Acta Crystallographica Section D
Biological Crystallography
ISSN 1399-0047

Lysine carboxylation: unveiling a spontaneous post-translational modification

David Jimenez-Morales,[a] Larisa Adamian,[a] Dashuang Shi[b] and Jie Liang[a]*

[a]Department of Bioengineering, University of Illinois at Chicago, 851 South Morgan Street, Room 218, Chicago, IL 60607, USA, and [b]Children's National Medical Center, Center for Genetic Medicine Research, 111 Michigan Avenue NW, Washington, DC 20010-2970, USA

Correspondence e-mail: jliang@uic.edu

Received 21 June 2013
Accepted 22 August 2013

The carboxylation of lysine residues is a post-translational modification (PTM) that plays a critical role in the catalytic mechanisms of several important enzymes. It occurs spontaneously under certain physicochemical conditions, but is difficult to detect experimentally. Its full impact is unknown. In this work, the signature microenvironment of lysine-carboxylation sites has been characterized. In addition, a computational method called *Predictor of Lysine Carboxylation* (*PreLysCar*) for the detection of lysine carboxylation in proteins with available three-dimensional structures has been developed. The likely prevalence of lysine carboxylation in the proteome was assessed through large-scale computations. The results suggest that about 1.3% of large proteins may contain a carboxylated lysine residue. This unexpected prevalence of lysine carboxylation implies an enrichment of reactions in which it may play functional roles. The results also suggest that by switching enzymes on and off under appropriate physicochemical conditions spontaneous PTMs may serve as an important and widely used efficient biological machinery for regulation.

1. Introduction

The addition of a carboxyl group to the ε-amino group of the lysine side chain is a post-translational modification (PTM) with remarkable consequences (Stec, 2012). Lysine is positively charged at physiological pH. However, the carboxyl group drastically changes the properties of the lysine side chain by providing negative charge. This modification occurs spontaneously under basic pH conditions involving carbon dioxide in solution (Park & Hausinger, 1995) and without the mediation of any other enzyme.

Lysine carboxylation is a chemical event that is required for the catalytic function of several enzymes. The modified lysine can play a direct role in the catalytic reaction as an acidic residue (Dementin *et al.*, 2001; Golemi *et al.*, 2001). More frequently, as a co-catalytic determinant, it bridges metal ion(s) (Meulenbroek *et al.*, 2009; Stec, 2012), which can trigger conformational and physicochemical changes that are essential for the function of the protein (Stec, 2012). Lysine carboxylation is also involved in rearrangement of the active site through the formation of strong hydrogen-bonding interactions, creating the appropriate orientation of side chains important for the function of the protein (Lorimer *et al.*, 1976; Wu *et al.*, 2008).

The extent of lysine carboxylation is currently unknown. Unlike other covalent modifications of lysine (Li *et al.*, 2010), carboxylation has not been fully investigated, partly owing to

difficulties associated with the highly unstable nature of this labile PTM, as the carboxyl group is spontaneously released in acidic conditions. Because of this complication, mass spectrometry cannot detect lysine carboxylation (Golemi *et al.*, 2001). This chemical instability has resulted in an incorrect perception of the scope of this PTM.

X-ray crystallography can identify carboxylated lysine residues, but is not exempt from problems. Firstly, the crystallization of a protein with a labile chemical modification is challenging. Secondly, it can be difficult to distinguish carboxylation from several other post-translational modifications based on the measured electron-density map alone. Finally, the most frequent damage produced by third-generation synchrotron radiation is the decarboxylation of acidic residues (Garman, 2010; Ravelli & McSweeney, 2000), which may alter the state of this PTM.

Computational methods can be valuable in overcoming the difficulties affecting the detection of spontaneous PTMs such as lysine carboxylation. However, there are currently no computational tools available for the prediction of lysine carboxylation. In this work, we describe a computational method for this task. Using currently available protein structures with carboxylated lysine residues, our method can detect modified lysine residues with a sensitivity and a specificity of 87.1% and 99.7%, respectively. It also predicts carboxylation of lysine residues in proteins with and without similarity to proteins in which lysine carboxylation is known to occur. We also found that some lysine residues that have been reported to be carboxylated are in fact non-carboxylated in the functional state of the protein. Finally, we assess the extent of lysine carboxylation in the protein structure database and discuss the implications of a broader prevalence of spontaneous PTMs as a biological regulatory system.

2. Methods

2.1. KCX and LYS data sets

A total of 251 protein structures were downloaded from the Protein Data Bank (Berman *et al.*, 2002) with at least one subunit containing a carboxylated lysine residue (denoted KCX proteins). Three different data sets were constructed. One of them contained the entire data set without removal of redundancy. The other two data sets consisted of KCX proteins with maximum similarities of 90 and 40%. To reduce the redundancy of multiple structures of the same protein, we selected that with the highest resolution from those sharing more than 90% sequence identity. The 90% sequence-identity cutoff was set to ensure that sufficient redundancy was removed (from 251 to 62 structures) while simultaneously retaining a sufficient amount of data. For example, of 24 structures of class D β-lactamases available in the Protein Data Bank, only six structures were included in our data set. All of them are different oxacillinase-type β-lactamases (OXAs; Poirel *et al.*, 2010), *i.e.* OXA-1, OXA-2, OXA-10, OXA-24, OXA-46 and OXA-48 (Supplementary Table S1[1]). To reduce redundancy, for each protein structure only the chain(s) solved with the carboxylated lysine residue (Kcx) was selected and redundancy reduction was performed by (i) clustering subunits sharing more than 90% sequence identity using *BLASTClust* (Altschul *et al.*, 1990; Wheeler & Bhagwat, 2007) and (ii) selecting the chain with the best resolution from each cluster. 65 structures met the 90% criterion, with an average resolution of 1.99 ± 0.5 Å. As will be explained below, three out of 65 were treated as non-carboxylated and were removed, resulting in a final set of 62 protein structures. Kcx residues (n = 62) were included in the data set denoted as KCX sites, with one per protein (Supplementary Table S1). In most cases, there is biological understanding of the role of the Kcx residue in the function of the protein (Supplementary Note S3). Details of each enzyme, its EC number and the role of the Kcx residue can be found in Supplementary Table S1. Since no proteins are known to exist with more than one carboxylated lysine residue in the same protein chain, we selected the remaining lysine residues from the same 62 proteins as a control (1275 residues). Since the number of buried lysine residues is generally limited in proteins and most of the Kcx residues were found to be buried, we added an equivalent number of buried lysine residues (62), obtained randomly from the set of high-resolution protein structures described below, to this data set. A total of 1337 amino acids were thus included in the data set of uncarboxylated lysine residues (denoted as LYS sites). The kcx90rr data set consisted of 62 KCX sites and 1337 LYS sites. The location of the Kcx and Lys residues on the surface or buried in the interior of the proteins was determined using *CASTp* (Dundas *et al.*, 2006), which uses weighted Delaunay triangulation and alpha-shape theory to measure the surface-accessible residues. The same steps were repeated using a 40% cutoff, resulting in 43 structures with an average resolution of 1.88 ± 0.45 Å. A total of 43 KCX and 954 LYS sites were included in the kcx40rr data set. The kcx251 data set was constructed using the entire set of KCX proteins without redundancy reduction, resulting in 251 KCX sites and 4259 LYS sites.

