Chalabi *et al. BMC Bioinformatics* (2019) 20:17
https://doi.org/10.1186/s12859-018-2573-8

BMC Bioinformatics

SOFTWARE Open Access

CoExpresso: assess the quantitative behavior of protein complexes in human cells

Morteza H. Chalabi[1], Vasileios Tsiamis[1], Lukas Käll[2], Fabio Vandin[3] and Veit Schwämmle[1*]

Abstract

Background: Translational and post-translational control mechanisms in the cell result in widely observable differences between measured gene transcription and protein abundances. Herein, protein complexes are among the most tightly controlled entities by selective degradation of their individual proteins. They furthermore act as control hubs that regulate highly important processes in the cell and exhibit a high functional diversity due to their ability to change their composition and their structure. Better understanding and prediction of these functional states demands methods for the characterization of complex composition, behavior, and abundance across multiple cell states. Mass spectrometry provides an unbiased approach to directly determine protein abundances across different cell populations and thus to profile a comprehensive abundance map of proteins.

Results: We provide a tool to investigate the behavior of protein subunits in known complexes by comparing their abundance profiles across up to 140 cell types available in ProteomicsDB. Thorough assessment of different randomization methods and statistical scoring algorithms allows determining the significance of concurrent profiles within a complex, therefore providing insights into the conservation of their composition across human cell types as well as the identification of intrinsic structures in complex behavior to determine which proteins orchestrate complex function. This analysis can be extended to investigate common profiles within arbitrary protein groups. CoExpresso can be accessed through http://computproteomics.bmb.sdu.dk/Apps/CoExpresso.

Conclusions: With the CoExpresso web service, we offer a potent scoring scheme to assess proteins for their co-regulation and thereby offer insight into their potential for forming functional groups like protein complexes.

Keywords: Protein complex, Statistics, Co-regulation

Background

Biological systems are governed by a multitude of entangled interactions between biomolecules with an immense number of physical and chemical properties. Protein complexes are large biomolecules with a wide range of tasks in the cell and consist of multiple subunits linked by non-covalent interactions. These interactions can lead to a variety of stable or transient states where the complexes display different compositions of their subunits or different structures that are often fine-tuned by post-translational modifications. An example of functional diversity are ribosomes that are known to contribute differentially to translation of distinct subpopulations of mRNAs [1]. There is a pressing need to investigate complex capabilities for regulatory control of cellular processes. To achieve this, a detailed map of protein complex composition, abundance, and behavior in different cell types and tissues is required. Such a map will considerably improve the characterization and the prediction of the functional states.

Various experimental methods exist to identify protein complexes and to determine and quantify which protein subunits they are composed of. Determination of protein interaction partners within a complex provides valuable knowledge about complex and protein function and thus their potential behavior [2]. Most prominent experimental methods to determine protein-protein interactions are based on the yeast-2-hybrid protocol or the application of affinity purification coupled with mass spectrometry [3, 4]. These methods however suffer from

*Correspondence: veits@bmb.sdu.dk
[1] Department of Biochemistry and Molecular Biology and VILLUM Center for Bioanalytical Sciences, University of Southern Denmark, Campusvej 55, 5230 Odense M, Denmark
Full list of author information is available at the end of the article

either large false identification rates or depend on purification steps that often lead to a strong bias in the results. More details about protein structure can be achieved by chemical cross-linking or hydrogen-deuterium exchange mass spectrometry [5]. Despite the power of these methods, they cannot yet be applied on entire proteomes. For an accurate, large-scale and general characterization, protein complex behavior should be studied across large numbers of samples without perturbations towards e.g. subgroups of proteins and additionally rely on highly confident identification of the proteins.

There is an increasing amount of evidence supporting the hypothesis that the majority of protein complexes are tightly controlled in the cell. Post-transcriptional regulation occurs predominantly for protein complex members, leading to strong co-regulation of complex subunits. This could be shown by systematic investigation of protein and gene expression levels in human cancer [6, 7], in a study comparing 11 cell types and 4 temporal states [8], based on the co-occurrence of protein pairs across human experiments in the PRIDE database [9], or generally in a selection of proteomics data sets [10]. In summary, these studies showed that only a fraction of complex composition and abundance is regulated at transcriptional level and therefore other mechanisms such as protein degradation contribute to protein complex stoichiometry. This highlights the power of directly measuring protein abundance profiles by common proteomics approaches such as bottom-up mass spectrometry to thoroughly study protein complexes and their variants across cell types and states.

In contrast to most proteomics data repositories where only raw data and identification results are available, ProteomicsDB [11, 12] is a large compendium of quantitative protein abundances, therefore highly useful to investigate general patterns of protein changes across more than 100 different human cell lines.

Here, we apply three scoring models on the ProteomicsDB data to assess the significance of subunit co-regulation in protein complexes. We compare and benchmark different randomization and scoring approaches on known complexes and reveal particular substructures of complex behavior for a few selected use cases. The scoring and extensive visualization is implemented in the web service CoExpresso that allows investigating co-regulatory patterns in any group of human proteins.

Implementation

Quantifications of proteins and IDs of known complexes were downloaded from ProteomicsDB [11, 12] and CORUM [13], respectively. We used three randomization approaches that differently resemble data structure within all protein abundance profiles. Scores were calculated for the co-regulation of proteins in a complex applying three different models for the comparison of protein profiles. The scores were stored in a database. For each protein in each complex, significance for their co-regulation was calculated and assessed on basis of the scores. A web service was implemented to allow interrogating the score database to test arbitrary protein groups for the significance of their co-regulation. Figure 1 provides an overview of the workflow and the web interface.

