Predicting the Oligomeric States of Fluorescent Proteins

Saw Simeon, Watshara Shoombuatong, Likit Preeyanon, Virapong Prachayasittikul, Chanin Nantasenamat

Currently, monomeric fluorescent proteins (FP) are ideal markers for protein tagging. The prediction of oligomeric states is helpful for enhancing live biomedical imaging. Computational prediction of FP oligomeric states can accelerate the effort of protein engineering to create monomeric FPs by saving time and money. To the best of our knowledge, this study represents the first computational model for predicting and analyzing FP oligomerization directly from their amino acid sequences. An exhaustive dataset consisting of 397 unique FP oligomeric states was compiled from the literature. FP were described by 3 classes of protein descriptors including amino acid composition, dipeptide composition and physicochemical properties. The oligomeric states of FP was predicted using decision tree (DT) algorithm and results demonstrated that DT provided robust performance with accuracies in ranges of 79.97-81.72% and 80.76-82.63% for the internal (e.g. 10-fold cross-validation) and external sets, respectively. This approach was also benchmarked with other common machine learning algorithms such as artificial neural network, support vector machine and random forest. A thorough analysis of amino acid sequence features was conducted to provide informative insights into FP oligomerization, which may aid in engineering novel monomeric fluorescent proteins. The following differentiating characteristics of monomeric and oligomeric fluorescent proteins were derived from DT: (i) substitution of any amino acid to Glu led to the reduction of aggregated proteins and (ii) oligomerization of FP appears to be stabilized by several hydrophobic contacts. Datasets and R source code are available at http://dx.doi.org/10.6084/m9.figshare.1348575.

PeerJ PrePrints | https://dx.doi.org/10.7287/peerj.preprints.922v1 | CC-BY 4.0 Open Access | rec: 24 Mar 2015, publ: 24 Mar 2015

Research Article

Predicting the Oligomeric States of Fluorescent Proteins

Saw Simeon[1,†], Watshara Shoombuatong[1,†], Likit Preeyanon[1], Virapong Prachayasittikul[2], Chanin Nantasenamat[1,*]

[1] *Center of Data Mining and Biomedical Informatics, Faculty of Medical Technology, Mahidol University, Bangkok 10700, Thailand*
[2] *Department of Clinical Microbiology and Applied Technology, Faculty of Medical Technology, Mahidol University, Bangkok 10700, Thailand.*

[†] Contributed equally

[*] Corresponding author
E-mail: chanin.nan@mahidol.ac.th, Tel: +66 2 441 4371, Fax: +66 2 441 4380

PeerJ PrePrints | https://dx.doi.org/10.7287/peerj.preprints.922v1 | CC-BY 4.0 Open Access | rec: 24 Mar 2015, publ: 24 Mar 2015

Abstract

Currently, monomeric fluorescent proteins (FP) are ideal markers for protein tagging. The prediction of oligomeric states is helpful for enhancing live biomedical imaging. Computational prediction of FP oligomeric states can accelerate the effort of protein engineering to create monomeric FPs by saving time and money. To the best of our knowledge, this study represents the first computational model for predicting and analyzing FP oligomerization directly from their amino acid sequences. An exhaustive dataset consisting of 397 unique FP oligomeric states was compiled from the literature. FP were described by 3 classes of protein descriptors including amino acid composition, dipeptide composition and physicochemical properties. The oligomeric states of FP was predicted using decision tree (DT) algorithm and results demonstrated that DT provided robust performance with accuracies in ranges of 79.97-81.72% and 80.76-82.63% for the internal (e.g. 10-fold cross-validation) and external sets, respectively. This approach was also benchmarked with other common machine learning algorithms such as artificial neural network, support vector machine and random forest. A thorough analysis of amino acid sequence features was conducted to provide informative insights into FP oligomerization, which may aid in engineering novel monomeric fluorescent proteins. The following differentiating characteristics of monomeric and oligomeric fluorescent proteins were derived from DT: (i) substitution of any amino acid to Glu led to the reduction of aggregated proteins and (ii) oligomerization of FP appears to be stabilized by several hydrophobic contacts. Datasets and R source code are available at http://dx.doi.org/10.6084/m9.figshare.1348575.

Keywords: fluorescent protein; FP; green fluorescent protein; GFP; oligomeric state; data mining

Introduction

Many coral fluorescent proteins (FP) are observed in anthozoans and because of their tertiary structures homologous to the *Aequorea victoria* jellyfish, they are termed green fluorescent protein (GFP)-like. These FPs represent an important class of bioluminescent proteins because of their immense utility for biomedical imaging in the life sciences. Such popularity lies in the diversity of their spectral colors and their lack of requiring co-factors because of the autocatalytic post-translational modifications of the chromophore from three or four amino acid precursors. Although the inherently weak dimerization of *Aequorea* GFP does not hinder its usage as a protein tag, the obligate tetramerization of DsRed has greatly impeded its utilization as a genetically encoded fusion tag because of possible perturbations to the tagged protein. Although oligomeric FPs in corals can serve as "sunscreen" to prevent coral bleaching, steric conflicts and stoichiometric clashes can occur when DsRed is tagged to oligomeric proteins of interest (i.e., actin, tubulin, connexin or histone) (Baird et al. 2000).

