Xjenza Online - Journal of The Malta Chamber of Scientists
www.xjenza.org
Doi: http://dx.medra.org/10.7423/XJENZA.2014.1.06

Review Article

The Structure of Protein Molecules: In Celebration of the International Year of Crystallography, 2014

Gary Hunter[1], **Marita Vella**[1], **Rosalin Bonetta**[1], **Diane Farrugia**[1], **Therese Hunter**[1]
[1]Department of Physiology and Biochemistry, University of Malta

Abstract. Many people, including laymen, are aware of the double helical nature of the DNA molecule. A few may actually realise that it was the technique of X-ray crystallography that was the key to solving this structure. Even fewer will understand the uses and applications of crystallography to the most diverse of biological materials; proteins. In this review we discuss the application of a number of methodologies required to progress from a cloned gene to protein expression and purification, crystallisation conditions and eventually to X-ray structure determination. We provide our own experience in the field as examples of the procedures required. Protein crystallographers worldwide are contributing to our understanding of how enzymes work, how our immune system defends us against viruses and are using structural information to design novel pharmaceutical reagents.

Keywords Protein Structure - Expression - Purification - X-ray crystallography.

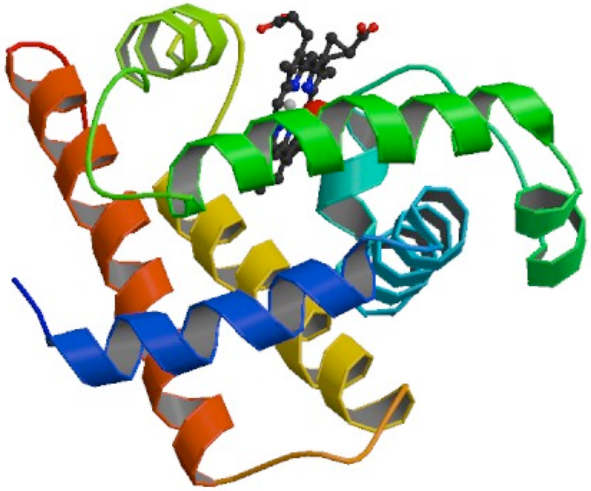

Figure 1: The first of many: the structure of the oxygen carrier protein Myoglobin from Sperm whale. Originally solved by Perutz, this is the structure from RCSB entry 1MBN, Watson, 1969. It clearly shows the position of the haem cofactor (ball and stick figure) and therefore where an oxygen molecule will bind.

1 Introduction

It is time for crystallography to step out of the shadows. Both to pay tribute to the contribution crystallographers have made to science and medicine, and to encourage young scientists in the field, UNESCO has declared 2014 as the International Year of Crystallography. Appropriately this celebration coincides with the centenary of the discovery of X-ray crystallography. The importance of X-ray crystallography to the advancement of science is apparent by the fact that since the first crystal was analysed by the Bragg father and son duo over 100 years ago at the University of Leeds, UK, 28 Nobel prizes, including the 2013 prize in Chemistry, have been awarded to X-ray crystallographers. Even in the early days of X-ray crystallography, interest was focused on biological molecules. The way in which structure helps to explain function was famously exemplified with the molecular model of DNA by Watson and Crick (Watson and Crick, 1953). Kendrew and Perutz were the first to apply the technology to proteins, namely myoglobin (Kendrew, 1962)(Figure 1; 1MBN, (Watson, 1969)) and haemoglobin, overcoming daunting obstacles to arrive at a three dimensional representation of these large molecules. Since then the field has exploded exponentially and now the repository of biological molecular structures, the RCSB (Research Collaboratory for Structural Biology, www.RCSB.org) which includes the protein data bank (PDB)(Berman et al., 2000) is fast

Correspondence to: T. Hunter (therese.hunter@um.edu.mt)

approaching one hundred thousand entries. The importance of proteins in biological systems underlies the need to understand more about them, and protein structure determination should not be underestimated.

