsymbol{b}_1\|^2 + \|\boldsymbol{a}_2 - \boldsymbol{b}_2\|^2 + \|\boldsymbol{a}_3 - \boldsymbol{b}_3\|^2\right)} \tag{4}$$

- Where $\boldsymbol{a}_i$ and $\boldsymbol{b}_i$ are 3×1 column vectors that resulted from transforming the basis vectors by $\mathcal{A}$ and $\mathcal{B}$ respectively.
- And $s_{\mathcal{A},\mathcal{B}}$ is the similarity score between transformations.
- This meets all our requirements.

Simple example

Science and Technology Facilities Council

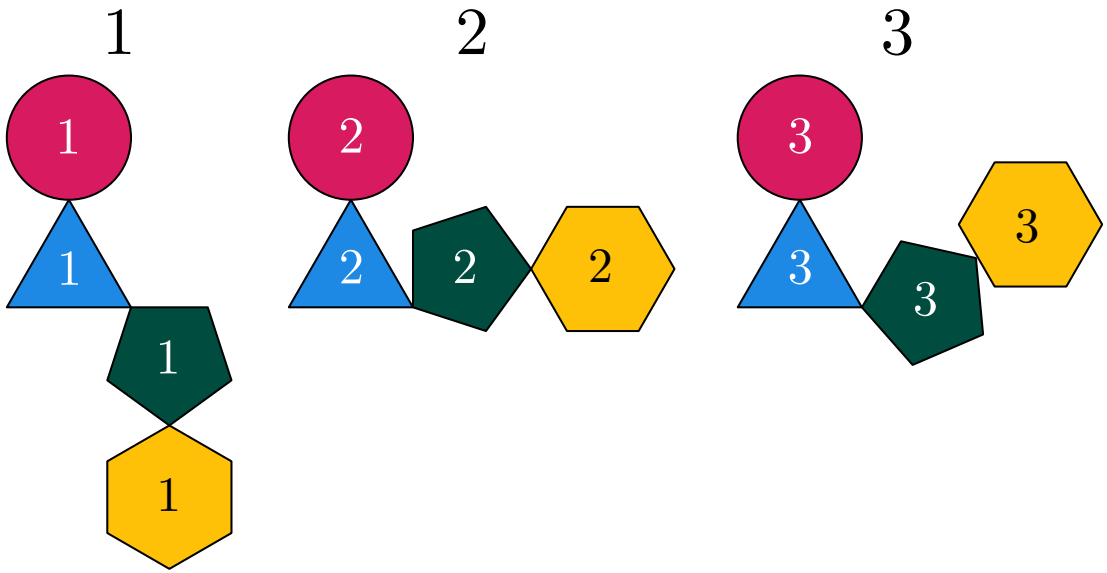
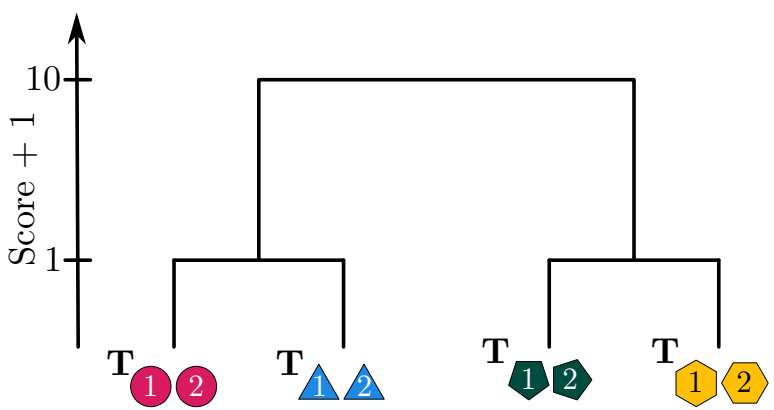


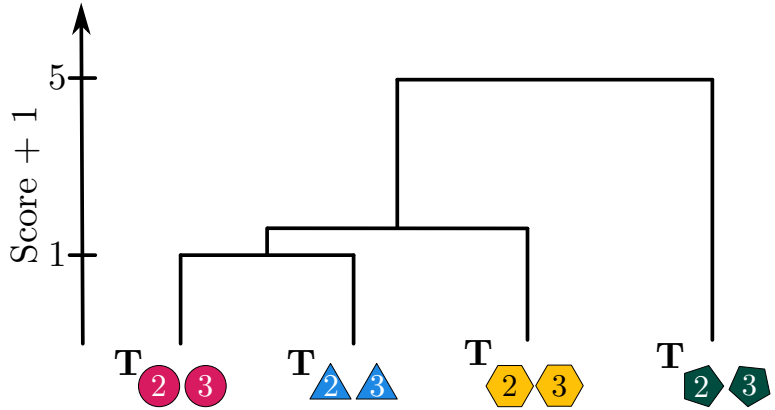
Figure 5: How does each chain need to move in order to go from one sample to another? Can we find commonalities in these movements? Can we arrange them according to how similar they are to one another?

Simple example

Science and Technology Facilities Council



(a) This dendrogram compares conformations 1 and 2.



(b) This dendrogram compares conformations 2 and 3.

Figure 6: Identification of complex conformations from pairwise comparisons. Hierarchical clustering can do this for us.

