Integration of new genes into cellular networks, and their structural maturation

György Abrusán

Present address:

Synthetic and Systems Biology Unit, Institute of Biochemistry, Biological Research Centre of the Hungarian Academy of Sciences,
Running title: Dynamics of proto-gene functionalization

Key words: *de novo* genes, protein-protein interaction, regulatory network, secondary structure, aggregation, *Saccharomyces cerevisiae*

Corresponding author: György Abrusán
Synthetic and Systems Biology Unit, Institute of Biochemistry, Biological Research Centre of the Hungarian Academy of Sciences, Temesváry krt. 62. Szeged, H-6701, Hungary
E-mail: abrusan@brc.hu
Tel.: +3612831653
Fax: +3662433506
ABSTRACT

It has been recently discovered that new genes can originate *de novo* from non-coding DNA, and several biological traits including expression or sequence composition form a continuum from non-coding sequences to conserved genes. In this paper, using yeast genes I test whether the integration of new genes into cellular networks, and their structural maturation shows such a continuum by analyzing their changes with gene age. I show that 1) the number of regulatory, protein-protein and genetic interactions increase continuously with gene age, although with very different rates. New regulatory interactions emerge rapidly within a few million years, while the number of protein-protein and genetic interactions increases slowly, with a rate of $2-2.25 \times 10^{-8}$ per year and $4.8 \times 10^{-8}$ per year, respectively. 2) Gene essentiality evolves relatively quickly: the youngest essential genes appear in proto genes ~14 my old. 3) In contrast to interactions, the secondary structure of proteins and their robustness to mutations indicate that new genes face a bottleneck in their evolution: proto-genes are characterized with high beta strand content, high aggregation propensity and low robustness against mutations, while conserved genes with lower strand content and higher stability, most likely due to the higher probability of gene loss among young genes and accumulation of neutral mutations.
INTRODUCTION

The established view is that new genes evolve primarily by duplications and recombination, i.e. by reorganization of existing domains of other genes (LONG et al. 2003; KAESSMANN 2010). Recent studies however highlighted that genes can also emerge from non-coding DNA: the de novo emergence of genes has been demonstrated in a number of cases in Drosophila (LEVINE et al. 2006; BEGUN et al. 2007; ZHOU et al. 2008), humans (TOLL-RIERA et al. 2009; KNOWLES and McLYSAGHT 2009; WU et al. 2011; XIE et al. 2012), rodents (HEINEN et al. 2009; MURPHY and McLYSAGHT 2012), yeast (CAI et al. 2008; CARVUNIS et al. 2012) and viruses (RANCUREL et al. 2009; SABATH et al. 2012). While orphan genes – genes with no homologs – are common in genomes (TAUTZ and DOMAZET-LOŠO 2011), proving that a gene emerged de novo is not straightforward, as the presence of a nonfunctional parent sequence also has to be demonstrated, thus their presence can be convincingly demonstrated only in species where also close relatives have been sequenced. However, the fact that they have been detected in several model organisms indicates that their emergence is not exceptionally rare, and may be an ongoing process in many, if not most genomes.

Although the de novo origination of genes is not questioned anymore, little is known about its frequency, and the subsequent fate of these genes in the genome: i.e. whether they are subject to turnover, how rapidly new protein-protein interactions are formed (LYNCH 2012) and integrated into regulatory networks (DING et al. 2010; CHEN et al. 2012), and also whether structural changes affect the novel proteins. Real proteins are not random structures, i.e. they can be characterized with distinct biophysical properties like stability or aggregation propensity (DEPRISTO et al. 2005; MONSELLIER and CHITI 2007), although, quite surprisingly, some of their basic features like the presence of secondary structural elements (alpha helices and beta strands) are already formed in random sequences (SCHAEFER et al. 2010). Genome-wide studies of transcription in several species indicate that a large fraction of genome, including the non-coding part is transcribed in most species (KAPRANOV et al. 2007;
NAGALAKSHMI et al. 2008; XU et al. 2009), and some of the non-coding transcripts are associated with ribosomes and occasionally translated (WILSON and MASEL 2011). A recent large scale study in yeast demonstrated the existence of a continuous change in the level of expression, selective constraints and codon adaptation index from recently emerged “proto-genes” to highly conserved, ancient genes (CARVUNIS et al. 2012), and suggested that genes can be placed in a continuum from non-genic sequences to conserved genes, and that de novo emergence of genes may be as common as emergence by the classic duplication-divergence mechanism. This is also supported by the phylogenetic pattern of gene emergence in vertebrates (NEME and TAUTZ 2013), indicating that young genes are typically short, and their length, number of exons and domains changes gradually over time.