Despite being the essential tagging tool for live biomedical imaging, the oligomerization of FPs hinders their utilization, problems have been reported, such as abnormal localizations, perturbing normal functions, interfering with signaling cascades, and preventing normal oligomerization fusion products within specific organelles. Shcherbo *et al.* (2007) stressed that Katushka, the dimeric far-red mutants of FPs from the sea anemone *Entacmaea quadricolor*, formed abnormal localization in Phoenix eco cells. Mizuno *et al.* (2001) demonstrated that aggregation of DsRed disturbs normal function of calmodulin in the cytosol. Zacharias (2002) stressed that oligomerization of FPs interfered with target protein signaling cascades when using them as tagging probes for Fluorescent Resonance Energy Transfer (FRET). Lauf *et al.* (2001) stressed that tetrameric DsRed tagged with connexins creates problems because DsRed cross-linked between different connexins, negative affecting connexin function. Jane *et al.* (2001) reported that in the secretory pathway of endocrine cells, EGFP oligomerized through the disulphide-linkage of Cys 49 and Cys 71. Typically, there are two ways to overcome oligomerization problems: to modify the FPs to monomeric states through rational and/or random mutagenesis and to look for natural monomeric FPs from other organisms. Zacharias *et al.* (2002) rationally created the monomeric FP of *A. victoria* by carefully looking at the crystal structure of GFP and modifying hydrophobic interactions into polar charged amino acids by changing Ala 206, Leu 221 and Phe 223 to Lys, Lys and Arg, respectively. Campbell *et al.* (2002) developed an mRFP1 containing 33 mutations in which Ile 125, a hydrophobic amino acid, was changed to Arg, a positively charged amino acid, to develop a dimeric FP before a cycle of random mutagenesis was performed. Shagin *et al.* (2004) screened for FPs from hydrozoan species from the ocean and observed that one in six discovered copepoda FPs were monomeric.

For computational investigations on protein oligomerization properties, Garian (2001) first implemented a DT algorithm using primary sequences from the SWISS-PROT database (Release 34) for classifying a particular protein into homodimers or non-homodimers. Afterwards, several computational models implementing support vector machines (SVM) (Qiu et al. 2011; Song & Tang 2005; Zhang et al. 2003), Function of Degree of Disagreement (FDOD) (Song & Tang 2004), *k*-NN algorithm (Song 2007), and probability approaches (Carugo 2007) were proposed to improve prediction results. Currently, although many predictive models have been proposed to enhance performances on various databases, no computational studies have been performed for specifically analyzing and investigating FPs. Details of existing methods for predicting protein oligomerization properties (Carugo 2007; Chou & Cai 2003; Garian 2001; Qiu et al. 2011; Shen & Chou 2009; Shi et al. 2005; Song 2007; Song & Tang 2004; Song & Tang 2005; Sun et al. 2012; Xiao & Lin 2009; Xiao et al. 2011; Zhang et al. 2007; Zhang et al. 2003) are provided in Table 1.

This study proposes the first computational model based on DT for predicting FP oligomeric states directly from protein sequences. Figure 1 illustrates the flowchart of the workflow used to predict and analyze the oligomerization of FPs. Three types of protein descriptors, including amino acid composition, dipeptide composition and physicochemical properties, were used to extract descriptors from primary sequences. The prediction results demonstrated that our proposed method performed well, with testing accuracy of 82.63%.

Furthermore, the use of a DT algorithm is easily interpretable and capable of predicting FP oligomerization, which is potentially helpful in engineering novel monomeric FPs.

Materials and Methods

Datasets

In this study, we compiled a large dataset consisting of 397 FP oligomeric states from the primary literature as provided in the supplementary file available at http://dx.doi.org/10.6084/m9.figshare.1348575. Monomeric FPs are ideal tools for fluorescent tagging in biomedical imaging, whereas dimeric, trimeric and tetrameric FPs hinder their usage as tagging labels because of their tendencies to aggregate. Therefore, we aimed to classify the 397 FPs as either monomeric or oligomeric states. To develop and validate the ability of the prediction model, the 397 FP oligomeric states were randomly divided into internal (80%) and external (20%) sets in which the former set was used for constructing predictive models as full training and 10-fold cross-validation (10-fold CV) while samples in the latter set was predicted using the aforementioned trained model. This data splitting was performed for 100 iterations followed by computing the mean prediction performance (as will be described in the subsequent section).

Protein descriptor extraction

There are many protein descriptors for the analysis of protein functions. In this study, easy and interpretable features consisting of amino acid composition (AAC), dipeptide composition (DPC) and physicochemical properties (PCP) were utilized to encode FPs. The potential ability of these descriptors to predict protein functions has been previously demonstrated (Huang et al. 2011; Liaw et al. 2013; Tung et al. 2011).

AAC is the proportion of each amino acid in a protein sequence, which was expressed as a fixed length of 20. Given a protein sequence of FP oligomeric states with length l, the occurrence frequency of the ith amino acid (a_i) is calculated as follows:

$$a_i = AA_i / l \qquad (1)$$

where AA_i is the number of occurrences in the sequence for the the ith amino acid.

DPC was used to provide global information for each protein sequence, which was a fixed length of 20×20 = 400. DPC encompassed information regarding amino acid composition along the local order of amino acids. In case of DPC, the occurrence frequency of the ith dipeptide (dp_i) is calculated as follows:

$$dp_i = DP_i / l \tag{2}$$

where DP_i is the number of occurrences in the sequence for the ith amino acid.

PCP has been demonstrated to be essential for the prediction and analysis of many protein structures in bioinformatics studies because of its interpretability (Charoenkwan et al. 2013; Liaw et al. 2013; Tung et al. 2011). Analysis of the correlation between PCPs and FP oligomeric states can therefore provide insights that can further our biological knowledge of these systems. Each physicochemical property is represented as a set of 20 numerical values for amino acids. After removing 13 physicochemical properties with 'NA' in their amino acid indices, a total of 531 physicochemical properties were attained (Kawashima et al. 1999).

