Annotation jargon – it's not too late to correct it

This is the era of power sequencing[1]. After completion of the genome sequencing project of an organism, we are interested in the decoding of information hidden in the sequence of bases. The decoding process known as annotation helps in the interpretation of genomic data. Often *in silico* methods have been used to annotate newly sequenced genomes and sequence similarity searches serve as a primary approach to assign function to such a sequence (genome/gene/protein). On the basis of these approaches, one assigns putative function to a sequence of interest, which can be further validated using wet lab experiments to confirm the exact function. To properly store and retrieve, the sequences (annotated/unannotated) are deposited in various biological databases. Most of these databases (NCBI, DDBJ, EMBL, etc.) are public repositories and can be accessed worldwide via the internet. Therefore, irrespective of the geographical location any user (researcher/academician/student) can easily retrieve desired information available in the database based on either the accession number of sequences or by selecting some other criteria, e.g. mainly the keyword search. For a biologist, a keyword may be the name of an organism or a gene/protein or gene/protein function; the name of a pathway or any other biological information provided by the submitter. However, the problem arises when a user is unable to retrieve all the desired information based on the keyword provided. There are no uniform standards/guidelines to assign such keywords while assigning the function. Thus, a sequence submitter is free to use his/her own convention to write gene/protein functions. Hence many-a-time, biological keywords may not provide complete information from the databases.

To check keyword ambiguity in databases, we manually prepared presence/absence table considering protein functions associated with the sequences of 14 complete chloroplast proteomes freely available at NCBI. These proteomes belong to various chloroplast containing organisms from algae to angiosperms. We successfully identified 51 keywords for different protein functions which were presented in various ways in these sequences. For example, the large subunit of Rubisco is represented by eight different ways in these proteomes, e.g.

Table 1. Few examples of ambiguous keywords identified in chloroplast genome

Organism name*	Cre	Cvu	Pye	Chara	Mpo	Afo	Pnu	Acave	Pko	Pth	Ath	Nsy	Osa	Tae
Functions/keywords														
photochlorophyllide reductase subunit B photochlorophyllide reductase subunit ChlB ChlB subunit of protochlorophyllide reductase protochlorophyllide reductase chlB chain light-independent protochlorophyllide reductase subunit B	+	+	+	+	+	+	–	+	+	+	–	–	–	–
photosystem I subunit VIII photosystem I reaction centre subunit VIII subunit VIII of photosystem I PSI I-protein PSI I subunit VIII PSI I protein	–	+	+	+	+	+	+	+	+	+	+	+	–	+
ribulose 1,5-bisphosphate carboxylase/oxygenase large chain ribulose 1,5-bisphosphate carboxylase/oxygenase large subunit ribulose 1,5-bisphosphate carboxylase/oxygenase subunit ribulose bisphosphate carboxylase large chain ribulose bisphosphate carboxylase large subunit ribulose-1,5-bisphosphate carboxylase/oxygenase large chain ribulose-1,5-bisphosphate carboxylase/oxygenase large subunit large subunit of Rubisco	+	+	+	+	+	+	+	+	+	+	+	+	+	+

*Cre, *Chlamydomonas reinhardtii*; Cvu, *Chlorella vulgaris*; Pye, *Porphyra yezoensis*; Chara, *Chara vulgaris*; Mpo, *Marchantia polymorpha*; Afo, *Anthoceros formosae*; Pnu, *Psilotum nudum*; Aca, *Adiantum capillus-veneris*; Pko, *Pinus koraiensis*; Pth, *Pinus thunbergii*; Ath, *Arabidopsis thaliana*; Nsy, *Nicotiana sylvestris*; Osa, *Oryza sativa* indica cultivar group and Tae, *Triticum aestivum*.

ribulose 1,5-bisphosphate carboxylase/oxygenase large chain, ribulose 1,5-bisphosphate carboxylase/oxygenase large subunit, ribulose 1,5-bisphosphate carboxylase/oxygenase subunit, ribulose bisphosphate carboxylase large chain, ribulose bisphosphate carboxylase large subunit, ribulose 1,5-bisphosphate carboxylase/oxygenase large chain, ribulose 1,5-bisphosphate carboxylase/oxygenase large subunit, and large subunit of Rubisco. A representative list of few ambiguous keywords is given in Table 1. Although all these keywords represent the same function of the protein sequence, it requires a meticulous effort by careful manual examination of the enormous amount of the data to confirm that this indeed is true. Merely by using computational approaches, it is not feasible to get all information accurately due to different biological semantics.

