Protein structure prediction for the male-specific region of the human Y chromosome

Krzysztof Ginalski*†, Leszek Rychlewski‡, David Baker§, and Nick V. Grishin*†¶

*Department of Biochemistry, University of Texas Southwestern Medical Center, 5323 Harry Hines Boulevard, Dallas, TX 75390-9038; ‡BioInfoBank Institute, ul. Limanowskiego 24A, 60-744, Poznan, Poland; §Howard Hughes Medical Institute and Department of Biochemistry, University of Washington, J Wing, Health Sciences Building, Box 357350, Seattle, WA 98195; and ¶Howard Hughes Medical Institute, University of Texas Southwestern Medical Center, 5323 Harry Hines Boulevard, Dallas, TX 75390-9050

Edited by John Kuriyan, University of California, Berkeley, CA, and approved December 22, 2003 (received for review September 30, 2003)

The complete sequence of the male-specific region of the human Y chromosome (MSY) has been determined recently; however, detailed characterization for many of its encoded proteins still remains to be done. We applied state-of-the-art protein structure prediction methods to all 27 distinct MSY-encoded proteins to provide better understanding of their biological functions and their mechanisms of action at the molecular level. The results of such large-scale structure-functional annotation provide a comprehensive view of the MSY proteome, shedding light on MSY-related processes. We found that, in total, at least 60 domains are encoded by 27 distinct MSY genes, of which 42 (70%) were reliably mapped to currently known structures. The most challenging predictions include the unexpected but confident 3D structure assignments for three domains identified here encoded by the *USP9Y*, *UTY*, and *BPY2* genes. The domains with unknown 3D structures that are not predictable with currently available theoretical methods are established as primary targets for crystallographic or NMR studies. The data presented here set up the basis for additional scientific discoveries in human biology of the Y chromosome, which plays a fundamental role in sex determination.

Due to an increasing gap between the overwhelming number of available protein sequences and experimentally determined protein structures, protein structure prediction has become an important venue with prolific applications in molecular biology (1). Continuous progress in this field has led to a variety of approaches applicable to structure-functional annotation of proteins. In particular, the recent advances in fold recognition (FR) and *ab initio* (AI) areas resulted in several methods that can reveal reliable but unexpected links between proteins (2, 3) defying standard approaches such as PSI-BLAST (4). FR/AI tools offer opportunities to advance annotation of poorly characterized proteins, providing valuable information to guide scientific discoveries.

Using a bouquet of state-of-art methods, we propose a coherent, semiautomatic strategy for structure-functional annotation of proteins and apply it to protein sequences encoded by the male-specific region of the human Y chromosome (MSY). For many years this distinctive segment of the human genome, which plays a critical role in sex determination, has been considered a functional wasteland. Complete sequence of the MSY, which comprises 95% of the length of the chromosome, revealed at least 78 protein-coding genes that collectively encode 27 distinct proteins (5). MSY genes participate in diverse processes such as skeletal growth, germ cell tumorigenesis, graft rejection, gonadal sex determination, and spermatogenic failure (6). The biological significance of the MSY has begun to surface in recent years; however, many protein-coding genes await more-detailed studies to understand their exact biological functions at the molecular level (7). Thus, comprehensive structural and functional annotation of the MSY-encoded proteins has a broad significance.

Methods

General Protocol. Sequences of all 27 distinct proteins demonstrated or hypothesized to be encoded by the MSY (5) first were subjected to Conserved Domain Database (CDD) (ref. 4; www.ncbi.nlm.nih.gov/Structure/cdd/wrpsb.cgi) and Simple Modular Architecture Research Tool (SMART) (ref. 8; http://smart.embl-heidelberg.de) searches to determine the conserved protein domains annotated in the SMART, Protein Families (Pfam) (9), and Clusters of Orthologous Groups (COG) (10) databases. This analysis also included identification of transmembrane segments [TMHMM2 (11)], signal peptides [SIGNALP (12)], low compositional complexity [CEG (13)], and coiled-coil [COILS2 (14)] regions, as well as regions containing internal repeats [PROSPERO (15)]. To define boundaries for regions with unknown structures that can be predicted easily by comparative modeling methods, the PDB-BLAST procedure [target sequence profile composed after five iterations of PSI-BLAST (4) on the nonredundant protein database run against the Protein Data Bank (PDB)] was applied. To avoid overprediction, which could mask other neighboring domains, regions containing multiple copies of the same structural motif and those that mapped to more than one domain in a template protein were also subjected to additional searches as single domains. Domains identified by CDD and/or SMART but not by PDB-BLAST, as well as all remaining regions, were subjected to the Structure Prediction Meta Server (ref. 16; http://bioinfo.pl/meta), which assembles various secondary structure prediction and top-of-the-line FR methods. These regions were divided further into single domains according to secondary structure predictions and preliminary results of FR searches and resubmitted to the Structure Prediction Meta Server. Collected models were screened with 3D-JURY (16), a consensus method of FR servers. Independently, all domains not annotated structurally with CDD and/or SMART but with clear predicted secondary structure patterns were also modeled AI by using the ROSETTA program (3). Final fold assignments were based on the similarity of ROSETTA and high-scoring 3D-JURY models, in addition to the compatibility of target-family-specific features (including predicted secondary structure) with characteristic features of the template/fold. Finally, domain boundaries for each region classified in Table 1 were assessed manually, taking into account all components of the performed analysis, which in many cases included 3D-model building.

