Supplemental Materials
*Molecular Biology of the Cell*

Borner et al.
Supplemental Figure Legends

Figure S1: Fractionation profiling accurately predicts the organisation of the 26S proteasome at the sub-complex level.
(A) 32 core subunits of the 26S proteasome were identified. PCA of subunit profiles shows their arrangement into the 20S core (red and orange subunits), and the 19S regulatory particle (light and dark blue subunits). The 20S core comprises an outer (red subunits) and an inner ring (orange subunits); the outer ring is clearly resolved. The regulatory particle is composed of a ‘base’ (dark blue subunits) and a ‘lid’ (light blue subunits); this division is also suggested by the analysis. Less tightly associated accessory proteins (grey) are found in the vicinity of the 19S and 20S particles. The two PSME subunits form a stable dimer, which is a known regulator of proteasome function; they form a separate, somewhat distal, cluster.
(B) Estimated stoichiometry, normalised to the median of all core subunits. The data are consistent with an equimolar stoichiometry for most core subunits. Similarly, all associated factors appear to be present in sub-stoichiometric quantities.
Please note that SHFM1, the 33rd known subunit of the 26S proteasome (Tomko and Hochstrasser, 2013), is very small (8kD), and does not produce tryptic peptides suitable for detection by mass spectrometry. Error bars of abundance scores correspond to MADs (n = 6 replicate measurements).

Figure S2: Predicted known and novel protein associations in HeLa cells.
To begin to identify novel protein complexes or functional associations, the HeLa Predictor was searched for pairs of extremely similar profiles (‘three star’ classification). In total, 989 such pairs were identified, which were derived from 269 proteins. PCA of the corresponding profiles shows that these proteins fall into 58 groups. 24 of these are known protein complexes; the remaining 34 represent potential novel association groups (see Table S5 for details). Please note that this analysis is only the ‘tip of the iceberg’; the HeLa Predictor contains ~7,000 pairs of profiles with very high similarity (‘two star’ classification). This level of similarity is commonly found between subunits of the same complex. Hence, the Predictor (Table S1) covers many more protein complexes than shown here.
Supplemental Items Summary

Supplemental Figures:

Figure S1: Fractionation profiling accurately predicts the organisation of the 26S proteasome at the sub-complex level.
Figure S2: Predicted known and novel protein associations in HeLa cells.

Supplemental Tables:

Table S1: The HeLa Predictor. A database predicting functional associations/protein complexes/clathrin-coated vesicle proteins from HeLa cells.
Table S2: The complete predicted HeLa CCV proteome.
Table S3: The Drosophila S2 Predictor. A database predicting functional associations/protein complexes/clathrin-coated vesicle proteins from Drosophila S2 cells.
Table S4: The complete predicted Drosophila S2 CCV proteome.
Table S5: Top predicted protein associations/protein complexes from HeLa cells.
Table S6: The reference dataset of CORUM complexes.

Supplemental Text File:

Predictor Manual: A complete description of the Predictor database, including detailed instructions on usage and data interpretation.
1. Overview

The fractionation profiling Predictor is an interactive database which contains abundance distribution profiles obtained through quantitative mass spectrometric comparisons of multiple subcellular fractions. Proteins with similar distribution profiles are likely to be functionally (and perhaps physically) associated. The interface allows users to search the database against one or more query proteins, and to specify various search parameters. The Predictor then retrieves the corresponding profiles from the raw data table, and calculates a consensus query profile. This query profile is automatically compared with all other profiles in the raw data table. Profiles are then sorted by similarity to the query profile (results table). In addition, the Predictor provides various graphical summaries for both query and output profiles, including estimates of relative protein abundance. Using the Predictor is very straightforward. This manual is a step-by-step guide to entering queries and understanding the output. There are currently two Predictor databases, for HeLa (human) and S2 (Drosophila) cells. The interface is very similar for both. The following manual is written for the HeLa Predictor, and explains features specific to the S2 Predictor at the end of each section.

Please Note:

a) The Predictor uses only basic Excel functions; macros do not have to be enabled.

b) The Predictor data and functions are protected against accidental changes by the user. Only cells which require user input can be changed. However, all contents can be viewed and copied by the user. To remove the protection, go to the ‘Review’ tab, and select the ‘Unprotect Sheet’ option. The password is ‘HeLaPred’ for Table S1, and ‘DrosPred’ for Table S3.

