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).