ELIXIR Data Platform Implementation Study 2018

**Increasing Interoperability between ELIXIR Protein Structure and Sequence Resources (CATH, SWISS-MODEL, PDBe and InterPro) and Expanding these Resources with 3D-Models of CATH Domains, built by SWISS-MODEL**

|  |  |  |  |
| --- | --- | --- | --- |
| Node | Name of PI | Role  (Lead or Member) | Person Months |
| UK | Christine Orengo | Lead | 8 |
| Swiss | Torsten Schwede | Lead | 8 |
| EBI | Rob Finn | Member | 1 |
| PDBe | Sameer Velankar | Member | 1 |
| Total | | | 18 |
| Work period | Start date 1st May 2018; End date Jan 31st 2019 | | |

### Abstract: This project will increase interoperability between four ELIXIR resources (CATH, SWISS-MODEL, InterPro and PDBe), three of which are Core Resources, by building APIs that facilitate the import and export of data between them. The ultimate goal is to improve provision of 3D-Models for protein domain sequences via CATH, SWISS-MODEL and InterPro. Less than 10% of known sequences have experimentally characterised 3D structural information and yet this data is often essential for understanding the protein’s molecular function and biological role and for determining whether residue mutations could damage the protein and lead to disease. So this integration is very timely as it will enhance links between sequence and structure data. APIs will be built using well-established protocols and as well as promoting interoperability, and therefore sustainability, we will expand the data in each resource to ensure they serve a wider community of biologists.

***Background/context:*** Thanks to revolutionary technologies like Next-Generation Sequencing, there is a vast expansion in sequence data and a rapid accumulation of data on residue mutations in proteins (ie non-synonymous single nucleotide polymorphisms nsSNPS) linked to phenotypes associated with human health (eg residue mutations linked to rare diseases) or food security (eg residue mutations affecting drought resistance in wheat). To determine whether a mutation is truly linked to a phenotypic effect it is important to assess the likely consequence of the mutation as a rationale for predicting whether it could be ‘driving’ the phenotype. The structure of a protein illuminates the mechanism by which a protein works and is often essential for interpreting the impact of genetic variations. Examining residue mutations, insertions and deletions in their 3D context helps explain how they could affect the protein stability or function and lead to disease. Although structural data in the PDB continues to increase, there is still more than 200-fold sequence data and for many important model organisms like human, <10% of the proteins have known structures. Furthermore, the information on known function or functional sites is much lower (<1% for some organisms) and therefore this is often inferred using family data eg inheriting characteristics from relatives.   
 *This proposal will establish data flows between key ELIXIR resources to enable high quality 3D structural models (SWISS-MODEL) to be predicted for protein domains (CATH) having a known functional role, and for these models to be integrated into the data pipelines of these resources (CATH, SWISS-MODEL) and made available on web pages (CATH, SWISS-MODEL, InterPro).*

* *CATH* combines known information about protein structure, sequence and function
* *SWISS-MODEL* predicts protein 3D structure from sequence
* InterPro provides family analyses of protein domains and information on important sites
* PDBe is the European resource for the collection, organisation and dissemination of data on biological macromolecular structures.

