BIOINFORMATICS APPLICATIONS NOTE Vol. 22 no. 5 2006, pages 637–640
doi:10.1093/bioinformatics/btk027

Databases and ontologies

ORegAnno: an open access database and curation system for literature-derived promoters, transcription factor binding sites and regulatory variation

S. B. Montgomery[1,†], O. L. Griffith[1,†], M. C. Sleumer[1], C. M. Bergman[2], M. Bilenky[1], E. D. Pleasance[1], Y. Prychyna[1], X. Zhang[1] and S. J. M. Jones[1,*]

[1]Canada's Michael Smith Genome Sciences Centre, British Columbia Cancer Agency, 100-570 West 7th Avenue, Vancouver, BC, Canada V5Z 4E6 and [2]Faculty of Life Sciences, University of Manchester, Michael Smith Building, Oxford Road, M13 9PT Manchester, UK

Received on October 25, 2005; revised on December 22, 2005; accepted on December 23, 2005
Advance Access publication January 5, 2006
Associate Editor: Nikolaus Rajewsky

ABSTRACT

Motivation: Our understanding of gene regulation is currently limited by our ability to collectively synthesize and catalogue transcriptional regulatory elements stored in scientific literature. Over the past decade, this task has become increasingly challenging as the accrual of biologically validated regulatory sequences has accelerated. To meet this challenge, novel community-based approaches to regulatory element annotation are required.

Summary: Here, we present the Open Regulatory Annotation (ORegAnno) database as a dynamic collection of literature-curated regulatory regions, transcription factor binding sites and regulatory mutations (polymorphisms and haplotypes). ORegAnno has been designed to manage the submission, indexing and validation of new annotations from users worldwide. Submissions to ORegAnno are immediately cross-referenced to EnsEMBL, dbSNP, Entrez Gene, the NCBI Taxonomy database and PubMed, where appropriate.

Availability: ORegAnno is available directly through MySQL, Web services, and online at http://www.oreganno.org. All software is licensed under the Lesser GNU Public License (LGPL).

Contact: sjones@bcgsc.ca

INTRODUCTION

The effectiveness of bioinformatics methods for identifying regulatory regions in genomic sequence is dependent on our understanding of gene regulation biology in its natural state. This is particularly evident in that models of transcription factor binding in regulatory regions have underpinned the development of such bioinformatics methods as phylogenetic footprinting, transcription factor binding matrices and motif clustering (Wasserman and Sandelin, 2004). However, the predictive ability of algorithms which implement these methods has been predominantly indeterminate, as their assessment has relied on datasets containing few biologically validated regulatory regions (Tompa *et al*., 2005). To enrich these datasets, several databases have been designed to independently organize the sites of promoter activity (Grienberg and Benayahu, 2005; Lescot *et al*., 2002; Pohar *et al*., 2004; Schmid *et al*., 2004; Shahmuradov *et al*., 2003; Zhu and Zhang, 1999), transcription factor binding (Bergman *et al*., 2005; Kanamori *et al*., 2004; Kolchanov *et al*., 2002; Matys *et al*., 2003) and regulatory variation (Stenson *et al*., 2003; Tahira *et al*., 2005; Zhao *et al*., 2004). Several challenges face the user when accessing these databases for the annotation of biologically validated regulatory regions. For many databases, considerable investigation can be required to collate its information, determine the original experimental techniques used, determine the 'genomic scope' of the annotation (i.e. what further annotation is in the vicinity and informative), obtain a sequence of sufficient length to map to new genome sequence assemblies, cross-reference or follow-up on specific annotation or access the annotation programmatically. Furthermore, as new regulatory sequences become characterized each database requires its own curators *ad infinitum* as few or no mechanisms currently exist in which a community of researchers can add to or comment on these annotations. The Open Regulatory Annotation (ORegAnno) database has been developed to address these issues and provide a unique platform for community annotation of experimentally verified regulatory regions.