Multivariate analysis

A DT algorithm was utilized for constructing a computational model to predict FP oligomeric states. Since, the DT method affords interpretable rules for estimating feature importance pertaining to FP oligomeric states, therefore it is helpful in revealing the different characteristics between monomeric and oligomeric states. The construction of a DT model requires the following: (i) all samples in the internal set belong to a single class; (ii) the tree depth is close to maximum; and (iii) the number of classes in the terminal node is less than the minimum number of classes of the parent nodes. In general, the root node is a variable with the highest information gain, whereas the other internal nodes provide the second and subsequent highest information gain thereafter. The information gain of variable $v\,(Gain_v)$ is calculated as follows:

$$Gain_v = \sum_{j=1}^{N} - p(C_j)\log_2 p(C_j) - \sum_{v\in V}\frac{|D_v|}{|D|}I(D_v) \tag{3}$$

where $Gain_v$ is the information gain of feature *v* on the remaining data $D_v \subset D$, and $p(C_j)$ is the probability of the relative frequency of class *j* (C_j). In this study, *j*=2 was denoted as monomeric or oligomeric states. The prediction model was constructed using the J48 algorithm as implemented using the RWeka package (Frank et al. 2004). Important parameters consisted of the confidence factor used for pruning (confidenceFactor), the minimum number of instances per leaf (minNumObj), and the amount of data used for reduced-error pruning (numFolds). Herein, the confidenceFactor, minNumObj, and numFolds were set to default values (Che et al. 2010; Neugebauer et al. 2007; Yan et al. 2007). Furthermore, our proposed prediction model was compared with other well-known computational methods, such as artificial neural networks (ANN), support vector machines (SVM) and random forests (RF). In ANN, the back-propagation algorithm was implemented using the Weka software package (Frank et al. 2004). For SVM, the optimal parameters (gamma and cost) must be estimated to build the optimal SVM model (Dimitriadou et al. 2008). The most frequently used radial basis function kernel was selected as

PeerJ PrePrints | https://dx.doi.org/10.7287/peerj.preprints.922v1 | CC-BY 4.0 Open Access | rec: 24 Mar 2015, publ: 24 Mar 2015

the kernel function (Frank et al. 2004). Finally, a RF model was constructed with 500 trees (Frank et al. 2004).

Validation of predictive model

For any empirical model, model validation is an important process. Four measurements were used to evaluate the prediction performances of the proposed model: accuracy (Acc), sensitivity (Sen), specificity (Spec), and the Matthews' correlation coefficient (MCC). These parameters are defined as follows:

$$\text{Accuracy} = \frac{TP + TN}{(TP + TN + FP + FN)} \times 100 \tag{4}$$

$$\text{Sensitivity} = \frac{TP}{(TP + FN)} \times 100 \tag{5}$$

$$\text{Specificity} = \frac{TN}{(TN + FP)} \times 100 \tag{6}$$

$$\text{MCC} = \frac{TP \times TN - FP \times FN}{\sqrt{(TP + FP)(TP + FN)(TN + FP)(TN + FN)}} \tag{7}$$

where TP, TN, FP and FN are the numbers of true positives, true negatives, false positives and false negatives, respectively. In this study, a 10-fold CV procedure was used to confirm the reliability and robustness of the QSPR models using the training set (Kohavi 1995). Additionally, an external validation set was used to assess the generalizability of our proposed model for predicting unknown samples.

Results and Discussion

Protein oligomerization

There are many advantages of association between subunits of proteins to exist in oligomeric states. Biological activities of proteins can depend on oligomerization because a single subunit alone can be unstable and unable to exert its functions. Oligomerization also extends protein flexibility by mutual coupling to enhance biological activities. Oligomerization is useful in regulating activities because it can prevent the binding of unnecessary substrates to their allosteric sites. However, disadvantages of oligomerization also exist. Oligomerization creates a larger protein that places evolutionary pressure on its substrate to become smaller, thus increasing the probability of unnecessary binding of related substrates. Oligomerization also slows down rotational and translational diffusion which are important factors for random collisions (Nooren & Thornton 2003). In some cases, protein oligomerization can have a devastating effect on health, as observed in many neurological diseases (Cleary et al. 2005).

Oligomeric protein states arise from interfacial residues that have great electrostatic, polar, hydrophobic geometrical shape, hydrophobic and hydrogen bonding complementarity, thus resulting in interaction specificity. Approximately one third of cellular proteins are oligomeric. Oligomeric proteins can be hetero-oligomeric, forming from different subunits, or homo-oligomeric, composed of the same subunits. For example, transthyretin (TTR), phenylalanine hydroxylase (PAH) and L-rhamnulose-1-phosphatase aldolase are homo-oligomeric proteins, whereas hemoglobin, immunotoxin and coagulation factor IX/X-binding protein are hetero-oligomeric proteins. The associations between subunits of oligomeric proteins depends on their strengths and durations which are influenced by factors including, pH, concentration and temperature (Ali & Imperiali 2005). Hydrophobic interactions are responsible for defining 'hot spots' because two-thirds of amino acid residues are nonpolar at oligomeric interfaces (Miller 1989). Conversely, increased polarity at interfaces is common with weakly associated and transient oligomerization of proteins. Particularly, these interfaces are rich in polar residues that are primarily noncovalent interactions (i.e. hydrogen bonding and electrostatic interactions) that stabilizes the oligomeric interfaces and which can be easily solubilized to individual subunits (Janin et al. 1988). Although oligomeric proteins can be broadly classified according to subunit type, strength and voracity of subunit association, Levy and Teichmann (2013) proposed a classification based on protein morphology. Thus, oligomeric states of FPs can be conceptually classified according to their number of units as being monomers or oligomers (e.g. dimers, trimers or tetramers).

