Computational Biology:

Structure Prediction and Analysis of DNA Transposon and LINE Retrotransposon Proteins

György Abrusán, Yang Zhang and András Szilágyi

J. Biol. Chem. 2013, 288:16127-16138.
doi: 10.1074/jbc.M113.451500 originally published online March 25, 2013

THE JOURNAL OF BIOLOGICAL CHEMISTRY VOL. 288, NO. 22, pp. 16127–16138, May 31, 2013
© 2013 by The American Society for Biochemistry and Molecular Biology, Inc. Published in the U.S.A.

Structure Prediction and Analysis of DNA Transposon and LINE Retrotransposon Proteins*[S]

Received for publication, January 8, 2013, and in revised form, March 21, 2013 Published, JBC Papers in Press, March 25, 2013, DOI 10.1074/jbc.M113.451500

György Abrusán[‡1], **Yang Zhang**[§], **and András Szilágyi**[¶]

From the [‡]Synthetic and Systems Biology Unit, Institute of Biochemistry, Biological Research Centre of the Hungarian Academy of Sciences, 6701 Szeged, Hungary, the [§]Center for Computational Medicine and Bioinformatics, Department of Biological Chemistry, Medical School, University of Michigan, Ann Arbor, Michigan 48109-2218, and the [¶]Institute of Enzymology, Research Center for Natural Sciences of the Hungarian Academy of Sciences, 1113 Budapest, Hungary

Background: No large-scale analysis of transposable element (TE) protein structures exists.
Results: We predicted and analyzed hundreds of proteins from a representative set of DNA and LINE transposable elements.
Conclusion: We provide new insights on TE evolution, structure, and frequency of sequence exchange events between TEs and their hosts.
Significance: This is the first large-scale analysis of TE protein structures.

Despite the considerable amount of 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. We show that 1) ORF1 and Gag proteins of retrotransposons contain high amounts of structural disorder; thus, despite their very low conservation, the presence of disordered regions and probably their chaperone function is conserved. 2) The distribution of SCOP classes in DNA transposons and LINEs indicates that the proteins of DNA transposons are more ancient, containing folds that already existed when the first cellular organisms appeared. 3) DNA transposon proteins have lower contact order than randomly selected reference proteins, indicating rapid folding, most likely to avoid protein aggregation. 4) Structure-based searches for TE homologs indicate that the overall frequency of TE domestication events is low, whereas we found a relatively high number of cases where horizontal transfer, frequently involving parasites, is the most likely explanation for the observed homology.

In recent years, it has become clear that transposable elements (TEs)[2] are not only a burden to their host but are also an important source of evolutionary innovation in eukaryotic (and probably also prokaryotic) genomes, by providing novel regulatory sites (1), modification of protein expression levels (2), or domestication of TE sequences, the acquisition of a TE fragment by host proteins (3, 4). Although several well described cases have been reported for these evolutionary scenarios, the extent to which sequences of TE origin become useful for the host is unclear. Recent studies show that in the human genome, the total amount of functional, conserved sequence is much higher (5%) than the amount of coding sequence (1.5%), and recent results from the ENCODE project (5) suggest that the amount of non-conserved, nevertheless functional sequence may even reach 80%, and a considerable fraction of it originates from TEs (6). However, in the case of domesticated TE proteins, estimates range from a few dozen to thousands, depending on how conservative the analysis was: the human genome project found 47 proteins with significant similarity to TEs (7), Zdobnov *et al.* (8) found only 35 proteins of retroviral/retrotransposon origin in mammals, whereas a considerably more relaxed analysis (9) estimated that the human genome itself contains almost 2000 proteins that contain remains of TEs. The key difficulty in identifying the actual number is that a very large fraction of the TE-like protein hits are in the "twilight zone" and show only remote similarity to a TE sequence; thus, they are either false positives or highly diverged, ancient cases of TE domestication where sequence similarity has deteriorated to the degree that it is barely detectable. One alternative approach is to use protein structures to identify distant homologs: the structures of proteins are typically much more conserved than their sequences; homologous protein sequences that have lost almost all sequence similarity can show a very high degree of structural similarity (although separating homologs from structural analogs in such cases is challenging because of the level of divergence overlap in the two), and protein sequences above 35% identity are typically structurally similar (10).

This work has two goals: (i) first, since a general structural characterization of proteins encoded by TEs is lacking, the first aim is to predict and characterize a representative set of TE structures, to identify common patterns among them like the overrepresentation of certain folds, their age, or signatures of selection that appear at the structural level. Because the num-

* This work was supported by Hungarian Scientific Research Fund OTKA Grants PD83571 (to G. A.) and K105415 and NK77978 (to A. S.), by the National Science Foundation Career Award DBI 0746198 (to Y. Z.), and the National Institute of General Medical Sciences Grants GM083107 and GM084222 (to Y. Z.).

[S] This article contains supplemental Tables 1–7 and predicted structures.

[1] To whom correspondence should be addressed: Synthetic and Systems Biology Unit, Institute of Biochemistry, Biological Research Center of the Hungarian Academy of Sciences, Temesvári krt. 62. Szeged H-6701, Hungary. E-mail: abrusan@brc.hu.

[2] The abbreviations used are: TE, transposable element; PDB, Protein Data Bank; TM, template modeling; PFP, Proteome Folding Project; MCM, mammoth confidence metric; Bya, billion years ago.

ber of experimentally solved TE protein structures is still very low, we used *in silico* methods and predicted the structures of DNA transposon and LINE proteins with I-TASSER, the protein structure prediction software that performed best in the last three CASP (critical assessment of techniques for protein structure prediction) experiments (11, 12). (ii) Second, our aim was to estimate the contribution of TE proteins to non-TE proteins, and we searched a recently published large-scale database of predicted protein structures (13) for proteins with structural and sequence homology to transposable elements and estimated the most likely evolutionary process that resulted in the observed homologies.

MATERIALS AND METHODS

Selection of a Representative Set of TEs—We selected a representative set of TE proteins for the structure prediction as follows. First, we selected all DNA transposons and LINE retrotransposons from the RepBase Database (version 15.12) (14) that were annotated as currently active, or their divergence from the consensus sequence was less than 3% and were thus active recently. Because the structure of proteins with sequence similarity higher than 35% is approximately similar, we clustered the selected proteins using a 35% identity threshold with UCLUST (15) and folded only the longest element of each cluster. In LINE retrotransposons, the ORF2 proteins show a much higher similarity to each other than ORF1 proteins (16), thus restricted the clustering to ORF2 proteins. Altogether, this procedure resulted in a set of 222 DNA transposon proteins and 232 LINE proteins, representing all major families of these repeats, including widely used transposon tools such as the hyperactive Sleeping Beauty (17), and piggyBac (18) transposases.

Domain Identification—*In silico* structure prediction is still most efficient for relatively short proteins. In many cases, the amino acid sequences of TEs were too long for building a reliable structure (*i.e.* the ORF2 proteins of LINEs, which are typically 1000–1400 amino acids long); thus, we split every protein longer than 500 residues into smaller regions, with a maximum length of 400–500 residues, which typically still contain several domains. First, we identified conserved domains in the amino acid sequences using the conserved domain search tool of NCBI (19). Next, we predicted domain boundaries with a method based on the domain prediction tool FiefDom (20). Briefly, FiefDom generates a PSSM using a query sequence and a reference database (nr) and searches for domain boundaries using the distribution of hits from a structure database (SCOP (21) or PDB (22)). Combining the coordinates of conserved domains and the distribution of SCOP/PDB hits on the TE protein sequence, we identified domain boundaries at the regions with the lowest sequence coverage, and when these boundaries collided with conserved domains, we adjusted them manually (Fig. 1). We split the proteins at the identified domain boundaries, and the resulting sequences, altogether 870 (see supplemental Table 1), were subsequently submitted to I-TASSER (11, 12).

Structure Prediction of TE Proteins—Detailed descriptions of I-TASSER can be found in Refs. 11 and 12. Briefly, I-TASSER first identifies suitable templates in the PDB database through sequence similarity searches (threading); second, using Monte Carlo simulations, the identified templates and regions modeled with *ab initio* methods are assembled into a large number of full-length conformations; third, by clustering the conformations, cluster centroids are identified, and the final models are built by additional refinements of the cluster centroids. All predicted structures are available for download as supplemental material.

C (Confidence) and Template Modeling (TM) Score Calculations—The C score of the predicted structures is the combination of the quality of the threading alignments (*Z* score) and the density of the clusters, and is defined by the formula shown in Equation 1,

$$\text{C-score} = \ln[Z \times (N_{\text{cl}}/N_{\text{tot}}) \times (1/\text{RMSD})] \qquad \text{(Eq. 1)}$$

where *Z* is the normalized *Z* score, N_{cl} is the number of conformations in a given cluster, N_{tot} is the total number of conformations used in the clustering, and r.m.s.d. is the average root mean square deviation of the trajectories from the cluster centroid in a given cluster. The local *C* score of domains was calculated using the same formula; first, we calculated the local r.m.s.d. for every amino acid position of the sequence, as the r.m.s.d. between a particular position of the cluster centroid sequence and the corresponding position of every conformation in the cluster. Next, we identified regions where the local r.m.s.d. was consistently below 5 Å (see Fig. 2, in cases where the average local r.m.s.d. of a structure was below 2 Å, we used a 3.5 Å cut-off) and reran the clustering only for these regions.

Another measure of the quality of a predicted structure is its estimated TM score. TM score measures the similarity between two structures (23); the estimated TM score is calculated from *C* score and estimates the similarity of a predicted structure to its experimentally determined structure (see Ref. 12 for details). A rule of thumb is that the quality of a structure prediction is good if the estimated TM score is at least 0.5.

