Preprint version of Ángyán et al., J. Mol. Evol. 78(5):263-274 (2014)
The final publication is available at Springer via http://dx.doi.org/10.1007/s00239-014-9622-3

Are proposed early genetic codes capable of encoding viable proteins?

Annamária Franciska Ángyán[1], Csaba Ortutay[2], Zoltán Gáspári[1]*

[1]Pázmány Péter Catholic University, Faculty of Information Technology, 1083 Budapest, Práter u. 50/A, Hungary Tel.: +36-1-886-4780; Fax: +36-1-886-4724.

[2] Institute of Biomedical Technology, FI-33014 University of Tampere, Tampere, Finland

***** Author to whom correspondence should be addressed; e-mail: gaspari.zoltan@itk.ppke.hu

Abstract

Proteins are elaborate biopolymers balancing between contradicting intrinsic propensities to fold, aggregate or remain disordered. Assessing their primary structural preferences observable without evolutionary optimization has been reinforced by the recent identification of *de novo* proteins that have emerged from previously non-coding sequences. In this paper we investigate structural preferences of hypothetical proteins translated from random DNA segments using the standard genetic code and three of its proposed evolutionarily predecessor models encoding 10, 6 and 4 amino acids, respectively. Our only main assumption is that the disorder, aggregation and transmembrane helix predictions used are able to reflect the differences in the trends of the protein sets investigated. We found that the 10-residue code encodes proteins that resemble modern proteins in their predicted structural properties. All of the investigated early genetic codes give rise to proteins with enhanced disorder and diminished aggregation propensities. Our results suggest that an ancestral genetic code similar to the proposed 10-residue one is capable of encoding functionally diverse proteins but these might have existed under conditions different from today's common physiological ones. The existence of a protein functional repertoire for the investigated earlier stages which is quite distinct as it is today can be deduced from the presented results.

Keywords: random-sequence proteins, protein evolution, protein intrinsic disorder, protein aggregation, structure prediction, genetic code evolution

Introduction

One of the most difficult problems related to the origin of life on Earth is the emergence and evolution of translation and the genetic code. Although there is considerable progress in understanding the mechanism of extant ribosomes and possible evolution of functional RNA and RNP components (Davidovich 2010, Harish 2012), we have limited knowledge on the structure and function of the earliest proteins. For the origin of the genetic code and the first proteins, a number of scenarios have been put forward and probably even more are possible (see e.g. Szathmáry 1993; Koonin 2009; Strulson 2012). These scenarios differ in the aspect whether proteins emerged with or without closely linked to RNA and whether template-based synthesis evolved after the first proteins emerged or that a primitive form of translation was a prerequisite for the birth of the first functional proteins. There are numerous considerations ranging from the availability of prebiotically synthesized amino acids (Higgs 2009) through some intrinsic structural properties of proteins (Greenwald 2012) to including functional requirements (Houen 1999). However, due to the complexity of

the issue, we argue that a number of aspects should be combined when reconstructing a viable earlier stage of biochemical evolution. An important aspect is that the earliest proteins produced by the first primitive translation machineries are expected to have some selective advantage to the host organism. In this study we aim to estimate the structural preferences and heterogeneity of proteins encoded by three proposed ancient genetic codes. These properties are then compared to each other and similarly derived predictions for the present-day universal genetic code (referred to as STANDARD below). Our study is designed to yield a first-approximation assessment of the potential of the proposed early codes to encode for a diverse set of functional proteins. The three proposed earlier stages of the genetic code investigated are detailed below.

- We have chosen a suggested 10-residue (EARLY10) stage (Higgs 2009) based on the availability of prebiotically synthesised amino acids and on the consideration that later additions should reduce the 'cost' of the code defined on the basis of the errors introduced by the incorporation of suboptimal amino acids in the encoded proteins. To define the cost of a specific codon reassignment, the authors utilize a number of physiochemical properties of the individual amino acids. The initial set of amino acids is taken as those formed most easily under abiotic conditions.
- The second code considered is a 6-residue (EARLY6) stage proposed (Di Giulio 1999; Di Giulio 2008) in line with the coevolution theory based on the assumption that the expansion of the genetic code occurred in parallel with the increasing availability of amino acids as biosynthetic pathways became more and more complex. Thus, the first set of encoded amino acids is defined by their availability by abiotic synthesis. Relationships between biosynthetic pathways and codon assignments find statistical support in the organization of the present code where biosynthetically related amino acids share codon boxes (Wong 2005).
- Third, a highly hypothetical 4-residue (EARLY4) stage (Houen 1999) was selected for comparison. This code was proposed based on the structure of the modern genetic code and on the physicochemical properties of the amino acids. Requirements for the ability to generate folded functional proteins are emphasized in the supporting argumentation. Among the three proposed early codes investigated here, this is the only one to encode for a positively charged amino acid (Arg) as an important requirement for the interaction with nucleic acids.
- Considering that for the first four amino acids Gly, Ala, Glu, and Val were proposed by independent authors (Higgs 2009, Oba 2005), and that short peptides consisting of these amino acids were found to possess catalytic (proteolytic) activity (Oba 2005), we have also generated random proteins with these four amino acids. This will be referred to as the GADV set below. It should be noted that both EARLY6 and EARLY10 codes above encode these four amino acids.

To compare the features of different proteins encoded by different early codes, we used random coding sequences that were translated by each code. This process mimics the flow of information at a stage of biochemical evolution where some form of a translation apparatus is present. Further, our approach acts as a highly simplified model for *de novo* protein generation where previously non-coding nucleic acid segments become translated. In this approach the frequency of the amino acids is determined by the codon tables and the GC-content of the underlying coding sequence and not by any independent *a priori* assumption. Random sequence generation was performed also for the STANDARD

code in order to evaluate the trends without bias that could have been introduced by investigating extant proteins. The use of random sequences imposed also a criterion for the predicted features and the algorithms as no approaches using evolutionary information can be applied in this scenario. It is also important to underline that all our investigations can be interpreted only in a comparative way as we do not expect that any of these predictions yields a reliable absolute estimate of the properties of the non-natural sequences investigated here. Thus, the single initial key assumption of this study is that the differences observed for the random sequence sets reflect the differences between extant and hypothetical extinct proteomes.

Most prediction programs are optimized for predicting structural features at ambient conditions, and we are aware of the fact that predictions usually do not reproduce experimental behavior completely and that the output of different algorithms might not be directly comparable. In line with this, we will always focus on the trends observed for a given property as predicted by the consensus of three methods with different theoretical basis.