DESCRIPTION OF THE ORegAnno DATABASE

ORegAnno permits the open annotation of regulatory regions by providing roles and secure user accounts to contributors. Three roles exist for ORegAnno contributors: user, validator and administrator. A user role enables a contributor to add individual annotations of promoters, transcription factor binding sites and regulatory mutations to the database. As a first step in validating a new annotation's authenticity, each submitted annotation is immediately

*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.

cross-referenced against PubMed (Wheeler *et al*., 2005), Entrez Gene (Maglott *et al*., 2005), dbSNP (Sherry *et al*., 2001), the NCBI Taxonomy database (Wheeler *et al*., 2005) and EnsEMBL (Hubbard *et al*., 2005). Once submitted, the record is added to the database and an email is generated containing an XML representation of this record to members of the ORegAnno developers' mailing-list (oreganno-guts@bcgsc.ca). As a second step in validating an annotation's authenticity, a validator role enables a contributor to score individual annotations in the database. Validators will modify an overall score for an annotation based on their ability to confirm the reliability of annotation from literature. Validators have the option of increasing the annotation score by one if they can confirm the record, leaving the score unchanged if their conclusions are indeterminate, or decreasing the score by one if an error has been found. Each observation and score modification of an annotation along with the associated validator user information is stored in ORegAnno. An administrator role enables a contributor to assign roles, add or define evidence (classes, types and subtypes) and batch upload large sets of annotations directly to the database. Both administrator and validator roles allow the modification of records; for a record modification, a new record is created and the old record is marked as being deprecated by the newer record. Each role is further permitted to add comments to individual annotations to improve subsequent users' understanding of a particular annotation. ORegAnno's usage of roles provides a level of accountability in the database as users become owners of their annotation and validators become responsible for verifying an annotation's authenticity.

For each type of annotation that is currently in ORegAnno, the database obeys the following rules:

(1) Each annotation describes a regulatory property of one target gene which is either user-defined, in Entrez Gene or in EnsEMBL.

(2) Each annotation must be attributed to a species which has a taxonomy id in the NCBI Taxonomy database.

(3) Each annotation can optionally be associated to a specific dataset. This functionality allows external curators to manage particular sets of annotation using ORegAnno's curation tools.

(4) Each annotation specifies an evidence type, subtype and class describing the biological technique cited to discover the regulatory sequence. Evidence classes are broken into two categories: the 'regulator' classes describe evidence for the specific protein(s) that bind a site. The 'regulatory site' classes describe evidence for the function of a regulatory sequence itself. These two categories are further divided into three levels of regulation (transcription, transcript stability and translation). Thus, a total of six evidence classes currently exist. Evidence types describe the generic assay used while subtypes define specific implementations of these assays (Table 1). Each annotation can have multiple entries from any evidence class, type and subtype describing each piece of experimental evidence for the regulatory sequence and/or binding protein.

(5) Each piece of experimental evidence is optionally associated to a specific cell type using the eVOC cell type ontology (Kelso *et al*., 2003).

Table 1. Evidence types and subtypes

Evidence type	Evidence subtype
Electrophoretic mobility shift assay (EMSA)	Direct gel shift
	Supershift
	Gel shift competition
Reporter gene assay	Transient transfection luciferase assay
	Chloramphenicol acetyltransferase (CAT) assay
	In vivo GFP expression assay
	Dual luciferase reporter gene assay
	In vivo LacZ expression assay
Protein binding assay	Chromatin immunoprecipitation (ChIP)
	DNase footprinting assay
	Yeast 1-hybrid assay
RNA expression assay	RNase protection assay (RPA)
	Reverse transcriptase polymerase chain reaction (RT–PCR)
	Allele-specific transcript quantification (ASTQ)
	Competitive PCR (cPCR)
	RNA ligase-mediated rapid amplification of cDNA ends (RLM_RACE)
	Whole-mount *in situ* hybridization
Protein expression assay	Western blot assay
	Enzyme-linked immunosorbent assay (ELISA)
	Luciferase expression assay
	Indirect Immunofluorescence
RNA stability assay	RNA synthesis blocking
Association study	Resequencing
	Single-stranded conformational polymorphism (SSCP)
	Restriction fragment length polymorphism (rflp) analysis
Orthologous gene conservation	Conservation found by alignment
	Conservation found by scanning with a motif model
Gene co-expression	Co-expressed genes determined through reporter gene experiments
	Co-expressed genes determined through microarray experiments
	Co-expressed genes determined through expression pattern

