D536–D544 *Nucleic Acids Research, 2013, Vol. 41, Database issue* *Published online 17 November 2012*
doi:10.1093/nar/gks1080

dcGO: database of domain-centric ontologies on functions, phenotypes, diseases and more

Hai Fang* and Julian Gough*

Department of Computer Science, University of Bristol, The Merchant Venturers Building, Bristol BS8 1UB, UK

Received August 15, 2012; Revised September 28, 2012; Accepted October 16, 2012

ABSTRACT

We present 'dcGO' (http://supfam.org/SUPERFAMILY/dcGO), a comprehensive ontology database for protein domains. Domains are often the functional units of proteins, thus instead of associating ontological terms only with full-length proteins, it sometimes makes more sense to associate terms with individual domains. Domain-centric GO, 'dcGO', provides associations between ontological terms and protein domains at the superfamily and family levels. Some functional units consist of more than one domain acting together or acting at an interface between domains; therefore, ontological terms associated with pairs of domains, triplets and longer supra-domains are also provided. At the time of writing the ontologies in dcGO include the Gene Ontology (GO); Enzyme Commission (EC) numbers; pathways from UniPathway; human phenotype ontology and phenotype ontologies from five model organisms, including plants; anatomy ontologies from three organisms; human disease ontology and drugs from DrugBank. All ontological terms have probabilistic scores for their associations. In addition to associations to domains and supra-domains, the ontological terms have been transferred to proteins, through homology, providing annotations of >80 million sequences covering 2414 complete genomes, hundreds of meta-genomes, thousands of viruses and so forth. The dcGO database is updated fortnightly, and its website provides downloads, search, browse, phylogenetic context and other data-mining facilities.

INTRODUCTION

Scientists are increasingly confronted with the grand challenge: how to convert sequenced genome information into higher-order knowledge on function (1), phenotype (2) and even human disease (3).

The domain-centric Gene Ontology (dcGO) database at http://supfam.org/SUPERFAMILY/dcGO is a comprehensive ontology resource that contributes to the aforementioned challenge through a new domain-centric strategy. Our method, dcGO (4), annotates protein domains with ontological terms. Ontologies are hierarchically organized controlled vocabularies/terms defined to categorize a particular sphere of knowledge (5). For example, 'Gene Ontology' (GO) was created to describe functions of proteins (6). Ontological labels are already available for full-length proteins, derived from experimental data for that protein. The dcGO approach takes the terms attached to full-length sequences, and combines them with the domain composition of the sequences on a large scale, to statistically infer for each term which domain is the functional unit responsible for it. The method has been formulated in a general way, enabling it to be applied to numerous ontologies. The dcGO database now contains a panel of ontologies from a variety of contexts: functions such as GO (6,7), enzymes (8), pathways (9) and keywords used by UniProt (10); phenotype and anatomy ontologies across major model organisms, including mouse (11), worm (12), yeast (13), fly (14), zebrafish (15), *Xenopus* (16) and *Arabidopsis* (17); human phenotypes (18), diseases (19) and drugs (20). In addition to complete sets of ontological terms, a collapsed subset (slim version) is also provided for each ontology. The automatically generated slim version of each ontology is based on annotation frequency (4), and it provides the user with a manageable and more coarse-grained list. This is analogous to the 'GO slim' provided by the Gene Ontology consortium, which has proven useful for enrichment analyses (21).

The domain definitions used in dcGO are taken from the structural classification of proteins (SCOP) (22) classified at both the superfamily and family levels. SCOP groups domains at the superfamily level if there is structure, sequence and function evidence for a common evolutionary ancestor. Some superfamilies are sub-divided into families, which often share a higher sequence

*To whom correspondence should be addressed. Tel: +44 792 9104281; Fax: +44 117 9545208; Email: hfang@cs.bris.ac.uk
Correspondence may also be addressed to Julian Gough. Tel: +44 117 3315221; Fax: +44 117 9545208; Email: gough@cs.bris.ac.uk

similarity and a related function. In addition to individual domains at these two different levels, dcGO also offers annotations for combinations of domains. We use the concept of supra-domains to describe combinations of two or more successive domains of known structure. In addition to providing ontology for SCOP domains, the generality of the method has enabled us to also include Pfam (23) domains in dcGO.

Our domain-centric ontology derived from proteins with experimental evidence can be turned and used as a predictor on proteins of unknown function but where the domain content is known. The 'dcGO Predictor' provides pre-computed functional annotation [using SUPERFAMILY hidden Markov models (24)] of all sequences in UniProt (25), 2414 completely sequenced genomes, thousands of viral genomes and hundreds of meta-genomes. The dcGO website also has a facility for the user to submit their own sequences for function prediction. The dcGO Predictor took part in the recent Critical Assessment of Functional Annotation experiment (http://biofunctionprediction.org); hence, for comparison with other non-domain-centric predictors, we refer the reader to this independent evaluation of its performance. A key result from the experiment, however, was that dcGO performs significantly better than the most commonly used method for GO annotation, Basic Local Alignment Search Tool searches against UniProt (25).

In the main body of the article later we describe the database contents in detail, doing so separately for GO and for other biomedical ontologies (collectively denoted as 'BO' hereinafter). Then, we provide an overview of various utilities available through the website that may interest users. Finally, we conclude with planned future developments.

DATABASE CONTENTS

Algorithm summary

To fully understand the content of the dcGO database (Table 1), it is necessary to give a description of the algorithm that is used to build it. Without loss of generality, we take the GO as an exclusive example, but in principle, the applications to other ontologies are the same. For more detail, the reader is referred to a previous publication of the algorithm (4).

