ICBAA2017-32

UNLOCKING THE UNKNOWNS IN PLANT GENOME VIA *IN SILICO* APPROACHES

Nur Syamimi Yusoff[1], Suhaila Sulaiman[1] and Lee Yang Ping[1]

[1]Felda Global Ventures Research and Development Sdn. Bhd., FGV Innovation Centre (Biotechnology), PT 23417 Lengkuk Teknologi, 71760 Bandar Enstek, Negeri Sembilan, Malaysia
syamimi.y@feldaglobal.com

Abstract: As we are entering the post genomic era, there are vast amount of newly sequenced genomes become available every day to be accessed from all over the world. However, it becomes a hindrance when only 50 to 60% of the genome content are accessible, while the rest remains unknown. A survey to an oil palm genome draft indicates more than 30% of proteins encoded by gene models that were annotated as ‘hypothetical proteins’, ‘uncharacterized’ and ‘unknown’. The large number of unknown proteins might contribute as important role in plant-related studies. Unfortunately, this valuable information is often overlooked by researchers due to the limitation of functional annotation tools to infer the genes function. Moreover, the conventional experiments in protein function assignment are not practical due to the high cost and laborious tasks. Despite all these caveats, here we recommend a series of *in silico* analyses and workflow to be implemented to unravel the unknowns based on current approaches and software. The proposed workflow consists of sequence and structural *in silico* analyses which applicable to unknown proteins from all type of organisms.

Keywords: hypothetical proteins, sequence, structure.

INTRODUCTION

Since the *Arabidopsis thaliana* genome was published more than 10 years ago, the number of sequenced crop genomes every year have keep on rising to sustain the plant-related studies with the aid of latest genomic technologies. As the number of sequenced plant genomes increase so do the number genes encode unknown proteins. These genes are unknown either because they encode new proteins that unique to the organism or due to insufficient information available in public database to match the sequences. Connecting the sequence with the biological values is not an easy task, thus researchers tend to ‘ignore’ these unknowns. These limitations urge the researchers to implement various computational tools and software as an alternative to save time and cost as the laboratory validation is more expensive and time consuming. Based on the oil palm (*Elaeis guineensis*) genome draft annotation, more than 30% of genes were identified as unknown by using ‘hypothetical protein’, ‘unknown’ or ‘uncharacterized’ keywords. These unknown will become a hindrance in the oil palm studies, hence, can affect the oil palm industry especially in Malaysia. In this paper, we suggest a series of *in silico* analyses that can be applied to uncover the biological roles of the unknown proteins. The framework includes the sequence search analyses, physicochemical analysis, sequence characterization analyses, multiple sequence alignment, domain and motif profile identification, phylogenetic tree construction, structure prediction and validation and patterns of amino acid side chains search in 3D space.

METHODS

Identification of hypothetical proteins

The annotated hypothetical proteins in the oil palm genome were distinguished using ‘hypothetical protein’, ‘unknown’, ‘unnamed’ or ‘uncharacterized’ keywords. All hypothetical proteins were selected to search for homologies using blast program.

Homology search

Sequence search against public databases is the initial step towards annotating the function of hypothetical protein. This was done using BLASTP program (Camacho et al., 2009) from NCBI’s Swiss-Prot database to detect any sequence or structural homologs of the queries. Proteins exhibit query coverage more than 30%, e-value less than 1e^{-5} and high sequence identities (>30%) with the queries are considered as close homologs. True hypothetical proteins are referring to protein queries which all the close homologs consist of annotated ‘hypothetical protein’, ‘unknown’ and ‘uncharacterized’ proteins or protein queries with no matches at all.

