1. Protein expression, purification and characterization

As defined earlier, the major objective of methods employed in proteomics is the total characterization of the protein. A thorough examination of the protein profile requires several steps, ranging from the proteins' identification and structural determination to the study of its post-translational modifications and from its quantification to the handling of the resulting proteomic data. In order to study any protein it is necessary to obtain it in a purified form. This is often a challenging task, particularly if proteins are present within the cell in low concentration. Frequently, this involves the purification of one single protein from a cell paste encompassing over 10.0000 different proteins. Two major alternatives are employed for isolating proteins. First, proteins can be isolated conventionally by obtaining the desired protein directly from the used source, such as a cell or tissue. Second, proteins can be expressed recombinantly, e.g. by introducing the DNA-sequence into a bacterial host.[108] In recent years, there have been numerous technical advances for proteomic technologies. Most commonly used methods for protein separation and identification are 2-D gel electrophoresis for protein separation and the proteome's analysis by mass spectrometry.[109] With the study of some proteins still being difficult to accomplish, further development of these tools is needed.

a) Gel electrophoresis

2D electrophoresis aims to separate proteins according to mass and overall charge. The technology is classified as the most common method for analyzing the purity of an isolated protein.[110] The principle of electrophoresis is the separation of proteins according to molecular mass by their movement through a polyacrylamide gel of closely defined composition under the influence of an electric field. The mobility of a protein through polyacrylamide gels is determined by a combination of overall charge, molecular shape, and molecular weight. The method is conducted by introducing a protein mixture to the top of a gel that proceeds through the matrix because of the electric field, with lighter components migrating faster than 'heavier' molecules. Over time, the component proteins are separated and the resolving power of

108 Whitford, David, Proteins: Structure and Function, Chichester, West Sussex, U.K., 313.

109 Another frequently employed method fort the purification of proteins is chromatography, see Whitford, David, Proteins: Structure and Function, Chichester, West Sussex, U.K., 326. There exist a number of different chromatographic methods.

110 See Gorg, Angelika/Weiss, Walter/Dunn, Michael J., Current two-dimensional electrophoresis technology for proteomics, 4 Proteomics 2004, 3665, 3665. The author considers two-dimensional gel electrophoresis (2-DE) with immobilized pH gradients (IPGs) combined with protein identification by mass spectrometry (MS) 'the workhorse' of proteomics.

the technique is sufficiently high that heterogeneous mixtures of proteins can be separated and distinguished from each other. The movement of proteins through the gel depends on the voltage/current conditions used as well as the temperature. Commonly, the mobility of an unknown protein or mixture of proteins is compared to that of a pure component of known molecular mass. Using this method for individual cell types or organisms makes it possible to identify large numbers of different proteins within proteomes of single-celled organism or individual cells. The technique allows the monomeric molecular mass to be determined with reasonable accuracy. [111]

b) Mass spectrometry

Mass spectrometry has emerged as the central analytical technique in proteomic analysis.[112] Like gel electrophoresis, the method is based on the discovery that the mass of a protein is one of the most useful characteristics for its identification.[113] Its observed parameter is the mass-to- charge ratio of gas phase ions, e.g. of electrically charged proteins in vapor state. Typically, mass spectrometers involve three basic components: an ion source, a mass analyzer and a detector. Ions are produced from samples generating charged states, which the mass analyzer separates according to their charge ratio. Simultaneously, a detector produces quantifiable signals. [114] Finally, the magnitude of these signals is recorded and converted into a mass spectrum. Early mass spectrometers required the sample to be a gas. Modern instrumentation, specifically the popular methods of "matrix assisted laser desorption time of flight (MALDI-TOF)" and "electrospray spectrometry", enable the analysis of ions embedded in a matrix or liquid solution samples. Over the last 20 years mass spectrometry has advanced rapidly and specifically in the area of proteomics. The importance of mass spectrometry for protein characterization was demonstrated by the award of the Nobel Prize for Chemistry in 2002.[115] The introduction of the techniques[116] described above enables accurate mass determination for primary sequences. Moreover, they allow for the detection of post-translational modifications

111 Whitford, David, Proteins: Structure and Function, Chichester, West Sussex, U.K., 333.

112 Recent successes demonstrate the role of mass spectrometry-based proteomics as an decisive tool in molecular and cellular biology, Aebersold, Ruedi/Mann, Matthias, Mass Spectrometry-based Proteomics, 422 Nature 2003, 198, 198.

