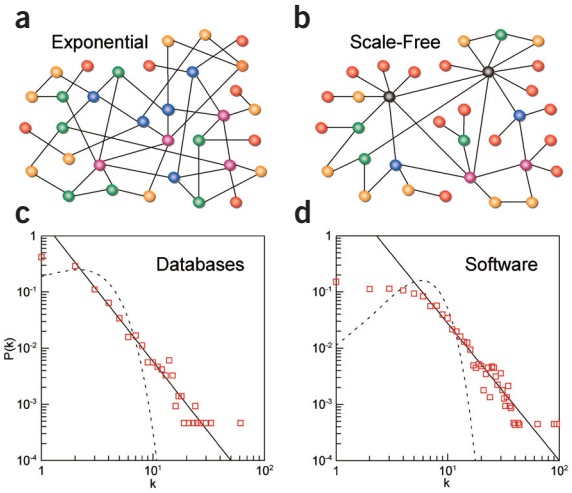

Figure 1 Network topologies and information systems. (**a**) An example of exponential network topology. With exponential networks, typical of classic random graph models, connections between any two nodes are equally probable, thus the number of connections to each node tends to cluster around the average. (**b**) An example of scale-free network topology. Many natural networks tend to adopt this form, in which most nodes have fewer than the average number of connections, while a few 'hub' nodes have a large number of connections. Scale-free networks, as seen in metabolic pathways and protein interactions, are typically more robust and scaleable than exponential networks. (**c**) Degree distributions of database schemas. Power law distributions, characteristic of degrees of connectivity of nodes in scale-free networks, are linear when plotted on log-log axes. Here, the number of connections *k* is plotted against the fraction *P(k)* of nodes of that degree. 30 database schemas with a total of 2,149 tables were analyzed. Solid lines indicate a scale-free model of the data fitted to the tail of the distribution, while for comparison dashed lines show a Poisson distribution, which would be characteristic of an exponential topology, for the observed average node degree of 2.64. (**d**) Degree distributions of software dependency graphs. Fourteen software systems with 2,236 code modules were analyzed, producing an average node degree of 6.34. In both cases, clear power-law tails indicate the presence of subpopulations of highly interconnected hubs, suggesting a natural tendency to scale-free structuring of information systems. A list of the systems analyzed, raw data counts and details of the methodology may be found in **Supplementary Tables 1–4** online.

would predict that the same behavior would extend in self-similar fashion to the next 'layer;' that of individual databases and programs. Can such a hypothesis be tested?

Software and databases are typically designed and analyzed using certain graph-based representations as engineering tools. Relational databases are depicted in schema diagrams, often based on conceptual 'entity-relationship' domain models, which connect database tables referencing each other. Similarly, the patterns by which software modules invoke each other in programs may be captured in 'dependency graphs.'

By searching the Web, I collected a number of 'real world' examples of dependency graphs and database schemas that are likely to have benefited not only from an expert design process but also from trial and optimization (**Supplementary Tables 1-4** online). For a wide variety of scientific, engineering, and business databases and applications, I counted and aggregated degrees of nodes with no attempt to apply any semantic criteria, filters, or rationales other than simple connectivity.

The resulting degree distributions indeed exhibit scale-free characteristics, in the form of hallmark power-law 'tails' that clearly diverge from Poisson exponential decay (**Fig. 1c,d**). This observation would suggest that, in one respect at least, engineering imitates life.

How might information systems have evolved in such a direction? Scale-free structuring of relational databases may in part reflect a trend to 'star schemas,' which connect many tables to central hubs for most efficient mining of data warehouses in decision-support applications[4]. Similarly, scale-free software architecture might be traced back to the 'structured programming' movement, which pointedly eschewed 'spaghetti code' that resulted from unconstrained jumps in flow-of-control, in favor of strictly modular, hierarchical programs that were more comprehensible and scaleable[5].

A 'structured integration' philosophy would resist the creation of spaghetti data that could ensue from simply connecting individual data sources indiscriminately. Instead, it would facilitate successive coalescence of natural clusters of data sources, taking advantage of their locality around hubs to focus attention and expertise optimally within a hierarchical framework. Network theory can offer mathematical metrics for evaluating architectures for robustness to the inevitable turnover of data sources, while guiding the principled insertion of new ones[2]. Thus, function would follow form—perhaps the better dictum in undertaking data integration.

ACKNOWLEDGMENTS

I thank J. Aaronson, P. Buneman, R. Reichard, K. Tatsuoka and N. Odendahl for helpful comments.

David B. Searls

Bioinformatics Division, Genetics Research, GlaxoSmithKline Pharmaceuticals, 709 Swedeland Road, PO Box 1539, King of Prussia, Pennsylvania 19406, USA.
e-mail: David_B_Searls@gsk.com

1. Barabási, A.-L. *Linked: The New Science of Networks* (Perseus, Cambridge, MA, 2002).
2. Albert, R. & Barabási, A.-L. *Rev. Modern Phys.* **74**, 47–97 (2002).
3. Adamic, L.A. *et al. Science* **287**, 2115 (2000).
4. Levene, M. & Loizou, G. *Info. Sys.* **28**, 225–240 (2003).
5. Dijkstra, E. *Comm. Assoc. Comput. Machinery* **11**, 147–148 (1968).

