Systems biology

Illoura™: a software tool for analysis, visualization and semantic querying of cellular and other spatial biological dataTim McComb^{1,2,†}, Oliver Cairncross^{1,2,†}, Andrew B. Noske^{1,2}, David L. A. Wood^{1,2}, Brad J. Marsh^{1,2,3} and Mark A. Ragan^{1,2,*}¹Institute for Molecular Bioscience, ²ARC Centre of Excellence in Bioinformatics and ³Centre for Microscopy and Microanalysis, The University of Queensland, Brisbane, Queensland 4072, Australia

Received on December 14, 2008; revised and accepted on February 28, 2009

Advance Access publication March 3, 2009

Associate Editor: Dmitrij Frishman

ABSTRACT

Summary: New high-resolution approaches for mapping ultra-structure of cells in 3D are leading to unprecedented quantities of spatial data. Here we present Illoura, a software tool for the integrated management, analysis and visualization of these data within a semantic context, and illustrate its capability by analysis of spatial relationships in mammalian beta cells.

Availability: <http://www.visiblecell.com/illoura>

Contact: m.ragan@uq.edu.au

Supplementary information: Supplementary data are available at *Bioinformatics* online.

1 INTRODUCTION

One of the most important challenges facing modern bioinformatics is the reconstruction and analysis of biomolecular interaction networks in cells. These networks are often represented as graphs in which the vertices represent biomolecules (e.g. proteins) and the edges depict functional relationships (catalysis, signaling). In reality, these networks do not exist in abstract coordinate space, but instead are physically embedded in subcellular structures (macromolecular complexes, membranes, organelles). Localization contributes to the emergence of function, e.g. by bringing together or separating different sets of biomolecules and thus differentially facilitating or constraining different interaction subgraphs. These structures, in turn, are often arrayed within the cell in characteristic, spatially complex ways that bear strong quantitative implications for the parameterization of dynamic models, e.g. of transport or diffusion. Thus a vertex should be described not only by the identity and inherent properties of the corresponding biomolecule (e.g. molecular weight, solubility), but also by the absolute and relative spatio-temporal coordinates of the cellular structure(s) in which that molecule is localized and functions *in vivo* (Bork and Serrano, 2005; Slepchenko *et al.*, 2002).

Properties and relationships can be described and organized by use of appropriate ontologies, e.g. hierarchies in GO (The Gene Ontology Consortium, 2000). Ontologies moreover support semantically based queries, especially when a framework such

as SPARQL (Prud'hommeaux and Seaborne, 2008) or SWRL (Horrocks *et al.*, 2004) is used to extract or infer relationships. The potential for knowledge discovery by these means is limited, however, by the requirement that all relevant data must be presented in a semantic format. This may be unreasonable or impossible, as in the case above: subcellular structures (*in sensu* sets of coordinates in 3D) are often topologically complex, do not have a native semantics, and furthermore bear innumerable spatial relationships of potential relevance (proximity, path reachability) with each other, not all of which can possibly be anticipated, calculated and instantiated in advance. On the other hand, relationships can often be calculated quickly when and if needed, and made available for inference.

Here we describe such a strategy and present a Java application, which we call Illoura™, which addresses the challenges of analysis, visualization and querying over spatial biological data. Illoura integrates and builds upon existing open-source technologies for handling and querying semantic data, visualizing 2D and 3D spatial data, and evaluating functional expressions. All data in the model are transformed into semantic format for integration, analysis and querying within this common environment. Queries can be issued against ontologies for which some of the underlying data are absent or incomplete but calculable; those parts of the ontology required to satisfy the query are then created based on analysis of the existing data. In this way, the amount of analytical work required is bounded by the question being posed rather than by the size of the model, and the output from any analysis is immediately embedded in a semantic context, and can persist as such.

2 OVERVIEW OF ILLOURA

Although described here in reference to biomolecular and cell-structural data, Illoura broadly supports the organization and integrated analysis of scientific data presented in Resource Description Framework (RDF) (Manola and Miller, 2004). The Illoura data format is publicly available, making the system applicable to a wide range of spatial data. Objects in our example model (derived by segmentation of the 3D density distribution of a mammalian insulin-producing beta cell reconstructed *in toto* at ~15–20 nm resolution: Noske *et al.*, 2008) are imported as IMOD (Kremer *et al.*, 1996) files to Illoura's data format using an import utility of wide potential applicability.

