

Type Provider for the UniProt Knowledge Base

Kseniia Popova | Faculty of Mathematics and Physics | Charles University | 2025

Introduction

The `UniProtProvider` is a domain-specific type provider in `F#` that integrates `UniProt` online knowledge base schema into the programming language type system, enabling *intuitive exploration* of protein data directly in the IDE. By abstracting `UniProt`'s REST API into a fluent, *type-safe* interface, the solution reduces boilerplate, therefore simplifying *programmatic access* for researchers in the field of bioinformatics.

F# Type Provider

An `F#` type provider is a mechanism that enables statically-typed access to external data sources from the `F#` programming language. It supports providing types, properties and methods for use in a program. Types and members are dynamically generated when they are explicitly referenced in the code. This approach allows *on-demand*, direct integration of *Internet-scale* data sources while maintaining a *strongly typed* representation in the code.



“`UniProt` is the world's leading high-quality, comprehensive and freely accessible resource of *protein* sequence and functional information”. It includes `UniProtKB` - the protein knowledge base.

Requirements

1. Provide static access to individual protein and taxonomy entries available from `UniProt`
2. Support multiple chained filtering operations to iteratively narrow the result set
3. Guide the user through available results and filtering options directly in the editor

Properties Generation

The properties are generated dynamically upon member access based on the provided organism and other available filters.

Entry	Variant viewer	Genomic coordinates	Publications	External links	History
Names & Taxonomy ⁱ					
Protein names ⁱ					
Recommended name	Clarin-2				
Gene names ⁱ					
Name	CLRN2 				
Organism names					
Taxonomic identifier ⁱ	9606 (NCBI )				
Organism ⁱ	Homo sapiens (Human)				
Taxonomic lineage ⁱ	cellular organisms > Eukaryota (eucaryotes) > Opisthokonta > Metazoa (metazoans) > Eumetazoa > Bilateria > Deuterostomia > Chordata (chordates) > Craniata > Vertebrata (vertebrates) > Gnathostomata (jawed vertebrates) > Teleostomi > Euteleostomi (bony vertebrates) > Sarcopterygii > Dipnotetrapodomorpha > Tetrapoda (tetrapods) > Amniota (amniotes) > Mammalia (mammals) > Theria > Eutheria (placentals) > Boreoeutheria > Euarchontoglires > Primates > Haplorrhini > Simiiformes > Catarrhini > Hominoidea (apes) > Hominidae (great apes) > Homininae > Homo				

