Final report on postdoctoral grant PD83571

“Structural, evolutionary and systems level analysis of the proteome of transposable elements”

by György Abrusán

Summary of the results

Published papers:

- Abrusán G. (2013) Integration of new genes into cellular networks, and their structural maturation. Genetics, 195: 1407-1417.

- Abrusán G., Zhang Y., Szilágyi A. (2013) Structure prediction and analysis of DNA transposon and LINE retrotransposon proteins. Journal of Biological Chemistry, 288: 16127-16l38.

- Abrusán G., Szilágyi A., Zhang Y., Papp B. (2013) Turning gold into 'junk': transposable elements utilize central proteins of cellular networks. Nucleic Acids Research 41:3190-3200.

- Abrusán G. (2012) Somatic transposition in the brain has the potential to influence the biosynthesis of metabolites in Parkinson's disease and schizophrenia. Biology Direct 7:41.

Submitted manuscripts:

- Abrusán G., Yant S., Mátés L., Izsvák Zs., Ivics Z. (2015) Structural determinants of transposase activity. Submitted.

- Abrusán G. (2015) New genes are lost rapidly in Drosophila. Submitted.

International acclaim:

- My 2013 paper on the evolution of new genes (“Integration of new genes into cellular networks, and their structural maturation, Genetics, 195:l407-1417”) was one of the four cover stories of the 2013 December issue of Genetics.

- The journal also published an educational primer on the above mentioned paper in the 2014 March issue of Genetics, to explain the evolutionary process of gene birth to a broader (student) audience. (Frietze S., Leatherman J. (2014) Examining the process of de novo gene birth: an educational primer on “Integration of new genes into cellular networks, and their structural maturation” Genetics 196: 593-599.)

Practical use of the results:

Although the research did not result in commercial products or gains, it did result in practically valuable results. The "Sleeping Beauty" transposon is one of the most commonly used tools in genome engineering, which is used for insertional mutagenesis, somatic gene transfer and other applications. Wild type transposases are not optimal for practical use, because they evolved to transpose at relatively low frequencies, as high transposition rates harm their hosts. In consequence, modifying their activity, or insertion patterns is of considerable practical importance. Using the predicted structure of Sleeping Beauty, a large mutant dataset, and simulations of protein folding energies, we provide qualitative and quantitative guidelines for rational planning of Sleeping Beauty mutants. (See the "Structural determinants of transposase activity" manuscript.)

Detailed description of results

I. Structural analysis and folding of transposable element (TE) proteins.

Despite the considerable research on transposable elements, no large-scale structural analyses of the TE proteome have been performed so far. We predicted the structures of hundreds of proteins from a representative set of DNA and LINE transposable elements, and used the obtained structural data to provide the first general structural characterization of TE proteins, and to estimate the frequency of TE domestication and horizontal transfer events.

A) *High amounts of disordered sequence in ORF1 proteins of retrotransposons.* As the first step of the structural analysis, we identified the regions of TEs that lack structure: the intrinsically disordered parts of the sequences. Intrinsically disordered proteins are proteins, or regions of proteins, which, in their native state, have no stable structure, except in the presence of their substrate or in complex. The existence of short, flexible regions of proteins that link rigid globular domains of proteins have been known for decades. In the last decade, however, it has been discovered that in some proteins a large fraction of the sequence, or even the entire sequence has no well defined tertiary structure in their native, functional form (1,2). We tested for the presence of disordered regions in more than 5000 TE protein sequences. While the overall level of protein disorder is comparable to other eukaryotic proteins, in retrotransposons, their ORF1 and 'gag' proteins contain strikingly high amounts of coiled coil and disordered sequence (Figure 1) (3). The few experimental systems where retrotransposon gag proteins have been analyzed (human L1 retrotransposon, mouse Ll retrotransposon) show that these proteins are necessary for transposition, and have chaperone function. Since disordered regions are characteristic for chaperones, our results indicate that, despite the fact that gag proteins show essentially no sequence conservation across different retrotransposon families, the chaperone function of these proteins is likely to be conserved.

