systems. Examples include the possibility of a resurgence in oil price from $30/bbl recently (as of December 2008; http://tonto.eia.doe.gov/dnav/pet/hist/rwtcd.htm) back to and beyond the previous high of $147/bbl (July 2008) in the near future (modeled here as $100 per barrel in both cases), the introduction of more stringent CO_2 emissions targets and carbon trading schemes (potentially rising to ~$200 per ton CO_2 by 2050; refs. 7,8) and the increased demand for food and fuel by a population rising from 6.8 billion in 2009 to 9.4 billion in 2050 (refs. 9,10). All of these factors appear to be strong drivers for the economic viability of this technology. This analysis suggests that although microalgal biofuel systems remain in an early stage of development, they are now approaching profitability if the co-production systems in the base case, and/or the increased productivities in the projected case can be attained. A recent report by Huntley and Redalje[11] estimates that current technology could produce oil for $84/bbl (with no value attributed to the non-oil fraction), with reasonable advancements in technology reducing this cost to $50/bbl or less. This supports our conclusion that co-production is required in the short term and that at increased oil prices (that is, $100 in this model) an IRR of 15% could be obtained.

Considerable synergies also exist between microalgae biofuel production and a wide range of other industries, including human and animal food production, veterinary applications, agrochemicals, seed suppliers, biotech, water treatment, coal seam gas, material supplies and engineering, fuel refiners and distributors, bio-polymers, pharmaceutical and cosmetic industries, as well as coal-fired power stations (CO_2 capture) and transport industries, such as aviation. Sound opportunities therefore exist for the development of a rapidly expanding sustainable industry base whose productivity is independent of soil fertility and less dependent on water purity. Thus, these technologies can conceivably be scaled to supply a substantial fraction of oil demand without increasing pressure on water resources while potentially contributing to food production. Furthermore, as this study was conservatively modeled on published data, excluding subsidies (which are actually commonly used to develop other renewable energy sectors, for example, photovoltaics) and proprietary technologies, it follows that strategic partnerships and government policy decisions will play a large part in facilitating a coordinated scale-up and deployment of these technologies to contribute to future energy security.

Note: Supplementary information is available on the Nature Biotechnology website.

ACKNOWLEDGMENTS

The authors gratefully acknowledge the support of the Australian Research Council, IMBcom, and the economic advice of Liam Wagner.

COMPETING INTERESTS STATEMENT

The authors declare competing financial interests: details accompany the full-text HTML version of the paper at http://www.nature.com/naturebiotechnology/.

Evan Stephens[1], Ian L Ross[1], Zachary King[2], Jan H Mussgnug[3], Olaf Kruse[3], Clemens Posten[4], Michael A Borowitzka[5] & Ben Hankamer[1]

[1]The University of Queensland, Institute for Molecular Bioscience, St. Lucia, Queensland, Australia. [2]IMBcom, St. Lucia, Queensland, Australia. [3]University of Bielefeld, Department of Biology, AlgaeBioTech Group, Bielefeld, Germany. [4]University of Karlsruhe, Institute of Life Science Engineering, Bioprocess Engineering, Karlsruhe, Germany. [5]Murdoch University, School of Biological Sciences and Biotechnology, Algae R&D Center, Murdoch, Western Australia, Australia. e-mail: b.hankamer@imb.uq.edu.au

