

Position Paper for Semantic Web Life Sciences Workshop

**J. Hunter (Distributed Systems Technology CRC), M. A. Ragan (ARC Centre in Bioinformatics and Institute for Molecular Bioscience), S. Little (Dept. ITEE)
The University of Queensland**

1. Introduction to the Visible Cell Project

The "[Visible Cell](#)" project is a research project being undertaken at the ARC Centre in Bioinformatics at the University of Queensland. The aim of this project is to significantly progress our understanding of the mammalian cell via the synthesis of physical data, models, mathematical and statistical simulations, and bioinformatics data. A single cell contains tens of thousands of molecules, each interacting with other molecules in complex ways as yet not fully understood. If we can understand, visualise, model, simulate and predict how normal cells behave, we will be that much closer to understanding how abnormal cells such as cancer cells behave. The ability to model and understand the interactions of biomolecules within cells will accelerate the design, discovery and development of biomolecules such as drugs, vaccines, protein therapeutics and gene therapies. It will also be important in understanding essential contemporary issues such as directing stem cell differentiation.

The objective of the Visible Cell project is to provide a visualization environment that seamlessly embeds macromolecular structures, networks and quantitative simulations based on mathematical and complex-system models into a 3D mammalian cell reconstructed from high resolution tomograms and electron micrographs. Using physical information gained by techniques such as high resolution tomography, NMR, electron and X-ray crystallography, it is possible to provide the scientists with a dynamic 3D visualization environment for hypothesis testing and integration of new discoveries. The challenge is to manage, integrate and assimilate the large amounts of information associated with the multiscale physical data, the related highly complex bioinformatics data and the mathematical and statistical simulations.

Instrument measurements provide microscopic image data describing physical geometry and location of sub-cellular components. These empirically derived geometries provide the setting for mathematical simulation. Computer algorithms to predict the precise 3D structures of, for example, an array of proteins can be compared directly with the experimentally gained results. Complex, high-resolution models and simulation systems can assist in the prediction of phenomena such as protein-protein interactions. Appropriate semantic representation of the images will allow tools to automatically access relevant biological databases and literature from around the world, and integrate this data and information within the simulations and visualizations which spatially and temporally map the data onto the model. Virtual reality environments may offer scientists even further immersion in the three-dimensional cell through the use of haptics, providing new innovative mechanisms for discovery.

2. Key Problems/Challenges

There are four main phases in developing a virtual Visible Cell environment that will enable distributed teams of scientists to better understand cell physiology and the behaviour of cells under different circumstances. Each phase has its own challenges and requirements:

1. Develop an underlying 3D spatial matrix from tomographic images

- Display reconstructed tomographic images of real mammalian cells in 3D to serve as the spatial matrix of the Visible Cell (an integrated data, modelling and visualisation

environment). In the first instance, this matrix (the cell in its structural detail, *i.e.* membranes, compartments, vesicles, microtubules, organelles etc.) is static in space and time.

- Render surfaces, volumes, lighting etc. of components of the Visible Cell (membranes, compartments etc.) only to the extent and detail required. Manage (distribute) heavy, real-time computation.

2. *Need to fit proteins into the matrix*

Different scenarios which need to be handled include:

1. protein or macromolecular assembly images available from microscopy, and protein structure known experimentally (x-ray crystallography or NMR);
 2. protein or macromolecular assembly images available from microscopy, but protein structure not known experimentally;
 3. protein or macromolecular assembly images not available from microscopy, but protein structure known experimentally;
 4. protein or macromolecular assembly images are not available from microscopy, and protein structure not known experimentally.
- Where the protein or macromolecular assembly image is available from microscopy, embed them into the appropriate part of the cell matrix (membrane, vesicle, etc.) and associate the underlying protein information accordingly.
 - Where the protein structure not known experimentally, predict as much as possible computationally. Where it is known experimentally, allow for deformation due to hydration, etc. when fitting into microscopic image. Update databases.
 - Where no localisation information is available from microscopy, use alternative rules (predicted charge, pI, hydrophilicity/hydrophobicity; computational prediction of localisation signals; PPI data; database mining for existing annotations; ontology; literature mining) to guide localisation.
 - Efficiently retrieve and exploit the relevant data from different databases in order to fit the proteins onto the cell - data models, images under Oracle Spatial, protein structures from PDB flat files, protein features from Interpro, etc..