2.2. High-resolution protein structures

We used *PDBselect* (Griep & Hobohm, 2010) to obtain a subset of high-resolution protein structures solved by X-ray crystallography (<1.5 Å). Each protein subunit was treated independently. Two different data sets were created with different levels of redundancy reduction (90 and 40%). For example, in the PDB90hr data set those subunits sharing more than 90% sequence identity were clustered using *CD-HIT* (Li & Godzik, 2006). The structure with the best resolution was selected from each cluster. Only proteins greater than 200 amino acids in length were selected, resulting in a final set of 575 structures (Supplementary List S3). The same steps were repeated for the 40% data set (PDB40hr), resulting in 381 structures (Supplementary List S4).

[1] Supporting information has been deposited in the IUCr electronic archive (Reference: LV5045).

2.3. Redundancy reduction of the PDB database

To further explore the incidence of lysine carboxylation in the PDB, we first reduced the number of similar proteins to 90 and 40% sequence identity. We combined *PDBselect* and *CD-HIT* as described earlier. As a result, the PDB90 data set consisted of 14 262 protein structures solved by X-ray crystallography with a resolution equal to or greater than 1.5 Å and more than 200 amino acids (Supplementary List S1). A total of 291 434 lysine residues were identified. The 40% data set (PDB40) consisted of 8508 proteins and 176 150 lysine residues (Supplementary List S2).

2.4. Measurements of the microenvironment of KCX sites and LYS sites

Every amino acid from both the KCX sites and LYS sites data sets was defined by the frequencies of the amino-acid side chains, water molecules and metal ions found within 5 Å. To avoid statistical bias as a consequence of the extra length of the Kcx residue, we did not include the carboxyl group of the Kcx residue. Instead, we projected an atom (named pCX for **p**rojected **C**arbo**X**yl group) onto the tip of the NZ atom of both the Kcx and the Lys side chains, taking as a direction the average direction of the remaining atoms of the side chain. The pCX atom was also used to measure distances to all of the other components of the microenvironment. Considering that most noncovalent bonding interactions occur at less than 4 Å (Hibbert & Emsley, 1991), and since we are not adding O atoms to the pCX, we used 5 Å as a distance cutoff to ensure that sufficient spatially proximal atoms are found in both the Kcx and the Lys microenvironments. The amino acids in the microenvironment were grouped according to the physicochemical properties of the side chains, *i.e.* negatively charged (NEG; Asp and Glu), positively charged (POS; Arg, His and Lys), small polar (ST; Ser and Thr), large polar (NQ; Asn and Gln), aromatic (ARO; Trp, Tyr and Phe) and hydrophobic (HYD; Met, Ile, Leu and Val). The metal ions considered were Zn, Mg, Co, Fe, Ni and Mn. Overall frequencies for the microenvironment of both KCX sites and LYS sites were finally calculated (Supplementary Tables S2 and S3).

2.5. The naïve Bayesian predictor

We used a naïve Bayesian probabilistic classifier to distinguished lysine residues that are and are not carboxylated. The naïve assumption is that all such structural features can effectively be treated as independent. It is known that Bayesian classifiers can achieve excellent results even if the independence of the features is questionable (Zhang, 2004). Two main parameters are calculated by our Bayesian method. The first is the probability distribution of the features, which we approximated with relative frequencies from the training set (F_1, F_2, ..., F_n; Supplementary Tables S1 and S2). The other is the prior probability, which we arbitrarily selected as a best reasonable guess of the frequency of lysine carboxylation (Supplementary Note S2). For any given lysine residue, the posterior probability of being carboxylated [$p(\mathrm{C_{Kcx}}|F_1, \ldots, F_n)$] and the posterior probability of not being carboxylated [$p(\mathrm{C_{Lys}}|F_1, \ldots, F_n)$] are calculated as follows,

$$p(\mathrm{C_{Kcx}}|F_1, \ldots, F_n) = \frac{p(\mathrm{C_{Kcx}})p(F_1, \ldots, F_n|\mathrm{C_{Kcx}})}{p(F_1, \ldots, F_n|\mathrm{C_{Kcx}}) + p(F_1, \ldots, F_n|\mathrm{C_{Lys}})} \quad (1)$$

and

$$p(\mathrm{C_{Lys}}|F_1, \ldots, F_n) = \frac{p(\mathrm{C_{Lys}})p(F_1, \ldots, F_n|\mathrm{C_{Lys}})}{p(F_1, \ldots, F_n|\mathrm{C_{Kcx}}) + p(F_1, \ldots, F_n|\mathrm{C_{Lys}})}, \quad (2)$$

where $p(\mathrm{C_{Kcx}})$ and $p(\mathrm{C_{Lys}})$ are the prior probabilities. The likelihood is calculated as follows,

$$p(F_1, \ldots, F_n|\mathrm{C_{Kcx}}) = p(F_1|\mathrm{C_{Kcx}}) \cdot p(F_2|\mathrm{C_{Kcx}}) \cdots p(F_n|\mathrm{C_{Kcx}}) \quad (3)$$

and

$$p(F_1, \ldots, F_n|\mathrm{C_{Lys}}) = p(F_1|\mathrm{C_{Lys}}) \cdot p(F_2|\mathrm{C_{Lys}}) \cdots p(F_n|\mathrm{C_{Lys}}) \quad (4)$$

where $p(F_n|\mathrm{C_{Kcx}})$ is the probability of the feature n when lysine is carboxylated and $p(F_n|\mathrm{C_{Lys}})$ is that when lysine is not carboxylated.

The predictor classifies any given lysine according to the largest posterior probability value, *i.e.* a Lys vector $L_x = (x_1, x_2, \ldots, x_n)$ will be classified as Kcx if

$$p(\mathrm{C_{Kcx}}|x_1, x_2, \ldots, x_n) > p(\mathrm{C_{Lys}}|x_1, x_2, \ldots, x_n), \quad (5)$$

and as non-carboxylated (Lys) otherwise. We consider the difference between the probability values to be a measure of confidence. We developed a tool implementing this method and named it *PreLysCar* (*Predictor of Lysine Carboxylation*).

2.6. Measures of performance

We assessed the performance of the Bayesian classifier by the technique of leave-one-out cross-validation (Supplementary Note S2).

2.7. Electron-density maps and remodeling

$2F_o - F_c$ and $F_o - F_c$ electron-density maps were downloaded from the Electron Density Server (Kleywegt *et al.*, 2004). Remodeling of lysine residues predicted to be Kcx was performed with *Coot* (Emsley *et al.*, 2010) and refinement was performed with *REFMAC* (Murshudov *et al.*, 2011).

2.8. Tool availability

PreLysCar is available at http://tanto.bioengr.uic.edu/prelyscar. *PreLysCar* uses the features that gave the best performance in the leave-one-out cross-validation test (the total number of elements and the numbers of positively charged, negatively charged, polar uncharged, hydrophobic and aromatic residues, ions and waters) and the frequencies from the entire KCX sites and LYS sites data sets. The user can submit a PDB file and choose the prior probability. As a result, *PreLysCar* lists the predicted Kcx residues (if any).

