20.453J / 2.771J / HST.958J Biomedical Information Technology
Fall 2008

Singapore–MIT Alliance

SMA 5304 Term Project Presentation

Constructing a Conformational Space of Pro-Ser-Thr Rich Non-Globular Domains

Liu Chengcheng
HT081976J
CSB,SMA

Outline

- Rationale

  Data integration

  Software

- Proposed Approach
- Proposed Architecture

  Data Retrieving

  Format Converter

  Data ontology

  Data Loading&wearhouse

Rationale

- Phosphorylation on serine/threonine plays an important role in signaling pathways.

- Checkpoint proteins contain sequence motifs (in P-S-T-rich domain) bound by SH2, SH3 etc.

- Data integration including usage of certain software to achieve a conformational space of such domain.

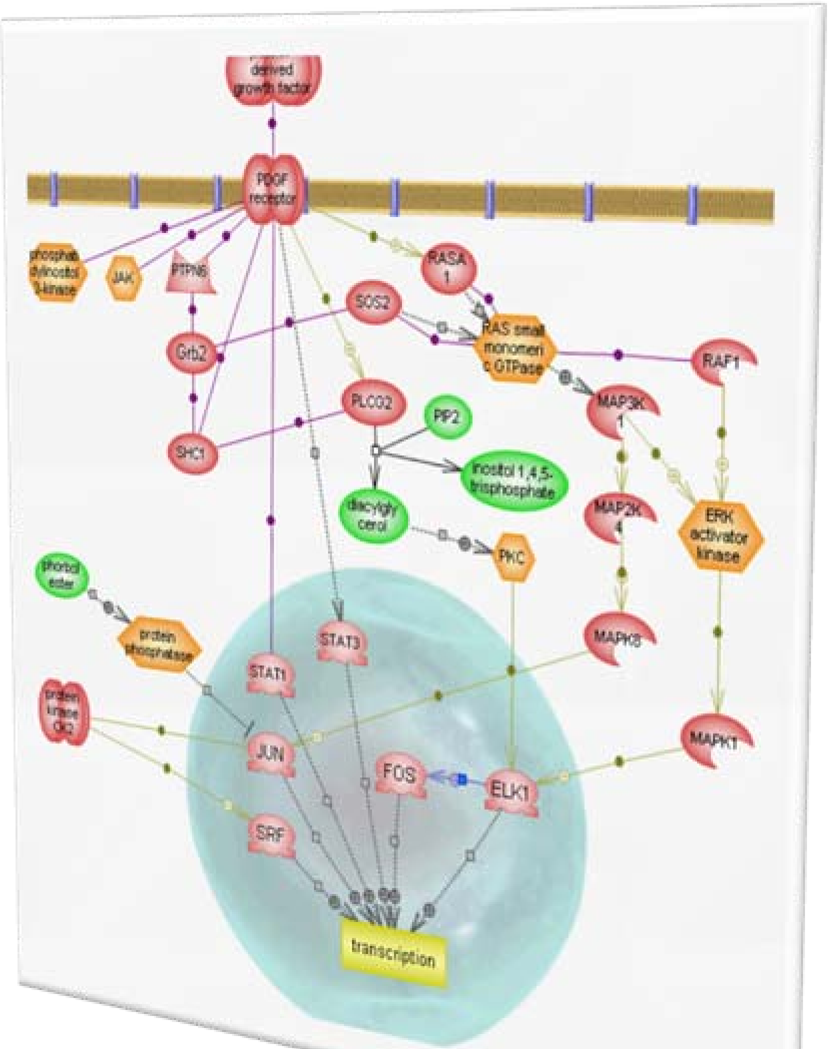


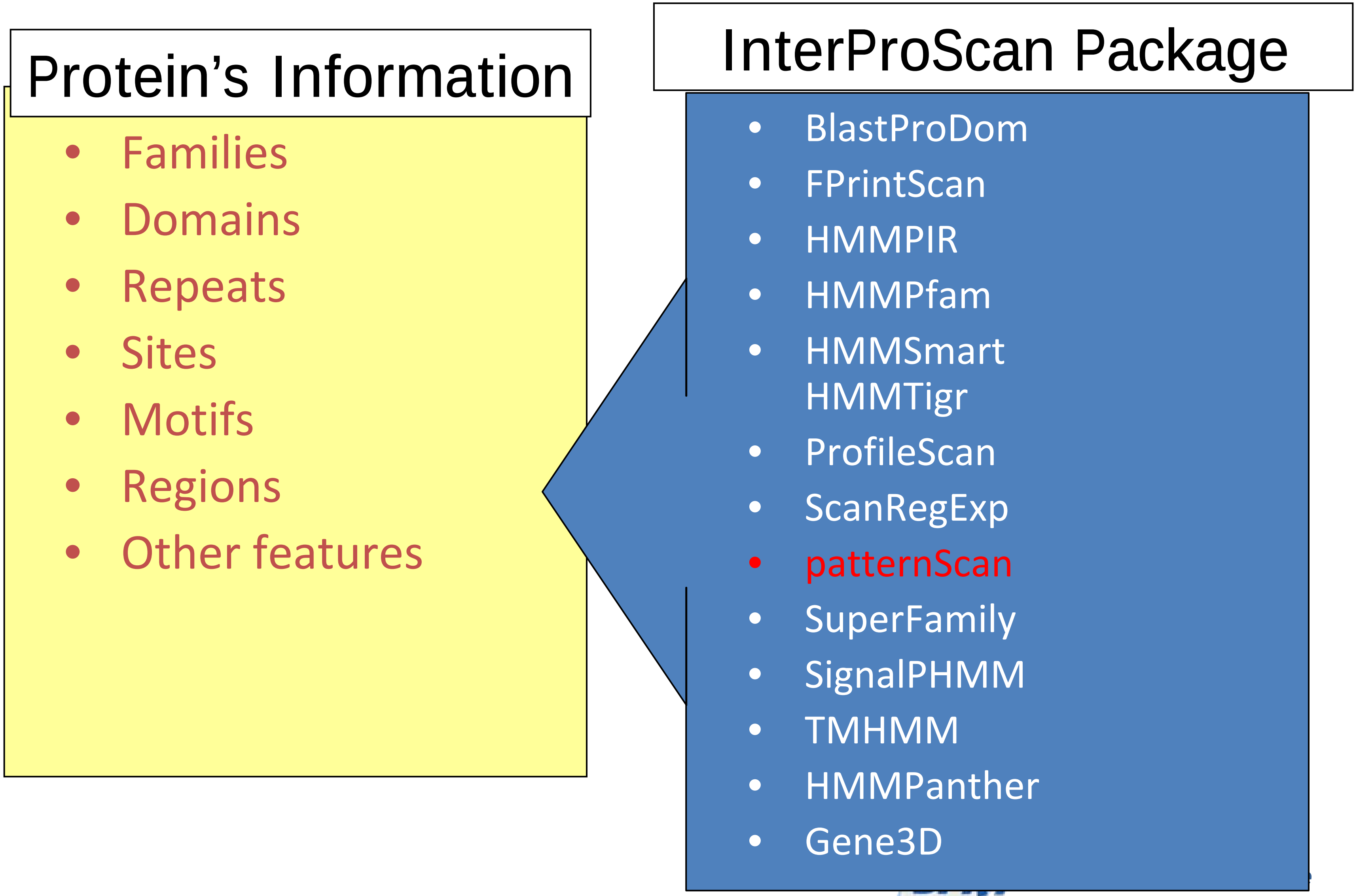
Diagram generated
in Pathway Studio®

Data Integration

Logos removed for copyright reasons.

