Biosequence annotation

High throughput sequencing efforts, regardless whether the genome or the transcriptome is targeted, aim at identifying the sequence of genes or transcripts - and ultimately of the functional gene products. However, biological sequences are, per se, of very little help. The succession of the four nucleotides along DNA and RNA. or of the 20 amino acids in the case of proteins provide, without further information, little insights into what the gene product is doing, how it is regulated, and what phenotype is connected to it. Two ways lead out of this situation, where the aim of either way is to add functional annotation to the gene product.
Before we continue, we would like to make the following specifications.

- Although many people talk and write about *gene function*, we refer here - for the sake of precision - to the function of the gene product. Reason is that one gene can code for more than one gene product, and the different products can vary in their function.
- The term *function* itself is not clearly defined. It is quite common to refer to the ***(biological) function*** of a gene as its activity in the context of a certain metabolic pathway. Thus, the same gene product can have different biological functions, depending on the pathway it is embedded into. Such people tend to refer to a gene's ***(biochemical) activity*** if they refer to what the gene product is actually doing, e.g. catalyzing a certain reaction in the case of an enzyme.

We try to adhere to this distinction whenever possible, although it is, from the context, pretty clear when a bioinformatician has *biological function* in mind, and when a gene's *biochemical activity*.

Experimental characterization

The one way, probably the harder one, to infer both a gene's biochemical activity, and its biological function, tries to assign function to the gene product *de novo*. In other words, people go into the lab and start doing experiments. Mass screens based on random mutagenesis, paired with massively parallelized phenotyping, are now increasingly used to rapidly establish a link between a gene and a phenotype de novo. The latter approach is commonly used in crop improvement (e.g. ***Garcia et al. (2016) Nature Protocols 11:2401–2418***).

In silico characterization

The *in silico* way to learn about what a gene might be doing is typically considered the easier one. We use the computer to tentatively assign a certain activity to the gene product, based on the results of various algorithms for biological sequence analysis. One of the most widely used evidence is the sharing of a significant sequence similarity to either an already functionally characterized gene product, or to statistical models that represent conserved regions in aligned sequences. In both cases, this similarity is interpreted as a signal that the sequences share a common ancestry, i.e. are homologous.
Basically, in silico analyses of gene function boil down to three general approaches

- The identification of functionally annotated sequences displaying a significant sequence similarity. The most common tool to do this is BLAST.
- The identification of evolutionary conserved subsequences -sometimes referred to as *domains* -

in the sequence of interest. One of the most widely used tools/databases is Pfam
- The identification of short motifs in a sequence. Note, contrary to the other two approaches, a motif search typically does not rely on the inference of an evolutionary relationship[1)]

Integrative approaches

Nowadays, experimental approaches, and those that are computer-based are far from being mutually exclusive. People rarely enter the lab without a prior idea about what a certain gene product could do. And likewise, *in silico* analyses on the computer heavily depend on the amount of existing information that can be exploited in the process annotating the function of an unknown gene product. In essence, annotating a novel gene product is equivalent to connecting it to the network of existing information about gene product function. Information is transferred, rather than generated *de novo*. If there is no network, then no connection can be established, no annotation transfer is possible, and the function of the gene product remains unknown.

1)

Motifs are short sub-sequences of DNA or a protein of defined length, e.g. the start codon ***ATG*** or the canonical splice donor and splice acceptor sites GT-AG. Due to their short length, it is feasible to assume that they can emerge independently more than once in the course of evolution.