ORIGINAL PAPER

Elena Papaleo · Piercarlo Fantucci · Marina Vai Luca De Gioia

Three-dimensional structure of the catalytic domain of the yeast β -(1,3)-glucan transferase Gas1: a molecular modeling investigation

Received: 27 September 2004 / Accepted: 23 June 2005 / Published online: 21 October 2005 © Springer-Verlag 2005

Abstract The three-dimensional (3D) structure of the catalytic domain of Gas1p, a protein belonging to the only family of β -(1,3)-glucan transferases so far identified in yeasts and some pathogenic fungi (family GH-72), has been predicted by combining results derived from threading methods, multiple sequence alignments and secondary-structure predictions. The 3D model has allowed the identification of several residues that are predicted to play a crucial role in structural integrity, substrate recognition and catalysis. In particular, the model of the catalytic domain can be useful for designing site-directed mutagenesis experiments and for developing inhibitors of Gas1p enzymatic activity.

Keywords Protein structure prediction · Computational methods · TIM barrel · Glycosidase

Introduction

Glycolipid anchored surface protein (Gas1p) of *Sac*charomyces cerevisiae is an exocellular glycoprotein endowed with β -(1,3)-glucan transferase activity [1, 2]. This enzyme catalyzes the splitting of an internal β -(1,3)glycosidic linkage in a donor glucan followed by the transfer of the new reducing end to the nonreducing end of an acceptor glucan, with the formation of another β -(1,3)-glycosidic bond in which the anomeric configuration of the linkage is conserved (retaining enzyme). The Gas1p plays a crucial role in the correct incorporation of glucan molecules in the cell wall [1, 2], which, in fungal

Electronic Supplementary Material Supplementary material is available for this article at http://dx.doi.org/10.1007/s00894-005-0025-7

E. Papaleo · P. Fantucci · M. Vai · L. D. Gioia (⊠) Dipartimento di Biotecnologie e Bioscienze, Università di Milano-Bicocca, Piazza della Scienza 2, 20126 Milano, Italy E-mail: luca.degioia@unimib.it Tel.: + 39-2-64483463 pathogens, is involved in interactions with host cells. This feature and the absence of analogous activities in mammalian cells make this enzyme an interesting molecular target for developing new antifungal drugs [3].

The identification of Gas1p homologues in yeast species, fungi and also in several human fungal pathogens has led to the definition of a new family of glycosyl hydrolases (family GH-72) (http://afmb.cnrs~mrs.fr/ ~cazy/CAZY/index.html). Sequence analysis of proteins belonging to family GH-72 revealed a modular organization. In particular, Gas1p has three different domains: an amino terminal catalytic (C) domain of about 300 residues, a cysteine-rich region of about 100 residues (Cys-box), the functional role of which is presently unknown, and a carboxy-terminal serine-rich region of variable length (Ser-box), which is the site for O-mannosylation and is not essential for catalytic activity [2, 4].

Hydrophobic cluster analysis led to the conclusion that the sequence of the C-domain of GH-72 proteins is compatible with a $(\beta/\alpha)_8$ barrel fold, even though it does not share significant sequence similarity to structurally characterized proteins [5]. This observation allowed the GH-72 members to be inserted in the so-called GH-A clan [6], which contains glycoside-hydrolase families characterized by the same global fold of the C-domain.

Other molecular details have been unraveled recently. Site-directed mutagenesis experiments have demonstrated the crucial role of E161 and E262 for the catalytic activity of Gas1p. In addition, it has been shown that C74 is necessary for the correct fold of the protein, whereas C103 and C265 are dispensable [7]. In spite of this evidence, the detailed three-dimensional (3D) structure of the C-domain of Gas1p is still unknown, hindering the full rationalization of experimental data and the design of targeted mutagenesis studies.

Several approaches to predict protein structures from sequences are now available [8–10]. The most reliable computational approach to predict the 3D structure of a protein is homology modeling, which, however, can only be safely used if at least one protein characterized by a significant (>25% sequence identity) similarity to the target protein has a known 3D structure [11]. When dealing with remote homologues (<25% sequence identity), the protein alignment and the subsequent construction of the 3D model are more problematic. In such cases, it has been shown that reliable results may sometimes still be obtained using fold-recognition (threading) methods. Notably, fold-recognition approaches often combine information obtained from many sources [9, 12, 13]. Along these lines, we have combined threading methods, sequence alignments, secondary-structure predictions and biochemical information to predict the 3D structure of the C-domain of Gas1p. The results disclose some key molecular characteristics of the C-domain of Gas1p and identify residues that play an important role in substrate recognition, maintenance of structural integrity and catalysis. Moreover, a map of the putative disulfide bridges of the entire protein is proposed. The 3D model is expected to be a useful tool for designing site-directed mutagenesis experiments and some possible inhibitors for Gas1p enzymatic activity.

Methods

Homologues of Gas1p were searched in the nonredundant database of protein sequences at NCBI, using both Blast and PSI-Blast [14, 15]. Multiple sequence alignments, as well as phylograms, were generated with Clustal W [16], using the Blosum scoring matrix. The gap insertion and extension penalties were set to 10 and 0.05, respectively. In order to highlight conserved regions, the alignment from Clustal W was submitted to ESPript [17].

Secondary structure was predicted by means of JPRED [18] and PSI-PRED [19]. Consensus secondary structures were obtained from comparison among PSI-PRED and individual JPRED results using a 75% stringency threshold.

Six threading methods (sequence-structure fitness) were used to detect remote similarities with proteins of known 3D structure: 3D-PSSM [20], mGen-THREA-DER [21], 123D+ [22], Fugue [23], Topits [24], SAM-T02 [25] and FFAS03 [26]. Only matches characterized by a high confidence level were used as templates to predict the structure of the C-domain of Gas1p. In particular, only proteins characterized by a confidence level higher than 70% or classified as CERTAIN/HIGH were taken from 3D-PSSM/TOPITS and MGen-THREADER, respectively. When the 123D+ and Fugue servers were used, only matches with a z-score equal or higher than 4 and 3.5, respectively, were considered, as suggested by Alexandrov et al. and Shi et al. [22, 23]. Accordingly, only FFAS03 matches characterized by a score lower than -9.5 were selected, as indicated by Rychlewsky et al. [26]. The SAM-T02 already shows only templates characterized by high reliability.

The 3D models were built using the Jackal proteinmodeling software package (http://trantor.bioc.columbia.edu/~xiang/jackal). In particular, the alignments between targets and templates were submitted to the subprogram Nest, which generates a 3D model on the basis of a given alignment, carries out geometry optimization in torsional space to remove clashes between atoms, and finally optimizes the loop regions that are characterized by the presence of gaps in the alignment. In particular, the prediction of loop regions was carried out as followed by Honig and coworkers [27].

As the final step, the models were submitted to molecular-mechanics optimization using the CVFF forcefield [28]. In particular, only the geometry of the protein side-chains was optimized initially (1,000 steps using the steepest-descent algorithm followed by 10,000 steps using the conjugate-gradient algorithm). Then, the optimization was restarted restraining only the α -carbons of the peptide chain. The quality of the final models was evaluated using the Whatif suite of programs [29].

Analysis of the models was carried out using Insight II tools (Accelrys, San Diego, CA, USA) and VMD [30].

Fingerprints for family GH-72 were derived according to the following procedure: (1) amino-acid sequences from the family GH-72 were submitted to PRATT [31], to generate possible patterns common to all probe sequences; (2) if necessary, the PRATT patterns containing functionally important residues were refined manually, using as reference the multi-sequence alignment of the family members; (3) to verify that the patterns selected identify only members of family GH-72, they were submitted to PHI-Blast, a tool developed to evaluate the significance of a specific pattern within a protein [32].

Results and discussions

With the aim of retrieving proteins sharing significant sequence similarity to Gas1p, its protein sequence was submitted to Blast, searching the nonredundant database at NCBI (http://www.ncbi.nlm.nih.gov/). The scan of the database resulted in 40 proteins characterized by high-sequence similarity (*E*-value lower than e^{-27}) to Gas1p. This protein set can be considered an "up to date" representation of family GH-72 [33]. However, it should be noted that among the 40 proteins, 11 correspond to fragments or hypothetical proteins and consequently were not analyzed further.

In order to disclose high-similarity regions, the sequence portions spanning the C-domains were aligned as described in Methods (Fig. 1). The alignment, which is consistent with previously reported data [7], allowed residues that are strictly conserved in all GH-72 members to be highlighted. In particular, 30 residues out of 314 (9.5%) are identical in all C-domains. Among the strictly conserved residues there are seven glycine residues (G159, G197, G243, G264, G290, G291, G304 and Gas1p numbering), six tyrosine residues (Y92, Y113, Y198, Y231, Y294 and Y303), two arginine residues