Domain/motif search

The presence of domain or motif in hypothetical protein sequence and classification of hypothetical protein sequence into families are performed using InterProScan. InterProScan is a software that integrates several different databases for protein signature search (Finn et al., 2017), such as CATH-Gene3D (Lam et al., 2016), HAMAP (Pedruzzi et al., 2015), CDD (Marchler-Bauer et al., 2015), PANTHER (Mi, Poudel, Muruganujan, Casagrande, & Thomas, 2016), Pfam (Finn et al., 2016), PIRSF (Wu et al., 2004), PRINTS (Attwood et al., 2012), ProDom (Bru et al., 2005), PROSITE Patterns (Sigrist et al., 2013), PROSITE profiles (Sigrist et al., 2013), SUPERFAMILY (Oates et al., 2015), TIGRFAMs (Haft et al., 2013), SMART (Letunic, Doerks, & Bork, 2015) and SFLD (Akiva et al., 2014).

Sequence analyses

Sequence analyses of hypothetical protein involve physicochemical analysis, prediction of signal peptide, sub-cellular localization and transmembrane helices. The physicochemical analysis is carried out using ProtParam tool in Expasy Server (Gasteiger et al., 2005), TargetP for prediction of signal peptide and sub-cellular localization (Emanuelsson, Brunak, von Heijne, & Nielsen, 2007) and TMHMM Server for transmembrane helices prediction (Moller, Croning, & Apweiler, 2001).

Multiple sequence alignment

The alignment between protein queries and their close homologs can be done using MUSCLE software (Edgar, 2004) embedded in Jalview platform (Waterhouse, Procter, Martin, Clamp, & Barton, 2009). The aligned output will be used for phylogenetic tree construction.

Phylogenetic tree construction

Different phylogenetic tree can be constructed from PHYLIP version 3.696 package (Shimada & Nishida, 2017) using different series of programs according to type of the tree, for example Neighbor-Joining (NJ) tree use protdist and neighbor programs, while Maximum-Parsimony (MP) and Maximum-Likelihood (ML) trees use protpars and protml program respectively. Seqboot and consense programs in PHYLIP package used for bootstrap analysis with 1000 bootstrap value. The generated consensus tree was viewed using MEGA 6.0 software (Tamura, Stecher, Peterson, Filipski, & Kumar, 2013).

Structure prediction

There are two ways of predicting the 3D structure of hypothetical protein**;** i) *ab initio* modelling using I-TASSER (Zhang, 2008) or Rosetta software (Rohl, Strauss, Misura, & Baker, 2004) when there is no structure homologues available; ii) homology modelling using SWISS-MODEL (Schwede, Kopp, Guex, & Peitsch, 2003) or MODELLER (Webb & Sali, 2014) if the template is available.

Structure validation

The predicted structure need to be validated using structure validation programs or servers such as PROCHECK (Laskowski, MacArthur, Moss, & Thornton, 1993), ERRAT (Colovos & Yeates, 1993) and Verify3D (Bowie, Luthy, & Eisenberg, 1991; Lüthy, Bowie, & Eisenberg, 1992) to ensure the protein 3D structure has a good quality.

Regardless of no sequence and fold similarities, the 3D patterns of amino acids side chains can be utilized to infer the actual function hypothetical protein. For this purpose, two different programs with different approach can be used; SPRITE (Nadzirin, Gardiner, Willett, Artymiuk, & Firdaus-Raih, 2012) and ASSAM (Spriggs, Artymiuk, & Willett, 2003). SPRITE program takes 3D structure as an input to identify functional sites against databases of sites while ASSAM program exploit the 3D amino acid pattern for search against protein structures database.

RESULTS & DISCUSSION

The in silico workflow for annotating the hypothetical functions in plant genome that we proposed in this paper basically comprises of two major parts: sequence and structure. In term of sequence, the workflow starts with identification of hypothetical proteins through annotation by using 'hypothetical protein', 'unknown' and 'uncharacterized' keywords. The close homologs of annotated hypothetical proteins were recognized from Swiss-Prot database through BLASTP program. Well defined homologs are needed for hypothetical proteins' homology search as a first step towards understanding the actual function of hypothetical proteins. Thus, Swiss-Prot is the most reliable database as it is a manually curated protein sequence database and it has its own advantages over other public databases. The advantages include minimal redundancy of sequences, incorporation of additional annotations and integration with other databases such as EMBL, PDB, OMIM, Pfam and PROSITE (Bairoch & Apweiler, 1999).

