

Kapitel 5: Protein-Datendanken

n Motivation und historische Entwicklung

n Proteomics

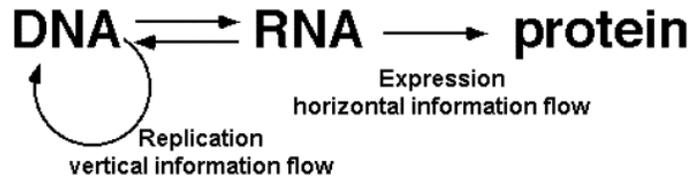
- Datengewinnung
- PEDRo-Projekt

n Protein-Datenbanken

- Anforderungen
- Sequenz-Datenbanken
- Domain/Familien-Datenbanken
- Struktur-Datenbanken



Vom Gen zum Protein



	U	C	A	G	
U	UUU Phenylalanine UUC UUA Leucine UUG	UCU Serine UCC UCA UCG	UAU Tyrosine UAC UAA Stop codon UAG Stop codon	UGU Cysteine UGC UGA Stop codon UGG Tryptophan	U C A G
C	CUU Leucine CUC CUA CUG	CCU Proline CCC CCA CCG	CAU Histidine CAC CAA Glutamine CAG	CGU Arginine CGC CGA CGG	U C A G
A	AUU Isoleucine AUC AUA Methionine, initiation codon AUG	ACU Threonine ACC ACA ACG	AAU Asparagine AAC AAA Lysine AAG	AGU Serine AGC AGA Arginine AGG	U C A G
G	GUU Valine GUC GUA GUG	GCU Alanine GCC GCA GCG	GAU Aspartic acid GAC GAA Glutamic acid GAG	GGU Glycine GGC GGA GGG	U C A G

Alanin	Ala	A	Serin	Ser	S
Valin	Val	V	Threonin	Thr	T
Phenylalanin	Phe	F	Tyrosin	Tyr	Y
Prolin	Pro	P	Histidin	His	H
Methionin	Met	M	Cystein	Cys	C
Leucin	Leu	L	Asparagin	Asn	N
Isoleucin	Ile	I	Glutamin	Gln	Q
Aspartat	Asp	D	Tryptophan	Trp	W
Glutamat	Glu	E	Glycin	Gly	G
Lysin	Lys	K			
Arginin	Arg	R			

hydrophob
 geladen
 polar



Proteine

n Proteinfunktionen

- Genregulation
- Verdauung / Metabolismus, enzymatische Steuerung
- Signalverarbeitung
- Zellstruktur und Organellen, Zellteilung, Vermehrung

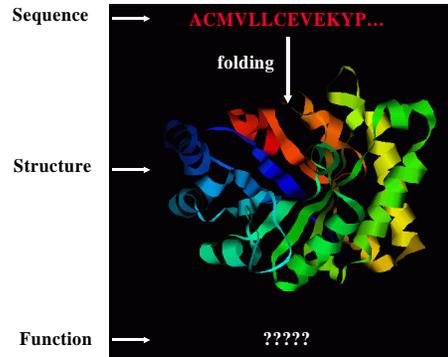
n Proteine im Menschen

- Durchschnittlich 447 Aminosäuren lang
- Kürzestes [Swiss-Prot]: ~ 40-50 Aminosäuren
- Längstes [SP]: ~ 8.700 Nesprin (Cytoplasma), [Swiss-Prot]: ~ 6.669 Nebulin (Muskel), [EMBL]: 34.350 Titin (Muskel)

n Die Funktion von 40-50% der Proteine ist unbekannt!

