BioPatML.NET – A bioinformatic pattern search engine

Samuel Yu Toh[1], Lawrence Buckingham[1], James M Hogan[1*], Stefan R Maetschke[2], Michael W Towsey[1]
[1]Queensland University of Technology, Brisbane, Australia
[2]University of Queensland, Brisbane, Australia
[*]Corresponding author: j.hogan@qut.edu.au

INTRODUCTION

The exponential growth of genomic data sets in recent years presents numerous analytical challenges to biologists seeking to identify functional motifs, particularly those for which the ubiquitous BLAST lacks sufficient precision. The emerging trend toward population based analyses makes it all the more pressing to support a broad range of complex pattern searches and to allow sharing, annotation and re-use. Existing bioinformatic pattern definition languages largely fall into two groups, namely those best described as extended regular expressions, and those that implement a form of stochastic grammar. In general, regular expressions are sufficient to solve many pattern searching problems. However their expressive power is insufficient to model complex structured pattern such as promoters, overlapping motifs or RNA stem–loops. In addition, the majority of existing languages offer minimal support for advanced search functionality like mismatch thresholds, weighted gaps, direct and inverted repeats, general similarity scoring and position weight matrices.

BioPatML [1] is a comprehensive bioinformatic pattern definition language and search engine that addresses the deficiencies of earlier systems, providing support for a broad range of atomic pattern types and a rich grammar that enables the construction of complex patterns via sequential and hierarchical aggregation of pattern fragments. In line with the increasing need for computational biologists to share sophisticated biological data and metadata, BioPatML uses a standard XML representation with a well-defined schema to represent pattern definitions and annotations.

In the present paper, we introduce BioPatML.NET, an application library for the Microsoft Windows .NET framework [2] that implements the BioPatML pattern definition language and sequence search engine. BioPatML.NET is integrated with the Microsoft Biology Foundation (MBF) application library [3], unifying the parsers and annotation services supported or emerging through MBF with the language, search framework and pattern repository of BioPatML. End users who wish to exploit the BioPatML.NET engine and repository without engaging the services of a programmer may do so via the freely accessible web-based BioPatML Editor, which we describe below.

BIOPATML LANGUAGE OVERVIEW

This section provides a brief overview of BioPatML; a detailed description of the language may be found in [1]. BioPatML is an XML based language which provides a common representation for structured patterns equally useful in pattern databases and pattern search. It allows precise specification of patterns that cannot be represented accurately by any simpler representation such as regular expressions or position weight matrices. A simple motivating example is the consensus description of the σ^{70} promoter in *E. coli*, which consists of two hexamers, **`ttgaca`** and **`tataat`**, separated by a gap with variable length containing arbitrary letters. A BioPatML description of this pattern consists of a simple hierarchical structure: a **`SERIES`** element represents the entire pattern; within this are nested, in sequence, a **`MOTIF`** element representing the first hexamer, a **`GAP`** element and a **`MOTIF`** element representing the second hexamer.

A BioPatML pattern definition contains two fundamental types of node. Primitive nodes represent such things as motifs and sequences (literal text), regular expressions, position weight matrices and aligned sequences. Introduction of pattern aggregation nodes then permits the development of a very rich modelling syntax. Supported structures include set, series, iteration, repeat and logic (disjunction and conjunction of pattern matches).