Identification of SCOP Domains in the TE Structures and Their Enrichment—To provide a functional characterization of the TE proteins, we searched the SCOP database (21) to identify domains that are similar to the predicted TE proteins with a minimum TM score of 0.5. We searched SCOP for structurally similar domains, excluding the sequences with higher sequence similarity than 95% (ASTRAL95) in an all *versus* all manner: all TE structures were compared with all SCOP structures with TMfold (24). From the hits, we kept only those with a TM score higher than 0.5 and a minimum number of aligned residues higher than 80, as the probabilistic background of detecting shorter matches is not well understood (24). Because there is large structural redundancy within SCOP domains, we applied a further filtering step; from the overlapping SCOP matches, we kept only those most similar to the query TE structure, *i.e.* with the highest TM score, and the highest number of aligned residues closer than 5 Å. This step removed the redundant hits and resulted in 403 different SCOP hits to DNA transposon structures and 521 SCOP domains that are similar to LINE structures (see supplemental Tables 3 and 4).

The enrichment of SCOP protein folds in TE structures (supplemental Table 5) was calculated as the ratio of the frequency of a fold in the TE structures and the frequency of the corre-

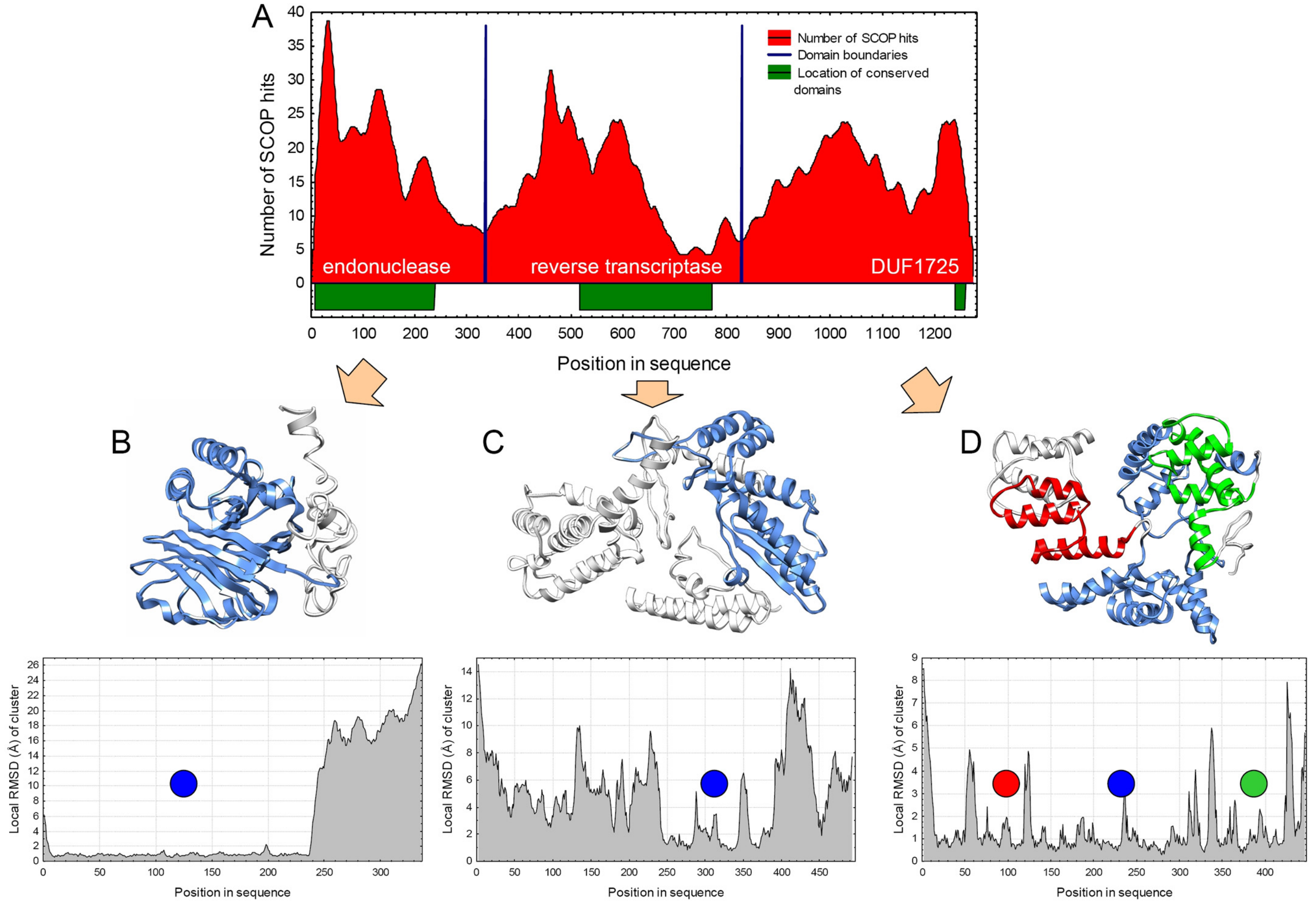

FIGURE 1. **Domain annotation, structure prediction, and quality determination of the protein models using the ORF2 protein of the human L1HS retrotransposon as an example.** *A*, the ORF2 protein of the human L1HS transposon is 1275 amino acids long, thus, a reliable model of the entire protein could not be built with current methods. Using a profile built from SCOP sequences similar to the target protein and annotation of conserved domains, we split the sequence into three regions, corresponding to the functional units of the protein: the endonuclease region, reverse-transcriptase region, and a cysteine-rich region with an unknown function (contains the Pfam conserved domain DUF1725). *B*, the I-TASSER protein model of the endonuclease region of the protein and the local r.m.s.d. distribution. Because a solved experimental structure for the human L1 endonuclease region is available in PDB, the structure of residues 1–235 is essentially similar to it, is characterized by very low r.m.s.d., and has a correct structure (TM score, 0.97; *blue*), whereas for the remaining 100 amino acids of the region, I-TASSER was not able to build a high quality structure. *C*, the predicted structure of the reverse-transcriptase region and the distribution of local r.m.s.d. values. The overall quality of the structure is low (estimated TM score, 0.33); however, the r.m.s.d. distribution shows a clear dip at the reverse transcriptase conserved domain, and the quality of the structure for this 170-residue region (residues 235–405) is essentially correct (highlighted with *blue*), with an estimated TM score of 0.59. *D*, the predicted structure of the region with the cysteine-rich domain and the distribution of the local r.m.s.d. values. The structure has a somewhat better overall quality than the reverse transcriptase region (TM score of 0.41) and can be split to several regions with low local r.m.s.d. (highlighted in *red*, *blue*, and *green*), which improves local TM scores to 0.45, 0.46, and 0.41, respectively.

sponding fold in the Uniprot database (25). For the latter, only those folds were used in the normalization which also occurred in TEs, *i.e.* the most common folds of TEs. This is necessary because Uniprot, due to its size, contains a much higher diversity of protein folds than a comparatively small set a proteins, and in consequence, the frequency of a particular fold in Uniprot is necessarily much lower if all folds are used in the normalization, which would result in a systematic overestimation of the enrichment in TEs. The abundances of particular folds in Uniprot were taken from the SUPERFAMILY database (26). Because the frequencies of low resolution protein structures (SCOP ID i.NN) are not present in the SUPERFAMILY database, these were not included in the analysis.

Search for Cases of TE Domestication and Protein Incorporations into TEs—We used the TE structures to provide an estimate of the frequency of TE domestication events using a recently published database of the Proteome Folding Project (PFP) (13). The database contains structural annotation of proteins on a genome scale from more than 94 organisms and provided novel structure predictions for 80,000 protein domains, which could not be annotated with SCOP, predicted with the Rosetta *de novo* structure prediction algorithm (28, 29). This data set is useful for detecting ancient TE homologies for two reasons: first, because *de novo* Rosetta does not use PDB templates for model building, the structural similarities are independent of the model building procedure of I-TASSER, *i.e.* any similarities are not due to using the same PDB structure as a template; second, due to the large number of organisms included in the PFP data set, it is possible to gain also information on the evolution/life style of organisms having proteins with TE homology.

Rosetta predicts an ensemble of structures for each submitted sequence; we used the structure with the highest quality score (Mammoth Confidence Metric, MCM) (30) for each domain and also excluded all domains where the highest MCM score was below 0.8; thus, the model was of low quality (13). This reduced the set of PFP structures used in the analysis to 16 000. We used the same all *versus* all approach for structural similarity searches as with the SCOP database (minimum required TM score of 0.5 and minimum length of aligned residues of 80).

Monte Carlo Simulations to Test for Homology/Analogy between Similar TE-PFP Structures—We tested whether the TE structures with similar topology (TM score > 0.5) to a PFP structure show any, at least remote sequence homology with a randomization procedure. Using a substitution matrix explicitly derived for homologous structures with remote sequence similarity (31), we calculated a sequence similarity score for the TE-PFP structural alignments obtained by TMfold. Next, we calculated 100,000 similarity scores between the TE sequence and random sequences, where the probability of selecting an amino acid for the random sequence was similar to the frequency of the amino acid in the PFP database. Significance was calculated with the formula shown in Equation 2,

$$p = (n + 1)/(N + 1) \qquad \text{(Eq. 2)}$$

where n is the number of random alignment scores equal or higher than that of the TE-PFP alignment, and N is the total number of random samples (100,000). Structure pairs with $p <= 0.001$ were accepted as homologous, whereas pairs with $p > 0.001$ were assumed to be analogous. The enrichment of particular taxonomic groups in the PFP homologs in comparison with the entire Proteome Folding Project database (Table 1) was calculated as the ratio of frequencies in the TE hits and their frequencies in the PFP database, and statistical significance was calculated with Chi square tests.

