Beilstein-Institut

Experimental Standard Conditions of Enzyme Characterizations,
September 13th – 16th, 2009, Rüdesheim/Rhein, Germany

Suggestions for a Protein Species Identifier System

Hartmut Schlüter[1,*], Hermann-Georg Holzhütter[2], Rolf Apweiler[3] and Peter R. Jungblut[4]

[1]Institute of Clinical Chemistry, University Medicine Hamburg-Eppendorf, Martinistr. 53, 20246 Hamburg, Germany

[2]Computational Systems Chemistry, Charité – University Medicine Berlin, Monbijoustr. 2, 10117 Berlin, Germany

[3]EMBL Outstation Hinxton, European Bioinformatics Institute, Wellcome Trust Genome Campus, Hinxton, Cambridge, CB10 1SD, U.K.

[4]Max Planck Institute for Infection Biology, Core Facility Protein Analysis, Charitéplatz 1, 10177 Berlin, Germany

E-Mail: *hschluet@uke.de

Received: 25th January 2010 / Published: 14th September 2010

Abstract

Protein variants, which vary in their exact chemical composition, and which are coded by one gene or by a paralogous or orthologous gene or alleles of that gene, are called protein species. The term protein species covers splicing variants, truncated proteins and post-translational modified proteins, and is defined chemically in contrast to the term isoform, which is defined genetically. The impact of the knowledge of the exact chemical composition of a protein species is determined by the relationship between its composition and its function. Since centuries it is known that post-translational modifications such as phosphorylation critically determine the activity status of enzymes. Proteolytic truncations can activate proteases, peptide hormones or receptors. However, despite of this knowledge, the relationship between the exact chemical composition of a protein and its function is not sufficiently considered in many protein investigations. In many of the current proteomics studies protein identification is based on sequence coverage significantly lower than 100%. Post-translational

modifications are more or less ignored. A second drawback concerning the comprehensive description of protein species derives from the absence of an identifier system, which describes their exact chemical composition. Therefore, up to now we have to deal with a huge ambiguity concerning the identity of a protein and its function. In the past, functions were assigned to genes, implicating that the full information for the function is encoded in the DNA sequence. Now it becomes obvious that both different modifications and different combinations result in different protein species with different functions. An identifier system for protein species allows the assignment of a defined function to a defined protein species, which is determined by its exact chemical composition. The protein species identifier system was introduced in 2009 by Schlüter *et al.* and is presented here.

Introduction

In the past twenty years there was a tremendous increase of knowledge about proteins. The progress in this area was accelerated by the development of new methods in both molecular biology and in chemical structure analysis of proteins. In particular, the development of the soft ionization techniques – ESI (electrospray ionization) [1] and MALDI (matrix assisted laser desorption/ionization) [2] – in mass spectrometry (MS) improved the analysis of the protein composition. Furthermore, the complete sequencing of genomes (see [3,4]) extended the knowledge about proteins. At the same time further questions about the functions of the gene products – the proteins – came up.

From proteomics approaches (see [5]) as well as classical biochemical investigations we know since many years that one gene codes not only for one but for many gene products. These different products can be created by alternative splicing, proteolytic processing of proteins subsequent to the protein synthesis at the ribosome and/or post-translational processing with regard to the covalent addition of functional groups towards residues of the amino acids within the protein. At present, the database UniMod lists more than 600 post-translational modifications (UniMod version in December 2009). This huge quantity of individual post-translational modifications is responsible for a large number of products, the protein species, which can arise from one single gene. The term protein species was introduced by Jungblut *et al.* in 1996 [6]. It was extended for proteomics in 2008 [7], because according to the nomenclature rules of IUBMB the term "protein isoform" does not describe proteins which are encoded by one single gene but proteins with the same function encoded by different genes [8]. Nielsen *et al.* [9] performed a study to evaluate the extent of protein modification. The authors stated that the current estimation of the number of different protein molecules in human beings is close to a million when combining the complexity generated by alternative splicing with that produced by PTMs. This is roughly 50 protein species per gene. How relevant is this huge protein diversity? Since centuries it is well

known that covalent modifications, such as phosphorylation, critically determine the activity status of enzymes [10]. The type of linkage of ubiquitins in polyubiquitin chains determines whether a protein is degraded (linkages via lysine 48) or acts as a signal (linkages via lysine 63) within the cell [11]. For proteins, which were investigated in depth often many different functions are listed. For example, it was found out that Hsp70 is involved not only in chaperoning but also in cell growth, apoptosis and genetic recombination [12]. As a result of covalent modification the function of a defined protein can change completely. Another enzyme, GAPDH which is integrated in glycolysis, will induce an apoptosis after being nitrosylated [13]. Besides the covalent modifications of protein side chains, the modification of the amino acid sequence itself is also important for the function. It is well known that enzymes such as thrombin [14] are activated by proteases, which remove one or more peptides from the enzyme by hydrolysis of the peptide bond. A difference in the amino acid sequence of a protein encoded by one single gene can also be induced by alternative splicing: Two splice variants are known from the angiotensin converting enzyme (ACE) gene. The somatic splice variant is involved in blood pressure regulation. In contrast, the ACE splice variant present in testis is responsible for sperm maturation [15].

