the FEBS Journal

How to remain nonfolded and pliable: the linkers in modular α-amylases as a case study

Georges Feller[1], Dominique Dehareng[1] and Jean-Luc Da Lage[2]

1 Center for Protein Engineering, University of Liège, Liège-Sart Tilman, Belgium
2 UPR9034 Evolution, Génomes et Spéciation, CNRS, Gif sur Yvette, France

Keywords
glycoside hydrolases; intrinsically disordered proteins; protein folding; protein unfolding; α-amylases

Correspondence
G. Feller, Laboratory of Biochemistry, Institute of Chemistry B6a, B-4000 Liège-Sart Tilman, Belgium
Fax: +32 4 366 33 64
Tel: +32 4 366 33 43
E-mail: gfeller@ulg.ac.be

(Received 17 December 2010, revised 18 April 2011, accepted 28 April 2011)

doi:10.1111/j.1742-4658.2011.08154.x

The primary structure of linkers in a new class of modular α-amylases constitutes a paradigm of the structural basis that allows a polypeptide to remain nonfolded, extended and pliable. Unfolding is mediated through a depletion of hydrophobic residues and an enrichment of hydrophilic residues, amongst which Ser and Thr are over-represented. An extended and flexible conformation is promoted by the sequential arrangement of Pro and Gly, which are the most abundant residues in these linkers. This is complemented by charge repulsion, charge clustering and disulfide-bridged loops. Molecular dynamics simulations suggest the existence of conformational transitions resulting from a transient and localized hydrophobic collapse, arising from the peculiar composition of the linkers. Accordingly, these linkers should not be regarded as fully disordered, but rather as possessing various discrete structural patterns allowing them to fulfill their biological function as a free energy reservoir for concerted motions between structured domains.

Introduction

Following its linear synthesis on the ribosome, a polypeptide must adopt its final and biologically active three-dimensional conformation. The forces driving protein folding are essentially based on the hydrophobic effect: the entropic cost of encaging nonpolar groups in the water molecule network is high and the system evolves towards the burial of these groups within a globular structure, away from the water molecules in the solvent. During the process of folding, as well as in its final fold, the stability of the molecular edifice is further modulated by interactions between groups that have been brought into contact. In proteins, van der Waals' interactions and hydrogen bonds are the most abundant, but salt bridges (or ion pairs), aromatic (or cation–π) interactions and some structural disulfide bonds make a substantial contribution to stability. Structural factors are also involved, such as the occurrence of Gly residues, which allow a large diversity of dihedral rotation, or Pro residues, which, by contrast, induce local rigidity in the polypeptide chain.

However, in certain specific cases, localized protein regions must remain nonfolded to fulfill their biological functions. Linkers found in carbohydrate-active enzymes are a typical example of such natively unfolded proteins. These linkers are amino acid segments of variable length, generally connecting a catalytic domain bearing the active site to a carbohydrate-binding module, which mediates attachment to the macromolecular substrate [1–5]. Significantly, in the crystal structure of these modular enzymes, no electron density is observed for the linker residues, indicating local disorder [6]. Nevertheless, small-angle X-ray scattering experiments have revealed that the linkers can adopt numerous nonrandom conformations, from sharply bended or compact to fully extended states [7–10]. Furthermore, it has been proposed that these modular enzymes can move on the substrate surface with a caterpillar-like displacement, a process in which the linker acts as a free energy reservoir [7].

In this study, we report a new group of α-amylases displaying a modular organization in which the linker sequences represent a biochemical paradigm that illustrates the structural parameters required to allow a polypeptide to remain unfolded, extended and flexible.

Results and Discussion

Identification of new modular α-amylases

α-Amylases are ubiquitous enzymes hydrolyzing α-1,4-glycosidic bonds of starch and related polysaccharides, such as glycogen, and belonging to family 13 in the glycoside hydrolase classification (http://www.cazy.org/). Amongst these enzymes, animal-type α-amylases are homologous enzymes present in all animals and in some rare bacteria [11,12]. They are nonmodular, consisting of a single globular catalytic domain, with the noticeable exception of the α-amylase from the bacterium *Pseudoalteromonas haloplanktis*, which displays an additional small (21 kDa) C-terminal domain having the size of a carbohydrate-binding module, but not previously reported in any other glycosidases. To delineate the occurrence and function of this new module, several animal cell extracts were screened using antibodies raised against the previously purified putative binding domain and six α-amylase genes from molluscan species were sequenced. Furthermore, we performed a search based on the primary structure of this domain in recently available sequence and genome databanks. Interestingly, this domain was found in some closely related bacterial species, but mainly in nonvertebrate animals, and invariably connected via a linker to an animal-type α-amylase (listed in Table S1), as shown in Fig. 1. More specifically, the primary structure of these linkers was remarkable if it is remembered that such polypeptides must be 'pliable' [13] and are expected to behave as a spring, allowing the nanomachine (catalytic domain–linker–binding module) to crawl on the substrate surface. The possible functional implications of the linker primary structures are presented in the following sections.