113 After a protein has been isolated by gel electrophoresis, it is typically analyzed further by mass spectrometry.

114 The development of MALDI-TOF, where a matrix assists in the formation of a gas phase protein ion enabled to overcome existing practical difficulties, such as the non-volatility of proteins.

115 The prize was given to John Fenn and Koichi Tanaka, two pioneers in the field of mass spectrometry, for their research related to the development of 'soft' desorption-ionization methods in mass spectrometry.

116 Such as MALDI-TOF and electrospray methods.

and single residue mutations in genetically engineered proteins. Large numbers of inventions focus on the improvement of mass spectrometry tools. [117]

2. Physical methods of determining the three-dimensional structure of proteins

a) Protein Crystallization

Another core element of proteomics is the development of methods leading to protein structure determination.[118] One of the most common methods of structure determination is protein crystallization. Protein crystals are characterized by a high degree of internal three-dimensional order and a definite overall chemical composition.[119] The crystallization process of molecules of any substance from its solution is characterized by a reversible equilibrium phenomenon, determined by the minimization of the free energy of the system. A solution in which the molecules are fully solvated[120] corresponds to the system at equilibrium; its free energy is minimized. If the amount of molecules in the solution is increased, the system goes through internal changes until the point is reached where there is insufficient liquid to maintain full hydration of the molecules. These conditions are called "the supersaturation state".[121] Crystallizing purified proteins is not only a time consuming process, but also requires a significant amount of protein sample. There are a number of techniques, which have been developed for bringing a protein solution into a supersaturation state. Among them, the most common methods are micro-batch, vapor-

117 Whitford, David, Proteins: Structure and Function, Chichester, West Sussex, U.K., 345. The development of "surface enhanced laser desorption" (SELDI) a technology that facilitates the fast monitoring of biomarkers for cancer diagnosis, is one example of a new mass spectrometry system, see Langbein, William, Mass Spec meets Oncology - A prolific pair of governments researchers developed a Proteomic Bar Code for detecting cancer, Genome Technology 2003, 42, 43. Another recently developed mass spectrometry tool is "Fourier transform mass spectrometry" (FT/MS). It is particularly used for the identification of post-translational modifications; see MacNeil, John S., Making things happen, Genome Technology 2003, 34, 34.

118 Whitford, David, Proteins: Structure and Function, Chichester, West Sussex, U.K., 347.

119 The term crystal comes from the Greek word "krustallos" (clear ice). Like clear ice, crystals are homogeneous solids, many of them having a transparant sparkling appearance and a well-defined geometrical shape, with regular faces and sharp edges, see Chirgadze, Dima, Protein Crystallization in Action, 3, available at http://daffy.bioc.cam.ac.uk/~dima/whitepapers/xtal-in-action/, last checked on July 5, 2005.

120 A liquid substance is considered as solvent if it is capable of dissolving other substances. One characteristic of a solvent is that the substance does not change its state in forming a solution. Solvation is a chemical process in which solvent molecules and molecules or ions of the solute combine to form a compound, see http://www.wordreference.com, last checked on January 21, 2008.

121 See Chirgadze, Dima, Protein Crystallization in Action, 3, available at http://daffy.bioc.cam.ac.uk/ ~dima/whitepapers/xtal-in-action/, last checked on July 5, 2005.

diffusion and dialysis.[122] Although supersaturation of a protein solution can be attained by means of each of these procedures, the underlying principles of these methods differ.