Gas1n		DDVPA	TEVV.GNKFF	YSNNG	SOFYTR	SVAVOADTA.	NETSG
Phr1p		.KFESSTPP	VEVV.GNKFY	FSNNG	SÕFLIR	GIAYOODAAG:	SVSSGYDADPN
Phr2p		AEDLPA	IEIV.GNKFF	YSNNG	SOFYIK	JIAYQON	NLDSN
CoGas1p		DDLPA	IEIK. GNKFF	FSNNG	SÕFYMK	GIAYOADTA.	NVTGG
CoGas2p			TETV. GNKFF	FSNNG	SOFYMR	TAYOADTA.	NATSG
Endin		EDLPP	TETV GNKER	VSNNG	SOFYTE	TAYOON	NLDSN
End2n		KEESATDD	TEWU GNKEY	VSNNG	TOFILE	TAVOODTSC	SVSNGVDSDPN
SpOrf 3n		SVSP	VHVD GRVE	EVENG	TREFIE	TAVOPNUDD	SDTEG
CdPhr1n		KEESCHDD	TEVI CNKEY	FSNNG	SORITR	TAVOODAAG	RUSSCUDADEN
CdPhr2n		FDIDA	TETU CNEPE	VENNC	COPPTE	TAYOON	NI DEN
		DVM DA	TKTV. GNKFF	DOBOO.		JINIQUN	
Gas2p		EKTPA	INIV.GNNFF	DSESG	EQFFIN	JIAIQLORSE	EELSNANGA
Gersp		TVPN	IQIKAGSKFF	YANNG		SVAYQEDYSG	GAGGTGQSE
CgGasJp	.QLADEVLI	RKWSLQLPT	IEIE.GNKFF	NSETG	EQFFMK	GIAYQQQVDQC	JSELYDG
PcPhrlp		RIQT	IYAY.GNKLF	Y.QNG	TQFFIA	GVAYQPPLRGI	
Gas3p		MLP	IHIK.NYRFI	KPSSATNSES	DNEV FFVK	GVDYQPGGSS(GYDAD
Gel2p		VVP	IEVK.GKDFV	NSKTG	DRFQIL	GVDYQPGGSS(3FTK
NcGel2p		VKP	LTVK.NQFFV	D PND	NIFQIV	GVAYQPGGSA(GYNAA
Gas5p		.SSNSSTPS	IEIK.GNAFF	NSESG	ERFYIR	GVDYQPGGSS	
Gellp	APSIKARD.	DVTP.	ITVK.GNAFF	KGD	ERFYIR	GVDYQPGGSS	
NcGel1p	TPTPTEREC	PHKRADLPA	ITVS.GNAFW	K GK	ERFYIR	GIDYQPGGSS	
SpOrf1p		DSASSA	IKIK.GNAFF	NSDTN	ERFYVR	GVDYQPGGSS	
SpOrf2p		OTHP	IVIK.GNAFF	DSKTN	ERFYIR	GVDYQPGGSS	
CiGel1p	.GVVTARG.	SKTPA	ITIK, GNAFF	OGD	ERFYIR	GLDYOPGGSS	
Gas4p		SVHT	IOIK. DKHFV	DTVTG	. KPFFIK	GVDY0PGGSSI	DVSE
Phr3p		NVHP	ITTH GHYFI	DSTTK	EPFYIK	TDYOPGGSS	VSE
Fa0rf1n		TSSSCEVPE	TVTK CNKER	VSNNG	TOFFMR	TAYOTDCH	DASAS
Bb0rf1n		MCCDDO	TITO NDMAR	N VIC	FDOFME	CTOOFIC	CNOTTOP
					DOBRIN	G.C.T.C.	
NeGelsp	• • • • • • • • •	ALOD	TUMU OOVER	V ENG			
NCG614D		· · · · · · ALQP	TAME . COVLL	gr. ENG	TQFIIK	JIAI QUDAAG	AGGETTT
	40	5 Q	бÖ	7 <u>0</u>	8 <u>0</u>	9 0	100
Gae1n	STVNDE		TDVI.KKI.N	TNVTEWATN	TTINESE	WKALINDADIS	
Phr1n	RKVNDPI	ADRDACKRD	VKVEKESN	TNTLBVVATD	PDKDHEFE	WKTESDAGTY	VARLSEPTV ST
Phr2p	RCEVODI	ANDEUGURD	TOVI FAUDVE			MOMIODACI	
	ATTNDD	ADWDTCCRD	TRAIDERVOID	TRVIATION	TOTOTOTOTO	WNATNDACT	
CgGasip	A DINDEL	ADECCOCRO	TETHQQUA	THE VIEW VIEW AVA	TODDORO	WIZATNDACT W	
CyGaszp	C D D D D D D D D D D D D D D D D D D D	ADI DIGUNO	IFILQQLA		UDOD UDDC		
Epaip Epaip	STFVDPI	ADADNCKRU	IPILEQVD				
Epazp	RKYNDPI	ADADAGKRD	VQYFIDIN	TRTLEVAGID	PDKDBEEG	MAIFSDAGI	
Sportsp	TLFVDPI	SDGDAGSRD	VPYFQELS	VNALRVMAVN	ASLD SAC	MQAFQDAGIM	LSDLAQPYE. AL
Caphrip	RKYTDPI	ADADACKRD	VKYFKESN	TRALEVALD	PERDEEEC.	MKIFSDAGIM.	VADLSEPTVSI
CdPhr2p	DSFVDPI	ANPEHCKRD	IPYLEAVD	TNVIRVYALD	TSQDETEC	MQMLQDAGIM	IADLSQPDESI
Gas2p	FETSYIDAI	ADPKICLRD	IPFLKMLG	. VNTLRVYAID	PTKSHDIC	MEALSAEGMY	/LLDLSEPDISI
Gel3p	ANYV DPI	ADGSKCERD	IPYLLQLR	. TNVIRTYAVN	PSLNHDAC	MQKLSDAGIY	/ITDLASPDESI
CgGas3p	TPYV DPI	ADPHICLRD	LPYLVELG	INTIRVYHID	PSSSHDTC	MKAFSDAGIY	/LIDLAEPEISI
PcPhr1p	. ADHYK NPI	ADPASCERD	LRYFLELG	INTIRVYTVN	PENNEDYC	MRLFENSGIY	/LLDLSEPRNSI
Gas3p	SDTDII	SDPEVCARD	AYAFQQLG	VNTVRIYSLN	PDLNHDKC	MTIFNNAGIYA	AILDVNSGNYGESL
Gel2p	DKDPI	SDPDACLRD	AALMÕR̃LG	VNTIRIVNLS	PSLNHDEC.	ASIFNAAGIVI	ILDVNSPLYGGYL
NcGel 2n	LGKDPI	SDGDIGERD	AATLOTLG	VNTTRVVNT.N	PDVNHDEC	ASTENAAGT	TLOVNSPOVAGST.
Gae5n	NLTOPI	ADASVODRO	VPVLKDLG.	TNTVDVVTVD	NSODUSHC	WKLLOFNCTV	TLOVNTOTS AT
Collin	DINDI	ADADCOVED	TAVEVELC.	TNUTTUWCUT	NCVNNDEG	NUT ADACT	WINNEDZY CT
Gerrp N=Colla	DIADFI	ADADGCARD	UNDERKLO	VND TRUVESVD	NCKWHDDG		
NCGellb	GMTDPI	ADLDNCKRU	VAEFAALG	TUTTRVNMVD	NSKKEDDE	ANLLADAGI	
sporrip	TLVDPI	ADTSICKRU	LPILQGLN		NSANDEG	SALQDAGI	/ILDLATSSNSI
spori2p	SLVDPI	ASRS. CKKD	VEIFKKLG	. INTVRVNQVD	NSADEDKC.	MNALSEAGIN	ILDLNTYRH. SI
CiGellp	KVEDPI	AETETCKRD	IEYFKELG	VNTIRIYTVD	NSKNHDEC	MKALSDAGIM	VLDVNTPKYSI
Gas4p	KQ DPI	SNPDACARD	ILLFQELG	INTVRIVSIN	PDLNIDAC	MTMLAMAGIMI	LIDVNSPLQNQHL
Phr3p	LNDPI	SDPDKCARD	IILFQELG	INTIRIYSIN	AHLNHDKC	MTMLAKAGIYI	JFLDVNSPLPHHHL
EgOrf1p	NKYV DPI	ADFKTCSRD	IPYLQQLR	. TNVIRVYALD	GKK DH TEC	MKALADAGIY	/IADLSEPSLSI
PbOrf1p	NGYK DPI	ANASACKRD	VPLPPMNSV.	TNTIRVYAIN	PKE DH KVCI	MDLLQKAGIY	VADLSEPNLSI
NcGel3p	ADPI	VDTEQCKRD	AALMEKLG	ANSIRVYHVD	PSANHDGC	MDVFEKAGIYI	PLIDLDTFDTYI
NcGel4p	GTFK DPI	ADEAACKRD	VPIMAAAG	TNAIRTYAID	PTADESAC	MKLLDEAGIY	/ISDLSEPSTSI
	110	12	0 13	14	0	1 5 0	160
	110		·	· · · · ·	-	130	
Gaslp				and the second sec		130	
Phr1p	NRDDPTV	TVDLFNSYK	TVVDTFANYI	NVLGFFAGN®	Υ Τ Ν ΝΥΤΝ.	. TDASA FVKA	IRDVRQYISDKNY
m10	NRDDPTV NRNNPEV	TVDLFNSYK NLDLYKRYT	T VVD TFANYI K VID K M QEYS	PNVLGFFAGND SNVLGFFAGND	V T N NYTN. VT N NRSN.	. TDASA FVKA . TDASA FVKA	AIRDVRQYISDKNY AIRDMKKYIKESDY
rnr2p	NRDDPTV NRNNPEV NRDDPSV	VTVDLFNSYK VNLDLYKRYT VDLDLFERYT	TVVDTFANYT KVIDKMQEYS SVVDLFHNYT	PNVLGFFAGNE SNVLGFFAGNE PNILGFFAGNE	VTNNYTN. VTNNRSN. VTNKKSN.	. TDASAFVKA . TDASAFVKA . TDASAFVKA	AIRDVRQYISDKNY AIRDMKKYIKESDY AIRDTKAYIKSKGY
rnr2p CgGas1p	NRDDPTV NRNNPEV NRDDPSV NRKSPSV	VTVDLFNSYK VNLDLYKRYT VDLDLFERYT VDLEIFDRYK	TVVDTFANYT KVIDKMQEYS SVVDLFHNYT SVVDMFANYS	PNVLGFFAGNE Snvlgffagne Pnilgffagne Snvlgffagne	VTNNYTN. VTNNRSN. VTNKKSN. VTNDATN.	. TDASAFVKA . TDASAFVKA . TDASAFVKA . TDASAFVKA . TDASAFVKA	AIRDVRQYISDKNY AIRDMKKYIKESDY AIRDTKAYIKSKGY AIRDTKSYIKEKGY
rnr2p CgGas1p CgGas2p	NRDDPTV NRNNPEV NRDDPSV NRKSPSV NRDSPSV	ITVDLFNSYK INLDLYKRYT IDLDLFERYT IDLEIFDRYK IDLELYERYT	TVVDTFANYT KVIDKMQEYS SVVDLFHNYT SVVDMFANYS SVVDMFANYS	PNVLGFFAGNE NVLGFFAGNE NVLGFFAGNE NVLGFFAGNE NVLGFFAGNE	VTNNYTN. VTNNRSN. VTNKKSN. VTNDATN. VTNDATN.	. TDA SAFVKA . TDA SAFVKA . TDA SAFVKA . TDA SAFVKA . TDA SAFVKA . TDA SAFVKA	A IRDVROYISDKNY A IRDMKKYIKESDY A IRDKKYKKSKGY A IRDTKSYIKEKGY A VRDTKSYIKEKGY
rnr2p CgGas1p CgGas2p Epd1p	NRDDPTV NRNNPEV NRDDPSV NRKSPSV NRDSPSV NRDSPSV	YTVDLFNSYK YNLDLYKRYT YDLDLFERYT YDLEIFDRYK YDLELYERYT YNLDLFERYT	TVVDTFANYT KVIDKMQEYS SVVDLFHNYT SVVDMFANYS SVVDMFANYS SVVDMFANYS SVVDKFHNYT	NVLGFFAGNE NVLGFFAGNE NILGFFAGNE NVLGFFAGNE NVLGFFAGNE	VTNNYTN. VTNNRSN. VTNKKSN. VTNDATN. VTNNSTN. VTNNSTN.	TDASAFVKA TDASAFVKA TDASAFVKA TDASAFVKA TDASAFVKA TDASAFVKA	A IRDVRQYISDKNY AIRDMKKYIKESDY AIRDTKAYIKSKGY AIRDTKQYIKEKGY AVRDTKQYIKSKGY AIRDTKAYIKAKGY
rnr2p CgGas1p CgGas2p Epd1p Epd2p	NRDDPTV NRNNPEV NRDDPSV NRKSPSV NRDSPSV NRDSPSV NRNDPQV	TVDLFNSYK NLDLYKRYT VDLDLFERYT VDLEIFDRYK VDLELYERYT NLDLFERYT NLDLYERYT	TVVDTFANYT KVIDKMQEYS SVVDLFHNYT SVVDMFANYS SVVDMFANYS SVVDKFHNYT KVIDNMQEYS	NVLGFFAGNE SNVLGFFAGNE NVLGFFAGNE SNVLGFFAGNE SNVLGFFAGNE SNVLGFFAGNE	VTNNYTN. VTNNRSN. VTNKKSN. VTNDATN. VTNDATN. VTNNSTN. VTNNVSN. VTNNRTN.	. TDASAFVKA . TDASAFVKA . TDASAFVKA . TDASAFVKA . TDASAFVKA . TDASAFVKA . TDASPFVKA	AIRDVRQYISDKNY AIRDMKKYIKESDY AIRDTKAYIKESDY AIRDTKAYIKEKGY AIRDTKQYIKEKGY AIRDTKQYIKAKGY AIRDTKAYI
rnr2p CgGas1p CgGas2p Epd1p Epd2p SpOrf3p	NRDDPTV NRNNPEV NRDDPSV NRKSPSV NRDSPSV NRNDPQV NRINPEV SSSDPTV	VTVDLFNSYK VDLDLYKRYT VDLDLFERYT VDLEIFDRYK VDLELYERYT VNLDLFERYT VNLDLYERYT VTVDLFSRYT	TVVDTFANYT KVIDKMQEYS SVVDLFHNYT SVVDMFANYS SVVDMFANYS SVVDKFHNYT KVIDNMQEYS EVVDSLAPYE	NVLGFFAGNE NVLGFFAGNE NVLGFFAGNE NVLGFFAGNE NVLGFFAGNE NVLGFFAGNE NVLGFFAGNE	VTNNYTN. VTNNRSN. VTNKKSN. VTNDATN. VTNNSTN. VTNNSTN. VTNNRTN. VTNNRTN.	TDASAFVKA TDASAFVKA TDASAFVKA TDASAFVKA TDASAFVKA TDASAFVKA TDASAFVKA TDASAFVKA	AIRDVRQYISDKNY AIRDMKKYIKESDY AIRDTKAYIKESDY AIRDTKAYIKSKGY AIRDTKAYIKSKGY AIRDTKAYIKAKGY AVRDTKYIKAKGY AVRDVKSYIKSSGY
rnr2p CgGas1p CgGas2p Epd1p Epd2p SpOrf3p CdPhr1p	NRDDPTV NRNNPEV NRDDPSV NRKSPSV NRDSPSV NRDSPSV NRNDPQV NRINPEV SSDPTV NRNSPEV	VTVDLFNSYK VDLDLYKRYT VDLDLFERYT VDLELYERYT VNLDLFERYT VNLDLYERYT VNLDLYERYT VNLDLYERYT	TVVDTFANYT KVIDKMQEYS SVVDLFHNYT SVVDMFANYS SVVDMFANYS SVVDKFHNYT KVIDNMQEYS EVVDSLAPYI KVVDNMOFYS	INVLGFFAGNE INVLGFFAGNE INVLGFFAGNE INVLGFFAGNE INVLGFFAGNE INVLGFFAGNE INVLGFFAGNE INVLGFFAGNE	VTNNYTN. VTNNRSN. VTNKKSN. VTNDATN. VTNNSTN. VTNNVSN. VTNNRTN. VIQNNTN. VTNNRSN.	TDASAFVKA TDASAFVKA TDASAFVKA TDASAFVKA TDASAFVKA TDASAFVKA TDASAFVKA TDASPFVKA TDASPFVKA	AIRDVRQYISDKNY AIRDMKKYIKESDY AIRDTKAYIKSKGY AIRDTKAYIKSKGY AIRDTKQYIKSKGY AIRDTKQYIKSKGY AIRDTKAYIKAKGY AVRDMKKYIKDNDY AVRDVKSYIKSSNY
rnr2p CgGas1p CgGas2p Epd1p Epd2p SpOrf3p CdPhr1p CdPhr2p	NRDDPT NRNNPEV NRDDPSV NRKSPSV NRDSPSV NRNDPQV NRINPEV SSSDPTU NRNSPEV NRNSPEV	VTVDLFNSYK VDLDLYKRYT VDLDEFRYT VDLELYERYT VDLDLFERYT VNLDLYERYT VNLDLYERYT VNLDLYERYT VNLDLYERYT	TVVDTFANYT KVIDKMQEYS SVVDLFHNYT SVVDMFANYS SVVDMFANYS SVVDKFHNYS KVIDNMQEYS SVVDSLAPYL KVVDNMQEYS SVVDLFHNY	NVLGFFAGNE NVLGFFAGNE NVLGFFAGNE NVLGFFAGNE NVLGFFAGNE NVLGFFAGNE NVLGFFAGNE NVLGFFAGNE NVLGFFAGNE	VTNNYTN. VTNNRSN. VTNKKSN. VTNNSTN. VTNNSTN. VTNNYSN. VTNNRTN. VINNRTN. VTNNRTN. VTNNRSN	TDASAFVKA TDASAFVKA TDASAFVKA TDASAFVKA TDASAFVKA TDASAFVKA TDASAFVKA TDASPFVKA TDASPFVKA TDASAFVKA	A IRDVR(YISDKNY A IRDMKKYIKESDY A IRDMKKYIKESDY A IRDTKAYIKSKGY A VRDTKQYIKSKGY A VRDKKYIKSKGY A VRDKKYIKSSGY A VRDVKSYIKSSGY A IRDTKAYIKSSY
rnr2p CgGas1p CgGas2p Epd1p Epd2p SpOrf3p CdPhr1p CdPhr2p Gas2p	NRDDPT NRNNPEV NRDPSV NRKSPSV NRKSPSV NRDP.QV SSSDP.TV NRNP.EV NRNP.EV NRNDP.SV NRNDP.SV NRNDP.SV	VUDLFNSYK VDLDLYRRYT VDLDLFERYT VDLELYERYT VDLDLFERYT VDLDLFERYT VDLDLYERYT VDLDLYERYT VDLDLYERYT VDLDLFERYT	TVVDTFANYT KVIDKMQEYS SVVDLFHNYT SVVDMFANYS SVVDMFANYS SVVDKFHNYT KVIDNMQEYS EVVDSLAPYL KVVDNMQEYS SVVDLFHNYT	INVLGFFAGNE INVLGFFAGNE INVLGFFAGNE INVLGFFAGNE INVLGFFAGNE INVLGFFAGNE INVLGFFAGNE INVLGFFAGNE INVLGFFAGNE	VTNNYTN. VTNNRSN. VTNKKSN. VTNDATN. VTNNSTN. VTNNYSN. VTNNRTN. VIQNNTN. VTNNRSN. VTNNRSN.	TDASAFVKA TDASAFVKA TDASAFVKA TDASAFVKA TDASAFVKA TDASAFVKA TDASAFVKA TDASPFVKA TDASPFVKA TDASAFVKA	I RDVROYISDKNY AIRDRKKYIKSSDY AIRDTKAYIKSKGY URDTKOYIKSKGY URDTKOYIKSKGY AIRDTKYIKSKGY AIRDTKAYIKSGY VRDMKKYIKSDNY VRDVKSYIKSSGY URDKKYIKSSGY IRDTKAYIKSKGY