Identification of Cases of TE Domestication or Protein Incorporation—To decide whether the homology between a PFP protein domain and the sequence of a TE is the effect of transposon domestication or a different process, *e.g.* the incorporation of a host protein fragment into a TE, we implemented a protocol similar to Ref. 32. First, we reconstructed a global taxonomic tree of ~180,000 taxa, using known taxonomic relationships defined by NCBI. Next, using the sequences from a homologous sequence pair of a PFP protein and a transposable element, we searched the Uniprot database with the sequence fragment of the PFP protein using the jackhmmer tool of HMMER (33) with a bit score threshold 27. Using the species of the resulting matches, we identified the branch of the global tree where the particular domain is present. We repeated the same procedure for the TE fragment, using the six-frame translated RepBase as the sequence database, and then compared the two branches. The branch that contains the other was assumed to be the source of the sequence; for example, if a TE domain was present in repeats across Metazoans, whereas the homologous PFP domain was restricted to primates, we then concluded that this is a signature of the domestication of a TE protein in primates. If, however, the phylogenetic distribution of the PFP was broader than that of the distribution of the TE domain we concluded that a host protein was incorporated into a TE. In cases where the two branches were unrelated (for example mammals and parasitic protists), we assumed horizontal transfer.

RESULTS AND DISCUSSION

High Amounts of Disordered Sequence in ORF1 Proteins of LINEs—As the first step of the structural analysis, we identified the regions of the 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 in proteins that link rigid globular domains has 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 its native, functional form (34–36). Genome-wide analyses show that disorder is more frequent in eukaryotes than in prokaryotes (37), disordered proteins evolve quickly (38) and are typically involved in molecular recognition and interactions (39).

We identified the amount of disordered sequence in the RepeatPeps library, a collection of more than 5000 TE proteins, which is distributed as part of the RepeatMasker software suite. First, we identified coiled-coil domains using the Marcoil tool (40, 41); next, in the remaining part (*i.e.* not coiled-coil) of the protein sequences, we identified disordered regions with IUpred (42) (we also used DISOPRED2 (37), another disorder prediction tool that works on very different principles than IUpred, which leads to similar conclusions as IUpred (data not shown)). We probably underestimate the amount of intrinsically disordered sequence this way because disordered and coiled-coil sequences are frequently linked (43), and also regions that are disordered in the absence of other proteins may assume a coiled-coil conformation in a protein complex. We find that similarly to the human ORF1 (44, 45), a large fraction of ORF1 proteins in the CR1, L2, and L1 clades of LINE retrotransposons have a coiled-coil domain located close to their N terminus (Fig. 2, *A* and *B*) and that ORF1 and Gag proteins of LINE and LTR retrotransposons contain 5-fold higher amounts of disordered sequence than the proteins of DNA transposons or the ORF2/pol proteins of LINEs and LTRs (Fig. 2, *C* and *D*).

Although the location of coiled-coil regions is fairly similar in the different LINE clades, the distribution of disordered regions shows little similarity between the ORF1 proteins of different LINEs (Fig. 2*D*), except that their amount is much higher than in ORF2/pol proteins or DNA transposons. In the few repeats where the structure and function of the ORF1 protein is known, such as the human L1 repeat (45, 46), these proteins have nucleic acid chaperone function (47, 48) and show considerable flexibility (45). It has been shown that high amount of disordered sequence (49, 50) is a characteristic of chaperones; thus, our findings suggest that despite the fact that ORF1 proteins of different LINEs have independent origins (44), most of them may have chaperone functions and evolved highly flexible structures characterized by disorder and coiled-coil domains via convergent evolution. In a recent study, Callahan *et al.* (51) showed that human ORF1 proteins readily polymerize and form large aggregates, in which ORF1p nevertheless remain

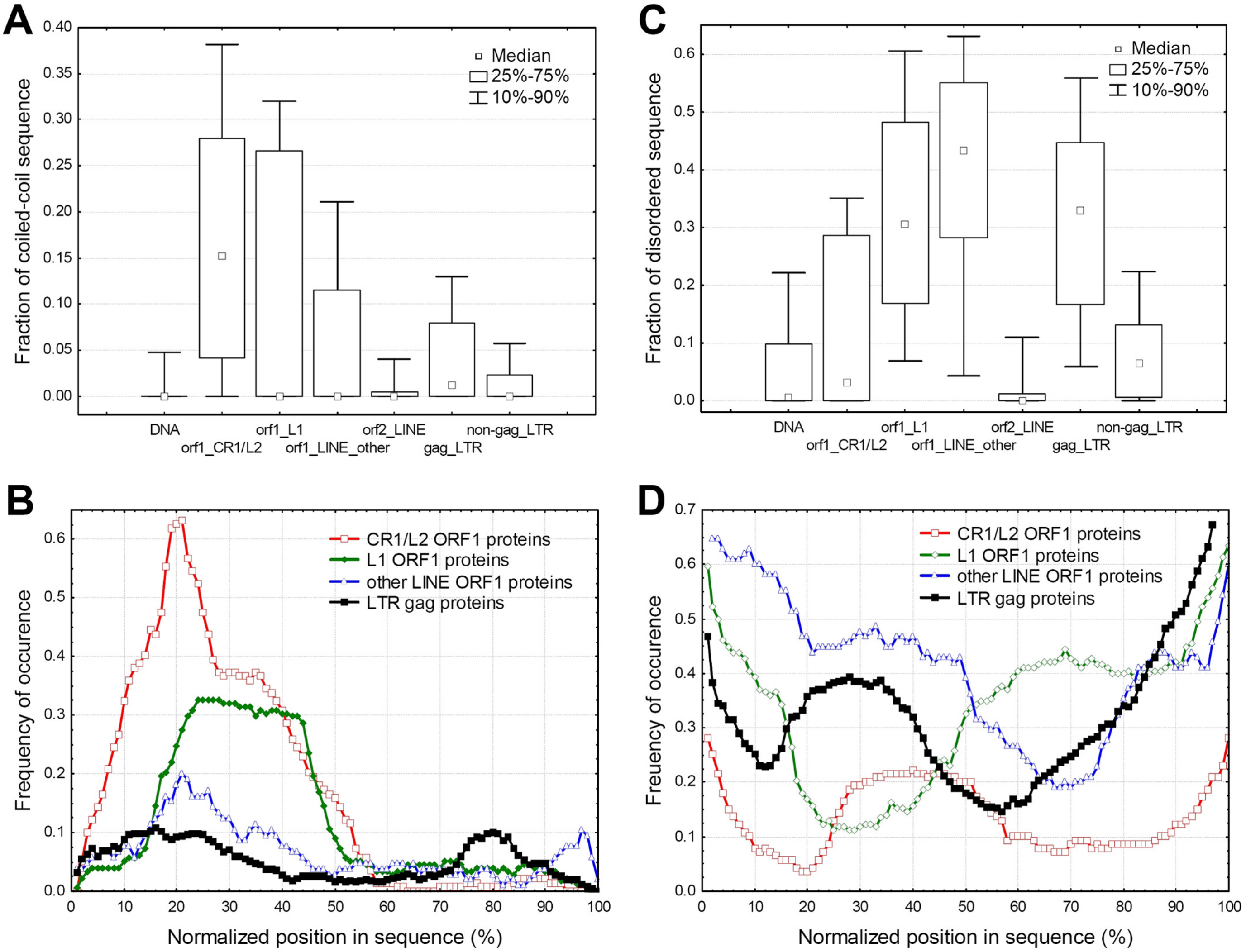

FIGURE 2. **Coiled-coil and intrinsically disordered regions in TE proteins.** *A*, the fraction of coiled-coil sequence in different TEs. ORF1 proteins of CR1, L1, and L2 families of LINEs are characterized by much higher amounts of coiled-coil sequence than ORF2 proteins or proteins of LTR retrotransposons. *B*, coiled-coil regions in the ORF1/Gag proteins are present near the N terminus of the sequence. *C*, the fraction of disordered sequence predicted wit IUpred in different TE protein types. ORF1/Gag proteins are characterized by ~5-fold higher amount of disordered sequence than ORF2 proteins of LINEs, LTR polyproteins, or DNA transposases. *D*, the distribution of disordered regions along the sequence of ORF1 proteins of LINEs and LTR Gag proteins.

functional. Because disordered protein fragments typically take a stable conformation when bound to their substrate proteins, we suggest that the high amount of flexible, disordered regions in ORF1 proteins indicates that the multimeric structure observed in human ORF1 can be generalized, may contribute to the formation of aggregates, and prevent their denaturation.

