doi:10.1093/nar/gkn844

dictyBase—a *Dictyostelium* bioinformatics resource update

Petra Fey[1], Pascale Gaudet[1], Tomaz Curk[2], Blaz Zupan[2,3], Eric M. Just[1], Siddhartha Basu[1], Sohel N. Merchant[1], Yulia A. Bushmanova[1], Gad Shaulsky[3], Warren A. Kibbe[1] and Rex L. Chisholm[1,*]

[1]dictyBase, Northwestern University Biomedical Informatics Center and Center for Genetic Medicine, Chicago, IL 60611, USA, [2]Faculty of Computer and Information Science, University of Ljubljana, Slovenia and [3]Department of Molecular and Human Genetics, Baylor College of Medicine, Houston, TX 77030, USA

Received September 10, 2008; Revised October 14, 2008; Accepted October 15, 2008

ABSTRACT

dictyBase (http://dictybase.org) is the model organism database for *Dictyostelium discoideum*. It houses the complete genome sequence, ESTs and the entire body of literature relevant to *Dictyostelium*. This information is curated to provide accurate gene models and functional annotations, with the goal of fully annotating the genome. This dictyBase update describes the annotations and features implemented since 2006, including improved strain and phenotype representation, integration of predicted transcriptional regulatory elements, protein domain information, biochemical pathways, improved searching and a wiki tool that allows members of the research community to provide annotations.

INTRODUCTION

Dictyostelium discoideum is a eukaryotic microorganism that exhibits a rather unusual life cycle characterized by a unicellular stage and a facultative multicellular stage, earning it the nickname ‘social amoeba’. This relatively complex behavior for a simple organism makes it an informative system in which to study cellular processes relevant to higher eukaryotes such as cell motility, cell to cell signaling, interspecies interactions and mechanism of drug action, to name a few recent important scientific contributions made using *Dictyostelium* (1).

dictyBase (http://dictybase.org) is the manually annotated genome database for *Dictyostelium*. It contains the entire 34 Mb nuclear genome sequence of the commonly used haploid laboratory strain, AX4 (2), the 55-kb mitochondrial genome (3), the extrachromosomal ribosomal RNA genes (4) and over 162 000 EST sequences (5, Urushihara, H., unpublished data). In addition, all relevant literature is integrated in the database, linked to the appropriate genes and used to annotate gene product functions, strains and mutant phenotypes, and to associate gene ontology terms with gene products.

Here, we describe the new annotations and features that have been implemented in dictyBase since our last report in 2006 (6): a new system for the annotation of strains and phenotypes, the integration of predicted transcriptional regulatory elements, the display of protein domains on the Gene Page and the annotation of biochemical pathways with the dictyCyc tool based on the Pathway Tools software (7). We have also improved search abilities and are now providing a wiki for researchers to share information about *Dictyostelium* genes with other users.

NEW DATA AND ANNOTATIONS

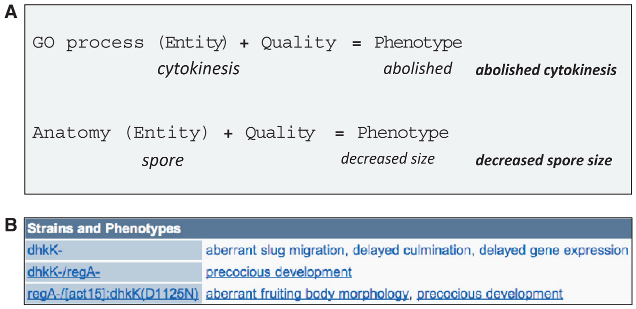
Gene annotations in dictyBase as of September 2008 are shown in Table 1. In addition to the extraction of biological information from the literature, one of the priorities at dictyBase is to manually review every gene model. All available evidence (ESTs, published sequences, sequence similarity) is taken into account to produce the best possible gene model, which is labeled ‘Curated Model’. More than 40% of the predicted genes have been individually inspected. During the gene model curation process, we occasionally encounter pseudogenes as well as splice variants, two types of genes that are difficult to detect by gene prediction softwares.