Amino acid bias: flexibility and rigidity

A close inspection of the linker sequences shown in Fig. 1 reveals a strong amino acid compositional bias, which is quantified in Table 1 in comparison with a subset of globular proteins [13] and with the whole Swiss-Prot databank. The 833 amino acid residues forming the 31 linkers are characterized by a significant enrichment in Pro, Gly, Thr and Ser (statistical data in Table S2). Gly and Pro constitute two extreme opposites for the dynamics of a polypeptide chain. The unusual abundance of Gly can be explained by the absence of a side chain, allowing dihedral angles not accessible to other residues and therefore promoting large-amplitude rotations around its α carbon. In Fig. 1, Gly has a strong propensity to be located near the N- and C-termini of the linkers: this suggests that a mobile connection with both the catalytic domain and the binding module is required for the function of the nanomachine. Furthermore, Gly repeats (*Corbicula*, *Haliotis*, *Strogylocentrotus*) and Gly-rich sequences (*Crassostrea*, *Mytilus*, *Acanthochitona* Amy1, etc.) within the linkers obviously provide additional flexibility. Pro is the most abundant residue in these linkers. As a result of the pyrrolidine cycle formed by its side chain bond to the terminal amino group, the dihedral angles with the preceding residue are severely restricted, introducing a rigid center in the polypeptide chain. No preferential location of Pro has been noted in the linker sequences, but some Pro repeats (*Ancylostoma*, *Daphnia* Amy1, *Platynereis*, *Venerupis*) can possibly adopt the stiff polyproline helix conformation. Overall, the distribution pattern of Gly and Pro in the linkers indicates, in many cases, a sequential arrangement of rigid peptides connected by mobile segments.

Polar versus nonpolar residues

As far as polar and nonpolar amino acids are concerned, the linkers are depleted in aliphatic residues (14.4% versus 28.9% in globular proteins, Gly excluded) and aromatic residues (3.6% versus 9% in globular proteins). Met, which possesses a marked hydrophobic character [14], is also avoided (Table 1). There is therefore a much weaker driving force for folding the connecting linkers when compared with a globular protein. In addition, the main polar uncharged side chains (Asn, Gln, Ser, Thr) are overrepresented in the linker sequences (34.6% versus 20.8% in globular proteins). Accordingly, extensive hydrogen bonding with the solvent should counteract the hydrophobic effect and favor an unfolded state of the linkers. In this context, we can wonder why aliphatic and aromatic residues are not totally avoided in linker sequences to prevent folding. These residues are either clustered (*Petrolisthes*) or randomly distributed (*Daphnia* Amy2) in the linker sequences. These hydrophobic residues presumably induce a local, transient and weak folding of the linkers, in agreement with small-angle X-ray scattering results showing occurrences of compact conformers [10]. This may be the physical basis of the postulated spring effect, with energy accumulation by a localized hydrophobic collapse when the linker shortens (bent, caterpillar-like state). It should be mentioned that the hydrophobic effect of a methylene group has been estimated to be approximately 5 $kJ \cdot mol^{-1}$ [15], whereas the enthalpy of α-1,4-glycosidic bond hydrolysis is 4.5 $kJ \cdot mol^{-1}$ [16]. If it is assumed that the catalytic domain processively

