Gene Ontology

THE IMPACT OF STANDARDISATION

The Gene Ontology (GO) project is a major initiative to develop a computational representation of biological function. GO provides a set of terms for describing gene product characteristics, and uses these standard terms to annotate genes and proteins. It also provides tools to discover, access and process these annotations.

The Gene Ontology Consortium (GOC) includes 34 groups from around the world that collaborate closely on developing and using the Gene Ontology. EMBL-EBI is a major contributor to this consortium across three key areas managed by distinct service teams; annotation, essential ontology development, tool development & provision. Tools such as Protein2GO, are used by a significant number of the consortium members for adding annotations. It is also used by an extensive range of external researchers in industry and academic institutions.

In this impact case study, we present short extracts from five interviews conducted with external users of the EMBL-EBI GO service. We use these extracts to illustrate how GO is underpinning scientific investment, improving R&D and increasing both productivity and performance.

www.ebi.ac.uk/goa

USER IMPACTS

Roche (Industry: Pharmaceutical and Healthcare)
Roche is a global, research-focused healthcare company headquartered in Switzerland.

"We rely on public databases for our analyses. Downloading data from EBI GO allows us to store the data locally and to run large or batch jobs on our server and GO updates are downloaded roughly every month [one unique visitor,12 downloads per year]. GO data is mostly used to query, for a given gene, the GO annotations in molecular function, cellular component and biological process. This information is part of our gene index tool which is used by around 1400 Roche scientists." *Isabelle Wells, Roche Pharma Research and Early Development.*

Counter-factual benefit: based on information supplied by EMBL-EBI and UK full economic costs, we estimate the cost of creating and maintaining EBI GO is around £498,500 per annum.

Cardiovascular Gene Ontology Annotation Initiative (Public Sector: Gene Annotation Project)
This collaborative project between UCL and EMBL-EBI provides more specific GO terms for, and annotations of, human genes that are implicated in heart development and cardiovascular processes.

"People are starting to use GO for Genome-Wide Association Study variants. In many cases a disease is caused by variants in a lot of different genes, but often these genes are all involved in a common pathway. There may, for example, be many genes required to keep the heart functioning, and the body can tolerate one gene variant within a single essential heart pathway but not (say) six. GO can help identify whether there are common variants causing the same disease by grouping genes together and improving the likelihood of the variants being detected by statistical analysis." *Ruth Lovering, Cardiovascular Annotation Coordinator, Centre for Cardiovascular Genetics, UCL*

Benefit: Enrichment analysis is a major application of the GO, and the addition of specific terms and annotations can increase the sensitivity of these analyses. In an experiment to determine the effect of new cardiovascular annotations it was shown that enrichment of metabolism for a key regulator of blood vessel constriction could be detected only with the additional annotations.

EFFICENCY SAVINGS

3 : 1

For every £1 invested in GO, an estimated £3 of efficiency savings are gained within academic and commercial settings.

HIDDEN REUSE

7000 GB

7000 GB of data was downloaded from GO in 2014. This data is fed into many different biological knowledge bases, allowing GO data to be used by thousands of additional users.

6 **18,525**

AgBase downloaded GO 6 times in 2013, alongside other relevant inputs. Agbase had 18,525 unique visitors in 2013.

12 **32,000**

dictyBase downloaded GO 12 times in 2013, alongside other relevant inputs. It had 32,000 users and 66,000 visits to the dictyBase gene page, which contained the relevant GO information.

Gene Ontology

⬤ USERS IMPACTS Cont.

REUTERS

Thomson-Reuters MetaCore (Industry: Bioinformatics database)

"Within MetaCore, we include both Thomson Reuters' and external ontologies including GO to enable our customers to choose their desired enrichment. Although our own ontologies are used by the significant majority of customers for enrichment and analytics, all entities within the databases are registered to differing GO processes. Some users will use different ontologies at different stages of the study workflow, and particularly academic users may apply the GO ontologies as a way to confer validation prior to external publication." *Gavin Coney and Mark Hughes, Thomson Reuters*

Benefit: Reproducibility or validation of scientific results add a vital assurance of quality for research. GO provides users with a good integration method between various ontologies and different datasets. When Thomson-Reuters supplement that with extra, highly curated information and advanced analytics it creates a lot of value to end users.

AgBase (Public sector database)
AgBase is a curated genomic database containing functional annotations of agriculturally important animals, plants, microbes and parasites.

"Using the EBI biocuration infrastructure means that we don't need to build our own data entry systems. Initial, one-off development of a biocuration interface I would estimate to take about 3 months of programmer time as an initial effort; about 1 week every 2 months afterward to do the quality control. Based upon my experience of data entry and number of papers curated over time, the Protein2GO curation interface increases [bio-curator] productivity by 10-15%..." *Fiona McCarthy, Associate Professor, School of Animal and Comparative Biomedical Sciences, University of Arizona, US*

Benefit: based on further information supplied in the interview and equivalent UK posts and full labour costs we estimate their use of the EMBL-EBI Protein2GO tool and infrastructure provides labour efficiencies to AgBase of around £50,700 per annum.

dictyBase (Public sector database)

dictyBase

dictyBase is an online bioinformatics database for the model organism *Dictyostelium discoideum*

"Protein2GO provides invaluable help to us, a small model organism database. It would have been possible to proceed without Protein2GO but at a high cost, severely delaying other projects that our users waited for, such as adding new genomes to the database." *Petra Fey, Senior Scientific Curator and Stock Center Manager, dictyBase, Northwestern University, Chicago, IL*

Benefit: based on further information supplied in the interview and equivalent UK posts and full labour costs we estimate their use of the EMBL-EBI Protein2GO tool and infrastructure provides labour efficiencies to dictyBase of around £14,300 per annum.

SCALE OF USAGE

2.8 MILLION WEB REQUESTS

There are over 50 different freely available functional analysis tools and 2 major commercial bioinformatics services that include GO annotations in their programs.

DIRECT IMPACT

Estimates of direct impact to interviewed users, using all the information supplied in interview and UK full labour costs, it was estimated that their use of the EMBL-EBI biocuration infrastructure provided labour efficiencies per annum as follows:

AgBase: £50,700

dictyBase: £14,300

Roche: £498,500*

*Estimated cost of developing and maintaining GO