Several types of nonprotein coding genes have been annotated. We have analyzed the *Dictyostelium* genome for putative tRNA genes using tRNAscan-SE software (8). We also collaborated with the Soderbom group (9) to do an all-automated load of nonprotein coding RNAs such as snoRNAs, signal recognition particle RNAs

[*]To whom correspondence should be addressed. Tel: +1 312 503 2303; Fax: +1 312 503 5603; Email: r-chisholm@northwestern.edu
Correspondence may also be addressed to Pascale Gaudet. Tel: +1 312 503 2303; Fax: +1 312 503 5603; Email: pgaudet@northwestern.edu

Table 1. Data and annotations in dictyBase (September 2008)

- 13 627 Predicted genes
- 5481 Curated gene models
- 19 Genes with alternative transcripts
- 123 Pseudogenes
- 418 tRNAs
- 87 Other ncRNAs
- Regulatory elements for 3600 genes
- 6500 PubMed references
- Gene products for 7594 genes
- Brief descriptions for 4842 genes
- GO annotations for 7141 genes
- Summary paragraphs for 643 genes
- Mutant phenotypes for 691 genes
- 4537 Strains associated with 810 genes
- Biochemical pathways for 491 genes
- 1463 Colleagues
- 111 Community annotations



Strains and Phenotypes	
dhkK-	aberrant slug migration, delayed culmination, delayed gene expression
dhkK-/regA-	precocious development
regA-/[act15]:dhkK(D1125N)	aberrant fruiting body morphology, precocious development

Figure 1. dictyBase phenotype and strain annotations. (**A**) Phenotype construction using the EQ model. See text for details. (**B**) Strain and phenotype annotations as they appear on the dictyBase gene page. Strains are listed by a 'strain descriptor' on the left, with associated phenotypes to the right. The strain dhkK− appears on the *dhkK* gene page, while dhkK−/regA− and regA−/[act15]:dhkK(D1125N) are linked to both the *dhkK* and the *regA* genes.

and snRNAs, most of which have been experimentally verified. The genome browser can be configured to view these various RNAs by turning on the tRNA and the ncRNA tracks.

STRAINS AND PHENOTYPES

Mutational analyses are widely used to study a gene's function or elucidate important biological pathways. Moreover, many diseases are caused by mutations in genes, and model organisms often provide essential insight into the molecular mechanisms of diseases. Mutations in conserved genes often cause similar phenotypes in all organisms that share comparable processes, while other phenotypic manifestations are distinctive to certain organisms. Thus, phenotype annotation poses a unique challenge: the annotations need to use a similar structure in order to be shared while describing the anatomical and behavioral features specific to every organism.

To provide consistent annotations, dictyBase employs a precomposed phenotype ontology based on the EQ syntax developed by the National Center for Biomedical Ontology (10). There are two parts to the phenotype ontology: the entity (E) changed in the mutant, and a quality (Q) describing that modification. For example, a 'small spore' phenotype qualifies the spore (entity) as having 'decreased size' (quality) (Figure 1A). The hierarchical relationships link different phenotypes through common parent terms. For example, 'decreased spore size' and 'increased cell size' are types (children) of 'aberrant cell morphology'. The Entity is derived from biological processes or cellular components in the Gene Ontology (GO) (11), from CheBI (12) or from the *Dictyostelium* Anatomy ontology (13). The qualities are based upon the Quality ontology (PATO) (10). Phenotype terms often contain synonyms to facilitate querying. The Entity–Quality system is increasingly being used by different databases to annotate phenotypes. The use of a consistent vocabulary aids searching and allows grouping of related phenotypes based on specific criteria: one can find all phenotypes related to a specific process, for example, *culmination* (abolished culmination, delayed culmination, etc.), or all phenotypes with a defined quality, such as *delay* in a process (*delayed* aggregation, *delayed* cytokinesis, etc.). Environment conditions and assays are also captured in the phenotype annotation.

Because phenotypes are characteristics of strains rather than genes, the database schema was modified so that phenotypes are linked to strains, which in turn are associated with the appropriate gene(s). Collecting all available information, dictyBase curators annotate strains from the published literature; subsequently, they annotate the phenotypes displayed by the mutant strain. For example, the dhkK$^-$ strain is associated with the phenotypes 'aberrant slug migration', 'delayed culmination' and 'delayed gene expression' (Figure1B). Researchers are encouraged to submit their strains to the Dicty Stock Center and to use controlled vocabularies in their publications.

