The Author(s) *BMC Systems Biology* 2016, **10**(Suppl 4):113
DOI 10.1186/s12918-016-0352-6

BMC Systems Biology

RESEARCH Open Access

Identifying the missing proteins in human proteome by biological language model

Qiwen Dong[1,2*], Kai Wang[3] and Xuan Liu[4*]

From The 27th International Conference on Genome Informatics
Shanghai, China. 3-5 October 2016

Abstract

Background: With the rapid development of high-throughput sequencing technology, the proteomics research becomes a trendy field in the post genomics era. It is necessary to identify all the native-encoding protein sequences for further function and pathway analysis. Toward that end, the Human Proteome Organization lunched the Human Protein Project in 2011. However many proteins are hard to be detected by experiment methods, which becomes one of the bottleneck in Human Proteome Project. In consideration of the complicatedness of detecting these missing proteins by using wet-experiment approach, here we use bioinformatics method to pre-filter the missing proteins.

Results: Since there are analogy between the biological sequences and natural language, the *n*-gram models from Natural Language Processing field has been used to filter the missing proteins. The dataset used in this study contains 616 missing proteins from the "uncertain" category of the neXtProt database. There are 102 proteins deduced by the *n*-gram model, which have high probability to be native human proteins. We perform a detail analysis on the predicted structure and function of these missing proteins and also compare the high probability proteins with other mass spectrum datasets. The evaluation shows that the results reported here are in good agreement with those obtained by other well-established databases.

Conclusion: The analysis shows that 102 proteins may be native gene-coding proteins and some of the missing proteins are membrane or natively disordered proteins which are hard to be detected by experiment methods.

Keywords: Human proteome, Missing protein, Biological language model

Background

Proteins play important roles in biology. The Human Genome Sequence Project [1] provides a comprehensive compendium about all the human protein encoding genes. However, due to the diversity of proteins and the under-development of current proteomics technology, there are many proteins which have not been identified and annotated.

The Human Proteome Project (HPP) [2] was launched by the Human Proteome Organization (HUPO) in 2011, which contains the Chromosome-centric HPP (C-HPP) [3] and Biology/Disease-Driven HPP (B/DHPP) [4]. This project tries to identify as more proteins as possible with the goal of covering all human protein-encoding genes. This great goal is cooperated by an international associations contains 25 members [5]. The baseline metrics for the HPP contains five annually updated data resources [5]: the Ensembl database [6] provides the possible genes coding proteins; Peptide Atlas [7] and GPMdb [8] separately screen high confident proteins from mass spectrometry data; the Human Protein Atlas [9] is in charge of extracting proteins by antibody-based research; and finally neXtProt [10] collects all human proteins and assigns confidence level (PE 1-5) by protein expression evidence. Proteins at the PE1 level are identified at protein expression level by mass spectrometry,

* Correspondence: qwdong@sei.ecnu.edu.cn; xliu@shou.edu.cn
[1]Institute for Data Science and Engineering, East China Normal University, Shanghai 200062, People's Republic of China
[4]College of Engineering, Shanghai Ocean University, Shanghai 201303, People's Republic of China
Full list of author information is available at the end of the article

immunohistochemistry, 3D structure, and/or amino acid sequencing. The proteins at PE2 level is detected by transcript expression but not by protein expression. At PE3 level, there are no protein or transcript evidence, but have homologies represented in related species. Proteins at PE4 level are speculated from gene models. Finally, the protein sequences at PE5 level are generated from "dubious" or "uncertain" genes which seemed to have some protein-level evidence in the past but such identifications are doubtful by curation.

Much progress has been achieved since 2011 by the proteomics community and the HPP. Based on the curation of neXtProt [11] database, currently 82% of the protein-coding genes in human have protein expressions with high-confidence. However, there are 3, 564 genes at levels PE2-5 which have no or insufficient evidence of identification by any experimental methods and are thus named as "missing proteins" [11]. Many of these missing proteins are hard to be detected because of low abundance, poor solubility, or indistinguishable peptide sequences within protein families. The missing of such a significant amount of proteins marks a significant problem about our current understanding of the human proteome, with particularly important questions including, e.g., whether these proteins are essential to the cell functions and if yes what biological roles they play in cell and why they are not detectable by the current instruments of both transcription and translation levels. Thus identifying the missing proteins will be a challenging task.

