Preprint of: *FEBS Letters* 586: 2468-72, http://dx.doi.org/10.1016/j.febslet.2012.06.007

Estimating intrinsic structural preferences of *de novo* emerging random-sequence proteins: is aggregation the main bottleneck?

Annamária F. Ángyán[1], András Perczel[1,2], Zoltán Gáspári[3,*]

[1] Eötvös Loránd University, Institute of Chemistry, Pázmány Péter s. 1/A, H-1117 Budapest, Hungary
[2] HAS-ELU Protein Modelling Group, Pázmány Péter s. 1/A, H-1117 Budapest, Hungary
[3] Pázmány Péter Catholic University, Faculty of Information Technology, Práter u. 50/A,
H-1083 Budapest, Hungary

[*]Corresponding author:
Zoltán Gáspári
Pázmány Péter Catholic University, Faculty of Information Technology, Práter u. 50/A, H-1083
Budapest, Hungary
Tel: +36-1-886-4780
E-mail: gaspari.zoltan@itk.ppke.hu

Classification: Biological sciences / Biophysics and Computational Biology

ABSTRACT

Present-day proteins are believed to have evolved features to reduce the risk of aggregation. However, proteins can emerge *de novo* by translation of non-coding DNA segments. In this study we assess the aggregation, disorder and transmembrane propensity of protein sequences generated by translating random nucleotide sequences of varying GC-content. Potential *de novo* random-sequence proteins translated from regions with GC content between 40-60% do not show stronger aggregation propensity than existing ones and exhibit similar tendency to be disordered. We suggest that *de novo* emerging proteins do not mean an unavoidable aggregation threat to evolving organisms.

KEYWORDS

Protein evolution, *de novo* proteins, aggregation propensity, random sequences, structure prediction

INTRODUCTION

The emerging consensus on protein aggregation is that it is an inherent property of any polypeptide chain and, regardless of their amino acid sequences, the amyloid fibril might be the most favored thermodynamic state of all proteins [1, 2, 3]. Even so, proteins display sequence-specific aggregation propensities that can be estimated by *in silico* methods [4, 5]. Thus, proteins can evolve to reduce the risk of aggregation and detailed studies of selected proteins revealed a number of such mechanisms [6]. However, proteins continuously emerge *de novo* by transcription and translation of previously non-coding DNA segments [7, 8, 9]. This poses the question whether novel proteins that did not yet have the chance to reduce their aggregation load by selection can seriously hinder molecular evolution: if the aggregation propensity of *de novo* proteins is generally high, leading to the aggregation of practically all *de novo* polypeptides, that might render the chances of the emergence of such proteins negligible.

De novo origin of coding sequences from non-coding ones is a rare but not improbable event, there are e.g. human- and primate-specific proteins thought to have arisen by this mechanism [9, 10, 11]. The overall low-level transcriptional activity of the human genome provides a plausible basis for such events [12]. Thus, the aggregation propensity of such proteins is worth to be explored. As the number of known genuine *de novo* proteins is fairly low, in this study we chose to use an *in silico* study on random, translated DNA sequences to i) have a dataset of sufficient size to observe trends, ii) assess the aggregation propensity before any - however short-time - selection could take place at the protein level and iii) have a standardized way to assess and compare trends for sequences with different GC-content that can be used as a benchmark for real *de novo* proteins.

MATERIALS AND METHODS

Detailed description of all the methods used and datasets can be found in the online supplementary material. Random DNA sequences of varying GC-content were generated with the restriction that in-frame STOP codons were avoided. Translated protein sequences were subjected to different algorithms (**Table S1**) to assess their tendency for aggregation (TANGO [13], WALTZ [14] and FoldAmyloid [15, 16]), forming disordered (IUPred [17 ,18], RONN [19] and VSL2B [20, 21]) or transmembrane structures (HMMTOP [22], DASTMfilter [23] and TMHMM [24]). The number of residues predicted to be in the given structural classes by the algorithms were averaged and used as a consensus prediction. The same algorithms were applied to a number of databases representing folded (ASTRAL40, version 1.75), unfolded (DISPROT, version 5.7), aggregation-prone (AmyPDB, last update on 7th April, 2008) and transmembrane proteins (PDBTM, version 2.3) as well as the complete human and mouse proteomes (from Uniprot release 2011_05). The obtained one- and two-dimensional distributions at the three properties (disorder, aggregation and transmembrane tendency defined as the percentage of residues falling to these categories in the consensus prediction) were compared by the appropriate variants of the Kolmogorov-Smirnov test. In addition, the area spanned by the sequences in the two- and three-dimensional plots and the overlap between the distributions obtained for different databases were estimated using a grid-based approach. The coding sequences of human orphan proteins were obtained by comparing the translated mRNA sequences to the available protein sequences and extracting the nucleotide sequences in the matching region.

