Alexandre Maes[1,a] / Xavier Martinez[2,a] / Karen Druart[2] / Benoist Laurent[3] / Sean Guégan[2] / Christophe H. Marchand[1] / Stéphane D. Lemaire[1] / Marc Baaden[4]

MinOmics, an Integrative and Immersive Tool for Multi-Omics Analysis

[1] Laboratoire de Biologie Moléculaire et Cellulaire des Eucaryotes, Institut de Biologie Physico-Chimique, UMR8226, CNRS, Sorbonne Université, 13 rue Pierre et Marie Curie, 75005, Paris, France, E-mail: stephane.lemaire@ibpc.fr
[2] Laboratoire de Biochimie Théorique, CNRS, UPR9080, Institut de Biologie Physico-Chimique, Univ Paris Diderot, Sorbonne Paris Cité, PSL Research University, 13 rue Pierre et Marie Curie, 75005, Paris, France
[3] Institut de Biologie Physico-Chimique, FRC 550, CNRS, Paris, France
[4] Laboratoire de Biochimie Théorique, CNRS, UPR9080, Institut de Biologie Physico-Chimique, Univ Paris Diderot, Sorbonne Paris Cité, PSL Research University, 13 rue Pierre et Marie Curie, 75005, Paris, France, E-mail: baaden@smplinux.de

Abstract:
Proteomic and transcriptomic technologies resulted in massive biological datasets, their interpretation requiring sophisticated computational strategies. Efficient and intuitive real-time analysis remains challenging. We use proteomic data on 1417 proteins of the green microalga *Chlamydomonas reinhardtii* to investigate physicochemical parameters governing selectivity of three cysteine-based redox post translational modifications (PTM): glutathionylation (SSG), nitrosylation (SNO) and disulphide bonds (SS) reduced by thioredoxins. We aim to understand underlying molecular mechanisms and structural determinants through integration of redox proteome data from gene- to structural level. Our interactive visual analytics approach on an 8.3 m^2 display wall of 25 MPixel resolution features stereoscopic three dimensions (3D) representation performed by UnityMol WebGL. Virtual reality headsets complement the range of usage configurations for fully immersive tasks. Our experiments confirm that fast access to a rich cross-linked database is necessary for immersive analysis of structural data. We emphasize the possibility to display complex data structures and relationships in 3D, intrinsic to molecular structure visualization, but less common for omics-network analysis. Our setup is powered by MinOmics, an integrated analysis pipeline and visualization framework dedicated to multi-omics analysis. MinOmics integrates data from various sources into a materialized physical repository. We evaluate its performance, a design criterion for the framework.

Keywords: Database, Display wall, Omics, Protein and proteome, Virtual Reality (VR)
DOI: 10.1515/jib-2018-0006
Received: January 28, 2018; **Revised:** May 3, 2018; **Accepted:** May 9, 2018

1 Introduction

The analysis of massive biological datasets available on public repositories is nowadays one of the key challenges biologists face due to the increasing size, variety and complexity of such data. Modern high-throughput and large-scale analytical technologies create an explosion of experimental datasets of quantitative and qualitative biological information, often simply called "omics" [1]. Besides, well-annotated and well-curated biological databases store and provide generic information on genomes (Genbank, EMBL, UCSC…) [2], [3], [4] and proteomes (Swiss-Prot, TrEMBL, PIR…) [5], [6] as well as extended information on protein 3D structure (PDB) [7], ontologies (GO) [8], classification (PANTHER, Pfam, SCOP) [9], [10], [11], cellular localization [12] and many more. Nowadays dedicated bioinformatics infrastructures are essential to supervise the sheer amount and diversity of datasets. One of many related challenges is to provide integrated and meaningful information for the biologist, accessible even without programming expertise. Visual analytics is a key strategy for this task [13]. Immersion and interactivity to delve into large biological datasets require the combination of ultra-fast access to databases, complex on-the-fly queries, analyses and efficient visualization tools.

Sophisticated open-source object-relational database management systems (ORDBMS) such as for instance PostgreSQL, enable a very fast and flexible control of relational databases. Relational databases provide a well-suited environment for omics datasets [14], [15], [16], [17] as they store various data types and provide a structured query language that confers filtering, joining, grouping and sorting abilities on subsets of very large

Stéphane D. Lemaire, Marc Baaden are the corresponding authors.
[a] **Alexandre Maes and Xavier Martinez:** These authors contributed equally.

datasets. The concept of subsets, or collections, in relational databases is highly relevant for biology, considering living organisms and omics datasets as collections of biological molecules with qualitative and quantitative parameters. To ensure performing tasks efficiently, ORDBMS require both an adapted indexing strategy and an optimized query-building system. Dedicated algorithms and software components are developed to assist structured language queries (Django ORM [18], Panda-SQL http://pandasql.free.fr/, SQL Search, ...) but the amount and heterogeneity of omics-datasets and associated biological tasks call for a dedicated query-building system. The efficiency of ORDBMS in handling biological datasets ensures real-time processing of numerous biological tasks, thereby providing the foundations to build an immersive access plunging the user right in the middle of the data through visualization and interactivity.

The rise of efficient and interactive data visualization libraries, for instance Javascript-based ones such as D3.js (https://d3js.org/) or Three.js (https://threejs.org/), provides access to powerful display technologies when combined with efficient low-level rendering approaches such as WebGL [19], the equivalent of the OpenGL standard on the web (https://www.opengl.org/). High-level tools such as for instance the Unity3D game engine (https://unity3d.com/) recently gave access to such technologies with convenient abstraction for scientific purposes as we have previously argued [20]. Modern game engines such as Unity3D or Unreal Engine (https://www.unrealengine.com/) provide an alternative to traditional development of visualization tools built from scratch that are usually platform dependent. The game engines offer a way to create multi-device and multi-platform projects based on a unique code base and give access to advanced rendering effects with a relatively small development effort and a short trial-and-error development cycle. The large community, the support and the wide adoption of those tools promote their use as a framework to build on. Furthermore, large developer communities supporting such engines usually ensure the support of new technologies and new devices, especially in the Virtual Reality (VR) and Augmented Reality (AR) fields. Recently, improvements on the WebGL library allowed developers to create new molecular viewers based on this modern web technology (e.g. NGL viewer by Rose and Hildebrand [21]). By providing direct access to graphics processing unit (GPU) capabilities in an efficient bundled library, this technology opens the way to web-based visualization tools, which can be used without installation, on mobile and desktop devices. Moreover, the Unity3D game engine provides a set of compilation tools based on Emscripten [22] to generate a WebGL and Javascript rendering system.

Several studies have investigated protein network visualization, focusing on improving the representation of these complex networks [23], [24] but also including the complexity of the temporal dimension involved in the link between the genotype and the phenotype [25]. In parallel, recent frameworks to cluster heterogeneous and voluminous biological data combined with text data mining approaches have been developed [26], especially to better understand the localization and the interactions of proteins in cells. Other solutions propose a fully web-based approach to combine molecular visualization with enriched analysis, linked to external databases [27]. Clustering biological information into a self-updating database was also experimented [28] to provide an up-to-date information based on a Java framework. In parallel, stereoscopic devices have been used to experiment how they can improve interaction and decision-making in an immersive visual analytic context [29].

In this manuscript we present an integrated pipeline, called MinOmics, a powerful tool to manage biological data from various sources, from storage to analysis and visualization. It combines PostgreSQL 9.6 ORDBMS as a data storage system, D3.js and UnityMol as visualization frameworks for two-dimensional (2D) and three dimensions (3D) data, respectively, and Django, a powerful high-level Python web framework that enables data analyses by linking all components together. The heterogeneity by nature of biological macromolecules (genes, transcripts, proteins) and their relations (transcription and translation) imposes the use of PostgreSQL ORDBMS that offers the storage and indexing of numerous data types such as strings, numerics, arrays, json and time. MinOmics is developed in Python, an object-oriented language, which possesses module adapters to communicate with ORDBMS. Furthermore, Django allows the deployment of MinOmics as a portable and multi-platform web service. Javascript, in particular with the D3.js library, brings interactivity, animations and information visualization capabilities, while UnityMol allows displaying 3D molecular structures on a web page with a WebGL and WebVR implementation. MinOmics produces an interactive visualization environment as a basis for data mining with the possibility of 3D stereoscopic immersion when appropriate. Both query builder and Unity3D modules allow the development of MinOmics either as a full web-service or as a web-independent cross-platform GUI. This visual analytics approach allows scientists to interactively manipulate large datasets on an 8.3 m^2 25 MPixel high-resolution display wall with stereoscopic 3D representations for a semi-immersive experience. Furthermore, it is possible to explore data inside a virtual environment in full-immersive VR, to better understand complex relationships, for instance related to protein networks or molecular structures.

As a study case, we will investigate through MinOmics the chemical and physical parameters governing the selectivity of three cysteine-based redox post translational modifications (PTM). We focus on the unicellular green alga *Chlamydomonas reinhardtii*, a major model for the study of fundamental biological processes and for exploitation as an industrial biotechnology host [30]. Although protein functions are encoded in genes,

the actual regulation of protein structure and function is generally executed by specific PTMs that enable a gigantic heterogeneity and diversity of gene products [31]. Emerging data indicate that redox networks coordinate large numbers of redox elements involved in a multitude of pathways and cellular processes to allow resistance and adaptation to environmental challenges [32]. These networks involve multiple redox PTMs that have emerged as important mechanisms of signaling and regulation in all organisms. Indeed, the thiol moiety of cysteine residues can evolve toward reversible redox states. The best studied ones are disulphide bond formation (SS, the formation of a disulphide bond between two protein cysteine residues), glutathionylation (SSG, the formation of a mixed disulphide with the major cellular antioxidant glutathione) and nitrosylation (SNO, the formation of a nitrosothiol by reaction of cysteine with nitric oxide, a major cellular messenger) [33]. These cysteine-based redox PTMs constitute molecular switches regulating protein functions. This cysteine proteome can be considered as an interface between the functional genome and the external environment [34]. It is a highly dynamic network of protein thiols with flexible reactivities [35], [36], [37]. Therefore, combinations of multiple redox PTMs act in concert throughout the cell and act as a network rather than as insulated elements. Gaining insights into the functioning of redox networks will require unraveling the determinants of the specificity of the diverse redox PTMs for specific proteins and cysteines. A better understanding of this specificity could allow predicting targeted proteins and modeling the functioning of the redox network. Preliminary results do not show any consensus primary sequence motif and suggest that the specificity primarily depends on the biochemical properties of the cysteine residue. These properties are largely linked to the cysteine microenvironment within the folded protein, which can notably influence the accessibility, the acidity and the nucleophilicity of the residue [38], [39], [40].

We will describe and use MinOmics for our first real-time explorations in an attempt to unravel the chemical and physical parameters governing the selectivity of three redox PTMs in *C. reinhardtii*: SSG [40], SNO [41] and SS reduced by thioredoxins [42]. For this purpose, both the visualization of proteomic networks and of related molecular structures in conjunction with the experimental and biological data is required. Stereoscopic visualization adds a precious dimension to this visual exploration and analysis process.