Performance of oligomeric states prediction

The internal set was used to construct a predictive model based on the J48 algorithm to discriminate FPs into either monomer or oligomer. The predictive model was performed using 10-fold CV as to prevent overtraining on the internal set and then tested on an external set in order to assess its ability to accurately predict unknown samples. Table 2 provides mean performance comparisons among the various types of features in the terms of 10-fold CV and external validation.

In the case of using single features, the highest performance on the external set using DPC exhibited Acc=82.63±4.06%, Sen=80.97±6.62%, Spec=84.20±6.57 and MCC=0.66±0.08. Meanwhile, using AAC afforded the second highest performance with Acc=81.23±4.06%, Sen=79.62±7.30%, Spec=82.76±6.51 and MCC=0.63±0.08. We also considered the four possible combinations of feature types in the construction of predictive models. Table 2 indicated that the combination of PCP with AAC and DPC could provide improvements to Acc (from 80.76% to 81.91%), Sen (from 80.92% to 81.74%), Spec (from 80.61% to 82.49%) and Mcc (from 0.62 to 0.64). These results demonstrated that using combination of three feature types could lead to better predictive performance. It was found that DPC is highly relevant for predicting the FP oligomeric states and was therefore used in further benchmarking studies with other machine learning algorithms.

The DT algorithm was benchmarked against other commonly used machine learning algorithms (using empirically determined optimal parameters) namely ANN (with 2 hidden nodes), SVM (with gamma=32 and cost =1) and RF (with 100 trees) using DPC. As described previously, 397 FPs were randomly divided into internal (80%) and external (20%) sets for 100 times. It can be seen from Table 3 that the best performance with Acc=86.54±3.30%, Sen=86.79±5.03%, Spec=86.29±5.49 and MCC=0.73±0.07 was obtained from the RF model. The second best model was obtained using SVM with Acc=84.45±3.57%, Sen=84.87±5.24%, Spec=84.05±6.46 and MCC=0.69±0.07. Thus, the first and second highest prediction result was reasonably obtained from RF and SVM models, respectively. It is well known that the RF model is established from several decision trees while SVM model is obtained from complex procedures both of which afford minimal interpretability (aside from the Gini index that could potentially afford evaluation of the feature importance but from several trees). It can be deduced from these results that our proposed DT model afforded comparable performance level with those of RF and SVM models while also maximizing both the prediction results and interpretability.

Identifying informative composition features and physicochemical properties

Investigating feature importance of each type of protein descriptor can provide insights into FP oligomerization. Herein, the efficient built-in feature importance selector of the J48 algorithm was used. In the J48 algorithm, the estimation of feature importance is calculated from feature usage based on information gain. The feature with the highest feature usage score is the most important feature because it maximizes the prediction performance. Feature importance is provided in Figure 2 as three subplots corresponding to the three classes of protein descriptors.

Figure 2 demonstrates that the top-three informative amino acids were Glu (100.00), Leu (36.52) and Gln (25.69). It could be well recognized that Glu is a negatively charged amino acid. Several previous studies in protein engineering suggested that the substitution of single-site amino acids X→Glu (replacing any given amino acid X by Glu) affected the reduction of aggregated proteins. Yanushevich *et al.* generated non-aggregating mutants of *Anthozoa* FPs from drFP583, zFP506, zFP538, amFP486 and asFP595 by using site-directed mutagenesis of Lys→Glu at the N-terminus of the protein (Yanushevich et al. 2002). Similarly, generation of the monomeric Azami-Green (mAG) FP from tetrameric AG of *Galaxeidae* also involved Lys→Glu to obtain reduced aggregated protein (Karasawa et al. 2003).

As for DPC, the top-four informative dipeptides were AS, FI, QP and MV with feature usage scores greater than 40. Notably, 2 out of 4 of these informative dipeptides (FI and MV) were composed of single hydrophobic amino acids. Results indicated that hydrophobic amino acids are important for oligomerization of FPs. This finding is well consistent with the experimental results of Yarbrough *et. al.* (2001) in which the crystal structure of DsRed from *Discosoma sp.* indicated that the oligomeric interfaces of subunits A and B consists of hydrophobic interactions with few hydrogen bonds and salt bridges. In a similar manner, the first discovered photoconvertible Kaede from *Trachyphyllia geoffroyi* displayed dominant hydrophobic interactions between the oligomeric interface at the A and C subunits (Hayashi et al. 2007). Additionally, *Heteractis crispa* HcRed, the commercially available dimeric FP from the company Clontech, was converted to a dimer from a tetramer by changing the hydrophobic Leu residue at position 123 to the aromatic His residue within the hydrophobic interface to perturb the tetrameric hydrophobic interface (Wilmann et al. 2005).

PCP has been shown to be crucial for predicting protein functions (Huang et al. 2011; Liaw et al. 2013; Tung et al. 2011). The identification of informative physicochemical descriptors of FPs was thus used for providing insights into the mechanism of FP oligomerization. The rank of feature importance for PCPs is shown in Figure 2, while Table 3 presents the 5 top-ranked informative PCPs according to their feature usage scores. It is observed that the five top-ranked PCPs (and their corresponding feature usage scores) consisted of AAindex IDs: FUKS010109 (100), ARGP820101 (15.62), OOBM770104 (15.62), ROSM880101 (ROSM880101) and FASG760102 (6.8). Remarkably, 3 out of 5 informative PCPs having AAindex IDs of ARGP820101, ROSM880101 and FASG760102 pertained to hydrophobicity properties (Huang et al. 2011). As mentioned above, the hydrophobicity was responsible for stabilizing the oligomeric states of FPs (Hayashi et al. 2007; Wilmann et al. 2005; Yarbrough et al. 2001). Particularly, ARGP820101, ROSM880101, and FASG760102 corresponded to (i) Hydrophobicity index, (ii) Side chain hydropathy (uncorrected for solvation) and (iii) Melting point, respectively. Thus, this suggests that FP oligomerization are stabilized by several hydrophobic contacts. These finding are consistent with the experimental work in which hydrophobic residues at the interface were substituted with polar residues in attempt to create monomic FPs (Campbell et al. 2002; Hayashi et al. 2007; Wilmann et al. 2005; Yarbrough et al. 2001). Along with hydrophobic contacts, several other interactions including the formation of coordination bonds, ionic interactions, van der Waals' contacts, electrostatic interactions, hydrogen bondings and π-π stackings may mediate the oligomerization of FP at the "hot spot" sites.