Quality of the Protein Structures—The quality of the protein models can be summarized with different quality scores: *C* (confidence) score (see "Materials and Methods"), or the estimated TM score with the native structure. The application of these scores to multidomain structures can be misleading (especially if the template coverage of the sequence is incomplete) because frequently individual domains within the structure are modeled correctly, but due to the uncertain structure of the linker regions (*i.e.* the uncertainty in the relative positioning of the domains), global confidence scores remain low. To account for this, besides the global confidence scores of the protein models, we also calculated a *C* score for 1283 regions (minimum length of 50 amino acids) in the protein models that

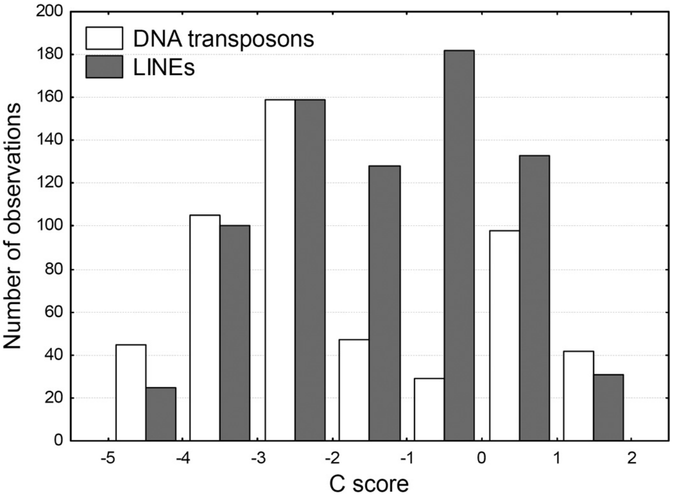

FIGURE 3. **Distribution of C scores of the low r.m.s.d. regions of the TE structures.** 61% of LINE low r.m.s.d. regions and 39% of DNA transposon low r.m.s.d. regions have a C-score higher than −1.78; thus, their estimated TM score is higher than 0.5.

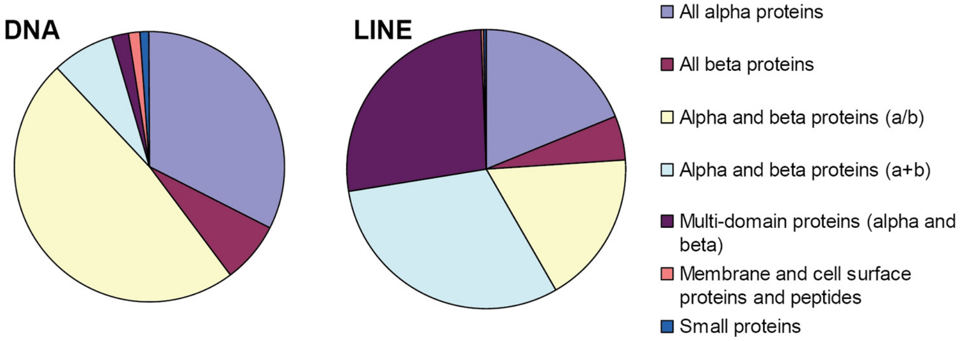

FIGURE 4. **The SCOP class composition of TE proteins.** DNA transposons are characterized mostly by all-α domains and α/β domains, whereas LINEs by multidomain hits and $\alpha+\beta$ domains (see also supplemental Table 5).

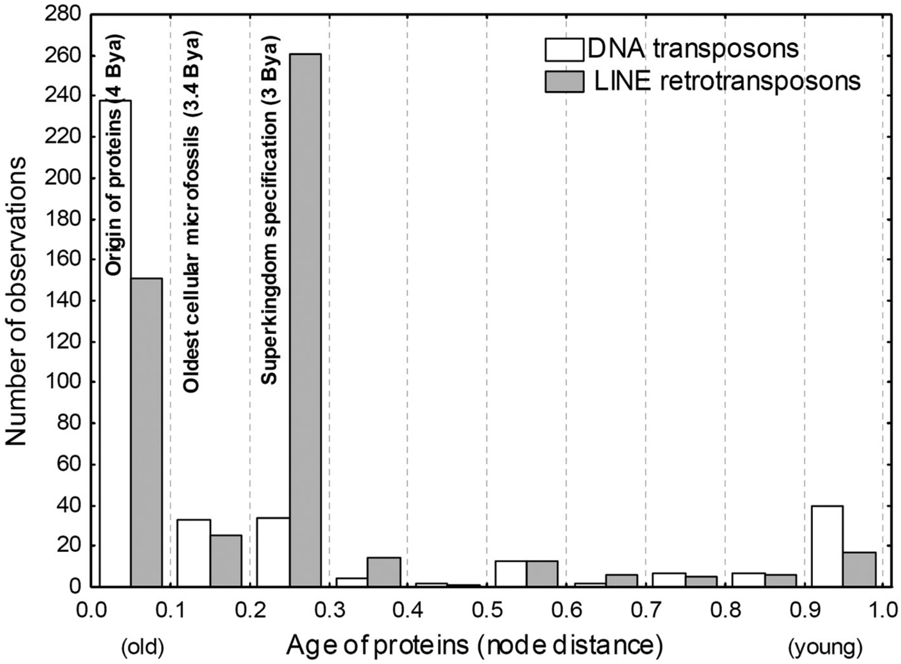

FIGURE 5. **The proteins of DNA transposons contain more ancient SCOP folds than LINE retrotransposons.** The age of protein folds is measured as node distance, a measure based on the phylogenetic spread of the fold; the larger the node distance, the younger the particular fold is, *i.e.* the more distant from the most ancient protein folds on the phylogenomic tree (see Ref. 54 for details). The histogram shows that in DNA transposons, the most abundant protein folds are among the most ancient ones, which were already present before the appearance of the first cellular organisms (~4 Bya), whereas the most frequent folds in LINEs were invented later, approximately at the time of the specification of the three superkingdoms (~3 Bya), suggesting that DNA substituted RNA as the carrier of genetic information already in the early Archean period.

are characterized by low r.m.s.d. within the I-TASSER clusters (see Fig. 1, *B–D*, and "Materials and Methods" for details). The C and TM scores are correlated; the C score of TASSER models typically falls in the range of −5 to +2, with scores above −1.78 indicating approximately correct topology, which means that the TM score between the predicted structure and the true structure is larger than 0.5. Overall, only 16% of LINE structures and 16.5% of DNA transposon proteins have their global *C* scores above −1.78 (supplemental Table 1); however, 61% of LINE low r.m.s.d. regions and 39% of DNA transposon low r.m.s.d. regions have the *C* score higher than −1.78, indicating that the quality of the structures at the domain level is much better than globally (Fig. 3 and supplemental Table 2).

Structural Composition of TE Proteins and the Time of Their Emergence—The distribution of SCOP classes in the matching TE domains reveals that the two types of TEs have different structural composition: DNA transposons are characterized predominantly by all-α domains and α/β domains, whereas LINEs with a more even distribution of SCOP classes, with and $\alpha+\beta$ domains being the most abundant (Fig. 4). Despite the differences in class composition, the examination of the most common folds of these two TE types show that they frequently use the same elements of the structural "alphabet" but with different frequencies (supplemental Table 5), which suggests that some of the domains shared by DNA transposons and LINEs have common origins (*i.e.* the DDE domain of transposases and retroviral integrases (52)).

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 folds of a protein contain 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 phylogenomic tree of folds, based on their abundance across different genomes (reviewed in Ref. 53). The observation that novel folds emerge at an approximately constant rate has even been used to date ancient and major evolutionary events such as the emergence of aerobic metabolism (54) or metal binding protein structures (55). To compare the age of proteins of DNA transposons and LINEs, we use the data from Wang *et al.* (54), which contain age estimates of protein folds (provided as node distance, the normalized number of nodes from the most ancient fold at the base of the global tree of folds) based on complete genomes from 749 species. The node distance distribution of the detected SCOP folds indicates that in DNA transposon proteins, the most abundant folds are among the most ancient known folds, which existed already before the appearance of the first cellular microfossils (3.4 Bya, Fig. 5) (56) and thus were probably present already 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 origins in the RNA world (reviewed in Ref. 57), the analysis of their protein structures indicates that although the most common protein folds of LINEs (SCOP IDs e.8, d.151) are indeed very ancient, they appeared approximately at the time of the specification of the three superkingdoms (Archaea, Bacteria, Eukaryota, 3 bya, Fig. 5) (54) after the most common folds of DNA transposons, and in consequence after the transition from RNA- to DNA-based replication.

Additional arguments for the ancient (Archean, > 2.5 Bya) origin of DNA transposons and LINEs come from their ligand binding. Both DNA transposon and LINE proteins bind metals, and different metal binding protein domains appeared at different periods in the history of life, which to a certain degree mirrored the presence of these metals in the environment. The history of metal utilization of proteins shows that the earliest metal binding protein domains evolved already in the Archean ocean and were either manganese-binding or bound to multiple metals (55, 58). Both the endonuclease domain of LINEs and the RNase H domain of DNA transposons bind manganese ions (and can also bind magnesium), which is in agreement with their very early origins. Additionally, the endonuclease domain of LINEs binds SO_4^{2-} ligands. Oxidized sulfur was absent before the first oxigenation of oceans ~2.9–2.5 Bya (59), which indicates that at least the current endonuclease domain of LINEs is adapted to the presence of oxygen in the environment (which does not necessarily mean oxygen in their microenvironment). It is currently unclear what are the cofactors of the reverse transcriptase (RT) domain of LINEs, but other RTs, *i.e.* telomerases (PDB code 3KYL) or HIV RT (PDB code 1VRT) are complexed with magnesium, and in retrotransposons, elevated manganese concentrations actually inhibit RT but not RNase H activity (60, 61).