Previous study has shown that there are analogies between biological sequences and natural language. In linguistics, some words and phrases can form a meaningful sentence; in biology, the tactic nucleotides denote gene, and the fixed protein sequences can determine its structure and function. Tsonis [12] discussed that whether DNA is a language or not. Many linguistic approaches have been used in computational biology [13–15]. Ganapathiraju et al. [16] analyzed the language feature of whole-genome protein sequence. Many techniques of Natural Language Pprocess have been used in bioinformatics, such as protein domain recognition based on language modeling [17], dictionary-driven protein annotation [18], protein remote homology detection by latent semantic analysis [19–22], identification of DNA-binding protein [23, 24], and so on.

In this study, the missing proteins in human proteome are identified by using biological language model. The amino acid *n*-gram models for human and non-human protein sequences are constructed. These models are subsequently used to discriminate whether the missing proteins are natively gene-coding proteins in human or not. The identified proteins are then analyzed by their predicted structures and functions, annotation from neXtProt database [10], HGNC database [25] and other mass spectrometry dataset [26].

Methods

Datasource

The native gene-coding proteins are downloaded from Swiss-Prot database [27]. To construct the reliable models, only the protein sequences with reviewed items are selected. Totally, there are 14565 human proteins and 70854 non-human proteins. The redundant sequences are then filtered by using CD-HIT program [28] with the sequence identity threshold of 90%. Finally, we get 14189 human proteins and 59060 non-human proteins, which are used to build the *n*-gram model for human and non-human respectively.

The "dubious" or "uncertain" missing proteins with confidence code "PE5" are extracted from the neXtProt database [10] that was released at Sep. 19, 2014. There are in total 616 proteins in this category with length ranging from 21 to 2, 252 residues. The structures of these proteins are predicted by using I-TASSER software [29]. The functions including the EC number, the GO terms and the binding sites are predicted by using COFACTOR software [30]. Both I-TASSER and COFACTOR are run in non-homology mode where all the homologous structures identified with sequence identities greater than or equal to 30% are removed. The subcellular localization is predicted by using Hum-mPLoc [31].

Biological language models to discriminate the native gene-coding human and non-human proteins

The protein sequences are composed of 20 native amino acids, while in natural language, the sentences are comprised by words. Such similarity has draw researchers attention and the language features of DNA and protein sequence have been investigated extensively [32–34]. In this study, the methods from natural language processing, especially, the statistical natural language processing methods which have been successfully solve many of the natural language task [35], are applied to identify the missing proteins. Formally, the problem of identifying missing proteins can be described as following: given a protein $P = (a_1, a_2, \ldots, a_L)$, is this protein a human protein or not:

$$P(H|P) > P(NH|P) \quad (1)$$

where $P(H|P)$ and $P(NH|P)$ represent the probability of this protein belongs to human and non-human. The above equation can be converted by conditional probability formulae:

$$\frac{P(H)P(P|H)}{P(P)} > \frac{P(NH)P(P|NH)}{P(P)} \quad (2)$$

Since the denominator is the same, it can be ignored during the comparison. $P(H)$ and $P(NH)$ are the prior probability of human and non-human proteins, and can be estimated from the training dataset by using maximum likelihood estimation based on the number of human and non-human protein. By applying the conditional probability formulae repeatedly, the probability of $P(H|P)$ or $P(NH|P)$ can be decomposed into:

$$\begin{aligned} P(P|H) &= P(a_1 \ldots a_L|H) \\ &= P(a_1|H)P(a_2 \ldots a_L|H, a_1) \\ &= P(a_1|H)P(a_2|H, a_1)P(a_3 \ldots a_L|H, a_1, a_2) \\ &= P(a_1|H)\sum_{i=2}^{L} P(a_i|a_1 \ldots a_{i-1}, H) \end{aligned} \quad (3)$$

The n-gram model supposes that the occurrence of each word is only dependent on the previous n-1 words, so the above equation can be recalculated as:

$$P(P|H) \approx \sum_{i=n}^{L} P(a_i|a_{i-n+1} \ldots a_{i-1}, H) \quad (4)$$

where the conditional probability can be estimated by maximum likelihood estimation:

$$P(a_i|a_{i-n+1} \ldots a_{i-1}) = \frac{C(a_{i-n+1} \ldots a_i)}{C(a_{i-n+1} \ldots a_{i-1})} \quad (5)$$

where $C(a_i \ldots a_j)$ is the number of occurrence of amino acid sequence $a_i..a_j$.

