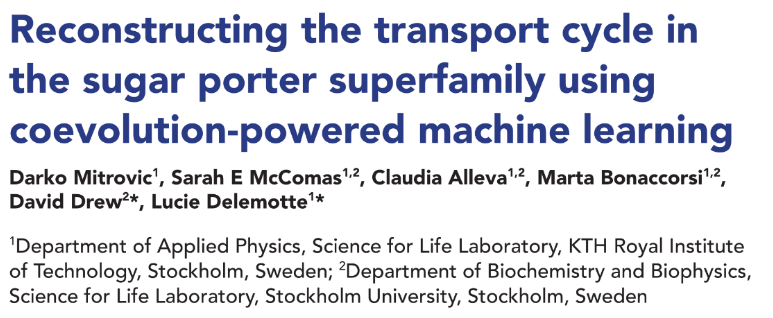
Short pres

Original paper

Reconstructing the transport cycle in the sugar porter superfamily using coevolution-powered machine learning



Reconstructing the transport cycle in the sugar porter superfamily using coevolution-powered machine learning

Darko Mitrovic[1], Sarah E McComas[1,2], Claudia Alleva[1,2], Marta Bonaccorsi[1,2], David Drew[2]*, Lucie Delemotte[1]*

[1]Department of Applied Physics, Science for Life Laboratory, KTH Royal Institute of Technology, Stockholm, Sweden; [2]Department of Biochemistry and Biophysics, Science for Life Laboratory, Stockholm University, Stockholm, Sweden

DOI: https://doi.org/10.7554/eLife.84805

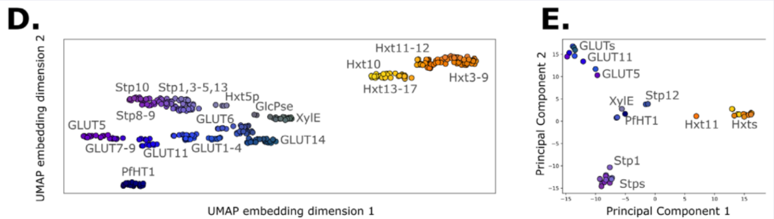
Short pres

Data

TODO

Problem

.

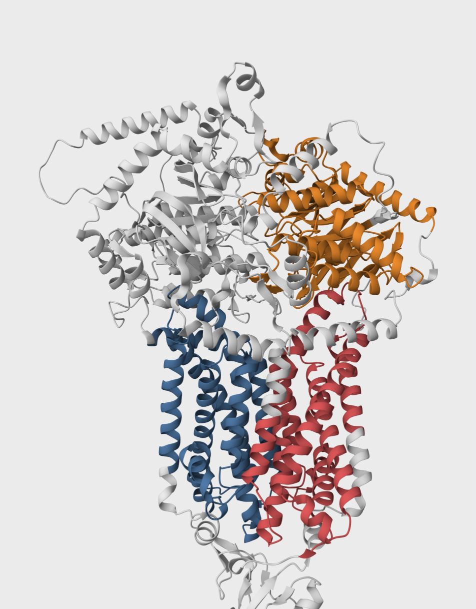


- Ⓓ UMAP of UniProtKB Sugar Porter sequences, using BLOSUM62-derived matrix for similarity scoring
- Ⓔ PCA projection of available AF2 Sugar Porters, using residue distance maps as features

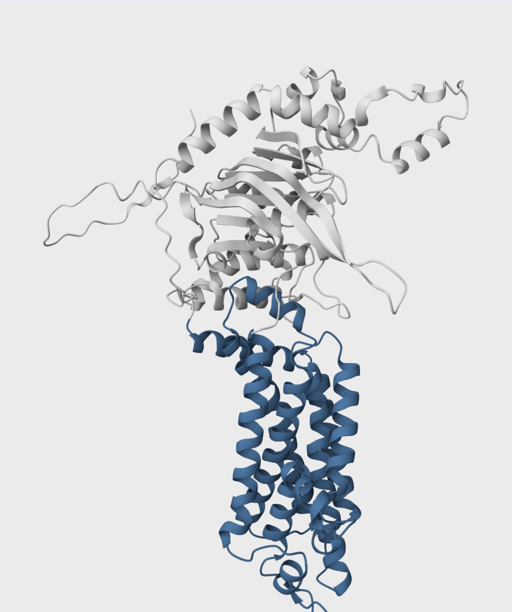
In both dots are coloured according to *phylogenetic proximity*

Data

ABCGs



ABCGs cont.