Sequence-to-Structure Mapping for Difficult Targets. For both target and template sequences, close homologs were collected with PSI-BLAST searches and aligned by using PCMA (17) with final manual adjustments. Sequence-to-structure alignments for the target-template families were obtained by using the consensus alignment approach and 3D assessment (18). Structural consistency between high-scoring 3D-JURY predictions and ROSETTA models was taken into account in defining structurally conserved

This paper was submitted directly (Track II) to the PNAS office.

Abbreviations: FR, fold recognition; AI, *ab initio*; MSY, male-specific region of the Y chromosome; CDD, Conserved Domain Database; SMART, Simple Modular Architecture Research Tool; PDB, Protein Data Bank; TPR, tetratricopeptide repeat; HTH, helix–turn–helix.

†To whom correspondence may be addressed. E-mail: kginal@chop.swmed.edu or grishin@chop.swmed.edu.

Table 1. Domain architecture for products of 27 distinct MSY genes demonstrated or hypothesized to encode proteins

MSY sequence class	Gene name	GI number[†]	Protein length	Region[‡]	Classification[§]	PDB template[¶]
X-transposed	*TGIF2LY*	13161078	185	1–50	Unstructured region	
				51–127	**Homeodomain* (HOX, S)**	1LFU_P
				148–178	*Possibly zinc-binding domain*	
	PCDH11Y	13161060	1340[‖]	4–55	Transmembrane region	
				56–812	**7 Cadherin repeats* (CA, S)**	1L3W_A
				845–867	Transmembrane region	
				882–1340	Unstructured region, internal repeats	
X-degenerate	*SRY*	36605	204	1–55	Unstructured region	
				56–140	**High-mobility group* (HMG, S)**	1J46_A[††]
				141–204	Unstructured region	
	RPS4Y1[‡‡]/ *RPS4Y2*[‡‡]	337512/ 20269885	263/ 263	4–115	**S4 RNA-binding domain* (S4, S)**	*1FJG_D*
				118–152, 234–263	*Possibly OB-fold domain*	
				155–231	**KOW motif* (KOW, S)**	*1FFK_Q*
	ZFY	340436	801	1–413	Zfx/Zfy transcription activation region (Zfx_Zfy_act, P)	
				1–166	α/β region	
				169–301	*Possibly β-sandwich domain*	
				302–413	α/β region	
				418–796	**13 Zinc fingers* (ZnF_C2H2, S)**	1MEY_C
	AMELY	178531	192	1–17	Signal peptide	
				18–192	Amelogenin (Amelogenin, P)	
	TBL1Y	13161069	522	3–68	Lissencephaly type-1-like homology motif (LisH, S), *possibly similar to 1b0n_A*	
				79–133	α-Helical region	
				134–167	Unstructured region	
				168–522	**8 WD40 repeats* (WD40, S)**	1ERJ_A
	PRKY	2696012	277	12–272	**S/T protein kinase, catalytic domain* (S_TKc, S)**	1CTP_E
	USP9Y	2580558	2555	1–70	Unstructured region	
				71–868	*Possibly right-handed superhelix*	
				884–971	**Ubiquitin-like (β-grasp) domain**	**1BT0_A**
				972–1007	Unstructured region	
				1008–1532	*Possibly right-handed superhelix*	
				1553–1996	**Ubiquitin C-terminal hydrolase* (UCH, P), additional zinc ribbon subdomain (C1726, C1729, C1773, C1776)**	1NBF_A
				2004–2476	*Possibly right-handed superhelix*	
				2477–2555	Unstructured region	
	DBY	2580556	660	20–141	Unstructured region	
				179–556	**DEAD-like helicase* (DEXDc, S) Helicase C-terminal domain* (HELICc, S)**	1HV8_A
				579–660	Unstructured region	
	UTY	2580574	1347[‖]	71–396	**9 Tetratricopeptide repeats* (TPR, S)**	1NA0_A
				451–536	Unstructured region	
				888–1003	α/β region	
				1039–1211	**Jumonji domain* (JmjC, S)**	*1MZE_A*
				1215–1268	α-Helical region	
				1275–1342	**Treble-clef zinc finger**	**1ZBD_B**
	TMSB4Y	2580564	44	2–41	**Thymosin β-actin-binding motif* (THY, S)**	1HJ0_A
	NLGN4Y	4589546	648	1–433	**Carboxylesterase* (Coesterase, P)**	1F8U_A
				446–502	Unstructured region	
				507–529	Transmembrane region	
				550–615	$\alpha + \beta$ region	
				616–648	Unstructured region	
	Cyorf15A	13161081	220		?[§§]	
	Cyorf15B	13161084	181	1–115	**Coiled-coil region**	*2TMA_A*
				116–181	Unstructured region	
	SMCY	1661016	1539	13–54	Small domain found in the jumonji family of transcription factors (JmjN, S), $\alpha + \beta$ region	
				67–185	**A/T-rich interaction domain* (BRIGHT, S)**	1KQQ_A
				186–221	Unstructured region	
				222–306	α/β region	
				317–362	**PHD zinc finger* (PHD, S)**	1F62_A