The same procedure can be applied to construct the n-gram model of non-human proteins. The missing proteins can be identified by using n-gram models to get whether it's native human proteins or not.

Results and discussions

Based on the n-gram models, there are 102 proteins in the neXtProt "PE5" category which have high probabilities to be native human proteins. In the following sections, these proteins are analyzed by the predicted structure and function and annotations from other databases.

The structure and function analysis of the high-probability proteins

Since the missing proteins have not been identified by experiment methods, their structures and functions are currently unknown. In this study, the structures and functions of the high-probability proteins inferred by n-gram models are predicted by I-TASSER and COFACTOR software and the confidence scores outputted by the software are used to indicate the reliability of the prediction. The I-TASSER confidence score (C-score) is computed by the accuracy of the threading programs and the simulation results of the structural assembly process. The value range is between -5 and 2, where the higher the C-score value is, the more confident the corresponding model is, and vice-versa. A model with I-TASSER C-score larger than -1.5 means that the structure topology is correct. The confidence score (C-score) of COFACTOR is calculated based on the confidence score of the structure prediction and the similarity between the predicted models and their native structures in the PDB. The COFACTOR C-score are normalized between 0 and 1, where a large value indicates a good prediction. Figure 1 shows the distribution of I-TASSER and COFACTOR C-score of the 102 high -probability missing proteins. The number of foldable proteins (with C-score higher than -1.5) are less than that of the un-foldable proteins (with C-score lower than -1.5) which is also confirmed in our previous study

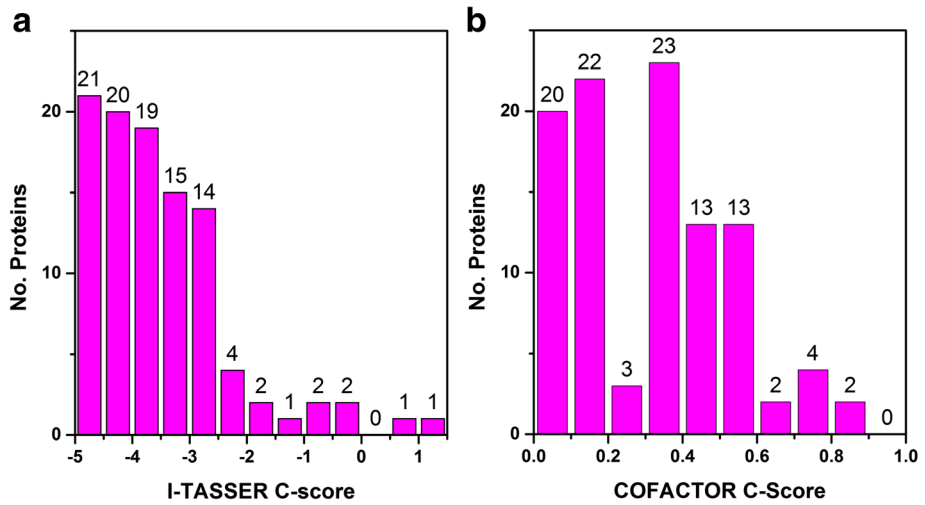

Fig. 1 The distribution of I-TASSER (**a**) and COFACTOR (**b**) C-score of the identified missing proteins

about missing proteins [36]. The reason for this phenomenon may be that the missing proteins are not gene-coding proteins or there is no homology templates used during prediction. Based on the results of structure prediction, there are 7 foldable proteins whose I-TASSER C-score are larger than the foldable threshold (-1.5). These proteins have good structure models with no homologous templates used during prediction, which means that they may be gene-coding proteins. Most of the COFACTOR C-scores are distributed between 0.3 and 0.6, where 8 proteins have very high COFACTOR C-score. As shown in the figure, the COFACTOR C-score for most of the missing proteins are larger than 0.2, which indicates a good function prediction based on experience evaluation.