The features of hypothetical proteins were characterized through identification of domain or motif, physicochemical characteristics, signal peptide, sub-cellular localization and transmembrane helices. Observation on the evolution of the hypothetical protein with its close homologs is referred to the phylogenetic protein tree constructed.

Normally not so much information can be obtained through sequence analyses of hypothetical protein. Hence, structural analyses are required since similar protein structures usually behave in similar manners (Lee, Redfern, & Orengo, 2007). In this workflow, we incorporate the 3D motif recognition-based for functional inference. If an unknown protein possesses known domain or motif, it will become easier as the site will carry the protein's function. For this purpose, two graph theoretical programs can be utilized, SPRITE and ASSAM. This two programs were developed using different approaches but they can be sequentially used if needed (Nadzirin, Gardiner, Willett, Artymiuk, & Firdaus-Raih, 2012).

SPRITE takes the 3D protein structure to search for protein sites in the sites' databases which were characterized from X-ray crystallographic structures in PDB. SPRITE, on the other hand uses the coordinates of 3D motif for search against protein structures database. Referring to Figure 1, if the structure of hypothetical protein has unknown motif predicted from sequence analysis, coordinate of the 3D motif will be search against databases in ASSAM. Else, if the structure of hypothetical protein does not has any predicted motif, then the whole structure will be used in SPRITE for functional sites identification. If the SPRITE returns the unknown functional site of the query, then ASSAM program will takes the turn to search the occurrence of the unknown motif in any protein structure.

CONCLUSIONS

The abundance of hypothetical proteins in crop genomes should not be ignored and intense workflow is needed in order to unravel the function. Combination of sequence and structural analyses discussed in this paper especially the 3D functional site search might help researchers out there to lessen the burden of the unknowns.

