**Automated interoperation and routing of protein sequence, structure and function information**

Erik Schultes, GO FAIR

December 2018

**Background and use case**

Genomic and proteomic sequence data are central to modern approaches in the life sciences and especially in biomedical research (most importantly in personalised medicine). Despite the conceptual simplicity behind modern protein biology (i.e., sequence implies structure implies function) the relevant data linking these levels of organisation are diverse, heterogeneous, and fragmented. Even when located, these complex data are often hard to interpret and can not be reused without tedious manual interventions. In this sparse and rugged data landscape, our ability to predict functional implications of sequence variants or to engineer novel protein function remains mediocre and *ad hoc*. The problem is systemic, and will not be solved by increasing amounts of data. To the contrary, emerging technologies such as automated laboratory systems and applications of machine learning will likely compound the problem of data fragmentation even further. Here, we propose a simple Digital Object approach to help render FAIR, all information about protein sequence-structure-function relations. The machine-assisted stewardship of protein data will have manifold implications throughout the life sciences, and will impact the routine exchange, combination, analysis and tracking of the collective knowledge of proteins, now and into the future.

**Protein sequences as digital objects**

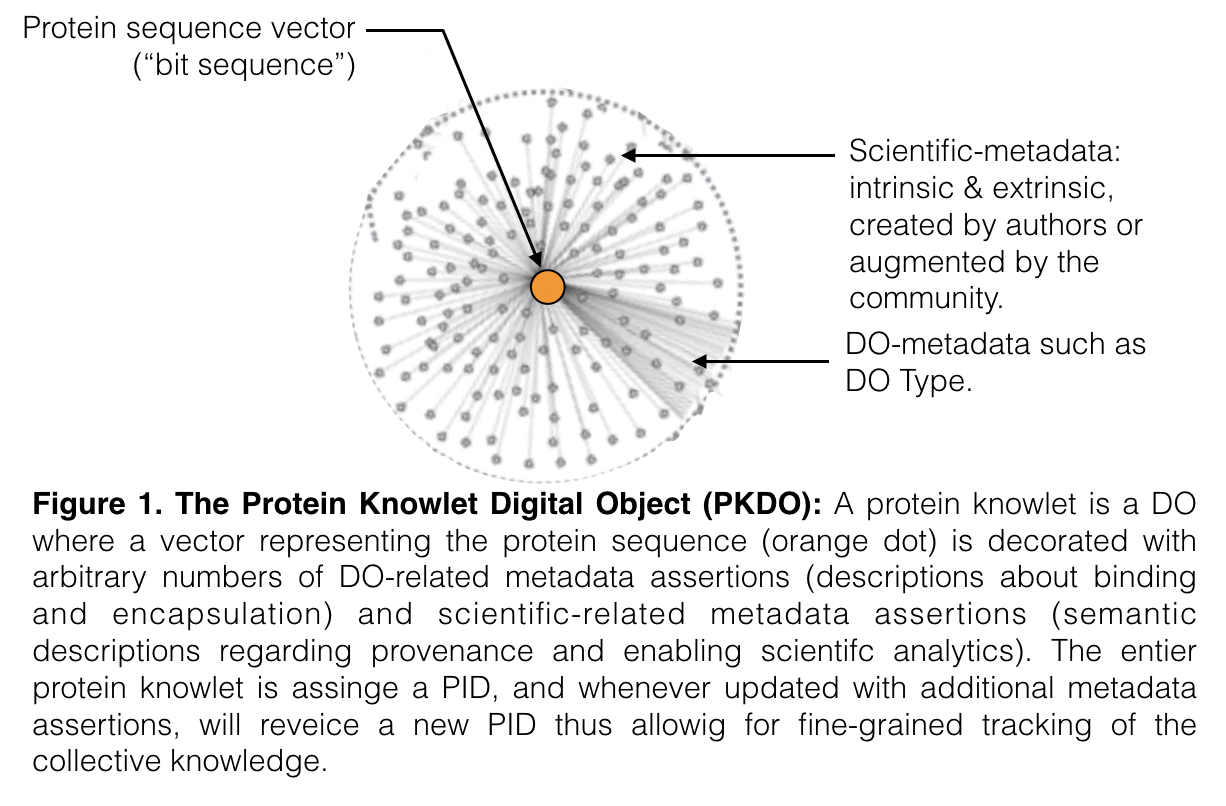
The core data model is itself inspired by the remarkable conceptual simplicity of proteins themselves as genetic entities. From a physical point of view a protein is a linear polymer whose sequence is composed from the 20 different kinds of amino acids. This physical polymer structure can be abstracted as a vector sequence over an alphabet of 20 symbols. Using an appropriate representation, this vector is susceptible to high-performance analytics (e.g., sequence alignment and homology assessments).

