Food Hydrocolloids 27 (2012) 503–516

Contents lists available at SciVerse ScienceDirect

Food Hydrocolloids

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

Bioinformatics predicts diverse *Aspergillus* hydrophobins with novel properties

Karen A. Littlejohn[a], Paul Hooley[b], Phil W. Cox[a,*]

[a] School of Chemical Engineering, Edgbaston, Birmingham B15 2TT, UK
[b] School of Applied Sciences, University of Wolverhampton, Wolverhampton WV1 1SB, UK

ARTICLE INFO

Article history:
Received 21 April 2011
Accepted 30 August 2011

Keywords:
Hydrophobin
Bioinformatics
Aspergillus
Protein

ABSTRACT

Hydrophobins, small, amphipathic proteins found only in filamentous fungi, have the ability to form elastic layers at air/water or oil/water interfaces. To identify potentially useful hydrophobins for technological applications, a bioinformatics approach has been employed. The methodology used is not only useful for the identification of hydrophobins but, can be used for the identification of other structurally similar proteins that have little amino-acid homology. The *Aspergillus* comparative database initially revealed 25 officially annotated hydrophobins from a keyword search. A detailed manual annotation using Blast searches and multiple alignments showed common errors in assigning descriptions centred upon difficulties in recognising the relatively small conserved motif of four pairs of cysteines. Searches across the databases gave two protein sequences from the Central *Aspergillus* Data Repository (CADRE) and two from Uniprot which did not match to any found on the *Aspergillus* comparative database. A total of 74 putative hydrophobin sequences were found across the different databases. The number per species ranges from six for *Aspergillus oryzae* to ten for *Aspergillus clavatus*, *Aspergillus fischeri*, *Aspergillus terreus*, *Aspergillus nidulans* and *Aspergillus niger*. Possible examples of proteins which appear to be an intermediate form between Class I and II were found. It is thought that this intermediate form may be of most use for food applications as hopefully it will have amenable solubility and assembly kinetics to be used in standard food production operations. The applications of computer aided bioinformatics for the selection and development of hydrophobins for model production systems are discussed.

1. Introduction

Hydrophobins are small proteins usually between 50 and 125 amino acids in length (Wösten & de Vocht, 2000) although they can be over 400 amino acids in length when including polyhydrophobins (de Vries, Moore, Arntz, Wessels, & Tudzynski, 1999). They are highly surface active proteins containing both hydrophobic and hydrophilic moieties. One of the most distinguishing features of a hydrophobin is its ability to form amphipathic membranes at interfaces between hydrophilic and hydrophobic surfaces. Thus, they will spontaneously form protein monolayers at air/water interfaces via self assembly (Kershaw & Talbot, 1998). Tcheunbou-Magaia, Norton, and Cox (2009), Cox, Aldred, and Russell (2008) and Cox, Cagnol, Russell, and Izzard (2007) have shown the potential application of hydrophobins for the long term stabilisation of food foams and emulsions. As these proteins can be derived from food grade genera this potentially makes them attractive to the food industry due to their long history of consumption. Although this is not the only industrial use for such unique proteins as there has also been interest from other fields such as biocompatibility, biosensors (Hektor & Scholtmeijer, 2005) and drug delivery systems (Akanbi et al., 2010).

To date hydrophobins have been found exclusively in ascomycota and basidiomycota, there have been multiple functions described, for example: sporulation, pathogenicity roles in wall structure and the breaking of air/water interfaces by reducing the surface tension (Wösten & Wessels, 1997). The hydrophobicity of the individual hyphae is changed by the coating of hydrophobins, this then allows them to grow through from the wet substrate into the air. Additionally, they have been associated with controlling evasion of immune responses in pathogenic species (Aimanianda et al., 2009). Genetically engineering fungi to block hydrophobin production limited the number of subsequent hyphae, thus showing that hydrophobins have an essential role in filamentous growth and for dissemination by means of spores (Wösten, 2001).

Traditionally two classes of hydrophobin have been described: Class I and Class II. Their hydropathy patterns and monomer and assembled interface solubility characteristics originally separated and distinguished each class (Wessels, 1994) and then later by their general sequences by both Kershaw and Talbot (1998) and Whiteford

* Corresponding author.
E-mail address: coxpw@bham.ac.uk (P.W. Cox).

0268-005X/$ – see front matter © 2011 Elsevier Ltd. All rights reserved.
doi:10.1016/j.foodhyd.2011.08.018

and Spanu (2002) who adapted the original sequence data from Wessels (1997). Class I assemblies are more insoluble and are more closely associated with the mycelial walls than Class II monomers which are found more freely in the culture media (Szilvay, Nakari-Setala, & Linder, 2006). Class I and Class II hydrophobins initially form comparable surface membranes, however, Class II have a smooth morphology whereas Class I films eventually develop a patterned rodlet layer (Askolin, 2006; Linder, 2009; Wösten, 2001). Class I hydrophobins can be glycosylated and are generally longer at 100–125 amino acids than Class II which are shorter with generally only 50–100 amino-acid residues (Hektor & Scholtmeijer, 2005). Each individual fungal genus can produce a large diversity of hydrophobins, it is generally thought that between two and seven different hydrophobins per species is the norm, although in *Coprinopsis cinerea* there are 34 identified hydrophobins (Stajich et al., 2010), which may still prove to be a more accurate number.