Simple example

Superpose 2 onto 1

Science and Technology Facilities Council

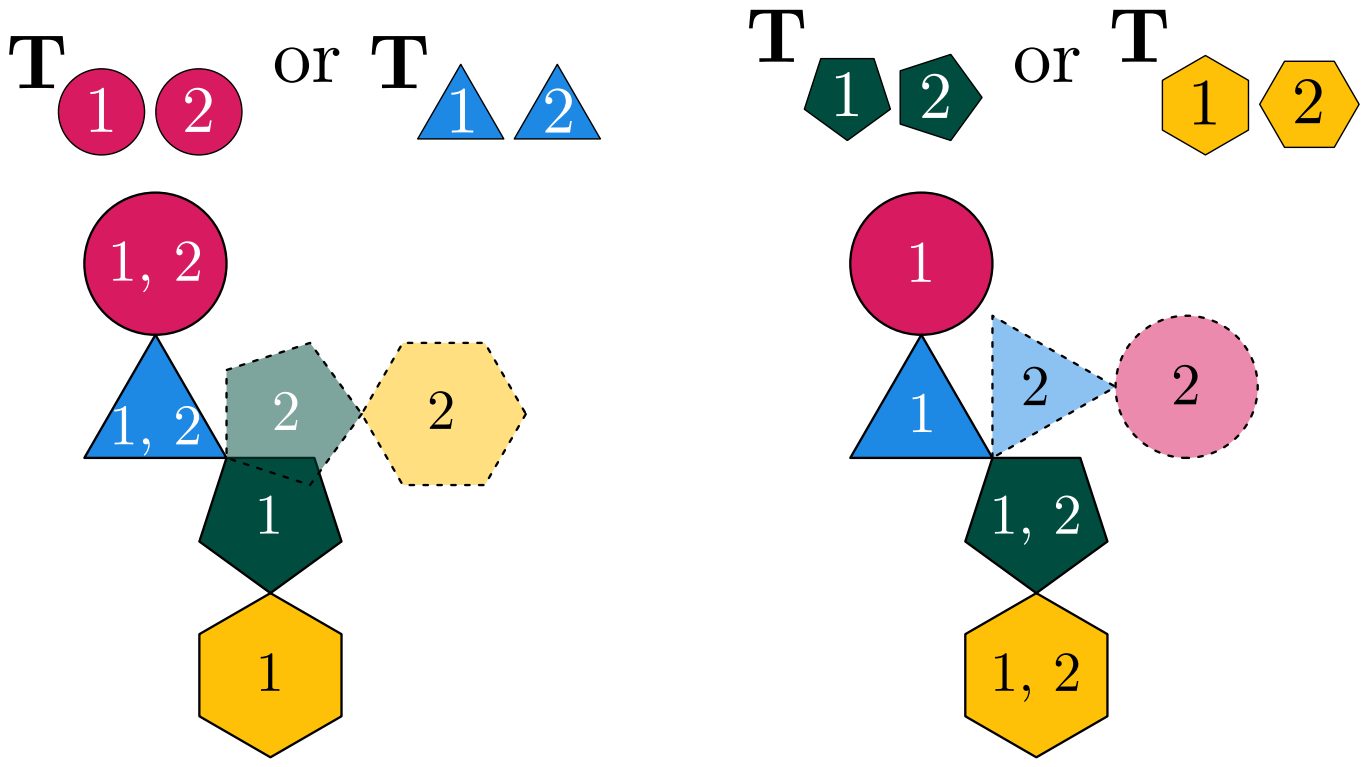


Figure 7: Superposing sample 2 onto 1.

Simple example

Superpose 3 onto 2

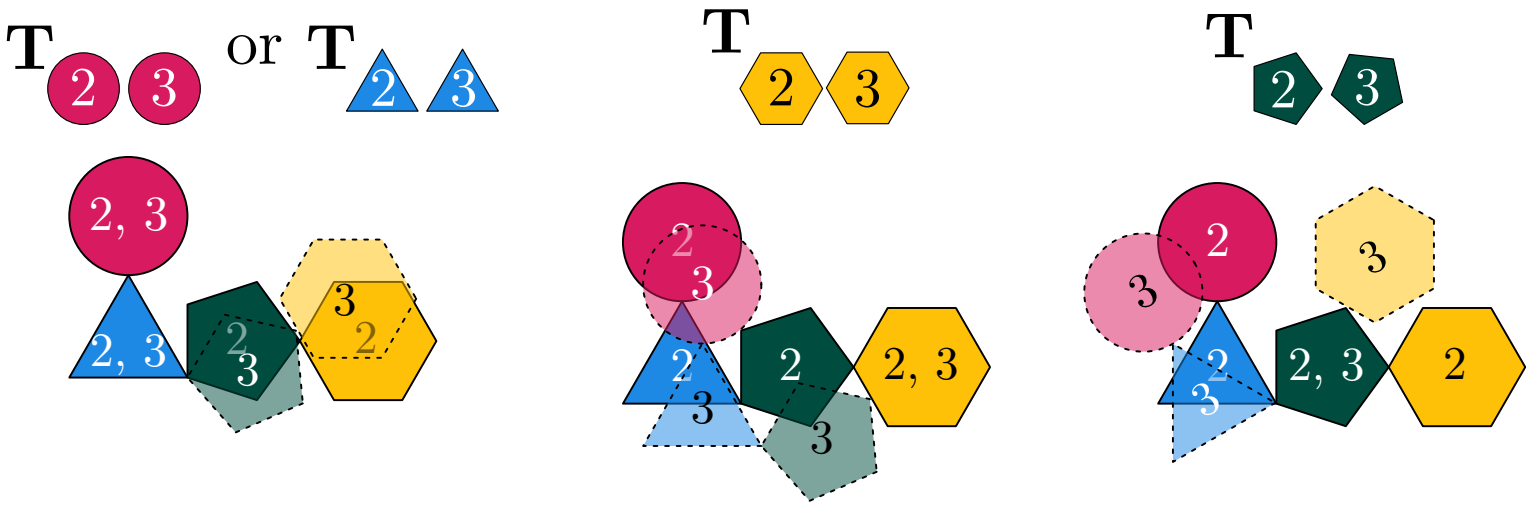


Figure 8: Superposing sample 3 onto 2.

Dealing with Multiple Matching Pairs

Science and Technology Facilities Council

- Each transformation uses two chains.
- Traverse the dendrogram from the smallest height to the largest.
- Check the chains involved in the transformations at every height (up to 2 per height).
- If a cluster is made up of two transformations, you have to ensure both transformations are valid before checking previous valid transformations.
  - If any chain appears in both transformations, this is an invalid cluster and move on to the next height.
  - Else continue to the next step.

Dealing with Multiple Matching Pairs

Science and Technology Facilities Council

- If the cluster only has one transformation, you can check the previous valid transformations straight away.
- If a cluster has no transformations (i.e. is made up of other clusters), you move on to the next height.
- Has any chain in the transformation being checked, been involved in a valid transformation we have previously encountered?
  - If no, this is a valid transformation, add it to the list, and move on to the next transformation/height.
  - If yes, this is an invalid transformation, move on to the next transformation/height.

Dealing with Multiple Matching Pairs

Science and Technology Facilities Council

- Clusterise again only with valid transformations to get a clean pairwise dendrogram.
- We can also take all valid transformations, for all comparisons and clusterise them. Giving a chain-level view of how all samples relate to each other.

From chains to complexes

Visualising the conformational space

Science and Technology Facilities Council

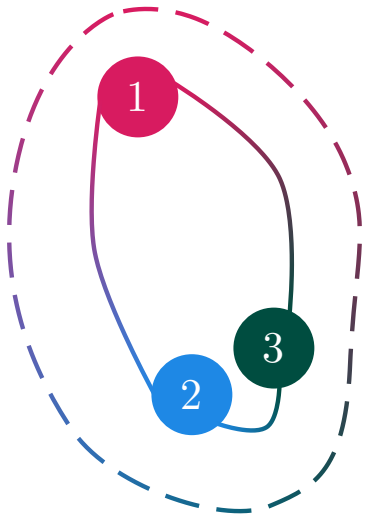


Figure 9: Each sample has a set of relationships with all others. The relationships are multidimensional. We can use these relationships to see where each one lies in the in-sample conformational landscape.

