Translating Biological Data Sets Into Linked Data

Mark Tomko
Simmons College, Boston MA
The Broad Institute of MIT and Harvard, Cambridge MA
September 28, 2011

Overview

- Why study biological data?
- UniProt & Pfam
- Translating Pfam into linked data
- Challenges for representing biological data

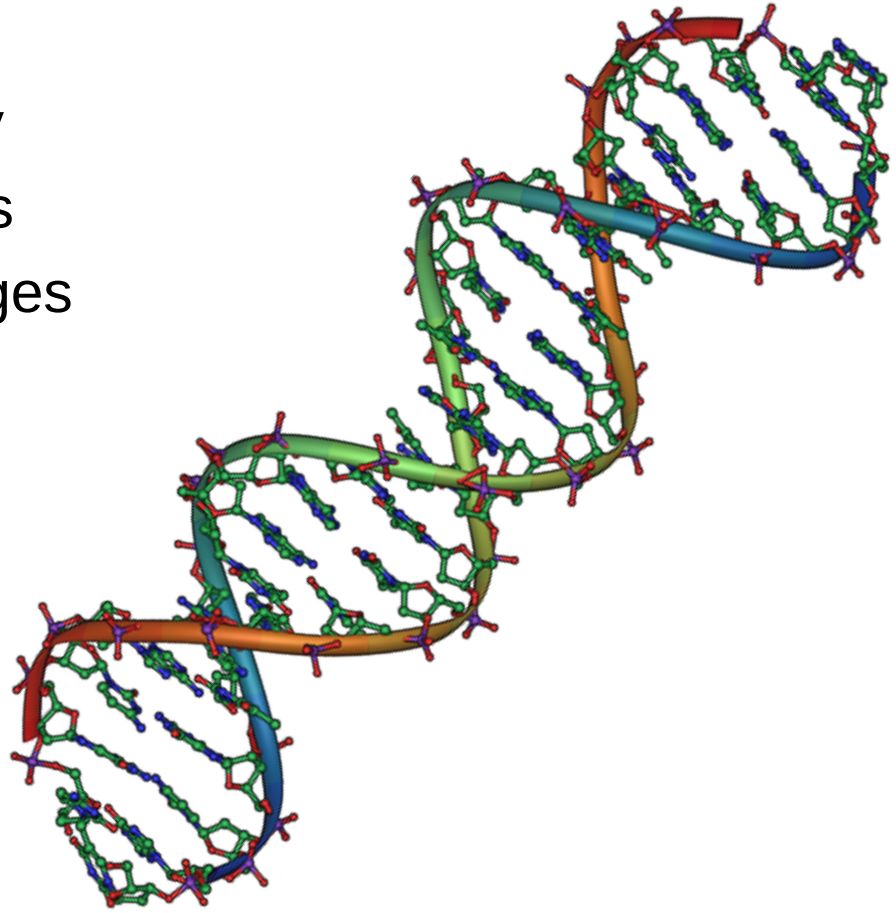
Perspective

- I am not a biologist
- I am not an expert on linked data
- I’m a software engineer interested in:
  - Scientific data and metadata
  - Scientific data sharing
  - Biology and bioinformatics

- This talk takes a pragmatic approach

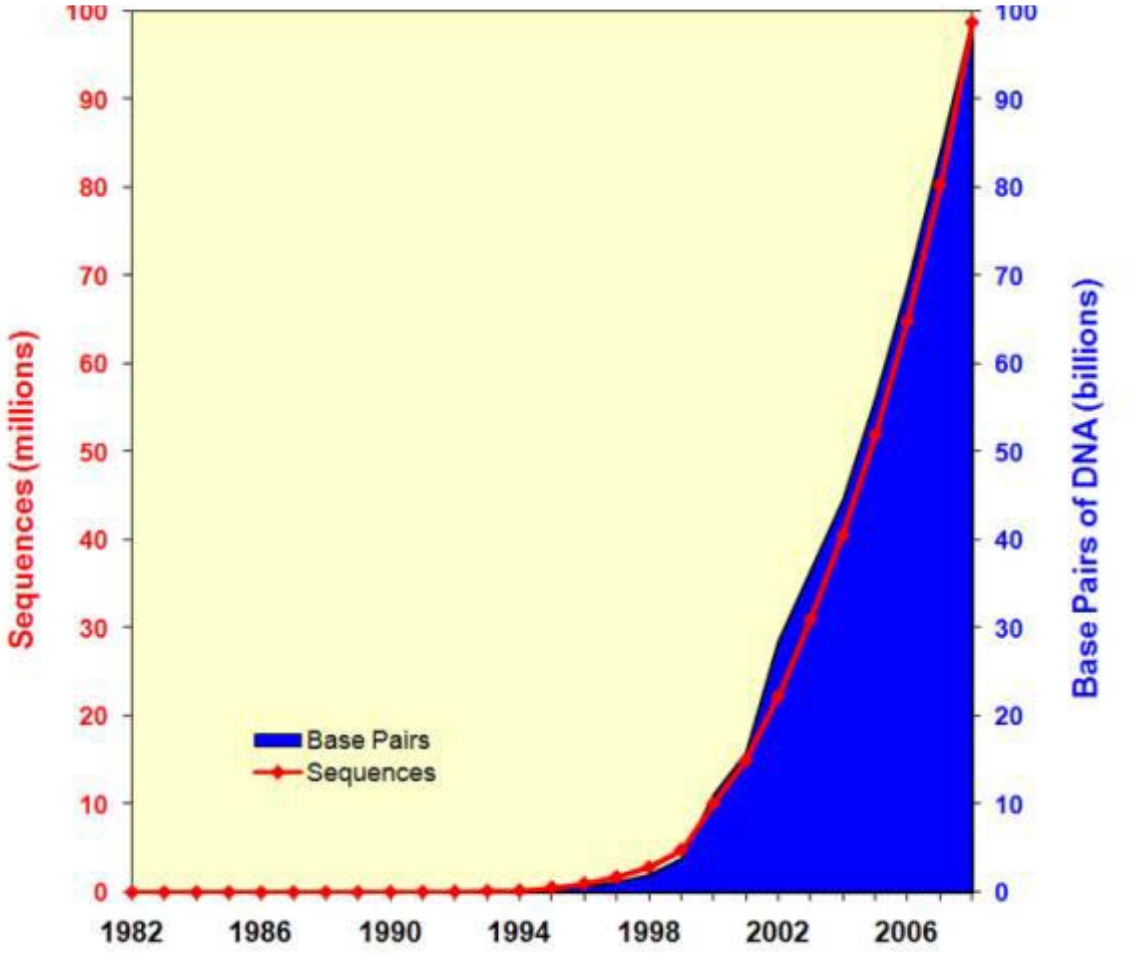
Why study biological data?

