Cluster analysis and visualization techniques for large datasets in complexome profiling

Heiko Giese[1], Jörg Ackermann[1], Ulrich Brandt[2], Ilka Wittig[3], and Ina Koch[1]

[1]**Molecular Bioinformatics, Institute of Computer Science, Faculty of Computer Science and Mathematics, Cluster of Excellence Frankfurt "Macromolecular Complexes", Johann Wolfgang Goethe-University Frankfurt am Main, Robert-Mayer-Str. 11-15, 60325 Frankfurt am Main, Germany**
[2]**Nijmegen Centre for Mitochondrial Disorders, Radboud University, Nijmegen Medical Centre, Geert Grooteplein Zuid 10, NL-6500 Nijmegen, The Netherlands**
[3]**Functional Proteomics, SFB815 core unit, Medical School, Johann Wolfgang Goethe-University Frankfurt am Main, Theodor-Stern-Kai 7, 60590 Frankfurt am Main, Germany**

ABSTRACT

Dysfunctional protein complexes are often associated with diseases. To develop effective treatments it is essential to understand the composition, formation and functionality of protein complexes. Novel techniques like complexome profiling give an overview of possible protein-protein interactions in an entire sample. In this approach intact protein complexes are separated using blue-native electrophoresis. The migration patterns of thousands of proteins are then uncovered using quantitative mass spectrometry and compared to find co-migrating proteins. Here, we present the concepts of our visualization approach for large complexome profiling datasets using our software NOVA. In agreement with recent literature we show that the protein NDUFA4, a previously known subunit of complex I of the mitochondrial respiratory chain, is instead a subunit of complex IV.

Keywords: complexome profiling, visualization, clustering, protein complexes, proteomics

INTRODUCTION

Proteins are very versatile and carry out a variety of functions. They preserve the structural integrity of cells, catalyze metabolic reactions, enable motion, transport of all kind of ions and molecules and are a major part of the organismic defense systems. To facilitate these diverse functions, proteins interact with each other, forming macromolecular assemblies that contain multiple proteins. These protein complexes can form highly organized structures, so-called super-complexes. Prominent examples are the mitochondrial complexes I - V of the oxidative phosphorylation (OXPHOS). Their main task is to produce ATP, the energy source for most processes in cells. Because of their important functionality, it is unsurprising that dysfunction of these complexes is associated with diseases. For instance, defects of OXPHOS complexes I and II are involved in human neurodegenerative diseases like Parkinson's (Tretter et al., 2004; Swerdlow et al., 1996) or Huntington's disease (Benchoua et al., 2006). To treat these diseases it is crucial to understand the composition, formation and functionality of protein complexes.

To identify the interaction partners of specific proteins, techniques like immunoprecipitation and tandem affinity protocols in combination with western blotting and mass spectrometry are commonly used. These approaches are limited to detect specific interactions. The recently introduced *complexome profiling* (Heide et al., 2012) has been developed to overcome these limitations. Complexome profiling uses blue-native electrophoresis to separate complex protein mixtures by size while leaving protein complexes intact. With quantitative mass spectrometry the migration pattern of each measured protein in the native gel is uncovered. Proteins with similar migration pattern, so–called migration profile, are likely to be part of the same protein complex. Typical complexome datasets contain thousands of such protein migration profiles. These large datasets give two major challenges for the researchers. Firstly, a suitable visual representation of the data is needed that allows easy comparison and interpretation. Secondly, an automated method for the comparison of the migration profiles is required.

Since none of the existing software solutions provided the functionality desired by the researchers we developed a "tailor-made" user-oriented software which is easy to use and provides an appropriate visualization and methods to analyze these large datasets. Here, we describe the challenges we faced and our approach to large data visualization in our software NOVA (Giese et al., 2015). We demonstrated the usefulness of NOVA for the analysis of complexome data, using a case study, in which we correctly assign a previously falsely predicted protein to a protein complex.