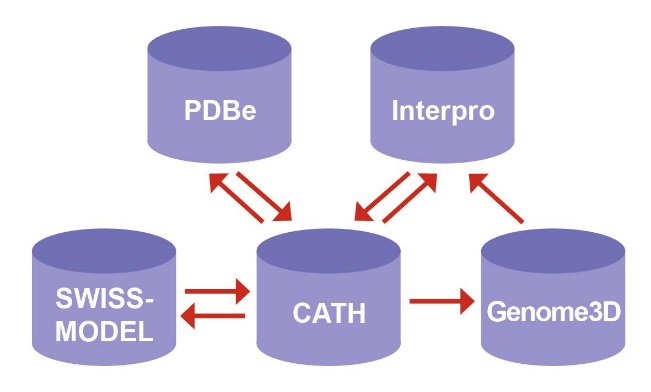
CATH ([www.cathdb.info](http://www.cathdb.info)) focuses on protein domains as these are primary evolutionary and functional units and uses protein structure comparison1 and sequence analysis algorithms2 to continuously classify protein structures from the PDB (using a machine-learning algorithm2 to recognise the majority of evolutionary relationships between domains and manual curation to validate very distant homologues). Currently, 450,000 domain structures are classified, accounting for ~90% of the PDB. Protein domain sequences from UniProt are assigned to superfamilies using HMM protocols2,3 (currently ~93 million sequences). Multiple functional annotations linked to these sequences are imported from a range of public resources (eg Gene Ontology (GO), Enzyme Classification (EC)). Currently, ~6000 CATH superfamilies account for ~70% of domains in completed genomes. CATH recently introduced a sub-classification of superfamilies - the functional family level (FunFams - sequences in a superfamily are clustered based on shared specificity determining residues). FunFams are structurally and functionally coherent4,5 and contain at least one experimentally characterised relative. There are ~64,000 CATH FunFams comprising ~35 million domain sequences.  
 The SWISS-MODEL server (https://swissmodel.expasy.org) was established nearly 25 years ago and pioneered the field as the first fully automated server for protein structure homology modelling on the Internet6 It employs the following tightly linked components: a user-friendly web-based graphical personal workbench6 a fully automated modelling workflow, a curated template database, a state-of-the-art modelling engine and a scoring method for model quality estimation7. The server is widely used by the scientific community worldwide - currently building around 2 models per minute. Model quality is continuously benchmarked together with other state-of-the-art modelling servers as part of the CAMEO project. The modelling service is complemented by the SWISS-MODEL Repository (https://swissmodel.expasy.org/repository) providing access to annotated high-quality protein models generated by SWISS-MODEL for sequences in UniProtKB8. Proteins from selected core proteomes are updated on a regular schedule to account for new template information, while updates for other UniProtKB entries can be initiated interactively by any user from the web interface. The repository contains ~1 million models from SWISS-MODEL and ~125,000 structures from PDB. This collaborative project will:

* Build an API that extracts 3D-template structures from CATH-FunFams for query sequences.
* Build an API that uses CATH 3D-templates to build 3D-models using SWISS-MODEL server.
* Build an API that allows SWISS-MODEL to extract predicted CATH functional site data, for the respective CATH-FunFams, to provide additional annotations for the 3D-models.
* Build workflows that allow CATH to import the SWISS-MODEL models, for selected organisms and Pfam families, into CATH, on a regular basis and export these to InterPro.
* Build an API that, for a given query structure in PDBe, provides information on the number of putative models that can be built with that structure.
* Disseminate the 3D-model data widely via CATH, SWISS-MODEL, InterPro and build training material for ELIXIR structural bioinformatics training workflows.

CATH-FunFams have been endorsed in several ways. Independent blind assessment (CAFA ie CASP for function) highly ranked FunFams in the last two rounds and sometimes top5. In silico validation confirmed functional purity using experimental data from Enzyme Classification and enrichment of known functional residues (eg catalytic from CSA) in highly conserved residues in FunFams5. Of more interest in the context of homology modelling, FunFam relatives with known structure typically superpose within 2A RMSD9 and we’ve demonstrated that using FunFams to select targets for homology modelling gives slightly more high quality models than protocols using state of the art homologue recognition methods, like HHpred, for template selection9.

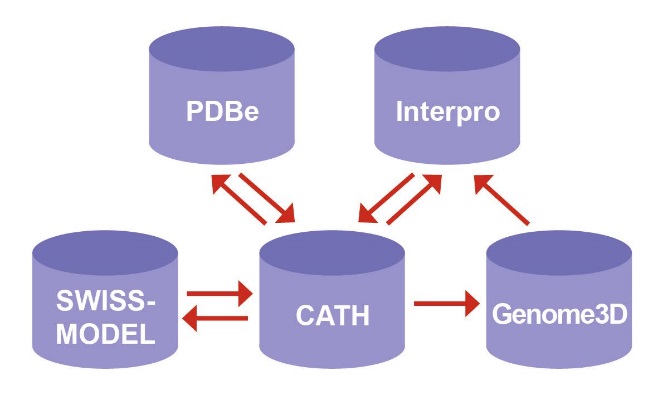
***Description of work***

***1. Build APIs that allow biologists to submit query sequences to CATH or SWISS-MODEL, to obtain 3D-templates subsequently used by SWISS-MODEL to build a 3D-model.*** CATH currently comprises 64,000 functional families (FunFams) containing ~35 million domain sequences. Relatives are highly structurally similar9 (see above) and pilot work building 3D-models for structurally uncharacterised human and fly sequences showed improvement in model quality over a protocol exploiting Hhpred for target selection9. Improvement is largely due to the fact that FunFams are built using a sensitive HMM-HMM strategy optimised to ensure relatives are only clustered if they share sequence patterns likely to be associated with functional properties4. Other template search tools are not constrained to functionally similar relatives and can select relatives more diverged in function - and therefore structure. The following API endpoints will be built:

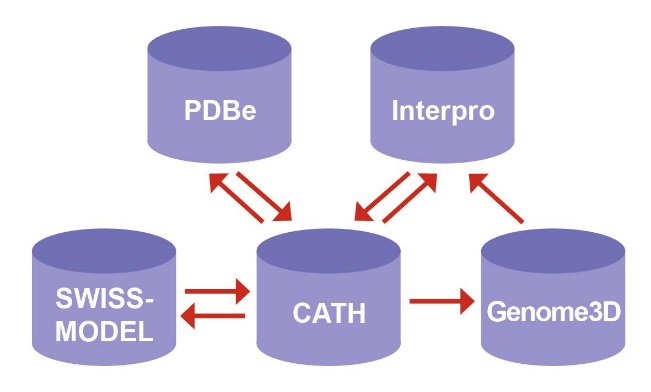
******API1:** **Get3DTemplate** - for a given query protein sequence, identify the most appropriate known structural domain to use for the 3D structural modelling

* *Input*: Protein Sequence
* *Output*: Sequence Alignment, Template PDB structure
* *Type*: RESTful (asynchronous, queued)

The input will be a query sequence for scanning against the library of CATH FunFam HMMs. For sequences matching a CATH FunFam (p value <0.001), the structural domain with highest sequence similarity (from a BLAST search) will be selected and returned to SWISS-MODEL as a 3D-template (ie in PDB file format). The query sequence and the multiple sequence alignment of the FunFam sequence relatives, generated by MAFFT, will also be returned for use by the modelling platform.

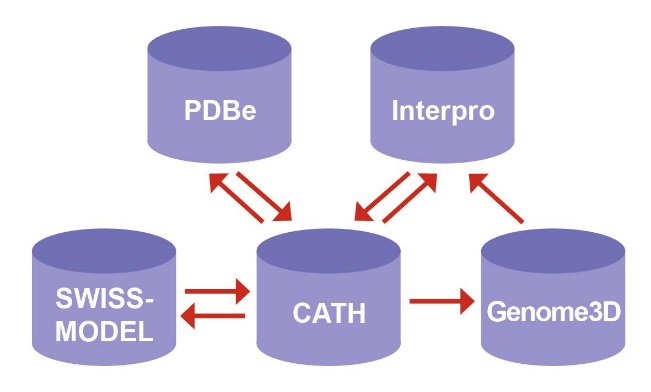
** API2: Get3DModel** - for a given sequence, alignment and template 3D structure, provide high quality 3D structural model(s)

* *Input:* Protein Sequence, Sequence Alignment, Template PDB Structure
* *Output:* Modelled PDB Structure
* *Type:* RESTful (asynchronous, queued)

****This API provides access to the core functionality of SWISS-MODEL: predicting a high quality 3D structural model from a query protein domain sequence.   
 **API3: GetFunData** - provide access to functional terms and functional site data for the respective functional families (to provide additional annotations for query sequences)

* *Input:* Protein Sequence
* *Output:* JSON data structure
* *Type:* RESTful (synchronous)

When CATH functional families have high information content, the multiple sequence alignments can be used to identify highly conserved residues likely to be important for function or folding. These sites – termed *FunSites* are downloadable from CATH (via annotated STOCKHOLM alignments). *GetFunData* takes a query sequence as input and returns the site data for the best matched FunFam (p value <0.001).

** API4: GetPutativeModelSequences** – provide information on the number of potential models that can be built by a query structure. Used by PDBe (see figure).

* *Input:* Protein Sequence (eg from a 3D structure in PDBe)
* *Output:* UniProt IDs for sequences for which domain models can be obtained via linked CATH and SWISS-MODEL APIs
* *Type:* RESTful (asynchronous, queued)