1. Waltz, E. *Nat. Biotechnol.* **27**, 15–18 (2009).
2. Mascarelli, A. *Nature* **461**, 460–461 (2009).
3. Melis, A. *Plant Sci.* **177**, 272–280 (2009).
4. Weissman, J.C. & Goebel, R.P. *Design and Analysis of Pond Systems for the Purpose of Producing Fuels* (Solar Energy Research Institute, Golden, Colorado, SERI/STR-231–2840, 1987).
5. Benemann, J.R. & Oswald, W.J. *Systems and Economic Analysis of Microalgae Ponds for Conversion of CO_2 to Biomass*. Final Report to the Pittsburgh Energy Technology Center. Grant no. DE-FG22–93PC93204 (1996).
6. Huggett, B. *Nat. Biotechnol.* **26**, 1208–1209 (2008).
7. McFarland, J.R., Reilly, J.M. & Herzog, H.J. *Energy Econ.* **26**, 685–707 (2004).
8. Zhu, Z., Graham, P., Reedman, L. & Lo, T.A. *Decis. Econ. Finance* **32**, 35–48 (2009).
9. Anonymous. *World Population Data Sheet* (Population Reference Bureau, Washington, DC, 2009). <http://www.prb.org/pdf09/09wpds_eng.pdf>
10. Anonymous. *Soziale und Demographische Daten zur Weltbevölkerung* (Deutsche Stiftung Weltbevölkerung, Hannover, Germany, 2009). <http://www.dsw-online.de/pdf/dsw_datenreport_09.pdf>
11. Huntley, M.E. & Redalje, D.G. *Mitig. Adapt. Strategies Glob. Change* **12**, 573–608 (2007).

Ontology engineering

To the Editor:

Gene Ontology (GO)[1] and similar biomedical ontologies are critical tools of today's genetic research. These ontologies are crafted through a painstaking process of manual editing, and their organization relies on the intuition of human curators. Here we describe a method that uses information theory to automatically organize the structure of GO and optimize the distribution of the information within it. We used this approach to analyze the evolution of GO, and we identified several areas where the information was suboptimally organized. We optimized the structure of GO and used it to analyze 10,117 gene expression signatures. The use of this new version changed the functional interpretations of 97.5% ($P < 10^{-3}$) of the signatures by, on average, 14.6%. As a result of this analysis, several changes will be introduced in the next releases of GO. We expect that these formal methods will become the standard to engineer biomedical ontologies.

Every year, over 400,000 new articles enter the biomedical literature[2], creating an unprecedented corpus of knowledge that is impossible to explore with traditional means of literature consultation. This situation motivated the development of biomedical ontologies, structured information repositories that organize biomedical findings into hierarchical structures and controlled vocabularies. GO is arguably the most successful example of a biomedical ontology. GO is a controlled vocabulary to describe gene and gene product attributes in any organism and includes 26,514 terms organized along three dimensions: molecular function, biological process and cellular component. GO has become even more intensively used with the introduction of high-throughput genomic platforms because of its ability to categorize large amounts of information using a controlled vocabulary to group objects and their relationships[1,3,4].

Today, GO and other biomedical ontologies are the result of a painstaking, costly and slow process of manual curation that requires reaching a consensus among many experts to implement a change. Furthermore, the topology of GO has become critically important because of the introduction of gene set enrichment methods. These methods have allowed investigators to characterize the results of a high-throughput experiment in terms of coherent, knowledge-defined sets of genes (e.g., pathways, functional classes or chromosomal locations) rather than in terms of anecdotal evidence about

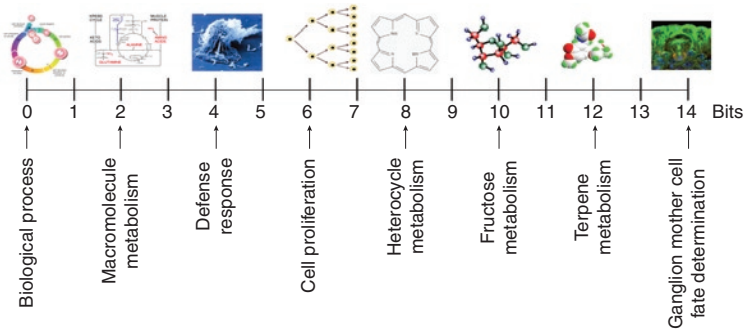

Figure 1 Spectrum of GO terms: examples ranging from 1 to 14 bits.

single genes[5,6]. GO has become a primary provider of these gene sets and researchers use its graphical structure to identify the specificity of a gene class so that they will compare classes of the same specificty[7]. Previous studies have found that the structure of GO does not conform to expected intuitions regarding the structure and distributions of ontology terms[8,9]. Gene enrichment methods typically use the structure of ontologies as a proxy for the specificity of a term[10,11] or, in some cases, use automated procedures to identify structural biases and to compensate for them in the analysis[7,8,12]. Unfortunately, in some cases, even these compensative methods are unable to reach the same conclusions of a well-calibrated ontology (**Supplementary Notes 1**).

