

The status of structural genomics



‘Structural genomics has been heralded as the follow on to the human genome project.’

Philip E. Bourne, Director,
Integrative Biosciences,
San Diego Supercomputer Center

Structural genomics has been heralded as the follow on to the human genome project. We interpret that to mean a large-scale project with scientific, engineering and technological components and the potential to have a large impact on the life sciences. A project that goes beyond the blueprint for life to the buildings defined by that blueprint — the three-dimensional protein structures. Whereas the human genome project was relatively well defined, with the aim to sequence the three billion nucleotides comprising the human genome, what constitutes structural genomics is open to different interpretations. To some, it aims to characterize all the protein structures in a given genome — *Arabidopsis thaliana*, *Thermotoga maritima* and *Mycobacterium tuberculosis* are examples under scrutiny. To others, the goal of structural genomics is to provide sufficient coverage of fold space to facilitate accurate homology modeling of the majority of proteins of biological interest. As structure determination has already taught us so much about biological function when undertaken as a functionally driven initiative, undertaking structure determination in a broader genomic sense is also likely to bring significant new understanding of living systems. Further, it is likely to lead to advances in the process of structure determination, whether by X-ray crystallography or NMR. With such promise, and with some projects already in their third or fourth year, an obvious question is, how are we doing?

How are we doing?

This has proven to be a somewhat controversial question. An initial report in *Science* [1] implied that the number of structures produced as of November 2002 was minimal. A response from the US Northeast Structural Genomics

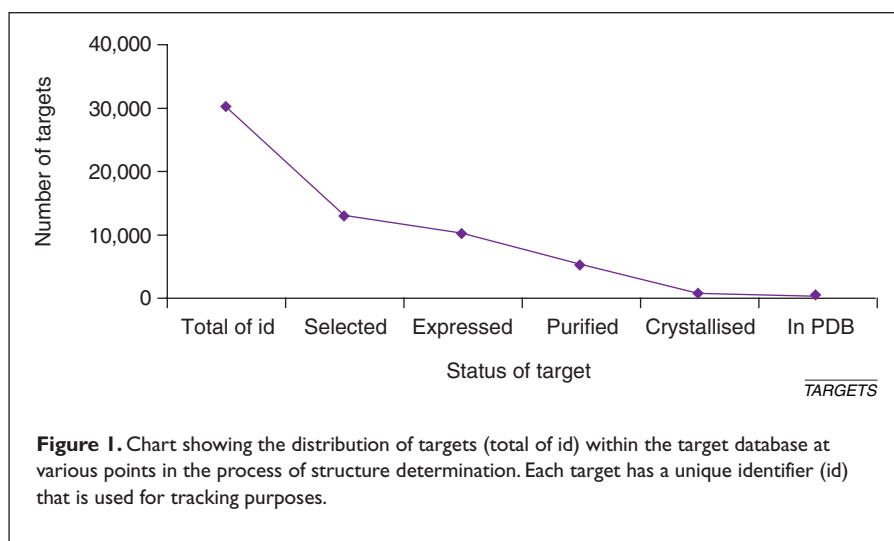
Consortium (NESG) [2] indicated it was early in the process and that the absolute number of structures produced may not be the best measure, but rather it is the value of those structures that is important. NESG indicated that a structure containing a novel fold would indeed provide a new template from which many sequences could be related and hence was a significant contribution. It is not my intent here to join this argument, but simply to point readers at some quantitative data from which they can draw their own conclusions.

Reporting progress

An important feature of structural genomics, laid out by the National Institutes of Health (NIH) as part of the awards made to the pilot centers engaged in this high-throughput structure determination, was the importance of reporting their progress on a regular basis. The 16 pilot centers in the US and worldwide do this by way of weekly updates made available through their individual centers and collated by the Protein Data Bank (PDB) into what is known as the target database (<http://targetdb.rcsb.org>) [3].

In the past year (May 1 2002 to May 31 2003) 316 structures resulting from structural genomics efforts were deposited with the PDB. At the same time, a total of 3324 structures were deposited with the PDB; thus, structural genomics is currently contributing approximately 10% of structures to the field of structural biology. The number of structures at each stage in the pipeline is shown in Figure 1.