Table 1. (continued)

MSY sequence class	Gene name	GI number†	Protein length	Region‡	Classification§	PDB template¶
				382–432	α/β region	
				458–627	**Jumonji domain* (JmjC, S)**	*1MZE_A*
				632–690	α-Helical region	
				691–774	C5HC2 zinc finger (zf-C5HC2, P), α/β region	
				779–1156	α-Helical region	
				1171–1239	**PHD zinc finger* (PHD, S)**	*1FP0_A*
				1240–1308	α-Helical region	
				1309–1354	Unstructured region	
				1355–1532	α-Helical region	
	EIF1AY	2580560	144	2–131	**Eukaryotic translation initiation factor 1A* (eIF1a, S)**	1D7Q_A††
Ampliconic	*TSPY*	292429	253	20–247	Nucleosome assembly protein (NAP, P), $\alpha + \beta$ region	
	VCY	2580544	125	1–125	Unstructured region	
	XKRY	2580580	159	1–159	Transmembrane protein	
	CDY	4558754	541	4–62	**Chromatin organization modifier domain* (CHROMO, S)**	1G6Z_A
				63–114	Unstructured region	
				115–162	α-Helical region	
				199–280	**Possibly β-sandwich domain**	
				282–541	**Enoyl-CoA hydratase/isomerase* (ECH, P)**	1DUB_A
	HSFY	13161090	401‖	76–194	**Heat-shock factor* (HSF, S)**	1HKS
				195–224	Unstructured region	
				225–356	$\alpha + \beta$ region	
				357–401	Unstructured region	
	RBMY	452367	496	8–82	**RNA recognition motif* (RRM, S)**	1CVJ
				83–496	Unstructured region, internal repeats	
	PRY	21270256	147	4–143	α/β region	
	BPY2	2580546	106	21–98	**Winged HTH-like domain**	**1AOY**
	DAZ	9651955	558¶¶	20–122	**RNA recognition motif* (RRM, S)**	2UP1_A
				123–540	Unstructured region, internal repeats	

†GI number of the corresponding protein product.
‡Region boundaries are estimated manually based on secondary structure prediction, tertiary fold recognition, and SMART/CDD searches. For regions that can be modeled by using available structural information, these can also include residues (present in template protein) that are located outside the structural domain. Regions <30 residues and those with the most ambiguous assignments are not listed.
§Regions for which 3D structure can be predicted with confidence are shown in bold type, possible structural assignments are denoted in italic type, and the most difficult but reliable are underlined. For domains annotated in SMART (S) or PFAM (P), names of entry and database are given in parentheses; the asterisk stands for available structural information. As a necessary disclaimer, the database entry name may not correspond to the exact function of the protein in question.
¶PDB ID codes of the template structures not detectable by PDB-BLAST but SMART and/or CDD are shown in italic type, those detected only by FR/AI techniques (3D-JURY/ROSETTA) are shown in bold type.
‖Length of the longest splice variant.
††Structure solved for the analyzed MSY protein sequence.
‡‡Protein products of these two isoforms display 93% of sequence identity.
¶¶Length of the longest family member.
§§Possible protein sequence errors due to incorrect assignment of intron/exon boundaries.

regions (for which alignment is meaningful) between target sequence and template(s).

Results and Discussion

Structure-Functional Classification of the MSY-Encoded Proteins. We analyzed the sequences of 27 distinct MSY-encoded proteins by using standard sequence-comparison tools such as PSI-BLAST, RPS-BLAST [CDD (4)], and profile hidden Markov modeling [SMART (8)], as well as the state-of-the-art approaches in FR [3D-JURY (16)] and AI [ROSETTA (3)], which are proven to be some of the best-performing methods in the fifth round of the Critical Assessment of Techniques for Protein Structure Prediction (CASP5) (19). The results of this structure-functional annotation are presented in Table 1 and summarized in Fig. 1. Table 1 illustrates what human expertise can accomplish with the aid of the currently available automatic methods and reports the key findings of our analysis. Importantly, because the majority of MSY proteins are modular, a complete understanding of the specific role played by each requires identification and characterization of all enclosed domains.

The application of PDB-BLAST allowed for detection of 31 domains of known structure, which in total encompass 4,446 (31%) of the analyzed 14,171 amino acids encoded by all 27 distinct MSY genes. In many of these cases, detailed sequence analysis combined with 3D-model building enabled us to redefine the exact number of repetitive domains or motifs contained within MSY-encoded proteins. In particular, we show that the ubiquitously transcribed tetratricopeptide repeat (TPR) protein on the Y chromosome (UTY) includes as many as nine TPRs. Interestingly, we have also detected as many as eight WD40 repeats in the C-terminal region of transducin β-like 1 Y protein (TBL1Y) that possibly forms an eight-bladed β-propeller in contrast to the structurally homologous protein most similar in sequence, the C-terminal WD40 domain of Tup1 (20), which has a seven-bladed β-propeller structure. SMART/CDD searches assigned 3D structure to eight more domains covering 905

