# Commentary: keeping biology in mind

doi:10.1038/ng1202

In working through the examples in the *User's Guide*, the reader is exposed to a number of databases, web sites and other resources of enormous value for performing *in silico* analysis of biological data. Familiarity with and use of this vast arsenal can help the researcher to plan and execute experiments more intelligently. In using these resources and, more importantly, in drawing biological conclusions based on the results gleaned from these sites, however, there are a number of caveats and potential pitfalls of which the user should be aware. Although some of the specific points we now discuss go beyond the sample questions included in this guide, the basic lessons to be learned apply to the full range of bioinformatic analyses.

The user must understand the capabilities—and limitations—of the programs being used. In the same way that molecular biologists need to understand the chemistry underlying a routine assay or the physics behind separation techniques, they must have a basic understanding of what search or analysis methods actually do once the 'Submit' button has been pressed. Understanding what the chemistry, physics or search methods can and cannot reveal is critical if the user is to extract the full meaning of the results but not overinterpret them. By understanding the methods, users can also optimize them and end up with a better set of results than if these sequence-based search methods were treated simply as a 'black box'.

A specific case in which the reader could have encountered difficulty deals with the detection of domains within a protein, as described in Question 10. Consider the part of the question that discussed the Conserved Domain Database (CDD) at the NCBI. The CDD is a 'secondary database', one in which the entries have been derived from other databases, in this case Pfam[23] and the Simple Modular Architecture Research Tool (SMART)[24]. Pfam provides collections of multiple sequence alignments that represent known, common protein domains. Pfam is subdivided into two parts: Pfam A, which is manually curated, and Pfam B, which is automatically generated. By virtue of being 'hand-crafted', the entries in Pfam A are of higher quality and are therefore more reliable than those in Pfam B. Nevertheless, both Pfam A and Pfam B provide broad coverage across the spectrum of known protein domains.

The second source database, SMART, provides information on 500 domain families, but with a specific emphasis on those domains that have been implicated in signaling or have been found in extracellular or chromatin-associated proteins. This was a deliberate choice by the developers, who wished to tackle what might be called 'tougher-to-detect' or 'tougher-to-define' domains. At the outset, simply knowing the scope of the target database tells the user whether or not it is an appropriate choice for a sequence of interest, especially when some biochemical data may already be available. If users were to search solely against SMART and find nothing, without understanding the limited scope of the data underlying the resource, they might erroneously conclude that the protein of interest had no known domains.

Continuing with this example, and assuming that the user now understands the scope of the underlying source databases, a second problem quickly surfaces. When searching Pfam and SMART through the CDD interface at the NCBI, the search is performed using a variation of the BLAST algorithm called RPS-BLAST[25]. If one were, however, to go directly to the Pfam or SMART web sites and issue the query there, the searches would be performed using a very different algorithm, a hidden Markov model[26]. Although a description of the two different methods is beyond the scope of this discussion, it is important to understand that they are fundamentally different and will therefore produce different results. An extended discussion on this point, using specific examples, is available[27]. The CDD front end will miss those SMART and Pfam entries that represent short domains, repeats and motifs[28]. To understand what the methods do does not mean having to comprehend advanced mathematical equations: basic explanations in layman's terms can be found in any one of a number of reviews or textbooks[7,8].

One can often carry out a search and become excited on the identification of a motif; frequently such a motif is rather small. The Lys-Asp-Glu-Leu motif is an example; it targets proteins to the endoplasmic reticulum. But one should beware the 'short-motif' pitfall. The level of sequence identity required for significant homology is much higher for smaller regions—they either match or they don't. For very short motifs, homology cannot be inferred by sequence identity, meaning that short motifs may not be at all helpful in describing what a protein does.

Longer motifs have greater power in identifying true positives and eliminating false positives. More importantly, the supporting information is made available by simply clicking past the first page of summary results provided by the search engine. Even, or especially, the newest of users is encouraged to click away and discover the information and assumptions underlying the results that the searches have produced. These are self-explanatory in many cases.

With respect to complete sequences, the reader is advised to recall that the preliminary analyses of the human genome sequence led to a large reduction in the estimated number of genes contained in the human genome. Earlier, numbers of the order of 80,000 to as high as 140,000 had been suggested[29]. With the draft sequence of the genome in hand, new estimates lie closer to 30,000–35,000 genes[11]. If this is correct, the human would have only twice as many genes as are observed in either the roundworm or the fruit fly[11]. At the same time, human genes appear (in general) to have a more complex structure.

This pronounced 'reduction' in the number of genes in the human genome obviously challenges the one-gene, one-protein hypothesis (or, more properly, the one-gene, one-enzyme hypothesis[30]), as the number of proteins in the human proteome is thought to be well in excess of 35,000 (ref. 11). One explanation of the large number of individual proteins that can be generated from this relatively small number of genes is alternative splicing, a process by which the transcripts from a single gene can be processed differently and thus give rise to several distinct proteins. Particularly germane to this discussion is that many proteins have more than one function, depending on where they are found in the cell or within the body as a whole.

