Protein Cell 2012, 3(5): 321–322
DOI 10.1007/s13238-012-2022-9

Protein & Cell

NEWS AND VIEWS

Protein construct optimization: data sharing strategy

Karen M. Polizzi[1], Rivka L. Isaacson[2] ✉

[1] Center for Synthetic Biology and Innovation, Division of Molecular Biosciences, Imperial College London, London SW7 2AZ, UK
[2] Center for Structural Biology, Division of Molecular Biosciences, Imperial College London, London SW7 2AZ, UK
✉ Correspondence: rivka.isaacson@imperial.ac.uk

The British Medical Journal devoted its first issue of 2012 to the problem of unreported clinical trial data, highlighting the serious impact that incomplete information can have on world health (for an editorial summary see Lehman and Loder, 2012). Adopted innovations in clinical intervention, inappropriately based on only a subset of trial data, can, at best, prove uneconomical and, at worst, lead to large-scale patient death. While its consequences may be most severe in medicine this problem applies to many scientific disciplines and we are now seeing a move towards greater transparency in data declaration. The Protein Data Bank (PDB), for example, requires (since 2011) that deposition of coordinates of solved nuclear magnetic resonance (NMR) structures of macromolecules is accompanied by concurrent deposition of NMR chemical shift data in the Biological Magnetic Resonance Bank (BMRB). Hence both are already prerequisites for publication in reputable journals, which inevitably require a PDB accession number.

In many cases the failure to report complete data reflects the fact that, in most fields, there is no forum in which negative conclusions and failed experiments are considered sufficiently interesting to be publishable. Even the new generation of 'negative results' journals (such as the *Journal of Negative Results in Biomedicine*, *The All Results Journals* and the *Journal of Interesting Negative Results*) specify in their instructions to authors that the results submitted, while negative, must also be interesting or novel. Boring, inconclusive or failed experiments, as well as the raw data from successful experiments, if made available, can save other researchers vast amounts of time but they are rarely reported.

Researchers in many fields such as synthetic biology, genetics and biochemistry often need to study sub-sections of proteins rather than the whole, either, because they are trying to dissect out the functions of different parts of the protein, or, because the limitations of current technology can't cope with the large size of some proteins. In particular, most techniques for structural study of proteins, including X-ray crystallography, NMR spectroscopy, electron microscopy (EM), circular dichroism (CD) and mass spectrometry (MS), at some point, require the truncation of a protein into its component sub-sections or domains for ease of handling, to overcome data acquisition limitations or to map its topology and interactions.

Defining domain boundaries is challenging and often relies on structure prediction algorithms such as Phyre (Kelley and Sternberg, 2009), or identification, by MS, of limited proteolysis products of the full-length protein, which often cannot define the exact ends. In practicality there is a lot of guesswork involved and researchers commonly try a variety of boundaries in the hope that at least one domain folds up correctly and reflects the protein's true physiological state. There are many ways to check this, such as CD, NMR or monitoring the proteins behaviour *in vivo* if the truncation were made in yeast, for example.

Small variations in construct design can make the difference between a physiologically correct fold and a protein that fails to express, rapidly degrades or irreversibly aggregates into inclusion bodies. Expression and solubility aside, a few too many amino acids can render a protein too flexible to crystallize whereas too few can cause fraying of secondary structure elements. Thanks to advanced cloning procedures it is now a trivial matter to manufacture domains of proteins at will. When the resources are at hand, such as in the well-funded structural genomics consortia, ten or more constructs for every domain are routinely tested. One of the most downloaded and cited papers from the journal *Protein Expression and Purification* discussed the benefits of casting the net wide when it comes to construct design (Gräslund et al. 2008). This is now becoming a reality for even the smallest

of labs with services such as the Oxford Protein Production Facility UK providing expertise and equipment for high throughput cloning and expression free of charge.