REGULATORY ELEMENTS

Putative transcriptional regulatory elements have been identified by analyzing genome-based motif information from promoter regions and expression data of about 3600 genes measured in wild-type cells and in fourteen different mutant strains. For every gene, the relation between promoter structure and gene expression in wild-type and mutant cells was identified using a data mining approach called rule-based clustering. This method finds groups of similarly expressed genes with distinct structural similarities in their regulatory regions. It uses a heuristic search, similar to that of the well-known CN2 algorithm (14). Regulatory element patterns are presented as logical expressions that include the assertions on the presence of regulatory elements, their orientation, their distance to other elements and to the first ATG. Details on the methodology can be found at http://dictybase.org/promoters/query.html.

The predicted transcriptional regulatory elements are integrated within the dictyBase Genome Browser, presented as an additional Genome Browser track called

'Putative TF binding sites' (Figure 2A). In addition, when predicted binding sites are present in a gene's upstream region, those are listed on the Gene Page in a section called 'Regulatory Elements'. Clicking on a regulatory element, either from the Genome Browser or the Gene Page, leads to a new page with detailed information on gene expression for that gene and a graphical overview of the promoter structure. From this page, the user can also explore clusters of genes with similar expression pattern and promoter structure (Figure 2B). Finally, there is a specific querying tool where groups of genes specified by the user can be analyzed which returns a ranked list of enriched gene clusters.

PROTEIN DOMAINS

dictyBase gene pages now contain a graphical display of InterPro protein domains (15); Figure 3. The domain information is generated real-time based on the InterPro matches. The retrieval process accesses the UniProt cross-reference (16) for that protein and subsequently leverages the Distributed Annotation System (DAS) protocol (17) to aggregate the domain information from a remote InterPro resource. InterPro integrates domain information from multiple databases such as Pfam, Prosite, Prodom, Smart and Prints. Once retrieved, the data is used to generate a genome browser style visual snapshot where each track represents a particular domain and source database. In the case of repeated domains within the protein, a segmented style track is displayed. Each track is hyperlinked to its source database and upon mouse-over, a tooltip pops up showing the descriptive name of the domain.

IMPROVED ACCESS TO DATA: dictyMart AND DOWNLOADS PAGE

We have expanded the database fields that can be searched using the 'Search dictyBase' tool to allow searching of ESTs, gene descriptions, plasmids, strains and phenotypes; in addition to gene names, gene product names, gene descriptions, Gene Ontology terms, dictyBase IDs, GenBank accession numbers, authors, colleagues and web pages.

For complex queries, we have implemented the open source European Bioinformatics Institute (EBI) package called BioMart (18) to create dictyMart that allows users to combine search criteria to generate custom data sets. dictyMart provides a graphical interface that allows searching for a gene list based on a defined set of dictyBase IDs, common Gene Ontology annotations or chromosomal location. The query output can be specified by selecting several combinations of gene names, identifiers and functional annotations or protein and DNA sequences. In the latter case, it is possible to view only upstream regions, coding sequences, or the complete genomic region. dictyMart is accessible at http://dictybase.org/biomart/martview.

In addition to custom data sets acquired through dictyMart, dictyBase has an extensive download environment where a large collection of up-to-date data can be readily accessed (http://dictybase.org/Downloads). Data include gene information such as gene names and protein products, sequences and sequence annotations (GFF3 format), protein domains, mutant phenotypes, GO annotations and publications.