These examples highlight the importance of the relationship between the chemical composition of a protein and its function. Although the significance of this relationship is not questioned any more, a system which allows an unambiguous assignment of the chemical composition of a protein to its function is still missing. Thus, we propose a protein species identifier system for the description of the chemical composition of every protein species. This system is based on the suggestion published recently [16] and is extended here according to the results of the discussions of the ESCEC meeting in 2009. This suggestion is a framework which has to grow and to be optimized according to the needs of the operators of the protein species identifier system (PSIS). In the near future software will be developed which simplifies the conversion of data about protein species into the PSIS-description.

The Protein Species Identifier System

The protein species identifier system (PSIS) consists of descriptors that allow the determination of every known aspect of the chemical composition of a protein species.

The description of a defined protein species starts on the level of the coding gene *(level A)* (Table 1). The descriptor consists of the entry gene name according to the UniProt knowledgebase followed by the species name. Both terms are preceded by G, which here stands for gene. In the case of the angiotensin-converting enzyme (ACE), the descriptor for the coding gene is [G_ACE_human]. The second descriptor (*level B*) provides information about nucleotide polymorphisms (NP). The descriptor starts with NP followed by the NP – accession number according to dbSNP of the NCBI (http://www.ncbi.nlm.nih.gov/sites/entrez? db = snp), *e. g.* [NP_rs4331]. On *level C* the initial amino acid sequence of the protein synthesized at the ribosome is described: This descriptor includes the accession

number (AC) according to UniProt. For instance, the *level C* descriptor for the angiotensin-converting enzyme is [AC_P12821]. *Level D* refers to splicing. A splicing variant with a deletion (SD) is explained by the localisation of the missing amino acids (*level D-1*). [SD_81 – 97] indicates that in a given protein sequence the amino acids between position 81 and 97 are missing because of splicing. The testis-specific angiotensin-converting enzyme can be described by [SD_1 – 640]. *Level D-2*: The splicing variant containing an insertion (SI), is explained a) by the number of the amino acid, behind which one or several amino acids were inserted. The inserted amino acid(s) are given by the one-letter amino acid code. A *level D-2* descriptor may be: [SI_43_LELFVMFL].

The *level E* descriptor contains information about truncated amino acids (T). For example, active thrombin is generated by the proteolytic removal of the amino acids 1 – 24 (signal peptide), the amino acids 25 – 43 (pro-peptide), the amino acids 44 – 198 (activation peptide fragment 1) and the amino acids 199 – 327 (activation peptide fragment 2). Therefore, the *level E* descriptor for active thrombin is [T_1 – 327].

Post-translational modifications (P) are explained by the *level F* descriptor. The number following P_ indicates the position of the modified amino acid. The second number names the type of the post-translational modification according to the UniMod accession number [17]. In the following example the number 21 indicates a phosphorylation. Seo *et al.* [18] reported that glyceraldehyde-3-phosphate dehydrogenase can be phosphorylated at Thr-75; Ser-122; Ser-148; Thr-229; Thr-237 and Ser-312. For a protein species, which is phosphorylated at all of these positions, the *level F* descriptor is [P_75 – 122 – 148 – 229 – 237 – 312_21].

Cofactors (C) are described by the *level G* descriptor. The cofactors of the human metalloprotease ACE (P12821), for example, are two Zn^{2+}-ions, bound by the amino acids at the positions 390, 394 and 418 as well as 988, 992 and 1016. Therefore, the descriptors of human ACE are [C_ 390 – 394 – 418_Zn] + [C_988 – 992 – 1016_Zn].

If additional descriptors are needed, further levels can be introduced. Every new descriptor should start with a short unambiguous letter describing the aspect of the chemical composition or another important property of a protein in relationship to its chemical composition.

The prefinal descriptor (*level Y*) lists the versions of the data bases (DB) which were used for the identification or description of the exact chemical composition of the protein species. For UniProt retrieved in January 2010 the descriptor is [DB_UniProt_15.12].