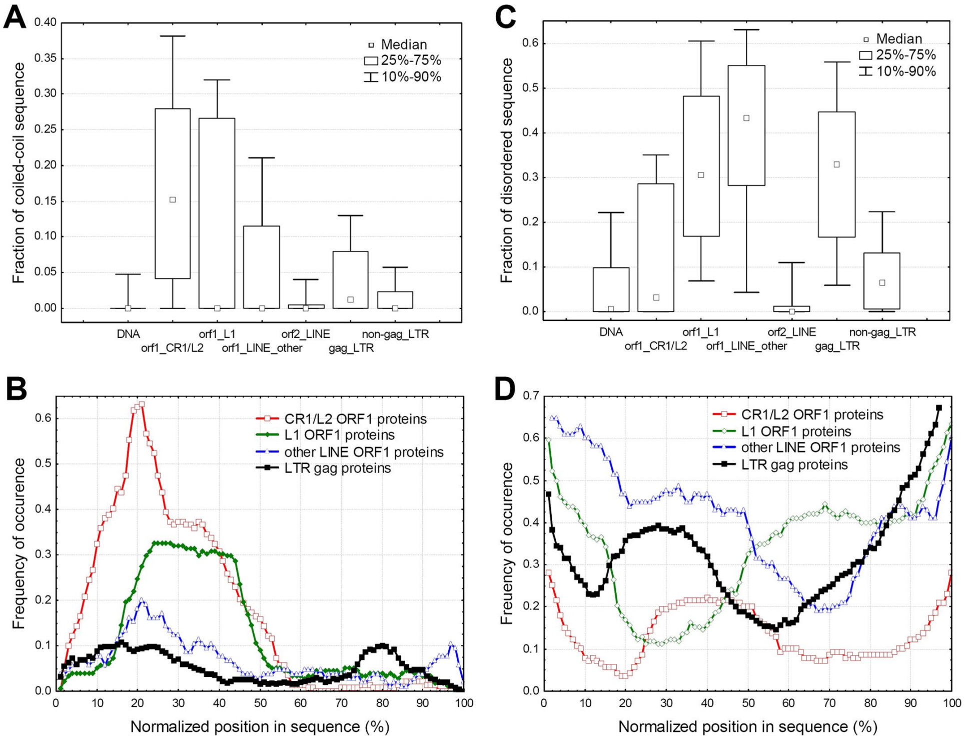


Figure 1. Coiled-coil and intrinsically disordered regions in TE proteins. **A)** The fraction of coiled-coil sequence (predicted with Marcoil) in different TEs. ORF1 proteins of the CR1 and L1 families contain much higher amounts of coiled coil sequence than other TE proteins. **B)** Coiled-coil regions in the ORF1/gag proteins are characteristically located near the N-terminus of the sequence. **C)** The fraction of disordered sequence (predicted with IUpred) in different TE protein types. ORFI/gag proteins are characterized by approx 5-fold higher amounts of disordered sequence than other TE proteins. **D)** The distribution of disordered regions along the sequence of ORF1/gag proteins of retrotransposons.

B) *Folding and analysis of TE proteins.* We folded 870 protein domains from a representative set of DNA and non-LTR transposable elements using I-TASSER (4), the protein folding tool that performed best on the last three CASP competitions. Overall, the folding took 100 CPU years (1year of running time on 100 processors). Since protein structure is much more conserved than sequence, we used structure-structure comparisons to gain insight on the evolution of domain composition of TEs, and to estimate the contribution of TE proteins to non-TE (host) proteins. The predicted TE structures were thus compared to SCOP protein domain database (5) and to the structures generated by the Proteome Folding Project (6), which is a large scale effort to fold all proteins from more than 94 complete proteomes and contains structures of more than 80 000 domains (16 000 of them high quality) that have no detectable similarity to SCOP. The comparison of TE domains with SCOP domains yielded an interesting insight about the age of TEs. Different protein structures were invented at different times during evolution, for example DNA/RNA polymerases are among the most ancient existing folds, whereas immunoglobulins are relatively young, and appeared after the emergence of the vertebrate immune system. In consequence, the SCOP fold composition of a protein contains also information on its age. In recent years a number of studies estimated the time of appearance of known protein folds using methods which rely on the reconstruction of a global phylogenetic tree of folds, based on their abundance across different genomes (7). The analysis of the age of SCOP folds across the various families of DNA transposons indicates that their most abundant folds are among the most ancient known folds, which existed already before the appearance of the first cellular microfossils (3.4 Bya), and thus were probably present in the oldest cellular organisms. Surprisingly, although reverse-transcriptases (and the process of retrotransposition) were suggested by many authors to be among the most ancient proteins, which may have their origin in the RNA world (8) the analysis of their structures indicates that although the most common protein folds of non-LTR retrotransposons are indeed very ancient, they appeared approximately at the time of the specification of the three superkingdoms (Archaea, Bacteria, Eukaryota), after the transition from RNA to DNA based replication. The dominance of the most ancient ~4 By old folds in DNA transposons leads to the controversial idea that DNA transposons are as old as the oldest proteins, and, because DNA transposons need DNA-based host for their replication, DNA was established as the carrier of genetic information early on in the evolution of life, implying that the RNA/RNP world did not last for billions of years. This finding suggests that genomic parasites are as ancient as proteins themselves, and from the two basic mechanisms of transposition (cut-and-paste vs. retrotransposition), the cut-and-paste mechanism is the older (Figure 2.) (3)