n Wissen über Proteine u.a. wichtig für

- Krankheitsverständnis
- Medikamenten-Design



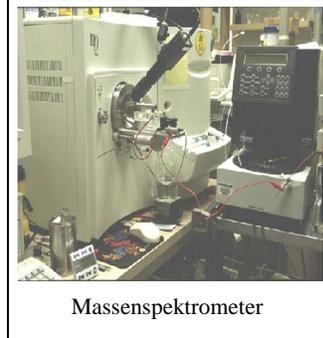
Protein-Sequenzen

- n Länger bekannt und untersucht als DNA-Sequenzen, da labortechnisch einfacher zugänglich
- n Erste Proteinsequenz 1951 (Sanger & Tuppy): Seitenkette von Insulin
- n Systematische Sammlung ab Anfang der 1960er (Dayhoff et al. 1965)
 - Protein Sequence Atlas: Buchform, 1968-1978
 - Motivation: Evolutionäre Untersuchungen
 - 1980: Protein Information Resource (seit 1988: PIR-International)
 - 1986: Swiss-Prot (Genf)



Proteomics

- n Ausgangsproblematik: Gensequenzen (und damit AS-Sequenzen) informieren nicht über
 - Proteinfunktionen
 - Protein-Protein-Interaktionen, Multiprotein-Komplexe
 - Post-translationale Modifikationen (z.B. bei Prionen)
- n Ziel von Proteomics: Bestimmung aller Proteine(komplexe) und ihrer Funktionen *
- n Wichtigstes Verfahren: Massenspektroskopie
- n Problematik
 - Noch keine öffentlichen Repositories für Spektren
 - Noch keine Standards zur Beschreibung von Experimenten und Protokollen
- n Derzeitige (noch nicht abgeschlossene) Standardisierungsmaßnahmen
 - HUPO: Human Proteome Organisation (mit wichtiger Unterorganisation JHUPO: Japan HUPO; Entwicklung von HUP-ML: Human Proteome Markup Language)
 - PSI: Proteomics Standard Initiative
 - PEDRo



* siehe auch: <http://us.expasy.org/> (ExPASy Molecular Biology Server)



Proteomics: Datengewinnung

The nature of proteomics experiment data

- **Sample generation**

- *Origin of sample*
 - hypothesis, organism, environment, preparation, paper citations

- **Sample processing**

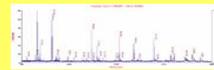
- *Gels (1D/2D), columns, other methods*
 - images, gel type and ranges, band/spot coordinates
 - stationary and mobile phases, flow rate, temperature, fraction details

- **Mass Spectrometry**

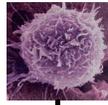
- machine type, ion source, voltages

- **In Silico analysis**

- peak lists, database name + version, partial sequence, search parameters, search hits, accession numbers



Proteomics: Workflow



Proteomics Workflow

Proteinextraktion

Aufbereitung aller Proteine einer Zelle

Proteintrennung

2D Gel-Elektrophorese

Proteinisolierung

Ausschneiden

Proteinidentifikation

Massenspektroskopie
Sequenzierung

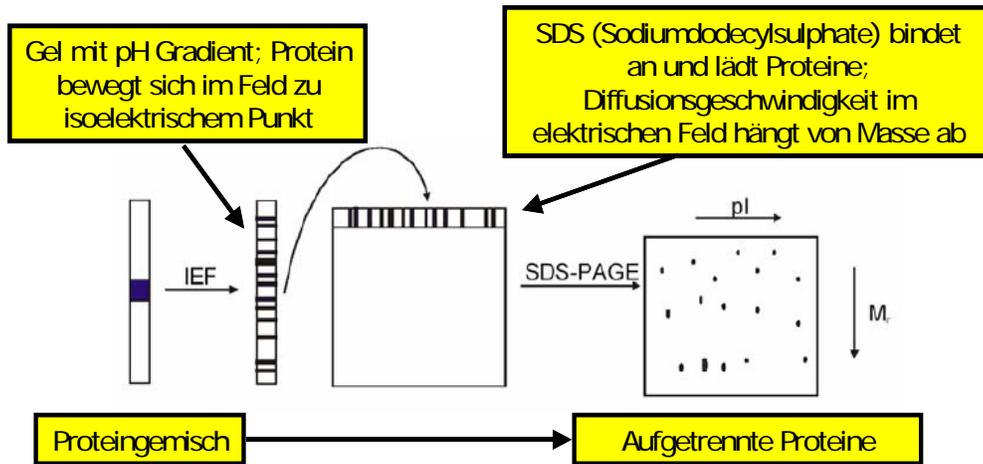
Analyse

Funktion, Struktur, Interaktion, ...