Conclusion

This study represents the first attempt in the development of a computational model for predicting and analyzing FP oligomerization from protein sequences using amino acid composition, dipeptide composition and physicochemical properties. The experimental results demonstrated that a DT algorithm utilizing dipeptide compositions and physicochemical properties performed well on both internal and external sets with accuracies of 81.69% and 82.63%, respectively. By identifying the informative features obtained from the feature usage scores of DT, the composition of Glu is important for the reduction of aggregated proteins. Moreover, we observed that hydrophobic amino acids were also important for FP oligomerization. Finally, the analysis of the most important physicochemical properties also revealed that hydrophobic properties are important for protein oligomerization. These findings can aid biologists in designing novel monomeric FPs.

PeerJ PrePrints | https://dx.doi.org/10.7287/peerj.preprints.922v1 | CC-BY 4.0 Open Access | rec: 24 Mar 2015, publ: 24 Mar 2015

References

Ali MH, and Imperiali B. 2005. Protein oligomerization: How and why. *Bioorg Med Chem* 13:5013-5020.

Baird GS, Zacharias DA, and Tsien RY. 2000. Biochemistry, mutagenesis, and oligomerization of DsRed, a red fluorescent protein from coral. *Proc Natl Acad Sci USA* 97:11984-11989.

Campbell RE, Tour O, Palmer AE, Steinbach PA, Baird GS, Zacharias DA, and Tsien RY. 2002. A monomeric red fluorescent protein. *Proc Natl Acad Sci USA* 99:7877-7882.

Carugo O. 2007. A structural proteomics filter: prediction of the quaternary structural type of hetero-oligomeric proteins on the basis of their sequences. *J Appl Crystallogr* 40:986-989.

Charoenkwan P, Shoombuatong W, Lee H-C, Chaijaruwanich J, Huang H-L, and Ho S-Y. 2013. SCMCRYS: predicting protein crystallization using an ensemble scoring card method with estimating propensity scores of P-collocated amino acid pairs. *PloS one* 8:e72368.

Che D, Hockenbury C, Marmelstein R, and Rasheed K. 2010. Classification of genomic islands using decision trees and their ensemble algorithms. *BMC Genomics* 11:S1.

Chou K-C, and Cai Y-D. 2003. Predicting protein quaternary structure by pseudo amino acid composition. *Proteins: Struct, Funct, Bioinf* 53:282-289.

Cleary JP, Walsh DM, Hofmeister JJ, Shankar GM, Kuskowski MA, Selkoe DJ, and Ashe KH. 2005. Natural oligomers of the amyloid-[beta] protein specifically disrupt cognitive function. *Nat Neurosci* 8:79-84.

Dimitriadou E, Hornik K, Leisch F, Meyer D, and Weingessel A. 2008. Misc functions of the Department of Statistics (e1071), TU Wien. *R package*:1.5-24.

Frank E, Hall M, Trigg L, Holmes G, and Witten IH. 2004. Data mining in bioinformatics using Weka. *Bioinformatics* 20:2479-2481.

Garian R. 2001. Prediction of quaternary structure from primary structure. *Bioinformatics* 17:551-556.

Hayashi I, Mizuno H, Tong KI, Furuta T, Tanaka F, Yoshimura M, Miyawaki A, and Ikura M. 2007. Crystallographic Evidence for Water-assisted Photo-induced Peptide Cleavage in the Stony Coral Fluorescent Protein Kaede. *J Mol Biol* 372:918-926.

Huang H-L, Lin I-C, Liou Y-F, Tsai C-T, Hsu K-T, Huang W-L, Ho S-J, and Ho S-Y. 2011. Predicting and analyzing DNA-binding domains using a systematic approach to identifying a set of informative physicochemical and biochemical properties. *BMC Bioinformatics* 12:S47.

Jain RK, Joyce PB, Molinete M, Halban PA, and Gorr SU. 2001. Oligomerization of green fluorescent protein in the secretory pathway of endocrine cells. *Biochem J* 360:645-649.

Janin J, Miller S, and Chothia C. 1988. Surface, subunit interfaces and interior of oligomeric proteins. *J Mol Biol* 204:155-164.

Karasawa S, Araki T, Yamamoto-Hino M, and Miyawaki A. 2003. A Green-emitting Fluorescent Protein from Galaxeidae Coral and Its Monomeric Version for Use in Fluorescent Labeling. *J Biol Chem* 278:34167-34171.

Kawashima S, Ogata H, and Kanehisa M. 1999. AAindex: Amino Acid Index Database. *Nucleic Acids Res* 27:368-369.

Kohavi R. 1995. A study of cross-validation and bootstrap for accuracy estimation and model selection. IJCAI. p 1137-1145.

Lauf U, Lopez P, and Falk MM. 2001. Expression of fluorescently tagged connexins: a novel approach to rescue function of oligomeric DsRed-tagged proteins. *FEBS Lett* 498:11-15.

