Biochimie 95 (2013) 1511–1524

Contents lists available at SciVerse ScienceDirect

Biochimie

journal homepage: www.elsevier.com/locate/biochi

Review

Structures composing protein domains

Jaroslav Kubrycht [a,*], Karel Sigler [b], Pavel Souček [c], Jiří Hudeček [d]

[a] Department of Physiology, Second Faculty of Medicine, Charles University, Plzenska 221, 150 00 Prague, Czech Republic
[b] Laboratory of Cellular Biology, Institute of Microbiology, Academy of Sciences of the Czech Republic, Videnska 1083, 142 20 Prague, Czech Republic
[c] Toxicogenomics Unit, National Institute of Public Health, Srobarova 48, 100 42 Prague, Czech Republic
[d] Department of Biochemistry, Charles University, Hlavova 2030, 128 40 Prague, Czech Republic

ARTICLE INFO

Article history:
Received 22 January 2013
Accepted 2 April 2013
Available online 10 April 2013

Keywords:
Catalytic
Disordered
Fold
Intradomain
Motif
Repeat

ABSTRACT

This review summarizes available data concerning intradomain structures (**IS**) such as functionally important amino acid residues, short linear motifs, conserved or disordered regions, peptide repeats, broadly occurring secondary structures or folds, etc. IS form structural features (units or elements) necessary for interactions with proteins or non-peptidic ligands, enzyme reactions and some structural properties of proteins. These features have often been related to a single structural level (e.g. primary structure) mostly requiring certain structural context of other levels (e.g. secondary structures or supersecondary folds) as follows also from some examples reported or demonstrated here. In addition, we deal with some functionally important dynamic properties of IS (e.g. flexibility and different forms of accessibility), and more special dynamic changes of IS during enzyme reactions and allosteric regulation. Selected notes concern also some experimental methods, still more necessary tools of bioinformatic processing and clinically interesting relationships.

1. Several notes to structural diversity, assembly and hierarchy

Domains are protein chain segments representing structural and functional units (building blocks), which have kept certain overall structural conservativeness during protein evolution. Binding interactions or enzyme reactions of domains and smaller structures are involved in multiple processes endowing proteins with new functions (cf. [1]). In fact, two main types of structures smaller than domains exist: (i) **interdomain connections** and (ii) **intradomain structures** (**IS**). The interdomain connections (denoted also **domain extensions**) constitute both conserved and structurally unstable disordered regions [2,3] linking or terminating domains. **IS** (**or otherwise structures composing domains**) dealt with in this review are more extensively diversified with respect to their extent, composition and stereo-chemical properties than the preceding structures. It is a question whether the assumed classification is definitively complete. Hence, for instance, we do not know whether some of mostly intradomain structures such as protein pockets (cf. Table 1) and still poorly described three-dimensional (**3D**) motifs (for related consideration see Ref. [4]; cf. also bi-domain catalytic sites [5,6]) can be formed by local spatial structural arrangements of segments present in two or more domains. Such structures appear to be interesting, because they would represent new types of spatially united oligochain and oligodomain substructures undergoing a unifying phylogenic pressure in contrast to currently considered evolution of single chain related structural units, e.g. domains and interdomain connections mentioned above.

Multiple important IS are shown in Table 1 (Refs. [7–86]). In accordance with Table 1, each functional site is formed in fact by a superposition (coincidence) of various structural levels mentioned in the upper part of the table. This means that physically interacting or chemically reacting specific (or sometimes alternative) amino acid residues (**aa**) as well as alternative simultaneously acting aa of short linear motifs (**SLM**) or catalytic motifs represent only the first "contact"-related structural level of IS important for triggering of functional events (cf. Chapters 3 and 4). Secondary structures then constitute spatial carriers of the preceding "contact" primary structures (cf. e.g. Ref. [87]). Moreover, cooperating "contact" structures can be differently present in single or several secondary structures of the same fold, or other folds located in the same or

Abbreviations: 3D, three-dimensional; aa, amino acid residue(s) in peptide; CDR, hypervariable complementarity determining region(s) of IgV; CDR*n*, CDR of *n*-th chain order (*n* is 1–3); FR, framework regions of IgV; FR*n*, FR of *n*-th chain order (*n* is 1–4); IS, intradomain structure(s); Ig, immunoglobulin(s); IgV, variable domain(s) of Ig; igvds, conserved variable Ig domain sequence(s) (derived by conserved domain search of BLAST); MSA, multiple sequence alignment(s); MSA-R, MSA derived record(s); PKSI, peptidic protein kinase substrates and inhibitors; SLM, short linear motif(s); TCR, T-cell receptor(s); TR, protein tandem repeat(s).

* Corresponding author. Tel.: +420 233 32 32 72.
E-mail address: jkub@post.cz (J. Kubrycht).

0300-9084/$ – see front matter © 2013 Elsevier Masson SAS. All rights reserved.
http://dx.doi.org/10.1016/j.biochi.2013.04.001

Table 1
Broadly occurring and some more special intradomain structures (IS).