The three structural properties predicted and investigated are:

- Intrinsic disorder: intrinsically disordered proteins (IDPs) and disordered segments do not adopt a stable three-dimensional fold, rather, they can be characterized by a high number of rapidly interconverting structural states (Tompa 2012). They are sometimes involved in biochemical tasks not easily performed by well-folded globular proteins (Ferreon 2013). The abundance of IDPs in proteomes was found to correlate with organism 'complexity' (Schád 2011), thus, it is an interesting question whether disorder could be prevalent in early proteomes. As today a number of translation-associated tasks are done by globular domains, our hypothesis is that the earlier codes should have been able to code for such proteins. For example, evolutionary analyses suggest that the most ancient ribosomal proteins have globular domains like the OB fold and SH3-like structure (Harish 2012), although a number of disordered segments contact ribosomal RNA in an extended conformation within the full ribosome (Ban 2000).
- Aggregation propensity: the ability to form amyloid-type aggregates is generally acknowledged to be an intrinsic property of proteins (Dobson 2003; Perczel 2007; Schnabel 2010). Aggregation is widely considered harmful for the cell, there is evolutionary pressure to reduce the risk of aggregation (Reumers 2009; Reumers 2009a; Pastore 2012; Villar-Pique 2012). However, if proteins need to be optimized during evolution to avoid aggregation, the question arises whether newly born ('*de novo*') proteins are viable and whether they pose a potential risk to the organism (Wu 2013). Considering the early stages of protein evolution, the problem might have been more serious for ancient proteins that existed in an environment devoid of today's elaborate cellular mechanisms acting to reduce the risk of aggregation (Monsellier 2007; Reumers 2009, Reumers 2009a; Stefani 2004). Based in this information, we expect that earlier codes are biased towards avoiding to produce proteomes with high aggregation propensities. However, it should be noted that a scenario has been proposed where the first functional proteins adopted amyloid-like aggregated structures (Greenwald 2012).
- Ability to form transmembrane helices: transmembrane (TM) proteins are essential in extant living cells for the organization of material transport and communication through lipid membranes. The transmembrane part can have a beta-barrel structure or can consist of a bundle of alpha-helices. This latter structure is more abundant and there are a number of algorithms available than can predict the presence of TM helices from amino acid sequence based on the presence of specific segments of predominantly hydrophobic residues able to form a

helix of sufficient length to span the membrane.

Our aim in this study is to compare the trends of properties varying with the size and composition of the amino acid alphabet regarding the aforementioned protein proteome properties.

Materials and methods

For this study we have chosen three different ancestral genetic codes at different stages of evolution, referred to as EARLY4, EARLY6 and EARLY10 (Figure S1 in Online Resource 1). Notice that arginine and leucine were encoded in the EARLY4 but not in the EARLY6 or EARLY10 codes. Thus, a chronological sequence of these three hypothetical ancestral codes is incompatible with sequential addition (and no other changes) of amino acids during genetic code evolution.

Nine sets of random DNA sequences with 10% to 90% GC-content incremented by 10% were generated using in-house Perl scripts. The resulting sequence sets are labeled as GC10 to GC90 . For each of them, 10,000 DNA sequences of 480 nucleotides in length were generated by adjusting the base probabilities *p(N)* to the desired values. For simplicity, the probabilities for complementer bases were kept equal (*p(C)=p(G)* and *p(A)=p(T)*). In-frame STOP (and START codons for EARLY4) were avoided. The generated DNA sequences were translated using the standard genetic code and the three proposed earlier ones. Thus, 4x9x10,000 *de novo* protein sequences with the length of 160 residues each were analyzed at this stage. The length of 160 amino acid residues (translated from 480 nucleotides) was chosen as a good estimate for average protein domain size (Lin 2012). In addition, for the GC40, GC50 and GC60 sets, protein sequences with lengths of 40, 80, and 120 residues were generated in order to analyze the possible effect of chain length on the results. The length of 40 residues is assumed to be around the limit where the applied structure predictions can be comparable

To further investigate the effect of amino acid distribution, we have generated additional data sets where the frequency of amino acids was not dictated by the GC-content of the coding sequence and the applied genetic code, but were adjusted directly. Besides random proteins with equal amino acid frequencies (designated Equal), additional protein sequence sets with lengths of 40, 80 120 and 160 residues were generated: one where the frequencies of charged, polar and hydrophobic residues were adjusted to match those in the standard genetic code at 50% GC-content (designated CHPfreq) and another in which the frequencies of the charged amino acids were set to match that in the standard code and the ratio of polar and hydrophobic residues matches that of the proposed early code in question (designated Cfreq). In addition, equal-frequency random proteins with the amino acids Gly, Ala, Asp and Val were generated (designated GADV) (Table S1 in Online Resource 1).

For the two sets of protein sequences obtained by the STANDARD and the EARLY10 code, we have performed a BLAST search against the 'nr' database to assess the similarity of the resulting random *de novo* proteins to known sequences. Sequences obtained with the EARLY6 and EARLY4 codes are not suitable for standard BLAST search because of their low complexity nature.

Data analysis was performed similarly to that described in our previous study (Ángyán 2012). We used sequence-based *in silico* prediction algorithms to estimate structural propensities of these random-sequence polypeptides (Ángyán 2012). An important aspect of the algorithms chosen is that they do not rely on evolutionary information, i.e. do not use alignments with homologous proteins as part of the prediction. This is necessary to avoid any bias as our investigation focuses on hypothetical proteins with no extant relative sequences.

Intrinsic disorder was estimated using the IUPred (Dosztányi 2005), RONN (Yang 2005) and VSL2B (Obradovic 2005) algorithms, transmembrane helix forming propensity using DAS-TMfilter (Cserző 2004), TMHMM (Krogh 2001) and PHOBIUS (Käll 2004) and aggregation propensity using FoldAmyloid (Garbuzynskiy 2010) and the TANGO/WALTZ (Fernandez-Escamilla 2004; Maurer-Stroh 2010) algorithms. None of these predictors use evolutionary information during data processing, thus we expect that they can be used for *de novo* sequences in an unbiased way.

To assess the consistency of the predictions, the segment overlap (SOV) measure (Zelma 1999) was calculated for all predicted properties on selected random protein sequences using the implementation of Balázs Szappanos (Szappanos 2010) We use the SOV(obs) measure which takes into account both positive and negative predictions, i.e. segments predicted to adopt the structural feature in question and also segments that are predicted not to be in that state.

Results

We have analyzed the intrinsic disorder, the tendency to form transmembrane helices and aggregation propensity for protein sequences translated from random DNA segments of varying length and GC-content with three proposed ancient genetic codes (EARLY10, EARLY6 and EARLY4 according to the number of amino acids coded, Figure S1 in Online Resource 1) besides the standard one. Our results obtained for the standard code using highly similar methodology were published before (Ángyán 2012). Besides, we have generated random protein sequences with adjusted amino acid frequencies for all codes and for the GADV residue set. It is important to stress that all our sequences analyzed are hypothetical and that we do not expect that we are able to provide a reliable estimate for the properties investigated for any given sequence.