- Quantity
- Diversity
- Changes
- Challenges



How much data?

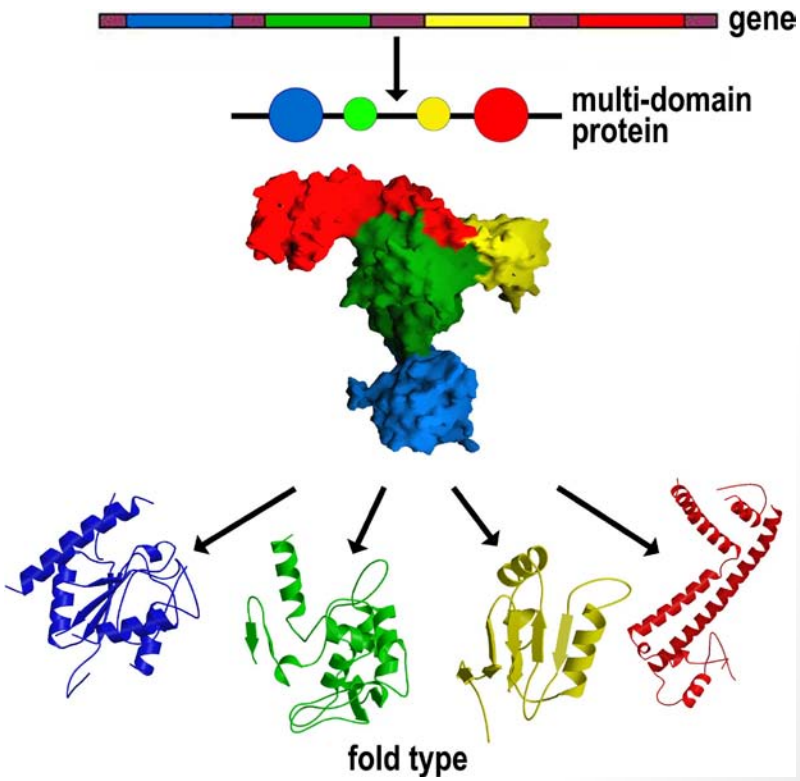
Growth of GenBank Genetic Sequence Database, 1982-2008



http://www.ncbi.nlm.nih.gov/genbank/genbankstats.html

What kind of data?

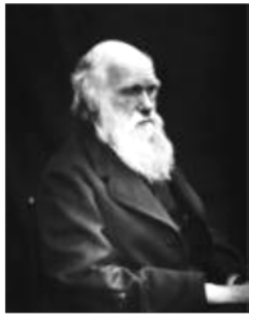
- Sequence data (nucleotide & amino acid)
- Genomic data
- Proteomic data
- Neuronal wiring
- Cell fates
- Phylogenic information
- Homologous molecules
- And so on …



http://www.mpg.de/english/illustrationsDocumentation/documentation/pressReleases/2002/news0204_bild1.jpg

What kinds of changes?

