

Minireview

Normalization of nomenclature for peptide motifs as ligands of modular protein domains

Rein Aasland^{a,*}, Charles Abrams^b, Christophe Ampe^c, Linda J. Ball^{d,*}, Mark T. Bedford^e, Gianni Cesareni^f, Mario Gimona^g, James H. Hurley^h, Thomas Jarchauⁱ, Veli-Pekka Lehto^j, Mark A. Lemmon^k, Rune Linding^l, Bruce J. Mayer^m, Makoto Nagaiⁿ, Marius Sudol^{n,*}, Ulrich Walterⁱ, Steve J. Winder^o

^aDepartment of Molecular Biology, University of Bergen, 5020 Bergen, Norway

^bDivision of Hematology and Oncology, University of Pennsylvania, Philadelphia, PA 19014, USA

^cDepartment of Biochemistry, Ghent University, B-9000 Gent, Belgium

^dDepartment of NMR and Structural Biology, F.M.P. – Research Institute of Molecular Pharmacology, D-13125 Berlin, Germany

^eDepartment of Carcinogenesis, University of Texas, Smithville, TX 78957, USA

^fDepartment of Biology, University of Rome, Tor Vergata, 00133 Rome, Italy

^gDepartment of Cell Biology, Austrian Academy of Sciences, 5020 Salzburg, Austria

^hLaboratory of Molecular Biology, National Institutes of Health, NIDDK, Bethesda, MD 20892, USA

ⁱInstitut für Klinische Biochemie und Pathobiochemie, D-97078 Würzburg, Germany

^jDepartment of Pathology, University of Oulu, 90410 Oulu, Finland

^kDepartment of Biochemistry and Biophysics, University of Pennsylvania School of Medicine, Philadelphia, PA 19104, USA

^lBiocomputing Unit, EMBL, D-69117 Heidelberg, Germany

^mDepartment of Genetics and Developmental Biology, University of Connecticut, Farmington, CT 06030, USA

ⁿDepartment of Medicine, Mt. Sinai Medical Center, New York, NY 10029, USA

^oDivision of Biochemistry and Molecular Biology, University of Glasgow, Glasgow G12 8QQ, UK

Received 8 November 2001; revised 20 November 2001; accepted 3 December 2001

First published online 21 December 2001

Edited by Gianni Cesareni and Mario Gimona

Abstract We propose a normalization of symbols and terms used to describe, accurately and succinctly, the detailed interactions between amino acid residues of pairs of interacting proteins at protein:protein (or protein:peptide) interfaces. Our aim is to unify several diverse descriptions currently in use in order to facilitate communication in the rapidly progressing field of signaling by protein domains. In order for the nomenclature to be convenient and widely used, we also suggest a parallel set of symbols restricted to the ASCII format allowing accurate parsing of the nomenclature to a computer-readable form. This proposal will be reviewed in the future and will therefore be open for the inclusion of new rules, modifications and changes. © 2002 Federation of European Biochemical Societies. Published by Elsevier Science B.V. All rights reserved.

Key words: Protein–protein interaction; Peptide motif; Protein domain, specificity and recognition; Proteomics; Glossary; Nomenclature; ASCII format

1. Introduction

At the recent conference on ‘Functional Protein Modules’¹ we discussed the urgent need of a unified glossary for describing protein–protein interactions of small, functional domains

with their peptide ligands. We propose a formalization of terms to enable the detailed description of residues involved in the protein:ligand interaction interface. This includes separate notations for (i) the consensus sequences of the cognate peptide ligands recognized by protein modules and (ii) the ‘epitopes’ (sometimes discontinuous) of the modules themselves, which interact directly with the ligand. This effort is aimed at unifying several diverse descriptions currently in use in the scientific literature, in order to facilitate communication in the rapidly developing fields of signal transduction and proteomics.

2. Definitions: modular protein domains and their peptide ligands

Modular protein domains (also known as protein modules or protein–protein interaction domains) are well demarcated and independently folded portions of proteins typically comprising 40–200 amino acids [1–3]. These domains are non-catalytic and bind specifically to short continuous peptide sequences in their binding partner(s) via one or more exposed, ligand-binding surfaces (‘ligand recognition pockets’). They are described as modular because of their frequent occurrence in various proteins and/or the fact that they often occur in multiple repeats within a single protein. They are autonomous in the sense that in most cases they may be removed from the original parent protein without compromising their ability to bind their cognate or target peptide ligands. The domain–ligand interactions are reversible and generally of low affinity (K_d values typically between 1 and 500 μ M). Complementary ‘epitopes’ present in both the domain and its target ligand act

*Corresponding authors.