Data retrieval

Quantitative abundance profiles of SwissProt proteins were extracted from ProteomicsDB hosting mass spectrometry based protein abundances for distinct human cell types including cell tissues, cell lines and fluids. In ProteomicsDB, proteins and samples are annotated according to UniProtKB and Brenda Tissue Ontology [14], respectively.

From the downloaded profiles (summer 2016), we retained only cell types with more than 1000 proteins and which were tagged by Brenda ontology terms. Proteins not available in at least 2 cell types were removed. This reduced the data to comprise 15,409 proteins and 140 tissues. Uniprot accession numbers for annotated human complexes were downloaded from CORUM and filtered for duplicates, leading to a total of 2175 reported complex compositions.

Complex abundance profiles

For each protein group C, only the n_t cell types with full coverage, $t = [1..n_t]$, i.e. having quantitative values for all proteins $p = [1..n_p]$, were considered, resulting in a n_t by n_p matrix $E_C(t, p)$.

Randomization techniques

We applied 3 different forms of randomization to obtain random protein groups being quantified in the same number of cell types as the proteins of protein group C. The often relatively low coverage of proteins over multiple cell types required creating randomized sets for each combination of number of cell types and number of proteins.

Independent sampling (IS): Randomization of quantitative values of all proteins in all tissues comprised sets with the same dimensions as the to be tested protein group. That is, the n_t by n_p randomized values were obtained by sampling, independently at random, $n_t n_p$ values from all quantitative values of all proteins in all tissues. 10,000 random groups were created for each combination of n_t and n_p.

Protein-centered sampling (PCS): Randomization of proteins and categorization into cell type coverage. This randomization type turned out to be more complex and a sufficient large coverage of random groups was achieved

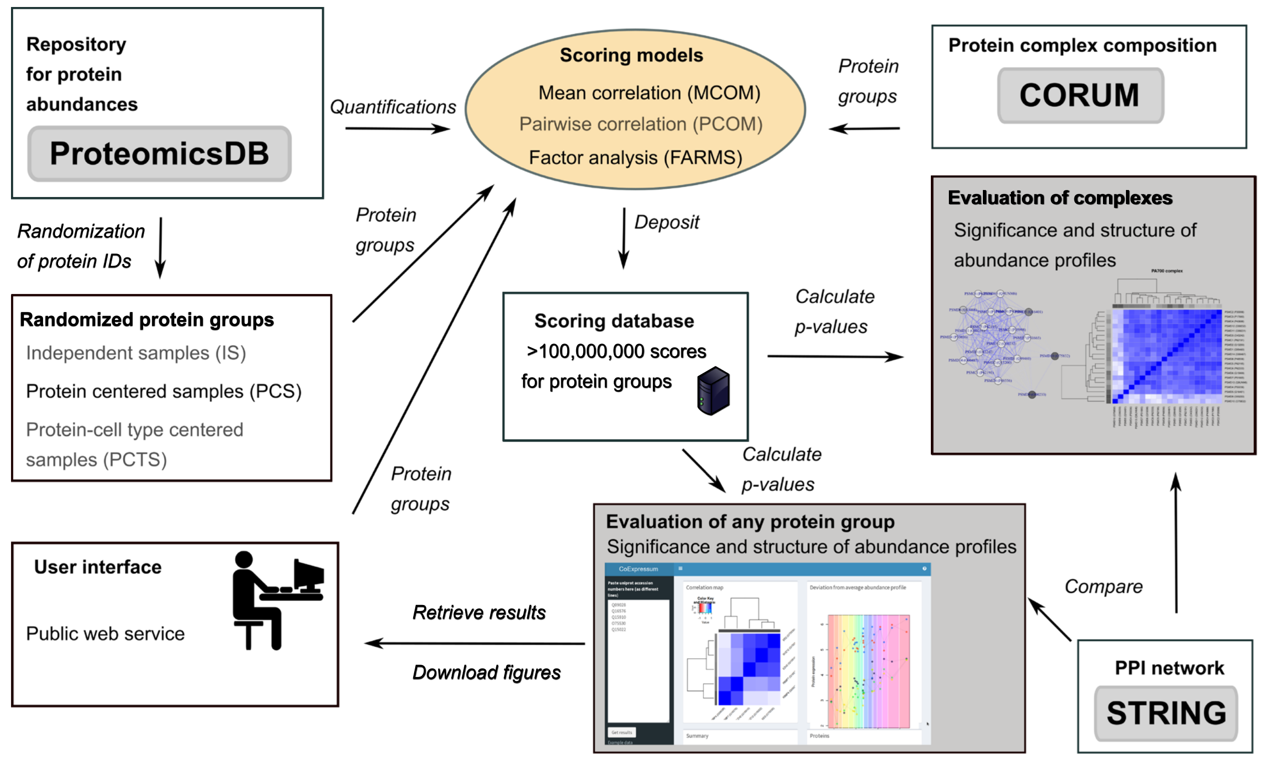

Fig. 1 Schema of entire workflow to investigate protein complex behavior. Models and randomization methods that were not used in the final assessment of CORUM complexes are shown in grey. For a more detailed description of the workflow, see Methods

by the following procedure. For each combination of number of proteins n_p and number of cell types n_t:

1 Take all proteins being each quantified in at least n_t cell types
2 Repeat the following 5000 times: sample n_p proteins IDs and count full cell type coverage of the protein group
3 Keep unique protein combinations with coverage over at least five cell types

With this procedure, we obtained 1000–20,000 unique and random protein groups for each relevant combination giving a total of more than 20,000,000 randomized groups. Obtaining random protein groups with low coverage was computationally most demanding. Our method is scalable with respect to data coverage and will also perform within a similar time frame when increasing the number of considered cell types.