2. Finding your protein of interest in the database

Go to the ‘Complete Data’ tab. This is the raw data, and hence the core of the Predictor. All profiled proteins are listed in alphabetical order (by gene name). The HeLa Predictor includes over 4,500 entries, representing >40% of the proteins expressed in HeLa cells. The UniProt gene name, protein name, and Fasta-header are listed, as well as a ‘gene number’ for easy reference. For each protein, the experimentally determined abundance ratios are given (log2 scale). For example, ‘H/L Frac1_35’ denotes the abundance ratio in the 35K sample of the first set of fractionation experiments. For each ratio, the number of “ratio counts” (ie quantified SILAC pairs) that was used to determine the ratio is indicated (minimum required count is 2). If no or only one ratio count is shown, then the protein was not quantified in this particular dataset. In total, there are six datasets, so proteins can have up to six data points (a ‘complete’ profile). Incomplete profiles have five or fewer data points, which may reduce the accuracy of the similarity assignment (see below, 3. Submitting a query). Importantly, accurate prediction requires that a protein has at least one complete ‘data triplet’, ie ratio values for all three Frac1 (columns E, G, I) or all three Frac2 (columns K, M, O) datasets.

To locate a protein of interest, you may either scroll down the alphabetical list, or use Excel’s search function. Press Ctrl + F to invoke the search box, and enter a key word, or partial gene or protein name. If your protein is not found, it may not be in the Predictor database. You may also try to find the protein in the UniProt database first, to determine its exact gene name (www.uniprot.org). Once you have found your protein, note or copy the exact gene name (column B), or gene number (column A).
Special notes for the S2 Predictor: Since many Drosophila proteins do not have UniProt gene or protein names, these are not suitable as unique identifiers. Hence, every protein in the S2 database was assigned a number (see ‘Complete Data’ tab, column A). Please find your protein of interest, and note or copy the exact gene number. In the ‘Submit Query’ tab, please enter this gene number (or gene numbers, in case of a multi-protein query) in cells A5-A48.

3. Submitting a query

Go to the ‘Submit Query’ tab. Here you can specify against which profile you wish to search the Predictor database. It is important that you know either the exact (UniProt) gene name, or the gene number of your protein of interest (see 2: Finding your protein of interest in the database). To submit a single query, simply enter the gene name in cell A5, or the gene number in cell A26. The Predictor will retrieve the corresponding profile from the database. If no values are retrieved, it means that the protein is either not in the database, or that the gene name is incorrect. Also make sure to delete any other gene names from a previous search from cells A5-A48.

Example: You are interested in the BLOC-1 complex, which consists of eight subunits: DTNB1, MUTED, SNAPIN, CNO, PLDN, BLOC1S1, BLOC1S2, and BLOC1S3. Enter ‘dtnbp1’ in cell A5. The profile of the protein dysbindin will be retrieved. Depending on system performance, this may take several seconds.

Input parameters:
The user can fine-tune several stringency settings of the search in column N (eg scoring method, number of valid datapoints, etc - see below, ‘Advanced applications’). It is however recommended to keep the default settings, at least initially, as they have the highest possible stringency.

Query evaluation:
The Predictor automatically evaluates the suitability of the query, and warns the user of inadequate settings or profiles. Details of the evaluation are given in rows 38-42.

1. Number of data points: Ideally, the query has six data points (a ‘complete’ profile). Incomplete profiles are permissible, but are likely to reduce the accuracy of prediction. A warning message appears in cell P38 if the query is incomplete.

2. Query abundance: If a single query is entered, the abundance of the protein is checked. If it is below a certain threshold, a warning is issued – low abundance proteins are likely to be less accurately quantified, and the accuracy of prediction may be reduced. A warning message appears in cell P39 if the query abundance is low.

NB: The low abundance filter only applies to single gene queries.

