Stockholm University

This is an accepted version of a paper published in *Journal of Molecular Biology*. This paper has been peer-reviewed but does not include the final publisher proof-corrections or journal pagination.

Citation for the published paper:

Ekman, D., Elofsson, A. (2010)

"Identifying and Quantifying Orphan Protein Sequences in Fungi"

Journal of Molecular Biology, 396(2): 396-405

URL: <http://dx.doi.org/10.1016/j.jmb.2009.11.053>

Access to the published version may require subscription.

Permanent link to this version:

<http://urn.kb.se/resolve?urn=urn:nbn:se:su:diva-49277>

<http://su.diva-portal.org>

AUTHOR QUERY FORM

ELSEVIER	Journal: YJMBI Article Number: 61939	Please e-mail or fax your responses and any corrections to: E-mail: corrections.essd@elsevier.spitech.com Fax: +1 61 9699 6721
---	---	--

Dear Author,

Any queries or remarks that have arisen during the processing of your manuscript are listed below and highlighted by flags in the proof. Please check your proof carefully and mark all corrections at the appropriate place in the proof (e.g., by using on-screen annotation in the PDF file) or compile them in a separate list.

For correction or revision of any artwork, please consult <http://www.elsevier.com/artworkinstructions>.

Articles in Special Issues: Please ensure that the words 'this issue' are added (in the list and text) to any references to other articles in this Special Issue.

Uncited references: References that occur in the reference list but not in the text – please position each reference in the text or delete it from the list.	
Missing references: References listed below were noted in the text but are missing from the reference list – please make the list complete or remove the references from the text.	
Location in article	Query / remark Please insert your reply or correction at the corresponding line in the proof
Q1	Running head: Orphan Protein Sequences in Fungi. Short title OK?
Q2	Data "95 or 76%" changed to "95% or 76%." Change correct?
Q3	"Rapidly evolving" changed to "rapidly evolving sequences." Modification correct?

Electronic file usage

Sometimes we are unable to process the electronic file of your article and/or artwork. If this is the case, we have proceeded by:

☐ Scanning (parts of) your article	☐ Rekeying (parts of) your article	☐ Scanning the artwork
---	---	---

Thank you for your assistance.

JMB

Available online at www.sciencedirect.com

ScienceDirect

Identifying and Quantifying Orphan Protein Sequences in Fungi

Diana Ekman and Arne Elofsson*

Stockholm Bioinformatics
Center/Center for Biomembrane
Research, Department of
Biochemistry and Biophysics,
Stockholm University, SE-10691
Stockholm, Sweden

Received 16 March 2009;
received in revised form
17 November 2009;
accepted 20 November 2009

For large regions of many proteins, and even entire proteins, no homology to known domains or proteins can be detected. These sequences are often referred to as orphans. Surprisingly, it has been reported that the large number of orphans is sustained in spite of a rapid increase of available genomic sequences. However, it is believed that *de novo* creation of coding sequences is rare in comparison to mechanisms such as domain shuffling and gene duplication; hence, most sequences should have homologs in other genomes.

To investigate this, the sequences of 19 complete fungi genomes were compared. By using the phylogenetic relationship between these genomes, we could identify potentially *de novo* created orphans in *Saccharomyces cerevisiae*. We found that only a small fraction, <2%, of the *S. cerevisiae* proteome is orphan, which confirms that *de novo* creation of coding sequences is indeed rare. Furthermore, we found it necessary to compare the most closely related species to distinguish between *de novo* created sequences and rapidly evolving sequences where homologs are present but cannot be detected.

Next, the orphan proteins (OPs) and orphan domains (ODs) were characterized. First, it was observed that both OPs and ODs are short. In addition, at least some of the OPs have been shown to be functional in experimental assays, showing that they are not pseudogenes. Furthermore, in contrast to what has been reported before and what is seen for older orphans, *S. cerevisiae* specific ODs and proteins are not more disordered than other proteins. This might indicate that many of the older, and earlier classified, orphans indeed are fast-evolving sequences. Finally, >90% of the detected ODs are located at the protein termini, which suggests that these orphans could have been created by mutations that have affected the start or stop codons.

© 2009 Published by Elsevier Ltd.

Edited by M. Sternberg

Keywords: evolution; protein domain; orphan protein; fungi

Introduction

Proteins that have no detectable homologs have been termed orphan proteins (OPs).¹ In analogy, parts of proteins that lack homologs may be referred to as orphan domains (ODs).² Novel genes and domains are believed to be created through gene duplication, shuffling of gene segments, modifications of existing sequences, combinations of short functional peptides,³ and *de novo* creation from

noncoding material. Proteins/domains created *de novo* could be defined as the true OPs/ODs. However, it is believed that *de novo* creation is a comparatively rare mechanism,⁴ implying that most gene segments should be homologous to other genes. Anyhow, less than half of the residues in a typical proteome can be matched to a domain in the Pfam-A database.² Even when the less well-characterized domains in Pfam-B are included, about one quarter of the residues within the proteome remain uncharacterized. Hence, for large parts of the genome, the origin cannot be identified. Thus, either *de novo* creation is more frequent than generally believed, or these regions are evolving so fast that current methods cannot detect the homology. It can indeed be difficult to identify homologs

*Corresponding author. E-mail address: arne@bioinfo.se.