EBI: http://www.ebi.ac.uk/
HSSP: http://swift.cmbi.kun.nl/swift/hssp/
RCSB PDB: http://www.rcsb.org/pdb/
UniProt: http://www.uniprot.org/
PIR: http://pir.georgetown.edu/pirwww/
RefSeq: http://www.ncbi.nlm.nih.gov/RefSeq/

Software for domain/motif scanning



Protein’s Information

- Families
- Domains
- Repeats
- Sites
- Motifs
- Regions
- Other features

InterProScan Package

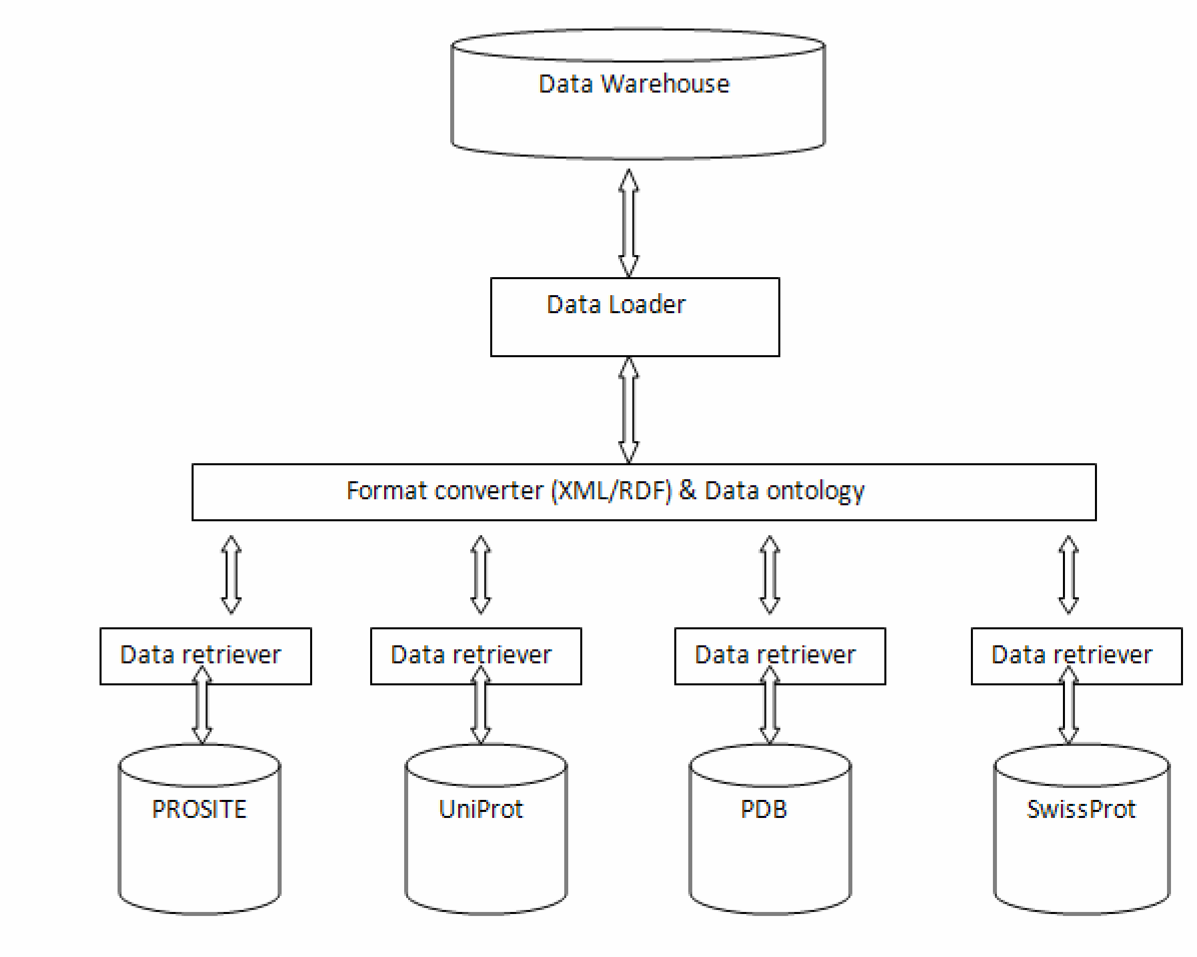
- BlastProDom
- FPrintScan
- HMMPIR
- HMMPfam
- HMMSmart
  HMMTigr
- ProfileScan
- ScanRegExp
- patternScan
- SuperFamily
- SignalPHMM
- TMHMM
- HMMPanther
- Gene3D

Proposed Approach: Data Warehousing

Image removed due to copyright restrictions.
See Fig. 5 in: Stein, L. D. "Integrating Biological Databases." *Nature Reviews Genetics* 4 (May 2003): 337-345. doi:10.1038/nrg1065.

NATURE REVIEWS| GENETICS VOLUME 4 | MAY 2003 |337-345

Proposed Architecture

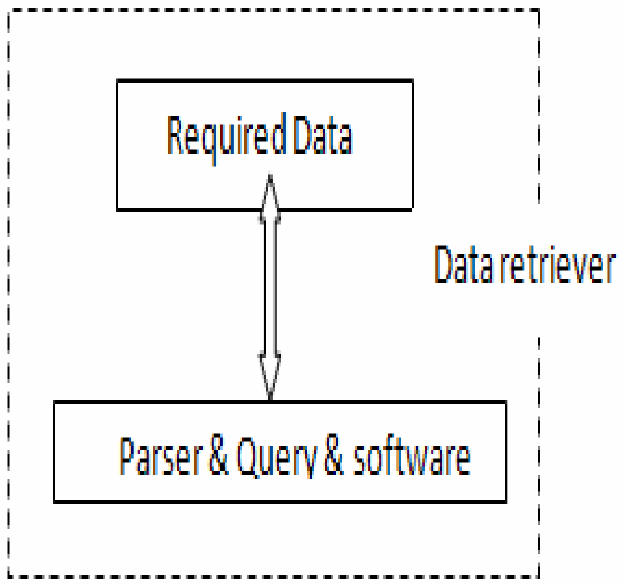


Data Retrieving

- Each database has a data retriever
- Parse and query the raw data

  (e.g. Nux Java toolkit & XQuery)
- Include using software

  (e.g. patternScan)

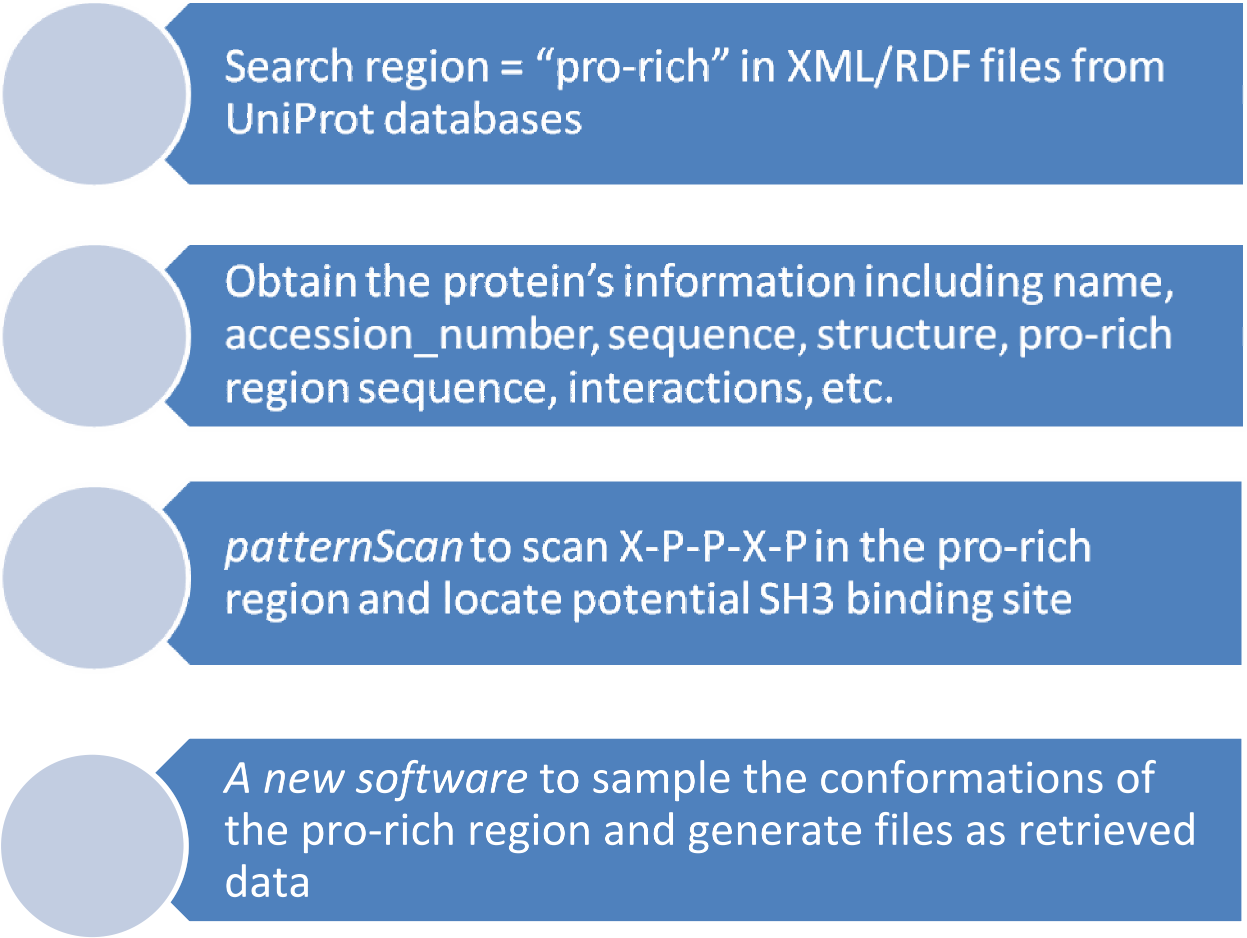


Example: retrieve data from UniProt

- Start with a single modular domain e.g. SH3 (Src Homology domain)
- About 60 amino-acid residues in several cytoplasmic protein tyrosine kinases (e.g. Src, Abl)
- five or six β-strands arranged as two tightly packed anti-parallel β sheets
- Binding to Pro-rich domain
- Binding pattern X-P-P-X-P or R-X-X-K