3. Results

3.1. Data sets analyzed

We obtained 62 protein structures (after 90% redundancy reduction) from the Protein Data Bank (Berman *et al.*, 2002) with at least one subunit containing a carboxylated lysine residue (denoted Kcx). In most cases, there is biological understanding of the role of the Kcx residue in the function of the protein (Supplementary Note S3). Details of each enzyme, its EC number and the role of the Kcx residue can be found in Supplementary Table S1. A total of 62 carboxylated lysine residues were included in the KCX site data set. Since no proteins with more than one carboxylated lysine residue in the same protein chain are known to exist, we selected the remaining lysine residues from the same 62 proteins as a control. A total of 1337 amino acids were included in the data set of uncarboxylated lysine residues, denoted LYS sites (see §2).

3.2. Signature microenvironment of KCX sites

We next examined the microenvironment of both KCX sites and LYS sites and determined the frequencies at which amino acids, metal ions and water molecules were found within 5 Å of carboxylated and uncarboxylated lysine side chains. LYS sites were further divided into subsets of buried and surface lysine residues (Fig. 1 and Supplementary Table S3).

A defining feature of KCX sites is the large number of packed atoms of other residues, water molecules and ions found in the proximity of the Kcx side chain. On average, 12.9 ± 1.1 of these are found within 5 Å. This is in part a consequence of the buried nature of the carboxylated lysine residue. All carboxylated lysine residues were found to be buried and inaccessible from the surface. Amino acids converging structurally at KCX sites were found to be dispersed in different regions along the primary sequence of the protein, although many of these fragments (from the first to the last amino-acid part of the KCX site) tended to be located towards the protein N-terminus (Fig. 2*b*).

The distribution of the total number of amino acids surrounding uncarboxylated lysine residues varies between surface and buried LYS sites. Only 7% of surface LYS sites showed more than nine, in contrast to 87% of buried LYS sites and 100% of KCX sites. As expected, most of the distributions of amino-acid groups vary significantly between KCX sites and surface LYS sites (Fig. 1 and Supplementary Tables S2 and S3).

3.2.1. Comparison of KCX sites and buried LYS sites. The distribution of the total number of residues found within 5 Å of buried LYS sites is similar to that of KCX sites. However, KCX sites tend to be more crowded. About 94% of KCX sites have more than 12, in contrast to 58% of buried LYS sites. Although some similarities exist in the distribution of the amino acids found in the microenvironments (*e.g.* hydrophobics and aromatics), there are important differences. Long polar residues (Asn and Gln) are most likely to be found in buried LYS sites (62 and 24% of buried LYS and KCX sites, respectively; Supplementary Tables S2 and S3). In addition, although water molecules are not uniformly resolved in X-ray structures, we found that the number of water molecules varies between KCX sites and buried LYS sites, with three or

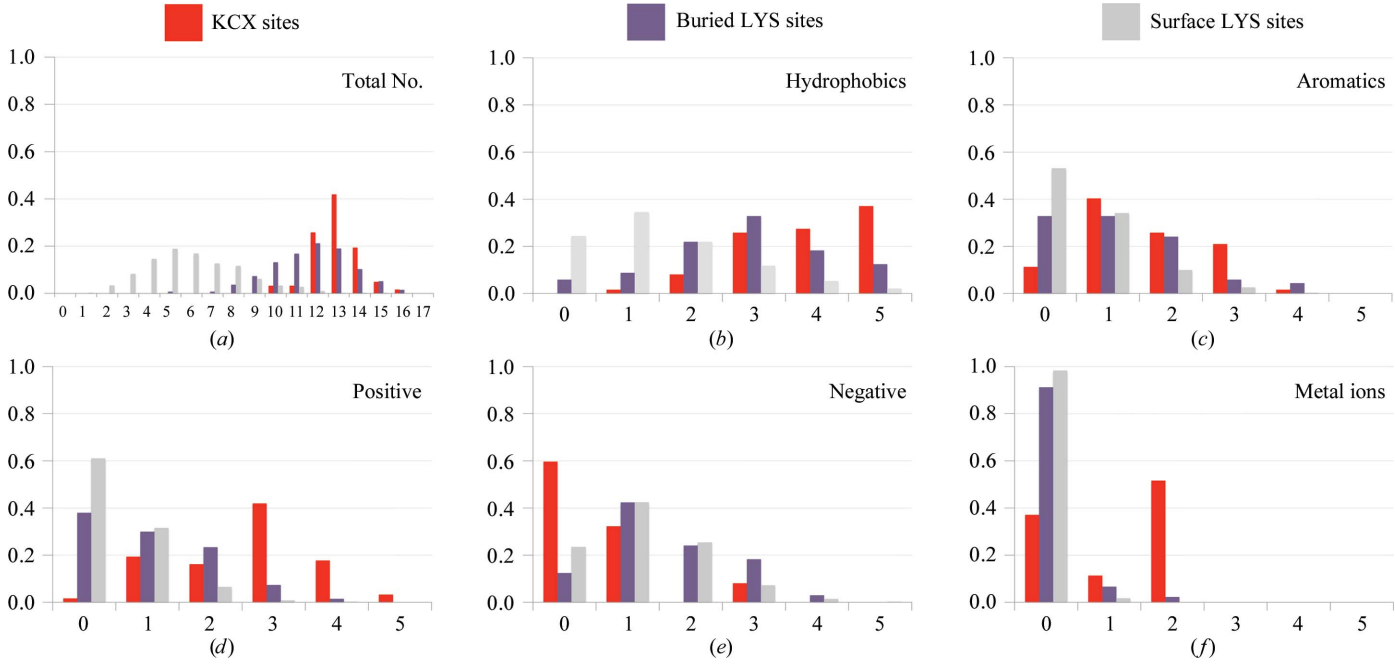


Figure 1
Composition of the microenvironment of KCX sites and LYS sites (buried and surface lysine residues). Frequency of amino acids, metal ions and water molecules found within 5 Å of the side chain of Kcx (black bars), buried Lys (magenta bars) and surface Lys (gray bars) residues (see also Supplementary Tables S2 and S3). The amino acids found in the microenvironments are grouped according to their main physicochemical properties, *i.e.* positively charged (Arg, Lys and His), negatively charged (Asp and Glu), aromatic (Trp, Phe and Tyr) and hydrophobic (Ile, Leu, Val and Met) residues. The *x* axis represents the number of amino acids around the KCX and LYS sites, while the *y* axis represents the frequencies of each count.

more water molecules found in 49% of LYS sites but only in 13% of KCX sites (Supplementary Tables S2 and S3).

An especially interesting difference between KCX sites and buried LYS sites is the number of ionizable residues. Lysine is a positively charged residue. The positive charge of lysine residues is expected to be balanced by a negative charge. The enrichment of negatively charged residues (Asp and Glu) near buried (and surface) LYS sites confirms this expectation. About 88% of the buried LYS sites are found to contain one or more negatively charged amino acids (Fig. 1*e*). In contrast, positively ionizable residues (Arg, Lys and His) are less likely to be found in the microenvironment of buried LYS sites (about 60%; Fig. 1*d*). The overall distribution of ionizable residues in the microenvironment of KCX sites is significantly different from that of buried LYS sites. After carboxylation, the modified lysine residue acquires a negative charge. This drastic change in the electrostatic properties of the modified lysine residue is reflected in the composition of ionizable residues found in its microenvironment. About 98% of the ionizable residues found within 5 Å of the Kcx side chain are positively charged, in contrast to the 40% of KCX sites with one or more negatively charged residues (Figs. 1*d* and 1*e*).