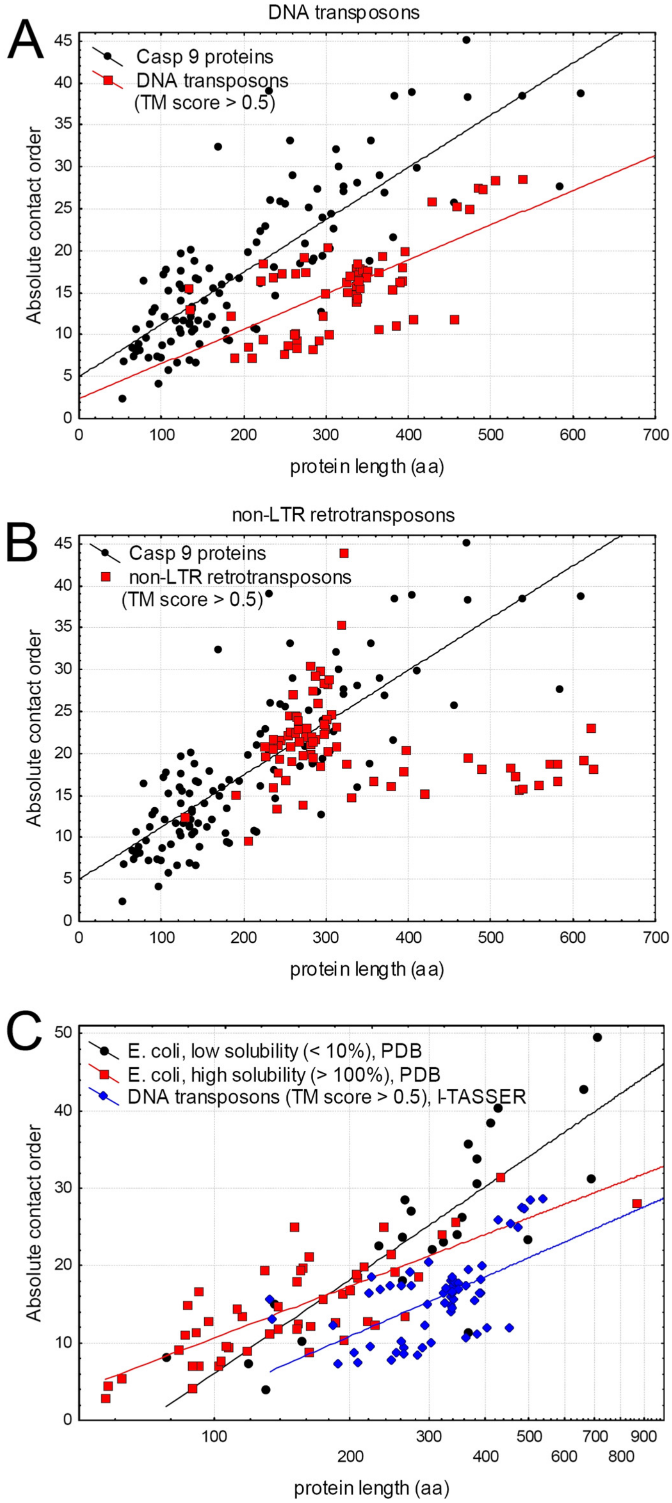

FIGURE 6. **Contact order of TE proteins.** *A*, the contact order of proteins of DNA transposons is significantly lower than the contact order of the reference CASP9 proteins ($p \ll 0.001$, ANCOVA), indicating that DNA transposons are under selection to fold rapidly. *B*, LINEs (non-LTR retrotransposons) do not show the same pattern ($p = 0.31$, ANCOVA). *C*, contact order of highly soluble and poorly soluble (prone to aggregation) *E. coli* proteins and DNA transposons.

The dominance of the most ancient ~4 Bya old folds in DNA transposons suggests that DNA transposons are as ancient as the oldest known proteins, implying that the RNA/RNP world did not last for billions of years, and, because DNA transposons need a DNA-based host organism to replicate, DNA was established as the carrier of genetic information early on in the evolution of life (although theoretically, it cannot be ruled out that transposases originally acted on RNA and were later adapted to DNA). Our findings indicate that genomic parasites are as old

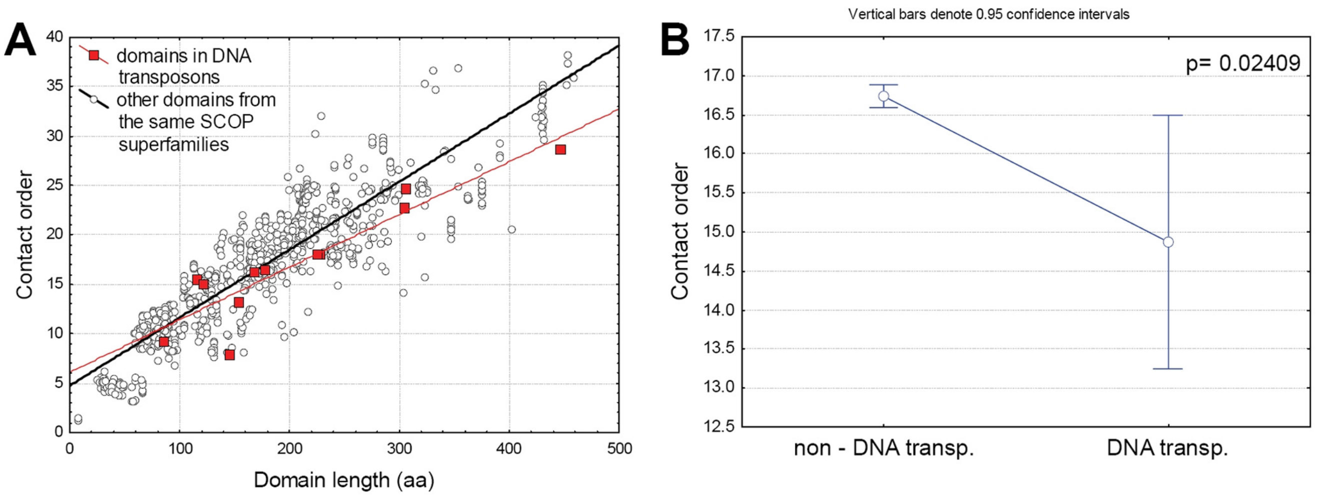

FIGURE 7. **The contact order of DNA transposon folds is low even within the same superfamilies.** *A*, correlations between length and absolute contact order, for SCOP families present in the high quality DNA transposon structures, and all other families from the same SCOP superfamilies. *B*, although the difference is small, SCOP families in DNA transposons have significantly lower contact order than other families from the same SCOP superfamilies (ANCOVA, $p = 0.02409$).

as genomes/proteins themselves, and from the two main mechanisms of transposition (cut-and-paste mechanism of DNA transposons and retrotransposition), the former is the older. Further insights into their origins could come from *in vitro* transposition assays, which could test the ability of TEs to "jump" in conditions that simulate Archeal ecosystems, be it hydrothermal vents, Archeal ocean, or other conditions (*i.e.* complete anoxia, lack of oxidized sulfur, presence of various cofactors, etc.).

DNA Transposon Proteins Evolved Low Contact Order, Possibly to Avoid Aggregation and Misfolding—Folding of proteins is a complex process and is subject to errors (62). Misfolded proteins are toxic due to aggregations, and misfolding has been proposed as one of the main causes for the variability in protein evolutionary rates (63–65). The probability of aggregation depends on the rate of folding, i.e., the amount of time the proteins spend in unfolded state; thus, one "strategy" to avoid it is simply to fold rapidly. In general, large, complex proteins fold slowly, and their folding is normally guided by other proteins, chaperones, that prevent aggregation and misfolding (reviewed in Ref. 62). The rate of protein folding has been measured for relatively few proteins; however, in these, a strong correlation has been established between folding rate and contact order (66), a measure of structural complexity, defined as the average distance along the sequence between non-hydrogen atoms that are physically closer than 6 Å in the folded protein.

Transposable elements are frequently present in large numbers in their hosts; thus, their high copy number may result in high protein abundance as well, at least in certain developmental stages when they are derepressed. Therefore, they may face a stronger evolutionary pressure to avoid protein aggregation than "regular" proteins, with only one or few copies in the genome. We calculated contact order for all TE structures with a global TM score higher than 0.5 and compared it with the contact order of proteins used in the CASP9 experiment, which were folded with I-TASSER.

We found that DNA transposons are characterized by significantly lower contact order than the reference proteins; thus, DNA transposons are under selection to fold rapidly (Fig. 6*A*). We did not find the same pattern for LINE proteins (Fig. 6*B*); most likely, the sizes of ORF2 proteins of LINEs are so big that they have to rely on chaperones for correct folding. Our finding indicates that DNA transposons are under selection for rapid folding *in vivo*.

The low contact order of DNA transposons can be the result of at least two phenomena: (*a*) DNA transposons might be built from domains with lower contact order than other proteins or (*b*) the domains in DNA transposons have lower contact order than in their homologs in other proteins. We tested the second hypothesis with the following method: we calculated the contact order for the SCOP domains present in the high quality (TM score > 0.5) DNA transposon structures and also for all other domains from the same SCOP superfamilies. Their comparison (with ANCOVA) shows that the SCOP domains present in DNA transposons have significantly lower contact order than other domains of the same superfamilies (Fig. 7; $p = 0.024$). However, the effect is small (Fig. 7); thus, the observed large difference in contact order between DNA transposons and other proteins is more due to a different overall domain composition and general topology of the proteins than due to differences in homologous domains.

One possible explanation for this pattern is that rapid folding is necessary to reduce aggregation, which is consistent with experimental observations of overproduction inhibition in DNA transposons (67, 68), *i.e.* that high amounts of transposases actually result in reduced levels of activity. We tested this hypothesis using *Escherichia coli* proteins, for which both solubility was measured (69), and an experimentally derived structure is also available in PDB (Fig. 6*C*). Although the soluble *E. coli* proteins (and DNA transposons) have significantly lower contact orders than insoluble ones (Fig. 6*C*), the results also show that contact order is only one of the factors determining solubility, and other properties of proteins (such as size or exposed hydrophobic residues) also have a significant effect; thus, aggregation propensity may not be the only factor causing the observed low contact order, although it probably contrib-

TABLE 1

Taxonomic distribution of TE hits to the PFP database (both sequence- and structure-based searches) and their over- or under-representation

Enrichment was calculated as the (frequency of TE hits)/(frequency in the database); significance was calculated with Chi square tests.