Abbreviations used: OP, orphan protein; OD, orphan domain; SCE, *Saccharomyces cerevisiae* level; SAC, *Saccharomyces* level.

of apparent OPs, since they are often both short and rapidly evolving.⁵⁻⁷

Further, orphanicity can be regarded as a relative concept (like lineage specificity⁵) in evolution, since it depends on the relation between the compared species. Hence, an orphan sequence may have homologs if more closely related species are sequenced. However, Fischer and colleagues noted early that the fraction of orphans decreased only slowly when new genome sequences became available.⁸ Further, although the large increase in sequences from environmental sequencing projects (metagenomics) revealed homologs to several orphans, it also resulted in the detection of many novel orphans.⁹ Finally, gene deletion and pseudo-genization might explain why no homologs can be detected. However, the effect of these uncertainties should decrease when additional genomes of closely related organisms have been sequenced.

Recent studies have provided evidence of *de novo* creation both of whole proteins¹⁰ and parts of proteins¹¹ in the *Saccharomyces cerevisiae* genome. Here, we characterize the OPs and ODs in *S. cerevisiae* and quantify the contribution of the aforementioned mechanisms of novel gene formation in the *S. cerevisiae* genome.

In earlier studies, we have defined orphans as sequence regions that remain after domain assignment by Pfam or other domain databases.^{2,12} In *S. cerevisiae*, 1372 unassigned proteins and almost 5000 unassigned domains (≥ 50 aa) remained after assignment of Pfam-A and Pfam-B domains (Supplementary Material). However, in this study, we found that only 179 (13%) of these unassigned proteins and 23 (<1%) of the unassigned domains lacked detectable homologs when compared to the most closely related genomes, and another 123 (9%) proteins and almost 500 (9%) domains have homologs only within the *Saccharomyces* family. Further, since about one third of the unassigned sequences have homologs in non-fungi species, we conclude that only a fraction of these unassigned sequences can be considered true orphans.

Although domain assignment is a powerful method for homology detection, this approach is not ideal for detection of orphans. First, many domain families are represented by the conserved core, leaving the borders unassigned. Further, the detection of orphan sequences will be dependent on the minimum length chosen for a region to be considered. In addition, different organisms can be more or less well represented in the domain database. Therefore, we have in this study used a strategy based on pairwise alignments between completely sequenced genomes. Comparison of sequences from closely related species should, to some degree, compensate for the fact that Blast is less sensitive than the HMMs used for Pfam assignments.

With this method, the amount of orphans was estimated at a number of reference points in the fungal evolution. We predicted that about 1% of the residues in the proteome are specific to *S. cerevisiae*, <5% percent is specific to the *Saccharomyces* family, and about 70% is ancient (i.e., have a pre-fungi origin). A characteristic of the detected orphans is that they are short; in addition, the domains are often disordered sequences located at the termini. However, we found significant differences between the species-specific orphans and the domains/proteins specific to the *Saccharomyces* family. First, species-specific proteins are unusually short, and second, species-specific domains are, in contrast to family-specific domains, not particularly disordered. These differences indicate that many of the orphans detected at the *S. cerevisiae* level are true *de novo* created orphans, whereas most of the older ones are instead sequences evolving too rapidly for the homology relationship to be found.

Results and Discussion

We have used a strategy based on Blast¹³ pairwise sequence alignments to identify and quantify OP

Fig. 1. A simplified species tree for the fungi included in this study. Each internal node represents a phylogenetic level. The three-letter abbreviations used for the levels are in boldface. A star marks the whole genome duplication event that distinguishes the species at the *Saccharomyces* level from the rest of the Hemiascomycota. The ancient level includes all non-fungi eukaryotes and prokaryotes. Fungi: AGO: *Ashbya gossypii*, AFU: *Aspergillus fumigatus*, CAL: *Candida albicans*, CGL: *Candida glabrata*, CNE: *Cryptococcus neoformans*, DHA: *Debaryomyces hansenii*, ECU: *Encephalitozoon cuniculi*, GZE: *Gibberella zeae*, KLA: *Kluyveromyces lactis*, MGR: *Magnaporthe grisea*, NCR: *Neurospora crassa*, SBA: *Saccharomyces bayanus*, SCA: *Saccharomyces castellii*, SCE: *Saccharomyces cerevisiae*, SKL: *Saccharomyces kluyveri*, SMI: *Saccharomyces mikatae*, SPO: *Schizosaccharomyces pombe*, UMA: *Ustilago maydis*, YLI: *Yarrowia lipolytica*. Branch lengths have no relevance; however, the species have been ordered within the groups so that species that are most closely related to *S. cerevisiae* are listed first.