This study has two goals. First, to examine as broadly as possible what is the fate of the new translatable open reading frames (ORFs) after their emergence, i.e. to test how rapidly they are integrated into regulatory, protein-protein or genetic interaction networks, whether they are stable in the genome, and whether there are characteristic sequence or structural changes in the process of their functionalization. The second aim is to test whether the observed patterns also support the theory proposed by (CARVUNIS et al. 2012), namely that non-genic sequences and well characterized genes form a continuum, and no sharp boundaries exist between them.

MATERIALS AND METHODS

Categorization of yeast genes into age groups. Similarly to (CARVUNIS et al. 2012), genes of Saccharomyces cerevisiae were divided into 11 groups according to their evolutionary conservation within fungi. Conservation level 0 corresponds to unannotated genes that were identified by (CARVUNIS et al. 2012), and are longer than 50 amino acids. Genes with conservation level 1 are genes that were annotated as genes by the Saccharomyces Genome Database (CHERRY et al. 2012) but have no orthologs in other species, higher conservation
levels indicate the presence of orthologs in other fungi (see Figure 1), i.e. conservation level N means that the orthologs of the gene are present in any species of the branch rooted at the Nth bifurcation (Figure 1), but not in species that split earlier from the *Saccharomyces* lineage. Following (CARVUNIS *et al.* 2012), I use the term proto-genes for genes that are present only in *S. cerevisiae* and its closest relatives (conservation levels 0-4).

The classification of genes into proto-genes (conservation levels 0-4) was taken directly from (CARVUNIS *et al.* 2012); the sequences of proto-genes with conservation level 0 (which are absent in the Saccharomyces Genome Database) were obtained using the genomic coordinates provided by the same study, using the R56 (20070406) release of the genome. Only sequences with a minimum length of 50 amino acids were used in the analysis. Conserved genes were classified into conservation groups (Figure 1) as in (CARVUNIS *et al.* 2012) using the fungal gene orthology dataset provided by (WAPINSKI *et al.* 2007) [http://www.broadinstitute.org/regev/orthogroups](http://www.broadinstitute.org/regev/orthogroups). The same orthology dataset was used to determine gene losses in *S. paradoxus* and *S. mikatae* (see Results).

The age estimates of the main branching events of the phylogeny of fungi (Figure 1) were obtained using TimeTree (HEDGES *et al.* 2006), except for *S. castellii*, which split from the *Saccharomyces* lineage 100-150 mya, after the whole genome duplication of yeasts (CLIFTEN *et al.* 2006). For nodes without age estimates such estimates are either unavailable, or conflict with the phylogeny of the species.

**Data sources.** The yeast genome and protein sequences were downloaded from the Saccharomyces Genome Database (CHERRY *et al.* 2012); recent duplications were removed from the protein dataset by clustering them at 80% amino acid similarity with uclust (EDGAR 2010).

The fitness effect of yeast genes (i.e. the effect of single gene deletion on growth) was obtained from (DEUTSCHBAUER *et al.* 2005). Yeast protein-protein interactions and genetic
interactions were downloaded from the BioGRID (v. 3.1.91) database (STARK et al. 2011), regulatory interactions were downloaded from the YEASTRACT database (ABDULREHMAN et al. 2011). The strength of genetic interactions (genetic interaction score, ε) was obtained from Costanzo et al. 2010 (COSTANZO et al. 2010); the largest interaction dataset with the “lenient cutoff” (p < 0.05) was used in the analysis, because it includes also weak interactions. Each gene was characterized with the average of the |ε| values of its significant genetic interactions, which was calculated separately for proto-genes, and conserved genes. The number and strength of protein-protein, genetic and regulatory interactions for each gene was determined with Perl scripts, developed in house. Due to their low expression and low conservation, the BioGRID database is probably significantly biased against proto-genes, thus I included only those genes in the analysis which have interactions in the databases, even if the majority of genes in the conservation group have no reported interactions (this is the case with protein-protein and genetic interactions for conservation levels 1-3).

Since in Saccharomyces cerevisiae RNAi is absent, the fitness effect of individual genes, genetic interactions, and epistasis between genes is determined with the deletion of genes (see (DEUTSCHBAUER et al. 2005)) as opposed to RNAi knockdown used in many multicellular organisms (DIXON et al. 2009). However, this also means that in the case of overlapping genes deletions target more than one gene, and thus neither fitness measurements, nor estimates of the strength of genetic interactions are accurate. In consequence, in the analyses involving fitness, genetic interactions and epistasis overlapping genes were excluded from the dataset.

