# The NewProt Self-Service Portal for Protein Engineering

Andreas Schwarte[1], Hanka Venselaar[2], Peter Haase[1], Gert Vriend[2]

1 fluid Operations AG, Altrottstraße 31,
69190 Walldorf, Germany
{andreas.schwarte, peter.haase}@fluidops.com

2 Nijmegen Centre for Molecular Life Sciences
CMBI, Nijmegen, The Netherlands
{Gerrit.Vriend, Hanka.Venselaar}@radboudumc.nl

**Abstract.** The NewProt portal is a platform that gives users access to a broad range of protein engineering tools and services. On an embedded website users can interactively work with these tools and perform their computations supported by integrated workflows. The portal provides a tight integration of standard tools for protein engineering and workflow development.
The concept of the NewProt project is to combine and integrate the best European softwares into a homogeneous portal for in silico protein engineering. The predictions made by the portal regarding the effect of mutations on protein stability, selectivity, activity, production level, etcetera, will be experimentally validated to iteratively improve the quality of the software.

**Keywords:** protein engineering, semantic technologies, self-service portal

## 1 Introduction

Researchers in industry and academia working on protein engineering can select from a broad variety of tools and services to address their molecular questions in common workflows. Many of these workflows consist of multiple tools that work consecutively with as consequence that intermediate results have to be passed between them, a process that would require manual operations in the absence of workflows. Although there exist standards for data formats as well as interfaces for exchange of intermediate results, a tight integration of the different tools addressing practical, *in silico* protein engineering problems is still missing.
The NewProt Self Service Portal (SSP) provides a platform that gives users access to a broad range of tools and services. On an embedded website users interactively work with these tools and perform their computations supported by integrated workflows. The SSP provides a tight integration of standard tools for protein engineering and workflow development. Throughout the workflows intermediate results can be retrieved and visualized in a unified user interface.

The SSP is based on the open source Information Workbench [1]. The SSP provides a working environment to easily interact with all NewProt resources. To this end, the SSP enables integrated database and software access. Software and databases are fully interoperable, i.e., users do not need to store in-between results and will not need to worry about file formats etc. This interoperability requires that all data types will be syntactically and semantically described in a common format: The NewProt Format (NPF).

## 2 Use Cases

We present an overview of typical use cases in interacting with the portal. Tutorial videos for these use cases are also available under https://www.youtube.com/NewProtTutorials.

***Creating a Project*** The user has several options to create a new project. The user can submit the unique Uniprot-code or sequence of the protein of interest and the corresponding project page is then generated automatically. Alternatively, the MRS search-software allows the user to find the protein of interest.

***Add an experimentally solved structure*** Experimentally solved structures can be added to the project page by using the "import PDB" option. This option performs a BLAST search using MRS against the PDB database and returns a list of all PDB-files that contain the experimentally solved protein. One of these files can be selected and will be imported in the portal.

***Add a homology model*** The user has three options to add a predicted homology model. 1) upload a selected file from the local computer. 2) find an external model in the Protein Model Portal. This option opens the PMP website for the protein of interest and shows all available models. Selected model(s) will be automatically imported. 3) Build a model using the YASARA software. In case the user has a YASARA license, a model for the protein can be built using the YASARA modelling server. The computed model is uploaded to the portal when finished.

***Protein Alignment*** A protein-family specific alignment can be obtained using the 3DM software. This alignment will only be shown when the user has a 3DM-license. Alternatively, an HSSP-alignment will be shown instead.

***Uploaded Resources, Project Management & Other*** The user can manage his own projects, invite others to projects and delete projects. Also, files can be easily uploaded and shared among contributors of the same project using the "upload Resources" option.

***Visualization with YASARA*** The YASARA software is used to visualize the results. YASARA scenes are generated in which the amino acids are coloured according to the values that were calculated by the Web Services. These YASARA scenes can be downloaded and opened with a local YASARA installation. The free YASARA_View software can do all visualisation for users who do not have a YASARA license.

***Hotspot computation using HotSpot Wizard*** The HotSpot Wizard analyzes a protein and will pinpoint those residues that are likely to change specificity, activity or selectivity of the protein when mutated.

***Computations using other services*** Many different calculations can be performed on the protein structure using the WHAT IF Web Services. These calculations include: metal and ligand contacts, the formation of saltbridges, symmetry contacts, variability, etc. Other compute services can be integrated through standardized extension points.

***Mutation analysis by HOPE*** The HOPE webserver can be used to predict the effects of a mutation. The user can indicate the position and mutant and a report will be generated that will describe the putative effects of that mutation on the proteins structure and function.

