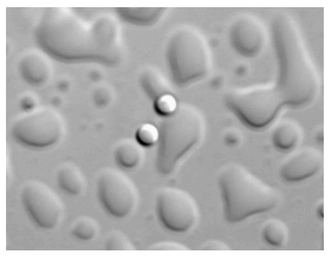
Multiscale biomolecular simulations – association rule mining in protein interactions

QUEST project from Antonia Niklasch, supervised by Lucia Baltz

What are protein granules?

- Dense, dynamic clusters of proteins that separate from the surrounding solution, forming a distinct membrane-less phase
- Relevant for a variety of biological processes, including epigenetic inheritance, signaling, stress response



Two protein granules. [1]

What are association rules?

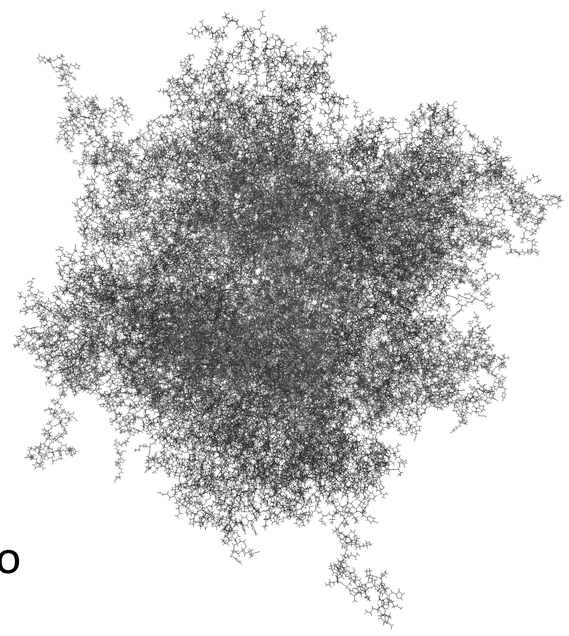
- Application in market analysis: If a customer already has bread and butter in their basket, is it more likely that they will add jam rather than detergent?
- Application in protein simulation: If amino acids A and B of any protein are currently in contact, how likely is it that amino acids C and D are as well?

Research Question

Can methods from market basket analysis be applied to molecular contact data to identify association rules that go beyond average-based contact analysis?

Investigated system: Tyr-Arg motif

- A system of MUT-8 and MUT-16 (two proteins) is simulated in a box.
- To increase specificity, only time steps in which a Tyr-Arg motif (interaction between three proteins) forms are considered.
- A distance based cutoff is introduced to define when two amino acids are considered to be in contact.



MUT-8 and MUT-16 in a box.

Method: Apriori Algorithm

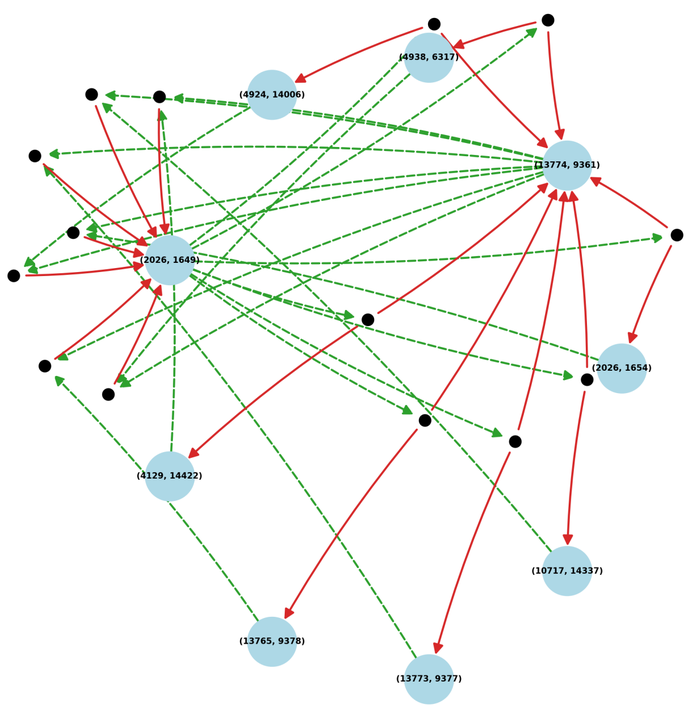
Apriori is an algorithm used to extract association rules from a dataset $D = \{L_1, L_2, \dots\}$, where $L_i = \{(X_1, X_2), (X_1, X_3), (X_2, X_2), \dots\}$ are the lists of contact tuples at different time steps.