*To whom correspondence should be addressed.

†The authors wish it to be known that, in their opinion, the first two authors should be regarded as joint First Authors.

Fig. 1. This screenshot shows (A) the objects of the model, organized hierarchically; (B) the properties ontologically associated with an object, in this case the plasma membrane; (C) the Lisp/SPARQL environment for interrogating the data; (D) the result of visualizing the granules and the plasma membrane; (E) histogram of the minimum distance of each granule from its nearest point on the plasma membrane, generated from the Lisp/SPARQL environment, for distinct cell models 1 and 2 (cells at different states of insulin-granule discharge); and (F) the result of visualizing one bin of the histogram relating to a single cell model (i.e. all granules that are within one bin-size, here 100 nm, of the same distance from the plasma membrane).

Large quantities of data not conveniently presented in RDF (e.g. 3D bitmap data, polygonal meshes) are stored separately and referenced via URLs. Specific properties can be associated with objects using the Illoura interface; these are arbitrary, but Illoura provides some native support for interpreting properties that relate to an object's visual representation (e.g. label, color, opacity, orientation) and for relationships that link objects with external ontologies, e.g. GO.

Manipulation and visualization of the model dataset is driven by SPARQL construct queries; the results of a query introduce or override existing properties of objects. Queries support embedded calls to Lisp functions, and functions are evaluated as soon as all of their input variables can be bound to values. As such, functions can constrain the search of the underlying model, significantly saving both time and storage space. Application of SPARQL and the Lisp/SPARQL environment in querying cell models that incorporate spatial data is described further in the Supplementary Material.

We have fully integrated Illoura with SRS (Etzold *et al.*, 1996, 2003) such that model data are readily linked with federated data sources; data can be analyzed with respect to SRS associations and visualized appropriately. Protein sequences and structures, pathways, subcellular locations, and any other data federated within SRS can be dynamically queried from within SPARQL via a plug-in extension; more generally, this illustrates the potential to increase capability and performance by distributing spatial and ontological

analyses to remote services on high-performance hardware. Basic charting (histogram, scatter plot) and tabulation facilities are built-in, in addition to 2D and 3D visualization and a plug-in interface that provides users with an option to extend the environment with custom methods for rendering data.

The screenshot (Fig. 1) shows views of models of two pancreatic beta cells reconstructed in 3D from electron tomographic data (Noske *et al.*, 2008). Our biological question centers on the mechanism by which insulin-granule contents are packaged at the Golgi complex, then transported through the cytoplasm to dock with the plasma membrane. We formulate a simple query to analyze and compare the spatial distributions of insulin granules in the two cell models (specifically, their distances from the plasma membrane) and display them in a histogram (Fig 1E). We currently calculate distributions on-the-fly in response to a user query, but intend to support precalculation of selected metrics. Figure 1D shows all the granules and the membrane. Figure 1F, showing only the set of granules within a specific distance range, was requested simply by dragging the corresponding bar from the histogram.

3 CONCLUSION

Spatial relationships are integral to understanding many biological systems, and there is a corresponding need for capable software environments that support the analysis, visualization and querying

of biological data in 2D, 3D or 4D. For example, PointCloudXplore (PCX: Rübel *et al.*, 2008) has been developed to visualize gene expression, e.g. in *Drosophila* embryos, in 3D and 4D; and extensions to the Visualization Toolkit VTK (Kitware, 2008) have been applied to integrate and visualize cell-lineage and gene-expression data (Cedilnik *et al.*, 2008).

Tools and techniques exist as well for applying inference rules to semantically annotated data, given the context of a query (Özacar *et al.*, 2007; Stuckenschmidt and Broekstra, 2005). Illoura combines key elements of both: it has been designed for application to complex biological data in native spatial context, but with the power and flexibility of semantic technologies for data management and querying. Through sourcing and importation from outside the boundary of the model, data of arbitrary complexity can be supplied in the required semantic format only as needed. Illoura thus enables systems biology in a spatial (cellular) context.