The approach we advocate here aims to solve the problem at its root by optimizing the structure of the ontology so that it will indeed be an accurate representation of the informational specificity of any term in the ontology. This approach would not only avoid the necessity to compensate for biases but also improve the semantic transparency of the ontology structure. To do so, we introduce an automated method for engineering the structure of GO based on the information content of each single term. The intuition behind this method is that ontologies are information systems and, as such, they can be optimized using the well-established mathematics of information theory. Given its mathematical nature, this optimization process can be automated, thus producing a principled and scalable architecture to engineer GO and, analogously, other biomedical ontologies.

Our approach starts from the quantification of information contained in the terms of the ontology. The information content of a term is computed from the amount of annotation available for it relative to all other terms, and it is a measure of the surprise caused by labeling a gene with this term rather than with any other term (**Supplementary Notes 2**). For instance, if a term contains all genes, then it is not surprising for a given gene to be labeled with it, so this term does not contain much information. Thus, the more genes or gene products associated with a term, the less specific the term is and the less information is conveyed by it. This 'surprise factor' is called 'self-information', and information theory provides a formal definition for it[13] (**Fig. 1**).

Using information theory, we analyzed the evolution of the information content of GO across time, examining 2 million genes across all the organisms encoded in the ontology annotations. This process highlighted information biases and inefficiencies that may affect the usage of GO and identified those areas of the ontology that were sub-optimally organized. The analysis identified three types of information inefficiencies in the structure of GO.

The first type of inefficiency arises from the variability of the information content among the terms within a given ontology level. By the principle of maximum entropy, an even a priori distribution of information (where all terms in a level are equally specific and hence equally informative) is most efficient because a random experiment is most informative if the probability distribution over outcomes is uniform[13]. Furthermore, gene set enrichment methods often use GO level (that is, distance from the top of the graph) as a proxy for degree of specificity[7,10,11]; this strategy implicitly relies on within-level uniformity of information content. Optimally, then, all the terms in a given level would have equal specificity and, therefore, the same information content. Our analysis revealed that the original version of GO contained a large degree of such intra-level variability of information content. For example, the term 'pilus retraction' was originally at level 2, at the same level of terms like 'cell cycle' and 'cell development' that are actually much more general.

The second type of structural inefficiency, inter-level variability, arises from deviations in information content between levels. In general, terms become more specific as the information content of a level increases with depth in the graph. In some areas of GO, however, the mean information content decreases from one level to the next, creating an information bottleneck. In this case, most of the annotation information of the previous level is transmitted to the next through only a few terms. The larger the decrease in information content, the more severe the bottleneck. The presence of these

