Evolution of structure and function of biological macromolecules

Wojciech Rypniewski

Center for Biocrystallographic Research, Institute of Bioorganic Chemistry, Polish Academy of Sciences, Poznań, Poland (wojtekr@ibch.poznan.pl)

Evolution is driven by changes of the genetic material. DNA undergoes mutations as the result of random or environmental factors. Even if we discount harmful environmental influences, some changes are inevitable due to “imperfect” copying of the genetic material. Mutations come in very different sizes. A single “point” mutation affects a very small area of the genome – perhaps a single nucleotide residue or a pair of residues, or it can involve a length of DNA sequence as large as a gene or larger, or it may even change the count and arrangement of the chromosomes.

A mutation can have no effect on the organism or it can be lethal, or it can change the properties of the organism, the way it functions, depending on how the mutated gene(s) are “expressed”. The products of gene expression are protein and RNA molecules and there is a direct correspondence between the DNA genetic sequence and the protein's sequence of amino acids, or the sequence of RNA's nucleotide residues.

A point mutation will generally result in a single amino acid substitution in the protein, unless it's a so-called “silent substitution” in which case the gene product (protein) is unchanged. A point mutation may also result in a premature termination of the protein chain. The rules are clearly defined and can be inferred easily from the genetic code.

The consequences of single amino acid substitution in the protein (or a nucleotide in RNA) depend on the location of the substitution in the structure of the molecule and on the nature of the substitution, i.e. what residue is replaced by what residue. The consequences of many such mutations can be readily studied and explained by crystallography, provided that they do not perturb the structure so as to make it unstable and thus unsuitable for crystallographic investigations. For example, an amino acid substitution within a protein's ligand-binding site can easily change the ligand specificity. We can see this in the so-called S1 pocket in serine proteases. It is a cavity near the catalytic site which binds an amino acid side chain of the protein (substrate) to be hydrolysed. The polypeptide chain of the substrate is then lysed next to this bound residue. An acidic residue (aspartate) at the bottom of the S1 pocket makes the pocket specific for binding basic side chains (arginine or lysine) with a perfect salt bridge formed between the side chains. This is why trypsin cuts the peptide bonds next to arginine or lysine residues. Related serine proteases have a common reaction mechanism but different substrate specificities because of differences in the shape and character of the S1 “specificity” pocket.

Trypsin-like serine proteases are an example of divergent evolution. Many mutations have accumulated over a long time – more than a billion years (a million million) – and today the amino acid sequences in the trypsin family are different in different species of organisms, the proteins have acquired different substrate specificities, they have different functions, may have become parts of larger assemblies – but they have retained the basic reaction mechanism and are still recognisably similar when we compare their amino acid sequences and, especially, when we compare their three-dimensional structures. Indeed, it seems that the 3-D structure, the “fold”, is more conserved than the amino acid sequence and the evolutionary relationship of distantly related proteins can only be discovered, or confirmed, after we've solved their structures.

The same useful device can be “invented” by nature many times independently. An example of convergent evolution is provided by subtilisins and trypsins. The two families of serine proteases are unrelated – the have no sequence similarities and have completely different folds – but they have the same enzymatic reaction mechanism based on the classical catalytic triad of serine, histidine and aspartate residues in the same relative orientation.

Comparing proteins within a related family gives a measure of the relationship between their respective organisms. A plausible phylogenetic tree can be constructed based on such a comparison. It shows that mutations for related proteins accumulate at an approximately constant rate between different species. We should take care, however, that in such cases we compare proteins that are not only related but also have retained similar functions. Otherwise we may find that different rates of mutations are tolerated for proteins that have different functions and so we are comparing “apples and oranges”.

When we compare related amino acid sequences and the corresponding three-dimensional protein structures, we realise that in some regions of the molecule changes are permitted easily, while other regions are more conserved. A few areas may be identified where mutations are never permitted. One can deduce from this which parts of the protein are crucial to its function or its stability. The surface of the molecule is usually the least conserved, excect for the sites of activity and binding of ligands or other partner biomolecules. The interior of the molecule is usually much better conserved. Substitutions are generally permitted there but they are likely to have a destabilising effect on the structure.

One of the most conserved structures is not a protein but the ribosome – the protein-making factory – which consists primarily of RNA. Every living organism has ribosomes and they are all related. Even eukaryotic organelles – mitochondria and chloroplasts – have their own ribosomes, together with the rest of their independent protein-making machinery. It's a remnant of their distant evolutionary past, when they were free-living organisms. The solving of the ribosome crystal structure was a major milestone of structural research and a proof of the catalytic role of RNA in the protein biosynthesis. Several ribosomal structures have been solved by now, as well as some other catalytic RNAs – remnants of the “RNA world” which is thought to pre-date the time when proteins emerged in early organisms as the main biocatalysts.

Apart from random point mutations, the major engine of evolution is gene duplication. Duplications arise during meiosis between imperfectly aligned homologous chromosomes. Once an additional copy of a gene arises it mutates quickly because it is not under selection pressure. In other words, the copy can mutate freely because it is not necessary for the organism, while the parent gene still functions. Multiple gene copies often function in parallel, or become tissue specific, and their products are known as isozymes. Sometimes they lose their functionality and become “silent genes”. Sometimes they acquire by chance some useful property and re-emerge in a new role. This is a common mechanism by which organisms become more complex.

It happens also that gene duplication results not in two copies of a protein but in a protein that is twice as large, in which the two repeats remain joined together within a single polypeptide chain. If such a “double protein” turns out to be viable, one half retains its basic function while the other half is free to mutate. This is analogous to the above case of releasing selection pressure from the “spare” isozyme, except that this is now happening within a protein molecule. In time the spare half can acquire some useful property. For example it can provide an additional means of regulating an enzyme's catalytic activity, like in the case of eukaryotic phosphofructokinase.

Evidence of gene duplications can be found within many proteins. Even when there is no longer any detectable sequence similarity, one may observe, knowing the three-dimensional structure, that the proteins consists of two or more domains with the same fold. Trypsin is a known example. It consists of two such domains.

Trypsin was also used to peer back in time to the beginning of the “protein world”. Baptista *et al.* (1) collected as many trypsin-like sequences as they could find and subjected them to Fourier analysis. The spectrum showed a signature of 15-to-18-nucleotide fragment, corresponding to a peptide five or six residues long, whose subsequent duplications and fusion eventually resulted in the trypsin molecule that we know.

Reference

1. Baptista, A.M., Jonson, P.H., Hough, E. &Petersen, S.B. The Origin of Trypsin: Evidence for Multiple Gene Duplications in Trypsins, *J. Mol. Evol.* (1998) 47:353–362