3.2.2. KCX metal-ion centers. The carboxyl group of the Kcx residue is commonly found in contact with positively charged ion(s). In 41 of the 62 protein structures, carboxylated lysine residues interact with metal ions in either single or multiple metal-ion centers. The majority (32) of the metal centers are binuclear, with nine mononuclear centers (Figs. 1*f* and 2*c* and Supplementary Table S1). Overall, we found that there are six different metal ions that can interact with the Kcx side chain: Zn^{2+}, Mg^{2+}, Co^{2+}, Fe^{3+}, Ni^{2+} and Mn^{2+}. Binuclear centers can contain either the same or different pairs of metal ions (Supplementary Table S1). Among the six ions, Ni^{2+} and Fe^{3+} are only found in binuclear metal centers. The divalent ions Zn^{2+} and Co^{2+} are found in both mononuclear and binuclear metal centers, while Mg^{2+} and Mn^{2+} are only found in mononuclear centers. His and Asp are almost invariably present in all analyzed Kcx-containing metal-binding sites. Some binuclear sites contain unusual zinc ligands such as an amide peptide backbone carbonyl (*e.g.* Gly in ureases) or a hydroxyl group from Tyr (*e.g.* adenine deaminase from *Enterococcus faecalis*, metal-dependent hydrolase from *Lactobacillus casei* and organophosphorus hydrolase from *Deinococcus radiodurans*). Mononuclear metal ions usually form catalytic complexes with the Kcx residue acting as an oxygen donor (Auld, 2001). Binuclear metal sites belong to the co-catalytic type of metal-binding site (Auld, 2001). They usually contain two metal ions in close proximity bridged by the side-chain moiety of the Kcx residue. Asp, Glu or His residues have also been found to carry out similar roles in

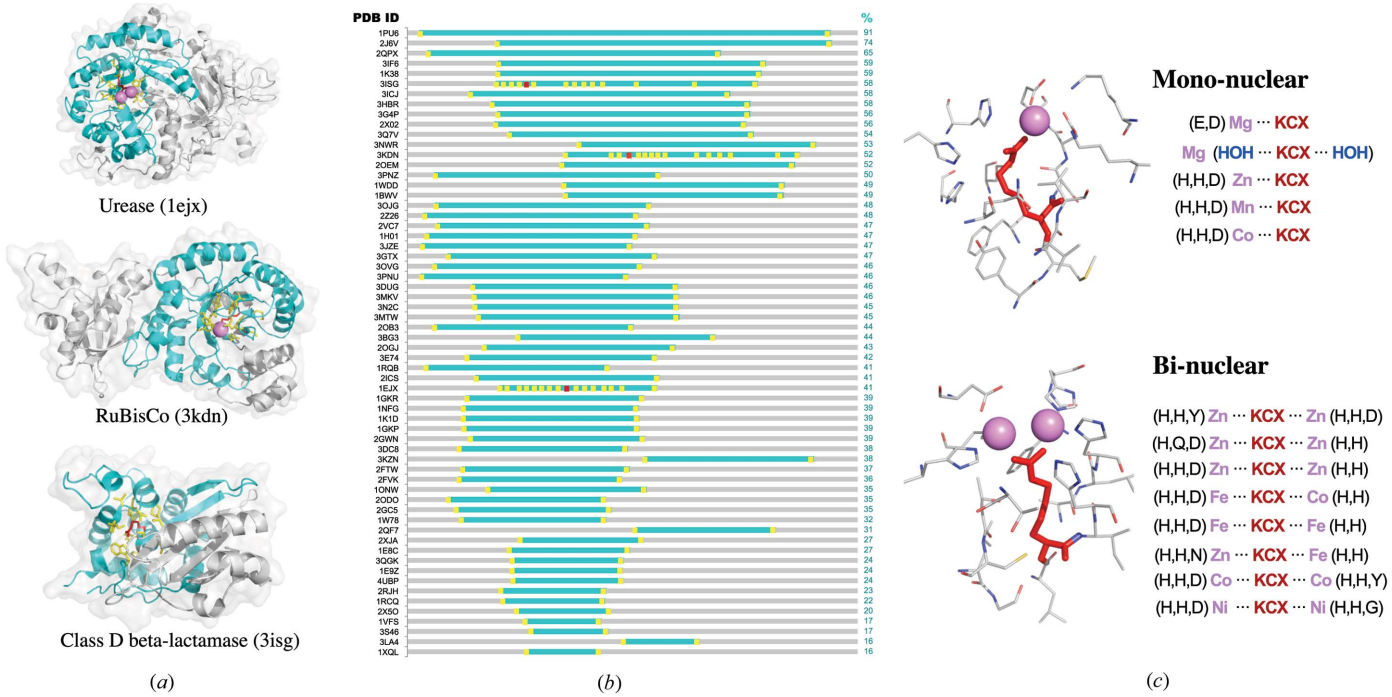


Figure 2
Defining features of KCX sites. (*a*) Examples of proteins requiring functional carboxylated lysine residues. Examples of Kcx as a co-catalytic determinant involved in the formation of binuclear and mononuclear metal-ion centers (*e.g.* urease and RuBisCo, respectively) and Kcx as a catalytic determinant (*e.g.* class D β-lactamase). (*b*) Dispersion along protein sequences of amino acids that are part of the KCX site. Gray bars represent protein sequences, which are scaled for comparison. Cyan portions represent the fragment from the first to the last of the amino acids found in KCX sites. Yellow dots represent amino acids that are part of the KCX site and the red dot represents the Kcx residue (only 1ejx, 3kdn and 3isg shown). The left axis shows the PDB code (PDB ID) and the right axis the percentage of the extent of the protein sequences that the KCX sites occupy. A lack of a sequence motif was concluded (see Supplementary Fig. S1) (*c*) Summary of metal-ion center motifs. The side chains of the residues interacting with the metal ion are given in parentheses. For example, (H,H,D)Zn represents the side chains of two His residues and one Asp residue interacting with a zinc ion. Kcx side chains can either interact with one ion, *e.g.* (H,H,D)Zn· · ·KCX, or can bridge two metal ions, *e.g.* (H,H,D)Zn· · ·KCX· · ·Zn(H,H). The stick representation shows detail of the metal-ion centers of urease and RuBisCo.

Table 1
Performance of the Bayesian classifier *PreLysCar* (*Predictor of Lysine Carboxylation*).