3. Profile distance to baseline: The discriminating power of the Predictor is best for proteins with strong changes in abundance across the profile. To quantify this, the Predictor calculates the profile’s distance to the baseline (which corresponds to an invariant ratio of 1 (or 0 on a log scale) across all six datasets). Profiles close to the baseline are less informative, and may have reduced accuracy of prediction. A warning message appears in cell P40 if this is the case. The query profile can be further evaluated graphically (see below, 4. Evaluating the query profile).

NB: The distance to baseline filter only applies to single gene queries.

A summary evaluation will appear in cell O40.
Example (continued): DTNBP1 has six data points, is sufficiently abundant, and the profile is not close to the baseline. The profile passes all three quality filters. This is a suitable query.

Furthermore, please note that that the number of data points in the profile must be equal to or larger than the minimum number of ‘output data points’ (set in cell N23). Otherwise an error message appears in cell P23, and no predictions are made. For complete profiles (ie six data points), this is never a problem.

When the Predictor is searched with clathrin heavy chain (CLTC), a special scoring mode is activated (see below, 10.). A blue message appears in row 44 to inform users of this.

Advanced application: Multiple input queries
If you enter multiple query genes, the Predictor will calculate a median or average consensus profile form all queries. Consensus building is particularly useful when you want to define the ‘typical’ profile of a protein complex, by entering several known members. You can also use this feature to compare two or more profiles graphically, as input profiles are automatically plotted (see 4. Evaluating the query profile).

If you wish to enter two or more query genes, enter gene names in cells A5, A6, etc up to A24. The consensus profile of all queries will be shown in rows 55 (average) and 57 (median), and will automatically be used to search the Predictor database.

Advanced application: User defined profile
It is also possible to specify an input profile directly. Simply enter the desired abundance ratios in cells E49-J49.

Advanced application: Specifying input parameters to fine-tune the analysis
Scoring method (select 1 or 2 in cell N9): For distance calculations, profiles are considered as two data points in n-dimensional space. Two alternative methods can be selected: the absolute ‘Manhattan’ distance, or the squared Euclidean distance. Both distances are normalised to the number of data points in a profile. Squared Euclidean distance scoring is more stringent (default setting), as it ‘punishes’ single large differences between profiles more severely than the absolute distance. Conversely, the squared Euclidean distance can be distorted by a single inaccurate data point; absolute distance scoring is more forgiving, and thus makes potentially more inclusive predictions.

Output filter (select 1 or 2 in cell N17): Accurate predictions require at least one complete data triplet (see above, 2.). You can deactivate this stringency setting to search with or for profiles without such a triplet. The accuracy of the predictions is likely to be drastically reduced though.

Minimum number of data points (specify 1-6 in cell N23): The default is to search only with and for complete profiles (6 data points). By reducing this number, it becomes possible to search with and for incomplete profiles. This may reduce the accuracy of prediction, but increases the sensitivity. In our experience, profiles with 5 data points are very suitable; profiles with 3 or 4 data points are acceptable, provided they contain one complete data triplet. While searches with other profiles are permissible with accordingly adjusted settings, they are not recommended.

Average/Median query profile: (specify 1 or 2 in cell N32): In case the query consists of multiple profiles, the Predictor calculates a consensus profile. The user can specify this to be the average (mean) or median of the input profiles (if there are only two profiles, the median equals the mean). Median consensus profiles are recommended, as they are more robust to outliers.
Special notes for the S2 Predictor: In the ‘Submit Query’ tab, please enter the gene number (or gene numbers, in case of a multi-protein query) in cells A5-A48.

4. Evaluating the query profile

Go to the ‘Query Profiles’ tab. Here you can evaluate the query profile. The Predictor automatically plots up to five input query profiles, as well as the consensus profile. (In case there is only one input profile, this will be identical to the consensus profile.)

Please note that profiles entered as gene numbers (cells A26-A48) and user defined profiles (rows 49-54) are not plotted. However, if queries are entered into these fields, they will still affect the consensus query profile, which can be plotted.

The Predictor also calculates the profile’s distance to the baseline (cells W5 and W8), compares it to all other profiles in the databases, and returns a percentile rank. Generally speaking, the higher the rank, the more distinctive is the profile. Distances below the 20th percentile are considered close to the baseline, and trigger a warning message in the ‘Submit Query’ tab (see above, 3.). Accurate predictions are still possible with such a profile, but unrelated profiles close to the baseline have a higher chance of blending into each other, as the differences are very small.