Structural topology analyses of the I-TASSER models

The SCOPe library [37] is used as the classification criterion of structure topology, which is an extended structure library integrated from the standard SCOP [38] and ASTRAL [39] databases. The structure class of the I-TASSER model is assigned as the corresponding structure class of the SCOPe domain which has the highest structural similarity with the model. The structure alignment program TM-align [40] is used to calculate the TM-score between the I-TASSER model and all structural domains from SCOPe. If there are multiple domains in target model, we selected the domain that has the maximum TM-score to SCOPe domain. Figure 2 shows the distribution of the SCOPe class of the identified missing proteins. It is interesting that some missing proteins have structure topology in 'membrane and cell surface proteins and peptides' and 'coiled coil proteins' class. Such phenomenon is reasonable since these kind of proteins are difficult to be identified by experiment method.

Evaluation of the function base on gene ontology

The GO molecular function of the high-probability missing proteins is predicted by the COFACTOR package and the number in each GO item is shown in Fig. 3. As shown in the Fig. 3 GO terms come from the first level of GO molecular function. Most of the high-probability missing proteins have the GO function of 'binding' (GO:0005488) and 'catalytic activity' (GO:0003824). However some missing proteins may be membrane proteins since they have the GO function of 'transporter activity' (GO:0005215) and 'receptor activity' (GO:0004872). The results are consistent with the structural topology analysis in which there are many membrane proteins in the missing proteins.

Comparision of subcellular localization

The subcellular localizations of proteins are critical for their biological functions. The Hum-mPLoc 2.0 program [31] is used to predict the subcellular localizations of the high-probability missing proteins. The types of subcellular localizations and the number of proteins in each types is illustrated in Fig. 4. Most of the proteins are located at extracellular and nucleus. The missing proteins are also observed at plasma membrane, which is also confirmed by the results from the structural topology analysis and function predictions.

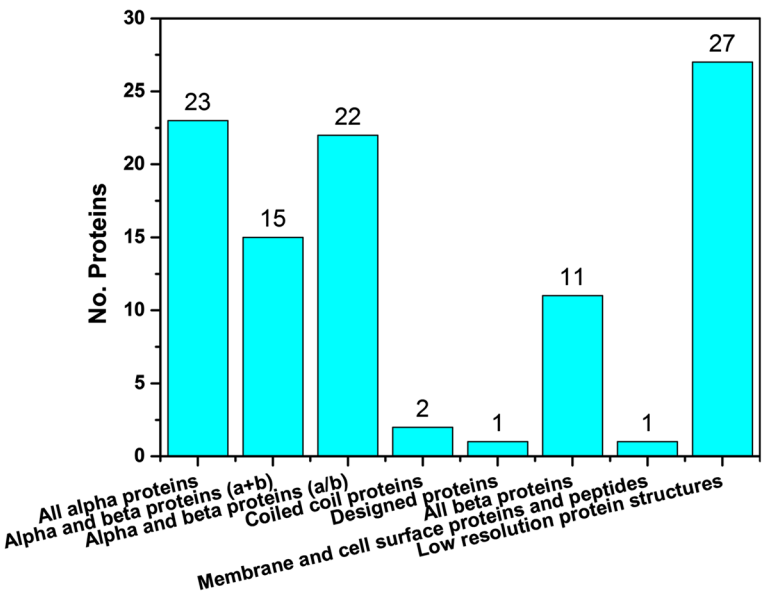

Fig. 2 Relative frequency distribution of SCOP classes for the identified missing proteins