Sequences in our random sets are distinct from extant proteins

The amino acids encoded by the different genetic codes can be classified based on their propensities for being disordered, aggregation-prone or typical for transmembrane helices. A simple grouping of the residue types based on literature data (Campen 2008; Zhao 2006; Pawar 2005) and considering their presence in each of the genetic codes, some basic considerations can be made. All proposed early codes investigated here show an over-representation of disorder-promoting residues relative to the standard one, whereas amino acids typical of aggregation-prone regions and transmembrane helices are almost completely absent from the EARLY6 and EARLY4 sets (Figure 1, Table S2 in Online Resource 1 ad Figure S2 in Online Resource 2). Thus, it can be expected that the predicted structural properties of the corresponding random proteins will also exhibit characteristic differences.

A BLAST search to identify potential similar sequences resulted in no hits below an E-value of 1^{-10} for sequences translated from GC50 sequences with the standard and the EARLY10 code.

Physicochemical properties of random proteins are distinct for all four genetic codes

As a first approximation, we prepared charge-hydropathy plots (Uversky 2002) for all the sequences generated in this study (Figure S3 in Online Resource 2). The plots show characteristic properties for all of the genetic codes investigated. Only STANDARD and EARLY10 sets exhibit sequences that are in the 'ordered' region of the plot, i.e. with sufficiently high average hydrophobicity and small average net charge. However, the trends according to the GC-content of the underlying coding sequences are different as STANDARD random proteins tend to be more hydrophobic with decreasing GC-content while EARLY10 proteins fall into both the ordered and disordered regions of the plot except for GC90 sequences, as both their charge and hydropathy increases largely in parallel with decreasing GC-content. The trends observed for 160-residue random sequences are closely reproduced by all sets of shorter sequences with higher deviation, for which the most straightforward explanation is the higher expected fluctuation of amino acid content of randomly generated short sequences.

Consistency of predictions varies with alphabet size

It is important to stress that as in this study we investigate random sequences, we do not use any sets of sequences with experimentally determined properties as a standard for assessing the performance of the methods. We used the percentage of residues predicted as disordered, TMH or aggregation-prone to compare trends between sequence sets (Ángyán 2012). However, we evaluate the consistency of the prediction algorithms for selected data sets by calculating the overlaps between segments (Zelma 1999) predicted to be in the particular state in question (disordered, transmembrane, aggregation-prone) and also segments predicted not to be in that state. A high segment overlap (SOV) value indicates that the two prediction algorithms yield highly similar results whereas low SOV values indicate that the two methods produce mutually exclusive results.

On Figure 2, we show SOV results obtained for sequences translated from coding sequences with 50% GC-content for each code investigated. These sequences basically have the same amino acid composition as deducible from the codon tables for each code adjusted for the absence of STOP codons (Figure S1 in Online Resource 2).

For the sequences translated with the standard code from DNA with a GC-content of 50%, disorder predictions show the least general overlap while all transmembrane helix (TMH) predictor algorithms yield practically identical results. Interestingly, the prediction of aggregation-prone segments by WALTZ overlaps almost completely with TMH predictions while FoldAmyloid yields results distinct from all 8 other programs used. The modest overlap between TANGO and WALTZ is expected based on their intentionally complementary nature (Ahmed 2013).

The picture for the GC50 segments translated with the EARLY10 code is similar with the TANGO-WALTZ predictions being closer to each other and that each one of the disorder predictions exhibiting less overlap with the other 8 algorithms.

For GC50 EARLY6 sequences, all 3 disorder predictions yield practically identical results and 5 of the remaining 6 algorithms producing similar output with FoldAmyloid being the outlier.

Sequences translated with the EARLY4 code are treated differently by the mutually consistent RONN-VSL2B algorithms and IUPred. All 3 TMH prediction methods together with TANGO and WALTZ show a high level of agreement with FoldAmyloid yielding again clearly distinct results.

Based on these results, we have chosen to compare results obtained by averaging all 3 prediction outputs for each feature, termed DIS3, TMH3 and AGR3 predictions. However, in all comparisons it is important to evaluate the possible discrepancies between all underlying predictions. We note that the extent of overlaps differ also according to

the GC-content of the underlying coding sequences as it also influences amino acid abundance (see Supplementary Dataset).

Structural preferences of proteomes depend on the GC-content of their underlying coding sequences

For random coding sequences, their GC-content drives the amino acid abundance in the protein sequences translated from them. Thus, it is expected that the predicted structural properties of our polypeptides show identifiable trends according to the alterations in the GC-content of the underlying coding sequences. Such relationships have been demonstrated quantitatively for random sequences obtained with the STANDARD code (Ángyán 2012) and have been observed here for all three proposed ancient genetic codes investigated. We note that, as in our previous study, we investigated sequences with GC-content ranging from 10% to 90% to be able to clearly identify trends despite that only the middle of this regime has broad biological significance.

The tendencies observed for different genetic codes with increasing GC-content show that intrinsic disorder generally increases at higher GC-content whereas aggregation propensity decreases – with noting that there are practically no trends in this regard for EARLY6 proteins as all sequences are predicted to be completely disordered and not prone to aggregation (Figures S4-S6 in Online Resource 2,). However, the trend to form TM helices exhibits a trend parallel to that of aggregation for the standard code but not for any of the proposed earlier ones. Close inspection of individual predictions reveals that this is caused by including the results of DAS-TMfilter in the average. Considering the predictions averaged for the selected pairs of algorithms as described above removes this discrepancy at the cost of predicting only a very low number of transmembrane helices compared to proteins obtained with the standard code ().

For the datasates with adjusted amino acid frequencies (Equal, CHPfreq and Cfreq) the effect of the size and nature of the alphabet is also prevalent. The properties of the random proteins with equal amino acid frequencies are mostly reminiscent to those of the corresponding GC40 datasets (Figure 3). For the EARLY10 and EARLY6 codes, both the CHPfreq and Cfreq sequence sets display properties similar to the GC50 and GC60 datasetes (for the EARLY10 code, the amino acid frequencies of the Cfreq set are identical to the GC50 set). EARLY6 Cfreq and CHPfreq data sets display more variability in disorder occupying more of the ordered regions than EARLY6 proteins with different amino acid frequencies. However, for the EARLY4 code, both CHPfreq and Cfreq random proteins exhibit properties that are largely outside of the regions occupied by standard proteins and which are also distinct from their Equal, GC40, GC50, and GC60 couterparts. It must be noted here that for the EARLY4, CHPfreq, and Cfreq sequences there was a high discrepancy between the outputs of the prediction algorithms resulting in abrupt limits in the averaged properties. VSL2b predicts all sequences composed only these four amino acids to be fully disordered, whereas IUPred predictions span the full range from complete order to disorder except for the CFreq and CHPfreq data sets where they indicate practically no disorder. This observation is in line with our analysis of the consistency of different prediction tools for sequences with a limited alphabet.

The GADV Equal sequences display a yet again different distribution of the predicted properties, with generally higher disorder and lower aggregation tendencies than the STANDARD data sets. Interestingly, the transmembrane helix-forming potential of GADV peptides seems similar to those of STANDARD proteins.