dictyCyc: BIOCHEMICAL PATHWAYS

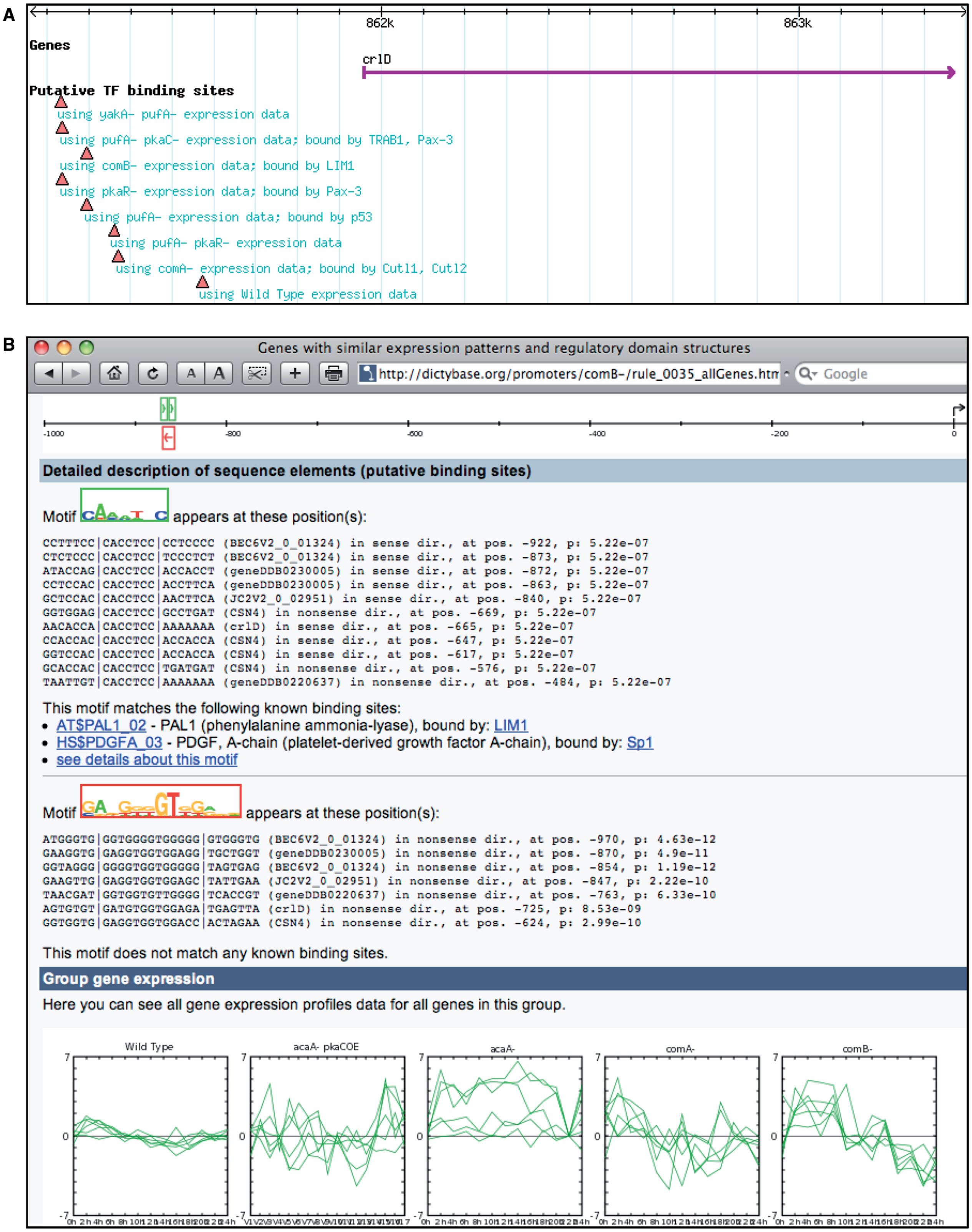
dictyCyc provides visualization of predicted biochemical pathways in dictyBase. Pathways are assigned based on curated gene product names matching entries in the MetaCyc database (7). These matches are used to generate associations between *Dictyostelium* genes and biochemical pathways shared among different organisms. dictyCyc includes software to generate a graphical depiction of the pathways featuring all the reactions, reactants, enzymes and protein complexes involved in a pathway as well as the genes encoding each known protein subunit. The Gene Page shows the names of biochemical pathways in which a gene's product is involved (Supplementary Figure S1A). The link takes the user to a clickable graphical interface displaying all other enzymes in the pathway and linking to genes encoding those enzymes (Supplementary Figure S1B). Pathways are predicted and displayed by the Pathway Tools software from SRI (7) and the dictyCyc main page is at http://dictybase.org/Dicty_Info/dictycyc_info.html. This page can also be accessed from the Biochemical Pathways link under the Research Tools menu of the dictyBase page.

COMMUNITY ANNOTATIONS

A new feature in dictyBase is the ability for members of the *Dictyostelium* research community to directly and immediately add annotations to genes in dictyBase. Each gene page is linked to a corresponding wiki page and a link in red alerts the user that a community annotation is available. The wiki can also be directly accessed at http://wiki.dictybase.org. Users have already entered information such as predicted protein function, suggestions for gene and protein nomenclature and unpublished experimental data such as pictures from mutants. We have found this to be a valuable forum for community input into dictyBase and whenever possible, curators use this information to improve the gene's annotations. The community annotation section was developed using the MediaWiki software, the same wiki software used to develop Wikipedia. The utility of wiki as an annotation tool in biology has recently been a topic receiving much discussion (19–21).