(*a*) Tests on KCX proteins (training data set). Three different data sets were constructed. Firstly, kcx251 was constructed using the entire data set of KCX proteins (251 and 4259 Kcx and Lys residues, respectively). The other two data sets contain different levels of redundancy reduction of the original data set (see §2). kcx40rr and kcx90rr contain 43 and 62 KCX proteins, respectively, after 40 and 90% redundancy reduction of the original data set (251 proteins). kcx40rr contains 43 KCX and 954 LYS sites. kcx90rr contains 62 KCX sites and 1337 LYS sites (Supplementary Table S1). Leave-one-out cross-validation tests were performed on each data set and measures of performance were calculated (see also Supplementary Table S4), *i.e.* sensitivity (SEN), specificity (SPF), accuracy (ACU), positive predictive value (PPV), negative predictive value (NPV) and Matthews correlation coefficient (MCC), according to the number of true positives (TP), true negatives (TN), false positives (FP) and false negatives (FN).

	SEN	SPF	ACU	PPV	NPV	MCC	TP	TN	FP	FN
kcx40rr	0.8372	0.9958	0.9890	0.9000	0.9927	0.8623	36	951	4	7
kcx90rr	0.8710	0.9970	0.9914	0.9310	0.9940	0.8961	54	1333	4	8
kcx251	0.8805	0.9991	0.9925	0.9822	0.9930	0.9261	221	4255	4	30

(*b*) Predictions on the PDB database. *PreLysCar* was executed on a subset obtained from the PDB database with a 90% reduction in redundancy (PDB90; see §2). The predicted KCXs (pKCX) in the high-resolution fraction (<1.5 Å) were assumed to be incorrect and the false discovery rate (FPR) was calculated. The FPR was further used to estimate the expected error (*e*-error) on protein structures with a resolution of >1.5 Å and the corresponding expected number of correct predictions (*e*-correct). See also Supplementary Table S5.

	PDB90, <1.5 Å		PDB90, >1.5 Å		
	pKCX	FPR (%)	pKCX	*e*-error	*e*-correct
kcx40rr	19/575	3.3	659/14262	471	188
kcx90rr	11/575	1.9	543/14262	271	272
kcx251	16/575	2.8	601/14262	399	202

other proteins (Auld, 2001). Residues interacting with metal ions often come from regions distributed along the entire length of the protein (Auld, 2001), as we observed for KCX sites (see Fig. 2*b*), which suggests that metal sites may be important not only for the catalytic function but also for the overall folding of the protein.

In summary, we have characterized the physicochemical environment of the KCX site, which is likely to be associated with the carboxylation event. The relatively large number of positively ionizable residues around the positively charged lysine side chain can be rationalized if the lysine residue becomes carboxylated and, consequently, negatively charged. In addition, the hydrophobic environment may also promote the carboxylation event.

3.3. *Predictor of Lysine Carboxylation* (*PreLysCar*)

Since there are no sequence motifs associated with lysine carboxylation (see Supplementary Note S1), the use of structural information is essential in identifying KCX sites. The components of the microenvironments of both KCX and LYS sites were used as features for our computational prediction method, which is based on a naïve Bayesian model (see §2). We implemented a Bayesian classifier called *PreLysCar* (*Predictor of Lysine Carboxylation*). The effectiveness of the classifier was assessed by performing leave-one-out cross-validation tests with the entire data set and at different thresholds of redundancy reduction: both 90 and 40% (Table 1*a* and Supplementary Note S2). For example, in the kcx90rr data set *PreLysCar* correctly classified 54 of 62 KCX sites (87% sensitivity) and 4255 of 4259 LYS sites (99.7% specificity). The positive and negative predictive values (93.1 and 99.4%, respectively) underline the high probability that the predictions are correct. The Matthews correlation coefficient (0.89), an indicator of performance when the classes have different sizes, reflects the overall excellent reliability of our predictor (Supplementary Note S2). Both kcx251 and kcx40rr performed similarly (see Table 2*a*).

3.4. Searching the Protein Data Bank

Owing to the difficulties in the experimental detection of carboxylated lysine residues, we hypothesized that undetected carboxylated lysine residues may exist in the Protein Data Bank. We searched three-dimensional protein models for lysine residues that could be carboxylated but were reported as non-carboxylated. In addition, we also examined lysine residues to identify those that could truly be non-carboxylated under the experimental conditions but that could become carboxylated under more favorable conditions.

We applied *PreLysCar* to a subset of protein structures from the PDB (>200 residues) solved by X-ray crystallography and consisting of 14 261 protein chains after 90% redundancy removal (Supplementary List S1). A total of 291 434 lysine residues were found in these proteins and their microenvironments were determined (§2). *PreLysCar* predicted that 543 protein structures contained a carboxylated lysine residue (denoted 'predicted'; pKcx; Supplementary List S5). Table 2 lists examples of pKcx. For some of these proteins, overwhelming evidence exists supporting our predictions. For example, there are lysine residues predicted to be carboxylated that belong to proteins for which structures had been solved in a different experiment in which the lysine residue was indeed carboxylated. This discrepancy generally occurred as a consequence of either the binding of an inhibitor (Radha Kishan *et al.*, 2005; Xiang *et al.*, 2010) or experimental conditions that prevented carboxylation (Okano *et al.*, 2002). We also identified lysine residues known to be carboxylated in multimeric proteins where one chain has an uncarboxylated lysine but it was reported as carboxylated in the remaining chains. For example, the phosphotriesterase from *Pseudomonas diminuta* contains Lys169, which was predicted to be carboxylated (PDB entry 1psc). The structure was described with this lysine residue carboxylated and bridging two atoms of cadmium at the active site of this holoenzyme (Benning *et al.*, 1995). However, it appears as uncarboxylated in the atomic coordinates available in the PDB. Another example is Lys84 of carbapenemase OXA-24 (PDB entry 2jc7). This enzyme belongs to the family of class D β-lactamases, which is well known to have a carboxylated lysine as part of the active site (Poirel *et al.*, 2010). The $F_o - F_c$ electron-density map showed a clear positive peak at the tip of Lys84. In this case, we can

Table 2
Examples of proteins with a predicted carboxylated lysine residue (pKcx).

The following are given: the PDB code and chain where the pKcx is predicted (PDB_chain), the residue number of the pKcx, the resolution of the protein structure in Å (RES) and the percentage similarity to any protein known to have a Kcx residue (SIM), where 'Fg' indicates that the similarity is detected in a fragment of a known KCX protein and 'No' indicates that no sequence similarity is detected.