METHODS

Complexome Profiling

Complexome profiling (Heide et al., 2012) uses blue-native gel electrophoresis (BNE) to separate complex protein mixtures (see Figure 1 step 1 - 2). BNE separates intact proteins and protein complexes up to a molecular weight of 10 MDa (Schägger and Jagow, 1991; Wittig et al., 2006). Larger protein complexes up to 60 MDa (Strecker et al., 2010) can be separated using special large pore gels (LP-BNE). After the electrophoresis the gel lane is cut into 60 gel slices of equal size (see Figure 1 step 3). Each slice is then separately analyzed by mass spectroscopy (LC-MS/MS) to identify the contained peptides. Using the Mascot search engine, the peptides are evaluated to identify the proteins they most likely originated from. Label-free LC-MS-based protein quantification is applied to compute the relative amount of each protein over all slices. The semi-quantitative information is used to create migration profiles for all proteins (see Figure 1 step 4). The migration profile for each protein is given by 60 values which indicate the amount of protein in the 60 gel slices. For easier visual inspection the resulting data matrix can be represented as a colored heat map (see Figure 1 step 5), where each row stands for one protein and each column for one slice. The abundance of a protein is then color-coded. We used black for slices in which no abundance was detected and a gradient from yellow to red for low to high abundance values, see Figure 1.

Subunits of the same protein complex co-migrated through the gel, and therefore are expected to show similar migration profiles. To identify the subunits of a protein complex the migration profiles are compared. A manual processing of all migration profiles, typically, hundreds to thousands, is not feasible. Statistical analysis techniques like cluster analysis are required to process such datasets. Using hierarchical clustering, groups of co-migrating proteins can be automatically recognized, indicating the composition of quaternary structures and functional complexes (Wessels et al., 2009; Foster et al., 2006; Andersen et al., 2003). Complexome profiling has been successfully applied to analyze mitochondrial complexes in rats (Heide et al., 2012) and humans (Wessels et al., 2013) as well as to explore complex formation in plants and bacteria (Takabayashi et al., 2013).

Hierarchical Clustering

The premise of complexome profiling is that co-migrating proteins are likely to have similar migration profiles. Manual comparison of several thousands of profiles is not efficient. Therefore, we need automated strategies for this task. Clustering is a useful tool for the comparison of large datasets. We used agglomerative hierarchical clustering. For a given dataset E containing m migration profiles $e \in E$ we define $C = \{c_1, c_2, \ldots, c_n\}$ as a partition of E into disjunct subsets, $c_i \subseteq E, i = 1, 2, \ldots, n$. A synonym for each element $c \in C$ is *cluster* and C is called the *set of clusters*. To calculate the dissimilarity between clusters a function D is used:

$$D\colon C \times C \to \mathbb{R}_+ \ .$$

D assigns any two clusters $c_i, c_j \in C$ to a positive real number representing the distance between the clusters. If $c_i = \{e_i\}, c_j = \{e_j\}$ are singleton clusters a distance function d:

$$d\colon E \times E \to \mathbb{R}_+$$

is used to calculate the distance. For singleton clusters, the measure of dissimilarity, D, is defined by:

$$D(c_i, c_j) = d(e_i, e_j) \ . \tag{1}$$

For clusters of larger size, i.e., for $|c_i| > 1$ and/or $|c_j| > 1$, we have to choose a so-called linkage function, l,:

$$l\colon C \times C \to \mathbb{R}_+$$

to compute the measure of dissimilarity, D,:

$$D(c_i, c_j) = l(c_i, c_j) \ . \tag{2}$$

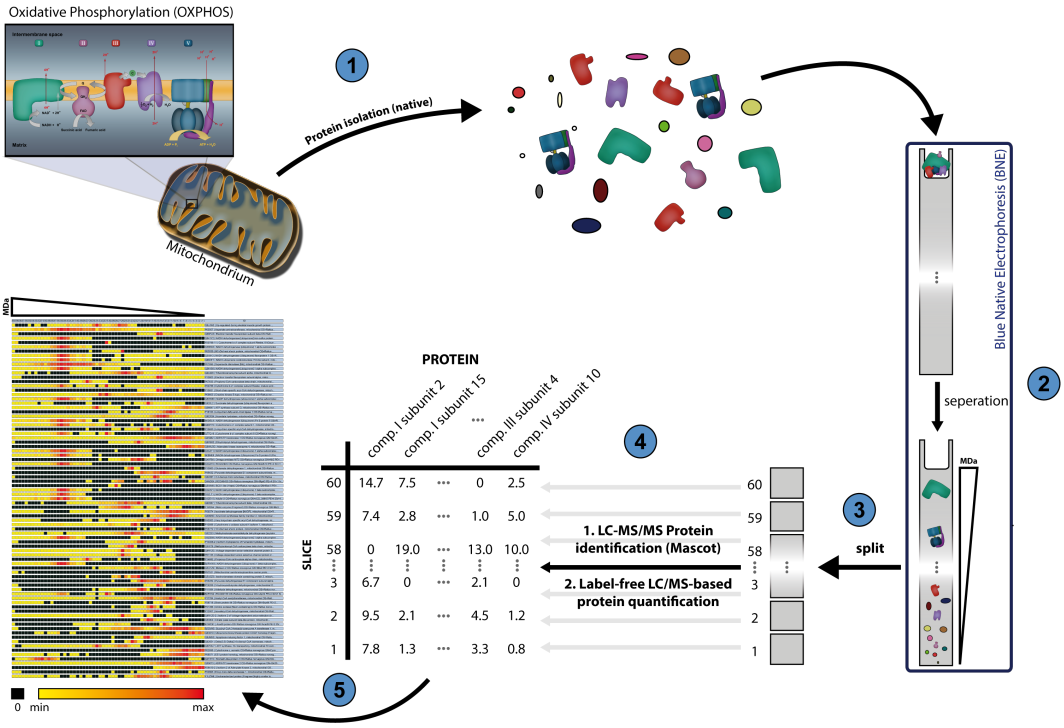

Figure 1. For complexome profiling, a protein mixture is separated by BNE (1-2). The gel lane is cut into 60 gel slices of equal size (3) and the protein content of each slice is analyzed by mass spectroscopy (LC-MS/MS). The semi-quantitative information on the abundance of proteins in each slice is then used to create migration profiles for all proteins (4). The migration profile for each protein is given by 60 values which indicate the amount of protein in the 60 gel slices. For easier visual inspection, the data matrices are displayed as a heat map (5). Here, we used black for values of zero and a gradient from yellow to red for the smallest to the highest values in a profile.

Distance function

To compute a measure of dissimilarity for singleton clusters, we have to choose a suitable distance function d. An example for a distance function is the Euclidean distance. Another commonly used distance function is the Pearson distance. It is based on the Pearson correlation coefficient, r, which calculates the correlation between two vectors. For defined vectors x and y, $|x| = |y| = n$ the Pearson correlation is defined by:

$$r = \frac{1}{n}\sum_{i=1}^{n}\left(\frac{x_i - \bar{x}}{\sigma_x}\right)\left(\frac{y_i - \bar{y}}{\sigma_y}\right) \tag{3}$$

with means

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i,\ \bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i$$

and standard deviation

$$\sigma_x = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2},\ \sigma_y = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \bar{y})^2}\ .$$

The Pearson correlation coefficient is in the interval $[-1, 1]$. Using r, we compute the Pearson distance by the following equation:

$$d_p(x,y) = 1 - r\ . \tag{4}$$

The resulting distances are in the interval $[0, 2]$. For the vectors x and y, $r_{x,y} = -1$, the distance is maximal $d_p(x,y) = 2$.
In a previous study (Giese, 2012) we have evaluated the Euclidean and Pearson distance functions as well as several other distance functions for the ability to correctly compute the distance between migration profiles. We have found preferable results by applying Pearson correlation.