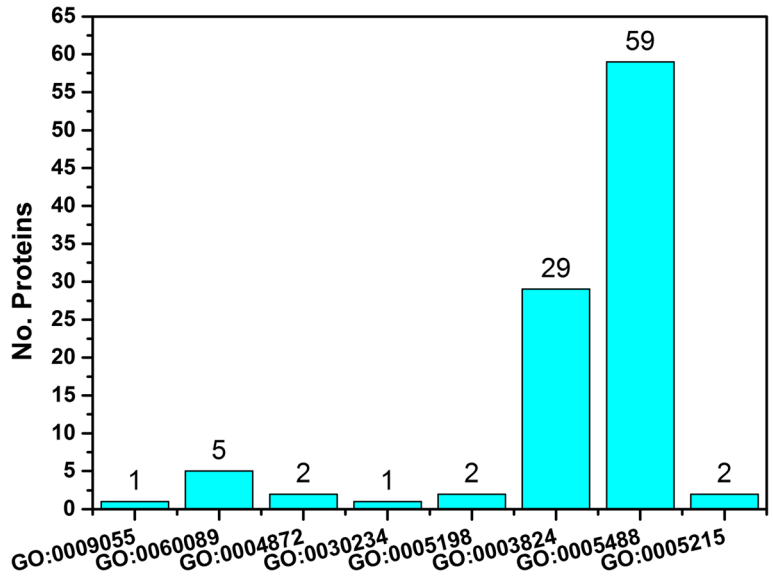

Fig. 3 The distribution of predicted GO items from the fist level of molecular function

HGNC mapping analysis

The HGNC [41] database is in charge of assigning a unique symbol and name for each gene loci from human genome. Most of the HGNC data are manually collected and carefully checked [41]. The information from HGNC gene loci provides valuable resource to identify the missing proteins. Based on the Gene mapping, 71 out of the 102 high-probability missing proteins can be mapped to one or more HGNC items. We collected the corresponding gene loci types for the 71 missing proteins and counted the number of missing proteins in each loci type. The results are shown in Table 1. There are 9 proteins confirmed by HGNC with gene loci type "gene with protein product". There are 26 pseudogenes. Since pseudogenes are the products of evolution. They usually have homologous proteins. That's the reason why there are many pseudogenes in the missing proteins.

Consistence analysis with other mass spectrometry dataset

Mass spectrometry is currently one of the efficient method to identify protein peptides. There are many mass spectrometry data deposited in public database, such as PeptideAtlas [42] and GPMDB [8]. The sketch

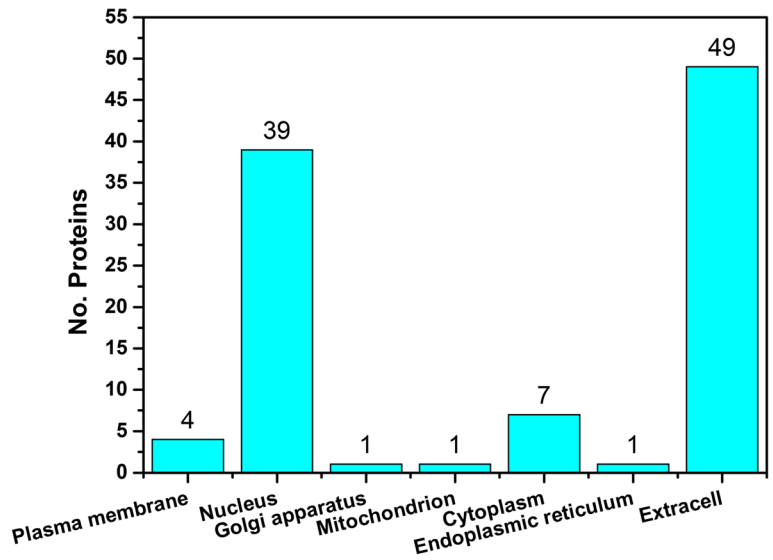

Fig. 4 The distribution of subcellular localizations for the identified missing proteins

Table 1 The gene loci types and the number of proteins for the high-probability missing proteins after hgnc mapping

Gene loci type	No. missing proteins
Gene with protein product	9
Pseudogene	26
RNA, long non-coding	30
Unknown	6

of human proteome is drawing by mass spectrometry data [26, 43]. Recently Kim et al [26] reported that about two-third (2535/3844) of the "missing proteins" [11] have been identified. Actually the "missing proteins" used by Kim et al. are constituted by the neXtProt proteins with evidence codes of "PE2", "PE3" or "PE4". This paper aims to identify the "PE5" missing proteins. By RefSeq [44] mapping, we found that there are 41 "PE5" proteins which are also in Kim's dataset. Among these 41 missing proteins, there are 6 proteins are foldable based on the structure prediction results in non-homology mode. These results indicate that our finding are in good consistent with Kim's results.

