Journal of Structural Biology xxx (2014) xxx–xxx

Contents lists available at ScienceDirect

Journal of Structural Biology

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

A first census of collagen interruptions: Collagen's own stutters and stammers

Jordi Bella [*]

Manchester Institute of Biotechnology, Faculty of Life Sciences, University of Manchester, UK

ARTICLE INFO

Article history:
Received 23 December 2013
Received in revised form 27 March 2014
Accepted 27 March 2014
Available online xxxx

Keywords:
Collagen
Interruptions
Triple helix
Protein structure
Hydrogen bonding
Protein conformation

ABSTRACT

The repetitive Gly-X-Y sequence is the telltale sign of triple helical domains in collagens and collagen-like proteins. Most collagen sequences contain sporadic interruptions of this pattern, which may have functional roles in molecular flexibility, assembly or molecular recognition. However, the structural signatures of the different interruptions are not well defined. Here, a first comprehensive survey of collagen interruptions on collagen sequences from different taxonomic groups is presented. Amino acid preferences at the sites of interruption and the flanking triplets are analysed separately for metazoan and prokaryotic collagens and the concept of commensurateness between interruptions is introduced. Known structural information from model peptides is used to present a common framework for hydrogen bonding topology and variations in superhelical twist for the different types of interruptions. Several collagen interruptions are further classified here as stutters or stammers in analogy to the heptad breaks observed in alpha-helical coiled coils, and the structural consequences of commensurate interruptions in heterotrimeric collagens are briefly discussed. Data presented here will be useful for further investigation on the relation between structure and function of collagen interruptions.

1. Introduction

Collagens and collagen-like proteins are immediately recognised in whole genome analyses by the characteristic repetitive $(\text{Gly-X-Y})_n$ sequence of their collagen triple helical domains. Collagen-containing proteins differ in their domain organisation and number of collagen domains, and appear now to be found in all kingdoms of life, from viruses to humans (Rasmussen et al., 2003; Kadler et al., 2007; Ghosh et al., 2012). The strictness with which the Gly-X-Y repetitive pattern is maintained varies from protein to protein. In humans, vertebrate fibrillar collagen types I, II, III, V and XI contain stretches of more than 1000 amino acids without interruption of the Gly-X-Y pattern (Kadler et al., 2007). Such structural regularity appears to be necessary for the formation of fibrils staggered with a 67-nm axial repeat. Missense mutations in fibrillar collagens where a Gly residue is replaced by another amino acid with a bulkier side chain lead to connective tissue disorders of differing degrees of severity (Myllyharju and Kivirikko, 2004). Non-fibrillar collagens, on the other hand, often contain a number of breaks of the Gly-X-Y repeating sequence (Brodsky et al., 2008; Thiagarajan et al., 2008). It is thought that some of these breaks may have functional significance by providing local flexibility points (Hofmann et al., 1984) or sites for molecular recognition (Miles et al., 1995). Yet, additional interruptions resulting from mutations in the Gly-X-Y sequences of human non-fibrillar collagen genes have also been reported in a broad variety of disorders (Parkin et al., 2011; Kuo et al., 2012).

While most efforts have been devoted to characterize collagen interruptions that arise from missense mutations in collagen genes, often resulting in hereditary diseases, a clear understanding of the structural consequences of collagen interruptions is missing. A widely extended common misconception is that all interruptions result in permanently kinked molecules, but that is not the only structural option. Other local distortions have been proposed, such as some residues looping out of the helix, or altered triple-helical conformations resulting in a slightly "bulged" helix (Long et al., 1992). These distortions might be compatible with straight molecules or could also introduce flexible bending points. Peptides designed to model breaks in the characteristic collagen sequence have been used to study the impact of several types of interruptions in the triple helical structure of collagen, its folding, its thermal stability, and its supramolecular association (Long et al., 1993; Bella et al., 1994, 1996, 2006; Baum and Brodsky, 1999; Mohs et al., 2006; Brodsky et al., 2008; Thiagarajan et al., 2008; Hwang et al., 2010); these studies have recently extended

[*] Address: 131 Princess Street, Manchester M1 7DN, UK.
E-mail address: jordi.bella@manchester.ac.uk

http://dx.doi.org/10.1016/j.jsb.2014.03.022
1047-8477/© 2014 Published by Elsevier Inc.

to engineered recombinant collagens (Hwang and Brodsky, 2012). All the interruptions introduced in reference collagen sequences such as (Pro-Hyp-Gly)$_{10}$ bring some destabilization to the triple helix conformation and lower the thermal denaturation temperature. Yet, conceptually identical interruptions often have a widely different impact on stability depending on the actual amino acid sequence at the interruption (Mohs et al., 2006; Thiagarajan et al., 2008).

