Journal of proteome research

Article

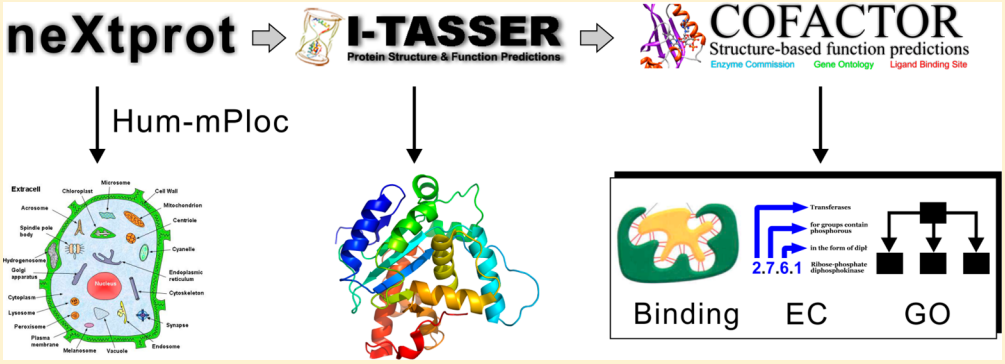
pubs.acs.org/jpr

Structural Bioinformatics Inspection of neXtProt PE5 Proteins in the Human Proteome

Qiwen Dong,[†,‡] Rajasree Menon,[†] Gilbert S. Omenn,*[,†,§] and Yang Zhang*[,†,∥]

[†]Department of Computational Medicine and Bioinformatics, [§]Departments of Internal Medicine and Human Genetics and School of Public Health, [∥]Department of Biological Chemistry, University of Michigan, Ann Arbor, Michigan 48109-2218, United States

[‡]School of Computer Science, Fudan University, Shanghai, 204433, China



ABSTRACT: One goal of the Human Proteome Project is to identify at least one protein product for each of the ~20 000 human protein-coding genes. As of October 2014, however, there are 3564 genes (18%) that have no or insufficient evidence of protein existence (PE), as curated by neXtProt; these comprise 2647 PE2−4 missing proteins and 616 PE5 dubious protein entries. We conducted a systematic examination of the 616 PE5 protein entries using cutting-edge protein structure and function modeling methods. Compared to a random sample of high-confidence PE1 proteins, the putative PE5 proteins were found to be over-represented in the membrane and cell surface proteins and peptides fold families. Detailed functional analyses show that most PE5 proteins, if expressed, would belong to transporters and receptors localized in the plasma membrane compartment. The results suggest that experimental difficulty in identifying membrane-bound proteins and peptides could have precluded their detection in mass spectrometry and that special enrichment techniques with improved sensitivity for membrane proteins could be important for the characterization of the PE5 "dark matter" of the human proteome. Finally, we identify 66 high scoring PE5 protein entries and find that six of them were reported in recent mass spectrometry databases; an illustrative annotation of these six is provided. This work illustrates a new approach to examine the potential folding and function of the dubious proteins comprising PE5, which we will next apply to the far larger group of missing proteins comprising PE2−4.

KEYWORDS: *Human Proteome Project, missing proteins, neXtprot, PeptideAtlas, protein folding, I-TASSER, COFACTOR, structure-based function annotation*

INTRODUCTION

Proteins are the workhorse molecules of life, participating in essentially every activity of various cellular processes. The near-completion of the Human Genome Sequence Project[1] generated a valuable blueprint of all of the genes encoding the amino acid sequences of the entire set of human proteins, providing an important first step toward interpreting their biological and cellular roles in the human body. However, due to the dynamic range and complexity of proteins and their isoforms as well as the sensitivity limits of current proteomics techniques, many predicted proteins have not yet been detected in proteomics experimental data.[2]

In 2011, the Human Proteome Organization (HUPO) launched the Human Proteome Project (HPP),[3] which includes the Chromosome-Centric HPP (C-HPP)[4] and Biology/Disease-Driven HPP (B/DHPP).[5] A major goal of the HPP is to identify at least one representative protein product and as many post-translational modifications, splice variant isoforms, and non-synonymous SNP variants as feasible for each human gene. This ambitious goal is being pursued through 50 international consortia for each of the 24 chromosomes, the mitochondria, and many organs, biofluids, and diseases.[2] Five extensive data resources contribute the baseline and annually updated metrics for the HPP:[2,6] the Ensembl database[7] and neXtProt[8] provide the number of predicted protein-coding genes (a total of 20 055 in neXtProt 2014-09-19); PeptideAtlas[9] and GPMdb[10] independently reanalyze, using standardized pipelines, a vast array of mass

Special Issue: The Chromosome-Centric Human Proteome Project 2015

Received: June 3, 2015
Published: July 21, 2015

ACS Publications © 2015 American Chemical Society

DOI: 10.1021/acs.jproteome.5b00516
J. Proteome Res. 2015, 14, 3750−3761

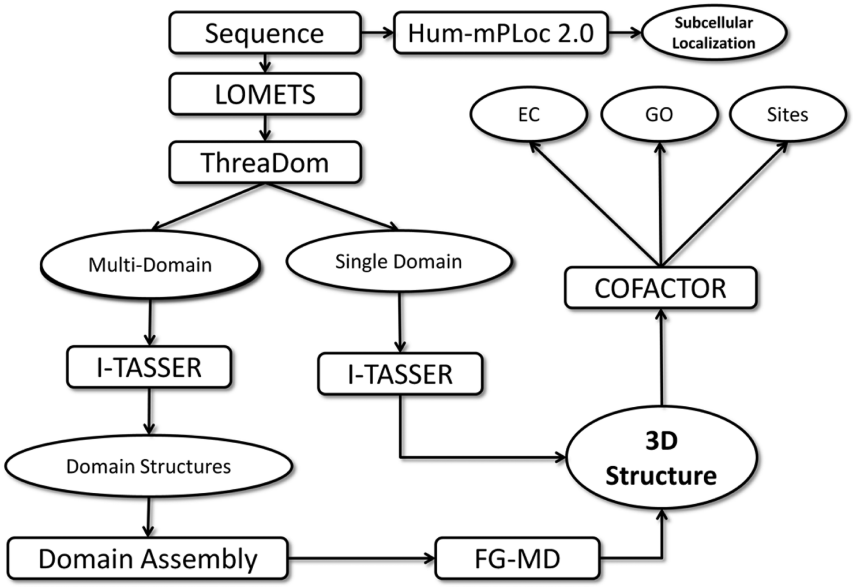


Figure 1. Flowchart of structure and function prediction for PE5 missing proteins.

spectrometry studies; the Human Protein Atlas[11,12] uses a huge antibody library to map the expression of proteins by tissue, cell, and subcellular location; and, finally, neXtProt[8] curates protein existence (PE) evidence and assigns one of five levels of confidence (PE1−5). Proteins at the PE1 level (16 491) have highly credible evidence of protein existence identified by mass spectrometry, immunohistochemistry, 3D structure, and/or amino acid sequencing. At the PE2 level (2647), there is evidence of transcript expression but not of protein expression. PE3 protein sequences (214) lack protein or transcript evidence in humans, but they have homologous proteins reported in other species. Proteins at the PE4 level (87) are hypothesized from gene models. Together, protein entries designated PE2−4 represent missing proteins in the HPP.[6] Finally, the predicted protein sequences at PE5 (616) have dubious or uncertain evidence; a small number of these seemed to have some protein-level evidence in the past, but curation has since deemed such identifications doubtful, primarily because of genomic information, such as lack of promoters or multiple mutations. Each year, a small number are nominated for re-evaluation in light of additional experimental data.

Since 2011, the proteomics community and the HPP have achieved steady progress in human proteome annotations. Now, 85% of putative human protein-coding genes have high-confidence PE1 protein existence, as curated by neXtProt.[6] The remaining 2948 genes at levels PE2−4 have no or insufficient evidence of identification by any experimental methods and are thus termed missing proteins.[6] Many of these missing proteins are presumed to be hard to detect because of low abundance, poor solubility, or indistinguishable peptide sequences within protein families, even in tissues in which transcript expression is detected. The HPP has begun a complementary process of closely examining the missing proteins to recognize those genes that are very unlikely to generate proteins at all or proteins detectable by current methods. PE5 protein entries are considered to be dubious proteins due to their lack of essential features for transcription and/or mutations of the sequence in the numerous cases of pseudogenes. At the HUPO2013 World Congress in Yokohama, it was decided to remove the PE5 entries from the denominator of protein-coding genes, but the community was invited to propose PE5 proteins that might have substantial new evidence or newly predicted features that might make them candidates for active protein expression.

To help address that challenge, we conducted a systematic bioinformatics inspection of the 616 PE5 predicted proteins by evaluating their potential for folding and generating biological functions using the cutting-edge structure folding and structure-based function prediction tools, I-TASSER[13,14] and COFACTOR.[15,16] One reason that we focused on PE5 proteins is that the PE5 sequences represent the most dubious set of missing proteins. Therefore, evidence of protein-coding genes from PE5 proteins will help to highlight their importance as the other categories, PE2−4 proteins (to which the next step of our analysis will be applied), are revisited. In addition, a critical study of these proteins from multiple approaches, including both proteomics and bioinformatics, is becoming increasingly urgent before these genes are removed from the coding-gene denominator. This study will help to demonstrate the analysis of PE5 proteins and lay the foundation for similar analysis of the much larger set of PE2−4 protein entries.