Protein- and tissue-centered sampling (PTCS): All proteins simultaneously found in the same cell types as the tested protein group were randomized to create 10,000 samples. That is, n_p proteins are sampled independently at random from all proteins that appear in the same cell types as the tested protein group, and their observed values in those cell types are considered.

Similarity models and scoring

Mean correlation model (MCOM): Protein abundances were averaged for each cell type, restricting to cell types covered by the entire protein group, $M(t) =< E_C(t,p) >_p$. For each protein p, Pearson's correlation to the means $M(t)$ provides a measure of how much the protein follows the common profile of the protein group, $S_{\mathrm{MCOM}}(p) = \mathrm{cor}(M(t), E_C(t,p))$, where $\mathrm{cor}(x(t), y(t))$ denotes Pearson's correlation between samples x and y.

Pairwise correlation model (PCOM): Pearson's correlation was calculated between all proteins pairs using the abundances in the cell types covered by all proteins. The score is then given by the sum,

$$S_{\mathrm{PCOM}}(p) = \sum_{p,q=1;q\neq p}^{n_p} \mathrm{cor}\left(E_C(t,p), E_C(t,q)\right)$$

Factor analysis model (FAMS): The model is based on factor analysis developed for microarray analysis [15] and recently modified to improve protein inference in bottom-up mass spectrometry data [16]. The following parameters were used: Weight $w = 0.1$, $\mu = 0.1$, 1000 maximal iterations and a minimal noise of 0.0001. The feature weights W were used to score each protein of a group: $S_{\mathrm{FARMS}}(p) = W(p)$.

Scores for protein groups: Overall scores per protein group were generated by simply averaging the scores of the individual proteins, $\hat{S} = \sum_{p=1}^{n_p} S_{\mathrm{MODEL}}(p)/n_p$, where MODEL stands for either MCOM, PCOM or FAMS.

Scoring statistics

For each model, randomization method, and a given combination of n_t and n_p, the aforementioned scores were calculated for randomized protein groups, and stored in a database. These scores, more than 100,000,000 in total, were then used to calculate the probabilities to reject the null hypothesis (of observing the score for a set of n_p proteins over n_t tissues) for both a single protein p and a group of proteins:

$$p_{\mathrm{MODEL}}(p) = \frac{N\left[S_{\mathrm{MODEL}}\left(p^{(\mathrm{random})}\right) > S_{\mathrm{MODEL}}(p)\right] + 1}{N\left[S_{\mathrm{MODEL}}\left(p^{(\mathrm{random})}\right)\right] + 1}$$

and

$$p_{\mathrm{MODEL}} = \frac{N\left[\hat{S}^{(\mathrm{random})}_{\mathrm{MODEL}} > \hat{S}_{\mathrm{MODEL}}\right] + 1}{N\left[\hat{S}^{(\mathrm{random})}_{\mathrm{MODEL}}\right] + 1},$$

where $p^{(\mathrm{random})}$ denotes a protein from a randomized protein group, $\hat{S}^{(\mathrm{random})}_{\mathrm{MODEL}}$ a score for a randomized protein group, and $N[\ldots]$ counting the number of all valid cases within the brackets.

For p-values from multiple protein groups, correction for multiple testing was carried out via the Benjamini-Hochberg procedure.

Results

Tight regulation of protein complexes by translational and post-translational control mechanisms may result in the degradation of more abundant proteins that do not form the complex. Then the proteins of known complexes, such as the ones collected in the CORUM database, will show similar abundance protein profiles when compared across different cell types. Given that proteins often have multiple functions, a protein complex might present itself in different compositions or the complex does not change in abundance, we did not assume all complexes to show highly similar abundance profiles of their proteins but merely investigated how much co-regulation can be observed.

We applied different scoring systems to evaluate whether proteins in human complexes exhibit similar regulatory behavior when compared over multiple cell types. Despite of having a large set of available protein abundances, coverage of the proteins over the 140 cell types was often sparse (Additional file 1: Figure S1), requiring scoring methods that account for missingness. In such a scenario, just calculating the similarity between protein abundance profiles, e.g. by calculating Pearson's correlation, will not provide statistically valid measures for their co-regulatory behavior. For instance, low coverage over cell types leads automatically to higher correlations between protein abundance profiles than for higher coverage (Additional file 1: Figure S2). Hence, confidence estimations of protein co-regulation require adapting the scoring methods to include effects coming from data coverage. This can be achieved by empirically calculating p-values from comparison of the score of a protein to the scores obtained from appropriately randomized data. Therefore, we investigated different ways of randomizing the ProteomicsDB data to identify the best performing combination of scoring scheme and randomization procedure.

Table 1 summarizes the used methods and randomizations. In short, MCOM compares each protein profile versus the averaged profile of protein group, allowing to assess how much a protein follows this common trend. PCOM is based on pairwise comparisons and summarizes them by their sum. This method was implemented to consider internal structures of protein subgroups with high correlations. The FAM model is based on factor analysis and calculates weights for each protein, giving a measure of how much each protein contributes to the profile of the entire protein group.