Conclusion

In this study, the human gene-coding proteins currently undetected are identified by using biological language models. The amino acid *n*-gram models of human and non-human proteins are constructed. These models are then used to identify the "uncertain" missing proteins with evidence code "PE5" from neXtProt database. The results show that 102 high probability proteins may be gene-coding proteins. The structure, function and subcellular localization of these proteins are then inferred by using the advanced programs. The identified missing proteins are then analyzed with the annotation from other database. Without using homology templates, 7 proteins have correct structure topology with I-TASSER C-score larger than -1.5. The predicted functions are mainly within GO items 'binding' (GO:0005488) and 'catalytic activity' (GO:0003824). 9 missing proteins are confirmed by HGNC with gene loci type "gene with protein product". 6 missing proteins are also detected by mass spectrometry experiment. The analysis also shows that many of the unknown proteins are membrane or natively disordered proteins which are difficult to be detected. The identified missing proteins need to be further validated by experimental approach. The results in this study provides valuable complementary resource for the human proteome.

Declarations

This article has been published as part of BMC Systems Biology Volume 10, Supplement 4, 2016: Proceedings of the 27th International Conference on Genome Informatics: systems biology. The full contents of the supplement are available online at http://bmcsystbiol.biomedcentral.com/articles/supplements/volume-10-supplement-4.

Funding

Financial support funding for publication costs was provided by the National Key Research and Development Program of China (Grant No. 2016YFB1000905) and National Natural Science Foundation of China (Grant No. 61672234, U1401256 and 61402177).

Availability of data and materials

The datasets of the current study are freely available by requesting on the corresponding authors.

Authors' contributions

DQW performed the experiment and wrote the paper. WK proposed the idea and polished the paper. LX designed the architecture and analyzed the results. All authors read and approved the paper.

Competing interests

The authors declare that they have no competing interest.

Consent for publication

Not applicable.

Ethics approval and consent to participate

Not applicable.

Author details

[1]Institute for Data Science and Engineering, East China Normal University, Shanghai 200062, People's Republic of China. [2]Key Laboratory of Network Oriented Intelligent Computation, Harbin Institute of Technology Shenzhen Graduate School, Shenzhen 518055, People's Republic of China. [3]College of Animal Science and technology, Jilin Agricultural University, Changchun 130118, People's Republic of China. [4]College of Engineering, Shanghai Ocean University, Shanghai 201303, People's Republic of China.