Structures	Short description	Experimental and database indications, predictions
Sequence related IS		
Short linear motif(s) (**SLM**)	*Short sequence structures (frequently 3–10 aa) of biological importance (proved in published papers and/or database annotated experiments) often achieving higher occurrence than expected →**SLM** are first of all investigated in case of interaction studies (cf. ch3). In addition, **SLM** can also determine other properties of proteins (e.g. flexibility or elasticity). →Only in case of sufficient spatial exposure **SLM** can interact. Such exposure follows from **SLM** superposition with proper conformation-related structures or new accessibility after proteolysis.	Discrimination of annotated **biologically relevant SLM** from stochastically occurring non-functional instances needs database agreement (e.g. Refs. [7,8]) or experiments confirming proposed biological function (SDM, binding assays, enzyme kinetics); interaction interfaces of **SLM** identified in range of human proteome [9]
Catalytic motifs	*Motifs composed of 2–10 aa including at least single catalytic aa (see ch4)	SDM; molecular docking; MD; see also ch5 and Ref. [10]
Sequence pattern(s) (**SP**)	***Short sequence structures** derived using searches in accessible or private sequence databases and occurring substantially more frequently than expected →**SP** are candidates for **SLM**, short **TR** and catalytic motifs or directly form these structures.	MSA columns with monotonous or almost monotonous occurrence of unique aa indicate **SP**, whose frequencies and specificity can be further reevaluated using e.g. ScanProsite or PHI BLAST.
Peptide signatures	*Sequence patterns generalized with respect to different extent of undetermined aa (recorded in usual **SP** as X) inserted between specific pattern aa (**UIaa**) →Increased distance variability follows from insertion-deletion changes on aa level, frequent in some protein families or familiar groups (see e.g. variable domains of immunoglobulins [11,12]).	Patterns regarding variation of gap numbers in MSA; ScanProsite with **SP** containing more diversified numbers of **UIaa** than original **SP** generated by MSA; CoPS – comprehensive peptide signature database [13]
Tandem repeat(s) of proteins (**TR**)	*Sequence structures repeating in single protein molecule and usually also in the majority of molecules of family or superfamily relationship →**TR** unit lengths range from a single aa (aa repetitions) to more than 100 residues (i.e. in the domain range) and the repeat number is sometimes over 100 [14]. **TR** carry fundamental functions frequently related to human diseases [15,16]. For protein repeats different from **TR** see ch2.	**TR** can be identified using recent Protein Repeat Database [17]. For other tools see ch5.
Conserved sequence regions (**CSR**)	*Sequence segments of highest domain/molecular similarity can be observed, when using a group of related representative sequences. →**CSR** regions are necessary for different functions or structural properties of proteins. →Structurally based analogs of **CSR** (i.e. structurally conserved regions) also exist [18,19].	Databases of **CSR** or conserved domain sequences [20–23]; PHI BLAST with (i) MSA-derived **SP** and (ii) the corresponding probabilistically restricted consensi [24]
IS structures derived based on conformation of peptide chains		
Secondary structure(s) of proteins (**SSP**)	*Minimum protein segments of uniform spatial arrangement and of at least five aa length →Only some types of secondary structures frequently participate in interactions (cf. ch3). →Knowledge based structural classes based on predicted **SSP** play an important role in understanding protein folding (see below; [25])	Web server related to eight experimental methods of **SSP** assignment [26]; PTGL – database for **SSP** [27]; **SSP** prediction using multistep learning (SPINE X [28]) or metapredictor (SymPsiPred [29])
Folds	*Folds are building blocks of domains. These blocks (of at least about twenty aa length) are composed of several SSP (as elements) forming together **supersecondary motif**. For instance supersecondary motifs ABAB, BABA, BABABB and BLBABA were described indicating A, B and L as alpha helix, beta sheet and loop, respectively [30–33]. → Folds are chain-related segments of certain structural autonomy. As it is well-known, folds can frequently keep phylogenic relationship even in cases, when superfamily relationship (indicated usually by a conserved domain similarity) is lost. Fold repertoire has not yet been enriched in the last two milliards years [34].	NMR techniques (e.g. Refs. [35–39]); X-ray methods SAXS and WAXS [40,41]; image reconstruction based on electron density evaluation [42,43]; predictions consisting in: SVM combining four descriptors including profile–profile alignment (DescFold [44]); template-based modeling [45]; knowledge-based approach (CONTSOR [46]); fold-specific PSSM libraries [47]; evaluation of multiple physicochemical aa properties [48]
(Intrinsically) disordered regions (**DR**)	*Segments without unique well-defined 3D structure →**DR** are mostly investigated in interactions, because their interactions can more rapidly evolve than those of conventional segments. This follows from the fact that **DR** are highly flexible and polyvalent in many cases due to existing alternative conformations (see ch3).	NMR and SAXS [49]; ANCHOR [50] estimates energy combining general disorder tendency with sensitivity to the structural environment; MetaDisorder – accurate meta-prediction method (based on 13 disorder predictors; [51])
Protein pockets (**PP**)	*__PP__ are formed by encapsulated protein surface separated from outside space and appear to be important for protein interactions including those with drugs [52,53]. →Inhibitors targeting the gp41 pocket has been developed for purposes of AIDS therapy [54].	SiMMap server statistically derives site-moiety map to recognize interaction preferences between protein pockets and compound (e.g. drug) moieties [55].