2 Protein Expression

Protein purification is an obvious primary requirement before crystallisation trials begin. Taken together, these are the two main bottlenecks frequently encountered in the process of determining the structure of a protein.

Isolation of protein from tissues may be cumbersome and laborious, requiring large quantities of source material and prone to high losses of protein. Although there may be good reasons for isolating proteins from their source organism (discovery of natural post-translational modifications, for example), the methodology of choice today is molecular cloning of the gene.

When information regarding the nucleotide sequence of the gene is available, proteins may be produced by recombinant DNA technology or by total gene synthesis. Expression of eukaryotic proteins starting with total messenger RNA entails the cloning and expression of the appropriate cDNA. This involves reverse transcription of messenger RNA, PCR amplification of the cDNA fragment of interest and cloning into an appropriate molecular vector. Even when genetic information of the gene of interest is absent, all is not lost, however. Short stretches of amino acid homology at the ends of a protein may be utilized to design degenerate primers for PCR experiments. Indeed this is how we cloned the gene for SOD-3 from *C. elegans*, and subsequently the corresponding cDNA for expression studies (Hunter et al., 1997a).

There are various expression platforms suitable for different proteins. With a large selection of expression vectors designed for specific expression protocols and purification systems now available, it should be possible to find a system that works for almost any protein. It should be stressed however that the best system is often found by trial and testing. Certain eukaryotic proteins require post-translational modifications for their biological activity and may have to be expressed in eukaryotic hosts capable of performing them. The yeast *Pichia pastoris* is such an expression platform that allows human-like glycosylation.

Bacterial expression systems produce large quantities of proteins and are easy to culture and remain the system of choice for many protein scientists. Over recent years the technology of expressing mammalian proteins in bacteria has advanced greatly permitting the production of a variety of proteins. Toxic and membrane proteins, inclusion body formation and limitations of codon usage are all problems which are now largely solved (Graslund et al., 2008).

3 Protein Purification

As each protein is unique, different purification procedures have to be designed and tested. Even similar proteins or mutant proteins differing by one amino acid residue may require different strategies in order to obtain protein of sufficient purity for later processing. The greatest challenge in purification is the elimination of background proteins whilst obtaining a high yield of soluble, pure target protein. Traditional protein purification procedures are often successful, and include salting out (ammonium sulfate precipitation), ion exchange chromatography and gel filtration. A combination of these procedures is usually required (Figure 2, (Vella et al., 2014)). When the protein is naturally abundant purification can be relatively easy, hence the use of Sperm whale as the source of myoglobin by Perutz.

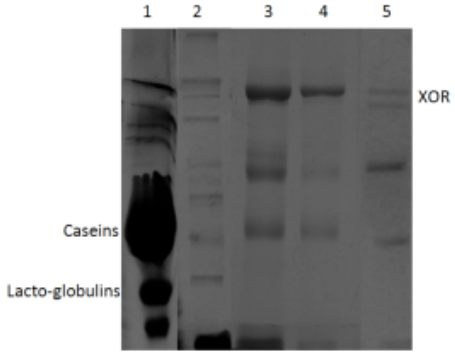

Figure 2: 15% SDS-PAGE illustrating the purification of xanthine oxidoreductase (XOR) from bovine milk. Lane 1 are proteins in fresh unpasteurized bovine milk, crude sample. Lane 2 is a protein sample following addition of 20% ammonium sulfate. Lanes 3 and 4 is a protein sample after chromatography on a heparin column. Lane 5 is a Protein sample after gel filtration chromatography. Purified XOR is shown as four bands on SDS-PAGE.

To expedite purification, affinity chromatography can be extremely effective, replacing a number of techniques with a single step and a plethora of affinity tags have been produced to utilize this powerful technique for almost any protein of choice. These tags are invariably incorporated into the target protein by genetic engineering, resulting in a chimeric fusion product. Removal of the tag after purification is also often an available option and again a number of protease restriction sites have been added to expression vector systems.