3. *The Visible Cell as a dynamic modelling environment*

- Associate molecules with user-specified and user-driven computational models/simulations so that the visualised molecules move in space and time as specified by the models and input parameters. Molecules embedded in the spatial matrix (above) could thereby deform the matrix - generating a matrix that is no longer static, but dynamic.
- Visually smooth the passage as the user navigates across large spatial and time scales, switching computational models where required (*e.g.* from stochastic to deterministic as concentration, volume or temporal scale increases).
- Automated representation of background - present not-immediately-relevant regions - in a visually realistic but computationally light manner.
- Impose gradients/fields (chemical, electrical etc.) across the cell or its parts, such that these gradients/fields interact with the computational models.
- Sparse datasets - mechanisms are required to detect and handle incomplete, inconsistent or redundant data.

4. *The Visible Cell as an environment for data exploration, hypothesis formulation and testing*

- Allow users to interactively specify preferred presentation modes and modify modelling

and simulation parameters;

- Allow the user to identify ("light up") sets of molecules, cell structures, interacting systems etc. according to functional annotation, ontological relationships etc. instead of sequentially.
- 3D data exploration - "zoom in/out" function - regulates level of molecular detail at any given time. Allow users to set parameters to enable/disable certain molecular features.
- Automate integration of relevant literature (PubMed). Extract key features and processes from Visual Cell environment to feed into literature mining. Mine literature for data that tests/validates/contradicts predictions of the model.
- Enable advanced single-user control (steering etc.) of simulations possibly including haptics.
- Enable collaborative multi-user steering of simulations and visualizations - both on-site and remotely.
- Enable logging/recording of the results of simulations. Associate and compare results with the appropriate database over a federation or Grid, and potentially updating that database - where the data are new or conflicting.
- Enable users to specify hypotheses and have the system automatically retrieve, integrate, model and visualize data so the user can determine whether the hypothesis is potentially valid or not;
- Enable hypotheses and their associated multimedia visualization, to be saved, indexed and retrieved so they can be shared and discussed;
- Enable users to attach metadata, annotations/interpretations - which can also be saved, searched and retrieved. Calling these comments, interpretations and metadata to the attention of subsequent users where appropriate, and not doing so if inappropriate - access constraints on annotations.
- Expand system from modelling and visualizing a single cell to groups of cells/tissues. Emphasis on cell surface properties, and systems involved in cell-cell interaction.

3. Progress to Date

Progress has been made in a number of areas. This work can be leveraged, extended and refined to develop new tools and services required by projects such as the Visible Cell project:

- Ontologies - biomedical ontologies such as [OpenGALEN](#) and [SNOMED CT](#); the Gene Ontology ([GO](#)); MPEG-7 ontology [1] for describing 2D images, 3D objects, animations and spatio-temporal relationships; [Open Microscopy Environment \(OME\)](#) for describing microscopic data and [MGED ontology](#) for describing microarray experiments.
- Harmonization of ontologies - a number of research and standards groups are working on the development of common conceptual models (or upper ontologies) to facilitate interoperability between metadata vocabularies and the integration of information from different domains e.g., the ABC Ontology/Model [2] and the Standard Upper Ontology ([SUO, 2002](#)) is being developed by the IEEE SUO Working Group.
- Inferencing rules and deductive query engines [3] - "[Ontology Storage and Querying](#)" published by ICS FORTH provides a good survey of the current state of ontology storage and querying tools. [RuleML](#) enables Web-based XML rule storage, interchange, retrieval and invocation. Possible candidates for RuleML inferencing engines include: [JESS](#) (Java Expert System Shell) and [Mandarax](#), a Java RuleML engine. These inferencing engines are currently limited by: slow processing speed; the need to convert data to in-memory RuleML facts and the lack of native RuleML support for applying standard string and mathematical relations (e.g., greater than, equal to, etc.).

- Analysis of scientific and engineering images and video. A key research challenge is to derive measurements of low-level features from 2D, 3D and multispectral images and video and to use this to derive semantic descriptions and to generate or evaluate new knowledge and hypotheses. [4]
- Hypothesis testing interfaces - which allow users to quickly and easily specify their hypotheses, see whether there was any evidence to support this theory or modify it based on the visual/graphical feedback. A prototype developed by DSTC [5] enables the formulation, storage, search and retrieval of past hypotheses - to enable domain expert knowledge capture and so they can be re-used and further refined as more research data is acquired. This will also help to keep track of past work, reduce duplication and provide evidence and provenance for experimental results.