**Determination of secondary structure and mutagenesis of proteins.** In the structural analysis, besides the sequences with conservation levels 0-10 I also used amino acid sequences that were generated randomly, with the same length and amino acid frequencies as the youngest proto-genes with conservation levels 0-1. The secondary structures of the
sequences were determined with PSSpred (http://zhanglab.ccmb.med.umich.edu/PSSpred/), aggregation propensity with Tango (FERNANDEZ-ESCAMILLA et al. 2004). Since protein sequences with different conservation levels have very different numbers of protein homologs, and secondary structure prediction includes homology search, to avoid biases PSSpred was run without the use of the homology search step (psi-blast), which reduces the accuracy of secondary structure prediction from 81% to 69%.

The mutagenesis of the yeast proteins was based on (SCHAEFER et al. 2010) and was performed as follows (see also Figure 6). Each sequence was mutated gradually in 70 steps, and in each step 1% of the residues were changed in the sequence. Since protein sequences do not evolve entirely randomly, to simulate realistic mutations, new residues were introduced in a context specific manner, using the csbuild tool of the CS-BLAST suite (BIEGERT and SÖDING 2009). In each step, the secondary structure of the mutated sequence was determined, and the fraction of residues with a similar secondary structure to the original secondary structure (Q3 value) was calculated. For each yeast protein sequence, the mutagenesis was repeated five times independently, and the average of the five replicates was used in the analyses. To save computing time, from genes with conservation level 10 only 10% of the sequences were mutated (which were chosen randomly).

RESULTS

Rapid emergence of regulatory interactions. The functionalization of a new gene means that, besides evolving a translatable open reading frame it needs to become transcriptionally active, i.e. acquire a promoter, and have to be co-regulated with the genes it interacts with, or other genes that are expressed in the same life stages or environmental conditions. Studies of novel retrogenes indicate that regulatory elements of new genes can be acquired from nearby genes or even from more distant locations (KAESSMANN et al. 2009; KAESSMANN 2010). Using the YEASTRACT database (ABDULREHMAN et al. 2011) I analyzed the integration of
proto-genes into the yeast regulatory network by quantifying the number of genes that share the same transcription factors (TFs) thus are co-regulated, the number of TFs regulating the genes, and the emergence of a simple network motif, the feed-forward loop. The results show that the integration of genes into the regulatory network happens gradually, nevertheless very quickly; while young proto genes with conservation levels 1-3 are co-regulated with a significantly smaller number of genes than genes with conservation level 5-10 (p < 0.05 for all comparisons, ANOVA, Bonferroni post hoc tests, Figure 2A), the genes in the youngest group (conservation level 1) are already co-regulated with thousands of other genes, and the oldest proto-genes with conservation level 4 (ca. 14 million years old) are not significantly different from most older groups (conservation levels 5, 6, 8, 10, p > 0.05, ANOVA, Bonferroni post hoc tests). A comparable pattern is visible for the number of transcription factors regulating the genes (Figure 2B), proto-genes with conservation level 1-3 are regulated by significantly less transcription factors than conserved genes, or even proto genes with conservation level 4 (p < 0.05, ANOVA, Bonferroni post hoc tests), however already the youngest group (cons. level 1) is regulated by several transcription factors.

Regulatory networks are characterized with significantly overrepresented patterns of TFs, termed network motifs, which are capable to perform distinct regulatory functions (Lee et al. 2002; Odom et al. 2006). One of the most important such motif is the feed-forward loop (FFL), which consists of three genes; two TFs, one of which regulates the other, and both regulating the same target gene. Depending on the type of interactions between their components, FFLs can speed up or delay the response of target genes (Mangan and Alon 2003). I identified all FFL-like motifs in the network of YEASTRACT TFs, and tested whether the number of FFLs regulating a gene changes with gene age. Due to the high connectivity of the network and overlaps between motifs a large number of FFLs can be identified in YEASTRACT, nevertheless I found a similar, increasing pattern as with the number of TFs; i.e. conserved genes are regulated by proportionally more FFLs than proto-
genes (p < 0.05, ANOVA, Bonferroni post hoc tests, Figure 2C), indicating a gradual and rapid acquisition of regulatory motifs that enable complex regulatory behaviors.

**Essential genes are already present among proto-genes.** Recent findings in *Drosophila* indicate that essential genes (i.e. genes where deletion mutants or knockouts are not viable), can evolve rapidly (CHEN et al. 2010) even though they are most common among ancient genes. Using the fitness dataset provided by (DEUTSCHBAUER et al. 2005) I tested how much time is necessary for the emergence of essential genes in yeast. The results show that the youngest essential genes (YEL035C, YPL124W) are present among the proto-genes with conservation level 4, with homologs in *S. bayanus* (Figure 3), which split from the *S. cerevisiae* lineage ca. 14 mya. However, essentiality may evolve even faster: the deletion of several other proto-genes (already 7 in genes with conservation level 1) have lethal phenotypes, which however is likely to be due to their overlap with conserved genes. Determining their independent fitness effect would require fitness estimates that are independent from the effect of the genes they overlap with, e.g. obtained by RNAi introduced to yeasts (DRINNENBERG et al. 2009).