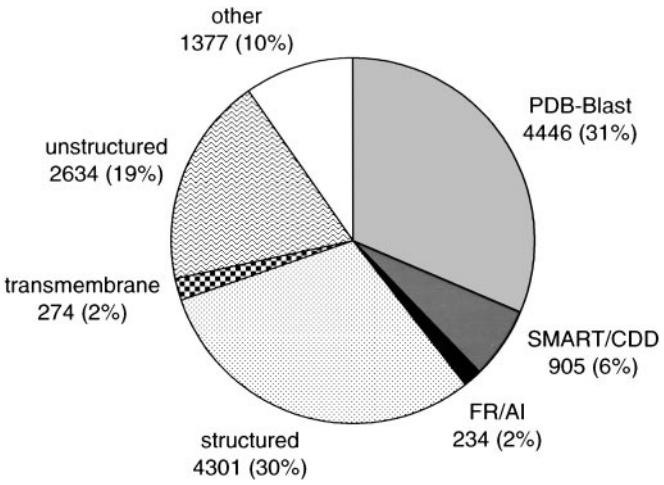

Fig. 1. Summary of structure prediction for the complete set of proteins encoded by 27 distinct MSY genes. The following classes together with the number and the percentage of encompassed amino acids are presented: 3D structure assigned with PDB-BLAST; 3D structure assigned with SMART/CDD; 3D structure assigned with FR/AI methods; structured regions with clear secondary structure prediction patterns but not annotated at the 3D level (include separate domains as well as regions that possibly pack on the neighboring domains); transmembrane regions; unstructured nonglobular regions; and other, remaining regions encompassing segments <30 residues (mainly linkers between domains) and regions that could not be assigned with confidence to any of the former classes.

residues (6%). Reliability of these hits was confirmed further with the consensus of FR methods, 3D-JURY meta predictor, which assigned an above-threshold confidence scores (2) in a majority of these cases. For an additional three domains [234 amino acids (2%)] identified in this study, the tertiary structure was predicted confidently by using both FR and AI approaches. Although these predictions appeared in the 3D-JURY system as weak hits with below-threshold scores, structures similar to the highest-scoring 3D-JURY models were obtained independently with the ROSETTA program. Transmembrane segments [274 amino acids (2%)] were detected in four proteins including a testis-specific XK-related protein Y (XKRY), a putative membrane transport protein. In addition, secondary structure-rich regions, which with all likelihood form compact globular domains, covered 4,301 amino acids (30%). Because no confident structural assignments could be made with currently available computational methods, these regions await experimental (NMR or crystallographic) studies. Importantly, a majority of these domains were identified in this study. For seven of these domains, hypotheses about their possible folds were suggested (Table 1). For example, taking into account potential domain insertions resulted in the detection of a previously uncharacterized domain in both Y isoforms of ribosomal protein S4 homologue (RPS4Y) that may form an oligonucleotide/oligosaccharide-binding (OB) fold structure. Interestingly, as much as 19% of encoded residues (2,634 amino acids) corresponds to potentially unstructured nonglobular regions, including the whole sequence of a testis-specific variably charged protein Y (VCY). The remaining 10% (1,377 amino acids) includes all segments <30 residues (mainly linkers between domains) as well as regions that could not be assigned with confidence to any of the former classes. In conclusion, 27 distinct MSY genes encode at least 60 domains, of which 42 (70%) were mapped reliably to currently known structure space.

Biological Significance of the MSY-Encoded Proteins. With the structure-functional annotation of MSY-encoded proteins, a coherent view of their specific biological roles begins to emerge. Importantly, a majority of these proteins, particularly those directly involved in sex determination or spermatogenesis, are responsible for regulation of gene expression on several different levels such as transcription, pre-mRNA processing, and translation. First, a number of DNA-binding domains were detected in several proteins encoded by MSY genes, such as SRY (HMG), HSFY (HSF), ZFY (Zfx_Zfy_act and ZnF_C2H2) or SMCY (BRIGHT) (see Table 1), which act as transcriptional regulators. MSY-encoded proteins such as UTY (TPR) or TBL1Y (WD40 repeats) participate in protein–protein interactions important for assembly and activity of multicomponent complexes involved in transcriptional repression (21). Some of the identified domains (e.g., JmjC) have a probable regulatory role in these complexes. Because eukaryotic gene regulation occurs within the context of chromatin, a few MSY genes encode domains taking part in histone binding (N-terminal region of TBL1Y) or histone acetylation (ECH, which in CDY protein is acetyltransferase) (22). In addition, the CDY protein contains a CHROMO domain, which by altering the structure of chromatin plays a critical role in mammalian spermatogenesis in histone-to-protamine transition. Second, several Y-linked proteins regulate gene expression at the level of pre-mRNA processing, including RBMY (RRM) and possibly DBY (helicase domain) (23). Third, some MSY-encoded proteins seem to be required for a maximal rate of protein biosynthesis [e.g., translation initiation factor 1A Y