The GO is designed to annotate full-length proteins in a species-independent manner for generality. The most comprehensive protein-level annotations are maintained by the Gene Ontology Annotation (GOA) project (7). Motivated by a domain-centric viewpoint, we developed a general algorithm (4) for revealing functional signals carried by protein domains (and supra-domains in the multi-domain proteins). Using protein domain architectures from SUPERFAMILY and protein GO annotations from UniProKB-GOA (respecting the GO hierarchy), we first prepare a correspondence matrix between domains/supra-domains and GO terms. Each entry has the observed number of UniProt proteins that contain a domain/supra-domain (columns) and that can be annotated by a GO term (rows). With this correspondence matrix, we then use Fisher's exact test to infer associations between the rows and columns. On top of this, we take advantage of the true path rule of the directed acyclic graph of the GO to determine the optimal level at which to make an association. We achieve this by comparing the significance of each term using two different backgrounds, one background using all analysable UniProt proteins (being annotatable by the GO), and one background using only those UniProt proteins annotated to direct parents of the term. If a GO term and its parent term are both significantly associated with a domain/supra-domain using the first background, and if the term is not significantly different from the parent term using the second background, then it is desirable to only associate the parent term. As a result of these dual constraints, only the most significant GO term associations to domains/supra-domains will be retained.

The significance of association is assessed by the method of false discovery rate (FDR) to account for multiple hypothesis tests, whereas the strength of association is measured by a hypergeometric distribution-based score. For a domain/supra-domain, the associated GO terms (i.e. direct annotations) are propagated to all ancestor terms (i.e. inherited annotations); together they constitute a complete GO annotation profile. Based on the information content of a GO term (i.e. negative logarithmic transformation of frequency of domains/supra-domains annotated to that term), a search procedure is applied to partition the directed acyclic graph structure of the GO, each partition reflecting the same or similar specificity but located in distinct paths. With four seeds of increasing information content, the procedure produces the 'GO slim' that contains GO terms classified into four levels of increasing granularity. These are highly general, general, specific and highly specific (Supplementary Table S1). The use of information content (as a measure of how specific and informative a term is) adds great value to the existing GO hierarchy for the user. The GO was created for annotating proteins, so some parts of GO structure are less valid for annotating domains/supra-domains than others. Rather than merely relying on the ontology graph depth to define the term specificity, our approach has taken into account actual usage of terms when determining the four-level depth classification of domains/supra-domains.

Domain-centric GO

Using the algorithm described earlier in the text, dcGO provides the user with two alternative versions of the GO associations with domains (Table 1). The high-quality version of associations includes only those that are supported by the unambiguous evidence (in terms of the causal domain) that comes from single-domain proteins of known function. The high-coverage associations also include those that are supported through statistical disambiguation from multi-domain proteins of known function. The high-quality associations are more reliable in their domain-centricity, but high-coverage associations are reliable enough for large-scale studies and provide a much greater coverage of function. Enrichment analyses

Table 1. A summary of the dcGO database contents (on 15 August 2012)

Ontology	Domains (superfamily level)			Domains (family level)			Supra-domains (superfamily level)			Data source
	Number of terms[a]	Number of domains[b]	Number of annotations[c]	Number of terms[a]	Number of domains[b]	Number of annotations[c]	Number of terms[a]	Number of supra-domains[b]	Number of annotations[c]	
Functions										
Gene Ontology (GO)	High-quality version[d]			High-quality version[d]						UniProtKB-GOA (6)
	4761	481	27 833	4652	657	29 529				
	High-coverage version[e]			High-coverage version[e]						
	10 497	1265	139 958	10 032	2026	157 618	13 306	7761	587 917	
Diseases										
Disease Ontology (DO)	357	115	1345	364	145	1702	402	276	2765	DO (19)
Phenotypes										
Human Phenotype (HP)	670	147	2079	605	141	1930	750	289	4104	HPO (18)
Mammalian Phenotype (MP)	1858	299	8555	2040	368	12 008	2202	844	23 149	MGI (11)
Worm Phenotype (WP)	556	296	4349	540	320	4572	571	507	6976	WormBase (12)
Yeast Phenotype (YP)	76	271	1070	72	256	1039	79	392	1529	SGD (13)
Fly Phenotype (FP)	64	140	268	62	167	314	69	283	557	FlyBase (14)
Fly Anatomy (FA)	502	191	3210	555	210	4183	551	349	8151	FlyBase (14)
Zebrafish Anatomy (ZA)	158	66	694	164	57	701	173	121	1316	ZFIN (15)
Xenopus Anatomy (XA)	243	474	8376	245	583	11 187	253	875	17 730	Xenbase (16)
Arabidopsis Plant (AP)	259	579	20 689	253	778	32 405	266	1093	45 311	TAIR (17)
Others										
Enzyme Commission (EC)	1918	830	8483	1973	1565	10 278	1947	2958	21 028	IntEnz (8)
DrugBank ATC_code (DB)	964	143	2801	904	145	2659	984	230	3950	DrugBank (20)
UniProtKB KeyWords (KW)	857	1573	19 312	840	2798	25 841	866	5579	84 815	UniProt (10)
UniPathway (UP)	664	474	6332	626	796	7395	665	1356	12 918	UniPathway (9)

[a] The total number of ontology terms used to annotate.
[b] The number of annotatable domains (or supra-domains).
[c] The number of domain-centric annotations.
[d] This version is truly domain-centric, supported both by single-domain proteins and all proteins (including multi-domain proteins).
[e] This version is only supported by all proteins, suitable for large-scale studies.

are improved more by annotation coverage than by annotation quality; hence, there is a strong justification for using the high-coverage version in such studies. Restricting the annotations to GO slim (described earlier in the text) is also highly recommended for domain-based enrichment analyses.

In addition to individual domains, dcGO also associates GO terms with supra-domains (Table 1). In general, supra-domains are defined as recurring combinations of two or more successive domains that could function together. In dcGO, we only include completely assigned supra-domains, without any significant gaps between domains, that is, supra-domains with regions not assigned to a known domain are excluded (26). GO associations to supra-domains hold great promise for understanding how domain combinations contribute to functional diversification and also in predicting the functions of multi-domain proteins.