Example (continued): The profile of DTNBP1 is in the 30th-40th percentile (for both scoring methods). Thus, its distance to the baseline is greater than 30% of the profiles in the database, which is acceptable.

A cross in Column AB indicates which scoring method was selected (see above, 3.). Sometimes the percentiles differ. Generally, it is recommended to select the scoring method with the higher percentile rank; usually they are similar though.

Advanced application: Comparing multiple queries

The profile plotting tool allows users to inspect the query profile visually. The input query profile(s) in cells A5-A24 (‘Submit Query’ tab) are plotted. You can specify how many input profiles to plot (enter 1-10 in cell W13). You can also toggle between showing and hiding the consensus profile (select 1 or 2 in cell W15). The main purpose of this tool is to evaluate the differences between multiple query profiles.

Example: You would like to compare the profiles of multiple subunits of the BLOC-1 complex. In the ‘Submit Query’ tab, you have already entered ‘DTNBP1’ in cell A5. Now enter ‘MUTED’ in cell A6, and ‘PLDN’ in cell A7. The Predictor will retrieve the profiles of the proteins MUTED and Pallidin (both members of BLOC-1). Go to the ‘Query Profiles’ tab. The plot shows the profiles of all three submitted proteins; they are very closely matched. It makes sense to build a consensus from these three profiles.

Now go back to the ‘Submit Query’ tab. Enter ‘AP2M1’ in cell A8, to retrieve the profile of the AP-2 medium subunit (a protein that is unrelated to BLOC-1). Go to the ‘Query Profile’ tab. You can see that AP2M1 has a completely different profile from the BLOC-1 subunits. It does not make sense to include AP2M1 for consensus building. Notice how the consensus is unaffected by addition of the AP2M1 profile though; since it is based on calculating the ‘median’, it simply ignores the AP2M1 profile. An ‘average’ consensus however would be grossly distorted by the AP2M1 profile. Change the setting in cell N32 (‘Submit Query’ tab) to 1, and look at the plot again. The new ‘average’ profile is a hybrid that neither reflects AP-2 nor BLOC-1 distribution, and is hence not a useful query.
Finally, the plotting tool can also be used to quickly assess if two proteins of interest have similar profiles, not just for consensus building, but simply to see how they behave in the profiling dataset.

**Special notes for the S2 Predictor:** Any query profile distance below the 30th percentile is considered to be ‘close to the baseline’. Furthermore, you can switch between displaying gene names and protein names in cell U24. Please note that many Drosophila proteins do not have a UniProt gene or protein name (annotated as ‘0’).

5. Results: Top 20 Scores and the ‘Star’ scoring system

Go to the ‘Results Top 20 – Score’ tab. The bar chart shows the gene names of the 20 profiles closest to the query, in ascending order of distance. In case of a single query, this will occupy rank 1 (distance = 0). The protein with rank 2 has the most similar profile to the query, rank 3 has the second most similar profile, etc. The cut-off at rank 20 has no particular significance – the function of this plot is to provide a quick overview of the top hits.

To guide the interpretation of the distance scores, the Predictor uses a simple ‘star’ system. This was derived by comparing profile distances between subunits of well-established protein complexes (see Materials and Methods for details). Profiles with a 3-star distance are almost identical, and indicate a tight functional or physical association. A 2-star distance denotes very similar profiles. Most subunits of stable complexes have 2- or 3-star distances. Hence, the 2-star classification is often a useful stringent cut-off when looking for members of the same protein complex. Of course even a two-star distance is no guarantee for a functional link between proteins; however, it is strongly suggested. A 1-star distance indicates similar profiles; many relevant hits are still to be expected in this range, but also many irrelevant hits. Finally, distances classified as ‘B’ (borderline) may still contain some relevant hits; beyond this classification, the ranking produced by the Predictor has little discriminating power.

In cell U18 you can select which cut-off you would like to see as part of the bar chart.