Image removed due to copyright restrictions.
“Ribbon diagram of the SH3 diagram, alpha spectrin, from chicken.”
http://en.wikipedia.org/wiki/File:1shg_SH3_domain.png

Flowchart



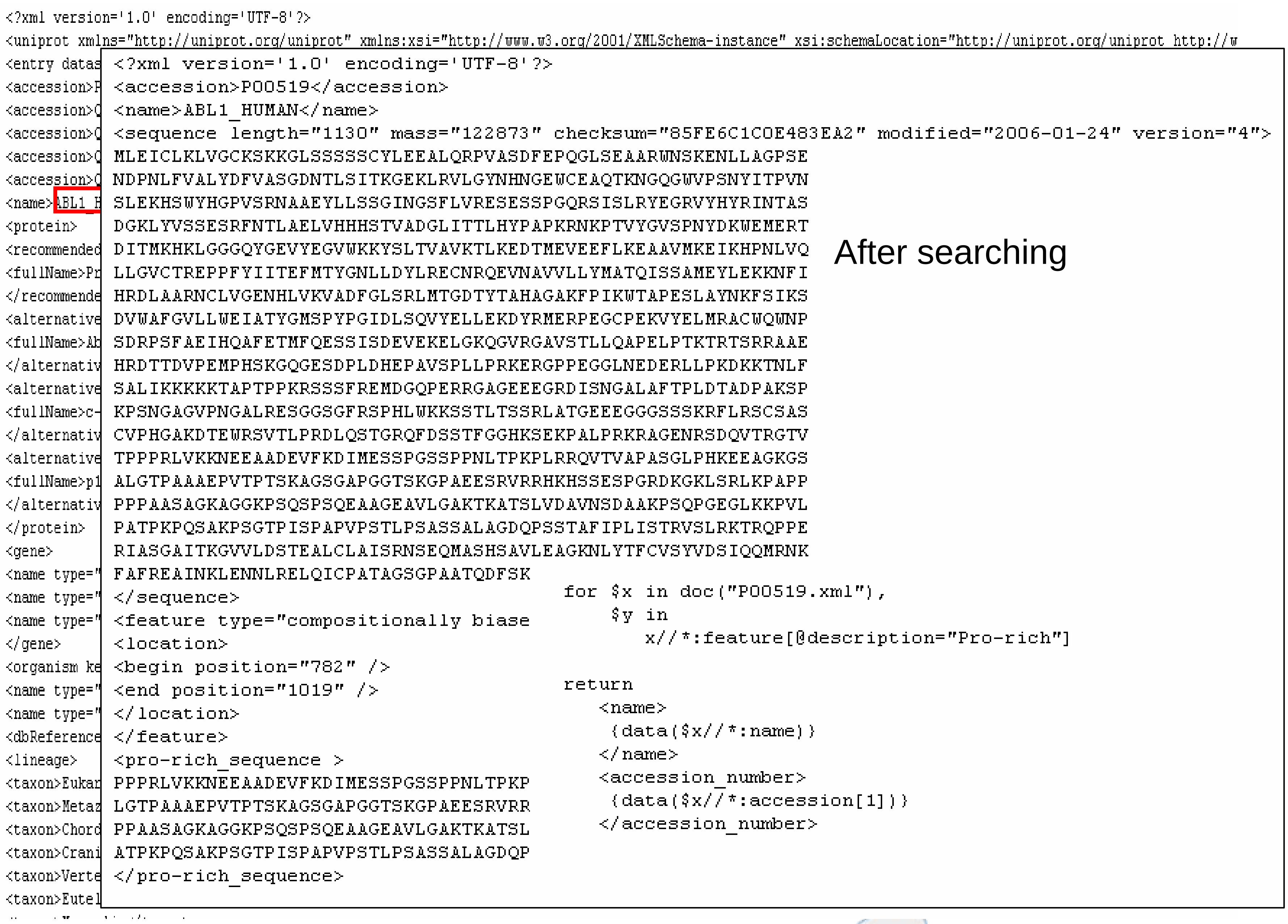
Search region = “pro-rich” in XML/RDF files from UniProt databases

Obtain the protein’s information including name, accession_number, sequence, structure, pro-rich region sequence, interactions, etc.

patternScan to scan X-P-P-X-P in the pro-rich region and locate potential SH3 binding site

A new software to sample the conformations of the pro-rich region and generate files as retrieved data

After searching

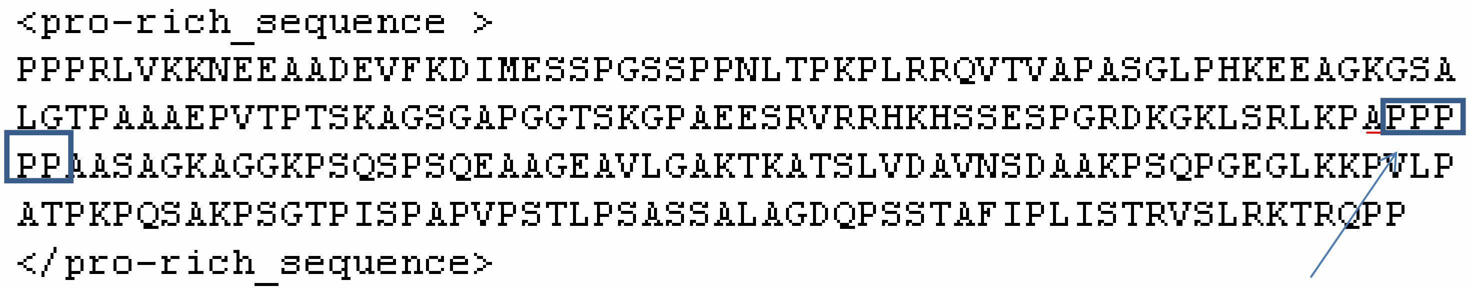


```
<?xml version='1.0' encoding='UTF-8'?>
<uniprot xmlns="http://uniprot.org/uniprot" xmlns:xsi="http://www.w3.org/2001/XMLSchema-instance" xsi:schemaLocation="http://uniprot.org/uniprot http://w
<entry datas
<accession>P
<accession>Q
<accession>Q
<accession>Q
<accession>Q
<name>ABL1_H
<protein>
<recommended
<fullName>Pr
</recommende
<alternative
<fullName>Ab
</alternativ
<alternative
<fullName>c-
</alternativ
<alternative
<fullName>p1
</alternativ
</protein>
<gene>
<name type="
<name type="
<name type="
</gene>
<organism ke
<name type="
<name type="
<dbReference
<lineage>
<taxon>Eukar
<taxon>Metaz
<taxon>Chord
<taxon>Crani
<taxon>Verte
<taxon>Eutel
```

```
<?xml version='1.0' encoding='UTF-8'?>
<accession>P00519</accession>
<name>ABL1_HUMAN</name>
<sequence length="1130" mass="122873" checksum="85FE6C1C0E483EA2" modified="2006-01-24" version="4">
MLEICLKLVGCKSKKGLSSSSSCYLEEALQRPVASDFEPQGLSEAARWNSKENLLAGPSE
NDPNLFVALYDFVASGDNTLSITKGEKLRVLGYNHNGEWCEAQTKNGQGWVPSNYITPVN
SLEKHSWYHGPVSRNAAEYLLSSGINGSFLVRESESSPGQRSISLRYEGRVYHYRINTAS
DGKLYVSSESRFNTLAELVHHHSTVADGLITTLHYPAPKRNKPTVYGVSPNYDKWEMERT
DITMKHKLGGGQYGEVYEGVWKKYSLTVAVKTLKEDTMEVEEFLKEAAVMKEIKHPNLVQ
LLGVCTREPPFYIITEFMTYGNLLDYLRECNRQEVNAVVLLYMATQISSAMEYLEKKNFI
HRDLAARNCLVGENHLVKVADFGLSRLMTGDTYTAHAGAKFPIKWTAPESLAYNKFSIKS
DVWAFGVLLWEIATYGMSPYPGIDLSQVYELLEKDYRMERPEGCPEKVYELMRACWQWNP
SDRPSFAEIHQAFETMFQESSISDEVEKELGKQGVRGAVSTLLQAPELPTKTRTSRRAAE
HRDTTDVPEMPHSKGQGESDPLDHEPAVSPLLPRKERGPPEGGLNEDERLLPKDKKTNLF
SALIKKKKKTAPTPPKRSSSFREMDGQPERRGAGEEEGRDISNGALAFTPLDTADPAKSP
KPSNGAGVPNGALRESGGSGFRSPHLWKKSSTLTSSRLATGEEEGGGSSSKRFLRSCSAS
CVPHGAKDTEWRSVTLPRDLQSTGRQFDSSTFGGHKSEKPALPRKRAGENRSDQVTRGTV
TPPPRLVKKNEEAADEVFKDIMESSPGSSPPNLTPKPLRRQVTVAPASGLPHKEEAGKGS
ALGTPAAAEPVTPTSKAGSGAPGGTSKGPAEESRVRRHKHSSESPGRDKGKLSRLKPAPP
PPPAASAGKAGGKPSQSPSQEAAGEAVLGAKTKATSLVDAVNSDAAKPSQPGEGLKKPVL
PATPKPQSAKPSGTPISPAPVPSTLPSASSALAGDQPSSTAFIPLISTRVSLRKTRQPPE
RIASGAITKGVVLDSTEALCLAISRNSEQMASHSAVLEAGKNLYTFCVSYVDSIQQMRNK
FAFREAINKLENNLRELQICPATAGSGPAATQDFSK
</sequence>
<feature type="compositionally biase
<location>
<begin position="782" />
<end position="1019" />
</location>
</feature>
<pro-rich_sequence >
PPPRLVKKNEEAADEVFKDIMESSPGSSPPNLTPKP
LGTPAAAEPVTPTSKAGSGAPGGTSKGPAEESRVRR
PPAASAGKAGGKPSQSPSQEAAGEAVLGAKTKATSL
ATPKPQSAKPSGTPISPAPVPSTLPSASSALAGDQP
</pro-rich_sequence>
```

```
for $x in doc("P00519.xml"),
    $y in
       x//*:feature[@description="Pro-rich"]

return
   <name>
    {data($x//*:name)}
   </name>
   <accession_number>
    {data($x//*:accession[1])}
   </accession_number>
```