Taxonomic group	Structure hits	Enrichment	*p*	Sequence hits	Enrichment	*p*	Nr in database
Mobile elements	4	1.52	0.400	17	41.305	<0.001	145
Viruses, phages	3	0.35	0.052	0	0.000	0.240	473
Prokaryotes	127	0.86	0.020	1	0.044	<0.001	8076
Protists	46	1.63	<0.001	1	0.228	0.089	1547
Fungi	18	0.93	0.735	1	0.330	0.227	1068
Metazoans	52	1.22	0.128	6	0.899	0.778	2351
Plants	45	0.97	0.830	20	2.768	<0.001	2546
Total	295			46			16,206

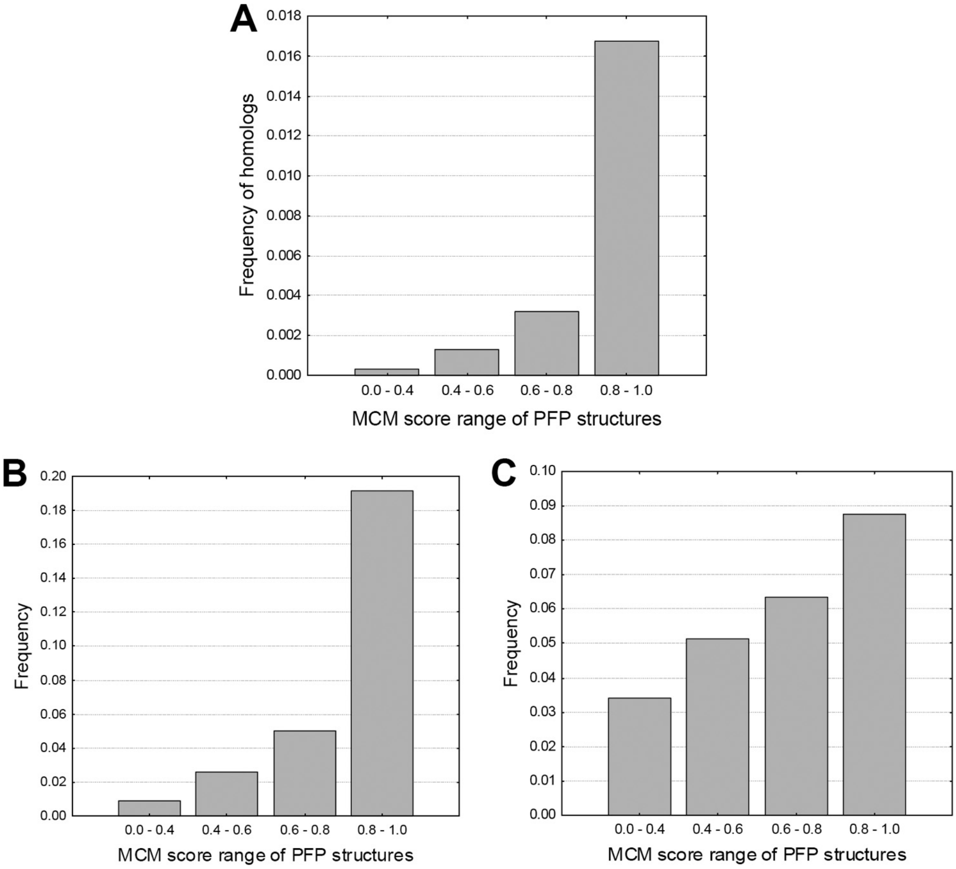

FIGURE 8. **The relationship between the quality of a PFP structure and the likelihood of detecting structural similarity/homology with a TE.** *A*, the probability of detecting homology between the TE structures and the proteome folding project structures depends largely on the quality (MCM score) of PFP decoys. Structures with an MCM score of 0.8 have mostly correct topology (two of three are correct), whereas below MCM score 0.4, their quality is low and are mostly incorrect. The quality of PFP structures has a very large effect on the number of detected homologs, which is the result of two independent processes: 1) the probability of detecting structural similarity (TM score > 0.5) between TE and PFP structures increases radically with the increasing quality (MCM score) of the PFP structure (*B*); 2) in the identified similar structure pairs, the fraction of pairs with significant sequence similarity ($p < 0.001$) also increases with the quality of the structures, although less dramatically (~2.5-fold; *C*). This has two consequences for the estimation of false positive rate of homolog detection. First, the fraction of incorrectly detected structure pairs due to modeling errors is probably low: we detect real analogs and homologs (*B*). However, based on the fraction of cases with significant sequence similarity ($p < 0.001$) where the structural similarity between a TE and a PFP decoy is likely to be an artifact (*C*; MCM score of PFP structures < 0.4), the number of homologs is probably overestimated, with up to 40%.

utes to it. The difference between DNA transposons and LINEs may to a certain degree be explained with the different age of the two types of repeats, as the low contact orders of DNA transposons may reflect ancient, chaperone-free environments.

Search for Novel Cases of TE Domestication and Estimation of Their Frequency—We compared the 16,000 high quality structures (MCM score > 0.8) of the Proteome Folding Project database with the TE structures (see "Materials and Methods"); we found as many as 3743 different PFP structures that show structural similarity to a TE protein. The high number of similar structures is partly due to a structural redundancy among the PFP hits: clustering the structure pairs from the 3743 PFP structures with a TM score threshold of 0.5 indicates that only three clusters contain more than 2600 of the hits. To separate homologous from analogous hits, we performed Monte Carlo simulations to detect remote sequence similarities ($p < 0.001$) between the TE and PFP structures, which reduced the set to 295 hits considered as homologous (Table 1 and see "Materials and Methods" for details).

To investigate what process caused the observed homologies (*i.e.* horizontal transfer, domestication by the host of TE, or incorporation of a host protein into a TE sequence) we compared the phylogenetic spread of the TE proteins and the homologous PFP proteins (see "Materials and Methods") to find out the most likely evolutionary scenario for each case (supplemental Table 6). We found that 25 cases support TE domestication, whereas the number of cases supporting the opposite process, the incorporation of a host protein domain into a TE is 100. In a large number of cases (80 cases), the lateral transfer of a TE domain is the most likely evolutionary scenario (supplemental Table 6). In 48 cases, it was not possible to find out in what direction the protein domain was transferred because both the TE and the PFP protein is present across the entire eukaryotic kingdom; thus, these cases may represent very ancient domestications or incorporations of host proteins; and in 41 cases, all homologs to the PFP hits were transposons or uncharacterized proteins; these were excluded from the analysis.

Comparison of Structure- and Sequence-based Searches—To compare the efficiency of structure searches with traditional sequence-based searches, we also performed sequence comparisons using hidden Markov models. The comparison shows that the two methods have different strengths and weaknesses and can be seen as methods that complement each other. Although we identified 295 different PFP structures that were similar at TM score $>$ 0.5 level to a PFP structure and also passed the homology filter (see "Materials and Methods"), searching the TE sequences against the same PFP sequences with the phmmer tool of the HMMER package (with an e-value cut-off of 0.001) resulted only in 46 PFP hits (Table 1). From these, the evolutionary analysis (see "Materials and Methods") identified only one case as domestication, 14 cases of protein incorporations into a TE, and in 29 cases, the origin of the sequence could not be identified (supplemental Table 7).

Overall, the structure-based search identified more than six times as many PFP hits as the sequence-based search. However, despite the higher sensitivity, structure searches are not without pitfalls: first, structure space is much more limited than sequence space (70), and in consequence, highly similar structures can be analogous not only homologous, and separating remote homology from analogy is not a straightforward task. We identified more than 10 times as many analogs as homologs, and our estimated false positive rate for homology detection is still high, possibly reaching 40% (see Fig. 8 for details). Second, currently, it is not possible to make iterative searches with structures, which offer much higher sensitivity for sequences. Third, the probabilistic background of structure comparisons is far less developed than for sequences, and structural similarity scores (including the TM score) typically represent global structural alignments, which are meaningful only above a certain protein size and do not introduce gaps in the structures (*i.e.* do not move structure fragments relative to each other). The comparison of PFP hits identified by structure and sequence searches shows that the two methods identify a very different set of sequences (Table 1): in the case of structure searches, most taxonomic groups are present relatively evenly, with parasitic protists significantly overrepresented, whereas bacterial proteins are underrepresented. In contrast, sequence searches identify mostly hits to transposable elements and plant proteins, which, due to the relatively incomplete annotation of TEs in most plant genomes, in many cases are also likely to be TEs (Table 1).

Our findings indicate that the frequency of sequence exchange between TEs and their hosts (either TE domestication or the incorporation of a host protein into a TE) is considerably higher than it is detected by sequence searches. However, their number is still not high: from 16,000 PFP structures, we identified slightly more than 2% as homologous to a TE. On the other hand the relatively large number of cases of putative horizontal transfer, many of which involves pathogenic bacteria and parasitic protists corroborates recent hypotheses on the importance of parasites in mediating horizontal transfer of TEs (27, 71).

Acknowledgments—We thank Ervin Welker and Balázs Papp for useful comments and suggestions, and Minglei Wang and Gustavo Caetano-Anolles for providing node distance data.