Levy ED, and Teichmann SA. 2013. Chapter Two - Structural, Evolutionary, and Assembly Principles of Protein Oligomerization. In: Jesús G, and Francisco C, eds. *Prog Mol Biol Transl Sci*: Academic Press, 25-51.

Liaw C, Tung C-W, and Ho S-Y. 2013. Prediction and Analysis of Antibody Amyloidogenesis from Sequences. *PLoS ONE* 8:e53235.

Miller S. 1989. The structure of interfaces between subunits of dimeric and tetrameric proteins. *Protein Eng* 3:77-83.

Mizuno H, Sawano A, Eli P, Hama H, and Miyawaki A. 2001. Red Fluorescent Protein from Discosoma as a Fusion Tag and a Partner for Fluorescence Resonance Energy Transfer†. *Biochemistry* 40:2502-2510.

Neugebauer A, Hartmann RW, and Klein CD. 2007. Prediction of Protein−Protein Interaction Inhibitors by Chemoinformatics and Machine Learning Methods. *J Med Chem* 50:4665-4668.

Nooren IMA, and Thornton JM. 2003. Structural Characterisation and Functional Significance of Transient Protein–Protein Interactions. *J Mol Biol* 325:991-1018.

Qiu J-D, Suo S-B, Sun X-Y, Shi S-P, and Liang R-P. 2011. OligoPred: A web-server for predicting homo-oligomeric proteins by incorporating discrete wavelet transform into Chou's pseudo amino acid composition. *J Mol Graph Model* 30:129-134.

Shagin DA, Barsova EV, Yanushevich YG, Fradkov AF, Lukyanov KA, Labas YA, Semenova TN, Ugalde JA, Meyers A, Nunez JM, Widder EA, Lukyanov SA, and Matz MV. 2004. GFP-like Proteins as Ubiquitous Metazoan Superfamily: Evolution of Functional Features and Structural Complexity. *Mol Biol Evol* 21:841-850.

Shcherbo D, Merzlyak EM, Chepurnykh TV, Fradkov AF, Ermakova GV, Solovieva EA, Lukyanov KA, Bogdanova EA, Zaraisky AG, Lukyanov S, and Chudakov DM. 2007. Bright far-red fluorescent protein for whole-body imaging. *Nat Meth* 4:741-746.

Shen H-B, and Chou K-C. 2009. QuatIdent: A Web Server for Identifying Protein Quaternary Structural Attribute by Fusing Functional Domain and Sequential Evolution Information. *J Proteome Res* 8:1577-1584.

Shi J, Pan Q, Zhang S, and Cheng Y. 2005. Classification of protein homo--oligomers using amino acid composition distribution. *Shengwu Wuli Xuebao* 22:49-56.

Song J. 2007. Prediction of homo-oligomeric proteins based on nearest neighbour algorithm. *Comput Biol Med* 37:1759-1764.

Song J, and Tang H. 2004. Accurate Classification of Homodimeric vs Other Homooligomeric Proteins Using a New Measure of Information Discrepancy. *J Chem Inf Comput Sci* 44:1324-1327.

Song J, and Tang H. 2005. Support vector machines for classification of homo-oligomeric proteins by incorporating subsequence distributions. *Comp Theor Chem* 722:97-101.

Sun X-Y, Shi S-P, Qiu J-D, Suo S-B, Huang S-Y, and Liang R-P. 2012. Identifying protein quaternary structural attributes by incorporating physicochemical properties into the general form of Chou's PseAAC via discrete wavelet transform. *Mol Biosyst* 8:3178-3184.

Tung C-W, Ziehm M, Kamper A, Kohlbacher O, and Ho S-Y. 2011. POPISK: T-cell reactivity prediction using support vector machines and string kernels. *BMC Bioinformatics* 12:446.

Wilmann PG, Petersen J, Pettikiriarachchi A, Buckle AM, Smith SC, Olsen S, Perugini MA, Devenish RJ, Prescott M, and Rossjohn J. 2005. The 2.1 Å Crystal Structure of the Far-red Fluorescent Protein HcRed: Inherent Conformational Flexibility of the Chromophore. *J Mol Biol* 349:223-237.

Xiao X, and Lin W-Z. 2009. Application of protein grey incidence degree measure to predict protein quaternary structural types. *Amino Acids* 37:741-749.

Xiao X, Wang P, and Chou K-C. 2011. Quat-2L: a web-server for predicting protein quaternary structural attributes. *Mol Divers* 15:149-155.

Yan X, Chao T, Tu K, Zhang Y, Xie L, Gong Y, Yuan J, Qiang B, and Peng X. 2007. Improving the prediction of human microRNA target genes by using ensemble algorithm. *FEBS Lett* 581:1587-1593.

Yanushevich YG, Staroverov DB, Savitsky AP, Fradkov AF, Gurskaya NG, Bulina ME, Lukyanov KA, and Lukyanov SA. 2002. A strategy for the generation of non-aggregating mutants of Anthozoa fluorescent proteins. *FEBS Lett* 511:11-14.

Yarbrough D, Wachter RM, Kallio K, Matz MV, and Remington SJ. 2001. Refined crystal structure of DsRed, a red fluorescent protein from coral, at 2.0-Å resolution. *Proc Natl Acad Sci USA* 98:462-467.

Zacharias DA. 2002. Sticky caveats in an otherwise glowing report: oligomerizing fluorescent proteins and their use in cell biology. *Sci Signal* 2002:pe23.

Zacharias DA, Violin JD, Newton AC, and Tsien RY. 2002. Partitioning of Lipid-Modified Monomeric GFPs into Membrane Microdomains of Live Cells. *Science* 296:913-916.