CONCLUSION/FUTURE DIRECTIONS

dictyBase provides a focal point for the integration of all information related to *Dictyostelium* research. The data presented aim to be comprehensive, accurate and easy to access. We will continue to provide new data sets and tools for the research community. Another goal is to expand the scope of external resources where *Dictyostelium* genes and gene products are represented,


A
862k
863k
Genes
crlD
Putative TF binding sites
using yakA- pufA- expression data
using pufA- pkaC- expression data; bound by TRAB1, Pax-3
using comB- expression data; bound by LIM1
using pkaR- expression data; bound by Pax-3
using pufA- expression data; bound by p53
using pufA- pkaR- expression data
using comA- expression data; bound by Cutl1, Cutl2
using Wild Type expression data
B
Genes with similar expression patterns and regulatory domain structures
http://dictybase.org/promoters/comB-/rule_0035_allGenes.htm
Google
-1000
-800
-600
-400
-200
0

Detailed description of sequence elements (putative binding sites)

Motif appears at these position(s):

```
CCTTTCC|CACCTCC|CCTCCCC (BEC6V2_0_01324) in sense dir., at pos. -922, p: 5.22e-07
CTCTCCC|CACCTCC|TCCCTCT (BEC6V2_0_01324) in sense dir., at pos. -873, p: 5.22e-07
ATACCAG|CACCTCC|ACCACCT (geneDDB0230005) in sense dir., at pos. -872, p: 5.22e-07
CCTCCAC|CACCTCC|ACCTTCA (geneDDB0230005) in sense dir., at pos. -863, p: 5.22e-07
GCTCCAC|CACCTCC|AACTTCA (JC2V2_0_02951) in sense dir., at pos. -840, p: 5.22e-07
GGTGGAG|CACCTCC|GCCTGAT (CSN4) in nonsense dir., at pos. -669, p: 5.22e-07
AACACCA|CACCTCC|AAAAAAA (crlD) in sense dir., at pos. -665, p: 5.22e-07
CCACCAC|CACCTCC|ACCACCA (CSN4) in sense dir., at pos. -647, p: 5.22e-07
GGTCCAC|CACCTCC|ACCACCA (CSN4) in sense dir., at pos. -617, p: 5.22e-07
GCACCAC|CACCTCC|TGATGAT (CSN4) in nonsense dir., at pos. -576, p: 5.22e-07
TAATTGT|CACCTCC|AAAAAAA (geneDDB0220637) in nonsense dir., at pos. -484, p: 5.22e-07
```

This motif matches the following known binding sites:

- AT$PAL1_02 - PAL1 (phenylalanine ammonia-lyase), bound by: LIM1
- HS$PDGFA_03 - PDGF, A-chain (platelet-derived growth factor A-chain), bound by: Sp1
- see details about this motif

Motif appears at these position(s):

```
ATGGGTG|GGTGGGGTGGGGG|GTGGGTG (BEC6V2_0_01324) in nonsense dir., at pos. -970, p: 4.63e-12
GAAGGTG|GAGGTGGTGGAGG|TGCTGGT (geneDDB0230005) in nonsense dir., at pos. -870, p: 4.9e-11
GGTAGGG|GGGGTGGTGGGGG|TAGTGAG (BEC6V2_0_01324) in nonsense dir., at pos. -854, p: 1.19e-12
GAAGTTG|GAGGTGGTGGAGC|TATTGAA (JC2V2_0_02951) in nonsense dir., at pos. -847, p: 2.22e-10
TAACGAT|GGTGGTGTTGGGG|TCACCGT (geneDDB0220637) in nonsense dir., at pos. -763, p: 6.33e-10
AGTGTGT|GATGTGGTGGAGA|TGAGTTA (crlD) in nonsense dir., at pos. -725, p: 8.53e-09
GGTGGTG|GAGGTGGTGGACC|ACTAGAA (CSN4) in nonsense dir., at pos. -624, p: 2.99e-10
```