**Protein secondary structure and aggregation propensity.** The analysis of secondary structure motifs in genes with different conservation level shows a surprising pattern. The abundance of alpha helices is not influenced by the age of the proteins; helices take up approximately 40% of the sequence, regardless of conservation level, even in random amino acid sequences (Figure 4A, p > 0.05 for comparisons between proto-genes and conserved genes, Bonferroni post hoc tests, ANOVA). In contrast, the amount of predicted beta sheets is highest in proto-genes and random sequences (ca. 20%), and shows a decline to ca. 10% with increasing level of conservation (Figure 4B, p << 0.001 for all comparisons between proto-genes and conserved genes, Bonferroni post hoc tests, ANOVA). This suggests that beta
strands are either evolutionarily unstable, and gradually disappear as the protein sequence accumulates neutral mutations, or their declining abundances are caused by natural selection. One selective force can be aggregation propensity (MONSELLIER and CHITI 2007), as it depends largely on the presence of beta strands (JAHN et al. 2010). Predicted aggregation propensity shows an even stronger pattern than beta strands; the average aggregation propensity of random sequences and proto-genes is almost three times higher than of conserved genes (Figure 4C, p << 0.001, for all comparisons between proto-genes and conserved genes, ANOVA).

**Turnover of proto-genes.** Functional non-coding sequences are short-lived, and subject to rapid turnover in mammalian genomes (PONTING et al. 2011). If natural selection is the cause of beta strand loss, it may manifest itself as a gradual loss of strands due to mutations, or the loss of entire genes with high strand content. I examined the importance of the latter mechanism by testing how the probability of gene loss depends on the age of the gene. I used genes from conservation levels 4-10, and using the orthogroup dataset provided by (WAPINSKI et al. 2007) I quantified gene losses in the youngest Saccharomyces species, S. paradoxus and S. mikatae (conservation levels 2-3). The results show that relatively young genes with conservation levels 4-5 are lost at significantly higher frequencies in these species than more conserved genes (40% is lost as opposed to 7-10%, p < 0.05 for all comparisons, Chi-square tests, Figure 5), which corresponds well with the pattern of beta strand loss, and indicates that the turnover of proto genes is likely to contribute to the observed structural change.

**Robustness of secondary structures against mutations.** To test the effect of mutation accumulation on the secondary structure and robustness of yeast proteins, I performed *in-silico* mutagenesis studies. I followed a procedure that was based on (SCHAEFER et al. 2010); mutations were introduced gradually into the protein sequences, and I examined how rapidly
the original secondary structure degrades with the change of the amino acid composition of
the sequence (see Figure 6 and methods for details). The results show that proto-genes are
more sensitive to mutations than older genes, i.e. their secondary structure changes faster with
sequence change than the structure of conserved genes (Figure 7A, \( p << 0.001 \), Bonferroni
post hoc tests, ANCOVA with conservation level as categorical predictor and sequence
divergence as continuous predictor). However, as their structural composition is also different
from more conserved genes (Figure 4), this may be a by-product of compositional differences.
The comparison of the mutational robustness of alpha-helices and beta-sheets indicates that
beta sheets decay considerably faster from random mutations than alpha helices (Figure 7B, \( p
<< 0.001 \), ANCOVA), and the robustness of proteins for mutations is negatively correlated
with their beta-strand content (Figure 7C, \( p << 0.001 \), \( R = -0.54 \)). This suggests that beta
strands are more difficult to maintain over longer evolutionary periods than alpha helices, and
their reduced amount in conserved genes may also be the result of the accumulation of neutral
mutations. Repeating the stability analysis only for the non-strand regions of the proteins
indicates, that proto genes remain more sensitive for mutations (Figure 7D, \( p << 0.001 \),
Bonferroni post hoc tests, ANCOVA), even after the exclusion of beta strands from the
analysis, thus the higher robustness of old genes compared to proto-genes is not merely a by-
product of their lower beta strand content.