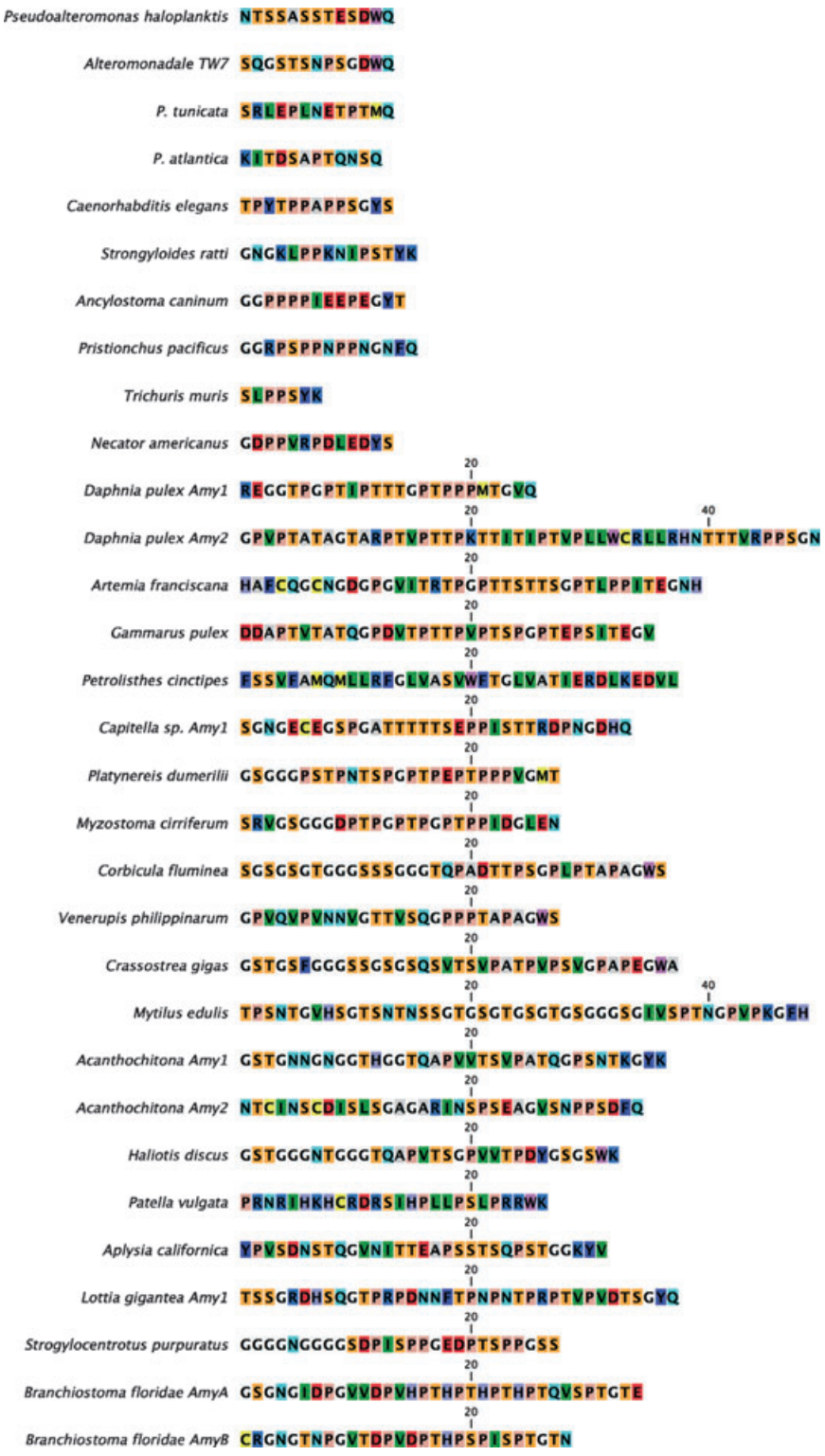

Fig. 1. Primary structures of linkers in modular animal-type α-amylases. Sequences are phylogenetically grouped and are not aligned by sequence similarity. The selection of both N- and C-terminal sequence limits is described in the Materials and methods section. The color code indicates side chains with similar chemical function according to the RasMol standard (Pro, flesh colored; Gly, white; Asp, Glu, red; Arg, Lys, blue; Cys, Met, yellow; Ser, Thr, orange; Asn, Gln, cyan; Phe, Tyr, mid-blue; Trp, purple; Leu, Val, Ile, green; Ala, gray; His, pale blue).

Table 1. Amino acid frequencies (%) in α-amylase linkers, in a set of globular proteins, in the Swiss-Prot databank and in intrinsically unstructured proteins.

Amino acid	Linkers[a]	Globular proteins[b]	Swiss-Prot[c]	Intrinsically unstructured proteins[b]
Ala	3.4	8.1	8.3	7.1
Arg	2.9	4.6	5.5	4.2
Asn	5.3	4.7	4.0	2.1
Asp	3.7	5.8	5.4	5.0
Cys	1.0	1.6	1.4	0.6
Gln	3.0	3.7	3.9	4.5
Glu	2.8	6.0	6.8	14.3
Gly	16.2	8.0	7.1	4.3
His	1.9	2.3	2.3	1.5
Ile	2.6	5.4	6.0	3.7
Leu	2.6	8.4	9.7	5.4
Lys	1.7	6.0	5.8	10.4
Met	0.6	2.0	2.4	1.3
Phe	1.2	3.9	3.9	1.7
Pro	16.7	4.6	4.7	12.1
Ser	12.0	6.3	6.5	6.9
Thr	14.3	6.1	5.3	5.1
Trp	1.1	1.5	1.1	0.3
Tyr	1.3	3.6	2.9	1.4
Val	5.8	7.0	6.9	8.0