Linkage method

We implemented various linkage functions as, e.g., the single–, complete–, and average–linkage function. According to our experience the average–linkage in combination with the Pearson distance performs best on complexome profiling datasets. For a given distance function, d, the average–linkage function, l, is defined as:

$$l(c_i, c_j) = \frac{1}{n_i \, n_j} \sum_{p \in c_i} \sum_{q \in c_j} d(p,q) \; . \tag{5}$$

The numbers of elements in $c_i \in C$ and $c_j \in C$ are given by $n_i = |c_i|$ and $n_j = |c_i|$, respectively. Agglomerative clustering using this average–linkage technique is commonly referred to as UPGMA (Unweighted Pair Group Method with Arithmetic Mean).

During the clustering process the two clusters with the smallest distance are joined into one cluster in each step. No further steps are possible after $n = |E| - 1$ steps because all elements, $e \in E$, will be united in one single cluster. For real–life datasets, the final partition, i.e., all elements in one cluster, will not be optimal in terms of intra-cluster homogeneity and inter-cluster heterogeneity. Therefore, many implementations provide a stop condition which determines when an *optimal* partition is reached. For our implementation we adopted another approach. NOVA performs the complete clustering and the user can interactively cut the cluster tree to choose a partition of interest.

RESULTS AND DISCUSSION

In the beginning, we visualized the migration profiles as heat maps, using Microsoft Excel. To cluster complexome profiling data, experimental scientists applied the software *Cluster 3.0* (de Hoon et al., 2004; Eisen et al., 1998). For the visualization of the cluster tree, the software *Java Treeview* (Saldanha, 2004) was used. Though these tools are in general suitable to each of the assigned tasks, transferring data from one tool to another, was time consuming and a potential source for errors. We were looking for a single software that combines the needed functionalities. Additionally, the software should enable the users to quickly create subsets of the data, compare multiple heat maps of various experimental conditions and allow for an easy visual inspection of migration profiles in e.g. a line chart. Because of the poor results of our search and the rather specific demands, we decided to develop our own software NOVA.

NOVA

We designed NOVA to be easy to use especially for scientists with no background in bioinformatics, mathematics, or computer science. Moreover, NOVA can handle a variety of file formats like xlsx, xls and csv. The format of most csv files can be recognized automatically. Results can be exported to files or as images.

Analysis

For the analyses of complexome profiling data, hierarchical clustering was implemented. Although averag–linkage gives good results, several other linkage methods, including single–, complete–, and Ward's–linkage, are provided. A variety of distance functions, including Pearson correlation-based distance functions can be selected for the clustering procedure. For faster performance, we applied a fast clustering approach which uses a queue to quickly find the closest clusters for the clustering implementation. The values of each migration profile can be normalized by a variety of normalization techniques, e.g. maximum and unit-vector normalization.

Visualization

For the visualization of migration profiles, we decided to use heat maps (see Figure 2 A). Each row of the heat map displays the profile of an individual protein. Rows and columns can be selected to, e.g., create subsets of the data according to the specific application or a specific hypothesis.

It is impossible to display a heat map with thousands of rows and still be able to distinguish each particular row on a PC monitor. Even if we draw each row with a height as less as one pixel we can only display a little more than thousand profiles on an average display with a resolution of 1920 x 1080 pixels. Furthermore, showing all the data on the screen, allows the user to see the larger picture, but, it also makes it difficult to identify subtle differences between profiles. To solve this, we made the heat map zoomable, allowing the user to see large parts of the data if they zoom out and also to focus on areas in more detail by zooming in. Once the data is clustered, the rows of the heat map are rearranged according to the cluster hierarchy. So, migration profiles of higher similarity are directly adjacent to each other (see Figure 3). Compared to the number of elements in the entire dataset,