2D Gel-Elektrophorese

- Zweidimensionale Trennung von Proteinen
 - 1. Dimension: Ladung
 - 2. Dimension: Masse

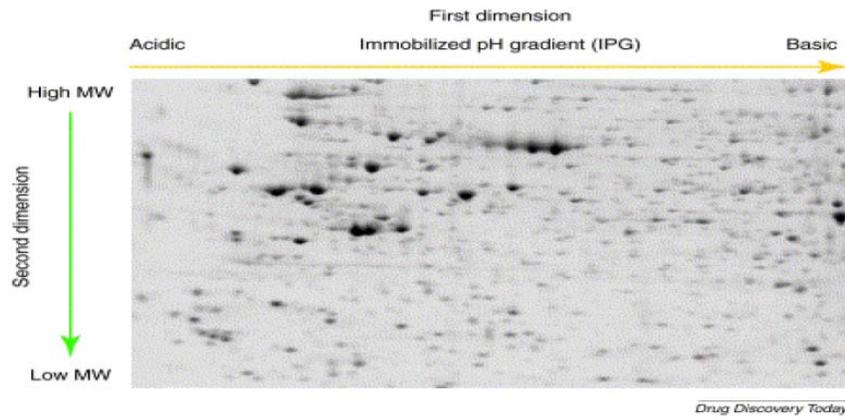


Isoelektrischer Punkt

- pH Wert einer Lösung = Anzahl freier H^+
 - Sauer = niedriger pH Wert = viele H^+ und wenige OH^-
 - Basisch = hoher pH Wert = wenig H^+ und viele OH^-
 - Neutral = $pH=7 = H^+$ und OH^- Gruppen ausgeglichen
- Proteine können an vielen Stellen H^+ verlieren (deprotonated) oder gewinnen (protonated)
- Protein in Lösung (mit bestimmten pH-Wert) verliert oder gewinnt Protonen
- Der **pKA Wert** eines Proteins gibt an, bei welchem pH Wert 50% des Proteins protonated sind
- **Isoelektrische Punkt**: pH-Wert, an dem das Protein Netto (im Vergleich zur Lösung) keine Ladung mehr hat
 - und sich damit auch nicht mehr bewegt