[a] Data for 833 amino acids in the 31 linkers shown in Fig. 1. [b] Data from ref. [13]. [c] Data from Swiss-Prot release 57.15 for 515 203 sequences.

hydrolyzes glycosidic linkages along the substrate chain, and that a full energy transfer occurs from the hydrolyzed bond to the nanomachine, each α-1,4 bond hydrolyzed has the theoretical capacity to disrupt a hydrophobic interaction in the linker, inducing or favoring its extension. As the catalytic constant k_{cat} of animal-type α-amylases is in the range of 300–600 α-1,4 bonds hydrolyzed per second [17], a single caterpillar-like motion could occur in the millisecond range, in agreement with the time scales observed for large concerted motions in polypeptide backbones [18,19]. The shorter size of linkers in bacteria and in some animals (Fig. 1) probably precludes a significant hydrophobic collapse, but nevertheless provides a mobile connection between the functional domains.

Polar and charged residues

Within the class of hydrophilic residues, the large excess of hydroxylated side chain (Ser, Thr; 26.3%) over the amide-containing Asn and Gln (8.3%) is appealing. This may be related to the selection for groups forming strong and stable hydrogen bonds with water molecules. Indeed, amongst the various stereochemical parameters involved in hydrogen bond strength (distance, coplanarity, etc.), the pK_a difference between heteroatoms is of importance: the smaller this difference, the stronger the hydrogen bond, as the hydrogen atom is equally shared between the donor and the acceptor [20]. As a result, hydrogen bonds formed by hydroxyl groups (O–H···O, $\sim$ 21 kJ·mol^{-1}) are twice as strong as those formed by the amide group [21]. Furthermore, hydrogen bonds formed by the single hydroxyl donor in Ser and Thr are expected to be more stable than those involving the amide group, which compete for various water molecules via possible bifurcated hydrogen bonds [22]. In this respect, 54% of Pro residues in the linkers are either preceded or followed by Ser or Thr: maintaining the hydroxyl donor in a rigid environment may possibly contribute to the stabilization of hydrogen bonds with the solvent. The four His–Pro–Thr repeats of *Amphioxus* AmyA are worth mentioning as they should form a rather rigid and hydrophilic peptide. The abundance of Ser and Thr residues in linkers from animals also provides numerous potential targets for *O*-linked glycosylation. By contrast, only five potential sites for *N*-glycosylation were detected [*Daphnia* Amy2, *Mytilus*, *Aplysia* (2 sites) and *Branchiostoma* AmyB]. Glycosylation is expected to modulate the linker dynamics [23], but this aspect cannot be addressed from the primary structure alone and requires further experimental evidence.

The linker sequences are typically depleted in charged residues (11.0% versus 22.4% in globular proteins, His excluded). This may be related to the avoidance of formation of stable salt bridges between oppositely charged residues brought into contact in the flexible conformers. However, the distribution of these residues is nonrandom in the linkers. Firstly, most linkers display either a net negative or a net positive charge. This is exemplified in *Gammarus* and *Capitella* (five acidic groups), *Patella* (eight basic groups) and *Daphnia* Amy2 (five basic groups). Secondly, identically charged residues are frequently adjacent (six occurrences) or at close proximity (12 occurrences) in the sequences. Both properties should result in strong electrostatic repulsions, promoting an extended conformation of the linkers. Furthermore, adjacent residues with opposite charges are observed in five linkers (*Daphnia* Amy1, *Capitella*, *Petrolisthes*, *Patella* and *Lottia*). In folded proteins, an ion pair between adjacent acidic and basic side chains is unlikely as a result of the steric constraints imposed on dihedral angles, but this limitation may be less relevant in an unfolded linker. Nevertheless, the strong electrostatic attraction between these neighboring charges should restrict the available dihedral angles between the participating

residues and induce local rigidity that complements the function of Pro. A similar limitation to dihedral angles should be obtained by adjacent identically charged side chains, but by repulsion in this case.

Cysteine and disulfide-linked loops

The occurrence of a single Cys residue in five linkers is intriguing as this residue is prone to oxidation, especially for these extracellular α-amylases. Therefore, it seems that the weakly polar sulfhydryl group is important for the linker structure and is protected from oxidation. Indeed, it should be noted that animal α-amylases are released in the intestinal tract where oxygen concentration is expected to be low, whereas bacterial linkers are devoid of Cys. However, *Artemia* and *Acanthochitona*2 linkers display two Cys residues at close proximity. This is reminiscent of a bacterial cellulase linker possessing 10 Cys residues, forming a series of five disulfide-linked small loops [9]. Such possible loops in *Artemia* and *Acanthochitona*2 linkers certainly provide steric hindrance to local folding. In addition, these covalently linked loops may also constitute a proteolytic trap. Unstructured chains are extremely susceptible to proteolytic cleavages [13,24] that definitively abolish the modular structure and its function. Proteolytic cleavages within such solvent-exposed loops should increase the probability to maintain the linker connectivity via the disulfide bond.

