

Fuelling discovery by sharing

The United States Materials Genome Initiative aims at accelerating the discovery, development and deployment of materials. Yet, finding data standards and sharing practices that can be leveraged by the disparate communities in materials science and technology may prove difficult.

Gene-sequencing, mass-spectrometry and drug-screening technologies, aided by the increasingly faster and cheaper computational power available, have greatly accelerated the discovery of drugs and the development of better diagnoses and therapies. Indeed, genomics and proteomics are arguably the most recent and high-impact successful stories made possible by synergetic collaborations between academia, government and industry (exemplified by the Human Genome and Proteome projects¹), and have enabled enormous progress in biology, medicine and healthcare.

Is it possible to spur similar rapid scientific and technological progress in materials science and technology by taking advantage of what has been learned in genomics and other biomics projects? The US Government certainly believes it is worth finding out. In June 2011, President Barack Obama announced the Materials Genome Initiative (MGI), a multi-year effort coordinated by the Office of Science and Technology Policy, and involving government agencies, universities, national laboratories and industry (more than 30 academic institutions and companies have committed to pursue collaborative efforts according to the vision of the MGI; ref. 2). Committed investments reached \$63 million in 2012, and the federal government has requested \$100 million for this year. The MGI's vision is to at least double the pace of discovery, development and deployment of new materials and related technologies (the time from discovery to application has been 10–20 years; Fig. 1), and bring them to market at a fraction of the cost³.

This is a lofty goal that could potentially bring enormous socioeconomical benefits. But the path ahead will probably be tortuous, and the obstacles could be significant. On the one hand, it is not clear that imitating the strategies for data standards and analysis used in genomics or proteomics will be all that helpful. The configuration space of genes and proteins lies in the myriad combinations of repetitive instances in the small dataset of 4 nucleotides and 20 amino acids. Instead, the components of alloys, topological

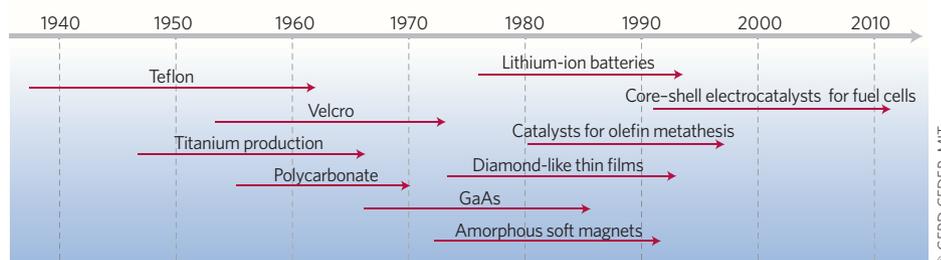


Figure 1 | Time frame from discovery to application for a few technologically important materials.

insulators, catalysts, superconductors, metal–organic frameworks, (bio)polymer composites or synthetic tissues can be made from hundreds — if not thousands or millions — of compounds, and be mixed at varying stoichiometries. Also, materials are structurally much more diverse, their properties may change with external conditions, time and size, and their relevant functionality can be of mechanical, chemical, biological, magnetic, electronic or optical nature. Therefore, for materials the space of possible relevant states and properties is, in general, vastly heterogeneous, and the type and amount of data necessary for their description is very much material dependent.

On the other hand, whereas in biomics efforts are focused on ‘finding’ (genes, proteins, biochemical interactions or functions), in materials the main job lies in ‘designing’ and/or ‘optimizing’ (composition, structure, properties or functions). Moreover, the spread of relevant time and length scales for some materials can be big. For instance, both the nano- and microstructure of metals can dramatically affect their macroscopic creep or, when used as biomedical implants, their resistance to wear and thus their lifetime. Furthermore, whereas genes in DNA and structural motifs in proteins are obvious descriptors of the functions of molecules made of nucleic and amino acids, what descriptors are best for the function of materials used in, for instance, solar cells, batteries, turbines or implants, is not immediately apparent.

In fact, as Stefano Curtarolo and colleagues argue in a Review in this issue⁴, finding appropriate and computationally fast

descriptors will be a key step in the quest for accelerating the discovery of novel materials. The researchers discuss recent advances in high-throughput computational materials design for a broad range of materials, and highlight the need for integrated and standardized communication protocols between data repositories.

This is perhaps the biggest challenge. It may be hard to find data standards and sharing practices that can be leveraged by the very disparate and monolithic communities in materials science and technology. A few recently created initiatives for data sharing and analysis, such as the Materials Project⁵ or the quantum materials AFLOWLIB.org repository⁶, testify the needs and goals of the communities involved. And it may well be that the nucleation of small and diverse community-driven initiatives, rather than any top-down, one-size-fits-all approach, ends up becoming more widespread.

There is, however, at least one important aspect that should be copied from the genomics story: the persistent efforts to share knowledge and data. In fact, the MGI intends to put incentives in place to encourage researchers from separate disciplines to join forces with industrial partners and government agencies to build shared and integrated platforms. And even though the journey ahead is uncertain, we are off to a good start. □

References

- <http://www.hupo.org/research/hpp>
- <http://www.tms.org/orlandoprinciples>
- <http://www.whitehouse.gov/mgi>
- Curtarolo, S. *et al. Nature Mater.* **12**, 191–201 (2013).
- <http://materialsproject.org>
- <http://afowlib.org>