Genomic databases yield novel bioplastic producers

To the editor:

Biodegradable plastics have been proposed as environmentally friendly alternatives to synthetic polymers, which often pose problems with disposal and persistence in the field. Polyhydroxyalkanoates (PHAs) are a group of biodegradable plastics produced by several microorganisms (*e.g.*, *Alcaligenes eutrophus, Pseudomonas oleovorans*) under nutritional stress[1]. Because of their high-production cost compared with synthetic plastics, a search is on to identify organisms capable of PHA production at high levels. Conventional methods for searching microbes for a particular characteristic demand high inputs of time and energy. On the other hand, simple bioinformatics searches that integrate multiple sources of data offer a faster and more rapid means of identifying new PHA producers[2].

We have carried out sequence analyses of 89 complete and 34 partially sequenced genomes in an attempt to identify domains of PHA genes using RPS-BLAST[3]. Our study reveals 13 putative PHA producers, belonging to 12 genera (**Table 1**). All the potential PHA-producing organisms we report here possess three key enzymes (β-ketothiolase, acetoacetyl Co-A and PHA synthase), except *Delftia*. Similarly, literature and National Center for Biotechnology Information (Bethseda, MD, USA) databases reveal organisms, such as *Rhodospirillum* that have alternative enzymes[1] and *Pseudomonas* spp. that lack acetoacetyl coenzyme A but still have the ability to produce PHA.

Among the identified organisms that look most promising for PHA production are *Microbulbifer degradans*, *Burkholderia fungorum, Novosphingobium aromaticivorans* and *Rhodopseudomonas palustris,* which have the ability to degrade a range of carbon compounds with low nitrogen content. It seems likely that these organisms can rather more easily divert their metabolism towards PHA formation. They assume higher significance since they (except *Microbulbifer*), also have the potential to produce hydrogen, a clean fuel[4].

Several novel organisms identified here (**Table 1**) have the ability to use a wide range of industrial wastewater, have magnetotactic ability, and dehalogenate and degrade environmental pollutants. This raises the intriguing possibility that they could be exploited to both breakdown wastes and produce PHAs.

ACKNOWLEDGMENTS

We are thankful to S. K. Brahmachari, Director, Institute of Genomics and Integrative Biology, CSIR for providing the necessary facilities and moral support.

Vipin Chandra Kalia, Ashwini Chauhan, Goutam Bhattacharyya & Rashmi

Institute of Genomics and Integrative Biology, CSIR, Delhi University Campus, Mall Road, Delhi 11 00 07, India.
e-mail: vckalia@igib.res.in and vc_kalia@yahoo.co.in

1. Madison, L.L. & Huisman, G.W. *Microbiol. Mol. Biol. Rev.* **63**, 21–53 (1999).
2. Caburet, S., Conti, C. & Bensimon, A. *Trends Biotechnol.* **20**, 344–350 (2002).
3. Altschul, S.F. *et al. Nucleic Acids Res.* **25**, 3389–3402 (1997).
4. Kalia, V.C., Lal, S., Ghai, R., Mandal, M. & Chauhan, A. *Trends Biotechnol.* **21**, 152–156 (2003).

Table 1 Potential polyhydroxyalkanoate producing organisms and their unique properties

Microorganism	β keto thiolase	Acetoacetyl-CoA reductase	PHA synthase	Properties
Agrobacterium tumefaciens C58 (Cereon/University of Washington)	+	+	+	Nonpathogenic bacterium already commonly applied in plant biotechnology.
Burkholderia fungorum[a]	+	+	+	Pollutant degrading capabilities, beneficial plant root colonizer, role in global C-cycle, commercially important for bioremediation.
Delftia acidovorans	+	–	+	Facultative pathogen grows preferably on organic acids as carbon source under aerobic and micro-aerobic conditions.
Magnetospirillum magnetotacticum[a]	+	+	+	Micro-aerophilic bacteria exhibit magnetotaxis. Found in both fresh and salt water.
Mesorhizobium loti	+	+	+	Nitrogen-fixing capacity.
Microbulbifer degradans 2-40[a]	+	+	+	Degrades at least 11 relatively insoluble complex polysaccharides, including agar, chitin, alginic acid, celllose, β-glucan, pectin, starch and xylan. Widely used for bioremediation.
Novosphingobium aromaticivorans[a]	+	+	+	Degrades aromatic hydrocarbons, including toluene, *p*-cresol, xylene, naphthalene, biphenyl, dibenzothiophene and fluorine. Ability to grow in a wide range of environments including soil, both marine and fresh waters.
Ralstonia solanacearum	+	+	+	Causative agent of Moko disease of banana.
Rhodopseudomonas palustris[a]	+	+	+	Degrades and recycles variety of aromatic compounds like lignin and toxic pollutants like chlorobenzoate. Capable of producing hydrogen.
Rickettsia prowazekii	+	+	+	Causative agent of epidemic typhus. Contains complete set of genes encoding component of tricarboxylic acid cycle and respiratory-chain complex. Small genome size (1.1×10^6 bp). No genes required for anaerobic glycolysis.
Vibrio cholerae	+	+	+	Agent of cholera. Relatively simple growth factor requirements and grows on glucose as a sole carbon and energy source.

[a]Unfinished genomes