**Aim**

*A good example is the best sermon*, claimed American polymath Benjamin Franklin. This worked-out example is intended as a step-by-step guide to assessing differential protein abundance from scratch using *iSanXoT*. For that, raw data from a high-throughput, quantitative proteomics experiment based on liquid chromatography coupled to tandem mass spectrometry (LC-MS/MS) analysis will be used.

**Project description and starting files**

|  |  |
| --- | --- |
| Table 1. **Isotopic sample labeling.** | |
| **Sample Name** | **iTRAQ label** |
| Jurkat WT 1 | 113 |
| Jurkat WT 2 | 114 |
| Jurkat WT 3 | 115 |
| Jurkat WT 4 | 116 |
| Jurkat KO 1 | 117 |
| Jurkat KO 2 | 118 |
| Jurkat KO 3 | 119 |
| Jurkat KO 4 | 121 |

We shall be working with a simple model consisting of four wild-type (WT) and four knock-out (KO) Jurkat cell preparations which we would like to compare for differential protein abundance. The protein extracts obtained from these eight samples were subjected to tryptic protein digestion, after which the resulting peptides were isotopically labelled using the iTRAQ Reagent 8 plex kit[[1]](#footnote-1) as depicted in Table 1. Then the labeled peptide samples were pooled and partly separated into five fractions using high-pH reversed-phase chromatography[[2]](#footnote-2). LC-MS/MS analysis of both the unfractionated material and the five peptide fractions resulted in the six raw LC-MS/MS files that can be found in …\WOE\Raw\_files. These raw files were used as inputs to Protein Discoverer 2.1[[3]](#footnote-3) to generate, among others, the six *PSM.txt* files stored in …\WOE\PSM.txt\_files upon database searching against the FASTA human protein database that can be found in …\WOE\Database\_and\_category\_files. Please note that this protein database contains both true protein sequences and the corresponding inverted (decoy) sequences to enable false discovery rate (FDR) estimation for peptide identification.

A close inspection of the *PSM.txt* files reveals that these plain text files hold 57 columns displaying information for every peptide-spectrum match (PSM) obtained upon database searching; however, only a reduced subset of these data will be accessed by *iSanXoT,* as shown in Table 2.

|  |  |
| --- | --- |
| Table 2. **Data accessed by iSanXoT in the *PSMs.txt* files.** | |
| **Header** | **Description** |
| Spectrum File | Name of the raw LC-MS/MS file |
| First Scan | Scan number identifier |
| Sequence | Peptide amino acid sequence |
| Modifications | Unimod[[4]](#footnote-4) chemical or posttranslational modifications to peptide sequence |
| Charge | Peptide charge |
| XCorr | Cross-correlation value as provided by the SEQUEST[[5]](#footnote-5) algorithm |
| MH+ [Da] | Measured monoisotopic protonated peptide mass in Da |
| Theo. MH+ [Da] | Theoretical monoisotopic protonated peptide mass in Da |
| DeltaM [ppm] | Difference between measured and theoretical monoisotopic mass in ppm |
| Protein Accessions | Accession codes for the proteins to which the peptide sequence is ascribed |
| 113-121 | Intensity of iTRAQ reporter ions 113-121 |

**Getting started: the *Inputs* menu**

Once the program has been successfully installed, executing the *isanxot.bat* script will bring you to *iSanXoT*’s main page, where a number of predefined workflows are displayed. The link *WSPP-SBT sample with iSanXoT databases* showing below the short description of the Basic Workflow will take you to the *Basic Workflow Inputs* window (Fig. 1), where the following information is provided:

* + *Input folder* specifies the location of the files containing PSM identification and quantification data (the *PSMs.txt* files in this example);
  + *Output folder* describes the path to the folder where *iSanXoT* output files will be stored. Selection of an *output folder* other than the *input folder* is strongly recommended;
  + *Select input files* indicates which of the *PSM.txt* files stored in the *input folder* must be considered by *iSanXoT* (all six files in this case). *PSMs.txt* file names are listed under Infile, while their experiment allocation is indicated in Experiment. All six PSMs.txt files originate from the same experiment (termed “TMT”) in our example.