Gregor Mendel



Charles Darwin

James D. Watson & Francis Crick

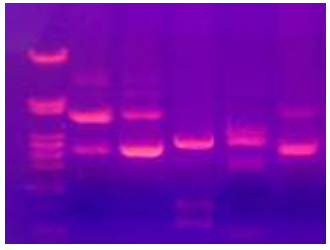
Gordon Moore

1800

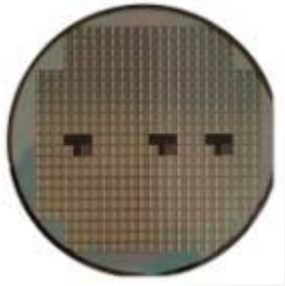
1900

2000

In vivo



In vitro



In silico

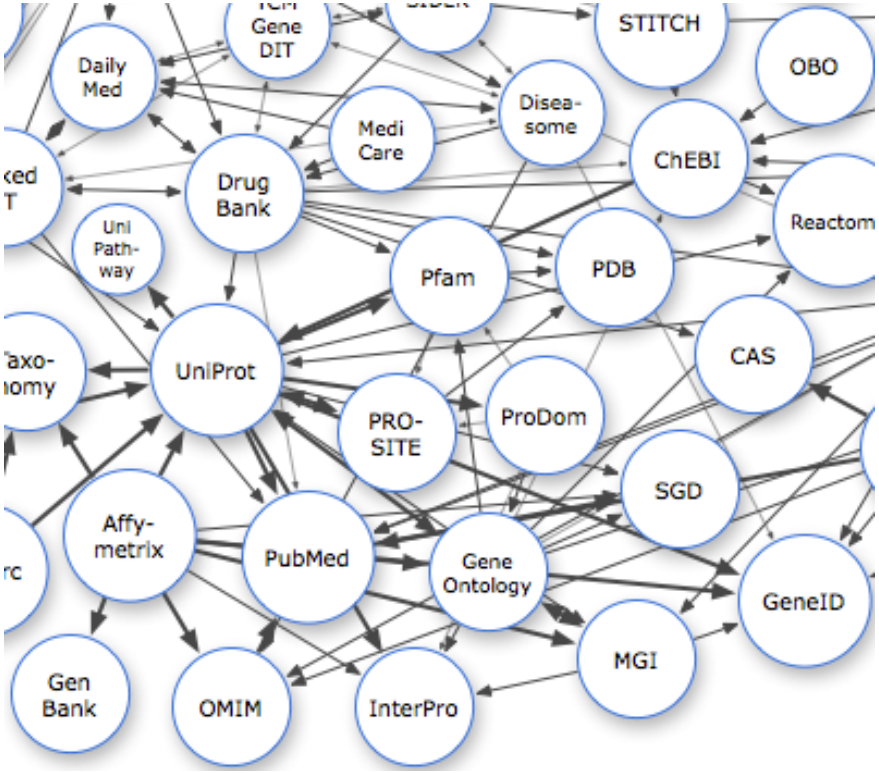
Challenges

- Very large data sets
- In a multitude of data formats
- Widely distributed
- Full of semantic linkages
- Yet:
  - Computational techniques are increasingly important
  - Experts may not have extensive backgrounds in:
    - Data modeling
    - Data management
    - Programming

Goals

- Organize the body of biological knowledge
- Link related knowledge
- Connect facts with research
- Make bioinformatics data discoverable and interoperable
- Facilitate data sharing between researchers

Existing linked biological data



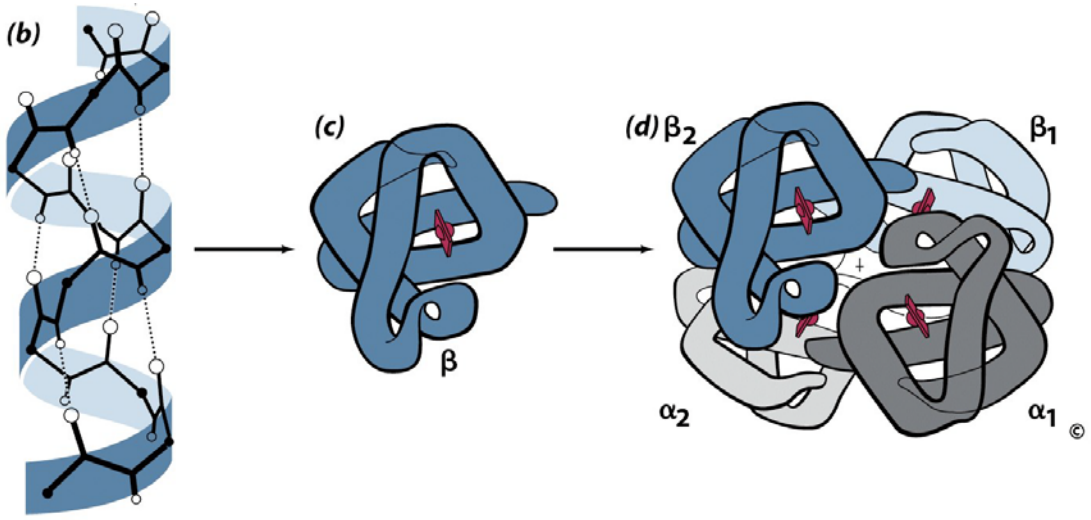
Linking Open Data cloud diagram, by Richard Cyganiak and Anja Jentzsch. http://lod-cloud.net/

A bit of biology

- Proteins are characterized by amino acid sequence and folding
- Similarity between analogous proteins suggests:
  - Functional similarity
  - Common evolutionary origin

--KAHGKKVLGA--

Single letter code
(Hemoglobin β chain)



UniProtKB

- Online repository of annotated protein sequences
- Derived from scientific literature & other databases
- Links to over 100 other data sets
- Supported by EBI, PIR, NIH
- Data available in several formats:
  - Online (HTML)
  - Flat files
  - RDF

http://www.uniprot.org/

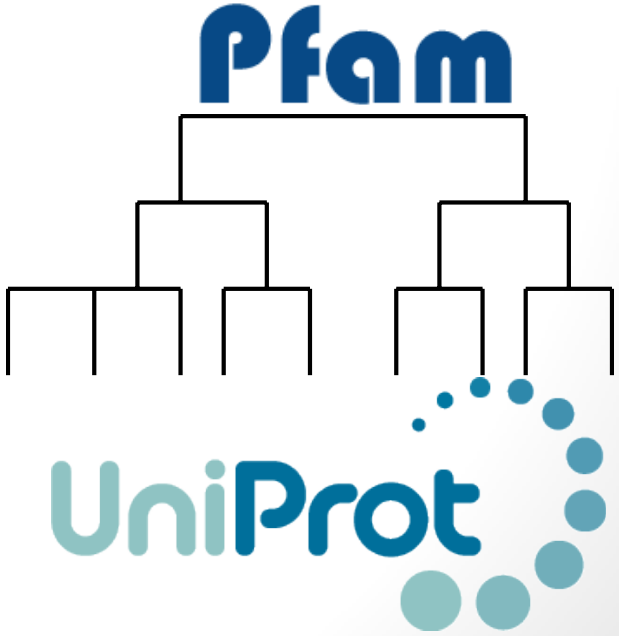
Pfam

- Organizes proteins from UniProt into families based on similarities in amino acid sequences
- Computes similarities via sequence alignments and Hidden Markov Models (HMMs)
- Organizes families into higher-level groups called clans
  - Clan membership is based on similarity between the characteristic sequences of the member families