For the following analysis, each protein complex reported in CORUM was assessed for coverage in ProteomicsDB and further evaluated by the different models when all protein subunits were available in at least 5 cell

Table 1 Summary of scoring models and randomization methods

Model	Abbrv.	Output
Mean correlation	MCOM	Similarity to averaged abundance profile
Pairwise correlation	PCOM	Sum of pairwise similarities
Factor analysis	FAM	Weights for protein contribution to full set
Randomization	Abbrv.	Basis
Independent sampling	IS	Mix all values
Protein-centered sampling	PCS	Keep protein profiles
Protein- and cell type-centered sampling	PTCS	Keep protein and cell type profiles

types. We tested a total of 1414 protein groups out of 2157 annotated in CORUM.

Scoring models

Empirical confidence estimation of co-regulatory behavior was carried out by representing the null distribution (i.e. cases of no co-regulation) by scores obtained from randomizations. By comparing the scores of the different models to scores from randomly sampled data, we obtained probabilities to discard the observed abundance profiles as result of randomly chosen proteins. Thus the false discovery rates (FDRs), represented by *p*-values corrected for multiple testing, provide a measure for significance of a given complex on basis of co-regulation of its subunits within human cell types. The different randomization techniques were applied to resemble the intrinsic data structure on different scales.

Figure 2a compares the *p*-values calculated for each model and randomization. More "realistic" randomization (IS< *PCS* <PTCS) resulted in lower number of complexes with significant abundance profiles. MCOM and PCOM, both models being based on Pearson's correlation, produced nearly the same results on complex level (see also Additional file 1: Figure S3). The FAMS approach however performed differently, reaching a higher number of significant complexes for the protein-centered randomization

On protein level (Fig. 2b), lower protein numbers with significant abundance profiles could be expected and were observed when using randomization methods that maintain protein and cell type properties. Here, PCOM displays a higher number of proteins than FAMS and MCOM for low false discovery rates.

Robustness

Recovery of proteins and complexes with significant abundance profiles does however not ensure robustness of the methods towards noise. As example, one could expect a protein complex to contain subunits that do not follow the general trend of the abundance profiles. This could be due to wrong assignment of a protein to a complex or due to different behavior of a subunit being heavily regulated by e.g. post-translational modifications or by forming transients regulating complex function.

Method robustness in handling differentially abundant proteins can be simulated by adding randomly chosen proteins to the CORUM complexes. In all complexes, we increased the number of proteins by 50%, 75% and 100%. Figure 3 shows ROC curves for these simulated complexes, where we compared the significance by counting true (actual complex subunits) and false positives (added proteins). Here, the different methods and randomization approaches showed consistent differences for their robustness. Randomization of the entire ProteomicsDB data lead to lower robustness for all methods. One the other hand, protein-centered (PCS) and protein-cell type centered (PTCS) randomization gave nearly identical performance results. Hence, the following analysis will focus on PCS randomization, although being the computationally mosts expensive one, as it yields higher counts of significant proteins. In addition, MCOM and FAM models had lower false positives rates at least in the lower range.

Use cases

The following use cases will provide detailed results of the scoring models and general complex behavior for three selected complexes that are representative for the investigated complexes. We obtained 60 CORUM complexes

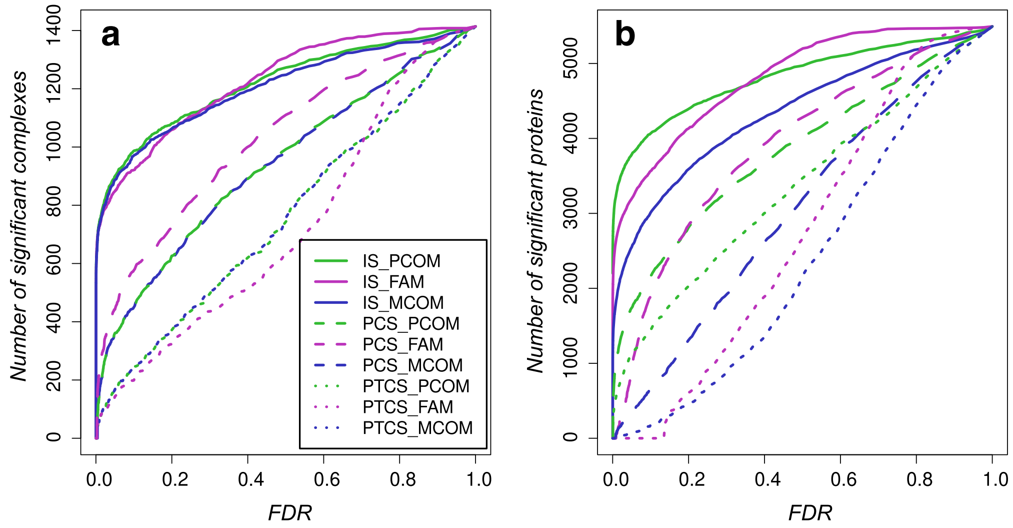

Fig. 2 Comparison of models for significant co-regulations. Number of complexes (**a**) and proteins (**b**) with significant abundance profiles according to the different scoring models and randomizations calculated for different thresholds for their false discovery rate (FDR)