An aromatic group at the C-terminus

Amongst the 31 identified linkers, 18 (58%) possess an aromatic side chain at the −2 position from the C-terminus. As the main bodies of these sequences are unrelated, this preferential position is apparently not fortuitous. It can be proposed that the large, planar aromatic group acts as a lubricant with the binding module surface for rotational motions of the linker, through, for instance, electrostatic repulsion from the δ^- π-electron cloud covering the face of the aromatic ring [25]. Alternatively, the ring may sterically disfavor extensive bending in this region, which could result in unwanted interactions between the linker and the binding module. In this respect, 72% of the C-terminal aromatic residues are preceded by Gly at the −3 or −4 position, indicating that mobility of the connecting region is required at the N-terminus of the aromatic side chain.

Modeling and molecular dynamics simulations

In order to address the possible conformations and motions of the linkers, model building and molecular dynamics simulations were performed on a subset of primary structures (*Pseudoalteromonas tunicata*, *Daphnia pulex* Amy2, *Platynereis dumerilii*, *Corbicula fluminea* and *Venerupis philipinnarum*). In a first step, the linker sequences were used as a query to screen the Protein Data Bank for similar sequences in proteins of known tridimensional structure using the program YASARA. In addition, the sequences were modeled by PEP-FOLD [26]. This approach does not retrieve a unique conformation, but rather a series of conformers either in an elongated state or in slightly collapsed or structured states (Fig. 2). This is a clear indication that sequences similar to the linkers are found in diverse and loosely packed conformations in known protein structures. It is worth mentioning that the various predicted linker conformations closely resemble the modeled conformational ensemble obtained from small-angle X-ray scattering experiments on a cellulase linker [9].

During molecular dynamics simulations, the linker total energy (peptide and solvation) of the conformers (for a given sequence) can vary significantly (up to

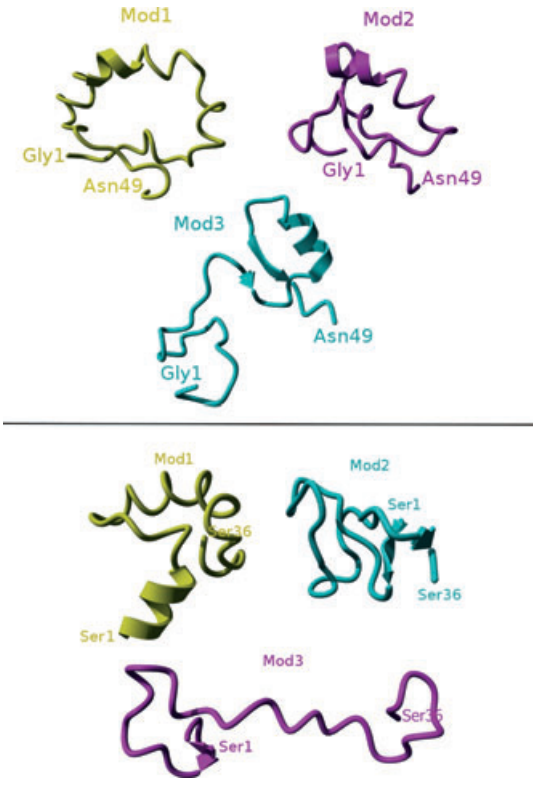

Fig. 2. Predicted conformers of α-amylase linkers. The models illustrated are from *Daphnia pulex* Amy2 (top panel) and *Corbicula fluminea* (bottom panel). In both cases, the three conformations with the lowest energy are shown as ribbon representations.

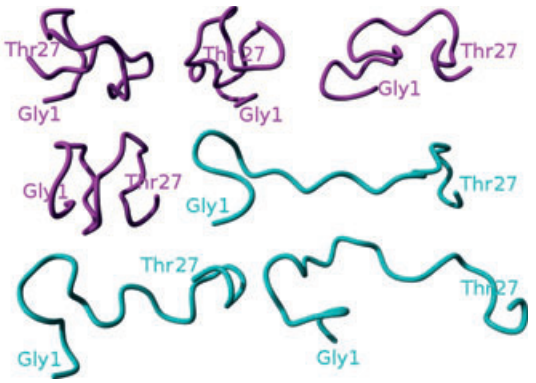

Fig. 3. Molecular dynamics simulations. Ribbon representation of Cα chain of four folded (magenta) and four extended (cyan) conformations in the *Platynereis dumerilii* linker in an 11 ns simulation.