**Validating peptide identification**

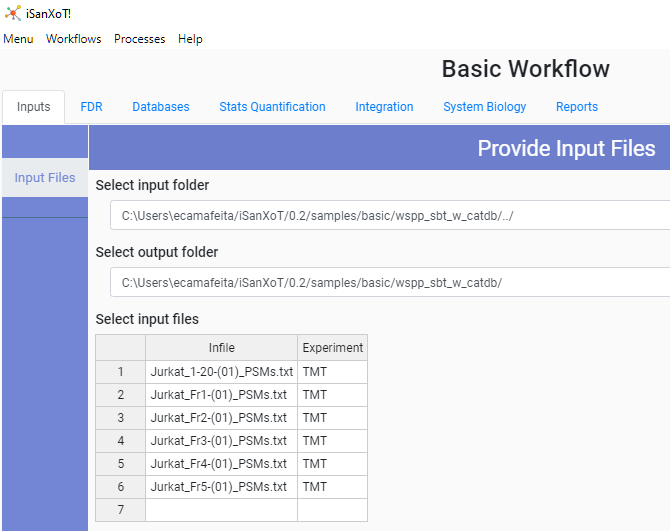
The next step is to validate the collection of PSMs that we have used as inputs in the previous Section. For that, iSanXoT relies on the probability ratio (pRatio) method[[6]](#footnote-6), which calculates the probability of random peptide matching and the corresponding FDR for peptide identification. The *FDR* menu displays the following fields:

Figure 1. **The Basic Workflow Inputs window after loading the sample data.**

* *Experiment*, the aforementioned experiment allocation;
* *Threshold (ppm)* isthe postscoring mass filtering cutoff to be applied after using wide mass windows in the database search[[7]](#footnote-7), as was the case with these sample data. Threshold is actually equivalent to the relative deviation experimentally observed for precursor ions in a particular LC-MS/MS run (15 ppm in this case);
* *Jump areas* is XXXXX. Possible values are 1, 3 and 5;
* *FDR* establishes the FDR cutoff for PSM validation. The value used here (0.01, *i.e.* 1% FDR) implies that one in every 100 validated PSMs is wrong;
* *XCorr type* determines whether XCorr (SEQUEST cross correlation score) or the corrected XCorr (cXCorr) will be used by pRatio for FDR calculation;
* *Decoy label* is the tag attached to decoy protein identifiers in the concatenated protein database used for peptide identification (INV\_ in this example).

**Relations matter: the *Databases* menu**

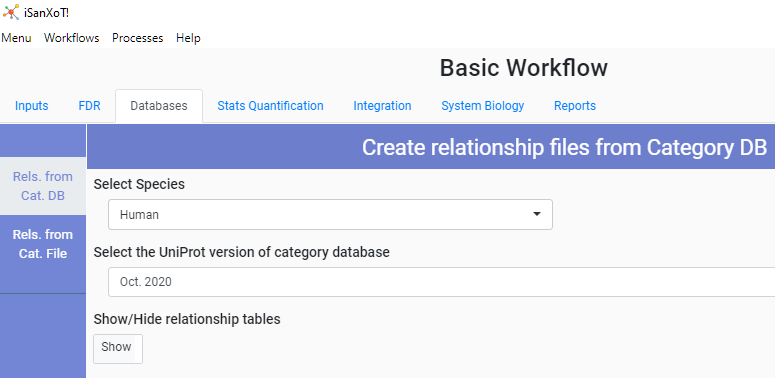
Two databases are necessary for *iSanXoT* to work: a protein sequence database and a protein-category relation database. The latter, a plain text file relating proteins to functional categories, is used by the Systems Biology Triangle (SBT) algorithm to estimate the experimental protein variance[[8]](#footnote-8). But don’t panic (yet): since you arrived here through the *WSPP-SBT sample with iSanXoT databases* link, your *Databases* menu should look like the one on Fig. 2. This means that you are having *iSanXoT* use its own databases downloaded from Uniprot, and therefore all you have to do is to select a species (human in this case, since we are using Jurkat cell preparations) from the *Select Species* drop-down menu and a database version from the *Select version of Uniprot category database* menu.