Experimental structures

prot	species	No. structs
ABCG8	HUMAN	8*
ABCG5	HUMAN	8*
ABCG2	HUMAN	41
ABCG1	HUMAN	9
PDR5	YEAST	4
AB25G	ARATH	15
AB16G	ARATH	4
CDR1	CANAL	3

Total No.

*All ABCG5/G8 in dataset are heterodimers. I found 69 distinct structures for 8 transporters.

Sequence library

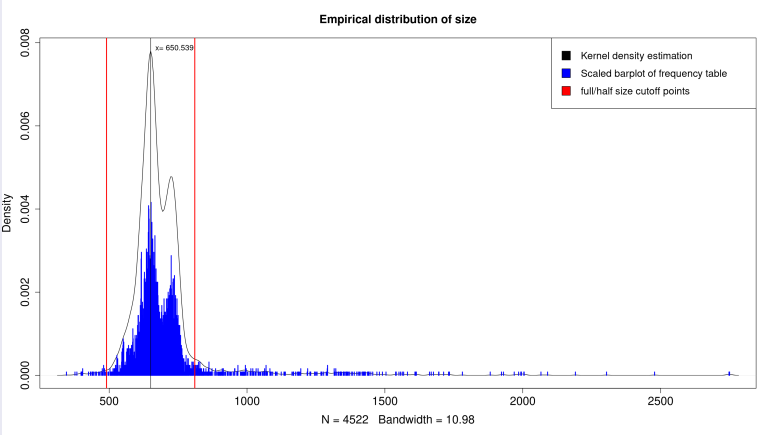
Querying NCBI Clustered-NR

I tried to use all distinct sequences from experimental structures, but only for **ABCG{1,2,8,16}** it was possible to run subsequent iterations of PSI-BLAST without exceeding NCBI CPU quota.
Yield: **4696**

Motif Filtering

All sequences were searched for PROSITE ABC motif on their web server.
Yield: **4522**

Size filtering



Vertical red bars represent the range taken as half-size ABCG transporter (roughly mean±2sd)

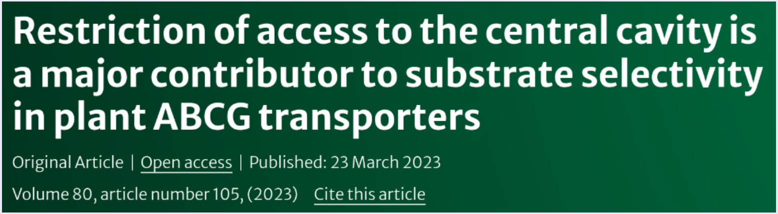
Yield: **4197**

Pseudo-full-size transporters

All half-size sequences selected from PSI-BLAST output were head-tail concatenated:

ACDKV → ACDKVACDKG

IChB Paper



Restriction of access to the central cavity is a major contributor to substrate selectivity in plant ABCG transporters