2 Materials and Methods

In this section, we describe the different databases, datasets and software components used or integrated within the MinOmics framework. All the data used are publicly accessible and we provide cross-references to different databases. Firstly, we depict the structural data that served as the basis for illustrating the framework, then we explain the choice of software and technical details in implementation and parameters used to refine, model and analyse proteomic data. We also provide a hardware reference for benchmarking MinOmics and a description of the specific devices such as the wall-sized display and VR headsets.

2.1 Proteomic Datasets

All proteins and redox PTM positions are extracted from published data: SSG [40], SNO ([43], ProteomeXchange accession: PXD000569), reduction by thioredoxins ([42]; ProteomeXchange accession: PXD006097 and PXD006116). In brief, these datasets comprise 41, 501 and 1188 protein descriptions respectively referenced in the Uniprot database (CHLRE accession). Among them, cysteine site's modification of 41, 302 and 602 proteins have been identified, respectively. Currently MinOmics stores 38 experiments (accessions and parameters such as *p*-values, fold changes, external file paths, peptide sequences, quantitative values...) performed on *C. reinhardtii*, *Saccharomyces cerevisiae* or *Arabidopsis thaliana.* In total it describes 12 organisms either from Uniprot, Refseq or phytozome, all the Gene Ontology identifiers and descriptions, all the Pfam annotations and available structural data from the PDB and primary sequences. The overall amount effectively stored in the MinOmics database with all cross-referenced descriptional data currently amounts to 1.7 GB, which excludes the large raw datasets from the experiments such as mass spectrometry data.

2.2 Protein Structural Modeling

The structures of all 1417 proteins for which mass spectroscopy data in the proteomic datasets indicates a cysteine modification were built through homology modeling with the @tome2 web server [44]. Based on their FASTA sequences the @tome2 protocol selects the supposedly best PDB template according to different alignment methods, in particular HHsearch [45], Fugue [46], psi-Blast [47], and Sp3 [48]. An initial homology model is then built with TITO [49] and scrwl [50] software components. Models are further refined with Modeler

[51] and their quality evaluated using the Qmean criteria [52]. In the whole set of models generated through this strategy, a subset of 731 protein models has been characterized to the point that we know which cysteine residue is modified.

This subset is thus particularly precious to attempt to infer rules about cysteine modifications. To identify an initial restricted set of models of high quality, we further restrained our 3D database by selecting models according to three main criteria. (1) cysteines of interest are modeled, (2) the percentage of identity between the sequence of interest and the PDB template is above 30 %, and (3) the protein structures are not aberrant (for instance not missing backbone atoms). At the end, we therefore generated a focused subset of 409 structural models that we considered of good enough quality, harboring 745 cysteines undergoing one or several PTMs, to derive initial hypotheses from.

2.3 Analysis of Molecular Model Properties

From our pool of structural models, we focus on cysteine sites to pre-calculate descriptors which can influence the reactivity. First, the pKa of each cysteine is calculated using PROPKA [53], keeping in mind the difficulty to estimate the pKa of thiols. For some cysteines, PROPKA yields a pKa value equal to 99.99 when it infers that a SS exists between two cysteines. For these particular cases, we do not have a pKa value. Then, we calculated the cysteine accessibility to the solvent thanks to naccess [54], dissecting the results into the accessibility of the whole cysteine residue and the accessibility of the thiol group alone. These accessibility descriptors allow us to estimate how much a given cysteine is buried in the protein 3D fold, especially if the cysteine is on the surface with the side chain directed toward the inside of the protein. In this case, it is reasonable to assume that a slight local movement may allow the thiol group to be exposed to the solvent and become reactive. Another parameter that we calculated is the secondary structure type of the protein backbone stretch comprising the cysteine of interest. For this purpose we used DSSP [55] to classify in α-helix, β-strand and random coil, taking into account the polarity of the secondary structure (N-ter or C-ter). Importantly, the polarity of α helices is known to have an effect on cysteine reactivity [56]. Finally, we use PyMol, a molecular viewer and analysis tool, to list residues located within 7 Å of a cysteine residue of interest. We then calculated the root mean square deviation (RMSD) between all models fitted on a structural alignment of the backbone Cα atoms. The whole set of descriptors is then injected into MinOmics for further data analysis.

Furthermore, the MinOmics framework itself can perform additional analyses of this structural data upon demand, for instance using R [57] scripts. After selecting a subset of cysteine residues based on different parameters (e.g. all nitrosylated cysteines), it is possible to plot the distribution of another parameter (e.g. the pKa of the selected subset). Other analyses are possible such as clustering of 2D data. As an example, on the already mentioned subset selection, MinOmics can generate a 2D matrix representation of pre-calculated RMSDs between all models. This similarity matrix can be used for further analysis and inference of sub-populations by applying clustering algorithms. The clustering is done by default using a Euclidean distance of the matrix and the hclust function with the "Ward.D2" method, generating a dendrogram. At this stage, the user can already choose a number of groups to cut the tree, or use different criteria such as Davies-Bouldin or Dunn ones to define the optimal number of clusters. A Silhouette criterion is functionally implemented but not yet accessible in MinOmics. Once the number of groups is defined, MinOmics generates a 2D picture showing each element colored according to its group, along with information for each group (e.g. the average, the name of the element at the centroid, the number of elements in the group...).

We can imagine extending clustering parameters to provide more flexibility to the user for exploration within the MinOmics framework, even if the less elegant solution to extract the raw matrix data to use another program for clustering calculations is already available.

2.4 Public Repositories

Data from seven public repositories was fed into the MinOmics store in order to enrich the information available about the studied proteins. In particular:

Uniprot: http://www.uniprot.org/uniprot/?query=organism:%s&format=xml,
Refseq: ftp://ftp.ncbi.nlm.nih.gov/genomes/,
Phytozome: https://phytozome.jgi.doe.gov/pz/portal.html,
Pfam: ftp://ftp.ebi.ac.uk/pub/databases/Pfam/current_release,
PDB: https://www.rcsb.org/,
Gene Ontology: http://purl.obolibrary.org/obo/go/go-basic.obo,
MapMan: https://mapman.gabipd.org/mapmanstore

2.5 MinOmics Software Framework

The software components and versions used to implement the core of the MinOmics system comprise PostgreSQL 9.6, Django 1.10/ backends.postgresql_psycopg2, Python 2.7/psycopg2, Javascript/D3.js v.4/jquery3.2.1 and Unity3D 2017.3 with UnityMol 0.9.8-webgl-beta.

2.6 Hardware Characteristics and Configurations

2.6.1 Workstations

The core MinOmics database and web applications are performed on an Intel(R) Xeon(R) CPU E5-2623 v3 @ 3.00 GHz Debian 4.9.65-3 (2017-12-03) x86_64 GNU/Linux server system with 16 GB RAM. For data storage we currently use standard 7200 rpm hard disks, which were also the basis for the provided benchmark data. On the client side, we used a dual Intel(R) Xeon(R) CPU E5-2630 v4 @ 3.10 GHz Windows 7 64 Pro system to power the display wall. It features three Quadro M4000 graphics cards. An Intel(R) Core i7-6700 @ 3.4 GHz Windows 10 64 Pro system with 32 GB RAM and a GeForce 1080 GTX graphics card was used to drive a head-mounted VR display, either an Oculus Rift or an HTC Vive.

2.6.2 Display Devices

We have access to two main classes of display devices, a semi-immersive display wall with a large surface, high resolution and stereoscopy, and fully immersive VR head-mounted displays (HMDs), which characteristics are detailed in Table 1.

Table 1: Characteristics of the stereoscopic display hardware setups.

	Display wall	**HTC vive HMD**	**Oculus rift HMD**
Display	DLP	OLED	OLED
Resolution	7680 × 3240 (Mono/Stereo)	1080 × 1200 (Stereo)	1080 × 1200 (Stereo)
FOV[a]	170°–115°	110°	110°
3D stereo (120 Hz)	o		
VR (90 Hz)		o	o
WebVR		o	o
Number of users	1–15	1	1
Head tracking		o	o
2D interaction[b]	o		
2.5D interaction[c]	o		
3D interaction (VR controller/LeapMotion)		o	o

[a]The display wall FOV for 3D stereo is imposed by the characteristics of the 3D glasses, which is 170° horizontal and 115° vertical. [b]2D interaction such as for instance through computer mice or equivalent devices. [c]2.5D interaction through devices such as a gyroscopic 3D mouse is an enhanced 2D interaction. The mouse can be used in full 3D space, which is translated into a final 2D signal for the operating system.

The display wall at the IBPC institute consists of 12 tiled backprojected EC-50-LHD CUBE-LED-SLIM series displays by EyeVis arranged in a four column by three row matrix setup. The dimensions are 4428 mm wide by 1866 mm high with a 0.58 mm pixel size yielding a resolution of 7680 by 3240 pixels. Interscreen bezel is less than 2 mm and the full wall is capable of active stereoscopy at 120 Hz refresh rate, in a window mixed with surrounding 3D content or full screen. The wall is addressed as a single screen through Windows, Mac OSX or Linux workstations. In the present case we used the Windows client to drive our MinOmics experiments.

The HMDs connected to our MinOmics setup are either a HTC Vive or an Oculus Rift. In their current version, both HMDs are using a 1080 × 1200 screen for each eye at 90 Hz refresh rate with a 110° field of view. The headset includes a gyroscope and an accelerometer. Two base stations emitting infrared light track the headset for the HTC Vive and a single base station emitting pulsed infrared light is needed with the Oculus. Controller positions are tracked by the same hardware for Vive and by an additional base station for the Oculus.

2.7 MinOmics User Studies and Feedback

At the present stage of development and implementation, we have not yet carried out formal user studies of the MinOmics system. Members of the current development and design team comprising eight individuals informally assessed individual MinOmics components, such as the relational database system and the visualization components, in an *ad hoc* manner. We plan to carry out validation studies on the full MinOmics system, based on selected tasks from the use cases described in the manuscript. Alternatively, a more theoretically oriented than empirical evaluation method can be used: Hierarchical Task Analysis may be performed to quantify the execution times of given tasks in a controlled manner [58].

The user feedback and experience we would particularly like to collect concerns the intuitiveness of the MinOmics user interface, potential fatigue after longer use, and the needs (as well as the opportunities) for efficient collaborative use.

3 Implementation

In the era of big data, numerous tools are now dedicated to gather and visualize large datasets, either through dedicated GUI designs or through web frameworks. Immersion and interactivity into biological datasets require a fast and efficient architecture that allows the communication between the user and the database (Figure 1). We first describe the implementation of the database (server) part of our framework, then focus on the visualization (client) implementation and the usage scenarios.

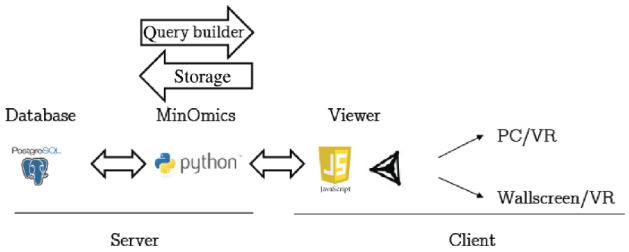

Figure 1: MinOmics web framework. Visual immersion into biological big data requires an efficient database management strategy. MinOmics assists the storage of massive datasets in PostgreSQL and building queries for an ultra fast exploration of biological data. Javascript and UnityMol-WebGL, a powerful molecular visualization framework, provide full immersion and real-time interactivity to the biologist.