PDB_chain	pKcx	RES	SIM	Protein name	Source
2jc7_*A*	84	2.50	100	Carbapenemase OXA-24	*Acinetobacter baumannii*
3msr_*A*	153	2.16	100	Amidohydrolase	*Mycoplasma synoviae*
3feq_*A*	188	2.63	98	Amidohydrolase Sgx9260c	Unknown
1iwa_*A*	201	2.60	97	RuBisCo	*Galdieria partita*
3igh_*X*	278	1.95	70	Uncharacterized metal-dependent hydrolase	*Pyrococcus horikoshii*
2vr2_*A*	159	2.80	60	Human dihydropyrimidinase	*Homo sapiens*
3hm7_*A*	150	2.60	40	Allantoinase	*Bacillus halodurans*
2cwx_*A*	186	2.00	40	RuBisCo	*Pyrococcus horikoshii*
2qs8_*A*	194	2.33	36	Dipeptidase	*Alteromonas macleodii*
2gok_*A*	155	1.87	30	Imidazolonepropionase	*Agrobacterium tumefaciens*
2bb0_*A*	149	2.00	30	Imidazolonepropionase	*Bacillus subtilis*
2oof_*A*	141	2.20	30	Imidazolonepropionase	Environmental sample
4fr4_*A*	184	2.29	26	Serine/threonine protein kinase	*Homo sapiens*
1ckj_*A*	237	2.46	24	Casein kinase 1δ	*Rattus norvegicus*
1iyx_*A*	390	2.80	23	Enolase	*Enterococcus hirae*
2p2s_*A*	96	1.25	Fg	Putative oxidoreductase	*Erwinia carotovora atroseptica*
3c4u_*A*	251	1.83	Fg	Class II fructose-bisphosphate aldolase	*Helicobacter pylori*
2jc7_*D*	205	2.00	Fg	Glucarate dehydratase	*Escherichia coli*
4g9p_*A*	204	1.55	Fg	GcpE (IspG)–MEcPP	*Thermus thermophilus*
1f3o_*A*	44	2.70	Fg	MJ0796 ATP-binding cassette	*Methanocaldococcus jannaschii*
2fiq_*A*	279	2.25	No	Putative tagatose 6-phosphate kinase	*Escherichia coli*
1qwr_*A*	96	1.80	No	Mannose 6-phosphate isomerase	*Bacillus subtilis*
3tcs_*A*	145	1.88	No	Putative racemase	*Roseobacter denitrificans*
3u4f_*A*	145	1.90	No	Mandelate racemase	*Roseovarius nubinhibens*
3uhj_*A*	298	2.34	No	Glycerol dehydrogenase	*Sinorhizobium meliloti*
1xe7_*A*	147	1.75	No	Yml079w	*Saccharomyces cerevisiae*
1mdo_*A*	188	1.70	No	Aminotransferase	*Salmonella typhimurium*
2as9_*A*	32	1.70	No	Serine protease	*Staphylococcus aureus*
2eb0_*B*	294	2.20	No	Mn-dependent inorganic pyrophosphatase	*Methanocaldococcus jannaschii*
2ejc_*A*	149	2.40	No	Pantoate-β-alanine ligase	*Thermotoga maritima*
2huo_*A*	127	2.00	No	Inositol oxygenase	*Mus musculus*
2ptx_*A*	343	1.90	No	Enolase	*Trypanosoma brucei*
2w8s_*B*	337	2.40	No	Phosphonate monoester hydrolase	*Burkholderia caryophylli*
3igy_*B*	357	2.00	No	Phosphoglycerate mutase	*Leishmania mexicana*

remodel Lys84 to Kcx84 with confidence (Fig. 3). See Table 2 and Supplementary Note S5 for more related examples and extended descriptions.

Among other predicted KCX proteins, the composition and the structural arrangement of the residues at the KCX sites showed the defining features of known KCX sites described earlier. For example, lysine residues that were predicted to be carboxylated were found buried in highly populated cavities, with three or more histidine residues located at the mouth of the cavity and a large number of hydrophobic and aromatic residues as part of the microenvironment, all within 5 Å from the pKcx side chain. A number of these pKCX proteins were found to share 90–20% sequence identity to known KCX proteins from different species, indicating that they are likely to be orthologs and therefore that the carboxylation event may occur in these homologous proteins as well (Table 2). For example, Lys278 from the uncharacterized metal-dependent hydrolase from *Pyrococcus horikoshii* (PDB entry 3igh) was predicted to be carboxylated. This protein shows 70% sequence identity to another metal-dependent hydrolase from *P. furiosus* (PDB entry 3icj), in which Lys294 is known to be carboxylated. Futhermore, both microenvironments show high sequence and structural similarity. Another example is the alanine racemase from *E. faecalis* (pKcx132; PDB entry 3e5p), which shares 50% sequence identity with the alanine racemase from *Streptococcus pneumoniae* (Kcx129; PDB entry 3s46). An extended description of these and many other selected predicted KCX proteins can be found in Table 2 and Supplementary Note S5.

Sequence similarity was also discovered along short fragments between predicted KCX proteins and KCX proteins. Although the overall protein topology was found to be different for some of them, these fragments share local structural similarity, generally consisting of α-helices. These fragments are enriched in the region where both Kcx and pKcx are found. Some examples include the putative oxidoreductase from *Erwinia carotovora atroseptica* (PDB entry 2p2s) and the class II fructose-bisphosphate aldolase from *Helicobacter pylori* (PDB entry 3c4u) (see Supplementary Note S5 for details).

The uncharacterized protein YML079w from *Saccharomyces cerevisiae* (PDB entry 1xe7) is a protein predicted to have a carboxylated lysine with no sequence similarity detected to any known KCX protein. Lys147 is found buried in a central location of the protein structure. At least four His residues and one Glu residue are in positions that resemble a typical KCX site. This protein of unknown function was solved at 1.75 Å resolution and pH 5.6, which could have prevented carboxylation. Despite the high resolution, the side chain of Lys147 was modeled in two different conformations, which may provide an indication of an unstable state of Lys147 (see Table 2 and Supplementary Note S5 for more predictions with no sequence similarity to known KCX proteins).

3.5. Unveiling incorrectly reported Kcx residues

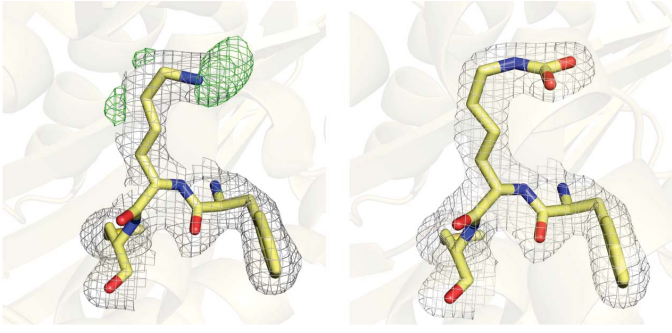
Our method initially predicted nine out of 65 Kcx residues as non-carboxylated (false negatives). A detailed analysis of the incorrect predictions unveiled that three of the nine Kcx residues are likely to be non-carboxylated and therefore the classification is likely to be correct. One of them is the class C β-lactamase from *Enterobacter cloacae* (PDB entry 2p9v). Lys315 was correctly reported as Kcx owing to carbamate cross-linking between Lys315 and Ser64 as a consequence of the action of an inhibitor (Wyrembak *et al.*, 2007). However,

carboxylation does not occur at Lys315 in functional class C β-lactamase proteins. Therefore, we consider our prediction to be correct.