RESULTS

Random sequences and predictions of structural features

We generated 10,000 random DNA sequences of 480 nucleotides without in-frame STOP codons for each of GC-content regime from 10% to 90% using steps of 10%. The 160-residue length of the translated polypeptides can be regarded as a reasonable estimate of average domain size in proteins [25, 26]. Although not the full GC-range explored is biological relevance, as for example the human genome has an average GC-content of 41% and ranges approximately from 20% to 60% [27], we chose our systematic scan to identify general trends. After translating all of the 9x10,000 nucleotide sequences, we have used BLAST [28] search to assess the similarity of the resulting random *de novo* proteins to known sequences. No hits were found below an E-value of $1*10^{-10}$, and only 30 hits were found below an E-value of 0.001 (**Table S2**). Thus, our random sequence set is sufficiently distinct from extant proteins. Next, we used a set of prediction algorithms to assess their aggregation loads (TANGO [13], WALTZ [14] and FoldAmyloid [15, 16]), their disorder (IUPred [17, 18], RONN [19] and VSL2B [20, 21]) and transmembrane propensities (HMMTOP [22], DASTMfilter [23]and TMHMM [24]). None of the applied methods uses evolutionary information during data processing like today's best-performing secondary structure prediction tools [29]), thus, we expect that they can be used for *de novo* sequences in an unbiased way. We have performed the same predictions on several databases representing folded, disordered, transmembrane and aggregation-prone proteins as well as the complete human and mouse proteomes. It is important to stress that we do not wish to assess the absolute aggregation propensity of any of the sequence sets, rather, in all evaluations below, we analyze trends and draw conclusions from comparisons of predictions made with the same toolkit.

General trends

Naturally, the amino-acid composition of our random datasets reflects the standard genetic code organization. At low GC-content, hydrophobic amino acids appear with higher frequency, typically representing 50-70% of all residues. At 90% GC-content, only 10% of all residues are hydrophobic and 20% is arginine **(Table S3)**. In present-day proteomes, acidic amino acids (Glu, Asp) are remarkably more frequent than expected from the codon distribution in the standard genetic code [30 ,31] (**Table S4**). At high GC-content, basic amino acids are overrepresented in the standard code-translated dataset relative to present-day natural proteins. The mean net charge of random *de novo* sequences exhibits a minimum at 40% GC-content and it is still higher than the highest value obtained for present-day proteins, corresponding to IDPs. The mean hydrophobicity shows a decreasing trend with increasing GC-content and covers a wider range than that of present-day proteins **(Fig. S1 and S2)**.

According to the averaged structural predictions, the GC-content of the underlying DNA sequences governs the structural preferences of the random proteins with clearly identifiable trends that are much more pronounced than the variations in the simple physico-chemical parameters. Intrinsic disorder is a dominant feature of sequences with coding regions of high GC-content (**Table 1**). Around 50% GC-content, 25% of all amino acid residues is predicted to be disordered. In this respect, only aggregation-prone and transmembrane present-day proteins have a lower average value. At 60% GC-content and above, random sequences are practically fully disordered containing on average one or two long disordered regions **(Fig. 1a)**.

The propensity to form transmembrane helices is relatively high at low GC-content and decreases rapidly to practically vanish over 60% GC-content. At 40% GC-content, the average ratio of residues in transmembrane segments is comparable to those in the complete human and mouse proteomes **(Fig 1b)**.

The aggregation load in random sequences is highest at low GC-content and drops quickly to an average 5% of all residues at 50% GC-content. At and above 60% GC, practically all parameters investigated are on average below those of present-day proteins **(Fig 1c)**.