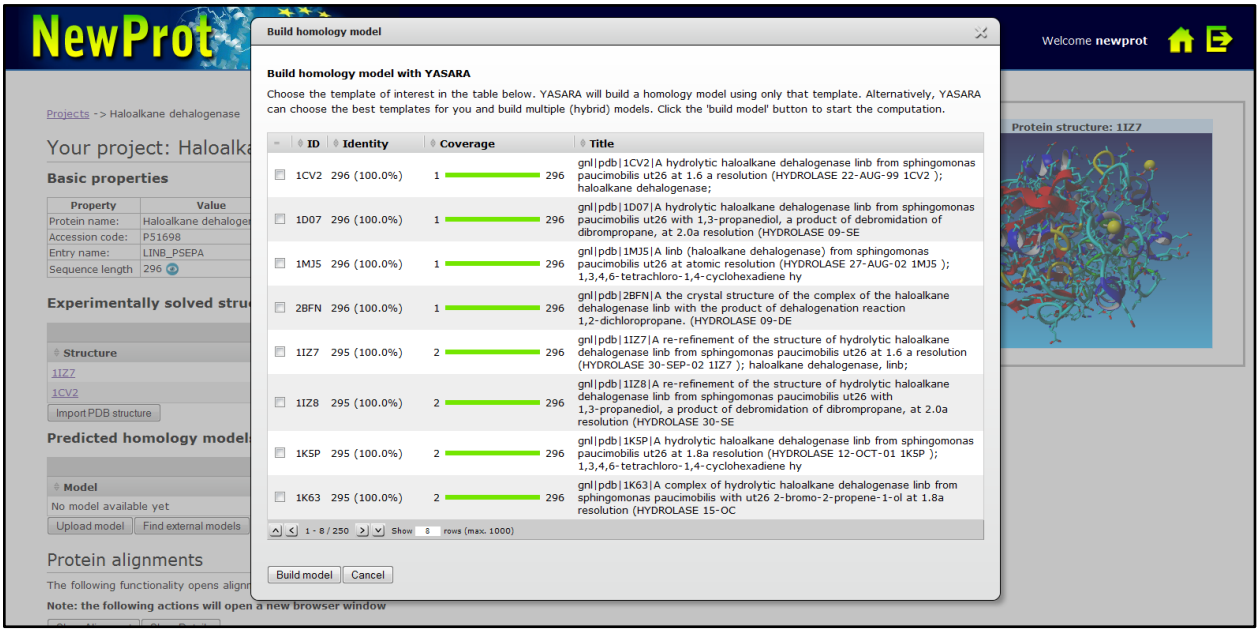


**Fig. 1.** Screenshot of the NewProt portal: Building a model with YASARA

## 3 Architecture

The architecture of the NewProt portal is designed using the Information Workbench. Serving as a platform for Linked Data, the Information Workbench allows for collaboration, integration of public as well as private data and services, and analytics on the data. Users benefit from a unified view of both the data and the integrated resources. The Information Workbench is the layer of integration, providing this unified view of the data and components. The architecture of the portal is divided in three layers centered around the core platform:

1. *Remote Data and Services*
2. *Integrated Processes and Workflow*
3. *Presentation Layer*

The first layer covers all NewProt software components, including the MRS database, a YASARA server instance, the HotSpot Wizard, RESTful or SOAP web services, and (access to) external databases. Through standardized web protocols and the common NPF Format described below additional external *compute services* can be integrated.

The second layer comprises the integrated services of the portal required for the interaction between the NewProt software components and the Information Workbench. In addition, this layer contains the local database that stores PDB models, NPF files, and metadata. All metadata is managed in RDF according to an ontology that describes projects, users, services, applications and protein metadata. This ontology serves as the structural backbone for the backend of the platform as well presentation layer.

The Presentation Layer is built with the UI components of the Information Workbench, in particular the Search, *Visualization*, *Exploration*, and *Authoring* Widgets. Templates for the concepts of the ontology are defined for the visualization of and interaction with the resources. Built-in mechanisms make it possible to access, extract and display information from the underlying RDF graph by means of SPARQL queries. In addition, the portal uses the visualization of the commercial software 3DM as well as the YASARA client**.**