**Example (continued):** You have entered a single gene query, DTNBP1, in cell A5 in the ‘Submit Query’ tab. Go to the ‘Results Top 20 – Score’ tab. The BLOC-1 complex has eight subunits: DNTBP1, MUTED, SNAPIN, PLDN, CNO, BLOC1S1, BLOC1S2, and BLOC1S3. The search with DTNBP1 returns all eight proteins at the top of the list. They are all above the 2-star cut-off typical of proteins in a stable complex. Furthermore, the plot indicates that several members of the AP-3 complex are very close to BLOC-1. Indeed, the AP-3 complex is known to interact with the BLOC-1 complex.

Adjust the floating star cut-off. Go to cell U18, and type in ‘1’. The plot now shows the cut-off for 3-star rankings. Only the top four proteins (DTNBP1, MUTED, BLOC1S3, CNO) get this ranking. Change the value in cell U18 it back to ‘2’; now, all eight subunits are included, as well as several AP-3 subunits.

Change the value to ‘3’; now all AP-3 subunits are included, but also other proteins of dubious relevance. In addition, more than 20 proteins have a 1-star ranking or higher; the plot alerts the user with a message in line 22.

Changing the parameter in cell U18 does not affect the output of the analysis; it is intended as a tool to see where the cut-offs lie.

**Special notes for the S2 Predictor:** You can switch between displaying gene names and protein names in cell U13. Please note that many Drosophila proteins do not have a UniProt gene or protein name (annotated as
‘0’). To see the identities of these proteins, you have to go the ‘Results Table - Complete’ tab and look at the Fasta header.

6. Results: Top 20 – Abundance

Go to the ‘Results Top 20 – Abundance’ tab. The column plot shows the gene names and relative abundance of the twenty hits closest to the query. In case of a single gene query, this will occupy the first position. You can view the abundance scores on a linear or on a log_{10} scale (toggled in cell U15).

Abundance scores are relative, and based on normalised iBAQ intensities (see Materials and Methods for details). To facilitate comparisons, Clathrin heavy chain (CLTC) was arbitrarily assigned a score of 1,000,000. Thus, a protein with a score of 10,000 is 1% as abundant as CLTC (ie 1 copy for every 100 copies of clathrin). Proteins with similar scores are present in similar copy numbers. This can be used for example to predict the stoichiometry within protein complexes.

The accuracy of the abundances estimation varies between proteins, and also depends on the absolute abundance (generally, the higher the better). It is not possible to provide an accurate confidence interval. However, based on our analysis of protein complexes with known stoichiometry, we estimate that values typically lie within a factor 1.5 of the true value (ie they scatter between ~0.7 to 1.5 around the “true” value 1.0). More extreme outlier can of course occur.

Each abundance score is calculated as the median from up to six repeat measurements. As a measure of the variability of estimates, the median absolute deviation (MAD) from the median score is shown for each protein (as error bars in the column plot, and as numerical values in the Complete Results table). Cell U21 indicates them minimum number of repeat measurements used to calculate the abundance score. This corresponds to the “minimum number of datapoints” setting in the Submit Query tab (cell N23).

Example (continued): You have entered a single gene query, DTNBP1, in cell A5 in the ‘Submit Query’ tab. Go to the ‘Results Top 20 – Abundance’ tab. The eight BLOC-1 subunits DNTBP1, MUTED, BLOC1S3, CNO, PLDN, BLOC1S1, BLOC1S2, and SNAPIN are among the top hits. Within the limits of the accuracy of the method, all appear to have similar abundance (~90,000 median, ranging from ~80,000 to ~140,000), consistent with an equimolar stoichiometry. The log-scale view makes this even more obvious (set cell U15 to ‘2’). Interestingly, the AP-3 subunits AP3D1, AP3B1, AP3S1, and AP3M are also present in very similar quantities (note that there are two AP3M subunits, 1 and 2, whose abundances have to be added). This suggests that BLOC-1 and AP-3 associate with a 1:1 stoichiometry.

Special notes for the S2 Predictor: You can switch between displaying gene names and protein names in cell V23. Please note that many Drosophila proteins do not have a UniProt gene or protein name (annotated as ‘0’). To see the identities of these proteins, you have to go the ‘Results Table - Complete’ tab and look at the Fasta header.