Figure 2. **The Basic Workflow Databases window in the WSPP-SBT sample with iSanXoT databases option.**

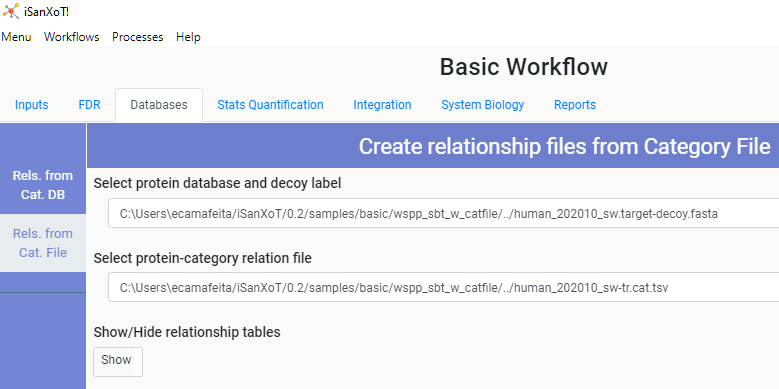
 Nevertheless, if for some reason you want to use your own databases, this is indeed possible. Select *Menu* on the top left corner and afterwards *Main Page* from the drop-down menu. Then follow the link *WSPP-SBT sample with user-provided databases* showing in the *Basic Workflow* field. In the *Databases* menu, both the *Select Species* and the *Select version of Uniprot category database* fields are now empty; but if you activate the *Rels. From Cat. File* option on the left, the *Protein database* and *Protein-category relation file* fields will display the path to a protein database and a protein-category relation file, respectively, stored in your hard drive (Fig. 3).

Figure 3. **The Basic Workflow Databases window in the WSPP-SBT sample with user-provided databases option.**

1. Ross P.L. *et al*. (2004) Multiplexed protein quantitation in Saccharomyces cerevisiae using amine-reactive isobaric tagging reagents*. Mol. Cell. Proteomics* **3**, 1154–69. [↑](#footnote-ref-1)
2. Stein D.R. *et al*. (2013) High pH reversed-phase chromatography as a superior fractionation scheme compared to off-gel isoelectric focusing for complex proteome analysis. *Proteomics* **13**, 2956-66. [↑](#footnote-ref-2)
3. https://www.thermofisher.com/order/catalog/product/OPTON-30810#/OPTON-30810. [↑](#footnote-ref-3)
4. https://www.unimod.org/ [↑](#footnote-ref-4)
5. Eng J.K. *et al.* (1994) An approach to correlate tandem mass spectral data of peptides with amino acid sequences in a protein database. *J. Am. Soc. Mass Spectrom*. **5**, 976–89. [↑](#footnote-ref-5)
6. Martínez-Bartolomé S. *et al*. (2008) Properties of average score distributions of SEQUEST: the probability ratio method. *Mol. Cell. Proteomics* **7**, 1135-45. [↑](#footnote-ref-6)
7. Bonzon-Kulichenko E. *et al*. (2015) Revisiting Peptide Identification by High-Accuracy Mass Spectrometry: Problems Associated with the Use of Narrow Mass Precursor Windows. *J. Proteome Res.* **14**, 700–10. [↑](#footnote-ref-7)
8. García-Marqués F. *et al*. (2016) A Novel Systems-Biology Algorithm for the Analysis of Coordinated Protein Responses Using Quantitative Proteomics. *Mol. Cell. Proteomics* **15**, 1740-60. [↑](#footnote-ref-8)