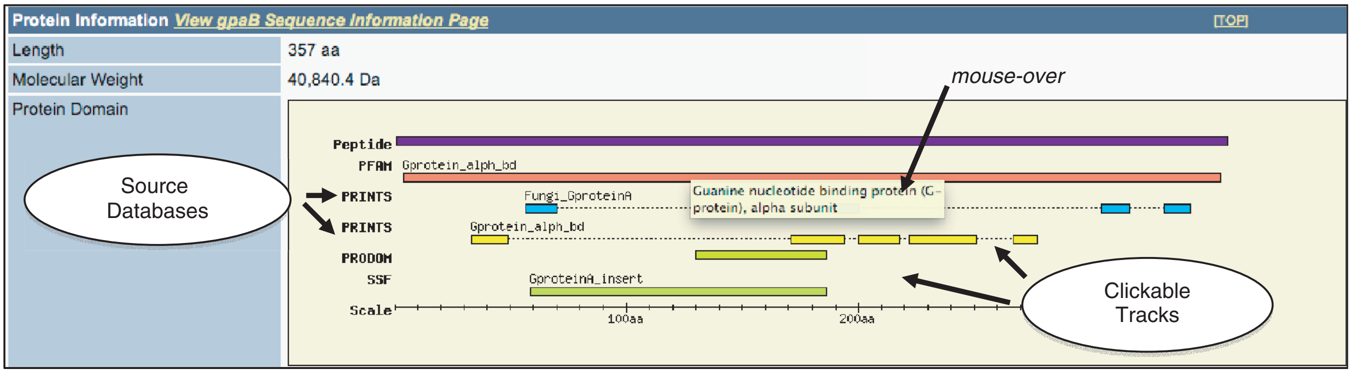
This motif does not match any known binding sites.

Group gene expression

Here you can see all gene expression profiles data for all genes in this group.

Wild Type
acaA- pkaCOE
acaA-
comA-
comB-
7
0
-7
0h 2h 4h 6h 8h 10h 12h 14h 16h 18h 20h 22h 24h
V1 V2 V3 V4 V5 V6 V7 V8 V9 V10 V11 V12 V13 V14 V15 V16 V17

Figure 2. Regulatory elements of *Dictyostelium* genes. (**A**) Gene *crlD* shown in the Genome Browser window with the 'Putative TF binding sites' track turned on and the upstream regulatory elements displayed. Eight putative elements have been detected based on the expression profile of *crlD* in wild-type cells and in seven mutant cell lines. (**B**) Part of a report on a query for genes similar in regulatory structure and expression to the *crlD* gene derived from expression profiles in the comB– mutant. Sequence of the two shared motifs and gene expression in wild-type and four mutant strains are displayed.



Protein Information View gpaB Sequence Information Page [TOP]

Length 357 aa

Molecular Weight 40,840.4 Da

Protein Domain

Figure 3. Graphical visualization of protein domains on a dictyBase gene page. The entire display is organized in the style of a 'generic genome browser' with the peptide on the top track and each track below representing a protein domain or domains and a source database.

which currently includes the Gene Ontology, GenBank, UniProt and orthology analysis tools (InParanoid, OrthoMCL).

As the availability of other amoebae genome sequences is close at hand, dictyBase looks forward to becoming the central genome database resource for those other amoebae. We have already created the infrastructure to house multiple genomes under the dictyBase umbrella and plan to develop tools for comparative genomics.

SUPPLEMENTARY DATA

Supplementary Data are available at NAR Online.

FUNDING

National Institutes of Health (GM64426 and HG0022) (R.L.C.), (P01-HD39691) (T.C., B.Z., G.S.) and the Slovenian Research Agency (P2-0209, J2-9699) (T.C., B.Z.) (partial). Funding for open access charge: National Institutes of Health.

Conflict of interest statement. None declared.

