Genomics

Updated catalogue of homologues to human disease-related proteins in the yeast genome

Miguel A. Andrade[1,a], Chris Sander[b], Alfonso Valencia[a,*]

[a] *Protein Design Group, CNB-CSIC, Campus Universidad Autonoma, Cantoblanco, E-28049 Madrid, Spain*
[b] *European Bioinformatics Institute, EMBL, Hinxton, Cambridge, UK*

Received 25 February 1998

Abstract The recent availability of the full *Saccharomyces cerevisiae* genome offers a perfect opportunity for revising the number of homologues to human disease-related proteins. We carried out automatic analysis of the complete *S. cerevisiae* genome and of the set of human disease-related proteins as identified in the SwissProt sequence data base. We identified 285 yeast proteins similar to 155 human disease-related proteins, including 239 possible cases of human-yeast direct functional equivalence (orthology). Of these, 40 cases are suggested as new, previously undiscovered relationships. Four of them are particularly interesting, since the yeast sequence is the most phylogenetically distant member of the protein family, including proteins related to diseases such as phenylketonuria, lupus erythematosus, Norum and fish eye disease and Wiskott-Aldrich syndrome.

© 1998 Federation of European Biochemical Societies.

Key words: Human disease; Yeast genome; Sequence analysis; Function prediction

1. Introduction

1.1. Human disease-related proteins

A number of human diseases are caused by malfunction of a gene or group of genes. One approach to the understanding of these human genetic diseases is the discovery and study of associated genes. As displayed in the last version of the SwissProt protein data base [1] (release 35.0, December 1997), about a tenth of the 6000 human genes known to date are annotated as disease-related. Although their degree of implication in the different diseases is not well understood in all cases, they constitute a large resource of computerised information on the molecular basis of human disease, whose characterisation may reveal valuable new knowledge.

1.2. Genes of equivalent function between different organisms

Equivalent genes in different organisms are routinely identified by sequence analysis on the assumption of a relationship between sequence similarity and a common evolutionary origin. The correspondence between protein sequence similarity and functional equivalence is not, however, straightforward. Whereas sequence similarity between two proteins can be measured with a scalar variable, e.g. percentage of sequence similarity in an alignment of the two sequences, functional similarity is a qualitative concept not well described in a single quantity. The level of sequence similarity that assures that two proteins have the same function changes for different families in a manner that is not well understood.

One approach to this problem is to consider two proteins from two different organisms as functionally equivalent (orthologues [2]) if there is more similarity between them than to any other protein of the two organisms (see [3,4]). Orthologous proteins are assumed to have originated before the phylogenetic split of the organisms compared. On the contrary, similar proteins of the same organism originated by a duplication event are called paralogues. Orthology and paralogy are therefore concepts relative to the organisms under comparison. The discrimination of orthologues from paralogues can only be strictly applied when the proteins belong to fully sequenced genomes, because only then it is clear that no other more directly related proteins are to be found.

The availability of the complete genome of the yeast *Saccharomyces cerevisiae* [5] offers, for the first time, the possibility of identifying systematically putative orthologues to human disease-related proteins (HDRPs) in a eukaryotic genome. The yeast genome contains more than 6000 non-overlapping genes (a small number compared to the 10^5 genes that are probably contained in the full human genome). Given that the human genome is far from being completely known, the orthology checking is necessarily incomplete. Analysis of the set of putative orthologues may, however, allow better and easier characterisation of the corresponding HDRPs and of the mechanisms by which they cause disease.

2. Materials and methods

We used the GeneQuiz system [6,7] for sequence analysis of 686 HDRPs annotated as disease-related in the SwissProt protein sequence data base [1] and of 6284 yeast proteins as provided by MIPS [5]. GeneQuiz is an automated system for large-scale sequence analysis that combines several sequence similarity and protein feature

*Corresponding author. Fax: (34) (1) 585 45 06.
E-mail: valencia@cnb.uam.es

[1] Present address: European Bioinformatics Institute, EMBL, Hinxton, Cambridge, UK.

Abbreviations: HDRP, human disease-related protein; WAS, Wiskott-Aldrich syndrome; WH, WAS homology; GBD, GTPase binding domain; snRNP, small nuclear ribonucleoproteins; aa, amino acids

Notation: nucleotide and protein sequences identifiers are given as DATABASE:IDENTIFIER where the codes for the corresponding data bases are SW (for SwissProt), EMBL, TREMBL, SPTREMBL, PIR, and OMIM. The corresponding sequences can be retrieved through the web using SRS (T. Etzold, http://www.ebi.ac.uk/srs/srsc), and the OMIM data base (B. Brylawski, http://www.ncbi.nlm.nih.gov/Omim/). All yeast sequences are tagged by their MIPS gene identifier (Yxynnnz, where x indicates chromosome, y stands for the arm, nnn is a numeral, and z indicates the direction of translation).

0014-5793/98/$19.00 © 1998 Federation of European Biochemical Societies. All rights reserved.
PII S0014-5793(98)00277-4