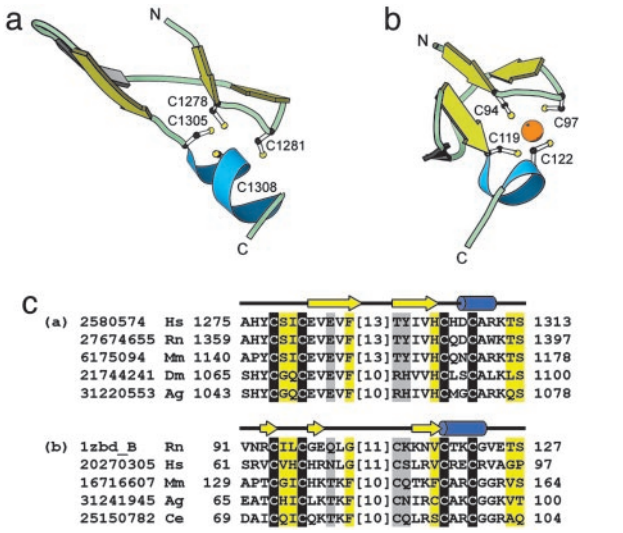

Fig. 2. THe C-terminal domain of UTY is a treble-clef zinc finger. (*a*) ROSETTA 3D model of C-terminal domain of UTY (GI:2580574). The side chains of Cys-1278, Cys-1281, Cys-1305, and Cys-1308 that are predicted to take part in coordination of zinc ion are shown. (*b*) Similar structure of *Rattus norvegicus* effector domain of Rabphilin-3A (PDB ID code 1zbd) (39) selected independently with the 3D-JURY method. Cys-94, Cys-97, Cys-119, and Cys-122 side chains, as well as coordinated zinc ion (orange), are presented. (*c*) The sequence alignment of representative sequences belonging to UTY and Rabphilin-3A families. Regions that could not be aligned with confidence by using the consensus alignment approach and 3D assessment, as well as those that may not be structurally conserved, are not shown. The numbers in square brackets specify the number of excluded residues. Uncharged residues in mostly hydrophobic sites are highlighted in yellow, polar residues at mostly hydrophilic sites are highlighted in light gray, and small residues at positions occupied by mostly small residues are shown in red letters. Conserved cysteine residues forming a zinc-binding site are highlighted in black. Locations of the secondary structure elements in UTY (consensus of secondary structure predictions) and Rabphilin-3A are marked above the sequences. The color shading of secondary structure elements corresponds to those in the respective structural diagrams. Secondary structure elements not shown in the alignment panel but presented in structural diagrams are colored white. The same presentation scheme is used for Figs. 3 and 4.

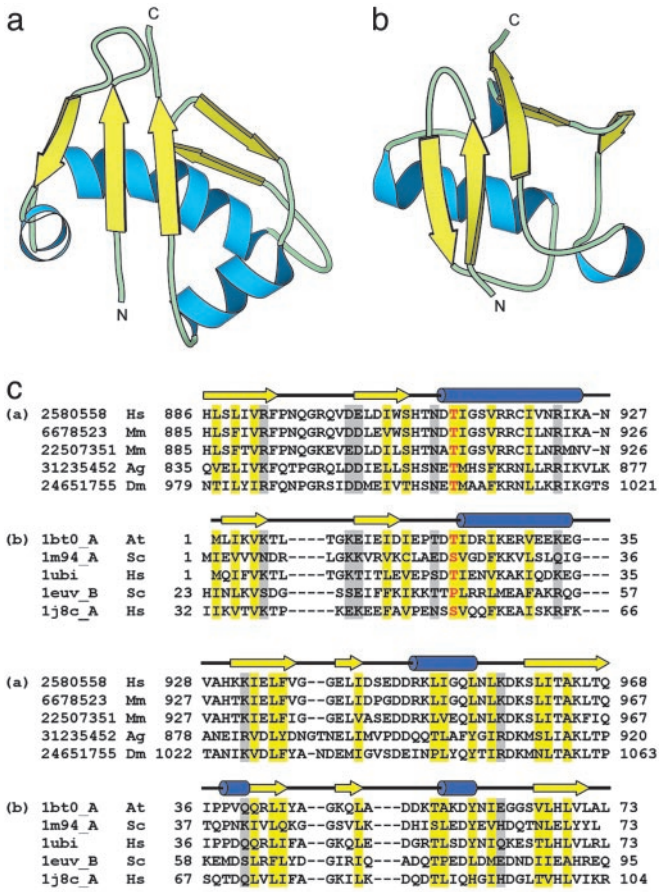

Fig. 3. USP9Y encloses the domain that belongs to the superfamily of ubiquitin-like proteins. (*a*) ROSETTA 3D model of the β-grasp (ubiquitin-like) domain of USP9Y (GI:2580558). (*b*) Similar structure of *Arabidopsis thaliana* ubiquitin-like protein 7 (Rub1) (PDB ID code 1bt0) (40) selected independently with the 3D-JURY method. (*c*) The sequence alignment of representative sequences belonging to USP9Y and Rub1 families.