Micro-Batch crystallization[123] involves the direct mixing of the undersaturated protein solution with a precipitant solution. The method aims to produce a final supersaturated concentration, which may eventually lead to crystallization. This is achieved with large amounts of solutions, and typically results in larger crystals owing to the larger volumes of solute present and the lower chance of impurities diffusing onto the face of the crystal. The main disadvantage of the micro-batch technique is that equilibration takes place very rapidly and therefore affects the rate of crystal growth and consequently the quality of the obtained crystals. Nevertheless, since the use of very small volumes of protein solution can be made, the method is quite useful as an early screening method.[124]

Vapor diffusion is the standard method utilized for protein crystallization.[125] It is the favored technique when screening large numbers of conditions.[126] Vapor diffusion is based on evaporation and diffusion of water between solutions of different concentrations as a means of approaching supersaturation of proteins. Typically, the protein solution is mixed in a 1:1 ratio with a solution containing the precipitant agent at the concentration required after vapor equilibration has occurred. A drop is then suspended and sealed over the well solution, which contains the precipitant solution at the target concentration. The difference in precipitant concentration between the drop and the well solution acts as the driving force. It leads to the vaporization of the drop until the concentration of the precipitant in the drop equals that of the well solution.

122 Chirgadze, Dima, Protein Crystallization in Action, 7, available at, http://daffy.bioc.cam.ac.uk/ ~dima/whitepapers/xtal-in-action/, last checked on July 5, 2005; as well as Ng, Joseph D./Gavira, Jose A./Garcia-Ruiz, Juan M., Protein crystallization by capillary counterdiffusion for applied crystallographic structure determination, 142 Journal of Structural Biology 2003, 218 who do not explicitly refer to the method of dialysis.

123 Ng, Joseph D./Gavira, Jose A./Garcia-Ruiz, Juan M., Protein crystallization by capillary counterdiffusion for applied crystallographic structure determination, 142 Journal of Structural Biology 2003, 218, 220. Micro-Batch is a variation of the simple batch crystallization technique.

124 Chirgadze, Dima, Protein Crystallization in Action, 7, available at http://daffy.bioc.cam.ac.uk/ ~dima/whitepapers/xtal-in-action/, last checked on July 5, 2005. This method was successfully conducted in order to obtain the initial NK1 protein crystallization conditions.

125 Ng, Joseph D./Gavira, Jose A./Garcia-Ruiz, Juan M., Protein crystallization by capillary counterdiffusion for applied crystallographic structure determination, 142 Journal of Structural Biology 2003, 218, 220; Whitford, David, Proteins: Structure and Function, Chichester, West Sussex, U.K., 359.

126 Chirgadze, Dima, Protein Crystallization in Action, 10, available at http://daffy.bioc.cam.ac.uk/ ~dima/whitepapers/xtal-in-action/, last checked on July 5, 2005. This technique can also be conducted to increase or decrease the concentration of proteins in the equilibrated state, relative to its initial concentration.

Dialysis techniques are typically conducted for proteins at low and high ionic strength.[127] They employ diffusion and equilibration of small precipitant molecules through a semipermeable membrane as a way of slowly approaching the concentration at which the macromolecule solute crystallizes. In a preliminary step, the protein solution is contained within the dialysis membrane, which is then equilibrated against a precipitant solution. Equilibration against the precipitant in the surrounding solvent slowly reaches supersaturation for the solute within the dialysis membrane, eventually resulting in crystallization. The improvement of dialysis over other methods is in the ease with which the precipitating solution can be varied, simply by shifting the entire dialysis button from one condition to another. Hence, the protein solution can be continuously recycled until the correct conditions for crystallization are obtained.[128]

b) X-ray crystallography

X-ray crystallography is a technique in which the pattern produced by the diffraction of x-rays through the closely spaced lattice of atoms in a crystal is recorded and then analyzed to reveal the nature of that lattice. It can provide an astonishingly fine visualization of protein structure, since it reveals the precise three-dimensional positions of most atoms in a protein molecule.[129]

The material and molecular structure of a substance can often be inferred by quantitative study of this pattern. It is widely used in chemistry and biochemistry to determine the structures of molecules, including DNA and proteins. The first protein structure of myoglobin was disclosed by Max Perutz and Sir John Cowdery Kendrew in 1958 and led to a Nobel Prize in Chemistry.[130] To determine a structure, one must obtain crystals of the protein of interest. This can be a painstaking procedure for macromolecules. Many proteins, such as hydrophobic or membrane-associated proteins, might not crystallize at all. Actually, it is generally possible to achieve crystalline forms of only 5-10 % of proteins, even though increasingly large and complex polypeptides are being crystallized.[131] Some proteins crystallize readi-

127 Whitford, David, Proteins: Structure and Function, Chichester, West Sussex, U.K., 359.

128 Chirgadze, Dima, Protein Crystallization in Action, 10, available at http://daffy.bioc.cam.ac.uk/~dima/whitepapers/xtal-in-action/, last checked on July 5, 2005. Under these conditions, the protein solution can be continiously recycled until the correct conditions for crystallization are obtained.