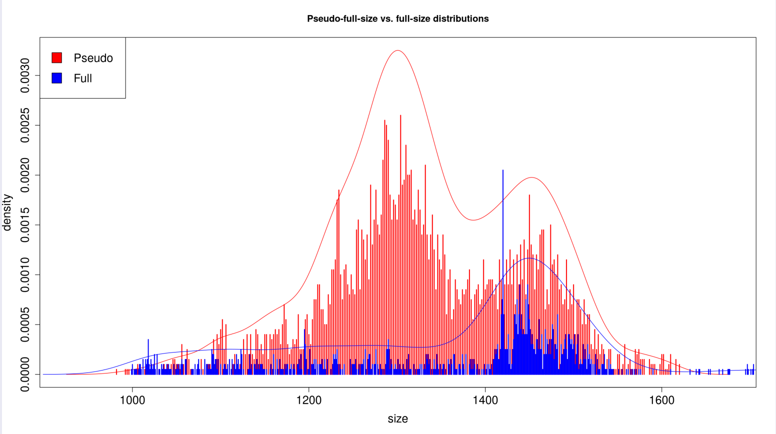
Original Article | Open access | Published: 23 March 2023

Volume 80, article number 105, (2023) Cite this article

Alphabeth filtering

I removed sequences which had `'?'` at any position, and `'*'` at any position other that the last one.
Yield: **1418**