From chains to complexes

How?

Science and Technology Facilities Council

- We have a dendrogram containing how all valid transforms from all pairwise comparisons compare to each other.
- For each sample we find where all of its transformations appear in the dendrogram, preserving the order in which they are found (smallest dendrogram height to largest).
- For each one we will have an L_i-dimensional point, where L_i is the number of valid transformations for sample i in the dendrogram.

From chains to complexes

How?

Science and Technology Facilities Council

- If we know we have good data.
  - All chains are represented in all samples.
  - All chain comparisons meet the sequence similarity threshold.
- We can use strict mode (default), meaning all L_i are equal.
- If any sample doesn't meet the criteria, the program throws a runtime error as soon as possible, explaining why the failure occurred and ways to fix it.

From chains to complexes

How?

Science and Technology Facilities Council

- If we don't have good data.
- Each L_i can be different.
- Turn off strict mode.
- Summarise each L_i-dimensional point as a 1D point,

$$p = \sqrt{\frac{1}{L_i} \sum_{j=1}^{L_i} h_j^2} \,. \tag{5}$$

- Where h_j is the value of the vector at position j, i.e. the height at which the corresponding transformation appears in the dendrogram.

From chains to complexes

How?

Science and Technology Facilities Council

- We can place the points into a vector.
- Clusterise using the Euclidian distance.

Determining optimal number of clusters

Science and Technology Facilities Council

- How many conformers does the data suggest we have?
- Use the 2-difference gap statistic [3] adapted to hierarchical clusters, i.e. discretised stable point analysis.
- Cuts the tree into as many trees as possible, and uses the difference in the information (heights) provided by each cut to determine the ideal cut level. There is a cap to avoid overfitting. Defaults to $c = \left\lceil \sqrt{N} \right\rceil$, where N is the number of samples.
- Each leaf at the ideal cut level can be viewed as a single conformation.

Real Examples

SARS-CoV 19 Spike protein

Science and Technology Facilities Council

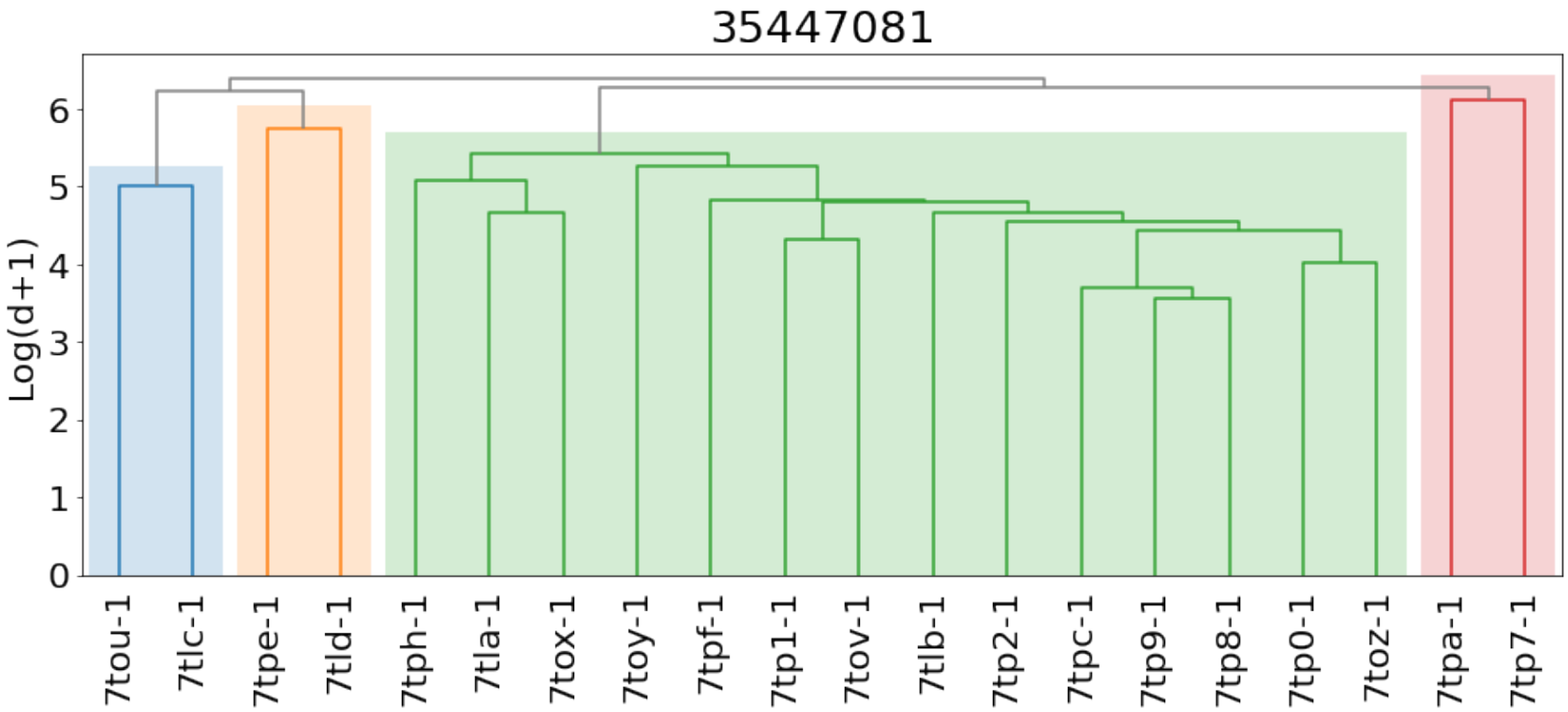
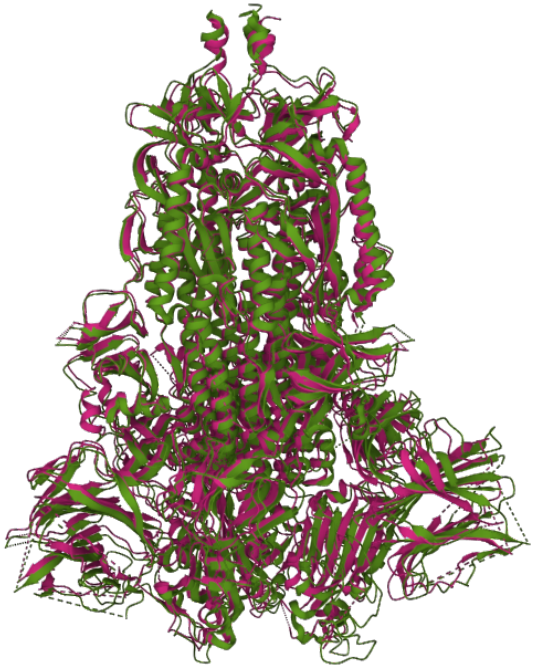


Figure 10: SARS-CoV 19 Spike protein, homo trimer. Ward's linkage.

