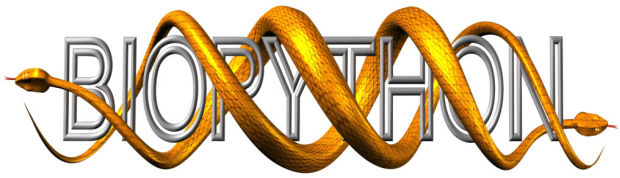

Biopython Project Update

Brad Chapman*, Peter Cock†, *et al.*
chapmanb@50mail.com

Bioinformatics Open Source Conference (BOSC) 2010, Boston, MA, USA

In this talk we present the current status of the Biopython project (www.biopython.org), described in a application note published last year (Cock *et al.*, 2009). Biopython celebrated its 10th Birthday last year, and has now been cited or referred to in over 150 scientific publications (a list is included on our website).

At the end of 2009, following an extended evaluation period, Biopython successfully migrated from using CVS for source code control to using git, hosted on github.com. This has helped our existing developers to work and test new features on publicly viewable branches before being merged, and has also encouraged new contributors to work on additions or improvements. Currently about fifty people have their own Biopython repository on GitHub.

In summer 2009 we had two Google Summer of Code (GSoC) project students working on phylogenetic code for Biopython in conjunction with the National Evolutionary Synthesis Center (NESCent). Eric Talevichs work on phylogenetic trees including phyloXML support (Han and Zamesk, 2009) was merged and included with Biopython 1.54, and he continues to be actively involved with Biopython. We hope to include Nick Matzkes module for biogeographical data from the Global Biodiversity Information Facility (GBIF) later this year. For summer 2010 we have Biopython related GSoC projects submitted via both NESCent and the Open Bioinformatics Foundation (OBF), and if all goes well we will again have summer students working on Biopython.

Since BOSC 2009, Biopython has seen four releases. Biopython 1.51 (August 2009) was an important milestone in dropping support for Python 2.3 and our legacy parsing infra-structure (Martel/Mindy), but was most noteworthy for FASTQ support (Cock *et al.*, 2010). Biopython 1.52 (September 2009) introduced indexing of most sequence file formats for random access, and made interconverting sequence and alignment files easier. Biopython 1.53 (December 2009) included wrappers for the new NCBI BLAST+ command line tools, and much improved support for running under Jython. Our latest release is Biopython 1.54 (May 2010), new features include Bio.Phylo for phylogenetic trees (GSoC project), and support for Standard Flowgram Format (SFF) files used for 454 Life Sciences (Roche) sequencing.

Biopython is free open source software available from www.biopython.org under the Biopython License Agreement (an MIT style license, `http://www.biopython.org/DIST/LICENSE`).

References

Cock, P.J.A., Antao, T., Chang, J.T., Chapman, B.A., Cox, C.J., Dalke, A., Friedberg, I., Hamelryck, T., Kauff, F., Wilczynski, B., de Hoon, M.J. (2009) Biopython: freely available Python tools for computational molecular biology and bioinformatics. *Bioinformatics* **25**(11) 1422-3. doi:10.1093/bioinformatics/btp163

Han, M.V. and Zmasek, C.M. (2009) phyloXML: XML for evolutionary biology and comparative genomics. *BMC Bioinformatics* **10**:356. doi:10.1186/1471-2105-10-356

Cock, P.J.A., Fields, C.J., Goto N., Heuer, M.L., and Rice, P.M. (2010) The Sanger FASTQ file format for sequences with quality scores, and the Solexa/Illumina FASTQ variants. *Nucleic Acids Res* **38**(6) 1767-71. doi:10.1093/nar/gkp1137

*Bioinformatics Core Facility, Molecular Biology Department, Massachusetts General Hospital, Boston, MA, USA
†Plant Pathology, SCRI, Invergowrie, Dundee DD2 5DA, UK