300 kJ·mol^{-1} observed in the simulations). Thus, temperature can affect significantly the geometry of the conformers, as expected for weakly structured peptides. Furthermore, on a 1 ns simulation time scale, the linker backbones are mobile and the short predicted secondary structures tend to move along the primary structure (Fig. S1), whereas α helices tend to stretch or to bend. This was confirmed by two longer lasting simulations performed on *D. pulex* Amy2 (15 ns) and *P. dumerilii* (11 ns) linkers, for which many different conformations were found (Fig. 3). The *D. pulex* linker (rich in aliphatic side chains) displayed versatile structures, remaining globular (Fig. S2), whereas the *P. dumerilii* linker (only one Val) moved from a series of folded to extended structures (Fig. 3). Together, these results suggest a dynamic ensemble of conformers, ranging from fully extended to loosely folded states, which are compatible with the proposed caterpillar-like motions of glycosidase nanomachines.

Conclusions

The above-mentioned amino acid bias in α-amylase linkers represents a specific and extreme trend of the bias observed in natively unfolded proteins (Table 1), as far as depletion in aliphatic/aromatic residues and enrichment in hydrophilic/Pro residues are concerned [13,24,27–30]. As a result, algorithms that have been developed as predictors of protein disorder (see Ref. [31] for compilation) invariably predict most α-amylase linkers to be intrinsically unstructured. However, the long linkers in Fig. 1 display a trend towards a minimal predicted disorder centered on the middle part of the sequences. This supports our suggestion that a weak and local fold can contribute to shorten or to bend these linkers, in agreement with modeling and molecular dynamics simulations. Accordingly, the linkers should not be regarded as fully disordered, but rather as polypeptides possessing various discrete structural patterns allowing them to remain extended, pliable and to function as an energy reservoir, possibly using localized hydrophobic collapse and torsional forces on the backbone during bending. The sequential organization into Pro-based rigid peptides and Gly-based mobile peptides can be considered as an elementary organization level, as well as the occurrence of Pro repeats, disulfide-linked loops and acidic/basic repeats, which can be tentatively regarded as pseudo-secondary structures. It is also worth mentioning that the linker primary structures closely resemble that of the Pro- and Gly-rich repeats in tropoelastin, a key component of vertebrate elastic fibers. Furthermore, the elastomeric properties have been related to the capacity to shift from a weakly globular structure to an extended form, mediated by the Pro- and Gly-rich repeats [32,33]. Accordingly, the α-amylase linkers have the additional potential to behave as elastic oligopeptides. It is expected that the present theoretical dissection of the linker sequences will stimulate further experimental approaches, such as the biophysical characterization of isolated linker peptides and the engineering of size and composition variations in order to address their function in activity, substrate binding and structural dynamics.

Materials and methods

Experimental data

The presence of a C-terminal putative binding domain in various animal cell extracts was checked experimentally by western blotting (not shown) using antibodies raised against the previously purified C-terminal domain from *P. haloplanktis* α-amylase [34]. This prompted us to sequence entirely the α-amylase genes from the bivalves *C. fluminea* and *Mytilus edulis*, and almost entirely the gene from the limpet *Patella vulgata*, using the Genome walker Universal kit (Clontech, Mountain View, CA, USA). The C-terminal domains were identified by BLAST search in the GenBank database. From the alignment of these domains with those of *P. haloplanktis* and *Caenorhabditis elegans*, PCR primers were designed from conserved parts of the domain, and various combinations were used for amplification of fragments showing attachment to the core α-amylase sequence, i.e. also using primers derived from the core enzyme. The reverse primers designed from the C-terminal domain were as follows: 2FIRREV, 5′-CCNCKNABRAAMANATCCT GTCC-3′; CTERMREV, 5′-TCNGCNCCRTACCARTC-3′.

The species assayed by PCR were the chiton *Acantochitona* sp. (Mollusca, Polyplacophora) and the oyster *Crassostrea gigas* (Mollusca, Bivalvia). Sequence data were deposited in GenBank (Table S1).

Searches in databases

Using the putative C-terminal domain of *C. fluminea* as a query, sequence databases were searched by BLASTP and TBLASTN for the occurrence of domains similar to the *P. haloplanktis* C-terminal domain. URLs of the relevant genome databases are given in Table S1. The linker between the core enzyme and its C-terminal domain was defined as the region between the end of the usual α-amylase sequence and the first conserved motif of the C-terminal putative binding domain, similar to the sequence RTVIF.