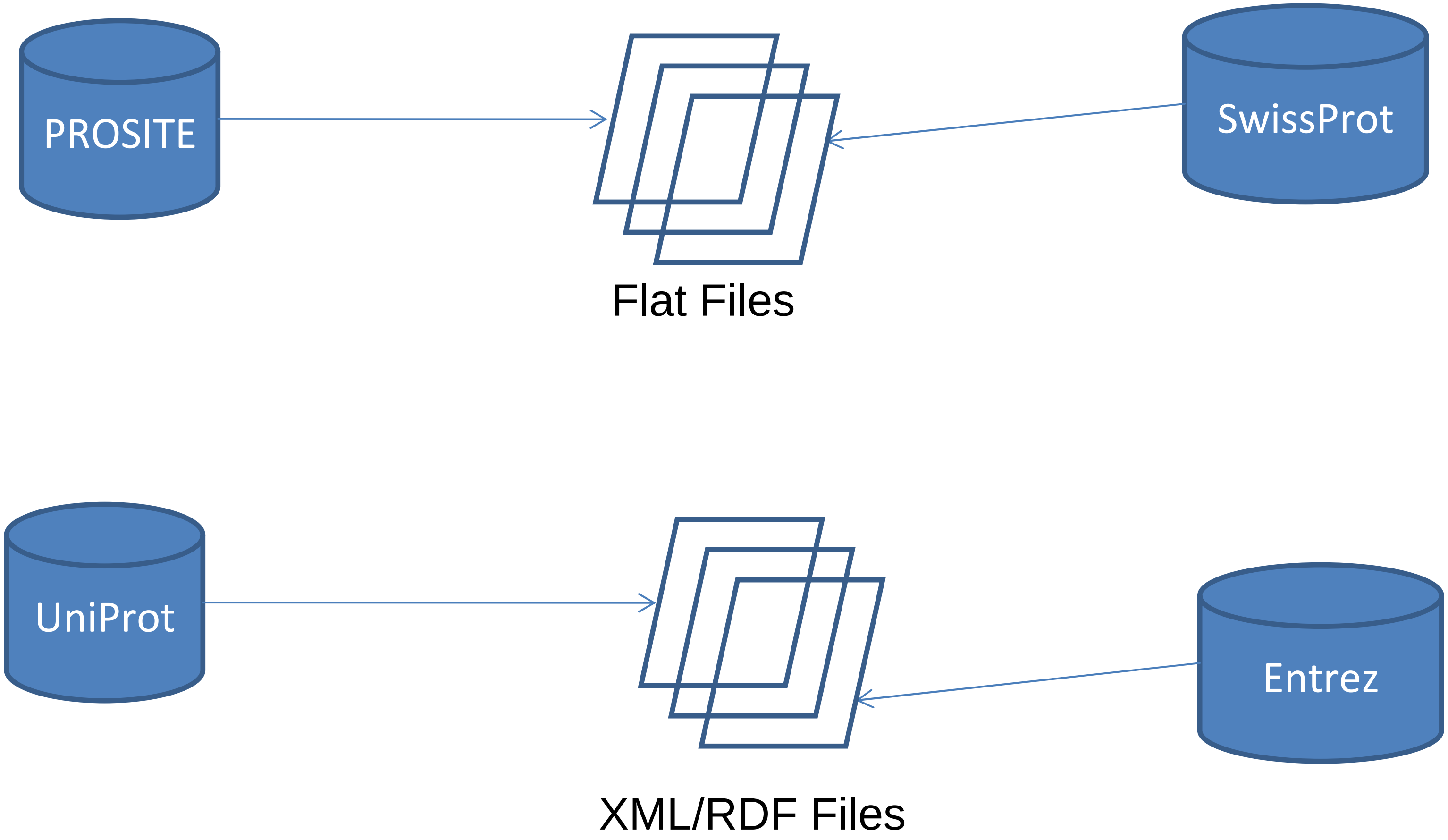


```
<pro-rich_sequence >
PPPRLVKKNEEAADEVFKDIMESSPGSSPPNLTPKPLRRQVTVAPASGLPHKEEAGKGSA
LGTPAAAEPVTPTSKAGSGAPGGTSKGPAEESRVRRHKHSSESPGRDKGKLSRLKPAPPP
PPAASAGKAGGKPSQSPSQEAAGEAVLGAKTKATSLVDAVNSDAAKPSQPGEGLKKPVLP
ATPKPQSAKPSGTPISPAPVPSTLPSASSALAGDQPSSTAFIPLISTRVSLRKTRQPP
</pro-rich_sequence>
```

patternScan identifies two potential SH3 binding sites

A new software takes input file of pro-rich sequence to generate sample conformations