Domain-centric BO

In dcGO, the 'BO' refers to all other Biomedical Ontologies that are not GO. They mainly consist of phenotype ontologies that have been developed to classify and organize information on model organisms and human. Similarly to GO, the BO is hierarchical going from general terms at the top to more specific terms at the bottom. As with domain-centric GO, dcGO has the associations of the BO terms to individual domains and supra-domains; each has its own slim version of the ontology at four levels of increasing granularity based on information content. Unlike the GO, the BO does not have the high-quality version of the associations. This is largely because of an insufficient number of single-domain proteins with annotations, especially for species-specific ontologies. As listed in Table 1, currently dcGO has eight phenotype and/or anatomy ontologies covering seven major model organisms. They include Mouse/Mammalian Phenotypes (MP) from Mouse Genome Informatics (MGI) (11), Worm Phenotypes (WP) from WormBase (12), Yeast/Ascomycete Phenotype (YP) from Saccharomyces Genome Database (SGD) (13), Fly Phenotype (FP) and Fly Anatomy (FA) from FlyBase (14), Zebrafish Anatomy (ZA) from ZFIN (15), *Xenopus* Anatomy (XA) from Xenbase (16) and *Arabidopsis* Plant (AP) ontology from TAIR (17). In addition to model organisms, dcGO also contains three ontologies with specific relevance to humans, including Human Phenotype (HP) (18), Disease Ontology (DO) (19) and DrugBank ATC codes (DB). The remaining ontologies have a fixed-length or much-simplified hierarchy. These include Enzyme Commission (EC) (8), UniProtKB UniPathway (UP) (9) and UniProtKB KeyWords (KW) (10).

DATABASE WEBSITE

Downloading data

The underlying data summarized in Table 1 are available for download on the dcGO website. For each ontology, the full and slim versions are provided separately for individual domains (i.e. superfamilies and families) and supra-domains. In addition, the user can download the MySQL relational database tables along with detailed documentation. All downloadable files are free for academic or commercial use and are automatically updated fortnightly.

Searching dcGO

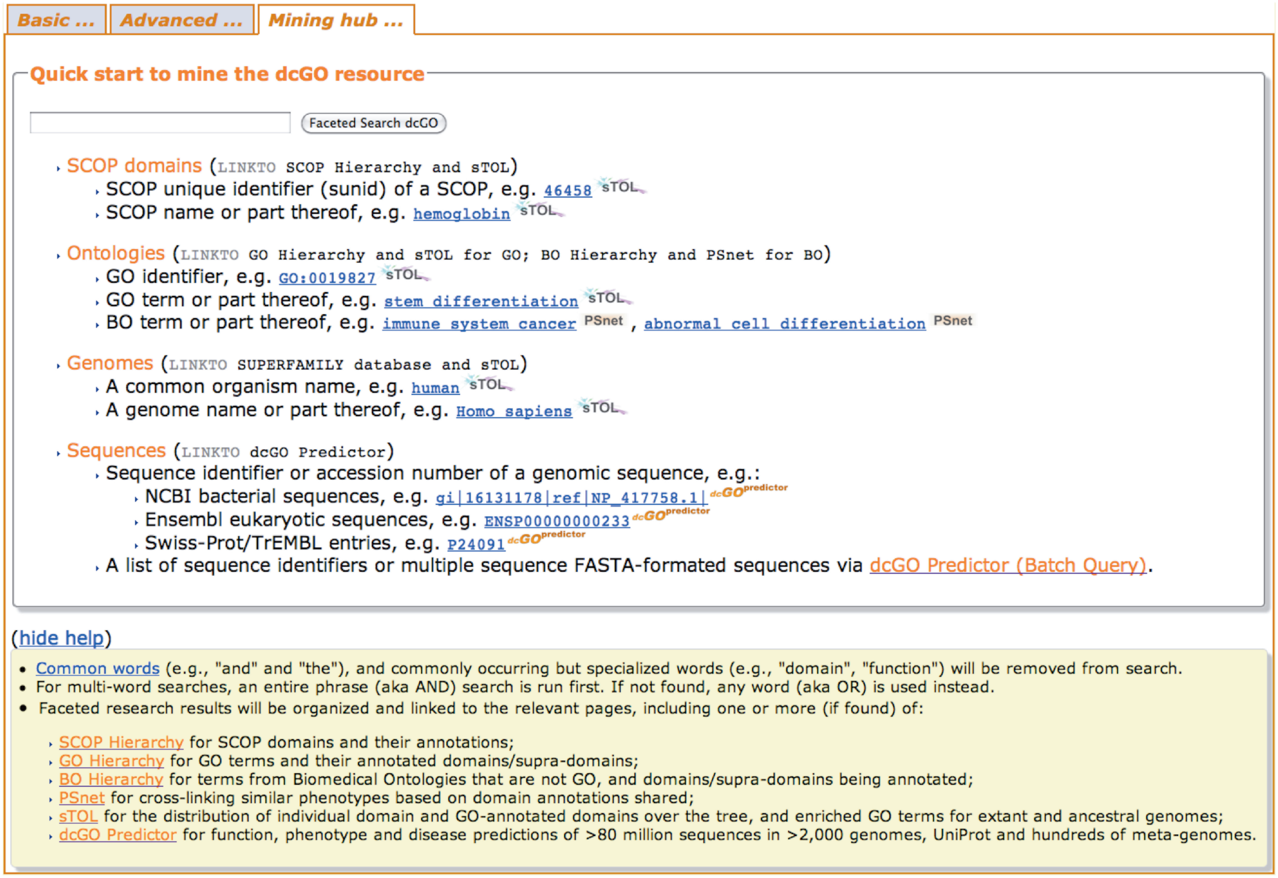
The faceted search on the dcGO website (Figure 1) is a mining hub for users, with additional bioinformatics tools hyperlinked from the search results. Full-text query is supported for SCOP domains, ontologies and genomes. Identifier or accession number lookup is supported for sequences. Ontologies and SCOP domains are linked to pages for browsing their respective hierarchies. Every genome is presented within its phylogenetic context by linking to a species tree of life (called sTOL, see 'Analysing GO terms over the species tree of life' section). There are also links from domains and ontological terms to the tree of life (to see their distribution across species). Search results returning BO terms are linked to a cross-ontology comparison tool, the phenotype similarity network (PSnet, see 'Cross-linking similar phenotypes' section). PSnet searches for terms from other ontologies with a similar profile of associations. For lookups returning a specific genome sequence, the user is provided with the facility to submit it automatically to the 'dcGO Predictor' for function, phenotype and disease prediction. In conclusion, the faceted search is designed for multi-tasking; it does not just provide search results but is intended to interconnect all the tools and cross-referencing abilities of dcGO.