rnr2p CgGas1p CgGas2p Epd1p Epd2p SpOrf3p CdPhr1p CdPhr2p Gas2p Gal35	NRDDPTW NRDDPSV NRDDPSV NRKSPSV NRKSPSV NRNDPQV NRINPEV SSSDPTV NRNDPEV NRNDPSV NRNDPSV NRNDPSV	VUDLFNSYK INLDLYRRYT VDLDLFERYT VDLEIFDRYK INLDLFERYT INLDLYERYT VUDLFSRYT VDLDLFERYT VDLDLFERYT VDLDLFERYT	TVVDTFANYT KVIDKMQEYS SVVDLFHNYT SVVDMFANYS SVVDMFANYS SVVDKFHNYT KVIDNMQEYS SVVDKHNYT SVIDAMSSF SVVDLFHNYT SVIDAMSSF	NVLGFFAGNE NVLGFFAGNE NVLGFFAGNE NVLGFFAGNE NVLGFFAGNE NVLGFFAGNE NVLGFFAGNE NVLGFFAGNE NILGFFAGNE NILGFFAGNE	VTNNYTN. VTNNRSN. VTNKKSN. VTNNATN. VTNNSTN. VTNNVSN. VTNNYSN. VTNNRSN. VTNKKSN. VTNKKSN. VTNKKSN.	TDASAFVKA TDASAFVKA TDASAFVKA TDASAFVKA TDASAFVKA TDASAFVKA TDASAFVKA TDASPFVKA TDASPFVKA TDASAFVKA TDASAFVKA	I R DVROYISDKNY I R DMKKYIK SEKSDY I R DTKSYIK SEKSDY I R DTKSYIK SEKSY I R DTKSYIK SEKGY I R DTKSYIK SEKGY I R DTKSYIK SEKSYI V R DMKKYIK SEKSNY I R DMKKYIK SEKSNY I R DAKEYISSSY I R DAKEYISSSY
rnr2p CgGas1p CgGas2p Epd1p Epd2p SpOrf3p CdPhr1p CdPhr2p Gas2p Ge13p	NRDDPTW NRNNPSV NRDSPSV NRKSPSV NRNDPQV NRINPEV SSSDPTW NRINSPEV NRNSPSV NRENPSV NRENPSV	VUDLFNSYK MULDLYKRYT VDLEIFDRYK NLDLFERYT NLDLFERYT NLDLYERYT NLDLYERYT NLDLYERYT VDLDLFERYT VDUFFERYK	TVVDTFANYT KVIDKMQEYS SVVDMFANYS SVVDMFANYS SVVDKFANYS SVVDKFHNYT KVIDNMQEYS SVVDLAPYL KVVDLAPYL SVVDLFHNYT SVIDANQSF SVIDAFQKYL	NVLGFFAGNE NVLGFFAGNE NVLGFFAGNE NVLGFFAGNE NVLGFFAGNE NVLGFFAGNE NVLGFFAGNE NLGFFAGNE NLGFFAGNE NLGFFAGNE NLGFFAGNE	VTNNYTN. VTTNNRSN. VTNNSTN. VTNNSTN. VTNNSTN. VTNNTN. VTNNTN. VTNNRTN. VTNNRSN. VTNNRSN. VTNNRSN. VTNNRSN.	TDASAFVKA TDASAFVKA TDASAFVKA TDASAFVKA TDASAFVKA TDASAFVKA TDASPFVKA TDASPFVKA TDASPFVKA TDASPFVKA SAGAAFVKA	I RDVROYISDKNY I RDVROYIKSSDY I RDTRISYIKSKGY V RDTRISYIKSKGY V RDTRISYIKSKGY V RDTRISYIKSKGY V RDVRSYIKSSGY A RDVRSYIKSSGY A RDVRSYIKSKSK N RDKAYIKSK N RDKAYIKSKGY
rnr2p CgGas1p CgGas2p Epd1p Epd2p SpOrf3p CdPhr1p CdPhr2p Gas2p Gel3p CgGas3p	NRDDDP. TW NRDDDP. SV NRDDP. SV NRDSP. SV NRDSP. SV NRDSP. SV NRDDP. QV NRINP. EV NRNSP. EV NRNSP. EV NRNSP. SV NRENP. SV VRNNP. SV VRNNP. SV	VUDLFNSYK INLDLYRRYT VDLEIFDRYK VDLEIFDRYK VDLELYERYT VNLDLFERYT VNLDLYERYT VDLDLFERYT VDLDLFERYT VDLDLFERYT VDVHIFERYK VDVHIFERYK	TVVDTFANYT KVIDKMQEYS SVVDMFANYS SVVDMFANYS SVVDKFHNYT KVIDNMQEYS SVVDKFHNYT KVIDNMQEYS SVVDLFHNYT SVIDAMSSFF SVIDAMSSF SVIDAMS	NVLGFFACNE NVLGFFACNE NVLGFFACNE NVLGFFACNE NVLGFFACNE NVLGFFACNE NVLGFFACNE NVLGFFACNE NVLGFFACNE NVLGFFACNE NLGFFACNE NLGFFACNE NVLGFFACNE	VTNNYTN. VTNNRSN. VTNKSS. VTNDATN. VTNNSTN. VTNNYSN. VTNNRTN. VTNNRSN. VTNNRSN. VTNKSN. VTNNRSN. VTNDHTN. VTNDHTN.	TDASAFVKA TDASAFVKA TDASAFVKA TDASAFVKA TDASAFVKA TDASAFVKA TDASAFVKA TDASPFVKA TDASPFVKA TDASPFVKA TDASAFVKA SAGAAFVKA TDASPFVKA	I R DVROYISDKNY I R DVROYISDKNY I R DRKKYIKSKOY I R DTKSYIKSKOY I R DTKSYIKSKOY I R DTKSYIKSKOY I R DTKSYIKSKOY I R DTKSYIKSKOY I R DKSYISSS I R DKSYISS I R DKSY I R DKSYISS I R DKSY I
rnr2p CgGas1p Epd1p Epd2p SpOrf3p CdPhr1p CdPhr2p Gas2p Gas2p Gas3p CgGas3p PcPhr1p	NRDDP .TV NRNNP.SV NRDDP .SV NRKSP .SV NRKSP .SV NRNDP .QV NRINP.EV SSSDP .TV NRNSP .SV NRNDP .SV NRNDP .SV NRENP .SV TSNSP .SV VRNNP .SV ISTDP .SV	VUDLFNSYK NULDLYKRYT VDLDIFERYT VDLEIFDRYK NULDLFERYT NULDLFERYT VDLDLYERYT VDLDLYERYT VDLDLFERYT VDLDLFERYT VDLDFERYT VDULFERYK VTVDQYARYT VDVKLFWRYS	TVVDTFANYT KVIDKMQEYS SVVDMFANYS SVVDMFANYS SVVDMFANYS SVVDKFHNYT KVIDNMQEYS SVVDLFHNYT SVIDAMSSF SVIDAMSSF SVIDAMSSF SVIDAMSKF	NVLGFFAGNE NVLGFFAGNE NVLGFFAGNE NVLGFFAGNE NVLGFFAGNE NVLGFFAGNE NVLGFFAGNE NLGFFAGNE NLGFFAGNE NLGFFAGNE NLGFFAGNE NLLGFFAGNE	VTNNYTN. VTNNRSN. VTNNRSN. VTNNATN. VTNNSTN. VTNNSTN. VTNNRTN. VTNNRSN. VTNNRSN. VTNUKSN. VTNUHTN. VTNUKSN. VTNUHTN. VVNQANQ. VTNUKYN. VINDTEN.	TDASAFVKA TDASAFVKA TDASAFVKA TDASAFVKA TDASAFVKA TDASAFVKA TDASPFVKA TDASPFVKA TDASPFVKA TDASPFVKA TFASPFVKA SAGAAFVKA	I RDVROYISDKNY IRDMKKYIKSSDY IRDTKSYIKSSGY VRDTKSYIKSSKY VRDTKSYIKSSKY VRDTKSYIKSSS VRDMKKYIKSSGY A IRDVKSYIKSSGY A IRDKKSYIKSSGY A IRDKKAYIKSNH A RDMKKYIKSNH IRDAKSYISTSSGY A IRDKKSYIKSSGY
rnr2p CgGas1p CgGas2p Epd1p Epd2p Sporf3p CdPhr1p CdPhr2p Gas2p Gel3p CgGas3p PcPhr1p Gas3p	NRDDP.TW NRDDP.SV NRDDP.SV NRKSP.SV NRDSP.SV NRDSP.SV NRDSP.SV SSDP.TW SSDP.TV NRDSP.SV VRNNP.SV VRNNP.SV VRNNP.SV VRNNP.SV VRNNP.SV	VUDLFNSYK RYT WILDLYRRYT IDLELFERYT IDLELYRRYT WILDLFERYT WILDLFERYT IDLLFERYT IDLDLFERYT IDLDLFERYT IDLDLFERYT IDVDLARYT IDVVLSRYRY IDVVLSRYRY INVRLFRRYS INVRLFRRYS	TVVDTFANYT KVIDKMQEYS SVVDLFHNYT SVVDMFANYS SVVDKFHNYT KVIDNMQEYS EVVDSLAPYT SVIDAFQKYT SVIDAFGKYT DVVDAMHFYN KVIDNMHKYF	NVLGFFAGNE NVLGFFAGNE NVLGFFAGNE NVLGFFAGNE NVLGFFAGNE NVLGFFAGNE NVLGFFAGNE NVLGFFAGNE NLGFFAGNE NLGFFAGNE NLGFFAGNE NLGFFAGNE NVLGFFAGNE NLLGFFAGNE NVLGFFAGNE	VTNNYTN. VTNNRSN. VTNNRSN. VTNNSTN. VTNNSTN. VTNNSTN. VTNNRTN. VTNNRSN. VTNNRSN. VTNNRSN. VTNNRSN. VTNNRSN. VTNDHTN. VTNDHTN. VTNDHTN. VINDKYN. VILDTEN. VILDTEN.	TDASAFVKA TDASAFVKA TDASAFVKA TDASAFVKA TDASAFVKA TDASAFVKA TDASAFVKA TDASAFVKA TDASAFVKA TDASPFVKA TASPFVKA TASPFVKA TASAFVKA THSAAYVKA	I R D VROYIS SDKNY I R D MKKY IK SSDY I R D TKSY IK SSGY VR D TKSY IK SKGY VR D TKSY IK SKGY VR D TKSY IK SKGY VR D MKKY IK SSNY I R D MKKY IK SSNY I R D KSY ISSSGY I R D KSY ISSSGY VR D VKSY MOOKGY VR D VKSY MOOKGY
rnr2p CgGas1p CgGas2p Epd1p Epd2p SpOrf3p CdPhr1p Gas2p Gel3p CgGas3p PcPhr1p Gas3p Gel2p	NRDDP.TW NRNNP.EV NRDSP.SV NRDSP.SV NRDSP.SV NRDDP.OV NRDP.OV NRDP.SV NRNDP.SV NRNDP.SV NRNP.SV SSDP.TV NRNSP.EV NRENP.SV NRENP.SV NRENP.SV NRENP.SV NRENP.SV NRENP.SV NRENP.SV NRENP.SV NRENP.SV NRENP.SV NRENP.SV	TYDLENSYK NILDLYKRYT IDLEIFDRYK IDLEFRRYT INLLFRRYT INLLFRRYT INLLFRYT IDULFSRYT IDULFRYT IDULFRYT IDULFRYT IDULFRYK IDVFVKVRYR IDVFVKVRYR IDSLYLSRVF	TVVDTFANT KVIDKMQEYS SVVDLFHNYT SVVDMFANYS SVVDKFHNYT KVIDNMQEYS SVVDSLAPYL SVVDSLAPYL SVIDAFQKYL SVIDAFQKYL KVIDNMHKYF KVIDNMHKYF KVIDAFKNYF	NVLGFFACNE NVLGFFACNE NVLGFFACNE NVLGFFACNE NVLGFFACNE NVLGFFACNE NVLGFFACNE NLGFFACNE NLGFFACNE NLGFFACNE NVLGFFACNE NVLGFFACNE NVLGFFACNE NVLGFFACNE NLGFFACNE NLGFFACNE NLGFFACNE NLGFFACNE NLGFFACNE	VTNNYTN. VTNKKSN. VTNKKSN. VTNNSTN. VTNNSTN. VTNNVSN. VTNNKSN. VTNKKSN. VTNKKSN. VTNKKSN. VTNKKSN. VTNLKSN. VILOTEN. VINDQSDY VINDQSDY	TDASAPVKA TDASAPVKA TDASAPVKA TDASAPVKA TDASAPVKA TDASAPVKA TDASAPVKA TDASAPVKA TDASAPVKA TDASPPVKA TASAPVKA SAGAPVKA TASAPVKA TASAPVKA TASAVVKA	I RDVRQYISDKNY IRDVRQYIKSKSDY IRDTKSYIKSKSDY IRDTKSYIKSKSKY VRDTKQYIKSKGY VRDTKQYIKSKGY VRDVKSYIKSSGY A IRDAKSYIKSSGY A IRDAKSYIKSSGY A IRDKKSYIKSSGY A IRDKKSYIKSSGY A IRDKKSYIKSKGY I IRDAKSYIKSKGY I IRDKKSYMQCKGY V / RDVKSYMCSKGY I IRDAKSYIKSKGY
rnr2p CgGas1p CgGas2p Epd1p Epd2p SpOrf3p CdPhr1p CdPhr2p Gas2p Gel3p CgGas3p PcPhr1p Gas3p Gel2p NcGel2p	NRDDP T NRNNP S NRDDP S NRDSP S NRDSP S NRDSP S NRDSP S NRDSP S SSSDP T NRNSP S NRNSP S SSSDP T NRNSP S NRNSP S NRNNP S NRANNP S NRANNP S NRANSP N NRANP N N N	TYDLENSYK NLLLYRXYT IDLELFRYT IDLEFRYT NLLLFRYT NLLLFRYT NLLLFRYT TYDLFRYT IDLLFRYT TYDLFRYT DLLLFRYT IDLLFRYT IDLLFRYT IDLFRYK IDSLYLSRYF NLVYRYS IDSLYLSRYF	TVUDTFANT SVUDLFHNT SVUDLFHNT SVUDMFANTS SVUDMFANTS SVUDMFANTS SVUDMFANTS SVUDLFHNT SVUDAMFANTS SVUDLFHNT SVUDLFHNT SVUDAMFST SVUDLFHNT SVUDAMFST RUDAMFST SVUDAMFST SVUDAMFST SVUDAMFST SVUDAMFST SVUDAMFST SVUDAMFST SVUDAMFST SVUDAMFST SVUDAMFST SVUDAMFST	NVLGFFAGNE NVLGFFAGNE NVLGFFAGNE NVLGFFAGNE NVLGFFAGNE NVLGFFAGNE NVLGFFAGNE NVLGFFAGNE NVLGFFAGNE NVLGFFAGNE NVLGFFAGNE NVLGFFAGNE NVLGFFAGNE NVLGFFAGNE NVLGFFAGNE NVLGFFAGNE NVLGFFAGNE	VTNNYTN. VTNKKSN. VTNAKSN. VTNNSTN. VTNNSTN. VTNNSTN. VTNNRTN. VTNNRSN. VTNKKSN. VTNKKSN. VTNDHTN. VVNQANQ. VTNDLYN. VILDTEN. VILDZSJ. VINDQSS.	TDASAPVKA TDASAPVKA TDASAPVKA TDASAPVKA TDASAPVKA TDASAPVKA TDASAPVKA TDASAPVKA TDASAPVKA TDASAPVKA TDASAPVKA TDASAPVKA TDASAPVKA THSAAVVKA	I R D VROYISDKNY I R D MKKYIK SEKSDY I R D TKSYIK SEGY I R D TKSYIK SEGY I R D TKSYIK SEGY VR D TKSYIK SEGY VR D MKKYIK SENDY I R D KSYIK SES I R D KSYIK SES I R D KSYKOK VR D VKSYKOK I R D KSYKOK I R D KSKOK I R D
rnr2p CgGas1p CgGas2p Epd1p Epd2p SpOrf3p CdPhr1p CdPhr2p Gel3p CgGas2p Gel3p CgGas3p PcPhr1p Gas3p Gel2p NcGel2p Gas5p	NRDDP.TW NRNNP.ED NRCDP.SV NRCSP.SV NRDSP.SV NRDSP.SV NRDDP.OV SSSDP.TV SSSDP.TV NRNDP.SV TSNSP.VV VRNNP.SV TSNSP.VV VRNNP.SV NRADPSGT SNRADPSGT QNYNPWEST QNYNPWEST	TYDLLFNSYK NLLLYKRYT DDLEFRYK NLLYFRYT INLLFFRYK NLLLYFRYT TYDLFSRYT TYDLFSRYT TYDLFSRYT TYDLFSRYK TYDLYFRYK INVRLFRYK INVRLFWRYS DSLYLSRYF NDVFFRYK	TVUDEFANT KVIDKAQ EXG SVUDAFANT SVUDAFANT SVUDKFANT SVUDKFANT SVUDKFANT SVUDKFANT SVUDKFANT SVUDAG SVUDAFANT SVIDAAFQKT SVIDAAFQKT SVIDAAFQKT SVIDAAFQKT SVIDAAFQKT SVIDAAFQKT SVIDAAFQKT SVIDAAFKNT SVIDAAFKNT SVIDAAFKNT SVIDAAFKNT SVIDAAFKNT SVIDAAFKNT SVIDAAFKNT	NVLCFFACND VLCFFACND VLCFFACND NVLCFFACND NVLCFFACND NVLCFFACND NLCFFACND NLCFFACND NLCFFACND NLCFFACND NLCFFACND NLCFFACND NLCFFFACND NLLCFFFACND NLLCFFFACND NLLFFFACND NTLAFFACND	VTNNYTN. VTNNRSN. VTNNKSN. VTNNSTN. VTNNSTN. VTNNSTN. VTNNRTN. VTNNRSN. VTNNRSN. VTNQANQ. VTNQANQ. VTNQANQ. VINQANQ. VINQASS. VINQSSS. VINQSSS.	TDASAFVKA TDASAFVKA TDASAFVKA TDASAFVKA TDASAFVKA TDASAFVKA TDASFVKA TDASFVKA TDASFVKA TDASFVKA TFASFVKA TFASFVKA TDASFVKA TASAVKA KIDPVITA	I R D VROYIS SCH NY I R D MKKY IKES SCY I R D TKSY IKES SCY V R D KKY IKES SCY I R D TKSY IKES SCY I R D TKSY I R D TSY I K D TSY