Retrieved data in different formats



PROSITE

SwissProt

Flat Files

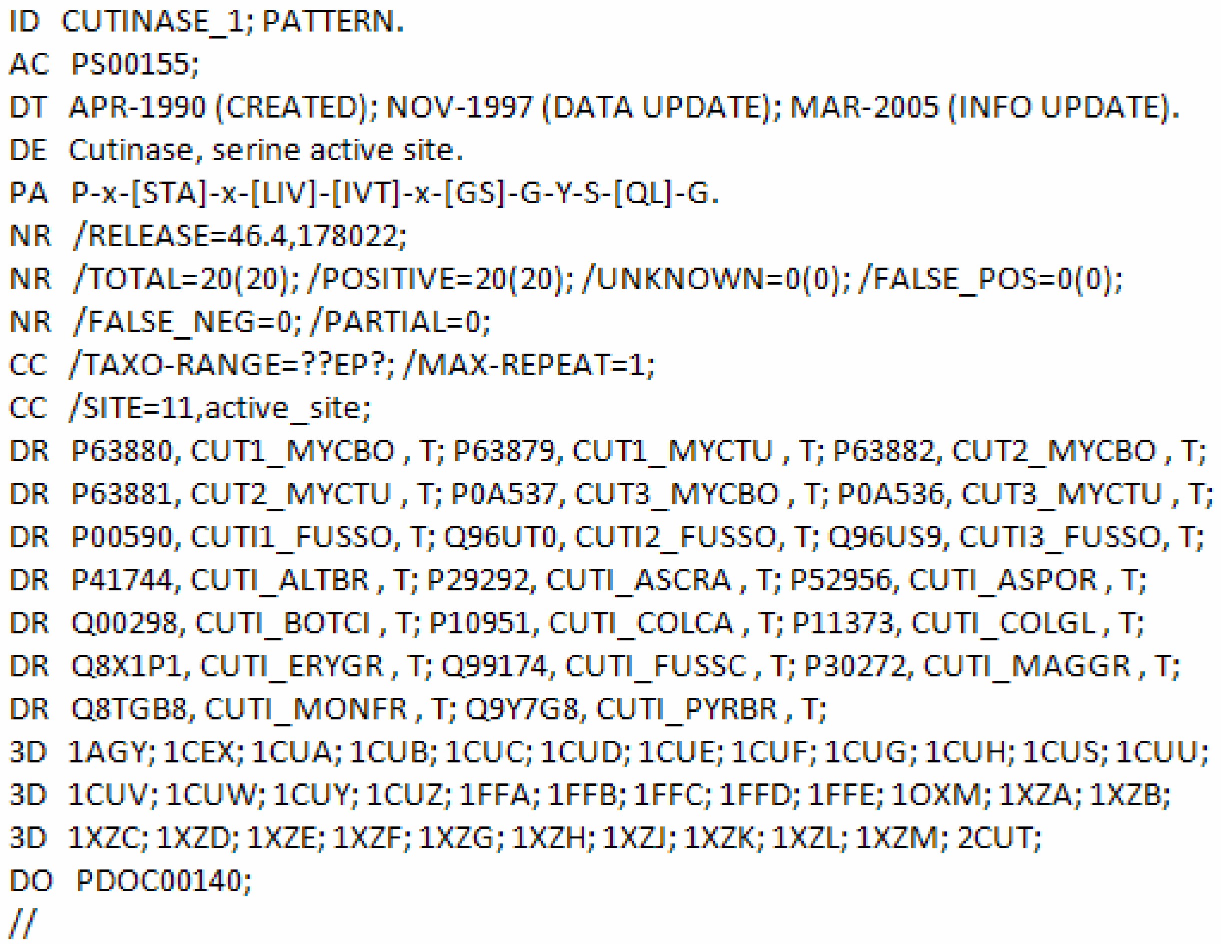
UniProt

Entrez

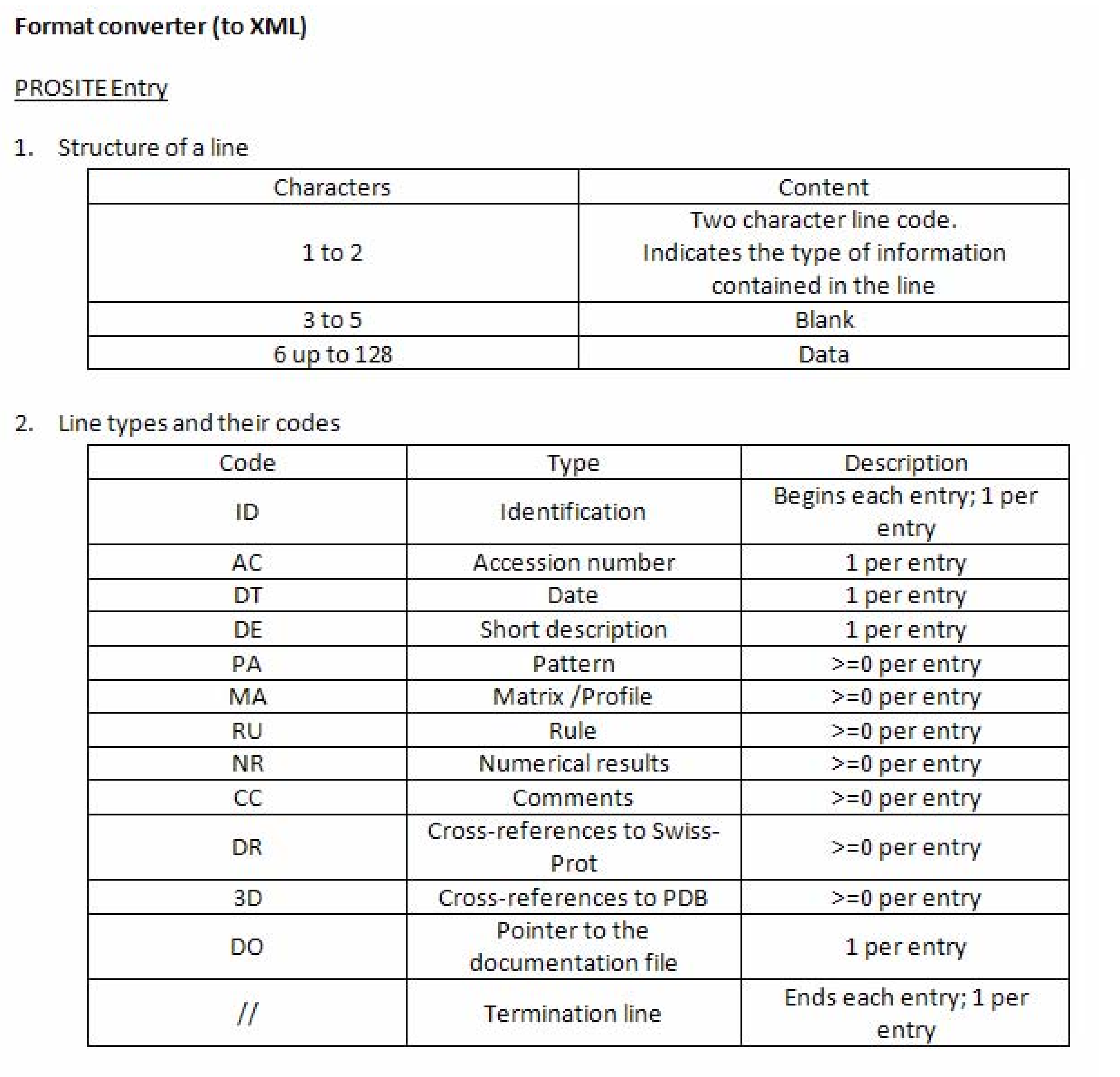
XML/RDF Files

Format Converter

- Specification of a DTD for the flat file
- Mapping attributes in the flat file to elements and attributes in the DTD
- Input flat files→ XML/RDF files by a format converter
- Example: converting PROSITE flat file to XML file



```
ID   CUTINASE_1; PATTERN.
AC   PS00155;
DT   APR-1990 (CREATED); NOV-1997 (DATA UPDATE); MAR-2005 (INFO UPDATE).
DE   Cutinase, serine active site.
PA   P-x-[STA]-x-[LIV]-[IVT]-x-[GS]-G-Y-S-[QL]-G.
NR   /RELEASE=46.4,178022;
NR   /TOTAL=20(20); /POSITIVE=20(20); /UNKNOWN=0(0); /FALSE_POS=0(0);
NR   /FALSE_NEG=0; /PARTIAL=0;
CC   /TAXO-RANGE=??EP?; /MAX-REPEAT=1;
CC   /SITE=11,active_site;
DR   P63880, CUT1_MYCBO , T; P63879, CUT1_MYCTU , T; P63882, CUT2_MYCBO , T;
DR   P63881, CUT2_MYCTU , T; P0A537, CUT3_MYCBO , T; P0A536, CUT3_MYCTU , T;
DR   P00590, CUTI1_FUSSO, T; Q96UT0, CUTI2_FUSSO, T; Q96US9, CUTI3_FUSSO, T;
DR   P41744, CUTI_ALTBR , T; P29292, CUTI_ASCRA , T; P52956, CUTI_ASPOR , T;
DR   Q00298, CUTI_BOTCI , T; P10951, CUTI_COLCA , T; P11373, CUTI_COLGL , T;
DR   Q8X1P1, CUTI_ERYGR , T; Q99174, CUTI_FUSSC , T; P30272, CUTI_MAGGR , T;
DR   Q8TGB8, CUTI_MONFR , T; Q9Y7G8, CUTI_PYRBR , T;
3D   1AGY; 1CEX; 1CUA; 1CUB; 1CUC; 1CUD; 1CUE; 1CUF; 1CUG; 1CUH; 1CUS; 1CUU;
3D   1CUV; 1CUW; 1CUY; 1CUZ; 1FFA; 1FFB; 1FFC; 1FFD; 1FFE; 1OXM; 1XZA; 1XZB;
3D   1XZC; 1XZD; 1XZE; 1XZF; 1XZG; 1XZH; 1XZJ; 1XZK; 1XZL; 1XZM; 2CUT;
DO   PDOC00140;
//
```



Format converter (to XML)

