BioTechnologia vol. 94(1) • pp. 9-14 • 2013
Journal of Biotechnology, Computational Biology and Bionanotechnology REVIEW PAPER

Identification of amino acid sequences *via* X-ray crystallography: a mini review of case studies

AGNIESZKA J. PIETRZYK [1], ANNA BUJACZ [2], MARIUSZ JASKOLSKI [1, 3], GRZEGORZ BUJACZ [1, 2*]

[1] Institute of Bioorganic Chemistry, Polish Academy of Sciences, Poznań, Poland

[2] Faculty of Biotechnology and Food Sciences, Lodz University of Technology, Łódź, Poland

[3] Faculty of Chemistry, Adam Mickiewicz University, Poznań, Poland

* Corresponding author: grzegorz.bujacz@p.lodz.pl

Abstract

The sequencing of a protein is a complicated issue. Several methods have been developed in order to establish the amino acid sequences of proteins, including Edman degradation, LC/MS/MS and cDNA sequencing. A number of examples confirm that X-ray crystallography could also be a useful tool for the identification of an amino acid sequence of unknown proteins. Here, we present a short review summarizing the application of crystallography for protein sequence determination and identification.

Introduction

The amino acid sequence of a particular protein can be determined using different chemical and biochemical methods, the most popular ones being Edman degradation, liquid chromatography and electrospray ionization tandem mass spectrometry (LC/MS/MS), and cDNA sequencing. Edman degradation was the first approach to be developed. In this method, the single N-terminal residue is chemically labeled, cleaved from the peptide and identified (Edman et al., 1950). About 40-50 N-terminal amino acid residues can be sequenced with this method; prior to this analysis, the protein is often cleaved into smaller peptides, which makes the determination of the complete protein sequence even more complicated. Moreover, there are some limitations to this method: if the N-terminus is blocked or buried in the protein molecule, the Edman procedure will not work. The other disadvantage of this technique is its high cost (Hou et al., 2007).

The first step in sequencing by LC/MS/MS is also enzymatic digestion of a protein polypeptide chain into short peptides. The peptides are separated using a reversed-phase LC (liquid chromatography) column and their tandem mass spectra are recorded and used to search protein sequence databases in order to identify the unknown sequence (McCornack et al., 1997). The main limitation of this method is the content of the databases. A protein can be identified with the LC/MS/MS approach only if its amino acid sequence has been previously deposited in the selected database(s).

cDNA sequencing is another popular method, but it deduces the protein amino acid sequence from the coding sequence of the corresponding DNA. Typically, the sequence information is obtained from the messenger mRNA transcript coding for the protein in question. The enzyme reverse transcriptase is used to obtain the complementary cDNA which is then sequenced with standard DNS sequencing techniques (Gubler and Hoffman, 1983). It should be emphasized that this method requires the isolation of the mRNA transcript.

Although the main application of macromolecular X-ray crystallography is the determination of protein 3D structure, it is also a powerful tool for protein sequencing and identification, especially at high resolution. This review presents several examples of successful identification of protein amino acid sequences using X-ray crystallography. Selected publications describing protein sequencing based on electron density maps are shown in Table 1.

Historical background

Almost from its inception as a method for protein structure determination, X-ray crystallography has also been used as an auxiliary method for resolving sequence-