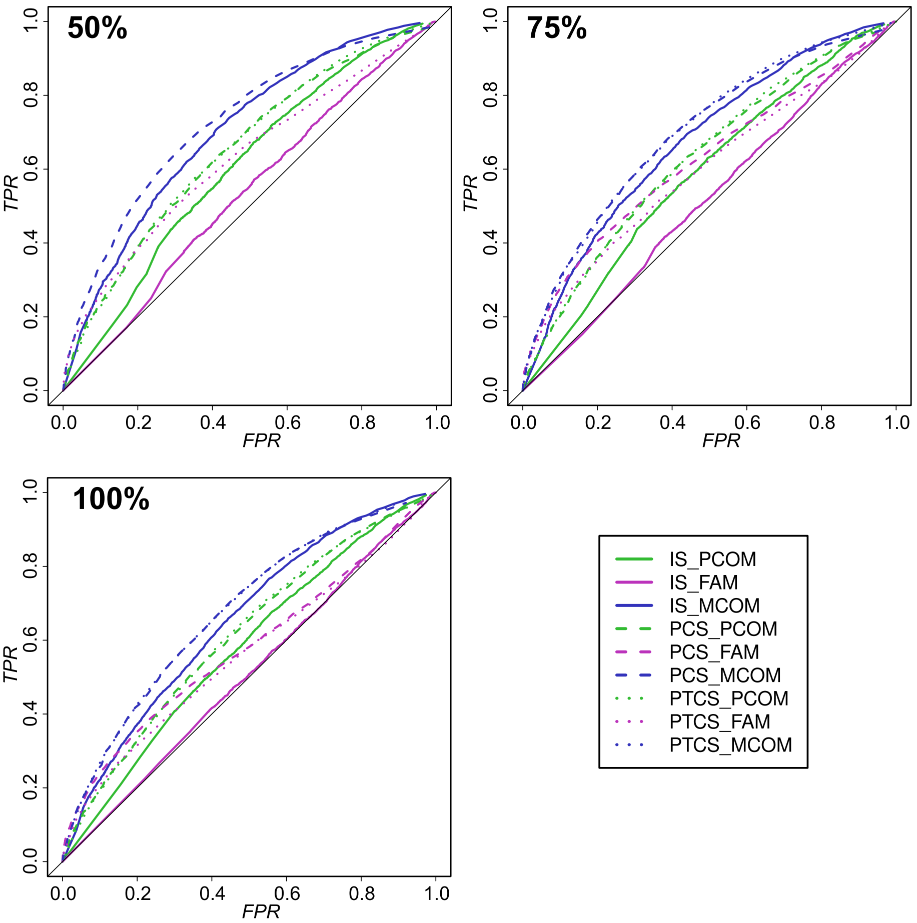

Fig. 3 Performance of scoring models measured by robustness to 50%, 75% and 100% artificially added random proteins. Proteins were categorized into complex subunits and random proteins. True positive and false positives rates (TPR and FPR) were given by the fraction of true positives and false positives at a given FDR threshold. MCOM and FAM models lead to better performance. Only slight difference between PCS and PTCS randomizations can be observed

with lowest FDR values (< 0.0003) for all three scoring models.

Use case A: Condensin I (Fig. 4a) represented the first of the complexes with lowest FDR values in all models (PCS randomization). All five proteins were commonly expressed in 75 cell types. Very high correlations between all proteins confirmed the high interaction evidence from STRINGdb [17]. However, Condensin subunit 2 (NCAPH) showed slightly lower correlation and lower scores. Indeed, NCAPH is known as regulatory subunit of Condensin I with different nucleolar localization during interphase [18]. We observed different abundance levels of NCAPH in several cell types leading to lower weight by factor analysis (Additional file 1: Figures S4 and S5A). Tissues with 2-fold lower abundance levels (compared to the mean of all proteins of the complex) were blood platelet and lung while 2-fold higher abundance levels where measured for lymph nodes and several cancer cell lines.

Use case B: 28S mitochondrial ribosomal subunit (Fig. 4b), being essential for ATP production, represents the complex with lowest FDR in all models and most proteins. The 30 proteins were commonly available in 23 cell types. Both our visualization and STRINGdb interactions suggest a more open structure or composition of the complex with a core component of heavily co-regulated proteins. The correlation map (upper figure) roughly distinguishes two slightly overlapping large subgroups (proteins MRPS22-MRPS2 and MRPS15-MRPS12) with higher correlations amongst their proteins. We found a strikingly different behavior of these groups in lung tissue (Additional file 1: Figure S5B). This suggests that the 28S ribosomal complex plays a different role in lung where it might break up into two functional units.

We looked into more examples of subgroups with highly co-regulated abundances. Higher coverage over cell types for these subgroups allows gathering further insight into their co-regulation. MRPS17, MRPS36 and MRPS12 show