Solution

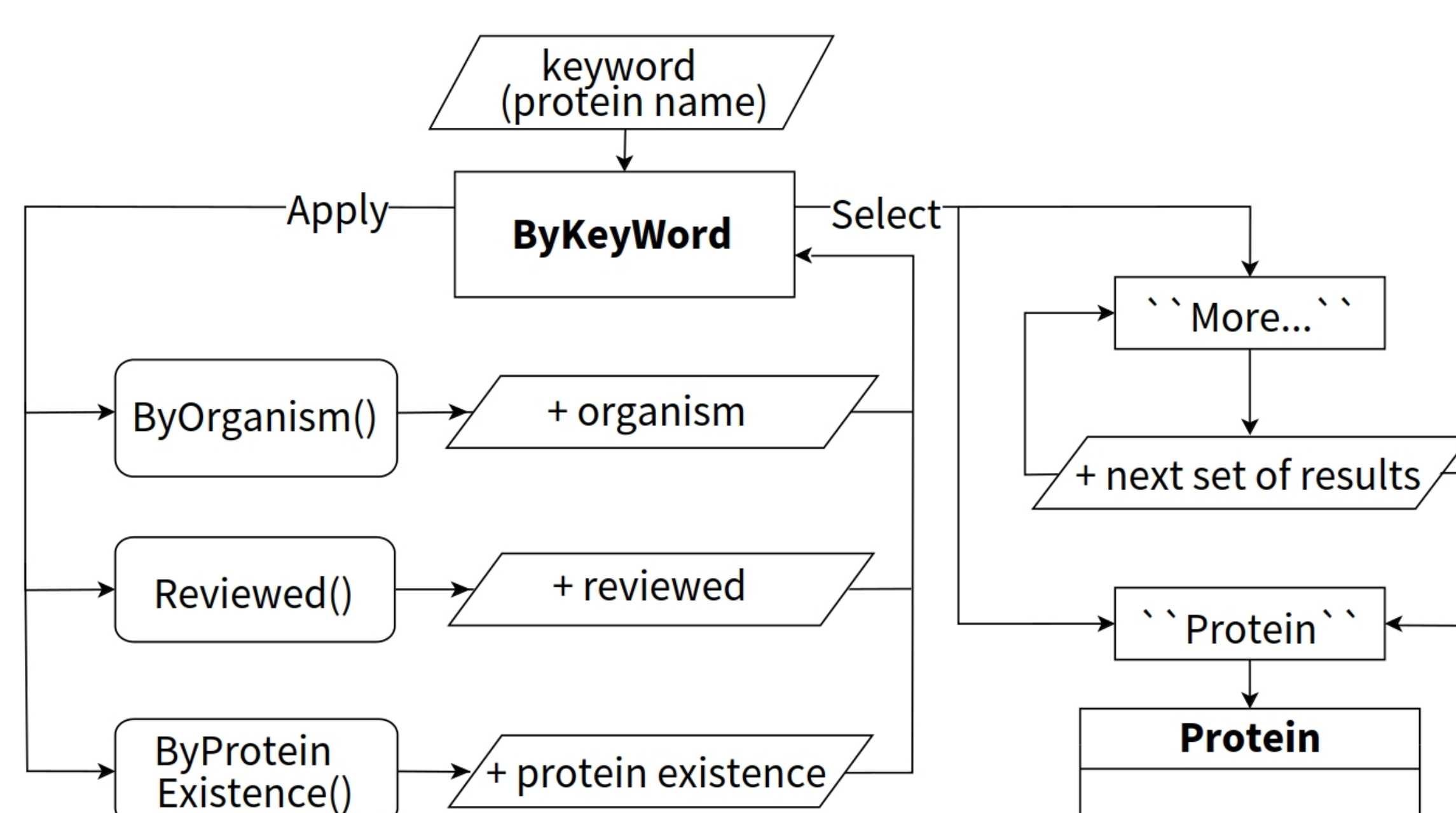
There are four *type providers* contained in the `UniProtProvider`:

- `ById` - Retrieves a single protein entry
- `ByKeyword` - Retrieves all protein results matching the specified keyword
- `ByTaxonId` - Retrieves a single taxonomy entry
- `ByOrganism` - Retrieves all taxonomy results that match the provided organism name

```
type Keratin = UniProtProvider.ByKeyword<"keratin">
let humanKeratin = Keratin().ByOrganism<"human">()
humanKeratin.
```

Keratin, type II cytoskeletal 78 (K2C78_HUM...
 Keratin, type II cytoskeletal 79 (K2C79_HUM...
 Keratin, type II cytoskeletal 8 (K2C8_HUMAN)
 Keratin, type II cytoskeletal 80 (K2C80_HUM...
 Keratin-associated protein 4-1 (KRA41_HUMAN)
 Keratin-associated protein 4-11 (KRA41_HUMA...
 Keratin-associated protein 4-6 (KRA46_HUMAN)
 Keratin-associated protein 5-7 (KRA57_HUMAN)
 Keratin-associated protein 5-9 (KRA59_HUMAN)
 Keratin-like protein KRT222 (KRT222_HUMAN)
 More...
 Reviewed

Among available filters for proteins are: `Reviewed`, `ByOrganism`, `ByProteinExistence`. `FindRelated` can be used to find proteins associated with selected organism. The `More...` property allows to retrieve the subsequent set of results.



```
UniProtProvider.ByOrganism<"human">()
  .Homo sapiens (9606)
  .FindRelated()
  .Clarin-2 (CLRN2_HUMAN)
    .Aldo-keto reductase family...
    .Alkylidihydroxyacetonephosp...
    .Bridge-like lipid transfer...
    .Ciliated left-right organi...
    .Clarin-2 (CLRN2_HUMAN)
    .Coiled-coil domain-contain...
    .Cysteine-rich tail protein...
    .Deoxyribonuclease-2-alpha ...
    .E3 ubiquitin-protein ligas...
```

Senya-P/UniProtProvider | UniProtProvider

Supervisor: Mgr. Tomáš Petříček, Ph.D

Department of Distributed and Dependable Systems

