# Consensus sequence

In molecular biology and bioinformatics, the **consensus sequence** (or **canonical sequence**) is the calculated order of most frequent residues, either nucleotide or amino acid, found at each position in a sequence alignment. It represents the results of multiple sequence alignments in which related sequences are compared to each other and similar sequence motifs are calculated. Such information is important when considering sequence-dependent enzymes such as RNA polymerase.[1]

## Biological significance

A protein binding site, represented by a consensus sequence, may be a short sequence of nucleotides which is found several times in the genome and is thought to play the same role in its different locations. For example, many transcription factors recognize particular patterns in the promoters of the genes they regulate. In the same way restriction enzymes usually have palindromic consensus sequences, usually corresponding to the site where they cut the DNA. Transposons act in much the same manner in their identification of target sequences for transposition. Finally splice sites (sequences immediately surrounding the exon-intron boundaries) can also be considered as consensus sequences.

Thus a consensus sequence is a model for a putative DNA binding site: it is obtained by aligning all known examples of a certain recognition site and defined as the idealized sequence that represents the predominant base at each position. All the actual examples shouldn't differ from the consensus by more than a few substitutions, but counting mismatches in this way can lead to inconsistencies.[2]

Any mutation allowing a mutated nucleotide in the core promoter sequence to look more like the consensus sequence is known as an **up mutation**. This kind of mutation will generally make the promoter stronger, and thus the RNA polymerase forms a tighter bind to the DNA it wishes to transcribe and transcription is up-regulated. On the contrary, mutations that destroy conserved nucleotides in the consensus sequence are known as **down mutations**. These types of mutations down-regulate transcription since RNA polymerase can no longer bind as tightly to the core promoter sequence.

## Sequence analysis

Developing software for pattern recognition is a major topic in genetics, molecular biology, and bioinformatics. Specific sequence motifs can function as regulatory sequences controlling biosynthesis, or as signal sequences that direct a molecule to a specific site within the cell or

regulate its maturation. Since the regulatory function of these sequences is important, they are thought to be conserved across long periods of evolution. In some cases, evolutionary relatedness can be estimated by the amount of conservation of these sites.

## Notation

The conserved sequence motifs are called **consensus sequences** and they show which residues are conserved and which residues are variable. Consider the following example DNA sequence:

A[CT]N{A}YR

In this notation, A means that an A is always found in that position; [CT] stands for either C or T; N stands for any base; and {A} means any base except A. Y represents any pyrimidine, and R indicates any purine.

In this example, the notation [CT] does not give any indication of the relative frequency of C or T occurring at that position. An alternative method of representing a consensus sequence uses a sequence logo. This is a graphical representation of the consensus sequence, in which the size of a symbol is related to the frequency that a given nucleotide (or amino acid) occurs at a certain position. In sequence logos the more conserved the residue, the larger the symbol for that residue is drawn; the less frequent, the smaller the symbol. Sequence logos can be generated using WebLogo, or using the Gestalt Workbench, a publicly available visualization tool written by Gustavo Glusman at the Institute for Systems Biology.[2]

## Software

Bioinformatics tools are able to calculate and visualize consensus sequences. Examples of the tools are JalView and UGENE.

## See also

## References