129 Pusey, Marc L./Liu, Zhi-Jie/Tempel, Wolfram/Praissman, Jeremy/Lin, Dawei/Wang, Bi-Cheng/Gavira, Jose A./Ng, Joseph D., Life in the Fast Lane for Protein Crystallization and X-ray Crystallography, 88 Progress in Biophysics & Molecular Biology 2005, 359 describes X-ray crystallography as the "foremost method" to acquire data relating to the three-dimensional structures for a multitude of proteins.

130 Today, X-ray crystallography is often used to determine how drugs, such as anti-cancer medications, can be improved to better influence their protein targets.

131 See Maggio, Edward T./Ramnarayan, Kal, Recent Developments in Computational Proteomics, 19 Trends in Biotechnology 2001, 266, 266.

ly, whereas others do so only after considerable effort has been spent in determining the optimal conditions. After the crystallization of the substance, the crystals are harvested and often frozen with liquid nitrogen. Freezing the crystals both reduces radiation damage incurred during data collection and decreases thermal motion within the crystal. Crystals are placed on a diffractometer, a machine that emits a beam of x-rays. The x-rays diffract off the electrons in the crystal. The crystal is rotated such that the beam can strike the crystal from many directions. This rotational motion results in an x-ray photograph consisting of a regular array of spots called reflections. The intensity of each spot is measured. These intensities and their positions are the basic experimental data of an x-ray crystallographic analysis. The observed intensities are then used to reconstruct an image of the protein. Furthermore, an electron density map is calculated, which serves for the determination of the density of electrons at a large number of regularly spaced points in the crystal.[132] In the next step, this density map is interpreted. The resolution of the x-ray analysis is determined by the number of scattered intensities. Once a model of a protein's structure has been determined, it is deposited in the Protein Data Bank (PDB).[133] The development of new methods for solving x-ray crystal structures is considered an important field of research.[134]

c) NMR structure determination

Structure Determination of Proteins with NMR Spectroscopy is another classic protein analysis technique. It is accomplished by the determination of the biological macromolecular structure at atomic resolution, but it is only possible with water-soluble proteins. The technique is based on the fact that energy levels of atomic nuclei are split by a magnetic field. Transitions between these energy levels can be achieved by exciting the sample with radiation whose frequency is equivalent to the energy difference between the two levels. The field of NMR spectroscopy has recently experienced an explosive growth, which started with the development of pulsed Fourier-transform NMR and multidimensional NMR spectroscopy and continues today. Progress in the theoretical and practical capabilities of NMR spectroscopy leads to an increasingly efficient utilization of the information content related

132 The term “electron density map” refers to the distribution of electron density in a crystal that is measured by the X-ray diffraction template.

133 Berg, Jeremy M./Tymoczko, John L./Stryer Lubert, Biochemistry, New York, NY, 2002, 110-112. Peters, Linde, http://www.boa-muenchen.org/linde.peters/postgen0.htm#top, Part IV, 1-14.

134 For a method relying on the introduction of iodine into proteins, see Dauter, Zbigniew, Phasing in Iodine for Structure Determination, 22 Nature Biotechnology 2004, 1239.

to it. Parallel developments in the biochemical methods (such as recombinant protein expression) allow simple and rapid preparation of protein samples.[135]

d) Protein modeling (homologous-comparison)