Comparison of size distributions



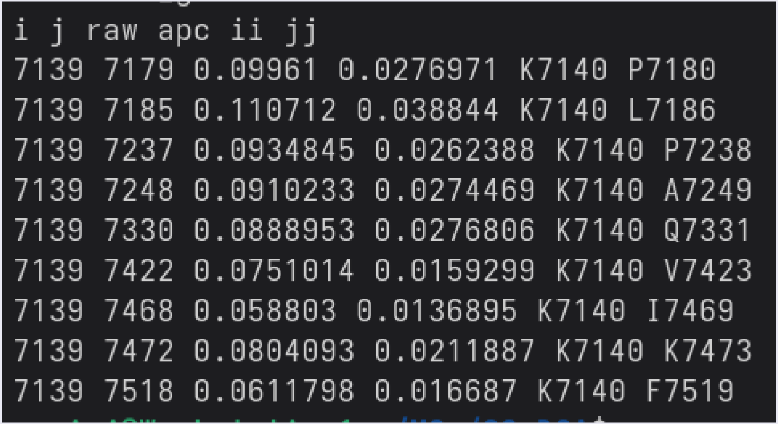
MSA

Input sequences

$4197 + 1418 + 8 = 5623$

DCA

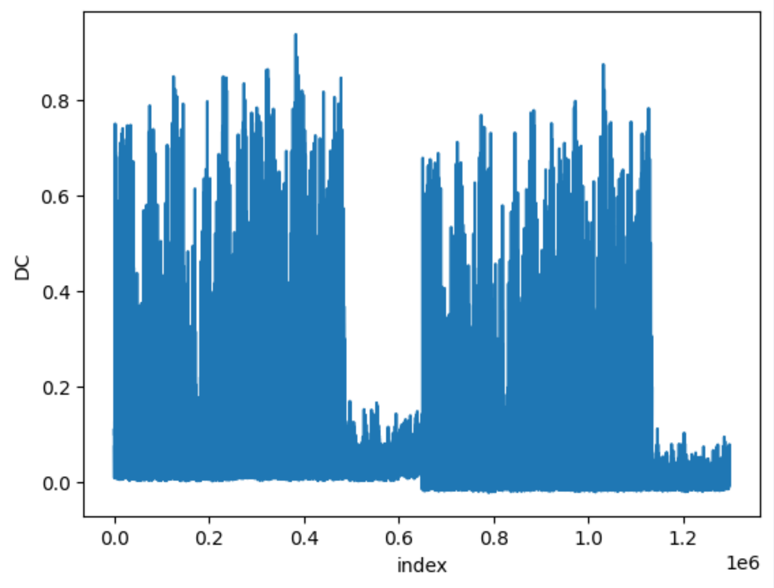
Output format



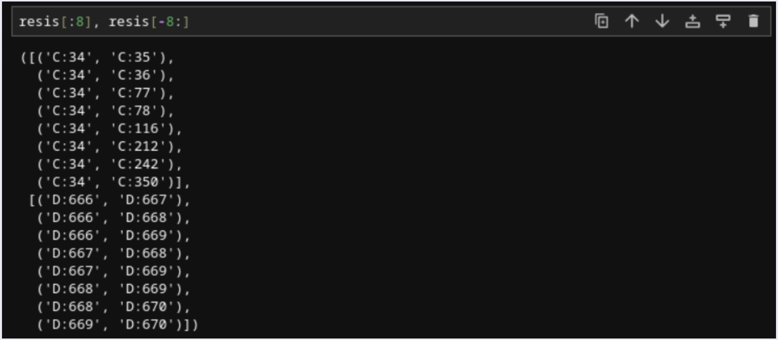
```
i j raw apc ii jj
7139 7179 0.09961 0.0276971 K7140 P7180
7139 7185 0.110712 0.038844 K7140 L7186
7139 7237 0.0934845 0.0262388 K7140 P7238
7139 7248 0.0910233 0.0274469 K7140 A7249
7139 7330 0.0888953 0.0276806 K7140 Q7331
7139 7422 0.0751014 0.0159299 K7140 V7423
7139 7468 0.058803 0.0136895 K7140 I7469
7139 7472 0.0804093 0.0211887 K7140 K7473
7139 7518 0.0611798 0.016687 K7140 F7519
```

- `raw`: l2norm of protein MRF (Markov Random Field (?))
- `apc`: `raw - mean(row) * mean(col) / mean(all)`

DCA raw vs apc



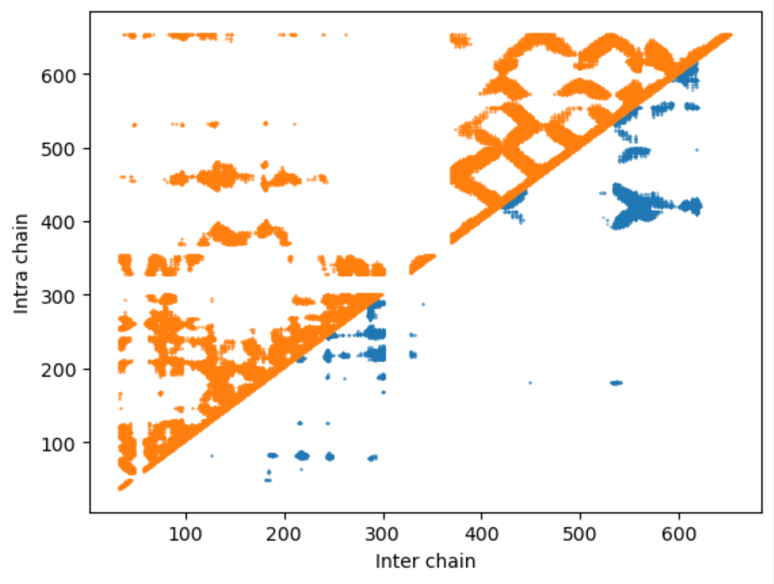
Contact maps



```
resis[:8], resis[-8:]
```

```
([('C:34', 'C:35'),
  ('C:34', 'C:36'),
  ('C:34', 'C:77'),
  ('C:34', 'C:78'),
  ('C:34', 'C:116'),
  ('C:34', 'C:212'),
  ('C:34', 'C:242'),
  ('C:34', 'C:350')],
 [('D:666', 'D:667'),
  ('D:666', 'D:668'),
  ('D:666', 'D:669'),
  ('D:667', 'D:668'),
  ('D:667', 'D:669'),
  ('D:668', 'D:669'),
  ('D:668', 'D:670'),
  ('D:669', 'D:670')])
```

Cont.