Molecular dynamics simulations

The preliminary step was the building of the tridimensional structure by homology, using YASARA (http://www.yasara.org/) and PSI-BLAST [35], as well as PEP-FOLD [26]. As PEP-FOLD can deal with a maximum length of 25 amino acids, the whole structure of larger linkers was built manually on the basis of overlapping results from PEP-FOLD. The second step was the soaking of the linker in a neutralized water box containing 0.9% NaCl. The box extended 3 Å around all atoms. The geometry of the whole system was optimized using the YAMBER3 force-field [36].The third step was the molecular dynamics simulation at 298 K from 500 to 1000 ps, the first 250 ps being considered as the equilibration step. The fourth step was the selection of several conformations randomly chosen among the molecular dynamics simulation snapshots, the optimization of their geometry and the determination of their total energy. For *D. pulex* Amy2 and *P. dumerilii* linkers, longer molecular dynamics simulations were performed, lasting 15 and 11 ns, respectively.

Acknowledgements

This work was supported by grants from the FRS-FNRS (Fonds National de la Recherche Scientifique, Belgium) to G.F. and from the Centre National de la Recherche Scientifique (France) to J.-L.D.L. D.D. was supported by the Poles of Attraction of the Belgian Science Policy (IAP No. P6/19).

References

1 Bourne Y & Henrissat B (2001) Glycoside hydrolases and glycosyltransferases: families and functional modules. *Curr Opin Struct Biol* **11**, 593–600.
2 Boraston AB, Bolam DN, Gilbert HJ & Davies GJ (2004) Carbohydrate-binding modules: fine-tuning polysaccharide recognition. *Biochem J* **382**, 769–781.
3 Hashimoto H (2006) Recent structural studies of carbohydrate-binding modules. *Cell Mol Life Sci* **63**, 2954–2967.
4 Machovic M & Janecek S (2006) Starch-binding domains in the post-genome era. *Cell Mol Life Sci* **63**, 2710–2724.
5 Shoseyov O, Shani Z & Levy I (2006) Carbohydrate binding modules: biochemical properties and novel applications. *Microbiol Mol Biol Rev* **70**, 283–295.
6 Receveur-Brechot V, Bourhis JM, Uversky VN, Canard B & Longhi S (2006) Assessing protein disorder and induced folding. *Proteins* **62**, 24–45.
7 Receveur V, Czjzek M, Schulein M, Panine P & Henrissat B (2002) Dimension, shape, and conformational flexibility of a two domain fungal cellulase in solution probed by small angle X-ray scattering. *J Biol Chem* **277**, 40887–40892.
8 Hammel M, Fierobe HP, Czjzek M, Kurkal V, Smith JC, Bayer EA, Finet S & Receveur-Brechot V (2005) Structural basis of cellulosome efficiency explored by small angle X-ray scattering. *J Biol Chem* **280**, 38562–38568.
9 Violot S, Aghajari N, Czjzek M, Feller G, Sonan GK, Gouet P, Gerday C, Haser R & Receveur-Brechot V (2005) Structure of a full length psychrophilic cellulase from *Pseudoalteromonas haloplanktis* revealed by X-ray diffraction and small angle X-ray scattering. *J Mol Biol* **348**, 1211–1224.
10 von Ossowski I, Eaton JT, Czjzek M, Perkins SJ, Frandsen TP, Schulein M, Panine P, Henrissat B & Receveur-Brechot V (2005) Protein disorder: conformational distribution of the flexible linker in a chimeric double cellulase. *Biophys J* **88**, 2823–2832.
11 D'Amico S, Gerday C & Feller G (2000) Structural similarities and evolutionary relationships in chloride-dependent alpha-amylases. *Gene* **253**, 95–105.
12 Da Lage JL, Feller G & Janecek S (2004) Horizontal gene transfer from Eukarya to bacteria and domain shuffling: the alpha-amylase model. *Cell Mol Life Sci* **61**, 97–109.
13 Tompa P (2002) Intrinsically unstructured proteins. *Trends Biochem Sci* **27**, 527–533.
14 Cornette JL, Cease KB, Margalit H, Spouge JL, Berzofsky JA & DeLisi C (1987) Hydrophobicity scales and computational techniques for detecting amphipathic structures in proteins. *J Mol Biol* **195**, 659–685.
15 Makhatadze GI & Privalov PL (1995) Energetics of protein structure. *Adv Protein Chem* **47**, 307–425.
16 Goldberg RN, Bell D, Tewari YB & McLaughlin MA (1991) Thermodynamics of hydrolysis of oligosaccharides. *Biophys Chem* **40**, 69–76.