Table 1
Forty pairs of possible new relations between yeast and HDRPs

HDRP				Yeast protein			
SwissProt identifier	Description	Related disease	OMIM	Similarity	MIPS	Data base identifier	Description
3O5B_HUMAN	3-oxo-5-β-steroid 4-dehydrogenase	neonatal cholestatic hepatitis	235555	2.1×10^{-86}	YHR104W	swiss|P38715|YHQ4_YEAST	hypothetical 37.1 kDa protein in NRK1-CDC12 intergenic region
AF17_HUMAN	AF-17 protein	acute leukaemias	600328	1.6×10^{-29}	YPR031W	trembl|Z49274|SC9367_11	product: 'unknown'; *S. cerevisiae* chromosome XVI cosmid 9367
AF9_HUMAN	AF-9 protein	acute leukaemias	159558	1.7×10^{-19}	YNL107W	swiss|P53930|YNK7_YEAST	hypothetical 26.0 kDa protein in CYB5-LEU4 intergenic region
ALDR_HUMAN	aldose reductase	diabetes and galactosaemia	103880	1.4×10^{-72}	YJR096W	swiss|P47137|YJ66_YEAST	hypothetical 32.3 kDa protein in ACR1-YUH1 intergenic region
				3.8×10^{-60}	YBR149W	swiss|P38115|YBZ9_YEAST	hypothetical 38.9 kDa protein in YSW1-RIB7 intergenic region
ANK1_HUMAN	ankyrin R	hereditary spherocytosis	182900	4.0×10^{-19}	YGR232W	swiss|P50086|YG4X_YEAST	hypothetical 25.6 kDa protein in SMI1-PHO81 intergenic region
AQP2_HUMAN	aquaporin-CD	nephrogenic diabetes insipidus	222000	4.6×10^{-17}	YFL054C	swiss|P43549|YFF4_YEAST	hypothetical 70.5 kDa protein in THI5-EMP47 intergenic region
BCR_HUMAN	breakpoint cluster region protein	chronic myeloid leukaemia, acute myeloid leukaemia, and acute lymphoblastic leukaemia	151410	1.0×10^{-20}	YBR260C	swiss|P38339|YB9G_YEAST	hypothetical 74.6 kDa protein in RIB5-SHM1 intergenic region
BPA1_HUMAN	bullous pemphigoid antigen 1	subepidermal blistering disease bullous pemphigoid	113810	2.4×10^{-10}	YMR159C	swissnew|Q03818|YM34_YEAST	hypothetical 17.2 kDa protein in IMP1-HLJ1 intergenic region
CGL_HUMAN	cystathionine γ-lyase	cystathioninuria	219500	2.6×10^{-44}	YGL184C	swiss|P53101|YGT4_YEAST	hypothetical 51.8 kDa protein in COX4-GTS1 intergenic region
				8.6×10^{-35}	YHR112C	swiss|P38716|YHR2_YEAST	hypothetical 42.4 kDa protein in CDC12-ORC6 intergenic region
CPT2_HUMAN	mitochondrial carnitine palmitoyltransferase II precursor	myoglobinuria	255110	5.5×10^{-35}	YER024W	swiss|P40017|YEL4_YEAST	hypothetical 103.3 kDa protein in PRO3-GCD11 intergenic region
DHAM_HUMAN	aldehyde dehydrogenase, mitochondrial precursor	acute alcohol intoxication	100650	1.1×10^{-165}	YER073W	trembl|U18814|SC3612_1	gene: 'YER073w'; product: 'Yer073p'; *S. cerevisiae* chromosome V lambda clone 3612 and cosmid 9747
DHB3_HUMAN	oestradiol 17β-dehydrogenase 3	male pseudohermaphrodism	264300	5.6×10^{-36}	YBR159W	swiss|P38286|YB09_YEAST	hypothetical 38.7 kDa protein in RPB5-CDC28 intergenic region
DHSA_HUMAN	succinate dehydrogenase	Leigh syndrome	600857	4.2×10^{-25}	YEL047C	swiss|P32614|YEF7_YEAST	hypothetical 50.8 kDa protein in PAU2-GLY1 intergenic region
EP15_HUMAN	epidermal growth factor receptor substrate 15	acute leukaemias	600051	5.4×10^{-32}	YBL047C	swiss|P34216|YBE7_YEAST	hypothetical 150.8 kDa protein in SEC17-QCR1 intergenic region
FVT1_HUMAN	follicular variant translocation protein 1 precursor	follicular lymphoma	136440	6.1×10^{-10}	YBR265W	swiss|P38342|YB9K_YEAST	hypothetical 36.0 kDa protein in SHM1-MRPL37 intergenic region
HIRA_HUMAN	Hira protein	DiGeorge syndrome	600237	2.7×10^{-14}	YML102W	swissnew|Q04199|YMK2_YEAST	hypothetical TRP-ASP repeats containing protein in NUP188-TSL1 intergenic region
HRX_HUMAN	zinc finger protein HRX	acute leukaemias	159555	5.8×10^{-32}	YHR119W	swiss|P38827|YHR9_YEAST	hypothetical 123.9 kDa protein in ORC6-MSH1 intergenic region
				6.9×10^{-14}	YJL168C	swiss|P46995|YJQ8_YEAST	hypothetical 84.5 kDa protein in CPS1-FPP1 intergenic region
LCAT_HUMAN	phosphatidylcholine-sterol acyltransferase precursor	Norum and fish eye diseases	245900	1.6×10^{-12}	YNR008W	swiss|P40345|YN84_YEAST	hypothetical 75.4 kDa protein in VPS27-CSE2 intergenic region

LIS1_HUMAN	platelet-activating factor acetylhydrolase γ subunit	lissencephaly syndrome	247200	3.0×10^{-33}	YPL151C	tremblnew|X96770|SCLACHX-VI_21	product: 'P2594 protein'; *S. cerevisiae* chromosome XVI, left arm DNA
				3.7×10^{-13}	YKR036C	swiss|P36130|YK16_YEAST	hypothetical TRP-ASP repeats containing protein in DAL80-GAP1 intergenic region
				7.2×10^{-21}	YNL317W	swiss|P42841|YN57_YEAST	hypothetical TRP-ASP repeats containing protein in HXT14-PHA2 intergenic region
				7.6×10^{-13}	YGL004C	swiss|P53196|YGA4_YEAST	hypothetical TRP-ASP repeats containing protein in PMC1-TFG2 intergenic region
MTHR_HUMAN	methylenetetrahydrofolate reductase	homocysteinaemia	236250	3.3×10^{-102}	YGL125W	swiss|P53128|YGM5_YEAST	hypothetical 68.5 kDa protein in SCS3-SUP44 intergenic region
MYOP_HUMAN	myo-inositol-1	related to the anti-manic and anti-depressant actions of Li^+		2.2×10^{-45}	YHR046C	swiss|P38710|YHK6_YEAST	hypothetical 32.8 kDa protein in DOG1-AAP1 intergenic region
NC5R_HUMAN	NADH-cytochrome b_5 reductase	methemoglobinaemia	250800	1.6×10^{-36}	YML087C	swissnew|Q04516|YMI7_YEAST	hypothetical 35.8 kDa protein in RPM2-TUB1 intergenic region
				3.8×10^{-53}	YML125C	swissnew|Q12746|YMM5_YEAST	hypothetical 35.3 kDa protein in HMGS-TUB3 intergenic region
PMGM_HUMAN	phosphoglycerate mutase, muscle form	myopathy	261670	1.3×10^{-85}	YKL152C	trembl|Z26877|SCDCHR11_13	product: 'unknown'; *S. cerevisiae* 36.2 kbp DNA fragment from chromosome 11
PMS2_HUMAN	PMS1 protein homologue 2	non-polyposis colon cancer	600259	1.3×10^{-28}	YPL164C	tremblnew|X96770|SCLACHX-VI_8	product: 'P2550 protein'; *S. cerevisiae* chromosome XVI, left arm DNA
PMSC_HUMAN	autoantigen PM-SCL	polymyositis and scleroderma	n/p	1.2×10^{-83}	YOR001W	trembl|U43491|SC43491_4	gene: 'UNC733'; product: 'hypothetical protein UNC733'; *S. cerevisiae* cosmid clone pEOA156 from chromosome XV
PTSR_HUMAN	peroxisomal targeting signal receptor 1	peroxisome biogenesis disorders	202370	1.8×10^{-28}	YMR018W	swissnew|Q04364|YMP8_YEAST	hypothetical 59.1 kDa protein in SOK2-FMS1 intergenic region
RFX1_HUMAN	MHC class II regulatory factor RFX1	MHC class II deficiency	600006	7.9×10^{-19}	YLR176C	swiss|P48743|RFXL_YEAST	hypothetical 85.7 kDa protein in CBF5-DKA1
RRA2_HUMAN	Ras-related protein R-Ras2	ovarian tumours	600098	5.2×10^{-32}	YCR027C	swiss|P25378|YCR7_YEAST	hypothetical Ras-related 23.4 kDa protein in MAK32-CRY1 intergenic region
RSMB_HUMAN	small nuclear ribonucleoprotein-associated proteins B and B′	systemic lupus erythematosus	182282	9.6×10^{-26}	YER029C	swiss|P40018|YEL9_YEAST	hypothetical 22.4 kDa protein in GAL83-YPT8 intergenic region
RUXE_HUMAN	U1 and U2 small nuclear ribonucleoprotein E	systemic lupus erythematosus	128260	8.0×10^{-12}	YER146W	swiss|P40089|YEX6_YEAST	hypothetical 10.4 kDa protein in UBP5-SPT15 intergenic region
SPRE_HUMAN	sepiapterin reductase	several	182125	1.0×10^{-11}	YIR035C	swiss|P40579|YIV5_YEAST	hypothetical oxidoreductase in LYS1-HYR1 intergenic region
WASP_HUMAN	Wiskott-Aldrich syndrome protein	Wiskott-Aldrich syndrome	301000	3.2×10^{-10}	YOR181W	trembl|D78487|SCLAS17_1	gene: 'LAS17'; product: 'proline-rich protein'; yeast LAS17 gene for proline-rich protein
XPG_HUMAN	DNA-repair protein complementing XP-G cells	xeroderma pigmentosum group G and Cockayne's syndrome	278780	2.2×10^{-18}	YDR263C	trembl|Z68290|SC9320B_2	gene: 'YD9320B.02c'; product: 'unknown'; *S. cerevisiae* chromosome IV cosmid 9320B