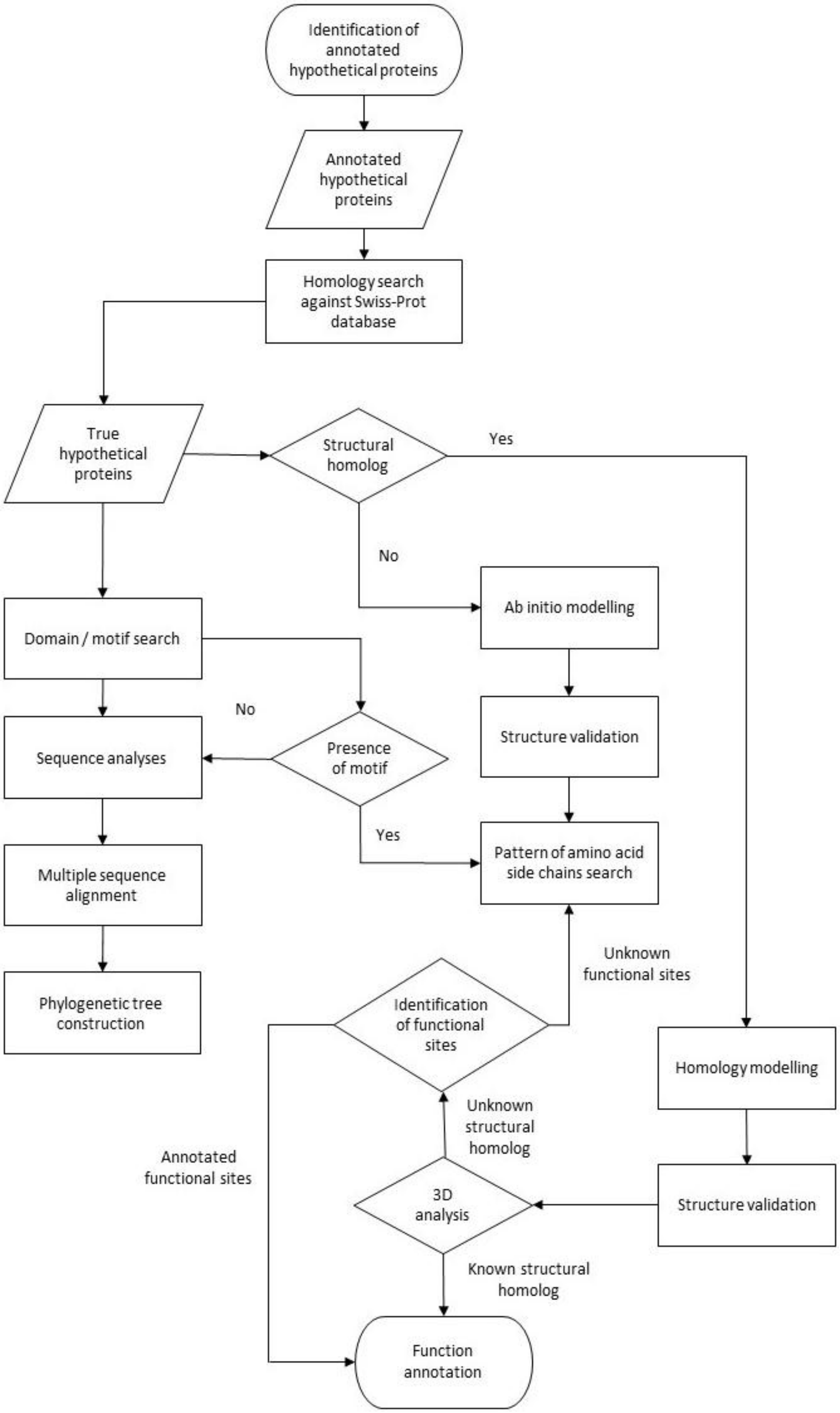
Identification of annotated hypothetical proteins
Annotated hypothetical proteins
Homology search against Swiss-Prot database
True hypothetical proteins
Structural homolog
Yes
No
Ab initio modelling
Domain / motif search
No
Presence of motif
Sequence analyses
Structure validation
Yes
Pattern of amino acid side chains search
Multiple sequence alignment
Phylogenetic tree construction
Unknown functional sites
Identification of functional sites
Homology modelling
Unknown structural homolog
Annotated functional sites
3D analysis
Structure validation
Known structural homolog
Function annotation

Figure 1. The proposed *in silico* workflow of annotating hypothetical proteins in plant genomes 3-D structure analysis.

REFERENCES

Akiva, E., Brown, S., Almonacid, D. E., Barber, A. E., Custer, A. F., Hicks, M. A., … Babbitt, P. C. (2014). The Structure-Function Linkage Database. Nucleic Acids Research, 42(D1). https://doi.org/10.1093/nar/gkt1130.

Attwood, T. K., Coletta, A., Muirhead, G., Pavlopoulou, A., Philippou, P. B., Popov, I., … Mitchell, A. L. (2012). The PRINTS database: A fine-grained protein sequence annotation and analysis resource-its status in 2012. Database, 2012. https://doi.org/10.1093/database/bas019.

Bairoch, A., & Apweiler, R. (1999). The SWISS-PROT protein sequence data bank and its supplement TrEMBL in 1999. Nucleic Acids Research. https://doi.org/10.1093/nar/27.1.49.

Bowie, J., Luthy, R., & Eisenberg, D. (1991). A method to identify protein sequences that fold into a known three-dimensional structure. Science, 253(5016), 164–170. https://doi.org/10.1126/science.1853201.

Bru, C., Courcelle, E., Carrère, S., Beausse, Y., Dalmar, S., & Kahn, D. (2005). The ProDom database of protein domain families: More emphasis on 3D. Nucleic Acids Research, 33(DATABASE ISS.). https://doi.org/10.1093/nar/gki034.