2D Gel-Ergebnisse

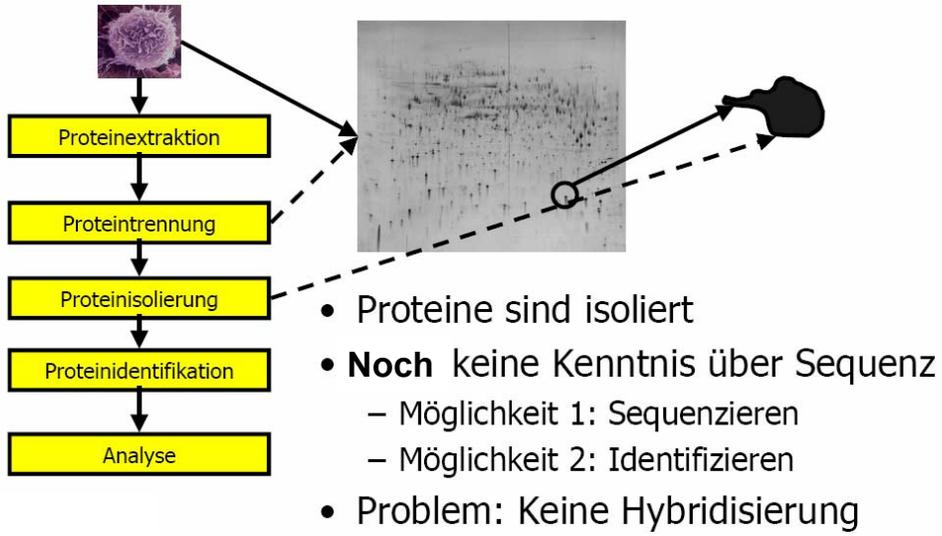


2D Gel Methode

- Einfache und billige Methode, sehr weit verbreitet
- Trennung bis zu 10.000 Proteine möglich
- Vergleich von Bildern begrenzt möglich (Gesund–krank)
- Nachteile
 - Ausschneiden aufwändig (manuell)
 - Einschränkungen
 - Keine Proteine <20KD oder >200KD
 - Keine extrem geladenen Proteine (sehr sauer / sehr basisch)
 - Schwierig bei geringen Konzentrationen (Low abundance)
 - Keine Membranproteine (Anderes Gel-Verhalten)
 - Keine Identifikation von Proteinen



Nächste Schritte



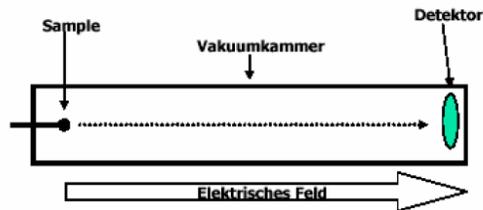
1. Sequenzierung, Edman Degradation

- Verfahren seit ca. 1980 bekannt
- Prinzip
 - Protein in hochreiner Konzentration vorhanden
 - Enzymatische Trennung einer Aminosäure vom N-Terminus
 - Identifikation durch chromatographische Verfahren
 - Zyklische Wiederholung
- Nachteile
 - Lange Dauer – ca. 30-60 Minuten pro Zyklus
 - Für Hochdurchsatz nicht verwendbar
 - Aber: wichtig zur Qualitätskontrolle



2. Identifikation: Massenspektroskopie

- Prinzipielle Idee
 - Beschleunigung von Proteinen in elektrischem Feld
 - Detektor misst Auftreffen der geladenen Teilchen (Ionen)
 - Flugzeit proportional zu Verhältnis Masse / Ladung (m/z)



Probleme

- Proteine empfindlich
 - Verdau
- Zu geringe Ladung
 - Ionisierung



Schritt 1: Verdau

- Problem: Proteine zu zerbrechlich für MS
- Lösung
 - Proteine vorab enzymatisch in Peptide zerbrechen
 - Enzymatischer Verdau
 - Peptide mit Massenspektroskop messen
 - Originalprotein aus Kombination der gemessenen Peptide bestimmen
- Annahme
 - Jedes Protein hat eindeutige Peptidsignatur (Fingerprint)
- Viele Peptidasen bekannt

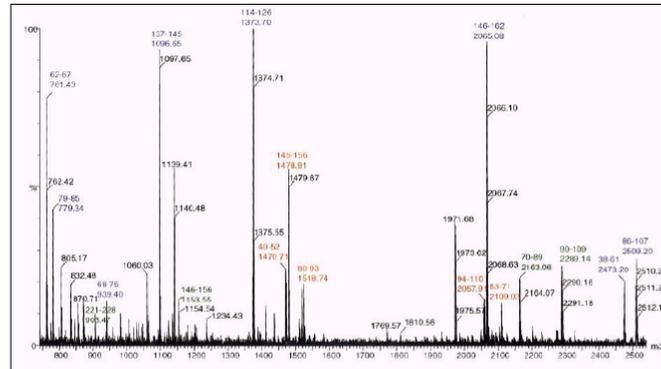


2. Ionisierung

- Problem: Peptide oft ohne Ladung – keine Beschleunigung
- Lösung
 - MALDI – Matrix Assisted Laser Desorption/Ionisation
 - Peptide in „Matrix“ einbetten – Kristallisierung mit lichtempfindlichen, geladenen Molekülen
 - Kristall mit Laser beschießen
 - Lichtempfindliche Moleküle verdampfen und reißen ionisierte Peptide in Gasphase mit
 - Beschleunigen in MS