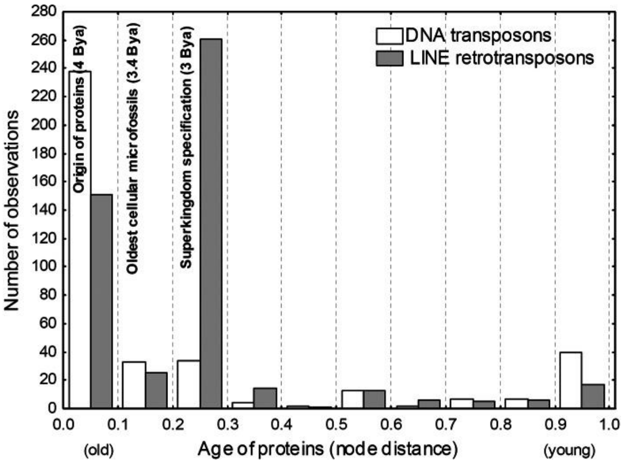


Figure 2. The proteins of DNA transposons contain more ancient SCOP folds than LINE retrotransposons. The age of protein folds is measured as node distance, a measure based on the phylogenetic spread of the fold. The larger the node distance the younger the particular fold is (see 9 for details). The histogram shows that in DNA transposons the most abundant folds are among the most ancient ones which were probably present before the first cellular organisms, whereas the most abundant folds in LINEs were invented later, approximately at the time of the specification of the three superkingdoms.

The comparison of the predicted TE structures to the 16 000 high quality structures of the Proteome Folding Project resulted in 295 significant hits, whereas only 46 hits were found when the same dataset was searched with a sequence based Hidden Markov method (HMMER), indicating that the total amount of sequence exchange events between TEs and their hosts is considerably higher than what can be seen by sequence comparisons alone. However their frequency is still low, as less than 2% of the PFP proteins show a significant hit to a TE structure (3). The taxonomic distribution of the hits show that proteins of parasitic protists and plants contain significantly higher amount of TE derived sequences than other taxa. Interestingly, when the number of TE domestication events (a TE protein is incorporated into a protein of the host organism) were compared with the TE incorporation events (a TE picked up a host protein) the latter turned out to be significantly more frequent, although protein incorporation and capturing is much less studied than TE domestication. Overall, the analysis demonstrated the significantly higher sensitivity of the structure-structure comparisons, however it also demonstrated that structure based and sequence based methods have different strengths and weaknesses, and they complement each other rather than substitute.

II . Structural determinants of transposase activity.

Recent studies indicate, that several key properties of proteins are determined by structural elements which do not follow the classic hierarchical organization into secondary and tertiary structural elements. These structures, named "sectors" form physically connected networks of coevolving residues within protein domains (10), and span across secondary structural units. Several important biological properties are determined by sectors; although they typically make up only 10-30% of the residues they were shown to significantly contribute to the specification of protein folds, allosteric communication in proteins, and evolution of novel functions. Using extensive mutagenesis data, simulations of folding energies, and protein modeling we tested whether sectors and other structural properties influence the transposition rate of the Sleeping Beauty transposase, an important tool in genome engineering. The Sleeping Beauty transposon was the first transposase developed for gene transfer, and was reconstructed from extinct TC1/mariner transposons in fish (11). Together with its hyperactive variant, it is still one of the most widely used transposon vector, which is applied in insertional mutagenesis, somatic gene transfer or cellular reprogramming (12). It is the only transposon vector being in in-human clinical trial for the treatment of patients with B-lymphoid malignancies (13,14).

We found that the Sleeping Beauty transposase contains three sectors, which, unlike the ones examined so far span across several conserved domains, indicating that the physically distinct domains of the protein (DNA binding domains and the DDE endonuclease) coevolve. Second, besides the hydrophobic core, sector residues are also significantly more sensitive for mutations than other residues, and point mutations of these residues usually strongly reduce transposition rate (Figure 3). However, the influence of sectors on protein function appears to depend on the tertiary structure – sector residues have significant effect on transposition rate in the globular, endonuclease domain of the protein, but not in the DNA binding and flexible helix-turn-helix domains.