(6) Each transcription factor binding site or regulatory mutation must specify a target transcription factor which is either user-defined, in Entrez Gene or in EnsEMBL. If there is no recorded gene target, a classification of 'unknown' is specified.

(7) Each transcription factor binding site or regulatory mutation must include sequence with at least 40 bases of flanking genomic sequence to allow the site to be mapped to any release of an associated genome.

(8) Where available, any annotation can provide search space information specifying the region that was assayed, not just the regulatory sequence.

(9) User information is recorded with each annotation.

(10) Each annotation must reference a valid PubMed article. To reduce the entry of redundant annotations, a warning is issued if an annotation is found with either an existing reference identifier or matching genomic sequence.

(11) For regulatory mutations, each variant that has been proven to cause a change in gene expression is a separate record. The sequences containing both the wild-type and mutant sequences must be specified. If available, a dbSNP cross-reference can also be specified. The type of variant is specified as either being germline, somatic or artificial.

(12) Each record is associated to a positive, neutral or negative outcome based on the experimental results from the primary reference. For instance, a sequence that was demonstrated not to bind a particular transcription factor could be annotated as a negative outcome; however, to be meaningful, the associated evidence must provide adequate information to determine the conditions assayed.

ORegAnno comes equipped with analysis tools to assist in annotation of new records. In many cases, extracting genome sequence from literature and identifying the corresponding sequences in genome databases is problematic (Frith *et al*., 2004). ORegAnno provides the tools ENSSCAN for finding one or more specific sequences within distances relative to the start of an EnsEMBL transcript, ENSFETCH for retrieving small sequences within distances relative to the start of an EnsEMBL transcript (i.e. from −34 to −40 of the transcription start site), NCBISCAN for finding one or more specific sequences within defined distances of a GenBank-reference sequence and NCBIFETCH for highlighting small (gapped) sequences within a GenBank-reference sequence.

CURRENT CONTENT OF THE ORegAnno DATABASE

At time of writing, the ORegAnno database housed a total of 2691 entries from over 20 users. These include 780 regulatory regions, 1804 transcription factor binding sites, and 107 regulatory mutations (polymorphisms and haplotypes) from 9 species (Table 2). A large fraction of these sites were obtained from previous large-scale collections such as the FlyReg resource (Bergman *et al*., 2005) and a large set of muscle/liver-specific regulatory sites curated by Wasserman and co-workers (Ho Sui *et al*., 2005; Wasserman and Fickett, 1998). Eleven regulatory polymorphism records were obtained from rSNP_DB (Ponomarenko *et al*., 2001); rSNP_DB records were filtered to include only those records which pertained to natural mutations or polymorphisms. In addition, over 200 new annotations were obtained by manual curation of literature. Thus, the ORegAnno resource represents an assembly of existing records, a significant addition of new records and provides an open-access system for continued, community-based accumulation of sites within a standardized framework.

ACCESS

The raw ORegAnno data are available directly over MySQL from db01.bcgsc.ca or through web services (Booth, 2004). Methods are exported using Web services to search for annotation by various fields enabling fetches by such fields as stable id, species, gene name, transcription factor name or cross-reference sources. ORegAnno also automatically maps each annotation to its relevant genome using Blast (Altschul *et al*., 1990); these mappings are viewable through the UCSC Genome Browser (Kent *et al*., 2002) or EnsEMBL using the Distributed Annotation System (Dowell *et al*., 2001). Finally, the entire database is converted to XML format and made available on the website daily. The ORegAnno web application is open-source under the Lesser GNU Public Licence thereby permitting all forms of modification and mirroring.