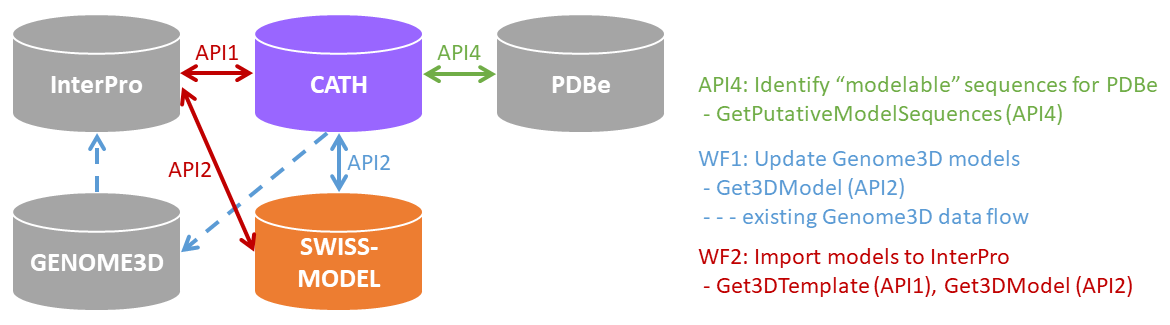
***2. Build workflows for importing data into CATH, Genome3D, InterPro:*** The API endpoints will allow SWISS-MODEL to exploit CATH FunFam data at any stage of the CATH release cycle and will also be available to other bioinformatics groups, downloadable via the CATH and SWISS-MODEL websites. In addition to these API endpoints, we will also create clients that will establish protocols for implementing CATH / SWISS-MODEL data into the following workflows:

**WF1: Update 3D structural models in Genome3D and export to InterPro**

CATH is a member of the Genome3D (http://genome3d.eu) consortium – involving 5 UK resources (PHYRE, GenTHREADER, FUGUE, SUPERFAMILY, CATH) disseminating structural data for selected model organisms (eg human, fly, mouse) and domain representatives from structurally uncharacterised Pfam families. We will implement clients for the *Get3DModel* API that will:

* Submit: CATH domain query sequences, alignment to the Functional Family and selected 3D-template to SWISS-MODEL.
* Return the 3D-model, together with information on model quality.

Once established, this pipeline will be used to regularly obtain 3D-models, built by SWISS-MODEL, for the Genome3D model organisms and Pfam families. We will make these models available from the CATH and Genome3D websites together with functional annotations and functional site data stored in CATH. BBSRC funding is currently supporting the development of an API for exporting 3D-models from Genome3D resources to InterPro so we will use this to regularly export the SWISS-MODEL models to InterPro. InterPro are currently building new webpages and a viewer to display structural information for a given UniProtKB sequence.

**WF2: Import 3D structural models into InterPro.** We will establish a workflow to provide InterPro with SWISS-MODEL models, for CATH domains, on demand, using the *Get3DTemplate* and *Get3Dmodel* APIs (see figure).

**3. Disseminate the data via CATH and SWISS-MODEL and build training material.** The 3D-models for selected model organisms and Pfam families will be disseminated widely via the CATH and SWISS-MODEL websites. Both resources are highly accessed (CATH has ~9000 unique users/per month (captured using awstats) and SWISS-MODEL has ~27,000 users/month (via google analytics). We will also generate training material including videos and web-based tutorials. Material will be developed for ELIXIR training workflows currently being built with funding from ELIXIR Excelerate. These guide the user on how to obtain homology models from European resources and use the models to view known and predicted functional sites and assess whether mutated residues in the protein lie near these sites and could affect function or stability. We will also develop a webinar explaining the data and present the data and demos of the APIs at an ELIXIR all-hands meeting.

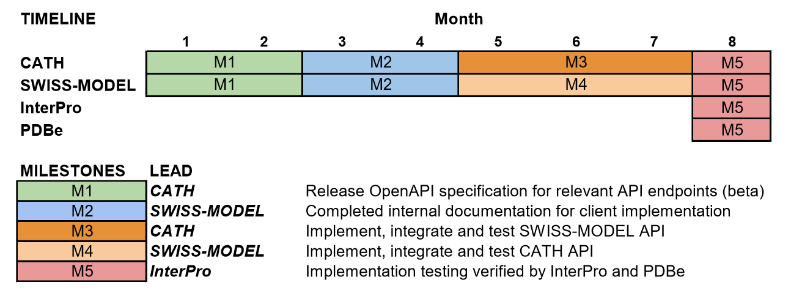
**Alignment with** [**Evaluation Criteria**](https://drive.google.com/file/d/0B60jEEGzhM72cHNXSHZQQ0NQS3c/view?usp=sharing)**:** This work will extend the structural data provided by three ELIXIR resources (CATH, SWISS-MODEL, InterPro). All are participating in the BioSchemas initiative, linking ELIXIR resources via standard ontologies, but this project will build much deeper levels of integration, promoting sustainability of the resources. All the resources are widely used (monthly web accesses of 1m (CATH), 36m (InterPro), 90,000 models/month (SWISS-MODEL), and have up-times >99%. CATH is presented regularly at EBI workshops and technology tracks in ISMB, ECCB. SWISS-MODEL is also regularly presented in conferences (eg [BC]2, ECCB). They are linked to other ELIXIR resources (eg UniProt, STRING) and will collaborate together and with other groups to develop FAIR protocols for increased interoperability and re-useability of data.