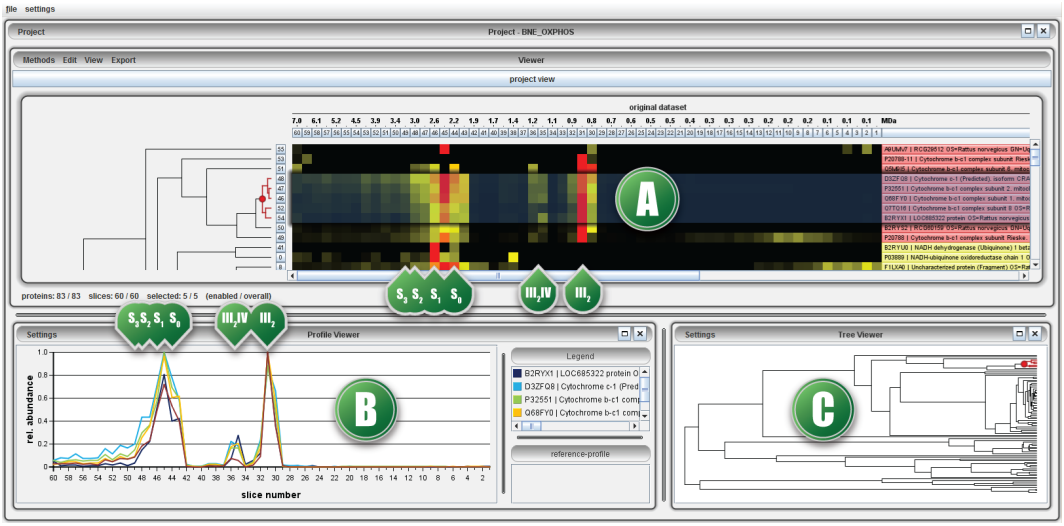

Figure 2. NOVAs graphical user interface (GUI): Displayed in the GUI is a clustered complexome profiling dataset from rat hearth mitochondria (Heide et al., 2012). (A) Gel migration profiles are represented as a colored heat map. Each row shows the profile of an individual protein. A label at the end of each row identifies the protein. The background color of the label indicates a known membership of the protein to a complex, e.g., yellow and red labels mark subunits of respiratory complex C I and C III, respectively. A mass scale on the top of the heat map shows the expected mass of proteins assembled in the gel slices. The migration profiles are clustered, here, the Pearson correlation-based distance and average–linkage were used. Left, the corresponding cluster subtree is aligned to the heat map. A cluster of proteins in the cluster tree is highlighted in red, corresponding to the selected rows of the heat map. (B) A line chart displays the migration profiles of the selected proteins. Each peak of the consensus profile corresponds to a protein assembly, functional complex, or super-complex. These assemblies are identifiable by their distinct masses. Here, the selected proteins are members of the respiratory homodimeric complex III_2, the complex assembly III_2IV, and the series of super-complexes $S_0 - S_3$. (C) The complete cluster tree is shown in the *tree viewer*. It allows to navigate through the heat map and to explore migration profiles of subgroups of proteins.

clusters of interest usually contain much fewer elements. Thus, the need to see all profiles decreases when the focus shifts onto particular clusters.

Profiles selected in the heat map can be displayed as a line chart (see Figure 2 B). This view allows a more detailed comparison of the migration profiles. A required feature for the comparison in the line chart was the one versus many comparison. While a line chart is particular suitable to compare a few profiles it can be crowded and confusing if several profiles are compared. The feature was mostly required to compare a specific profile against the entire ensemble of profiles in a cluster. Based on that, we integrated the option to assign profiles to a reference profile. The reference profile is the average of all profiles that are assigned to it. It is always displayed in the line chart even if the profiles assigned to it are not selected in the heat map. This enables the user to quickly select any other profile and compare it against the reference profile.

To visualize the assembly of the clusters, the cluster tree is displayed on the left side of the heat map (see Figure 2 A). Parts of the tree can be selected which in turn selects the associated profiles in the heat map. The tree can be cut interactively. Additionally, the tree can be explored in a separate viewer (see Figure 2 C). The viewer provides the same functionality and additionally supports zooming. This is particular useful for very large trees.

For some experiments multiple complexome profiling datasets need to be compared, for example, to compare the wild type vs. a knock-out type. To facilitate this, heat maps of multiple complexome profiling datasets can be displayed side by side with the heat map of the initial dataset, the reference heat map. The arrangement of the profiles in the other heat maps is synchronized to the reference heat map.