The number and strength of genetic and protein-protein interactions increases with the
age of genes. Proteins typically perform their function in association with other proteins, and
form protein complexes. Physical interactions between proteins are largely determined by
their structure (ZHANG et al. 2012), and it has been demonstrated that the interacting surfaces
can frequently be characterized by a similar arrangement of secondary structure elements
(TUNCBAG et al. 2008; ZHANG et al. 2010; GAO and SKOLNICK 2010). The size and age of
protein complexes are highly variable, and currently there is no consensus on their rate of
evolution (Qian et al. 2011; Lewis et al. 2012a). In a previous study (Capra et al. 2010) have demonstrated that yeast genes younger than the whole genome duplication are less integrated into protein-protein interaction networks than older genes. I tested how the age (conservation level) of yeast proteins influences the number of their physical interactions, and whether it mirrors changes in the secondary structure of proteins. Unlike secondary structure elements, protein-protein interactions increase gradually with the age of proteins (Figure 8A) and appears to be a monotonous process; using conservation levels 6 and 10 for the estimate, the median number of interactions increases at a rate of $2-2.25 \times 10^{-8}$ per year (7/311 my and 15/760 my).

Genetic interactions mean that the synergistic effect of genes $i$ and $j$ on fitness ($f_{ij}$) is different than it is expected by their multiplicative effects: $f_{ij} = f_i f_j + \varepsilon$, and $\varepsilon \neq 0$; where $f_j$ denotes the effect of gene $j$ on fitness, and $\varepsilon$ the magnitude of epistasis between the two genes (Dixon et al. 2009; Costanzo et al. 2010). An extreme example is the synthetic lethal phenotype, where the independent deletion of two genes in both cases results in viable phenotypes, but their double deletion mutant is lethal. I tested whether the number of genetic interactions depends on gene age, and also found a positive relationship between the two (Figure 8B): the number of genetic interactions increase at a rate of $4.8 \times 10^{-8}$ per year (15/311 mya), however it appears to level off at conservation level 6 (311 mya). Similarly to the number of genetic interactions, also their strength changes with the age of the genes: the magnitude of epistasis ($|\varepsilon|$) of the interactions of proto genes is significantly lower than of conserved genes ($p << 0.001$, Mann-Whitney U test, Figure 9).

**DISCUSSION**

Overall, the integration of new genes into cellular networks supports the hypothesis of (Carvunis et al. 2012) that a continuum exists between conserved genes and proto-genes. The fastest process is the integration of proto-genes into regulatory networks (Figure 2):
already the youngest proto-genes are regulated by several TFs, co-regulated with a large number genes, and many are regulated by network motifs (FFLs). The oldest proto-genes, which appeared ~14 mya are almost as well integrated into the regulatory network as the much more ancient fraction of the yeast proteome, which appeared before the \textit{S. cerevisiae} – \textit{S. pombe} split. Although this analysis focused on the topological features of the entire regulatory network (i.e. ignored that in different environmental conditions or life stages only subsets of the genes are expressed), this is in agreement with the observations that the acquisition of regulatory sequences by new (retro)genes can happen rapidly (KAESSMANN et al. 2009), horizontally transferred genes in bacteria acquire regulatory interactions within few million years (LERCHER and PÁL 2008), and also with the findings that the proportion of proto-genes in proximity to known TF binding sites changes little with conservation level (CARVUNIS et al. 2012). The fact that youngest essential genes are present already among proto-genes (Figure 3), which is in agreement with experimental findings reporting differences between the sets of essential genes between yeast strains (DOWELL et al. 2010) also indicates that proto-genes gain functions rapidly. Unlike regulatory evolution, the acquisition of protein-protein interactions (PPI) is a dramatically slower process, which does not appear to level off with protein age, at least not within the range of examined conservation levels (Figure 8). The observed rate of PPI gain is at least an order of magnitude faster than the loss of interactions, which was estimated to occur at a rate of $10^{-9}$ (LEWIS et al. 2012b) to $1.6-2.6 \times 10^{-10}$ (QIAN et al. 2011). In addition, as the interaction databases are incomplete, the rate of PPI gain is likely to be underestimated. In the case of genetic interactions the increase in the number of new interactions is less pronounced above conservation level 6 (311 mya, Figure 8), although the oldest group (cons. level. 10) is still characterized with significantly higher number of interactions than genes with conservation levels 6-9. Unlike PPIs, genetic interactions allow also the measurement of the strength of interactions ($\varepsilon$), which indicates
stronger epistasis between conserved genes than proto-genes, thus the change is not only quantitative but also qualitative (Figure 9).

