Chemometrics and Intelligent Laboratory Systems 120 (2013) 42–52

Contents lists available at SciVerse ScienceDirect

Chemometrics and Intelligent Laboratory Systems

journal homepage: www.elsevier.com/locate/chemolab

Quantitative structure–property relationship study of spectral properties of green fluorescent protein with support vector machine

Chanin Nantasenamat [a,b,*], Kakanand Srungboonmee [a], Saksiri Jamsak [a], Natta Tansila [c], Chartchalerm Isarankura-Na-Ayudhya [b], Virapong Prachayasittikul [b]

[a] Center of Data Mining and Biomedical Informatics, Faculty of Medical Technology, Mahidol University, Bangkok 10700, Thailand
[b] Department of Clinical Microbiology and Applied Technology, Faculty of Medical Technology, Mahidol University, Bangkok 10700, Thailand
[c] The Establishment Project of Faculty of Medical Technology, Prince of Songkla University, Hat Yai 90112, Thailand

ARTICLE INFO

Article history:
Received 30 June 2012
Received in revised form 6 November 2012
Accepted 6 November 2012
Available online 16 November 2012

Keywords:
Green fluorescent protein
GFP
QSPR
Support vector machine
Chromophore
Imidazolinone

ABSTRACT

Green fluorescent protein (GFP) is an autofluorescent protein that has been widely used in the biomedical sciences for molecular imaging applications. Computational approach for predicting the spectral properties of GFP offers great benefit for the design and engineering of novel color variants. Herein, we present a quantitative structure–property relationship (QSPR) study to model the spectral properties (e.g. excitation and emission maxima) of GFP chromophores using support vector machine (SVM). The data set is composed of 19 chromophores from GFP color variants and 29 synthetic GFP chromophores based on the imidazolinone scaffold. Quantum chemical descriptors were used to provide information on the physicochemical properties of the chromophores. Such descriptors were mapped onto a higher dimensional space via kernel functions (e.g. linear, polynomial and radial basis function kernels) and learning is then performed using SVM. The predicted spectral properties were well correlated with their experimental values as observed from correlation coefficient in the range of $r = 0.953–0.979$. Predictive performance of excitation maxima ($r = 0.967–0.979$) outperformed that of the emission maxima ($r = 0.953–0.961$). The present strategy holds great promise for expanding the spectral repertoire of GFP by facilitating the rational design of novel color variants.

1. Introduction

Green fluorescent protein (GFP) is an autofluorescent protein expressed in the outer dermal layer of the bioluminescent Pacific Northwest jellyfish *Aequorea victoria*. The crystal structure of *A. victoria* GFP (accession code 1EMA) had first been determined at a resolution of 1.9 Å by Ormö et al. in 1996 [1]. GFP is a protein composing of 238 amino acids where 11 β-sheets form the β-barrel and the centrally located α-helix harbors the tripeptide Ser–Tyr–Gly which autocatalytically cyclizes to form the hyperconjugated *p*-hydroxybenzylidene-imidazolinone ring system [2]. Native GFP expresses two bands of excitation maxima: major peak at 395 nm and minor peak at 475 nm. Such excitation peaks is a consequence of the existence of two protonation states: protonated (neutral) and deprotonated (anionic) forms where the Tyr residue at position 66 exists as a phenol or phenolate form, respectively. The advantage of GFP is that it does not require any substrate to become fluorescent hence it is autofluorescent upon UV illumination. Owing to this bioluminescent property, GFP has heavily been employed in the life sciences as fluorescent markers of gene expression [3], protein localization [4], protein–lipid interaction [5], protein–protein interaction [6] and as analytical sensors [7]. The ubiquity of GFP in the life sciences has led to the award of the 2008 Nobel Prize in Chemistry to Osamu Shimomura [8], Martin Chalfie [9] and Roger Y. Tsien [10] for isolating GFP from the *A. victoria* jellyfish in 1962, demonstrating that GFP can be used as a genetic tag along with the finding that it can be expressed in other organisms in 1994 and expanding the spectral repertoire of GFP since 1994, respectively. These achievements were greatly facilitated by the sequencing and cloning of GFP by Douglas Prasher [11] in 1992.

