FEBS Letters 579 (2005) 3342–3345 FEBS 29534

Minireview

Linear motifs: Evolutionary interaction switches

Victor Neduva, Robert B. Russell*

EMBL, Meyerhofstrasse 1, 69117 Heidelberg, Germany

Accepted 7 April 2005

Available online 18 April 2005

Edited by Péter Friedrich

Abstract Linear motifs are short sequence patterns associated with a particular function. They differ fundamentally from longer, globular protein domains in terms of their binding affinities, evolution and in how they are found experimentally or computationally. In this Minireview, we discuss various aspects of these critically important functional regions.

Keywords: Protein evolution; Protein–ligand interaction; Binding motifs; Motif conservation

1. Introduction

Proteins are probably the most important component of the cell's functional repertoire and are involved in virtually all critical processes. Over time, proteins diverge, and are often duplicated extensively. This means that nearly all proteins can be grouped into families where all members share a similar function. Proteins are usually *modular*, containing discrete regions each of which performs a different sub-function. The most widely known modular element is the protein domain. These are typically more than 30 residues in length and fold into an independent compact structure. More than 7000 domains are known [1], performing an enormous diversity of functions from catalysis during metabolism to cell–cell recognition in the immune system. Domain duplication is now an accepted mechanism of evolution, and differences in domain architecture are often responsible for critical differences between organisms. Duplications are thought to be followed either by loss of one copy or the evolution of a new function by point mutations [2].

However, domains are only part of the picture. Many studies have shown that they cover only a fraction of the protein sequence contained in an organism. The remaining parts of the sequence only rarely contain undiscovered domains, and indeed have been shown to be low-complexity (i.e., dominated by a few amino acids) or *intrinsically disordered* (see [3] for review). A fraction of these regions are likely linkers that permit the correct spacing of domains in a functional protein, though many others are known to play pivotal functional roles. Critical sites for phosphorylation, or other modifications often lie within them, as do regions important for interactions with other proteins. These very short, functional regions, though not globular domains, often conform to particular sequence patterns or *linear motifs* indicative of a particular function [4]. Known examples include phosphorylation sites (e.g. [5]), localization signals like KDEL (e.g. [6]), and binding regions such as the canonical SH3 ligand PxxP (e.g. [7]).

2. Linear motifs are hard to find

Domains were first defined in the 1960s with the arrival of the first protein crystal structures (see [8] for a review). The original definition was largely structural: domains were thought to be spatially distinct, probably independently folding entities. The advent of modern molecular biology gave rise to many thousands of DNA and protein sequences. Sequence alignments showed that many long proteins shared shorter regions of homology with others, and this gave rise to a definition of domains based more on sequence recurrence, usually also associated with some common function (e.g., catalysis or binding). Domains are now readily detectable with sequence searching programs (e.g., Blast [9] or HMMer [10]) and readily alignable by standard methods (e.g., ClustalW [11] or MUSCLE [12]). Known domains are now stored in a number of databases including Pfam [1], SMART [13], CDD [14] and InterPro [15] and remain a critical component of genome annotation procedures. Particular domains and domain architectures are well conserved over the course of evolution (e.g., Fig. 1). The sequences diverge, but the overall domain architecture remains the same.

Although the notion of linear motifs has been around since the mid-1970s (see [16–18] for review), the first clear example of a motif paired to its receptor molecule was not described until 1990, when the targeting signal KDEL was paired to the ERD2 receptor [19]. Moreover, despite the availability of many thousands of sequences, the discovery of linear motifs, in contrast to domains, has remained difficult. Their short length makes them difficult to detect using sequence comparison procedures that aid domain discovery. They are typically discovered by difficult and time-consuming experimental procedures. This usually involves first identifying a set of proteins sharing a common function (e.g., a common interaction partner or targeting within the cell), and then gradually delineating a short, common segment associated with this function through a variety of experimental techniques. For instance, the SH3 ligand was first identified as a recurring sequence feature in signaling proteins [20]. As interacting partners of SH3 containing proteins were gradually identified, the interacting

*Corresponding author. Fax: +49 6221 387 517.
E-mail address: russell@embl.de (R.B. Russell).