Forty pairs of possible new relations yeast-HDRPs, for which the function of the yeast sequences has not been yet characterised, are part of a complete list of 285 HDRP-yeast pairs accessible at http://www.ebi.ac.uk/–andrade/papers/hdr/. The server contains the full GeneQuiz sequence analysis corresponding to the 686 HDRPs. The regularly updated GeneQuiz analysis of the yeast genome is accessible from the GeneQuiz web home-page (http://www.sander.ebi.ac.uk/genequiz/). Description is taken from the corresponding data base entry; related disease is extracted from the 'DISEASE' SwissProt field of the corresponding entry; OMIM identifiers are given for the HDRPs (note that SW:PMSC_HUMAN is not present in the OMIM data base); similarity between the HDRP and the yeast protein is indicated by the *P*-value of the BLASTP sequence comparison; MIPS identifiers are given for the yeast proteins, as well as a data base identifier with the format 'data base|accession number|identifier'.

searches in up-to-date sequence data bases, followed by evaluation of the results using expert rules.

Family members were identified using the GeneQuiz definition of 'clear' sequence similarity. This is equivalent in simple terms to a BLAST [8] *P*-value of 10^{-10} or to a FASTA [9] score of 135, both for protein or nucleotide searches in protein data bases. In addition, GeneQuiz applies a special masking procedure of the query sequence that avoids spurious hits to amino acid-biased composition regions (Casari and Ouzounis, unpublished).

Our definition of similarity is chosen to be conservative, as needed for the automatic exploration of large data sets. With this decision, we risk missing a number of possible relations but increase the reliability of the predictions for the set of HDRPs.

To discriminate pairs of paralogues from pairs of putative orthologues, we tested whether for a yeast protein similar to an HDRP there were other human proteins more similar than the HDRP. The similarity levels were taken from the BLASTP *P*-value (after application of the above-mentioned filtering scheme). A pair was considered to denote a paralogous relationship if the BLASTP comparison of the yeast and human sequences had a *P*-value at least 10-fold lower than the HDRP-yeast pair, i.e. the second human protein was clearly more similar than the HDRP. The 10-fold factor was used as a very restrictive threshold for the assignment of paralogous pairs.

Multiple sequence alignments and phylogenetic trees were generated with the ClustalW package [10]. Sequence alignments were represented using the program BOXSHADE (Hofmann, Baron and Schirmer; ISREC, Switzerland; http://ulrec3.unil.ch/software/BOX_form.html).

3. Results

Of the 686 proteins annotated as HDRP in SwissProt, we detected 285 pairs HDRP-yeast protein. All cases were validated by manual inspection and the results are deposited in the web appendix of this paper at http://www.sander.ebi.ac.uk/–andrade/papers/hdr/. In 98 cases, the yeast proteins are enzymes (containing an EC number in the description), 17 are transmembrane, 25 are mitochondrial, and 30 are nuclear (15 of them DNA-binding). The main differences as compared with the total set of HDRPs are a higher representation of mitochondrial proteins (16% vs 9%) and ATP-binding proteins (11% vs 7%), and a smaller representation of transmembrane proteins (11% vs 22%), partly caused by the absence of relevant similarity to human receptors.

In 46 of the 285 pairs (involving 25 HDRPs), the human-yeast pairs can be described as paralogous; that is, the yeast protein was closer to other human proteins than to the HDRP. The most probable origin of these human paralogues is from ancestral sequences that expanded with different isoforms. These isoforms presumably represent higher specialisation levels, that are not present in yeast.

The remaining pairs (239 of 285, involving 142 human proteins, since some HDRPs are similar to more than one yeast