Table 2. Current content of ORegAnno database

	Regulatory haplotype	Regulatory polymorphism	Regulatory region	Transcription factor binding site
Caenorhabditis briggsae	0	0	0	24
Caenorhabditis elegans	0	0	8	117
Danio rerio	0	0	2	0
Drosophila melanogaster	0	0	0	1331
Gallus gallus	0	0	0	13
Homo sapiens	4	103	765	196
Mus musculus	0	0	1	87
Rattus norvegicus	0	0	4	35
Xenopus tropicalis	0	0	0	1
Totals	4	103	780	1804

CONCLUSIONS

The ORegAnno resource represents the first open-access, community-based forum for annotation of regulatory sequences. ORegAnno is currently the largest collection of functionally validated regulatory annotations available with unrestricted access. To our knowledge, it is the first resource to incorporate regulatory regions, binding sites and variation into a single resource. It is also the first system to incorporate a structured system for experimental evidence and allow both negative and positive results. The requirements for sufficient flanking sequence and verified gene identifiers (Ensembl or Entrez) ensure maximum compatibility with the community's various research needs, both currently and in the future. The intention of ORegAnno is not to replace any regulatory element databases. Many of the well-targeted databases have domain- or species-specific information that would be impractical to incorporate into a single resource. Instead, we hope to create a single multi-species database and curation system for some of the most essential information (target gene, binding protein, binding site sequence, etc.). Thus, we believe ORegAnno should exist in collaboration with the more specific databases as a central warehouse of data, with the ultimate goal of incorporating all experimentally verified regulatory annotation. We anticipate that this growing library of regulatory elements will prove an important resource for the validation of computational methods of motif detection, investigations of regulatory element evolution and an

essential resource for the appraisal and validation of genome-wide regulatory predictions (Robertson *et al*., 2006; Xie *et al*., 2005).

ACKNOWLEDGEMENTS

We would like to acknowledge the Wasserman lab (http://www.cisreg.ca/tjkwon/) and James Fickett (http://www.cbil.upenn.edu/MTIR/HomePage.html) for generously making their regulatory element catalogues publicly available. We thank the ORegAnno users for their continuing efforts to improve this resource through manual curation and record validation. We gratefully acknowledge funding from Genome Canada, Genome British Columbia and the BC Cancer Foundation. S.B.M. was supported by the Natural Sciences and Engineering Research Council (NSERC) and the Michael Smith Foundation for Health Research (MSFHR). O.L.G. was supported by the Canadian Institutes of Health Research (CIHR), NSERC and MSFHR. E.D.P. was supported by CIHR. M.C.S. and S.J.M.J. were supported by MSFHR.Funding to pay the Open Access publication charges for this article was provided by Genome Canada.

Conflict of Interest: none declared.

REFERENCES

Altschul,S.F. *et al*. (1990) Basic local alignment search tool. *J. Mol. Biol*., **215**, 403–410.

Bergman,C.M. *et al*. (2005) *Drosophila* DNase I footprint database: a systematic genome annotation of transcription factor binding sites in the fruitfly, *Drosophila melanogaster*. *Bioinformatics*, **21**, 1747–1749.

Booth,D., Haas,H., McCabe,F., Newcomer,E., Champion,M., Ferris,C. and Orchard,D. (2004) Web Services architecture, *W3C working group note, W3C*.

Dowell,R.D. *et al*. (2001) The distributed annotation system. *BMC Bioinformatics*, **2**, 7.

Frith,M.C. *et al*. (2004) Site2genome: locating short DNA sequences in whole genomes. *Bioinformatics*, **20**, 1468–1469.

Grienberg,I. and Benayahu,D. (2005) Osteo-Promoter Database (OPD)—promoter analysis in skeletal cells. *BMC Genomics*, **6**, 46.

Ho Sui,S.J. *et al*. (2005) oPOSSUM: identification of over-represented transcription factor binding sites in co-expressed genes. *Nucleic Acids Res*., **33**, 3154–3164.