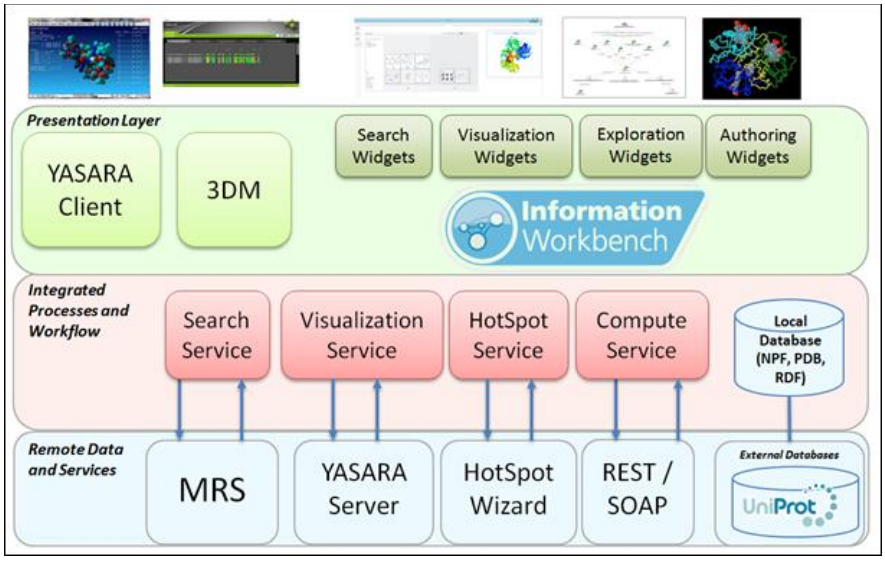


**Fig. 2.** Architecture of the NewProt portal

In the following, we focus on the NewProt format for achieving interoperability between all software components.
**NPF - The NewProt Format for Achieving Interoperability** The NewProt Format (NPF) is an XML-based file format for data exchange between the different software tools. The format allows residues to be represented as a mapping from Residue to Residue-Value. The NewProt Format is relevant for the following software components: HotSpot Wizard [2], YASARA [3], 3DM [4], HOPE [5], WHAT IF Web services [6]. In addition the NewProt format is used by external computational services to communicate the results.

The NewProt Format is defined by the XML Schema Definition[1] and allows specifying metadata (such as e.g., the source, the title, or the creation date) using the Dublin Core vocabulary. In the body of the XML document a list of residues can be represented, each of which is associated with a set of values. The supported values and their format are defined in the schema, and include

- mut: Target value expressing the degree of mutability
- flags: Flags for residue, i.e. Catalytic, Pocket, Tunnel
- alignments: Definition of alignments of this residue

[1] http://newprot.fluidops.net/npf

In addition the NewProt format supports the definition of custom values and metadata for computational services to provide additional information. In the head-declaration of the NewProt format it is possible to specify the relevant metadata about the custom value, including a range specification for valid values. The metadata declaration for a value called “prolineMutation”, for example, is:

```
<customValueDeclarations>
        <customValueDeclaration name="prolineMutation" rangeMin="-5" rangeMax="5" />
</customValueDeclarations>
```

This custom value definition can then be used by services to provide data as part of the body:

```
<residueValue isHotspot="true">
        <residue>
                <number>7</number>
                <chain>A</chain>
                <type>ILE</type>
                <pdb_number>7</pdb_number>
                <insertion_code/>
                <model_number>0</model_number>
        </residue>
        <value>
                <customValue name="prolineMutation">1.557</customValue>
        </value>
</residueValue>
```

Computational results provided using the NewProt format are attached to the respective proteins in the NewProt portal’s RDF database. More particularly, the provided information is made accessible to the Information Workbench and its visualization facilities through an RDF lifting.

## 4 Conclusions

We have presented the NewProt Portal as a platform to provide integrated access to a broad range of protein engineering tools and services. We have described typical use cases in working with the platform, the underlying architecture based on the Information Workbench as integration platform for the software components as well as the role of the NewProt format for achieving extensible interoperability across services and components on the data level.

At the current stage, the NewProt portal is a fully functional platform accessible to members of the NewProt consortium. Future work will focus on integrating additional services, opening the portal towards a larger user base and developing a sustainable model for the operations of the portal for both academic research and commercial uses.

**Acknowledgments.** The research presented in this paper was financed by the Seventh Framework Program (FP7) of the European Commission under Grant Agreement 318338, the NewProt project.

## References

1. Haase, P., Schmidt, M., Schwarte, A.:The Information Workbench as a Self-service Platform for Linked Data Applications. In: 2. Intl. Workshop on Consuming Linked Data (COLD), Bonn (Deutschland)
2. Pavelka A., Chovancova E., Damborsky J..: HotSpot Wizard: a web server for identification of hot spots in protein engineering., Nucleic Acids Res. 2009
3. Krieger, E., Vriend, G. (2014) YASARA View - molecular graphics for all devices - from smartphones to workstations. Bioinformatics in press, doi:10.1093/bioinformatics/btu426.
4. Kuipers RK, Joosten HJ, van Berkel WJ, Leferink NG, Rooijen E, Ittmann E, van Zimmeren F, Jochens H, Bornscheuer U, Vriend G, dos Santos VA, Schaap PJ.: 3DM: systematic analysis of heterogeneous superfamily data to discover protein functionalities. Proteins. 2010 Jul;78(9):2101-13
5. Venselaar, H., te Beek, T., Kuipers, R., Hekkelman, M, Vriend, G.: Protein structure analysis of mutations causing inheritable diseases. An e-Science approach with life scientist friendly interfaces. BMC Bioinformatics. 2010 Nov 8;11(1):548
6. Vriend, G,: WHAT IF: a molecular modeling and drug design program, Journal of molecular graphics 8 (1), 52-56, 2009.