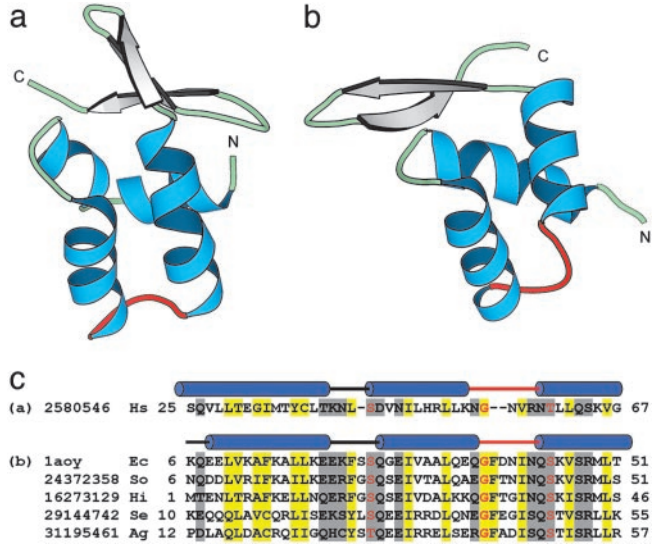

Fig. 4. BPY2 forms a winged HTH-like structure. (*a*) ROSETTA 3D model of BPY2 (GI:2580546). (*b*) Similar structure of *Escherichia coli* N-terminal DNA-binding domain of arginine repressor (PDB ID code 1aoy) (37) selected independently with the 3D-JURY method. (*c*) The sequence alignment of BPY2 and representative sequences belonging to the arginine repressor family. The turn in an HTH-like motif is shown in red.

(EIF1AY)]. Regulatory roles of the genes implicated in spermatogenesis also can be achieved at the level of the protein turnover, which is controlled by ubiquitin-specific protease 9 Y (USP9Y) (24).

Genes, which do not seem to be directly involved in sex determination or spermatogenesis, seem to play crucial roles in developmental processes. These genes are likely to enhance reproductive function and performance in sperm competition, because their functions may provide an advantage in male-to-male contest (25). In particular, genes such as *AMELY* and *TMSB4Y* play important roles in tooth development and the organization of the cytoskeleton, respectively (26). Two other genes expressed predominantly in the brain (*PCDH11Y* and *NGLN4Y*) encode cell-surface proteins involved in cell–cell interactions and cell adhesion (27). These genes thus may provide a basis for sexually dimorphic features such as stature, tooth development, or behavior (brain), which could influence the ability to attract a partner.

Prediction Highlights. The most challenging domain predictions for us were unexpected but confident structural assignments for three domains (encoded by the *UTY*, *USP9Y*, and *BPY2* genes) identified in this study. Prediction of the tertiary structure for these domains adds to their functional characterization; however, exact roles and detailed mechanisms of their action need to be elucidated through additional biochemical experiments. Discussion of these three domains follows.

The C-terminal domain of UTY is a treble-clef zinc finger. UTY protein encoded by the X-degenerate *UTY* gene starts from nine TPRs shown to be responsible for the protein–protein interactions with the N-terminal Q domain of TLE1 (28). UTY also contains the Jumonji (JmjC) domain, which is homologous to an aspartyl hydrolase enzyme [factor-inhibiting HIF-1 (FIH-1)] of known structure (29). In addition to these previously described domains, we identified an uncharacterized C-terminal domain as a treble-clef zinc finger (30) with conserved cysteine residues (Cys-1278, Cys-1281, Cys-1305, and Cys-1308) taking part in the coordination of a zinc ion (Fig. 2). With the evidence that mammalian UTY and TLE proteins may form a transcription repressor complex and mediate repression mechanisms to some extent similar to those performed by SSN6-TUP1 in yeast; a unique biological role of the SSN6 mammalian counterpart, UTY, mediated through the JmjC and zinc-finger domains emerges. While JmjC has a probable regulatory function, the treble-clef zinc-finger domain may be responsible for direct DNA binding or for interactions with other proteins such as DNA-binding factors or other elements of the repressor complex. Interestingly, another MSY protein, TBL1Y, displays structural and functional similarities to TUP1 and Groucho/TLE corepressors, sharing with them WD40 repeats as well as the ability to interact with histones. In addition, SSN6 has been shown to interact through its TPR repeats with the DNA-binding homeodomain (HOX) of the protein Matα2 (31), and this domain is also encoded by one of the MSY genes, *TGIF2LY*. This rather unlikely coincidence raises the exciting possibility that these three MSY-encoded proteins could form a common repression complex.

USP9Y encloses ubiquitin-like domain. Widely expressed in embryonic and adult tissues, the *USP9Y* (32) gene is known to encode ubiquitin-specific protease 9 Y (USP9Y), which contains a ubiquitin C-terminal hydrolase domain. Involvement of *USP9Y* in male infertility emphasizes a special requirement for certain components of the ubiquitin system in spermatogenesis. *USP9Y*, a member of a family of deubiquitinating genes, thus may play an important