Fig. 2. Orphan assignment using a pairwise sequence alignment method. (a) Each protein is aligned to proteins from other species groups. (b) Each residue is assigned an age according to the most distantly related sequence it is aligned to. Continuous non-ancient regions ($\geq$ cutoff) are defined as orphan regions at the SCE (*S. cerevisiae*), SAC (*Saccharomyces*), FUN (fungi), and ANC (ancient) levels. If the orphan region covers the entire protein, it is defined as an orphan protein; if it covers only parts of the protein, it is defined as an orphan domain.

sequences in *S. cerevisiae* (i.e., protein sequences that lack homologs in other species). The *S. cerevisiae* genome was selected because it has gene annotations of high quality, and in addition, many closely related genome sequences are available. In this study, we searched for homologs of the *S. cerevisiae* proteins in 18 fungi species at different evolutionary distances from *S. cerevisiae*. To simplify the estimation of protein age, these species were divided into five phylogenetic levels, from the least specific level (fungi) to the species specific [i.e., *S. cerevisiae* specific level (SCE)] (Fig. 1). We defined the *Saccharomyces* level (SAC) to be for all species that evolved after a whole genome duplication in a common ancestor about 100 million years ago,¹⁴ which distinguishes them from the rest of the Hemiascomycota. Next, we estimated the amount of novel, or orphan, genetic material (i.e., sequences that lack homologs in more distantly related species) at these different points in evolution (Fig. 2). Finally, the orphan regions were divided into OPs (i.e., entire proteins that lack homologs) and ODs (i.e., orphan regions), which are adjacent to non-ODs. The distinction between ODs and OPs was introduced in part due to the expected differences in mechanisms involved in their creations and to the problems associated with their detection. For example, ODs are expected to be found in proteins known to be functional, whereas OPs may be results of gene annotation errors.

Detection of orphans

Innovation at the residue level

We used Blast to align each *S. cerevisiae* protein with sequences from several complete fungi genomes as well as non-fungi eukaryotes and prokaryotes. Each residue was thereby assigned to one of the six phylogenetic levels in Fig. 1, depending on how distant the homologs it aligned with (Fig. 2). To reduce the problem associated with alignments of low-complexity sequences, we run Blast twice, first with the low-complexity filter to identify hits and then without the filter to realign low-complexity regions.

According to our alignments, about 70% of the residues in the proteome of *S. cerevisiae* originate from a pre-fungi ancestor (Table 1), while approximately 1% of the residues are defined as species-specific orphans because they did not align to any protein. Another 3% to 5% aligned only to proteins in the most closely related *Saccharomyces* species and were hence assigned to the *Saccharomyces* level, while the remaining proteins are related to proteins in more distantly related fungi. This result was robust at different length cutoffs (between 25 and 100 aa) for inclusion of regions (Fig. 3) (i.e., although the cutoff has a major effect on the number of detected orphans, it has only a moderate impact on the fraction of residues belonging to orphan regions). It can be noted that the orphan residues were quite evenly divided between ODs and OPs. Finally, these numbers were similar when we instead used a domain-based approach (Supplementary Material).

The results in other fungi were similar to those in *S. cerevisiae* (Fig. 4). As expected, longer evolutionary distances between a species and its closest relative result in a larger proportion of orphans. Worth noting is the comparatively large fraction of species/lineage-specific residues in the genome of the filamentous fungi *Neurospora crassa*. *N. crassa* also contains almost twice as many genes as *S. cerevisiae*, suggesting that many orphans appear when genomes increase in size.

Defining orphan domains and orphan proteins

To identify ODs and OPs, we extracted continuous regions that lack homologs at each level (Fig. 2).

Table 1. Orphans

Level	Residues (%)	OP	OPcorr	OPnr	OD	ODcorr	ODnr
<i>S. cerevisiae</i> (SCE)	1.3	188	158	149	22	22	20
<i>Saccharomyces</i> (SAC)	4.9	158	125	116	255	239	209
Hemiascomycota (HEM)	13	379	—	293	1040	—	841
Fungi (FUN)	13	203	—	179	889	—	767
Ancient (ANC)	68	2466	—	1882	2150	—	1715

Residues assigned to each orphan level with the pairwise alignment method. Corrections were made at the SCE and SAC levels after detection of Pfam domains, hits to non-annotated open reading frames in other species as well as with HHpred for OPs. OP, numbers of orphan proteins; OPnr, nonredundant OPs; OPcorr, corrected number of OPs; OD, orphan domains (length ≥ 50 aa); ODnr, nonredundant OD sets; ODcorr, corrected number of ODs.

Fig. 3. Quantification of orphans. The proportion of the proteome that is orphan at different phylogenetic levels for different length cutoffs (25, 50, and 100 aa) using a pairwise sequence alignment method. Residues in OPs are shown separately.

This resulted in 188 proteins defined as OPs and 22 domains (longer than 50 aa) defined as ODs specific to *S. cerevisiae* (SCE level), or 346 OPs and 277 ODs at the *Saccharomyces* level (Table 1). A minimum length was required for ODs, since we only wanted to include sequences that were long enough to be likely to contain a domain.

However, there are a number of caveats that can make it difficult to detect homologs and therefore result in false positives in the orphan assignments. The problem of aligning sequences of low complexity was mentioned above. Furthermore, the results are dependent on the sensitivity of homology detection tools, the quality of the gene annotations, pseudogenization, and the evolutionary rates of individual genes. In the following sections, we have therefore used more sensitive alignment methods and genomic sequences to limit the number of false positives.

More sensitive methods reduce the number of orphans

To increase the sensitivity of the homology detection and detect more distantly related matches, we used HMMER¹⁵ to find homologs in the Pfam domain database. In addition, the HHpred¹⁶ web server was used to find even more distantly related homologs in a large number of protein and domain databases.