Despite the many constructs that are manufactured, we tend only to hear the details of successfully designed protein domains as these are the ones that make it into scientific literature. Unfortunately there is no catalogue for the years of wasted effort researchers put into producing slightly varied protein fragments that turn out to be of limited scientific use. A colleague recently made 15 viable shorter versions of an unstable protein only to discover that none of them interacted with the original binding partner. They are thus useless in his present study but may be valuable for alternative applications or other scientists in his field who have no way of knowing they exist. Access to this information would be helpful on several counts both to researchers, who could avoid hours of work repeating failed experiments, and to bioinformaticians or molecular dynamics simulators.

With the vast profusion of online databases it is unfortunate that no web-based repository exists for protein sub-domain or construct design attempts, regardless of the degree of success. This information, on what has and hasn't worked, could provide a unique resource for scientists in all areas of protein research, particularly as fuel for the burgeoning fields of systems and synthetic biology. The engineering approach adopted by synthetic biologists is a key area as it often requires the construction of modular proteins not encountered in nature. Examples of these include synthetic, two-component, signalling proteins composed of the sensory domain of one protein fused to the signalling domain of another, or fusion proteins which are active in multiple networks. Although there is scope for variation in the way these domains are fused together, each specific case will require a particular flexibility or rigidity in the linking region to ensure appropriate interplay while maintaining the integrity and relative functions of the components.

Analysis of subtly different domain ends could help to calculate why those few amino acids make such a practical difference, and potentially enable the prediction of more accurate boundaries for future experiments. Comparing subtly different constructs or even just the domains' loose ends in terms of their secondary structure, sequence conservation, surface accessibility, contact order and other measures that depend on structure (predicted or experimentally solved) should provide new insight into what makes a successful construct. Inputting all the favourable and unfavourable changes into a machine learning system like a support vector machine or neural network is another way to predict favourable protein fragments.

A web-based repository for construct design could provide fields to input data on construct design including basic domain boundaries and sequence information as well as more qualitative information on the success of the finished protein. Depending on the scope of the application, sample questions might include: Was it expressable in *E. coli* or any other expression system? Were different expression temperatures tried? Was the protein soluble? If not, was it possible to refold it from denaturing conditions? How stable was the protein? By which methods was this established? Did it aggregate or precipitate over time? What was its oligomeric state? Did it crystallize? Were any other parameters established such as melting temperature or helical content? Upload any spectra you may have recorded, stating the technique used. What made you abandon this construct?

At a later stage, the website could be expanded to include other useful 'failures' along the path of synthetic biology, genetics, biochemistry or molecular biology research. It is failure or 'noise' that drives evolution and many processes in science and technology are modelled after evolution, i.e. the 'engineering cycle' and other design methodologies whose iterations represent stylized trial and error. This web resource could be expanded to contain all kinds of negative results; the sort you sometimes hear about in talks but can never revisit as the information is rarely published. The potential for trouble-shooting benefits to experimentalists is immense.

The question of how to incentivize researchers to input their failed construct design attempts would need addressing. There are many examples such as the Critical Assessment of protein Structure Prediction (CASP) competition where experimentalists and theoreticians cooperate for no financial gain in the interests of scientific progress. Perhaps deposition of data could come with some sort of benefit such as entry into a prize draw or access to protein design services that ultimately arise from the resource.

From a social science perspective such a database could launch social and cultural investigations of a 'failure tolerant society' and explore how failure-driven learning attitudes can spur innovation and motivate much of current scientific research. Examining how this can be communicated and captured into a database service could have implications for learning/experiment design and education. The authors offer this gap in the market as a challenge to the readers of *Protein and Cell* and hope to see it filled in the near future.

REFERENCES

Gräslund, S., Sagemark, J., Berglund, H., Dahlgren, L.G., Flores, A., Hammarström, M., Johansson, I., Kotenyova, T., Nilsson, M., Nordlund, P. and Weigelt, J. (2008). The use of systematic N- and C-terminal deletions to promote production and structural studies of recombinant proteins. Protein Expr Purif 2, 210–221.

Kelley, L.A. and Sternberg, M.J.E. (2009). Protein structure prediction on the web: a case study using the Phyre server. Nat Protoc 4, 363–371.

Lehman, R. and Loder, E. (2012). Missing clinical trial data: A threat to the integrity of evidence based medicine. BMJ 344, d8158.