Hubbard,T. *et al*. (2005) Ensembl 2005. *Nucleic Acids Res*., **33**, D447–D453.

Kanamori,M. *et al*. (2004) A genome-wide and nonredundant mouse transcription factor database. *Biochem. Biophys. Res. Commun*., **322**, 787–793.

Kelso,J. *et al*. (2003) eVOC: a controlled vocabulary for unifying gene expression data. *Genome Res*., **13**, 1222–1230.

Kent,W.J. *et al*. (2002) The human genome browser at UCSC. *Genome Res*., **12**, 996–1006.

Kolchanov,N.A. *et al*. (2002) Transcription Regulatory Regions Database (TRRD): its status in 2002. *Nucleic Acids Res*., **30**, 312–317.

Lescot,M. *et al*. (2002) PlantCARE, a database of plant *cis*-acting regulatory elements and a portal to tools for *in silico* analysis of promoter sequences. *Nucleic Acids Res*., **30**, 325–327.

Maglott,D. *et al*. (2005) Entrez Gene: gene-centered information at NCBI. *Nucleic Acids Res*., **33**, D54–D58.

Matys,V. *et al*. (2003) TRANSFAC: transcriptional regulation, from patterns to profiles, *Nucleic Acids Res*., **31**, 374–378.

Pohar,T.T. *et al*. (2004) HemoPDB: Hematopoiesis Promoter Database, an information resource of transcriptional regulation in blood cell development. *Nucleic Acids Res*., **32**, D86–D90.

Ponomarenko,J.V. *et al*. (2001) rSNP_Guide, a database system for analysis of transcription factor binding to target sequences: application to SNPs and site-directed mutations. *Nucleic Acids Res*., **29**, 312–316.

Robertson,A.G. *et al*. (2006) cisRED: a database system for genome-scale computational discovery of regulatory elements. *Nucleic Acids Res*., **34**, D68–D73.

Schmid,C.D. *et al*. (2004) The Eukaryotic Promoter Database EPD: the impact of *in silico* primer extension. *Nucleic Acids Res*., **32**, D82–D85.

Shahmuradov,I.A. *et al*. (2003) PlantProm: a database of plant promoter sequences. *Nucleic Acids Res*., **31**, 114–117.

Sherry,S.T. *et al*. (2001) dbSNP: the NCBI database of genetic variation. *Nucleic Acids Res*., **29**, 308–311.

Stenson,P.D. *et al*. (2003) Human Gene Mutation Database (HGMD): 2003 update. *Hum. Mutat*., **21**, 577–581.

Tahira,T. *et al*. (2005) dbQSNP: a database of SNPs in human promoter regions with allele frequency information determined by single-strand conformation polymorphism-based methods. *Hum. Mutat*., **26**, 69–77.

Tompa,M. *et al*. (2005) Assessing computational tools for the discovery of transcription factor binding sites. *Nat. Biotechnol*., **23**, 137–144.

Wasserman,W.W. and Fickett,J.W. (1998) Identification of regulatory regions which confer muscle-specific gene expression. *J. Mol. Biol*., **278**, 167–181.

Wasserman,W.W. and Sandelin,A. (2004) Applied bioinformatics for the identification of regulatory elements. *Nat. Rev. Genet*., **5**, 276–287.

Wheeler,D.L. *et al*. (2005) Database resources of the National Center for Biotechnology Information. *Nucleic Acids Res*., **33**, D39–D45.

Xie,X. *et al*. (2005) Systematic discovery of regulatory motifs in human promoters and 3′ UTRs by comparison of several mammals. *Nature*, **434**, 338–345.

Zhao,T. *et al*. (2004) PromoLign: a database for upstream region analysis and SNPs. *Hum. Mutat*., **23**, 534–539.

Zhu,J. and Zhang,M.Q. (1999) SCPD: a promoter database of the yeast *Saccharomyces cerevisiae*. *Bioinformatics*, **15**, 607–611.