Browsing the hierarchies

The 'BROWSE' navigation on the website (aforementioned) provides browsing for the SCOP, GO and various BO hierarchies. The hierarchy-like structure of the SCOP (or ontology) has a domain (or term) as a node and its relations to parental nodes as directed edges. To navigate this hierarchy, we display all the paths from the current node upwards to the root ordered by the shortest distances. Also, all direct children of the current node are listed underneath to enable browsing downwards. In addition to the hierarchy itself, a tabbed interface is used to aid the display of domain-centric annotations in a subject-specific manner. The SCOP-orientated hierarchy shows terms used to annotate a domain, and vice versa, the ontology-orientated hierarchy shows domains/supra-domains annotated by a term.

Analysing GO terms over the species tree of life

The dcGO website is integrated with a species tree of life (called sTOL), which is provided by SUPERFAMILY (27). The sTOL is a fully resolved binary tree of species of completely sequenced organisms providing a phylogenetic context. Within the sTOL, the presence/absence of domains and supra-domains are pre-computed and stored, both for extant genomes and for reconstructed ancestral genomes in eukaryotes. The integration enables



Quick start to mine the dcGO resource

- SCOP domains (LINKTO SCOP Hierarchy and sTOL)
  - SCOP unique identifier (sunid) of a SCOP, e.g. 46458 sTOL
  - SCOP name or part thereof, e.g. hemoglobin sTOL
- Ontologies (LINKTO GO Hierarchy and sTOL for GO; BO Hierarchy and PSnet for BO)
  - GO identifier, e.g. GO:0019827 sTOL
  - GO term or part thereof, e.g. stem differentiation sTOL
  - BO term or part thereof, e.g. immune system cancer PSnet , abnormal cell differentiation PSnet
- Genomes (LINKTO SUPERFAMILY database and sTOL)
  - A common organism name, e.g. human sTOL
  - A genome name or part thereof, e.g. Homo sapiens sTOL
- Sequences (LINKTO dcGO Predictor)
  - Sequence identifier or accession number of a genomic sequence, e.g.:
    - NCBI bacterial sequences, e.g. gi|16131178|ref|NP_417758.1| dcGO predictor
    - Ensembl eukaryotic sequences, e.g. ENSP00000000233 dcGO predictor
    - Swiss-Prot/TrEMBL entries, e.g. P24091 dcGO predictor
  - A list of sequence identifiers or multiple sequence FASTA-formated sequences via dcGO Predictor (Batch Query).

(hide help)

- Common words (e.g., "and" and "the"), and commonly occurring but specialized words (e.g., "domain", "function") will be removed from search.
- For multi-word searches, an entire phrase (aka AND) search is run first. If not found, any word (aka OR) is used instead.
- Faceted research results will be organized and linked to the relevant pages, including one or more (if found) of:
  - SCOP Hierarchy for SCOP domains and their annotations;
  - GO Hierarchy for GO terms and their annotated domains/supra-domains;
  - BO Hierarchy for terms from Biomedical Ontologies that are not GO, and domains/supra-domains being annotated;
  - PSnet for cross-linking similar phenotypes based on domain annotations shared;
  - sTOL for the distribution of individual domain and GO-annotated domains over the tree, and enriched GO terms for extant and ancestral genomes;
  - dcGO Predictor for function, phenotype and disease predictions of >80 million sequences in >2,000 genomes, UniProt and hundreds of meta-genomes.

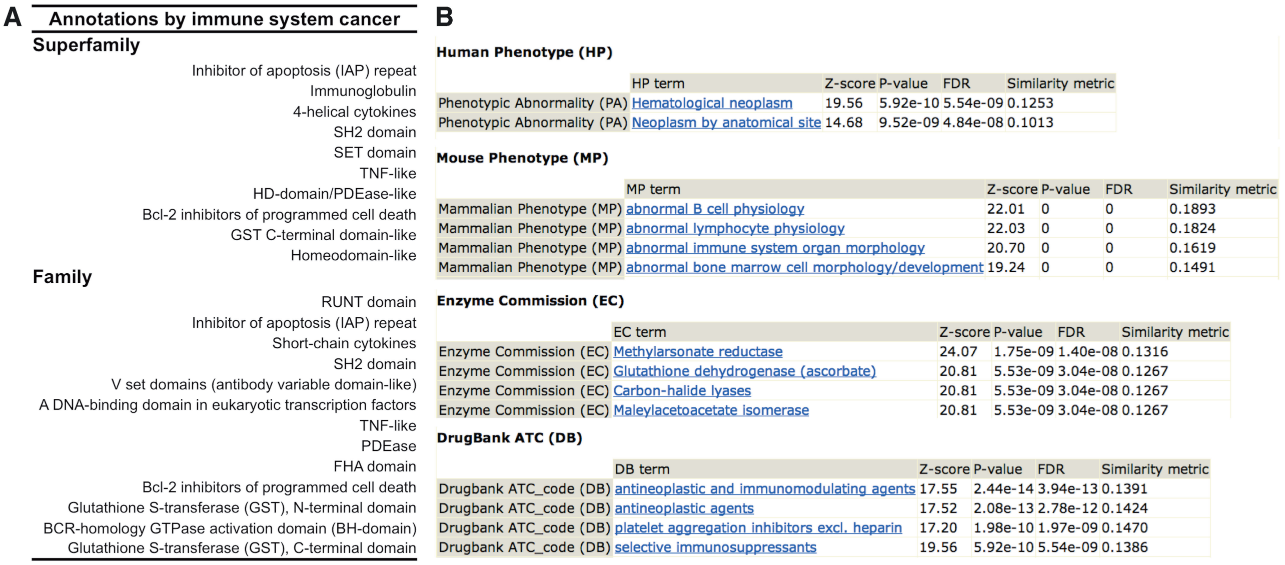
Figure 1. The dcGO website has the 'Faceted Search' interface as a hub to mine the resource. By searching against keywords of interest, the user can access the resource in an organized manner and can link to additional analysis tools.

cross-comparison between both resources for understanding the evolutionary context of the functional associations. The distribution of GO terms can be explored over the sTOL tree, which can be accessed either by links from the GO hierarchy pages or from the faceted search. GO term enrichment for extant and ancestral genomes (using domain-based GO enrichment analysis) can also be explored when browsing the tree. Thus, the sTOL adds an extra dimension to the dcGO resource utility.

