Aquaria: Integrating Sequence and Structure

Seán I. O'Donoghue*† ‡
sean@odonoghuelab.org

Kenneth Sabir† ‡
k.sabir@garvan.org.au

Maria Kalemanov§
kalemanovm@rostlab.org

Christian Stolte*
christian.stolte@csiro.au

Nelson Perdigão¶
npereira74@gmail.com

Fabian A. Buske† ‖
f.buske@garvan.org.au

Julian Heinrich*
julian.heinrich@csiro.au

Burkhard Rost‡
rost@rostlab.org

Andrea Schafferhans‡
andrea.schafferhans@rostlab.org

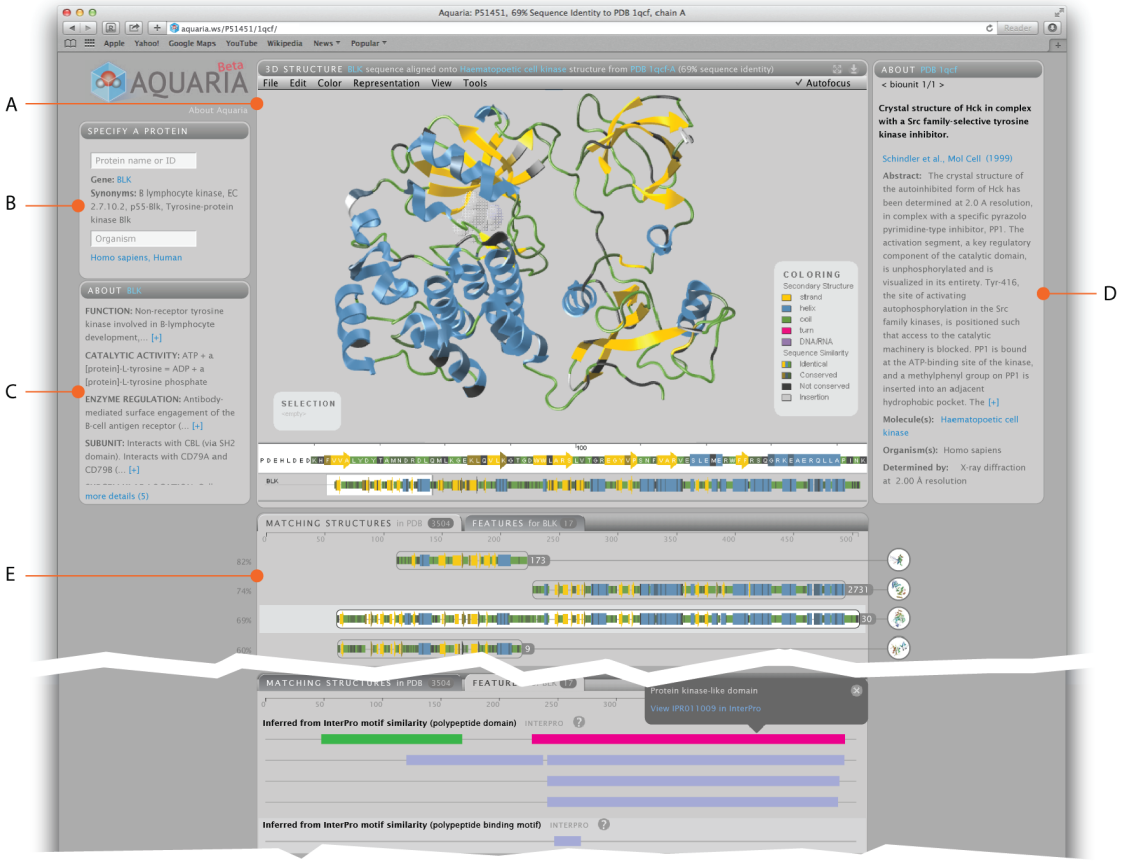

Figure 1: The Aquaria user interface comprises five panels: a 3D view (A) shows the currently selected 3D structure with various rendering modes using the same colour scheme as applied for all structures and aggregates (E) that match a given user query entered through the search field (B). A white background is used to visually connect the sequence being rendered in the 3D view and its cluster in the matching sequences panel (E). Panels on either side give information about the Uniprot entry corresponding to the query (B) as well as details of the structure (D) being shown in the 3D view.

*CSIRO Computational Informatics
†Garvan Institute of Medical Research
‡The University of Sydney
§Technische Universitaet Muenchen
¶University of Lisbon
‖St. Vincents Clinical School

Abstract

To understand the molecular mechanisms that give rise to a protein's function, biologists often need to (i) find and access all related atomic-resolution 3D structures, and (ii) map onto these structures sequence-based features (e.g., domains, SNPs, or posttranslational modifications). We recently developed Aquaria, a resource offering unprecedented access to protein structure information. Aquaria allows biologists to query by gene name or protein synonym and explore (often large numbers of) sequence-similarity based matches of solved structures, ranked and grouped according to sequence similarity and alignment position. Annotated features are tightly

integrated into one interactive user interface and can be viewed in their spatial context. In this poster, we give an overview of the system and its visualisation features.

Index Terms: H.3.5 [Online Information Services]: Web-based services—

1 Introduction

The number of protein sequences collected in public databases such as UniProt [6] has been growing exponentially over the last decade, and is expected to grow even faster with the advance of new sequencing technologies such as next generation sequencing (NGS). To give an example, UniProt lists more than fifty million protein sequence entries at the time of writing this manuscript. In fact, the total number of known protein sequences is substantially larger, since individual UniProt entries typically document multiple sequence variants deriving either from single-nucleotide polymorphisms (SNPs) or from splicing.