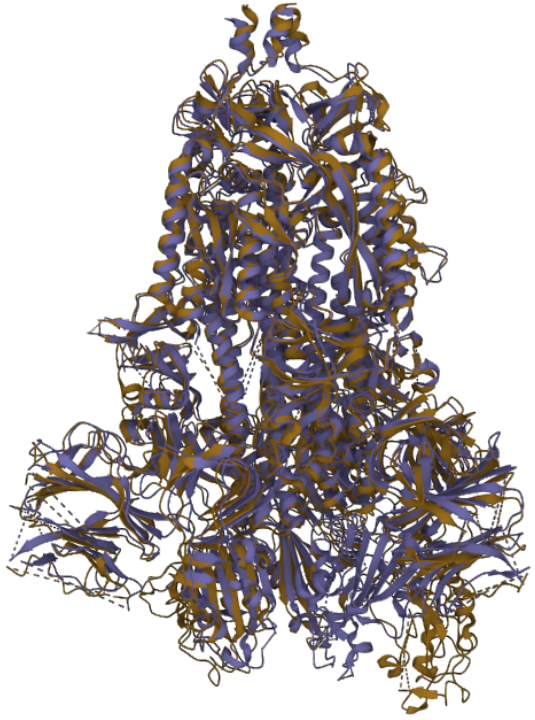
Real Examples

SARS-CoV 19 Spike protein

Science and Technology Facilities Council



(a) 7tou-1, 7tlc-1.

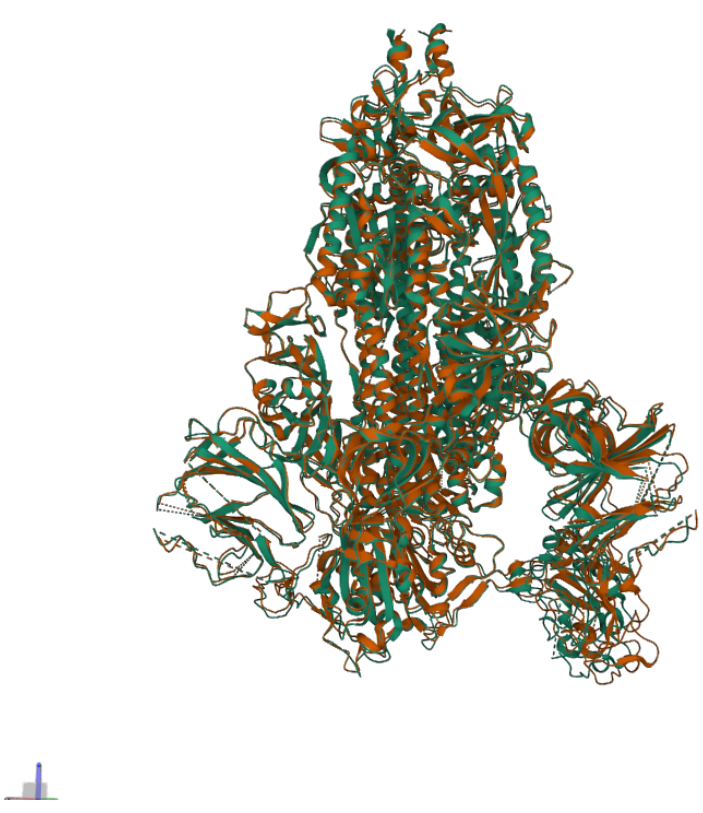


(b) 7tpe-1, 7tld-1.

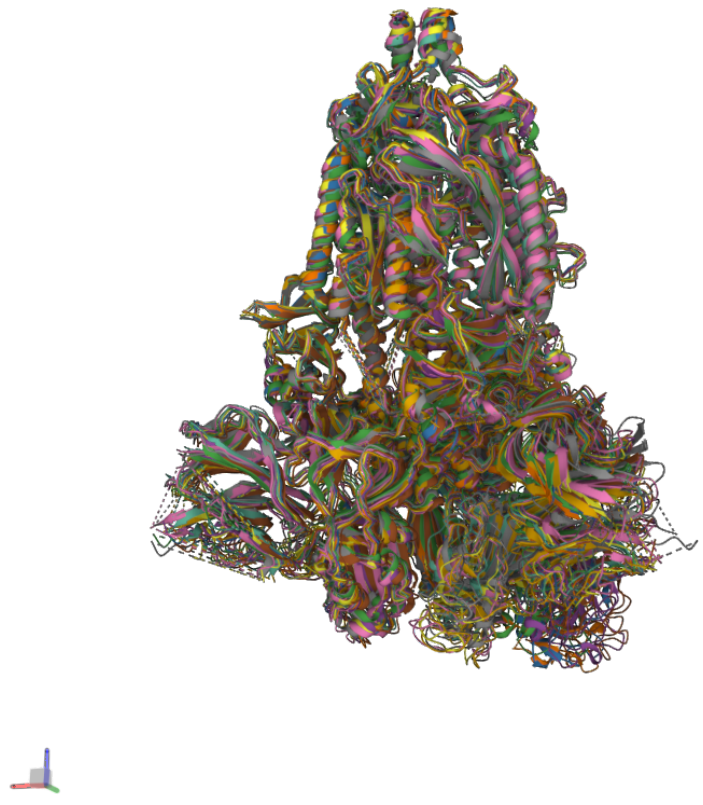
Real Examples

SARS-CoV 19 Spike protein

Science and Technology Facilities Council



(a) 7tpa-1, 7tp7-1.



(b) All the rest.