In contrast, the structural maturation of new genes shows a different pattern: the secondary structures and structural stability of proteins indicates that proto-genes and conserved genes form two, relatively distinct groups: proto-genes with high beta strand content, high aggregation propensity and low robustness against mutations, and conserved genes with lower strand content and higher stability. Proto-genes with conservation levels 0-3 are not much different from random sequences (with similar amino acid frequency), and the transition between the two states falls between conservation levels 4 (~14 mya) and 6 (~311 mya) thus requires considerably more time than regulatory evolution (Figures 4, 7). There are at least two mechanisms that may cause this shift; selection against high beta strand content, or a neutral process. Although selection favoring mutations that reduce aggregation propensity of proto-genes may seem as an attractive explanation for the reduction seen in beta strands, and the fitness cost of misfolding-induced aggregation was even demonstrated experimentally in yeast (Geiler-Samerotte et al. 2011), the tests of selection by (Carvunis et al. 2012) reported weak purifying selection in proto-genes, and not strong adaptive evolution. Alternatively, the reduction of strand-content and aggregation propensity may reflect that, despite being already partly integrated into regulatory networks, a considerable fraction of proto-genes does not “make it” to higher conservation levels, and are lost relatively quickly after their appearance. This is also indicated by phylostratigraphic studies, which show that the highest number of founder genes typically form a peak in the youngest evolutionary strata (Tautz and Domazet-Lošo 2011), implying that proto-genes are subject to some form of turnover, similarly to what has been recently shown for functional non-coding sequence in mammals (Meader et al. 2010). Gene deletion and inactivation studies show that 80-90% of genes in eukaryotes and prokaryotes can be lost individually without a significant fitness effect (Korona 2011), at least in laboratory conditions, and it has been
suggested that genes that are lost easily during evolution are less important, i.e. have lower expression levels, fewer protein-protein interactions (Krylov et al. 2003), or higher evolutionary rate (Zhang and He 2005). The analysis of gene loss in this study indicates that proto-gene turnover exists; young genes with lower conservation levels (4-5) are lost considerably more easily in Saccharomyces paradoxus and Saccharomyces mikatae than ancient genes (Figure 5), which are more integrated into cellular networks. An alternative explanation for the decay of beta strands that it is a neutral process, and while the amount of alpha helices is maintained by selection, the loss of strands due to neutral mutations is not compensated.

The robustness of secondary structure to mutations shows a similar shift between conservation levels 4-6 as beta strand content and aggregation propensity: proto-genes are less robust than conserved genes, even if one takes into consideration their higher beta strand content (Figure 7). The amount of mutations proteins can accumulate without a change in their structure is one of the key determinants of their capacity to produce evolutionary innovations, because more robust proteins can better tolerate innovative but destabilizing mutations (Bloom et al. 2006b; Ferrada and Wagner 2008). The pattern observed in the yeast proteome is in line with experimental findings on individual proteins, which suggest that robustness against mutations promotes evolvability (Bloom et al. 2006b), and that evolution favors mutational robustness (Bloom et al. 2007; Bershtein et al. 2008). Additionally, large scale analyses of folding energies (ΔG) of protein structures also indicate that ancient, eukaryote-wide proteins are thermodynamically more stable and robust against mutations than younger ones present across metazoans or vertebrates (Toll-riera et al. 2012). Moreover, the lower tolerance of beta strands for mutations predicts that strands, at least in conserved proteins, should evolve at lower rates than helices or coils, what was indeed found in yeast (Bloom et al. 2006a), although this effect is weaker than the effects of other structural traits like solvent accessibility.
To conclude, the results show a somewhat contradictory picture of the evolution of new genes: the integration of proto-genes into cellular networks shows a continuum, although the rates of regulatory evolution and the gain of protein and genetic interactions are very different. In contrast, from the structural point of view new genes do seem to face a bottleneck: proto-genes and conserved genes form relatively distinct groups, with different beta-strand content, aggregation propensity, and robustness for mutations. This, together with the finding that young genes are lost much more easily than conserved ones indicates that, even if they already have some functionality, young genes are still unstable in the genome.

COMPETING INTERESTS

The author declares that he has no competing interests.

SUPPLEMENTARY DATA

Supplementary material with the data used in the calculations is available in the online version of the paper: Supplementary Table 1.

ACKNOWLEDGMENTS

I thank Balázs Papp and the referees for useful comments and suggestions. This study was supported by the Hungarian Scientific Research Fund (OTKA) grant PD83571 to G.A.

REFERENCES

access to curated transcriptional regulatory associations in Saccharomyces cerevisiae through a web services interface. Nucleic Acids Res. 39: D136–140.


CHEN S., NI X., KRINSKY B. H., ZHANG Y. E., VIBRANOVSKI M. D., WHITE K. P., LONG M.,
2012 Reshaping of global gene expression networks and sex-biased gene expression
by integration of a young gene. EMBO J. 31: 2798–2809.