0014-5793/$30.00
doi:10.1016/j.febslet.2005.04.005

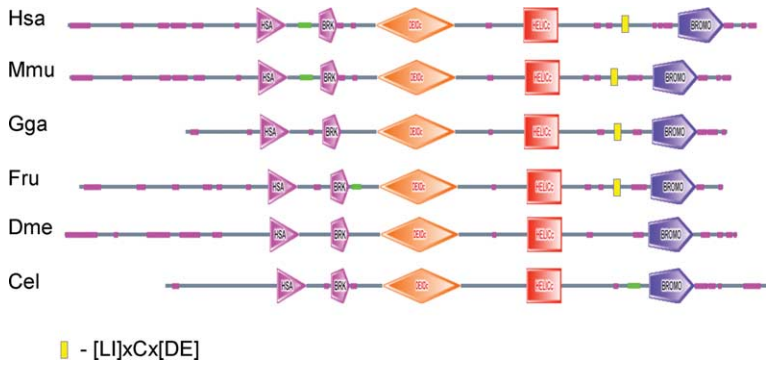

Fig. 1. Domains in contrast to linear motifs. Domain architectures for SNF2-alpha like proteins as determine by SMART [13]. Domains are shown as colored shapes, coiled-coil containing regions are shown in green, and low-complexity sequences are colored magenta. The location of a vertebrate specific instance of the retinoblastoma (RB) linear motif [LI]xCx[DE] is shown in yellow. Species names are abbreviated as follows: Hsa, *Homo sapiens* (swissprot: Q9HBD4); Mmu, *Mus musculis* (ensemble: ENSMUSP0000030821); Gga, *Gallus gallus* (ensemble: ENSGALP00000016509); Fru, *Fugu ripes* (ensemble: SINFRUP00000139269); Dme, *Drosophila melanogaster* (swissprot: P25439); Cel, *Caenorhabditis elegans* (swissprot: Q19106).

region was eventually reduced to a 9–10 residue, proline-rich segment [21], which suggested a mode of action ultimately confirmed by 3D structures (e.g. [21,22]). Today it is known that SH3 domains bind a short sequence PxxP and there are several variations that confer specificity for particular SH3 containing proteins (e.g. [23,24]). New motif discovery remains a difficult but rewarding task: for example, a recent combination of bioinformatics and functional studies in *Plasmodium falciparum* revealed the secretion motif RxLxE, which provided fascinating insights into the physiology of this malaria parasite [25,26]. Despite such findings, the difficulties in their discovery mean that only about two hundred linear motifs are known compared to thousands of domains that might bind them. Known motifs are now being catalogued by several resources (elm.eu.org [4]; scansite.mit.edu [27]).

3. Differing affinities

A key difference between linear motifs and domains is their affinity for their binding partners. Domains, when they bind to each other, tend to do so with relatively strong affinities: low-nanomolar or even picomolar affinities are known (e.g. [28]). The short length of linear motifs means that they rarely have such strong affinities: typically between 0.5 and 1 μM, with the low-affinity in the 10 μM range [7,29]. This has certain implications for their function. They tend generally to be the mediators of transient interactions, thus, making them popular in signaling networks [7]. Even the most casual glance at signaling pathways shows linear motifs to be of paramount importance. The pathways almost always contain kinase phosphorylation sites, SH2 or PTB domain specific phosphotyrosines, SH3 proline-rich ligands, or 14-3-3 domain interacting motifs.

4. Differing evolution

Linear motifs are short: between three and ten amino acids, of which usually just two or three are important for function. This makes them fundamentally different from domains in terms of how they arise or how long they tend to be conserved over evolutionary lineages. Domains face tough requirements of being able to fold into a stable, globular structure and enact a specific function. When duplicated they face the additional challenge of evolving a new function before being lost [2]. Linear motifs, in contrast, are very likely to arise or disappear by chance: just one mutation can change an inert stretch of sequence into a functional linear motif, or cause a functional site to become inactive. This gives them a certain evolutionary plasticity missing from protein domains.

A good example of the contrast between domains and linear motifs can be seen in SNF2-alpha homologues from human to worm. These proteins are very similar in terms of domain architecture throughout all lineages, though linear motif in this family shows different behavior. The retinoblastoma (RB) binding motif shown in Fig. 1 is seen in the vertebrates but disappears in fly and worm. There are many examples confirming the plasticity of linear motifs. For instance, a single Trp to Cys mutation eliminates a C-mannosylation site in rodents Interleukin-12 homologues proteins, which found in other metazoans (Fig. 2A). This is also apparent within multiple copies of a similar protein in a single species. For example in two closely related paralogues of mouse Actin-like protein 6B (ACTL6B), only one contains the motif for binding C-terminal Binding Protein (CtBP), while another has apparently lost it (Fig. 2B). For critical interactions, this plasticity is risky, since a single mutation can remove a critical interaction. The Clathrin coat assembly protein AP180, which is important for vesicles formation during the edocytosis, contains two proximal AP-2 binding sites about fifty amino acids apart (Fig. 2C). One of the motifs is lost in rat while the other is intact. Thus, the reoccurrence of the motif could attenuate the sensitivity to deleterious mutations.

The overall consequence of the above is a poor general conservation of instances of linear motifs. Diverse species use the same *kinds* of linear motifs, for example, SH3 domains bind proline-rich sequences in all known Eukaryotes, but the particular instance of the motif is rarely conserved over long evolutionary distances. We looked at all known instances of linear motifs in humans, and charted their conservation in orthologues from other eukaryotes with complete genomes (Fig. 3A), and found that the motifs were often not conserved outside of the vertebrates (~65%). There are some exceptions to this, for example some instances of the KDEL targeting motif are conserved in virtually all eukaryotes (from Human to Plants). In contrast, domains within orthologous proteins stay largely conserved (Fig. 3B).