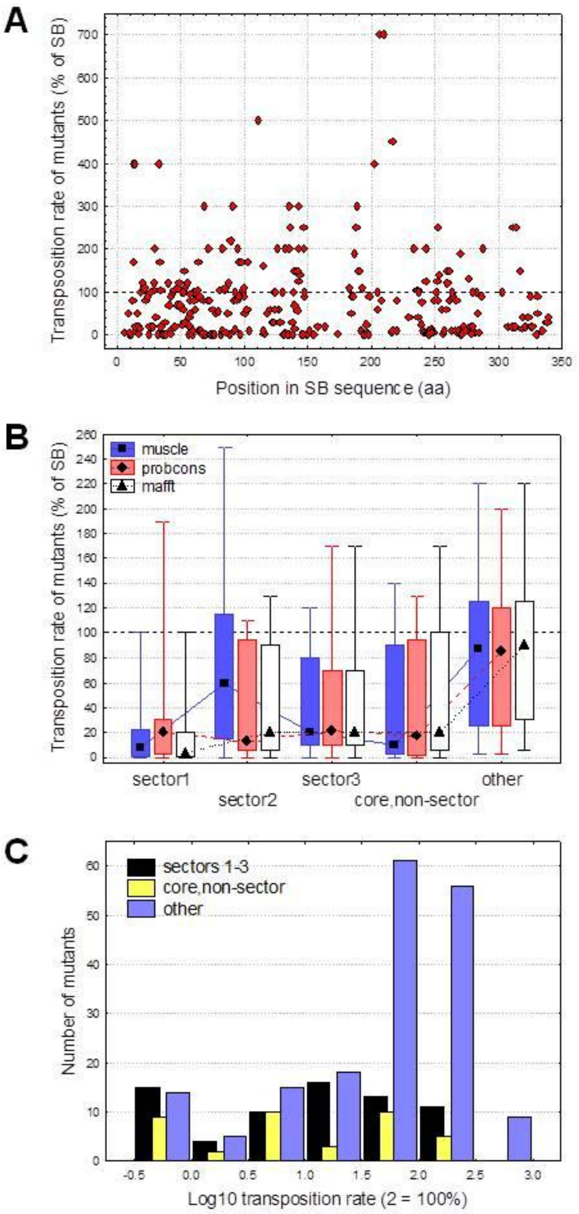


Figure 3. Effect of residue location on the transposition rate of Sleeeping Beauty mutants. A) The location of 286 point mutants along the transposon sequence, and their effect on the transposition rate. The mutations are distributed approximately evenly across the sequence; the vast majority of mutants reduces transposition rate. B) The effect of sectors and core residues on transposition rate (median; box: 25-75%; whiskers: 10-90%). Sector residues are identified from multiple alignments, using the covariance between different locations. To correct for any biases introduced by different multiple alignment methods, we calculated sector residues from three different alignments of 250 transposon sequences, each made with a different tool: muscle, mafft and probcons. C) Frequency distributions of $\log_{10}$ transformed transposition rates. Residues were classified as 'sector', 'core' and 'other' categories using the muscle alignment; the three sectors were grouped for clarity.

Most proteins can function only in a narrow range of folding energies (15), as unstable proteins may not fold properly, whereas very stable ones may be to rigid to perform their functions. Mutating a residue in a protein can have significant effect on its overall stability and function, thus we tested whether the differences in transposition rate between sector, core and other residues are caused by their different effects on protein stability, measured as the difference in folding energies between the wild type and mutant. Simulations of the effect of mutations on folding energies show that, mutations with a strong effect on van der Waals forces in the globular DDE domain significantly reduce transposition rates. Taken together, our findings highlight the importance of sectors and folding energies in determining transposition rate and can be used for a rational planning of mutants in Sleeping Beauty (and other transposon) mutagenesis studies.