Cross-linking similar phenotypes

Traditionally, phenotype ontologies are developed for within-species comparisons. Recently, many attempts have been made at cross-species comparisons (28–30), these studies mostly focusing on text mining and formal definitions. Within dcGO is a tool called 'PSnet' that cross-references terms between ontologies (mostly phenotypes in different model organisms). Given one phenotype, PSnet can be used to search for a similar candidate phenotype on the basis of their shared domain annotations (both at superfamily and family levels). The statistical significance of the shared domains versus the expected overlap by chance is evaluated by Fisher's exact test. PSnet reports a *Z*-score for the strength of the overlap, and a *P*-value and false discovery rate (accounting for multiple hypothesis tests) for the significance. An information content-based similarity metric is used to rank the phenotype similarities; if a certain domain is more frequently annotated (less informative) than others, then its contribution to the phenotype similarity is less. In this way, for any given phenotype, PSnet will suggest the best-correlated terms from other ontologies.

As a proof of principle, we consider the disease term 'immune system cancer' [DOID: 0 060 083] from the Disease Ontology (19). In Figure 2, we illustrate with this example, how PSnet displays cross-links between correlated terms, which facilitates the development of hypotheses. In dcGO, 10 superfamilies and 13 families are associated to this term (Figure 2A). Supplementary Figure S1B shows the numerical pathway of associating immune system cancer with the immunoglobulin superfamily, following the general procedure shown in Supplementary Figure S1A. Given this disease and its domain-centric annotation profile in Figure 2A, PSnet searches for phenotypes with similar domain annotations. As shown in Figure 2B, PSnet cross-references this disease term with closely related terms from the Human Phenotype ontology (18), suggests possible links to the abnormal counterparts in the Mouse Phenotype ontology (11), reveals the mechanisms by listing top



Annotations by immune system cancer

Superfamily

Inhibitor of apoptosis (IAP) repeat
Immunoglobulin
4-helical cytokines
SH2 domain
SET domain
TNF-like
HD-domain/PDEase-like
Bcl-2 inhibitors of programmed cell death
GST C-terminal domain-like
Homeodomain-like

Family

RUNT domain
Inhibitor of apoptosis (IAP) repeat
Short-chain cytokines
SH2 domain
V set domains (antibody variable domain-like)
A DNA-binding domain in eukaryotic transcription factors
TNF-like
PDEase
FHA domain
Bcl-2 inhibitors of programmed cell death
Glutathione S-transferase (GST), N-terminal domain
BCR-homology GTPase activation domain (BH-domain)
Glutathione S-transferase (GST), C-terminal domain

Human Phenotype (HP)

	HP term	Z-score	P-value	FDR	Similarity metric
Phenotypic Abnormality (PA)	Hematological neoplasm	19.56	5.92e-10	5.54e-09	0.1253
Phenotypic Abnormality (PA)	Neoplasm by anatomical site	14.68	9.52e-09	4.84e-08	0.1013

Mouse Phenotype (MP)

	MP term	Z-score	P-value	FDR	Similarity metric
Mammalian Phenotype (MP)	abnormal B cell physiology	22.01	0	0	0.1893
Mammalian Phenotype (MP)	abnormal lymphocyte physiology	22.03	0	0	0.1824
Mammalian Phenotype (MP)	abnormal immune system organ morphology	20.70	0	0	0.1619
Mammalian Phenotype (MP)	abnormal bone marrow cell morphology/development	19.24	0	0	0.1491

Enzyme Commission (EC)

	EC term	Z-score	P-value	FDR	Similarity metric
Enzyme Commission (EC)	Methylarsonate reductase	24.07	1.75e-09	1.40e-08	0.1316
Enzyme Commission (EC)	Glutathione dehydrogenase (ascorbate)	20.81	5.53e-09	3.04e-08	0.1267
Enzyme Commission (EC)	Carbon-halide lyases	20.81	5.53e-09	3.04e-08	0.1267
Enzyme Commission (EC)	Maleylacetoacetate isomerase	20.81	5.53e-09	3.04e-08	0.1267

DrugBank ATC (DB)

	DB term	Z-score	P-value	FDR	Similarity metric
Drugbank ATC_code (DB)	antineoplastic and immunomodulating agents	17.55	2.44e-14	3.94e-13	0.1391
Drugbank ATC_code (DB)	antineoplastic agents	17.52	2.08e-13	2.78e-12	0.1424
Drugbank ATC_code (DB)	platelet aggregation inhibitors excl. heparin	17.20	1.98e-10	1.97e-09	0.1470
Drugbank ATC_code (DB)	selective immunosuppressants	19.56	5.92e-10	5.54e-09	0.1386

Figure 2. Using 'PSnet' to cross-link phenotypes and other ontologies based on shared domain-centric annotations. (**A**) A list of superfamilies and families annotated by a disease term 'immune system cancer'. (**B**) The top well-correlated ontological terms are returned for the disease term in this query.