Table 1. List of publications discussed in this review

Protein	Protein source	Resolution of X-ray data [Å]	PDB code	N-terminal sequencing	Edman degradation [a]	LC/MS/MS	References
Hexokinase B	*Saccharomyces cerevisiae*	2.1	2YXH	–	+	–	Anderson et al., 1978
Hen egg lysozyme	*Gallus gallus*	2.0	1LYZ 2LYZ 3LYZ 4LYZ 5LYZ 6LYZ	+	+	–	Blake et al., 1965 Diamond, 1974
Trichomaglin	*Trichosanthes lepiniana*	2.2	1SGL	+	–	+	Gan et al., 2004
Concanavalin A	*Canavalia ensiformis*	2.4	3CNA	+	+	–	Hardman and Ainsworth, 1972
β-mannanase	*Thermomonospora fusca*	1.5	1BQC	+	–	–	Hilge et al., 1998 Hilge et al., 2001
Luffaculin I	*Luffa acutangula*	1.4	2OQA	+	–	–	Hou et al., 2007
Chondroitin lyase AC (ArthroAC)	*Arthrobacter aurescens*	1.4 1.3 1.9 1.5 1.5 1.3	1RW9 1RWA 1RWC 1RWF 1RWG 1RWH	+	–	+	Lunin et al., 2004
Xylanase	*Thermoascus aurantiacus*	1.8	1TUX	–	–	–	Natesh et al., 1999
B. mori lipoprotein 7 (Bmlp7)	*Bombyx mori*	1.3 1.9 2.5	4EFP 4EFQ 4EFR	+	–	+	Pietrzyk et al., 2012
B. mori lipoprotein 3 (Bmlp3)	*Bombyx mori*	2.4 2.1	4IY8 4IY9	+	–	–	Pietrzyk et al., 2013
β-galactosidase	*Penicillium* sp.	1.9 2.1	1TG7 1XC6	–	–	–	Rojas et al., 2004
Fab fragment of RU5 antibody	*Mus musculus*	2.0	1FE8	–	–	–	Romijn et al., 2001
Brefeldin A esterase (BFAE)	*Bacillus subtilis*	1.9	1JKM	–	–	–	Wei et al., 1999

[a] Sequencing of internal peptides by Edman degradation

related ambiguities. A case in point is the story of the hen egg lysozyme (HEL) which has been extensively studied over recent decades. The complete amino acid sequence of this protein was reported simultaneously but independently by two research groups (Canfield, 1963; Jolles et al., 1963). However, there were several differences at a number of key positions. The determination of the lysozyme crystal structure (Blake et al., 1965; Diamond, 1974) confirmed that the proper sequence was that described by Canfield (1963). It is noteworthy that the hen egg lysozyme was only the third protein to have had its three-dimensional structure determined with X-ray crystallography.

Another interesting example is the case of concanavalin A, a well-known lectin isolated from jack-bean. Its crystal structure was solved in 1972 at 2.4 Å resolution (Hardman and Ainsworth, 1972). Not much had been known about its amino acid sequence before the 3D structure determination. Four of its peptides had been sequenced, but this covered only 15% of the total protein sequence. The crystallographic electron density maps provided the missing information about the unknown parts of the amino acid sequence (Hardman and Ainsworth, 1972). A similar situation was described for yeast hexokinase B, for which only two short peptides had been sequenced chemically, and the first complete amino acid sequence was established *via* visual inspection of the electron density maps (Anderson et al., 1978).

Identification of an amino acid sequence from electron density maps

At present, advancement in the technology of recombinant protein expression has made this approach the method of choice for protein production. In this scheme, especially in high-throughput genomic approaches, a protein sequence deduced from cDNA sequence is a convenient by-product of the methodology. However, some projects still rely on proteins isolated from their natural source, especially when the coding sequence is unknown. In such cases, X-ray crystallography provides not only information on the 3D structure, but often also on the amino acid sequence. In recent years, X-ray crystallography has significantly contributed to the amino acid sequence determination of proteins isolated from different sources, including microbes (Hilge et al., 1998; 2001; Lunin et al., 2004; Natesh et al., 1999; Rojas et al., 2004; Wei et al., 1999), plants (Gan et al., 2004; Hou et al., 2007), and animals (Pietrzyk et al., 2012; 2013; Romijn et al., 2001). It is noteworthy that for most of the proteins mentioned above, information on their amino acid sequence was not provided in the available databases. On the other hand, it does sometimes happen that the available sequences are not accurate, as illustrated by two cases studied in our laboratory. In those cases, the two most abundant proteins were isolated from the hemolymph of the mulberry silkworm (*Bombyx mori*) as unknown proteins. Furthermore, the result of LC/MS/MS analyses for one of them was incorrect (Pietrzyk et al., 2011), because the database search did not include the Silkworm Knowledgebase (Duan et al., 2010), where the proper sequence was deposited. The final identification of both was performed according to electron density maps, which classified these proteins as lipoproteins Bmlp7 (Pietrzyk et al., 2012) and Bmlp3 (Pietrzyk et al., 2013). Although the amino acid sequences of both proteins share 94% similarity and their structural similarity is also high (Fig. 1), the electron density maps indicated differences in residue side chains (Pietrzyk et al., 2012; 2013).