17 D'Amico S, Sohier JS & Feller G (2006) Kinetics and energetics of ligand binding determined by microcalorimetry: insights into active site mobility in a psychrophilic alpha-amylase. *J Mol Biol* **358**, 1296–1304.
18 Gurd FR & Rothgeb TM (1979) Motions in proteins. *Adv Protein Chem* **33**, 73–165.
19 Baldwin AJ & Kay LE (2009) NMR spectroscopy brings invisible protein states into focus. *Nat Chem Biol* **5**, 808–814.
20 Hibbert F & Emsley J (1990) Hydrogen bonding and chemical reactivity. *Adv Phys Organ Chem* **26**, 255–379.
21 Weiss MS, Brandl M, Suhnel J, Pal D & Hilgenfeld R (2001) More hydrogen bonds for the (structural) biologist. *Trends Biochem Sci* **26**, 521–523.
22 Rozas I (2007) On the nature of hydrogen bonds: an overview on computational studies and a word about patterns. *Phys Chem Chem Phys* **9**, 2782–2790.
23 Beckham GT, Bomble YJ, Matthews JF, Taylor CB, Resch MG, Yarbrough JM, Decker SR, Bu L, Zhao X, McCabe C *et al*. (2010) The *O*-glycosylated linker from the *Trichoderma reesei* Family 7 cellulase is a flexible, disordered protein. *Biophys J* **99**, 3773–3781.
24 Dunker AK, Silman I, Uversky VN & Sussman JL (2008) Function and structure of inherently disordered proteins. *Curr Opin Struct Biol* **18**, 756–764.
25 Burley SK & Petsko GA (1988) Weakly polar interactions in proteins. *Adv Protein Chem* **39**, 125–189.
26 Maupetit J, Derreumaux P & Tuffery P (2010) A fast method for large-scale de novo peptide and miniprotein structure prediction. *J Comput Chem* **31**, 726–738.
27 Cortese MS, Uversky VN & Dunker AK (2008) Intrinsic disorder in scaffold proteins: getting more from less. *Prog Biophys Mol Biol* **98**, 85–106.
28 Tompa P (2005) The interplay between structure and function in intrinsically unstructured proteins. *FEBS Lett* **579**, 3346–3354.
29 Uversky VN (2002) Natively unfolded proteins: a point where biology waits for physics. *Protein Sci* **11**, 739–756.
30 Uversky VN (2003) Protein folding revisited. A polypeptide chain at the folding–misfolding–nonfolding cross-roads: which way to go? *Cell Mol Life Sci* **60**, 1852–1871.
31 Uversky VN & Dunker AK (2010) Understanding protein non-folding. *Biochim Biophys Acta* **1804**, 1231–1264.
32 Matsushima N, Yoshida H, Kumaki Y, Kamiya M, Tanaka T, Izumi Y & Kretsinger RH (2008) Flexible structures and ligand interactions of tandem repeats consisting of proline, glycine, asparagine, serine, and/or threonine rich oligopeptides in proteins. *Curr Protein Pept Sci* **9**, 591–610.
33 Wise SG & Weiss AS (2009) Tropoelastin. *Int J Biochem Cell Biol* **41**, 494–497.
34 Feller G, D'Amico S, Benotmane AM, Joly F, Van Beeumen J & Gerday C (1998) Characterization of the C-terminal propeptide involved in bacterial wall spanning of alpha-amylase from the psychrophile *Alteromonas haloplanctis*. *J Biol Chem* **273**, 12109–12115.
35 Altschul SF, Madden TL, Schaffer AA, Zhang JH, Zhang Z, Miller W & Lipman DJ (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res* **25**, 3389–3402.
36 Krieger E, Darden T, Nabuurs SB, Finkelstein A & Vriend G (2004) Making optimal use of empirical energy functions: force-field parameterization in crystal space. *Proteins* **57**, 678–683.

Supporting information

The following supplementary material is available:
Fig. S1. Molecular dynamics simulations of the *Corbicula fluminea* linker in a 1 ns simulation.
Fig. S2. Molecular dynamics simulations of the *Daphnia pulex* Amy2 linker in a 15 ns simulation.
Table S1. Accession numbers and genome coordinates of the sequences used in this study.
Table S2. Chi-squared test showing the weight of each amino acid in the compositional bias of the linkers, sorted by decreasing bias.

This supplementary material can be found in the online version of this article.

Please note: As a service to our authors and readers, this journal provides supporting information supplied by the authors. Such materials are peer-reviewed and may be re-organized for online delivery, but are not copy-edited or typeset. Technical support issues arising from supporting information (other than missing files) should be addressed to the authors.