Computational Biology and Chemistry 47 (2013) 215–220

Contents lists available at ScienceDirect

Computational Biology and Chemistry

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

Research Article

Prediction of protein modification sites of gamma-carboxylation using position specific scoring matrices based evolutionary information

Jianzhao Gao[a,1], Ning Zhang[b,*,1], Jishou Ruan[a,c,**]

[a] School of Mathematical Sciences and LPMC, Nankai University, 300071, PR China
[b] Department of Biomedical Engineering, Tianjin University, Tianjin Key Lab of BME Measurement, Tianjin 300072, PR China
[c] State Key Laboratory of Medicinal Chemical Biology, Nankai University, Tianjin 300071, PR China

ARTICLE INFO

Article history:
Received 23 August 2013
Received in revised form
12 September 2013
Accepted 12 September 2013

Keywords:
Gamma-carboxylation
Evolutionary conservation
Position specific scoring matrix
Support vector machine

ABSTRACT

Gamma-carboxylation, one type of post-translational modifications, is involved in many human disease. However, very few computational methods for gamma-carboxylation site prediction are available. In this paper, we develop a novel method *CarboxySVM* which is based on support vector machine with radial basis function kernel to identify the gamma-carboxylation sites. In this method, we combine position specific scoring matrices (PSSM)-based evolutionary conservation scores and other sequences-derived descriptors. As a result, an accuracy of 91.2% is achieved on training dataset with fivefold cross validation, and 91.8% on the independent test dataset. It is demonstrated by empirical evaluation on benchmark datasets that our method outperforms several other modern predictors. Our model reveals that evolutionary conservation is higher in carboxylation sites, compared to non-carboxylation sites. The composition of arginine in carboxylation sites is higher than that of non-carboxylation sites. *CarboxySVM* can be downloaded from http://code.google.com/p/gamma-carboxylation/source/browse/trunk.

1. Introduction

Gamma-carboxylation means that glutamic acid (Glu) transforms into gamma-glutamate (Gla). It is catalyzed by a vitamin K-dependent gamma-glutamate carboxylase in the lumen of the endoplasmic reticulum (Bristol et al., 1996; Wu et al., 1991). In this modification, carbon dioxide (CO_2) is added at gamma-carbon of glutamic acid residue. Gamma-carboxylation plays important role in vascular calcification (Schurgers, 2013), bone fracture (Fusaro et al., 2011), and warfarin embryopathies (Neacsu et al., 2011).

In lab, researchers used mass spectrometry to determine gamma-carboxylation acid residues (Morris et al., 1995; Niiranen et al., 2002). Nevertheless, it is time-consuming and expensive to identify gamma-carboxylation sites. It is necessary to develop a computational method to predict the gamma-carboxylation. Lee et al. (2011) built a support vector machine model based on composition of amino acids, predicted solvent accessible surface area (ASA), predicted secondary structure and positional weighted matrix. Zhang et al. (2012) built a random forest model which fused compositions of amino acids, sequence conservations, disorder regions, secondary structures, solvent accessibilities, and physicochemical/biochemical properties to predict the gamma-carboxylation sites.

Although these methods performed accurately and robustly, some factors, for example evolutionary conservation, structural affinities of carboxylation sites, should be considered (Lee et al., 2011). Zhang et al. (2012) utilized one row of position specific scoring matrix (PSSM), which is a 20-dimensional vector, as conservation scores of a residue. Secondary structure was denoted by 3-bit binary string. However, these features are too simple. One motivation of this paper is to examine whether the improvement can be made by combining more evolutionary conservation. To this end, we investigated more evolutionary conservation of carboxylation sites and secondary structure affinities of carboxylation sites. We carefully designed two new features on PSSM profiles and some other sequence-derived descriptors to improve the performance for identifying gamma-carboxylation sites. Our method outperforms other modern predictors by empirical evaluation on benchmark datasets.

2. Materials and methods

2.1. Datasets

We retrieved 287 protein sequences, which contain 448 gamma carboxylated glutamate (Glu, E) from UniProt (version 2012_03) (The UniProt Consortium, 2012). All annotated carboxylated Glu residues are regarded as positive samples, while non-annotated

* Corresponding author. Tel.: +86 132 9991 7132; fax: +86 22 23506423.
** Corresponding author. Tel.: +86 22 23501449; fax: +86 22 23506423.
E-mail addresses: gaojz@nankai.edu.cn (J. Gao), zhni@tju.edu.cn (N. Zhang), jsruan@nankai.edu.cn (J. Ruan).
[1] These authors contributed equally to this work.

1476-9271/$ – see front matter © 2013 Elsevier Ltd. All rights reserved.
http://dx.doi.org/10.1016/j.compbiolchem.2013.09.002

Table 1
Statistics of training dataset *Train70* and test dataset *Test30*.

Data set	Non-redundant dataset	Training dataset	Test dataset
Positive samples[a]	436	307	129
Negative samples[b]	816	569	247
Total	1252	876	376

[a] Postive samples: gamma-carboxylated Glu segments (21mer).
[b] Negative samples: non-gamma-carboxylated Glu segments (21mer).

carboxylated Glu sites are regarded as negative samples. There are totally 825 non-carboxylated Glu sites which were regarded as negative.