Camacho, C., Coulouris, G., Avagyan, V., Ma, N., Papadopoulos, J., Bealer, K., & Madden, T. L. (2009). BLAST+: architecture and applications. BMC Bioinformatics, 10(1), 421. https://doi.org/10.1186/1471-2105-10-421.

Colovos, C., & Yeates, T. O. (1993). Verification of protein structures: Patterns of nonbonded atomic interactions. Protein Science, 2(9), 1511–1519. https://doi.org/10.1002/pro.5560020916.

Edgar, R. C. (2004). MUSCLE: a multiple sequence alignment method with reduced time and space complexity. BMC Bioinformatics, 5, 113. https://doi.org/10.1186/1471-2105-5-113.

Emanuelsson, O., Brunak, S., von Heijne, G., & Nielsen, H. (2007). Locating proteins in the cell using TargetP, SignalP and related tools. Nature Protocols, 2(4), 953–971. https://doi.org/10.1038/nprot.2007.131.

Finn, R. D., Attwood, T. K., Babbitt, P. C., Bateman, A., Bork, P., Bridge, A. J., … Mitchell, A. L. (2017). InterPro in 2017 - beyond protein family and domain annotations. Nucleic Acids Research, 45, D190–D199. https://doi.org/10.1093/nar/gkw1107.

Finn, R. D., Coggill, P., Eberhardt, R. Y., Eddy, S. R., Mistry, J., Mitchell, A. L., … Bateman, A. (2016). The Pfam protein families database: Towards a more sustainable future. Nucleic Acids Research, 44(D1), D279–D285. https://doi.org/10.1093/nar/gkv1344.

Gasteiger, E., Hoogland, C., Gattiker, A., Duvaud, S., Wilkins, M. R., Appel, R. D., & Bairoch, A. (2005). Protein Identification and Analysis Tools on the ExPASy Server. In The Proteomics Protocols Handbook (pp. 571–607). https://doi.org/10.1385/1592598900.

Haft, D. H., Selengut, J. D., Richter, R. A., Harkins, D., Basu, M. K., & Beck, E. (2013). TIGRFAMs and genome properties in 2013. Nucleic Acids Research, 41(D1). https://doi.org/10.1093/nar/gks1234.

Lam, S. D., Dawson, N. L., Das, S., Sillitoe, I., Ashford, P., Lee, D., … Lees, J. G. (2016). Gene3D: Expanding the utility of domain assignments. Nucleic Acids Research, 44(D1), D404–D409. https://doi.org/10.1093/nar/gkv1231.

Laskowski, R. A., MacArthur, M. W., Moss, D. S., & Thornton, J. M. (1993). PROCHECK: a program to check the stereochemical quality of protein structures. Journal of Applied Crystallography, 26(2), 283–291. https://doi.org/10.1107/S0021889892009944.

Lee, D., Redfern, O., & Orengo, C. (2007). Predicting protein function from sequence and structure. Nature Reviews Molecular Cell Biology, 8(12), 995–1005. https://doi.org/10.1038/nrm2281.

Letunic, I., Doerks, T., & Bork, P. (2015). SMART: Recent updates, new developments and status in 2015. Nucleic Acids Research, 43(D1), D257–D260. https://doi.org/10.1093/nar/gku949.

Lüthy, R., Bowie, J. U., & Eisenberg, D. (1992). Assessment of protein models with three-dimensional profiles. Nature, 356(6364), 83–85. https://doi.org/10.1038/356083a0.

Marchler-Bauer, A., Derbyshire, M. K., Gonzales, N. R., Lu, S., Chitsaz, F., Geer, L. Y., … Bryant, S. H. (2015). CDD: NCBI's conserved domain database. Nucleic Acids Research, 43(D1), D222–D226. https://doi.org/10.1093/nar/gku1221.

Mi, H., Poudel, S., Muruganujan, A., Casagrande, J. T., & Thomas, P. D. (2016). PANTHER version 10: Expanded protein families and functions, and analysis tools. Nucleic Acids Research, 44(D1), D336–D342. https://doi.org/10.1093/nar/gkv1194.