Zhang S-W, Chen W, Zhao C-H, Cheng Y-M, and Pan Q. 2007. Predicting Protein Quaternary Structure with Multi-scale Energy of Amino Acid Factor Solution Scores and Their Combination. In: Zhang D, ed. *Medical Biometrics*: Springer Berlin Heidelberg, 65-72.

Zhang S-W, Pan Q, Zhang H-C, Zhang Y-L, and Wang H-Y. 2003. Classification of protein quaternary structure with support vector machine. *Bioinformatics* 19:2390-2396.

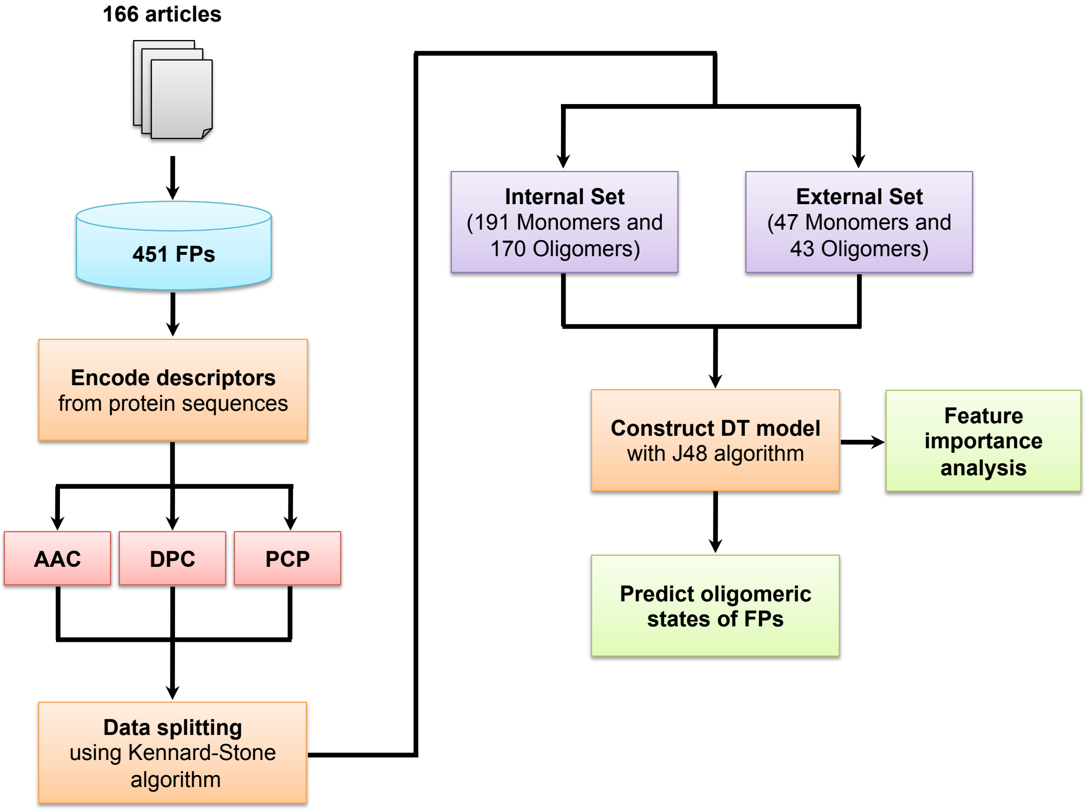
166 articles
451 FPs
Encode descriptors from protein sequences
AAC
DPC
PCP
Data splitting using Kennard-Stone algorithm
Internal Set (191 Monomers and 170 Oligomers)
External Set (47 Monomers and 43 Oligomers)
Construct DT model with J48 algorithm
Feature importance analysis
Predict oligomeric states of FPs

Figure 1. Schematic representation of the computational workflow.