The final descriptor (*level Z*) describes the function (F) of an enzyme with the appropriate EC number. The *level Z* descriptor for *e. g.* Glyceraldehyde-3-phosphate dehydrogenase is [F_EC = 1.2.1.12].

Table 1. Listing of the descriptors and their terms of the protein species identifier system (PSIS).

	1st Term: Defined aspect of the chemical composition of the protein species	2nd Term: Name or description Recommended data base	3rd Term Further description
Descriptor-Level	*Symbol*	*Example*	*Example*
A	Gene	Gene Name UniProt	Species
	G	*ACE*	*human*
B	Nucleotide polymorphisms	Accession number dbSNP (NCBI)	–
	NP	*rs4331*	
C	Initial amino acid sequence of the protein synthesized at the ribosome	Accession number UniProt	–
	AC	*P12821*	
D-1	Splicing variant	Number of the first and the last amino acid within the sequence which is deleted by splicing	–
	SD	*1 – 640*	
D-2	Splicing variant	Number of amino acid, which precedes the sequence, which was inserted by splicing	Sequence of the inserted peptide
	SI	*43*	*LELFVMFL*
E	Truncated amino acids	Sequence described by the first & the last number of the amino acids within the removed sequence	–
	T	*1 – 29*	
F	Post-translational modifications	Amino acid(s), which are modified	Accession number of the post-translation modification UniMod
	P	*75 – 122 – 148 – 229 – 237 – 312*	*21*
G	Cofactors	Amino acid(s), which bind the cofactor	Symbol describing the co-factor
	C	*390 – 394 – 418*	*Zn*
Y	Data base	Name of the data base	Version number
	DB	*UniProt*	*15.12*
Z	Function	EC Number	–
	F	*EC = 1.2.1.12*	

Rules for the Protein Species Identifier System

The general rules for the protein species identifier system (PSIS) are:

1. Descriptors must rely on experimental data, *e.g.* mass spectrometric analysis, immunological methods, knock-out experiments, over-expression experiments.

1.1. At least one descriptor needs to be given, which is either the descriptor for the gene (*level A* descriptor) or the descriptor for the protein (*level C* descriptor) or the function (*level Z* descriptor). Please note that the protein species level is cannot be obtained until information about *level C* and/or level *D* and *E* are given.

1.2. Every descriptor is given in square bracket characters.

1.3. One and the same descriptor can be used several times. For example, if a protein species carries post-translational modifications each post-translational modification is described by its own descriptor. For example the post-translational modification descriptors for a protein, phosphorylated (UniMod accession number: 21) at amino acid number 165 and sulfonated (UniMod accession number: 40) at amino acid number 223, are [P_165_21] + [P_223_40].

1.4. Descriptors are separated by a plus sign.

1.5. Every descriptor is composed by one or more terms.

2.1. A term, for example, is a symbol (*e.g.* NP), an accession number, a number indicating the position of an amino acid or two numbers separated by a hyphen indicating a partial amino acid sequence.

2.2. The underscore character separates terms.

2.3. The first term is a symbol, which is an abbreviation of the individual descriptor explaining a defined aspect of the chemical composition of the protein species, *e.g.* the identity of the gene (symbol: G) coding the protein species.

2.4. The second term refers to *e.g.* the gene name (*level A* descriptor), an accession number (*level B* descriptor and *level C* descriptor) and a partial sequence (*level D-1* or *E* descriptor) described by the first and the last number of the amino acids within the sequence, which is a part of a complete sequence of a defined protein or the position (number within the sequence) of an amino acid (*level D-2*, *E* or *F* descriptors). If several amino acids within one protein species are modified by the same moiety (*level F* descriptor), or involved in the binding of a cofactor (*level G* descriptor) every number of the concerned amino acids is listed, each number separated by a hyphen.

2.5 The third term refers *e.g.* to the species (*level A* descriptor), to the amino acid sequence of the inserted peptide (*level D-2* descriptor), to the type of post-translational modification (*level F* descriptor) or to a symbol for the cofactor (*level G* descriptor).

2.6 If necessary further terms can be added.

Optional

If the experimental data were obtained by immunological methods (ia = immunological analysis) such as Western Blots the protein accession number (*level C*) must be given. In this case the epitope(s), which is (are) recognized by the antibody, should be described (give

the numbers of the amino acids within the epitope), provided the epitopes are known. The term describing an immunological analysis and the epitope of the antibody is for example ia_456 – 464.

References

[1] Fenn, J.B., Mann, M., Meng, C.K., Wong, S.F., Whitehouse, C.M. (1989) Electrospray ionization for mass spectrometry of large biomolecules. *Science* **246**:64 – 71. doi: http://dx.doi.org/10.1126/science.2675315.