To date, high-resolution structural information has only been obtained for two collagen peptides containing interruptions of the consensus sequence (Bella et al., 1994, 2006). Additional structural insight on a few types of interruption has been obtained experimentally from NMR studies of peptides combined with molecular dynamics simulations (Mohs et al., 2006; Li et al., 2007, 2009; Thiagarajan et al., 2008). The sequence features of the most common types of interruption in human non-fibrillar collagens have been explored (Li et al., 2007; Thiagarajan et al., 2008). With the continuous increase of available collagen and collagen-like protein sequences from ongoing whole genome sequencing efforts, it is now becoming necessary to develop tools for systematic detection, classification and annotation of collagen interruptions on sequence databases. Here, a first comprehensive survey of collagen interruptions of a range of lengths on collagen-containing sequences from metazoa, bacteria, viruses and archaea is presented. Amino acid preferences at and around the sites of interruption are analysed separately for metazoan collagens and two groups of prokaryotic collagens. Additionally, the concept of commensurateness between collagen interruptions of different lengths is introduced. Finally, the structural consequences from interruptions observed in high resolution crystal structures are combined with a description of the collagen triple helix based on residues at the same level (Bella, 2010), to make inferences about hydrogen bonding topology and superhelical parameters at the sites of collagen interruption. It emerges from this analysis that some collagen interruptions can be referred as “stutters” or “stammers” in analogy to the well-established interruptions in α-helical coiled coils (Brown et al., 1996). The possible use of these inferences to make predictions on the local chain register of heterotrimeric collagens at the sites of interruption is discussed.

2. Methods

2.1. Defining sequence patterns associated to the different collagen interruptions

Regular expression motifs for each type of interruption were defined using appropriate pattern syntax rules for the sequence motif search tool ScanProsite (de Castro et al., 2006). As interruptions only make sense in a collagen context, motifs were defined by embedding the appropriate Gly-Z_n-Gly sequence between a number of flanking Gly-X-Y triplets conforming to a collagen repeating pattern. For example, a regular expression motif for a G1G interruption could be defined as (Gly-X-Y)$_2$-Gly-Z-Gly-(X-Y-Gly)$_2$, which translated into the ScanProsite syntax as G-{g}-{g}-G-{g}-{g}-G-{g}-G-{g}-{g}-G-{g}-{g}-G, where {g} indicates any amino acid other than Gly. Exploratory searches were conducted with motifs containing one, two or three Gly-X-Y triplets at each side of the Gly-Z_n-Gly sequence. Motifs with one triplet gave too many false positive matches to non-collagen domains; patterns with three triplets were too strict and missed neighbouring interruptions separated by only two triplets. Patterns with two triplets were therefore adopted as compromise and used in all searches reported here. Patterns defined as the example above would still be too strict, as they would miss interruptions were Gly residues occur in normal Gly-X-Y triplets (that is, without creating an interruption). To minimize chances of false positive matches to non-collagenous Gly-rich repetitive sequences (such as those from silk proteins), searches were conducted in several steps where Gly residues were at first completely avoided from the X and Y positions of the motif, and then introduced at specific positions one at a time, to a maximum of two Gly residues per side and not allowing Gly$_3$ or longer stretches. Last, Gly residues were allowed in the internal positions of the Gly-Z_n-Gly motifs but the resulting tandem interruptions (defined later) were not analysed separately (for example a Gly-Z_3-Gly-Z_4-Gly motif was considered a G8G interruption). This complicated strategy was devised to minimize both the number of missed true matches and that of false positive matches.

2.2. Protein sequence databases: taxonomic collagenomes

The UniProt KnowledgeBase protein database (Boutet et al., 2007) was used for all searches, applying specific taxonomic filters when needed (human, metazoa, invertebrates, bacteria, etc.). To minimize false positive matches, subsets of sequences annotated as containing collagen triple helix repeats by InterPro (entry IPR008160, release 46.0, 27th January 2014) (Hunter et al., 2012) were created as user-defined databases. The term “collagenome” is used here to designate the collection of sequences of collagens and collagen-like proteins of a group of organisms. For example, the group of sequences from invertebrates that are annotated to contain at least one collagen triple helix is collectively referred as the “invertebrate collagenome”.

A typical search run in ScanProsite would involve a sequence motif defined as above (e.g. G1G) against one of these user-defined databases (e.g. invertebrate collagenome) uploaded into the ScanProsite website.

2.3. Lists of matches

Sequence patterns and complete lists of matches from the searches reported here are available directly from the author.

2.4. Residue frequency plots

Positive matches for the different interruption motifs were used to produce frequency-based sequence logos in WebLogo (Crooks et al., 2004), using its own default amino acid colour scheme (black, AFILMPVW; green, GCSTY; blue, KHR; red, DE; purple, NQ).

3. Results

3.1. A systematic nomenclature for collagen interruptions

Table 1 illustrates several ways in which the consensus collagen sequence can be interrupted. A typical collagen sequence around any arbitrarily chosen triplet can be represented with the (Gly-X-Y)$_n$-Gly-Z_2-Gly-(X-Y-Gly)$_n$ notation, where Z is any amino acid. If one Z residue from the central Gly-Z_2-Gly group is deleted, a Gly-Z-Gly interruption arises, represented here by the abbreviation G1G. Deletion of both Z residues produces a Gly-Gly interruption, represented as G0G. Insertion of an extra Z amino acid in the central group results in a Gly-Z_3-Gly interruption (G3G). Deletion of one single Gly residue produces a Gly-Z_4-Gly interruption (G4G), which is formally equivalent to inserting two additional Z residues into the central group. Similarly, deletion of two Gly residues results in a Gly-Z_6-Gly interruption (G6G) equivalent to inserting four additional Z residues into the central group. Substitution of one Gly residue to a Z amino acid results in a Gly-Z_5-Gly interruption (G5G). Increasingly longer interruptions can be thought as the

Please cite this article in press as: Bella, J. A first census of collagen interruptions: Collagen’s own stutters and stammers. J. Struct. Biol. (2014), http://dx.doi.org/10.1016/j.jsb.2014.03.022