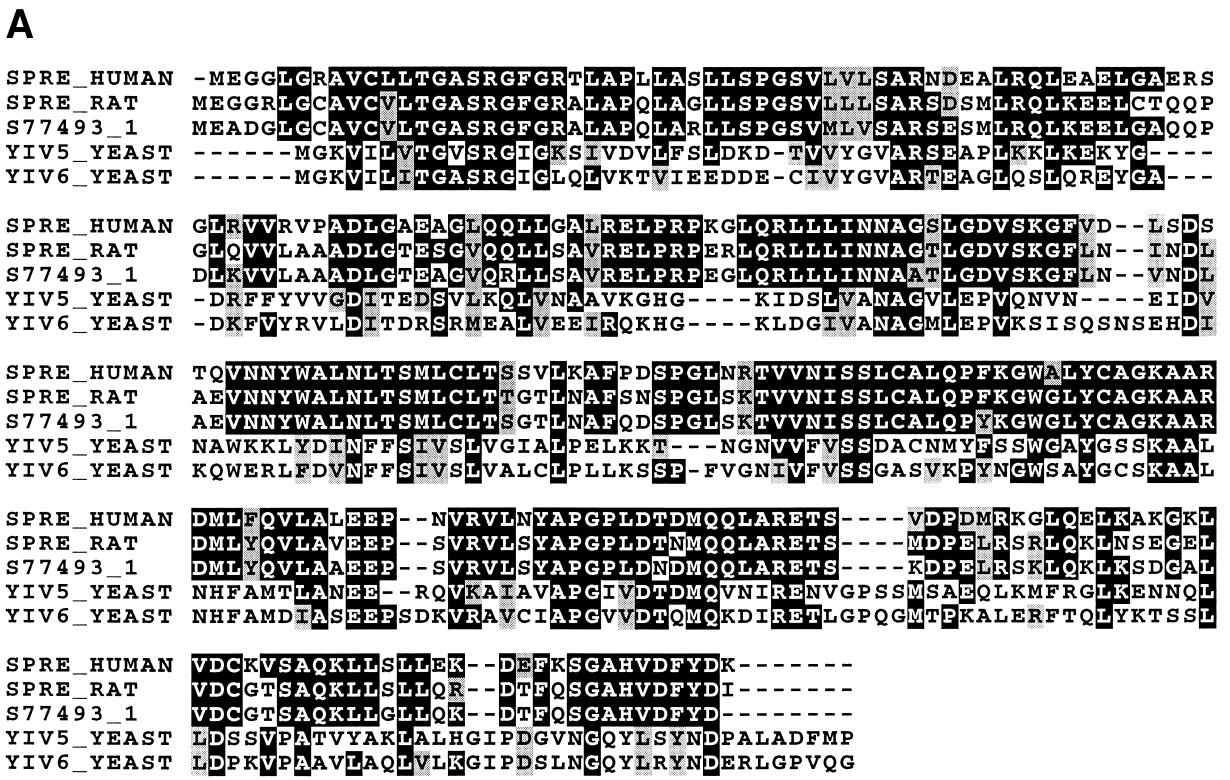

Fig. 1. A: Alignment of sepiapterin reductase (SW:SPRE_HUMAN), related to several human diseases [11], with the almost identical sequences in rat (SW:SPRE_RAT) and mouse (TREMBL:S77493_1, EMBL:S77493) and with two sequences in yeast (SW:YIV5_YEAST, MIPS:YIR035C; and SW:YIV6_YEAST, MIPS:YIR036C). Black boxes indicate residues completely conserved in at least 50% of the aligned sequences at a given alignment position. Grey boxes indicate conservation of residue type in at least 50% of the aligned sequences. The closer yeast sequence YIV5_YEAST scores a BLASTP *P*-value of 6.1×10^{-12}. The two yeast sequences have higher similarity between them (51% identity) than to the chordata sequences (about 25% identity). A recent genome duplication event in yeast cannot be discarded. B: Phylogenetic tree of the five sequences from A together with other representatives of the short-chain alcohol dehydrogenase family. Protein identifiers are from SwissProt. Bootstrapping values are shown (1000 indicates reliable branches, lower values indicate lesser reliability). The yeast sequences are shown to form a separate sub-family together with sepiapterin reductases (SPRE*) and corticosteroid 11-β-dehydrogenases (DHI1*), closely related to the former. Other outstanding protein sub-families represented in the tree are: FABG*, acyl-carrier protein reductases; DHG*, glucose 1-dehydrogenases; and BPHB*, biphenyl-*cis*-diol dehydrogenases.

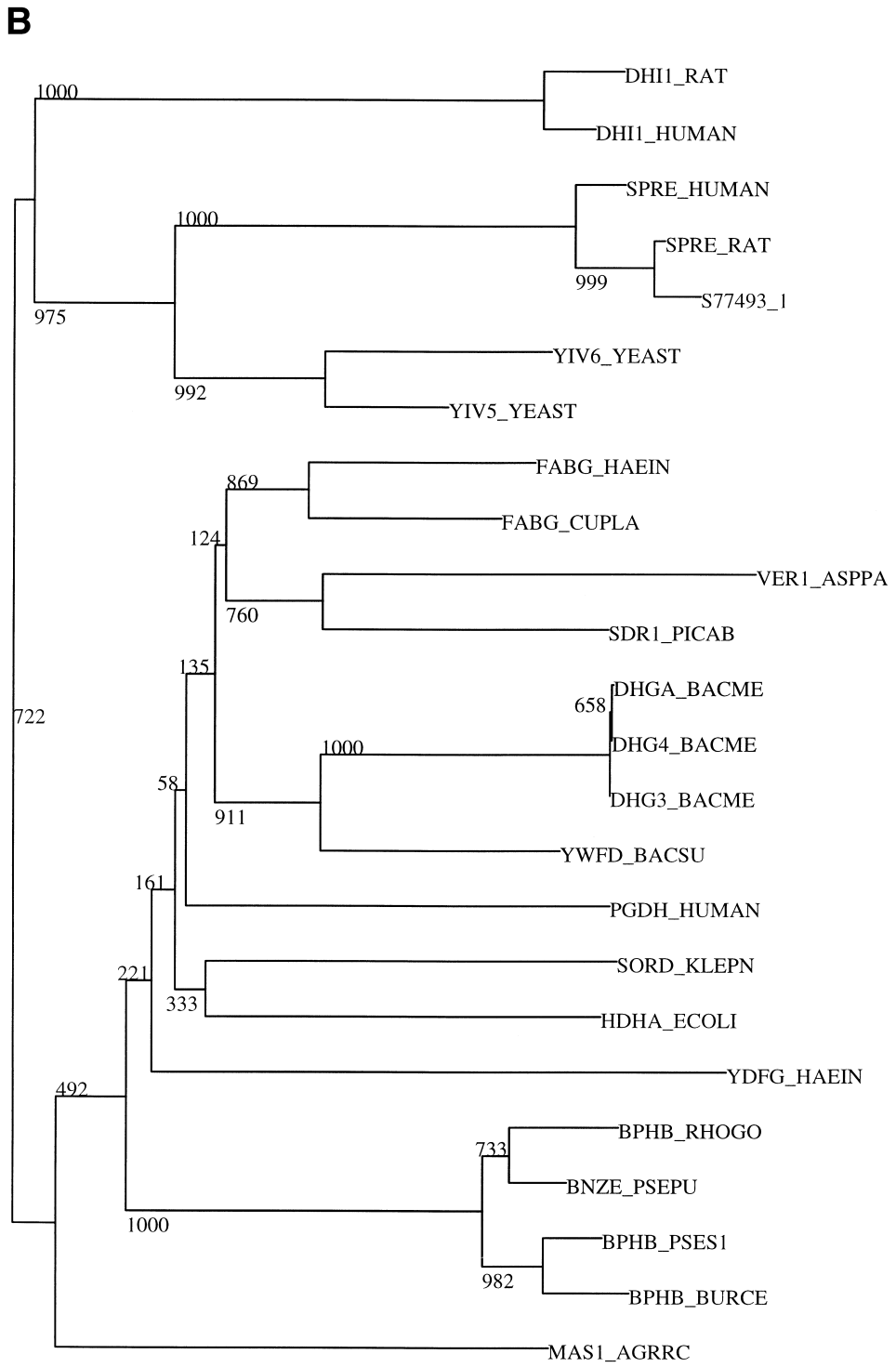

Fig. 1. (*continued*)

sequence) indicate putative orthologues. With the discovery of new human sequences, the fraction of orthologous cases will decrease, with the corresponding increase in the number of paralogues.

