The International Journal of Biochemistry & Cell Biology 45 (2013) 2392–2401

Contents lists available at ScienceDirect

The International Journal of Biochemistry & Cell Biology

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

Prediction and analysis of higher-order coiled-coils: Insights from proteins of the extracellular matrix, tenascins and thrombospondins

Thomas L. Vincent[a,b,1], Derek N. Woolfson[b,c,*], Josephine C. Adams[c,**]

[a] Bristol Centre for Complexity Sciences, University of Bristol, Queen's Building, University Walk, Bristol BS8 1TR, UK
[b] School of Chemistry, University of Bristol, Cantock's Close, Bristol BS8 1TS, UK
[c] School of Biochemistry, University of Bristol, Medical Sciences Building, University Walk, Bristol BS8 1TD, UK

ARTICLE INFO

Article history:
Received 7 May 2013
Received in revised form 14 June 2013
Accepted 9 July 2013
Available online 24 July 2013

Keywords:
Computational biology
Protein-structure prediction
Coiled coil
Extracellular matrix
Oligomer

ABSTRACT

α-Helical coiled-coil domains (CCDs) direct protein oligomerisation in many biological processes and are of great interest as tools in protein engineering. Although CCDs are recognizable from protein sequences, prediction of oligomer state remains challenging especially for trimeric states and above. Here we evaluate LOGICOIL, a new multi-state predictor for CCDs, with regard to families of extracellular matrix proteins. Tenascins, which are known to assemble as trimers, were the first test case. LOGICOIL out-performed other algorithms in predicting trimerisation of these proteins and sequence analyses identified features associated with many other trimerising CCDs. The thrombospondins are a larger and more ancient family that includes sub-groups that assemble as trimers or pentamers. LOGICOIL predicted the pentamerising CCDs accurately. However, prediction of TSP trimerisation was relatively poor, although accuracy was improved by analyzing only the central regions of the CCDs. Sequence clustering and phylogenetic analyses grouped the TSP CCDs into three clades comprising trimers and pentamers from vertebrates, and TSPs from invertebrates. Sequence analyses revealed distinctive, conserved features that distinguish trimerising and pentamerising CCDs. Together, these analyses provide insight into the specification of higher-order CCDs that should direct improved CCD predictions and future experimental investigations of sequence-to-structure functional relationships.

1. Introduction

The α-helical coiled-coil domain (CCD) is found in all organisms and accounts for approximately 3% of all protein-encoding regions of genomes (Rackham et al., 2010). Coiled coils are important in mediating protein–protein oligomerisation interactions in a multitude of biological processes that underpin cell organization and morphology and, in multicellular organisms, tissue stability. All CCDs have two key features in common: first, at the structural level, they comprise bundles of two or more amphipathic α-helices; secondly, underlying this structure at the sequence level, they have a *heptad*, or similar, sequence pattern of hydrophobic (*H*) and polar (*P*) residues, *HPPHPPP*, often designated *abcdefg*. Because the α-helix has 3.6 residues per turn, and the average spacing of *H*-type residues in the heptad repeat is 3.5 residues, the helices necessarily wrap around each other to form rope-like structures. These common and apparently straightforward features mask considerable sequence and structure diversity: coiled coils form oligomers of between 2 and 5 helices, though larger assemblies are known; the helices can be parallel or antiparallel; the assemblies can be homo- or hetero-typic, and the lengths of the rod-like domains formed can range from nm – μm, spanning 10 s to 100 s of residues, respectively (Lupas, 1996; Mason and Arndt, 2004; Lupas and Gruber, 2005; Moutevelis and Woolfson, 2009).

Because of their relatively straightforward sequence characteristics, and the clear link between protein sequence and 3D structure, considerable effort has gone into predicting CCDs and their associated oligomeric states and stabilities from sequence data alone. This has met with mixed success. Over the past two decades, it has become possible to recognize with some confidence potential coiled-coil-forming regions within protein sequences (Lupas et al., 1991; Berger et al., 1995; Delorenzi and Speed, 2002). However, prediction of coiled-coil oligomer state has been

Abbreviations: CCD, coiled-coil domain; ECM, extracellular matrix; TN, tenascin; TSP, thrombospondin.