Since the default I-TASSER folding simulation uses fragments from the Protein Data Bank (PDB), the results of which can be easily contaminated by the existence of homologous proteins, we have exploited a stringent filter (sequence identity > 25% or PSI-BLAST *E*-value < 0.5) to exclude all homologous proteins from the template structure library. In fact, PE5 genes have homology with few entries in current structure and function databases from our threading search results (this holds even for many PE5 pseudogenes since we found that most pseudogenes do not have homology in the PDB library); therefore, the exclusion of homologous templates did not result in observable differences in the I-TASSER folding results. In this context, the results of folding simulations are more sensitive to the physical components of the I-TASSER force field that is used to justify the foldability of the sequences than they are to the existence of homologous templates.

It is important to recognize that there are many pseudogenes in DNA that have lost their protein-coding ability due to the accumulation of multiple mutations. However, these genes often

DOI: 10.1021/acs.jproteome.5b00516
J. Proteome Res. 2015, 14, 3750−3761

have a very similar sequence to that of their original functional protein ancestors, which makes it difficult to use sequence homology-based bioinformatics approaches (like BLAST) to distinguish the pseudogenes. An advantage of the combined I-TASSER and COFACTOR procedure over traditional sequence-based homologous approaches is that the I-TASSER folding results are less dependent on homologous proteins after homologous templates are excluded. Moreover, the follow-up COFACTOR algorithm conducts functional annotations based on a function library derived from canonical protein products, assisted with composite examinations from biochemical feature matching and physics-based fitting calculations, including steric testing and ligand-docking scores. This functional analysis ensures further discrimination of distantly related pseudogenes, which face no functional selection during the accumulation of random mutations. These pseudogenes usually do not satisfy the stringent requirements for biological functions, such as subtle binding pockets and functional sites with appropriate physicochemical characteristics.

All of the I-TASSER modeling and COFACTOR annotation results for the PE5 proteins are made publicly available at http://zhanglab.ccmb.med.umich.edu/HPSF/. We expect that the availability of these high-resolution and structure-based annotations from bioinformatics approaches will provide useful insights complementary to other proteome investigations and will help to guide further experimental designs for the characterization of dubious and missing proteins.

EXPERIMENTAL SECTION

Computational modeling of protein sequences in this study consists of three general steps: threading and domain parsing, structure folding simulation, and structure-based function annotations (Figure 1).

First, the query sequence is threaded through a nonredundant set of PDB structures by LOMETS, which is designed to detect possible structural template and super secondary structure fragments using nine state-of-the-art threading algorithms.[17] To avoid homologous contaminants, all homologous proteins that have a sequence identity >25% or are detectable by PSI-BLAST with an *E*-value < 0.5 were excluded from the LOMETS template library. Starting from the multiple threading alignments, the query sequence is parsed into individual domains by ThreaDom,[18] which decides the domain boundary and linker regions of the query sequence based on the conservation and gap and insertion scores in the multiple template alignments.

For each domain, I-TASSER is used to conduct the folding simulations by reassembling the continuous structure fragments excised from the continuous threading alignments through replica-exchange Monte Carlo simulations, under the guidance of a highly optimized knowledge-based force field.[13,14] For proteins with multiple domains, the quaternary structure is constructed by docking the models of the individual domains based on the full-length I-TASSER models, followed by fragment-guided molecular dynamics (FG-MD) refinement.[19] I-TASSER has been recognized as being one of the most robust methods for nonhomologous protein structure prediction in the community-wide CASP experiments.[20−22] The confidence of the folding simulations is evaluated by the C-score,[23] which is calculated by combining the significance score of the threading alignment and the extent of the convergence of the Monte Carlo simulations. C-score is normally in the range of [−5,2], with a C-score > −1.5 indicating confident models with correct fold according the former large-scale benchmark test experiment,[23] where a Pearson correlation coefficient = 0.91 was found between the C-score and the actual accuracy of the I-TASSER models. In a recent computational protein design folding experiment, it was found that the I-TASSER C-score is also highly correlated with the likelihood of the computationally designed sequences folding in the physiological environment.[24]

Starting from the I-TASSER models, the enzyme commission (EC), gene ontology (GO), and ligand-binding site functional annotations are generated using COFACTOR.[15,16] The COFACTOR algorithm has been designed to derive functional insights by global and local (binding-pockets and active sites) structure comparisons of the target with known proteins in the BioLip function library.[25] The functional insights are then translated from known proteins to the target sequences according to a scoring function that combines the structural and evolutionary matches between the target and template proteins. For ligand-binding and enzyme commission assignments, the scoring function of the COFACTOR annotations also combines a chemical feature match and physical fit of the ligand and cofactors with the putative binding/active sites on the I-TASSER structure models. COFACTOR was ranked as the most sensitive algorithm for ligand-binding recognition in the recent CASP experiment.[26]

Finally, the subcellular localizations of the query proteins are predicted by the widely used Hum-PLoc2.0 software,[27] which derives protein locations through the clustering of gene ontology annotations. Hum-PLoc2.0 can generate predictions for 14 subcellular locations (centriole, cytoplasm, cytoskeleton, endoplasmic reticulum, endosome, extracellular, Golgi apparatus, lysosome, microsome, mitochondrion, nucleus, peroxisome, plasma membrane, and synapse) and has a success rate of 70% in large-scale jackknife cross-validation tests.[27]

RESULTS AND DISCUSSION

Data Sets

The dubious or uncertain missing proteins comprising confidence code PE5 were extracted from the neXtProt database[8] of 19 September 2014. There are 616 predicted proteins in this category, with lengths ranging from 21 to 2252 residues. As a control study, we collected all of the high-confidence PE1 proteins from neXtProt for which a structure is solved in the PDB library. A random list of 616 proteins was then chosen that has a distribution of lengths that is similar to that for the PE5 proteins.

Benchmark Test of Structure and Function Predictions on Control Proteins in PE1

As part of the effort to test the I-TASSER and COFACTOR scoring function, as well as to establish a control set for the PE5 proteins, we first conducted structure and function modeling simulations on the 616 highly confident PE1 proteins selected from neXtProt. The structural accuracy of the I-TASSER models can be measured by their TM-score[28] in comparison to that of the known experimental structures. The TM-score has a range of [0,1]; a TM-score > 0.5 generally corresponds to structural similarity in the same SCOP/COTH fold family.[29] Although no homologous templates from the PDB library were employed, 515 of the 616 PE1 proteins have been correctly folded by I-TASSER, with an average TM-score = 0.78. The I-TASSER simulations generally refined the threading templates closer to the native structure. If we account for the best templates from threading from which the I-TASSER simulations start, then there are only 285 targets that have a TM-score > 0.5 and the average TM-score

DOI: 10.1021/acs.jproteome.5b00516
J. Proteome Res. 2015, 14, 3750−3761

= 0.69. Such a significant increase in the folding rate and TM-score of the I-TASSER models from the threading templates is mainly attributed to the highly optimized I-TASSER force field, which has the capacity to reassemble unrelated fragments into a correct global fold.[30]

Here, we have employed the TM-score to assess the accuracy of the modeling using PE1 proteins for which an experimental structure has been solved. For PE5 proteins, however, none of the sequences has an experimental structure available, so we will use the confidence score (C-score) of the I-TASSER simulations to estimate the accuracy of the modeling and foldability. In Figure 2, we present a histogram of the I-TASSER C-score of the

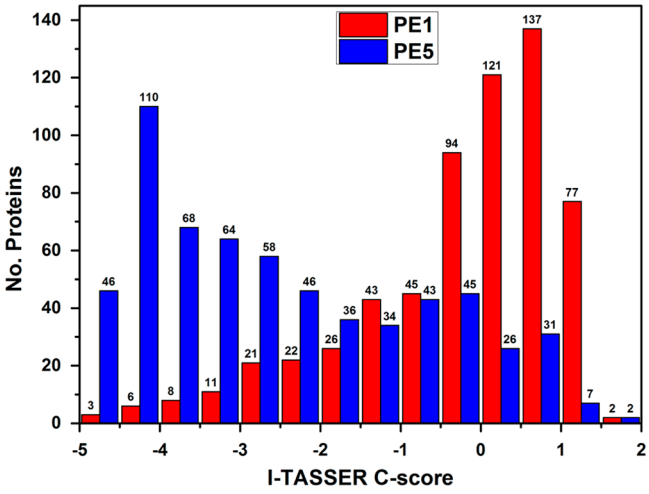


Figure 2. Histogram distribution of I-TASSER C-scores for PE1 and PE5 proteins.

