

# BioPatML - Pattern Sharing for the Genomic Sciences

Stefan Maetschke, Michael Towsey and James M. Hogan

Microsoft QUT eResearch Centre  
Queensland University of Technology  
GPO Box 2434, Brisbane, QLD, 4001  
AUSTRALIA  
[j.hogan@qut.edu.au](mailto:j.hogan@qut.edu.au)

## Abstract

Computational biology increasingly demands the sharing of sophisticated data and annotations between research groups. Web 2.0 style sharing and publication requires that biological systems be described in well-defined, yet flexible and extensible formats which enhance exchange and re-use. In contrast to many of the standards for exchange in the genomic sciences, descriptions of biological sequences show a great diversity in format and function, impeding the definition and exchange of sequence patterns. In this presentation, we introduce BioPatML, an XML-based pattern description language that supports a wide range of patterns and allows the construction of complex, hierarchically structured patterns and pattern libraries. BioPatML unifies the diversity of current pattern description languages and fills a gap in the set of XML-based description languages for biological systems. We discuss the structure and elements of the language, and demonstrate its advantages on a series of applications, showing lightweight integration between the BioPatML parser and search engine, and the SilverGene genome browser. We conclude by describing our site to enable large scale pattern sharing, and our efforts to seed this repository.

## 1. Introduction

Modern computational biology requires standardized, machine-readable descriptions of biological systems to enable the computational evaluation and the unhampered exchange of biological data. Many XML-based formats have been introduced to support the exchange of genomic data, each benefitting from the range of parsing and transformation technologies available for XML. BioPatML is an XML-based pattern description language intended to provide a unifying format for the sharing of biological patterns. Most importantly, 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, a community may encode and share each improvement in its understanding of the complex motif relationships which underpin binding and transcriptional regulation.

Existing bioinformatic pattern description languages are best seen as subsets or minor extensions of regular expressions, and offer limited modelling power. Good examples of this type are Kangaroo [1], which extends plain regular expressions to allow IUPAC ambiguity symbols, and PatMatch [2]. Many other pattern languages exist, and may be categorized as Extended Regular Expressions or Stochastic Grammars, such as Profile Hidden Markov Models. A number of specializations of these approaches were developed for particular domains, with PROSITE [3] and PepPat [4] being typical examples. BioPatML is more closely aligned to PatSearch [5], but permits more general pattern definitions. These approaches vary in their support beyond the basic regular expression matches, but BioPatML provides a superset of their functionality, including mismatch thresholds, weighted gaps, direct and inverted repeats, overlapping patterns, a general similarity scoring, position weight matrices, and pattern aggregation through sets and series operators.

A more detailed description of the language, and the sharing site described below may be found by following the links from <http://www.mquter.qut.edu.au/bio>.

## 2. Example Patterns

We here illustrate the use of the language through the problem of predicting transcription start sites (TSS) in bacteria. Bacterial gene transcription is initiated when an enzyme complex, consisting of the RNA polymerase and a sigma factor, binds to a promoter upstream of the actual transcription start site. The canonical model of a sigma-70 promoter specifies two hexamers, separated by a variable gap of approximately 15-21bp, and located near positions -35 and -10 relative to the TSS. The consensus sequence for the -35 motif is TTGACA, and that for the -10 motif is TATAAT. The classical promoter model for *E. coli* shows an optimal gap length of 17bp between the two hexamers (we refer to this gap as the spacer) and a distance of 7bp between the -10 motif and the TSS (we refer to this gap as the discriminator). Patterns of this form are readily described in BioPatML as shown below, through a combination of PWMs, weighted Gaps, and a motif. The Series element is used to describe an ordered sequence of sub-patterns that may be composed of other patterns itself.

```
<Series mode="BEST" threshold="0.0">
  <PWM name="-35element" alphabet="DNA" threshold="0.0">
    <Row letter="a">-1.5 -2.3 -1.5 0.8 -0.8 0.9</Row>
    <Row letter="g">-1.5 -1.8 1.3 -0.6 -1.1 -0.0</Row>
    <Row letter="c">-0.8 -1.6 -0.5 -0.7 1.3 -1.0</Row>
    <Row letter="t"> 1.2 1.4 -0.5 -0.4 -0.5 -1.0</Row>
  </PWM>
  <Gap name="Spacer" minimum="15" maximum="21" impact="0.2">
    <Weights>
      0.15 0.16 0.20 0.16 0.09 0.12 0.11
    </Weights>
  </Gap>
  <PWM name="-10element" alphabet="DNA" threshold="0.0">
    <Row letter="a">-2.2 1.3 -0.8 1.1 1.2 -2.1</Row>
    <Row letter="g">-1.2 -1.1 -0.8 -0.5 -1.4 -2.0</Row>
    <Row letter="c">-0.5 -2.1 -0.9 -1.2 -0.4 -1.5</Row>
    <Row letter="t"> 1.2 -1.3 1.0 -1.1 -1.4 1.4</Row>
  </PWM>
  <Gap name="Discriminator" minimum="4" maximum="12" impact="0.2">
    <Weights>
      0.05 0.11 0.27 0.22 0.09 0.05 0.08 0.06 0.08
    </Weights>
  </Gap>
  <Motif name="TSS" impact="0.2" alphabet="DNA" motif="R" threshold="0.0"/>
</Series>
```

## 3. Conclusion

BioPatML increases the power of classical pattern definition languages through principled aggregation. It simplifies the compilation of pattern libraries and promotes exchange of complex patterns. The language provides a convenient format to encapsulate pattern definitions and their annotations for integrated bioinformatic analyses. We have attempted to foster this culture of sharing through the provision of a BioPatML pattern exchange, a site at which researchers may contribute new patterns and tag existing entries, increasing the utility of the database based on their experience. Searching is supported based on pattern name and tags, and the existing BioPatML parser and search engine have recently been integrated with the SilverGene genomic visualiser, allowing interactive exploration of the sequence and visualisation of matches returned.

## 4. References

- [1] Betel D, Hogue CWV: Kangaroo—a pattern-matching program for biological sequences. *BMC Bioinformatics* 2002, 3:20.
- [2] Yan T, Yoo D, Berardini TZ, Mueller LA, Weems DC, Weng S, Cherry JM, Rhee SY: PatMatch: a program for finding patterns in peptide and nucleotide sequences. *Nucleic Acids Res* 2005, 33(Web Server issue):W262–W266, [<http://dx.doi.org/10.1093/nar/gki368>].

[3] Hulo N, Sigrist CJA, Saux VL, Langendijk-Genevaux PS, Bordoli L, Gattiker A, Castro ED, Bucher P, Bairoch A: Recent improvements to the PROSITE database. *Nuc Acid Res* 2004, 32(Database issue):D134–D137, [<http://dx.doi.org/10.1093/nar/gkh044>].

[4] Jiang Y, Gao G, Fang G, Gustafson EL, Laverty M, Yin Y, Zhang Y, Luo J, Greene JR, Bayne ML, Hedrick JA, Murgolo NJ: PepPat, a pattern-based oligopeptide homology search method and the identification of a novel tachykinin-like peptide. *Mamm Genome* 2003, 14(5):341–9, [<http://dx.doi.org/10.1007/s00335-002-3061-y>].

[5] Grillo G, Licciulli F, Liuni S, Sbis`a E, Pesole G: PatSearch: A program for the detection of patterns and structural motifs in nucleotide sequences. *Nucleic Acids Research* 2003, 31(13):3608–3612.