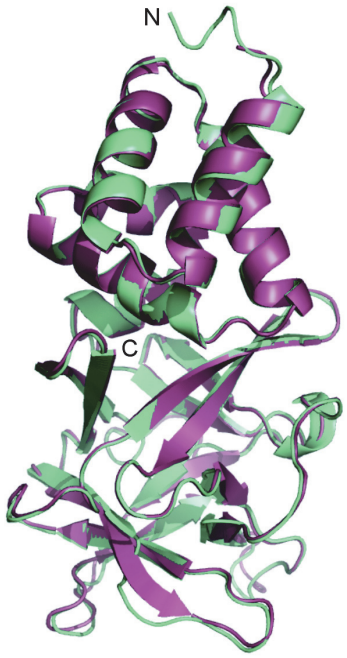

Fig. 1. The Cα-superposition of Bmlp3 (green; PDB code: 4IY8; Pietrzyk et al., 2013) and Bmlp7 (red; PDB code: 4EFP; Pietrzyk et al., 2012) indicates high structural similarities

Another advantage of crystallographic sequencing is that electron density maps of natural macromolecules can provide crucial information on post-translational mo-

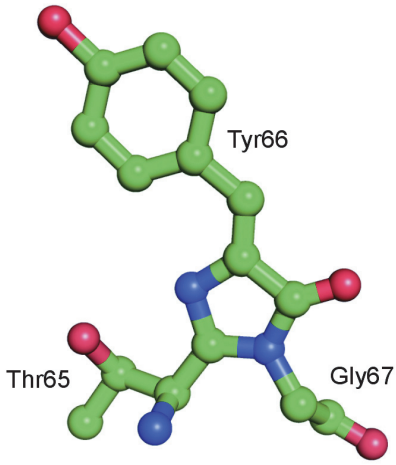

Fig. 2. The chromophore of GFP is formed from three residues: threonine, tyrosine and glycine (PDB code: 1EMA; Ormo et al., 1996)

difications of the polypeptide chain. These include conjugation of the side chains with signaling molecules, such as phosphate, carbohydrate or methyl groups, or even more radical chemical modifications. An excellent example of such a case is the elucidation from a crystallographic structure of the post-translational chromophore (Fig. 2) formation in the green fluorescent protein, GFP (Ormo et al., 1996; Yang et al., 1996; Palm et al., 1997).

It is also possible to deduce the nature of chemical modifications by a combination (usually complicated) of spectroscopic and mass spectrometric methods. However, the crystallographic evidence, especially at high resolutions, is straightforward and even if ambiguous, it can provide hints for further analyses by complementary methods.

Furthermore, some sequences available in the databases contain errors. The amino acid sequence of brefeldin A esterase was determined based on its cDNA sequencing. Brefeldin A esterase is an enzyme isolated from *Bacillus subtilis,* which is capable of hydrolyzing brefeldin A, a lactone antibiotic produced by fungal organisms. The protein was obtained using the original cDNA and its crystal structure revealed significant discrepancies between the sequence from the database and the side chains visible in the electron density maps. The errors were tracked down to erroneous insertions or deletions of C or G nucleotides during the cDNA sequencing experiment (Wei et al., 1999).

In addition, the initial sequence obtained from electron density maps can be used for designing specific primers to be utilized in further gene recombination experiments. The gene sequence can be then derived from genomic DNA following a standard PCR reaction. Such an approach worked perfectly for β-mannanase from *T. fusca* (Hilge et al., 1998; 2001) and β-galactosidase from *Penicillium* sp. (Rojas et al., 2004).

Nevertheless, it should be emphasized that sequencing from electron density maps should be confirmed by alternative methods. The simplest way for protein identity confirmation is N-terminal sequencing or LC/MS/MS analysis. The database hits so obtained can then be checked against the sequence derived from crystallography to clarify ambiguities, correct errors, or provide final confirmation. Table 1 shows whether additional analyses were performed for each of the discussed proteins.