Per-sequence distribution of predicted properties is characteristic of the genetic code The largest variety of structural features is exhibited by proteins with the STANDARD code with all the earlier codes resulting in random proteins with

less complex features. STANDARD proteins span the largest range for all structural properties investigated. EARLY10 proteins, although also spanning the full range of intrinsic disorder from fully ordered to completely disordered, show a characteristically different distribution across the sequences generated with higher intrinsic disorder that for STANDARD proteins (75.01±15.30 vs. 24.01±13.71 for GC50 sequences for DIS3 predictions,). In addition, EARLY10 proteins have a very low number of predicted transmembrane helices and show practically no propensity to aggregate. All EARLY6 proteins are predicted to be almost fully disordered by all the predictors investigated and they show only a limited number of aggregation-prone segments. Only DAS-TMfilter predicts some transmembrane helices for these sequences. The largest discrepancies in the predictions are observed for random EARLY4 proteins. Whereas both VSL2B and RONN predicts them to be completely disordered, IUPred suggests only a limited extent of disorder for sequences translated from low-GC coding segments, which contain almost exclusively Arg and Leu (over 75% in average for GC10-GC30). Similarly, only DAS-TMfilter predicts more TM segments and FoldAmyloid more aggregation-prone regions than for EARLY6 sequences.

Interplay between structural features varies with the genetic code

The structural preferences are not expected to be independent of each other. Figure 3 shows the interdependence of aggregation propensity with disorder and TMH-forming tendency for random sequences translated from GC40-50-60 segments with each of the 4 genetic codes investigated. For STANDARD proteins, there is a clear trend for less disordered proteins to potentially exhibit higher aggregation propensity, which is also observed for EARLY10 and EARLY4 proteins. We note that these trends are much less clear when FoldAmyloid and IUPred predictions are not considered. The predictions for the EARLY4 CHPfreq and Cfreq sets revealed properties distinct from all other sets investigated. For STANDARD and EARLY10 proteins, higher TMH propensity implies higher aggregation potential but not vica versa. Interestingly, this trend is also observable for the Cfreq and CHPfreq EARLY4 data sets but not for any other EARLY set. In general, the low TMH propensity observed for EARLY6 and EARLY4 sequences precludes the identification of clear trends

Discussion

We have generated and analyzed hypothetical protein sequences by translating random DNA segments of varying GC content using different proposed ancestral genetic codes. Our approach corresponds to a first-approximation simulation of a scenario for a protein emerging *de novo* from a previously non-coding DNA segment (Knowles 2009, Guerzoni 2011, Tautz 2011) at a specific stage of the evolution of the genetic code. Naturally, the first translated RNA molecules might have had other functions besides encoding proteins, e.g. we speculate that some of them could be reminiscent of today's tmRNAs (Janssen 2012) etc., thus they almost certainly were non-random. Nevertheless, we assume that our random sequence sets are suitable to investigate the differences between proteins encoded by different early genetic codes investigated here.

It is important to note that the evolutionary stages investigated here are incompatible with each other as the four-residue code includes Arg and Leu that is missing from all other codes but the standard one. Thus, it is highly unlikely that the four codes investigated here represent snapshots of the actual history of the evolving code. Nevertheless, we do not question the relevance of any the proposed ancient codes investigated *a priori*, rather, as a starting point, we accept that all are based on well-reasoned solid considerations and each of them (or stages highly similar to them) might have

actually existed at some stage. We believe that our results and conclusions highlight the importance of considering the expected structural properties of the proteins encoded, thus, help to elucidate relevant aspects of early protein evolution and also the prediction approaches used.

Our results show high variability both in the performance of the prediction algorithms used and for the evaluated genetic codes. Specifically, predictions yield much more detailed results than considering only the described propensities of the individual amino acids (Figure 1) and different algorithms might yield different results for the very same sequence, the discrepancy also depending on the set of residues used. Thus, we separate our first part of the discussion according to two different assumptions.

Assumption 1: our general structural predictions are valid

Here we assume that our averaged structure predictions are valid in the sense that 1) they reproduce the trends for proteins translated with a given genetic code according to varying GC-content of the coding sequence and 2) they reproduce the differences between protein sets obtained with different genetic codes. We stress again that the we do not use the absolute propensities predicted for any given random sequence for drawing conclusions Moreover, one has to be well aware of the fact that the performance of prediction algorithms varies and the same sequence motif might be predicted to have different preferences by specific algorithms, which might or might not reflect the ability of the given segment to interconvert between structural states (Szappanos 2010). Also, the common interpretation of predictions, also used here, that segments can be assigned distinct, mutually exclusive states is most likely not generally valid because of the dynamic nature of proteins (Andersen 2002).

Nevertheless, accepting the above assumptions leads to the conclusion that all three hypothetical earlier genetic codes investigated here determine proteins with properties markedly different from extant ones. Only EARLY10 proteins show comparable trends and variability to existing ones in terms of the properties investigated. Even EARLY10 proteins show higher disorder tendency and lower aggregation propensity than STANDARD ones, and this trend is even more marked for EARLY6 and EARLY4 proteins (). Thus, we either accept that there was a stage in protein evolution were protein disorder was prevalent (regardless of which of the EARLY6 and EARLY4 codes actually existed), or we question the relevance of the proposed 6- and 4-residue codes as they are almost fully incapable of producing globular and transmembrane proteins, which is in apparent contradiction with the emergence of membrane-surrounded living entities. However, this picture is more complicated by the observation that EARLY4, Cfreq, and CHFreq sets show a higher TMH abundance even than the STANDARD data sets. These contradictory results give no clear clue whether a compartmentalization process precluded the appearance of a primitive translation machinery (Maynard-Smith 1995) or coding for transmembrane proteins is a late addition in the evolution of the genetic code.

We note that the relevance of intrinsic disorder in early stages of biochemical evolution is supported by a growing number of observations that the ability to perform enzymatic catalysis is not unique to well-folded globular enzymes. Proteins in a molten globule state, as well as partially or fully disordered polypeptides have been shown to catalyze reactions with an efficiency sometimes comparable to that of globular enzymes (Vendruscolo 2010, Schulenburg 2013). Thus, despite the fact that today protein disorder seems to be associated with higher organism complexity, assumptions about the prevalence of IDPs in early proteins can not be immediately rejected on the basis of their limited biochemical functionality.

Assumption 2: At least some of the proposed early genetic codes actually existed

In this frame we accept that at least the EARLY10 code represents an actual stage of the evolution of the protein translation machinery. If so, we can conclude that EARLY10 proteins exhibited a lower propensity to aggregate than STANDARD proteins, thus, aggregation was a lower risk for *de novo* protein evolution and gradual optimization to avoid aggregation could have coevolved with the extension of the genetic code.

Assuming that at least one of the earlier codes actually existed and requiring functional diversity comparable to extant proteins, we might call into question the validity of predictions used. The sometimes substantial discrepancies between different predictions already proves the difficulty of yielding reliable estimates of such properties even if the underlying principles have a solid physicochemical/biochemical basis. Moreover, predictions are inherently optimized to perform well for extant proteins, i.e. nonrandom sequences potentially containing all 20 amino acids encoded in the standard code as their training / parameter optimization was based on the properties of such segments. Thus, it might well be that there are some inherent systematical errors in our predictions when considering proteins with limited amino acid alphabets.