E-mail addresses: aasland@mbi.uib.no (R. Aasland), linda@fmp-berlin.de (L.J. Ball), marius.sudol@mssm.edu (M. Sudol).

¹ EURESCO conference held in Seefeld, Austria, 6–11 October 2001.

together to encode specificity and affinity in a co-operative manner.

In general, a number of highly conserved residues in the domain make direct contacts to the amino acids of the ligand. In cases where the presence of a ligand induces folding of part of a domain, the directly interacting ‘contact’ residues are also important for maintaining the structure of the complex.

Ligands interact with their complementary domains through short and generally continuous sequence motifs (the core motifs), composed of three to six amino acids (Fig. 1). In some cases certain amino acids of the ligands must be post-translationally modified before recognition and binding can occur (e.g. phosphorylation or acetylation for ligands of SH2 and bromo domains, respectively). In other cases post-translational modification of target sequences downregulates binding by preventing interactions, usually sterically or via charge repulsion (e.g. phosphorylation or methylation of the ligands of certain WW and SH3 domains).

3. Rules for peptide ligands and protein domains

The following rules propose a general syntax for the representation of consensus sequences and conserved residues when describing complementary binding ‘epitopes’ involved in protein:peptide interactions. Single-character strings would thus be generated using specific symbols for different subsets of amino acids, defined on the basis of their chemical properties, including, where relevant, status of post-translational modification.

1. Continuous peptide ligands are represented by a linear string of amino acids in single-character notation as described by The International Union of Pure and Applied Chemistry (IUPAC) and International Union of Biochemistry and Molecular Biology (IUBMB) [4]. All letters symbolizing amino acids are capital. Lower-case letter ‘x’ is proposed to denote ‘unknown, other or any amino acid’.
2. Unless stated otherwise, for specific reasons, the default ‘0’ (zero) position should be assigned to the most ‘important’ amino acid within the ligand core and the remaining amino acids are designated ‘–1’, ‘–2’, and ‘1’, ‘2’, toward the (N-) amino- and (C-) carboxy-terminal directions, respectively (Fig. 1). Where there are multiple ‘important’ or *critical* residues (e.g. PxxP core of SH3 ligands), the most N-ter-

minal of them should be designated ‘0’. In cases where structures of complexes are not available and mutational analysis of ligands is missing, an initial assignment would be arbitrary.

3. With two or three known exceptions where the free carboxy-terminal end is an integral part of the ligand (e.g. ligands of PDZ domains), the core motifs are surrounded by flanking amino acids at their amino- and carboxy-termini, whose sequences are not given in detail but are represented by general symbols ‘*fn*’ or ‘*fc*’ respectively (Fig. 1). Absence of *fn* or *fc* in notation would indicate that the N- or C-terminal residue of the core motif is the respective terminus. Flanking sequences in concert with core motifs dictate the specificity of interaction within a given family of modules and are an important part of ligand characterization. Whenever possible they should be specified as direct and/or consensus sequences before and after *fn* and *fc* symbols.
4. Hydrophobic amino acids are represented by ‘Φ’ (Phi) because of phonetic relation between ‘phi’ and ‘phobic’. The following amino acids are considered hydrophobic: V, I, L, F, W, Y and M.
5. Aromatic amino acids are represented by ‘Ω’ (Omega) and they include F, W, Y. The letter Ω is derived from the Greek word *Ωπατο*, pronounced *Oreo*, meaning pleasant, and frequently used to describe aroma. The visual mnemonic here is the round shape of ‘Oreo’ cookie that is reminiscent of the closed aromatic ring.
6. Hydrophilic amino acids are represented by lower-case (not capital) ‘ζ’ (zeta). The letter is derived from the Greek word *ζωη*, pronounced *zoi*, which means life, a simile to water (*hydro*). N, Q, S and T are hydrophilic uncharged amino acids and hydrophilic charged amino acids are: E, D, K, R and H.
7. Amino acids with large aliphatic side chains have been traditionally marked with, and should continue to be designated, ‘Ψ’ (Psi); again because of the phonetic similarity between ‘aliphatic side’ and ‘psi’. Side chains of V, I, L and M belong to this group. (Annotations are recommended whenever symbols for groups of amino acids described in rules 4–7 are used.)
8. Charged residues will be marked with [+] (positive) and [–] (negative), the square brackets to be included in the notation. Direct specification of amino acids is recommended where appropriate. For example, for a position occupied by negatively charged amino acids the alternative choice would be (E/D).
9. Amino acids with small chains, namely P, G, A and S, will be named with letter π (pi) after representative *proline*.
10. Phosphorylated amino acids are indicated by lower-case ‘*po*’ in italics without a dash connecting it to the modified amino acid, e.g. *poY* is phospho-tyrosine (Fig. 1). Italics are suggested for symbols of post-translational modification because in certain conventions lower-case symbols for certain amino acids are used for highly but not completely conserved positions. For symbols of other post-synthetic modifications, not indicated in this glossary, refer to examples listed in the supplement to reference [5].
11. Sulfated amino acid, namely tyrosine, is indicated by ‘*su*’ – *suY*.
12. *O*-Glycosylated serine and threonine on hydroxyl groups by a single β-*N*-acetylglucosamine have been identified