Extract frequent itemsets: Sets of contacts are selected when they exceed a predefined support (fraction of all time steps in which it appears).

↓

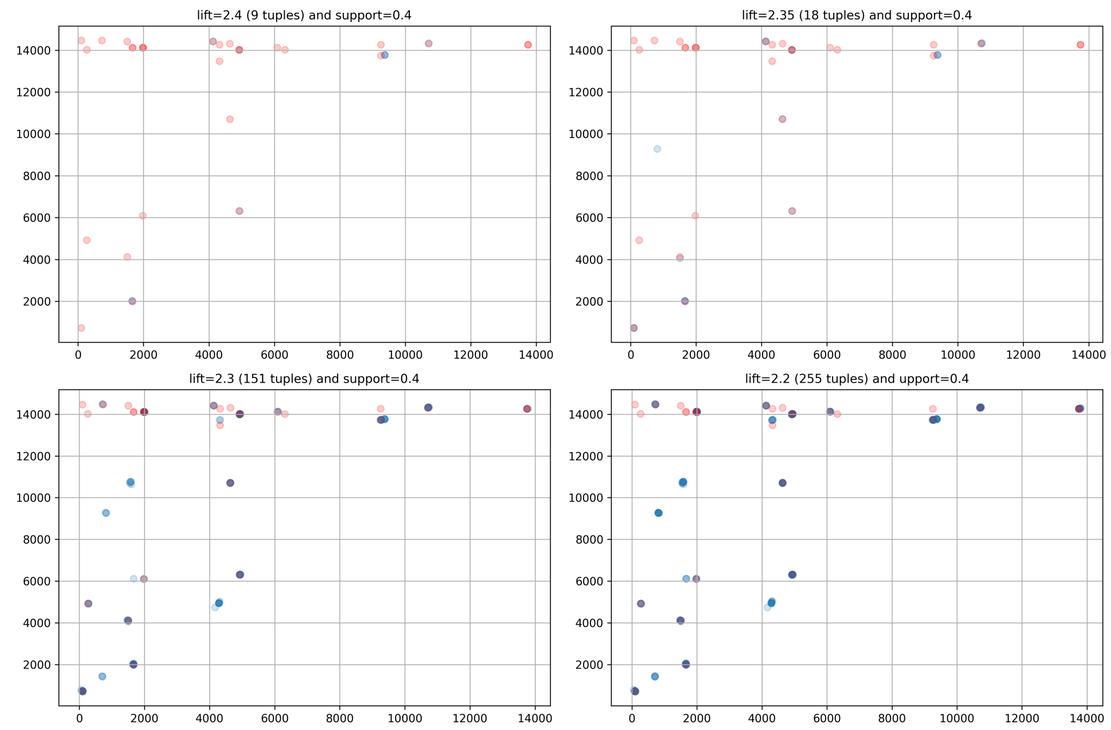
Form association rules: From each frequent items, rules of the form $A \to B$, with disjoint frequent itemsets A and B are generated and assessed by their lift $= \frac{P(A \cap B)}{P(A)P(B)}$ (which quantifies deviation for statistical independence between A and B).

Results



Left: Visual representation of association rules with lift $= 2.4$ andsupport $= 0.4$; an arrow pointing to another arrow indicates two contacts in either the cause (green outgoing) or the effect (red incoming) of a given rule.

Right: Contact tuples occurring in association rules with given lift and support (blue) and Tyr-Arg motif contacts (red).



Conclusion

- The Apriori algorithm can identify association rules from molecular contact data.
- For a conditioned contact dataset of a Tyr-Arg motif formation –as expected– the resulting rules with the strongest lift include the tuples of the motif.
- Limitations: many similarly well-supported rules were found; the number decreases when applying a stricter distance-based cutoff.

Outlook

- Biochemical interpretation can assess emergence of specific rules and contact tuples.
- Method extensible to: residue grouping, amino-acid type analysis, cross-time-step analysis (e.g., one contact influencing the likelihood of another in subsequent steps).
- Includes time as a dimension into contact analysis, especially valuable for systems with low data correlation.

Contact aniklasc@students.uni-mainz.de

Sources [1] Gaurav, K., Busetto, V., Páez-Moscoso, D. J., *et al.*, *Biophysical Journal* (2025).

QUEST