CHERRY J. M., HONG E. L., AMUNDSEN C., BALAKRISHNAN R., BINKLEY G., CHAN E. T.,
CHRISTIE K. R., COSTANZO M. C., DWIGHT S. S., ENGEL S. R., FISK D. G.,
HIRSCHMAN J. E., HITZ B. C., KARRA K., KRIEGER C. J., MIYASATO S. R., NASH R.
S., PARK J., SKRZYPEK M. S., SIMISON M., WENG S., WONG E. D., 2012
Saccharomyces Genome Database: the genomics resource of budding yeast. Nucleic
Acids Res. 40: D700–705.

CLIFTEN P. F., FULTON R. S., WILSON R. K., JOHNSTON M., 2006 After the duplication: gene

COSTANZO M., BARYSHNIKOVA A., BELLAY J., KIM Y., SPEAR E. D., et al., 2010 The

DEPRISTO M. A., WEINREICH D. M., HARTL D. L., 2005 Missense meanderings in sequence

DEUTSCHBAUER A. M., JARAMILLO D. F., PROCTOR M., KUMM J., HILLENMEYER M. E.,
DAVIS R. W., NISLOW C., GIAEVER G., 2005 Mechanisms of haploinsufficiency


GEILER-SAMEROTTE K. A., DION M. F., BUDNIK B. A., WANG S. M., HARTL D. L.,

DRUMMOND D. A., 2011  Misfolded proteins impose a dosage-dependent fitness cost

HEDGES S. B., DUDLEY J., KUMAR S., 2006  TimeTree: a public knowledge-base of

HEINEN T. J. A. J., STAUBACH F., HÄMING D., TAUTZ D., 2009  Emergence of a new gene

727.

KAESSMANN H., 2010  Origins, evolution, and phenotypic impact of new genes. Genome
Res. 20: 1313–1326.

KAESSMANN H., VINCKENBOSCH N., LONG M., 2009  RNA-based gene duplication:

KAPRANOV P., WILLINGHAM A. T., GINGERAS T. R., 2007  Genome-wide transcription and

KNOWLES D. G., MCLYSAGHT A., 2009  Recent de novo origin of human protein-coding


MONSELLIER E., CHITI F., 2007  Prevention of amyloid-like aggregation as a driving force of protein evolution. EMBO reports 8: 737–742.


SCHAEFER C., SCHLESSINGER A., ROST B., 2010  Protein secondary structure appears to be robust under in silico evolution while protein disorder appears not to be. Bioinformatics 26: 625–631.


FIGURE LEGENDS

Figure 1. A schematic phylogenetic tree of fungal species, and conservation of yeast genes (after CARVUNIS et al. 2012, modified). The main bifurcation events on the Saccharomyces lineage are numbered, from 0 to 10, age estimates were obtained using TimeTree (HEDGES et al. 2006), except for S. castellii, which split from the Saccharomyces lineage 100-150 mya, after the whole genome duplication of yeasts (CLIFITEN et al. 2006). Yeast genes were classified according to their conservation level, which corresponds to the phylogenetic spread of their orthologs; for example a yeast gene with conservation level 5 means that it has orthologs in S. castellii but not in the fungal species that split earlier from the Saccharomyces lineage, while conservation level 7 means that the yeast gene has orthologs either in D. hansenii or C. albicans. Conservation level 0 marks the putative ORFs identified by (CARVUNIS et al. 2012), excluding sequences shorter than 50 aa), while conservation level 1 indicates genes annotated by the Saccharomyces genome consortium, having no orthologs in any other species.

Figure 2. Integration of new genes into regulatory networks. Proto-genes acquire regulatory interactions rapidly, already genes with conservation level 1 are co-regulated with thousands of genes (A), are regulated by several transcription factors (B), and also rapidly gain regulatory motifs (feed-forward loops, C). The difference between proto-genes and conserved genes largely disappears by conservation level 4, representing ca. 14 million years old genes (p > 0.05 for conservation levels 5,6,8,10, ANOVA, Bonferroni post-hoc tests).

Figure 3. The percent of essential genes among genes with different conservation level; proportions are indicated above the bars. The youngest essential genes which do not overlap with older genes (YEL035C, YPL124W) appear in conservation level 4. (Note that among
proto-genes that overlap with conserved genes essentiality is present already in conservation level 1, however in these cases their fitness effect is not independent from the overlapping conserved gene).

Figure 4. Changes in secondary structure and aggregation propensity with gene age. While the amount of alpha helices does not depend on protein age (A), the amount of beta strands declines significantly between conservation levels 4-6. (B). Aggregation propensity, which is partly caused by the presence of beta strands shows an even stronger trend than beta strands, with random amino acid sequences and proto genes being much more prone to aggregation than conserved genes (C).