Commonly used tags include glutathione-S-transferase (GST) (Figure 3) (Hunter et al., 1997b), maltose-binding protein (MBP) or hexahistidine tags (Figure 4) (Hunter and Hunter, 2013) incorporated onto the N or C terminus of the protein being purified. Columns with immobilised glutathione, maltose or metals such as nickel are used respectively for pu-

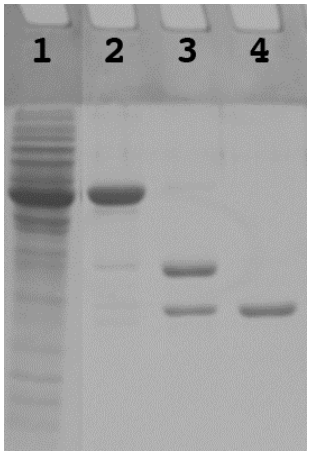

Figure 3: Expression and purification of a GST-SOD fusion protein. Lane 1 is the cell lysate, lane 2 is GST-SOD purified by GSH–sepharose chromatography, lane 3 is thrombin-cleaved GST-SOD showing both GST (upper band) and released SOD, lane 4 rechromatographed sample to remove the GST, leaving pure SOD as a single band.

rification of these fusions. Specific protease cleavage sites, including thrombin (Figure 3) and Factor Xa (Figure 4), may be engineered between the tag and the recombinant protein. This enables the cleavage of the tag away from the protein under study. A second affinity column removes the released tag (Figure 4) and many restriction proteases can be similarly removed to leave highly pure native proteins (Hunter and Hunter, 1998; Hunter et al., 2002). One disadvantage is that many vectors with engineered fusion tags will leave extra amino acids at the end of the protein, and to this end we developed a hexahistidine tag vector which leaves only authentic protein sequence after cleavage of the product (Hunter and Hunter, 2013).

4 Protein Characterisation

Once sufficient amounts of protein have been purified, which may be in the range of 20 to 50 mg, characterisation is carried out to confirm the identity of the isolated protein and to determine its biochemical properties to decipher the mechanism of function or catalysis. Absolute identity of the protein is confirmed by immunoblot with a specific antibody for the protein. Mass Spectrometry such as MALDI-TOF-TOF gives vital molecular weight information. This may confirm the occurrence of proteolysis or post-translational modifications. Purity is often assessed by SDS-PAGE and protein concentration is recorded using the BCA (Smith et al., 1985) or the Bradford method (Bradford, 1976), standard curves being made using BSA. A more accurate method of determining the concentration of pure protein is measuring the absorbance at 280nm and then calculating the concentration using the extinction coefficient by the Beer-Lambert law. Biological activity of the protein after purification is of paramount importance and often used to monitor the progress of purification procedures. Precisely what is measured will depend on the protein. With respect to metalloenzymes such as SOD it is necessary to measure cofactor metal content in order to calculate the specific enzyme activity. ICP-MS, MP-AES and AAS-GF are methods sensitive enough to detect such low levels of metals. Various spectrophotometric assays exist for the measurement of enzyme activity or native PAGE followed by zymography may be employed (Figure 5). Circular dichroism spectroscopy can provide information of any structural changes that may have occurred during purification, which could be detrimental to the activity of the protein. Gel filtration may be used to determine the molecular weight of the native protein which when combined with mass spectrometry or SDS-PAGE data can be used to calculate quaternary structure.

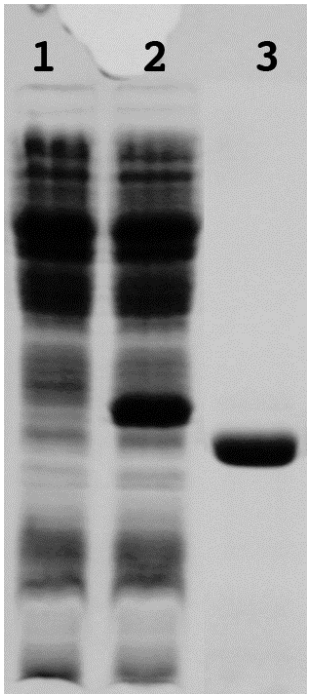

Figure 4: Expression and purification of a Hexahistidine-tagged SOD protein. Lane 1 is an E.coli cell lysate without the expression plasmid, Lane 2 is a cell lysate showing overexpression of the H6-SOD protein, Lane 3 shows the pure SOD band after nickel chelation affinity chromatography and cleavage by Factor Xa to remove the tag.