An interesting example of this phenomenon is the multifunctional protein phosphoglucose isomerase[31]. This protein catalyzes the interconversion of D-glucose-6-phosphate and D-fructose-6-phosphate. It is identical to neuroleukin, a protein secreted by T cells that promotes the survival of some embryonic spinal neurons and sensory nerves. It is also identical to an autocrine motility factor that might be involved in metastasis, and to a differentiation and maturation mediator implicated in the *in vitro* differentiation of human myeloid leukemia HL-60 cells to terminal monocytes. This therefore appears to be a single soluble protein that can take on four distinct cellular roles.

A more extreme example of one protein being used in alternative contexts involves an outright phase shift: the proteins known as α-enolase and τ-crystallin are encoded by a single gene and have the same amino-acid sequence. In the liver, the protein functions as α-enolase, a soluble glycolytic enzyme, whereas within the lens of the eye, it functions as τ-crystallin, a structural protein[32]. Proteins for which alternative functions have been identified have been given the playful name 'moonlighting proteins' (see ref. 33 for a review).

Why is this biological finding important to anyone who uses comparative sequence information? In the early days of sequence comparison, it was assumed that if a sequence of unknown function matched a sequence of known function, one knew, by extension, the function of the unknown; the conclusions of many published papers were based on this assumption. In light of these and similar, more recent findings, does sequence similarity still imply common function? The answer is: maybe yes and maybe no. In any case, more evidence than just sequence similarity is needed to draw any conclusion about sequence function.

Moving up in conceptual complexity to the level of structure, an entire class of molecular modeling techniques is available to consider similarities between proteins whose relationship might not be obvious from looking strictly at the nucleotide or amino-acid sequence. The reason one would want to perform such analyses was stated early in a relatively short history of bioinformatics[34]: structure is conserved to a greater extent than sequence. This stands to reason, as there is evolutionary pressure to maintain the three-dimensional shape of proteins, particularly those critical to the basic functions of a cell.

Inferring common function from structural similarity, however, is more problematic. Consider the TIM barrel. It defines a structural superfamily whose members show a high degree of structural similarity over a substantial number of residues. The TIM-barrel fold is a good example of possible divergent evolution, because this same basic structure mediates a wide variety of chemical reactions critical to biological survival. The TIM barrel is associated with one non-enzymatic and fifteen enzymatic functions[35], and transcripts encoding TIM-barrel proteins account for over 8% of the yeast transcriptome[36]. The roles of TIM-barrel proteins are diverse, ranging from isomerases to oxidoreductases and hydrolases. This generic versatility is economical for the cell but can make the job of assigning function to structures or substructures difficult. In deciding whether structural similarity implies common function, one needs to consider the subcellular localization of the proteins, when they are expressed, and the presence or absence of cofactors that might significantly alter their structure.

A final point to be considered relates to annotations in the public databases. Although these are of great value, most are made in an automated fashion, without the benefit of human curation. This is a matter of practicality, as it would be difficult to verify every annotation in the human genome, let alone those of every sequenced organism. Although some sequence-based annotations, such as the positions of genome, are determined experimentally and are therefore quite reliable, others are no more than predictions. The most notable of these are the predictions of gene structure that can be found at the NCBI, Ensembl and UCSC. Question 7 in this guide provides an excellent example of inconsistencies in gene predictions obtained using methods; the user should use such information carefully, particularly when designing experiments.

The second type of annotation—functional annotation—can be even more problematic. Even when similarity can be reliably detected, the functional annotations currently found in the public databases are often incorrect. For example[37], the functional annotations of 340 *Mycoplasma* genes were assessed: 8% were found to be incorrect, and, in many cases, did not logically connect to the known biology and metabolism of *Mycoplasma*. So never use database annotation as evidence of function when there are few homologs or when the annotations are inconsistent between homologs. And remember that annotations are intransitive[38]: if protein A and protein B share a common functional annotation, and so do proteins B and C, proteins A and C do not necessarily have the same function. Use functional annotations as a first step, and confirm the annotations by going back into the primary literature.

Biology is complex, and we still do not understand it very well. Although performing searches and finding data are not difficult, the intelligent use of all of the accumulated facts from databases is. It is always necessary to take a step backwards and ask a very simple question: do the search results actually make biological sense? Even when one is able to make biological sense of a prediction of function, it may turn out to be incorrect. As science is increasingly undertaken in a 'sequence-based' fashion, using sequence data to underpin the experimental design and interpretation of experiments, it becomes increasingly important that computational results are cross-checked in the laboratory, against the literature and with more robust computational analysis, so that the conclusions not only make sense, but are also correct.

**Acknowledgments**

*David Haussler (University of California, Santa Cruz), Ewan Birney (The Wellcome Trust Sanger Institute) and David J. Lipman (National Center for Biotechnology Information) served as advisors during the development of this guide.*

*The authors would also like to thank the following people for their contributions: K.N. Lazarides, S.K. Loftus, E.H. Margulies, K.L. Mohlke, P.M. Pollock, R.B. Sood and J.W. Touchman (National Human Genome Research Institute); D. Karolchik and J. Kent (University of California, Santa Cruz); D. Church and K. Pruitt (National Center for Biotechnology Information); and M. Hammond and E. Schmidt (Ensembl).*