7. Results: Top 10 profiles

Go to the ‘Results Top 10- Profiles’ tab. The plot shows the profiles of the top ten proteins closest to the query. You can specify how many profiles are plotted (1-10, default is 5; cell W10). You can also show the (consensus) query that was used to generate the results (toggle in cell W15). The purpose of the plot is to
help evaluate visually how the profiles of the top hits relate to each other, i.e., how similar or different they are.

NB: The Excel graph function substitutes missing values with 0, and connects the graph with the baseline. This can generate very misleading plots. Hence, the Predictor flags profiles with missing data points (row 20-32). If only complete profiles are analysed, this is never an issue. If however incomplete profiles are included, the Predictor will alert the user to this fact (cell W20), and also indicate which profiles have missing data points (and how many).

Example (continued): You have entered a single gene query, DTNBP1, in cell A5 in the ‘Submit Query’ tab. Go to the ‘Results Top 10 – Profiles’ tab. The profiles of the top five hits are shown, DNTBP1, MUTED, BLOC1S3, CNO and PLDN. They are all extremely similar, consistent with the notion that they are part of the same complex. Set cell W15 to ‘1’, to see the query profile. Set cell W10 to 9, to see all members of the complex.

8. Results: Complete table

Go to the ‘Results Table – complete’ tab. This table shows all profiles in the database sorted by similarity to the query (which has rank 1 in case of single gene query).

The columns include the following information: A: Similarity rank of the protein.
B: The ‘star’ classifier (see above, 5.).
C: Gene name (UniProt).
D: Protein name (UniProt).
E: Fasta header of the UniProt entry.
F: An indication whether or not this protein was previously identified by proteomics as a component of HeLa clathrin-coated vesicles (CCV). References in parentheses refer to (1) Borner et al., 2012 (Journal of Cell Biology, 197:141-160), or (2) Hirst et al., 2012 (Current Biology, 22:1711-1716).
G-R: The experimentally determined abundance ratios (log2 scale) from all six datasets, and the corresponding number of quantification events (SILAC pairs). The minimum number of quantification events required to calculate a ratio was set to 2. Blank cells indicate that the protein was either not detected in a particular dataset, or quantified only once.
T: The number of data points available in a protein’s profile (1-6).
V: Average absolute (‘Manhattan’) distance of the protein to the query profile (see above, 3.).
X: Average squared Euclidean distance of the protein to the query profile (see above, 3.).
AA/AB: Relative abundance score/median absolute deviation (MAD) (linear scale; see above, 6.)
AC: Relative abundance score (log10 scale; see above, 6.)
AF-AS: A compact output of the results, with some less important information trimmed.

Cell A2 is a status window which highlights potential problems with the analysis. If there are no issues, it remains blank. If the special CCV-scoring mode is active, it will alert the user to this (see below, 10.).

Please note that the results table consist of automatically filled in cells, and is hence not searchable with the Excel ‘find’ function. If you wish to search the output with ‘find’, you first need to copy the sheet, and ‘paste special’ as ‘values’ into a new spread sheet.
9. Compact output – Top 100 hits

Go to the ‘Results Top 100 - compact’ tab. This provides key data for the 100 hits most similar to the query profile (see above, 8., for a detailed explanation of the annotation). You may wish to copy-paste NB: as values only) this output into a separate Excel-document, to keep a record of your Predictor search.

If you find the top 100 hits insufficient, please go to the ‘Results Table –complete’ tab. A compact output including all profiles in the database is given in rows AF to AU.

10. Special query: Clathrin scoring

The Predictor was originally designed to identify proteins associated with clathrin-coated vesicles (CCVs). The only absolutely essential and universal component of CCVs is clathrin heavy chain (CLTC). A search of the database against this protein is the best way to reveal the composition of CCVs. Whenever CLTC is the top hit of a Predictor search, for example by entering ‘CLTC’ as a single gene query in cell A5 in the ‘Submit Query’ tab, the interface automatically switches to a different scoring mode. A message in row 44 will alert users to this.