3.1 Server-Side MinOmics Database

The MinOmics framework allows (1) the integration and updating of seven public biological repositories into the MinOmics database (2) the management of multiple omics datasets and (3) the building of efficient SQL queries. Our chosen data integration paradigm is a materialized one, creating an integrated physical repository of selected data extracted from the aforementioned collection of information sources. The database is designed in four layers corresponding to different levels of biological datasets numbered from level 0 to 3 (Figure 2). A dedicated module stores and indexes the datasets provided by public biological repositories (Uniprot, Refseq, Phytozome, PDB, Pfam, Gene ontology, MapMan) and the datasets provided by high-throughput technologies (proteomics, RNASeq). Currently, seven databases are integrated to MinOmics, but their number is easily extensible. The heterogeneity of omics datasets is contained in a *jsonb* data-type, a dictionary-like and indexable data-type managed by PostgreSQL. The structural models and their properties are similarly classified and accessed as level 1 datasets. Finally, both the presence of arrays and *jsonb* elements requires an efficient and adapted *btree* and *gin* indexing mechanism to maintain fast performances.

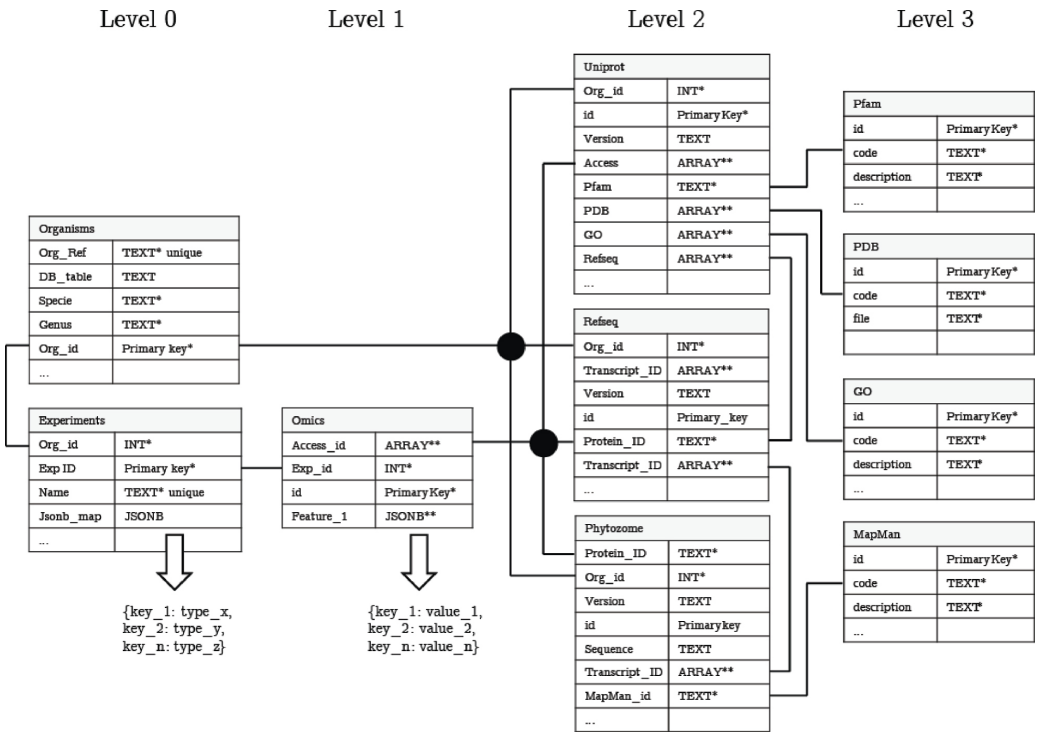

Figure 2: Design of the MinOmics ORBDMS. Collections of biological datasets are stored in the Organisms and Experiments tables (level 0). Table key relations unleash the efficiency of querying (black lines). The Omics table (level 1) stores the heterogeneous biological data obtained by high throughput technologies (RNASeq, Proteomic). Level 2 tables are the main public biological repositories stored on MinOmics (Uniprot, Refseq, Phytozome). Secondary databases (PDB, GO, Pfam, MapMan) are the level 3 tables. The nature and heterogeneity of biological datasets (data types) require an adapted indexing strategy (*btree indexing, **GIN indexing).

Querying MinOmics is performed in two steps. First, one gets and sets the parameters required by the query building module: the data types stored in the level 1 (*jsonb*) Feature_1 key (e.g. Structural modeling and post-processed parameters) and the identity and data types of the level 2 tables (db_table key) corresponding to the collections (Org_id or Exp_id) (Figure 2). This task is performed once, specific to the collection, and stored for further queries. The second step executes the biological query. The relations between biological macromolecules and the data available on public repositories permit and orient the establishment of the ORBDMS linkage between level 1 and level 2, between level 2 tables (genes, transcripts and proteins) and between level 2 and level 3 tables (secondary biological databases). The querying module combines the various data types, indices and links in order to produce an adapted and efficient SQL query (see an example in Supplemental Material S1).

The query efficiency through the different layers of MinOmics is reported in Figure 3. The SQL transaction time needed to identify all the proteins in the glutathionylome, nitrosylome or thioredoxome from the level 1 table ranges between 17 ms and 30 ms. This performance measured without indexing can be brought down between 0.9 ms and 1.7 ms with a btree indexing method, an efficient index for text data types (Figure 3A). Even if ultra-fast responses are achieved with an appropriate indexing method for relatively small datasets, the query efficiency decreases when the size of datasets increases. The largest soluble proteome of *C. reinhardtii* available nowadays (Multiconsensus, 2198 detected proteins, unpublished data) does not exceed 5 ms to collect, fast enough to allow interactivity with such 'large' biological datasets. Nevertheless, with one or several orders of magnitude larger datasets, such as for example, the human proteome (Human proteome, 161521 proteins in Uniprot), a lower reactivity can be observed, in this case roughly 100 times slower (Figure 3B).

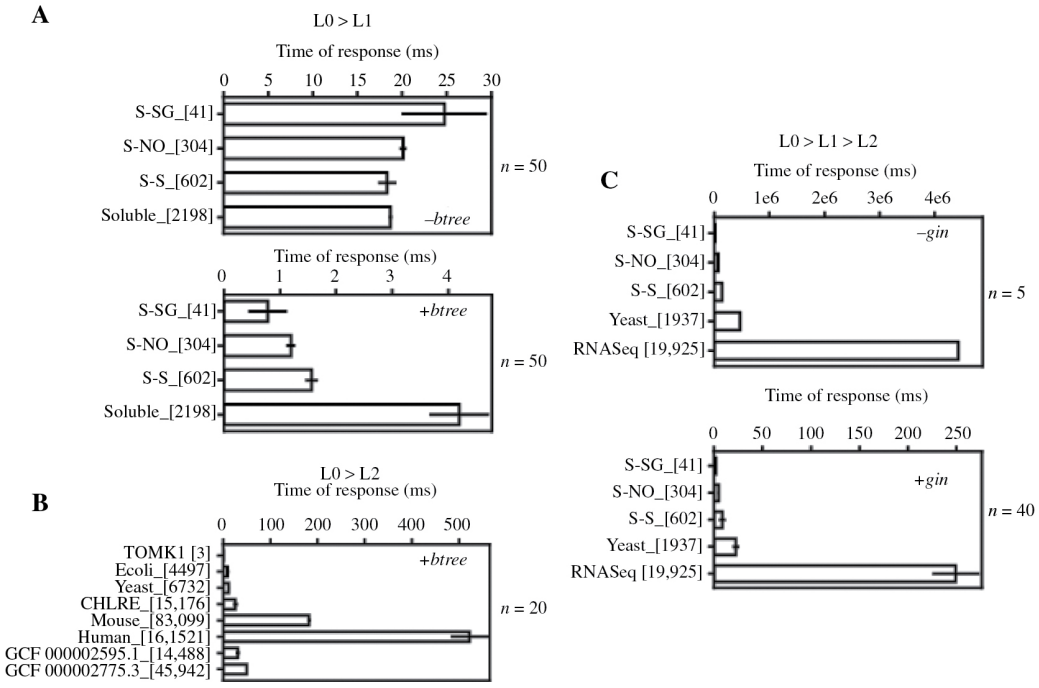

Figure 3: MinOmics querying performances. The speed and efficiency of navigation into the layers of the MinOmics database is measured as the time of transaction for fetching all elements either from experimental datasets through Exp_id (A and C), or organisms datasets through Org_id (B) keys. S-SG stands for glutathionylome, S-NO for nitrosylome, S-S for reductome and soluble for *C. reinhardtii* soluble proteome. An adapted indexing strategy allows faster queries (lower panels A and C). Standard error of N tests is indicated and the queryset size is given in between brackets.

The path to level 2 is of major extent to retrieve external annotations from locally stored public repositories (Figure 2). The intrinsic presence of multiple elements belonging to one biological element (protein or transcript) in proteomic (and to a smaller extent in RNASeq) data is due to the intrinsic principles of these technologies. Identified peptides (or reads) can belong to different proteins (or RNA sequences, respectively) without clearly discriminating which one is really present in the biological sample. Therefore, information about other proteins or transcripts that cannot be differentiated from the master hit has to be conserved. This condition forces the use of GIN indexing, which permits to decrease the time of response by 104, for almost all the tested queries (Figure 3C). This gain is essential to maintain interactivity, again predominantly for larger datasets, regarding RNASeq data queries performed on *C. reinhardtii*.

Finally, the filter and lateral join SQL operators allow to retrieve the embedded datasets in the level 3 containing the secondary biological repositories (Figure 4). Partial pattern matching on b-tree indexed text data in the level 2 (Figure 4A) or level 3 (Figure 4B and C) confers good query efficiency although an important latency for large datasets in the level 3 is observed (Figure 4B). Remarkably, embedded jsonb key:value data remains highly accessible to the selection with time of response of around 3 ms (Figure 4D). Molecular modeling and structural parameters of redox-modified sites are stored in such json data types. The raw server file path is also recorded in this field in the case of structural studies, so that the molecular structural models can be sent to the molecular viewer upon request.