Interplay between structural properties

We have investigated whether the predicted structural preferences are independent of each other or there are some associations. We have investigated this aspect at the sequence level, calculating correlations between the percentage of residues predicted to be disordered, transmembrane and aggregation-prone **(Tables 2 and S5)**. Both these values and two-dimensional plots of these features indicate that these features are loosely interdependent and not all regions of the disorder-transmembrane-aggregation space are accessible either for random or for existing protein sequences **(Figure 1d, 1e, 1f)**.

Sequences with higher percentage of disordered residues tend to have less transmembrane helices and lower aggregation propensity. However, the nature of the interdependence is different, with a large range of variation in transmembrane propensity at low disorder tendency, whereas aggregation load seems to be more strictly negatively associated with disorder. On the other hand, the tendency to form transmembrane helices shows a positive association with aggregation propensity. These trends suggest that amino acid composition plays a decisive role in defining these structural features.

Comparison to databases

We stress that we do wish to assess the absolute propensity of any sequence set to be disordered, form transmembrane helices and being prone to aggregation, rather use consensus predictions for comparative purposes. Our results allow to compare the trends observed for random-sequence proteins to those observed in natural ones. Below, unless noted otherwise, we will focus on random proteins translated from the physiologically most relevant range of GC-content, between 40 and 60%.

Intrinsic disorder depends heavily on the underlying GC-content of the random sequences, at 60% the random sequences show clearly higher disorder propensity than even DISPROT, whereas at 40 and 50% of the translated proteins are predicted to contain less disordered residues than those in extant proteomes **(Table 1)**.

The tendency to form transmembrane helices is much lower for random sequences translated from DNA of 50% or higher GC content than for extant proteins except globular and disordered ones **(Table 1)**.

Interestingly, the highest aggregation potential can be attributed to transmembrane proteins in PDBTM and not aggregation-prone proteins in AmyPDB which do not show higher aggregation propensity than globular proteins (ASTRAL40) or those in the human and mouse proteomes.

For all three properties investigated, standard deviation for existing proteins is higher than for random ones, corresponding to higher variability in selected, functional proteins than in potential *de novo* ones.

To further elaborate and compare the properties of random *de novo* and extant proteins, we plotted the investigated structural preferences in two dimensions for each sequence and compared the resulting two-dimensional distributions and the area covered by the sequences in the disorder-transmembrane-aggregation space (**Fig S3, S4 and S5)**.

Statistical tests (one- and two-dimensional Kolmogorov-Smirnov tests) reveal that the distributions obtained for the disorder, transmembrane and aggregation tendencies of the human proteome and the proteins translated from random DNA segments with 40-50-60% GC-content are totally dissimilar with a P-value of 0 (**Table 2**). This is due to the different local densities of the data points in the investigated data sets. However, when estimating the (2D or 3D) space covered by the data points corresponding to the random sequence set above, it is apparent that more than 95% of this space overlaps with that spanned by proteins in the human proteome. In contrast, the overlap is only around 35% when calculated relative to the human proteome **(Tables S6, S7, S8)**. The non-overlapping part of the space covered by the random sequences corresponds to low aggregation propensity.

It should be noted that the most striking difference between the random protein sets and the human proteome is in their tendency to form transmembrane helices, as natural sequences exhibit higher propensity for this than those translated from random DNA segments.

De novo *proteins in the human genome*

We have investigated three *de novo* human proteins [8, 10] using the same methodology as for random sequences. Interestingly, these are in accordance with the trends observed for random *de novo* proteins with respect to the dependence of structural features on the GC-content of the underlying DNA segment. The DNAH10OS (P0CZ25) and C22ORF45 (P86434) proteins are predicted to have disordered stretches and the GC-content of their coding segments is around 60% for the coding segment (**Table 3**). In contrast, CLLU1 (Q5K131) is predicted to have a significant aggregation tendency, as expected for a protein translated from a low-GC DNA segment. More detailed prediction results can be found in the online Supplementary Material.

DISCUSSION

In this *in silico* study we generated and analyzed random-sequence hypothetical *de novo* proteins from DNA with varying GC-content. This method differs from generally applied ones for investigating random protein sequences where amino acid frequencies are set *a priori*, whereas in our approach these are dictated by the GC-content of the underlying coding segments and the (standard) genetic code. Although the connection between genomic GC-percentage and the amino acid composition of the translated proteins is a finding neither novel nor surprising, our approach yields a solid basis and benchmark to estimate the structural properties of newly emerging proteins. Besides, it corresponds to a realistic scenario shown to be operative for a few proteins even recently in the human lineage. We were able to identify trends in the structural properties of potential *de novo* proteins as a function of the features of the hypothetical genomic sequences translated. We have shown that increasing GC-content implies higher tendency to form disordered segments and lower transmembrane and aggregation potential for the translated sequences.