616 PE1 proteins, where 519 proteins have a C-score above −1.5, which largely corresponds to the number of proteins with a TM-score > 0.5. The average TM-scores for the proteins with C-score > −1.5 and < −1.5 are 0.86 and 0.32, respectively, which confirms the strong correlation of the C-score and the quality of the I-TASSER models, as observed in previous benchmark tests.[23]

Starting from I-TASSER models, COFACTOR can generate three aspects of functional annotations: enzyme commission, gene ontology, and ligand-binding site predictions. Among the 616 PE1 proteins, 582, 585, 556, and 224 proteins have GO molecular function, GO biological process, GO cellular component, and enzyme commission annotations in neXtProt database, respectively; 276 proteins have ligand-binding sites annotated in the BioLip database.[25] Although there are no homologous templates used, the COFACTOR models have 508, 515, 432, and 161 proteins for which the GO molecular function, GO biological process, GO cellular component, and enzyme commission are correctly assigned, which corresponds to an accuracy of 87, 88, 77, and 72%, respectively. Here, a correct EC assignment is defined as having the first three digits correctly predicted, and a correct GO assignment is defined as having the GO item at the first level correctly identified. Among the 276 proteins of the ligand-binding data, 172 (62%) have more than 70% of their binding sites correctly predicted. The majority of targets with a correct functional assignment are also correctly folded with the I-TASSER model, i.e., TM-score > 0.5, showing the dependence of the functional annotations on the correctness of the structure's folding.

Figure 3 presents a histogram of the confidence score (F-score) of the COFACTOR predictions. If we account for the 85

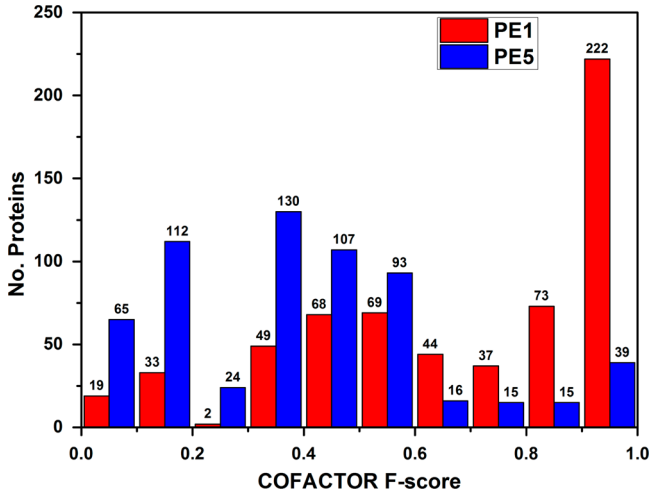


Figure 3. Histogram distribution of COFACTOR F-scores for PE1 and PE5 proteins.

proteins that have a F-score above 0.6, then the success rates of the functional assignments are 93, 95, 93, 94, and 91% for GO molecular function, GO biological process, GO cellular component, enzyme commission, and ligand-binding sites, respectively, which are significantly higher than those with a F-score below 0.6 (i.e., 79, 81, 78, 65, and 53%). These data show the efficiency of COFACTOR for structure-based functional assignments and the ability of the F-score to distinguish correct from incorrect functional assignments.

Summary of the Predicted Structure and Function of the Putative Proteins in PE5

Because PE5 proteins have not been validated by any proteomics experimental method, the native structure and function of these proteins are unknown. In Figure 2, we show the C-score histograms of PE5 proteins in comparison with those of PE1 proteins. As expected, the population of proteins with a high-confidence folding score is much lower in the PE5 group than that in the PE1 group. For example, there are 519 PE1 proteins that have a C-score > −1.5, whereas the number for the PE5 proteins is only 188. This is understandable because most PE1 proteins are well-characterized proteins with regular structural folds, whereas, by definition, PE5 proteins are dubious or uncertain and their gene sequences may not code for expressable proteins. Here, we note that the best C-score of all domains for multidomain proteins is reported in Figure 2 for PE5 proteins since the existence of one domain from a protein sequence can be sufficient to confirm that the corresponding protein is a gene-coding protein.

Nevertheless, the data seems to suggest that not all PE5 proteins are from noncoding genes. If we consider a stringent C-score cutoff > 0.0, in which all proteins have I-TASSER models with a correct fold in our benchmark test on the PE1 proteins as well as in the former benchmark experiment,[23] then there are 66 PE5 proteins that meet this criterion; these are the most likely to correspond to gene-coding proteins from the viewpoint of non-homology-based structure folding. A summary of these proteins is listed in Table 1; the data are also available at http://zhanglab.ccmb.med.umich.edu/HPSF/66.html. We acknowledge that proteins with a lower C-score may also be correctly folded in I-TASSER, but the likelihood of success is lower than for those with a higher C-score.

DOI: 10.1021/acs.jproteome.5b00516
J. Proteome Res. 2015, 14, 3750−3761

Table 1. List of 66 PE5 Proteins That Have C-Score >0 in I-TASSER Folding Simulations

	ID[a]	Chr[b]	Name[c]	C[d]	F[e]	Dm[f]	Loc[g]	Class[h]	HGNC[i]	K[j]	P[k]
1	NX_A6NI03	11	TRIM64B	1.61	0.23	Y	Cytoplasm	All beta proteins	gene with protein product	Y	N
2	NX_A6NLI5	11	TRIM64C	1.51	0.23	Y	Cytoplasm	All beta proteins	gene with protein product	N	N
3	NX_Q6ZN08	19	ZNF66	1.24	0.42	Y	Nucleus	Small proteins	unknown	N	N
4	NX_Q5NE16	9	CTSL3P	1.12	0.9	N	Extracell	Alpha and beta proteins (a+b)	pseudogene (ID = 0.45)	N	N
5	NX_A6NK02	4	TRIM75P	1.12	0.37	Y	Cytoplasm	All beta proteins	pseudogene (ID = 0.47)	N	N
6	NX_A6NMB9	12	FIGNL2	1.12	0.7	Y	Nucleus	Alpha and beta proteins (a+b)	unknown	Y	N
7	NX_P48741	1	HSPA7	1.09	0.98	Y	Nucleus	Alpha and beta proteins (a+b)	pseudogene (ID = 0.93)	N	N
8	NX_A6NGE7	13	URAD	1.07	0.67	N	Extracell	All alpha proteins	gene with protein product	Y	N
9	NX_A6NHM9	7	MOXD2P	1.03	0.06	Y	Extracell	All beta proteins	pseudogene (ID = 0.40)	N	N
10	NX_Q96TA0	5	PCDHB18	0.92	0.48	Y	Plasma membrane	Low resolution protein structures	pseudogene (ID = 0.78)	N	N
11	NX_A4D2B8	7	PMS2P1	0.91	0.57	Y	Nucleus	Alpha and beta proteins (a+b)	pseudogene (ID = 0.38)	N	N
12	NX_Q9H560	9	ANKRD19P	0.9	0.96	N	Plasma membrane	Alpha and beta proteins (a+b)	pseudogene (ID = 0.88)	N	N
13	NX_Q8N7Z5	5	ANKRD31	0.9	0.12	Y	Cytoplasm	Alpha and beta proteins (a+b)	gene with protein product	Y	N
14	NX_O95397	12	PLEKHA8P1	0.89	0.44	Y	Cytoplasm	All alpha proteins	pseudogene (ID = 0.97)	N	N
15	NX_Q6ZTB9	19	ZNF833P	0.89	0.91	N	Nucleus	Designed proteins	pseudogene (ID = 0.69)	N	N
16	NX_B5MCN3	22	SEC14L6	0.88	0.49	Y	Cytoplasm	Alpha and beta proteins (a+b)	gene with protein product	Y	Y
17	NX_A0PJZ0	18	ANKRD20A5P	0.87	0.97	N	Plasma membrane	Alpha and beta proteins (a+b)	pseudogene (ID = 0.92)	N	N
18	NX_C9J798	7	RASA4B	0.86	0.34	Y	Plasma membrane	All alpha proteins	gene with protein product	N	N
19	NX_A8MWD9	19	SNRPGP15	0.8	0.41	N	Nucleus	All beta proteins	pseudogene (ID = 0.95)	N	N
20	NX_A4QPH2	22	PI4KAP2	0.79	0.69	Y	Nucleus	Alpha and beta proteins (a+b)	pseudogene (ID = 0.93)	N	N
21	NX_Q6ZT77	19	ZNF826P	0.76	0.81	N	Nucleus	Designed proteins	pseudogene (ID = 0.50)	N	N
22	NX_A6NIE9	16	PRSS29P	0.73	0.97	Y	Extracell	All beta proteins	pseudogene (ID = 0.46)	N	N
23	NX_Q8NGA4	19	GPR32P1	0.73	0.59	N	Plasma membrane	Membrane and cell surface proteins and peptides	pseudogene (ID = 0.83)	N	N
24	NX_P0C7Q3	1	FAM58BP	0.73	0.9	N	Nucleus	All alpha proteins	pseudogene (ID = 0.85)	N	N
25	NX_P0CB33	7	ZNF735P	0.71	0.12	Y	Nucleus	Designed proteins	pseudogene (ID = 0.81)	Y	Y
26	NX_E5RG02	3	PRSS46	0.71	0.81	Y	Extracell	All beta proteins	gene with protein product	N	Y
27	NX_A6NEY8	2	PRORSD1P	0.69	0.49	N	Cytoplasm	Alpha and beta proteins (a+b)	pseudogene (ID = 0.27)	N	N
28	NX_Q9HAU6	8	TPT1P8	0.65	0.76	N	Cytoplasm	All beta proteins	pseudogene (ID = 0.69)	N	N
29	NX_A8MUV8	7	ZNF727P	0.65	0.43	Y	Nucleus	Designed proteins	pseudogene (ID = 0.71)	Y	Y
30	NX_P12525	X	MYCLP1	0.63	0.31	Y	Nucleus	All alpha proteins	pseudogene (ID = 0.73)	N	N
31	NX_P0C7 × 4	X	FTH1P19	0.62	0.9	Y	Plasma membrane	All alpha proteins	pseudogene (ID = 0.54)	N	N
32	NX_Q63ZY6	7	NSUN5P2	0.6	0.82	N	Nucleus	Alpha and beta proteins (a+b)	pseudogene (ID = 0.76)	N	N
33	NX_A8MV57	1	MPTX1	0.6	0.68	N	Extracell	Low resolution protein structures	pseudogene (ID = 0.50)	N	N
34	NX_Q6NSI1	16	ANKRD26P1	0.6	0.94	N	Nucleus	Alpha and beta proteins (a+b)	pseudogene (ID = 0.54)	N	N
35	NX_Q8IWF7	X	UBE2DNL	0.59	0.56	N	Nucleus	Alpha and beta proteins (a+b)	pseudogene (ID = 0.71)	N	N
36	NX_A8MUU1	7	FABP5P3	0.58	0.46	N	Cytoplasm	All beta proteins	pseudogene (ID = 0.91)	N	N