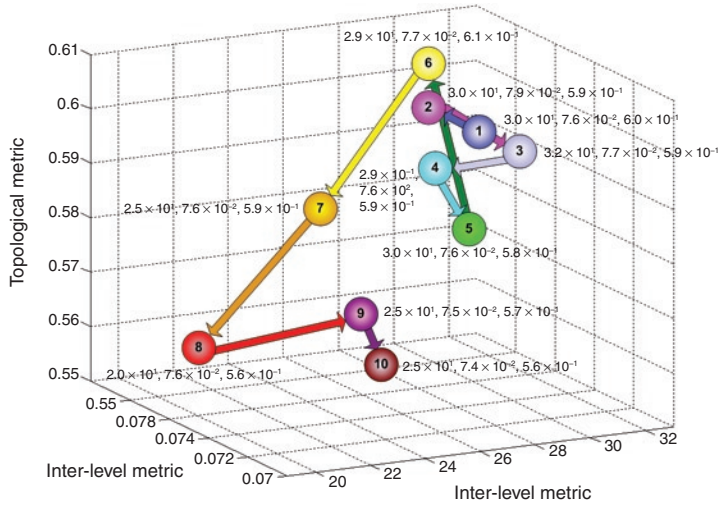

Figure 2 Three-dimensional evolution of GO over ten releases from 2005 to 2007 along the three dimensions of structural inefficiency. An ontology with no inefficiency across these metrics would be at the origin (0,0,0).

areas of suboptimal information distribution violate the assumption of gene set enrichment analysis methods[7,12] that the specificity in GO terms effectively increases from one level to the next (**Supplementary Notes 3**).

The third type of structural inefficiency, topological variability, arises from the suboptimal organization of the branches. The principle of maximum entropy dictates that the closer a topological structure is to uniform, the greater is the information that experiments can derive from it[8]. We used entropy rate to quantify the uniformity of the GO branch structure (**Supplementary Notes 4**) so that a higher entropy rate indicates that the ontology structure is closer to uniform.

We analyzed the evolution of GO along these three dimensions of structural inefficiency using ten releases of GO containing over 2 million unique genes[14]. **Figure 2** plots their structural inefficiencies for each release of GO and illustrates how they have been decreasing over time (**Supplementary Notes 5**). For instance, with time point 8 (February 1, 2007), inter-level variability and topological variability saw substantive improvements, coinciding with introduction of the ['is_a complete'] property in GO[15]. In contrast, intra-level variability saw comparatively modest improvements over the evolution of GO.

One of the greatest dangers of structural inefficiencies in GO is the impact they can have on the functional interpretation of the results of high-throughput experiments. We thus optimized the information distribution of GO by introducing single-level changes and modifying 1,001 relationships and 11% of GO terms, thus significantly improving the overall intra-variability ($P < 10^{-3}$) (**Supplementary Notes 6**).

We used this optimization method to create a modified, improved GO and we compared it to the current GO in the interpretation of 10,117 gene expression signatures from DNA microarray experiments[16]. Each signature contains genes differentially expressed between two biological conditions, and we compared the results of gene enrichment analysis of these signatures obtained by the original and the modified GO. We found that these changes significantly affected the functional interpretations of 97.5% ($P < 10^{-3}$) of the experimental gene signatures and altered the resulting set of GO categories by 14.6% on average (**Supplementary Notes 7**). On the basis of this analysis, we presented 14 recommendations to the GO Consortium and most of these new annotations (12) will be introduced in the next release of GO (**Supplementary Notes 8**).

Finally, as a result of our analysis, we applied this approach to more complicated multi-level structural changes. We suggested the GO Consortium move 12 terms. The terms all underwent the standard curatorial validation of the GO consortium, and 11 of them are now included in the current release of GO. The twelfth term, pigmentation (GO:0043473) had few annotations at the time but was not moved as it was expected that many more genes would be annotated with that term in the future.

The most striking result of our experiment was to show the convergence of mathematical optimality and biological validity and that a formal, automated analysis is able to uncover sound biological information hidden in the structure of the ontology. By altering the ontology itself, our approach improves gene enrichment results in ways that cannot be obtained by simply changing the underlying gene enrichment method (**Supplementary Notes 1**).

Our analysis also reveals that GO contains more information than is currently used. By optimizing the distribution of information within GO, our method can be used to aid the design of more efficiently organized knowledge repositories—leading to a more effective use of biological information. This method is already being used to achieve this aim by the GO Consortium and other ontologies, such as the Phenotypic Quality Ontology (PATO)[17] in the OBO Foundry[18]. We expect that formal and automated methods will become the standard for the engineering of biomedical ontologies.