III. Origin of novel protein domains in TEs.

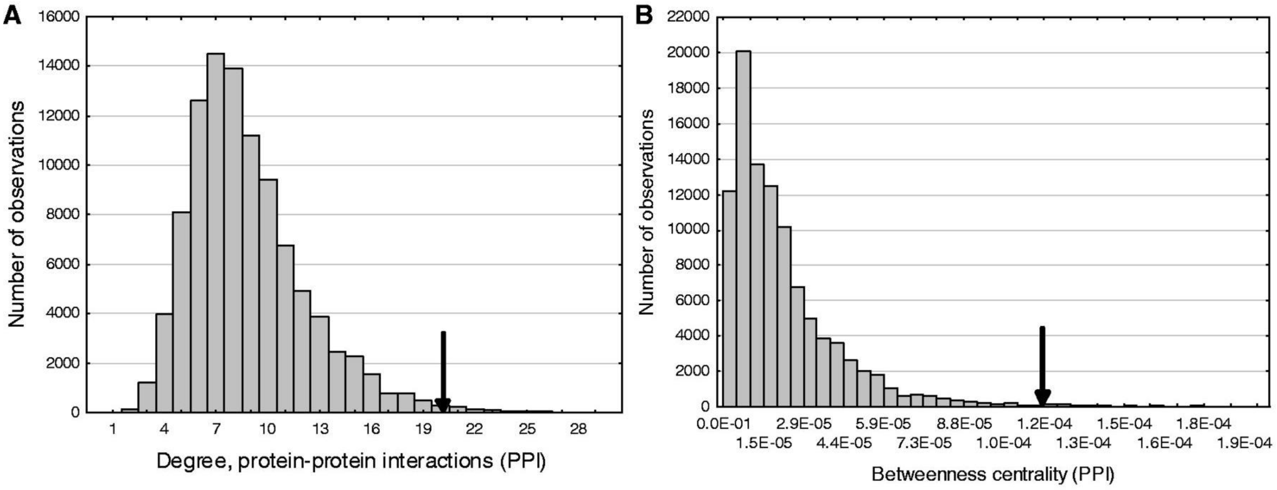
In recent years the traditional view that transposable elements are only a burden to their host organism has shifted. Although their parasitic nature is not questioned, the discoveries that many host proteins originate from TEs, and that TEs contributed to the invention of several key cellular machineries of multicellular organisms highlighted their significance in evolutionary innovations (16-18). The best-known cases of TE domestications include the RAG protein of the immune system of vertebrates, CENP-B protein of centromeres, light sensing proteins in plants, regulation of telomere length, or the PAX6 gene. As the global sequencing effort completed the genomes of most classic model organisms, attention has turned towards several less well researched eukaryotic genomes, either owing to their phylogenetic importance, or due to being important to a narrower research community. Besides providing key insights into genome organization and function, these studies have also revealed a large diversity of TEs that previously had not been appreciated: repeat classes that had been thought to be extinct were found (e.g. DNA transposons in bats), entire novel classes of TEs were discovered (e.g. Polintons), and the diversity of known families was greatly expanded.

Unlike the domestication of transposon proteins, much less in known about the reverse process, i.e. whether, and to what degree the evolution of TEs is influenced by the genome of their hosts. We addressed this issue by searching for cases of incorporation of host genes into the sequence of TEs, and examined the systems level properties of these genes: whether the incorporated sequences originate from a random selection of host proteins, or TEs selectively incorporate genes with distinct properties within the cellular networks. Currently cellular networks are well characterized only for a small number of model organisms, therefore we used the budding yeast (*Saccharomyces cerevisiae*) and the fruitfly (*Drosophila melanogaster*) interactomes in the analysis. We identified 51 cases where the evolutionary scenario was the incorporation of a host gene fragment into a TE consensus sequence, and we show that both the yeast and fly homologues of the incorporated protein sequences participate in significantly more protein-protein interactions than expected by chance, and are significantly more central in the cellular networks than randomly selected proteins. (Figure 4).

An analysis of selective pressure (Ka/Ks ratio) detected significant selection in 37% of the cases, indicating that at least some of the cases of protein incorporation are beneficial for the repeat. Recent research on retrovirus-host interactions shows that virus proteins preferentially target the hubs of the host interaction networks, enabling them to take over the host cell using only a few proteins. As autonomous TEs encode only few proteins, and viruses and TEs interact with overlapping sets of proteins (19) we propose that TEs face similar evolutionary pressure to evolve proteins with high interacting capacities and take some of the necessary protein domains directly from their hosts (20).

IV. Possible effects of somatic transposition on the metabolism of human brain.

Transposable elements make up at least half of the human genome, however the vast majority of the approximately 3.5 million TE insertions are fixed, ancient repeats. The overall impact of this enormous number of insertions on the evolution of the genome is still under intensive research; it has been shown that 25% of mammalian promoters contain a transposable element (2l), a number of genes and protein domains were derived from transposable elements, the presence of TEs can result in alternative splicing (22) and structural variation of the genome (23). Despite the very high number of fixed insertions, the number of transpositionally active TEs in the human genome is surprisingly small. The only active autonomous transposon is the L1 retrotransposon, with approximately 80-100 active copies contributing the majority of the insertions in the genome (24).