Our main finding is that the random proteins translated from DNA with 40-60% GC occupy a region in the space of the properties considered that is almost entirely within the span of those of extant proteins in the human proteome. Random *de novo* proteins are not expected to have a larger aggregation potential than existing ones, nor display a higher degree of disorder. However, they clearly display a lower propensity to form transmembrane helices, meaning that from the three properties investigated, this is the one that most likely needs the most serious optimization during further evolution. This finding is also in line with the notion that probably transmembrane helices represent the most regular type of structural elements investigated here, with requirements on length and composition etc., thus these are the least expected to arise by chance in random sequences. The situation is similar to that observed for coiled coils with underlying specific repeats and disordered segments [32].

It should be stressed that our study corresponds to a first approximation of the problem and can rather be viewed as a benchmark study than an accurate model of real evolutionary processes. Genomic sequences are non-random, and real proteins can display bias relative to the expected properties based on the GC-content of the underlying sequences. For example, the amino acid composition of proteins in the PDBTM database shows the highest similarity to our random proteins translated from DNA with GC-content of 70-80% **(Table S3, S4)**. However, these proteins have higher aggregation propensity and transmembrane tendency than other extant proteins, contrary to the trends observed for random proteins.

Our finding that *de novo* proteins are not particularly prone to aggregation might appear contradictory to claims that proteins are optimized against aggregation during evolution. However, our methods addressing three basic structural properties do not reveal any detailed structural, let alone functional features. On the other hand, we feel that as the prediction programs used here consider sequence only and no evolutionary relationships, their results are suitable for comparing the features of extant and hypothetical proteins. In particular, the fact that the presence of structured parts can influence aggregation properties of proteins is taken onto account by disorder predictions, approximating globularity with the inverse of disorder. It is expected that after the birth of a *de novo* protein it is optimized by selection to perform its function and to adjust its structure, stability and dynamics. During this process the maintenance or even lowering of the aggregation potential present in the newly born protein is one of the pressures operative during evolution. It can even be expected that initially requirements for the presence of a suitable hydrophobic core or transmembrane helices can even render these proteins more prone to aggregation, and the described mechanism to lessen this load can be operative after the suitable structure and stability are reached. So long as the benefits of the new protein outweigh its hazards for the organism, especially if its expression level is low, such scenarios can be plausible. Our results do not contradict the presence and nature of selection pressures present at any later stages of protein evolution, but they suggest that the appearance of novel coding sequences is not expected to be hampered by unusually high aggregation propensity of the translated proteins.

ACKNOWLEDGMENTS

This research was supported by he Hungarian Scientific Research Fund (OTKA NK101072). The European Union and the European Social Fund provided financial support for the project under grant agreements no. TÁMOP 4.2.1./B-09/1/KMR-2010-0003 and TÁMOP 4.2.1/B-11/2/KMR-2011-0004. The authors thank Zsuzsanna Dosztányi and the VIB for providing standalone versions of the programs IUPred and TANGO/WALTZ, respectively.

APPENDIX A

Supplementary data associated with this article can be found in the online version.

REFERENCES

1 Dobson, C.M. (2003) Protein folding and misfolding. Nature 426, 884-890

2 Baldwin, A.J., Knowles, T.P., Tartaglia, G.G., Fitzpatrick, A.W., Devlin, G.L., Shammas, S.L., Waudby, C.A., Mossuto, M.F., Meehan, S., Gras, S.L., Christodoulou, J., Anthony-Cahill, S.J., Barker, P.D., Vendruscolo, M., Dobson, C.M. (2011) Metastability of native proteins and the phenomenon of amyloid formation. JACS 133, 14160-14163

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

4 Caflish, A. (2006) Computational models for the prediction of polypeptide aggregation propensity. Curr Op Chem Biol. 10, 437-444

5 Reumers, J., Rousseau, F., Schymkowitz, J. (2009) Multiple evolutionary mechanisms reduce protein aggreation. The Open Biology Journal 2, 176-184

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

7 Toll-Riera, M., Castelo, R., Bellora, N., Albà, M.M. (2009) Evolution of primate orphan proteins. Biochem Soc Trans 37, 778-782

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

9 Li, C-Y, Zhang Y, Wang Z, Zhang Y, Cao C, et al. (2010) A human-Specific *De Novo* Protein-Coding Gene Associated with Human Brain Functions. PLoS Comput Biol 6(3): e1000734.