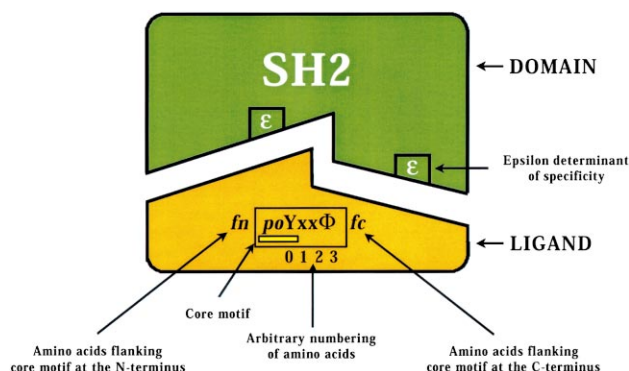


Fig. 1. Formalization of terms to describe protein–protein interactions mediated by modules and their cognate peptide ligands using the SH2 domain and its ligand as example.

and characterized in cytoplasmic and nuclear proteins. Past abbreviations were S-(O-GlcNAc) or T-(O-GlcNAc). We suggest to use *g/S* and *g/T* symbols. N-Glycosylated amino acids should also be indicated with the *gl* prefix (e.g. *glN* may indicate glycosylated asparagine).


13. Methylated amino acids will be indicated by '*me*' in italics whether it is mono- or dimethyl- form (e.g. mono- or dimethyl-arginine will be: *meR*). Since symmetrical and asymmetrical methylations have been shown to modulate protein–protein interaction differentially, we suggest '*sme*' for symmetrical and '*ame*' for asymmetrical methylation (e.g. symmetrically dimethylated arginine will be indicated by *smeR*, whereas the asymmetrically dimethylated variant will be marked by *ameR*).
14. Acetylated amino acids will be indicated with '*ac*' in italics (e.g. acetyl-lysine will be: *acK*).
15. Hydroxylated proline will be indicated by *hyP*.
16. Amino acids to be excluded from a given position in a consensus motif should be marked by '^' (e.g. ^P – no proline at this position allowed, or ^*(R/K/W)* – no arginine or lysine or tryptophan allowed).
17. For figures, tables and graphical representations of domain–ligand complexes, the residues of the ligand, which pack closely into the domain surface, are underlined with yellow boxes or marked with yellow background (Fig. 1).
18. Variations in primary and ultimately in tertiary structures of domains contribute to the specificity of interaction with the ligand. A term, 'epsilon determinant(s)' (ϵ , from the Greek word for specificity: $\epsilon\delta\iota\kappa\omicron\tau\eta\tau\alpha$, pronounced *ede-kohteta*), was proposed for those amino acids within domains which determine ligand predilections [6]. The ϵ determinant is represented by one or several amino acid positions located mainly within the conserved structure of the domain, and usually in the ligand-binding interface. Notation for ϵ determinants was previously described: (*Substructure Letter Number*) [6–8]. To illustrate it by example: The ϵ determinant for one of the classes of SH2 domains was represented by the β D5 position (the notation indicated fifth amino acid in the fourth β -strand; D indicated the fourth) and occupied by aromatic residues, tyrosine or phenylalanine. To avoid any confusion with one-letter symbols of amino acids, we propose to change this notation by replacing non-Greek letters referring to the order of secondary structural element with numbers. Previously the ϵ determinant for class I of SH2 domains was β D5=(Y/F). We propose to use: β 4-5=(Y/F). For another class of SH2 domains, the ϵ determinant resides in the EF1 position (the first amino acid position in the loop between E [fifth] and F [sixth] β -strands). We propose 5:6-1 notation; a loop would be indicated by ':'. For more details see [6–8].