Of the 239 pairs of possible orthologues, we have selected 40 uncharacterised relations in which the yeast sequences are not functionally characterised; most are annotated as 'hypothetical' in the data bases (Table 1). These new sequence relations add a new perspective to the HDRP family, facilitating its experimental study in a simpler system.

3.1. Relevant examples

The above-mentioned public availability of the automatic analysis for all the HDRP-yeast pairs will drastically reduce the effort required for detailed analysis of any of the cases. It is nevertheless important to note that the validation of each of

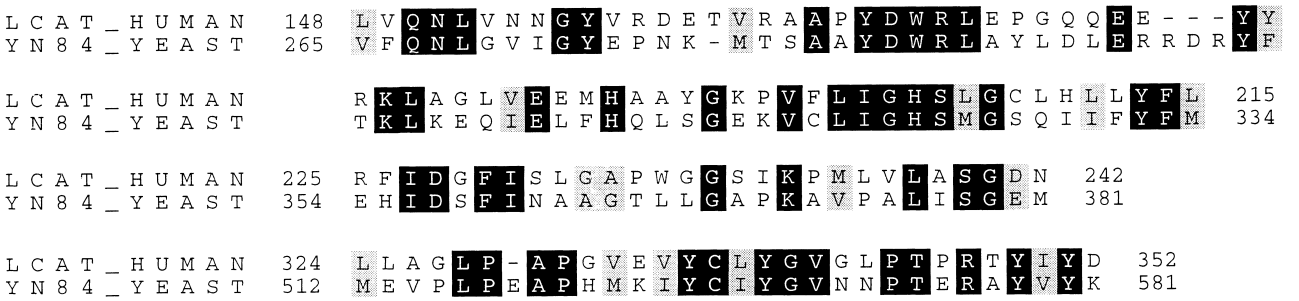

Fig. 2. Partial alignment of the human phosphatidylcholine-sterol acyltransferase precursor 440 aa long (LCAT) (SW:LCAT_HUMAN) with a yeast sequence 661 aa long (SW:YN84_YEAST, MIPS:YNR008W). The other similar sequences in the databases from rabbit (TREMBL:OCLCAT_1, EMBL:D13668), rat (SW:LCAT_RAT), mouse (SW:LCAT_MOUSE), chicken (TREMBL:GGLECCHAC_1, EMBL:X91011), and an identical sequence from baboon, have no less than 63% identity among them and to LCAT, whereas the yeast sequence has 16% identity and therefore highlights better the conserved regions of the family. The corresponding positions in LCAT may be functionally relevant. Residue conservation is represented as in Fig. 1A.

these cases requires extensive additional human intervention. In the following sections, a detailed analysis is presented for four of the 40 newly discovered orthologous sequences.

These four cases were selected because the yeast sequences are the most distant members of the HDRP family, entering the 'twilight zone'. They represent a radical advance in the comprehension of the family, sharpening aspects such as sequence conservation and secondary structure prediction.

3.1.1. Phenylketonuria-related protein. Human sepiapterin reductase (SW:SPRE_HUMAN) participates in the biosynthesis of tetrahydrobiopterin [11], a cofactor of phenylalanine hydrolase. Its deficiency affects the function of phenylalanine hydrolase, leading to the accumulation of phenylalanine. Acute phenylketonuria is related to mental retardation.

We found two yeast proteins with remote sequence similarity to this human protein (SW:YIV5_YEAST, MIPS:YIR035C; and SW:YIV6_YEAST, MIPS:YIR036C) (Fig. 1A), which may help in the characterisation of the human protein, although phenylalanine hydrolase is apparently not present in yeast.

These sequences belong to the family of short-chain alcohol dehydrogenases, which has other human members. The tree of this family (depicted in Fig. 1B) shows that the five sequences belong to a subfamily that includes human corticosteroid 11-β-dehydrogenase (SW:DHI1_HUMAN) and a similar sequence in rat (SW:DHI1_RAT).

YIV5_YEAST and YIV6_YEAST have considerable sequence similarity between them. Taking into account that they belong to contiguous genome positions on chromosome VIII (421 026–421 787 and 422 074–422 862, respectively), a tandem duplication seems plausible. Whether the two yeast genes are responsible for overlapping functions has to be elucidated by experimentation.

3.1.2. Norum- and fish-eye disease-related protein. The lecithin-cholesterol acyltransferase human protein (LCAT, SW:LCAT_HUMAN) [12] is secreted by the liver and esterifies cholesterol in plasma. This is a key step in the process of cholesterol transport and metabolism, since cholesterol is soluble but the esterified form is insoluble. When either LCATase activity is inhibited or the protein is defective, cholesterol is no longer transported and accumulates in the tissues, producing the Norum and fish-eye diseases.

There are sequences with very high sequence similarity to the human protein in other chordata (baboon, rabbit, rat, mouse and chicken). We found a less related yeast sequence (SW:YN84_YEAST, MIPS:YNR008W) defined as a hypothetical protein (open reading frame N2052 in [13]). One of the possibilities offered by the analysis of distant sequences is the better characterisation of important conserved residues in alignment (see Fig. 2).

3.1.3. Lupus erythematosus-related protein. The human U1/U2 small nuclear ribonucleoprotein E (snRNP) (SW:RUXE_HUMAN) is involved in systemic lupus erythematosus [14]. snRNP of the U family are required for several RNA-processing reactions in eukaryotic cells and participate in formation of nuclear ribonucleoprotein complexes. The protein components of the snRNP are recognised by antibodies produced by patients with autoimmune disorders, e.g. lupus erythematosus. RUXE_HUMAN is known to be an autoimmune antigen.

There are two possible homologues of the HDRP in yeast, one closest and already characterised as 'core snRNP protein E' (SW:SME1_YEAST, MIPS:YOR159C, TREMBL:SCSME1GEN_1) and a hypothetical protein with a low sequence similarity level (SW:YEX6_YEAST, MIPS:YER146W) (see Fig. 3a).

A previous publication [14] reported sequence similarity of the human protein to *S. cerevisiae* mitochondrial 38S ribosomal protein var1 (EMBL:SCer38SRP, SW:RMAR_YEAST). The quality of the alignment to YEX6_YEAST with the HDRP is better, with greater identity and fewer gaps (17 identities with five gaps vs 23 identities with one gap, Fig. 3b). YEX6_YEAST and SME1_YEAST, and not SCer38SRP, therefore seem to be the yeast homologues to the HDRP RUXE_HUMAN, and provide a better experimental model.