Real Examples

HIV transmembrane protein

Science and Technology Facilities Council

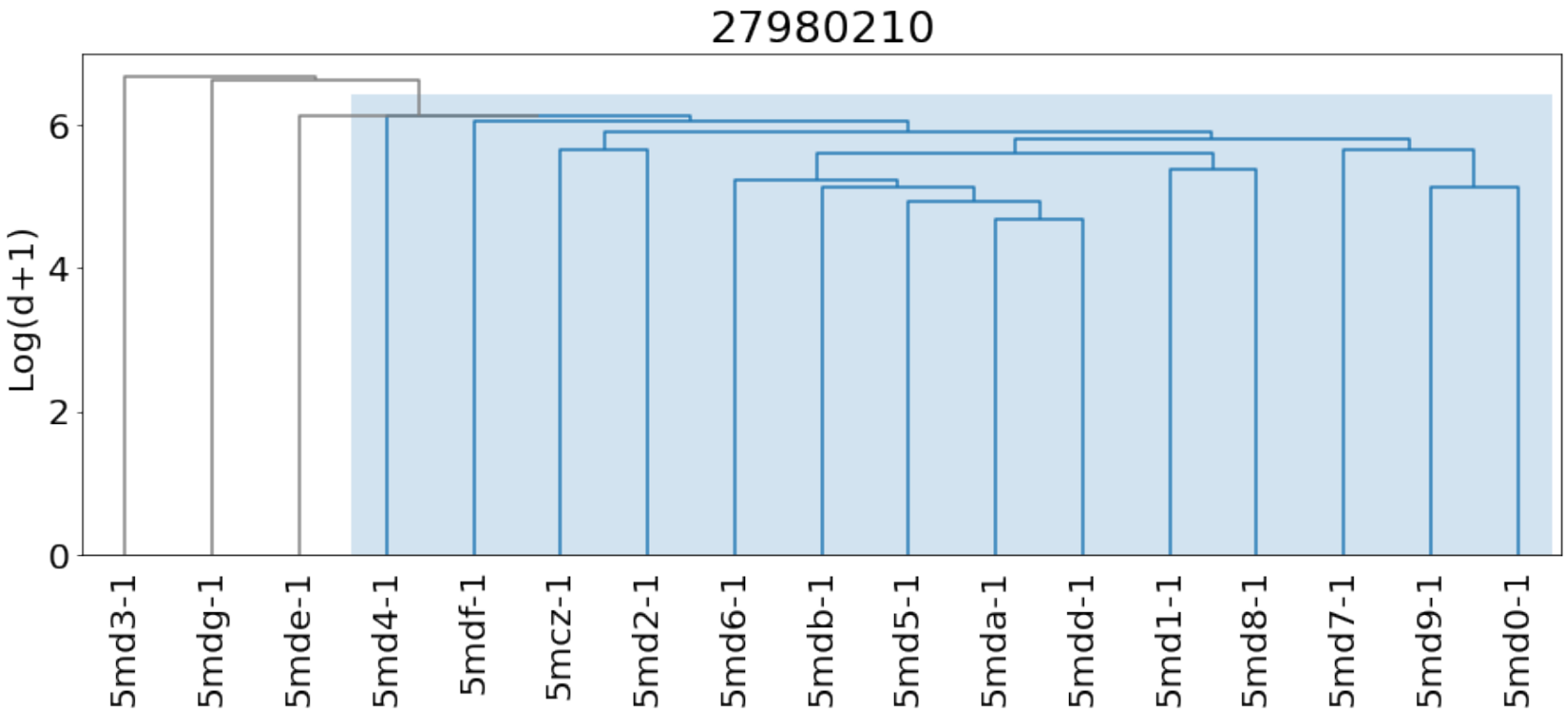


Figure 13: HIV transmembrane protein, hetero tetradecamer. Ward's linkage.

Real Examples

Pancreatic ATP-sensitive potassium channel

Science and Technology Facilities Council

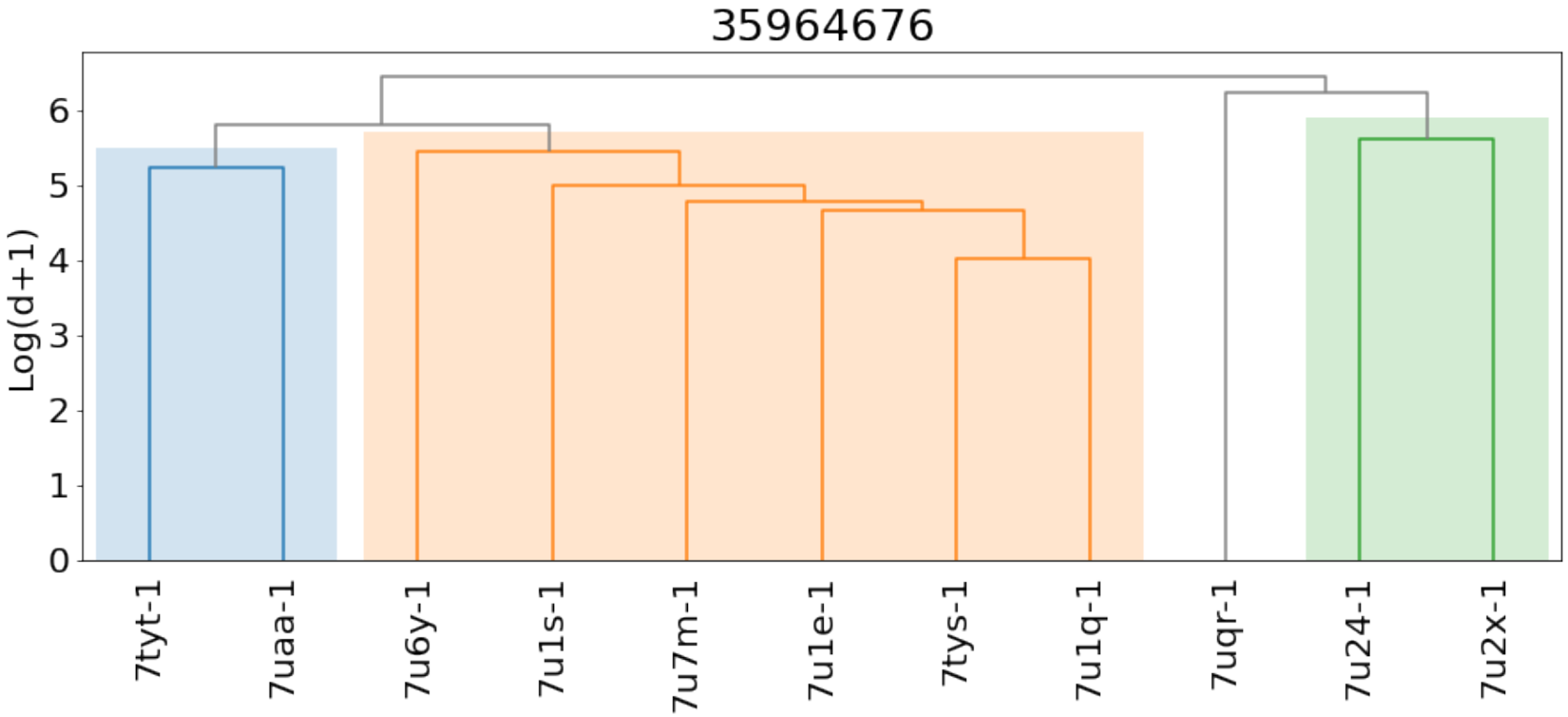


Figure 14: Pancreatic ATP-sensitive potassium channel, hetero pentamer. Ward's linkage.