To fully understand the biological functions of a protein, however, life scientists need to know much more than just its amino acid sequence — one very rich source of additional knowledge is the three-dimensional (3D) structures adopted by a protein across a range of physiologically relevant conditions. Where available, such structures can give detailed insight into the molecular mechanisms underlying a protein's function. Unfortunately, the experimental determination of protein structures lags significantly behind sequencing; currently, the protein data bank [2] (PDB) holds less than $100,000$ structures, thus comprising less than 1% of the available sequences in UniProt.

When using structural models derived from *homology modelling* or *ab initio* methods, scientists need to be very aware that the models contains regions with variable levels of uncertainty, and will often contain considerable inaccuracies, especially *ab initio* models. The quality of a model depends on details of the often very complex method used to derive it — communicating this information clearly to end-users can be difficult, with the very real danger that incorrect conclusions may be drawn by inexpert users.

The avoidance of such misinterpretations was the motivation behind the development of a related approach taken by the SRS 3D system [3], which was recently superseded by the Aquaria resource developed in our laboratory. Instead of calculating model structures, the Aquaria approach simply matches all known protein sequences onto all known protein structures [5], and displays experimentally-determined structures overlaid with abstract data to indicate the quality of the sequence match [3]. These data are provided in a web-based resource that facilitates visual exploration, allowing users to rapidly query and visualise *all* available experimentally determined molecular structures in the PDB that match a given protein sequence. Although the models presented in Aquaria are not as refined as those derived from modelling [1], the uncertainties and inaccuracies can be much more easy understood by molecular biologists and biochemists who are not experts in structures or homology modelling — this is aligned with our goal in creating Aquaria, which was to make structural information more accessible to a much wider audience.

2 Aquaria

Aquaria is available online at `www.aquaria.ws`. Providing the software through a browser interface was an important design decision made in order to make protein structures accessible to a wide range of researchers. The user interface (Figure 1) comprises five panels outlined in the following sections.

2.1 Information Panels

Panels (B), (C), and (D) are used for querying and the display of information obtained from the PDB and UniProt.

The starting point for a query is Panel (B), which accepts a wide range of protein identifiers such as protein names, gene names, or UniProt primary accessions. Upon entering a search query, the system suggests possible matches to facilitate the selection of available proteins. All queries are restricted to the organism specified in the respective input field within the same Panel.

Panels (C) and (D) summarise information about the query protein gathered from UniProt and the PDB. These are very important as they provide many properties of a protein and the current structure from resources that many users are familiar with.

2.2 Sequence View

The sequence view (Panel (E)) provides a visual interface to structures that align with the query protein. All matching sequences are organised in clusters determined by the position and length of the alignment with the query sequence. The best matching structure of each cluster is used as a visual representative of the cluster, color-coded by its secondary structure. The number attached to the right of each cluster indicates its size. The vertical position of a cluster reflects the degree of sequence identity (with the query sequence, decreasing from top to bottom), which is further shown as a percentage on the left side of the panel. All clusters can be expanded by clicking on the cluster size label, revealing the cluster members in a navigable tree with two levels for (i) conformation and binding partner as well as (ii) individual structures.

2.3 3D Viewer

For each query protein, the best matching structure (in terms of length of the match and sequence identity) is rendered in the 3D Viewer panel (A) using cartoon rendering, coloured by secondary structure. The viewer is based on SRS3D [3] and allows for a variety of configuration options, including representation of substructures (such as individual chains or amino acids) as wireframe, surface, space filling, or ball and stick. Support for gestural input via LeapMotion or Microsoft Kinect devices has been added using the Molecular Control Toolkit [4].

2.4 Features

Annotated features (such as SNPs or post-translational modifications (PTMs)) are retrieved from Interpro, UniProt, and other databases. In Aquaria, they are accessible in the `FEATURES` tab in Panel (E) and can be loaded into the 3D Viewer by clicking any of the feature tracks. This will map the selected feature onto the current structure, replacing the viewer's colouring scheme for all affected amino acids.

References

[1] K. Arnold, F. Kiefer, J. Kopp, J. N. D. Battey, M. Podvinec, J. D. Westbrook, H. M. Berman, L. Bordoli, and T. Schwede. The protein model portal. *Journal of Structural and Functional Genomics*, 10(1):1–8, 2009.

[2] H. M. Berman, J. Westbrook, Z. Feng, G. Gilliland, T. N. Bhat, H. Weissig, I. N. Shindyalov, and P. E. Bourne. The protein data bank. *Nucleic Acids Research*, 28(1):235–242, 2000. PMID: 10592235.

[3] S. I. O'Donoghue, J. E. W. Meyer, A. Schafferhans, and K. Fries. The SRS 3D module: integrating structures, sequences and features. *Bioinformatics*, 20(15):2476–2478, 2004. PMID: 15087318.

[4] K. Sabir, C. Stolte, B. Tabor, and S. O'Donoghue. The molecular control toolkit: Controlling 3D molecular graphics via gesture and voice. In *2013 IEEE Symposium on Biological Data Visualization (BioVis)*, pages 49–56, 2013.

[5] A. Schafferhans, J. E. W. Meyer, and S. I. O'Donoghue. The PSSH database of alignments between protein sequences and tertiary structures. *Nucleic Acids Research*, 31(1):494–498, 2003. PMID: 12520061 PMCID: PMC165557.

[6] The UniProt Consortium. Activities at the universal protein resource (UniProt). *Nucleic Acids Research*, 42(D1):D191–D198, 2014.