5 Crystallisation

The ultimate analysis for any biological molecule including proteins is the determination of its three-dimensional structure. X-ray Crystallography has been described as the technology that marries art with science. The objective is to produce crystals that are composed of regular, repeated arrangements of, in this case, a protein molecule. It should not

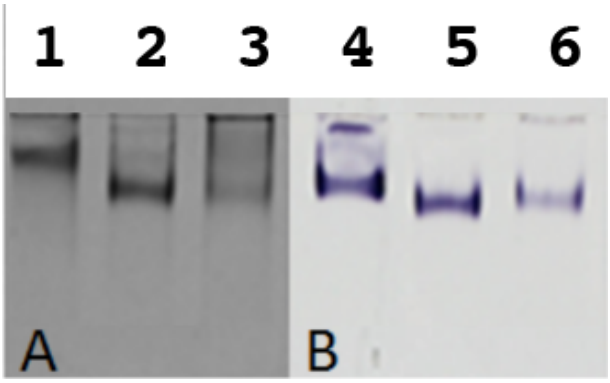

Figure 5: 8% Native PAGE followed by zymography. **A**: Native PAGE showing bovine, caprine and ovine XOR band (lanes 1 to 3 respectively) stained with Coomassie Brilliant Blue. **B**: Native PAGE after 2 hours of incubation in active solution showing bovine, caprine and ovine XOR band (lanes 4 to 6 respectively). Active XOR present on the gel catalyzes the conversion of xanthine to uric acid, producing hydrogen peroxide and superoxide anions. The latter reduce NBT to form a purple coloured insoluble formazan product

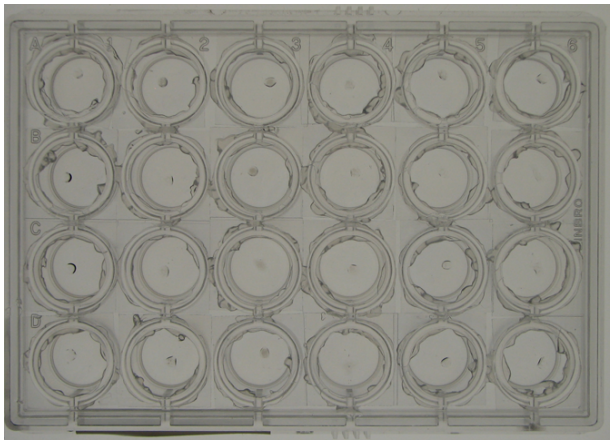

Figure 7: The hanging drop method of crystallisation. A drop containing reservoir solution and protein (1:1) is placed on an upside down coverslip over a well containing reservoir solution (precipitant, buffer and additives) and sealed. Vapour diffusion helps to produce crystals. Here a 24 well plate is used to test 24 different conditions for the same protein.

be forgotten that protein crystals are unnatural states for any protein, which helps to explain the difficulty in predicting the requirements to produce them. A 0.5mm cuboid crystal may contain as many as 10^{15} molecules of protein (Figure 6). The major ingredients that encourage protein crystal formation are a buffer and a precipitant. Initially the protein is diluted with this mixture and allowed to slowly equilibrate by vapour diffusion in a sealed chamber. This is most commonly achieved by the hanging drop method (Figure 7). Proteins may be co-crystallised with other molecules such as inhibitors, agonists, cofactors, nucleic acids and even other proteins. It may take months if not years to grow the perfect crystal.