Ergebnis der MS

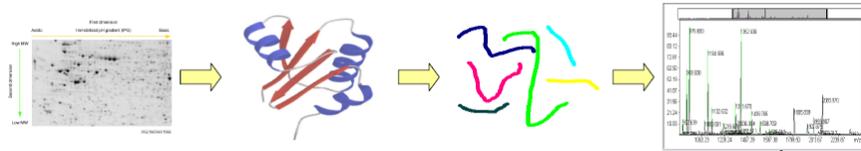


- Jedes Peptid ist ein Peak
- Peakhöhe mit heutiger Technik nicht relevant
- Algorithmisches Problem: Protein aufgrund des Peak-Fingerprints bestimmen

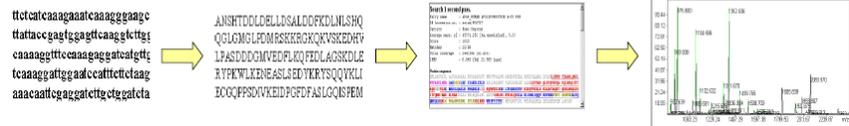


Proteinidentifikation

Experimentelle Messung



Theoretische Vorhersage



Vergleich



Probleme

- Geringe Gewichtsänderungen
 - Isotope – Peptidmassen sind nicht fest
 - Modifikationen: Phosphorylierung, Glycosylierung, ...
 - Messfehler, ungenaue MS-Kalibrierung
- Probleme/Fehler in Datenbanken
 - Frameshifts in DNA – Protein Übersetzung
 - SWISS-PROT speichert Consensussequenzen – ohne Variationen. Besser: Non-Redundant Datasets NLR-3D, OWL, ...
- Statistischer Bias
 - Proteinlänge nicht berücksichtigt
 - Lange Proteine haben höhere Grundwahrscheinlichkeit, ein bestimmtes Peptid zu enthalten
 - Relative Häufigkeit von Peptiden nicht beachtet
- Keine Einschätzung der Güte des Ergebnis
 - Keine Garantie, das Spektrum in DB enthalten



Praktische Algorithmen

➤ Heuristische Korrekturfaktoren

- MOWSE
- Matches mit Score
- Beachtung Peptidhäufigkeit und Proteinmasse

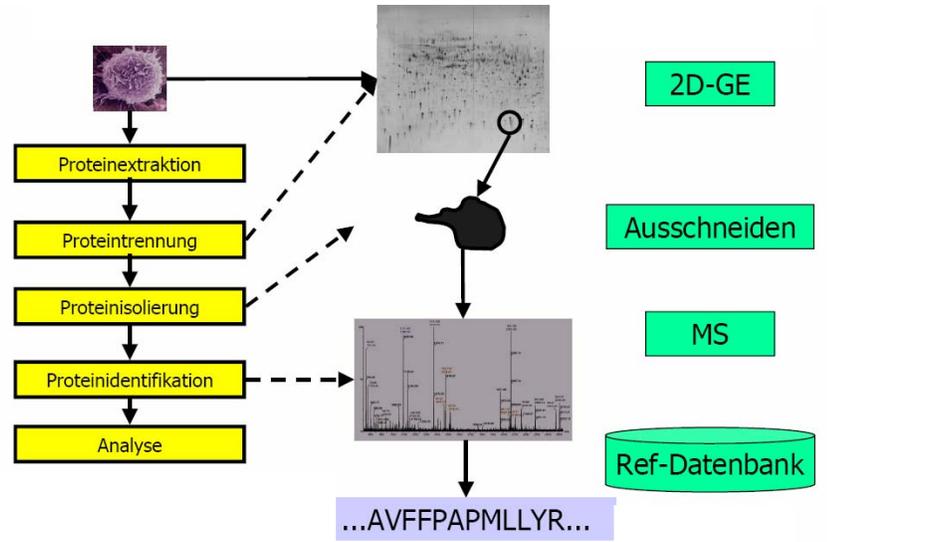
➤ Wahrscheinlichkeitsbasierte Verfahren

- ProFound
- Bayes-basiert
- Beachtung Messfehler, Proteingröße und Peptidhäufigkeiten

- Diverse weitere Algorithmen:
MASCOT, PeptIdent, ProteinProspector, ...