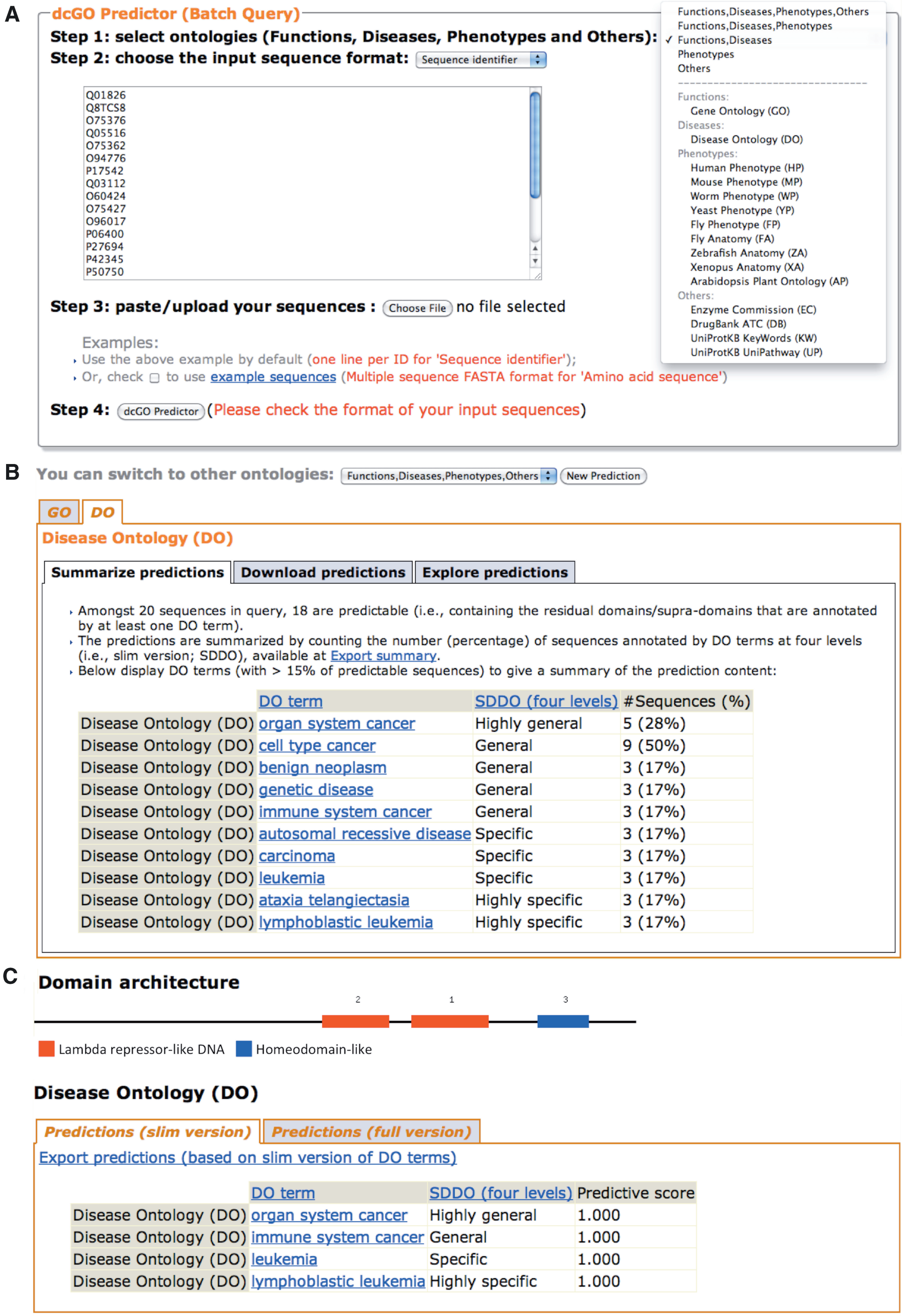
enzymes from the Enzyme Commission (8) and implicates the treatment agents through DrugBank ATC codes (20). The multiple layers of information revealed by PSnet provide a powerful tool for hypothesis generation. PSnet brings additional understanding to the essential roles that protein domains can play in functions, phenotypes and diseases.

Predicting functions, phenotypes and diseases for >80 million sequences

Using domain-centric GO annotations as a functional predictor, we entered the Critical Assessment of Function Annotation competition and came in the top 10 of >50 methods. Considering that only domain information is involved and natively used as a single direct prediction, its relative success validates the quality of this resource for widespread use. We provide pre-computed annotations for >80 million sequences (at the time of writing) stored in the SUPERFAMILY database that includes 2414 genomes, UniProt and hundreds of meta-genomes. Through the dcGO Predictor (Figure 3), functions and other higher-order knowledge (phenotypes, diseases and more) can be predicted for user-submitted sequences. The implementation is fairly straightforward: first the domain architecture of the query protein is determined, and then the ontological terms associated with its component domains/supra-domains are transferred to the query protein. A score is provided for ranking the confidence of such predictions/transfers. In addition to access through the faceted search, a batch query mode is provided, which allows the submission of up to 1000 sequences at a time (Figure 3A). The prediction results are summarized to give an overview of the prediction content and are also available for download (Figure 3B). Figure 3C shows the results for the example input sequence 'Q01826', i.e. special AT-rich sequence-binding protein 1 (SATB1). As a chromatin regulator, this protein has been reported to promote tumour growth and metastasis (31), which is consistent with the prediction. For a sequence to receive annotation, first there must be domains detected by the SUPERFAMILY hidden Markov model library search, and then those domains/supra-domains must have ontological associations in dcGO. Our coverage will improve as new structures are deposited in the Protein Data Bank (32) and as more sequences have ontological terms experimentally determined.

CONCLUSION AND FUTURE DEVELOPMENTS

With the rate of growth of biological (e.g. sequence) data increasing rapidly, the only realistic way to analyse biology in a holistic way is computationally. Gene Ontology has become a widely adopted medium for handling biological concepts in a structured way that can be processed computationally. With this unique database 'dcGO', treating domains and supra-domains as functional units, we provide GO plus a growing number of other ontologies in a probabilistic framework. The results of the Critical Assessment of Function Annotation experiment show that this domain-centric approach performs significantly better than simple whole-sequence pair-wise homology on the task of labelling sequences of unknown function with GO terms; by 'simple whole-sequence homology', we mean the strategy of annotating a sequence with GO terms by searching it against UniProt using Basic Local Alignment Search Tool and transferring any GO terms associated with significant hits. Thus, the dcGO database, in providing full functional annotations of all completely sequenced genomes in addition to the domain-ontology associations themselves, makes a massive contribution to the body of computer-readable biological knowledge. In the future, the intention is to expand the ontologies included in the



dcGO Predictor (Batch Query)

Step 1: select ontologies (Functions, Diseases, Phenotypes and Others):

Step 2: choose the input sequence format: Sequence identifier

```
Q01826
Q8TCS8
O75376
Q05516
O75362
O94776
P17542
Q03112
O60424
O75427
O96017
P06400
P27694
P42345
P50750
```

Step 3: paste/upload your sequences : Choose File no file selected

Examples:
- Use the above example by default (one line per ID for 'Sequence identifier');
- Or, check □ to use example sequences (Multiple sequence FASTA format for 'Amino acid sequence')