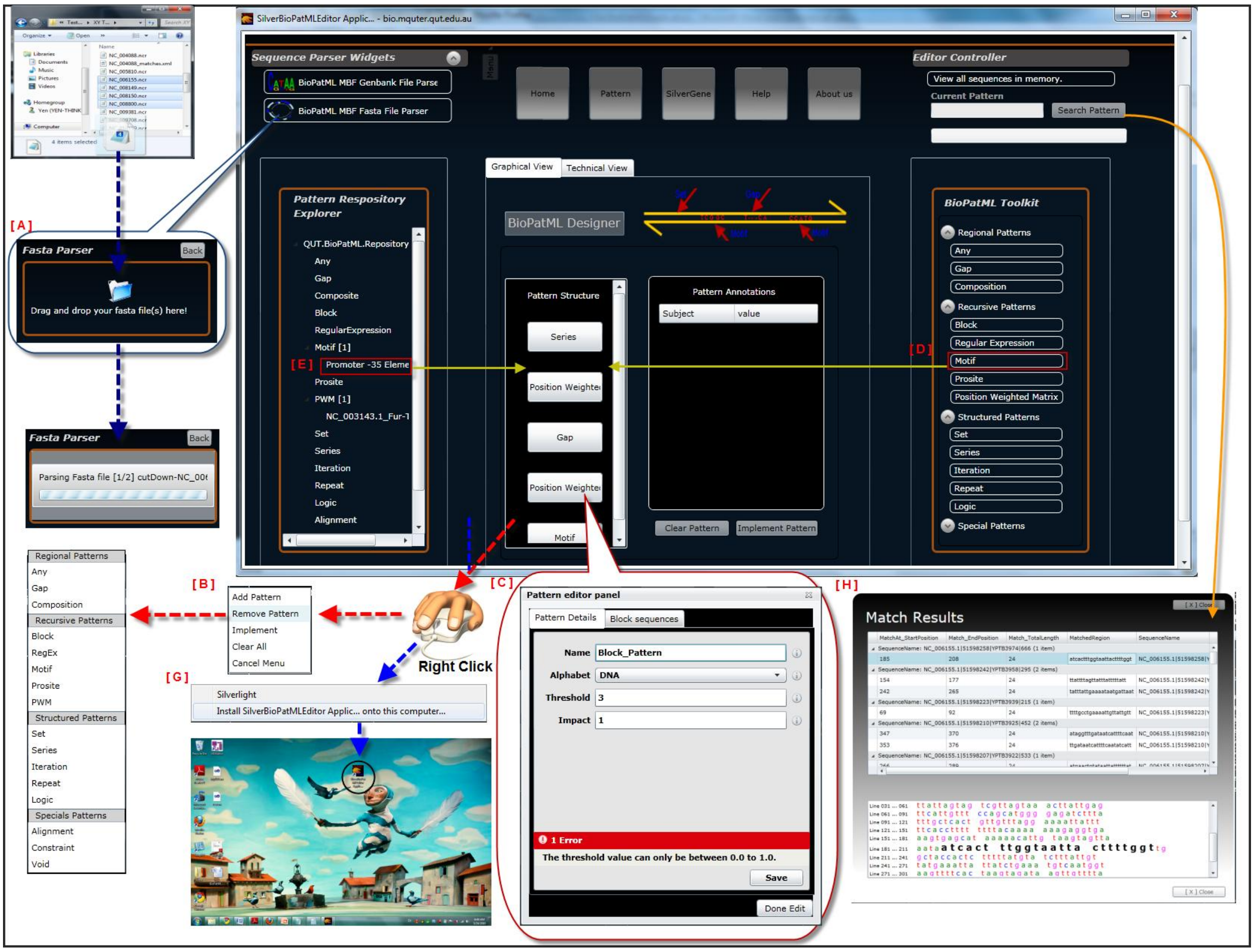
THE BIOPATML EDITOR

The editor provides a graphical user interface running under Microsoft Silverlight [4] through which a user may:

- Load and parse sequence data;
- Create, edit and annotate pattern descriptions;
- Search sequences to locate pattern occurrences;
- Visualize and export search results in a range of formats; and
- Save and share pattern descriptions with fellow researchers.

The editor takes advantage of facilities introduced in version 4 of Silverlight, including streamlined access to the local file system and the ability to install the application to the desktop, enabling a certain level of offline functionality.

Figure 1 (below) illustrates several of the more prominent features of the editor, which operates on application conventions familiar to users of Windows or other GUI-based environments. Sequence data files may be parsed by dragging them from Windows Explorer and dropping them onto the MBF parser panel of the editor (Figure 1 #A). BioPatML patterns are constructed by dragging pattern fragments onto the design surface; once a fragment has been added to the pattern under construction, its properties may be updated via the pattern configuration window (Fig 1 #C).



Name Block_Pattern

Alphabet DNA

Threshold 3

Impact 1

1 Error

The threshold value can only be between 0.0 to 1.0.

Figure 1: Major features of the BioPatML Editor

The editor provides two sets of pattern fragments for this purpose. The toolbox (Figure 1 #D) provides empty pattern fragments corresponding to each of the BioPatML node types; these may be used to construct patterns from scratch, or combined with patterns obtained from the repository (Figure 1 #E) to build extended patterns based on previously defined patterns. This makes our vision of a biological pattern community accessible and practical, enabling users to repeat and confirm the analyses of previous authors and to incorporate their patterns as components in more sophisticated searches. Once a search has been executed the results may be viewed in a track-based sequence browser, SilverGene [5], or displayed in a dynamic rich text box (Figure 1 #H). The user may export the result to any word processing tool through familiar copy and paste operations.

CONCLUSION AND FUTURE WORK

The grammar of BioPatML, which supports the hierarchical combination of existing pattern definitions, allows researchers to rely on shared patterns as a basis for more complex pattern structures. In this way, the community may encode and share each improvement in its understanding of complex motif relationships which underpin binding and transcriptional regulation, through the use of the BioPatML Editor. Work in the near future for BioPatML will include further exploration of other pattern languages. We also aim to further optimize the processing speed for both the BioPatML.NET and its editor, and to make the browser BioPatML Editor work purely on the local machine independent of internet connectivity.

Availability:

- Editor: http://bio.mquter.qut.edu.au/BioPatML2010/SilverBioPatMLEditorTestPage.aspx
- Video: http://bio.mquter.qut.edu.au/BioPatML2010/MyBioPatMLDemo.html

REFERENCES

1. Maetschke, S.R., Towsey, M.W., Hogan, J.M., *BioPatML - an XML description language for patterns in biological sequences*, QUT FaST Technical Report, 2007, available from http://eprints.qut.edu.au/7730, accessed 29 June 2010.
2. Microsoft .NET Framework: http://www.microsoft.com/net/, accessed 29 June 2010.
3. The Microsoft Biology Foundation: http://research.microsoft.com/en-us/collaboration/tools/mbf.aspx, accessed 29 June 2010.
4. Microsoft Silverlight: http://www.silverlight.net/, accessed 29 June 2010.
5. Microsoft QUT e-Research Centre web site: http://www.mquter.qut.edu.au/bio , accessed 29 June 2010.