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

11 Toll-Riera, M., Bosch, N., Bellora, N., Castelo, R., Armengol, L., Estivill, X., Albà, M.M. (2009) Origin of primate orphan genes: a comparative genomics approach. Mol Biol Evol 26, 603-612

12 ENCODE Project Consortium (2007) Identification and analysis of functional elements in 1% of the human genome by the ENCODE pilot project. Nature 447, 799-816

13 Fernandez-Escamilla, A.M., Rousseau, F., Schymkowitz, J., Serrano, L. (2004) Prediction of sequence-dependent and mutational effects on the aggregation of peptides and proteins. Nat Biotechnol. 22(10):1302-1306

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

15 Galzitskaya, O.V., Garbuzynskiy, S.O., Lobanov, M.Y. (2006) Prediction of amyloidogenic and disordered regions in protein chains. PloS Comput Biol 2, e177

16 Garbuzynskiy, S.O., Lobanov, M.Y., Galzitskaya, O.V. (2010) FoldAmyloid. A method of prediction of amyloidogenic regions from protein sequence. Bioinformatics 26, 326-332

17 Dosztányi, Z., Csizmók, V., Tompa, P., Simon, I. (2005) The pairwise energy content estimated from amino-acid composition discriminates between folded and intrinsically unstructured proteins. J Mol Biol 347, 827-839

18 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

19 Yang, Z.R., Thomson, R., McNeil, P., Esnouf, R.M. (2005) RONN: the bio-basis function neural network technique applied to the detection of natively disordered regions in proteins. Bioinformatics

21, 3369-3376

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

21 Peng, K., Radivojac, P., Vucetic, S., Dunker, A.K., Obradovic, Z. (2006) Length-Dependent Prediction of Protein Intrinsic Disorder, BMC Bioinformatics 7, 208

22 Tusnády, G.E., Simon, I. (2001) The HMMTOP transmembrane topology server. Bioinformatics 17, 849-850

23 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

24 Krogh, A., Larsson,B., von Heijne, G., Sonnhammer, E.L.L. (2001) Predicting transmembrane protein topology with a hidden Markov model: Application to complete genomes. Journal of Molecular Biology, 305, 567-580

25 Shen, M.-Y, Davis, F.P., Sali, A. (2005) The optimal size of a globular protein domain: A simple sphere-packing model. Chem Phys Letters 405, 224-228

26 Gáspári, Z., Vlahovicek, K., Pongor, S. (2005) Efficient recognition of folds in protein 3D structures by the improved PRIDE algorithm. Bioinformatics 21, 3322-3323.

27 International Human Genome Sequencing Consortium (2001) Initial sequencing and analysis of the human genome. Nature 409, 860-921

28 Altschul, S. F., Gish, W., Miller, W., Myers, E. W., Lipman, D. J. (1990) Basic local alignment search tool. J Mol Biol 215, 403-410

29 Pirovano, W., Heringa, J. (2010) Secondary Structure Prediction Data Mining Techniques for the Life. Sciences Methods in Molecular Biology 609, 327-348

30 Bogatyreva, N.S., Finkelstein, A.V., Galtziskaya, O.V. (2006) Trend of amino acid composition of proteins of different taxa. J Bioinform Comput Biol 4, 597-608

31 Lobanov, M.Y., Galzitskaya, O.V, (2012) Occurrence of disordered patterns and homorepeats in eukaryotic and bacterial proteomes. Mol BioSyst 8, 327-337.

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

Figure legend

Fig. 1. Comparison of predicted properties of the human proteome with the 40, 50 and 60% GC-based random protein sets.

