Gene 543 (2014) 93–100

Contents lists available at ScienceDirect

Gene

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

Functional annotation of putative hypothetical proteins from *Candida dubliniensis*

Kundan Kumar, Amresh Prakash, Munazzah Tasleem, Asimul Islam, Faizan Ahmad, Md. Imtaiyaz Hassan *

Center for Interdisciplinary Research in Basic Sciences, Jamia Millia Islamia, Jamia Nagar, New Delhi 110025, India

ARTICLE INFO

Article history:
Received 23 November 2013
Received in revised form 27 March 2014
Accepted 28 March 2014
Available online 1 April 2014

Keywords:
Candida dubliniensis
Hypothetical protein
Sequence analysis
Functional annotation
Functional genomics

ABSTRACT

An extensive analysis of *C. dubliniensis* proteomics data showed that ~22% protein are conserved hypothetical proteins (HPs) whose function is still not determined precisely. Analysis of gene sequence of HPs provides a platform to establish sequence–function relationships to a more profound understanding of the molecular machinery of organisms at systems level. Here we have combined the latest versions of bioinformatics tools including, protein family, motifs, intrinsic features from the amino acid sequence, sequence–function relationship, pathway analysis, etc. to assign a precise function to HPs for which no any experimental information is available. Our results show that 27 HPs have well defined functions and we categorized them as enzyme, nucleic acid binding, transport protein, etc. Five HPs showed adhesin character that is likely to be essential for the survival of yeast and pathogenesis. We also addressed issues related to the sub-cellular localization and signal peptide identification which provides an idea about its colocalization and function. The outcome of the present study may facilitate better understanding of mechanism of virulence, drug resistance, pathogenesis, adaptability to host, tolerance for host immune response, and drug discovery for treatment of *C. dubliniensis* infections.

1. Introduction

Candida dubliniensis is a germ tube-positive yeast, that acts as an opportunist pathogen (Sebti et al., 2001). Normally, this species of *Candida* is harmless in different body parts but it may become virulent under certain conditions (Sullivan et al., 2005). *Candida* present in most of the body parts including the oral cavity, urine, vagina, lung, feces and sputum, especially in immunocompromised individuals/HIV-infected patient (Sebti et al., 2001). Clinically, 2 to 7% of candidemia caused by *C. dubliniensis,* showed their presence in the gastrointestinal tract. *Candida* showed a wide range of infections from superficial vaginal and oral mucosa to serious systematic infections (Sullivan et al., 2005). These infections are usually countered with the administration of antifungal drugs, however, treatment becomes more difficult with the development of resistance to antifungal agents (Moran et al., 1997). Furthermore, a close phenotypic resemblance of *C. dubliniensis* with *Candida albicans* makes the clinical diagnosis more difficult (O'Connor et al., 2010). Although, *C. dubliniensis* is less pathogenic than *C. albicans*, its ability to produce hyphae and having more survival time pronounce its pathogenicity (Jackson et al., 2009). Hence, this species is a prime target of investigation of fungal infection, especially for the condition of low immunity and frequent development of resistance to antifungal agents (Sullivan and Coleman, 1998).

Recently, the genome of *C. dubliniensis* has been sequenced, and open a new promising channel for extensive research (Jackson et al., 2009). The genome of *C. dubliniensis* is composed of eight chromosomes containing 262288 reads with a total length of 14.6 Mb. An extensive analysis of *C. dubliniensis* genome leads to the identification of 1323 proteins as hypothetical out of 5860 open reading frames (Jackson et al., 2009). HPs are predicted from open reading frame, having no experimental evidence of translation and from their functional annotation (Nimrod et al., 2008). Nearly, half of the proteins in most genomes belong to HPs, and have an absolute importance to complete genomic and proteomic information (Loewenstein et al., 2009). Recent studies suggest many significant roles of HPs because it constitutes a considerable fraction of proteomes and has a reasonable probability that these proteins are novel with uncharacterized biological roles (Adams et al., 2007; Desler et al., 2009; Eisenstein et al., 2000). HPs generally contain low identity compared to other known or annotated proteins (Galperin and Koonin, 2004). However, recent studies showed that a large fraction of genes encoding HPs have strong phylogenetic linkages with known proteins (Mazandu and Mulder, 2012; Shahbaaz et al., 2013). Furthermore, we have been working on the structure based drug design and

Abbreviations: HP, hypothetical protein; BLAST, basic local alignment search tool; PSI-BLAST, position specific iterative basic local alignment search tool; HMMTOP, prediction of transmembrane helices and topology of proteins; TMHMM, membrane protein topology prediction method based on a hidden Markov model; CATH, class, architecture, topology and homology; GRAVY, grand average of hydropathicity; CDD, Conserved Domain Database; SMART, simple modular architecture research tool; PANTHER, Protein ANalysis THrough Evolutionary Relationships; SVM, Support Vector Machine; PP2C, protein phosphatase 2C; SAM, S-adenosyl methionine; DGK, diacylglycerol kinase; CMD, carboxymuconolactone decarboxylase; MFS, major facilitator superfamily.

* Corresponding author.
E-mail address: mihassan@jmi.ac.in (M.I. Hassan).

http://dx.doi.org/10.1016/j.gene.2014.03.060
0378-1119/

searching for a novel therapeutic targets (Hassan et al., 2007a, 2007b; Thakur and Hassan, 2011; Thakur et al., 2013). Therefore, HPs may also serve as markers and a potential drug target for drug design, discovery and screen. A precise annotation of HPs of a particular genome leads to the discovery of new functions, and helps in bringing out a list of additional protein pathways and cascades, thus completing our fragmentary knowledge on biological significance of many novel proteins.