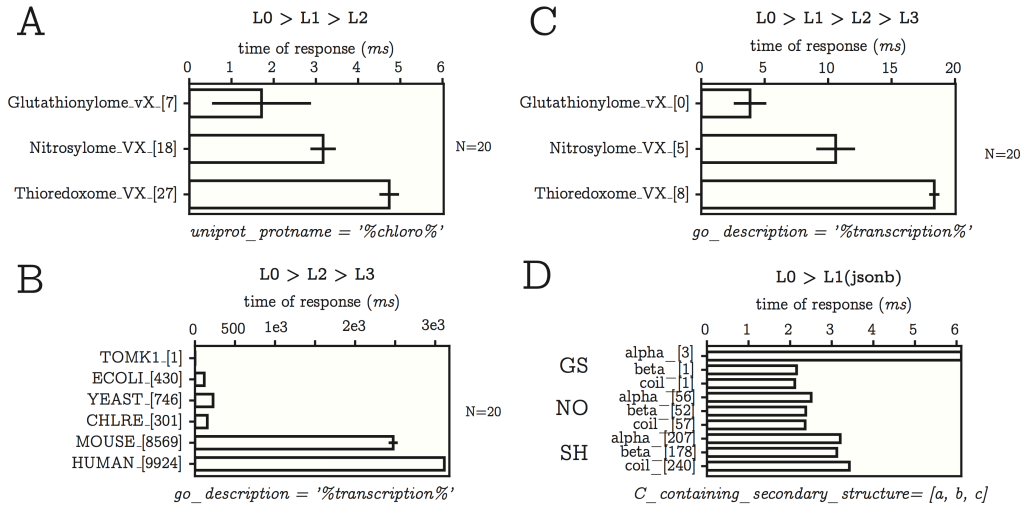

Figure 4: MinOmics filtering performances. The efficiency of filtering is measured as the time of response of fetching filtered data sets through the different layers of the MinOmics database. Timing for partial pattern matching on text data types is reported in (A, B, C). Complete pattern matching is performed on jsonb data-type as reported in (D). Patterns and queried keys are indicated below each panel. The standard error for N tests is indicated and the queryset size is given between brackets.

We use SQL clauses and operators for navigation and sub-selection of proteins by their numerous parameters, either from the public databases or by the omic experiments (Figure 5A). They can be graphically and interactively manipulated by an intuitive and dynamic Javascript interface where each dataset is represented as a colored node (Figure 5B). Filtering and grouping operations allow the user to refine the node content while merging and overlapping operations allow comparing different datasets. These tools will be used throughout different biological use cases as described below.

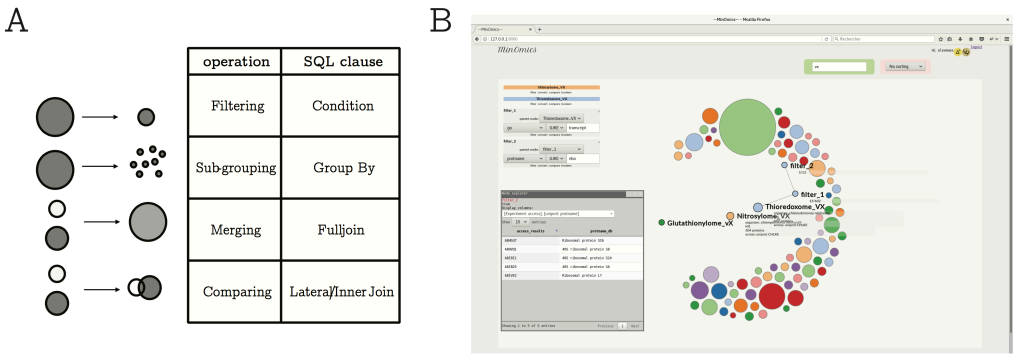

Figure 5: MinOmics query builder and web visualization. Structured query language offers relevant clauses to explore biological datasets as summarized in panel (A). Diving into subsets of big data is performed through interactive web tools that control the relations and constraints of the query builder depicted in panel (B).

3.2 Client-Side Data Viewer Implementation

Visualization of MinOmics data beyond the classical textual web interface is built upon two frameworks. The first one is JavaScript-based D3.js for interactive graphs up to two dimensions directly embedded in the web application. The second one comprises variants of our UnityMol software for 3D objects, such as complex proteomic networks and protein structures. The latter modules are capable of mono- and stereoscopic visualization and fully immersive data exploration in VR. For efficient integration with the server backend, we particularly leveraged a new WebGL version of UnityMol that will be briefly described. Then, we detail the visualization scenarios that we have envisioned and explored so far, as the overall hardware and software setup allows for a range of possible configurations that can be custom tailored to scientists' needs. An important aspect that has not yet been explored in much detail is the way and means to interact with the system. These questions will be addressed in future studies on this Human Computer Interaction sub-topic.

3.3 Integration of a WebGL UnityMol Instance with the Server Backend

Thanks to the Unity3D features mentioned in the introduction, the UnityMol code base [20], [59], a Unity3D based molecular visualization framework can now be used from within a web application such as MinOmics, retaining most functionalities from the standalone version. For example, the guided navigation feature [60] can be used to find the best point-of-view of an atom and determine the appropriate camera path to navigate around the molecule or a protein network. All standard molecular representations are available (CPK, liquorice, space-filling, HyperBalls, surface, cartoon). Proteomic networks are typically depicted through HyperBalls [61]. We implemented this WebGL version by stabilizing and improving a previous beta release of UnityMol for WebGL, as the web context imposes some constraints, for instance related to memory handling and to the ability to couple with other web components. The WebGL version of UnityMol designed for the MinOmics application communicates with the web page by sending messages to and from the game engine so that every UnityMol feature can be triggered from outside of the game engine by Javascript code. This communication is possible because the Unity game engine is embedded as a Javascript library. The module further offers access to immersive and stereoscopic WebVR visualization from within the web application. Recent advances in the WebVR library allow to experiment with immersive molecular visualization directly from the web browser. UnityMol is one of the first molecular viewers that can be used in a VR mode inside a web browser thanks to WebVR (Figure 6). Nowadays, the WebVR standard is still very young and we expect many evolutions in the near future. The WebVR integration in Unity3D is implemented by adding Javascript functions compiled during the Unity3D build process. These functions provide a communication layer between the Unity3D game engine renderer and WebGL. This experimental yet working implementation was achieved by using sample code provided by an independent developer (https://github.com/gtk2k/Unity-WebVR-Assets/).

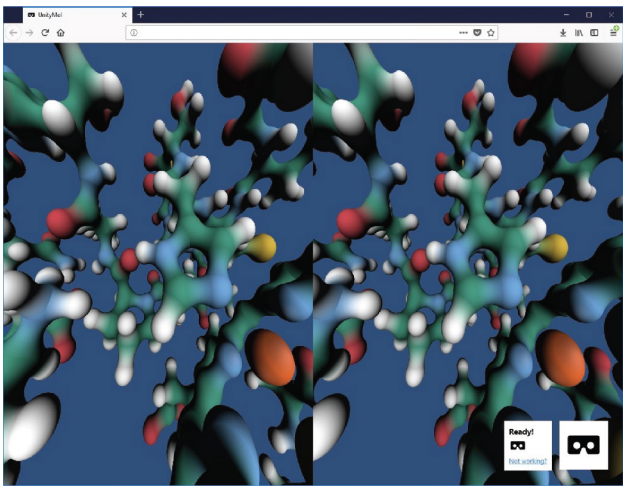

Figure 6: Screen capture of UnityMol WebGL+WebVR. The Firefox web browser (version 55 and up) hosts the WebVR version of UnityMol as integrated to MinOmics. Different devices can be addressed, here full-fledged HMDs, but possibly also Google cardboard-like simple setups could be used to share immersive MinOmics views outside of our local setup. Interaction devices yet need to be implemented and configured for a more efficient use of this functionality.

The most obvious way to integrate UnityMol in MinOmics was by embedding the WebGL version as just described. The restrictions of Unity3D implementation of message sharing between Javascript code and in-game C# script still required to develop a specific set of functions exposed to any Javascript code. This mechanism only allows sending one parameter per function message. Another issue to be addressed is the capture of input devices by the UnityMol WebGL component. Therefore UnityMol was embedded as an "iframe" in MinOmics, isolating it, thereby preventing mouse and keyboard input events from being captured. Through this mechanism, several instances of UnityMol WebGL can be created easily. This integration allows visualizing large datasets managed by MinOmics and displayed by UnityMol on a large high-resolution display wall, giving access to the easy filtering capability and high-quality rendering on the same screen. The analysis is then directly linked to the visualization to provide a visual analytics loop.

A remaining issue with the embedded WebGL version is access to stereoscopic rendering (other than through fully immersive WebVR). As a workaround, we explored an alternative way for MinOmics and UnityMol to communicate by using a standalone version of UnityMol outside a web context receiving data via a web socket or checking for new files to read in a folder. This approach does not suffer from memory limitations inherent to WebGL, and exposes the full power of OpenGL or DirectX functionalities and optimizations including stereoscopic rendering. Platform specific and optimized code can be deployed. The following paragraph explains the main visualization scenarios we explored so far.

4 Results

4.1 Visualization Scenarios

The intended scientific applications and first results form the core of this section and are detailed based on three use cases. These are closely linked to the available hardware setups and usage scenarios. Considering the MinOmics framework and the different devices available to explore and analyse our dataset, we identified different scenarios for different tasks in which multiple users can interact. We then present several relevant applications of these scenarios in a biological context to draw conclusions and extract new knowledge from the complex dataset.

4.1.1 Scenario 1: Full-Screen MinOmics with Embedded UnityMol WebGL in Monoscopy

The first scenario we developed was the integration of UnityMol WebGL to depict and manipulate 3D objects in a MinOmics web page displayed on the wall-sized display (Figure 7, panel 1). Everything is depicted in monoscopy. Only one UnityMol instance was used to display molecular data from the MinOmics database but the layout can be customized and several instances can be started to show or possibly compare side-by-side different types of information. In this particular setup, multiple users can benefit from the same view. A single user is interacting, for instance through a gyroscopic mouse, to perform both 2D and 3D tasks. Multiple users can watch and discuss with the main user to refine the analysis or change the camera point of view in UnityMol for instance. The interaction device can be passed on among the users.

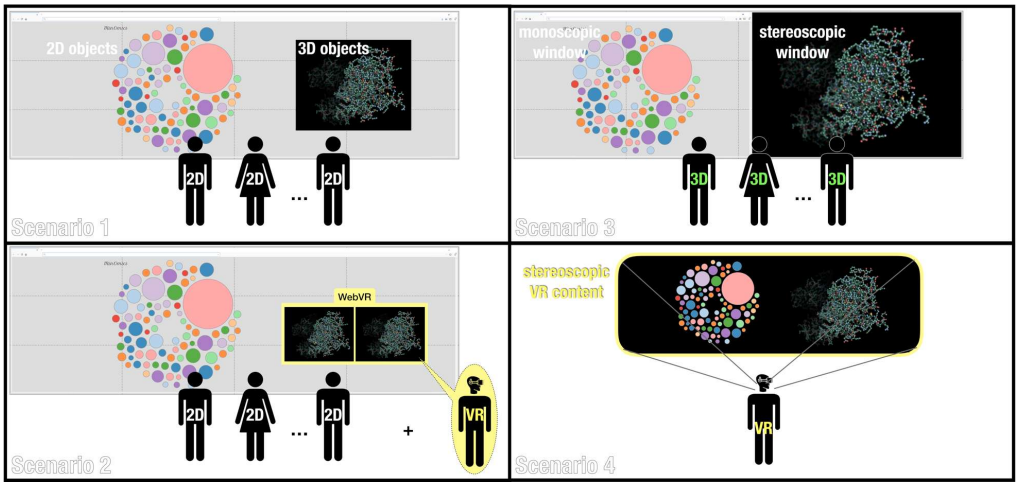

Figure 7: Schematic depiction of four possible visualization scenarios. The way UnityMol is integrated with MinOmics for direct visual feedback using our wall-sized display and VR headsets determines the stereoscopic features and environments accessible to MinOmics users, ranging from pure monoscopic (scenario 1) to fully immersive VR (scenario 4) visualization.