We found that Pfam domains overlapped 6 SCE OPs, 27 SAC OPs, and 16 SAC ODs. With HHpred, we detected 2 additional SCE OP homologs and 7 new SAC OP homologs (HHpred was not applied to ODs). These ODs and OPs were removed from the orphans; see the “corrected values” in Table 1. Overlapping domains were more common at the fungi level, indicating that the sensitivity of Blast can be insufficient for detection of distantly related homologs, while it can be used to align the closely related sequences in the *Saccharomyces* family.

Homologs detected in genomic sequences

Further, incomplete gene annotations and pseudogenization in closely related species might result in overestimations of the number of orphans. Therefore, we searched for homologs of the species-specific OPs in the genome sequences of *Saccharomyces paradoxus*, *Saccharomyces mikatae*, and *Saccharomyces bayanus*. Hits were detected for 98 (54%) of the SCE OPs, and we found that 29 of these matched a non-annotated open reading frame of a similar length. This suggests that these sequences are also present in closely related species. Hence, the

Fig. 4. Protein innovation. Residue age distribution in seven species at increasing evolutionary distance from *S. cerevisiae* (SCE). CGL: *Candida glabrata*; KLA: *Kluyveromyces lactis*; DHA: *Debaryomyces hansenii*; NCR: *Neurospora crassa*; UMA: *Ustilago maydis*. Legend: Residue age from the youngest (species specific) to the oldest (ancient) level, SAC/EQ: *Saccharomyces* level or equivalent families; HEM/PEZ: Hemiascomycota/Pezizomycotina; ASC/BAS: Ascomycota/Basidiomycota. The interpretation differs depending on the species; however, all species have the fungi level in common. Blast was run with low-complexity filtering and residues in low-complexity regions were ignored in the calculations.

number of species-specific OPs is probably smaller than first estimated (corrected values in Table 1).

The orphans detected with our method are listed in Supplementary Material. However, it should be noted that homologs are likely to be found for many of these orphans; hence, this is not intended to be the final list of orphans present in *S. cerevisiae*.

Most orphans are singletons

Most of the detected orphans are singletons (i.e., most OPs have no paralogs within the genome and the domains occur in only one protein in *S. cerevisiae*). For example, 17 (78%) and 147 (93%) of the SCE ODs and OPs, respectively, lack homologs within the genome. Corresponding numbers at the SAC level are 206 (86%) and 94 (75%). Orphan domains that have more than one copy are almost exclusively found in homologous proteins, and have probably multiplied through gene duplication rather than through domain duplication and recombination. Finally, by removing homologs in the orphan sets, the number of unique orphans was further reduced (nonredundant values in Table 1).

Orphan domains

Most ODs are short sequences, in particular, the most recent ones, and more than 90% are located at the protein termini (Fig. 5). In addition, we predicted

the amount of low complexity and disorder in these sequences using Seg¹⁷ and IUPred¹⁸, which showed that ODs have larger fractions of residues in low-complexity segments and more predicted disorder than ancient regions (χ^2 test, $P < 10^{-5}$). However, this might reflect the terminal location of orphans, since the amount of disorder is elevated at the N terminus in general and not only in sequences identified as orphan. On the other hand, we did not see any such bias for low-complexity regions.

Furthermore, the few *S. cerevisiae* specific ODs have a comparatively small amount of low complexity and disorder (Fig. 5). It is well known that both low-complexity regions and disordered sequences can have biased residue compositions, which complicates sequence alignments.¹⁹ Hence, this difference between the SCE ODs and other ODs could reflect that many of the older ODs have low complexity or contain disordered sequences that cannot be aligned to their homologs, whereas many SCE ODs are *de novo* created orphans.

Distinguishing between true orphan domains and unaligned sequences

Due to the limitation of sequence alignments, it is difficult to distinguish between novel ODs and rapidly evolving sequences that are homologous but where homology cannot be detected. However, if we assume that creation makes the protein longer,

Fig. 5. Orphan domain characteristics. *Saccharomyces cerevisiae* protein sequence regions that lack homologs in other species were termed orphan domains. Orphans were defined at four different levels with increasing phylogenetic distance from *S. cerevisiae* [fungi (FUN), Hemiascomycota (HEM), *Saccharomyces* (SAC), and *S. cerevisiae* (SCE)]. The orphans were characterized and compared to ancient (ANC) domains (i.e., domains with homologs in non-fungi species). (a) Length distributions. (b) Relative positions within the protein chain, from the N terminus to the C terminus. (c) Proportion of residues in low-complexity regions. (d) Proportion of residues that are predicted to be disordered.

whereas rapid evolution does not, we may compare the lengths of the *S. cerevisiae* protein and its orthologs, thereby identifying true ODs. For example, a potential *de novo* created sequence under this assumption is the 102-aa-long N-terminal OD that was predicted in the alanine transaminase YLR089C (Fig. 6).

We found that 67 (28%) of the proteins with SAC ODs were longer than their orthologs. Of the remaining proteins, 88 (37%) had not increased, 31 (13%) belonged to protein families with complex length variations, and for 49 proteins (21%) no orthologs were detected. On the other hand, as many as 16 of the 17 proteins with SCE OD and detected orthologs were longer than these orthologs. Hence, a much larger fraction of the ODs at the *S. cerevisiae* level are classified as *de novo* created orphan sequences than at the *Saccharomyces* level. Again, this indicates that novel ODs were detected

when studying the most closely related species, while many ODs at the *Saccharomyces* level are probably results of a failure to align homologs correctly. This implies that we need to compare very closely related species to identify ODs reliably. Anyhow, our results agree with those of Giacomelli *et al.*¹¹ that short sequences can be added to proteins, leading to gradual extensions, preferably at the termini.

