Predicting protein complexes:

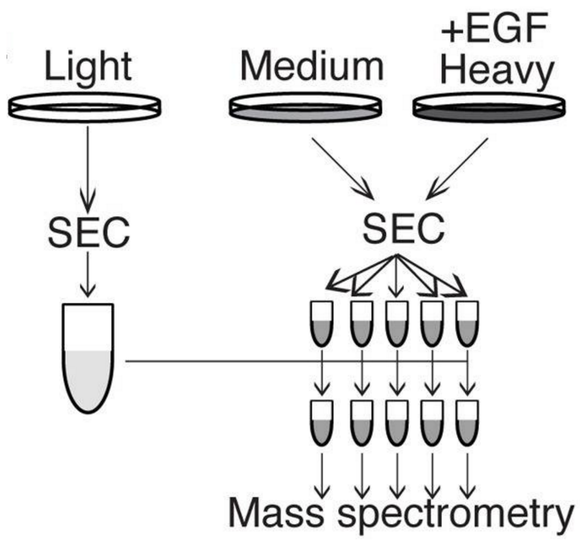
an improved pipeline for co-fractionation data

R. Greg Stacey,[1] Nichollas E. Scott,[1,2] Leonard J. Foster [1]

1 Michael Smith Laboratories, University of British Columbia 2 Doherty Institute, University of Melbourne

Introduction

- Protein complexes are ubiquitous and central to the function of cellular life.
- One method to identifiy protein complexes is protein correlation profiling (PCP) [1].
- In PCP, complexes are identified as proteins that "travel together" across a separation gradient.
- SILAC labeling [2] allows **conditional experiments** to be analyzed simultaneously
- The bioinformatics pipeline below uses PCP-SILAC data to 1) detect changes in protein levels between conditons, 2) predict interactions between protein pairs, and 3) predict protein complexes across conditions.



1 Fold changes

The pipeline detects chromatogram peaks and reports the relative protein amounts between experimental conditions. Changes in protein amounts between conditions ("fold changes") are reported, and statistical significance is assessed using hypothesis testing (t-test, rank-sum test) across replicates.

Feature extraction

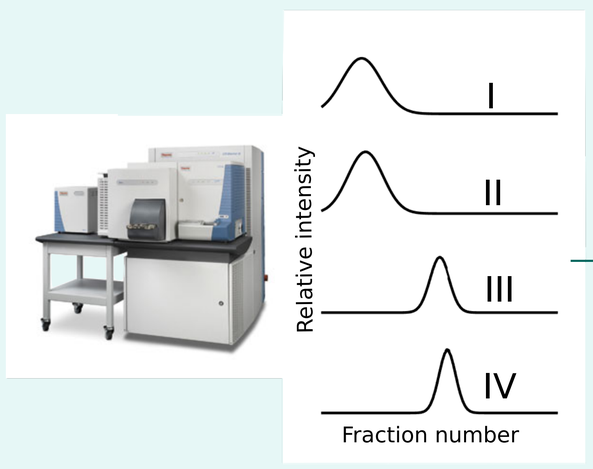


Fig. 1: Raw chromatograms are measured by SEC-MS. Multiple conditions can be run simultaneously.

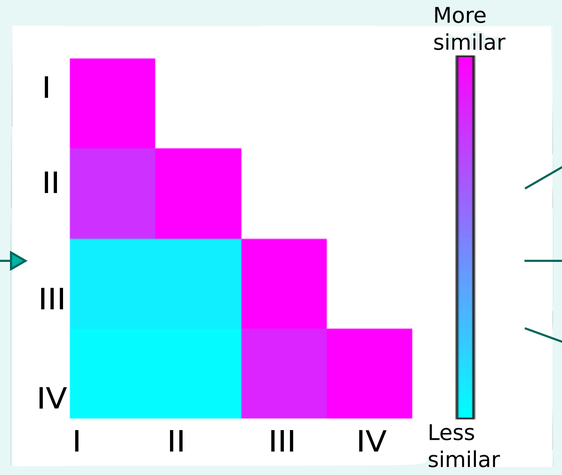
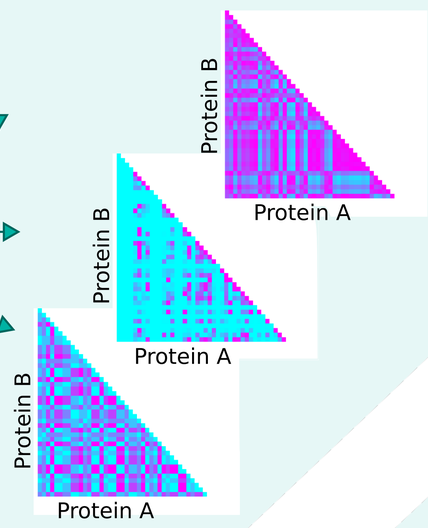


Fig. 2: A "similarity matrix" for chromatogram pairs.

Fig. 3: Multiple similarity matrices are the "extracted features".