rnr2p CgGas1p CgGas2p Epd1p Epd2p SpOrf3p CdPhr1p CdPhr2p Gas2p Gel3p CgGas3p FcPhr1p Gas3p Gel2p NcGel2p Gas5p Gas1p	NRDDP.TW NRNNP.ED NRDSP.SV NRDSP.SV NRDSP.SV NRDSP.SV NRDP.QV SSSDP.TW NRDP.SV NRDP.SV NRDP.SV NRDP.SV NRDP.SV NRDP.SV NRDP.SV NRDP.SV SSSDP.TV SSSP.VV VRNNP.SV SSSTDP.ST	TYDLENSYK NULDLYKRYT DDLEIFDRYK NDLEIFDRYK NULDLYERYT NULDLYERYT DULLYERYT DDLLYERYT DULJYERYT DULJYERYT DULLYERYT DUVHIFERYK NVRLFWRYS NUVRLFWRYS DSLYLSRVF NDVYFQVF NDVYFQVF NDVYLQNFF	TVUD LIFLANT KVID KMQ EYG SVUD MFANNS SVUD MFANNS SVUD KFANNS SVUD KFANNS SVUD LIFLANT KVID NMQ EYG SVUD LIFLANT SVID LAFK SVUD LIFLANT SVID LAFKNS SVID LAFKNS SVUD LIFLANT KVID ANFKNS SVID AFKNS SVID SSA	NVLCFFAGNE NVLCFFAGNE NVLCFFAGNE NVLCFFAGNE NVLCFFAGNE NVLCFFAGNE NVLCFFAGNE NVLCFFAGNE NVLCFFAGNE NVLCFFAGNE NVLCFFAGNE NVLCFFAGNE NVLCFFAGNE NVLCFFAGNE NTLAFFAGNE NTLAFFAGNE	VTNNYTN VTNKKSN VTNKKSN VTNKKSN VTNNSTN VTNNSTN VTNNRSN VTNNRSN VTNKKSN VTNKKSN VTNKKSN VTNKKSN VTNKKSN VTNKKSN VTNKKSN VTNKKSN VTNKKSN VTNKKSN VINCOSO VINCOSO VINCOSO VINCOSO VINCOSO	TDASAPVKA TDASAPVKA TDASAPVKA TDASAPVKA TDASAPVKA TDASAPVKA TDASAPVKA TDASAPVKA TDASAPVKA TDASAPVKA TDASAPVKA TDASAPVKA TSAAVVKA	I R D VROYISDKNY I R D MKKY IKESDY I R DTKSYIKESDY I R DTKSYIKESGY VRDTKSYIKESGY VRDTKSYIKESGY VRDTKAYIKSKGY I R DTKAYIKSSNY I R DKKYIKSSNY I R DKKYIKSSNY I R DKKSYMOCGY I R DKKSYMOCGY I R DKKSYMOCGY I R DKKSYMOCGY I R DKKSYMOCGY I R DKKSYMOCGY I SKLAN I R DKKYXKAR I SKLAN I R I SKLAN
rnr2p CgGas1p CgGas2p Epd1p Epd2p SpOrf3p CdPhr1p CdPhr2p Gel3p CgGas2p Gel3p CgGas3p PcPhr1p Gas3p Gel2p NcGel2p Gas5p Gel1p NcGel1p	NRDDP. TW NRNNP. ED NRCDP. SW NRCSP. SW NRDSP. SW NRDSP. SW NRDDP. OW SSSDP. TW SSSDP. TW NRNDP. SW TSNSP. VW VRNNP. SW TSNSP. VW VRNNP. SW NRADPSGT SNRADPSGT QNYNPWESS NRAKPKES NRAKPKES	TYDLLFNSYK NLLLYKRYT DDLEFRYK NLLYFRYT INLLFRYT INLLFRYT TYDLFRYT TYDLFRYT TYDLFRYT TYDLFRYT TYDLFRYT TYDLYKWSFYR NVRLFWRYS DSLYLSRYT NDVYFKCVF YDTYVNRTF YDTYVNRTF	TVUDIFANT KVIDKUQEX SVUDMFANT SVUDMFANT SVUDKFHNY KVIDNMGEX EVUDSLAPT SVIDANQEX SVUDLFHNY SVIDANQEX SVUDANQEX SVUDANGEX SVID SVID SVIDANGEX SVID S	NVLGFFAGNS VLGFFAGNS NVLGFFAGNS NVLGFFAGNS NVLGFFAGNS NVLGFFAGNS NVLGFFAGNS NLGGFFAGNS NLLGFFAGNS NLLGFFAGNS NLLGFFAGNS NLLGFFAGNS NVLGFFAGNS NTLAFFGAS NTLAFFGGNS NTLAFFGGNS	VTNNYTN VTNKKSN VTNKKSN VTNKKSN VTNNSTN VTNNSTN VTNNRSN VTNKRSN VTNKRSN VTNKKSN VTNLKSN VTNL VTNDLYN VTNDLYN VTNDLYN VTNDLYN VILDTEN VINDQSD VINDQSS VINDQSP VINDQSP VINDQSP VINDQSP VINDQSP VINDQSP	TDASA PVKA TDASA PVKA TDASA PVKA TDASA PVKA TDASA PVKA TDASA PVKA TDAS PVKA TDAS PVKA TDAS PVKA TDAS PVKA TDAS PVKA TASA PVKA TASA VVKA KIDP VIT KKIP VVV SSSAA PVVKA	I R D VROYISDKNY I R D MKKYIKSSDY I R DTKSYIKSSGY I R DTKSYIKSKGY VR D TKGYIKSKGY VR D TKGYIKSKGY VR D KKYIKSSGY I R DTKAYIKSKGY I R DAKEYIKSSGY I R DAKEYIKSKGY I R DAKEYIKSKGY I R DAKEYIKSKGY I R DKKSYMOCKGY V R DVKSYMOCKGY V R DVKSYMOKGY I R DAKEYIKSKGY I R DAKEY I R DAKEY I R DTKAYIKTKGY I R DIKYYKSKA K M S K M C K KAN I R DIKYYKSKA V R DVKSYMOKGY V R DVKSYMOKGY V R DVKSYMOKGY I R DAKEYIKSK I R D K K K K K K K K K K K K K K K K K K
rnr2p CgGas1p CgGas2p Epd1p Epd2p SpOrf3p CdPhr1p Gas2p Gel3p CcGas3p PcPhr1p Gas3p Gel2p NcGel2p Gas5p Gel1p NcGel1p	NRDDP.TW NRNNP.EW NRESP.SW NRDSP.SW NRDSP.SW SSDP.TW SSSDP.TW SSSDP.TW SSSDP.TW NRNSP.W NRNSP.W NRNDP.SW TSNSP.VW VRNNP.SW ISTDP.SW TSNSP.VV VRNNP.SS SRYDPACS SRYDPACS SRYDPACS	TYDLENSYK NLLLYRRYT DLLEFRYK DLEIFRYK NLLLYRRYT NLLLYRRYT NLLLYRRYT DLEIFRYK NLLLYRRYT DLLIYRRYT DLLYRRYT DLLLYRRYT DULJYRRYK TYDQVARYT NVRLFWRYS NNVRLFWRYS NLYRYF NLYLSRYF NLYLSRYF NLYLCYTF	TVUDETFANT KVIDKMQEYS SVUDMFANTS SVUDMFANTS SVUDKFHNYT KVIDKHNYT KVIDKASSFH SVUDLFHNYT SVIDANS SVIDAFS SVIDAFS SVIDAFS SVIDAFS SVIDAFS KVIDAFS SVID SVID SVIDAFS SVID SVIDAFS	NVLCFFAGNE NVLCFFAGNE NVLCFFAGNE NVLCFFAGNE NVLCFFAGNE NVLCFFAGNE NVLCFFAGNE NVLCFFAGNE NVLCFFAGNE NVLCFFAGNE NVLCFFAGNE NVLCFFAGNE NVLCFFAGNE NTLAFFAGNE NTLAFFAGNE NTLAFFSGNE NTLAFFSGNE	VTNNYTN VTNNRSN. VTNKKSN. VTNNRSN. VTNNSTN. VTNNRSTN. VTNNRSTN. VTNNRSN. VTNNRSN. VTNNKKSN. VTNKKSN. VTNKKSN. VTNLDHTN. VINDHTN. VINDLYN. VINDQSD. VINDQSS. VINDQSS. VINDQS. VINDQS. VINDQS. VINDQS. VINDQS.	TDASAPVKA TDASAPVKA TDASAPVKA TDASAPVKA TDASAPVKA TDASAPVKA TDASAPVKA TDASAPVKA TDASAPVKA TDASAPVKA TDASAPVKA TDASAPVKA TDASAPVKA KIDPPIIRA KIDPPIIRA KIDPPIIRA KIDPPIIRA KIDPPIIRA	I R D VROYISDEKNY I R D MKKY IKES DY I R D TKSY IKES DY I R D TKSY IKES GY V R D TKSY IKES KGY V R D KKSY IKES KGY V R D KKSY IKES ST I R D TKAY IKES ST I R D TKAY IKES ST I R D TKAY IKES ST I R D KSY IKES ST I R D KSY IKES ST I R D KSY ISSE V R D VKSY ST I R D KSY ISSE V R D VKSY ST I R D KSY
rnr2p CgGaslp Epd1p Epd2p SpOrf3p CdPhr1p CdPhr1p CdPhr1p Gas2p Gel3p CgGas3p PcPhr1p Gas3p NcGel2p NcGel2p NcGel2p NcGel1p NcGel1p SpOrf1p SpOrf1p	NRDDP.TW NRNNP.ED NRCDP.SW NRCSP.SW NRDSP.SW NRDSP.SW NRDDP.W SSSDP.TW SSSDP.TW NRNDP.SW SSDP.TW NRNDP.SW YSNSP.W VRNNP.SW YSNDP.SW NRADPSGT SSTDP.SS NRADPSGT SSTDP.SS NRAKPKES SRLDAASS SRLDAASS	TY DLLFNSYK NLLLYKRYT DLLFFRYK DLEIFPRYK NLLYFRYT NLLLYFRYT TY DLFFRYT TY DLFFRYT TY DLFFRYK TY DLJFFRYK TY DLDFFRYK TY DLYFRYS DSLYLFRYS DSLYLSRYF NDVFK CVFF YDTY VNRTF YDTV NRLFYRYS NLYFRCYF YDTY VNRTF NDVYFC VF	TVUDIFANT KVIDKUQEX SVUDMFANT SVUDMFANT SVUDKFHNY VUDKFHNY KVIDNMQEX SVUDKFHNY SVUDKFHNY SVUDALFHNY SVIDANGES SVIDANGES SVIDANGKS SVIDANGKN KVIDNMHKY KVIDNMHKY KVIDNMHKY ATIDAFKNY ATIDAFANY ATIDAFANY	NVLGFFAGNS VLGFFAGNS NVLGFFAGNS NVLGFFAGNS NVLGFFAGNS NVLGFFAGNS NVLGFFAGNS NLGGFAGNS NLGGFAGNS NLGFFAGNS NLLGFFAGNS NLLGFFAGNS NVLGFFAGNS NTLAFFGAS NVLGFFAGNS NVLGFFAGNS NVLGFFAGNS NVLGFFAGNS NVLGFFAGNS NVLGFFAGNS	VTNNYTN VTNNRSN VTNKKSN VTNKKSN VTNNSTN VTNNSTN VTNNRSN VTNNRSN VTNKRSN VTNKKSN VTNKKSN VTNDLYN VTNDLYN VTNDLYN VTNDLYN VILDTEN VINDQSD VINDQSD VINDQSS VINDQSD VINDQS VIND VINDQS VIND VINDQS VIND VIND VIND VIND VIND VIND VIND VIND	Τ DASA PVKA T DASPVKA T DASPVKA T DASA PVKA T DASA PVKA T DASA PVKA K D P PVKA K D P PVKA K V P PVKA K V P PVKA S S A A PVKA S S S A PVKA S S A PVKA S S S A PVKA S S S A PVKA S S S A PVKA S S S S S S S S S S S S S S S S S S S	I R D VROYISDKNY I R D MKKY IKES DY I R D TKSY IKES GY I R D TKSY IKES KGY V R D TKSY IKES KGY V R D TKSY IKES KGY V R D KSY IKES SGY A R D MKKY IKEN GY V R D VKSY IKES KGY I R D TKAY IKES KGY I R D TKAY IKES KGY I R D KSY MCOKGY V KSY MCOKGY V R D VKSY MESKGY I R D KSY MCOKGY V R D VKSY MESKGY I R D KSY MESKGY I R D KSY KSK KSY SS KS I R D KSY SS KS I R D K
rnr2p CgGas1p Epd1p Epd2p SpOrf3p CdPhr1p CdPhr1p CdPhr1p Gas2p Gel3p CgGas3p PcPhr1p Gas2p Gel2p NCGel2p Gas5p Gel1p NCGel1p SpOrf1p SpOrf2p Cice11-	NRDDP.TW NRNNP.EW NRDDP.SW NRDDP.SV NRDDP.SV SSDP.TW SSDP.TW SSDP.TW SSDP.TW NRNSP.W NRNSP.EW NRNDP.SV TSNSP.VV VRNNP.SV ISTDP.SST JSTDP.SST SRYDPACS SRYDPACS SRYDPACS SRYDPACS SRAUPACS SRAUPACS	TYDLENSYK NULDLYKRYT DDLEIFDRYK NDLEIFDRYK NDLLYFRYT NULDLYFRYT NULDLYFRYT DDLUYFRYT DDLUYFRYT DDUUYFRSYK TYDQVARYT NULLYFRYT TYDQVARYT NUVFYRSYF NDVYFSVF NDVYLSYF NDVYLQYF NDVYLQYF NDVYLQYF NDVYLQYF	TVUDETFANT KVIDKMQEYS SVUDMFANTS SVUDMFANTS SVUDKHNYT KVIDNMQEY SVUDKHNYT KVIDNMQEY SVIDANS SV	N VLCFFAGNE N VLCFFAGNE N VLCFFAGNE N VLCFFAGNE N VLCFFAGNE M VLCFFAGNE N VLC	VTNNYTN VTNKKSN VTNKKSN VTNKKSN VTNNSTN VTNNSTN VTNNRSTN VTNNRSN VTNNKSN VTNNKSN VTNKKSN VTNKKSN VTNKKSN VTNLDHTN VILDTEN VILDTEN VILDES VINDQSD VINDQSS VINDQS VINDQS VINDQS VINDQVN VINDVN VINDVN	TDASAPVKA TDASAPVKA TDASAPVKA TDASAPVKA TDASAPVKA TDASAPVKA TDASAPVKA TDASAPVKA TDASAPVKA TDASAPVKA TDASAPVKA TDASAPVKA TDASAPVKA SAGAAPVKA KIDPPIIRA VKNVPTVKA SSAAPVKA SSAAPVVKA TTAAT	I R D VROYISDEKNY I R D MKKY IKES DY I R D TKSYIKES BY I R D TKSYIKES BY I R D TKSYIKES BY V R D KSYIKES BY I R D TKAYIKS KGY V R D MKKYIKS SY S I R D TKAYISS BY I R D TKAYISS BY I R D KSYISS I R D KSYISS V R D VKSYISS V R D VKSYISS V R D VKSYISS I R D KSYISS I R D