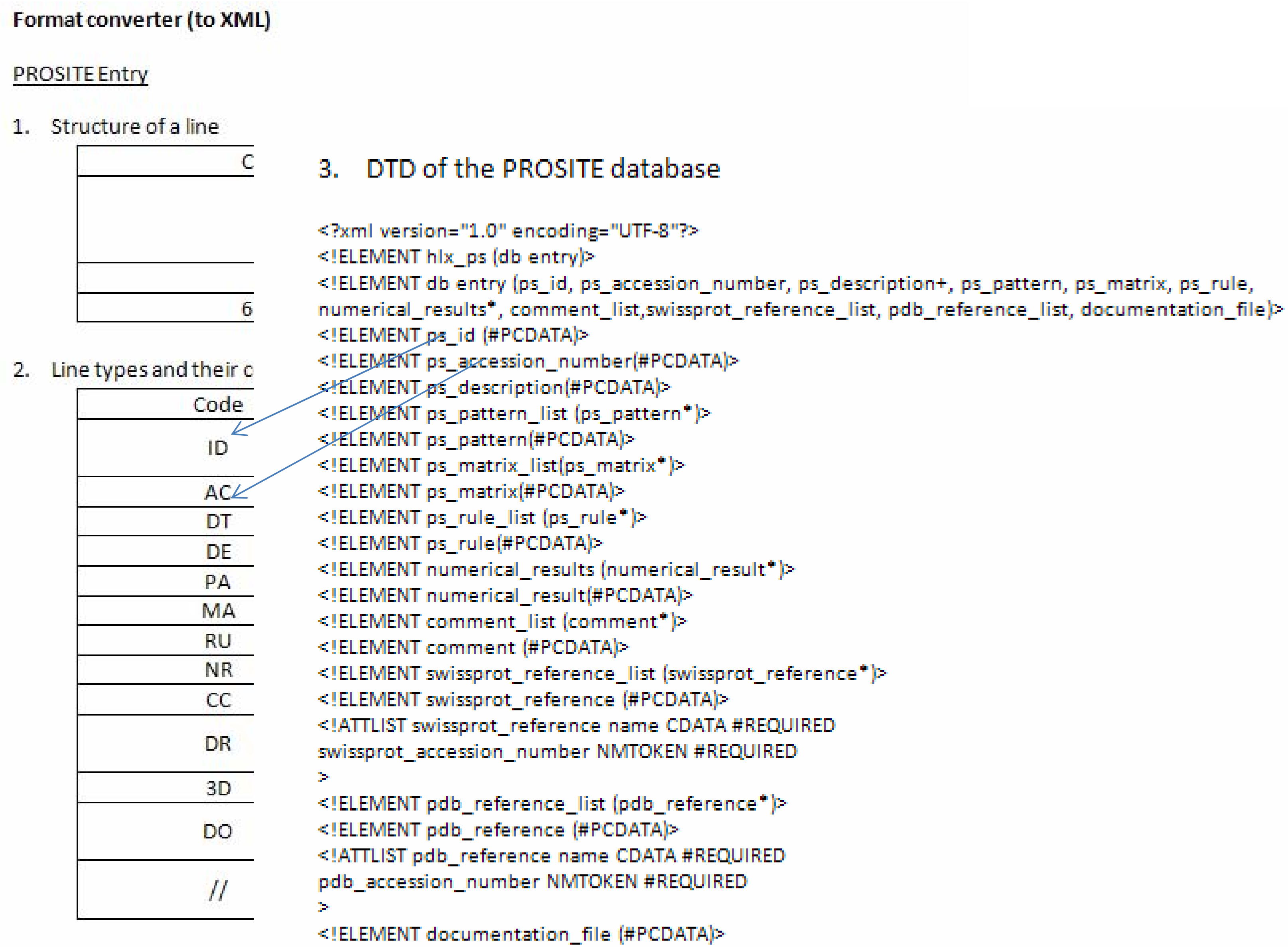
PROSITE Entry

1. Structure of a line

Characters	Content
1 to 2	Two character line code. Indicates the type of information contained in the line
3 to 5	Blank
6 up to 128	Data

2. Line types and their codes

Code	Type	Description
ID	Identification	Begins each entry; 1 per entry
AC	Accession number	1 per entry
DT	Date	1 per entry
DE	Short description	1 per entry
PA	Pattern	>=0 per entry
MA	Matrix /Profile	>=0 per entry
RU	Rule	>=0 per entry
NR	Numerical results	>=0 per entry
CC	Comments	>=0 per entry
DR	Cross-references to Swiss-Prot	>=0 per entry
3D	Cross-references to PDB	>=0 per entry
DO	Pointer to the documentation file	1 per entry
//	Termination line	Ends each entry; 1 per entry



Format converter (to XML)