The Protein Data Bank (http://www.pdb.org/pdb/home/home.do) currently (16/6/11) holds eight structures under the keyword "hydrophobin". Tertiary structure prediction is a lengthy enterprise relying upon the ability to produce reliable crystal structures for NMR or X ray diffraction. Bioinformatics tools such as Genthreader (http://bioinf.cs.ucl.ac.uk/psipred/) may only compare primary sequences to previously resolved structures. Hence a very limited number of hydrophobins (EAS, *Neurospora crassa* and HFBI and HFBII, *Trichoderma reesei*) provide tertiary structures for analysis. For example Kwan et al. (2006) reported the three dimensional structure of the Class I hydrophobin EAS (PDB accession 2FMC) with the monomer forming a beta barrel structure containing disordered regions but strictly segregating areas of charged and hydrophobic amino acids to facilitate the amphipathic character. The crucial element is the formation of the four pairs of cysteine bonds to produce the correct folds. Though the EAS and HFBII monomers produce similar folds, Class II hydrophobins like HFBII do not form ordered rodlets either because the Class II members are less hydrophobic or they lack flexibility to fit the monomers together (Kwan et al., 2006).

The different classes of hydrophobin can also be distinguished by their rate of self association and layer maturation. Class I show typically slow rates of self association and are amenable for incorporation into processes for food production or other functional structures. However, their insolubility in aqueous solvents makes them unsuitable for food and potential pharmaceutical applications as well. Conversely, Class II hydrophobins have a greatly improved solubility, suitable for the process industries, but display such a rapid rate of layer formation that their use in standard chemical engineering unit operations, such as large scale mixing, heat and mass transfer and thermodynamic processes (McCabe, Smith, & Harriott, 2004) is extremely limited (Askolin, 2006; Cox et al., 2007; Tcheunbou-Magaia et al., 2009). An ideal situation would be to find a hydrophobin with the combined characteristics of both classes, i.e. good solubility with a controllable rate of layer formation and maturation. The identification of such a molecule from within the fungal genera commercially encountered in food production may be advantageous for use in the FMCG (Fast Moving Consumer Goods) sector, pharmaceutics, biosensors and biocompatibility.

Computer aided bioinformatics now provides a rapid means for identifying hydrophobin encoding genes, but is reliant upon accurate data sets from completed fungal genome projects. Algorithms can then be employed to recognise patterns of similarity from conserved residues and predict the translation of putative genes into hypothetical proteins. Some preliminary predictions of likely characteristics may also be made and realised *in silico* and used as a guide for the design of laboratory models. The best annotated fungal database for comparative genomics focuses upon the genus *Aspergillus*, which includes a number of industrially important species such as *Aspergillus oryzae* and *Aspergillus sojae* (Galagan et al., 2005). One of the major metabolites of *Aspergillus niger* is the commercially produced citric acid (Baker, 2006). Moreover, *Aspergillus* has a high capacity for producing and secreting extracellular enzymes, it therefore plays an important role as the work horse producer organism for industrial enzymes (de Vries et al., 1999). *Aspergillus* species also produce a variety of amylases and proteases in the fermented food industry (MacKenzie, Jeenes, Gou, & Archer, 2000) whilst others cause important human diseases e.g. *Aspergillus fumigatus* causing farmer's lung.

A very recent bioinformatics study by Jensen, Anderson, Pedersen, Frisvad, and Søndergaard (2010) has identified 50 hydrophobin proteins in the *Aspergilli*, 20 of which had previously been unidentified within this group. This work interestingly used the CADRE database (http://www.cadre-genomes.org.uk/) but lacked the cross referencing across all the currently available databases allied with manual annotation to check for errors and duplicates. The work presented here uses the self checking process of manual annotation and the cross checking provided by re-interrogation of the *Aspergillus* comparative database at each stage. The constant re-interrogation of sequence data is required as the identification of hydrophobin encoding genes presents a particular challenge for a bioinformatics approach; this is because beyond the conserved pattern of eight cysteine amino acids which defines the classes, there is very little similarity across the rest of the protein. This makes automated searches somewhat difficult as algorithms for gene identification only work on an overall similarity at the primary amino-acid sequence of a predicted protein and not a "holistic" view of the general protein structure.

Protein sequence data may be found on several databases such as – Interpro (http://www.ebi.ac.uk/interpro/), Uniprot (http://www.ebi.ac.uk/uniprot/), NCBI (http://www.ncbi.nlm.nih.gov/) and CADRE (http://www.cadre-genomes.org.uk/) – each of which may represent different stages of the gene annotation process. Confusingly, different databases may use different accession numbers to refer to even the common proteins. Moreover, the best characterised genomes may also carry large numbers of errors. For example, a manual reappraisal of the *Candida albicans* genome identified around 4% of the 6354 genes as containing sequencing errors or mutations affecting the reading frame and therefore encoded protein structure predictions (Braun et al., 2005). Poptsova and Gogarten (2010) emphasise that even recently completed genomes may have sequencing error rates of up to 0.33%. Indeed, automated gene search programmes may add a further layer of error, for example by missing open reading frames, stop codons or introns altogether (Brachat et al., 2003). In principle the interrogation of genomic data to look for silent hydrophobins is a sound approach and Yang, Deng, Zhang, and Elasri (2006) showed how the application of a simple motif search (SMART/MEME/MAST) could identify several new hydrophobins in existing databases, it should also be noted that the genomic data is not a static entity as the continual process of re-annotation and error correction takes place. Here we demonstrate how simple, focused BLAST searches based on primary structure similarity, coupled with manual annotation can accurately extend the range of known hydrophobins. As said our searches have been looped back to the *Aspergillus* comparative database at each point to verify each protein and maintain accuracy of the data and the definition of the protein class we are looking for.

The traditional division of hydrophobins has produced polarised expectations of rodlet forming Class I molecules that are relatively stable in detergents and ethanol and a second group of more water soluble but less robust Class II counterparts. The suggestion in the present work of the existence of hydrophobins that do not fit easily into either class encourages a focus on molecules that may combine