DOI: 10.1021/acs.jproteome.5b00516
J. Proteome Res. 2015, 14, 3750−3761

Table 1. continued

	ID[a]	Chr[b]	Name[c]	C[d]	F[e]	Dm[f]	Loc[g]	Class[h]	HGNC[i]	K[j]	P[k]
37	NX_Q9NSJ1	21	ZNF355P	0.54	0.36	Y	Nucleus	Small proteins	pseudogene (ID = 0.67)	N	N
38	NX_Q96P88	1	GNRHR2	0.54	0.51	N	Plasma membrane	Membrane and cell surface proteins and peptides	pseudogene (ID = 0.36)	N	N
39	NX_Q6ZUV0	4	PDXDC2P	0.53	0.3	N	Cytoplasm	Alpha and beta proteins (a+b)	NE	N	N
40	NX_Q58FG1	4	HSP90AA4P	0.5	0.36	Y	Cytoplasm	Alpha and beta proteins (a+b)	pseudogene (ID = 0.87)	N	Y
41	NX_Q6P474	16	PDXDC2P	0.48	0.82	N	Cytoplasm	Alpha and beta proteins (a+b)	pseudogene (ID = 0.97)	N	N
42	NX_Q3KNT7	7	NSUN5P1	0.41	0.6	N	Nucleus	Alpha and beta proteins (a+b)	pseudogene (ID = 0.91)	N	N
43	NX_A6NGU5	22	GGT3P	0.4	0.47	Y	Extracell	Alpha and beta proteins (a+b)	pseudogene (ID = 0.98)	N	N
44	NX_Q6ZSU1	19	CYP2G1P	0.4	0.68	N	Endoplasmic reticulum	All alpha proteins	pseudogene (ID = 0.60)	N	N
45	NX_Q8IZP2	13	ST13P4	0.4	0.73	Y	Cytoplasm	All alpha proteins	pseudogene (ID = 0.97)	N	N
46	NX_Q7RTY9	16	PRSS41	0.39	0.98	Y	Extracell	All beta proteins	unknown	N	N
47	NX_B5MD39	22	GGTLC3	0.34	0.92	N	Extracell	Alpha and beta proteins (a+b)	unknown	N	N
48	NX_Q9BZ68	X	FRMD8P1	0.32	0.4	N	Cytoplasm	All beta proteins	pseudogene (ID = 0.94)	N	N
49	NX_D6RBM5	4	USP17L23	0.31	0.7	N	Nucleus	All alpha proteins	gene with protein product	N	N
50	NX_Q8NHW5	2	RPLP0P6	0.29	0.78	Y	Nucleus	Low resolution protein structures	pseudogene (ID = 0.98)	N	N
51	NX_Q58FF6	15	HSP90AB4P	0.28	0.57	Y	Centrosome	Alpha and beta proteins (a+b)	pseudogene (ID = 0.82)	N	Y
52	NX_Q99463	5	NPY6R	0.27	0.68	N	Plasma membrane	Membrane and cell surface proteins and peptides	pseudogene (ID = 0.51)	N	N
53	NX_O60774	1	FMO6P	0.26	0.47	Y	Endoplasmic reticulum	Alpha and beta proteins (a+b)	pseudogene (ID = 0.71)	N	N
54	NX_Q7RTZ2	8	USP17L1P	0.23	0.91	Y	Nucleus	Alpha and beta proteins (a+b)	pseudogene (ID = 0.99)	Y	N
55	NX_A8MVU1	7	NCF1C	0.2	0.12	Y	Cytoplasm	All beta proteins	pseudogene (ID = 0.99)	N	N
56	NX_P01893	6	HLA-H	0.2	0.97	Y	Plasma membrane	Alpha and beta proteins (a+b)	pseudogene (ID = 0.87)	N	N
57	NX_Q15940	19	ZNF726P1	0.19	0.34	N	Nucleus	Designed proteins	pseudogene (ID = 0.76)	N	N
58	NX_Q9BYX7	2	POTEKP	0.19	0.99	N	Cytoskeleton	Low resolution protein structures	pseudogene (ID = 0.95)	N	N
59	NX_A4D1Z8	7	GRIFIN	0.13	0.98	N	Extracell	All beta proteins	gene with protein product	N	N
60	NX_O95744	7	PMS2P2	0.13	0.91	Y	Nucleus	Alpha and beta proteins (a+b)	pseudogene (ID = 0.54)	N	N
61	NX_Q5VTE0	9	EEF1A1P5	0.08	0.95	N	Cytoplasm	Low resolution protein structures	pseudogene (ID = 1.00)	N	N
62	NX_P0CG00	19	ZSCAN5DP	0.06	0.12	Y	Nucleus	Small proteins	pseudogene (ID = 0.76)	N	N
63	NX_P0CF97	4	FAM200B	0.06	0.46	Y	Nucleus	Alpha and beta proteins (a+b)	gene with protein product	Y	N
64	NX_Q8WTZ4	X	CA5BP1	0.03	0.9	N	Cytoplasm	All beta proteins	pseudogene (ID = 0.35)	N	N
65	NX_Q9NRI7	17	PPY2	0.02	0.14	N	Extracell	Peptides	pseudogene (ID = 0.29)	N	N
66	NX_Q6ZRF7	19	ZNF818P	0	0.85	N	Nucleus	Designed proteins	pseudogene (ID = 0.47)	N	N

[a]ID, neXtProt ID. [b]Chr, order number of chromosome. [c]Name, gene name from HGNC symbol. [d]C, I-TASSER C-score. For multidomain proteins, the highest C-score of all domains is listed. [e]F, F-score of COFACTOR prediction on GO molecular function. [f]Dm: Y, multidomain protein; N, single-domain protein. [g]Loc, subcellular localization predicted by Hum-mPloc. [h]Class, fold class. [i]HGNC, HGNC annotation retrieved on 2014/9/5; the number in parentheses is the sequence identity (ID) between the pseudogene and the closest PE1−4 protein. [j]K: Y, detected by Kim et al;[38] N, not detected by Kim et al. [k]P: Y, included in PeptideAtlas 2014-08; N, not included in PeptideAtlas 2014-08.

In Figure 3, we show the F-score distribution of PE5 proteins in comparison with that for PE1 proteins. Again, there is a much lower population of high F-score proteins in PE5 than there is in PE1. There are 85 PE5 proteins that have a F-score above 0.6, of which 32 are also in the list of the 66 high C-score proteins from the I-TASSER folding simulations (Table 1); this agreement partly confirms the coincidence of the structure and function annotation data.

We also examined and compared the intrinsically disordered regions of the PE1 and PE5 sequences using the DisEMBL

DOI: 10.1021/acs.jproteome.5b00516
J. Proteome Res. 2015, 14, 3750−3761

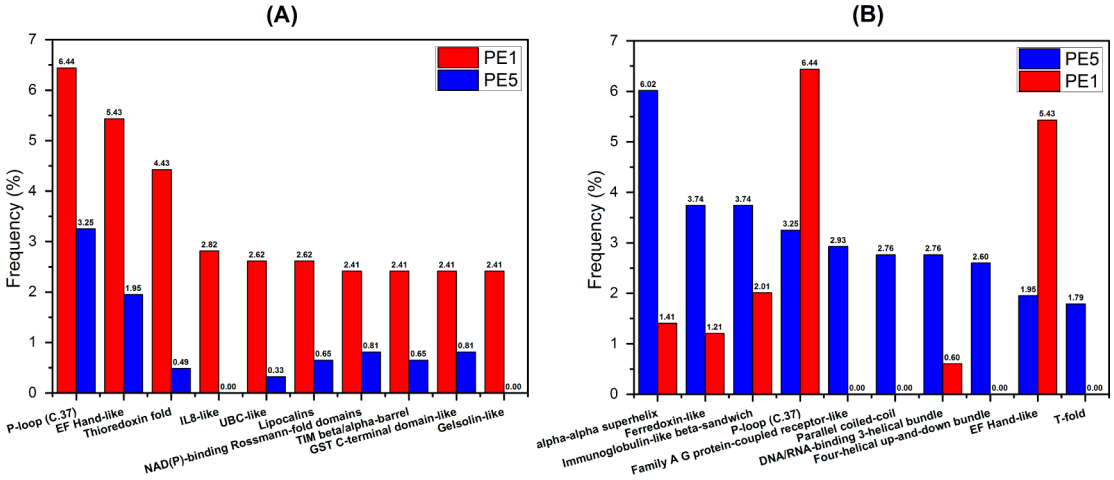


Figure 4. Relative frequency distributions of the top-ten fold families assigned for (A) PE1 and (B) PE5 proteins. The corresponding frequencies from proteins in the opposite protein sets are listed as a control.

program.[31] There are only two out of 616 PE1 sequences for which more than 40% of their regions are predicted to be disordered by DisEMBL, whereas the corresponding number of PE5 sequences is 79. Since intrinsically disordered regions do not have regular 3D structures, the average I-TASSER C-score for the disordered proteins is −3.44, which is 35% lower than that of other PE5 proteins. Such a high fraction of disordered sequences in PE5 proteins should also contribute to the low folding rate compared to that of PE1 proteins.