[2] Karas, M., Hillenkamp, F. (1988) Laser desorption ionization of proteins with molecular masses exceeding 10,000 daltons. *Anal. Chem.* **60**:2299 – 2301. doi: http://dx.doi.org/10.1021/ac00171a028.

[3] Lander E.S., *et al.* (2001) Initial sequencing and analysis of the human genome. *Nature* **409**:860 – 921. doi: http://dx.doi.org/10.1038/35057062.

[4] Finishing the euchromatic sequence of the human genome. *Nature* (2004) **431**:931 – 945. doi: http://dx.doi.org/10.1038/nature03001.

[5] Klose, J., Nock, C., Herrmann, M., Stuhler, K., Marcus, K., Bluggel, M., Krause, E., Schalkwyk, L.C., Rastan, S., Brown, S.D., Bussow, K., Himmelbauer, H., Lehrach, H. (2002) Genetic analysis of the mouse brain proteome. *Nat. Genet.* **30**:385 – 393. doi: http://dx.doi.org/10.1038/ng861

[6] Jungblut, P., Thiede, B., Zimny-Arndt, U., Muller, E.C., Scheler, C., Wittmann-Liebold, B., Otto, A. (1996) Resolution power of two-dimensional electrophoresis and identification of proteins from gels. *Electrophoresis* **17**:839 – 847. doi: http://dx.doi.org/10.1002/elps.1150170505.

[7] Jungblut, P.R., Holzhütter, H.G., Apweiler, R., Schlüter, H. (2008) The Speciation of the Proteome. *Chem. Cent. J.* **2**:16. doi: http://dx.doi.org/10.1186/1752-153X-2-16

[8] Joint Commission on Biochemical Nomenclature IUPAC-IUBMB: Nomenclature of multiple forms of enzymes. In: *Biochemical Nomenclature and Related Documents* 2nd edition. Edited by: Liébecq C. Colchester: Portland Press 1992.

[9] Nielsen, M.L., Savitski, M.M., Zubarev, R.A. (2006) Extent of modifications in human proteome samples and their effect on dynamic range of analysis in shotgun proteomics. *Mol. Cell Proteomics* **5**:2384 – 2391. doi: http://dx.doi.org/10.1074/mcp.M600248-MCP200.

[10] Riou, J.P., Claus, T.H., Pilkis, S.J. (1978) Stimulation of glucagon of *in vivo* phosphorylation of rat hepatic pyruvate kinase. *J. Biol. Chem.* **253**:656 – 659.

[11] Hochstrasser, M. (2009) Origin and function of ubiquitin-like proteins. *Nature* **458**:422 – 429.
doi: http://dx.doi.org/10.1038/nature07958.

[12] Morishima, N. (2005) Control of cell fate by Hsp70: more than an evanescent meeting. *J. Biochem.* (Tokyo) **137**:449 – 453.
doi: http://dx.doi.org/10.1093/jb/mvi057.

[13] Hara, M.R., Cascio, M.B., Sawa, A. (2006) GAPDH as a sensor of NO stress. *Biochim. Biophys. Acta* **1762**:502 – 509.

[14] Lane, D.A., Philippou, H., Huntington, J.A. (2005) Directing thrombin. *Blood* **106**:2605 – 2612.
doi: http://dx.doi.org/10.1182/blood-2005-04-1710.

[15] Woodman, Z.L., Schwager, S.L., Redelinghuys, P., Chubb, A.J., van der Merwe, E.L., Ehlers, M.R., Sturrock, E.D. (2006) Homologous substitution of ACE C-domain regions with N-domain sequences: effect on processing, shedding, and catalytic properties. *Biol. Chem.* **387**:1043 – 1051.
doi: http://dx.doi.org/10.1515/BC.2006.129.

[16] Schlüter, H., Apweiler, R., Holzhutter, H.G., Jungblut, P.R. (2009) Finding one's way in proteomics: a protein species nomenclature. *Chem. Cent. J.* **3**:11.
doi: http://dx.doi.org/10.1186/1752-153X-3-11.

[17] http://www.unimod.org/modifications_list.php?:

[18] Seo, J., Jeong, J., Kim, Y.M., Hwang, N., Paek, E., Lee, K.J. (2008) Strategy for comprehensive identification of post-translational modifications in cellular proteins, including low abundant modifications: application to glyceraldehyde-3-phosphate dehydrogenase. *J. Proteome Res.* **7**:587 – 602.
doi: http://dx.doi.org/10.1021/pr700657y.