Case study

To test our implementation, we reprocessed the dataset used by Heide et al. (2012) in their original publication with NOVA. The dataset was a complexome profiling of rat heart mitochondria and

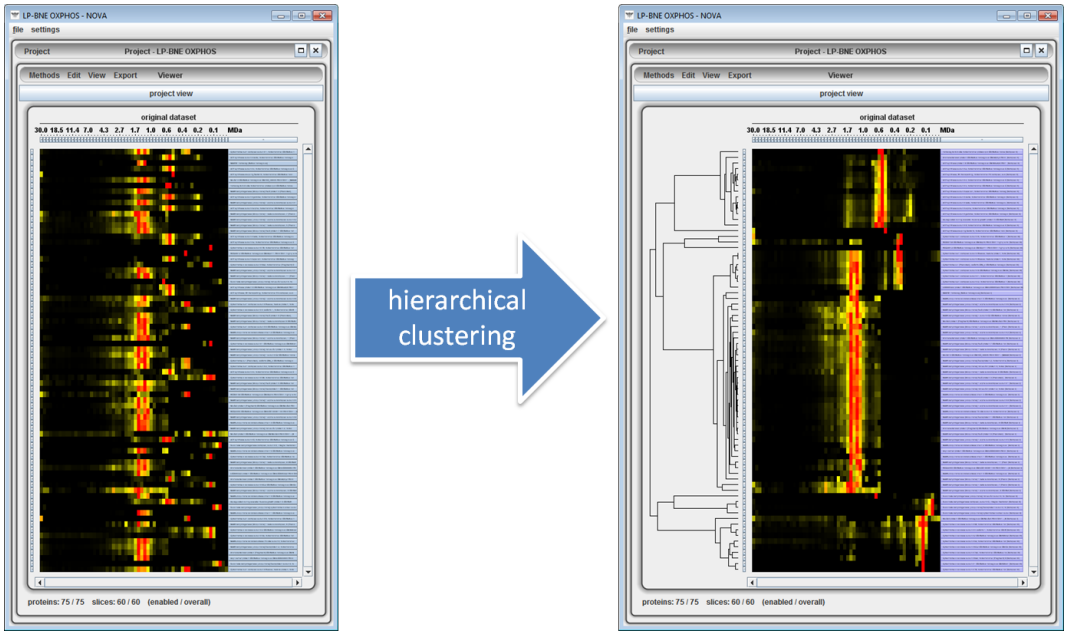

Figure 3. On the left side, an unclustered complexome profiling dataset is displayed as a heat map in the NOVA software. Though many profiles seem to have similar profiles, no clear patterns can be recognized. On the right side, the same dataset is displayed after hierarchical clustering was applied. Structures of high similarity are revealed. The cluster tree is displayed on the left side of the heat map.

contained more than 500 migration profiles. We reduced the dataset to about 80 by selecting only migration profiles of proteins that are known subunits of the OXPHOS complexes. To easily identify proteins which are part of a certain complex, we assigned colors to the labels of the proteins. We assigned yellow for subunits of complex I, orange for complex II, red for complex III, green for complex IV and purple for complex V (see Figure 4 A). The data was then hierarchically clustered, using UPGMA and the Pearson correlation distance. The cluster tree is displayed on the left side of the heat map in Figure 4 A and B. On a first glance, we can see that proteins of the same complex are nicely grouped together. A closer look at the cluster tree reveals that though the overall arrangement is good there are some outliers. Here, we will focus on one of them, a protein called NDUFA4. To the best of our knowledge, the function of NDUFA4 is still unknown. To get an idea of its functionality, it is important to know which protein complexes contain NDUFA4. Though NDUFA4 is marked as a subunit of complex I (yellow label), it clearly grouped with complex IV. We took a closer look at the profile of NDUFA4 in comparison to the profiles of complex I and complex IV. Judging from the profile comparison, it is evident that NDUFA4 matches the migration profile of complex IV much better than complex I. This suggested that NDUFA4 might have been falsely classified as a complex I subunit. The assumption was validated by independent work of Balsa et al. (2012).