Published: 23 December 2016

References

1. Venter JC, Adams MD, Myers EW, Li PW, Mural RJ, Sutton GG, Smith HO, Yandell M, Evans CA, Holt RA, et al. The sequence of the human genome. Science. 2001;291(5507):1304–51.
2. Legrain P, Aebersold R, Archakov A, Bairoch A, Bala K, Beretta L, Bergeron J, Borchers CH, Corthals GL, Costello CE, et al. The human proteome project: current state and future direction. Mol Cell Proteomics. 2011;10(7):M111 009993.
3. Paik YK, Jeong SK, Omenn GS, Uhlen M, Hanash S, Cho SY, Lee HJ, Na K, Choi EY, Yan F, et al. The Chromosome-Centric Human Proteome Project for cataloging proteins encoded in the genome. Nat Biotechnol. 2012;30(3):221–3.
4. Aebersold R, Bader GD, Edwards AM, van Eyk JE, Kussmann M, Qin J, Omenn GS. 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.
5. Marko-Varga G, Omenn GS, Paik YK, Hancock WS. A first step toward completion of a genome-wide characterization of the human proteome. J Proteome Res. 2013;12(1):1–5.
6. Flicek P, Amode MR, Barrell D, Beal K, Billis K, Brent S, Carvalho-Silva D, Clapham P, Coates G, Fitzgerald S, et al. Ensembl 2014. Nucleic Acids Res. 2014;42(Database issue):D749–755.
7. Farrah T, Deutsch EW, Hoopmann MR, Hallows JL, Sun Z, Huang CY, Moritz RL. The state of the human proteome in 2012 as viewed through PeptideAtlas. J Proteome Res. 2013;12(1):162–71.
8. Craig R, Cortens JP, Beavis RC. Open source system for analyzing, validating, and storing protein identification data. J Proteome Res. 2004;3(6):1234–42.
9. Uhlen M, Oksvold P, Fagerberg L, Lundberg E, Jonasson K, Forsberg M, Zwahlen M, Kampf C, Wester K, Hober S, et al. Towards a knowledge-based Human Protein Atlas. Nat Biotechnol. 2010;28(12):1248–50.
10. Lane L, Argoud-Puy G, Britan A, Cusin I, Duek PD, Evalet O, Gateau A, Gaudet P, Gleizes A, Masselot A, et al. neXtProt: a knowledge platform for human proteins. Nucleic Acids Res. 2012;40(Database issue):D76–83.
11. Lane L, Bairoch A, Beavis RC, Deutsch EW, Gaudet P, Lundberg E, Omenn GS. Metrics for the Human Proteome Project 2013-2014 and strategies for finding missing proteins. J Proteome Res. 2014;13(1):15–20.