C

	Saccharomyces cerevisiae					*Drosophila melanogaster*				
	nr of genes in dataset	mean of random samples	SD	sample	P	nr of genes in dataset	mean of random samples	SD	sample	P
fitness	20	0.977	0.061	0.972	0.748	-	-	-	-	-
degree, PPI	23	13.891	4.625	30	**0.005**	35	8.796	3.41	20	**0.001**
degree, GI	23	27.055	9.65	23	0.638	14	3.321	1.526	6.5	**0.046**
betweenness, PPI	23	4.75E-05	2.74E-05	1.11E-04	**0.032**	35	2.28E-05	2.36E-05	1.26E-04	**0.008**
betweenness, GI	23	4.15E-05	3.68E-05	4.57E-05	0.306	14	1.41E-04	2.41E-04	1.81E-04	0.234
closeness, PPI	23	0.382	0.012	0.393	0.145	35	0.325	0.011	0.348	**0.021**
closeness, GI	23	0.366	0.01	0.369	0.361	14	0.249	0.012	0.276	**0.011**
direct PPI (nodes)	23	2.199	2.278	9	**0.016**	35	4.471	3.135	13	**0.012**
direct PPI (edges)	23	1.323	1.687	6	**0.025**	35	2.964	2.495	14	**0.002**

PPI - protein-protein interactions
GI - genetic interactions

Figure 4. Characteristics of genes that were incorporated into TEs. We performed Monte Carlo simulations to test whether fitness and network characteristics like degree, betweenness centrality and closeness centrality are significantly different in the incorporated genes than in the random expectation. **A)** Distribution of the median degree (protein-protein interactions) in 100 000 random samples each containing 35 proteins from the Drosophila proteome; the arrow represents the median of the 35 Drosophila genes that have a homologue in a transposon and for which PPI information was available. **B)** Distribution of the median betweenness centrality in 100 000 random samples of 35 Drosophila genes, and the ones with a homologue in a transposon. **C)** Statistical summary of Monte Carlo simulations.

Although the millions of fixed insertions in our genome have neutral or nearly neutral effect on fitness, the recent polymorphic insertions of active TEs are likely to be harmful, and were shown to be responsible for a number of diseases including hemophilia, leukemia, colon cancer or breast cancer (25). In order to reduce their negative effects on the host, TEs were assumed to "jump" only in the germline, as somatic insertions are not inherited, and may seriously harm their host. This view was challenged by recent findings, which show that the rate of somatic transposition is higher than expected. Particularly high levels of transposition were detected in the human brain, including the hippocampus, middle temporal gyrus or caudate nucleus (26). Currently it is unknown how the observed pervasive transposition in the brain influences its functioning, whether it is responsible for any pathological processes, and why neural tissue is so permissive for retrotransposition. I computationally investigated what effect somatic transposition can have on the metabolism of the brain and the biosynthesis of its key metabolites like neurotransmitters. I used flux balance analysis to investigate which metabolic pathways can be disrupted by TE insertions, and to estimate whether the elimination of these metabolic pathways can have significant influence on metabolism. The analysis shows that somatic transposition in the brain can influence the biosynthesis of more then 250 metabolites, including dopamine, serotonin and glutamate, and shows large variation between individuals. Additionally, the metabolic "fingerprints" that TE insertions have on metabolism indicate that their activity may contribute to the development of Parkinson's disease and schizophrenia. Overall, the analysis shows the ability of TEs to interfere with the biosynthesis of several metabolites, particularly ones linked with tyrosine metabolism, and it suggests their contribution to diseases of the central nervous system (27).

V. Integration of novel genes into cellular networks.

Besides research on transposable elements, in the last year of the project I also worked on a topic that was unrelated to the original goals of the grant; however, I considered it so novel and interesting that I decided to pursue it nevertheless.

The established view is that new genes evolve primarily by duplications and recombination, i.e. by recombination of existing domains of other genes (28). Recent studies however highlighted that genes can also emerge from noncoding DNA: *de novo* emergence of genes has been demonstrated in a number of species like *Drosophila* (29), humans (30), yeasts (31) and viruses (32). Although traditionally de novo emergence of genes was assumed to be extremely rare, the fact that they have been detected in several model organisms indicates that this phenomenon is not exceptionally rare and may be an ongoing process in many, if not most genomes. Although de novo origination of new 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, and integrated into regulatory networks, and also whether structural changes affect the novel proteins. 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 (31), and suggested that genes can be placed in a continuum from nongenic sequences to conserved genes, and de novo emergence of genes may be as common as emergence by the classic duplication-divergence mechanism.

Using yeast genes I tested whether the integration of new genes into cellular networks and their structural changes show such a continuum, by analyzing their changes with gene age. I show that 1) the number of regulatory, protein-protein and genetic interactions increases continuously with gene age, although at 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 2-2.25e-08/year and 4.8e-08/year, respectively.
2) gene essentiality evolves relatively quickly: the youngest essential genes appear in proto genes approx. 14 my old.