4.1.2 Scenario 2: Full-Screen MinOmics with Embedded UnityMol WebVR in Immersive VR

With MinOmics, wall-sized displays provide a complete overview of different omics data along with 3D molecular structures. For an in-depth and immersive visualization, Virtual Reality systems such as Cave Automatic Virtual Environment (CAVEs) or VR headsets are relevant solutions as they provide suitable interaction metaphors and high-quality adaptive stereoscopic rendering. Scenario 2 provides a way for multiple users to collaborate on the MinOmics analysis. Stereoscopic visualization is enabled for a single user through wearing a VR headset. This requirement is somewhat disruptive, as this user loses access to the initial MinOmics data view in the virtual scene once he puts on the headset. This limitation constitutes the main drawback of scenario 2, as visual analytics requires a direct visual feedback of the operations carried out during analysis. A possible usage scenario can be imagined in a collaborative context, where the main MinOmics user (or group of users) provides and updates the raw data that the VR user can explore and manipulate. As both are co-located in our setup, vocal communication is straightforward. To stay "connected", the MinOmics users have a restricted 2D view of the 3D scene the main VR user is visualizing (Figure 7, panel 2).

4.1.3 Scenario 3: Split-Screen Featuring MinOmics Connected to a Stereoscopic UnityMol View

Stereoscopic rendering allows benefiting from the 3D effect to visualize complex 3D objects like protein structures or protein networks. Due to current limitations in the WebGL implementation of the Unity3D engine, 3D stereoscopic rendering cannot be triggered from this context, which we hope to be able to address in the future by adapting the WebVR implementation for stereoscopic rendering. Our current solution is to use a standalone version of UnityMol with stereoscopic rendering enabled on part of the display wall (for example, half the screen, but this is fully configurable), which is communicating with the MinOmics web components. TCP sockets and WebSockets between the browser and a standalone application can be easily established via the localhost address. However, it should be mentioned that this approach might imply security issues, depending on the context where such a solution is deployed. Although the system is heavier to install and set up, high-performance rendering is possible thanks to OpenGL and DirectX access. In the setup we experimented with, half of the wall was dedicated to the MinOmics web page (in monoscopy) and the other half to the UnityMol stereoscopic visualization. The operating system's window manager allows the user to resize these windows on the fly (Figure 7, panel 3 and Figure 8).

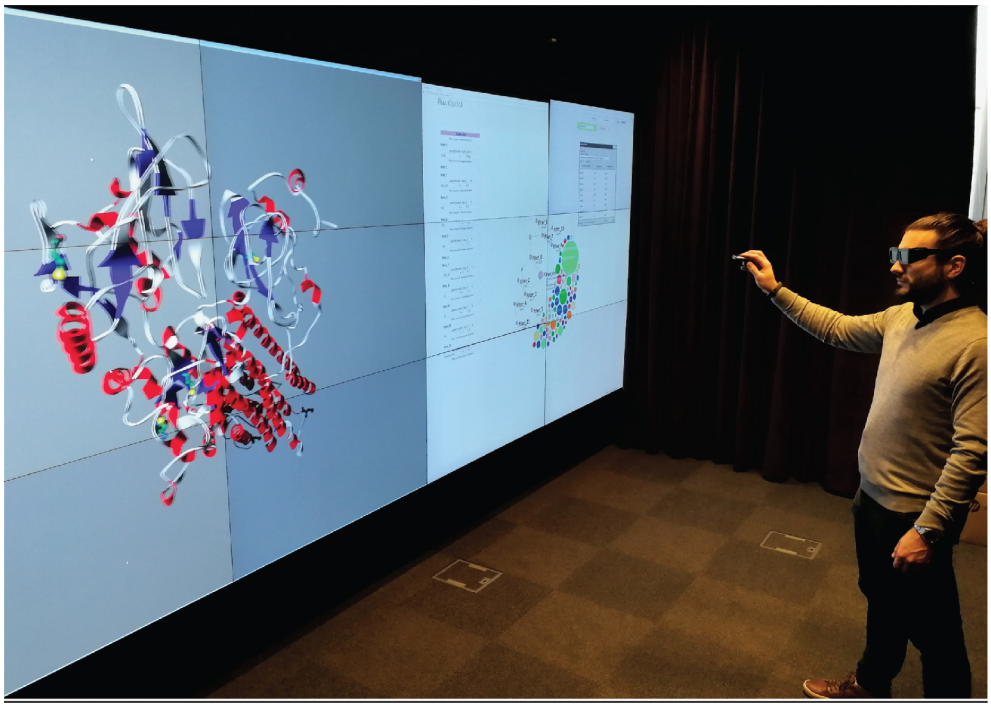

Figure 8: Split-screen setup with stereoscopy. Example of MinOmics running on a web browser on the right half of the display wall, linked to a 3D stereoscopic rendering via UnityMol filling the left half. A gyroscopic mouse is used to operate the system and interact with all elements.

A gyroscopic mouse is used to interact with both the MinOmics analysis window and the UnityMol visualization window providing a unified but limited interaction metaphor. This simple solution is convenient when a single user is analysing data, but troublesome in a multi-user context, as the gyroscopic mouse is managed by the operating system, which allows using only one mouse at a time. To overrule this limitation, a solution would be to develop a custom input device management system able to process multiple mouse inputs, shunting the operating system layer.

4.1.4 Scenario 4: Full VR Context by Bringing MinOmics Inside UnityMol VR

Using a standalone software instance that communicates with or even embeds the MinOmics server part enables to access the workstation's full performance and features, leveraging recent advances in the VR field. The VR implementation of UnityMol provides an immersive way to visualize 3D molecular data coming from the MinOmics workflow and to interact with suitable 3D interaction metaphors using a Leap Motion or common VR controllers (Figure 7, panel 4 and Figure 9). To bring and integrate MinOmics within the VR context, a 2D web browser interface can be mapped to a customizable billboard placed anywhere in the scene.

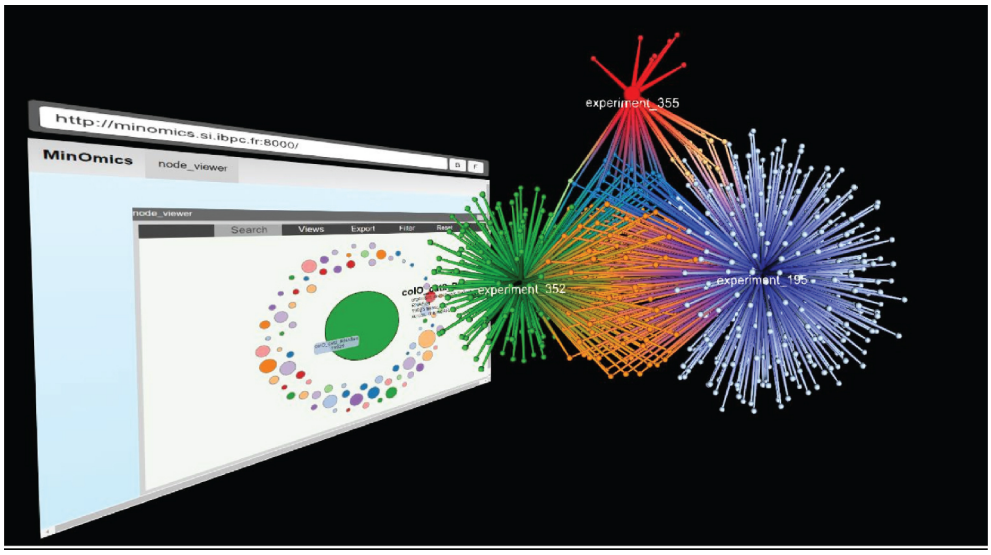

Figure 9: Example of a MinOmics webpage inside a Unity3D application. A similar embedding can be achieved inside a UnityMol application, where a web browser could render the MinOmics web page alongside the 3D molecular data or a proteomic network as illustrated here.

The user can interact with this interface using a VR controller to mimic a mouse with a pointing metaphor and process omics data using MinOmics features. Visualization of 3D data can thus be directly done and controlled from within the virtual environment. Note that this approach is currently restricted to a single user. Adding multiple users implies synchronizing scenes and molecular structure data for each computer attached to a VR headset. The user manipulating the 3D object already benefits from relevant 3D interaction metaphors like VR or Leap Motion controllers. To go further and overcome this limitation, multiple users and potentially distant users could connect to a shared virtual space and perform collaborative analysis over a network.

4.2 Scientific Applications

MinOmics allows visual exploration of multiple omics datasets based on user-defined criteria. We will employ the three redox proteomic datasets generated in the green alga *C. reinhardtii* for SSG, SNO and SS. These ensembles are used to illustrate how MinOmics can be employed to explore the specificity of redox PTMs at the proteome scale through several specific use case.

4.2.1 Protein Properties

The three datasets correspond to a list of unique proteins, identified by their accession number (usually Uniprot ID) and represent proteins undergoing each modification (SS, SNO, SSG). The properties of the proteins in each dataset can be easily visualized using user-defined criteria based on the features determined by/through the MinOmics pipeline. These features are inferred by MinOmics from other external and public databases or prediction algorithms and are not included in the initial proteomic dataset files provided. They allow visualization of the data according to biological features (e.g. subcellular localization, functional annotation) or physico-chemical properties of the proteins (e.g. molecular weight, number of cysteines, hydrophobicity, ...). This type of visualization allows user and data-driven exploration of the results. The three redox PTMs studied are mainly triggered under stress conditions associated with the production of reactive oxygen and nitrogen species (ROS/RNS). The production of these species that drives redox PTMs mainly occurs in specific subcellular compartments, especially electron transfer chains of the mitochondria and chloroplasts. Therefore, one could wonder whether the proteins from these compartments are more susceptible to undergo redox PTMs due to increased local concentrations of ROS/RNS. This question can easily be answered using MinOmics. For example, after loading the largest dataset containing 1188 proteins harboring thioredoxin-regulated disulphides (SS), the user can represent the proteins according to their subcellular localization determined by MinOmics using the PredAlgo software, a multi-subcellular localization prediction tool dedicated to algae [62]. The distribution can be visualized by the user as a pie chart that reveals that the 1188 proteins are distributed for 30 % to the chloroplast, 9 % to mitochondria, 6 % to the secretory pathway and 55 % to other compartments (see Figure 10 and Supplemental Material S2). For comparison, the user can load our control dataset containing

an experimental total soluble proteome of *C. reinhardtii*, and visualize its subcellular distribution (unpublished data). This view immediately reveals that the distributions of the two datasets (SS and soluble proteome) are comparable in percentages. This result suggests that the propensity of proteins to undergo SS oxidoreduction is not influenced by their cellular localization. It is consistent with the established importance of thioredoxins in multiple subcellular compartments, organs and developmental stages of photosynthetic organisms [42].