PROSITE Entry

1. Structure of a line

C
6

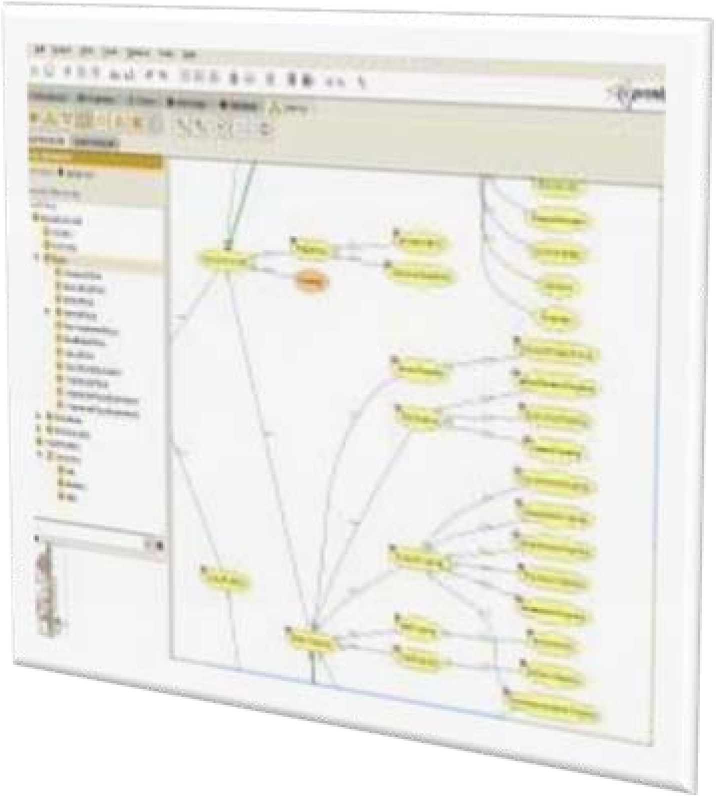
2. Line types and their c

Code
ID
AC
DT
DE
PA
MA
RU
NR
CC
DR
3D
DO
//

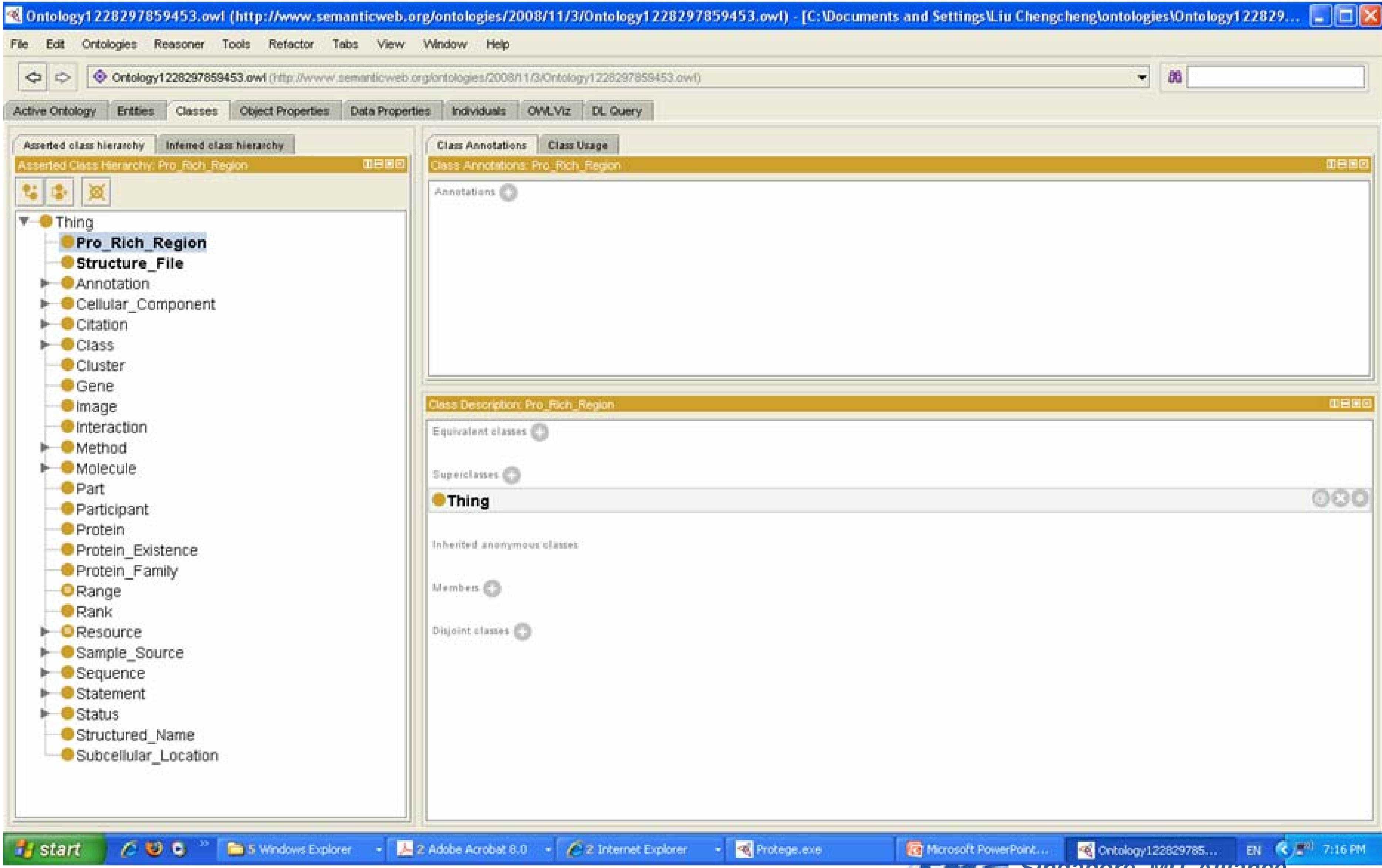
3. DTD of the PROSITE database

```
<?xml version="1.0" encoding="UTF-8"?>
<!ELEMENT hlx_ps (db entry)>
<!ELEMENT db entry (ps_id, ps_accession_number, ps_description+, ps_pattern, ps_matrix, ps_rule,
numerical_results*, comment_list,swissprot_reference_list, pdb_reference_list, documentation_file)>
<!ELEMENT ps_id (#PCDATA)>
<!ELEMENT ps_accession_number(#PCDATA)>
<!ELEMENT ps_description(#PCDATA)>
<!ELEMENT ps_pattern_list (ps_pattern*)>
<!ELEMENT ps_pattern(#PCDATA)>
<!ELEMENT ps_matrix_list(ps_matrix*)>
<!ELEMENT ps_matrix(#PCDATA)>
<!ELEMENT ps_rule_list (ps_rule*)>
<!ELEMENT ps_rule(#PCDATA)>
<!ELEMENT numerical_results (numerical_result*)>
<!ELEMENT numerical_result(#PCDATA)>
<!ELEMENT comment_list (comment*)>
<!ELEMENT comment (#PCDATA)>
<!ELEMENT swissprot_reference_list (swissprot_reference*)>
<!ELEMENT swissprot_reference (#PCDATA)>
<!ATTLIST swissprot_reference name CDATA #REQUIRED
swissprot_accession_number NMTOKEN #REQUIRED
>
<!ELEMENT pdb_reference_list (pdb_reference*)>
<!ELEMENT pdb_reference (#PCDATA)>
<!ATTLIST pdb_reference name CDATA #REQUIRED
pdb_accession_number NMTOKEN #REQUIRED
>
<!ELEMENT documentation_file (#PCDATA)>
```

Data Ontology



Data Ontology



Ontology1228297859453.owl (http://www.semanticweb.org/ontologies/2008/11/3/Ontology1228297859453.owl) - [C:\Documents and Settings\Liu Chengcheng\ontologies\Ontology122829...

Asserted Class Hierarchy: Pro_Rich_Region

- Thing
  - **Pro_Rich_Region**
  - **Structure_File**
  - Annotation
  - Cellular_Component
  - Citation
  - Class
  - Cluster
  - Gene
  - Image
  - Interaction
  - Method
  - Molecule
  - Part
  - Participant
  - Protein
  - Protein_Existence
  - Protein_Family
  - Range
  - Rank
  - Resource
  - Sample_Source
  - Sequence
  - Statement
  - Status
  - Structured_Name
  - Subcellular_Location

Class Annotations: Pro_Rich_Region

Annotations

Class Description: Pro_Rich_Region

Equivalent classes

Superclasses

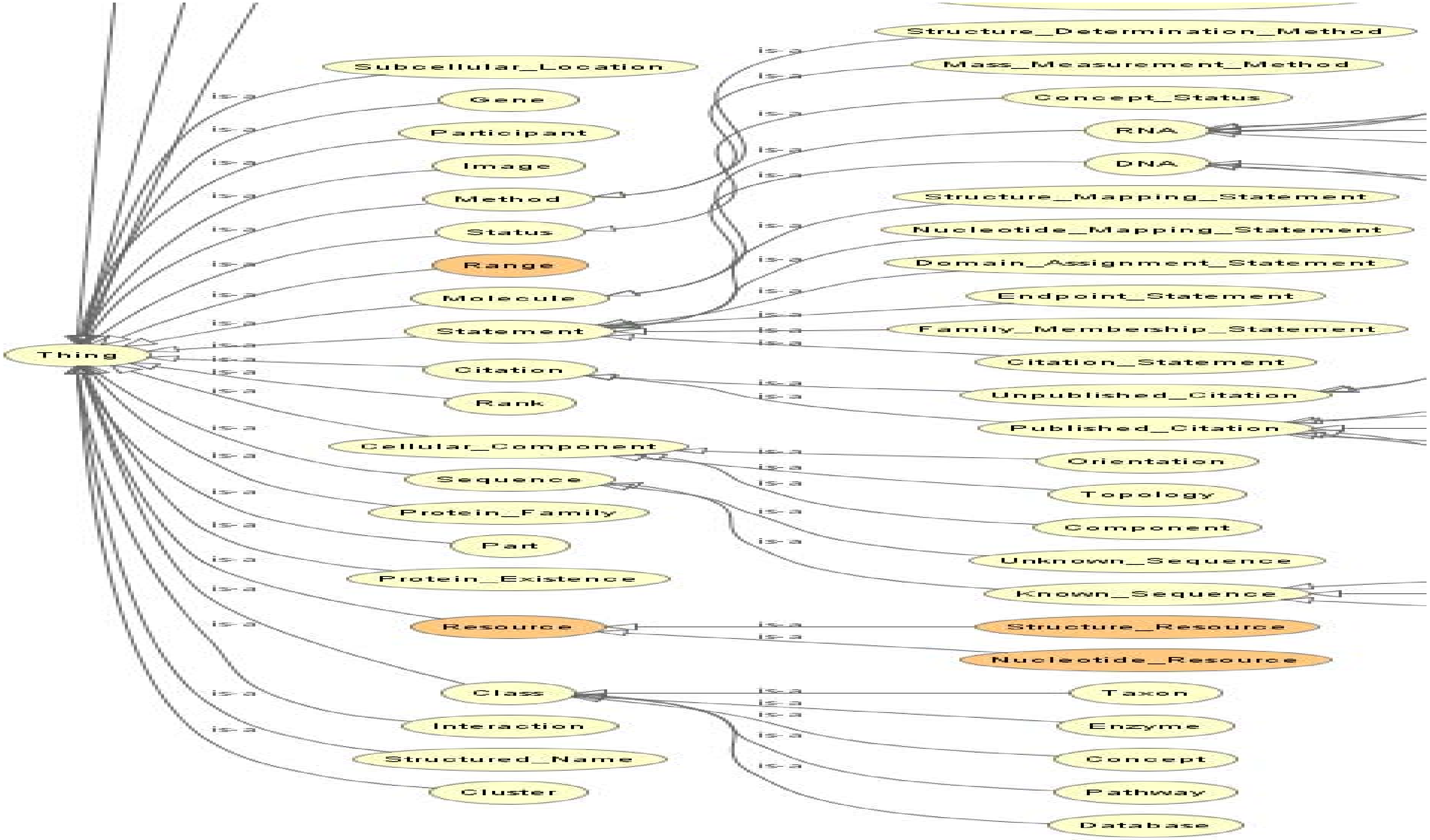
Thing

Inherited anonymous classes

Members

Disjoint classes

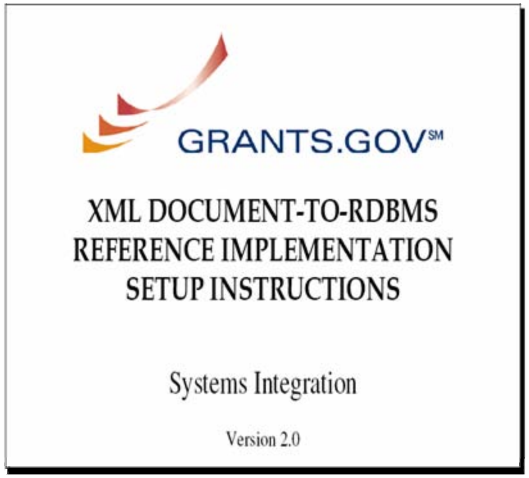
Part of Data Ontology



Singapore–MIT Alliance

Data Loading & Warehouse

- Loading approaches:
- Store and query XML using RDBMS?