The suite of APIs built by the project will also enable deeper integration with other protein resources across Europe and internationally. CATH provides superfamily HMMs to InterPro for the ELIXIR metagenome case study. The work proposed here will allow query sequences in metagenomes to be annotated with CATH based SWISS-MODEL models to aid interpretation of sequence divergence in different communities. As part of the Genome3D-InterPro collaboration we are building a 3D-viewer for domain structures, using BioJS, that will allow biologists to benefit from intuitive rendering of the SWISS-MODEL models.  
 This project will use an accurate template selection protocol, based on CATH-FunFams, together with a state-of-the-art modelling protocol established by SWISS-MODEL, to build 3D models for selected organisms and uncharacterised Pfam families, disseminated by CATH, SWISS-MODEL, Genome3D and InterPro. This work is timely as Genomics England, the Swiss Personalized Health Network (SPHN) and other related initiatives across Europe are accumulating vast amounts of genotype data for rare diseases and cancers which would benefit significantly from interpretation of residue mutations in the context of structural data. The new data and availability through well-established websites will extend the value of these resources to a wider biomedical, clinical and pharmaceutical community (for drug design) thereby increasing the sustainability of all these resources through much greater demand and therefore more sustained funding from national and international funding initiatives. All the resources have >20 years track record of sustained funding.

**Expected outcome:** This collaboration will increase interoperability between Core ELIXIR resources and promote deeper integration of these protein sequence and structure resources and mechanisms that can be exploited to integrate other related resources in the future. In addition, it will establish workflows that expand the data available in all resources and lead to wider dissemination of the data and training workflows on accessing and using the data. The project is timely in benefiting from existing collaborations between CATH, Genome3D, InterPro, PDBe, ELIXIR TESS. No person months are being requested for work done on other projects, that will also benefit this project.

**References:** 1) Redfern, OC et al. (2007). PLoS Comput. Biol., 3(11), e232. 2) Dawson, NL et al. (2017). Protein Bioinformatics, 79-110. 3) Dawson, NL et al. (2017). Nucleic Acids Res., 45(D1), D289-D295. 4) Das, S et al. Bioinformatics, 31(21), 3460-3467.11) 5) Jiang, Y et al. (2016). Genome biology, 17(1), 184. 6) Guex, N. et al. (2009) Electrophoresis, 30(S1), S162-S173. 7) Benkert, P et al. (2011) Bioinformatics, 27(3), 343-50. 7) Studer G et al. (2017) Nucleic Acids Res. 45(D1):D313-D319. 9) Lam, SD et al. (2017) Acta Cryst Section D: Structural Biology, 73(8), 628-640.

**Project Timeline**

We will hold fortnightly calls for CATH/SWISS-MODEL teams and with InterPro/PDBe as required.  
**Milestones from project start, for each node:**

M1: Completed internal release of OpenAPI specification covering relevant API endpoints for CATH + SWISS-MODEL (Month 2, UCL Lead)

M2: Completed internal documentation for client implementation of CATH + SWISS-MODEL annotations (Month 4, SWISSMOD Lead)

M3: Implement and test SWISS-MODEL API (Month 7, CATH Lead)

M4: Implement and test CATH API (Month 7, SWISS-MODEL Lead)

M5: Implementation testing verified by InterPro (Month 8, InterPro Lead)

M6: Implementation testing verified by PDBe (Month 8, PDBe Lead)

**Deliverables:**

D1: OpenAPI specification for CATH / SWISSMODEL API (Month 8, CATH Lead)

D2: Documentation / scripts for client implementation of CATH + SWISS-MODEL annotation pipeline (Month 8, SWISS-MODEL Lead)

D3: Implementation and documentation tested and verified (Month 8, InterPro Lead)

D4: Implementation and documentation tested and verified (Month 8, PDBe Lead)