3.1.4. Wiskott-Aldrich syndrome human protein. The Wiskott-Aldrich syndrome human protein (WASP, SW:WASP_HUMAN) [15] is related to an X-linked recessive disorder associated with thrombocytopenia, eczema, bloody diarrhoea, immunodeficiency, and risk of malignancies. WASP mRNA is expressed in cells related to platelet production, and in T and B cell lines. The protein is known to interact with the GTPase cdc42 through a GTPase binding domain (GBD) [15].

WASP_HUMAN has reliable sequence similarity over a length of 100 aa to TREMBL:Sclas17_1 (SW:LA17_YEAST, MIPS:YOR181W), a hypothetical yeast protein succinctly annotated as gene LAS17 product 'proline-rich protein' (YSCLAS17). The similarity between the human and the yeast sequences has already been reported as marginal (http://

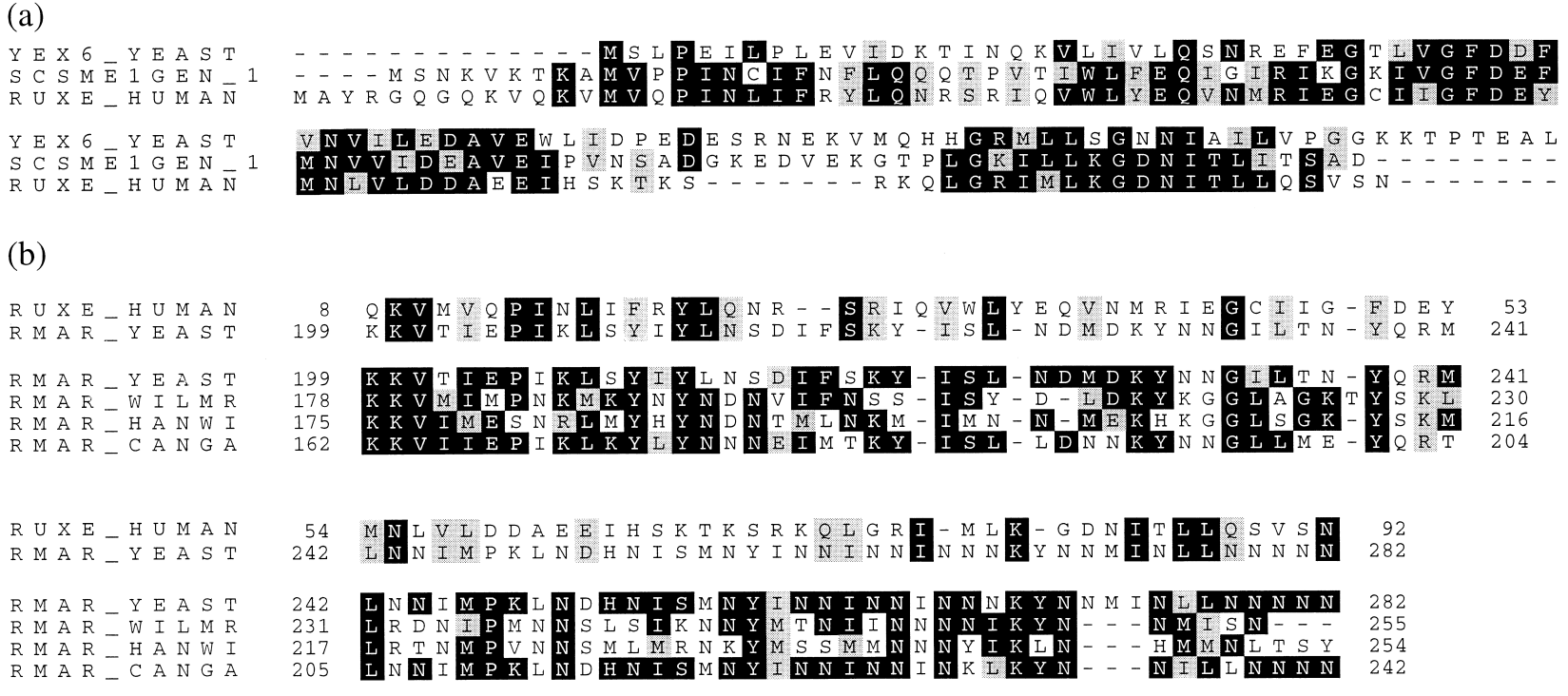

Fig. 3. a: Full-length alignment of the human small ribonucleoprotein (SW:RUXE_HUMAN) with two yeast homologues, a sequence characterised as core snRNP protein E (SW:SME1_YEAST, MIPS:YOR159C, TREMBL:SCSME1GEN_1) and a yeast hypothetical protein (SW:YEX6_YEAST, MIPS:YER146W). The most distant homologue, YEX6_YEAST, shows better the conserved residues of the family. Residue conservation is represented as in Fig. 1A. b: Alignment of RUXE_HUMAN with *S. cerevisiae* Scer38SRP, mitochondrial 38S ribosomal protein var1 (SW:RMAR_YEAST), as proposed in [14]. For comparison we show the alignment of RMAR_YEAST with the other known 38S ribosomal var1 proteins. The conservation patterns are markedly different between the two alignments. Residue conservation is represented as in Fig. 1A.