The CCV scoring mode does not influence the ranking of proteins, but it adjusts the star-classifiers. A protein with 3-star classifier is an extremely strong candidate CCV protein (in fact, most of them are known CCV proteins). A 2-star classifiers denotes a very likely CCV protein; a 1-star classifier is still a likely CCV protein, but some false positives may be in this range; a few genuine CCV proteins will have B-classifiers, among many less relevant proteins.

*Special notes for the S2 Predictor:* The Drosophila clathrin heavy chain gene name is Chc.

11. Some notes on interpreting the output

The Predictor finds proteins with similar fractionation behaviour. Proteins in a stable complex tend to have very similar fractionation behaviour; but so do proteins which are in the same vesicle or membrane domain. Similar profiles may imply a physical association, but may also result from physical proximity without direct interaction. Furthermore, similar profiles can arise by chance; two protein particles with very similar size and density will have similar fractionation properties, and may thus not be resolved by the method. Similarly, if a protein is found in more than one subcellular compartment, its profile will show a hybrid behaviour which may be quite different from the profiles of the host compartments, unless one particular localisation predominates strongly. The profiles of integral membrane proteins may be difficult to interpret, since many of them travel on circular routes and are hence more or less spread out over the endomembrane system. Matched profiles of integral membrane proteins may suggest similar trafficking itineraries or steady-state localisations, rather than physical association. For example, many plasma membrane proteins have fairly similar profiles.

In general, the Predictor output should be treated like data from most other large-scale screening methods: as suggestive, but not as definitive. Predictions may guide future experiments, by suggesting candidate binding partners or functionally related proteins. In addition, the Predictor data may be used in conjunction with other data, such as yeast-two-hybrid screens; overlap between orthogonal methods is usually a strong indicator of useful data.
12. Troubleshooting & error messages

No retrieved profile: If on entering a gene name in the ‘Submit Query’ tab no profile is retrieved, it means that no gene of this name is in the database. It’s best to check if the gene is in the database, but has a different (UniProt) name. Go to the ‘Complete data’ tab and use Excel’s find function (Ctrl + F) to search for a keyword or partial gene name.

Unexpected predictions: Make sure the ‘Submit Query’ tab did not contain any previous searches in rows 5-54.

Error messages: Error messages will be displayed in the ‘Submit Query’ tab. Generally, there are two common types of problems: Queries that may lead to low-accuracy predictions, and parameter configurations that lead to no predictions.

The default configuration of the search parameters uses the highest possible stringency. Users may decrease the stringency (eg by allowing output proteins with incomplete (<6 data points) profiles), to increase the coverage/sensitivity of the analysis. The Predictor will warn users if any settings or query properties may reduce stringency.

Some parameter configurations prevent the Predictor from producing any output. For example, if your query profile has only five data points, but you specify that all output profiles must have at least 6 data points, no predictions will be made.

No protein name/ ‘0’ protein name: Some proteins do not have a UniProt protein name; by default, they are annotated as ‘0’. In addition, some proteins occur in multiple isoforms (eg splice variants) in the database. A feature of the MaxQuant 1.3.0.5 software (Cox and Mann, Nature Biotechnology 26, 1367- 1372, 2008), which was used to calculate the abundance ratios, is to assign no protein name to multiple variants of the same protein (only the first entry will get the protein name). Look at the Fasta header in the ‘Results Table - complete’ tab to identify proteins without protein name.

Special notes for the S2 Predictor: Many Drosophila proteins do not have a UniProt gene or protein name (annotated as ‘0’). To see the identities of these proteins, go the ‘Results Table - Complete’ tab and look at the Fasta header.

Tied ranks: Two proteins may be tied for rank, ie they may have exactly the same distance to the query profile. This is not a problem in recent versions of Excel (eg Excel 2013). Some older versions of the Excel RANK function however cannot deal with the situation, and an “#N/A” appears across the output cells in the ‘Results Table - complete’, and the ‘Results Top 20 – Scores’ plot shows no output. This may happen for example when a consensus profile from two queries is used to search the predictor (the consensus profile will be the exact average of the two queries, and they will hence have the same distance). A simple remedy is to enter the consensus query profile as a ‘user defined query’ in row 50 of the ‘Submit Query’ tab, and fractionally change one of the ratios (eg the 6th digit after the decimal point), to break the tie.