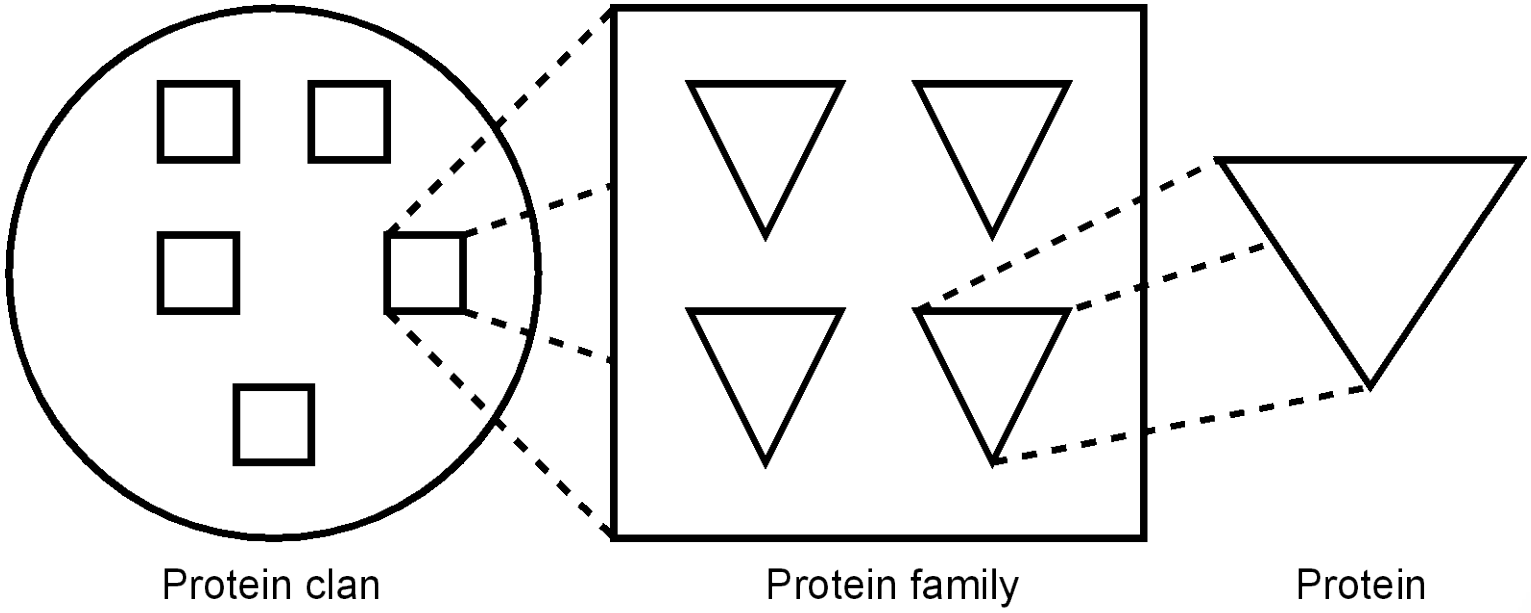
http://pfam.sanger.ac.uk

Pfam organizes UniProt

- Provides alternate access points for UniProt sequences
- Clusters UniProt entries
  - Domain specific similarity metrics
  - Non-obvious without domain knowledge
  - Cluster membership helps to predict useful properties
    - Function
    - Evolutionary origin
    - Shapes & features

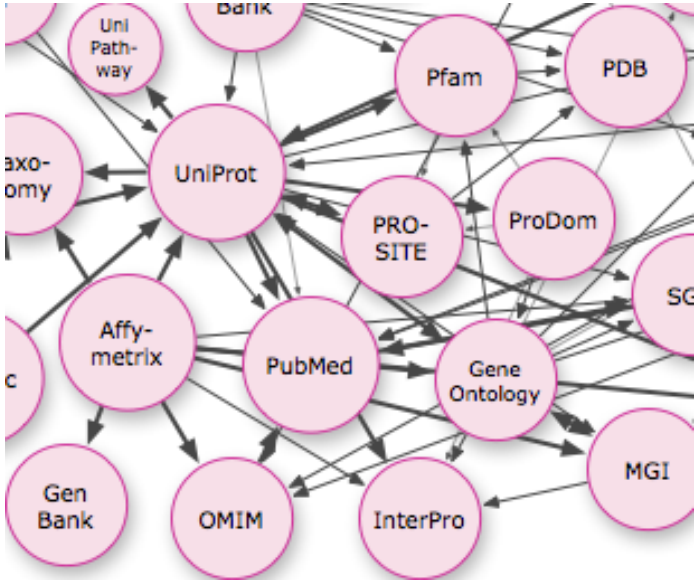


Pfam families and clans



Pfam provides context

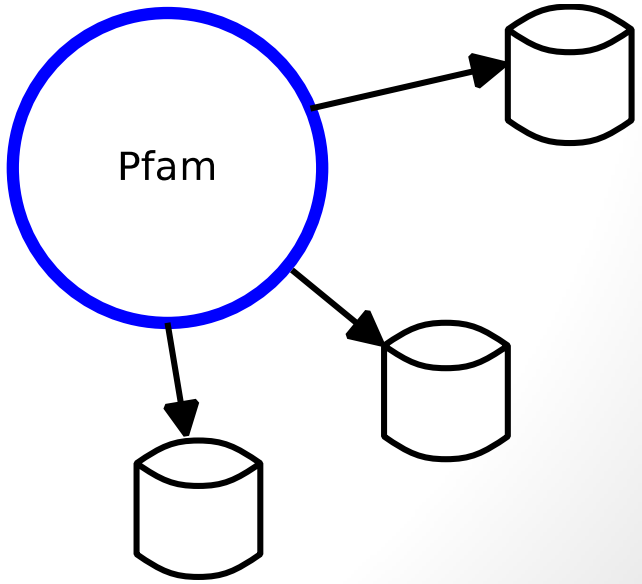
- Collects findings from disparate literature
- Establishes critical links that cannot be inferred automatically without specific domain knowledge



Linking Open Data cloud diagram, by Richard Cyganiak and Anja Jentzsch. http://lod-cloud.net/

Pfam references other data

- Pfam links to existing databases such as:
  - InterPro (http://www.ebi.ac.uk/interpro/)
  - SCOP (http://scop.mrc-lmb.cam.ac.uk/scop/)
  - PROSITE (http://prosite.expasy.org/)
  - HOMSTRAD (http://tardis.nibio.go.jp/homstrad/)
- Also links to related publications in PubMed
  - http://www.ncbi.nlm.nih.gov/pubmed/