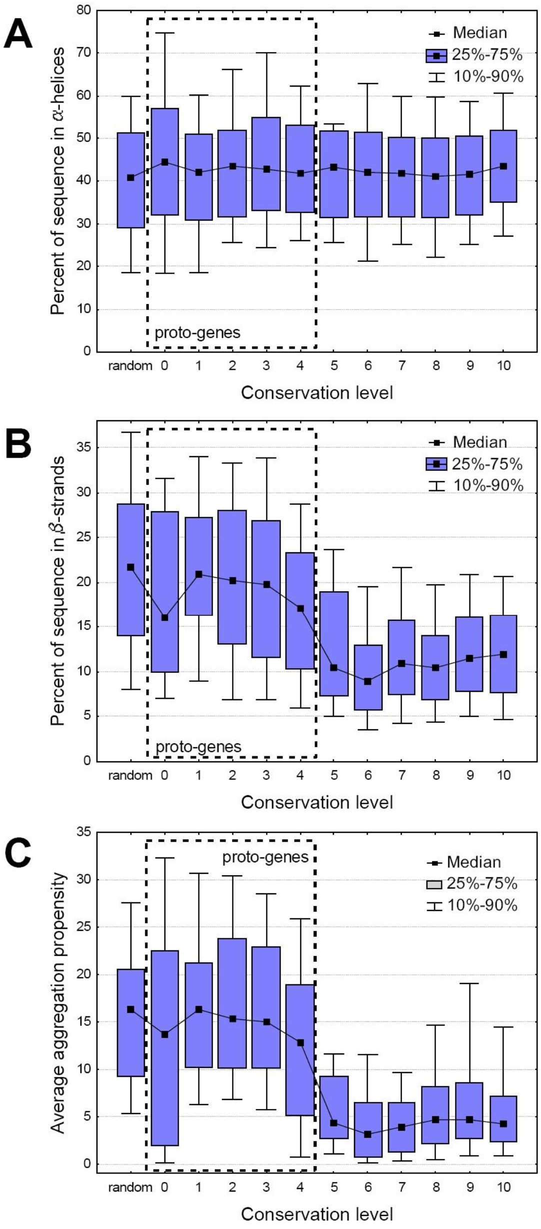


Figure 5. Changes in secondary structure and aggregation propensity with gene age (0: proto genes, present only in *Saccharomyces cerevisiae*; 10: genes present in the common ancestor of Saccharomyces cerevisiae and *Schizosaccharomyces pombe* [older than 760 My]). While the amount of alpha helices does not depend on protein age **(A)**, the amount of beta strands declines significantly between conservation levels 4 and 6 **(B)**. Aggregation propensity, which is partly caused by the presence of cross 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)**.

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 by high beta strand content, high aggregation propensity, and low robustness against mutations, while conserved genes are characterized by lower strand content and higher stability, most likely due to higher probability of gene loss among young genes and accumulation of neutral mutations. (Figure 5.)

Overall the results show a somewhat contradictory picture of the evolution of new genes in yeast: their integration into cellular networks shows a continuum, although the rates of regulatory evolution and the gain of protein and genetic interactions are very different; while from the structural point of view proto-genes and conserved genes form two distinct groups, with different beta strand content, aggregation propensity and robustness for mutations. This, 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 (33).

References

1. Dyson H.J, Wright, P.E. (2005) Intrinsically unstructured proteins and their functions. Nat. Rev. Mol. Cell. Biol. 6:197-208.

2. Tompa P. (2012) Intrinsically disordered proteins: a 10 year recap. Trends Biochem. Sci. 37: 5O9-516.

3. Abrusán G., Zhang Y., Sziláqyi A. (2013) Structure prediction and analysis of DNA transposon and LINE retrotransposon proteins. Journal of Biological Chemistry, 288: 16127 -16138.

4. Roy A., Kucukural A., Zhang Y. (2010) I-TASSER: a unified platform for automated protein structure and function prediction. Nat. Protoc. 5:725-738.

5. Andreeva A., Howorth D., Chandonia J. M., Brenner S.E., Hubbard T.J., Chothia C., Murzin A.G. (2008) Data growth and its impact on the SCOP database. Nucleic Acids Res. 36: D419-D425.

6. Drew K., Winters P., Butterfoss G.L., Berstis V., Uplinger K., et al. (2011) The Proteome Folding Project: Proteome scale prediction of structure and function. Genome Res. 21: 1981-1994.

7. Caetano-Anolles G., Wang M., Caetano-Anolles D., Mittenhal J.E., (2009) The origin, evolution and structure of the protein world. Biochem. J . 417: 621-637.

8. Brosius J., (2005) Echoes from the past - are we still in an RNP world? Cytogenet. Genome Res. 11O: 8-24.

9. Wang M., Jiang Y.Y., Kim K.M., Qu G., Ji H.F., Mittenhal J.E., Zhang H.Y., Caetano-Anolles G. (2011) A universal molecular clock of protein folds and its power in tracing the early history of aerobic metabolism and planet oxygenation. Mol. Bio1. Evol 28: 567-582.