Is this percentage likely to increase in the near future? To answer this question requires that we review the number of targets under consideration and the status of those targets. These data are available at <http://spam.sdsc.edu/sgtdb>, a version of the target database that includes multiple theoretical models for targets being determined by structural genomics. As pointed out by NESG, these data must also be considered in context. If one looks at the targets under scrutiny by each group it is clear that the numbers are highly variable. It is not necessarily that one group is more active than another, but rather that each group uses different criteria to define an active target. Without tighter definitions the data remains nothing more than a useful indicator at best. Moreover, the database contains no indication of how individual centers approach the structure determination of a target. One group might rigorously pursue a difficult target, whereas another group will leave it and look for easier candidates. What is clear is that



equal to 30% identity. The value of 30% sequence identity is important because, as a rule of thumb, two proteins that are this similar in primary sequence are likely to have the same fold. Again there are multiple interpretations of this information. Individual groups are either operating without regard for other groups, or there is interest in the same targets by different groups, perhaps indicating some important functional significance for a particular target.

Concluding remarks

Structural genomics is still in the prototype stage, so reading too much into

across all projects there is not a common bottleneck to structure determination (see Figure 2). By tracking the progress of all targets through the structural genomics pipeline over time, it is apparent that the steps that define the process of structure determination take approximately the same amount of time.

Target significance

Another question of interest is what is the significance of the targets themselves? A review of the over 30,000 targets in the target database indicates that 13% are duplicates (that is identical in primary protein sequence). This duplication increases to 38% for sequences with greater than or

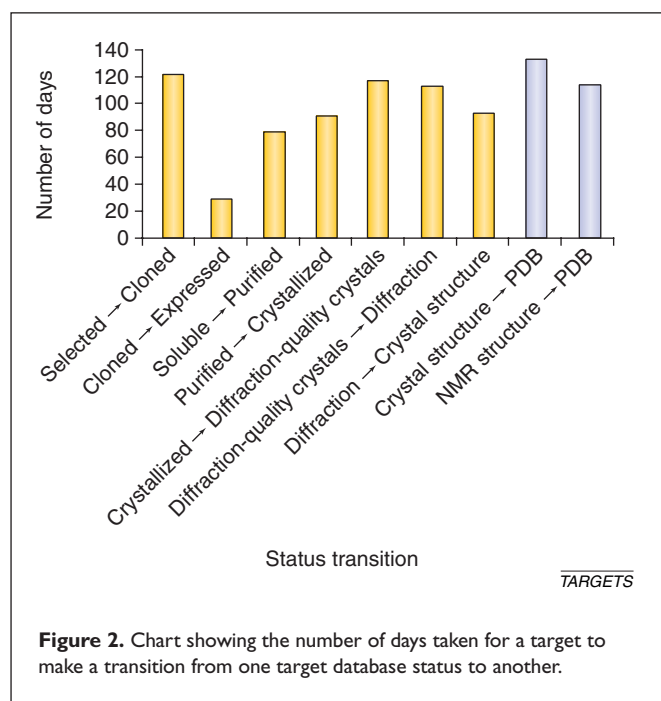
these early indicators of progress is not constructive. Cynics would say that structural genomics will contribute relatively little on a cost basis when compared with functionally driven structure determination. In short, simply repeating a large number of similar structures to achieve numbers, rather than spending time on the difficult structures that often come from much sweat and human input, will have less reward. Advocates will say that significant success has already been achieved, that the promise is there, and that the engineering needed is now coming together. Only time will tell, and we will continue to monitor this progress via the database found at <http://spam.sdsc.edu/sgtdb> as scientists interested in the wonders that macromolecular structures reveal continue their work.

Acknowledgements

The sgtdb database is the work of Charles Allerston, Philip Bourne, Li Chen, Iddo Friedberg, Zoubin Ghahramani, Adam Godzik, Werner Krebs, Wilfred Li, Tong Liu, Ilya Shindyalov, John Westbrook and David Wild.

References

- 1 Service, R.F. (2002) Tapping DNA for structures produces a trickle. *Science* 298, 948–950
- 2 Gerstein, M. *et al.* (2003) Structural genomics: current progress. *Science* 299, 1663
- 3 Westbrook, J. *et al.* (2003) The Protein Data Bank and structural genomics. *Nucleic Acids Res.* 31, 489–491



Philip E. Bourne,
 Director, Integrative Biosciences,
 San Diego Supercomputer Center,
 University of California,
 9500 Gilman Drive, La Jolla, San Diego, USA
 e-mail: bourne@sdsc.edu