Methods for quantitatively predicting the spectral properties (e.g. excitation and emission maxima) of GFP from their chemical structures would be of great benefit for the design of novel GFP color variants. Current approaches for theoretical study of GFP chromophores rely on computationally-intensive *ab initio* and quantum chemistry calculations [12–19]. Quantitative structure–property relationship (QSPR) is an approach that correlates descriptors derived from the molecular structures of compounds with their respective spectral properties. QSPR assumes that the structural variations in a set of compounds will directly contribute to changes in the numerical values of the molecular descriptors, which is consequently correlated with the observed spectral properties [20,21]. The successful utilization of QSPR in predicting the biological activities or chemical properties for a diverse range of compounds has previously been demonstrated by us [22–33]. In our

* Corresponding author at: Center of Data Mining and Biomedical Informatics, Faculty of Medical Technology, Mahidol University, Bangkok 10700, Thailand. Tel.: +66 2 441 4371; fax: +66 2 441 4380.
E-mail address: chanin.nan@mahidol.ac.th (C. Nantasenamat).

0169-7439/$ – see front matter © 2012 Elsevier B.V. All rights reserved.
http://dx.doi.org/10.1016/j.chemolab.2012.11.003

previous study [22], we demonstrated for the first time the utilization of QSPR in predicting the spectral properties of GFP color variants. Results suggested that artificial neural network was the best performing method from among the employed approaches (e.g. multiple linear regression and partial least squares regression) for predicting the spectral properties of GFP chromophores.

In this study, the QSPR study of the spectral properties of 19 GFP color variants and 29 synthetic GFP chromophores is revisited. Molecular descriptors derived from B3LYP/6-31g(d)//HF-3-21G calculations were used to account for the physicochemical properties of the GFP chromophores. Multivariate analysis using support vector machine in conjunction with various kernels was performed in order to establish a correlation between the molecular structures of GFP color variants and synthetic GFP chromophores with their spectral properties. The obtained predictive models developed using support vector machine provide marked improvement over our previous QSPR models built using multiple linear regression, partial least squares regression and artificial neural network [22].

2. Materials and methods

2.1. Data sets

The spectral properties of 19 GFP color variants and 29 synthetic GFP chromophores were previously collected as described in our previous work [22]. Briefly, the structures of 19 GFP color variants were derived from the post-translationally cyclized forms of the GFP chromophore where backbone carbons were replaced with hydrogen atoms. The molecular structure of the 19 GFP color variants as obtained from several experimental reports is shown in Fig. 1 and their spectral properties are summarized in Table 1. The set of 19 GFP color variants are composed of the following chromophores: AWG [34], AYG [35], CYG [35], GYG [36], LYG [35], S(*p*-amino-F)G [37], S(*p*-bromo-F)G [37], S(*p*-iodo-F)G [37], S(*p*-methoxy-F)G [37], SFG [35], SHG [35], SWG [35], SYA [38], SYG [39], T(3-fluoro-Y)G [40], T(4-amino-W)G [41], THG [42], TWG [43] and TYG [35]. The molecular structure of the 29 synthetic GFP chromophores based on the imidazolone derivatives as obtained from the study of Follenius-Wund et al. [44] is shown in Fig. 2 while their spectral data are summarized in Table 2. In the set of 19 GFP color variants, the appropriate protonation states of the Tyr side chain of the 66th residue was assigned the phenol (protonated) form if the chromophore's excitation maxima was ~400 nm whereas the side chain was assigned the phenolate (deprotonated) form if the excitation maxima was ~480 nm. A schematic overview of the computational methodology carried out in this study is summarized in Fig. 3.