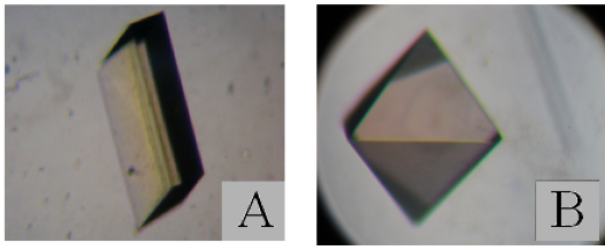

Figure 6: Crystals of superoxide dismutase. Crystal structures of similar proteins can produce surprisingly different crystals. *E.coli* FeSOD (A) and *C.elegans* MnSOD (B).

There are many variables involved in the crystallisation process and parameters that must be tested include protein concentration, temperature of the environment, humidity, pH, type and concentration of both precipitant and buffer and the inclusion of additives (almost any chemical compounds in existence). Consequently endless permutations may be possible and must be tested. For this reason hundreds of screening trials are carried out to identify the stringent optimal conditions that will reproducibly form stable and diffraction-quality crystals. Maybe out of sheer frustration, there are those who claim that other factors such as music and even supernatural phenomena affect the growth of that ever-elusive perfect crystal. The goal however is for a single crystal to form under reproducible conditions, which is large enough to diffract an X-ray beam effectively as the greater the diffraction angle, the higher the resolution of the final structure.

6 X-ray Diffraction

The X-ray diffraction pattern is essentially a series of spots of different intensities in different positions on a two dimensional detector (Figure 8A). A series of diffraction patterns has to be produced with the crystal mounted in the X-ray beam rotated by some small amount between data collection. Powerful computer programs utilizing Fourier transformation algorithms are used to convert this data into an electron density map (Figure 8B). The latter is effectively a three dimensional map of the position of all the electrons in the molecules.

More computing is required using the known sequence of amino acids in the protein to effectively fit the heavy atoms of the string of amino acids into the determined electron density (Figure 8C). This process, known as refinement, includes a reiterative technique that reduces errors to a minimum and produces what is accessible to all from the databank, the coordinates for the heavy atoms that make up the protein. A typical small protein like superoxide dismutase (MnSOD3) which has 194 amino acids produced 49,346 reflections and yielded the coordinates for 3569 heavy atoms including waters (pdb 3DC5) (Trinh et al., 2008). The later molecules are an intrinsic part of any protein, often determining how in-

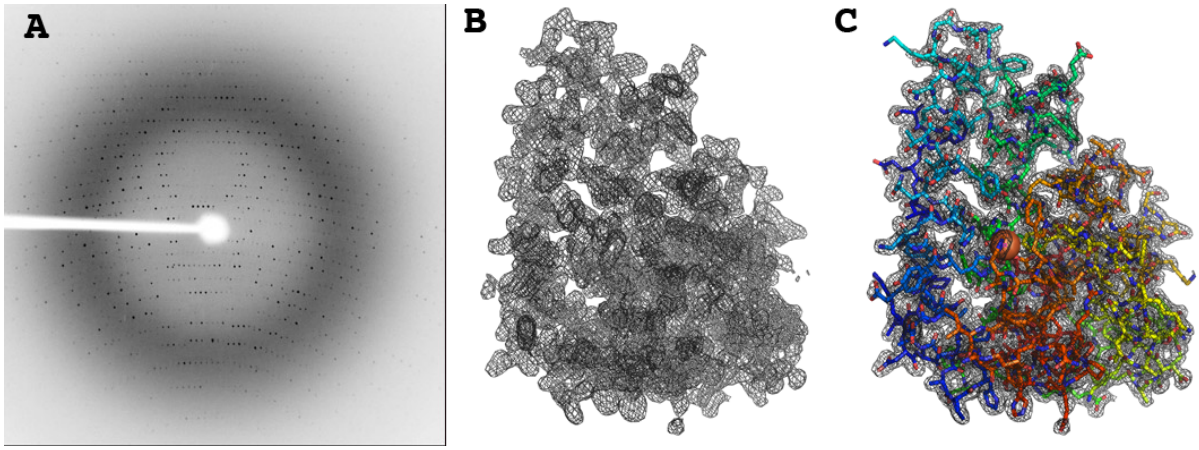

Figure 8: Generating a protein structure. The process begins with (A) the diffraction pattern produced by a protein crystal of FeSOD. (B) the electron density map (EDM) generated by Fourier transform of (A), and (C) the fitted amino acid sequence with the EDM.