Intra chain
100
200
300
400
500
600
Inter chain
100
200
300
400
500
600

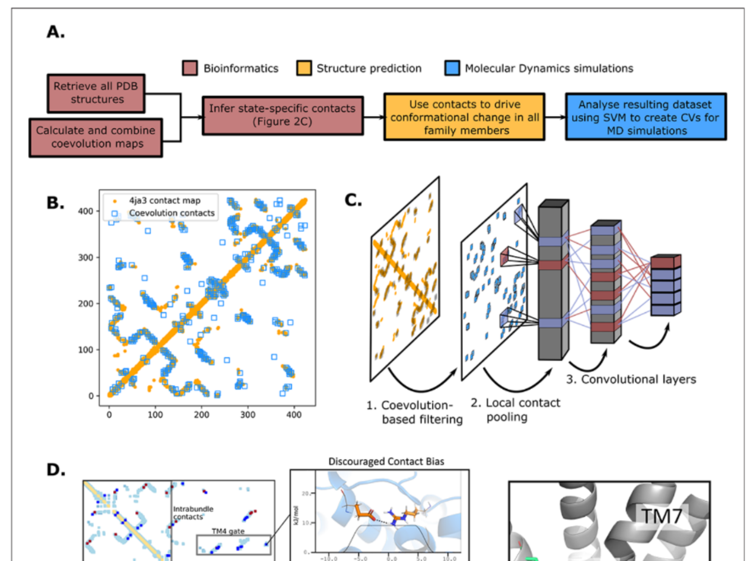
TODO

Clustering structures by state

.

1. MSA of 8 unique sequeneces for which we have any structure
2. Calculate contact/distance map for each of them
3. Select indices within contact/distance maps which correspond to MSA columns with 0 gaps
4. Which similarity measure can be used for clustering on this data?

Finding state specific contacts from DCA (with ML?)


A.
Bioinformatics
Structure prediction
Molecular Dynamics simulations
Retrieve all PDB structures
Calculate and combine coevolution maps
Infer state-specific contacts (Figure 2C)
Use contacts to drive conformational change in all family members
Analyse resulting dataset using SVM to create CVs for MD simulations
B.
4ja3 contact map
Coevolution contacts
0
100
200
300
400
C.
1. Coevolution-based filtering
2. Local contact pooling
3. Convolutional layers
D.
Intrabundle contacts
TM4 gate
Discouraged Contact Bias
kJ/mol
TM7

Cont.

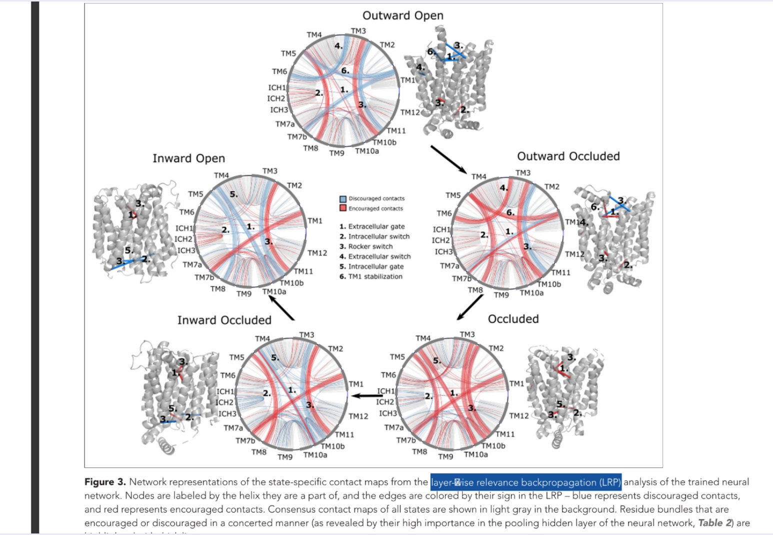


Figure 3. Network representations of the state-specific contact maps from the layer-wise relevance backpropagation (LRP) analysis of the trained neural network. Nodes are labeled by the helix they are a part of, and the edges are colored by their sign in the LRP – blue represents discouraged contacts, and red represents encouraged contacts. Consensus contact maps of all states are shown in light gray in the background. Residue bundles that are encouraged or discouraged in a concerted manner (as revealed by their high importance in the pooling hidden layer of the neural network, ***Table 2***) are