Moller, S., Croning, M. D. R., & Apweiler, R. (2001). Evaluation of methods for the prediction of membrane spanning regions. Bioinformatics, 17(7), 646–653. https://doi.org/10.1093/bioinformatics/17.7.646.

Nadzirin, N., Gardiner, E. J., Willett, P., Artymiuk, P. J., & Firdaus-Raih, M. (2012). SPRITE and ASSAM: Web servers for side chain 3D-motif searching in protein structures. Nucleic Acids Research, 40(W1). https://doi.org/10.1093/nar/gks401.

Oates, M. E., Stahlhacke, J., Vavoulis, D. V., Smithers, B., Rackham, O. J. L., Sardar, A. J., … Gough, J. (2015). The SUPERFAMILY 1.75 database in 2014: A doubling of data. Nucleic Acids Research, 43(D1), D227–D233. https://doi.org/10.1093/nar/gku1041.

Pedruzzi, I., Rivoire, C., Auchincloss, A. H., Coudert, E., Keller, G., De Castro, E., … Bridge, A. (2015). HAMAP in 2015: Updates to the protein family classification and annotation system. Nucleic Acids Research, 43(D1), D1064–D1070. https://doi.org/10.1093/nar/gku1002.

Rohl, C. A., Strauss, C. E. M., Misura, K. M. S., & Baker, D. (2004). Protein Structure Prediction Using Rosetta. Methods in Enzymology, 383(2003), 66–93. https://doi.org/10.1016/S0076-6879(04)83004-0.

Schwede, T., Kopp, J., Guex, N., & Peitsch, M. C. (2003). SWISS-MODEL: An automated protein homology-modeling server. Nucleic Acids Research, 31(13), 3381–3385. https://doi.org/10.1093/nar/gkg520.

Shimada, M. K., & Nishida, T. (2017). A modification of the PHYLIP program: A solution for the redundant cluster problem, and an implementation of an automatic bootstrapping on trees inferred from original data. Molecular Phylogenetics and Evolution, 109, 409–414. https://doi.org/10.1016/j.ympev.2017.02.012.

Sigrist, C. J. A., De Castro, E., Cerutti, L., Cuche, B. A., Hulo, N., Bridge, A., … Xenarios, I. (2013). New and continuing developments at PROSITE. Nucleic Acids Research, 41(D1). https://doi.org/10.1093/nar/gks1067.

Spriggs, R. V., Artymiuk, P. J., & Willett, P. (2003). Searching for patterns of amino acids in 3D protein structures. In Journal of Chemical Information and Computer Sciences (Vol. 43, pp. 412–421). https://doi.org/10.1021/ci0255984.

Tamura, K., Stecher, G., Peterson, D., Filipski, A., & Kumar, S. (2013). MEGA6: Molecular evolutionary genetics analysis version 6.0. Molecular Biology and Evolution, 30(12), 2725–2729. https://doi.org/10.1093/molbev/mst197.

Waterhouse, A. M., Procter, J. B., Martin, D. M. A., Clamp, M., & Barton, G. J. (2009). Jalview Version 2-A multiple sequence alignment editor and analysis workbench. Bioinformatics, 25(9), 1189–1191. https://doi.org/10.1093/bioinformatics/btp033.

Webb, B., & Sali, A. (2014). Comparative protein structure modeling using MODELLER. Current Protocols in Bioinformatics, 2014, 5.6.1-5.6.32. https://doi.org/10.1002/0471250953.bi0506s47.

Wu, C. H., Nikolskaya, A., Huang, H., Yeh, L.-S. L., Natale, D. a, Vinayaka, C. R., … Barker, W. C. (2004). PIRSF: family classification system at the Protein Information Resource. Nucleic Acids Research, 32(Database issue), D112-4. https://doi.org/10.1093/nar/gkh097.

Zhang, Y. (2008). I-TASSER server for protein 3D structure prediction. BMC Bioinformatics, 9(1), 40. https://doi.org/10.1186/1471-2105-9-40.