Previously, an effort had been made to unify biology by creating the Gene Ontology (GO) Consortium[2]. Along with its several useful features it also contains synonyms of a biological keyword used in biological literature. But it is again a tedious job for a surfer to look for all the synonyms and explore the databases. Even for a computer program it is difficult to fetch all information by providing the entire set of keywords related to any function because we do not know what type of keywords belonging to a function are present in a database. Further, we have noted several keywords in our dataset that were not present in the GO synonyms list. Hence, it is essential to carefully examine the annotation (keyword) problem and to formulate specific guidelines, which will provide unification of keywords for the nomenclature in such cases. Moreover, in the absence of any guidelines for assigning gene/protein functions, more and more of such synonyms will be encased in future by the scientist community using their self-defined guidelines. All this will lead to further propagation of ambiguous keywords in biological databases. Still a large number of genomes have to be sequenced; therefore it is not too late to correct this annotation jargon.

1. Graveley, B. R., *Nature*, 2008, **453**, 1197–1198.
2. Ashburner, M. *et al.*, *Nat. Genet.*, 2000, **25**, 25–29.

ASHEESH SHANKER*
VINAY SHARMA

Department of Bioscience and Biotechnology,
Banasthali University,
Banasthali 304 022, India
**e-mail: ashomics@gmail.com*

Placing the scientist ahead of the science

The recent growth in funding for science and its continuation have the potential to make scientific projects carried out in India globally competitive. To get there, we need more students pursuing a career in science – a challenge world over. Science is primarily a human endeavour, and increased investments in institutions and equipment are blunted without corresponding incentives for people.

This issue is particularly pressing in India where most scientists are compensated by standardized packages linked to corresponding government pay-scales. This compensation is broadly revised about once or twice every decade as a result of deliberations by a pay commission. So far, this approach has helped foster a small, well-established and internationally well-regarded scientific community. Scientific progress today occurs at a much faster pace. The relatively lethargic compensation process for scientists eventually reflects on the progress of science within the country by keeping academic fields alive longer than due, and not catching emerging areas early.

While government-financed institutions are constrained in their ability to drastically alter salaries, many are autonomous in building policies that make a scientific career attractive to those with the requisite skills and motivation. The decision by the Department of Scientific and Industrial Research in allowing faculty to own equity in companies they form is in this spirit. Academic institutions, however, typically have a policy of clawing back a portion of external compensation their faculty may obtain. The share claimed by the institution applies to initiatives faculty take in generating funds beyond what is required in routine academic pursuits, such as consultancy fees, or profits a faculty-run-enterprise might incur.

While having institutions sharing the profits generated by faculty is certainly justified, often such policies are created in the relative lack of substantial revenue in comparison to an institution's operating costs. Most start-up companies fail within the first few years of being established, and imposing administrative and financial burden on them from the onset opposes the very culture the institutions try to promote in their grab for alternate funding. New ventures and collaborations are to be nurtured at nascent stages when they are most vulnerable. A more generous package of holding back until a stable company or consulting practice is formed (say, exceeding a certain revenue per year or some other metric) and then incrementally clawing back may help bring in steady revenue over a longer term.

A lot more thought needs to be put into a suitable approach governing personal incentives for scientists that obtain external funding. These funds may come from royalties on books, patent licence fees, consulting fees, entrepreneurial enterprises or grants from various organizations including the government. If we aim to attract scientific talent from across the world, our current institutions cannot compete using only the salaries paid as an incentive. The goal must be to make the salary irrelevant to those with the skills and drive to flourish financially by choosing Indian institutions as their base. Grant giving agencies too need to rethink incentives for the investigators for carrying out projects they need executed, beyond rigid rates offered to students and research scholars.

Policies that shift doing science in India from being a destination for those with personal interests to those with profitable ones will help ameliorate many of the challenges we currently face in making scientific careers a mainstream rather than an exceptional choice.

KAPILANJAN KRISHAN

School of Physical Sciences,
Jawaharlal Nehru University,
New Delhi 110 067, India
e-mail: kkrishan@mail.jnu.ac.in