rnr2p CgGas1p Epd1p Epd2p SpOrf3p CdPhr1p CdPhr2p Gas2p Gel3p CgGas3p Gel2p NcCel1p Gas5p Gel1p NcCel1p SpOrf1p SpOrf1p SpOrf1p Ciael1p	NRDDP. TW NRDNDP. SW NRDSP. SW NRDSP. SW NRDSP. SW NRDDP. O SSDP. TW SSSDP. TW NRNDP. SW SSDP. TW NRNDP. SW TSNSP. VW NRDP. SW TSNSP. VW NRDP. SW TSNSP. VW NRDP. SW TSNSP. VW NRDP. SW STSNSP. VW NRDP. SW STSNSP. VW NRDP. SW STSNSP. VW NRDP. SW STSNSP. VW NRDP. SW STSNSP. VW STSNSP. VW NRDP. SW STSNSP. VW NRDP. SW STSNSP. VW STSNSP. VW STSN	TYDLEFNSYK NLLLYKRYT DDLEFRYK NDLEFRYT NLLYFRYT NLLYFRYT NLLDFFRYT TYDLFFRYT TYDLFFRYK NULLFRYT TYDLFFRYK TYDLYFRYS DSLYLSRYT NUVFKCVFF NDVFKCVF YDTYVNRT YDLYKKCVF NDVFKCVF NDVFKCVF	TVUDIFANT KVIDKMQEYS SVUDMFANTS SVUDMFANTS SVUDMFANTS KVIDMFANTS KVIDMFANTS SVUDKFHNYT KVIDMMQEYS SVUDLFHNYT SVIDAMQEYS SVUDAFQKT DVUDAMFYM KVIDMMFX FIDAFKNYT ATIDAFANY ATIDAFANY ATIDAFANY	NVLCFFACND VLCFFACND VLCFFACND VLCFFACND NVLCFFACND NVLCFFACND NVLCFFACND NLCFFACND NLCFFACND NLCFFACND NLCFFACND NLCFFACND NLCFFACND NLCFFACND NTLAFFSCND NVLCFFACND NTLAFFSCND NULCFFACND NTLAFFSCND NULCFFACND NTLAFFSCND	VTNNYTN VTNNRSN. VTNKKSN. VTNKKSN. VTNNSTN. VTNNSTN. VTNNRSTN. VTNNRSN. VTNKRSN. VTNKKSN. VTNDKYN. VTNDLYN. VTNDLYN. VINDQSD. VINDQSD. VINDQSD. VINDQSD. VINDQSP. VINDQSP. VINDQSP. VINDQSP. VINDQED.	TDASA PVKA TDASA PVKA TDASA PVKA TDASA PVKA TDASA PVKA TDASA PVKA TDAS PVKA TDAS PVKA TDAS PVKA TDAS PVKA TDAS PVKA TFAS PVKA TFAS PVKA SAGAA PVKA SAGAA PVKA TSILAAK VKA SSSAA PVKA SSSAA PVVKA SSSAA PVVKA SSSAA PVVKA SSAA PVVKA SATTT IVVKA	I R DVROYISDKNY I R DVROYISDKNY I R DRKY I K SSOY I R DTKSYIKSKY V R DTKOYISSKY V R DTKOYISSKY V R DVKSYIKSSGY V R DVKSYIKSSGY V R DVKSYIKSSGY V R DVKSYIKSSGY I R DAKEYIKSKY V R DVKSYMOVG V R VKSYMOVG V R V R V R V R V R V R V R V R V R V R
rnr2p CgGas1p Epd1p Epd1p SpOrf3p CdPhr1p CdPhr1p Gas2p Gel3p CgGas3p PcPhr1p Gas3p Gel2p NcGel2p Gas5p Gel1p NcGel1p SpOrf1p Cicel1p SpOrf2p Cicel1p Gas4p	NRDDP.TW NRNNP.EW NRCDP.SV NRCDP.SV NRCSP.SV NRDSP.SV SSDP.TW SSSDP.TW SSSDP.TW NRNSP.EV NRNDP.SV TSNSP.VV VRNNP.SV ISTDP.SV TSNSP.VV VRNNP.SS SRYDP.SST NRADPSST DRTDPEST DRTDPEST NRAKKESS NRAKPKES NRAKPKES NRAKPKAS	TYDLENSYK NLLLYRKYT DLLEFRYK DLLEFRYK NLLLYRRYT NLLLYRRYT NLLLYRRYT DLLIFRYK NLLLYRRYT DLLLYRRYT DLLLYRRYT DLLLYRRYT DLLLYRRYT TYDLYRRYF NVRLFRYK NVRLFRYK NVRTFRY NDVYLOXFY NDVYLOXFY NAJLONT NAVFLOCT NAVFLOCT NAVFLOCT NAVFLOCT NAVFLOCT	TVUDEFANT KVIDKMQEYS SVUDMFANTS SVUDMFANTS SVUDKFANTS SVUDKFHNYT KVIDNMQEY SVUDKFHNYT KVIDNMQEY SVIDAFQKU SVIDAFQKU SVIDAFQKU KVIDNMHKY SVIDAFQKNY KVIDAFQKNY ATIDAFKNYT ATIDAFKNYT ATIDAFKNYT	N VLGFFAGNS VLGFFAGNS N VLGFFAGNS N VLGFFAGNS	VTNNYTN VTNKKSN VTNKKSN VTNKKSN VTNNSTN VTNNSTN VTNNRSN VTNNRSN VTNNRSN VTNNRSN VTNKKSN VTNKKSN VTNKKSN VTNKKSN VTNKKSN VINDGS VINDQSD VINDQSS VINDSS VINDSS VINDSS VINDSS VINDSS VINDSS VINDSS VINDSS VINDSS VINDSS VINDSS VINDSS VINDSS VINDSS VINDS	TDASAPVKA TDASAPVKA TDASAPVKA TDASAPVKA TDASAPVKA TDASAPVKA TDASAPVKA TDASAPVKA TDASAPVKA TDASAPVKA TDASAPVKA TDASAPVKA TDASAPVKA SGAAPVKA TASAPVKA TSAAPVKA TSAAPVKA SSAAPVVKA TTAIT.WVKA	I RDVRQYISDKNY I RDMKKYYKESDY I RDTKSYYKESGY VRDTKSYIKESGY VRDTKSYIKESGY VRDTKSYIKESGY VRDTKSYIKESGY VRDVKSYIKSSY I RDTKAYISSGY I RDTKAYISSGY VRDVKSYKSS VRDVG VRDVKSYMSGGY VRDVKSYMSGGY VRDVKSYMSGGY VRDVKSYMSGGY VRDVKSYMSGGY VRDVKSYMSGGY VRDVKSYMSGS VRDVKSYMSGS VRDVKSYMSGS VRDVKS VRDVS VRDVS VRDVS VRDVS VRDVS VRDVS VRDVS VRDVS VRDVS VRDVS VRDVS VRDVS VRDVS VRDVS
rnrzp CgGaszp Epd1p Epd2p SpOrf3p CdPhr1p CdPhr1p Gas2p Gel3p CgGas3p Gel2p McGel2p McGel2p McGel1p SpOrf1p SpOrf1p SpOrf1p SpOrf1p SpOrf1p SpOrf1p	NRDDP. TW NRDNP. BW NRDSP. SV NRDSP. SV NRDSP. SV NRDSP. SV NRDDP. O SSDP. TW SSSDP. TW SSSDP. TW NRNDP. SV TSNSP. VV NRDP. SV TSNSP. VV NRDP. SV TSNSP. VV NRDP. SV TSNSP. VV NRDP. SV SSDP. TW SSTDP. SV SSP. VDP. SV SSP. VDP. SV SSP. VDP. SV SSP. VDP. SV SSC DP. TW SSSDP. TW	TY DLLFNSYK NLLLYRRYT DDLEFRYK NDLEFRYT NLLYRRYT NLLYRRYT NLLLYRRYT TY DLFSYT TY DLFSYT TY DLFSYT TY DLFSYT TY DLYRSYT NDVHFRYR SLYLFRY NDVYK VSFYR NDVYFKQVF NDVYFKQVF NDVYFKQVF NDVYLQYL NDVYLQYF NDVYLQYF NDVYLQYF NDVYLQYF NDVYLQYF NDVYLQYF NDVYLQSVF NAVLQSVF NAVLQSVF NAVLQSVF NAVLQEVF NAVLQEVF	TVUDIFANT SVUDIFANT SVUDIFANT SVUDIFANT SVUDIFANT SVUDIFANT KVIDNMFANT SVUDIFANT KVIDNMFANT SVUDIFANT SVUD	NVLCFFAGNS VLCFFAGNS VLCFFAGNS VLCFFAGNS NVLCFFAGNS NVLCFFAGNS NVLCFFAGNS NVLCFFAGNS NLCCFFAGNS NLCGFFAGNS NLCGFFAGNS NLCGFFAGNS NTLAFFGGNS NVLCFFAGNS NTLAFFGGNS NVLCFFAGNS NTLAFFGGNS NVLCFFAGNS NTLAFFGGNS	VTNNYTN VTNKKSN. VTNKKSN. VTNKKSN. VTNNSTN. VTNNSTN. VTNNKTN. VTNNKTN. VTNKKSN. VTNKKSN. VTNDKYN. VTNDLYN. VTNDLYN. VINDQSD. VINDQSD. VINDQSD. VINDQSD. VINDQS.	TDASA PVKA TDASA PVKA TDASA PVKA TDASA PVKA TDASA PVKA TDASA PVKA TDAS PVKA TDAS PVKA TDAS PVKA TDAS PVKA TDAS PVKA TASA PVKA SAGAA PVKA KIDP YIKA AGOV P P YIKA AGOV PX	I R D VROYISDKNY I R D MKKYIKSDY I R DTKSYIKSKY I R DTKSYIKSKY VR D TKOYIKSYIKSKY VR D TKOYIKSKY VR D KKYIKSKSY VR D KKYIKSKSY VR D KKYIKSK VR D KKYISK VR D KKYIKSK VR D KKYIK VR D
rnr2p CgGas1p Epd1p Epd2p SpOrf3p CdPhr1p CdPhr1p Gas2p CgGas3p PcPhr1p Gas3p Gel1p NCGel1p SpOrf1p SpOrf1p Gas4p Phr3p BgGvf1p	NRD DP. TW NRDNP. SW NRDNP. SW NRD DP. SW NRD DP. SW NRD DP. SW NRD DP. SW SS DP. TW SS DP. TW SS DP. TW NRN SP. W NRN DP. SW TSN SP. VW NRD DP. SW TSN SP. TDP. SW SS ST DP. SW SS ST DP. SW SS SW NRA KPKSS NRA KPKSS SR YDPACS SR YDPACS SR YDPACS SR YDPACS SR XDPACS SR XDPACS SR AKPKES NRA KPKSS NRA KPKSS SR AKPKSS SR STALASS	TYDLENSYK NULDLYKRYT DDLEIFPRYK NDLDLYFRYT NDLLYFRYT NULDLYFRYT INLLLYFRYT INLLLYFRYT DDLULYFRYT INLLLYFRYT DDLULYFRYT TYDLYK NUVLYFRYT NUVLYFRYT NUVLYFRYT NUVLYFR NUVYLOSAP NUVYLOSAP NUVYLOSAP NUVYLOSAP NUVYLOSAP NEVYLFRYT	TVUDUFFANT KVIDKMQEYS SVUDMFANTS SVUDMFANTS SVUDKFANTS SVUDKFHNYT KVIDNMQEYV SVUDLFHNYT SVIDAFQKUL SVIDAFQKUL SVIDAFQKUL FUVDAFQKNYT GVIEAFKNYT ATIDAFGKNY ATIDAFAKNYT ATIDAFAKNYT ATIDAFAKNYT ATIDAFKNYT	N VLCFFAGNE VLCFFAGNE N VLCFFAGNE N VLCFFAGNE	UTINNYTN VTINNKKSN VTINKKSN VTINKKSN VTINKSTN VTINNSTN VTINNKSTN VTINNKKSN VTINKKSN VTINKKSN VTINKKSN VTINKKSN VINDQSD VIND VINDQSD VIND VIND VIND VIND VIND VIND VIND VIN	TDASAPVKA TDASAPVKA TDASAPVKA TDASAPVKA TDASAPVKA TDASAPVKA TDASAPVKA TDASAPVKA TDASAPVKA TDASAPVKA TDASAPVKA TASAPVKA SGAAPVKA SGAAPVKA SSAAPVVKA SSAAPVVKA SSAAPVVKA	I R D VROYISDEKNY I R D MKKY IKES DY I R D TKSYIKES BY I R D TKSYIKES BY I R D TKSYIKES BY I R D TKSYIKES BY I R D TKAYIKES BY I R D TKAYIK
rnrzp CgGaslp Epd1p Epd2p SpOrf3p CdPhrlp CdPhrlp CdPhrlp CdPhrlp Gas2p Gel3p Cel2p NCGel2p Gas1p SpOrf1p SpOrf1p SpOrf1p SpOrf1p SpOrf1p EgOrf1p Phorf1p	NRDDP. TW NRDNDP. 50 NRDSP. 50 NRDSP. 50 NRDSP. 50 NRDSP. 50 NRDDP. 60 SSDP. TW SSSDP. TW SSSDP. TW NRNSP. 60 SSDP. 70 NRDP. 50 TSNSP. 70 NRDP. 50 STSNSP. 70 NRDP. 50 STSNSP. 70 NRDP. 50 STSNSP. 70 STSNSP. 70 NRDP. 50 STSNSP. 70 NRDP. 50 STDP. 50 SR 1DP. 50 SR 1DP	TY DLLFNSYK NLLLYKRYT DDLEFRYT NLLFRYT NLLFRYT NLLFRYT NLLFRYT TY DLSFRYT TY DLSFRYT TY DLFFRYK NT DLLFRYT TY DDLFFRYK NDVYFRYT DDUFFRYK NDVYFRYT NDVYKSFYR NDYFRYF SUELYRYT	KVUDKUR SVVDLFHNY SVVDMFANNS SVVDMFANNS SVVDKFHNY KVIDNMFANNS SVVDKFHNY KVIDNMFANNS SVVDLFHNY KVIDNMFAN SVVDLFHNY SVIDAF SVID SVIDAF SVIDAF SVID SVIDAF SVID SVID SVID SVID SVID SVID SVID SVID	NVLCFFAGNS VLCFFAGNS VLCFFAGNS VLCFFAGNS NVLCFFAGNS NVLCFFAGNS NVLCFFAGNS NVLCFFAGNS NULCFFAGNS NULCFFAGNS NVLCFFAGNS NVLCFFAGNS NTLAFFSGNS NVLCFFAGNS NTLAFFSGNS NTLAFFSGNS NTLAFFSGNS NTLAFFSGNS NTLAFFSGNS NTLAFFSGNS NVLCFFAGNS NTLCFFAGNS NTLCFFAGNS NTLCFFAGNS NTLCFFAGNS NTLCFFAGNS NVLCFFAGNS NVLCFFAGNS NVLCFFAGNS NVLCFFAGNS NVLCFFAGNS NVLCFFAGNS	VTNNYTN VTNNKSN. VTNKKSN. VTNKKSN. VTNNSTN. VTNNSTN. VTNNKTN. VTNNKTN. VTNNKKSN. VTNKKSN. VTNNKKSN. VTNDLYN. VTNDLYN. VINDQSDY. VINDQSDY. VINDQSSN. VINDQSS. VINDQSS. VINDQS. VINDS. V	TDASAPVKA TDASAPVKA TDASAPVKA TDASAPVKA TDASAPVKA TDASAPVKA TDASPVKA TDASPVKA TDASPVKA TDASPVKA TDASPVKA TDASAPVKA SAGAAPVKA KIDPYIKA SAGAAPVKA KIDPYIKA SSAAPVVKA SNTAPVVKA QOYPPAVKA	I R DVROVI SDKNY I R DVROVI SDKNY I R DTRISVI SSKY I R DTRISVI SSKY I R DTRISVI SSKY V R DTRISVI SSKY I R DVRSVI SSKY I R DVRSVI SSKY I R DVRSVI SSKY V R DVRSVI SSKY I R DAKEY I R DVRSVI SSKY I R DAKEY I R DVRSV V R DVRSV I R V R DVRSV I R DVRSV I R V R V R V R V R V R V R V R V R V R
rnrap CgGaslp Epd1p Epd2p SpOrf3p CdPhr1p CdPhr1p CdPhr1p Gas2p Gel2p Res5p Gel2p NcGel2p NcGel2p NcGel2p SpOrf1p SpOrf2p Ciellp SpOrf2p Ciellp Bas4p Phr3p EgOrf1p Phorf1p NcGel3p	NRDDP. TW NRDNP. SW NRDNP. SW NRDSP. SW NRDSP. SW NRDSP. SW NRDP. OU SSSDP. TW SSSDP. TW SSSDP. TW SSSDP. TW NRNSP. SW NRNDP. SW TSNSP. VW NRNDP. SW ISTDP. SW TSNSP. VW NRDP. SW SSSTDP. S	TYDLENSYK NULDLYKRYT DDLEIFDRYK NDLDLYFRYT NDLLYFRYT NULDLYFRYT NULDLYFRYT DDLULYFRYT DDLDLFFRYT NDLLYFRYT DDLDLYFRYT DDLDLYFRYT NDVLOVKUGRYR NVRLFWRYS NDVYLGYF NDVYLGYF NDVYLGYF NDVYLGYF NDVYLGYF NDVYLGYF NDVYLGYF NDVYLGYF NDVYLGYF NAFLGGT NNVYLGLYF NAVFLGGT NNVYLGHYF NAVFLGGT NNVYLGRYF	TVUDEFANT KVIDKMQEYS SVUDMFANTS SVUDMFANTS SVUDKFHNYT KVIDNMQEY SVUDKFHNYT KVIDNMQEY SVIDLFHNYT SVIDAMGSSF SVIDAFGXU SVIDAMGXSSF SVIDAFGXU SVIDAFGXU KVIDNMHXT SVIDAFGXU KVIDAFGXU ATIDAFGXU ATIDAFGXU ATIDAFAXU ATIDAFAXU ATIDAFAXU ATIDAFAXU	N VLGFFAGNS VLGFFAGNS N VLGFFAGNS N VLGFFAGNS	UTINNYTN VTINNKKSN VTINKKSN VTINKKSN VTINNSTN VTINNSTN VTINNKTN VTINNKKSN VTINKKSN VTINKKSN VTINKKSN VINDQSD V	TDASAPVKA TDASAPVKA TDASAPVKA TDASAPVKA TDASAPVKA TDASAPVKA TDASAPVKA TDASAPVKA TDASAPVKA TDASAPVKA TDASAPVKA TDASAPVKA TDASAPVKA TASAPVKA TASAAVVKA TASAAVVKA SSAAPVVKA TASAPVKA STAAPVKA TASAPVKA TASAAVVKA	I R D VROVI SDKNY I R D MKKY IKES DY I R D TKSYIK SKGY VR D TKSYIK SKGY VR D TKSYIK SKGY VR D TKSYIK SKGY I R D TKAYIKAKGY I R D TKAYIKAKGY I R D TKAYIK SKGY I R D KSYIK SKGY I R D KSKAYIK SKGY I R SKAYIK SKGY I R SKAYIK SKGY I R D KSKAYIK SKGY I R D KSKAYIK SKGY I R D KSKAYIK SKGY I R D KKAYI SKKGY I R D KKAYIK SKGY I R D KKAYIK SKKY I R D KKY I R