REFERENCES

1. Feschotte, C. (2008) Transposable elements and the evolution of regulatory networks. *Nat. Rev. Genet.* **9,** 397–405
2. Zhang, Y., Romanish, M. T., and Mager, D. L. (2011) Distributions of transposable elements reveal hazardous zones in mammalian introns. *PLoS Comput. Biol.* 7, e1002046
3. Jurka, J., Kapitonov, V. V., Kohany, O., and Jurka, M. V. (2007) Repetitive sequences in complex genomes: structure and evolution. *Annu. Rev. Genomics Hum. Genet.* **8,** 241–259
4. Feschotte, C., and Pritham, E. J. (2007) DNA transposons and the evolution of eukaryotic genomes. *Annu. Rev. Genet.* **41,** 331–368
5. ENCODE Project Consortium, Dunham, I., Kundaje, A., Aldred, S. F., Collins, P. J., Davis, C. A., Doyle, F., Epstein, C. B., Frietze, S., Harrow, J., Kaul, R., Khatun J, Lajoie, B. R., Landt, S. G., Lee, B. K., Pauli, F., Rosenbloom, K. R., Sabo, P., Safi, A., Sanyal, A., Shoresh, N., Simon, J. M., Song, L., Trinklein, N. D., Altshuler, R. C., Birney, E., Brown, J. B., Cheng, C., Djebali, S., Dong, X., Dunham, I., Ernst, J., Furey, T. S., Gerstein, M., Giardine, B., Greven, M., Hardison, R. C., Harris, R. S., Herrero, J., Hoffman, M. M., Iyer, S., Kelllis, M., Khatun, J., Kheradpour, P., Kundaje, A., Lassman, T., Li, Q., Lin, X., Marinov, G. K., Merkel, A., and Mortazavi, A. (2012) An integrated encyclopedia of DNA elements in the human genome. *Nature* **489,** 57–74
6. Lindblad-Toh, K., Garber, M., Zuk, O., Lin, M. F., Parker, B. J., Washietl, S., Kheradpour, P., Ernst, J., Jordan, G., Mauceli, E., Ward, L. D., Lowe, C. B., Holloway, A. K., Clamp, M., Gnerre, S., Alföldi, J., Beal, K., Chang, J., Clawson, H., Cuff, J., Di Palma, F., Fitzgerald, S., Flicek, P., Guttman, M., Hubisz, M. J., Jaffe, D. B., Jungreis, I., Kent, W. J., Kostka, D., Lara, M., Martins, A. L., Massingham, T., Moltke, I., Raney, B. J., Rasmussen, M. D., Robinson, J., Stark, A., Vilella, A. J., Wen, J., Xie, X., Zody, M. C., Broad Institute Sequencing Platform and Whole Genome Assembly Team, Baldwin, J., Bloom, T., Chin, C. W., Heiman, D., Nicol, R., Nusbaum, C., Young, S., Wilkinson, J., and Worley, K. C. (2011) A high-resolution map of human evolutionary constraint using 29 mammals. *Nature* **478,** 476–482
7. Lander, E. S., Linton, L. M., Birren, B., Nusbaum, C., Zody, M. C., Baldwin, J., Devon, K., Dewar, K., Doyle, M., FitzHugh, W., Funke, R., Gage, D., Harris, K., Heaford, A., Howland, J., Kann, L., Lehoczky, J., LeVine, R., McEwan, P., McKernan, K., Meldrim, J., Mesirov, J. P., Miranda, C., Morris, W., Naylor, J., Raymond, C., Rosetti, M., Santos, R., Sheridan, A., Sougnez, C., Stange-Thomann, N., Stojanovic, N., Subramanian, A., Wyman, D., Rogers, J., Sulston, J., Ainscough, R., Beck, S., Bentley, D., Burton, J., Clee, C., Carter, N., Coulson, A., Deadman, R., Deloukas, P., Dunham, A., Dunham, I., Durbin, R., French, L., Grafham, D., and Gregory, S. (2001) Initial sequencing and analysis of the human genome. *Nature* **409,** 860–921
8. Zdobnov, E.M., Campillos, M., Harrington, E. D., Torrents, D., and Bork, P. (2005) Protein coding potential of retroviruses and other transposable

elements in vertebrate genomes. *Nucleic Acids Res.* **33,** 946–954

9. Britten, R. (2006) Transposable elements have contributed to thousands of human proteins. *Proc. Natl. Acad. Sci. U.S.A.* **103,** 1798–1803
10. Tramontano, A. (2006) *Protein Structure Prediction: Concepts and Applications*, pp. 37–39, 1st Ed. Wiley-VCH, Weinheim, Germany
11. Roy, A., Kucukural, A., and Zhang, Y. (2010) I-TASSER: a unified platform for automated protein structure and function prediction. *Nat Protoc.* **5,** 725–738
12. Zhang, Y. (2008) I-TASSER server for protein 3D structure prediction. *BMC Bioinformatics* **9,** 40
13. Drew, K., Winters, P., Butterfoss, G. L., Berstis, V., Uplinger, K., Armstrong, J., Riffle, M., Schweighofer, E., Bovermann, B., Goodlett, D. R., Davis, T. N., Shasha, D., Malmström, L., and Bonneau, R. (2011) The Proteome Folding Project: Proteome-scale prediction of structure and function. *Genome Res.* **21,** 1981–1994
14. Jurka, J., Kapitonov, V. V., Pavlicek, A., Klonowski, P., Kohany, O., and Walichiewicz, J. (2005) Repbase Update, a database of eukaryotic repetitive elements. *Cytogenet. Genome Res.* **110,** 462–467
15. Edgar, R.C. (2010) Search and clustering orders of magnitude faster than BLAST. *Bioinformatics* **26,** 2460–2461
16. Malik, H. S., Burke, W. D., and Eickbush, T. H. (1999) The age and evolution of non-LTR retrotransposable elements. *Mol. Biol. Evol.* **16,** 793–805
17. Mátés, L., Chuah, M. K., Belay, E., Jerchow, B., Manoj, N., Acosta-Sanchez, A., Grzela, D. P., Schmitt, A., Becker, K., Matrai, J., Ma, L., Samara-Kuko, E., Gysemans, C., Pryputniewicz, D., Miskey, C., Fletcher, B., VandenDriessche, T., Ivics, Z., and Izsvák, Z. (2009) Molecular evolution of a novel hyperactive Sleeping Beauty transposase enables robust stable gene transfer in vertebrates. *Nat. Genet.* **41,** 753–761
18. Yusa, K., Zhou, L., Li, M. A., Bradley, A., and Craig, N. L. (2011) A hyperactive piggyBac transposase for mammalian applications. *Proc. Natl. Acad. Sci. U.S.A.* **108,** 1531–1536
19. Marchler-Bauer, A., Lu, S., Anderson, J. B., Chitsaz, F., Derbyshire, M. K., DeWeese-Scott, C., Fong, J. H., Geer, L. Y., Geer, R. C., Gonzales, N. R., Gwadz, M., Hurwitz, D. I., Jackson, J. D., Ke, Z., Lanczycki, C. J., Lu, F., Marchler, G. H., Mullokandov, M., Omelchenko, M. V., Robertson, C. L., Song, J. S., Thanki, N., Yamashita, R. A., Zhang, D., Zhang, N., Zheng, C., and Bryant, S. H. (2011) CDD: a Conserved Domain Database for the functional annotation of proteins. *Nucleic Acids Res.* **39,** D225–D229
20. Bondugula, R., Lee, M. S., and Wallqvist, A. (2009) FIEFDom: a transparent domain boundary recognition system using a fuzzy mean operator. *Nucleic Acids Res.* **37,** 452–462
21. Andreeva, A., Howorth, D., Chandonia, J. M., Brenner, S. E., Hubbard, T. J., Chothia, C., and Murzin, A. G. (2008) Data growth and its impact on the SCOP database: new developments. *Nucleic Acids Res.* **36,** D419–D425
22. Berman, H., Henrick, K., Nakamura, H., and Markley, J. L. (2007) The worldwide Protein Data Bank (wwPDB): ensuring a single, uniform archive of PDB data. *Nucleic Acids Res.* **35,** D301–D303
23. Zhang, Y., and Skolnick, J. (2004) Scoring function for automated assessment of protein structure template quality. *Proteins* **57,** 702–710
24. Xu, J., and Zhang, Y. (2010) How significant is a protein structure similarity with TM-score = 0.5? *Bioinformatics* **26,** 889–895
25. UniProt Consortium (2011) Ongoing and future developments at the Universal Protein Resource. *Nucleic Acids Res.* **39,** D214–D219
26. Wilson, D., Pethica, R., Zhou, Y., Talbot, C., Vogel, C., Madera, M., Chothia, C., and Gough, J. (2009) SUPERFAMILY–sophisticated comparative genomics, data mining, visualization and phylogeny. *Nucleic Acids Res.* **37,** D380–D386
27. Gilbert, C., Schaack, S., Pace, J. K., 2nd, Brindley, P. J., and Feschotte, C. (2010) A role for host-parasite interactions in the horizontal transfer of transposons across phyla. *Nature* **464,** 1347–1350
28. Das, R., and Baker, D. (2008) Macromolecular modeling with rosetta. *Annu. Rev. Biochem.* **77,** 363–382
29. Leaver-Fay, A., Tyka, M., Lewis, S. M., Lange, O. F., Thompson, J., Jacak, R., Kaufman, K., Renfrew, P. D., Smith, C. A., Sheffler, W., Davis, I. W., Cooper, S., Treuille, A., Mandell, D. J., Richter, F., Ban, Y. E., Fleishman, S. J., Corn, J. E., Kim, D. E., Lyskov, S., Berrondo, M., Mentzer, S., Popović, Z., Havranek, J. J., Karanicolas, J., Das, R., Meiler, J., Kortemme, T., Gray, J. J., Kuhlman, B., Baker, D., and Bradley, P. (2011) ROSETTA3: an object-oriented software suite for the simulation and design of macromolecules. *Methods Enzymol.* **487,** 545–574
30. Malmström, L., Riffle, M., Strauss, C. E., Chivian, D., Davis, T. N., Bonneau, R., and Baker, D. (2007) Superfamily assignments for the yeast proteome through integration of structure prediction with the gene ontology. *PLoS Biol.* **5,** e76
31. Prlić, A., Domingues, F.S., and Sippl, M. J. (2000) Structure-derived substitution matrices for alignment of distantly related sequences. *Protein Eng.* **13,** 545–550
32. Abrusán, G., Szilágyi, A., Zhang, Y., and Papp, B. (2013) Turning gold into 'junk': transposable elements utilize central proteins of cellular networks. *Nucleic Acids Res.* **41,** 3190–3200
33. Johnson, L. S., Eddy, S. R., and Portugaly, E. (2010) Hidden Markov model speed heuristic and iterative HMM search procedure. *BMC Bioinformatics* **11,** 431
34. Dyson, H. J., and Wright, P. E. (2005) Intrinsically unstructured proteins and their functions. *Nat. Rev. Mol. Cell Biol.* **6,** 197–208
35. Wright, P. E., and Dyson, H. J. (1999) Intrinsically unstructured proteins: re-assessing the protein structure-function paradigm. *J. Mol. Biol.* **293,** 321–331
36. Tompa, P. (2012) Intrinsically disordered proteins: a 10-year recap. *Trends Biochem. Sci.* **37,** 509–516
37. Ward, J. J., Sodhi, J. S., McGuffin, L. J., Buxton, B. F., and Jones, D. T. (2004) Prediction and functional analysis of native disorder in proteins from the three kingdoms of life. *J. Mol. Biol.* **337,** 635–645
38. Brown, C. J., Johnson, A. K., Dunker, A. K., and Daughdrill, G. W. (2011) Evolution and disorder. *Curr. Opin. Struct. Biol.* **21,** 441–446
39. Dunker, A. K., Silman, I., Uversky, V. N., and Sussman, J. L. (2008) Function and structure of inherently disordered proteins. *Curr. Opin. Struct. Biol.* **18,** 756–764
40. Delorenzi, M., and Speed, T. (2002) An HMM model for coiled-coil domains and a comparison with PSSM-based predictions. *Bioinformatics* **18,** 617–625
41. Gruber, M., Söding, J., and Lupas, A. N. (2006) Comparative analysis of coiled-coil prediction methods. *J. Struct. Biol.* **155,** 140–145
42. Dosztányi, Z., Csizmok, V., Tompa, P., and Simon, I. (2005) IUPred: web server for the prediction of intrinsically unstructured regions of proteins based on estimated energy content. *Bioinformatics* **21,** 3433–3434
43. Anurag, M., Singh, G. P., and Dash, D. (2012) Location of disorder in coiled coil proteins is influenced by its biological role and subcellular localization: a GO-based study on human proteome. *Mol. Biosyst.* **8,** 346–352
44. Khazina, E., and Weichenrieder, O. (2009) Non-LTR retrotransposons encode noncanonical RRM domains in their first open reading frame. *Proc. Natl. Acad. Sci. U.S.A.* **106,** 731–736
45. Khazina, E., Truffault, V., Büttner, R., Schmidt, S., Coles, M., and Weichenrieder, O. (2011) Trimeric structure and flexibility of the L1ORF1 protein in human L1 retrotransposition. *Nat. Struct. Mol. Biol.* **18,** 1006–1014
46. Januszyk, K., Li, P. W., Villareal, V., Branciforte, D., Wu, H., Xie, Y., Feigon, J., Loo, J. A., Martin, S. L., and Clubb, R. T. (2007) Identification and solution structure of a highly conserved C-terminal domain within ORF1p required for retrotransposition of long interspersed nuclear element-1. *J. Biol. Chem.* **282,** 24893–24904
47. Martin, S. L. (2006) The ORF1 protein encoded by LINE-1: structure and function during L1 retrotransposition. *J. Biomed. Biotechnol.* **2006,** 45621
48. Nakamura, M., Okada, N., and Kajikawa, M. (2012) Self-Interaction, Nucleic Acid Binding, and Nucleic Acid Chaperone Activities Are Unexpectedly Retained in the Unique ORF1p of Zebrafish LINE. *Mol. Cell. Biol.* **32,** 458–469
49. Tompa, P., and Csermely, P. (2004) The role of structural disorder in the function of RNA and protein chaperones. *FASEB J.,* **18,** 1169–1175
50. Tompa, P., and Kovacs, D. (2010) Intrinsically disordered chaperones in plants and animals. *Biochem. Cell Biol.* **88,** 167–174
51. Callahan, K. E., Hickman, A. B., Jones, C. E., Ghirlando, R., and Furano, A. V. (2012) Polymerization and nucleic acid-binding properties of human L1 ORF1 protein. *Nucleic Acids Res.* **40,** 813–827