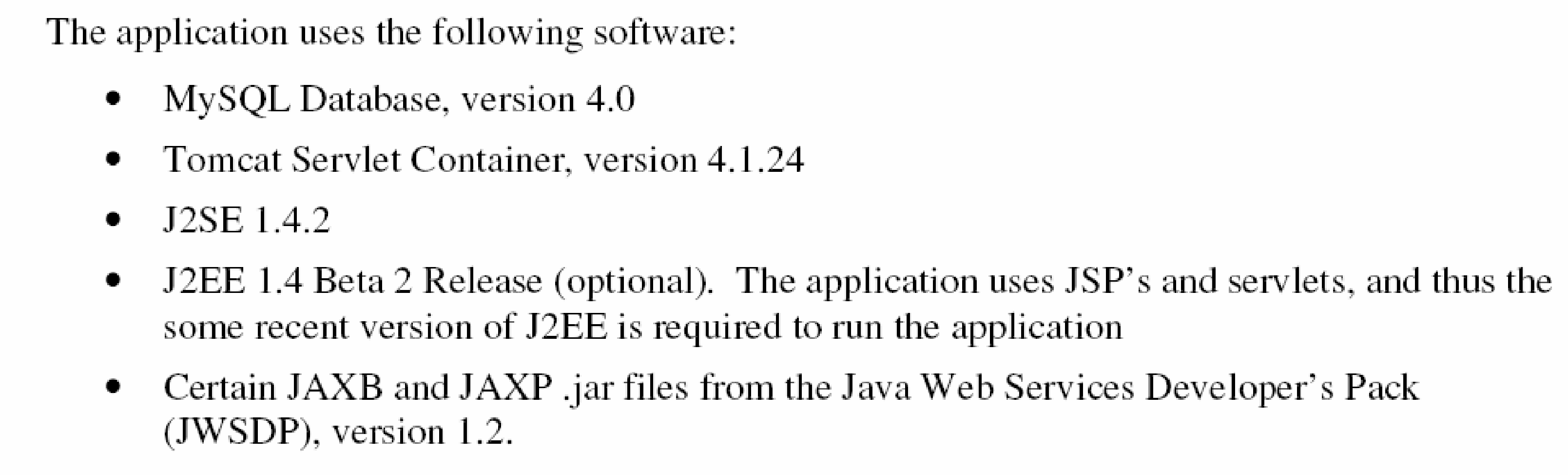


GRANTS.GOV℠

XML DOCUMENT-TO-RDBMS
REFERENCE IMPLEMENTATION
SETUP INSTRUCTIONS

Systems Integration

Version 2.0

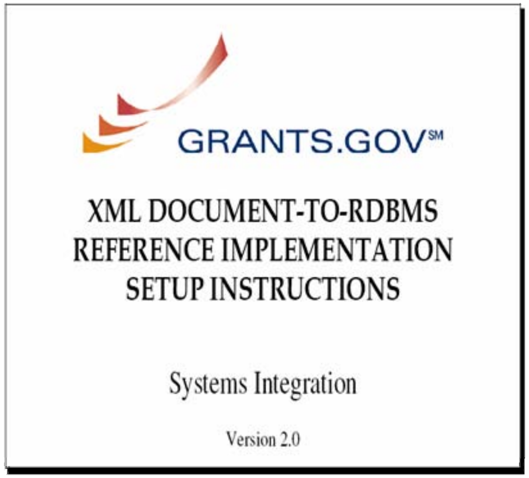


The application uses the following software:

- MySQL Database, version 4.0
- Tomcat Servlet Container, version 4.1.24
- J2SE 1.4.2
- J2EE 1.4 Beta 2 Release (optional). The application uses JSP's and servlets, and thus the some recent version of J2EE is required to run the application
- Certain JAXB and JAXP .jar files from the Java Web Services Developer's Pack (JWSDP), version 1.2.

Data Loading & Warehouse

- Loading approaches:
- Store and query XML using RDBMS?

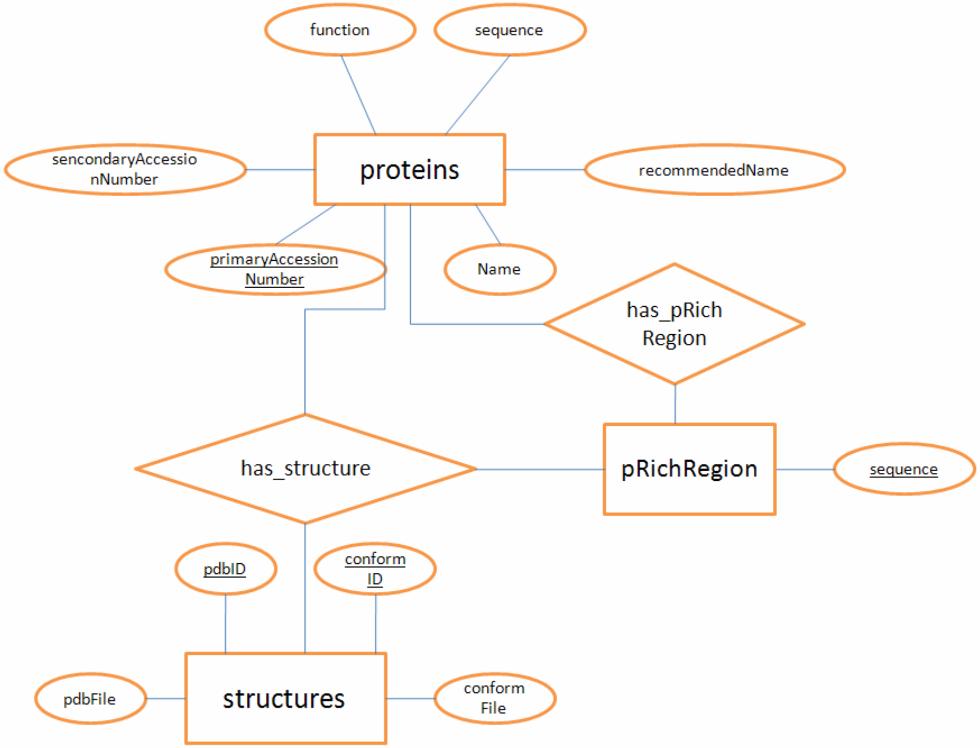


GRANTS.GOV℠

XML DOCUMENT-TO-RDBMS
REFERENCE IMPLEMENTATION
SETUP INSTRUCTIONS

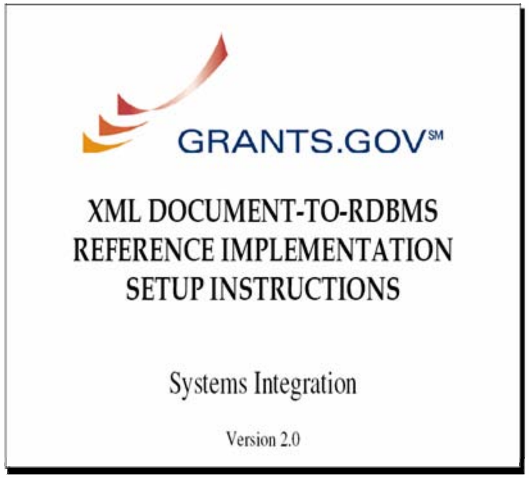
Systems Integration

Version 2.0



Data Loading & Warehouse

- Loading approaches:
- Store and query XML using RDBMS?
- New architecture to store and query XML/RDF?



GRANTS.GOV℠

XML DOCUMENT-TO-RDBMS
REFERENCE IMPLEMENTATION
SETUP INSTRUCTIONS

Systems Integration

Version 2.0

issues

- Updates in source database
- Changes in warehouse maintainance
- Versions of software for converting and loading data

References

Florescu et.al*, Storing and Querying XML Data using an RDBMS*, Bulletin of the IEEE Computer Society Technical Committee on Data Engineering 1999

Stein, *Integrating Biological Databases*, Nature reviews genetics, volume 4, 2003, 337-345

Davidson et al, *The Kleisli Approach to Data Transformation and Integration,*

Broekstra et al, Sesame: A Generic Architecture for Storing and Querying RDF and RDF Schema

Wilkinson et al, Efficient RDF Storage and Retrieval in Jena2

Horridge, *A Practical Guide To Building OWL Ontologies Using The Prot´eg´e-OWL Plugin and CO-ODE Tools,* 2004

XML document-to-RDBMS reference implementation setup instructions Version 2.0, 2003

NUS National University of Singapore

THANK YOU

SMA Singapore–MIT Alliance

TraDES

- TraDES (Trajectory Directed Ensemble Sampling)
- Being composed to sample conformations of Pro-rich non-globular domains
- To analyze statistically the accessibility of kinases or other binding proteins on such domains
- To output conformations of pro-rich non-globular domains that are most accessible