Real Examples

Insect flight muscle

Science and Technology Facilities Council

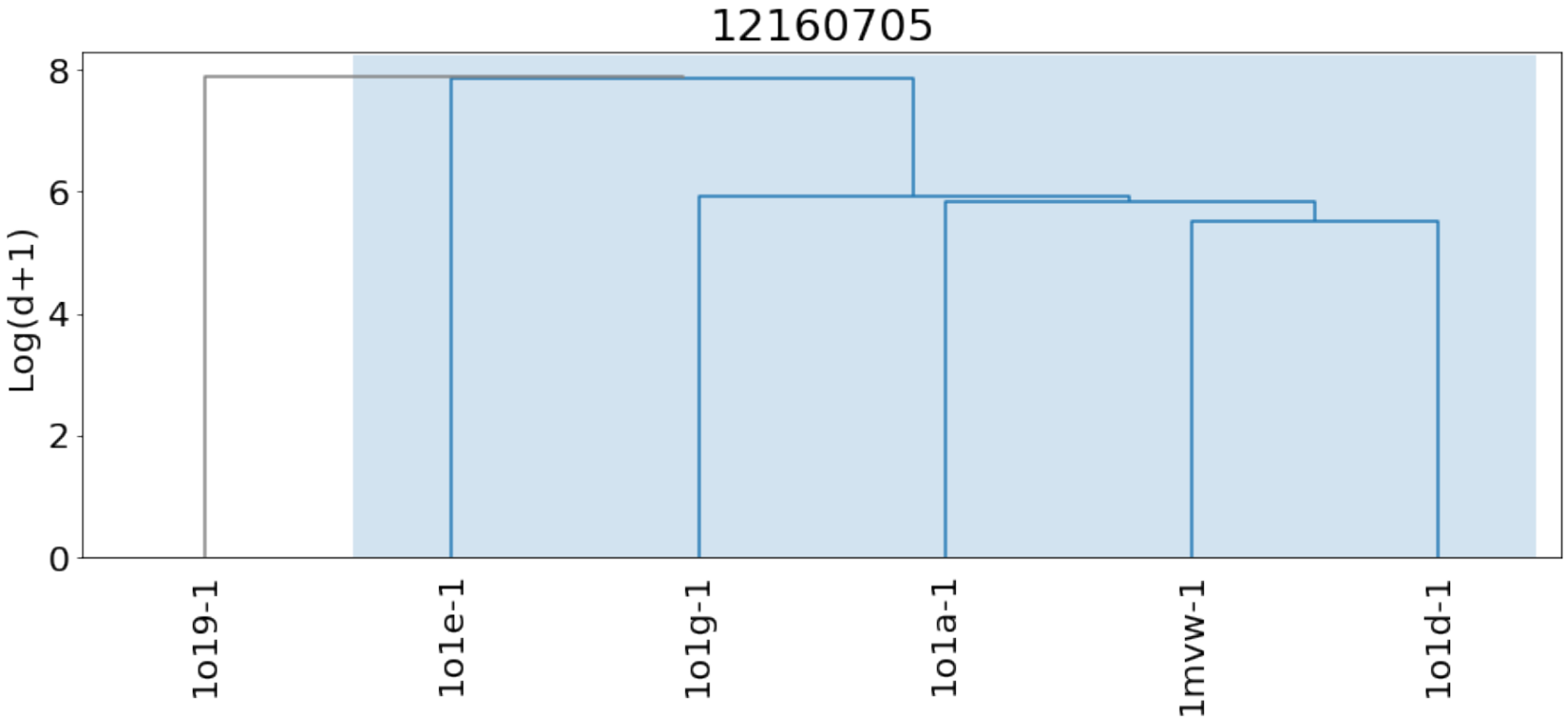


Figure 15: Insect flight muscle, hetero 32-mer. Ward's linkage.

Real Examples

Human insulin receptor

Science and Technology Facilities Council

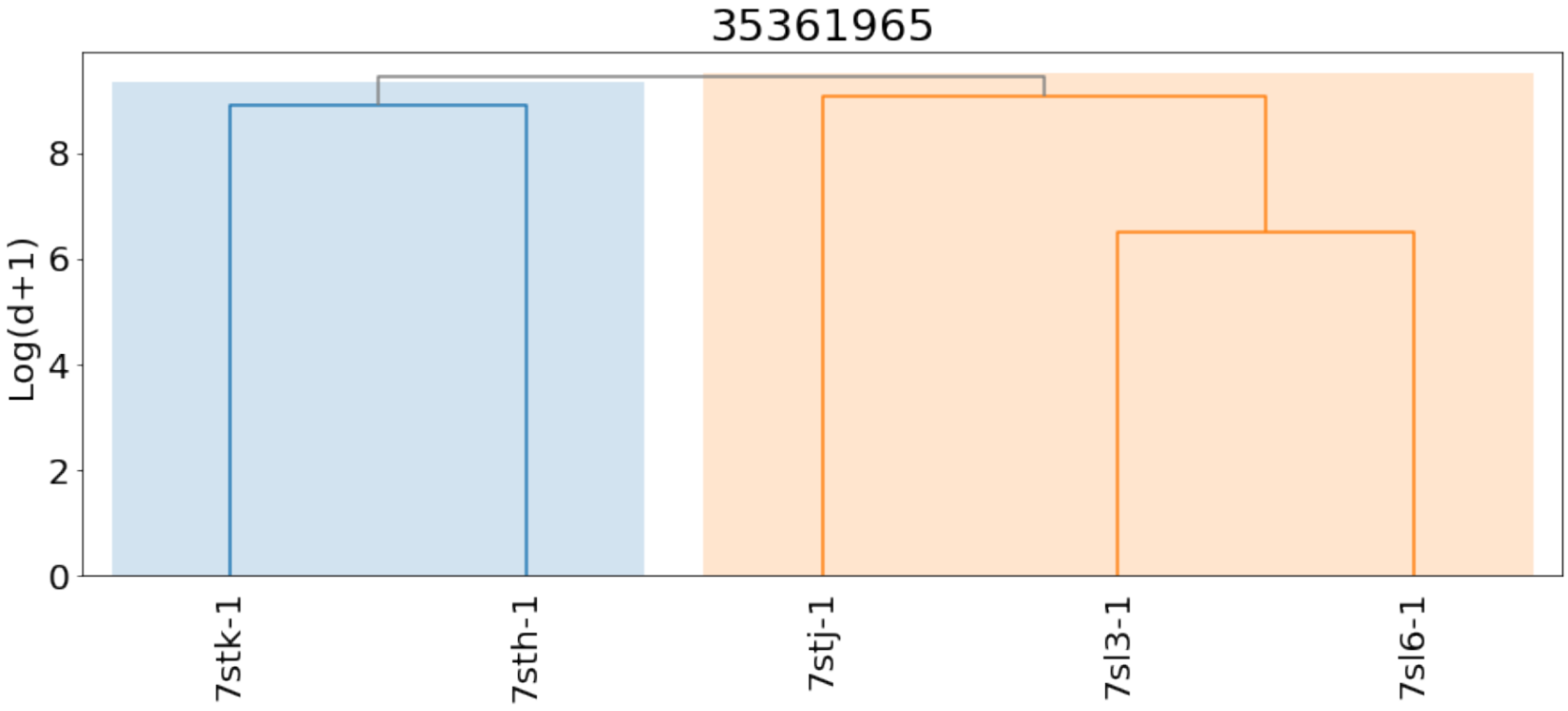


Figure 16: Human full-length insulin receptor, hetero hexamer. Ward's linkage.