Structure Classification Analyses of the I-TASSER Models

To assign analogous structure families, we match the I-TASSER models of the target proteins with the structure domains in the SCOPe library,[32] an extended structure fold-family library integrated from the standard SCOP[33] and ASTRAL[34] databases. We use the structure alignment program TM-align[35] to align the I-TASSER model with all structural domains in SCOPe; the fold family of the SCOPe protein that has the highest TM-score in the I-TASSER model is then assigned to the target sequence. In the case where the target protein contains multiple domains, the domain that has the maximum TM-score for the top-ranked SCOPe domain is used.

We first applied the structure-based threading approach to PE1 proteins, where 98.6% of PE1 proteins have an I-TASSER model that matches the correct SCOPe fold families despite the fact that no homologous templates were used in the I-TASSER modeling. Overall, the 616 PE1 proteins were assigned to 168 fold families, with an average TM-score between the I-TASSER model and the SCOPe domain of 0.94.

When we applied the same structural matching procedure to PE5 proteins, a more divergent set of 202 families was associated with the 616 PE5 proteins. This divergence in the fold family assignment might be partly due to uncertainty in the TM-align structural assignments at the low-fold-similarity range because the average TM-score between the I-TASSER model and the SCOPe domain for PE5 proteins is much lower than that for PE1 proteins (0.72 vs 0.94). If we focus only on the assignments with a TM-score above 0.5, then there are 318 targets that have been reliably assigned to 152 fold families, with an average TM-score increasing to 0.86. This protein set was considered in our domain assignment analysis for the PE5 proteins.

Although a different number of proteins has been assigned, it is of interest to examine the relative distributions of the top fold families assigned to the PE1 and PE5 proteins. Figure 4A,B shows the top 10 folds for proteins in PE1 and PE5, respectively. As shown in the figure, the P-loop (C.37) and EF hand-like fold are among the most popular folds for both PE1 and PE5 proteins. However, the three largest fold families for PE5 proteins, alpha−alpha superhelix, ferredoxin-like, and immunoglobulin-like beta-sandwich, have quite a low population among PE1 proteins. Noticeably, PE5 proteins are overexpressed in four of the 10 families, i.e., family A G-protein-coupled receptor-like, parallel coiled-coil, four-helical up-and-down bundle, and T-fold, in which there are no PE1 proteins.

Since protein folds in the SCOPe database are specific to the architecture of secondary structure arrangements, the limited proteins tested may result in variations in the above comparisons. In Figure 5, we compare PE1 and PE5 proteins based on their

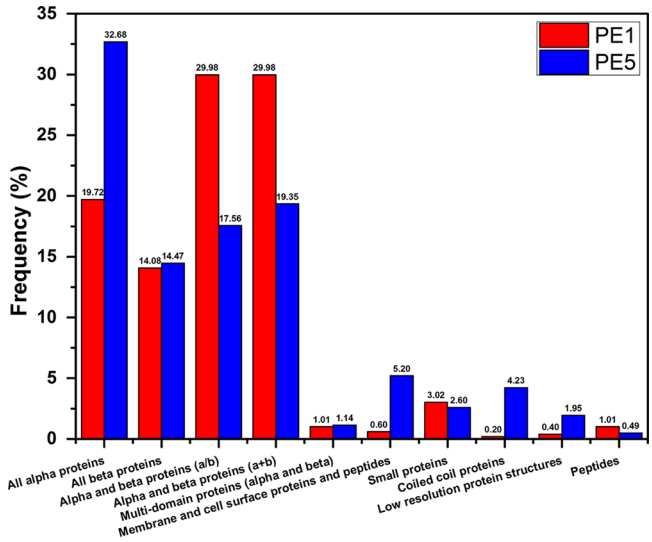


Figure 5. Relative frequency distribution of SCOPe classes for PE1 and PE5 proteins.

class, which is a higher-level protein structure classification. The populations observed for the PE1 and PE5 proteins are more consistent in this comparison due to the lower level of coarse-grained classification. However, again, the proteins in PE5, if expressed, would be over-represented in the membrane and cell

DOI: 10.1021/acs.jproteome.5b00516
J. Proteome Res. 2015, 14, 3750−3761

surface proteins and peptides and coiled coil proteins classes compared with PE1 proteins.

GO Function Prediction Evaluation

In Figure 6, we show a histogram distribution of the gene ontology (GO) predictions generated by COFACTOR. There

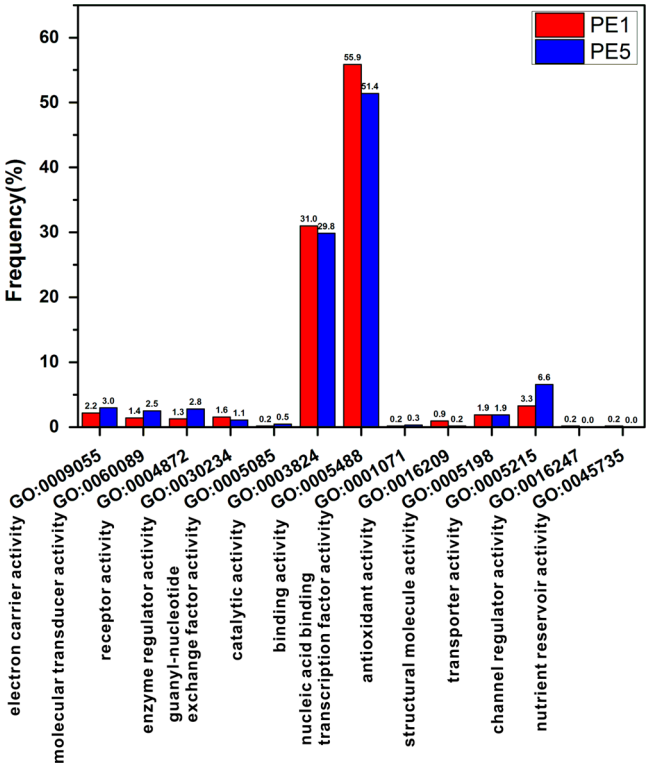


Figure 6. Relative frequency distribution of predicted GO items from the first level of molecular function for PE1 and PE5 proteins.

are 13 GO terms from the first level of molecular function. PE1 and PE5 proteins have a similar GO population distribution, both of which are dominated by two major groups: binding (GO:0005488) and catalytic activity (GO:0003824). Nevertheless, PE5 proteins are over-represented in the transporter activity (GO:0005215) and receptor activity (GO:0004872) GO items in comparison with PE1 proteins. The *p*-values of the difference in the A/B split significance test are 0.034 and 0.004, respectively, for GO:0005215 and GO:0004872. The differences between PE1 and PE5 for all other GO terms are statistically insignificant. The data are consistent with the structure-based fold family annotations in which PE5 shows a higher rate of membrane proteins than PE1.

Subcellular Localization Analyses

Proteins conduct different functions in specific cellular compartments. Information regarding the subcellular localization of proteins is therefore important for understanding their biological functions. In Figure 7, we present a comparison of the subcellular localizations of PE1 and PE5 proteins generated by the HumPLoc 2.0 program.[27] Most PE1 and PE5 proteins are located in the cytoplasm, extracellular, nucleus, and plasma membrane, but the relative rates are very different. More than 40% of PE1 proteins (vs 14% of PE5 proteins) are located in the cytoplasm; the relative populations of PE5 in extracellular, nucleus, and plasma membrane are much higher than those of PE1 proteins,

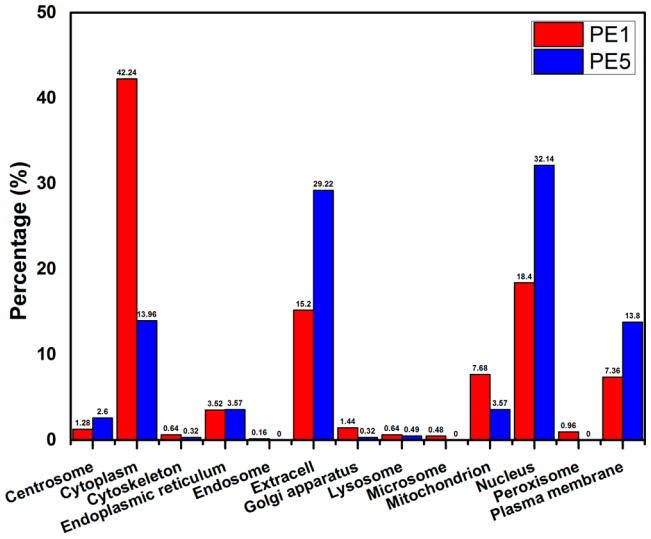


Figure 7. Comparison of subcellular localizations of PE1 and PE5 proteins.

partly consistent with the structure fold assignment and function predictions.