Figure 5. The probability of gene loss in *S. paradoxus* or *S. mikatae*. Only proteins that emerged before the *S. cerevisiae* - *S. mikatae* split were examined. Genes with conservation levels 4-5 are lost at significantly higher frequencies than more conserved genes (p < 0.05 for all comparisons between conservation levels 4-5 vs. 6-10, Chi-square tests).

Figure 6. An overview of the analysis of protein structural robustness, on the example of yeast ORF YDR103W. A) The tertiary structure of the protein (PDB id: 4F2H). Alpha helices are highlighted with blue, beta strands with yellow. B) The sequence of the protein was gradually mutated in 70 steps, in each step 1% of the residues were changed, and in each step the secondary structure was determined. The change in the location of helices and sheets that occurs with the mutagenesis is indicted with the respective colors. As sequence similarity to the original sequence declines, less and less residues are part of the same secondary structure as in the original protein, particularly in beta strands. C) For every protein the mutagenesis was repeated five times, and the Q3 value – the percent of residues with the same secondary
structure as in the original structure – was calculated for each step. Every line represents one mutagenesis-path (replicate); in the analyses the average of the five replicates was used.

Figure 7. Structural robustness of proteins. A) The robustness of secondary structures for mutations depends on their conservation level. Proto-genes and ancient genes show a highly significant difference (ANCOVA, p << 0.001 for comparisons between proto genes and conserved genes, Bonferroni post-hoc tests, whiskers represent 95% confidence intervals); the secondary structure of ancient genes is less sensitive for mutations (i.e. the Q3 value is higher). B) Beta strands decay faster from random mutations than alpha helices (p<< 0.001, ANCOVA). C) The amount of beta strands in proteins correlates negatively with the structural stability of the protein. Q3 values were calculated at 50% sequence similarity with the original sequence. D) The structural stability of proteins, excluding the regions with beta-strands. The differences between proto-genes and ancient genes are still highly significant (p << 0.001, ANCOVA), indicating that it is not merely a by-product of compositional differences between ancient and proto-genes (see Figure 4).

Figure 8. The dependence of protein-protein and genetic interactions from gene age. Note that the y axis is logarithmic, and that only those genes were included that have interactions, to correct for research biases. A) The number of protein-protein interactions increases continuously with conservation level; new protein-protein interactions emerge at a rate 2-2.25 \times 10^{-8} per year. B) The number genetic interactions increase at a rate of \sim 4.8 \times 10^{-8}, however the rate of change slows down above conservation level 6.

Figure 9. Genetic interactions between proto genes show weaker epistasis than between conserved genes (p << 0.001, Mann-Whitney U test). For each proto- and conserved gene the
mean of their absolute genetic interaction scores (|\epsilon|) were calculated, thus the histograms represent both the positive and negative epistatic interactions.
Conservation levels

<table>
<thead>
<tr>
<th>Conservation level</th>
<th>Fraction of S. cerevisiae or S. mikatae genes lost in S. paradoxus or S. mikatae (conservation levels 2-3)</th>
</tr>
</thead>
<tbody>
<tr>
<td>4</td>
<td>17/43</td>
</tr>
<tr>
<td>5</td>
<td>5/12</td>
</tr>
<tr>
<td>6</td>
<td>33/227</td>
</tr>
<tr>
<td>7</td>
<td>21/136</td>
</tr>
<tr>
<td>8</td>
<td>13/123</td>
</tr>
<tr>
<td>9</td>
<td>36/408</td>
</tr>
<tr>
<td>10</td>
<td>208/2618</td>
</tr>
</tbody>
</table>
A. Effect of conservation on robustness (Q3, %) against conservation level:

- Random: 69
- Conservation level 0: 71
- Conservation level 1-4: 71
- Conservation level 5: 72
- Conservation level 6: 73
- Conservation level 7-10: 73

B. % of residues with unaltered structure against pairwise sequence identity:

- Conservation level: 10
- Helices:
  - Random: 90
  - Pairwise sequence identity 100: 80
- Sheets:
  - Random: 90
  - Pairwise sequence identity 100: 80

C. Secondary structure similarity (Q3) at 50% sequence similarity:

- Correlation: $r = -0.544$
- Significance: $p = 0.0000$
- Equation: $y = 66.181 - 0.308x$

D. Effect of conservation on robustness (Q3, %) excluding beta strands against conservation level:

- Random: 71
- Conservation level 0: 72
- Conservation level 1-4: 73
- Conservation level 5: 74
- Conservation level 6: 75
- Conservation level 7-10: 76
Proto-genes
Conserved genes

mean absolute epistasis (\( \epsilon \))

Nr of conserved genes

Nr of proto-genes