The other two Kcx residues are found in TdcF from *Escherichia coli* and α-L-fucosidase from *Thermotoga maritima*. Detailed analysis of these Kcx residues suggests that they are likely to be non-carboxylated. For example, TdcF is a member of the highly conserved YjgF/YER057c/UK114 family, although its biological function is unknown (Burman *et al.*, 2007). Kcx58 of TdcF (PDB entry 2uyn) is likely to be a non-carboxylated lysine residue for the following reasons. Firstly, the authors suggested that the Kcx modification could be an artifact with no biological importance (Burman *et al.*, 2007). Secondly, the ambiguity of the electron-density map at the tip of Lys58 compelled the authors to model the carboxylated lysine in two different positions. Thirdly, known Kcx residues are mostly buried, but Kcx58 is on the surface and is quite accessible to solvent. Finally, no evidence exists for this modification in any other TdcF structure (Burman *et al.*, 2007). Therefore, we believe that Lys58 was likely to have been incorrectly reported as carboxylated. As a consequence, it was eliminated from the KCX sites data set. Further details of the analysis of the three proteins can be found in Supplementary Note S4.

4. Discussion

We have developed a computational method that is capable of identifying lysine residues that undergo spontaneous carboxylation. The task, however, is not devoid of complications. Firstly, the microenvironment of non-carboxylated lysine residues buried in the protein often resembles that of the KCX site. Although local packing is the most prominent feature of KCX sites, alone it only provides modest predictive power. Secondly, correct classification is more challenging for Kcx residues that are not part of metal-ion centers, *i.e.* either involved in hydrogen-bond interactions with other residues of the active site or directly involved in the catalytic reaction, owing to a larger variability of the amino acids. Thirdly, the protein conformational changes that occur before and after carboxylation of the lysine residue (Stec, 2012) may have an impact on the local environment of the affected lysine residue. If the protein was crystallized in a non-carboxylated state, the microenvironment of a positively ionizable residue could be different from the carboxylated lysine residue and consequently more difficult to detect. Finally, the prediction results may also be influenced by the biased nature of the PDB. For example, the number of structures solved with magnesium and manganese as part of KCX metal-ion centers is relatively small. Such issues will be likely to become less problematic as the PDB expands. In spite of all of these difficulties, our computational method generally works well in distinguishing lysine residues that become carboxylated from those that do not.



(*a*) (*b*)

Figure 3
Remodeling of Lys84, predicted to be carboxylated, to Kcx84 in carbapenemase OXA-24 (PDB entry 2jc7). (*a*) shows Lys84 highlighted in stick representation as reported in the structural coordinates (Santillana *et al.*, 2007). The $2F_o - F_c$ electron-density map around Lys84 is shown as a gray mesh and is contoured at 1σ. The $F_o - F_c$ map is shown in green and is contoured at 3σ. (*b*) shows the $2F_o - F_c$ map around Lys84 after refinement with *REFMAC* (Murshudov *et al.*, 2011) and with the Lys84 modified to include the PTM. This figure was created with *PyMOL* (Schrödinger; http://www.pymol.org).

Lysine residues that become carboxylated are mostly buried. The ionizable groups found in the highly hydrophobic neighborhood of these buried lysine residues may promote their chemical modification. However, this modification is also sensitive to changes in the surrounding environment, which has been found to be an intrinsic component of the regulatory mechanism of this PTM as a functional switch (Che *et al.*, 2012; Buchman *et al.*, 2012; Borbulevych *et al.*, 2011; Kumar *et al.*, 2011; Vercheval *et al.*, 2010; Meulenbroek *et al.*, 2009). Carboxylation occurs when the carboxyl group can form at least one hydrogen bond or other forms of interactions (*e.g.* with metal ions). Carboxylated lysine residues on protein surfaces would be likely to be in an ordered state and detected by X-ray crystallography. However, we cannot rule out the possibility that disordered carboxylated lysines undetected by X-ray crystallography may exist.

Although there are a number of characteristic physical properties of residues at the KCX site, it would be desirable to derive different classes of structural motifs using techniques such as signature binding pockets (Dundas *et al.*, 2007) with evolutionary information incorporated (Tseng & Liang, 2006; Jimenez-Morales & Liang, 2011). Further research will be likely to be useful.

4.1. Estimation of the prevalence of lysine carboxylation

We assessed the extent of lysine carboxylation. In addition to the tests carried out using the training data set, we further evaluated the false-positive rate (FPR; see §2) in a set of high-resolution protein structures (<1.5 Å resolution). High-resolution protein structures allow us to make the reasonable assumption of a carboxylated lysine being less likely to be incorrectly reported, *i.e.* any prediction of Kcx was assumed to be incorrect in this set of proteins. Using the frequencies of kcx90rr, a total of 11 of 575 proteins were incorrectly predicted to have a Kcx residue, resulting in an FPR of about 1.9% (Table 1*b*). The same test was carried out using kcx40rr, which resulted in 19 misclassifications of 575 proteins, an FPR of about 3.3%. Based on the calculated FPRs (1.9 and 3.3%),

we should expect about 271 and 471, respectively, of the 14 261 proteins to be incorrect predictions (Table 1*b*). *PreLysCar* predicted 543 and 659 proteins to have a KCX site, respectively, which implies that 272 and 188 predicted Kcxs are expected to be correct, respectively. Considering the most conservative scenario with an FPR value of 3.3% (kcx40rr), about 1.3% of the proteins with more than 200 amino acids in the PDB may potentially be subject to spontaneous lysine carboxylation, which implies a much broader prevalence of Kcx than is currently known.

4.2. Implications

Taking into consideration the importance of post-translational modifications and how they contribute to enrich the functional variability of proteins (Jensen, 2004; Walsh *et al.*, 2005), we believe that spontaneous PTMs, such as the carboxylation of lysine studied here, may have broad implications both in eukaryotes and prokaryotes. Spontaneous PTMs can switch enzymes on and off under the appropriate physicochemical conditions (Stec, 2012), which may provide yet another elegant and efficient biological mechanism of regulation. As shown by our study of lysine carboxylation, computational analysis may help to overcome experimental challenges in detecting spontaneous PTMs owing to a large number of experimental complications.

5. Conclusions

We have developed a method for the prediction of lysine carboxylation. This PTM is difficult to detect using common experimental techniques such as mass spectrometry and X-ray crystallography. We have characterized the KCX microenvironment and its defining features. The computational method *Predictor of Lysine Carboxylation* (*PreLysCar*) developed here can provide useful predictions with robust positive and negative predictive values (93.1 and 99.4%, respectively); namely, the odds of correctly predicting both KCX sites and LYS sites are high. It also has excellent performance in sensitivity and specificity. *PreLysCar* predicted Kcx residues in proteins within different ranges of sequence identity to homologs of known KCX proteins. It also unveiled carboxylation sites in proteins without similarities to any known cases. In addition, a few lysine sites incorrectly modeled as carboxylated in the Protein Data Bank have also been uncovered. Our results indicate that the scope of this spontaneous PTM, which does not require the action of specific enzymes, might be larger than expected. We conservatively estimate that about 1.3% of protein structures available in the PDB contain a carboxylated lysine residue, which implies a threefold increase over those currently known. Spontaneous post-translational modifications of functional residues under certain physicochemical conditions may provide an efficient mechanism of biological regulation (Che *et al.*, 2012; Buchman *et al.*, 2012; Borbulevych *et al.*, 2011; Kumar *et al.*, 2011; Vercheval *et al.*, 2010; Meulenbroek *et al.*, 2009).