Comparison of I-TASSER Folding with Mass Spectrometry Data

Mass spectrometry is an effective tool to identify proteins and peptides. Large-scale mass spectrometry data have been accumulated in various public databases, such as PeptideAtlas[36] and GPMDB.[10] The HUPO Human Protein Project and other groups[37,38] use mass spectrometry to develop draft human proteomes. Kim et al.[38] recently reported the identification of about two-thirds (2535/3844) of the neXtProt 2013 missing proteins[12] using high-resolution Fourier-transform mass spectrometry. Although these reported proteins are mainly PE2−4 proteins, using RefSeq mapping,[39] we found 41 of the 616 PE5 proteins that are also in Kim's original data set. Among the 41 proteins, nine are foldable by I-TASSER simulations with a C-score above 0. Similarly, we found 40 PE5 proteins that were identified in the PeptideAtlas 2014-08 data set using the Trans-Proteomic Pipeline with a 1% protein FDR filter,[9] where six entries have an I-TASSER C-score > 0. These proteins are marked in the last two columns of Table 1.

In Figure 8, we present the I-TASSER 3D structure of three illustrative high C-score examples that have been detected by the mass spectrometry. A stable fold was seen in each of the examples due to the selection of high-confidence folding scores.

Annotation of the Six High-Scoring PE5 Proteins in PeptideAtlas

As shown in Table 1, six of the 66 targets with an I-TASSER C-score > 0 are found in the PeptideAtlas 2014-08 data set. These protein entries may be of high priority for further experimental evaluation. Although no homologous templates were used for either structure or function prediction, the top GO-term predictions of these targets are found to match the specific functional regions/residues observed in these proteins. Here, we present a brief summary of the available information on these proteins when the prediction data is manually compared with major proteomics data libraries. The annotations are presented in descending order of C-score in Table 1.

SEC14L6 (Putative SEC14-like Protein 6) (Line 16). This is a protein-coding transcript found in both Ensembl and Refseq

DOI: 10.1021/acs.jproteome.5b00516
J. Proteome Res. 2015, 14, 3750−3761

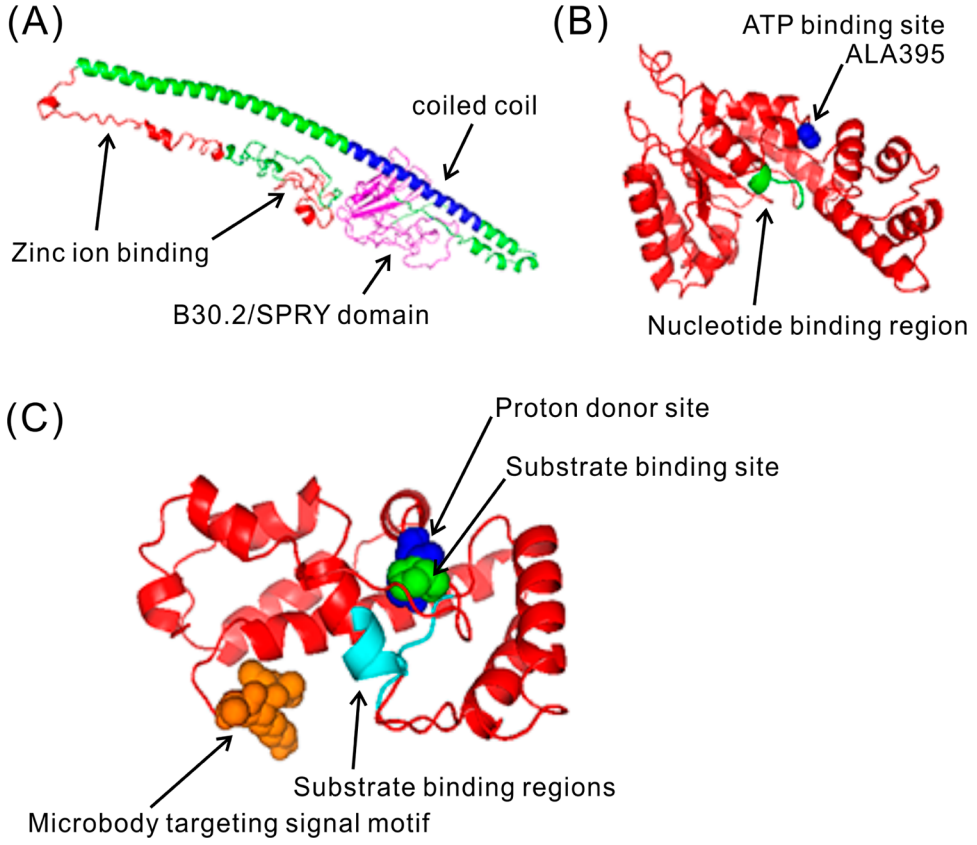


Figure 8. I-TASSER structural models for the three genes that have the highest C-score and are detected by mass spectrometry data. (A) TRIM64b with the two zinc finger regions (15−56 and 87−128 aa) (red), one coiled-coil region (189−225 aa) (blue), and a B30.2/SPRY domain (268−449 aa) (magenta). (B) C-Terminal domain (380−653 aa) of FIGNL2 containing the ATP binding site (ALA395) (blue) and the nucleotide binding region (435−440 aa) (green). (C) URAD protein with a proton donor site (His67) (blue), a substrate binding site (Pro68) (green), two substrate binding regions (84−88 and 119−123 aa) (cyan), and a microbody targeting signal motif (171−173 aa) (orange).

translated to a protein product of 397 aa containing a CRAL-TRIO domain (76−249 aa) and a GOLD domain (252−383 aa). The CRAL-TRIO domain binds small lipophilic molecules; GOLD is a β-strand-rich domain found in several proteins involved in Golgi dynamics as well as intracellular protein trafficking. The top GO molecular function term by COFACTOR for the full-length protein is transporter activity (GO:0005215); for the N-terminal domain (C-score = 0.65), it is Vitamin E binding (GO:0008431). Several studies have linked Vitamin E with proteins with CRAL-TRIO and GOLD domains.[40,41] HPA detected impressive expression in lung and kidney at the RNA level but not at the protein level. This protein has seven peptides in PeptideAtlas, but none is proteotypic, so the protein appears to be subsumed under SEC14L4 (Q9UDX3) (PE1), known as tocopherol (vitamin E)-associated protein 3, or possibly SEC14L6.

ZNF735P (Putative Zinc Finger Protein 735) (Line 25). A protein-coding transcript found in both Ensembl and Refseq translates to a protein product of 412 aa containing a KRAB domain (16−87 aa) and five zinc-finger regions (157−179, 185−207, 213−235, 241−263, and 269−291 aa). KRAB and zinc-finger regions are involved in transcriptional regulation. The COFACTOR predicted molecular function terms nucleic acid binding (GO:0003676) and zinc ion binding (GO:0008270) are both linked to transcription regulation. Ten peptides shared by ZNF735P and other proteins are found in PeptideAtlas. It is called a pseudogene by neXtProt, but HGNC recently upgraded it to a gene with a protein product from its previous status as a pseudogene (see HGNC annotation in Table 1, retrieved on 2014/9/5).

PRSS46 (Putative Serine Protease 46) (Line 26). A protein-coding transcript found in both Ensembl and Refseq translates to a protein product of 174 aa containing a peptidase S1 domain (43−174 aa). Serine-type endopeptidase activity (GO:0004252) was the top molecular function predicted by COFACTOR. This protein is found in PeptideAtlas, but there were no proteotypic peptides. HPA reports RNA expression of this gene in testis and skeletal muscle. Three peptides are shared by PRSS46 and many other proteins found in PeptideAtlas. The high C-score may reflect normal folding of an N-terminal domain; note in Table 1 that this predicted protein has multiple domains.

ZNF727P (Putative Zinc Finger Protein 727) (Line 29). A protein-coding transcript found in both Ensembl and Refseq translates to a protein product of 499 aa containing a KRAB domain (4−75 aa) and five Krueppel C2H2-type zinc-finger regions (143−167, 200−222, 228−250, 256−278, and 284−306 aa). COFACTOR predicted molecular function terms nucleic acid binding (GO:0003676) and zinc ion binding (GO:0008270). HPA reports protein detection at high or medium expression level in 10 normal tissue cell types. At least five nonproteotypic peptides are shared by ZNF727P and other proteins found in PeptideAtlas.