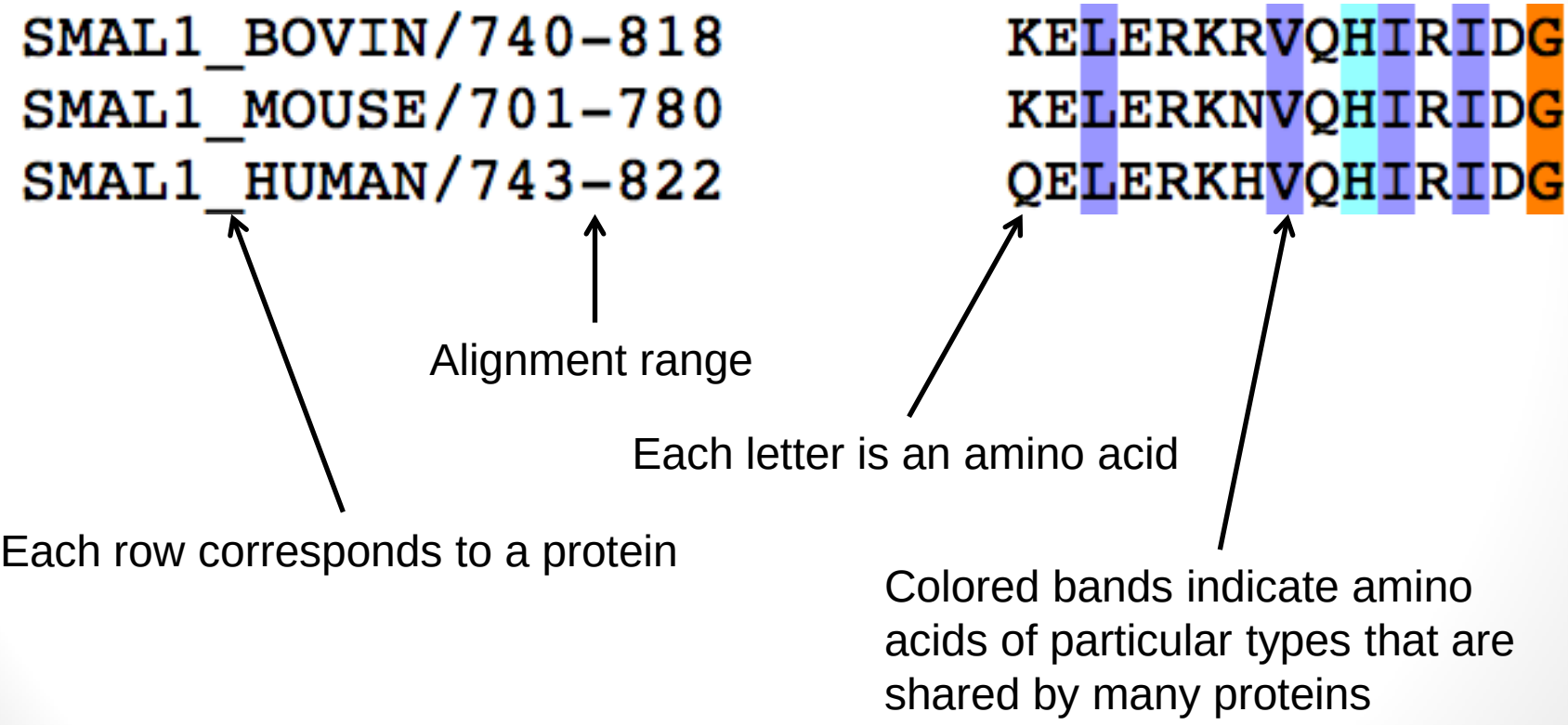


Example Pfam data

```
# STOCKHOLM 1.0
#=GF ID   Helicase_C
#=GF AC   PF00271.25
#=GF DE   Helicase conserved C-terminal domain
#=GF PI   helicase_C;
#=GF AU   Sonnhammer ELL
#=GF SE   Published_alignment
#=GF GA   20.90 11.60;
#=GF TC   20.90 11.60;
#=GF NC   20.80 11.50;
#=GF BM   hmmbuild HMM.ann SEED.ann
#=GF SM   hmmsearch -Z 11384036 -E 1000 --cpu 4 HMM pfamseq
#=GF TP   Family
#=GF DR   INTERPRO; IPR001650;
#=GF DR   PROSITE; PDOC00039;
#=GF DR   SCOP; 1d2m; fa;
#=GF DR   HOMSTRAD; helicase_C;
#=GF DR   HOMSTRAD; helicase_NC;
#=GF CC   The Prosite family is restricted to DEAD/H helicases, whereas
#=GF CC   this domain family is found in a wide variety of helicases and
#=GF CC   helicase related proteins. It may be that this is not an
#=GF CC   autonomously folding unit, but an integral part of the helicase.
#=GF SQ   491
```

Pfam sequence alignments

Alignment for PF00271 (Helicase C domain):



Bio2RDF Pfam translation

```
<rdf:RDF
        xmlns:rdfs="http://www.w3.org/2000/01/rdf-schema#"
        xmlns:rdf="http://www.w3.org/1999/02/22-rdf-syntax-ns#">
<rdf:Description rdf:about="http://bio2rdf.org/pfam:PF00023">
        <linkedToFrom xmlns="http://bio2rdf.org/bio2rdf_resource:"
                        rdf:resource="http://bio2rdf.org/pfam:PF08344"/>
        <xPfam xmlns="http://bio2rdf.org/bio2rdf_resource:"
                        rdf:resource="http://bio2rdf.org/pfam:PB178448"/>
        <xPfam xmlns="http://bio2rdf.org/bio2rdf_resource:"
                        rdf:resource="http://bio2rdf.org/pfam:PB179386"/>
        <xPfam xmlns="http://bio2rdf.org/bio2rdf_resource:"
                        rdf:resource="http://bio2rdf.org/pfam:PB177829"/>
        <xPfam xmlns="http://bio2rdf.org/bio2rdf_resource:"
                        rdf:resource="http://bio2rdf.org/pfam:PB178369"/>
        <xPfam xmlns="http://bio2rdf.org/bio2rdf_resource:"
                        rdf:resource="http://bio2rdf.org/pfam:PB177261"/>
        <xPfam xmlns="http://bio2rdf.org/bio2rdf_resource:"
                        rdf:resource="http://bio2rdf.org/pfam:PB177264"/>
```