How can the existence of orphan domains be explained?

As described above, ODs are typically short sequences located at the protein termini. Thus, innovation might be explained by mutations that result in new translation start and stop codons^{20,11} and thereby incorporation of previously noncoding sequences in the coding region. This was studied by

Fig. 6. Orphan domain. (a) Species tree with the lengths (aa) of the orthologs of YLR089C in each branch. Abbreviations are shown for the species where orthologs were found. (b) N-terminal part of a multiple sequence alignment of YLR089C and a selection of orthologs. The predicted orphan domain is in boldface.

Table 2. Orphan protein characteristics

Proteins	Number	Length (aa)	Uncharacterized	Essential	Fitness	PPI
All	5879	496	0.18	0.18	0.83	0.86
<i>Saccharomyces</i> OP	125	215	0.59	0.01	0.76	0.54
<i>S. cerevisiae</i> OP	158	71	0.88	0.01	0.18	0.08

Length: average length of these proteins. Fitness: fraction of proteins with a fitness effect in a deletion study. PPI: fraction of proteins detected as prey in protein-protein interactions.

aligning the genomic sequences of the ODs with the untranslated regions of its orthologous genes in closely related genomes (*S. paradoxus*, *S. mikatae*, and *S. bayanus*). This confirmed that the open reading frames of the detected homologs were shorter in all sequences except two, and shifts of start or stop codons could potentially explain 16 of the *S. cerevisiae* ODs (Supplementary Material). These observations could be due to errors in the genome sequences, but in a recent study, Giacomelli *et al.* found such sequencing errors to be rather rare.¹¹ Finally, it can be noted that an indel close to the C terminus can cause a frameshift that creates an OD with comparatively small changes in protein length (which might explain the OD in YKR028W). Our discrimination of true ODs based on the length of orthologous proteins might fail to detect such examples.

Further, many ODs are disordered. Since short nucleotide motifs have been observed in disordered sequences,²¹ a mechanism for nucleotide duplication could have been used to increase the amount of coding sequences in some of these genes, as suggested by Kellis *et al.*²⁰ Thus, disordered sequences could appear to be orphan because they often evolve more rapidly than other sequences,^{19,22} or they might be true orphans if disorder expansion causes protein elongation.

Orphan proteins

Similar to observations in other species,^{6,7} OPs in *S. cerevisiae* are short (Table 2). However, the length distribution is shifted towards longer proteins at the less specific levels (i.e., SAC OPs are longer than SCE OPs) (Fig. 7). Further, the OPs contain somewhat more low-complexity sequences than ancient proteins, whereas the predicted amount of disorder and secondary structure content, as well as the average codon adaptation index, differ very little from older proteins (not shown). Hence, OPs are in many respects similar to older proteins, with the exception of *S. cerevisiae* specific OPs, which are exceptionally short, sometimes <50 residues. However, homology detection is less sensitive for shorter sequences; hence, some of these apparent OPs could be fast-evolving proteins where no homology could be found.

Are orphan proteins functional?

A problem that appears for OPs at the species-specific level is that it may be questioned if they are expressed and functional. However, this problem

should be comparatively small in *S. cerevisiae*, since its genome is well studied and many spurious genes have been removed after a comparative genomic study of closely related species.²⁰ Nonetheless, we examined the evidence for the functionality of these proteins using data from large-scale experimental studies.

As expected, OPs are rarely essential and most of them are uncharacterized (Table 2). However, a recent study demonstrated that deletion of most genes results in a fitness effect under at least some conditions.²³ A majority (95% or 76%) of the *Saccharomyces* OPs showed a fitness effect in this study,²³ and so did 28 (18%) of the *S. cerevisiae* OPs. However, it should be noted that most of the remaining OPs were not included in the deletion study and, hence, might also prove to be functional.

Fig. 7. Orphan protein characteristics. Proteins defined as orphan (i.e., lacking homologs) at different phylogenetic distances (levels) from *Saccharomyces cerevisiae* were characterized. Ancient proteins with homologs in non-fungi species were included for comparison. (a) Length distribution. (b) Disorder and low complexity. Proportion of residues that are predicted to be disordered or in low-complexity regions. Levels: ancient (ANC), fungi (FUN), Hemiascomycota (HEM), *Saccharomyces* (SAC), and *S. cerevisiae* (SCE).

Another experimental evidence for functionality of proteins is identification in protein complexes. Here, 67 (54%) of the *Saccharomyces* OPs were detected (as prey) in interactions,^{24,25} while only 13 (8%) of the *S. cerevisiae* OPs were detected (Table 2). In summary, a majority of the *Saccharomyces* level OPs have been experimentally verified to be functional, whereas (in these studies) only some *S. cerevisiae* OPs have been verified. It has been shown that the species-specific protein YNL269W evolved through a number of point mutations.¹⁰ Whether random mutations in noncoding sequences (i.e., *de novo* creation) are a general mechanism for protein innovation remains to be discovered. Perhaps, these short open reading frames can then combine with other genes or increase in length through *de novo* creation of domains.