teractions between the protein and substrate occur as well as protein:protein interactions in the quaternary structure. This is not quite the end of the process however as the structure returned by the above processes is the asymetric unit. In other words it is the most minimal structure to be found within the crystal that is duplicated many times in order to compose the crystal itself. Sometimes this is more than the biological unit. For example the structural information deposited for the MnSOD from *E.coli* contains seven protein subunits even though the biological unit is a dimer. In the case of MnSOD2, the asymetric unit is two subunits while the biologically active protein is tetrameric. Further manipulations are therefore required to arrive at the all-important biologically active form of the protein (Figure 9).

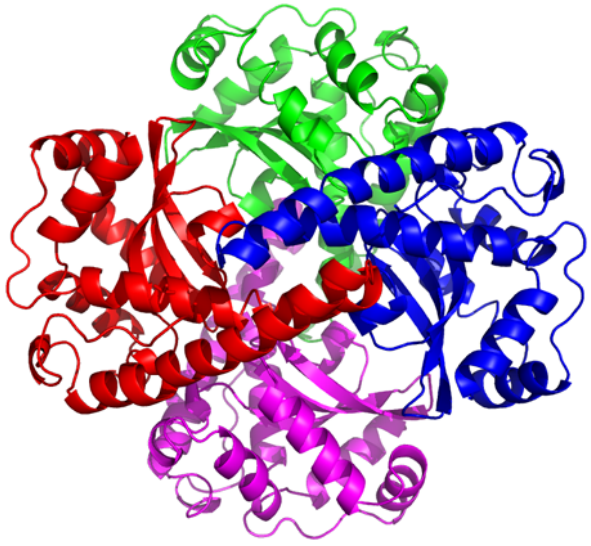

Figure 9: The biologically active three-dimensional structure of SOD-2, one of two MnSODs from *C.elegans* is homotetrameric. Each of the four chains is shown in a different colour. Only the protein backbone is shown in ribbon representation.

6.1 Conclusion

The determination of the structure of proteins is now considered to be the last hurdle to unlocking the molecular secrets of their mode of action. Over recent years, for example, the Center for Structural Genomics of Infectious Diseases (CSGID)(http://www.csgid.org/2014) and the Seattle Structural Genomics Center for Infectious Disease (SSGCID)(http://www.ssgcid.org/2014) have worked together to determine the structure of over one thousand proteins from forty pathogenic organisms responsible for diseases such as leprosy, cholera, TB and influenza. Other groups have obtained the structure of the HIV protein responsible of hijacking human cells (Tahirov et al., 2010). And recently scientists from Scripps Institute and Cornell School of Medicine have determined the structure of a number of G-protein-coupled receptors that are a vitally important component of many signaling pathways in many different types cells (gpcr.scripps.edu/2014) (Huang et al., 2013). Collectively these and similar breakthroughs are considered as important as the completion of the human genome sequence. The availability of this type of information for proteins marks the start of an new era in the advancement of science and medicine, enabling not only the understanding of protein function in health and disease but also providing opportunities for better diagnostic tools and development of novel drugs designed for specific targets.

References

Berman H., Westbrook J., Feng Z., Gilliland G., Bhat T., Weissig H., Shindyalov I., Bourne P. (2000). The Protein Data Bank. *Nucelic Acids Res.*, 28, 235–242.

Bradford M. (1976). Rapid and sensitive method for quantification of microgram quantities of protein utilizing the principle of protein-dye binding. *Anal. Biochem.*, 72(248-254).