10. Halabi N., Rivoire O., Leibler S., Ranganathan R. (2009) Protein sectors: evolutionary units of three-dimensional structure. Cell 138: 774-786 .

11. Ivics Z., Hackett P.B., Plasterk R., Izsvák Zs. (1997) Molecular reconstruction of Sleeping Beauty, a Tc1-like transposon from fish, and its transposition in human cells. Ce1l 91: 501-510.

12. Ivics Z., Izsvák Zs. (2011) Nonviral gene delivery with the Sleeping Beauty transposon system. Hum. Gene Ther. 22:1043-1051.

13. Hackett P.B., Largaespeda D.A., Switzer K.C., Cooper L.J. (2013) Evaluating risks of insertional mutagenesis by DNA transposons in gene therapy. Transl. Res. 161: 265-283.

14. Williams D.A. (2008) Sleeping Beauty vector system moves toward human trials in the United States. Mol. Ther. 16: 1515-1516.

15. DePristo M.A., Weinreich D.M., Hartl D.L. (2005) Missense meanderings in sequence space: a biophysical view of protein evolution. Nat. Rev. Genet. 6: 678-687.

16. Volff J. (2006) Turning junk into gold: domestication of transposable elements and the creation of new genes in eukaryotes. Bioessays 28: 913-922.

17. Jurka J., Kapitonov V.V., Kohany O., Jurka M.V. (2007) Repetitive sequences in complex genomes: structure and evolution. Annu. Rev. Genomics Hum. Genet. 8:241-259.

18. Feshotte C., Pritham E.J. (2007) DNA transposons and the evolution of eukaryotic genomes. Annu. Rev. Genet. 41: 331-368.

19. Irwin B., Aye M., Baldi P., Beliakova-Bethell N., Cheng H., Dou Y., Liou W., Sandmeyer S. (2005) Retroviruses and yeast retrotransposons use overlapping sets of host genes. Genome Res. 15: 641-654.

20. Abrusán G., Szilágyi A., Zhang Y., Papp B. (2013) Turning gold into 'junk': transposable elements utilize central proteins of cellular networks. Nucleic Acids Res. 41:3190-3200.

21. Jordan I.K., Rogozin I.B., Glazko G.V., Koonin E.V. (2003) Origin of a substantial fraction of regulatory sequences from transposable elements. Trends Genet. 19:68-72.

22. Sorek R., Ast G., Graur D. (2005) Alu containing exons are alternatively spliced. Genome Res. 12: 1060-1067.

23. Xing J., Zhang Y., Han K., Salem A.K., Sen S.K., Huff C.D., Zhou Q., Kirkness E.F., Levy S., Batzer M.A., Jorde L.B. (2009) Mobile elements create structural variation: analysis of a complete human genome. Genome Res. 19:1516-1526.

24. Brouha B., Schustak J., Badge R.M., Lutz-Prigge S., Farley A.H., Moran J.V., Kazatian H.H. (2003) Hot L1s account for the bulk of retrotransposition in the human population. Proc. Natl. Acad. Sci. USA 100: 5280-5285.

25. Hancks D.C., Kazazian H.H. (2012) Active human retrotransposons: variation and disease. Curr. Opin Genet. Dev.

26. Bailie J.K., Barnett M.W., Upton K.R., Gerhardt D.J., Richmond T.A., et al. (2011) Somatic retrotransposition alters the genetic landscape of the human brain. Nature 479: 534-537.

27. Abrusán G. (2012) Somatic transposition in the brain has the potential to influence the biosynthesis of metabolites in Parkinson's disease and schizophrenia. Biology Direct 7:4l.

28. Kaessmann H. (2010) Origins, evolution and phenotypic impact of new genes. Genome Res. 20: 1313-1326.

29. Begun D.J., Lindfors H.A., Kern A.D., Jones C.D. (2007) Evidence for de novo evolution of testis expressed genes in the *Drosophila yakuba/Drosophila erecta* clade. Genetics 176: 1131-1137.

30. Knowles D.G., McLysaght A. (2009) Recent de novo origin of human protein coding genes. Genome Res. 19:1752-1759.

31. Carvunis A.R., Rolland T., Wapinski I., Calderwood M.A., Yildrim M.A., et al. (2012) Proto-genes and de novo gene birth. Nature 487: 370-374.

32. Sabath N., Wagner A., Karlin D. (2012) Evolution of viral proteins originated de novo by overprinting. Mol. Biol. Evol. 29: 3761-3780.

33. Abrusán G. (2013) Integration of new genes into cellular networks and their structural maturation. Genetics 195: 1407-1417.