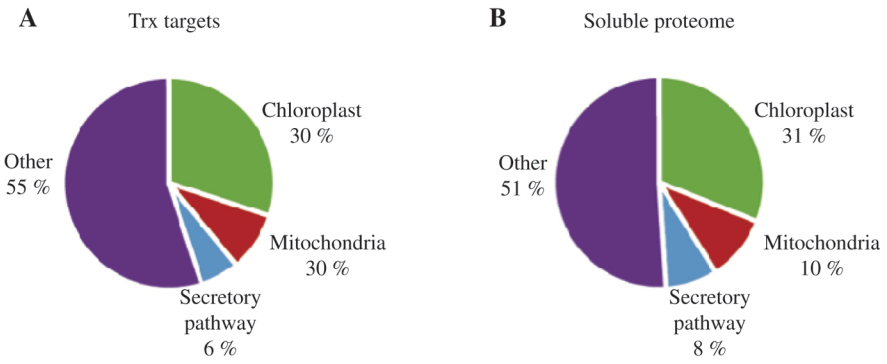

Figure 10: Subcellular localization of Chlamydomonas proteins. Pie charts of the subcellular localization of Trx-targets (A) are shown in comparison to proteins of the soluble proteome (B).

With a large visualization screen, a huge number of such distributions can be compared and visualized simultaneously to rapidly identify potential outliers, i.e. PTM proteomes that would not follow the general rule and would deserve further analysis. This type of scrutiny can be applied to any biological or physicochemical property of interest for the MinOmics user.

4.2.2 Use Case 1: Exploring the Protein Specificity of Multiple Redox PTMs

Numerous proteins are known to be regulated by multiple redox PTMs such as the *Escherichia coli* transcription factor OxyR [63] or the 11 enzymes of the Calvin-Benson cycle in photosynthetic organisms [64]. This feature may not be true for all redox modified proteins and we can wonder whether multiple redox PTMs occur on a limited number of proteins containing reactive cysteines or if each modification targets a distinct redox network. The first use case will address this question using the three datasets available for *C. reinhardtii* (SS, SSG, SNO). The aim is to explore the specificity of redox PTMs at the proteome scale. MinOmics allows fast and easy comparison of multiple proteomes generated in the same organism. After loading the datasets generated, the user can select the proteomes to be compared and activate the grouping function of MinOmics that will analyse the overlap between the datasets and provide a graphical representation of the resulting network for further visual analysis. With the three *Chlamydomonas* datasets, the user obtains a comprehensive map of the algal redox network (Figure 11). Although some proteins are clearly targeted by multiple PTMs, the overlap appears limited since 68.8 % appear regulated by a single type of modification. A similarly high specificity was observed when comparing 193 sulfenylated proteins with previously identified targets of SS formation, SNO and SSG [65]. The limited overlap between the protein targets suggests that the different redox PTMs are specific toward different subsets of the proteome.

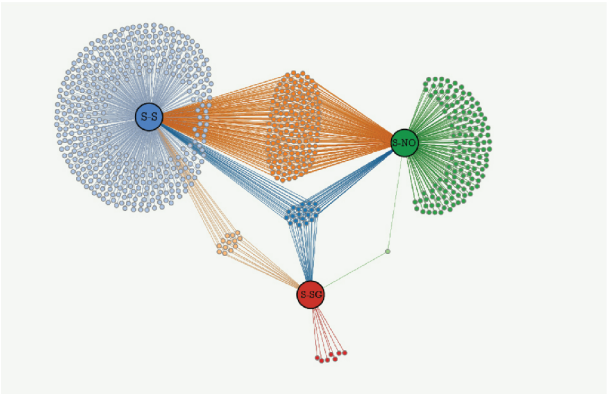

Figure 11: Redox PTMs network in *C. reinhardtii*. Proteins undergoing glutathionylation (S-SG), nitrosylation (S-NO) and/or reduction by thioredoxin (S-S) are grouped by the nature of their redox modification either in single-modification clusters or intersecting multiple modification groups.

4.2.2.1 Constituting and Exploring a First Network Representation in 2D

The redox network or Cys proteome can be easily visualized with MinOmics (Figure 11). Visual exploration and hovering over nodes can display additional information. Clicking on a node opens the corresponding structural model in UnityMol for exploration. More advanced structural grouping and superposition features have been explored manually so far, and remain to be implemented in an automated fashion. In any case, this representation is rather static while the redox network probably involves spatial and temporal regulation of several redox PTMs on 100 of proteins in a highly dynamic manner. This network is likely a major component of signal integration and constitutes the molecular signature of the ROS/RNS crosstalk. Understanding this complex network requires to determine the stoichiometry and dynamics of multiple redox PTMs under diverse physiological conditions or in different genetic backgrounds using time-resolved quantitative proteomics. This capacity should be favored in the future by the development of sensitive and accurate redox quantitative mass spectrometry approaches. Such time-resolved analyses will generate big data which analysis could take advantage of the MinOmics framework and its visual analytics approach powered by UnityMol. The use of a large-scale display will be crucial to allow visual analysis of time series in diverse physiological conditions. In addition, exploring this dynamic network with MinOmics will allow integrating the Cys proteome at the structural level in order to gain insights into the molecular mechanisms and the structural determinants governing each type of redox modification. Moreover, besides redox PTMs, the integration of the signal implicates a myriad of other molecules and processes acting at multiple levels. The multi-omics capacities of MinOmics could be crucial to integrate redox networks with other signaling pathways. This integration will be crucial to understand how environmental challenges are encoded into a biochemical signal that can trigger the appropriate responses in terms of localization, duration and intensity, at the genome, transcriptome, proteome and metabolome level to allow adaptation and survival.

4.2.2.2 Towards Immersive Protein Network Exploration in 3D

The power of MinOmics combined with the 3D UnityMol visualization enables us to envisage an extension of the exploration possibilities of the Redox PTM network shown in Figure 11 in 2D. MinOmics offers the possibility to enrich the network description with analyses from many different omics data. With the user immersed in a virtual environment depicting a 3D representation of the network (our 3rd visualization scenario described above) further hypotheses on the network grouping can be assessed. In accordance with Shneiderman's principles [66], several actions will be mandatory to explore and analyse the network. The actions include zooming out to gain an overview of the complete network, zooming in on items of interest, filtering out uninteresting items, adding details on demand when needed, viewing relationships between network nodes, store a history on the refinement and be able to extract sub-collections and query parameters. The literature exposes many different possibilities to display such networks, both in 2D and in 3D. Here, we propose to extend the MinOmics 2D approach into 3D by enriching the data used for network arrangement and visualization. We envisage two approaches to do so, a first simple one is implemented in the current version by assigning the third dimension of each node based on a user defined parameter extracted from the MinOmics data server (Figure 12).

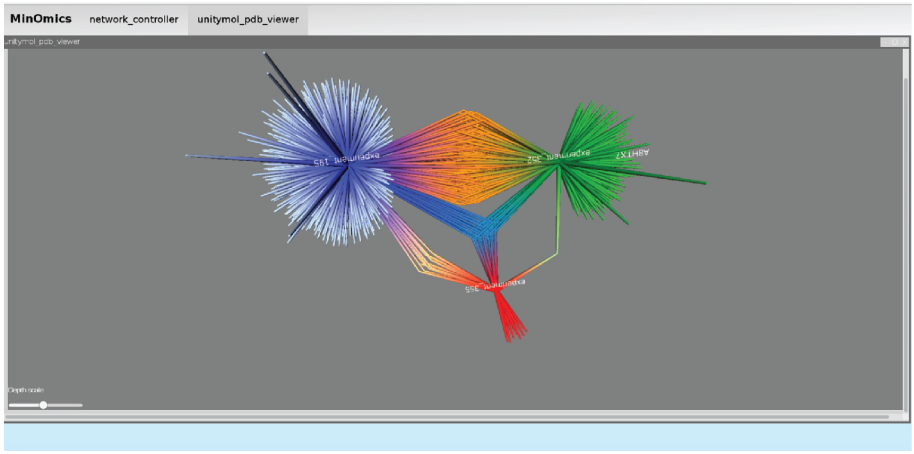

Figure 12: Redox PTMs 3D network in *C. reinhardtii*. The same PTM network as in Figure 11 is shown here in UnityMol after extending it to 3D. We can now continuously navigate from the 2D to the 3D network. The z-axis represents the molecular weight as an example.

Thereby, we ensure a match between the 2D and the 3D color mapping, while adding valuable information through the 3rd dimension in the VR view. Nodes that are close in the 2D representation of the network would

be expected to stay nearby the same nodes in VR, if a pertinent descriptor is found, colors and relative scales are kept. To change the scale and weight of this dynamically chosen depth parameter, a slider can be added to the MinOmics web page or included in the VR environment as illustrated in Figure 12.

An alternative approach to explore 3D would rely on exploiting the Unity3D physics engine, where relationships between nodes are translated into either attraction or repulsion forces, so that the ensemble of nodes would auto-organize based on the chosen mapping. We will explore this approach later on.

4.2.3 Use Case 2: Exploring the Cysteine Site Specificity of Multiple Redox PTMs

Despite the specificity observed at the proteome scale, 100 of proteins are regulated by multiple redox PTMs. Nevertheless, the fact that these proteins undergo two or three distinct modifications does not necessarily imply that the same cysteine is targeted. In the second use case we will use MinOmics to explore the site specificity of the different redox PTMs at the proteome scale. After loading the datasets generated, the user can first select the proteomes to be compared and, after selecting redox-modified cysteines, can activate the grouping function of MinOmics that will analyse the overlap between the residues undergoing each redox PTM (SS, SSG, SNO). The visualization reveals a strikingly high specificity of each modification for specific cysteine residues. Indeed, when all modified cysteines are considered, more than 87 % are found to undergo a single modification while less than 1 % are regulated by the three PTMs. However, this analysis is considerably biased by the fact that most proteins (68.8 %, as revealed by use case 1) undergo a single modification. Therefore, to truly explore cysteine specificity, the user selects only cysteines from proteins undergoing at least two different redox PTMs and visualizes the overlap with the grouping function of MinOmics. Astonishingly, despite this filtering the proportion of proteins specifically modified by one modification remains very high (86 %). This observation indicates that when a protein undergoes several modifications, in most cases the target cysteine residues are different, even if they belong to the same polypeptide. These results indicate that the Cys proteome does not represent a small subset of highly reactive cysteines that would be modified through indiscriminate interaction with the reactive molecules they encounter (e.g. ROS and RNS) but represents a complex organized network. The different redox PTMs appear to control different subnetworks that are largely interconnected. Strikingly, a recent analysis of 1319 mouse liver proteins and four cysteine modifications (SNO, SSG, sulfenylation and S-acylation) also revealed a very high specificity of redox PTMs with limited overlap [67]. These results suggest that the different redox modifications are specific toward distinct interconnected protein networks.