In addition to protein crystallization or synchronization that is not possible with all existing proteins, new computational methodologies have recently yielded modeled structures that are, in many cases, quantitatively comparable to crystal structures.[136] The method of homology modeling of proteins relies on the structural knowledge of proteins for which 3-D structures have been determined, in order to infer the structure of other proteins for which only the sequence is known. The sequence of a polypeptide of unknown structure is combined with the template of another polypeptide in an attempt to predict the unknown structure. Obtained data helps to determine the structure of homologous proteins. For comparative protein modeling, at least one sequence of known 3-D structure with significant similarity to the target sequence is required. Accordingly, protein modeling is only limited by the need for at least one crystal structure within each fold-class to be modeled. In order to determine whether a modeling request can be carried out, one compares the target sequence with a database of sequences derived from a sequence-related protein database using the corresponding bioinformatics program. This method might lead to the selection of several suitable templates for a given target sequence. Commonly, up to ten templates are used in the modeling process. The best template structure, which is the one with the highest sequence similarity to the target, will be chosen as the reference. As a next step, the target sequence needs to be aligned with the template sequence. Residues that are unimportant for the model building will be ignored during the modeling process. Thus, the common core of the target protein and the loops defined by at least one supplied template structure are simulated. Further, the position of each atom in the target sequence is averaged with the help of the location of the corresponding atoms in the template. Those loops for which no structural information is available in the template structure are not defined and thus must be simulated. Most of the known 3-D structures available do not share complete similarity with the template. However, there may be similarities in the loop regions, which can be simulated as loop structure of the new protein. The loop fragments are extracted from the searched protein database. Since each loop is defined by its length and particular atom co-ordinates of its residues, they can be indicated by using particular algo-

135 For a detailed description see Griesinger, Christian, Proteinstruktur-Aufklärung durch 3D-NMR-Spektroskopie, Laborwelt 2003, 10, 10ff; another description is provided by the Max-Planck-Institute for biochemistry, available at: http://www.cryst.bbk.ac.uk/PPS2/projects/schirra/ html/ home.htm, last checked October 12, 2004.

136 Maggio, Edward T./Ramnarayan, Kal, Recent developments in computational proteomics, 19 Trends in Biotechnology 2001, 266-272, providing a general review about computational proteomics.

rithms. Since the templates fail to provide the structural information of many protein side chains, these can therefore not be assimilated in a first attempt and must be added later by further simulation. The amount of side chains required to be constructed depends on the degree of sequence identity between target and template protein. The method has its major weakness regarding protein divergences, which are not connected to any typical homologous family. [137]

III. Data and Bioinformatics for proteomics

1. Databases

To catalogue all human proteins and reveal their function and interaction is an immense challenge for scientists. The number of databases providing biological information is steadily increasing.[138] A new discipline, bioinformatics, has emerged to utilize the information in these databases for a better understanding of biological processes. Bioinformatics scientists work to interpret experimental data. Several sequence and sequence-related databases are available for public use. Additionally, a number of specialized databases exist which focus on a single enzyme, protein family or disease.[139] For amino acid sequences, the databases differ in their content. SWISS-Prot, established in 1986, strives to provide a high level of protein annotation with several cross-links to other databases.[140] Since 2003, it has been carried out by the UniProt Consortium, a collaboration between the Swiss Institute of Bioinformatics (SIB) and the Department of Bioinformatics and Structural Biology of the Geneva University, the European Bioinformatics Institute (EBI) and the Georgetown

137 Maggio, Edward T./Ramnarayan, Kal, Recent Developments in Computational Proteomics, 19 Trends in Biotechnology 2001, 266, 271.

138 Detailed overview of bioinformatics techniques of importance in protein analysis is provided by: Persson, Bengt, Bioinformatics in protein analysis, In: Proteomics in Functional Genomics - Protein Structure Analysis; Jollès, P./Jörnvall, H. Ed. Basel, Boston, Berlin, 2000; 215. Access to most protein databases is free. Recently, however, many providers stopped granting open access and started requiring licenses from commercial users. In the future, even academic users might have to register and pay. Goodman, Phillip, Access Ability, Genome Technology 2004, 21. Carugo, Oliviero/Pongor, Sàndor, The Evolution of Structural Databases, 20 Trends in Biotechnology 2002, 498 emphasize that the evolution of structural databases has been driven by the practical application of structural knowledge.

139 Links to biologically relevant databases are available at the web pages of EBI (European Bioinformatics Institute, Hinxton, England; http://www.ebi.ac.uk), the University of Geneva, Switzerland (http://www.expasy.ch), and NCBI (National Center for Biological Information, Bethesda, MD, USA; http://ncbi.nlm.nih.gov).

140 An annotation gives a narrative description to the formal structure of a protein, Carugo, Oliviero/Pongor, Sàndor, The Evolution of Structural Databases, 20 Trends in Biotechnology 2002, 498, 498.