* Corresponding author at: School of Chemistry, University of Bristol, Cantock's Close, Bristol BS8 1TS, UK. Tel.: +44 0117 954 6347; fax: +44 0117 92 77985.
** Corresponding author at: School of Biochemistry, University of Bristol, Medical Sciences Building, University Walk, Bristol BS8 1TD, UK. Tel.: +44 0117 3312361; fax: +44 0117 3312168.
E-mail addresses: D.N.Woolfson@bristol.ac.uk (D.N. Woolfson), jo.adams@bristol.ac.uk (J.C. Adams).
[1] Department of Genetic Medicine, Weill Cornell Medical College, 1305, York Avenue, New York, NY 10021, USA.

1357-2725/$ – see front matter © 2013 Elsevier Ltd. All rights reserved.
http://dx.doi.org/10.1016/j.biocel.2013.07.011

much more challenging. Most current algorithms function as two-state predictors of oligomer state, distinguishing between parallel dimers and trimers (Woolfson and Alber, 1995; Wolf et al., 1997; Armstrong et al., 2011; Mahrenholz et al., 2011; Trigg et al., 2011). Yet, parallel dimers and trimers account for only ≈30% of experimentally verified coiled-coil structures (Mason and Arndt, 2004). To address this gap in oligomer-state prediction, we developed the LOGICOIL algorithm, which has capacity for multi-state prediction of parallel or antiparallel dimers, parallel trimers and parallel tetramers, thus covering ≈90% of known coiled-coil-structure space (Vincent et al., 2013).

Understanding sequence-to-structure relationships is of particular interest for oligomer-states above dimers, because of the above-mentioned gap in our ability to predict these states and also because of the great potential for protein engineering or clinically relevant applications. Trimerising, tetramerising or pentamerising CCDs can be used to generate engineered polypeptide clusters, for example as "super-ligands" for interactions that are based in nature on monomeric ligands (Voulgaraki et al., 2005; Kim et al., 2009), or to generate potent blocking reagents (Wang et al., 2008). Due to geometric constraints, pentamer, hexamer and other higher-order coiled-coil assemblies contain a central channel, raising the possibility of additional applications such as internal loading with a cargo molecule or use as pore structures (Guo et al., 1998; Ozbek et al., 2002; Zaccai et al., 2011). To improve predictions and develop these possible applications, a better understanding of natural higher-order coiled coils is needed.

The metazoan extracellular matrix (ECM) is a compartment rich in oligomeric proteins, many of which assemble through the action of coiled coils (Engel, 2004). From these, we selected the tenascin and thrombospondin families as particularly suitable for analysis of sequence determinants for oligomer-state assembly. Both families have relatively short CCDs of 3–5 heptad repeats; the domains assemble exclusively as parallel coiled coils, and appear in nature predominantly as homo-oligomers. Experimental analyses of oligomer status have been carried out for multiple members of both families, allowing for comparison of bioinformatics predictions with native molecules.

Tenascins (TNs) are large multi-domain glycoproteins that evolved in the chordate lineage (Tucker and Chiquet-Ehrismann, 2009). TNs are highly expressed during embryogenesis and also contribute to the adult ECM during tissue regeneration or remodelling, such as during wound healing or inflammation (Jones and Jones, 2000; Chiquet-Ehrismann and Tucker, 2011). Vertebrates encode four TNs, designated TN-C, TN-R, TN-X and TN-W, which have arisen from a single ancestor through a series of gene duplication events (Tucker et al., 2006). TNs oligomerise by a complex, two-step process in which CCDs located close to the *N*-terminus mediate assembly of homo-trimers. In TN-C, and likely also in TN-W, trimerisation increases the affinity of a so-called *N*-terminal assembly domain to enable a dimer of trimers to be assembled "head-to-head". Cysteine residues within the *N*-terminal assembly domain of TN-C may stabilize the resulting hexabrachion, but are not essential for its formation (Nies et al., 1991; Kammerer et al., 1998; Scherberich et al., 2004). Based on electron microscope images of purified proteins, TN-R and TN-X assemble predominantly as trimers (Pesheva and Probstmeier, 2000; Lethias et al., 2006).