Another possibility to corroborate X-ray sequence determination is through multiple sequence alignment, using either proteins of homologous sequences or those belonging to the same family. For example, the monoclonal antibody RU5 was produced in mice and its complete amino acid sequence was unknown, since information on the variable domain was missing. When the crystal structure of the complex between RU5 and domain A3 of the von Willebrand factor was solved, the amino acid sequence of the variable domain of RU5 was determined by "electron density sequencing" combined with sequence alignment of 110 different antibodies (Romijn et al., 2001). A comparison of a number of related sequences enables the identification of conserved residues, as demonstrated by the cases of luffaculin I (Hou et al., 2007), xylanase (Natesh et al., 1999), β-galactosidase (Rojas et al., 2004) and silkworm lipoproteins (Pietrzyk et al., 2012; 2013).

The main criteria, the advantages and disadvantages of sequencing based on electron density maps

The amino acid sequence obtained with X-ray crystallography is reliable if several conditions are fulfilled. Most importantly, the resolution of the X-ray diffraction data should be reasonable; the resolution of the cases discussed in this review ranges from 2.5 to 1.3 Å. However, even 3.0 Å data could also be useful for verification of whether large side chains (mainly aromatic residues) have been modeled correctly. The differences between electron density maps at high and low resolution are shown in Figure 3. The quality of X-ray diffraction data is another important factor. By eliminating noise from

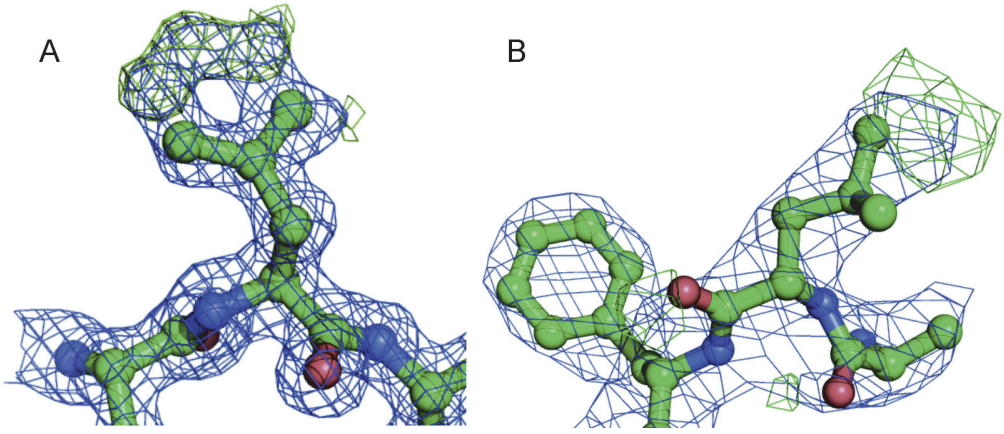

Fig. 3. This figure presents different shapes of electron density maps at 1.3 Å (A) and 2.9 Å resolution (B). The 2Fo-Fc maps are displayed in blue at the 1.0σ level and the Fo-Fc maps are displayed in green at the 3.0σ level. In both cases, the central leucine residue should be changed to phenylalanine. Both the presented protein fragments are initial models of two silkworm proteins: Bmlp7 (A) (Pietrzyk et al., 2012) and a high molecular weight lipoprotein (unpublished data)

the input data, one tends to obtain clearer electron density maps, which obviously aids the interpretation process. Finally, a comparison of electron density corresponding to two or more protein molecules, if present in the asymmetric unit, or a comparison of protein molecules in different crystal forms is extremely helpful at ambiguous sites. At a high resolution, some ambiguities, such as for the Val/Thr pairs, can be resolved when the chemical environment of the residue is being analyzed. Taking everything into account, on average, more than 80% of all amino acid residues can be unequivocally assigned on the basis of electron density maps (Hou et al., 2007).

The main advantage of sequencing based on electron density maps is that this approach does not require additional experiments, since each 3D structure also contains information on the protein's primary structure.

Considering the disadvantages of sequencing by X-ray crystallography, it should be pointed out that the most problematic positions are Asx and Glx. A clear-cut distinction between Asn and Asp, or Gln and Glu is often almost impossible, although even in such close cases help can be obtained from the analysis of H-bonding networks (the NH_2 group can only act as a donor) or (with high resolution data) from the analysis of the atomic displacement parameters (ADPs or temperature factors), which should have a balanced distribution (for properly assigned atom types) rather than a jumpy pattern. A serious problem arises from the fact that residues located at the protein surface are often disordered and have poor electron density, making their identification extremely difficult. Nonetheless, X-ray crystallography has significantly contributed to amino acid sequence assignment and to the identification of novel proteins with unknown primary structure.