However, borrowing from the ‘knowlet’ model of Digital Objects [1], this vector can be extensively ‘decorated’ with arbitrary numbers of metadata assertions relating to all DO-related and scientific-related metadata. The result is a (potentially) long list of metadata assertions (subject-predicate-object triples), each having the same subject (the protein sequence vector) (**Figure 1**).

The protein sequence vector can be annotated with the appropriate abstraction, binding and encapsulation descriptions allowing automated routing of data, services on those data, and access to required compute and storage resources as envisioned by the FAIR Principles.

The protein sequence vector can also be annotated with scientific metadata, both intrinsic (describing the vector as protein sequence, its origins as observational and experimental data, etc) and extrinsic (describing features about the protein as they are subsequently discovered by others, such as interactions with other proteins, relations to diseases, predicted secondary structures via various neural network algorithms, by x-ray diffraction or cryo-electron microscopy). Scientific metadata assertions can refer to any aspect of proteins, including folded structure (represented in any format) and functions (be they biochemical functions or biological phenotypes). The Protein Knowlet Digital Object (PKDO) can thus unify, for both humans and machines, what is presently a fragmented landscape of diverse protein information.

Furthermore, although the protein sequence vector and metadata assertions are themselves DOs, taken together a protein knowlet is also a DO. As such, the PKDO could also be assigned a PID. One approach to creating a system of PIDs for protein knowlets, is by hashing the total content of the knowlet at any given point time [2]. This hash code can then be used as part of a handle PID system using appropriate IDs. This PID system automatically and intrinsically links a unique and persistent identifier to the unique content of a PKDO. Any future discoveries, revisions, disputes regarding the scientific understanding of the protein, captured as additions or deprecations of the knowlet metadata assertions, will be automatically tracked by re-hashing the knowlet and creating novel PIDs. The sequence of PIDs would then preserve a detailed conceptual provenance train of that protein over time.

****

**Proposal**

To build an Open platform for the collective creation of protein knowlets as DOs. Specifically:

1. To implement the PKDO data model described above;
2. To populate the PKDO with existing data (sequence) and metadata (DO and scientific);
   1. Peptides
      1. Some known peptide sequences
      2. Large numbers of theoretical peptide sequences
   2. Human proteins
      1. And their known variants
      2. Large numbers of theoretical human variants
3. To demonstrate DOIP functionality on PKDOs: automatic routing of new protein knowlets to appropriate services such as;
   1. Analytics
      1. Sequence similarity searches (“has this variant been observed before?”);
      2. Structure prediction services (every sequence will now have a detailed structure profile based on the consensus of an ensemble of prediction methods)
      3. Biomedical analytics (using the Euretos Platform).
   2. Compute: These services may be run on a distributed network of public and commercial compute resources   
      1. NOMAD
      2. Amazon
      3. Nerdalize
4. To allow PKDO updates and corresponding automatic minting of new PIDs;
   1. Edits by the creator
   2. Edits by the community at large (including those by machines)
   3. Edits via a FAIR Compliant and Flexible Semantic Interoperability Framework (F2SIF)
      1. Enables semantic annotation and encoding of scientific content used to carry out scientific analytics especially between knowledge domains such as the “sequence”, “structure” and “function” knowledge domains where methods, data formats and jargon tend to be siloed.
      2. Enables relations to be drawn between semantic categories defined in different semantic domains
      3. Enables these relationships to be stored in FAIR compliant ways and registries so they are trackable, and can be shared with others
      4. Enables identifiable collections of such relationships, addressing certain specific needs
      5. Enables the reuse of these registered collections of relationships (user can modify them and register them again to enable ongoing reuse)

**References**

[1] B. Mons: FAIR Science for Social Machines: Let’s Share Metadata Knowlets in the Internet of FAIR Data and Services;<http://www.data-intelligence-journal.org/static/publish/B5/D1/28/A85A05492CA4E681827A4F8BF7/Barend_Mons.pdf>

[2] Trusty URI approach;<http://trustyuri.net>