Thrombospondins (TSPs) are multi-domain, low abundance proteins of adult ECM that are important for cell-ECM and cell-cell interactions (Adams and Lawler, 2011). All TSPs contain a conserved set of domains in their *C*-terminal regions, and the majority are assembled and secreted as oligomers through the action of a CCD adjacent to an *N*-terminal, laminin G-like domain. TSPs have been studied extensively in mammals, where they constitute a five-member gene family, and individual family members function in angiogenesis, vascular biology, immune response, wound repair, synaptogenesis and connective tissue organization. The TSPs of vertebrates are divided into two structural sub-groups, A and B, which differ in overall domain architecture and oligomer status (Adams and Lawler, 2011). Biochemical and electron microscopy data have demonstrated that TSP1 and TSP2 assemble as trimers and TSP3, -4 and -5 assemble as pentamers (Lawler et al., 1985, 1995; Sottile et al., 1991; Mörgelin et al., 1992; Qabar et al., 1995). The CCD of TSP5 has been resolved at atomic level to reveal a pentameric α-helical bundle, which is 7.3 nm long and has a central pore of diameter 0.2–0.6 nm (Malashkevich et al., 1996). The recent appreciation that TSPs are amongst the most highly conserved of metazoan ECM components, present from sponges to human (Bentley and Adams, 2010; Ozbek et al., 2010), has provided many additional sequences for analysis. Most invertebrates encode a single TSP with a domain organization similar to the vertebrate B subgroup of TSPs; correspondingly, *Drosophila* TSP assembles as a pentamer (Bentley and Adams, 2010; Adams et al., 2003). Given that pentamers represent <1% of all known coiled-coil structures in nature, and that most of these are membrane-inserted proteins, the TSP family is an important context in which to identify essential attributes of pentamerising CCDs in secreted proteins and more generally.

The studies presented here apply LOGICOIL to the prediction of natural higher order coiled coils and combine additional forms of sequence analysis to expand our understanding of sequence-to-structure relationships in natural trimers and pentamers.

2. Materials and methods

2.1. Datasets of tenascin and thrombospondin sequences

The dataset of 32 TN sequences was compiled from those listed in the Supplementary Information file of Ozbek et al. (2010), plus the complete protein sequences of *Branchiostoma floridae* TN and *Ciona intestinalis* TN as reported in Tucker and Chiquet-Ehrismann (2009) and Tucker et al. (2006), respectively. The dataset of TSPs, extended from that presented in Bentley and Adams (2010), included 30 subgroup A TSP sequences and 47 TSP sequences with subgroup B domain architecture from early diverging metazoa, protostomes and deuterostomes.

2.2. Identification of coiled-coil domains

CCDs of TNs and TSPs were identified and annotated according to consensus-based decisions of the MARCOIL, SOSUIcoil, COILS and PrOCoil prediction algorithms (Delorenzi and Speed, 2002; Mahrenholz et al., 2011; Lupas et al., 1991; Tanizawa et al., 2008). MARCOIL and SOSUIcoil algorithms were used with default parameters; MARCOIL was used from source code (not the web interface). A cutoff window length of 21 residues was used for COILS. PrOCoil is set up such that positive and negative scores indicate trimer and dimer predictions, respectively. For consistency of presentation in Fig. 1, we inverted the sign of the ProCoil scores, i.e., dimer predictions are presented as positive scores and trimer predictions as negative scores. MARCOIL predictions for coiled coils >13 residues in length and with >10% confidence score for each residue were included in the dataset. All predictions were inspected manually to ensure that the position of the coiled coil in each full-length sequence was as expected based on prior knowledge of TSP and TN domain architectures; this process also ensured that predictive coverage was maximized but not at the expense of accuracy. The *N*- and *C*- termini of each CCD were also curated manually depending on heptad repeat register shifts, sharp drop-offs in scores (generally attributed to the occurrence of helix breaking glycine or proline