The use of advance bioinformatics tools for sequence analysis is an initial step to identify homology shared between proteins, which could lead to a robust function prediction. Here, we have successfully characterized 43 HPs of *C. dubliniensis* using various computational tools. Preliminary sequence analysis of all 43 HPs was carried using BLAST-P, PSI-BLAST, Pfam and CDD search. Their functions were inferred on the basis of the presence of specific motifs, important region(s) and specific folds, using InterProScan, InterPro, ScanProsite and PFP-FunDSeqE. Other bioinformatics tools such as ProtParam, HMMTOP, TMHMM, SOSUI and CATH have been used precisely to precisely define physicochemical property, subcellular localization and their family. Furthermore, adhesin like proteins, or human pathogenic fungal adhesins were identified with FungalRV. Furthermore, *C. dubliniensis* is one of the major causative agents of infection in immuno-compromised individual, especially in HIV/AIDS patients. Therefore, functional annotation of HPs may lead to identification of novel targets for better treatment and understanding of *C. dubliniensis* infections.

2. Materials and methods

2.1. Sequence retrieval and homology search

Search for HP sequences of *C. dubliniensis* was carried out on UniProt database (http://www.uniprot.org/). The FASTA sequence along with their UniProt ID and primary accession number of 43 HPs were taken separately to perform sequence analysis. UniProt ID of protein has been used to identify the protein sequence to perform sequence analysis. Table 1 provides list of all tools and software that were used for the functional annotation of HPs from *C. dubliniensis*. We used BLAST-P and PSI-BLAST for searching similar sequences with known function (Altschul and Koonin, 1998; Altschul et al., 1997). Top hits were selected and further analyzed using ClustalW to find the alignment of functional residues of protein of known function with the sequence of HPs (Thompson et al., 2002).

2.2. Physicochemical characterization

Theoretical physiochemical parameters such as molecular weight, isoelectric point, aliphatic index, instability index and grand average of hydropathicity (GRAVY) of each protein was carried out on Expasy's ProtParam server (http://web.expasy.org/protparam/). Results of this analysis are listed in Table S1.

2.3. Sub-cellular localization

In order to identify a protein as a drug or vaccine target, sub-cellular localization of the protein is essentially important. Surface membrane protein can be used as a potential vaccine target while cytoplasmic proteins may act as promising drug targets (Vahisalu et al., 2008). We used PSORT II tool (Nakai and Horton, 1999) for the prediction of sub-cellular localization protein. Online tools, TMHMM, SOSUI and HMMTOP were used for predicting the propensity of a protein for being a membrane protein, based on Hidden Markov Model (Chen et al., 2003; Hirokawa et al., 1998). SingnalP 4.1 (Petersen et al., 2011) was used to predict the signal peptide and location of cleavage site in the peptide chain based on neural network method. Results of these predictions are summarized in Table 2.

2.4. Function prediction

In order to assign a precise function to HPs from *C. dubliniensis*, we first analyzed all sequences on Conserved Domain Database (CDD) (Marchler-Bauer et al., 2011), SMART (Letunic et al., 2012), ScanProsite, CATH and PANTHER. CDD includes manually curated domain model based on the tertiary structure of the protein to provide sequence/structure/function relationship in an organized hierarchy of family and superfamily (Marchler-Bauer et al., 2011). SMART compares

Table 1
List of bioinformatics tools and databases used for function prediction.

S. N	Tools	URL	Uses
1. Sequence similarity search tool			
i	BLAST	http://blast.ncbi.nlm.nih.gov/Blast.cgi	To find the similar sequence in the gene database
ii	ClustalW2	https://www.ebi.ac.uk/Tools/msa/clustalw2/	Sequence comparison to compare homologous region
2. Biophysical &chemical characterization			
i	ProtoParam	http://web.expasy.org/protparam/	To calculate various physical and chemical parameters for a given protein sequence
3. Function prediction			
i.	Conserved Domain	http://www.ncbi.nlm.nih.gov/Structure/cdd/wrpsb.cgi	Used to search Conserved Domain in the sequences
ii.	InterProScan	http://www.ebi.ac.uk/Tools/pfa/iprscan/	For functional analysis of the amino acid sequences by finding the specific motif in the sequences
iii.	Interpro	http://www.ebi.ac.uk/interpro/	For functional analysis of proteins on the basis of protein family categorization by predicting domains and important sites
iv	ScanProsite	http://prosite.expasy.org/scanprosite/	Used to scan profile based on domains, motifs and pattern
v	Panther	http://www.pantherdb.org/	Classify proteins on the basis of evolutionary relation and biological process
vi	Pfam	http://pfam.sanger.ac.uk/	Classify protein into family on the basis of multiple sequence alignment
vii	SMART	http://smart.embl-heidelberg.de/	Allow analysis of the domain in the protein sequences
5. Sub-cellular localization of the protein			
i.	SOSUI	http://bp.nuap.nagoya-u.ac.jp/sosui/sosui_submit.html	Used to identify weather the given protein sequences is of soluble protein or of trans-membrane protein
ii.	TMHMM	http://www.cbs.dtu.dk/services/TMHMM/	Used to predict the transmembrane topology of the protein
iii.	Psort II	http://psort.hgc.jp/form2.html	Used to predict sub-cellular localization with a good reliability
iv.	SignalP	http://www.cbs.dtu.dk/services/SignalP/	Predict cleavage site of signal protein
v.	HMMTOP	http://www.enzim.hu/hmmtop/index.php	Predict transmembrane helix and topology of the protein
6. Prediction of fold pattern			
i.	PFP-FunDSeqE	http://www.csbio.sjtu.edu.cn/bioinf/PFP-FunDSeqE/	Used to find the type of protein fold in the protein sequence
7. Virulence prediction			
i.	FungalRV	fungalrv.igib.res.in/query.php	Used in adhesin prediction