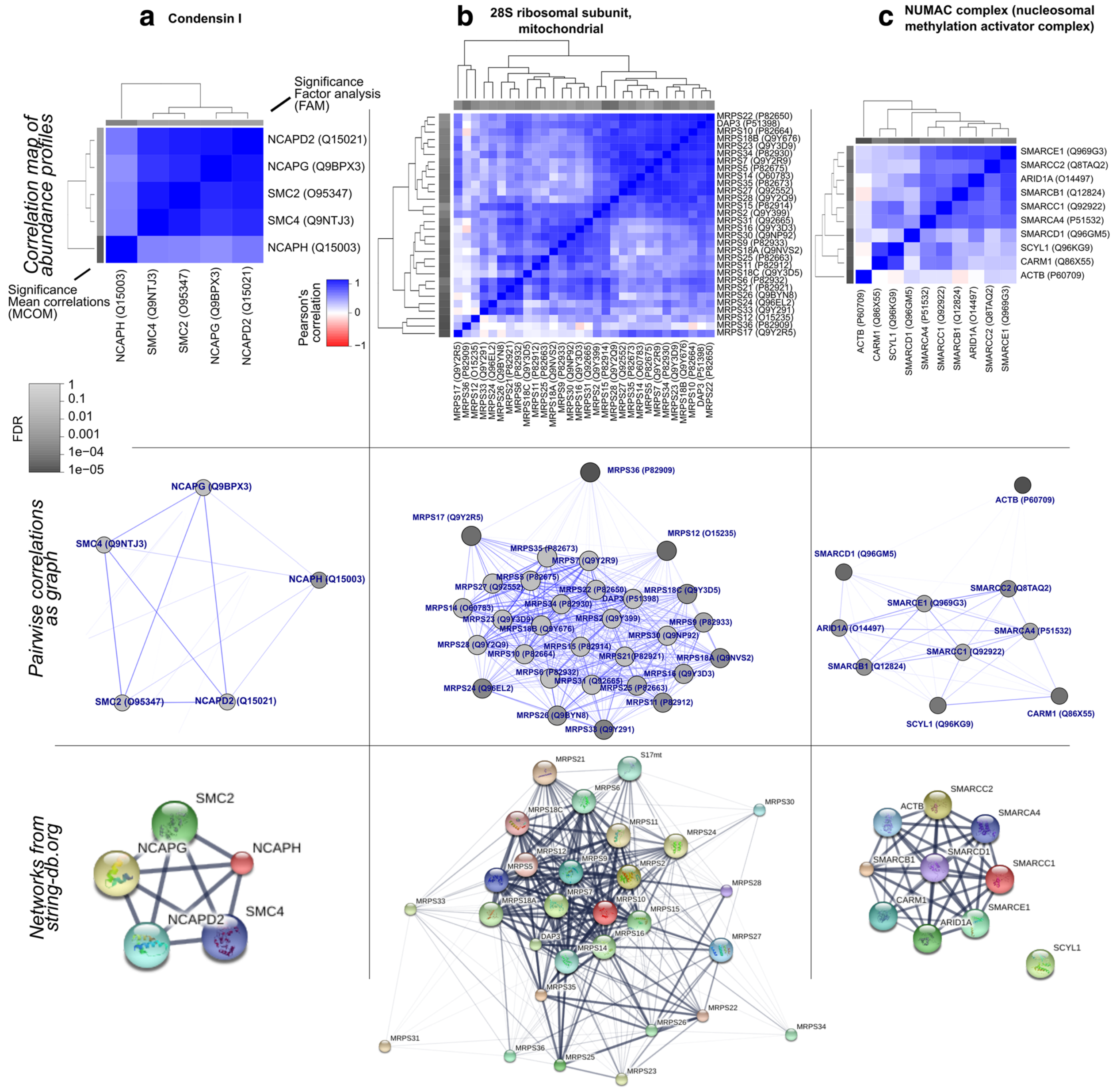

Fig. 4 Examples for complexes with highly co-regulated proteins. Upper panels: hierarchical clustering of pairwise correlations between protein abundance profiles. The sidebars show the significance of MCOM and FAM models (PCS randomization). Middle panels: Network visualization of profile similarities. Edge widths correspond to pairwise correlations. Grey tones of the proteins depict FDR significance calculated by PCOM (PCS randomization). Lower panels: STRINGdb (version 10) networks of proteins. Edge width is given by interaction confidence

very low correlation (Additional file 1: Figure S6A) and this is confirmed when estimating their significance over 48 cell types (individual proteins $p > 0.05$ and all overall scores $p > 0.1$). MRPS12 and MRPS12 are known to be altered in many and different cancer types [19], which could explain their particularly different behavior.

A group of the five proteins MRPS21, MRPS24, MRPS26, MRPS6 and MRPS33 exhibited highest correlations and reasonably high significance. We investigated their co-regulation as a protein group on their own where their abundance profiles were available in a higher number of 34 cell types (Additional file 1: Figure S6B) and confirmed highly significant co-regulation. A literature search did not identify any functional behavior for this protein subgroup.

Moreover, the correlation map of 28S mitochondrial ribosomal subunit exihibits a large subgroup of proteins (DAP3, MRPS2, MRPS5, MRPS7, MRPS10, MRPS14,MRPS15, MRPS18B, MRPS22, MRPS23, MRPS27, MRPS28, MRPS34 and MRPS35) with high correlations (Additional file 1: Figure S6C). When investigating these proteins as subgroup, all proteins but MRPS15 showed high significance for co-regulation. All of them were consistently lower abundant in lung tissue when compared to the other proteins of the complex. This confirms that this subgroup might play a particular role in lung tissue.

Use case C: NUMAC complex (nucleosomal methylation activator complex, Fig. 4) denotes a case with slightly lower significance. All scoring models suggest high significance with an FDR below 0.5%. The 10 proteins were found in 33 cell types with ACTB distinct behavior and drastically higher abundance than the other proteins. Strong evidence for interactions of all components but SCYL1 in STRINGdb suggests that ACTB plays a crucial role in complex composition but might still have other functions in the cell. We assume that this protein is not actively degraded when not forming the complex. All 3 models agreed in having high FDR values for ACTB and SMARCD1 (FDR >0.1), suggesting that the latter plays a particular role in this complex.

A data source for tightly co-regulated proteins

Given the strong co-regulation in annotated protein complexes, we asked whether our randomly sampled protein groups with highly significant co-regulation could determine novel but yet not well characterized complex compositions in human cells. Random protein groups with the highest scores did however not provide evidence for these proteins to be arranged as complexes but showed an increase in protein interactions. We calculated network enrichment scores in STRINGdb for the top scoring 100 protein groups and found the majority to be consistently higher than for randomly chosen protein groups (Additional file 1: Figures S7-S8). This means that highly significant protein groups do potentially have particular common biological functions such as co-regulation on transcriptional level or being represented by common members of a known or unknown pathway. We implemented CoExpresso that interrogates groups of human proteins to assess their co-regulation strength. Therefore, our CoExpresso web service can be highly useful for the interested researcher to test their hypothesis on the basis of human cell types in general. Figures and statistical measures can be obtained for any list of (mixed) human protein accession numbers and gene names given that there is sufficient data coverage in ProteomicsDB.