HSP90AA4P (Putative Heat Shock Protein HSP 90-alpha A4) (Line 40). Annotated as a processed pseudogene with no protein product in Ensembl, this translated protein sequence

DOI: 10.1021/acs.jproteome.5b00516
J. Proteome Res. 2015, 14, 3750−3761

would contain four ATP binding sites (33, 52, 78, and 204 aa). COFACTOR predicts ATP binding (GO:0005524) as the top molecular function. Eight proteotypic peptides matching only to HSP90AA4P are reported in PeptideAtlas, with striking sequence differences from related HSP90 proteins.

HSP90AB4P (Putative Heat Shock Protein HSP 90-alpha B4) (Line 51). Annotated as a processed pseudogene in neXtProt, the translated protein sequence would contain three ATP binding sites (22, 83, and 109 aa). COFACTOR predicted ATP binding (GO:0005524) and chaperone activity (GO:0051082) as the top molecular functions. Ten proteotypic peptides matching only to HSP90AB4P, of a total of 155 peptides, are reported in PeptideAtlas; however, it has high homology to four other HSP proteins, with many apparent single amino acid substitutions.

Overall, these brief annotations illustrate how the I-TASSER/COFACTOR results can be utilized to select protein candidates for further evaluation by proteomic databases, including reconciliation of different conclusions by the different databases, and then critical experiments. It suggests that certain PE5 entries may be priority candidates for reclassification. From time to time, PE5 entries may be promoted to higher categories, especially PE1, or may be sent to the UniProtArchive UniParc. In a private communication, we were informed by the neXtProt director, Lydie Lane, that 10 of the 616 PE5 genes have been recently nominated to UniProt/SwissProt for reassessment of the PE level assignments in light of additional data, of which two are in our top 66. When the SwissProt and neXtProt reassessment is completed and the results made public for those 10 entries, it will be of interest to examine how I-TASSER and COFACTOR characterized each of them.

Pseudogenes in PE5 Proteins

Pseudogenes are facsimiles of protein-coding genes but have lost the ability to produce functional proteins.[42,43] Pseudogenes are nearly as numerous as coding genes in the human genome, with an estimated range from 10 000 to 20 000.[44] Most pseudogenes are the consequence of gene duplication and reverse transcription events; there are also unitary pseudogenes (<100) that arose from direct mutation from existing coding genes.[45] Despite their high population and multiple resources of origin, a common feature of pseudogenes is their lack of ability to code functional proteins due to the failure of transcription and/or translation. Since there is no evolutionary negative selective pressure from coding function, mutations of pseudogenes are essentially random, which can drive the sequences of pseudogenes far from those of the original protein-coding genes, although many pseudogenes may still remain homologous with a high sequence identity to the original coding genes, depending on the distance of evolution.

According to the HGNC annotation, there are 252 pseudogenes among the 616 PE5 proteins, including 51 of our top 66 in Table 1 and four of the six selected for annotation. To obtain a rough examination of the evolutionary distance, we performed a PSI-BLAST search[46] of the PE5 sequences against all PE1−4 proteins in the human genome that are supposed to be protein-coding genes, and we calculated sequence identity with the NW-align program.[47] We found that 135 of the 252 pseudogenes (or 23 in the list of 66 and three in the list of six) have a sequence identity above 80% with PE1−4 coding genes. It might be difficult for current methods to discriminate these pseudogenes because their sequence identity to coding genes is high and their functional features (including the binding pockets and catalysis sites) may still be well-conserved.

However, for more distantly related putative pseudogenes, the foldability and especially the subtle functional characteristic of the coding genes are likely to be spoiled by the accumulation of random mutations. The high confidence models from the combined I-TASSER and COFACTOR simulations might help to examine their coding potential. Here, the sequence identity cutoff of 80% is somewhat arbitrary for defining closely related pseudogenes. If we reduce the cutoff to 70% (or 50%), then the number of the closely related pseudogenes will increase to 31 (or 40) in the list of 66 and four in the list of six. We conclude that attention should be concentrated on nonpseudogenes and putative pseudogenes with a distant relation to the original coding genes. The sequence identity data relative to the original coding gene is listed in column 10 of Table 1 for all of the putative pseudogenes.

Web Interface of the HPSF Database

We established a new webpage to deposit and make available the structure folding and function annotation results of the PE5 proteins: http://zhanglab.ccmb.med.umich.edu/HPSF/. Users can browse or search specific proteins by clicking the Browse&Search link. The protein entries can be searched by inputting the gene name, neXtProt ID, or HGNC symbol, where both partial and full values are accepted. To facilitate the search by users who are not familiar with the query ID and protein names, the input box can automatically provide up to 20 suggestions if any record matches the inputted text.

For each protein entry, HPSF provides structure-folding information, including secondary structure assignment, solvent accessibility, and the I-TASSER structure models that are associated with the confidence assessment from the C-score and residue-level error estimations. The output page also provides structure-based functional annotations, including gene ontology, enzyme commission, ligand-binding site, subcellular localization, and the associated COFACTOR confidence measure, the F-score.

To facilitate comparative analyses, up to 10 homologous hits with PE1 level proteins are listed that have a PSI-BLAST *E*-value below 0.001 for the PE5 protein. Meanwhile, links are provided to other important databases, including neXtProt, HGNC, and ENSEMBL, for each PE5 protein entry.

■ CONCLUSIONS

We have performed a systematic examination of the 616 PE5 proteins that were dubious or uncertain based on experiments curated in neXtProt 2014-09-19 using the cutting-edge structural bioinformatics tools, I-TASSER and COFACTOR, for protein folding and structure-based functional annotations. The I-TASSER simulations show that PE5 proteins, overall, have a significantly lower folding rate than that of PE1 proteins that have been detected confidently in proteome experiments. Nevertheless, there are 66 proteins from the PE5 data set where at least one domain can be folded by I-TASSER with high confidence without using any homologous templates; 32 proteins/domains are further shown to have a high possibility of being functional by structure-based COFACTOR functional annotation. Of the 66 highest scoring PE5 protein candidates, six PE5 entries appear in PeptideAtlas 2014-08 and nine PE5 entries (three in common with PeptideAtlas) were reported by Kim et al. Among the six PE5 entries in PeptideAtlas, only HSP90AA4P and HSP90AB4P have proteotypic peptides (with 8 and 10

DOI: 10.1021/acs.jproteome.5b00516
J. Proteome Res. 2015, 14, 3750−3761

proteotypic peptides, respectively), which are probably more likely to be functional. These proteins may represent the most likely set of PE5 proteins predicted to be functional, and we recommend them for further experimental investigation. In fact, some of the proteins in the list have already been updated to a higher level after this work was completed (e.g., USP17L1P in line 54 of Table 1 is now in PE3 in the UniProtKB and neXtProt databases). Nevertheless, since closely related pseudogenes have a high possibility of maintaining their folding and functional characteristics on their own due to having close sequence identity with that of their original coding genes, more attention should probably be paid to the nonpseudogene and the distantly related pseudogene entries. In the 66 high-scoring genes (or the six PE5 genes in PeptideAtlas), there are 43 (or three) that belong to this category of non- or distantly related pseudogene entries under a sequence identity cutoff of 80% relative to the PE1−4 protein partners (Table 1).

By matching the structural models with the SCOPe domain family database, 318 PE5 proteins have been assigned to 152 domain families with high structural similarity. Compared with normal (PE1) protein-coding genes, PE5 proteins are found to be over-represented in the membrane and cell surface and peptides and coiled coil fold families. Detailed structure-based functional analyses show that most of these over-represented PE5 proteins belong to transporter and receptor and, if expressed, would be localized in the plasma membrane compartment. These data suggest that, besides unfavorable genomic features, the nondetection of such PE5 proteins may be attributed to the experimental difficulty in identifying membrane-bound proteins and peptides. The use of enrichment techniques with improved solubilization and higher sensitivity for membrane-embedded proteins may be necessary for characterization of this "dark matter" in the human proteome. The annotations provided here for the six selected protein candidates give good illustrations of the complexity of declaring pseudogenes and the variety of data available.

Finally, we note that, while this study is focused on the PE5 genes that have the lowest level of confidence to code proteins, the approach can, in principle, be applied to examine the much larger set of missing proteins in PE2−4. Such analyses are underway.

AUTHOR INFORMATION

Corresponding Authors

*(G.S.O.) E-mail: gomenn@umich.edu.
*(Y.Z.) E-mail: zhng@umich.edu.

Notes

The authors declare no competing financial interest.

ACKNOWLEDGMENTS

We thank Lydie Lane of neXtProt and Eric Deutsch of PeptideAtlas for assistance with the protein annotations. This work was supported in part by the NIH National Institute of General Medical Sciences (GM083107 and GM084222), the National Natural Science Foundation of China (30700162), and the National Institute for Environmental Health Sciences (U54ES017885).