Moreover, we – as most prediction algorithms by default – assumed ambient conditions which might not have been valid for earlier stages of the evolution of the genetic code. A recent study suggested that EARLY10 proteins might have been optimized for a high-salt environment (Longo 2013). It has also been suggested that current algorithms might overestimate disorder content for proteins in extremophiles (Pancsa R, Kovacs D, Bhomwick P, Tompa P, personal communication). . It is also documented that ion concentrations have substantial effects on the aggregation propensities of proteins (Baussay 2004; Song 2013), thus, the actual risk of aggregation for ancestral proteins, especially the EARLY10 and EARLY6 sets might be substantially higher in a high-salt environment than predicted here.

Considering the charge-hydrophobicity plots for EARLY6 and EARLY4 proteins, we have to assume an even larger difference to present-day ambient conditions for that stage of molecular evolution (Figure S3 in Online Resource 2). Thus, the structural properties, especially the folded-unfolded equilibrium might have been quite different for such proteins than inferred from predictions optimized for extant proteins. Nevertheless, these considerations still make it unlikely that proteins with the proposed EARLY4 and EARLY10 compositions could have provided sophisticated interactions with lipid membranes similar to extant ones. Another question is the possible emergence of globular enzymes responsible for the biosynthesis of more and more complex amino acids, as these could only consist of amino acids already available. It is also unlikely that a highly disordered proteome could have been a direct ancestor of modern ones, especially that today intrinsic disorder is prevalent in organisms regarded as having 'higher complexity'. Thus, if stages of the genetic code like those investigated here indeed existed, the exact nature and role of peptides and proteins encoded by them as well as the conditions under these existed and provided selective benefits to the organisms remain elusive. Again, if translation was not the only source of proteins at these stages, the problems outline here might have been not entirely relevant.

Conclusion: relevance of predictions and genetic codes

Investigations of protein sequences lacking the diversity of known wild-type proteins have shown that foldable and functional proteins can be constructed from simplified random amino acid alphabets (Watters 2004) or sequence-independent peptides (Milner-White 2011), providing additional support for the theory of genetic code evolution by expansion. Our results show that the structural diversity for proteins encoded by the EARLY10 code approaches that of

extant proteins, and this might have been more prevalent under conditions different from recent physiological ones. However, none of the proposed earlier codes has this property and, if existed, they likely exhibited significantly limited functionality compared to extant proteins in an environment supposedly quite different from physiological conditions today. However, accepting the RNA world as a preceding state, disordered proteins even with narrower functional arsenal could have been a useful functional addition to the living systems either as separate molecules or fused to RNA (Szathmáry 1993). For example, RNA-bound small hydrophobic peptides might have added the ability to interact with membranes to the molecules of the RNA world. Finally, a highly hypothetical scenario to reconcile the high disorder content of early translated proteins with structural and functional requirements can be that a different (e.g. abiotic) source of proteins existed in parallel with the earliest translation apparatus. This scenario might also be expanded as instead of stepwise addition to a limited set, a rich source of prebiotic amino acids could have been available for protein synthesis and the current code reflects an optimized and narrowed state of the original pool. Such a hypothesis can still be compatible with all the considerations about the optimization of the genetic code when it reached a functional state comparable to its extant one. A similar scenario might also partially decouple biosynthetic problems from the evolution of the code and is largely compatible with Carl Woese's early suggestion on a preliminary ambiguous code which evolved through ambiguity reduction (Woese 1995).

One hypothetical early environment, which could also be relevant for the EARLY10 stage of genetic code evolution, might have been one with high salt concentrations. This is supported by a recent experimental study where mutations were introduced to resemble the amino acid distribution of the EARLY10 code better and the result was a halophilic protein. We note that to achieve this, the authors have increased also the Arg content of the protein, although neither Arg nor any other positively charged amino acid is encoded by the EARLY10 code. The charged single alpha-helix (CSAH) structural motif that is also stable under a wide range of salt concentrations exhibits a comparable abundance of negatively and positively charged amino acids (Süveges 2009, Gáspári 2012). In a wider perspective which we will not discuss further here, the requirement for the presence of Arg in primordial proteins is a matter of debate centered on the nucleic acid-binding capabilities of such proteins (McDonald 2010).

In summary, we propose that a genetic code similar to the EARLY10 one could have actually existed and give rise to proteins with a relatively wide range of structural variability. In a transition process from that stage to the present one which most likely included an increase of the abundance of hydrophobic residues, a gradual optimization of aggregation propensity, structural stability and function could have taken place. For the proposed earlier stages, however, either a protein/peptide functional repertoire or an environment different from today's is most likely required for a viable hypothesis, or even both of these.

Acknowledgments

Financial support of the Hungarian Scientific Research Fund (OTKA 104198) and TÁMOP-4.2.1.B-11/2/KMR-2011-0002 is acknowledged. The authors thank Dr. Péter Tompa and Dr. András Perczel for helpful discussions.

Conflicts of interest

The authors declare that they have no conflict of interest.

References

Ahmed A, Kajava AV (2013) Breaking the amyloidogenicity code: methods to predict amyloids from amino acid sequence. FEBS Lett 587:1089–1095

Andersen CA, Palmer AG, Brunak S, Rost B (2002) Continuum secondary structure captures protein flexibility. Structure 10:175-184

Ángyán AF, Perczel A, Gáspári Z (2012) Estimating intrinsic structural preferences of de novo emerging random-sequence proteins: Is aggregation the main bottleneck? FEBS Lett 586:2468–2472

Ban N, Nissen P, Hansen J, Moore PB, Steitz T A (2000) The complete atomic structure of the large ribosomal subunit at 2.4 Å resolution. Science 289: 905-920

Baussay K, Bon CL, Nicolai T, Durand D, Busnel JP (2004) Influence of the ionic strength on the heat-induced aggregation of the globular protein β-lactoglobulin at pH 7. Int J Biol Macromol 34:21-28

Campen A, Williams RM, Brown CJ, Meng J, Uversky VN, Dunker AK (2008) TOP-IDP-scale: a new amino acid scale measuring propensity for intrinsic disorder. Protein Peptide Lett 15:956-963

Cserző M, Eisenhaber F, Eisenhaber B, Simon I (2004) TM or not TM: transmembrane protein prediction with low false positive rate using DAS-TMfilter. Bioinformatics 20:136-137

Davidovich C, Belousoff M, Wekselman I, Shapira T, Krupkin M, Zimmerman E, Bashan A, Yonath A (2010) The proto-ribosome: an ancient nano-machine for peptide bond formation. Isr J Chem 50:29-35

Di Giulio M (2008) An extension of the coevolution theory of the origin of the genetic code. Biology Direct 3:37

Di Giulio M, Medugno M (1999) Physicochemical optimization in the genetic code origin as the number of codified amino acids increases. J Mol Evol 49:1-10