Note: Supplementary information is available on the Nature Biotechnology website.

ACKNOWLEDGMENTS

This work was supported in part by the National Library of Medicine (NLM/NIH) under grants 1K99LM009826 and 5T15LM007092 and by the National Human Genome Research Institute (NHGRI/NIH) under grants 2P41HG02273, 1R01HG003354, and 1R01HG004836. The authors are grateful to the anonymous reviewers for their helpful suggestions.

COMPETING INTERESTS STATEMENT

The authors declare no competing financial interests.

Gil Alterovitz[1–3], Michael Xiang[1,2], David P Hill[4], Jane Lomax[5], Jonathan Liu[6], Michael Cherkassky[2], Jonathan Dreyfuss[1,2], Chris Mungall[7], Midori A Harris[5], Mary E Dolan[4], Judith A Blake[4] & Marco F Ramoni[1,2]

[1]Children's Hospital Informatics Program, Harvard-MIT Division of Health Sciences and Technology, Harvard Medical School, Boston, Massachusetts, USA. [2]Partners Healthcare Center for Personalized Genetic Medicine, Boston, Massachusetts, USA. [3]Department of Electrical Engineering and Computer Science, Massachusetts Institute of Technology, Cambridge, Massachusetts, USA. [4]Jackson Laboratory, Bar Harbor, Maine, USA. [5]EMBL-EBI, Wellcome Trust Genome Campus, Hinxton, UK. [6]Department of Biology, Massachusetts Institute of Technology, Cambridge, Massachusetts, USA. [7]Lawrence Berkeley National Laboratory, Berkeley, California, USA.
e-mail: gil_alterovitz@hms.harvard.edu or marco_ramoni@harvard.edu.

1. Ashburner, M. *et al. Nat. Genet.* **25**, 25–29 (2000).
2. Davis, D.A., Ciurea, I., Flanagan, T.M. & Perrier, L. *Med. J. Aust.* **180**, S68–S71 (2004).
3. Camon, E. *et al. Nucleic Acids Res.* **32**, D262–D266 (2004).
4. Harris, M. *et al. Nucleic Acids Res.* **32**, D258–D261 (2004).
5. Subramanian, A. *et al. Proc. Natl. Acad. Sci. USA* **102**, 15545–15550 (2005).
6. Doniger, S.W. *et al. Genome Biol.* **4**, R7 (2003).
7. Al-Shahrour, F., Diaz-Uriarte, R. & Dopazo, J. *Bioinformatics* **20**, 578–580 (2004).
8. Alterovitz, G., Xiang, M., Mohan, M. & Ramoni, M.F.G.O. *Nucleic Acids Res.* **35**, D322–D327 (2007).
9. Ogren, P.V., Cohen, K.B. & Hunter, L. *Pac. Symp. Biocomput.* 174–185 (2005).
10. Dennis, G. Jr. *et al. Genome Biol.* **4**, 3 (2003).
11. Zhou, M. & Cui, Y. *In Silico Biol.* **4**, 323–333 (2004).
12. Raychaudhuri, S., Chang, J.T., Sutphin, P.D. & Altman, R.B. *Genome Res.* **12**, 203–214 (2002).
13. MacKay, D.J.C. *Information Theory, Inference, And Learning Algorithms*, xii (Cambridge University Press, Cambridge, U.K.; New York, 2003).
14. Wu, C.H. *et al. Nucleic Acids Res.* **34**, D187–D191 (2006).
15. The Gene Ontology Consortium. *Nucleic Acids Res.* **36**, D440–D444 (2008).
16. Yi, Y., Li, C., Miller, C. & George, A.L. Jr. *Genome Biol.* **8**, R133 (2007).
17. Gkoutos, G.V. *et al. Comp. Funct. Genomics* **5**, 545–551 (2004).
18. Smith, B. *et al. Nat. Biotechnol.* **25**, 1251–1255 (2007).