REFERENCES

(1) Venter, J. C.; Adams, M. D.; Myers, E. W.; et al. The sequence of the human genome. *Science* **2001**, *291* (5507), 1304−51.
(2) Marko-Varga, G.; Omenn, G. S.; Paik, Y. K.; et al. A first step toward completion of a genome-wide characterization of the human proteome. *J. Proteome Res.* **2013**, *12* (1), 1−5.
(3) Legrain, P.; Aebersold, R.; Archakov, A.; et al. The human proteome project: current state and future direction. *Mol. Cell. Proteomics* **2011**, *10* (7), M111.009993.
(4) Paik, Y. K.; Jeong, S. K.; Omenn, G. S.; et al. The Chromosome-Centric Human Proteome Project for cataloging proteins encoded in the genome. *Nat. Biotechnol.* **2012**, *30* (3), 221−3.
(5) Aebersold, R.; Bader, G. D.; Edwards, A. M.; et al. The biology/disease-driven human proteome project (B/D-HPP): enabling protein research for the life sciences community. *J. Proteome Res.* **2013**, *12* (1), 23−7.
(6) Lane, L.; Bairoch, A.; Beavis, R. C.; et al. Metrics for the Human Proteome Project 2013−2014 and strategies for finding missing proteins. *J. Proteome Res.* **2014**, *13* (1), 15−20.
(7) Flicek, P.; Amode, M. R.; Barrell, D.; et al. Ensembl 2014. *Nucleic Acids Res.* **2014**, *42*, D749−55.
(8) Lane, L.; Argoud-Puy, G.; Britan, A.; et al. neXtProt: a knowledge platform for human proteins. *Nucleic Acids Res.* **2012**, *40*, D76−3.
(9) Farrah, T.; Deutsch, E. W.; Hoopmann, M. R.; et al. The state of the human proteome in 2012 as viewed through PeptideAtlas. *J. Proteome Res.* **2013**, *12* (1), 162−71.
(10) Craig, R.; Cortens, J. P.; Beavis, R. C. Open source system for analyzing, validating, and storing protein identification data. *J. Proteome Res.* **2004**, *3* (6), 1234−42.
(11) Uhlen, M.; Oksvold, P.; Fagerberg, L.; et al. Towards a knowledge-based Human Protein Atlas. *Nat. Biotechnol.* **2010**, *28* (12), 1248−50.
(12) Uhlen, M.; Fagerberg, L.; Hallstrom, B. M.; et al. Proteomics. Tissue-based map of the human proteome. *Science* **2015**, *347* (6220), 1260419.
(13) Roy, A.; Kucukural, A.; Zhang, Y. I-TASSER: a unified platform for automated protein structure and function prediction. *Nat. Protoc.* **2010**, *5* (4), 725−38.
(14) Yang, J.; Yan, R.; Roy, A.; et al. The I-TASSER Suite: protein structure and function prediction. *Nat. Methods* **2014**, *12* (1), 7−8.
(15) Roy, A.; Zhang, Y. Recognizing protein-ligand binding sites by global structural alignment and local geometry refinement. *Structure* **2012**, *20* (6), 987−97.
(16) Roy, A.; Yang, J.; Zhang, Y. COFACTOR: an accurate comparative algorithm for structure-based protein function annotation. *Nucleic Acids Res.* **2012**, *40*, W471−7.
(17) Wu, S.; Zhang, Y. LOMETS: a local meta-threading-server for protein structure prediction. *Nucleic Acids Res.* **2007**, *35* (10), 3375−82.
(18) Xue, Z.; Xu, D.; Wang, Y.; et al. ThreaDom: extracting protein domain boundary information from multiple threading alignments. *Bioinformatics* **2013**, *29* (13), i247−56.
(19) Zhang, J.; Liang, Y.; Zhang, Y. Atomic-level protein structure refinement using fragment-guided molecular dynamics conformation sampling. *Structure* **2011**, *19* (12), 1784−95.
(20) Tai, C. H.; Bai, H.; Taylor, T. J.; et al. Assessment of template-free modeling in CASP10 and ROLL. *Proteins: Struct., Funct., Genet.* **2014**, *82* (S2), 57−83.
(21) Montelione, G. T. Template based modeling assessment in CASP10, 10th Community Wide Experiment on the Critical Assessment of Techniques for Protein Structure Prediction, Gaeta, Italy, 2012.
(22) Kryshtafovych, A.; Fidelis, K.; Moult, J. CASP10 results compared to those of previous CASP experiments. *Proteins: Struct., Funct., Genet.* **2014**, *82* (S2), 164−74.
(23) Zhang, Y. I-TASSER server for protein 3D structure prediction. *BMC Bioinf.* **2008**, *9*, 40.
(24) Mitra, P.; Shultis, D.; Brender, J. R.; et al. An Evolution-Based Approach to De Novo Protein Design and Case Study on Mycobacterium tuberculosis. *PLoS Comput. Biol.* **2013**, *9* (10), e1003298.
(25) Yang, J.; Roy, A.; Zhang, Y. BioLiP: a semi-manually curated database for biologically relevant ligand-protein interactions. *Nucleic Acids Res.* **2013**, *41* (D1), D1096−103.

DOI: 10.1021/acs.jproteome.5b00516
J. Proteome Res. 2015, 14, 3750−3761

(26) Schmidt, T.; Haas, J.; Gallo Cassarino, T.; et al. Assessment of ligand-binding residue predictions in CASP9. *Proteins: Struct., Funct., Genet.* **2011**, *79* (S10), 126−36.
(27) Chou, K. C.; Shen, H. B. Hum-PLoc: a novel ensemble classifier for predicting human protein subcellular localization. *Biochem. Biophys. Res. Commun.* **2006**, *347* (1), 150−7.
(28) Zhang, Y.; Skolnick, J. Scoring function for automated assessment of protein structure template quality. *Proteins: Struct., Funct., Genet.* **2004**, *57*, 702−710.
(29) Xu, J.; Zhang, Y. How significant is a protein structure similarity with TM-score = 0.5? *Bioinformatics* **2010**, *26* (7), 889−95.
(30) Zhang, Y. Template-based modeling and free modeling by I-TASSER in CASP7. *Proteins: Struct., Funct., Genet.* **2007**, *69* (S8), 108−117.
(31) Linding, R.; Jensen, L. J.; Diella, F.; et al. Protein disorder prediction: implications for structural proteomics. *Structure* **2003**, *11* (11), 1453−9.
(32) Fox, N. K.; Brenner, S. E.; Chandonia, J.-M. SCOPe: Structural Classification of Proteins—extended, integrating SCOP and ASTRAL data and classification of new structures. *Nucleic Acids Res.* **2014**, *42* (D1), D304−D309.
(33) Murzin, A. G.; Brenner, S. E.; Hubbard, T.; et al. SCOP: a structural classification of proteins database for the investigation of sequences and structures. *J. Mol. Biol.* **1995**, *247* (4), 536−40.
(34) Chandonia, J. M.; Hon, G.; Walker, N. S.; et al. The ASTRAL Compendium in 2004. *Nucleic Acids Res.* **2004**, *32*, D189−92.
(35) Zhang, Y.; Skolnick, J. TM-align: a protein structure alignment algorithm based on the TM-score. *Nucleic Acids Res.* **2005**, *33* (7), 2302−9.
(36) Desiere, F.; Deutsch, E. W.; King, N. L.; et al. The PeptideAtlas project. *Nucleic Acids Res.* **2006**, *34*, D655−8.
(37) Wilhelm, M.; Schlegl, J.; Hahne, H.; et al. Mass-spectrometry-based draft of the human proteome. *Nature* **2014**, *509* (7502), 582−7.
(38) Kim, M. S.; Pinto, S. M.; Getnet, D.; et al. A draft map of the human proteome. *Nature* **2014**, *509* (7502), 575−81.
(39) Pruitt, K. D.; Brown, G. R.; Hiatt, S. M.; et al. RefSeq: an update on mammalian reference sequences. *Nucleic Acids Res.* **2014**, *42*, D756−63.
(40) Zimmer, S.; Stocker, A.; Sarbolouki, M. N.; et al. A novel human tocopherol-associated protein: cloning, in vitro expression, and characterization. *J. Biol. Chem.* **2000**, *275* (33), 25672−80.
(41) Johnson, K. G.; Kornfeld, K. The CRAL/TRIO and GOLD domain protein TAP-1 regulates RAF-1 activation. *Dev. Biol.* **2010**, *341* (2), 464−71.
(42) Pink, R. C.; Wicks, K.; Caley, D. P.; et al. Pseudogenes: pseudo-functional or key regulators in health and disease? *RNA* **2011**, *17* (5), 792−8.
(43) Mighell, A. J.; Smith, N. R.; Robinson, P. A.; et al. Vertebrate pseudogenes. *FEBS Lett.* **2000**, *468* (2−3), 109−14.
(44) Zhang, Z.; Gerstein, M. Large-scale analysis of pseudogenes in the human genome. *Curr. Opin. Genet. Dev.* **2004**, *14* (4), 328−35.
(45) Zhang, Z. D.; Frankish, A.; Hunt, T.; et al. Identification and analysis of unitary pseudogenes: historic and contemporary gene losses in humans and other primates. *Genome Biol.* **2010**, *11* (3), R26.
(46) Altschul, S. F.; Madden, T. L.; Schaffer, A. A.; et al. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res.* **1997**, *25* (17), 3389−402.
(47) Yan, R.; Xu, D.; Yang, J.; et al. A comparative assessment and analysis of 20 representative sequence alignment methods for protein structure prediction. *Sci. Rep.* **2013**, *3*, 2619.

DOI: 10.1021/acs.jproteome.5b00516
J. Proteome Res. 2015, 14, 3750−3761