Acknowledgments

Part of this work was supported by the European Union within the framework of the European Regional Development Fund and by grant 2011/03/B/NZ1/01238 from the National Science Centre to GB.

References

Anderson C.M., Stenkamp R.E., Steitz T.A. (1978) J. Mol. Biol. 123: 15-33.

Blake C.C.F., Koenig D.F., Mair G.A., North A.C.T., Phillips D.C., Sarma V.R. (1965) Nature 206: 757-761.

Canfield R.E. (1963) J. Biol. Chem. 238: 2698-2707.

Diamond R. (1974) J. Mol. Biol. 82: 371-391.

Duan J., Li R., Cheng D., Fan W., Zha X., Cheng T., Wu Y., Wang J., Mita K., Xiang Z., Xia Q. (2010) Nucl. Acids Res. 38: D453-D456.

Edman P. (1950) Acta Chem. Scand. 4: 283-293.

Gan J.H., Yu L., Wu J., Xu H., Choudhary J.S., Blackstock W.P., Liu W.Y., Xia Z.X. (2004) Structure 12: 1015-1025.

Gubler U., Hoffman B.J. (1983) Gene 25: 263-269.

Hardman K.D., Ainsworth C.F. (1972) Biochemistry 11: 4910-4919.

Hilge M., Gloor S.M., Rypniewski W., Sauer O., Heightman T.D., Zimmermann W., Winterhalter K., Piontek K. (1998) Structure 6: 1433-1444.

Hilge M., Perrakis A., Abrahams J.P., Winterhalter K., Pionteka K., Gloor S.M. (2001) Acta Cryst. D57: 37-43.

Hou X., Chen M., Chen L., Meehan E.J., Xie J., Huang M. (2007) BMC Struct. Biol. 7: 29.

Jolles J., Jauregin-Adell J., Bernier I., Jolles P. (1963) Biochim. Biophys. Acta 78: 668.

Lunin V.V., Li Y., Linhardt R.J., Miyazono H., Kyogashima M., Kaneko T., Bell A.W., Cygler M. (2004) J. Mol. Biol. 337: 367-386.

McCormack A.L., Schieltz D.M., Goode B., Yang S., Barnes G., Drubin D., Yates J.R. (1997) Anal. Chem. 69: 767-776.

Natesh R., Bhanumoorthy P., Vithayathil P.J., Sekar K., Ramakumar S., Viswamitra M.A. (1999) J. Mol. Biol. 288: 999-1012.

Ormo M., Cubitt A.B., Kallio K., Gross L.A., Tsien R.Y., Remington S.J. (1996) Science 273: 1392-1395.

Palm G.J., Zdanov A., Gaitanaris G.A., Stauber R., Pavlakis G.N., Wlodawer A. (1997) Nat. Struct. Biol. 4: 361-365.

Pietrzyk A.J., Bujacz A., Łochyńska M., Jaskolski M., Bujacz G. (2011) Acta Cryst. F67: 372-376.

Pietrzyk A.J., Panjikar S., Bujacz A., Mueller-Dieckmann J., Lochynska M., Jaskolski M., Bujacz G. (2012) Acta Cryst. D68: 1140-1151.

Pietrzyk A.J., Bujacz A., Mueller-Dieckmann J., Lochynska M., Jaskolski M., Bujacz G. (2013) Plos One 8(4): e61303.

Rojas A.L., Nagem R.A., Neustroev K.N., Arand M., Adamska M., Eneyskaya E.V., Kulminskaya A.A., Garratt R.C., Golubev A.M., Polikarpov I. (2004) J. Mol. Biol.343: 1281-92.

Romijn R.A., Bouma B., Wuyster W., Gros P., Kroon J., Sixma J.J., Huizinga E.G. (2001) J. Biol. Chem. 276: 9985-9991.

Wei Y., Contreras J.A., Sheffield P., Osterlund T., Derewenda U., Kneusel R.E., Matern U., Holm C., Derewenda Z.S., (1999) Nat. Struct. Biol. 6: 340-345.

Yang F., Moss L.G., Phillips G.N. (1996) Nat. Biotechnol. 14: 1246-1251.