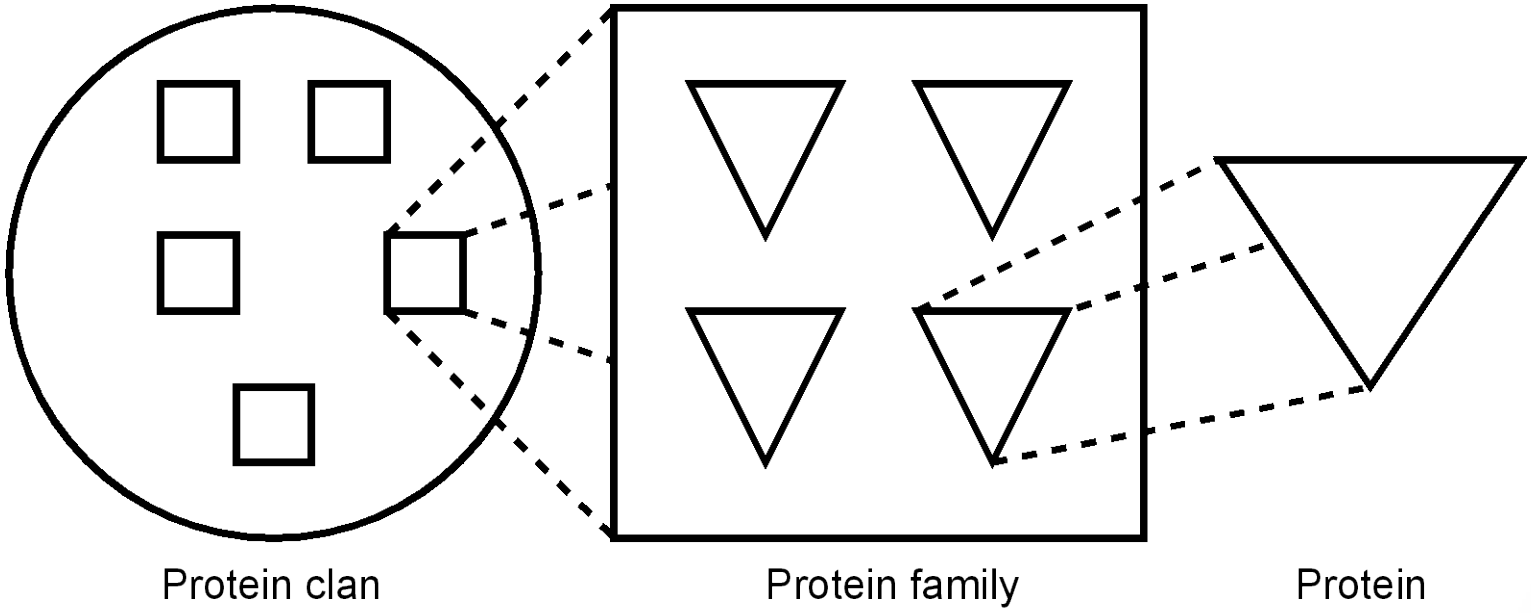
This work

- Retains original UniProt and Pfam URIs
- Captures:
  - Clans
  - Families
  - Annotations
  - Sequence alignments
  - Links to PubMed
  - Database references (InterPro and PROSITE)
- Uses existing vocabularies
  - (in some cases, this may not be a feature)