20

30

1

10

Fig. 1 Multiple sequence alignment of the portions spanning the Cdomain of proteins belonging to family GH-72 of the GH-A clan. The sequences of the C-domain of Gas1p (D23-T336) and of other members of Family GH-72 are presented. The identical residues (in the *black boxes*), similar residues (*bold*) and regions with consecutive similar residues (*white boxes*) are indicated. In order to improve the alignment, the N-terminal signal peptides were not included. Therefore, the numbering of the sequences starts from the putative or experimentally determined amino acid of the mature proteins. Thus, the D22 is D1 and E161 and E262 of Gas1p correspond to E139 and E240 in this figure. (Sequences from *Saccharomyces cerevisiae* are: Gas1p (SwissProt code P22146), Gas2p (Q06135), Gas3p (Q03655), Gas4p (Q08271), Gas5p (Q08193); from *C. albicans*: Phr1p (P43076), Phr2p (O13318), Phr3p (Q9P8R2); from *A. fumigatus*: Gel1p (O74687), Gel2p (Q9P8U4), Gel3p (Q9P8U3); from *C. glabrata*: CgGas1p (Q8X0Z7), CgGas2p (Q8X0Z6), CgGas3p (Q8X0Z5); from *P. carinii*: PcPhr1p (Q9UVL7); from *C. maltosa*: Epd1p (P56092), Epd2p (O74137); from *S. pombe*: SpORF1p (O13692), SpORF2p (Q9Y7Y7), SpORF3p (Q9P378); from *C. dubliniensis*: CdPhr1p (Q9HG19), CdPhr2p (Q9HG18); from *N. crassa*: NcGel1p (Q8X0X4), NcGel2p (Q8X094), NcGel3p (Q8X1E8); from *P. brasiliensis*: PbOrf1p (Q7Z8M3), from *E. gossypii*: EgOrf1p (GenPep code AAS51046.1)). The alignment is truncated since the C-terminal region of the C-domain (327–381) of NcGel4p presents an extra tail, which is not aligned to the other sequences

Fig. 1 (contd.)