2.2. Geometry optimization and descriptor calculation

Geometry optimization and descriptor calculation were performed as previously described [22]. Briefly, the three-dimensional molecular

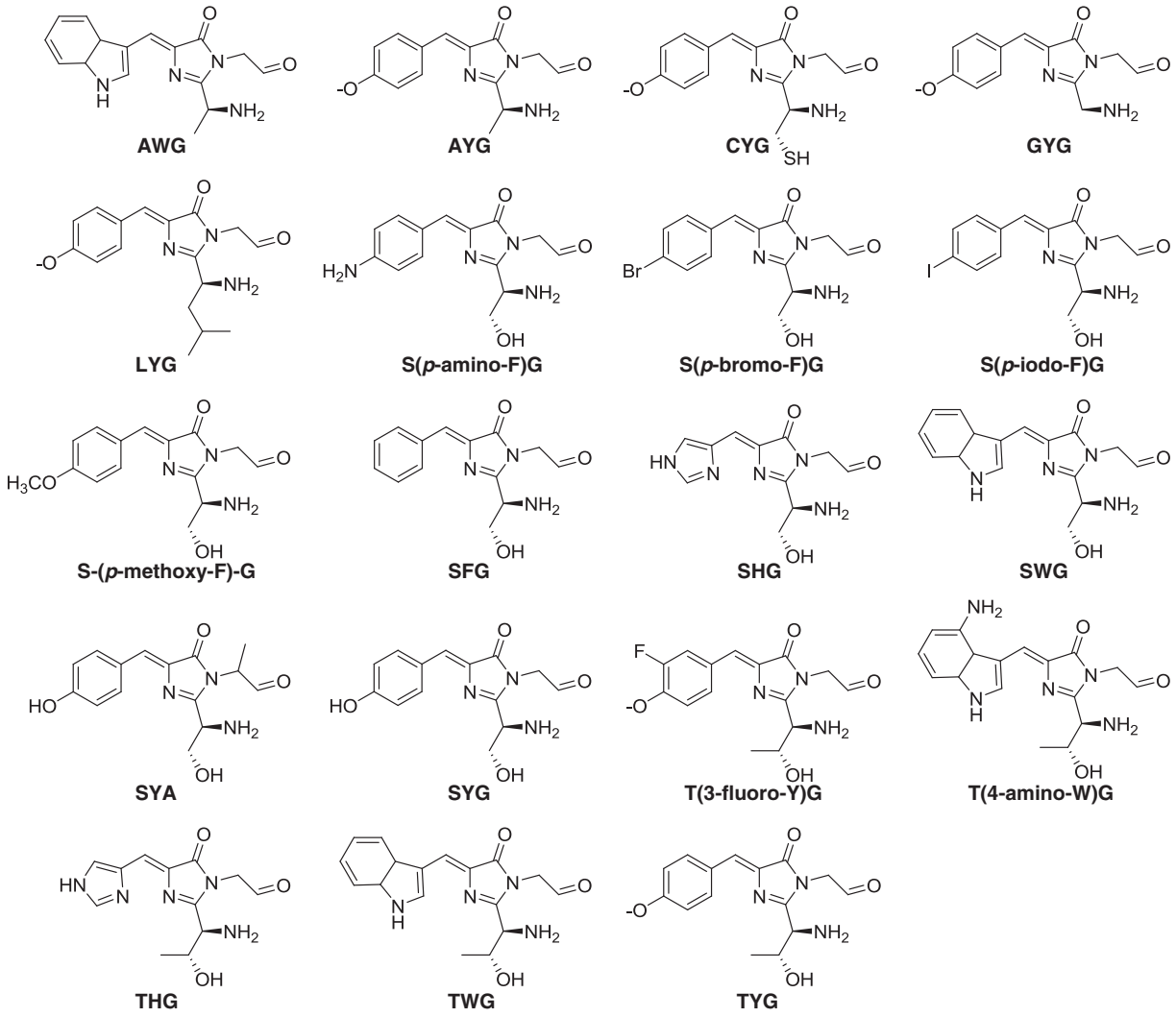

Fig. 1. Chemical structures of 19 GFP color variants.