Conclusion and outlook

From out test study and extensive feedback from external testers, we concluded that NOVA is a very useful tool for the evaluation of complexome profiling data. Intermediates of protein complexes can be measured using complexome profiling. Thus the results of this type of analysis can not only be applied for the assignment of proteins to a complex, but also give insight into protein assembly.

Researchers have been combining complexome profiling with other techniques. For example Pulse-SILAC (stable isotope labeling with amino acids in cell culture) with complexome profiling has been used to study turnover of single proteins within protein complexes. For such new approaches, suitable visualization and analysis methods need to be developed and integrated into NOVA.

ACKNOWLEDGMENTS

We gratefully thank Lea Bleier and Stefan Dröse for supplying us with the datasets we described in this paper. Furthermore, we would like to thank Valentina Strecker, Kim-Kristin Prior, Jens Einloft, Leonie Amstein and Jörg Kuharev for testing and many valuable suggestions.

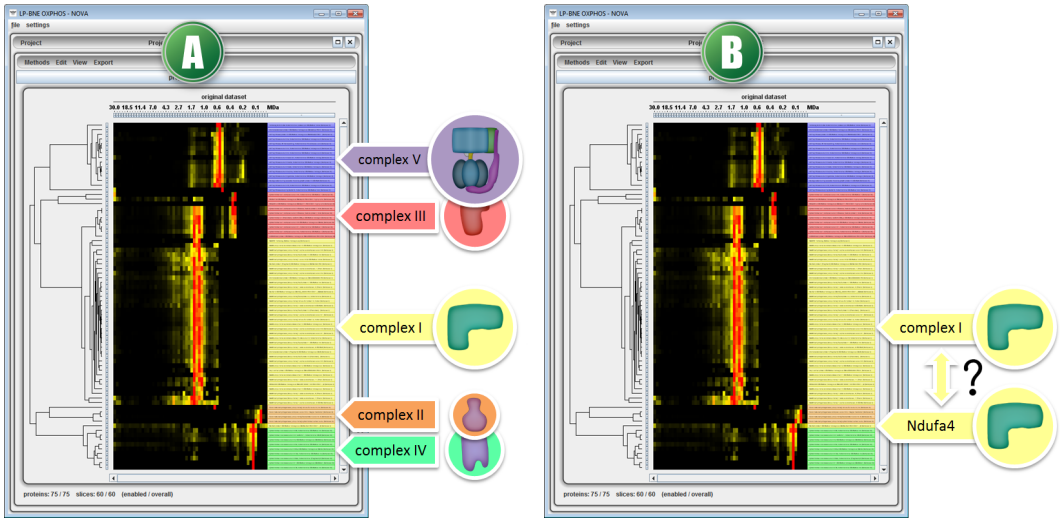

Figure 4. A: A clustered selection containing only known subunits of OXPHOS complexes of a complexome profiling dataset from rat heart mitochondria (Heide et al., 2012). The labels of the proteins are colored according to the complex they belong to. Proteins of complex I are yellow, complex II orange, complex III red, complex IV green and complex V purple. B: Most of the complex I subunits are clustered together. NDUFA4 a, previously known subunit of complex I, clusters with complex IV rather than complex I (Balsa et al., 2012).

REFERENCES

Andersen, J. S., Wilkinson, C. J., Mayor, T., Mortensen, P., Nigg, E. A., and Mann, M. (2003). Proteomic characterization of the human centrosome by protein correlation profiling. *Nature*, 426(6966):570–574.

Balsa, E., Marco, R., Perales-Clemente, E., Szklarczyk, R., Calvo, E., Landázuri, M. O., and Enríquez, J. A. (2012). NDUFA4 Is a Subunit of Complex IV of the Mammalian Electron Transport Chain. *Cell Metabolism*, 16(3):378–386.