1	70 180	190		200	210 2	20
Gas1p Phr1p Phr2p CgGas2p Epd1p Epd1p Epd2p CgGas2p Gal3p CgGas2p Gel3p CgGas3p PcPhr1p Gas2p Gel2p NcGel2p NcGel2p NcGel1p CiGel1p Gas4p Phr3p EgOrf1p Pborf1p NcGel3p NcGel4p	RK. IPVGYSSND RG. IPVGYSSND RG. IPVGYSSND RG. IPVGYSSND RG. IPVGYSSND RT. IPVGYSSND RT. IPVGYSSND RT. IPVGYSSND RT. IPVGYSSND RK. IPVGYSSND RK. IPVGYSSND RK. IPVGYSAND RS. IPVGYSAND RS. IPVGYSAND RS. IPVGYSAND RC. IPVGYSAND	DEDTRVKMTDV DEETRVALADY DEETRVALADY DADTRVJLADY DADTRVJLADY DADTRVJLADY DSDIRVSLARY DSDIRVSLARY DETRVALADY DETRVALADY DETRVALADY DATRVSLAEY DAMTRDNLARY DATRDNLARY DATRDLATFKY INPEIRLPLSDY DAETRINLSKY INPEIRLPLSDY DAETRINLSKY INTDLALATFKY INTDLALATFKY INTDLALATFKY INTDLALATFKY INTDLALATFKY INTDLALATFKY INTDLALATFKY INTDLALASY VAENRLOLASY VAENRLOLASY DANTRDEITA	TACGD TSCGS. TSCGS. TACGD. TACGD. TACGD. TACGD. TACGD. TACGS. TSCGS.	DDVKADFYGIN. LDDRADFFGIN. LDRADFFGIN. EDEAADFFGIN. EDESADFFGIN. EDESADFFGIN. EDESADFFGIN. DDRVDFYGIN. LDDRADFFGIN. DDRVDFYGIN. V. KADFFGIN. N. SADFYGIN. N. SADFYGIN. SICNIFFLKNI SICNIFFLKNI DFFAFN. DFFAFN. DFFAFN. DFFAFN. DFFAFN. SVDFFGVN. SVDFYGVN. SSENVDFFALN. SSENVDFFALN. QSAIDFFN. SSENVDFFN.	MYEWCG KSDFK MYEWCG DSSYK MYEWCG DSSYK MYEWCG DSSYK MYEWCG NSFYC MYEWCG NSFYC MYEWCG SSSFK MYEWCG DSSYK MYEWCG DSSYK MYEWCG VSSFY MYEWCG VSSFY MYEWCG VSSYK MYEWCG VSSYK MYEWCG VSSYK SKSWC GSSSK SKSWC GSSSF SYSK GBSSYK DYSWC SSSSF SYSK GBSSFY DYSWC SSSSF DYSWC SSSSFY DYSWC SSSSFY SYSK SSS SYSK GBSSFX SYSK GBSSFX	TTANKATSSATTTTSKKVTDIVTTSKEEG
	230	240	250	260	270 280	
Gasip Phrip Phrip CgGasip CgGasip Epdip Epdip Epdip CdPhrip CdPhrip Gasip CdPhrip Gasip CgGasip CgGasip CgGasip CgGasip NcGelip NcGelip NcGelip Sporfip Sporfip Sporfip Sporfip NcGelip NcGelip NcGelip NcGelip NcGelig Phorfip NcGelig NcGelig NcGelig Sporfip NcGelig NcGelig Sporfip NcGelig NcGelig	CYADRTAEFKNL CYADRTAEFKNL CYEATNDYKNL CYADRTKEFANL CYADRTKEFSNL CYADRTKEFSNK CYADRTEFKKF CYARTEFKNL CYARTEFKNL CYEATTEFKNL CYEATTEFKNL CYEATTEFKNL CYEATTEFKNL CYEATTEFKNL CYCAURTFCY CYCAURTFCY CYCAURTFCY CYDLIKNFTGY CYDQRWEFANY CYDQRWEFANY CYDQRKKFTGY CYDQRKKFTGY CYDQRKKFTGY CYDQRKKFTGY CYDQRKFSSY CYDQRKFSSY CYDQRKFSSY CYDQRKFSSY CYDQUVKFSSY CYDQUVKFSSY CYDQUVKFSSY CYDQUDFFKNY	SIEVFFSEYGC TIPAFFSEYGC SIPLFFSEYGC GIPIFFSEYGC GIPIFFSEYGC GIPIFFSEYGC TVPMIFSEYGC TVPMIFSEYGC FIPIFFSEYGC PIPVFFSEYGC PIPVFFSEYGC SVPVFFSEYGC GIPIFFSEYGC SVPVFFSEYGC GIPIFFSEYGC SVPVFFSEYGC GIPIFFSEYGC SVPVFFSEYGC SVPVFFSEYGC SVPVFFSEYGC SIPLFLSEYGC SIPLFLSEYGC SVPVFFSEYGC SVPVFFSEYGC SVPVFFSEYGC SVPVFFSEYGC SVPVFFSEYGC SVPVFFSEYGC	WEYT. PRL VANR. PRL VANR. PRL WEVR. PRL WEVT. PRL WEVT. PRL WEVT. PRL WEVT. PRL WEVS. PRT WEVS. PRT WEVS. PRT WEVR. RD WEVR. RD WEVR. PRT WEVR. <th>TEVEA. LYG DIMIT TEVAL LFGDQMT TEVAL LFGDQMT TEVAL LFGDQMT TEVGAL LFGDQMT TEVGAL LFGDQMT TEVGAL LFGDQMT OEVGT. LFGDQMT OEVGT. LFGDQMT OEVGT. LFGDQMT TEVGAL LFSDKMT TEVGAL LFSDKMT TEVGAL LFSDKMT TEVGAL LYGQLEMT TEVGAL LYGQMT OEVSSL VSTDMT OEVSL USTDMT TEVGAL LYGQMT TEVGAL LYGQMT TEVAA LYGQMT TEVAA LYGQMT TGQQNA LFGPEMU</th> <th>UW SGGIV MYFEERM VW SGGIV MYFEERM VW SGIV MYFEERM VW SGIV MYFEEN VW SGIV MYFEEN VY SGIV MYFEEN VY SGIV MYFEEN VY SGIV MYFEEN VY SGIV VYFEEN VY SGIV YEY VY SGIV VYFEEN VY SGIV YEY SARNA VFSGIV YEY SARNA VY SGIV YEY SARNA VY SGIV YEY SGEN VY SGEN VY SGEN</th> <th>XXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXX</th>	TEVEA. LYG DIMIT TEVAL LFGDQMT TEVAL LFGDQMT TEVAL LFGDQMT TEVGAL LFGDQMT TEVGAL LFGDQMT TEVGAL LFGDQMT OEVGT. LFGDQMT OEVGT. LFGDQMT OEVGT. LFGDQMT TEVGAL LFSDKMT TEVGAL LFSDKMT TEVGAL LFSDKMT TEVGAL LYGQLEMT TEVGAL LYGQMT OEVSSL VSTDMT OEVSL USTDMT TEVGAL LYGQMT TEVGAL LYGQMT TEVAA LYGQMT TEVAA LYGQMT TGQQNA LFGPEMU	UW SGGIV MYFEERM VW SGGIV MYFEERM VW SGIV MYFEERM VW SGIV MYFEEN VW SGIV MYFEEN VY SGIV MYFEEN VY SGIV MYFEEN VY SGIV MYFEEN VY SGIV VYFEEN VY SGIV YEY VY SGIV VYFEEN VY SGIV YEY SARNA VFSGIV YEY SARNA VY SGIV YEY SARNA VY SGIV YEY SGEN VY SGEN VY SGEN	XXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXX
Gas1p Phr1p CgGas2p Epd1p Epd1p Epd2p Gas2p Gas1p CdPhr1p CdPhr1p Gas2p Gel1p CgGas2p Gel2p NcGel1p NcGel1p NcGel1p SpOrf1p SpOrf1p SpOrf1p CiGel1p Gas4p Phr3p EgOrf1p NcGel3p NcGel3p NcGel3p	LVS	290 GND	Z VKTLDDF VKTLDDF VKTLDDF VKTLDDF VKTLDDF VKTLDDF VSTLDPF VSTLDF VKTLSDY VSTLDF VKTLSDY VTLADY VTLDPF VTLDPF VFPDF VFPDF VFPDF VFFDF VFFDF VFEDF VFLKDF VTLKDF VTLKDF VKEPDF VFSKNT FRAGTFFVLPDF AKGRTDF AKGRTDF AKGKT	NYSSEINKIS NYKSEMIKIS YYSSEIKDIH YYSSEIHSIS NYSKQIHSIS NYSKQIHSIS NYKSEILDIS NYKSEILDIS NYKSENKSIS NYKSENKSIS NYKSENKSIS NYKSENKSIS NYKSENKSIS NIKKEFAKAD YLSSEIGAT NILEQLRNVS NILEQLKNVS NILEQLKNVS NILEQLKNVS NILEQLKNVS NILEQLKVS NILEQLKVS NILEQLKVS NILEQLKVS NILEQLKVS NILEQLKVS NILEQLKVS NILEQLKVS NILEQLKVS NILEQLKVS NILEQLKVS NILEQLKVS NILEQLKVS NILEQLKVS NILEQLKYS NILEQLKYS NILEQLKYS NILEXAY NILEQLKYS NILEXAY NILEQUKY NILEQUKKFDTLPELD YYSAQIAKAS NILEQUWATIT	310 PTSANT PSLAHT PSSKA PSSVNS PSSVNS PSSVQA PSLAHT PSLAHT PSLAHT PSLAHT PKGVNL PKGVNL PKGVNL PKGVNL PKGVNL PKGVNL PKGVNL PKGVNL PKGVNL PKGVNL PKGVNL PKGVNL PKGVNL PKGVNL PKGVN PKGVN PKGVS PKGVN PTGLKRS PKGVNS PK	· · · · · · · · · · · · · · · · · · ·

(R90 and R271) and five cysteine residues (C74, C103, C216, C234 and C265), which might be involved in intra- or inter-domain disulfide bridges. Moreover, there are three residues (G48, Y51 and Q52) that are only missing in the NcGel3p sequence (Fig. 1).

Analysis of the alignment among the C-domains revealed that proteins from different species, but charac-

terized by the same modular architecture, are more closely related evolutionarily than proteins belonging to the same organism but featuring a different modular organization (Figs. 2 and 3), suggesting that the appearance of different modular organizations preceded speciation. Interestingly, members of the family characterized by the presence of the Cys-box domain have

DOMAIN COMPOSITION



Fig. 2 The domain composition of family GH-72 members, obtained by similarity with Gas members from S. cerevisiae. The C-domain, the pro-rich motif, the Cys-box, the Ser-box (or an

the LP [T, I] PP pro-rich motif, the only exceptions being Gas2p (LPETP) and NcGel3p (TPTPV) (not shown). In contrast, members in which the Cys-box is missing show more variability in the pro-rich motif, usually conserving only the LPXXP motif (which is missing in Gas4p and Phr3p).

The alignment also allowed us to define two fingerprints, which include the two catalytic residues E161 and E262, and are strictly specific for the family GH-72 of glycoside hydrolases:

- N-x(1,3)-[L, I]-[G, A]-[F, Y]-x-G-N-E-[I, V]
- P-x-[F, I]-x-[S,A]-E-[Y, F, T]-G-C

Fig. 1

In fact, the two patterns correctly identified all proteins belonging to family GH-72 in the nonredundant database of NCBI, with no hits corresponding to false positive protein sequences (see Methods).

As expected, the Blast search did not reveal any homologous proteins with known 3D structure and this ruled out the possibility of using standard homologymodeling approaches. With the aim of finding possible remote homologues to Gas1p, its sequence was also

PROTEIN

Gas1p, CgGas1p, CgGas2p, Phr2p, CdPhr2p, Epd1p, Phr1p, CdPhr1p, Epd2p, EgOrf1p, SpOrf3p, Gas2p, CgGas3p, PcPhr1p, NcGel3p, Gel3p, PbOrf1p, NcGel4p

Gas3p, Gel2p, NcGel2p, Gas4p, Phr3p, Gas5p, SpOrf1p, SpOrf2p, Gel1p, CiGel1p, NcGel1p

aspecific box) are indicated in light gray, black, dark gray and white, respectively. The members of the family from S. cerevisiae are indicated in *bold* [48]

submitted to PSI-Blast, obtaining statistically significant similarity to the LacZ domain of β -galactosidase (belonging to family GH-2 of the GH-A clan), in agreement with the previous observations made for other family GH-72 members [34], and also to the cellulase domain of glycosyl hydrolase (belonging to family GH-5 of the GH-A clan). To search for other possible remote homologues of Gas1p, we submitted the aminoacid sequences of the C-domain of Gas1p to six threading servers that use different methods to find suitable templates and generate the corresponding alignments: 3D-PSSM, mGen-THREADER, TOPITS, 123D+, Fugue, SAM-T02 and FFAS03. It should be noted that all hits found by TOPITS were characterized by low z-scores and therefore not analyzed further (not shown). Protein scaffolds retrieved by the majority of servers came from three families (GH-2, GH-5 and GH-17) belonging to the GH-A clan (Table 1), confirming and extending the results obtained by PSI-Blast, and strongly suggesting remote evolutionary relationship (and common fold) with family GH-72 members. Consequently, these proteins could be used as scaffolds to



Classification	Template (PDB code)						
	3D-PSSM	MG-THREADER	123D+	Fugue	FFAS03		
GH-A 5	7A3H , 1BQC , 1ECE, 1QN_ , 1GHS, 1EGZ , 1EDG, 1G0C, 1GZJ	7A3H, 1BQC , 1ECE, 1EDG, 1GZJ	7A3H, 1BQC, 1ECE, 1QN_, 1EGZ, 1EDG	7A3H, 1BQC, 1ECE, 1QN_, 1EGZ, 1GZJ, 1CZ1	7A3H , 1BQC , 1QN_ , 1EGZ , 1EDG, 1G0C, 1GZJ, 1CZ1, 1LF1, 1CEC, 1H4P, 1NOF		
GH-A 2	1BHG, BGAL	1BHG, BGAL	1BHG, BGAL	1BHG, BGAL	1BHG, BGAL		
GH-A 17	n.f.	1AQ0	1AQ0	n.f.	1AQ0,1GHS		
GH-A 42	n.f.	n.f.	n.f.	1KWG	1KWG		
GH-A 10	n.f	n.f.	n.f.	1CLX, 1XYZ, 1BG4, 1TAX	1CLX, 1XYZ, 1BG4, 1TAX. 1E0X, 1US2, 1N82, 1NQ6,1ISY, 1HIZ, 1UQY		
GH-A 26	n.f.	n.f.	n.f.	1J9Y	n.f.		
GH-A 39	n.f.	n.f.	n.f.	n.f.	1PX8		
GH-A 53	n.f.	n.f.	n.f.	n.f.	1HJQ, 1HJS, 1FHL		
GH-A 1	n.f.	n.f.	n.f.	n.f.	1NP2, 1BGG, 10D0, 1QOX, 1PBG, 1E1E, 1GNX, 1CBG, 1MYR		
GH-A 51	n.f.	n.f.	n.f.	n.f.	1PZ2		
GH-A 30	n.f.	n.f.	n.f.	n.f.	10GS		
$(\beta/\alpha)_8$	n.f.	n.f.	n.f.	1K6W	1UG6, 1QVB		
OTHER	n.f.	n.f.	1GCA	n.f.	n.f.		

Table 1 Template structures obtained by the threading servers using as probe the amino-acid sequence of the C-domain of Gas1p

Note that for some proteins more than one structure is deposited in the Protein data bank. In particular, 7A3H stands for 7A3H, 1A3H and 1E5J; 1QN_stands for 1QNR, 1QNO and 1QNS; BGAL stands for 1F49, 1F4A, 1DPO, 1JZ8, 1JZ7 and 1BGL. In *bold* the templates for which structural alignment were considered for the 3D-model generation are indicated. n.f. stands for not found

predict the 3D structure of the C-domain of Gas1p. However, it should be noted that the very low sequence similarity between Gas1p and these proteins (less than 15% identity) is expected to make room for errors due to local misalignment, which eventually can affect the quality of the 3D model. In particular, it is well known that alignments to the same scaffold produced by different threading methods can be affected by local errors, making the derivation of a good structural model a nontrivial task [35]. In fact, even though many threading servers converge on the same scaffolds (Table 1), corresponding alignments can be quite different (see below).

With the aim of selecting the most reliable sequencestructure alignments, we started from the observation that some amino acids, such as N160, E161 and E262 (Gas1p numbering) are strictly conserved in the GH-A clan [6] and therefore were expected to be aligned with the corresponding residues of the templates. In fact, the glutamic acid residue corresponding to E240 in Gas1p was always misaligned by Fugue and SAM-T02 (see Supporting information). It should also be noted that the corresponding alignments obtained with Clustal W [16] or T-Coffee [36] were affected by similar problems, even when using different scoring matrices and gap penalties (not shown), confirming the non-applicability of classical homology modeling approaches.

The surviving alignments were pruned further considering results from secondary-structure predictions. Indeed, we are aware that scores from some threading methods (mGen-THREADER, 3D-PSSM and 123D+) already take into account secondary-structure prediction

results. However, the prediction of the $(\beta/\alpha)_8$ fold for the C-domain of these proteins is so well grounded [5] that secondary-structure prediction data are expected to be more easily evaluated than other parameters entering the scoring function of the threading methods, such as the solvation potential. In fact, due to the multidomain architecture of Gas1p and congeners, the evaluation of solvation potential might be partially misleading. If some interactions among the different domains take place, hydrophobic residues could be exposed on the surface of the C-domain, thus producing low values of the solvatation potential. Therefore, with the aim of evaluating the different alignments in light of secondarystructure predictions, we submitted the sequences of all C-domains of family GH-72 members to the JPRED [18] and PSI-PRED [19] servers, obtaining a consensus prediction according to the procedure outlined in Methods (Fig. 4). As expected, the general $(\beta/\alpha)_8$ architecture was predicted with high confidence, in agreement with previous proposals [5]. However, some irregularities are predicted to characterize the C-domain of Gas1p and congeners. In particular, $\beta 1$ is preceded by two extra β strands, $\beta 5$ is very short, $\alpha 7$ might be missing or very short and one extra β -strand is present before $\beta 8$. In fact, slightly irregular $(\beta/\alpha)_8$ folds are quite common [37, 38] and also characterize some members of the GH-A clan for which the 3D structure has been solved by X-ray diffraction [39, 40].

In light of these results, alignments where the secondary-structure elements of the templates were badly aligned or largely different from those predicted for the Fig. 4 Secondary-structure prediction for Gas1p. The secondary structure of the regions predicted with high confidence are *highlighted*. The catalytic glutamic residues are underscored. Examples of alignments that have been discarded because secondarystructure elements are badly aligned are shown



1AQ0 aligned to Gas1p (3D-PSSM)			1ECE aligned to Gas1p (G. THREADER)			
$\beta 1$		$\alpha 1$	$\alpha 5$	eta 6		
EEEEEEEE		ННННН	НН ННННННННННН	EEEE E		
20	40		180	200		
Gas1p DDVPIEVVGNKFFYSNN	GSQFYTIRGV	AYQADTAN	Gas1p NDDEDTRVKMTDYFAC	GDDDVKADFYGINMY		
1AQ0	IGVCYGMS	ANNLPAAS	1ECE YWWGGNLQGAGQYPV-V	/LNVPNRLVYSAHDYATS		
120		140	220	240		
	EEE	нн		EEEEEE		
	$\beta 1$	$\alpha 1$		$\beta 6$		

C-domain were discarded. Note that in some cases the misalignment was not due to failure of the threading method but simply due to the structural features of the template, which did not fit properly to the predicted secondary structure of the C-domain. As an example, the analysis of the alignment between Gas1p and 1AQ0 reveals that the first β -strand (β 1) in the template is too short and that the two extra β -strands preceding β 1 are missing (Fig. 4). Similarly, α 5 is missing in the templates 1ECE, 1EDG, 1BHG and 1BGL, whereas its presence was predicted with high confidence for Gas1p and congeners (Fig. 4). Moreover, the extra β -strands at the *N*-terminal are missing in 1CZ1 and 1GZJ (data not shown).

