

Data Capture and Data Analysis

Hupo Plasma Proteome Project workshop
Bethesda, July 2003, Henning Hermjakob, EBI

Aims

- **Ensure data comparability to allow**
 - **Comparative analysis of results**
 - **Presentation of results**
 - **User-friendly public access to results**
- **Ensure data quality**

Overview

- **Data submission review**
 - Submission tools
 - Database standardization
 - Representation of PTMs
 - Future perspectives
- **Data analysis**

Data submission tools

- **Excel spreadsheets**
 - Easy, accessible technology
 - More effort for central analysis
 - Sent out in June:
 - [Protein summary](#)
 - [Abundance summary](#)
 - [Resources summary](#)

Data submission tools

- **XML-based:**
 - Better data integration
 - Easier central analysis
 - Better for data presentation
 - Input tools: Pedro
 - Experimental status

Database standardisation

- **Comparison of results obtained from searches against different databases is difficult**
- **Proposal: Define one default database as a basis for all searches**
- **Proposal: IPI**

International Protein Index (IPI)

- **Merged set of all major protein sequence data sources**
- **First created for:**
Initial sequencing and analysis of the human genome.
Lander, ES., et al., Nature. 2001 Feb 15;409(6822):860-921.
- **Monthly updated**
- **Stable, versioned identifiers**
- **Detailed documentation**
- **Statistics: 56530 human entries as of July 2.**
- **<http://www.ebi.ac.uk/IPI>**

IPI formats

- **Fasta format:**

```
>IPI:IPI00000005.1|SWISS-PROT:P01111-3|REFSEQ_NP:NP_002515|TREMBL:P54111|  
REFSEQ_XP:XP_032698;XP_001317|ENSEMBL:ENSP00000261444 Tax_Id=9606 Transforming  
protein N-Ras  
MTEYKLVVVGAGGVGKSALTIQLIQNHFVDEYDPTIEDSYRKQVVIDGETCLLDILDITAG  
QEEYSAMRDQYMRTGEGFLCVFAINNSKSFADINLYREQIKRVKDSDDVPMVLVGNKCDL  
PTRTVDTKQAHELAKSYGIPFIETSAKTRQGVEDAFYTLVREIRQYRMKKLNSSDDGTQG  
CMGLPCVVM
```

- [Swiss-Prot style format](#)

IPI algorithm outline

- **Expand all annotated SP splice variants into separate entries**
- **Inter-database similarity searches**
- **Clustering based on pairwise reciprocal best matches and subfragment matches (95% cutoff)**
- **Each cluster represents one IPI entry**
- **Identifier tracking based on cluster member accession numbers**

Representation of PTMs

- **How much detail do we need?**
- **High-detail proposal:
PSI PTM classification:**
 - **Controlled vocabulary for PTMs in GO format**
 - **Covers both general (“phosphorylation”) and detailed PTMs**
 - **Fully cross-referenced with RESID database**

Future work

- **Controlled vocabulary development:
Hierarchical classification of experimental techniques**
- **Representation of complexes and interactions**
- **Partially done for Protein Interactions within HUPO Proteomics Standards Initiative**

Data analysis

- **How to verify protein identifications?**
- **How to integrate cross-specimen, cross-laboratory, and cross-technology data?**

Acknowledgements

- **Richard Simpson, Ludwig Institute for Cancer Research**
- **Paul Kersey, EBI**
- **Luisa Montecchi-Palazzi, University of Rome**
- **Rolf Apweiler, EBI**

- **You!**

Questions for the Data Management/Analysis Breakout Session

Data Management/Analysis Session

- **Timeline:**
 - **XML schema by Monday, July 21**
 - **Preliminary data submissions until September 15**
 - **Data mapping and joining**
 - **Comparative analysis until HUPO congress**
 - **More detailed round for Jambouree Spring 2004**
- **HELP!!!!**

Data Management/Analysis Session

- **Format questions:**
 - Shared database possible?
 - Which one?
 - PTM representation?
- **Access to joint dataset:**
 - Only for PPP consortium until Spring 2004?
 - To all as soon as possible?

Data Management/Analysis Session

- **Quality checks:**

- Which checks to run?
- Who?
- What needed?

- **Analysis on joined set**

- Which questions do you want to ask?
- Do you want to participate?
- In which form?
- **Teaser: Clustering identification lists, do you get stronger correlation with samples or labs?**

Data Management group report

- **Modifications to the Excel forms:**
 - **Add mass, pi, intensity (2D)**
 - **Confidence (high, low)**
- **XML preferred format, Excel ok**

Data Management group report

- **Data submission support:**
 - **Univ. Manchester and Yale offer assistance with formatting/writing Mascot parser**

Data Management group report

- **Timeline:**

- **Next week: XML schema and documentation**
- **August 15: Test submission**
- **September 15: Real submission**
- **October 10: Preliminary analysis results**

Data Management group report

- **Analysis aims:**
 - **How/Which proteins identified by abundance**
 - **Consistency of identification**
 - **Comparison of samples**
 - **Comparison of methods**
 - **Consistency of abundance (by rank order)**

Data Management group report

- **Open questions:**
 - **8 labs indicated “Seldi only”**
 - **This data is not captured by the proposed spreadsheet**
 - **We need Seldi experts!**

Data Management group report

- **Open questions:**
 - **Data release:**
 - **Data will be anonymised, but be aware it can be “attributed” to sources by experts**
 - **Publications will be anonymous**
 - **“First right” of publication for participants.**