4. Conclusions and Requirements

The Visible Cell framework needs to integrate physical data, simulation data and bioinformatics data, in order to construct 3D models of cells and cellular processes, with the capability to extract, record and reuse new information and ideas. Consequently it requires mechanisms for storing, indexing, searching, accessing, retrieving, sharing, reusing and tracking and integrating resources which may include:

- biological/cellular data (at different scales), spatial data, 3D models and images, spectral analyses, numerical arrays and matrices, computational models, hypotheses and publications;
- electron microscopes;
- image and data analysis and processing services;
- modelling, simulation and predictive components;
- distributed computing power;
- scientists and experts in diverse fields of biology, chemistry, mathematics, computer science and IT.

Providing a semantically rich framework for the Visible Cell project will depend on the availability of semantic descriptions of these resources. Semantic descriptions will reduce subjectivity, enhance resource discovery and interoperability and allow sophisticated semantic querying and knowledge mining. Semantic inferencing rules offer potential to generate high-level semantic descriptions of cell components from automatically extracted low-level features and to correlate data from across disciplines, media types and formats. However projects such as the Visible Cell project are going to require both extensions to existing Semantic Web technologies as well as the development of new Semantic Web technologies. More specifically it will require tools and services that include:

- Metadata schemas and indexing, search, browse and retrieval interfaces - for storing, indexing, retrieving complex data types such as: arrays, matrices, spectral data, 3D models, mathematical models (both stochastic and deterministic), equations, analytical services, computing power etc.
- Intelligent identification of relevant datasets for pre-processing, as a function of user history, current task, etc.
- Inferencing rules that support more than just the logic-based inferencing provided by RuleML - need to combine RuleML with [CellML](#), [MathML](#), Bioinformatics Sequence ML ([BSML](#)), etc.
- Ability to save and reuse models and hypotheses which combine RuleML, CellML etc.
- Inferencing engines which support the mathematical functions embedded in rules.

- Ontologies are required that support relationships between entities that are not semantic or part-of relationships - mathematical, chemical, electrical, gravitational, spatial and temporal relationships.
- Top-level ontologies for integrating multi-disciplinary ontologies.
- The ability to determine optimum methods for visually representing related data, information and knowledge bases e.g., graphs (2D and 3D), animations, virtual reality, hypermedia, map interfaces and combinations of these.
- Integration of machine-learning techniques for pattern recognition with interactive user-defined rules to improve models of biological processes.
- Simulation, predictive modelling and visualization (2D, 3D, animations) of cellular processes using high performance parallel processing techniques.
- Semantic web/grid services - semantic descriptions of parallel, distributed web/grid services that will enable automatic scheduling, distribution, choreography and sequencing of computationally intensive pre-processing and analytical services and pipelines to accomplish complex scientific tasks.
- Large-scale highly dimensional data aggregation and reduction. Because of the size and highly dimensional nature of a significant proportion of the data, a key prerequisite is to aggregate and reduce the size and dimensionality of the data in order to enable speedy analysis and visualization without loss of vital information. Parallel processing methods are required to do this efficiently or dynamically.
- Mechanisms are required to both detect and handle new, contradictory, redundant, incomplete, uncertain or inconsistent information and to alert the appropriate agents (human or machine)
- Semantic annotations (based on domain-specific ontologies) of visualizations - enabling their retrieval and re-use for further knowledge mining.

References

- [1] Jane Hunter. ["Adding Multimedia to the Semantic Web - Building an MPEG-7 Ontology"](#) International Semantic Web Working Symposium (SWWS). Stanford. July 2001.
- [2] Jane Hunter. ["Enhancing the Semantic Interoperability of Multimedia through a Core Ontology"](#) IEEE Transactions on Circuits and Systems for Video Technology, Special Issue on Conceptual and Dynamical Aspects of Multimedia Content Description. February 2003.
- [3] Jane Hunter and Suzanne Little. "A Framework to enable the Semantic Inferencing and Querying of Multimedia Content" International Journal of Web Engineering and Technology (IJWET) Special Issue on the Semantic Web. *to appear 2005*
- [4] Suzanne Little and Jane Hunter, ["Rules-By-Example - a Novel Approach to Semantic Indexing and Querying of Images"](#), 3rd International Semantic Web Conference (ISWC2004). Hiroshima, Japan, November 2004.
- [5] Jane Hunter, Katya Falkovych and Suzanne Little. ["Next Generation Search Interfaces - Interactive Data Exploration and Hypothesis Testing"](#) 8th European Conference on Digital Libraries (ECDL2004). Bath, UK, September 2004.