Conclusions

In contrast to earlier work using domain-based methods, we have here used a pairwise sequence alignment approach for detection of ODs and OPs as well as to estimate the rate of protein innovation in *S. cerevisiae*. In agreement with the assumption that *de novo* creation is rare, less than 2% of the residues in *S. cerevisiae* were found to be species specific and up to 5% of the residues appear to be specific to the *Saccharomyces* lineage.

All methods used to identify orphan regions are dependent on sequence alignments, and short sequences as well as rapidly evolving sequences and sequences with biased amino acid compositions can be difficult to align correctly. Hence, a major problem when detecting orphans is to distinguish between these unalignable sequences and new genetic material created by some *de novo* mechanism. By studying the predicted ODs, we conclude that most orphans at the *S. cerevisiae* level are longer than their orthologs, and thus appear to have been created from noncoding DNA by a *de novo* mechanism. However, already at the *Saccharomyces* level, this length difference is not detectable, indicating that many unalignable sequences are included among the potential orphans. Hence, we conclude that to detect *de novo* created orphan sequences, it is necessary to compare proteins from very closely related species.

At the *S. cerevisiae* level, the OPs outnumber the ODs, and only about 20 ODs were found. Most of these ODs exist in one single copy and are mainly short domains located at the protein termini. The terminal location suggests a possible mechanism for their creation, involving changes of start and stop codons. In contrast, the *Saccharomyces* specific ODs contain a high amount of low-complexity regions and disorder, which suggests that nucleotide repetitions could have been involved in their creation. However, increased evolutionary rates are associated with disorder, and therefore, we believe that many of these are not true orphans but just rapidly evolving sequences.

Orphan proteins appear to be structurally similar to older proteins, but shorter. In particular, *S. cerevisiae* specific orphans are extremely short. Further, the functions of most orphans are uncharacterized. However, experimental studies have shown that at least some of these proteins are functional, since their deletion affects the fitness of *S. cerevisiae* or, alternatively, they are found in protein complexes. In addition, we found that the structure of two SAC-specific OPs have been determined. One of these, YNR034W-A (2grg), is a yeast protein of unknown function, whose structure is solved by a structural genomics project, and the other, YMR174C (chain B of 1dp5), is a proteinase inhibitor. Hence, at least some of the OPs appear to be *de novo* created functional proteins.

Finally, it is clear from this study that it is difficult to predict the exact number of true ODs and OPs. However, by comparing the *S. cerevisiae* proteins with proteins from several closely related species, we could identify some domains and proteins that appear to be truly orphan. Although many of the older orphans most likely are rapidly evolving sequences, it seems likely that the process of *de novo* protein creation is responsible also for some of the orphans detected at other levels.

Methods

Data

Fungi protein sequences were retrieved from the National Center for Biotechnology Information[†] and from the *Saccharomyces* Genome Database (SGD)[‡]. The following genomes were included in the study: *Ashbya gossypii*, *Aspergillus fumigatus*, *Candida albicans*, *Candida glabrata*, *Cryptococcus neoformans*, *Debaryomyces hansenii*, *Encephalitozoon cuniculi*, *Gibberella zeae*, *Cluyveromyces lactis*, *Magnaporthe grisea*, *Neurospora crassa*, *Saccharomyces bayanus*, *Saccharomyces castellii*, *Saccharomyces cerevisiae*, *Saccharomyces kluyveri*, *Saccharomyces mikatae*, *Schizosaccharomyces pombe*, *Ustilago maydis*, and *Yarrowia lipolytica* (for species tree, see Supplementary Material). Genomic sequences were also downloaded from SGD for *S. paradoxus*, *S. mikatae* and *S. bayanus*. The UniRef50 data set from UniProt was used for non-fungal proteins§.

Detection of orphans

Six different phylogenetic levels were defined in accordance with a published phylogenetic tree.²⁶ These levels ranged from ancient (non-fungi eukaryotes and prokaryotes) to species specific (i.e., *S. cerevisiae* specific) (Fig. 1). Each species belongs to the level of the last ancestral node shared with *S. cerevisiae*.

Each residue was assigned to one of the phylogenetic levels based on pairwise sequence alignments between the protein and its homologs, created with Blast (Fig. 2b). The low-complexity filter was used and high-scoring segment

[†] <http://ftp.ncbi.nih.gov/genomes/Fungi/>, June 2007.

[‡] <http://www.yeastgenome.org/>

[§] <http://www.uniprot.org/>, June 2007.

pairs with an *E*-value <0.001 were identified. To align the low-complexity regions, all proteins with low-complexity regions were realigned to the identified proteins without the low-complexity filter. Each residue was then assigned an age according to the most distantly related protein to which it was aligned. Results achieved with PSI-Blast¹³ were similar to those presented.

Defining orphan domains and proteins

From the residue assignments, orphan regions of ≥ 50 consecutive orphan residues were extracted. This was repeated for each level, and regions where all the residues were assigned to this, or a more specific level, were defined as orphan. If the entire protein was included in the region, it was defined as an orphan protein, whereas regions that covered only a part of the sequence (coexisting with older domains) were defined as orphan domains.