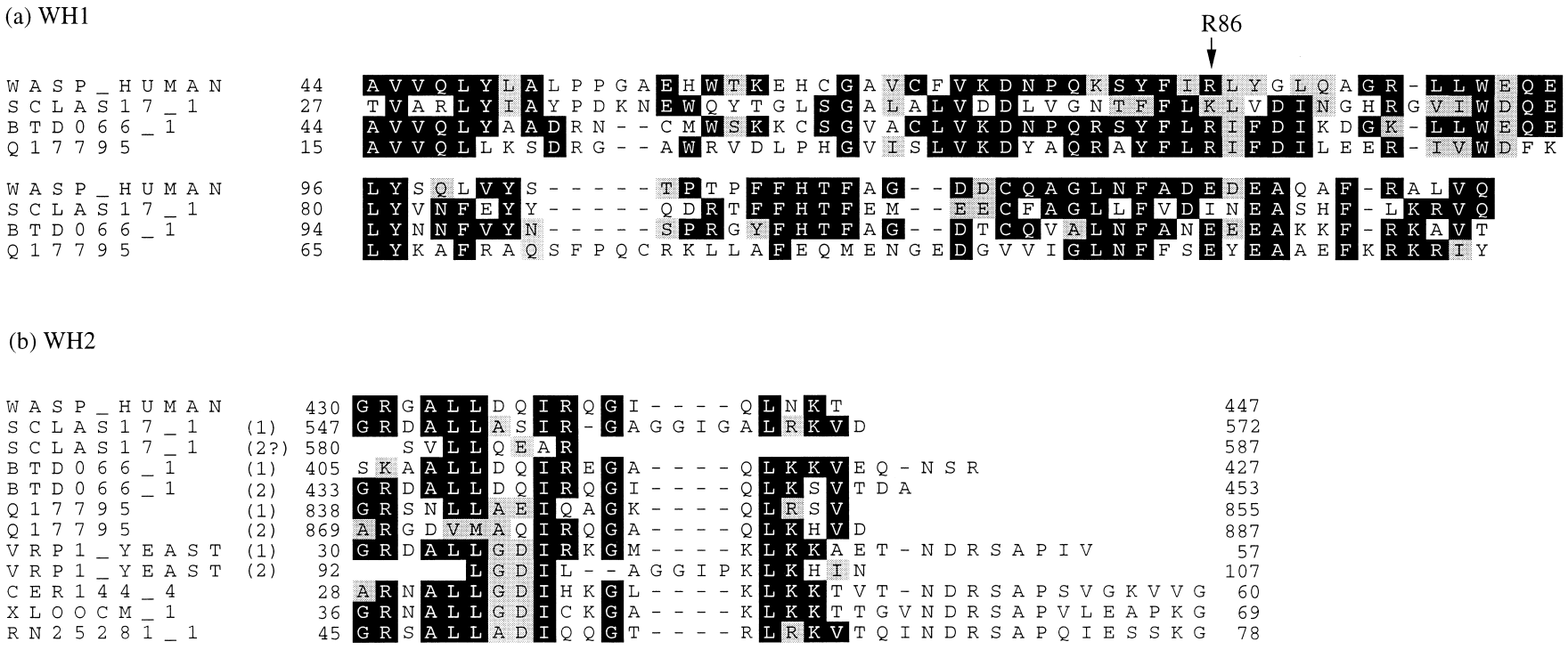

Fig. 4. Alignment of the Wiskott-Aldrich syndrome protein (SW:WASP_HUMAN) with a yeast hypothetical protein (TREMBL:SCLAS17_1, SW:LA17_YEAST, MIPS:YOR181W), bovine actin-depolymerising protein (EMBL:BTD066) and *Caenorhabditis elegans* hypothetical protein (SWTREMBL:Q17795). Other very similar human and bovine sequences are not shown. Residue conservation is represented as in Fig. 1A. a: WH1 domain. Arginine-86 in the human WASP, known to be mutated in WAS patients [15], is marked with an arrow. b: WH2 repeat. Other sequences lacking the WH1 domain were used to extend the alignment in this region: *S. cerevisiae* verprolin (SW:VRP1_YEAST, MIPS:YLR337W), *C. elegans* hypothetical protein (EMBL:CER144), frog *Xenopus laevis* hypothetical protein (EMBL:XLOOCM) and rat hypothetical protein (EMBL:RN25281). We propose that WH2 is a tandem repeat of 20 aa in a number of proteins. This fact seems to have escaped the attention of previous investigators [16]. It is unclear whether SCLAS17 has the WH2 repeat twice, whereas human WASP seems to have it only once.

www.ncbi.nlm.nih.gov/Bassett/Yeast, [16]). Here we characterise this similarity in more detail.

Interestingly, the yeast protein YSCLAS17 seems not to have a GBD, and it would therefore not interact with the ras cdc42 protein. Both WASP and YSCLAS17 have two domains, called WH1 and WH2 (after WASP homology 1 and 2) and a characteristic proline-rich region (Fig. 4). Mutations known to induce the WAS occur inside the WH1 domain (two point mutants, R86L and R86H, and a frame-shift that caused truncation at position 74) [15]. The yeast homologue thus provides an appealing model system, since it shares a WASP region implicated in the disease and different from the GBD.

4. Discussion

4.1. Related studies of human disease-related proteins

One of the more interesting implications originated by the genome projects is the possibility of finding proteins related to potential HDRPs in organisms in which experimentation is feasible. This is possible for the first time with the availability of the *S. cerevisiae* genome.

4.1.1. Analysis of positionally cloned human genes. Those HDRP genes that have been identified by positional cloning are candidates for direct implication in the corresponding disease, while for a broader spectrum of human genes, their involvement in the corresponding diseases has not been fully demonstrated.

The set of 70 known positionally cloned genes has been previously analysed [17,18]. Bassett et al. [17] first reported 15 cases of clear homology between human and yeast proteins (web page at http://www.ncbi.nlm.nih.gov/Bassett/Yeast, last update November 1996). The proportion of HDRPs analysed for which a relevant similarity with a yeast protein was reported (15/70, 21%) was similar to the proportion we found for a set of HDRPs obtained from SwissProt (285 HDRP-yeast pairs involving 155, 23%, out of 686 human proteins), pointing to a similar quality of the similarity searches carried out in both studies.

The second publication on the same subject [18] reported similar sequences for 60% of the 70 positionally cloned human disease genes, corresponding to 18% pairs of orthologues. Some of these relationships are new and based on weak sequence similarities detected by a manual iterative strategy of linked sequence searches [19].

An interesting example is the relationship between SW:MLH1_HUMAN (OMIM:120436), involved in hereditary non-polyposis colon cancer, and two yeast sequences, a putative orthologue SW:MLH1_YEAST (MIPS:YMR167W) and a paralogue SW:PMS1_YEAST (MIPS:YNL082W) (Fig. 2A in [18]). We confirmed this finding with our automatic searching strategy. The number of other cases in which the proposed twilight sequence similarities hold true functional relationships remains to be determined.

4.1.2. Manual analysis of HDRPs annotated in OMIM. Two hundred and fifty HDRPs were previously analysed by Foury [20], including 80 positionally cloned genes, 10 more than in the previous studies. Significant human-yeast pairs were reported for approximately 41% of them.

This analysis provides interesting evidence about many human proteins. For example, the human adenine phosphoribosyltransferase SW:APT_HUMAN (OMIM:102600), related to urolithiasis, is reported to be similar to yeast APT1 (MIPS:YML022W) and, to a lesser extent, to APT2 (MIPS :YDR441C). Our automatic procedure was also able to find these cases; full information on several sequence searches and multiple sequence alignments is available on the web appendix of this paper.

Unfortunately, different technical shortcomings diminish the quality of Foury's analysis. First, the basic data underlying the conclusions are not provided, making it impossible to validate the functional assignments without repeating the analysis. This is something that most readers will not be able to do systematically, as in the case above.

Second, there is no discussion of what are the likely orthologous sequences. Some of the functional assignments in families of paralogous sequences are therefore misleading in the sense that a single yeast sequence should not be assigned as functionally equivalent to the HDRP.

One example of such an error is the analysis of the ankyrine defect SW:ANK1_HUMAN (OMIM:182900) for which the yeast protein MIPS:YIL112W was assigned as a possible homologue. It resulted that YIL112W is more similar to another human protein (EMBL:HS439286) than to the one proposed and three other yeast proteins are more similar to the HDRP (MIPS:YGR232W, MIPS:YGR233C, and MIPS:YOR034C, see Table 1 and web appendix) than YIL112W. The correct conclusion would have been that the HDRP belongs to a family of paralogues with different representatives in yeast and humans, and it would be incorrect to point to only one of them as directly related.