Discussion

The literature provides at least hundreds proteomics experiments per year from which a large percentage have their raw data deposited in the major data repositories (e.g. PRIDE nearly reaching 10,000 projects to date [20]). Availability of protein abundances is however still very rare also because the comparison of protein abundance across experiments and projects is still a major bottleneck in the proteomics field. ProteomicsDB provides a large catalogue of protein abundances in human cell types which we used to thoroughly investigate protein complex behavior. Despite the large number of characterized cell types, data coverage is rather low, where more than 20% of the proteins were detected in only 2–5 cell types. Such low coverage hindered straightforward application of e.g. simple correlation and we therefore compared a variety of different scoring models and randomizations that reproduce the inherent data structure.

Our comparison showed that appropriate randomizations are crucial to achieve results with simultaneously high recall and considerable robustness to noise. The results speak against complete randomization of all values, where global differences amongst cell types and proteins are neglected. We found that protein identities (PCS method) needed to be maintained to reach robust results. On the other hand, maintaining the identity of the tissues (PCTS method) in the investigated protein group did not lead to lower robustness. We therefore conclude that testing properties of protein profiles in general should be compared to a randomized set where protein identity is kept. In data with many missing values, this randomization requires categorizing the random protein groups into their tissue coverage which can be computationally expensive. We therefore provide a web service that stores the randomizations and where arbitrary protein groups can be tested for their significance. By testing annotated complexes from the CORUM database for the significance of their concurrent protein abundance profiles, we could confirm almost 50% (500–600 depending on scoring model) of the protein groups being co-regulated with an FDR below 0.1. This confirms the tight regulation of complex proteins previously reported and extends this observation to be valid generally in human cells. Given the lack of coverage over sufficient cell types in many cases, resulting in rather low statistical power, we predict that most protein complexes will be found to be translationally and post-translationally regulated. While most insight into dysregulation of complex subunits comes from gene expression data, our tool allows extending the analysis by determination and comparison of complex behavior on protein level. Instead of analyzing and comparing protein behavior alone, our user-friendly tool characterizes protein changes with respect to the complex or a in general to a protein group. Thus we provide direct insight into the functional behavior of a protein group.

Our analysis additionally confirmed and extended details about protein complex substructure that indicates regulatory features that orchestrate complex function by changes in complex composition or by here not investigated post-translational modifications.

We furthermore tested whether the large database of randomized protein groups could be used to identify novel protein assemblies that represent highly interacting functional modules such as complexes. We did not find enrichment for known protein-protein interactions in the most significant protein groups. This means that investigating protein co-regulation by random sampling alone is not a good source to search for novel complexes but remains highly valuable to test for complex behavior and confirm their composition across cell types. Given the combinatorial explosion when considering the number of possible protein groupings, the random sampling strategy used here considers only a small fraction of all protein groups that contain highly co-regulated proteins. Novel protein assemblies could still be found by selective and iterative algorithms that determine protein groups with highest co-regulation within all possible combinations.

Conclusion

The here presented study provides deep insight into protein complex behavior in human cells. The data for all 1414 investigated protein groups can be accessed via the CoExpresso web service. Arbitrary protein groups can be tested for their significance with respect to their co-regulation in human cell, such as investigating prior hypotheses about protein groups with common strongly co-regulated functional behavior. With more data on hand, we expect to improve statistical power and accuracy by including more data sets and by characterizing the role of quantified post-translational modifications.

Availability and requirements

Project name: CoExpresso
Project home page: http://computproteomics.bmb.sdu.dk/Apps/CoExpresso and https://bitbucket.org/veitveit/coexpresso for source code and R scripts.
Operating system(s): Platform independent (web service)
Programming language: R and javascript
Other requirements: We recommend a modern web browser (e.g. Firefox or Chrome)
License: Apache 2.0

Additional file

Additional file 1: Supplementary Figures to CoExpresso: Assess the quantitative behavior of protein complexes in human cells. (PDF 5647 kb)

Abbreviations

FAM: Factor analysis model; FDR: False discovery rate; IS: Independent sampling; MCOM: Mean correlation model; PCOM: Pairwise correlation model; PCS: Protein-centered sampling; PTCS: Protein- and tissue-centered sampling

Funding

MCH acknowledges support from the University of Southern Denmark. VS acknowledges support from The Danish Council for Independent Research, the Danish Danish National Research Foundation (DNRF82) and the EU ELIXIR consortium (Danish ELIXIR node). This work is supported, in part, by University of Padova project SID 2017.

Authors' contributions

FV and VS designed the work. MCH and VS collected, analyzed and interpreted the data. VT and VS implemented the software. FV, OK and VS drafted and revised the article. All authors gave final approval of the version to be published.

Ethics approval and consent to participate

Not applicable.

Consent for publication

Not applicable.

Competing interests

The authors declare that they have no competing interests.

Publisher's Note

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Author details

[1]Department of Biochemistry and Molecular Biology and VILLUM Center for Bioanalytical Sciences, University of Southern Denmark, Campusvej 55, 5230 Odense M, Denmark. [2]KTH - Science for Life Laboratory, School of Biotechnology, Royal Institute of Technology, Solna, Sweden. [3]Department of Information Engineering, University of Padua, Padua, Italy.

