Generalization is a term used is computer science to refer to the process of reducing complexity of a problem while still being able to produce a solution. It does this by replacing many similar functions with one broadly defined function that includes all of the similar functions (in python this can be represented by a lambda function). This idea can be used many times in biology related fields such as proteomics, bioinformatics, pharmacology and more, and is therefore a necessary concept to comprehend. In fact generalization has been used extensively in biology for dividing species into classes and subclasses based on underlying similarities, for example that all mammals fit into the same class because of the possession of hair, three middle ear bones, mammary glands in females, and a neocortex even though the similarities between a bat and a whale originally seem scant.

The applications for generalization are many. A more recent application for generalization is in the field of bioinformatics. As the field develops, a growing number of individuals are having their genomes sequenced and studied to find genetic solutions to diseases. A problem that arises is how to access this information without compromising the identification of the individual and invasion of privacy, which has already been documented in an anonymous DNA study.([2](#_ENREF_2)) Therefore a generalization technique is usually used so that the individual will be indistinguishable from the other patients in the study while still being different enough to have valuable data (k-anonymity). For example when looking at a patients DNA one technique replaces similar nucleotides, pyrimidine or purines, with a common letter([3](#_ENREF_3)), or substituting an age range instead of an actual age when reporting age (for example age 30 for all patients ages 30-39):



Another application is in the field of proteomics and pharmacology. Determining the secondary structure of a protein is a huge unsolved problem in proteomics. Techniques such as Xray crystallography and NMR are hard to do and take time, and offer no help in predicting a novel protein. A method based on over-generalization is being developed, based on known proteins to try to predict the structure of an undetermined sequence. It starts with finding similar known proteins based on sequence, and then generalizing common structures of those proteins based on common sequences. These predicted structures are then applied to the novel sequence to give a broad definition of the structure.([4](#_ENREF_4)) The same researchers are utilizing the same technique to predict possible drug candidates without the need to synthesize them. This technique can save time and money, and allow multiple simulated trials to start the actual synthesis with a compound that has a higher chance of success from the get go.

In conclusion the generalization is a term in computer science that can be utilized extensively in the field of biology, bioinformatics proteomics and more. It is a concept that can turn complex problems into simple ones, while getting rid of unnecessary detail. Anyone interested in the fields, even if not intending to write actually code, should be familiar with the term to communicate with someone is the computer science field.

1. **Gusfield D.** 1997. Algorithms on strings, trees, and sequences : computer science and computational biology. Cambridge University Press, Cambridge England ; New York.

2. **Sweeney L, Abu A, Winn J.** 2013. Identifying Participants in the Personal Genome Project by Name, Data Privacy Lab, IQSS, Harvard University.

3. **Malin BA.** 2005. Protecting genomic sequence anonymity with generalization lattices. Methods of information in medicine **44:**687-692.

4. **Eschrich S, Chawla NV, Hall LO.** 2002, p 25-32. BIOKDD.