Real Examples

E. coli RNA polymerase

Science and Technology Facilities Council

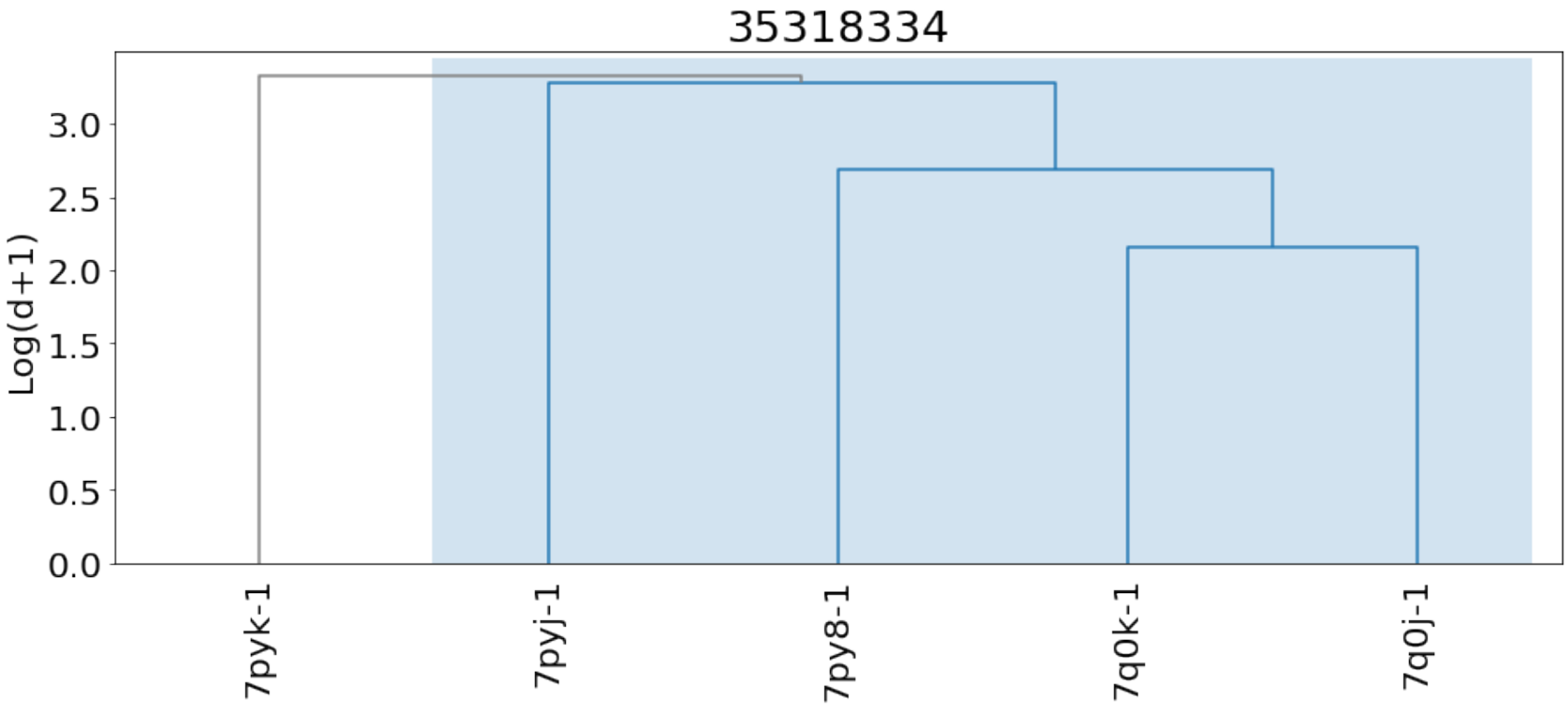


Figure 17: E. coli RNA polymerase elongation complex, hetero decamer. Ward's linkage.