Table 1. Summary of averaged prediction results on random sequences and selected databases

Values refer to the percentage of residues predicted to be in the structural state investigated

database		**Disorder**				**Transmembrane**				**Aggregation**			
	No. of sequences			**percentile**				**percentile**				**percentile**	
		Avg ± stdev		**25%**	**75%**	**Avg ± stdev**		**25%**	**75%**	**Avg ± stdev**		**25%**	**75%**
AmyPDB	247	33.62	23.62	44.90	47.29	4.20	6.74	0.00	6.73	19.18	6.55	9.29	23.87
ASTRAL40	10175	16.26	13.45	7.10	20.77	1.15	5.45	0.00	0.00	21.26	5.49	18.09	24.33
DISPROT	529	43.89	28.22	21.11	64.48	2.61	6.03	0.00	2.08	16.51	6.80	11.72	21.07
HUMAN	20899	34.74	24.14	14.99	51.11	6.10	11.91	0.00	5.00	21.05	8.41	15.25	25.07
MOUSE	18525	33.31	23.95	13.69	49.56	7.03	13.03	0.00	6.10	21.69	8.88	15.60	25.69
PDBTM	429	12.59	9.65	5.75	16.88	31.86	20.11	12.00	48.02	34.47	10.79	25.61	41.97
GC=10%	10000	1.10	0.95	0.42	1.46	42.47	9.53	36.04	48.96	54.75	4.48	51.88	57.81
GC=20%	10000	2.08	1.78	1.04	2.71	29.85	10.48	22.92	37.08	47.95	5.36	44.38	51.56
GC=30%	10000	4.51	3.69	2.08	6.04	16.83	10.91	8.75	24.58	39.26	5.64	35.31	43.12
GC=40%	10000	10.26	7.32	4.79	13.96	6.98	8.08	0.00	12.08	30.29	5.25	26.56	33.75
GC=50%	10000	24.01	13.71	13.75	31.67	2.65	4.68	0.00	3.96	22.31	4.58	19.06	25.31
GC=60%	10000	50.57	19.25	36.04	64.58	1.01	2.63	0.00	0.00	15.50	3.81	12.81	17.81
GC=70%	10000	81.48	14.45	73.07	92.92	0.24	1.06	0.00	0.00	9.72	3.05	7.50	11.60
GC=80%	10000	96.68	4.69	95.42	100.00	0.09	0.52	0.00	0.00	5.18	2.23	3.75	6.56
GC=90%	10000	99.77	0.79	100.00	100.00	0.22	0.77	0.00	0.00	1.84	1.31	0.94	2.50

Table 2. Correlation between structural properties for selected data sets

Properties correlated (number of residues predicted to be in the states below)	Random-nucleotide translated				Proteome data	
	40% GC	50% GC	60% GC	40-50-60% GC combined	Human	Mouse
disorder-transmembrane	-0.205	-0.193	-0.248	-0.375	-0.350	-0.380
disorder-aggregation	-0.572	-0.692	-0.771	-0.834	-0.780	-0.781
transmembrane-aggregation	0.587	0.433	0.331	0.588	0.742	0.780

All correlations are significant at the 0.05 level

Table 3. Predictions of structural features and GC-content of three recently identified *de novo* human orphan genes [8].The GC-content was calculated for the whole mRNA segment (%GC mRNA) and for the protein-coding RNA segment (%GC exon).

UniProt ID	length	%disorder	%transmembrane	%aggregation	%GC mRNA	%GC exon
P0CZ25	163	92.95	0	9.51	54.92	63.60
P86434	159	50.31	0	15.41	58.89	59.12
Q5K131	121	7.02	10.95	41.74	37.11	31.96