Kompletter Workflow

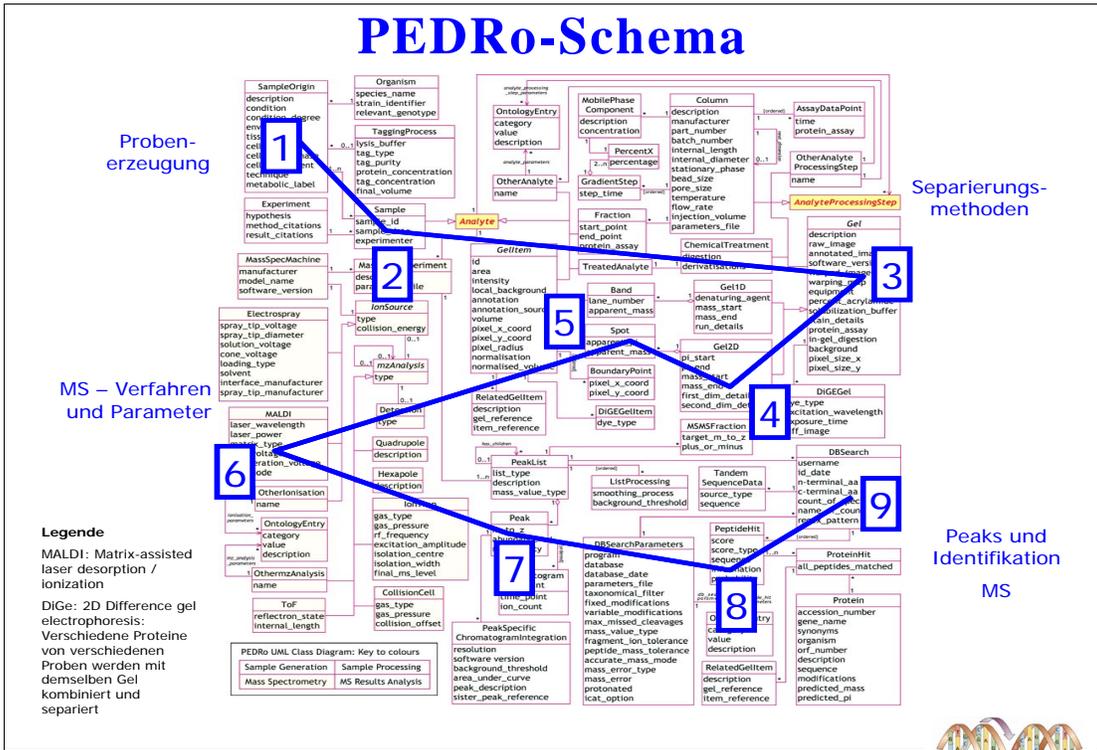


PEDRo

- n Proteomics Experiment Data Repository
- n Systematischer Ansatz zur
 - Modellierung
 - Speicherung und
 - Distributionvon Proteomics-Experimentaldaten und deren Interpretation
- n University Manchester (Taylor, Paton & Goble)
- n <http://pedro.man.ac.uk/home.shtml>



PEDRo-Schema



(C) Prof. R. Müller, Prof. E. Rahm



PEML: Proteomics Markup Language

```
<xs:element name="Protein">
  <xs:complexType>
    <xs:sequence>
      <xs:element name="accession_number" type="xs:string"/>
      <xs:element name="gene_name" type="xs:string" minOccurs="0"/>
      <xs:element name="synonyms" type="xs:string" minOccurs="0"/>
      <xs:element name="organism" type="xs:string" minOccurs="0"/>
      <xs:element name="orf_number" type="xs:string" minOccurs="0"/>
      <xs:element name="description" type="xs:string" minOccurs="0"/>
      <xs:element name="sequence" type="xs:string" minOccurs="0"/>
      <xs:element name="modifications" type="xs:string" minOccurs="0"/>
      <xs:element name="predicted_mass" type="xs:decimal" minOccurs="0"/>
      <xs:element name="predicted_pi" type="xs:decimal" minOccurs="0"/>
    </xs:sequence>
  </xs:complexType>
</xs:element>
<xs:element name="ProteinHit">
  <xs:complexType>
    <xs:sequence>
      <xs:element name="all_peptides_matched" type="xs:boolean" minOccurs="0"/>
      <xs:element name="component_peptides" type="xs:string" minOccurs="0"/>
      <xs:element name="masses_matched" type="xs:string" minOccurs="0"/>
      <xs:element name="score" type="xs:string" minOccurs="0"/>
      <xs:element name="score_type" type="xs:string" minOccurs="0"/>
      <xs:element ref="Protein"/>
      <xs:element ref="RelatedGelItem" minOccurs="0" maxOccurs="unbounded"/>
    </xs:sequence>
  </xs:complexType>
</xs:element>
```