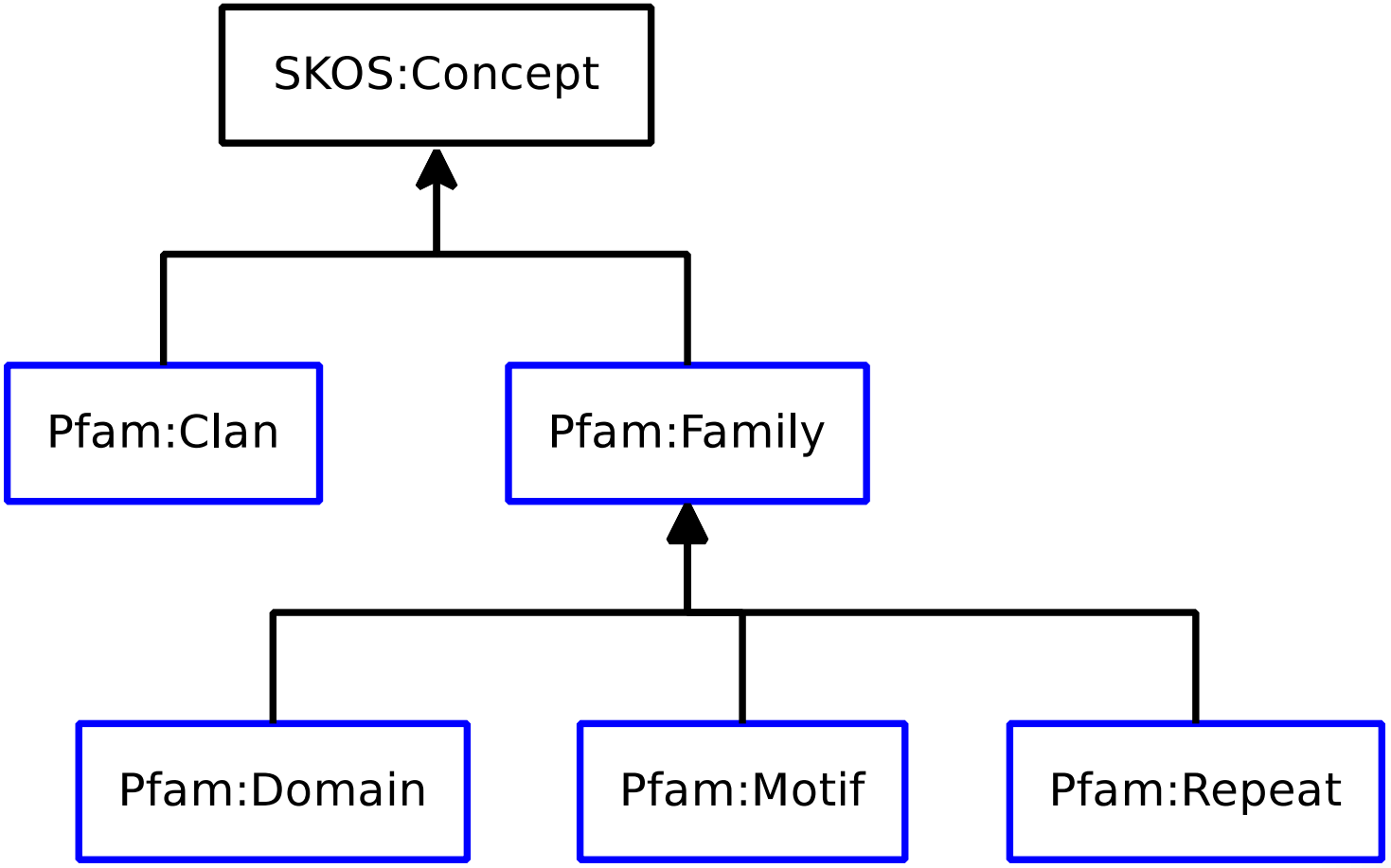
Vocabulary usage

- Uses SKOS vocabulary
  - narrower / broader
  - prefLabel / altLabel
  - definition / scopeNote
  - related
- Uses UniProt core vocabulary
  - Citations
  - Sequence alignments