According to this analysis, only four templates survived the pruning procedure. Remarkably, they all belong to the GH-5 family even if they share low-sequence similarity. This suggests a closer evolutionary relationship between the GH-72 and GH-5 families. However, it is known that GH-5 family members can be characterized by significantly different sequences, an observation that led to the definition of different subfamilies [41]. In

fact, the four templates belong to three different subfamilies. In particular, 1EGZ [42] and 7A3H [43] belong to the subfamily 5-2 and are characterized by endo-1,4glucanase activity, whereas 1QNS [44] and 1BQC [45], which have been classified in the subfamilies 5-7 and 5-8, respectively, are both characterized by β -mannanase activity.

The alignments among the four template proteins and Gas1p, as obtained by the different threading methods, are extremely similar for the location of the amino acids conserved in the GH-72 family (see Supporting information), indicating that, starting from a specific template, similar 3D models are obtained even considering alignments from different fold-recognition servers. The structures of the C-domain of Gas1p, obtained using the alignments produced by 3D-PSSM, and using the four templates from family GH-5, are shown in Figs. 5, 6 and 7.

Considering the general structural features of the Cdomain in the four templates, and consequently in Gas1p model, the $-NH_2$ and -COOH termini of the domain are located at the bottom of the barrel (with



Fig. 5 Three-dimensional models of the Gas1p C-domain as predicted using as templates 1QNS (a), 1EGZ (b), 1BQC (c) and 7A3H (d). For the sake of clarity only the β -strands forming the barrel and the side chains of R90, D117, E161 and E262 have been explicitly shown

respect to the catalytic glutamic residues). In particular, the two short N-terminal β -strands preceding β 1 reduce the accessibility to the bottom of the barrel, as also observed in other members of the GH-A clan [39, 42, 44, 45].

As discussed above, there are several residues that are strictly conserved in GH-72 family members. Residues corresponding to R90, G264 and Y231 (Gas1p numbering) are strictly conserved both in families GH-72 and GH-5. In particular, the functional role of the amino acids corresponding to R90 and Y231 has already been investigated in members of family GH-5. In the retaining cellulase Cel5A from *Bacillus agaradhaerens*, which corresponds to the template 7A3H, it has been argued that the hydrogen bond formed between the arginine residue and the catalytic nucleophile E228 is

crucial to maintaining the proper orientation and protonation state of the nucleophile in the glycosylation step [42, 43]. The Y231 residue has been shown to be important for substrate recognition and orientation/ activation of the nucleophilic glutamic catalytic residue [43, 46]. In both cases, a similar role in Gas1p and the other members of family GH-72 can be predicted confidently on the basis of the 3D models (Figs. 5, 7). The functional relevance in the GH-5 family of the glycine residue corresponding to G264 has never been investigated. Our structural analysis shows that this glycine residue is located near the two catalytic glutamic acid residues, suggesting a role in substrate recognition in both families (Fig. 7). Among the glycine residues conserved only in C-domains belonging to the family GH-72, G243 and G304 are solvent exposed and located at



Fig. 6 Three-dimensional models of the Gas1p C-domain as predicted using as templates 1QNS (a), 1EGZ (b), 1BQC (c) and 7A3H (d). For the sake of clarity only the side chains of E161 and E262 and cysteine residues have been explicitly shown

the beginning of $\alpha 6$ and in the loop following $\beta 8$, respectively. All other glycine residues are buried inside the barrel and are predicted to play a structural role.

The residues F44 and Y113, which are strictly conserved in family GH-72 members, are buried in the protein core and form a hydrophobic cluster with F35 or Y37, which are conserved in many members of the family (not shown).

The side chain of R247, which is strongly conserved or substituted by a lysine residue in the family, points toward a negatively charged region formed by the peptide segment DDED (residues 202-205). The first aspartate residue (D202) is conserved in a large number of family GH-72 members and it might form a salt bridge with R247 (not shown). The D117, which is conserved in GH-72 and GH-5 members (not shown), is placed at the end of β 3 and its side chain can interact with Arg90 (Fig. 5), possibly forming a network of electrostatic interactions also involving the nucleophilic glutamate residue (E262). This is analogous to the corresponding role proposed in the cellulase Cel5A (7A3H) from *B. agaradhaerens* [43].

The analysis of the spatial location of cysteine residues is particularly important for predicting the possible formation of disulfide bridges. The Gas1p contains 14 cysteine residues: five residues are located in the C-domain, eight in the Cys-box and one in the linker region located between the C-domain and the Cys-box. Ten cysteine residues out of 14 are involved in intra-domain



Fig. 7 Three-dimensional models of the Gas1p C-domain as predicted using as templates $1QNS(\mathbf{a})$, $1EGZ(\mathbf{b})$, $1BQC(\mathbf{c})$ and $7A3H(\mathbf{d})$. For the sake of clarity only the side chains of E161 and E262 as well as glycine, tryptophan, tyrosine residues predicted to be involved in substrate recognition have been explicitly shown

disulfide bonds in the native state of the protein [7]. Moreover, on the basis of the high similarity between the Cys-box of Gas1p and the corresponding domain in plants, it has been predicted that three disulfide bridges are formed in this domain [47]. These observations suggest that the other two disulfide bridges could involve cysteine residues in the C-domain (Fig. 6). The analysis of our 3D models of Gas1p suggests that a disulfide bridge can be formed between C234 and C265, which are located in loop regions in the proximity of the catalytic

site between $\beta6-\alpha6$ and $\beta7-\alpha7$, respectively. The C74 and C103 are localized on $\alpha1$ and $\alpha2$, respectively, and their distance suggests that the formation of a disulfide bridge is possible only upon some rearrangement of the protein backbone. Finally, C216, which is located in the loop connecting $\alpha5$ and $\beta6$, is far from the other cysteine residues and mainly solvent exposed. Remarkably, a cysteine residue (C348), which is conserved in the GH-72 family, is present in the sequence portion linking the C-domain and the Cys-box. This linker region is predicted

to assume a coil conformation according to secondarystructure predictions carried out with the JPRED and PSI-PRED servers. To investigate the possibility that C348 might be involved in a disulfide bridge with a cysteine residue of the C-domain, we linked a peptide spanning the amino-acid sequence of the linker (KSYSATTSDVAC) to the carboxy-terminal end of the C-domain, evaluating, by computer-aided graphical analysis (not shown), the possibility that C348 could interact with C74, C103 and C216. It turned out that C216 is too far from the C-terminal end of the C-domain to allow the formation of a disulfide bridge with C348. On the other hand, C348 can interact with C74 and C103, suggesting possible formation of a disulfide bridge. In light of our modeling results and the experimental observation that C74 is crucial for proper folding and maturation of Gas1p (while mutation of C103 and C265 have only slight effects) it can be inferred that the two disulfide bridges involving residues of the C-domain should be C234-C265 and either C74-C348 or C103-C348. It should also be noted that the phenotype observed following C74 mutation might be due to its involvement in a transient disulfide bridge formed during the folding process [7]. Site-directed mutagenesis of C348 is predicted to be a crucial experiment to distinguish among these possibilities.

The spatial localization of tyrosine and tryptophan residues is particularly relevant because these residues are often involved in substrate recognition in GH-A members [39, 43, 44]. Most of the tyrosine residues conserved in Gas1p and congeners are located in the β strands forming the barrel (Fig. 7). The Y294 and Y303 are predicted to be localized at the gate of the barrel, implying a role in substrate recognition. Remarkably, Y294 corresponds to a Trp residue that is conserved in family GH-5 members and is known to be involved in substrate recognition [42-45], suggesting that this position might be important for tuning substrate selection. Also, Y92 can interact with substrates, even if its spatial location is less conserved in the different models. On the other hand, Y51, Y113 and Y198, which are located in β 1, β 3 and β 5, respectively, are deeply buried in the barrel and are predicted to play a structural role. Finally, W233 is placed on the top of the barrel and can be involved in substrate recognition.

In conclusion, the merging of biochemical data with results from threading methods, multiple sequence alignments and secondary-structure predictions has allowed to predict the 3D-structure of the C-domain of Gas1p, in spite of its low-sequence similarity to structurally characterized proteins. The inferred structural properties of the C-domain have been used to generate a working hypothesis about the structural and functional role of key residues. The model could also be relevant for designing specific inhibitors of Gas1p and therefore new antifungal agents. In addition, it opens the possibility for targeted mutagenesis experiments.

Supporting information

The sequence alignments among templates and Gas1p and *xyz* coordinates for Gas1p models.

Acknowledgments The authors thank Popolo L. for helpful discussions, Stocchetti E., Gonella Diaza R., Mereghetti P. and Negroni J. for fruitful discussions and comments and Santarossa G. for assistance in preparing the figures.

References

- Mouyna I, Fontaine T, Vai M, Monod M, Fonzi WA, Diaquin M, Popolo L, Hartland RP, Latge JP (2000) J Biol Chem 275:14882–14889
- 2. Popolo L, Vai M (1999) Biochim Biophys Acta 1426:385-400
- 3. Klis F (1994) Yeast 10:851–869
- Gatti E, Popolo L, Vai M, Rota N, Alberghina L (1994) J Biol Chem 269:19695–19700
- Mouyna I, Monod M, Fontaine T, Henrissat B, Lechenne B, Latge JP (2000) Biochem J 347:741–747
- 6. Henrissat B, Davies G (1997) Curr Opin Struct Biol 7:637-644
- Carotti C, Ragni E, Palomares O, Fontaine T, Tedeschi G, Rodriguez R, Latge JP, Vai M, Popolo L (2004) Eur J Biochem 271:3635–3645
- 8. Forster MJ (2002) Micron 33:365-384
- 9. Godzik A (2003) Methods Biochem Anal 44:525–546
- 10. Tramontano A, Morea V (2003) Biotechnol Bioeng 84:756-762
- Contreras-Moreira B, Fitzjohn PW, Bates PA (2002) Appl Bioinform 1:177–190
- Rigden DJ, Jedrzejas MJ, De Mello LV (2003) FEBS Lett 544:103–111
- Karplus K, Karchin R, Draper J, Camper J, Mandel-Gutfreund Y, Diekhans M, Hughey R (2003) Proteins 53:491–496
- Altschul SF, Madden TL, Schäffer AA, Zhang J, Zhang Z, Miller W, Lipman DJ (1997) Nucleic Acids Res 25:3389–3402
- Gribskov M, McLachlan AD, Eisenberg D (1987) Proc Natl Acad Sci USA 84:4355–4358
- Higgins D, Thompson J, Gibson T, Thompson JD, Higgins DG, Gibsor TJ (1994) Nucleic Acids Res 22:4673–4680
- 17. Wootton JC (1996) Methods Enzymol 266:554-571
- Cuff JA, Clamp ME, Siddiqui AS, Finlay M, Barton GJ (1998) Bioinformatics 14:892–893
- McGuffin LJ, Bryson K, Jones DT (2000) Bioinformatics 16:404–405
- Kelley LA, MacCallum RM, Sternberg MJE (2000) J Mol Biol 299:499–520
- 21. Jones DT (1999) J Mol Biol 287:797-815
- Alexandrov N, Nussinov R, Zimmer R (1996) Pac Symp Biocomput. In: Hunter L, Klein TE (eds) World Scientific Publishing Co., Singapore, pp 53–72
- 23. Shi J, Blundell TL, Mizuguchi K (2001) J Biol Mol 301:243-257
- Rost B (1995) In: The third international conference on intelligent system for molecular biology (ISMB). CA: AAAI Press, Menlo Park, Cambridge, UK, pp 314–321
- Karplus K, Barrett C, Hughey R (1998) Bioinformatics 14:846– 56
- Rychlewsky L, Jaroszewski L, Li W, Godzik A (2000) Protein Sci 9:232–241
- Xiang Z, Soto C, Honig B (2002) Proc Natl Acad Sci USA 99:7432–7437
- Dauber-Osguthorpe P, Roberts VA, Osguthorpe DJ, Wolff J, Genest M, Hagler AT (1988) Proteins 4:31–47
- 29. Vriend G (1990) J Mol Graph 8:52–56
- Humphrey W, Dalke A, Schulten K (1996) J Mol Graph 14:33– 38

- Zhang Z, Schaffer AA, Miller W, Madden TL, Lipman DJ, Koonin EV, Altschul SF (1998) Nucleic Acids Res 26:3986– 3990
- Coutinho PM, Henrissat B (1999) In: Gilbert HJ, Davies G, Henrissat B, Svensson (eds) The Royal Society of Chemistry, Cambridge, pp 3–12
- 34. Fonzi WA (1999) J Bacteriol 181:7070-7079
- 35. Lemer CM, Rooman MJ, Wodak SJ (1995) Proteins 23:337– 355
- Notredame C, Higgins D, Heringa J (2000) J Mol Biol 302:205– 217
- 37. Nagano N, Orengo CA, Thornton JM (2002) J Mol Biol 321:741–765
- Henrissat B, Callebaut I, Fabrega S, Lehn P, Mornon JP, Davies G (1995) Proc Natl Acad Sci USA 92:7090–7094
- Sakon J, Adney WS, Himmel ME, Thomas SR, Karplus PA (1996) Biochemistry 35:10648–10660
- Jacobson RH, Zhang XJ, DuBose RF, Matthews BW (1994) Nature 369:761–766

- 41. Wang Q, Tull D, Meinke A, Gilkes NR, Warren RAJ, Aebersold R, Withers SG (1993) J Biol Chem 268:14096–14102
- 42. Chapon V, Czjzek M, Hassouni E, Py B, Juy M, Barras F (2001) J Mol Biol 310:1055–1066
- Davies GJ, Mackenzie L, Varrot A, Dauter M, Brzozowski AM, Schulein M, Withers SG (1998) Biochemistry 37:11707– 11713
- 44. Sabini E, Schubert H, Murshudov G, Wilson KS, Siika-Aho M, Penttila M (2000) Acta Crystallogr D Biol Crystallogr 56:3– 13
- Hilge M, Gloor SM, Rypniewski W, Sauer O, Heightman TD, Zimmermann W, Winterhalter K, Piontek K (1998) Structure 6:1433–1444
- 46. Shirai T, Ishida H, Noda J, Yamane T, Ozaki K, Hakamasa Y, Ito S (2001) J Mol Biol 310:1079–1087
- Palomares O, Villalba M, Rodriguez R (2003) Biochem J 369:593–601
- 48. In Gas4p and Phr3p the pro-rich motif is absent