REFERENCES

1. Gaudet,P., Fey,P., Chisholm,R.L. (2008) Dictyostelium discoideum: The Social Amoeba. In *Emerging Model Organisms*. Cold Spring Harbor Laboratories Press, Cold Spring Harbor, NY.
2. Eichinger,L., Pachebat,J.A., Glockner,G., Rajandream,M.A., Sucgang,R., Berriman,M., Song,J., Olsen,R., Szafranski,K., Xu,Q. *et al.* (2005) The genome of the social amoeba *Dictyostelium discoideum*. *Nature*, **435**, 43–57.
3. Ogawa,S., Yoshino,R., Angata,K., Iwamoto,M., Pi,M., Kuroe,K., Matsuo,K., Morio,T., Urushihara,H., Yanagisawa,K. *et al.* (2000) The mitochondrial DNA of *Dictyostelium discoideum*: complete sequence, gene content and genome organization. *Mol. Gen. Genet.*, **263**, 514–519.
4. Sucgang,R., Chen,G., Liu,W., Lindsay,R., Lu,J., Muzny,D., Shaulsky,G., Loomis,W., Gibbs,R. and Kuspa,A. (2003) Sequence and structure of the extrachromosomal palindrome encoding the ribosomal RNA genes in *Dictyostelium*. *Nucleic Acids Res.*, **31**, 2361–2368.
5. Urushihara,H., Morio,T. and Tanaka,Y. (2006) The cDNA sequencing project. *Methods Mol. Biol.*, **346**, 31–49.
6. Chisholm,R.L., Gaudet,P., Just,E.M., Pilcher,K.E., Fey,P., Merchant,S.N. and Kibbe,W.A. (2006) dictyBase, the model organism database for *Dictyostelium discoideum*. *Nucleic Acids Res.*, **34**, D423–D427.
7. Caspi,R., Foerster,H., Fulcher,C.A., Kaipa,P., Krummenacker,M., Latendresse,M., Paley,S., Rhee,S.Y., Shearer,A.G., Tissier,C. *et al.* (2008) MetaCyc Database of metabolic pathways and enzymes and the BioCyc collection of Pathway/Genome Databases. *Nucleic Acids Res.*, **36**, D623–D631.
8. Lowe,T.M. and Eddy,S.R. (1997) A program for improved detection of transfer RNA genes in genomic sequence. *Nucleic Acids Res.*, **25**, 955–964.
9. Aspegren,A., Hinas,A., Larsson,P., Larsson,A. and Soderbom,F. (2004) Novel non-coding RNAs in *Dictyostelium discoideum* and their expression during development. *Nucleic Acids Res.*, **32**, 4646–4656.
10. Mabee,P.M., Ashburner,M., Cronk,Q., Gkoutos,G.V., Haendel,M., Segerdell,E., Mungall,C. and Westerfield,M. (2007) Phenotype ontologies: the bridge between genomics and evolution. *Trends Ecol. Evol.*, **22**, 345–350.
11. Consortium,T.G.O. (2008) The Gene Ontology project in 2008. *Nucleic Acids Res.*, **36**, D440–D444.
12. Degtyarenko,K., deMatos,P., Ennis,M., Hastings,J., Zbinden,M., McNaught,A., Alcántara,R., Darsow,M., Guedj,M. and Ashburner,M. (2008) ChEBI: a database and ontology for chemical entities of biological interest. *Nucleic Acids Res.*, **36**, D344–D350.
13. Gaudet,P., Williams,J.G., Fey,P. and Chisholm,R.L. (2008) An anatomy ontology to represent biological knowledge in *Dictyostelium discoideum*. *BMC Genomics*, **9**, 130–141.
14. Clark,P. and Niblett,T. (1989) The CN2 induction algorithm. *Machine Learning*, **3**, 261–283.
15. Mulder,N.J., Apweiler,R., Attwood,T.K., Bairoch,A., Bateman,A., Binns,D., Bork,P., Buillard,V., Cerutti,L., Copley,R. *et al.* (2007) New developments in the InterPro database. *Nucleic Acids Res.*, **35**, D224–D228.
16. Consortium,T.U. (2007) The Universal Protein Resource (UniProt). *Nucleic Acids Res.*, **35**, D193–D197.
17. Dowell,R.D., Jokerst,R.M., Day,A., Eddy,S.R. and Stein,L. (2001) The distributed annotation system. *BMC Bioinformatics*, **2**, 7–13.
18. Kasprzyk,A., Keefe,D., Smedley,D., London,D., Spooner,W., Melsopp,C., Hammond,M., Rocca-Serra,P., Cox,T. and Birney,E. (2004) EnsMart: a generic system for fast and flexible access to biological data. *Genome Res.*, **14**, 160–169.
19. Giles,J. (2007) Key biology databases go wiki. *Nature*, **445**, 691.
20. Osborne,J.D., Lin,S. and Kibbe,W.A. (2007) Other riffs on cooperation are already showing how well a wiki could work. *Nature*, **445**, 856.
21. Mons,B., Ashburner,M., Chichester,C., van Mulligen,E., Weeber,M., den Dunnen,J., van Ommen,G.J., Musen,M., Cockerill,M., Hermjakob,H. *et al.* (2008) Calling on a million minds for community annotation in WikiProteins. *Genome Biol.*, **9**, R89.