We are very grateful to Drs Erin Adams, Andrew Binkowski, Pedro Brugarolas, Zhao Jieling, Andrzej Joachimiak, Linda Kenney, Jacinto Lopez-Sagaseta, Boguslaw Nocek, Karl Volz and an anonymous referee for comments, discussions and suggestions. We are also very grateful to Dr Lopez-Sagaseta for assistance in generating Fig. 3. This work was supported by National Institutes of Health grants GM-079804 and GM-086145, National Science Foundation grants DMS-0800257 and DBI-1062328 and a grant from the Chicago Biomedical Consortium (CBC) supported by the Searle Funds at The Chicago Community Trust. DJM is very grateful for the support from a Becas Talentia Excellence Grant (Andalusian Ministry of Innovation, Science and Enterprise, Junta de Andalucia, Spain).

References

Altschul, S. F., Gish, W., Miller, W., Myers, E. W. & Lipman, D. J. (1990). *J. Mol. Biol.* **215**, 403–410.
Auld, D. S. (2001). *Biometals*, **14**, 271–313.
Benning, M. M., Kuo, J. M., Raushel, F. M. & Holden, H. M. (1995). *Biochemistry*, **34**, 7973–7978.
Berman, H. M. *et al.* (2002). *Acta Cryst.* D**58**, 899–907.
Borbulevych, O., Kumarasiri, M., Wilson, B., Llarrull, L. I., Lee, M., Hesek, D., Shi, Q., Peng, J., Baker, B. M. & Mobashery, S. (2011). *J. Biol. Chem.* **286**, 31466–31472.
Buchman, J. S., Schneider, K. D., Lloyd, A. R., Pavlish, S. L. & Leonard, D. A. (2012). *Biochemistry*, **51**, 3143–3150.
Burman, J. D., Stevenson, C. E., Sawers, R. G. & Lawson, D. M. (2007). *BMC Struct. Biol.* **7**, 30.
Che, T., Bonomo, R. A., Shanmugam, S., Bethel, C. R., Pusztai-Carey, M., Buynak, J. D. & Carey, P. R. (2012). *J. Am. Chem. Soc.* **134**, 11206–11215.
Dementin, S., Bouhss, A., Auger, G., Parquet, C., Mengin-Lecreulx, D., Dideberg, O., van Heijenoort, J. & Blanot, D. (2001). *Eur. J. Biochem.* **268**, 5800–5807.
Dundas, J., Binkowski, T. A., DasGupta, B. & Liang, J. (2007). *BMC Bioinformatics*, **8**, 388.
Dundas, J., Ouyang, Z., Tseng, J., Binkowski, A., Turpaz, Y. & Liang, J. (2006). *Nucleic Acids Res.* **34**, W116–W118.
Emsley, P., Lohkamp, B., Scott, W. G. & Cowtan, K. (2010). *Acta Cryst.* D**66**, 486–501.
Garman, E. F. (2010). *Acta Cryst.* D**66**, 339–351.
Golemi, D., Maveyraud, L., Vakulenko, S., Samama, J.-P. & Mobashery, S. (2001). *Proc. Natl Acad. Sci. USA*, **98**, 14280–14285.
Griep, S. & Hobohm, U. (2010). *Nucleic Acids Res.* **38**, D318–D319.
Hibbert, F. & Emsley, J. (1991). *Adv. Phys. Org. Chem.* **26**, 255–379.
Jensen, O. N. (2004). *Curr. Opin. Chem. Biol.* **8**, 33–41.
Jimenez-Morales, D. & Liang, J. (2011). *PLoS One*, **6**, e26400.
Kleywegt, G. J., Harris, M. R., Zou, J., Taylor, T. C., Wählby, A. & Jones, T. A. (2004). *Acta Cryst.* D**60**, 2240–2249.
Kumar, V., Saxena, N., Sarma, M. & Radha Kishan, K. V. (2011). *Protein Pept. Lett.* **18**, 663–669.
Li, W. & Godzik, A. (2006). *Bioinformatics*, **22**, 1658–1659.
Li, Y., Yu, X., Ho, J., Fushman, D., Allewell, N. M., Tuchman, M. & Shi, D. (2010). *Biochemistry*, **49**, 6887–6895.
Lorimer, G. H., Badger, M. R. & Andrews, T. J. (1976). *Biochemistry*, **15**, 529–536.
Meulenbroek, E. M., Paspaleva, K., Thomassen, E. A., Abrahams J. P., Goosen, N. & Pannu, N. S. (2009). *Protein Sci.* **18**, 549–558.
Murshudov, G. N., Skubák, P., Lebedev, A. A., Pannu, N. S., Steiner, R. A., Nicholls, R. A., Winn, M. D., Long, F. & Vagin, A. A. (2011). *Acta Cryst.* D**67**, 355–367.
Okano, Y., Mizohata, E., Xie, Y., Matsumura, H., Sugawara, H., Inoue, T., Yokota, A. & Kai, Y. (2002). *FEBS Lett.* **527**, 33–36.

Park, I. S. & Hausinger, R. P. (1995). *Science*, **267**, 1156–1158.
Poirel, L., Naas, T. & Nordmann, P. (2010). *Antimicrob. Agents Chemother.* **54**, 24–38.
Radha Kishan, K. V., Vohra, R. M., Ganesan, K., Agrawal, V., Sharma, V. M. & Sharma, R. (2005). *J. Mol. Biol.* **347**, 95–105.
Ravelli, R. B. G. & McSweeney, S. M. (2000). *Structure*, **8**, 315–328.
Santillana, E., Beceiro, A., Bou, G. & Romero, A. (2007). *Proc. Natl Acad. Sci. USA*, **104**, 5354–5359.
Stec, B. (2012). *Proc. Natl Acad. Sci. USA*, **109**, 18785–18790.
Tseng, Y. Y. & Liang, J. (2006). *Mol. Biol. Evol.* **23**, 421–436.
Vercheval, L., Bauvois, C., di Paolo, A., Borel, F., Ferrer, J.-L., Sauvage, E., Matagne, A., Frère, J.-M., Charlier, P., Galleni, M. & Kerff, F. (2010). *Biochem. J.* **432**, 495–504.
Walsh, C. T., Garneau-Tsodikova, S. & Gatto, G. J. (2005). *Angew. Chem. Int. Ed. Engl.* **44**, 7342–7372.
Wheeler, D. & Bhagwat, M. (2007). *Methods Mol. Biol.* **395**, 149–176.
Wu, D., Hu, T., Zhang, L., Chen, J., Du, J., Ding, J., Jiang, H. & Shen, X. (2008). *Protein Sci.* **17**, 1066–1076.
Wyrembak, P. N., Babaoglu, K., Pelto, R. B., Shoichet, B. K. & Pratt, R. F. (2007). *J. Am. Chem. Soc.* **129**, 9548–9549.
Xiang, D. F., Patskovsky, Y., Xu, C., Fedorov, A. A., Fedorov, E. V., Sisco, A. A., Sauder, J. M., Burley, S. K., Almo, S. C. & Raushel, F. M. (2010). *Biochemistry*, **49**, 6791–6803.
Zhang, H. (2004). *Proceedings of the Seventeenth International Florida Artificial Intelligence Research Society Conference*, edited by V. Barr & Z. Markov, pp. 562–567. Palo Alto: AAAI Press.