52. Capy, P., Langin, T., Higuet, D., Maurer, P., and Bazin, C. (1997) Do the integrases of LTR-retrotransposons and class II element transposases have a common ancestor? *Genetica* **100,** 63–72
53. Caetano-Anollés, G., Wang, M., Caetano-Anollés, D., and Mittenthal, J. E. (2009) The origin, evolution and structure of the protein world. *Biochem. J.* **417,** 621–637
54. Wang, M., Jiang, Y. Y., Kim, K. M., Qu, G., Ji, H. F., Mittenthal, J. E., Zhang, H. Y., and Caetano-Anollés, G. (2011) A universal molecular clock of protein folds and its power in tracing the early history of aerobic metabolism and planet oxygenation. *Mol. Biol. Evol.* **28,** 567–582
55. Dupont, C. L., Butcher, A., Valas, R. E., Bourne, P. E., and Caetano-Anollés, G. (2010) History of biological metal utilization inferred through phylogenomic analysis of protein structures. *Proc. Natl. Acad. Sci. U.S.A.* **107,** 10567–10572
56. Wacey, D., Kilburn, M. R., Saunders, M., Cliff, J., and Brasier, M. D. (2011) Microfossils of sulphur-metabolizing cells in 3.4-billion-year-old rocks of Western Australia. *Nat. Geosci.* **4,** 698–702
57. Brosius, J. (2005) Echoes from the past–are we still in an RNP world? *Cytogenet. Genome Res.* **110,** 8–24
58. Ji, H. F., Chen, L., Jiang, Y. Y., and Zhang, H. Y. (2009) Evolutionary formation of new protein folds is linked to metallic cofactor recruitment. *Bioessays* **31,** 975–980
59. Anbar, A. D. (2008) Oceans. Elements and evolution. *Science* **322,** 1481–1483
60. Yarrington, R. M., Chen, J., Bolton, E. C., and Boeke, J. D. (2007) Mn^{2+} suppressor mutations and biochemical communication between Ty1 reverse transcriptase and RNase H domains. *J. Virol.* **81,** 9004–9012
61. Bolton, E. C., Mildvan, A. S., and Boeke, J. D. (2002) Inhibition of reverse transcription in vivo by elevated manganese ion concentration. *Mol. Cell.* **9,** 879–889
62. Dobson, C. M. (2003) Protein folding and misfolding. *Nature* **426,** 884–890
63. Drummond, D. A., and Wilke, C. O. (2008) Mistranslation-induced protein misfolding as a dominant constraint on coding-sequence evolution. *Cell* **134,** 341–352
64. Drummond, D. A., and Wilke, C. O. (2009) The evolutionary consequences of erroneous protein synthesis. *Nat. Rev. Genet.* **10,** 715–724
65. Pál, C., Papp, B., and Lercher, M. J. (2006) An integrated view of protein evolution. *Nat. Rev. Genet.* **7,** 337–348
66. Ivankov, D. N., Garbuzynskiy, S. O., Alm, E., Plaxco, K. W., Baker, D., and Finkelstein, A. V. (2003) Contact order revisited: influence of protein size on the folding rate. *Protein Sci.* **12,** 2057–2062
67. Grabundzija, I., Irgang, M., Mátés, L., Belay, E., Matrai, J., Gogol-Döring, A., Kawakami, K., Chen, W., Ruiz, P., Chuah, M. K., VandenDriessche, T., Izsvák, Z., and Ivics, Z. (2010) Comparative analysis of transposable element vector systems in human cells. *Mol. Ther.* **18,** 1200–1209
68. Ni, J., Clark, K. J., Fahrenkrug, S. C., and Ekker, S. C. (2008) Transposon tools hopping in vertebrates. *Brief. Funct. Genomics Proteomics* **7,** 444–453
69. Niwa, T., Ying, B. W., Saito, K., Jin, W., Takada, S., Ueda, T., and Taguchi, H. (2009) Bimodal protein solubility distribution revealed by an aggregation analysis of the entire ensemble of Escherichia coli proteins. *Proc. Natl. Acad. Sci. U.S.A.* **106,** 4201–4206
70. Zhang, Y., Hubner, I. A., Arakaki, A. K., Shakhnovich, E., and Skolnick, J. (2006) On the origin and highly likely completeness of single-domain protein structures. *Proc. Natl. Acad. Sci. U.S.A.* **103,** 2605–2610
71. Schaack, S., Gilbert, C., and Feschotte, C. (2010) Promiscuous DNA: horizontal transfer of transposable elements and why it matters for eukaryotic evolution. *Trends Ecol. Evol.* **25,** 537–546