4. Alternative nomenclature in ASCII format

Considering the rapid progress in fields of molecular signaling and proteomics, we have also identified a need for a parallel set of symbols in ASCII format, i.e. a complete set of symbols found on any standard computer keyboard. In Table 1 we summarize all terms in non-ASCII style – the terms described above in 18 rules, and their equivalents in ASCII format.

Table 1

Summary of symbols for representing consensus sequences of peptide ligands and protein modules

BRIEF SUMMARY OF RULES	SYMBOL	SYMBOL IN ASCII
Unknown, other or any other amino acid	x	x
Sequences flanking ligand core at the N-terminus	<i>fn</i>	fn
Sequences flanking ligand core at the C-terminus	<i>fc</i>	fc
Hydrophobic amino acid	Φ	%
Aromatic amino acid	Ω	@
Hydrophilic amino acid	ξ	&
Positively charged residue	[+]	[+]
Negatively charged residue	[-]	[-]
Aliphatic side chains	Ψ	#
Small chain amino acids	π	~
Phosphorylated amino acid (Tyr)	<i>poY</i>	[Y:po]
Sulfated amino acid (Tyr)	<i>suY</i>	[Y:su]
O-glycosylated amino acid (Ser)	<i>glS</i>	[S:gl]
N-glycosylated amino acid (Asp)	<i>glN</i>	[N:gl]
Methylated amino acid (Arg)	<i>meR</i>	[R:me]
Symmetrically methylated Arg	<i>smeR</i>	[R:sme]
Asymmetrically methylated Arg	<i>ameR</i>	[R:ame]
Acetylated amino acid (Lys)	<i>acK</i>	[K:ac]
Hydroxylated Proline	<i>hyP</i>	[P:hy]
Excluded amino acid (Pro)	^P	(^P)
Amino acid that packs against the domain (e.g., <i>poY</i> with SH2)		{Y:po}
ϵ -determinants	β 4-5 α 2-2	(beta4)5 (alpha2)2

5. Three examples of nomenclature in both formats

1. Tyrosine phosphorylated motif: *fnAGpoY(G/A/P/S)HFfc* or *fnAGpoYπHFfc*; and in ASCII format: *fnAG[Y:po](G/A/P/S)HFfc* or *fnAG[Y:po]~HFfc*. Note that when several amino acids are allowed at a certain position, they are placed in parentheses and separated by slashes.
2. A peptide core terminating with methylated (or carboxy-methylated) leucine: *fnPALPPAmeL*; and in ASCII format: *fnPALPPA[L:me]*.
3. One of the ϵ determinants for class II WW domains is β B6 position occupied by aromatic amino acids: β 2-6= Ω ; in ASCII format: (beta2)6=@.

6. Concluding remarks

Our proposal is open for the inclusion of new rules, modifications and changes, and we hope to update this nomenclature again at the next international gathering of researchers in the field of protein modules. Several principles guided us in compiling this glossary: (i) previous and/or prevalent symbols should be primary choices and if needed only minimally modified [1,5–8]; (ii) the phrases to describe consensus sequences should be as simple as possible while being unequivocal; (iii) symbols for post-translational modifications and other annotations should be set in lower-case characters and placed in front of the modified amino acid residue. For ASCII format we place modifications after the modified amino acid, with ':' in between, and close the composite symbol in brackets, following recommendations of Bader et al. [5].

Acknowledgements: Supported by grants from Human Frontier Science Program Organization and National Institutes of Health. We would like to thank Konstadinos Moissoglou for advice on Greek terms used in the nomenclature.

References

- [1] Pawson, T. (1995) *Nature* 373, 573–580.
- [2] Bork, P., Downing, A.K., Kieffer, B. and Campbell, I. (1996) *Q. Rev. Biophys.* 29, 119–167.
- [3] Das, S. and Smith, T.F. (2000) *Adv. Protein Chem.* 54, 159–183.
- [4] IUPAC web address: <http://www.chem.qmw.ac.uk/iupac>
- [5] Bader, G.D. (2001) *Nucleic Acids Res.* 29, 242–245.
- [6] Sudol, M. (1998) *Oncogene* 17, 1469–1474.
- [7] Kuriyan, J. and Cowburn, D. (1997) *Annu. Rev. Biophys. Biomol. Struct.* 26, 259–288.
- [8] Songyang, Z. and Cantley, L. (1995) *Trends Biochem. Sci.* 20, 470–475.