4.2.4 Use Case 3: Exploring Structural Determinants of Redox-Modified Cysteines

Bioinformatics tools to reliably predict redox modification sites are currently lacking, presumably because they depend on the environment of the target cysteine in the folded protein rather than on the primary sequence alone. The specificity of the different redox PTMs primarily depends on the biochemical properties of the target cysteine residue that are largely linked to its microenvironment within the folded protein, which can notably influence the accessibility, the acidity, and the nucleophilicity of the residue [38], [39], [40]. The MinOmics framework coupled to UnityMol enables structural interpretation of proteomic datasets. The third study case will focus on the structural analysis of the microenvironment of redox-modified cysteines (Figure 13).

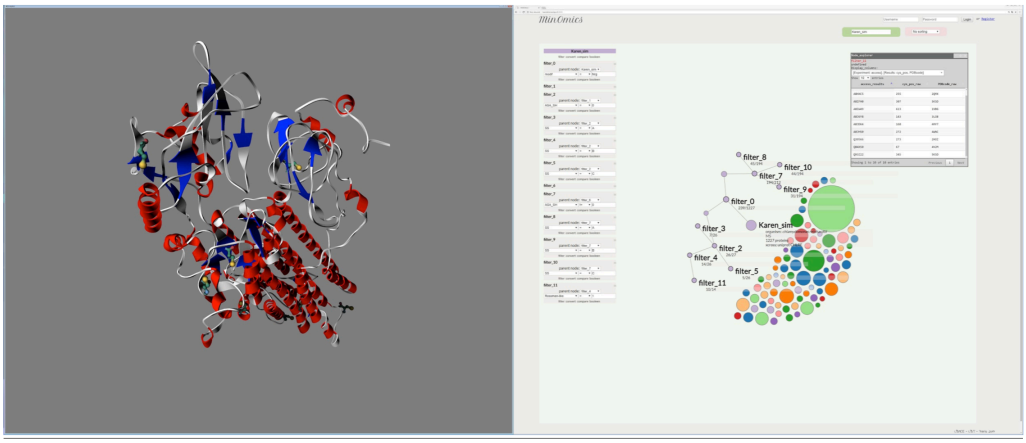

Figure 13: Navigation through MinOmics datasets. To illustrate navigation through datasets available in MinOmics, we investigate the subset of nitrosylated proteins. A combination of filtering clauses permits to i) focus on nitrosylated proteins (filter 0), ii) located on a β-sheet (filter 4), iii) with a cysteine and thiol buried inside the protein surface (filter 1 and 2) and iv) located on a Rossman-like fold (filter 11). This part is shown on the right half split-screen. The left half split-screen, a UnityMol stereoscopic view, allows to explore the molecular structures that were selected through the filtering process, here one of the proteins selected through the combined filters with highlighted cysteines.

The user loads the *Chlamydomonas* SNO dataset, selects redox-modified cysteines and splits the nitrosylome network in two groups according to cysteine accessibility (buried vs. accessible). The user can then select to visualize these two groups according to the type of secondary structure containing these cysteines (α-helix, β-strand or random coil). This visual analysis reveals that the majority of buried cysteines are positioned on β-strands (46.7 %), a proportion significantly higher compared to exposed cysteines (26.6 %) (Figure 14A and B). It is then possible for the user to investigate whether this particularity can have an impact on the cysteine modification. The user selects the 14 cysteine residues that are both buried and located on β-strands. He then opens the corresponding three-dimensional structures that can be visualized simultaneously next to each other on the large display wall, or superposed with each other within the VR headset. This visual analysis reveals that nine proteins out of 14 adopt a similar fold. Further inspection of these nine structures revealed that they all harbor a Rossman or sandwich fold and that the modified cysteine shares a similar position within the structure (Figure 14C). The user can perform the same type of analysis on a distinct control dataset, such as one comprising cysteines involved in a TRX regulated disulphide (SS). This analysis yields results comparable to those obtained for nitrosylated cysteines, since 26 proteins out of 49 possess a modified cysteine buried on a β-strand in a Rossman or sandwich fold. These results suggest that buried cysteines are more likely to be redox modified if they are located on a β-strand within a Rossman fold protein. This fold may increase cysteine reactivity but does not appear to contribute to the specificity of the modification, at least not for SNO and SS. Additional analyses with other datasets for different redox PTMs and from diverse organisms will need to be explored to confirm and generalize this observation. This use case illustrates how MinOmics can lead from the genome level to detailed molecular-level structural interpretations of proteomic data. Future developments of MinOmics will allow automated analysis of the structural environment of modified residues in order to perform statistical analysis of the physico-chemical environment of modified cysteines (charges, hydrophobicity, accessibility, nature of neighboring amino acids, ...). These data could be employed to train an algorithm, which would be able to predict potential redox modification sites for a given protein, based on its actual or predicted structure.

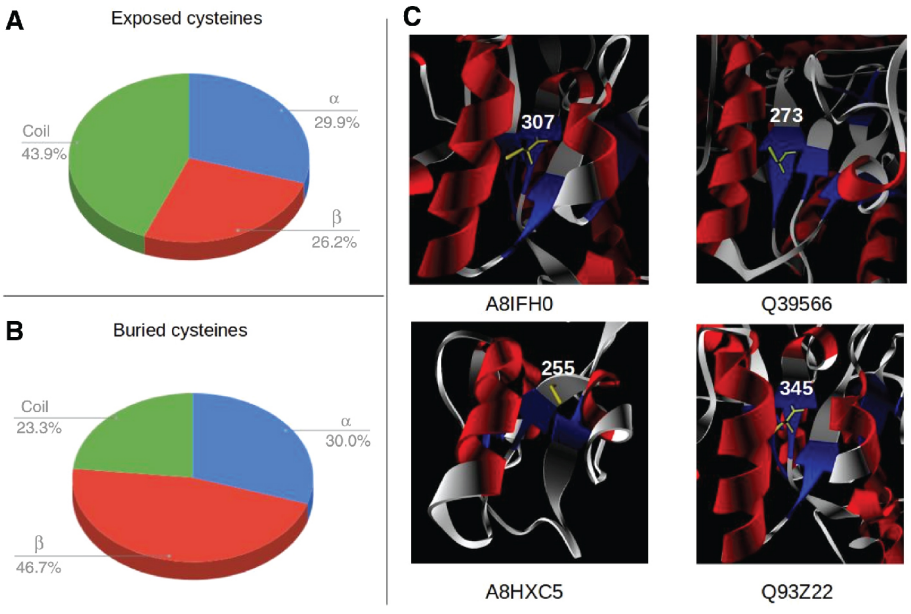

Figure 14: Structural analysis of buried PTM modified cysteine residues. (A) and (B) Overall distribution of secondary structure motifs for the protein backbone stretches hosting exposed and buried cysteines, respectively. (C) Visualization of four cysteine sites that are all nitrosylated and located on β-strands within a Rossman fold.

5 Conclusion and Perspectives

In this paper we present MinOmics, an innovative analysis pipeline for multi-omics data that benefits from the latest data management, web and 3D technologies. The Python Django framework provides clean, fast, secure structure to the web interface and query builder while the efficient design of the PostgreSQL database allows retrieving the query results in a matter of milliseconds. UnityMol takes advantage of the latest progress in game engines to provide advanced 3D representations of molecular and network models embedded in a WebGL version. Integrated in virtual reality environments, those tools provide an accessible and fully immersive comprehensive picture of multi-omics datasets.

MinOmics combined with UnityMol offers a web-based tool using a visual analytics approach to manage and analyse numerous omics data types, providing an accessible and clear tool to achieve in-depth insight of complex and interconnected scientific data. Here, we have presented the main concepts and design principles. The full-fledged application is still in development and requires further stabilization of the different elements to achieve the robustness expected for everyday use. Some parts are already quite advanced in that respect, such as the database server backend. The link with the visualization components and the full implementation of the visual analytics features are still work in progress. Some general challenges in the field concern MinOmics as all other applications of that type, for instance in terms of Human Computer Interaction capacities. Efficiently interacting with a large-scale, high-resolution, stereoscopic display wall is generally still a largely unsolved problem.

5.1 Perspectives on UnityMol WebGL

To go even further in terms of 2D and 3D integration, the MinOmics 2D interface could be fully refactored inside a VR environment. In such a scenario, the users would manage omics data in real time thanks to the MinOmics architecture, while analysing and visualizing relevant data plots and molecular structure relationships in a multi-user and collaborative context. This combination would open the way for immersive visual analytics in VR, benefiting from interaction metaphors suitable for molecular structures but also taking advantage of stereoscopic rendering and complete immersion.

6 Supplementary Material

Supporting information is provided to illustrate several aspects of this paper in more detail, in particular through movies and screen captures.

Acknowledgements

We would particularly like to thank our colleague H. Santuz for extensive assistance with the display wall implementation. G. Labesse and J.L. Pons were of precious help by modeling the full Chlamydomonas protein dataset structures with their @tome2 server, and through stimulating subsequent exchanges on the homology modeling part of this work.

Funding

This work was supported by the Paris Ile-de-France Region and by the Fondation pour la Recherche Médicale, grant number FRM DBI20141231801, to SL. XM and MB thank UCB Biopharma for support. We further acknowledge support by the "Initiative d'Excellence" program from the French State (Funder Id: 10.13039/501100001665, Grant "DYNAMO", ANR-11-LABX-0011-01 and Funder Id: 10.13039/501100001665, Grant "CACSICE", ANR-11-EQPX-0008)".

Conflict of interest statement: Authors state no conflict of interest. All authors have read the journal's Publication ethics and publication malpractice statement available at the journal's website and hereby confirm that they comply with all its parts applicable to the present scientific work.