Protein interactions

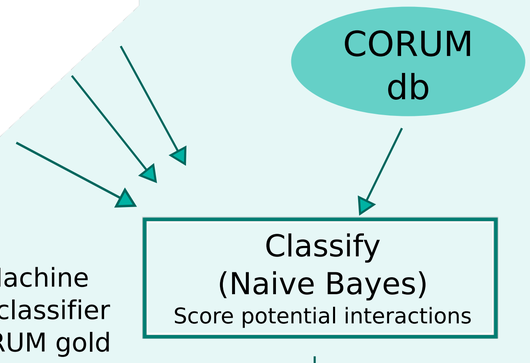
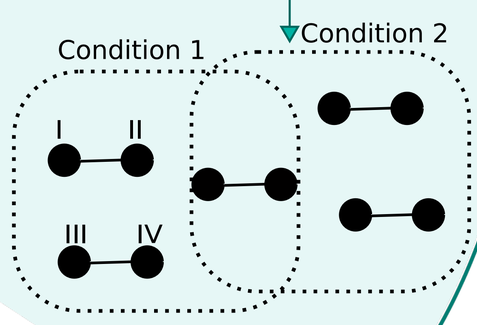


Fig. 4: Machine learning classifier uses CORUM gold standards to score all pairs of chromatograms.

Fig. 5: Pairs with score > threshold are predicted interactions.

	Protein A	Protein B	Score
	P62277	P62753	0.999
	E9PQ61	Q96PK6	0.998
	J3KTC1	Q96PK6	0.998
	O43290	Q96PK6	0.989
	P46013	Q96PK6	0.971
	Q5BKZ1	Q96PK6	0.952
	Q96I24	Q96PK6	0.937
	P23396	P25787	0.912
Interacting	P62753	P62829	0.853
NOT Interacting	P49720	Q9Y230	0.739
	P60866	Q9Y230	0.643
	P62753	Q9Y230	0.631
	E9PEZ3	P35244	0.582
	B4DS61	Q96SI9	0.531



3 Protein complexes

- Complexes are predicted using pairwise interactions and ClusterONE.
- Parameters are optimized to produce "CORUM-like" complexes.
- Complexes are predicted for each experimental condition, allowing the direct comparison of complexes between different treatments,

Fig. 7: Protein complexes are predicted for each experimental condition.

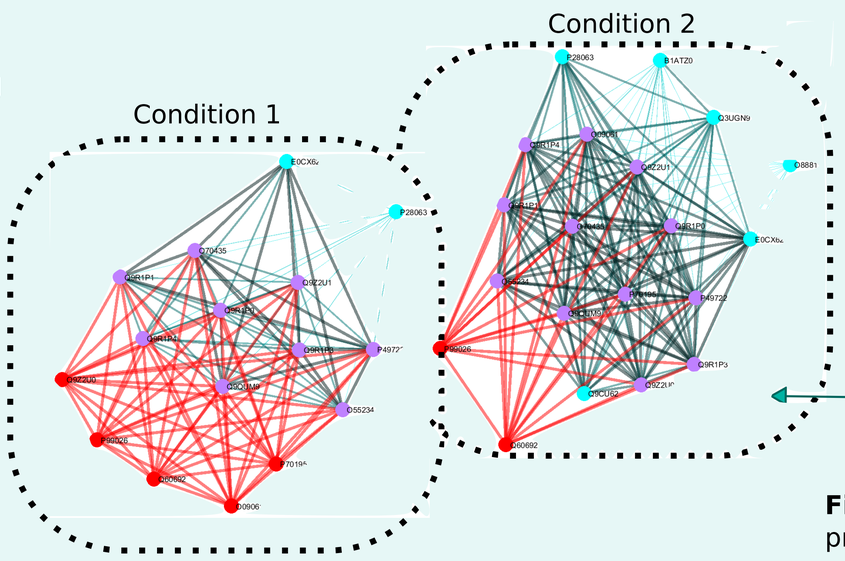


Fig. 6: The ClusterONE algorithm predicts complexes from PPIs.

Conclusions

- The bioinformatics pipeline above predicts protein interactions and complexes in conditional co-fractionation experiments.
- The CORUM database is used as a gold standard reference, although any database of complexes could be used, e.g. STRING.
- False positive interactions are minimized by controlling the precision, TP / (TP + FP), of the final interaction list.

References

[1] Andersen et al. Proteomic characterization of the human centrosome by protein correlationproling. Nature, 426(6966):570-574, 2003.
[2] Kristensen et al. A high-throughput approach for measuring temporal changes in the interactome. Nature methods, 9(9):907-909, 2012.
[3] Scott et al. Developmentof a computational frame-work for the analysis of protein correlation proling and spatialproteomics experiments. Journal of proteomics, 118:112{129, 2015.

2 Pairwise interactions

- Protein-protein interactions are predicted by a machine learning classifier.
- Each pair of chromatograms is compared to CORUM gold standards.
- A Naive Bayes classifier assigns an interaction score to each chromatogram pair.
- Finally, a threshold is applied to all scores to determine interacting pairs.

richard.greg.stacey@gmail.com