BIOPHYSICS

regulatory role at the level of protein turnover by preventing degradation of proteins by the proteasome through the removal of ubiquitin from protein–ubiquitin conjugates, similar to its *Drosophila melanogaster* homolog *FAF* (33). Interestingly, we found four cysteine residues (Cys-1726, Cys-1729, Cys-1773, and Cys-1776) in the Fingers domain of USP9Y ubiquitin C-terminal hydrolase that may coordinate a zinc ion. These cysteines present in the region forming the zinc ribbon-like structure are absent in the structurally homologous protein most similar in sequence, the catalytic core domain of HAUSP (34). We also detected three previously uncharacterized, long α-helical regions located on both sides of the ubiquitin C-terminal hydrolase domain, which may form a right-handed superhelical structure. The most unexpected finding was detection and structural characterization of another previously unknown domain located in the N-terminal region of USP9Y between the first two α-helical regions. This domain has a β-grasp fold characteristic of ubiquitin-like proteins (Fig. 3). Moreover, we argue that this ubiquitin-like domain is a distant homolog of other ubiquitin-like proteins, and we hypothesize that its function is to target the USP9Y protein to its specific cellular localization. Taking a possible regulatory role of USP9Y in protecting proteins from being degraded by the proteasome, the β-grasp domain may tether the ubiquitin C-terminal hydrolase to the proteasome through an interaction with ubiquitin-binding sites; however, without additional experimental evidence, other possible roles (including direct inhibition of ubiquitin hydrolase domain) cannot be excluded unequivocally.

BPY2 forms a winged helix–turn–helix (HTH)-like structure. Expressed exclusively in testis basic protein Y 2 (BPY2) is likely to function in male germ cell development because of its specific localization in germ cell nuclei. Involvement of the *BPY2* gene in the pathogenesis of male infertility (35) as well as in prostate cancer (36) has been suggested, but little is known about the specific role of its encoded protein. Importantly, this protein represents singleton without detectable sequence homologs. We predicted that BPY2 forms a winged HTH-like domain with a 3D structure similar to the N-terminal DNA-binding domain of arginine repressor (37) (Fig. 4). The sequence-to-structure alignment in Fig. 4*c* encompasses only the N-terminal region of BPY2 with the HTH-like motif formed by the second and third α-helices, because considerable ambiguity exists in obtaining a reliable mapping within the C-terminal β-hairpin. However, this finding points to a possible role of this highly charged protein in DNA or RNA binding through the HTH-like motif. In addition, previous experimental studies show that the BPY2 protein interacts with the HECT domain of ubiquitin-protein ligase E3A (UBE3A) and that UBE3A ubiquitination may be required for BPY2 function (38).

Conclusions

The data presented in this study provide a comprehensive view of the proteins encoded by MSY genes, which have been implicated in several human diseases such as Turner syndrome, gonadal sex reversal, spermatogenic failure, and gonadoblastoma. Importantly, knowledge of 3D structure for MSY-encoded proteins is a prerequisite for a better understanding of Y-specific biological processes, providing some level of insight into their molecular functions, mechanisms of action, and substrate specificities and aiding in the design of experiments. In addition, identification of domains for which tertiary structure is not (confidently) predictable with the currently available theoretical approaches is of importance for crystallographers or NMR spectroscopists. These domains including, among others, whole proteins encoded by *TSPY* and *PRY* genes become primary targets for structural studies and may encompass new folds. The structural and functional description of the MSY-encoded proteins presented here sets up a basis for additional biological discoveries in human biology.

We thank Lisa N. Kinch for critical reading of the manuscript. This work was supported by National Institutes of Health Grant GM67165 (to N.V.G.).

