BioSolr: Building better search for bioinformatics

Tony Burdett[1], Matt Pearce[2], Tom Winch[2], Charlie Hull[2], Helen Parkinson[3] and Sameer Velankar[3]

[1] European Bioinformatics Institute (EMBL-EBI), European Molecular Biology Laboratory, Wellcome Trust Genome Campus, Hinxton, Cambridge CB10 1SD, United Kingdom: tburdett@ebi.ac.uk

[2] Flax, St Johns Innovation Centre, Cowley Road, Cambridge, CB4 0WS, United Kingdom

[3] European Bioinformatics Institute (EMBL-EBI), European Molecular Biology Laboratory, Wellcome Trust Genome Campus, Hinxton, Cambridge CB10 1SD, United Kingdom.

BioSolr Wiki:	http://tinyurl.com/biosolr
Source Code:	https://github.com/flaxsearch/BioSolr
License:	http://www.apache.org/licenses/LICENSE-2.0
Mailing list:	solr-users@ebi.ac.uk

Data retrieval is common in bioinformatics databases, however, optimal strategies for indexing rich biomedical data are less well understood. Biomedical data often contains hierarchical components, such as annotations to ontologies, and therefore do not conform to the flattened document-based model imposed by most search technologies. BioSolr advances the state of the art with regard to indexing and querying biomedical data with open source software. This unique BBSRC funded collaboration between Flax, an open source search specialist company based in Cambridge, and The European Bioinformatics Institute (EMBL-EBI), brings together experts in biological data management and experts in utilising the world-leading Apache Lucene/Solr search engine framework to address the challenges of making biomedical data more accessible. Challenges include integrating ontology-enabled search and searching by common classification systems (taxonomy, enzyme classifications, protein families etc).

BioSolr is developing software to facilitate indexing of ontologies, ontology driven faceting, searching Solr indexes with SPARQL, FASTA Solr search components, and an “x-join” search component to integrate external data resources with Solr. Development is in the form of Solr patches, plugins or clients, and all code developed as part of the BioSolr project is available as open source software on GitHub. In addition, BioSolr is building a wide community of users of Lucene/Solr for bioinformatics via international workshops and the open source search developer community. BioSolr is also working to identify a series of best practices for working with Solr in bioinformatics. Requirements and common usage scenarios across the full spectrum of bioinformatics have been collated and cover a range of domains, including searching over protein structures and sequences, ontologies and literature.