Benchoua, A., Trioulier, Y., Zala, D., Gaillard, M., Lefort, N., Dufour, N., Saudou, F., Elalouf, J., Hirsch, E., Hantraye, P., Déglon, N., and Brouillet, E. (2006). Involvement of Mitochondrial Complex II Defects in Neuronal Death Produced by N-Terminus Fragment of Mutated Huntingtin. *Annals of Neurology*, 17:1652–1663.

de Hoon, M., Imoto, S., Nolan, J., and Miyano, S. (2004). Open source clustering software. *Bioinformatics*, 20(9):1453–1454.

Eisen, M. B., Spellman, P. T., Brown, P. O., and Botstein, D. (1998). Cluster analysis and display of genome-wide expression patterns. *Proceedings of the National Academy of Sciences*, 95(25):14863–14868.

Foster, L. J., de Hoog, C. L., Zhang, Y., Zhang, Y., Xie, X., Mootha, V. K., and Mann, M. (2006). A Mammalian Organelle Map by Protein Correlation Profiling. *Cell*, 125(1):187–199.

Giese, H. (2012). Bioinformatische Untersuchungen in der Proteomik: Analyse von Komplexbildungsprozessen. *Diploma thesis, J. W. Goethe-University Frankfurt am Main*.

Giese, H., Ackermann, J., Heide, H., Bleier, L., Dröse, S., Wittig, I., Brandt, U., and Koch, I. (2015). Nova: a software to analyze complexome profiling data. *Bioinformatics*, 31(3):440–441.

Heide, H., Bleier, L., Steger, M., Ackermann, J., Dröse, S., Schwamb, B., Zörnig, M., Reichert, A. S., Koch, I., Wittig, I., and Brandt, U. (2012). Complexome Profiling Identifies TMEM126B as a Component of the Mitochondrial Complex I Assembly Complex. *Cell Metabolism*, 16(4):538–549.

Saldanha, A. J. (2004). Java Treeview–extensible visualization of microarray data. *Bioinformatics*, 20(17):3246–3248.

Schägger, H. and Jagow, G. (1991). Blue native electrophoresis for isolation of membrane protein complexes in enzymatically active form. *Analytical Biochemistry*, 199:223–231.

Strecker, V., Wumaier, Z., Wittig, I., and Schägger, H. (2010). Large pore gels to separate mega protein complexes larger than 10 MDa by blue native electrophoresis: Isolation of putative respiratory strings or patches. *Proteomics*, 10:3379–3387.

Swerdlow, R. H., Parks, J. K., Miller, S. W., Davis, R. E., Tuttle, J. B., Trimmer, P. A., Sheehan, J. P., Bennett, J. P., and Parker, W. D. (1996). Origin and functional consequences of the complex I

defect in Parkinson's disease. *Annals of Neurology*, 40:663–671.

Takabayashi, A., Kadoya, R., Kuwano, M., Kurihara, K., Ito, H., Tanaka, R., and Tanaka, A. (2013). Protein co-migration database (PCoM -DB) for *Arabidopsis* thylakoids and *Synechocystis* cells. *SpringerPlus*, 2(1):148.

Tretter, L., Sipos, I., and Adam-Vizi, V. (2004). Initiation of Neuronal Damage by Complex I Deficiency and Oxidative Stress in Parkinson's Disease. *Neurochemical Research*, 29:569–577.

Wessels, H. J. C. T., Vogel, R. O., Lightowlers, R. N., Spelbrink, J. N., Rodenburg, R. J., van den Heuvel, L. P., van Gool, A. J., Gloerich, J., Smeitink, J. A. M., and Nijtmans, L. G. (2013). Analysis of 953 Human Proteins from a Mitochondrial HEK293 Fraction by Complexome Profiling. *PLoS ONE*, 8(7):e68340.

Wessels, H. J. C. T., Vogel, R. O., van den Heuvel, L., Smeitink, J. A., Rodenburg, R. J., Nijtmans, L. G., and Farhoud, M. H. (2009). LC-MS/MS as an alternative for SDS-PAGE in blue native analysis of protein complexes. *Proteomics*, 9(17):4221–4228.

Wittig, I., Braun, H. P., and Schägger, H. (2006). Blue-Native PAGE. *Nature Protocols*, 1:418–428.