Dobson CM (2003) Protein folding and misfolding. Nature 426:884-890

Dosztányi Z, Csizmók V, Tompa P, Simon I (2005) IUPred: web server for the prediction of intrinsically unstructured regions of proteins based on estimated energy content. Bioinformatics 21:3433-3434

Fernandez-Escamilla AM, Rousseau F, Schymkowitz J, Serrano L (2004) Prediction of sequence-dependent and mutational effects on the aggregation of peptides and proteins. Nat Biotechnol 22:1302-1306

Ferreon ACM, Ferreon JC, Wright PE, Deniz AA (2013) Modulation of allostery by protein intrinsic disorder. Nature 498:390-394

Garbuzynskiy, SO, Lobanov MY, Galzitskaya OV (2010) FoldAmyloid. A method of prediction of amyloidogenic regions from protein sequence. Bioinformatics 26:326-33

Gáspári, Z, Süveges, D, Perczel, A, Nyitray, L, Tóth, G (2012) Charged single alpha-helices in proteomes revealed by a consensus prediction approach. Biochem. Biophys. Acta - Proteins and Proteomics 1824:637-646

Greenwald J, Riek R (2012) On the possible amyloid origin of protein folds. J Mol Biol 421:417-426

Guerzoni D, McLysaght A (2011) De novo origins of human genes. PLoS Genet.7:e1002381

Harish A, Caetano-Anollés G (2012) Ribosomal history reveals origins of modern protein synthesis. PLoS ONE 7:e32776

Higgs PG (2009) A four-column theory for the origin of the genetic code: tracing the evolutionary pathways that gave rise to an optimized code. Biology Direct 4:16

Houen G (1999) Evolution of the genetic code: the nonsense, antisense, and antinonsense codes make no sense. BioSystems 54:39-46

Janssen, BD, Hayes, CS (2012) The tmRNA ribosome rescue system. Adv Prot Chem Struct Biol 86:151-191

Käll L, Krogh A, Sonnhammer EL (2004) A combined transmembrane topology and signal peptide prediction method. J Mol Biol 338:1027-1036

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

Koonin, EV, Novozhilov AS (2009) Origin and evolution of the genetic code: the universal enigma. IUBMB Life 61:99-111

Krogh A, Larsson B, von Heijne G, Sonnhammer ELL (2001) Predicting transmembrane protein topology with a hiddn Markov model: Application to complete genomes. J Mol Biol 305:567-58

Lin MM, Zewail AH (2012) Hydrophobic forces and the length limit of foldable protein domains. PNAS 109:9851-9856

Longo LM, Lee J, Blaber M (2013) Simplified protein design biased for prebiotic amino acids yields a foldable, halophilic protein. PNAS 110:2135-2139

Maurer-Stroh S, Debulpaep M, Kuemmerer N, Lopez de la Paz M, Martins IC, Reumers J, Morris KL, Copland A, Serpell L, Serrano L, Schymkowitz JW, Rousseau F (2010) Exploring the sequence determinants of amyloid structure using position-specific scoring matrices. Nat Methods 7:237-42

Maynard-Smith J, Szathmáry E (1995) The major transitions in evolution. Oxford University Press, Oxford, UK.

McDonald GD, Storrie-Lombardi MC (2010) Biochemical constraints in a protobiotic earth devoid of basic amino acids: the "BAA (-) World". Astrobiology 10:989-1000

Milner-White EJ, Russell, MJ (2011) Functional capabilities of the earliest peptides and the emergence of life. Genes 2:671-688

Monsellier E, Chiti F (2007) Prevention of amyloid-like aggregation as a driving force of protein evolution. EMBO Rep 8:737-742

Oba, T, Fukushima J, Maruyama M, Iwamoto R, Ikehara K (2005) Catalytic activities of [GADV]-peptides. Origins of Life and Evolution of Biospheres 35: 447-460

Obradovic Z, Peng K, Vucetic S, Radivojac P, Dunker AK (2005) Exploiting Heterogeneous Sequence Properties Improves Prediction of Protein Disorder. Proteins 61:176-182

Pastore A, Temussi PA (2012) The two faces of Janus: functional interactions and protein aggregation. Curr Opin Struct Biol 22:30-37

Pawar AP, Dubay KF, Zurdo J, Chiti F, Vendruscolo M, Dobson CM (2005) Prediction of "aggregation-prone" and "aggregation-susceptible" regions in proteins associated with neurodegenerative diseases. J Mol Biol 350:379-392

Perczel A, Hudáky P, Pálfi VK (2007) Dead-end street of protein folding: thermodynamic rationale of amyloid fibril formation. JACS 129:14959-14965

Reumers J, Maurer-Stroh S, Schymkowitz J, Rousseau F (2009) Protein sequences encode safeguards against aggregation. Hum Mutat 30:431-437

Reumers J, Rousseau F, Schymkowitz J (2009) Multiple evolutionary mechanisms reduce protein aggregation. Open Biol 2:176-184

Schád E, Tompa P, Hegyi H (2011) The relationship between proteome size, structural disorder and organism complexity. Genome Biol 12:R120

Schnabel J (2010) Protein folding: the dark side of proteins. Nature 464:828-829

Schulenburg C, Hilvert D (2013) Protein Conformational Disorder and Enzyme Catalysis. In: Dynamics in Enzyme Catalysis . Springer Berlin Heidelberg, pp. 41-67

Song J (2013) Why do proteins aggregate?"Intrinsically insoluble proteins" and "dark mediators" revealed by studies on "insoluble proteins" solubilized in pure water. F1000Research 2

Stefani M (2004) Protein misfolding and aggregation: new examples in medicine and biology of the dark side of the protein world. BBA-Mol Bas Dis1739:5-25

Strulson CA, Molden RC, Keating CD, Bevilacqua PC (2012) RNA catalysis through compartmentalization. Nat Chem 4:941-946

Süveges D, Gáspári Z, Tóth G, Nyitray L (2009) Charged single α-helix: A versatile protein structural motif. Proteins 74:905-916

Szappanos B, Süveges D, Nyitray L, Perczel A, Gáspári Z (2010) Folded-unfolded cross-predictions and protein evolution: the case study of coiled-coils. FEBS Lett 584:1623-1627

Szathmáry E (1993) Coding coenzyme handles: a hypothesis for the origin of the genetic code. PNAS 90:9916-9920

Tautz D, Domazet-Lošo T (2011) The evolutionary origin of orphan genes. Nat Rev Gen 12:692-702

Tompa P (2012) Intrinsically disordered proteins: a 10-year recap. Trends Biochem Sci 37:509-516

Uversky VN (2002) Natively unfolded proteins: a point where biology waits for physics. Protein Sci 11:739-756

Vendruscolo M (2010) Enzymatic activity in disordered states of proteins. Curr Opin Chem Biol 14:671-675