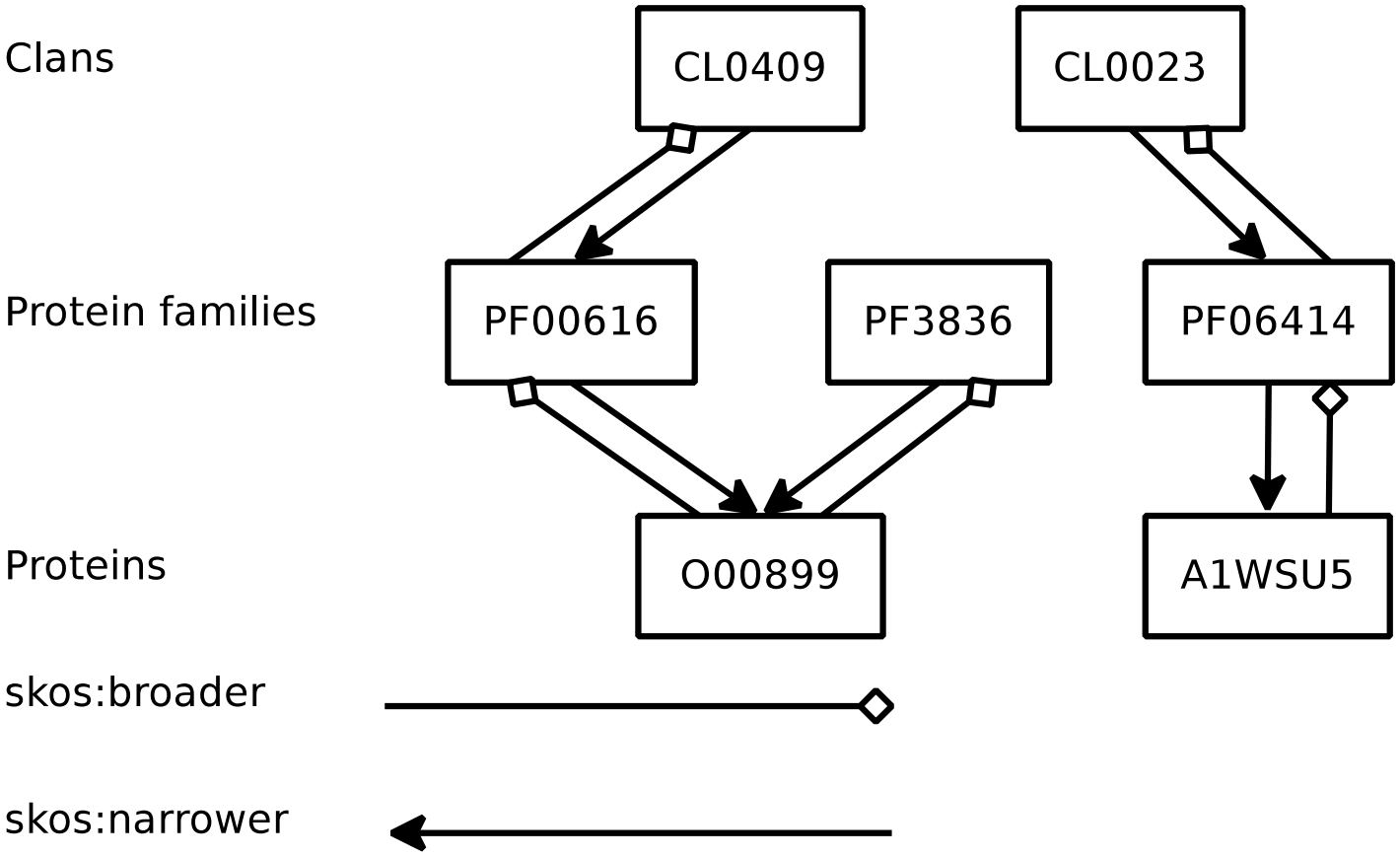
Clan & family structure



Modeling Pfam entities

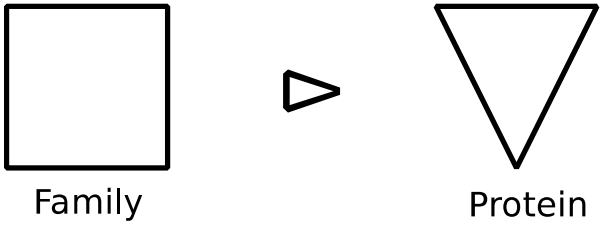


Class membership

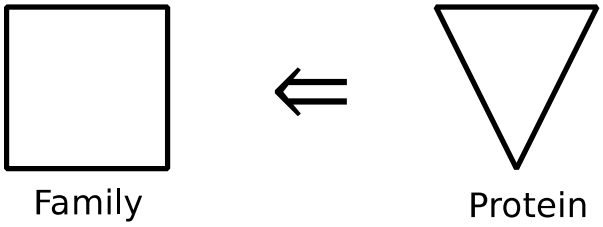


Membership

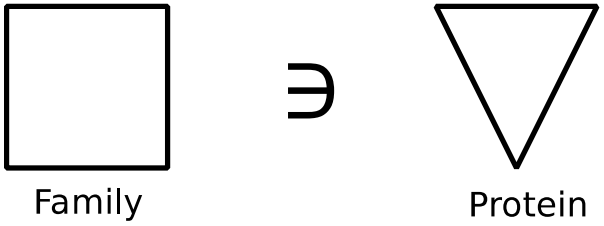
- Broader/narrower might not sufficiently precise



Protein is an example of a family



Protein is a subtype of a family



Protein belongs to a family

- SKOS no longer contains instantive relationships

Sequence alignments

```
<rdf:Description rdf:about="http://www.uniprot.org/uniprot/Q9TTA5">
  <rdf:type rdf:resource="http://purl.uniprot.org/core/Sequence"/>
  <uni:sequenceFor rdf:about="http://pfam.sanger.ac.uk/family/PF00271"/>
  <uni:begin>740</uni:begin>
  <uni:end>818</uni:end>
  <uni:sequence>KELERKRVQHIRIDG.....................STSSADRETSASSFSCPRA.....
.....................LRGVLSITAANMGLTFSSADLVVFGEL.....................FWNPGV..
.........LMQAEDRVHRIG</uni:sequence>
</rdf:Description>
<rdf:Description rdf:about="http://www.uniprot.org/uniprot/Q52902">
  <rdf:type rdf:resource="http://purl.uniprot.org/core/Sequence"/>
  <uni:sequenceFor rdf:about="http://pfam.sanger.ac.uk/family/PF00271"/>
  <uni:begin>239</uni:begin>
  <uni:end>333</uni:end>
  <uni:sequence>GRFGDD.TAIVPLYG.....................NLSQKEQDAAIRPAPKGTR.....
.....................KIVLATSIAETSITIDGVRIVVDSGLQRLPVFEAA..TGITRLETVRVSKAS..
.........ADQRAGRAGRTE</uni:sequence>
</rdf:Description>
```

Automated translation

- Developed translation program in Scala
- Input:
  - Pfam-C (clans) (≈360 kb)
  - Pfam-A (curated families) (≈190 mb)
  - uniprot_sprot (proteins) (≈350 mb compressed)
- Output:
  - Single RDF file (≈333 mb)
- Source is available on GitHub
  - http://github.com/mtomko/pfamskos

http://www.scala-lang.org/

Open problems for Pfam

- Need vocabulary for class membership:
  - skos:narrowerInstantive and skos:broaderInstantive
    - Deprecated after SKOS-Core 1.0 Guide
- Need a better model for sequence alignments
- Both of these could be easily defined using OWL
  - But should they have to be?
  - Does something similar already exist?
  - How do I find it?

Future work

- Extract all external database references
- Capture HMM parameters
- Infrastructure improvements
  - Hosting
  - Separate URLs for entities
  - Improved codebase

Problems for biology data

- Existing linked data vocabularies are too general or too specific
- Vocabularies are hard to find
- Insufficient or inadequate software tools
- Linked data specifications are daunting to outsiders

Acknowledgements

Simmons College:

Kathy Wisser and Candy Schwartz

Graduate School of Library and Information Science

Tufts University:

Caroline L. Dahlberg

Sackler School of Graduate Biomedical Sciences

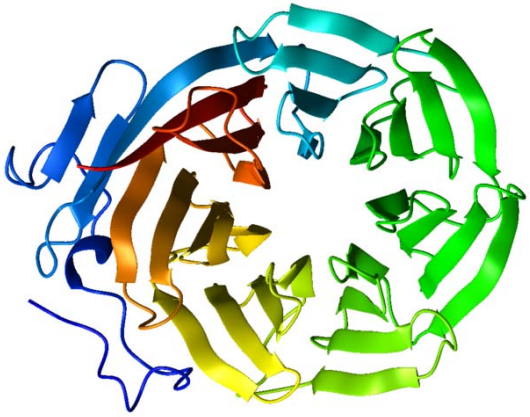
The Broad Institute:

Tom Green and David Root

The Broad Institute RNAi Platform

Thanks!

Project	http://web.simmons.edu/~tomko/pfam
Code	http://github.com/mtomko/pfamskos
Contact	mark.tomko@simmons.edu



Adapted from Lehninger, 3rd Ed.
Structure image from the PDB

BROAD INSTITUTE

http://www.broadinstitute.org/

SIMMONS Graduate School of Library and Information Science

http://www.simmons.edu/gslis/

How much data?

- Human genome contains 20-25K genes
- Human DNA contains 3 billion base pairs (A,G,C,T)
- UniProt/TrEMBL database contains 16,504,022 protein annotations as of August 2011
- Pfam contains:
  - 458 clans
  - 12,273 families
  - 8,729,906 sequences

(http://www.intel.om)