Computational and Structural Biotechnology Journal 14 (2016) 271–282

COMPUTATIONAL ANDSTRUCTURAL BIOTECHNOLOGY JOURNAL

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

Sequence comparison, molecular modeling, and network analysis predict structural diversity in cysteine proteases from the Cape sundew, *Drosera capensis*

Carter T. Butts [a,b,c,]*, Xuhong Zhang [c], John E. Kelly [e], Kyle W. Roskamp [e], Megha H. Unhelkar [e], J. Alfredo Freites [e], Seemal Tahir [e], Rachel W. Martin [e,f,]**

[a] Department of Sociology, UC Irvine, USA
[b] Department of Statistics, UC Irvine, USA
[c] Department of Electrical Engineering and Computer Science, UC Irvine, USA
[e] Department of Chemistry, UC Irvine, USA
[f] Department of Molecular Biology & Biochemistry, UC Irvine, Irvine, CA, 92697 USA

ARTICLE INFO

Article history:
Received 22 March 2016
Received in revised form 11 May 2016
Accepted 17 May 2016
Available online 14 June 2016

Keywords:
Cysteine protease
Carnivorous plant
Protein sequence analysis
Protein structure prediction
Protein structure network
Rosetta
Molecular dynamics
Digestive enzyme
In silico maturation

ABSTRACT

Carnivorous plants represent a so far underexploited reservoir of novel proteases with potentially useful activities. Here we investigate 44 cysteine proteases from the Cape sundew, *Drosera capensis*, predicted from genomic DNA sequences. *D. capensis* has a large number of cysteine protease genes; analysis of their sequences reveals homologs of known plant proteases, some of which are predicted to have novel properties. Many functionally significant sequence and structural features are observed, including targeting signals and occluding loops. Several of the proteases contain a new type of granulin domain. Although active site residues are conserved, the sequence identity of these proteases to known proteins is moderate to low; therefore, comparative modeling with all-atom refinement and subsequent atomistic MD-simulation is used to predict their 3D structures. The structure prediction data, as well as analysis of protein structure networks, suggest multifarious variations on the papain-like cysteine protease structural theme. This *in silico* methodology provides a general framework for investigating a large pool of sequences that are potentially useful for biotechnology applications, enabling informed choices about which proteins to investigate in the laboratory.

1. Introduction

The proteases of carnivorous plants present attractive targets for exploitation in chemical biology and biotechnology contexts. Carnivorous plants, such as *Drosera capensis*, whose prey capture functions take place in the open have been rigorously selected by evolution for the ability to digest large prey over long time periods, without assistance from physical disruption of prey tissue, and in competition with ubiquitous fungi and bacteria. These evolutionary constraints have selected for highly stable enzymes with a different profile of substate specificities and cleavage patterns from those found in animal digestive enzymes. Carnivorous plant digestive enzymes function at pH values ranging from 2–6, depending on the species [1,2]. They also function over a wide range of temperatures; *Drosera* are endemic to every continent except Antarctica and both tropical and temperate species exist. In particular, the pH of *D. capensis* mucilage is around 5 [3], and temperatures in the Western Cape region of South Africa where these plants are found typically range from 5–30 °C.

Characterization of carnivorous plant digestive enzymes could lead to their use in a variety of laboratory and applications contexts, including analytical use in proteomics studies as well as preventing fouling on the surface of medical devices that cannot be treated under harsh conditions. New proteases may also prove useful for cleaving amyloid fibrils, such as those responsible for the transmission of prion diseases or the formation of biofilms by pathogenic bacteria. The characterization of aspartic proteases from the tropical pitcher plants (*Nepenthes* sp.) [4–6], has already led to useful advances in mass spectrometry-based proteomics applications, where the ability to digest proteins using a variety of cut sites is essential for identifying proteins and peptides from complex mixtures. Proteases from plant and animal sources are also important components of pharmaceutical preparations for gluten

* Correspondence to: C. Butts, Calit2, UC Irvine, CA, 92697, UCI.
** Correspondence to: R. Martin, Department of Molecular Biology & Biochemistry, UC Irvine, Irvine, CA, 92697, USA.
E-mail addresses: buttsc@uci.edu (C.T. Butts), rwmartin@uci.edu (R.W. Martin).

http://dx.doi.org/10.1016/j.csbj.2016.05.003

intolerance, arthritis, and pancreatic disease [7]. Characterizing proteases from carnivorous plants has the potential to diversify the toolbox of proteases with different functional properties that are available for these and other applications.

Plant cysteine proteases form a large and diverse family of proteins that perform cellular housekeeping tasks, fulfill defensive functions, and, in carnivorous plants, digest proteins from prey. It is typical for plants to contain many different cysteine protease isoforms; for instance, in the case of tobacco (*Nicotiana tabacum*), more than 60 cysteine protease genes have been identified [8]. Many of the cysteine proteases of interest are classified by the MEROPS database as family C1 [9], a broad class of enzymes including cathepsins and viral proteases as well as plant enzymes that function to deter herbivory. C1 proteases can operate as endopeptidases, dipeptidyl peptidases, and aminopeptidases [10]. In plants, many C1 enzymes are used to degrade proteins in the vacuole, playing many of the same roles as their lysosomal counterparts in animals [11]. They are also found in fruits, particularly unripe ones; this protease activity impedes insect feeding and also serves to cleave endogenous proteins during fruit ripening. Some families of cysteine proteases in plants have been subject to diversifying selection due to a molecular arms race between these plants and their pathogens; as plants produce proteases that suppress fungal growth, fungi evolve inhibitors specific to these proteases, driving the diversification of plant proteases involved in the immune response [12].