Villar-Pique A, Ventura S (2012) Protein Aggregation Acts as Strong Constraint During Evolution. In: Evolutionary Biology: Mechanisms and Trends. Springer Berlin, Heidelberg, pp 103-120

Watters AL, Baker D (2004) Searching for folded proteins in vitro and in silico. Eur J Biochem 271:1615-1622

Wong J (2005) Coevolution theory of the genetic code at age thirty. BioEssays 27:416-425

Woese CR (1965) On the evolution of the genetic code. PNAS 54:1546

Wu DD, Zhang YP (2013) Evolution and Function of De Novo Originated Genes. Mol Phylogenet Evol 67:541-545

Yang ZR, Thomson R, McNeil P, Esnouf RM (2005) RONN: the bio-basis function neural network technique applied to the detection of natively disordered regions in proteins. Bioinformatics 21:3369-3376

Zemla A, Venclovas Č, Fidelis K, Rost B (1999) A modified definition of Sov, a segment-based measure for protein secondary structure prediction assessment. Proteins 34:220-223

Zhao G, London E (2006) An amino acid "transmembrane tendency" scale that approaches the theoretical limit to accuracy for prediction of transmembrane helices: relationship to biological hydrophobicity. Protein Sci 15:1987-2001

FIGURE LEGENDS

Figure 1. Amino acids in the four genetic codes investigated as classified according their disorder (DIS +/-), transmembrane helix forming (TMH+/-) and aggregation propensity (AGR+/-). (See also Figure S2 in Online Resource 2).

Figure 2. Segment overlaps (SOVs) of prediction outputs for 160-residue long GC50 sequences translated with each of the four genetic codes investigated. Overlaps for three selected algorithms with reference to all nine used are shown in each panel. The SOV measure is unity for identical predictions but is not symmetric between different algorithms.

Figure 3. Distribution of predicted aggregation propensity as a function of disorder for all data sets investigated. Sequences with amino acid distribution based on identical considerations and with matching length are shown in each panel. Each data point corresponds to the predicted properties of a single sequence. Note that data points are overlapping, labels point to regions where the coloring is clearly visible. Numerical values of the overlaps of the property areas can be found in Online Resource 3..

FIGURES

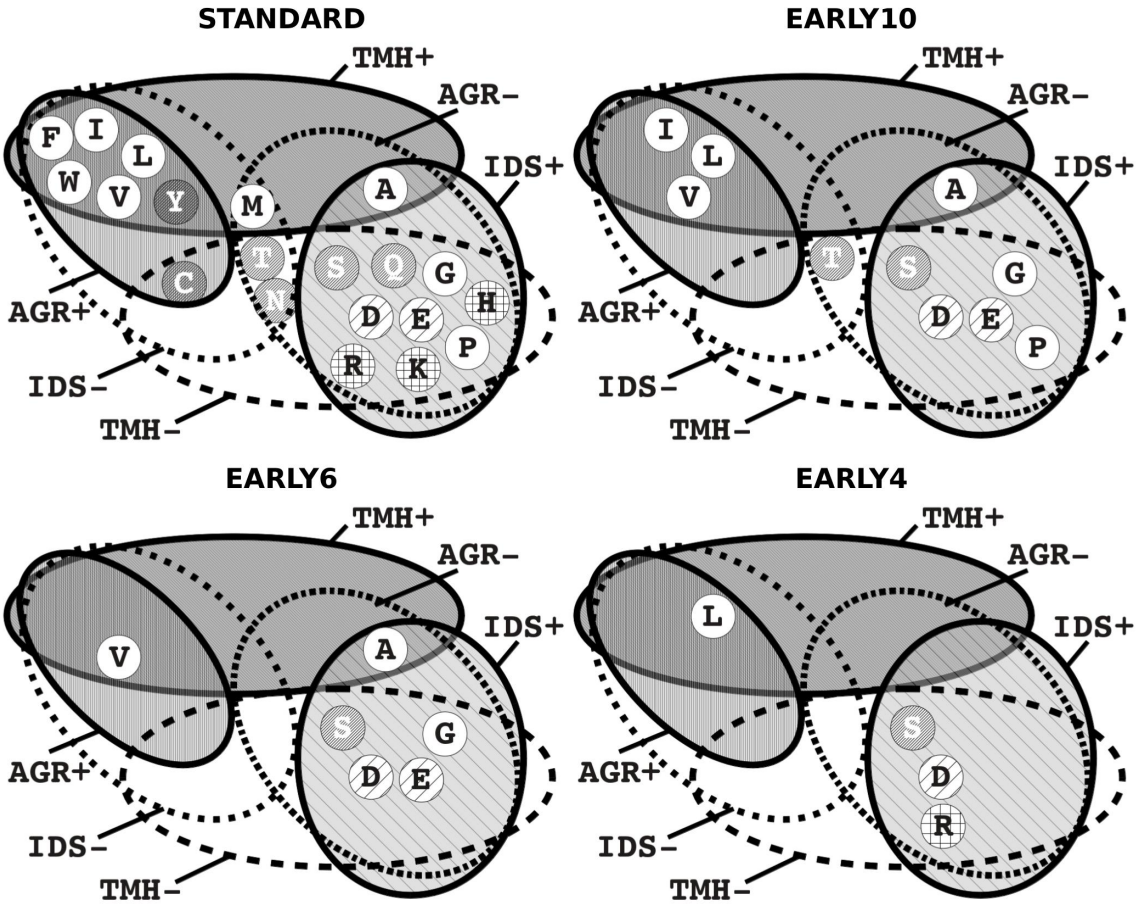

Figure 1.