Step 4: dcGO Predictor (Please check the format of your input sequences)

You can switch to other ontologies: Functions,Diseases,Phenotypes,Others New Prediction

GO | DO

Disease Ontology (DO)

Summarize predictions | Download predictions | Explore predictions

- Amongst 20 sequences in query, 18 are predictable (i.e., containing the residual domains/supra-domains that are annotated by at least one DO term).
- The predictions are summarized by counting the number (percentage) of sequences annotated by DO terms at four levels (i.e., slim version; SDDO), available at Export summary.
- Below display DO terms (with > 15% of predictable sequences) to give a summary of the prediction content:

	DO term	SDDO (four levels)	#Sequences (%)
Disease Ontology (DO)	organ system cancer	Highly general	5 (28%)
Disease Ontology (DO)	cell type cancer	General	9 (50%)
Disease Ontology (DO)	benign neoplasm	General	3 (17%)
Disease Ontology (DO)	genetic disease	General	3 (17%)
Disease Ontology (DO)	immune system cancer	General	3 (17%)
Disease Ontology (DO)	autosomal recessive disease	Specific	3 (17%)
Disease Ontology (DO)	carcinoma	Specific	3 (17%)
Disease Ontology (DO)	leukemia	Specific	3 (17%)
Disease Ontology (DO)	ataxia telangiectasia	Highly specific	3 (17%)
Disease Ontology (DO)	lymphoblastic leukemia	Highly specific	3 (17%)

Domain architecture

🟥 Lambda repressor-like DNA 🟦 Homeodomain-like

Disease Ontology (DO)

Predictions (slim version) | Predictions (full version)

Export predictions (based on slim version of DO terms)

	DO term	SDDO (four levels)	Predictive score
Disease Ontology (DO)	organ system cancer	Highly general	1.000
Disease Ontology (DO)	immune system cancer	General	1.000
Disease Ontology (DO)	leukemia	Specific	1.000
Disease Ontology (DO)	lymphoblastic leukemia	Highly specific	1.000

Figure 3. Converting genome sequences to knowledge about function, phenotype and disease using the 'dcGO Predictor'. (**A**) A batch query facility allows the user to upload up to 1000 sequences for the prediction on function, disease, phenotype and other information, such as enzyme classification, drugs and pathways. (**B**) The result page provides a summary of the prediction content. New predictions are supported by instantly switching to other ontologies. In addition to the download, the user can also explore predictions for each of the input sequences, such as Q01826 (human SATB1 protein; see next). (**C**) The domain architecture of the human SATB1 protein is graphically displayed using the SCOP domains at the superfamily level, whereas the bottom panel shows the predicted Disease Ontology terms.

database, expand the domain collection as more domains are classified, and expand the collection of functionally annotated sequences as new genomes, meta-genomes and so forth are released.

In addition to the value of the large-scale raw annotations in the dcGO database, the anticipated potential for comparative analyses is already reflected in the sTOL evolutionary context and PSnet cross-referencing tools. Other than the data expansion aforementioned, other future developments will focus on introducing more comparative tools and increasing the use cases of the existing ones. These will include network-based infrastructures, for example, of domains and of terms spanning different ontologies. The construction of functional domain networks with respect to GO is already on the agenda.

SUPPLEMENTARY DATA

Supplementary Data are available at NAR Online: Supplementary Table 1 and Supplementary Figure 1.

ACKNOWLEDGEMENTS

The authors acknowledge the numerous providers of the primary ontologies. Special thanks go to all of the contributors to the SUPERFAMILY resource, which is tightly linked to the dcGO database.

FUNDING

Funding for open access charge: Biotechnology and Biological Sciences Research Council [BB/G022771/1 to J.G.].

Conflict of interest statement. None declared.