In the third place, Foury carried out homology searches in only one direction, that is, searching in the yeast genome with the human proteins. Given the asymmetry of the sequence space, a double check is necessary, running the yeast sequences against the human data to eliminate possible equivocal assignments.

An example of this type of error is the reported analysis of the LIM kinase (OMIM:248610, SW:LIK1_HUMAN, 647 aa long). This human protein was reported as a homologue to MIPS:YOL113W (655 aa long) with a BLASTP *P*-value of 10^{-24}. The double check reveals, however, that YOL113W is most likely the orthologue of SW:PAK1_HUMAN (545 aa long), which reaches a BLASTP *P*-value of 8.9×10^{-135}.

4.2. Analysis of HDRPs annotated in SwissProt

A note of caution should be introduced here, since we used those human proteins broadly defined as HDRPs and manually annotated as such in the SwissProt data base. In the current release of this data base it is possible that some of the annotations are not completely accurate and that the data base is not completely up-to-date. For example, eight of the nine HDRPs related to yeast sequences in the work of Mushegian et al. [18] were not present in SwissProt.

An example of possible inaccuracy of the SwissProt annotations is the human mitochondrial carrier protein (SW:GDC_HUMAN, OMIM:139080) annotated in SwissProt as related to human Graves' disease. In this case, the initial data [21] were not later confirmed by other studies.

The set of 686 HDRPs will undoubtedly contain other possible annotations that will be falsified by further experiments. It is also true that SwissProt is still the best available source of functional annotations, and most of the functional implica-

tions of the 686 HDRPs will be supported by solid experimental evidence.

4.2.1. A possible distinction between cases of orthology and paralogy. Our analysis introduces the orthology/paralogy distinction that is not used in any of the three previous publications [17,18,20]. We show that yeast contains as many as 285 proteins similar to HDRPs, corresponding to 11% of all known human genes and 4% of the complete yeast genome. Of these pairs, 239 are possible yeast orthologues of the HDRP, since there is no other human sequence significantly closer to the yeast protein than the HDRP. They constitute a selected set of examples in which functional similarity is very likely.

We have included an in-depth analysis of four cases. (i) We found two distant members of the human sepiapterin reductase subfamily and we propose that the yeast members are the product of a recent duplication. (ii) A distant member of the family of phosphatidylcholine-sterol acyltransferase is reported. This finding is crucial for the definition of the functionally conserved regions of the family. (iii) We corrected a previously reported homology of a human snRNP and predicted a similarity with a different yeast protein. (iv) We detected a yeast sequence with both WH1 and WH2 domains. These two domains are not generally associated, but interestingly they are also present in the HDRP, with possible functional implications.

The results presented here point to yeast proteins as possible models for the study of the HDRPs and may help advance the comprehension of the biochemistry associated with diseases like phenylketonuria, lupus erythematosus, Norum and fish-eye syndrome.

Acknowledgements: We are thankful to the anonymous referee for detailed comments and especially for pointing out the new experimental evidence on the molecular origin of Graves' disease. This work was supported by EC-TMR Grant 'GeneQuiz'. M.A.A. holds a postdoctoral fellowship from the same program.

References

[1] Bairoch, A. and Apweiler, R. (1997) Nucleic Acids Res. 25, 31–36.
[2] Fitch, W. and Markowitz, E. (1970) Biochem. Genet. 4, 579–593.
[3] Bork, P., Ouzounis, C., Casari, G., Schneider, R., Sander, C., Dolan, M., Gilbert, W. and Gillevet, P. (1995) Mol. Microbiol. 16, 955–967.
[4] Koonin, E.V., Tatusov, R.L. and Rudd, K.E. (1996) Methods Enzymol. 266, 295–322.
[5] Goffeau, A., Barrell, B.G., Bussey, H., Davis, R.W., Dujon, B., Feldmann, H., Galibert, F., Hoheisel, J.D., Jacq, C., Johnston, M., Louis, E.J., Mewes, H.W., Murakami, Y., Philippsen, P., Tettelin, H. and Oliver, S.G. (1996) Science 274, 546–567.
[6] Casari, G., Ouzounis, C., Valencia, A. and Sander, C. (1996) in: Proceedings of the First Annual Pacific Symposium on Biocomputing, pp. 707–709, World Scientific, Hawaii.
[7] Scharf, M., Schneider, R., Casari, G., Bork, P., Valencia, A., Ouzounis, C. and Sander, C. (1994) in: Second International Conference on Intelligent Systems for Molecular Biology (Altman, R., Brutlag, D., Karp, P., Lathrop, R. and Searls, D., Eds.), pp. 348–353, AAAI Press, Menlo Park, CA.
[8] Altschul, S.F., Gish, W., Miller, W., Myers, E.W. and Lipman, D.J. (1990) J. Mol. Biol. 215, 403–410.
[9] Pearson, W.R. (1996) Methods Enzymol. 266, 227–258.
[10] Thompson, J.D., Higgins, D.G. and Gibson, T.J. (1994) Nucleic Acids Res. 22, 4673–4680.
[11] Ichinose, H., Katoh, H., Sueoka, T., Titani, K., Fujita, K. and Nagatsu, T. (1991) Biochem. Biophys. Res. Commun. 179, 183–189.
[12] McLean, J., Fielding, C., Drayna, D., Dieplinger, H., Baer, B., Kohr, W., Henzel, W. and Lawn, R. (1986) Proc. Natl. Acad. Sci. USA 83, 2335–2339.
[13] Verhasselt, P., Aert, R., Voet, M. and Volckaert, G. (1994) Yeast 10, 1355–1361.
[14] Stanford, D.R., Kehl, M., Perry, C.A., Holicky, E.L., Harvey, S.E., Rohleder, A.M., Jr, K.R., Luhrmann, R. and Wieben, E.D. (1988) Nucleic Acids Res. 16, 10593–10605.
[15] Derry, J.M.J., Ochs, H.D. and Francke, U. (1994) Cell 78, 635–644.
[16] Symons, M., Derry, J.M., Karlak, B., Jiang, S., Lemahieu, V., Mccormick, F., Francke, U. and Abo, A. (1996) Cell 84, 723–734.
[17] Bassett, D.J., Boguski, M. and Hieter, P. (1996) Nature 379, 589–590.
[18] Mushegian, A., Bassett, D., Boguski, M., Bork, P. and Koonin, E. (1997) Proc. Natl. Acad. Sci. USA 94, 5831–5836.
[19] Mushegian, A. and Koonin, E. (1996) Genetics 144, 817–828.
[20] Foury, F. (1997) Gene 195, 1–10.
[21] Zarrilli, R.A., Oates, E.L., McBride, O.W., Lerman, M.I., Chan, J.Y., Santisteban, P., Ursini, M.V., Notkins, A.L. and Kohn, L.D. (1989) Mol. Endocrinol. 3, 1498–1508.