Received: 13 August 2018 Accepted: 10 December 2018
Published online: 09 January 2019

References

1. Gilbert WV. Functional specialization of ribosomes? Trends Biochem Sci. 2011;36:127–32. https://doi.org/doi:10.1016/j.tibs.2010.12.002..
2. Bauer A, Kuster B. Affinity purification-mass spectrometry. Powerful tools for the characterization of protein complexes. Eur J Biochem. 2003;270: 570–8.
3. Gingras AC, Gstaiger M, Raught B, Aebersold R. Analysis of protein complexes using mass spectrometry. Nat Rev Mol Cell Biol. 2007;8: 645–54. https://doi.org/doi:10.1038/nrm2208.
4. Musso GA, Zhang Z, Emili A. Experimental and computational procedures for the assessment of protein complexes on a genome-wide scale. Chem Rev. 2007;107:3585–600. https://doi.org/doi:10.1021/cr0682857.
5. Zhang Z, Vachet RW. Kinetics of Protein Complex Dissociation Studied by Hydrogen/Deuterium Exchange and Mass Spectrometry. Anal Chem. 2015;87:11777–83. https://doi.org/doi:10.1021/acs.analchem.5b03123.
6. Gonçalves E, Fragoulis A, Garcia-Alonso L, Cramer T, Saez-Rodriguez J, Beltrao P. Widespread Post-transcriptional Attenuation of Genomic Copy-Number Variation in Cancer. Cell Syst. 2017;5:386–98e4. https://doi.org/doi:10.1016/j.cels.2017.08.013.
7. Ryan CJ, Kennedy S, Bajrami I, Matallanas D, Lord CJ. A Compendium of Co-regulated Protein Complexes in Breast Cancer Reveals Collateral Loss Events. Cell Syst. 2017;5:399–409e5. https://doi.org/doi:10.1016/j.cels.2017.09.011.
8. Ori A, Iskar M, Buczak K, Kastritis P, Parca L, Andrés-Pons A, et al. Spatiotemporal variation of mammalian protein complex stoichiometries. Genome Biol. 2016;17:47. https://doi.org/doi:10.1186/s13059-016-0912-5.
9. Gupta S, Verheggen K, Tavernier J, Martens L. Unbiased Protein Association Study on the Public Human Proteome Reveals Biological

Connections between Co-Occurring Protein Pairs. J Proteome Res. 2017;16:2204–12. https://doi.org/doi:10.1021/acs.jproteome.6b01066.

10. Rogowska-Wrzesinska A, Wrzesinski K, Fey SJ. Heteromer score-using internal standards to assess the quality of proteomic data. Proteomics. 2014;14:1042–7. https://doi.org/doi:10.1002/pmic.201300457.
11. Wilhelm M, Schlegl J, Hahne H, Gholami AM, Lieberenz M, Savitski MM, et al. Mass-spectrometry-based draft of the human proteome. Nature. 2014;509:582–7. https://doi.org/doi:10.1038/nature13319.
12. Schmidt T, Samaras P, Frejno M, Gessulat S, Barnert M, Kienegger H, et al. ProteomicsDB. Nucleic Acids Res. 2018;46:D1271–81. https://doi.org/doi:10.1093/nar/gkx1029.
13. Ruepp A, Waegele B, Lechner M, Brauner B, Dunger-Kaltenbach I, Fobo G, et al. CORUM: the comprehensive resource of mammalian protein complexes–2009. Nucleic Acids Res. 2010;38(Database issue): D497–501. https://doi.org/doi:10.1093/nar/gkp914.
14. Gremse M, Chang A, Schomburg I, Grote A, Scheer M, Ebeling C, et al. The BRENDA Tissue Ontology (BTO): the first all-integrating ontology of all organisms for enzyme sources. Nucleic Acids Res. 2011;39:D507–13. https://doi.org/doi:10.1093/nar/gkq968.
15. Hochreiter S, Clevert DA, Obermayer K. A new summarization method for Affymetrix probe level data. Bioinformatics (Oxford, England). 2006;22: 943–9. https://doi.org/doi:10.1093/bioinformatics/btl033.
16. Zhang B, Pirmoradian M, Zubarev R, Käll L. Covariation of Peptide Abundances Accurately Reflects Protein Concentration Differences. Mol Cell Proteomics MCP. 2017;16:936–48. https://doi.org/doi:10.1074/mcp.O117.067728.
17. Szklarczyk D, Morris JH, Cook H, Kuhn M, Wyder S, Simonovic M, et al. The STRING database in 2017: quality-controlled protein-protein association networks, made broadly accessible. Nucleic Acids Res. 2017;45:D362–8. https://doi.org/doi:10.1093/nar/gkw937.
18. Cabello OA, Eliseeva E, He WG, Youssoufian H, Plon SE, Brinkley BR, et al. Cell cycle-dependent expression and nucleolar localization of hCAP-H. Mol Biol Cell. 2001;12:3527–37. https://doi.org/doi:10.1091/mbc.12.11.3527.
19. Gopisetty G, Thangarajan R. Mammalian mitochondrial ribosomal small subunit (MRPS) genes: A putative role in human disease. Gene. 2016;589: 27–35. https://doi.org/doi:10.1016/j.gene.2016.05.008.
20. Vizcaíno JA, Csordas A, del Toro N, Dianes JA, Griss J, Lavidas I, et al. 2016 update of the PRIDE database and its related tools. Nucleic Acids Res. 2016;44:D447–56. https://doi.org/doi:10.1093/nar/gkv1145.