We extracted one 21-residue peptide fragment for each annotated/non-annotated gamma-carboxylated Glu site. The gamma-carboxylation site is at the center, plus 10 residues upstream and 10 residues downstream of the modification site. The peptides with length less than 21 residues were complemented by "X". We removed the 21mer segments, which have the identity of more than 25% using BLASTCLUST (ftp://ftp.ncbi.nih.gov/blast/documents/blastclust.html). The non-redundant dataset contains 1252 sites, where there are 816 negative sites and 436 positive sites. We randomly selected 70% of the non-redundant data set as the training dataset (*Train70*), 30% as test dataset (*Test30*) (Table 1).

In this paper, *Train70* is used to build the model, and *Test30* is used to validate the model. Two datasets Zh_{train} and Zh_{test}, which were first introduced by Zhang et al. (2012), are also used to compare our model with existing methods. In Zh_{train} dataset, there are 40 proteins, where there are 115 positive sites and 251 negative sites. In Zh_{test} dataset, there are 7 proteins, where there are 34 positive sites and 84 negative sites.

2.2. Feature vectors

We designed features based on followings groups: amino acid composition (AA), predicted disorder (DS), predicted relative solvent accessibility (RSA), predicted secondary structure (SS), properties of amino acids (PA), and PSSM-based evolutionary conservation (PEC). We designed following features to encode each 21-mer peptide, which were described as following.

2.2.1. Amino acid composition (AA)

Amino acid composition is the number of residues divided by the sequence length. The properties of amino acid residues are also considered as the new features to improve the prediction performance.

AA_i composition of the 20 amino acid (AA) types, i.e., the count divided by the sequence length, where AA_i stands for the *i*th amino acid, $i = 1, \ldots, 20$ (20 features).

2.2.2. Predicted disorder (DS)

Disorder regions were predicted by DISOPRED (standalone Version 2.4) (Ward et al., 2004).

(1) Num_DIS, total number of predicted disorder and non-disorder segments (1 feature).
(2) Num_Dis/NonDis, number of predicted disorder/non-disorder segments, where Dis stands for disorder, NonDis stands for non-disorder (2 features).
(3) Percent_min/max_Dis/NonDis, predicted minimum/maximum of disorder/non-disorder segments divided by sequence length ($2 \times 2 = 4$ features).
(4) Percent_Dis/NonDis, composition of predicted disorder/non-disorder segment (2 features).

2.2.3. Relative solvent accessibility (RSA)

Solvent accessible surface area was predicted from the sequence using the Real-Spine (Faraggi et al., 2009). Relative solvent accessibility (RSA) is defined as the ratio of solvent accessible surface area of a residue observed in its three dimensional structure to that observed in an extended (Gly-X-Gly or Ala-X-Ala) tripeptide conformation. The residue is considered to be buried if its predicted RSA is less than 25%; otherwise, it is assumed to be exposed. We computed the RSA normalized using Ala-X-Ala tripeptide as suggested by Ahmad and colleagues (Ahmad et al., 2003, 2004).

(1) Num_RSA, total number of buried/exposed residue fragment (1 feature).
(2) Num_Bd/Ed, number of predicted buried, exposed segments, where Bd stands for buried residues, Ed stands for exposed residues (2 features).
(3) Percent_min/max_Bd/Ed, composition of predicted minimum/maximum of buried/exposed segments ($2 \times 2 = 4$ features).
(4) Percent_Bd/Ed composition of predicted buried/exposed segment (2 features).
(5) RSA_site_k, the RSA values on the *k*th site, where $k = 1, 2, \ldots, 21$ (21 features).

2.2.4. Predicted secondary structure (SS)

In order to explore the character of the secondary structure and disorder region surrounding carboxylation sites, secondary structures were predicted by PSIPRED(3.3) (McGuffin et al., 2000).

(1) Num_SS, total number of secondary structure (SS) fragment (1 feature).
(2) Num_C/H/E, number of predicted secondary structure segments, where C stands for coil, H stands for helix and E stands for strands (3 features).
(3) Percent_min/max_C/H/E, predicted minimum/maximum of coil/helix/strand segments divided by the sequence length ($2 \times 3 = 6$ features).
(4) Percent_C/H/E, composition of predicted coil/helix/strand. That is the number of coil/helix/strand divided by the sequence length (3 features).

2.2.5. Properties of amino acid (PA)

We used 49 selected physical chemical, energetic and conformational properties of the amino acids in this work. These properties have been used in pre-work (Gromiha and Selvaraj, 1999; Gromiha, 2005; Gromiha et al., 2006). More detailed descriptions can be found at http://www.cbrc.jp/~gromiha/fold_rate/property.html. We computed the average and standard variance of the 49 properties for each sequence.

AAproperty_i_mu/sd, the average and standard variance of the *i*th amino acids property over the whole sequence, where $i = 1, \ldots, 49$ ($49 \times 2 = 98$ features).

2.2.6. PSSM-based evolutionary conservation (PEC)

PSSM profiles have been widely used in bioinformatics problems with significant improvement (Gao et al., 2012; Jones, 1999; Xie et al., 2005). The program "blastpgp" (PSI-BLAST) was used to calculate the PSSM profiles with three iterations ($-j$ 3) and no redundant (nr) protein database. We designed three types of evolutionary conservation scores based on the PSSM profile: (a) for each residue in 21mer segment, one row (20-dimensional vector) of PSSM matrix is used as the features of the residue. This is motived by Zhang et al. (2012); (b) for each residue, we computed expected