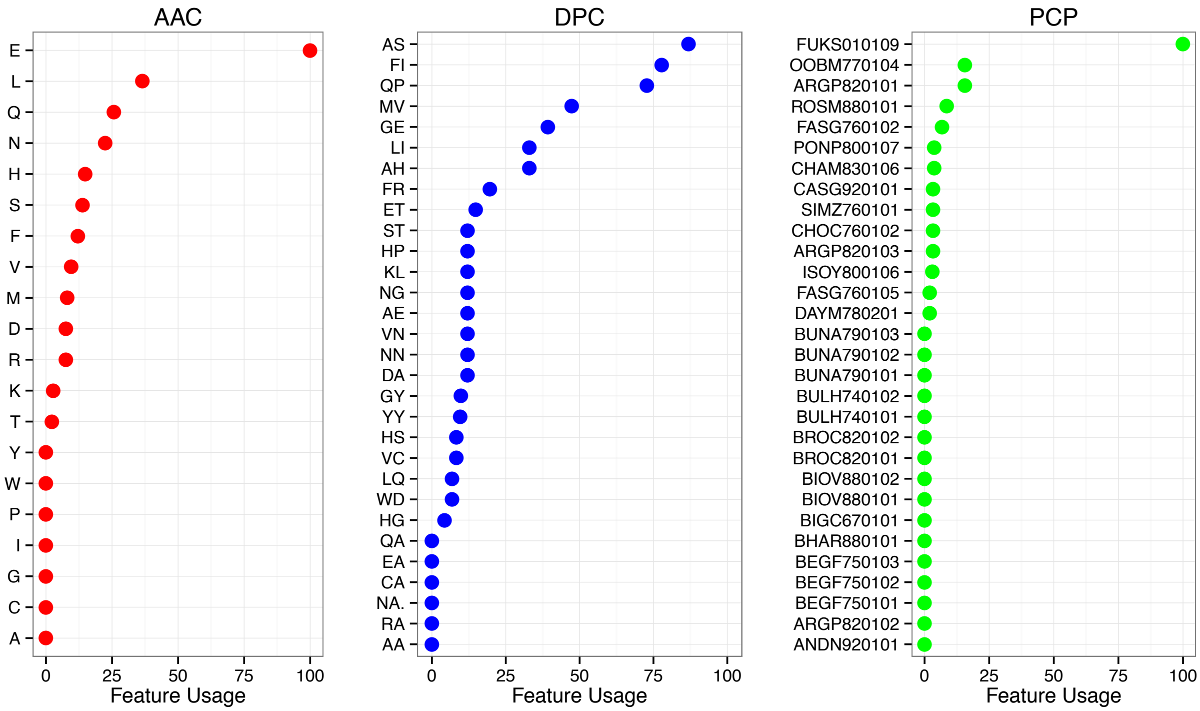

Figure 2. Feature importance analysis of amino acid composition, dipeptide composition and physicochemical properties.

Table 1. Summary of existing studies for predicting oligomeric states from protein sequences.

Dataset	Method	Internal set size	External set size	Sequence Features	Source
SWISS-PROT database (Release 34)	DT	1639	N/A	PCP	Garian 2001
	SVM	1639	N/A	AAC, AC	Zhang 2003
	FDOD	1639	N/A	Quasi-Sequence-Order	Song 2004
SWISS-PROT database (Release 34) after removing similar protein sequence	SVM	1568	N/A	Quasi-Sequence-Order	Song 2005
	SVM	1568	N/A	AAC, DPC, AACD	Shi 2005
	k-NN	1568	N/A	Quasi-Sequence-Order	Song 2007
	SVM	1568	1283	PseAAC	Qiu 2011
SWISS-PROT databank	DA	3174	332	PseAAC	Chou 2003
	SVM	3174	N/A	Factor Scores, MSE	Zhang 2007
	NN	3174	332	PseAAC	Xiao 2009
UniProtKB (Release 15.6)	Probability	5495	N/A	AAC, DPC	Carugo 2007
	Fuzzy *k*-NN	5495	N/A	PseAAC,	Xiao 2011
SWISS-PROT database (Release 55.3)	OET- *k*-NN	6702	N/A	FunD, PsePSSM	Shen 2009
	DWT_DT	6702	N/A	PseACC, PCP	Sun 2012
FP dataset	DT	318	79	ACC, DPC,PCP	This study

AAC is defined as amino acid composition.
DPC is defined as dipeptide composition.
PCP is defined as physicochemical properties.
PseAAC is defined as pseudo amino acid composition.
PsePSSM is defined as pseudo position-specific scoring matrix.
FDOD is defined as function of degree of disagreement.
FunD is defined as functional domain composition.
AACD is defined as amino acid composition distribution.
MSE is defined as multi-scale energy.
OET- *k*-NN is defined as optimized evidence-theoretic *k*-NN algorithm.
DWT_DT is defined as discrete wavelet transform and decision tree.

Table 2. Summary of the predictive performance as a function of various protein descriptor class as assessed by10-fold CV and external validation.

	10-fold CV				External validation			
Descriptors	Acc (%)	Sen (%)	Spec (%)	MCC	Acc (%)	Sen (%)	Spec (%)	MCC
AAC	80.65±1.82	78.99±2.75	82.23±2.67	0.61±0.04	81.23±4.06	79.62±7.30	82.76±6.51	0.63±0.08
DPC	81.69±2.19	79.95±3.13	83.35±2.69	0.63±0.04	82.63±4.06	80.97±6.62	84.20±6.57	0.66±0.08
PCP	79.97±1.91	80.01±2.48	79.94±2.51	0.60±0.04	80.76±4.13	80.92±6.23	80.61±6.24	0.62±0.08
AAC+DPC	81.72±2.19	80.30±3.09	83.08±2.66	0.63±0.04	82.60±4.10	80.74±6.36	84.37±6.38	0.65±0.08
AAC+PCP	80.66±1.85	80.79±2.57	80.54±2.62	0.61±0.04	81.68±3.79	81.56±6.62	81.78±6.31	0.64±0.08
DPC+PCP	80.82±2.15	80.52±2.78	81.11±2.81	0.62±0.04	81.58±3.64	81.74±6.54	81.41±6.18	0.63±0.07
AAC+DPC+PCP	81.55±2.04	80.91±3.04	82.17±2.49	0.63±0.04	81.91±3.82	81.31±6.02	82.49±6.27	0.64±0.08

Table 3. Comparison of the proposed method with related computational models.

Methods	10-fold CV				External validation			
	Acc (%)	Sen (%)	Spec (%)	MCC	Acc (%)	Sen (%)	Spec (%)	MCC
ANN	82.11±1.51	82.26±2.34	81.98±2.41	0.64±0.03	83.01±4.25	81.97±6.32	84.00±6.40	0.66±0.09
SVM	82.64±1.19	82.97±2.31	82.33±2.38	0.65±0.02	84.45±3.57	84.87±5.24	84.05±6.46	0.69±0.07
RF	86.36±1.32	87.51±1.72	85.27±1.61	0.73±0.03	86.54±3.30	86.79±5.03	86.29±5.49	0.73±0.07
DT	81.72±2.19	80.30±3.09	83.08±2.66	0.63±0.04	82.60±4.10	80.74±6.36	84.37±6.38	0.65±0.08

ANN, SVM and RF were constructed with 2 hidden nodes, gamma=32/cost =1 and 100 trees, respectively.