Graslund S., Nordlund P., Weigelt J., Hallberg B., Bray J., Gileadi O., Knapp S., Oppermann U., Arrowsmith C., Hui R., Ming J., Dhe-Paganon S., Park H., Savchenko A., Yee A., Edwards A., Vincentelli R., Cambillau C., Kim R., Kim S., Rao Z., Shi Y., Terwilliger T., Kim C., Hung L., Waldo G., Peleg Y., Albeck S., Unger T., Dym O., Prilusky J., Sussman J., Stevens R., Lesley S., Wilson L., Joachimiak A., Collart F., Dementieva I., Donnelly M., Eschenfeldt W., Kim Y., Stols L., Wu R., Zhou M., Burley S., Emtage J., Saunder J., Thompson D., Bain K., Luz J., Gheyi T., Zhang F., Atwell S., Almo S., Bonanno J., Fiser A., Swaminathan S., Studier F., Chance M., Sali A., Acton T., Xiao R., Zhao L., Ma L., Hunt J., Tong L., Cunningham K., Inouye M., Anderson S., Janjua H., Shastry R., Ho C., Wang D., Wang H., Jiang M., Montelione G., Stuart D., Owens R., Daenke S., Schutz A., Heinemann U., Yokoyama S., Bussow K., Gunsalus K. (2008). Protein production and purification. *Nat. Methods*, 5(2), 135–146.

Huang J., Chen S., Zhang J., Huang X. (2013). Crystal structure of oligomeric β1-adrenergic G protein-coupled receptors in ligand-free basal state. *Nat. Struc. Mol. Biol.*, 20(4), 419–25.

Hunter G., Hunter T. (2013). GroESL protects superoxide dismutase (SOD)-deficient cells against oxidative stress and is a chaperone for SOD. *Health*, 5, 1719–1729.

Hunter T., Bannister J., Hunter G. (2002). Thermostability of Manganese-and Iron-Superoxide Dismutases from *Escherichia Coli* is Determined by the Characteristic Position of a Glutamine Residue. *Eur. J. Biochem.*, 269, 5137–5148.

Hunter T., Bannister W., Hunter G. (1997a). Cloning, expression and charaterisation of two manganese superoxide dismutases from *Caenorhabditis elegans. J. Biol. Chem.*, 272, 28652–28659.

Hunter T., Hunter G. (1998). GST fusion protein expression vector for in-frame cloning and site-directed mutagenesis. *BioTechniques*, 24, 194–196.

Hunter T., Ikebukuro K., Bannister W., Bannister J., Hunter G. (1997b). The Conserved Residue Tyrosine 34 is Essential for Maximal Activity of Iron-Superoxide Dismutase from *Escherichia coli*. *Biochemistry*, 36(16), 4925–4933.

Kendrew J. (1962). Myoglobin and the structure of proteins. Nobel Lecture.

Smith P., Krohn R., Hermanson G., Mallia A., Gartner F., Provenzano M., Fujimoto E., Goeke N., Olsen B., Klenk D. (1985). Measurement of protein using bicinchoninic acid. *Anal. Biochem.*, 150(1), 76–85.

Tahirov T., Babayeva N., Varzavand K., Cooper J., Sedore S., Price D. (2010). Crystal structure of HIV-1 Tat complexed with human P-TEFb. *Nature*, 465(7299), 747–751.

Trinh C., Hunter T., Stewart E., Phillips S., Hunter G. (2008). Purification, crystallization and X-ray structures of two manganese superoxide dismutase from *Caernohabditis elegans*. *Acta. Cryst.*, F64, 1110–1114.

Vella M., Hunter T., Farrugia C., Pearson A., Hunter G. (2014). Purification and Characterisation of xanthine oxidoreductases from local bovids in Malta. *Ad. Enzyme Res.*, 2, 54–63.

Watson H. (1969). The sterochemistry of the protein myoglobin. *Prog. Stereochem.*, 4, 299.

Watson J., Crick F. (1953). A Structure for Deoxyribose Nucleic Acid. *Nature*, 171, 737–738.