To estimate copy numbers of orphans, Blast was run with each orphan sequence against all *S. cerevisiae* proteins. Hits (*E*-value <0.001) that aligned at least half of the query sequence were clustered into families using single linkage clustering.

Blast (tblastn) was also used to search for homologs in the genomic sequences of *S. paradoxus*, *S. mikatae*, and *S. bayanus*. The length of the longest open reading frame in the matching region was compared with the length of the coding sequence in *S. cerevisiae*. These were considered similar if the length difference was <5% of the OP length. Similarly, the creation of SCE ODs through mutations resulting in new start/stop codons was studied by aligning the genomic sequences of these four species. The alignments were analyzed manually to confirm that start/stop codons were found at different positions in the other genomes.

Orphan domains and length variations between orthologs

To distinguish between true ODs and rapidly evolving sequences, we compared the lengths of proteins with ODs to the length of its orthologs. Proteins with *S. cerevisiae* specific ODs were compared to orthologs at the *Saccharomyces* level, whereas proteins with *Saccharomyces* ODs were compared with orthologs at the Hemiascomycota level. First, a protein was considered unchanged if the protein had changed by <30 residues compared to all orthologs. Otherwise, changes where the length had increased compared to all orthologs were defined as increasing, and other changes as complex. For proteins with identical domain architectures, the domains were used to align the proteins, and the proteins divided into separate regions corresponding to domains and unassigned regions. If the region that contained the OD had increased by $\geq 50\%$ of the OD length, the protein was classified as increasing.

Sequence analysis

The relative positions of ODs within proteins were calculated by dividing the protein into 10 parts of equal length. The average fraction of residues in each of the 10 parts was calculated. The multiple sequence alignment was created with ClustalW²⁷ and orthology was predicted with Inparanoid 1.0²⁸ between *S. cerevisiae* and each of the fungi species. The quality of the multiple sequence alignments was confirmed using kalign2.²⁹ Disorder was predicted with IUPred,¹⁸ and secondary structure with

PSIPRED.³⁰ Low-complexity sequences were predicted with SEG and codon usage was analyzed with the codon adaptation index³¹ from the emboss package. All programs were used with their default parameters, unless stated differently. A list of dubious and uncharacterized ORFs in *S. cerevisiae* was obtained from SGD. In addition, a list of essential *S. cerevisiae* genes was retrieved.³² A χ^2 test was used to determine if disorder and low complexity was significantly overrepresented in ODs compared to ancient domains. Here, we compared the number of disordered residues and the total number of residues in the two data sets. We also tested the number of regions with >40% disorder in the two sets; the results of the tests were similar.

HHpred

HHpred¹⁶ was run through the web interface, and all available databases (PDB, SCOP, CDD, interpro, Pfam, Smart, Panther, TIGRFAM, COG, KOG, CATH, SUPFAM, CD) in May 2009 were included in the HMM-HMM search. All hits with an *E*-value <0.001 were assumed to be significant. The results for these hits are available in Supplementary Material.

Acknowledgements

This work was supported by grants from the Swedish Research Councils, SSF (the Foundation for Strategic Research). The EU 6Th Framework Program is gratefully acknowledged for support to the GeneFun project, contract no. LSHG-CT-2004-503567, and the EMBRACE project, contract no. LSHG-CT-2004-512092. We are grateful to Johannes Söding for the use of the HHpred web server and to Marcin Skwark for creating the script that allowed the use of this server automatically.

Supplementary Data

Supplementary data associated with this article can be found, in the online version, at doi:10.1016/j.jmb.2009.11.053

References

- Fischer, D. & Eisenberg, D (1999). Finding families for genomic ORFans. *Bioinformatics*, **15**(9), 759–762.
- Ekman, D., Bjorklund, A. K., Frey-Skott, J. & Elofsson, A. (2005). Multi-domain proteins in the three kingdoms of life: orphan domains and other unassigned regions. *J. Mol. Biol.* **348**(1), 231–243.
- Lupas, A., Ponting, C. & Russell, R. (2001). On the evolution of protein folds: Are similar motifs in different protein folds the result of convergence, insertion, or relics of an ancient peptide world? *J. Struct. Biol.* **134**, 191–203.
- Long, M., Betran, E., Thornton, K. & Wang, W (2003). The origin of new genes: glimpses from the young and old. *Nat. Rev. Genet.* **4**(11), 865–875.
- Cai, J. J., Woo, P. C., Lau, S. K., Smith, D. K. & Yuen, K. Y. (2006). Accelerated evolutionary rate may be