References

[1] Gomez-Cabrero D, Abugessaisa I, Maier D, Teschendorff A, Merkenschlager M, Gisel A, et al. Data integration in the era of omics: current and future challenges. BMC Syst Biol. 2014;8(Suppl 2):I1.
[2] Benson DA, Cavanaugh M, Clark K, Karsch-Mizrachi I, Ostell J, Pruitt KD, et al. GenBank. Nucleic Acids Res. 2018;46(D1):D41–7.
[3] Kanz C, Aldebert P, Althorpe N, Baker W, Baldwin A, Bates K, et al. The EMBL Nucleotide Sequence Database. Nucleic Acids Res. 2005;33:D29–33.
[4] Karolchik D, Baertsch R, Diekhans M, Furey TS, Hinrichs A, Lu YT, et al. The UCSC Genome Browser Database. Nucleic Acids Res. 2003;31:51–4.
[5] Barker WC, Garavelli JS, Huang H, McGarvey PB, Orcutt BC, Srinivasarao GY, et al. The protein information resource (PIR). Nucleic Acids Res. 2000;28:41–4.
[6] Pundir S, Martin MJ, O'Donovan C. UniProt Protein Knowledgebase. Methods Mol Biol. 2017;1558:41–55.
[7] Berman HM, Westbrook J, Feng Z, Gilliland G, Bhat TN, Weissig H, et al. The Protein Data Bank. Nucleic Acids Res. 2000;28:235–42.
[8] Ashburner M, Ball CA, Blake JA, Botstein D, Butler H, Cherry JM, et al. Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. Nat Genet. 2000;25:25–9.
[9] Finn RD, Bateman A, Clements J, Coggill P, Eberhardt RY, Eddy SR, et al. Pfam: the protein families database. Nucleic Acids Res. 2014;42:D222–30.
[10] Hubbard TJ, Murzin AG, Brenner SE, Chothia C. SCOP: a structural classification of proteins database. Nucleic Acids Res. 1997;25:236–9.
[11] Mi H, Poudel S, Muruganujan A, Casagrande JT, Thomas PD. PANTHER version 10: expanded protein families and functions, and analysis tools. Nucleic Acids Res. 2016;44(D1):D336–42.
[12] Binder JX, Pletscher-Frankild S, Tsafou K, Stolte C, O'Donoghue SI, Schneider R, et al. COMPARTMENTS: unification and visualization of protein subcellular localization evidence. Database (Oxford). 2014;2014:bau012.
[13] Kielman J, Thomas J, May R. Foundations and frontiers in visual analytics. Inf Vis. 2009;8:239–46.
[14] Khushi M. Benchmarking database performance for genomic data. J Cell Biochem. 2015;116:877–83.
[15] Kozanitis C, Heiberg A, Varghese G, Bafna V. Using Genome Query Language to uncover genetic variation. Bioinformatics. 2014;30:1–8.
[16] Latendresse M, Karp PD. An advanced web query interface for biological databases. Database (Oxford). 2010;2010:baq006.
[17] Vilaplana J, Solsona F, Teixido I, Usie A, Karathia H, Alves R, et al. Database constraints applied to metabolic pathway reconstruction tools. ScientificWorldJournal. 2014;2014:967294.
[18] Holovaty A, Kaplan-Moss J. The definitive guide to Django: Web development done right: Apress; 2009.
[19] Marrin C. Webgl specification. Khronos WebGL Working Group. 2011.
[20] Lv Z, Tek A, Da Silva F, Empereur-mot C, Chavent M, Baaden M. Game on, science – how video game technology may help biologists tackle visualization challenges. PLoS One. 2013;8:e57990.
[21] Rose AS, Hildebrand PW. NGL Viewer: a web application for molecular visualization. Nucleic Acids Res. 2015;43:W576–9.
[22] Zakai A, editor Emscripten: an LLVM-to-JavaScript compiler. Proceedings of the ACM international conference companion on Object oriented programming systems languages and applications companion; 2011: ACM.
[23] Fung DC, Hong SH, Koschutzki D, Schreiber F, Xu K. 2.5D visualisation of overlapping biological networks. J Integr Bioinform. 2008;5:337–42.
[24] Widjaja YY, Pang CN, Li SS, Wilkins MR, Lambert TD. The Interactorium: visualising proteins, complexes and interaction networks in a virtual 3-D cell. Proteomics. 2009;9:5309–15.
[25] Secrier M, Pavlopoulos GA, Aerts J, Schneider R. Arena3D: visualizing time-driven phenotypic differences in biological systems. BMC Bioinformatics. 2012;13:45.
[26] Sommer B, Tiys ES, Kormeier B, Hippe K, Janowski SJ, Ivanisenko TV, et al. Visualization and analysis of a cardio vascular disease- and MUPP1-related biological network combining text mining and data warehouse approaches. J Integr Bioinform. 2010;7:148.
[27] O'Donoghue SI, Sabir KS, Kalemanov M, Stolte C, Wellmann B, Ho V, et al. Aquaria: simplifying discovery and insight from protein structures. Nat Methods. 2015;12:98–9.
[28] Topel T, Kormeier B, Klassen A, Hofestadt R. BioDWH: a data warehouse kit for life science data integration. J Integr Bioinform. 2008;5:93–102.
[29] Sommer B, Barnes DG, Boyd S, Chandler T, Cordeil M, Czauderna T, et al. 3D-stereoscopic immersive analytics projects at Monash University and University of Konstanz. Electronic Imaging. 2017;2017:179–87.
[30] Scaife MA, Nguyen GT, Rico J, Lambert D, Helliwell KE, Smith AG. Establishing Chlamydomonas reinhardtii as an industrial biotechnology host. Plant J. 2015;82:532–46.
[31] Choudhary C, Weinert BT, Nishida Y, Verdin E, Mann M. The growing landscape of lysine acetylation links metabolism and cell signalling. Nat Rev Mol Cell Biol. 2014;15:536–50.
[32] Go YM, Chandler JD, Jones DP. The cysteine proteome. Free Radic Biol Med. 2015;84:227–45.
[33] Couturier J, Jacquot JP, Rouhier N. Toward a refined classification of class I dithiol glutaredoxins from poplar: biochemical basis for the definition of two subclasses. Front Plant Sci. 2013;4:518.
[34] Go YM, Jones DP. Redox biology: interface of the exposome with the proteome, epigenome and genome. Redox Biol. 2014;2:358–60.
[35] Paulsen CE, Carroll KS. Cysteine-mediated redox signaling: chemistry, biology, and tools for discovery. Chem Rev. 2013;113:4633–79.
[36] Poole LB, Schoneich C. Introduction: What we do and do not know regarding redox processes of thiols in signaling pathways. Free Radic Biol Med. 2015;80:145–7.
[37] Weerapana E, Wang C, Simon GM, Richter F, Khare S, Dillon MB, et al. Quantitative reactivity profiling predicts functional cysteines in proteomes. Nature. 2010;468:790–5.
[38] Reddie KG, Carroll KS. Expanding the functional diversity of proteins through cysteine oxidation. Curr Opin Chem Biol. 2008;12:746–54.
[39] Winterbourn CC, Hampton MB. Thiol chemistry and specificity in redox signaling. Free Radic Biol Med. 2008;45:549–61.

[40] Zaffagnini M, Bedhomme M, Groni H, Marchand CH, Puppo C, Gontero B, et al. Glutathionylation in the photosynthetic model organism Chlamydomonas reinhardtii: a proteomic survey. Mol Cell Proteomics. 2012;11:M111.014142.
[41] Morisse S, Zaffagnini M, Gao XH, Lemaire SD, Marchand CH. Insight into protein S-nitrosylation in Chlamydomonas reinhardtii. Antioxid Redox Signal. 2014;21:1271–84.
[42] Perez-Perez ME, Mauries A, Maes A, Tourasse NJ, Hamon M, Lemaire SD, et al. The deep thioredoxome in chlamydomonas reinhardtii: new insights into redox regulation. Mol Plant. 2017;10:1107–25.
[43] Morisse S, Michelet L, Bedhomme M, Marchand CH, Calvaresi M, Trost P, et al. Thioredoxin-dependent redox regulation of chloroplastic phosphoglycerate kinase from Chlamydomonas reinhardtii. J Biol Chem. 2014;289:30012–24.
[44] Pons JL, Labesse G. @TOME-2: a new pipeline for comparative modeling of protein-ligand complexes. Nucleic Acids Res. 2009;37:W485–91.
[45] Soding J. Protein homology detection by HMM-HMM comparison. Bioinformatics. 2005;21:951–60.
[46] Shi J, Blundell TL, Mizuguchi K. FUGUE: sequence-structure homology recognition using environment-specific substitution tables and structure-dependent gap penalties. J Mol Biol. 2001;310:243–57.
[47] Altschul SF, Madden TL, Schaffer AA, Zhang J, Zhang Z, Miller W, et al. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res. 1997;25:3389–402.
[48] Zhou H, Zhou Y. SPARKS 2 and SP3 servers in CASP6. Proteins. 2005;61(Suppl 7):152–6.
[49] Labesse G, Mornon J. Incremental threading optimization (TITO) to help alignment and modelling of remote homologues. Bioinformatics. 1998;14:206–11.
[50] Canutescu AA, Shelenkov AA, Dunbrack RL, Jr. A graph-theory algorithm for rapid protein side-chain prediction. Protein Sci. 2003;12:2001–14.
[51] Sali A, Potterton L, Yuan F, van Vlijmen H, Karplus M. Evaluation of comparative protein modeling by MODELLER. Proteins. 1995;23:318–26.
[52] Benkert P, Tosatto SC, Schomburg D. QMEAN: A comprehensive scoring function for model quality assessment. Proteins. 2008;71:261–77.
[53] Li H, Robertson AD, Jensen JH. Very fast empirical prediction and rationalization of protein pKa values. Proteins. 2005;61:704–21.
[54] Hubbard S, Thornton J. NACCESS: Department of Biochemistry and Molecular Biology, University College London. Software available at http://www.bioinf.manchester.ac.uk/naccess/nacdownload.html. 1993.
[55] Kabsch W, Sander C. Dictionary of protein secondary structure: pattern recognition of hydrogen-bonded and geometrical features. Biopolymers. 1983;22:2577–637.
[56] Kortemme T, Creighton TE. Ionisation of cysteine residues at the termini of model alpha-helical peptides. Relevance to unusual thiol pKa values in proteins of the thioredoxin family. J Mol Biol. 1995;253:799–812.
[57] RDevelopment CORE TEAM R. R: A language and environment for statistical computing. Austria: R foundation for statistical computing Vienna; 2008.
[58] Annett J. Hierarchical task analysis. Handbook of cognitive task design. 2003;2:17–35.
[59] Perez S, Tubiana T, Imberty A, Baaden M. Three-dimensional representations of complex carbohydrates and polysaccharides–SweetUnityMol: a video game-based computer graphic software. Glycobiology. 2015;25:483–91.
[60] Trellet M, Ferey N, Baaden M, Bourdot P, editors. Content and task based navigation for structural biology in 3D environments. Virtual and Augmented Reality for Molecular Science (VARMS@ IEEEVR), 2015 IEEE 1st International Workshop on; 2015: IEEE, 2015.
[61] Chavent M, Vanel A, Tek A, Levy B, Robert S, Raffin B, et al. GPU-accelerated atom and dynamic bond visualization using hyperballs: a unified algorithm for balls, sticks, and hyperboloids. J Comput Chem. 2011;32:2924–35.
[62] Tardif M, Atteia A, Specht M, Cogne G, Rolland N, Brugiere S, et al. PredAlgo: a new subcellular localization prediction tool dedicated to green algae. Mol Biol Evol. 2012;29:3625–39.
[63] Seth D, Hausladen A, Wang YJ, Stamler JS. Endogenous protein S-Nitrosylation in E. coli: regulation by OxyR. Science. 2012;336:470–3.
[64] Michelet L, Zaffagnini M, Morisse S, Sparla F, Perez-Perez ME, Francia F, et al. Redox regulation of the Calvin-Benson cycle: something old, something new. Front Plant Sci. 2013;4:470.
[65] Leonard SE, Reddie KG, Carroll KS. Mining the thiol proteome for sulfenic acid modifications reveals new targets for oxidation in cells. ACS Chem Biol. 2009;4:783–99.
[66] Shneiderman B. The eyes have it: a task by data type taxonomy for information visualizations. the craft of information visualization. San Francisco: Morgan Kaufmann; 2003. p. 364–71.
[67] Gould NS, Evans P, Martinez-Acedo P, Marino SM, Gladyshev VN, Carroll KS, et al. Site-Specific Proteomic Mapping Identifies Selectively Modified Regulatory Cysteine Residues in Functionally Distinct Protein Networks. Chem Biol. 2015;22:965–75.

Supplemental Material: The online version of this article offers supplementary material (https://doi.org/10.1515/jib-2018-0006).