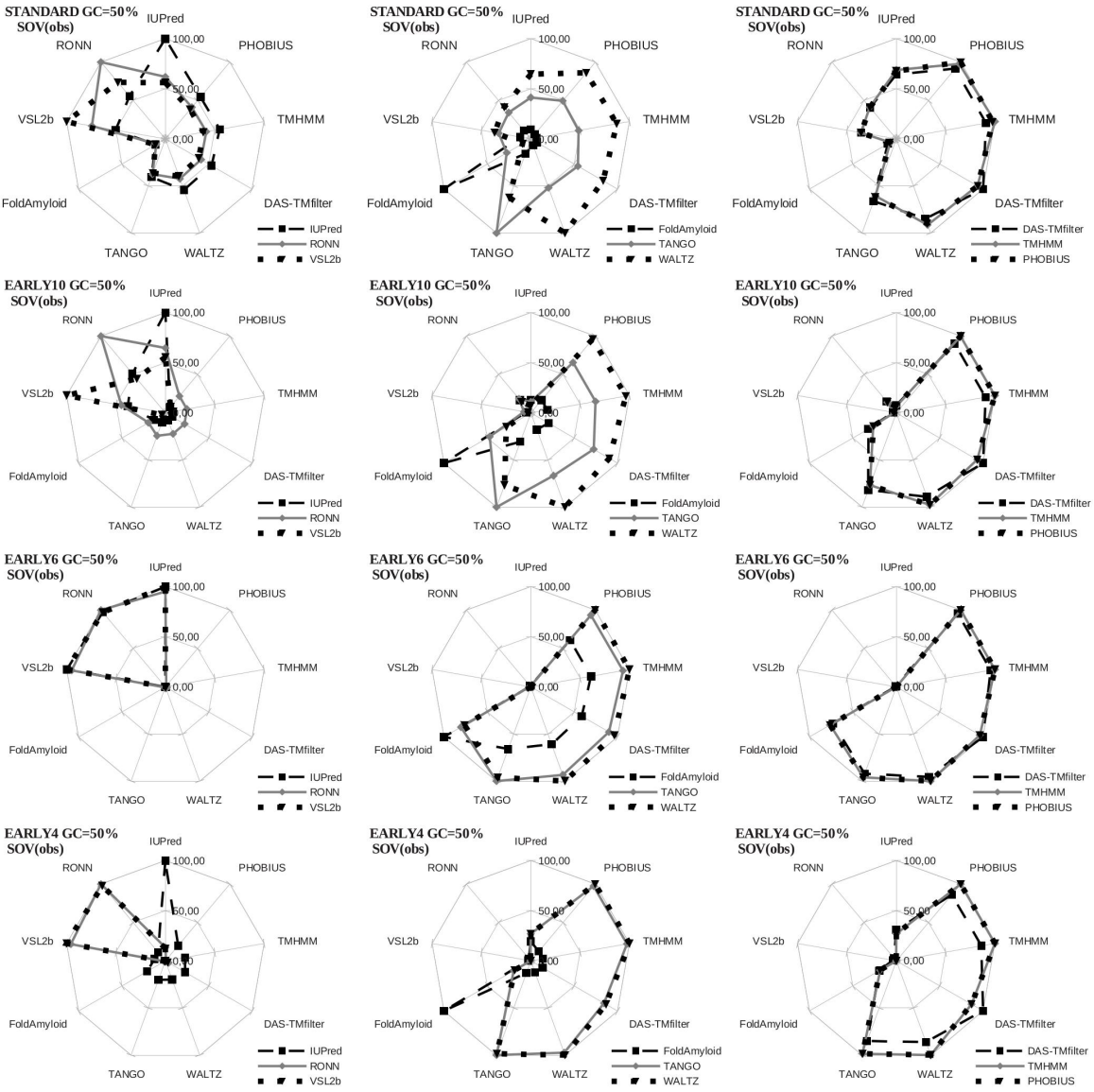
STANDARD GC=50% SOV(obs)
EARLY10 GC=50% SOV(obs)
EARLY6 GC=50% SOV(obs)
EARLY4 GC=50% SOV(obs)
IUPred
PHOBIUS
TMHMM
DAS-TMfilter
WALTZ
TANGO
FoldAmyloid
VSL2b
RONN
100,00
50,00
0,00
IUPred RONN VSL2b
FoldAmyloid TANGO WALTZ
DAS-TMfilter TMHMM PHOBIUS

Figure 2.

Preprint version of Ángyán et al., J. Mol. Evol. 78(5):263-274 (2014)
The final publication is available at Springer via http://dx.doi.org/10.1007/s00239-014-9622-3

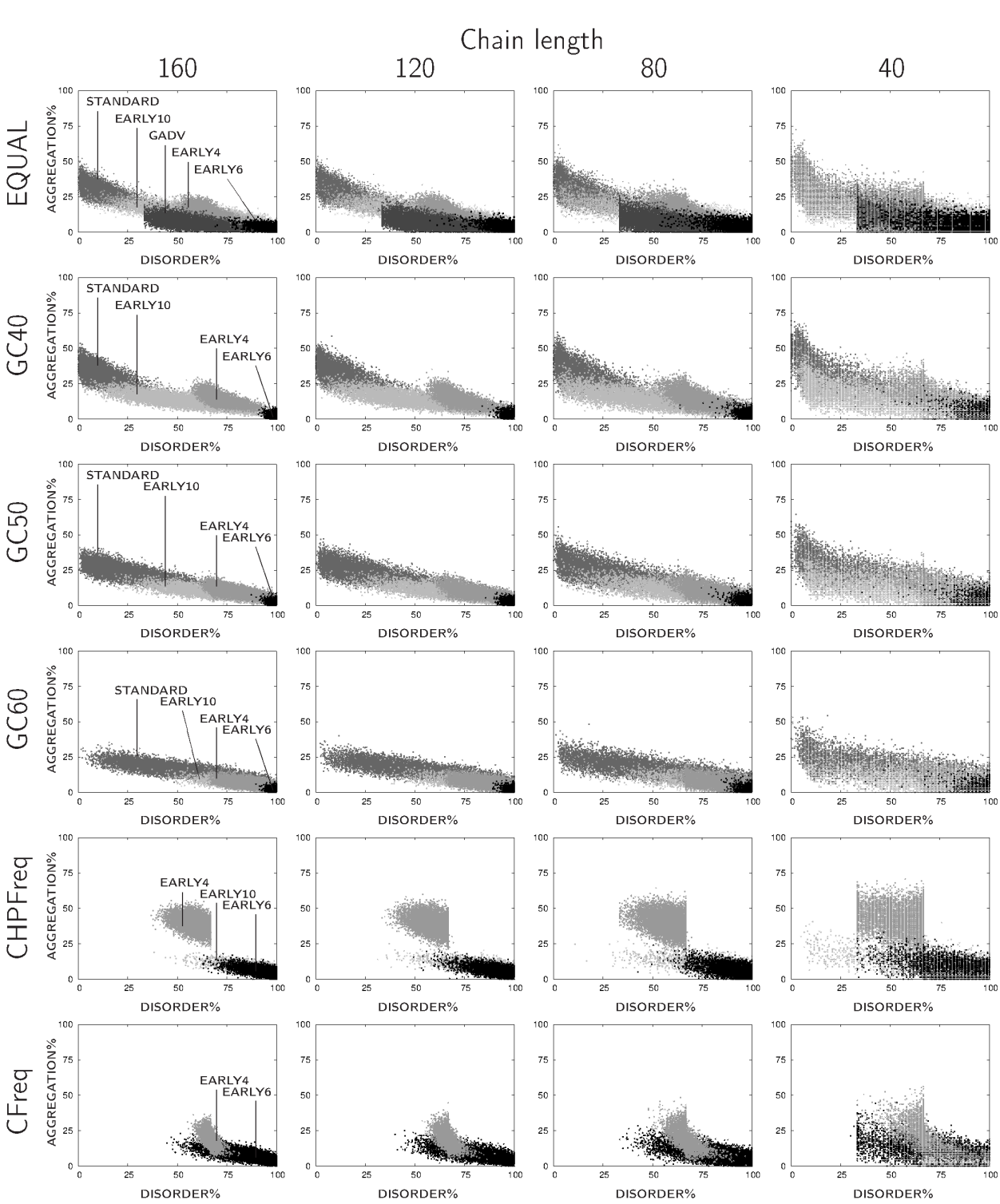

Figure 3.