PEDRo: SQL-Schema (Ausschnitt)

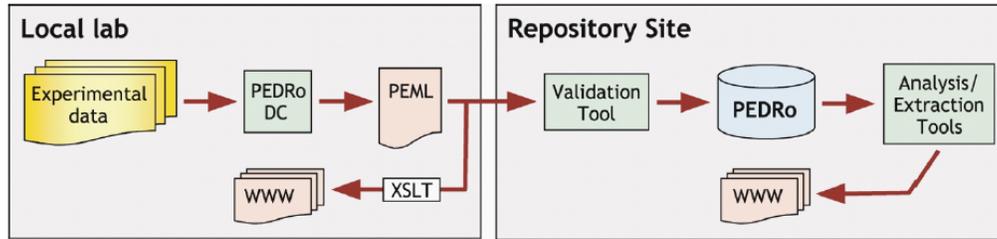
```
CREATE TABLE Experiment /* Describes the overall motivation for the proteomics experiment */
(
  id          integer PRIMARY KEY          ,
  hypothesis  varchar(500) NOT NULL        , /* Summary of the motivation for the work */
  method_citations varchar(200)          , /* References to method paper(s) */
  result_citations varchar(200)          /* References to results paper(s) */
);

CREATE TABLE Sample /* Identifiers for the sample to be analysed */
(
  sample_id   varchar(50) PRIMARY KEY      , /* Unique (lab assigned) identifier */
  sample_date date NOT NULL                , /* Date on which sample was obtained */
  experimenter varchar(200) NOT NULL       , /* Name of experimenter who produced the sample */
  experiment  integer REFERENCES Experiment ON DELETE CASCADE
);

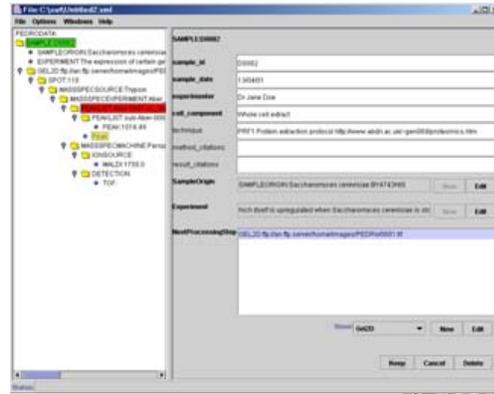
CREATE TABLE Organism /* Identifiers for the organism used */
(
  id          integer PRIMARY KEY          ,
  species_name varchar(100) NOT NULL       , /* Full systematic name of species */
  strain_identifier varchar(100) NOT NULL  , /* Identification string for particular strain */
  relevant_genotype varchar(200)          /* List of relevant gene names */
);
```



PEDRo: Datenfluss



- PedroDC: Pedro Graphical User Interface
- XSLT: eXtensible Stylesheet Language for Transformation



PEDRo: Bewertung

- n Noch keine Vergleiche / Erfahrungsberichte verfügbar
- n Keine Modellierung der Dynamik der Prozesse
 - Constraints
 - Zustandsübergänge
 - Workflow
- n Kein Standard (PEDRo ist ein Versuch)
- n Unklarer Nutzen vieler Metadaten
 - Experimente sind oft nicht vergleichbar
 - Und werden auch bei genauerer Beschreibung nicht vergleichbar