Development of the Illoura environment is ongoing, with immediate goals including (i) implementation of precomputation and storage of selected spatial metrics, particularly surface-area maps, organelle volumes and inter-organelle Euclidean distances and results of user queries on the data; (ii) ability to cache results of spatial queries, thereby improving response times in many common use cases; (iii) more capable import functionality for image data from 3D and 4D fluorescence confocal microscopy of cells, including TIFF stacks and vendor-specific formats; and (iv) greater export functionality for graphics, e.g. in JPEG and TIFF formats.

ACKNOWLEDGEMENTS

The authors acknowledge the facilities of Queensland Facility for Advanced Bioinformatics.

Funding: Australian Research Council (grant CE0348221); Australian Postgraduate Award (to A.B.N.); Juvenile Diabetes Research Foundation International (grant 2-2004-275); National Institutes of Health (grant DK-71236 to B.J.M).

Conflict of Interest: none declared.

REFERENCES

- Bork,P. and Serrano,L. (2005) Towards cellular systems in 4D. *Cell*, **121**, 507–509.
- Cedilnik,A. *et al.* (2008) Integration of information and volume visualization for analysis of cell lineage and gene expression during embryogenesis. In *Proceedings of SPIE Visualization and Data Analysis* 6809. Available at <https://wiki.med.harvard.edu/pub/SysBio/Megason/Publications/cedilnik2008.pdf> (last accessed date March 17, 2009).
- Etzold,T. *et al.* (1996) SRS: information retrieval system for molecular biology data banks. *Meth. Enzymol.*, **266**, 114–128.
- Etzold,T. *et al.* (2003) SRS: an integration platform for databanks and analysis tools in bioinformatics. In Lacroix,Z. and Critchlow,T. (eds) *Bioinformatics: Managing Scientific Data*. Morgan Kaufmann, San Francisco, pp. 109–145.
- Horrocks,I. *et al.* (2004) SWRL: a semantic web rule language combining OWL and RuleML. Available at <http://www.w3.org/Submission/SWRL> (last accessed date March 17, 2009).
- Kitware Inc. (2008) VTK visualization toolkit. Available at <http://www.vtk.org> (last accessed date March 17, 2009).
- Kremer,J.R. *et al.* (1996) Computer visualization of three-dimensional image data using IMOD. *J. Struct. Biol.*, **116**, 71–76.
- Manola,F. and Miller,E. (eds) (2004) RDF primer (W3C recommendation 10 February 2004). Available at <http://www.w3.org/TR/REC-rdf-syntax/> (last accessed date March 17, 2009).
- Noske,A.B. *et al.* (2008) Expedited approaches to whole cell electron tomography and organelle mark-up *in situ* in high-pressure frozen pancreatic islets. *J. Struct. Biol.*, **161**, 298–313.
- Özacar,T. *et al.* (2007) A pragmatic approach for RDFS reasoning over large scale instance data. *Lect. Notes Comput. Sci.*, **4806**, 1155–1164.
- Prud'hommeaux,E. and Seaborne,A. (eds) (2008) SPARQL query language for RDF (W3C recommendation 15 January 2008). Available at <http://www.w3.org/TR/rdf-sparql-query/> (last accessed date March 17, 2009).
- Rübel,O. *et al.* (2008) PointCloudXplore 2: visual exploration of 3D gene expression. Lawrence Berkeley National Laboratory Paper LBNL-249E. Available at <http://repositories.cdlib.org/lbnl/LBNL-249E/> (last accessed date March 17, 2009).
- Slepchenko,B.M., *et al.* (2002) Computational cell biology: spatiotemporal simulation of cellular events. *Annu. Rev. Biophys. Biomol. Struct.*, **31**, 423–441.
- Stuckenschmidt,H. and Broekstra,J. (2005) Time-space trade-offs in scaling up RDF schema reasoning. *Lect. Notes Comput. Sci.*, **3807**, 172–181.
- The Gene Ontology Consortium (2000) Gene Ontology: tool for the unification of biology. *Nature Genet.*, **25**, 25–29.