REFERENCES

1. Friedberg,I. (2006) Automated protein function prediction—the genomic challenge. *Brief. Bioinform.*, **7**, 225–242.
2. Thorisson,G.A., Muilu,J. and Brookes,A.J. (2009) Genotype-phenotype databases: challenges and solutions for the post-genomic era. *Nat. Rev. Genet.*, **10**, 9–18.
3. Butler,D. (2010) Human genome at ten: science after the sequence. *Nature*, **465**, 1000–1001.
4. de Lima Morais,D.A., Fang,H., Rackham,O.J., Wilson,D., Pethica,R., Chothia,C. and Gough,J. (2011) SUPERFAMILY 1.75 including a domain-centric gene ontology method. *Nucleic Acids Res.*, **39**, D427–D434.
5. Smith,B., Ashburner,M., Rosse,C., Bard,J., Bug,W., Ceusters,W., Goldberg,L.J., Eilbeck,K., Ireland,A., Mungall,C.J. *et al.* (2007) The OBO Foundry: coordinated evolution of ontologies to support biomedical data integration. *Nat. Biotechnol.*, **25**, 1251–1255.
6. Ashburner,M., Ball,C.A., Blake,J.A., Botstein,D., Butler,H., Cherry,J.M., Davis,A.P., Dolinski,K., Dwight,S.S., Eppig,J.T. *et al.* (2000) Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nat. Genet.*, **25**, 25–29.
7. Barrell,D., Dimmer,E., Huntley,R.P., Binns,D., O'Donovan,C. and Apweiler,R. (2009) The GOA database in 2009—an integrated Gene Ontology Annotation resource. *Nucleic Acids Res.*, **37**, D396–D403.
8. Fleischmann,A., Darsow,M., Degtyarenko,K., Fleischmann,W., Boyce,S., Axelsen,K.B., Bairoch,A., Schomburg,D., Tipton,K.F. and Apweiler,R. (2004) IntEnz, the integrated relational enzyme database. *Nucleic Acids Res.*, **32**, D434–D437.
9. Morgat,A., Coissac,E., Coudert,E., Axelsen,K.B., Keller,G., Bairoch,A., Bridge,A., Bougueleret,L., Xenarios,I. and Viari,A. (2012) UniPathway: a resource for the exploration and annotation of metabolic pathways. *Nucleic Acids Res.*, **40**, D761–D769.
10. Bairoch,A., Apweiler,R., Wu,C.H., Barker,W.C., Boeckmann,B., Ferro,S., Gasteiger,E., Huang,H., Lopez,R., Magrane,M. *et al.* (2005) The Universal Protein Resource (UniProt). *Nucleic Acids Res.*, **33**, D154–D159.
11. Smith,C.L. and Eppig,J.T. (2009) The Mammalian Phenotype Ontology: enabling robust annotation and comparative analysis. *Wiley Interdiscip. Rev. Syst. Biol. Med.*, **1**, 390–399.
12. Schindelman,G., Fernandes,J.S., Bastiani,C.A., Yook,K. and Sternberg,P.W. (2011) Worm Phenotype Ontology: integrating phenotype data within and beyond the C. elegans community. *BMC Bioinformatics*, **12**, 32.
13. Engel,S.R., Balakrishnan,R., Binkley,G., Christie,K.R., Costanzo,M.C., Dwight,S.S., Fisk,D.G., Hirschman,J.E., Hitz,B.C., Hong,E.L. *et al.* (2010) *Saccharomyces* Genome Database provides mutant phenotype data. *Nucleic Acids Res.*, **38**, D433–D436.
14. Grumbling,G. and Strelets,V. (2006) FlyBase: anatomical data, images and queries. *Nucleic Acids Res.*, **34**, D484–D488.
15. Bradford,Y., Conlin,T., Dunn,N., Fashena,D., Frazer,K., Howe,D.G., Knight,J., Mani,P., Martin,R., Moxon,S.A. *et al.* (2011) ZFIN: enhancements and updates to the Zebrafish Model Organism Database. *Nucleic Acids Res.*, **39**, D822–D829.
16. Bowes,J.B., Snyder,K.A., Segerdell,E., Jarabek,C.J., Azam,K., Zorn,A.M. and Vize,P.D. (2009) Xenbase: gene expression and improved integration. *Nucleic Acids Res.*, **38**, D607–D612.
17. Ilic,K., Kellogg,E.A., Jaiswal,P., Zapata,F., Stevens,P.F., Vincent,L.P., Avraham,S., Reiser,L., Pujar,A., Sachs,M.M. *et al.* (2007) The plant structure ontology, a unified vocabulary of anatomy and morphology of a flowering plant. *Plant Physiol.*, **143**, 587–599.
18. Robinson,P.N., Kohler,S., Bauer,S., Seelow,D., Horn,D. and Mundlos,S. (2008) The Human Phenotype Ontology: a tool for annotating and analyzing human hereditary disease. *Am. J. Hum. Genet.*, **83**, 610–615.
19. Schriml,L.M., Arze,C., Nadendla,S., Chang,Y.W., Mazaitis,M., Felix,V., Feng,G. and Kibbe,W.A. (2012) Disease Ontology: a backbone for disease semantic integration. *Nucleic Acids Res.*, **40**, D940–D946.
20. Knox,C., Law,V., Jewison,T., Liu,P., Ly,S., Frolkis,A., Pon,A., Banco,K., Mak,C., Neveu,V. *et al.* (2011) DrugBank 3.0: a comprehensive resource for 'omics' research on drugs. *Nucleic Acids Res.*, **39**, D1035–D1041.
21. Davis,M.J., Sehgal,M.S. and Ragan,M.A. (2010) Automatic, context-specific generation of Gene Ontology slims. *BMC Bioinformatics*, **11**, 498.
22. Andreeva,A., Howorth,D., Chandonia,J.M., Brenner,S.E., Hubbard,T.J., Chothia,C. and Murzin,A.G. (2008) Data growth and its impact on the SCOP database: new developments. *Nucleic Acids Res.*, **36**, D419–D425.
23. Punta,M., Coggill,P.C., Eberhardt,R.Y., Mistry,J., Tate,J., Boursnell,C., Pang,N., Forslund,K., Ceric,G., Clements,J. *et al.* (2012) The Pfam protein families database. *Nucleic Acids Res.*, **40**, D290–D301.
24. Gough,J., Karplus,K., Hughey,R. and Chothia,C. (2001) Assignment of homology to genome sequences using a library of hidden Markov models that represent all proteins of known structure. *J. Mol. Biol.*, **313**, 903–919.
25. The UniProt Consortium. (2012) Reorganizing the protein space at the Universal Protein Resource (UniProt). *Nucleic Acids Res.*, **40**, D71–D75.
26. Wilson,D., Madera,M., Vogel,C., Chothia,C. and Gough,J. (2007) The SUPERFAMILY database in 2007: families and functions. *Nucleic Acids Res.*, **35**, D308–D313.
27. Wilson,D., Pethica,R., Zhou,Y., Talbot,C., Vogel,C., Madera,M., Chothia,C. and Gough,J. (2009) SUPERFAMILY—sophisticated comparative genomics, data mining, visualization and phylogeny. *Nucleic Acids Res.*, **37**, D380–D386.

28. Washington,N.L., Haendel,M.A., Mungall,C.J., Ashburner,M., Westerfield,M. and Lewis,S.E. (2009) Linking human diseases to animal models using ontology-based phenotype annotation. *PLoS Biol.*, **7**, e1000247.
29. Hoehndorf,R., Schofield,P.N. and Gkoutos,G.V. (2011) PhenomeNET: a whole-phenome approach to disease gene discovery. *Nucleic Acids Res.*, **39**, e119.
30. Mungall,C.J., Gkoutos,G.V., Smith,C.L., Haendel,M.A., Lewis,S.E. and Ashburner,M. (2010) Integrating phenotype ontologies across multiple species. *Genome Biol.*, **11**, R2.
31. Han,H.J., Russo,J., Kohwi,Y. and Kohwi-Shigematsu,T. (2008) SATB1 reprogrammes gene expression to promote breast tumour growth and metastasis. *Nature*, **452**, 187–193.
32. Velankar,S., Alhroub,Y., Alili,A., Best,C., Boutselakis,H.C., Caboche,S., Conroy,M.J., Dana,J.M., van Ginkel,G., Golovin,A. *et al.* (2011) PDBe: Protein Data Bank in Europe. *Nucleic Acids Res.*, **39**, D402–D410.