Real Examples

ATP dependent folding chaperone

Science and Technology Facilities Council

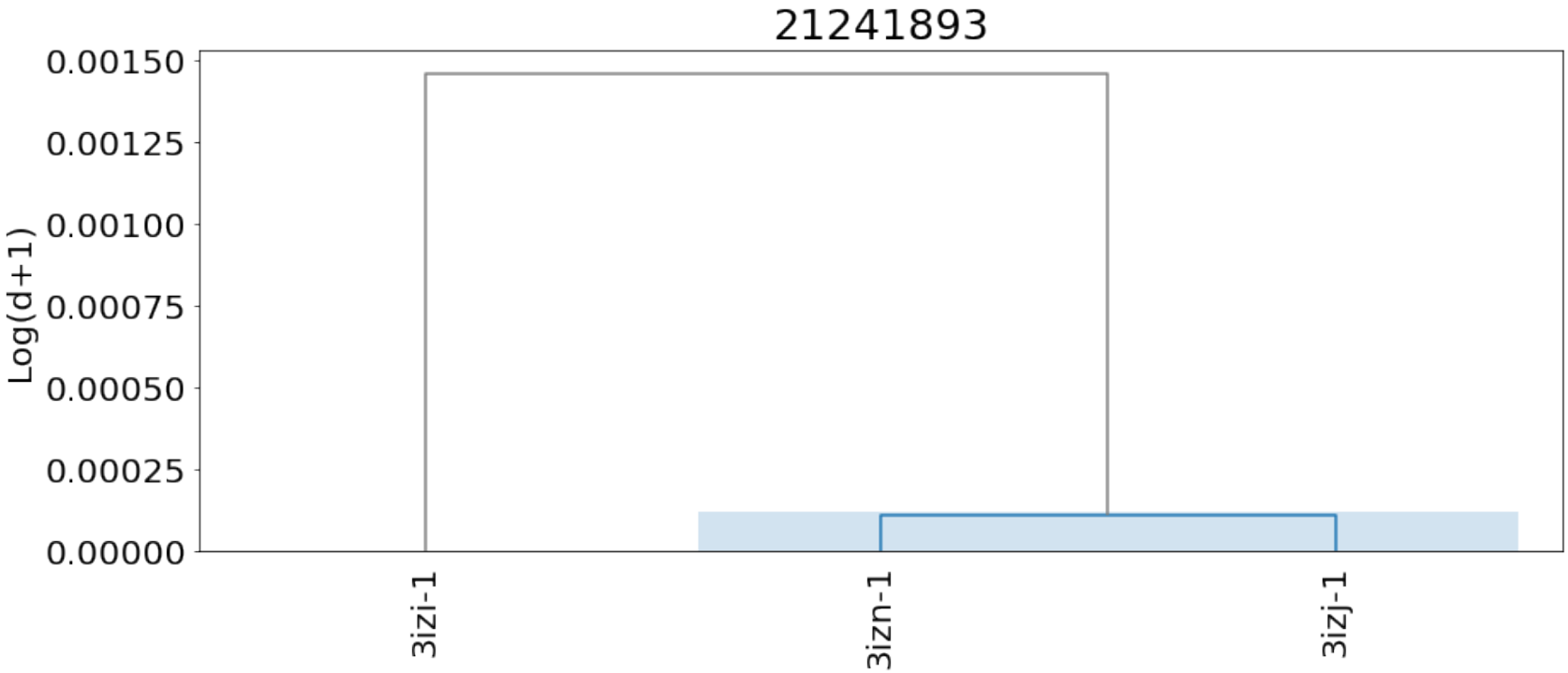


Figure 18: ATP-dependent protein folding chaperone, homo hexadecamer, non-strict mode. Ward's linkage.

Real Examples

Rabbit RyR1

Science and Technology Facilities Council

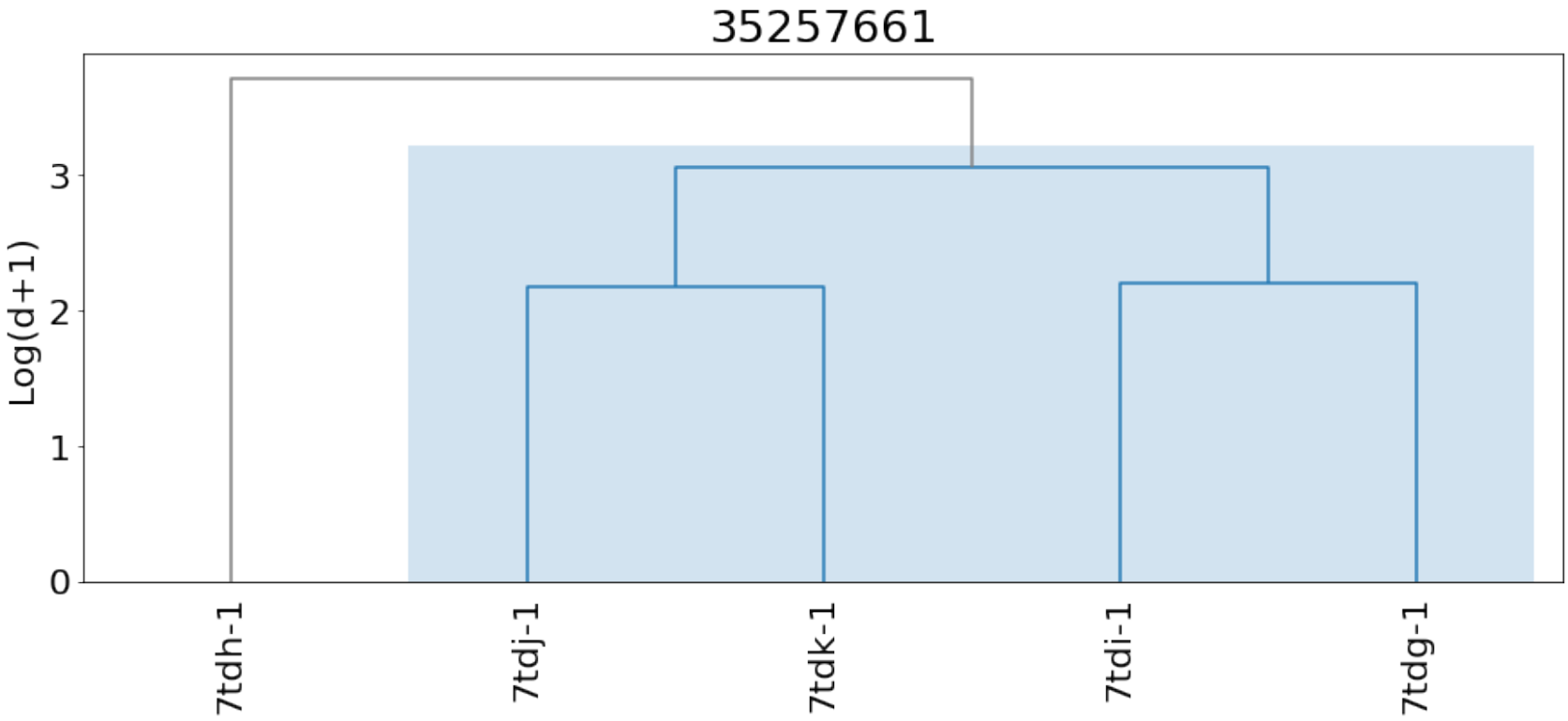


Figure 19: Rabbit RyR1, homo tetramer. Ward's linkage.

Where can I get it from, and what can I use it for?

Science and Technology Facilities Council

- Will make its way into CCP4.
- Experimental method design and validation, i.e. synthesis, crystallisation, isolation, preservation, measurement.
- Computational method design and validation, i.e. complex predictions, model building, ligand interaction modelling.
- Research integrity, do you really have what you say you have?

Acknowledgements

Science and Technology Facilities Council

- BBSRC grant BB/V015591/1.
- Eugene Krissinel (CCP4)
- Sameer Velankar (PDBe)
- Joseph Ellaway (PDBe)
- Sri Devan Appasamy (PDBe)
- ALGLIB [4] for hierarchical clustering.
- GEMMI [5] for file manipulation, sequence matching and structure superposition.

References I

Science and Technology Facilities Council

[1] Emmanuel D Levy, Jose B Pereira-Leal, Cyrus Chothia, and Sarah A Teichmann. 3d complex: a structural classification of protein complexes. *PLoS computational biology*, 2(11):e155, 2006.

[2] Eugene Krissinel and Ville Uski. Desktop and web-based gesamt software for fast and accurate structural queries in the pdb. *Journal of Computer Science Applications and Information Technology*, 2:1–7, 2017.

[3] Shihong Yue, Xiuxiu Wang, and Miaomiao Wei. Application of two-order difference to gap statistic. *Transactions of Tianjin University*, 14:217–221, 2008.

[4] Sergey Bochkanov. URL `https://www.alglib.net/`.

[5] Marcin Wojdyr. Gemmi: A library for structural biology. *Journal of Open Source Software*, 7(73):4200, 2022. doi: 10.21105/joss.04200. URL `https://doi.org/10.21105/joss.04200`.