- responsible for the emergence of lineage-specific genes in Ascomycota. *J. Mol. Evol.* **63**(1), 1–11.
6. Domazet-Lošo, T. & Tautz, D (2003). An evolutionary analysis of orphan genes in *Drosophila*. *Genome Res.* **13** (10), 2213–2219.
 7. Guo, W. J., Ling, J., Ye, S. P. Significant comparative characteristics between orphan and nonorphan genes in the rice (*Oryza sativa* L.) genome. *Comp. Funct. Genomics*, 21676.
 8. Siew, N. & Fischer, D (2003). Analysis of singleton ORFans in fully sequenced microbial genomes. *Proteins*, **53**(2), 241–251.
 9. Rusch, D., Halpern, A., Sutton, G., Heidelberg, K., Williamson, S., Yoosheph, S. *et al.* (Mar 2007). The sorcerer II global ocean sampling expedition: north-west atlantic through eastern tropical pacific. *PLoS Biol.* **5**(3), e77.
 10. Cai, J., Zhao, R., Jiang, H. & Wang, W (2008). De novo origination of a new protein-coding gene in *Saccharomyces cerevisiae*. *Genetics*, **179**, 487–496.
 11. Giacomelli, M. G., Hancock, A. S. & Masel, J (2007). The conversion of 3' UTRs into coding regions. *Mol. Biol. Evol.* **24**(2), 457–464.
 12. Bjorklund, A. K., Ekman, D., Light, S., Frey-Skott, J. & Elofsson, A. (2005). Domain rearrangements in protein evolution. *J. Mol. Biol.* **353**(4), 911–923.
 13. Altschul, S., Madden, T., Schaffer, A., Zhang, J., Zhang, Z., Miller, W. & Lipman, D (1997). Gapped blast and psi-blast: a new generation of protein database search programs. *Nucleic Acids Res.* **25**, 3389–3402.
 14. Wolfe, K. H. & Shields, D. C (1997). Molecular evidence for an ancient duplication of the entire yeast genome. *Nature*, **387**(6634), 708–713.
 15. Eddy, S (1998). Profile hidden Markov models. *Bioinformatics*, **14**, 755–763.
 16. Soding, J., Biegert, A. & Lupas, A (Jul 2005). The HHpred interactive server for protein homology detection and structure prediction. *Nucleic Acids Res.* **33**(Web Server issue), W244–W248.
 17. Wootton, J. & Federhen, S (1996). Analysis of compositionally biased regions in sequence databases. *Methods Enzymol.* **266**, 554–571.
 18. Dosztányi, Z., Csizmadák, V., Tompa, P. & Simon, I. (2005). The pairwise energy content estimated from amino acid composition discriminates between folded and intrinsically unstructured proteins. *J. Mol. Biol.* **347**(4), 827–839.
 19. Romero, P., Obradovic, Z., Li, X., Garner, E. C., Brown, C. J. & Dunker, A. K (2001). Sequence complexity of disordered protein. *Proteins*, **42**(1), 38–48.
 20. Kellis, M., Patterson, N., Endrizzi, M., Birren, B. W. & Lander, E. S (2003). Sequencing and comparison of yeast species to identify genes and regulatory elements. *Nature*, **423**(6937), 241–254.
 21. Tompa, P (2003). Intrinsically unstructured proteins evolve by repeat expansion. *BioEssays*, **25**(9), 847–855.
 22. Brown, C., Takayama, S., Campen, A., Vise, P., Marshall, T., Oldfield, C. *et al.* (2002). Evolutionary rate heterogeneity in proteins with long disordered regions. *J. Mol. Evol.* **55**(1), 104–110.
 23. Hillenmeyer, M. E., Fung, E., Wildenhain, J., Pierce, S. E., Hoon, S., Lee, W. *et al.* (2008). The chemical genomic portrait of yeast: uncovering a phenotype for all genes. *Science*, **320**(5874), 362–365.
 24. Gavin, A., Aloy, P., Grandi, P., Krause, R., Boesche, M., Marzioch, M. *et al.* (2006). Proteome survey reveals modularity of the yeast cell machinery. *Nature*, **440** (7084), 631–636.
 25. Krogan, N., Cagney, G., Yu, H., Zhong, G., Guo, X., Ignatchenko, A. *et al.* (2006). Global landscape of protein complexes in the yeast *Saccharomyces cerevisiae*. *Nature*, **440**(7084), 637–643.
 26. Fitzpatrick, D. A., Logue, M. E., Stajich, J. E. & Butler, G. (2006). A fungal phylogeny based on 42 complete genomes derived from supertree and combined gene analysis. *BMC Evol. Biol.* **6**, 99.
 27. Thompson, J. D., Higgins, D. & Gibson, T (1994). Clustal w: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, positions-specific gap penalties and weight matrix choice. *Nucleic Acids Res.* **22**, 4673–4680.
 28. Remm, M., Storm, C. E. & Sonnhammer, E. L. (2001). Automatic clustering of orthologs and in-paralogs from pairwise species comparisons. *J. Mol. Biol.* **314**(5), 1041–1052.
 29. Lassmann, T., Frings, O. & Sonnhammer, E (Feb 2009). Kalign2: high-performance multiple alignment of protein and nucleotide sequences allowing external features. *Nucleic Acids Res.* **37**(3), 858–865.
 30. McGuffin, L., Bryson, K. & Jones, D (Apr 2000). The PSIPRED protein structure prediction server. *Bioinformatics*, **16**(4), 404–405.
 31. Sharp, P. M. & Li, W. H (1987). The codon adaptation index—a measure of directional synonymous codon usage bias, and its potential applications. *Nucleic Acids Res.* **15**(3), 1281–1295.
 32. Winzeler, E. A., Shoemaker, D. D., Astromoff, A., Liang, H., Anderson, K. & Andre, B. (1999). Functional characterization of the *S. cerevisiae* genome by gene deletion and parallel analysis. *Science*, **285**(5429), 901–906.