12. Tsonis AA, Elsner JB, Tsonis PA. Is DNA a language? J Theor Biol. 1997;184(1):25–9.
13. Dyrka W, Nebel JC. A stochastic context free grammar based framework for analysis of protein sequences. BMC Bioinformatics. 2009;10:323.
14. Ganapathiraju M, Balakrishnan N, Reddy R, Klein-Seetharaman J. Computational Biology and Language. Ambient Intell Sci Discov LNAI. 2005;3345:25–47.
15. Searls DB. Linguistic approaches to biological sequences. Comput Appl Biosci. 1997;13(4):333–44.
16. Ganapathiraju M, Weisser D, Rosenfeld R, Carbonell J, Reddy R, Klein-Seetharaman J: Comparative n-gram analysis of whole-genome protein sequences. In: Proceedings of the second international conference on Human Language Technology Research. San Diego: Morgan Kaufmann Publishers Inc; 2002. pp. 76–81.
17. Coin L, Bateman A, Durbin R. Enhanced protein domain discovery by using language modeling techniques from speech recognition. Proc Natl Acad Sci. 2003;100(8):4516–20.
18. Rigoutsos I, Huynh T, Floratos A, Parida L, Platt D. Dictionary-driven protein annotation. Nucleic Acids Res. 2002;30(17):3901–16.
19. Dong Q-W, Wang X-L, Lin L. Application of latent semantic analysis to protein remote homology detection. Bioinformatics. 2006;22(3):285–90.
20. Liu B, Wang X, Lin L, Dong Q. A discriminative method for protein remote homology detection and fold recognition combining Top-n-grams and latent semantic analysis. BMC Bioinformatics. 2008;9:510.
21. Liu B, Xu J, Zou Q, Xu R, Wang X, Chen Q. Using distances between Top-n-gram and residue pairs for protein remote homology detection. BMC Bioinformatics. 2014;Suppl 2:S3.
22. Liu B, Zhang D, Xu R, Xu J, Wang X, Chen Q, Dong Q, Chou KC. Combining evolutionary information extracted from frequency profiles with sequence-based kernels for protein remote homology detection. Bioinformatics. 2014;30(4):472–9.
23. Liu B, Xu J, Fan S, Xu R, Zhou J, Wang X. PseDNA-Pro: DNA-Binding Protein Identification by Combining Chou's PseAAC and Physicochemical Distance Transformation. Mol Inform. 2015;34(1):8–17.
24. Liu B, Xu J, Lan X, Xu R, Zhou J, Wang X, Chou KC. iDNA-Prot|dis: identifying DNA-binding proteins by incorporating amino acid distance-pairs and reduced alphabet profile into the general pseudo amino acid composition. PLoS ONE. 2014;9(9):e106691.
25. Gray KA, Yates B, Seal RL, Wright MW, Bruford EA. Genenames. org: the HGNC resources in. Nucleic Acids Res. 2015;43(D1):D1079–85.
26. Kim M-S, Pinto SM, Getnet D, Nirujogi RS, Manda SS, Chaerkady R, Madugundu AK, Kelkar DS, Isserlin R, Jain S. A draft map of the human proteome. Nature. 2014;509(7502):575–81.
27. Consortium U. UniProt: a hub for protein information. Nucleic Acids Res. 2015;43(Database issue):D204–212.
28. Fu L, Niu B, Zhu Z, Wu S, Li W. CD-HIT: accelerated for clustering the next-generation sequencing data. Bioinformatics. 2012;28(23):3150–2.
29. 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.
30. Roy A, Yang J, Zhang Y. COFACTOR: an accurate comparative algorithm for structure-based protein function annotation. Nucleic Acids Res. 2012;40(Web Server issue):W471–477.
31. Shen HB, Chou KC. A top-down approach to enhance the power of predicting human protein subcellular localization: Hum-mPLoc 2.0. Anal Biochem. 2009;394(2):269–74.
32. Liu B, Liu F, Fang L, Wang X, Chou KC. repDNA: a Python package to generate various modes of feature vectors for DNA sequences by incorporating user-defined physicochemical properties and sequence-order effects. Bioinformatics. 2015;31(8):1307–9.
33. Liu B, Liu F, Wang X, Chen J, Fang L, Chou KC. Pse-in-One: a web server for generating various modes of pseudo components of DNA, RNA, and protein sequences. Nucleic Acids Res. 2015;43(W1):W65–71.
34. Liu B, Liu F, Fang L, Wang X, Chou KC. repRNA: a web server for generating various feature vectors of RNA sequences. Mol Genet Genomics. 2016;291(1):473–81.
35. Hristea FT. Statistical Natural Language Processing. In: International Encyclopedia of Statistical Science. Heidelberg: Springer; 2011. pp. 1452–1453
36. Dong Q, Menon R, Omenn GS, Zhang Y. Structural Bioinformatics Inspection of neXtProt PE5 Proteins in the Human Proteome. J Proteome Res. 2015;14(9):3750–61.
37. Fox NK, Brenner SE, 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):D304–9.
38. Murzin AG, Brenner SE, Hubbard T, Chothia C. SCOP: a structural classification of proteins database for the investigation of sequences and structures. J Mol Biol. 1995;247(4):536–40.
39. Chandonia JM, Hon G, Walker NS, Lo Conte L, Koehl P, Levitt M, Brenner SE. The ASTRAL Compendium in 2004. Nucleic Acids Res. 2004;32(Database issue):D189–192.
40. Zhang Y, Skolnick J. TM-align: a protein structure alignment algorithm based on the TM-score. Nucleic Acids Res. 2005;33(7):2302–9.
41. Gray KA, Yates B, Seal RL, Wright MW, Bruford EA. Genenames.org: the HGNC resources in 2015. Nucleic Acids Res. 2015;43(Database issue):D1079–85.
42. Desiere F, Deutsch EW, King NL, Nesvizhskii AI, Mallick P, Eng J, Chen S, Eddes J, Loevenich SN, Aebersold R. The PeptideAtlas project. Nucleic Acids Res. 2006;34(Database issue):D655–658.
43. Wilhelm M, Schlegl J, Hahne H, Moghaddas Gholami A, Lieberenz M, Savitski MM, Ziegler E, Butzmann L, Gessulat S, Marx H, et al. Mass-spectrometry-based draft of the human proteome. Nature. 2014;509(7502):582–7.
44. Pruitt KD, Brown GR, Hiatt SM, Thibaud-Nissen F, Astashyn A, Ermolaeva O, Farrell CM, Hart J, Landrum MJ, McGarvey KM, et al. RefSeq: an update on mammalian reference sequences. Nucleic Acids Res. 2014;42(Database issue):D756–763.