1. Baker, D. & Sali, A. (2001) *Science* **294,** 93–96.
2. Ginalski, K. & Rychlewski, L. (2003) *Nucleic Acids Res*. **31,** 3291–3292.
3. Bonneau, R., Tsai, J., Ruczinski, I., Chivian, D., Rohl, C., Strauss, C. E. & Baker, D. (2001) *Proteins*, Suppl. 5, 119–126.
4. Altschul, S. F., Madden, T. L., Schaffer, A. A., Zhang, J., Zhang, Z., Miller, W. & Lipman, D. J. (1997) *Nucleic Acids Res*. **25,** 3389–3402.
5. Skaletsky, H., Kuroda-Kawaguchi, T., Minx, P. J., Cordum, H. S., Hillier, L., Brown, L. G., Repping, S., Pyntikova, T., Ali, J., Bieri, T., *et al*. (2003) *Nature* **423,** 825–837.
6. Vogt, P. H., Affara, N., Davey, P., Hammer, M., Jobling, M. A., Lau, Y. F., Mitchell, M., Schempp, W., Tyler-Smith, C., Williams, G., *et al*. (1997) *Cytogenet. Cell Genet*. **79,** 1–20.
7. Lahn, B. T. & Page, D. C. (1997) *Science* **278,** 675–680.
8. Schultz, J., Milpetz, F., Bork, P. & Ponting, C. P. (1998) *Proc. Natl. Acad. Sci. USA* **95,** 5857–5864.
9. Bateman, A., Birney, E., Cerruti, L., Durbin, R., Etwiller, L., Eddy, S. R., Griffiths-Jones, S., Howe, K. L., Marshall, M. & Sonnhammer, E. L. (2002) *Nucleic Acids Res*. **30,** 276–280.
10. Tatusov, R. L., Natale, D. A., Garkavtsev, I. V., Tatusova, T. A., Shankavaram, U. T., Rao, B. S., Kiryutin, B., Galperin, M. Y., Fedorova, N. D. & Koonin, E. V. (2001) *Nucleic Acids Res*. **29,** 22–28.
11. Sonnhammer, E. L., von Heijne, G. & Krogh, A. (1998) *Proc. Int. Conf. Intell. Syst. Mol. Biol*. **6,** 175–182.
12. Nielsen, H., Engelbrecht, J., Brunak, S. & von Heijne, G. (1997) *Protein Eng*. **10,** 1–6.
13. Wootton, J. C. (1994) *Comput. Chem*. **18,** 269–285.
14. Lupas, A., Van Dyke, M. & Stock, J. (1991) *Science* **252,** 1162–1164.
15. Mott, R. (2000) *J. Mol. Biol*. **300,** 649–659.
16. Ginalski, K., Elofsson, A., Fischer, D. & Rychlewski, L. (2003) *Bioinformatics* **19,** 1015–1018.
17. Pei, J., Sadreyev, R. & Grishin, N. V. (2003) *Bioinformatics* **19,** 427–428.
18. Ginalski, K. & Rychlewski, L. (2003) *Proteins* **53,** 410–417.
19. Kinch, L. N., Wrabl, J. O., Krishna, S. S., Majumdar, I., Sadreyev, R. I., Qi, Y., Pei, J., Cheng, H. & Grishin, N. V. (2003) *Proteins* **53,** 395–409.
20. Sprague, E. R., Redd, M. J., Johnson, A. D. & Wolberger, C. (2000) *EMBO J*. **19,** 3016–3027.
21. Yoon, H. G., Chan, D. W., Huang, Z. Q., Li, J., Fondell, J. D., Qin, J. & Wong, J. (2003) *EMBO J*. **22,** 1336–1346.
22. Lahn, B. T., Tang, Z. L., Zhou, J., Barndt, R. J., Parvinen, M., Allis, C. D. & Page, D. C. (2002) *Proc. Natl. Acad. Sci. USA* **99,** 8707–8712.
23. Venables, J. P., Elliott, D. J., Makarova, O. V., Makarov, E. M., Cooke, H. J. & Eperon, I. C. (2000) *Hum. Mol. Genet*. **9,** 685–694.
24. Lee, K. H., Song, G. J., Kang, I. S., Kim, S. W., Paick, J. S., Chung, C. H. & Rhee, K. (2003) *Reprod. Fertil. Dev*. **15,** 129–133.
25. Roldan, E. R. & Gomendio, M. (1999) *Trends Ecol. Evol*. **14,** 58–62.
26. Salido, E. C., Yen, P. H., Koprivnikar, K., Yu, L. C. & Shapiro, L. J. (1992) *Am. J. Hum. Genet*. **50,** 303–316.
27. Blanco, P., Sargent, C. A., Boucher, C. A., Mitchell, M. & Affara, N. A. (2000) *Mamm. Genome* **11,** 906–914.
28. Grbavec, D., Lo, R., Liu, Y., Greenfield, A. & Stifani, S. (1999) *Biochem. J*. **337,** 13–17.
29. Dann, C. E., III, Bruick, R. K. & Deisenhofer, J. (2002) *Proc. Natl. Acad. Sci. USA* **99,** 15351–15356.
30. Krishna, S. S., Majumdar, I. & Grishin, N. V. (2003) *Nucleic Acids Res*. **31,** 532–550.
31. Smith, R. L., Redd, M. J. & Johnson, A. D. (1995) *Genes Dev*. **9,** 2903–2910.
32. Brown, G. M., Furlong, R. A., Sargent, C. A., Erickson, R. P., Longepied, G., Mitchell, M., Jones, M. H., Hargreave, T. B., Cooke, H. J. & Affara, N. A. (1998) *Hum. Mol. Genet*. **7,** 97–107.
33. Huang, Y., Baker, R. T. & Fischer-Vize, J. A. (1995) *Science* **270,** 1828–1831.
34. Hu, M., Li, P., Li, M., Li, W., Yao, T., Wu, J. W., Gu, W., Cohen, R. E. & Shi, Y. (2002) *Cell* **111,** 1041–1054.
35. Tse, J. Y., Wong, E. Y., Cheung, A. N., O, W. S., Tam, P. C. & Yeung, W. S. (2003) *Biol. Reprod*. **69,** 746–751.
36. Perinchery, G., Sasaki, M., Angan, A., Kumar, V., Carroll, P. & Dahiya, R. (2000) *J. Urol*. **163,** 1339–1342.
37. Sunnerhagen, M., Nilges, M., Otting, G. & Carey, J. (1997) *Nat. Struct. Biol*. **4,** 819–826.
38. Wong, E. Y., Tse, J. Y., Yao, K. M., Tam, P. C. & Yeung, W. S. (2002) *Biochem. Biophys. Res. Commun*. **296,** 1104–1111.
39. Ostermeier, C. & Brunger, A. T. (1999) *Cell* **96,** 363–374.
40. Rao-Naik, C., delaCruz, W., Laplaza, J. M., Tan, S., Callis, J. & Fisher, A. J. (1998) *J. Biol. Chem*. **273,** 34976–34982.