The plethora of paralogs found in a typical plant is indicative of the need for a range of different substrate specificities; this is particularly important in the case of carnivorous plants, which must digest prey proteins to their component amino acids. Aspartic proteases have long been implicated in *Nepenthes* pitcher plant digestion [4,13], and more recently the cysteine protease dionain 1 has been confirmed as a major digestive enzyme in the Venus flytrap (*Dionaea muscipula*) [14]. In *D. capensis*, proteins from prey consititute the major nitrogen source for producing new plant tissue [15]. Given that plant carnivory appears to have evolved from defensive systems in general [16], and that the feeding responses are triggered by the same signaling pathway as is implicated in response to wounding [17], one would expect cysteine proteases to play a major role; here we investigate some of the many cysteine protease genes in *D. capensis* with the objective of adding to the portfolio of cleavage activities available for chemical biology applications. The *D. capensis* enzymes are particularly appealing for mass spectrometry-based proteomics applications, due to their ability to operate under relatively mild conditions, i.e. at room temperature and pH 5.

This study focuses on the C1 cysteine proteases from the Cape sundew (*D. capensis*), whose genome we have recently sequenced [18]. Here we use sequence analysis, comparative modeling with all-atom refinement and atomistic molecular dynamics (MD) simulation, and investigation of protein structure networks to identify structurally distinct subgroups of proteins for subsequent expression and biochemical characterization.

C1 cysteine proteases share a common papain-like fold, a property also predicted for the proteins studied here. Despite this conservation of the papain fold and critical active and structural residues, sequence analysis of the *D. capensis* cysteine proteases indicates that they represent a highly diverse group of proteins, some of which appear to be specific to the Droseraceae. In particular, a large cluster of proteases containing dionains 1 and 3 as well as many homologs from *D. capensis* has particular sequence features not seen in papain or other reference enzymes. Finally, a new class of granulin domain-containing cysteine proteases is identified, based on clustering of the granulin domains themselves.

Molecular modeling was performed in order to translate this sequence diversity into predicted structural diversity, which is more informative for guiding future experimental studies. Examination of the predicted enzyme structures potentially suggests diversity that may imply a variety of substrate preferences and cleavage patterns. The relationships between the shape of the substrate-recognition pockets and variation in substrate cleavage activity have been established for other plant cysteine proteases, including the ervatamins [19], the KDEL-tailed CysEP protease from the castor bean [20], and in dionain 1 [21]. The sequence-structure relationships outlined here suggest hypotheses that can be tested in the laboratory, providing a starting point for discovering novel enzymes for use in biotechnology applications. In most cases, the sequences have only weak identity to known plant proteases, making traditional homology modeling of dubious utility. Instead, we use Rosetta [22,23] to perform comparative modeling with all-atom refinement, combining local homology modeling based on short fragments with de novo structure prediction. We then employ atomistic MD simulation of these initial structures in explicit solvent to produce equilibrated structures with corrected active site protonation states; these equilibrated structures serve as the starting point for further analysis.

Quality control was performed using both sequence alignment and inspection of the Rosetta structures; proteins that are missing one of the critical active residues (C 158 or H 292, papain numbering) were discarded, as were some lacking critical disulfide bonds or other structural features necessary for stability. After winnowing out sequences that are unlikely to produce active proteases, 44 potentially active proteases were chosen for further analysis. This methodology allows the development of hypotheses based on predicted 3D structure and activity, in contrast to focusing on the first discovered or most abundantly produced enzymes, enabling selection of the most promising targets for structural and biochemical characterization based on the priorities of technological utility rather than relative importance in the biological context.

2. Methods

2.1. Sequence Alignment and Prediction of Putative Protein Structures

Sequence alignments were performed using ClustalOmega [24], with settings for gap open penalty = 10.0 and gap extension penalty = 0.05, hydrophilic residues = GPSNDQERK, and the BLOSUM weight matrix. The presence and position of a signal sequence flagging the protein for secretion was predicted using the program SignalP 4.1 [25], while other localization sequences were identified using TargetP [26]. Structures were predicted using a three-stage process. First, an initial model was created for each complete sequence using the Robetta server [22]; the Robetta implementation of the Rosetta [23] system generates predictions from sequence information using a combination of comparative modeling and all-atom refinement based on a simplified forcefield. Second, any residues not present in each mature protein were removed, disulfide bonds were identified by homology to known homologs, and the protonation states of active site residues were fixed to their literature values. Finally, in the third phase, each corrected structure was equilibrated in explicit solvent under periodic boundary conditions in NAMD [27] using the CHARMM22 forcefield [28] with the CMAP correction [29] and the TIP3P model for water [30]; following minimization, each structure was simulated at 293 K for 500 ps, with the final conformation retained for subsequent analysis. This process was performed for the 44 protease sequences from *D. capensis*, as well as 10 reference sequences from other organisms (see below); where published structures were available, these were used as the initial starting model (following removal of heteroatoms and protonation using REDUCE [31] as required). For the 5 *D. capensis* sequences with granulin domains (as well as the two references with such a domain), steps (2) and (3) of the above were repeated after removal of the domain and any linking residues. This process provides predicted structures both with and without the domain in question. The PDB files corresponding to the equilibrated structures for all the proteins discussed in this manuscript are available in the Supplementary Information.