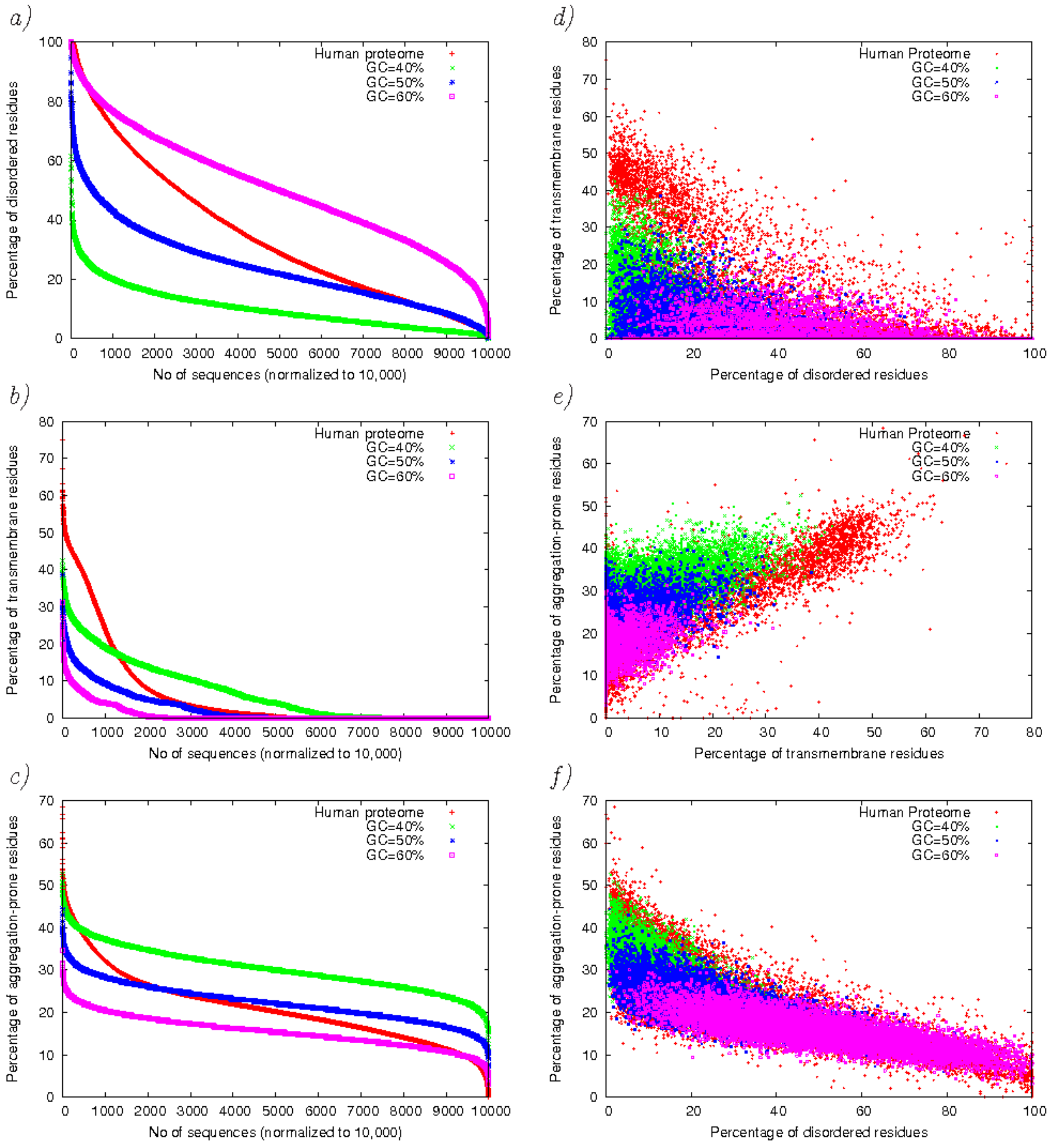
a)
Percentage of disordered residues
No of sequences (normalized to 10,000)
0
20
40
60
80
100
0
1000
2000
3000
4000
5000
6000
7000
8000
9000
10000
Human proteome
GC=40%
GC=50%
GC=60%
b)
Percentage of transmembrane residues
No of sequences (normalized to 10,000)
0
10
20
30
40
50
60
70
80
Human proteome
GC=40%
GC=50%
GC=60%
c)
Percentage of aggregation-prone residues
No of sequences (normalized to 10,000)
0
10
20
30
40
50
60
70
Human proteome
GC=40%
GC=50%
GC=60%
d)
Percentage of transmembrane residues
Percentage of disordered residues
0
20
40
60
80
100
Human Proteome
GC=40%
GC=50%
GC=60%
e)
Percentage of aggregation-prone residues
Percentage of transmembrane residues
0
10
20
30
40
50
60
70
80
Human Proteome
GC=40%
GC=50%
GC=60%
f)
Percentage of aggregation-prone residues
Percentage of disordered residues
Human Proteome
GC=40%
GC=50%
GC=60%

Figure 1.